Implementación de un sistema Big Data que permita alertar en tiemporeal posibles irregularidades en la contratación del gasto público

Juan Pablo Arévalo Merchán y Laura Milena Ramos Bermúdez

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría en Informática
Bogotá D.C., 22 de abril de 2023

Implementación de un sistema Big Data que permita alertar en tiemporeal posibles irregularidades en la contratación del gasto público

Juan Pablo Arévalo Merchán y Laura Milena Ramos Bermúdez

Trabajo de investigación para optar al título de
Magíster en Informática

Director
Andrés Gómez Casanova

Jurados
Francisco Eliecer Sarmiento
Oswaldo Castillo Navetty

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría en Informática
Bogotá D.C., 22 de abril de 2023

Publicado en 2023 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Página de aceptación del jurado

El trabajo de grado de maestría titulado “Implementación de un sistema Big Data que permita alertar en tiempo real posibles irregularidades en la contratación del gasto público”, presentado por Laura Milena Ramos y Juan Pablo Arévalo Merchán, cumple con los requisitos establecidos para optar al título de Magíster en Informática.

Francisco Eliecer Sarmiento

Oswaldo Castillo Navetty

Director: Andrés Gómez Casanova

Bogotá, D.C., 06 de 07 de 2023

Dedicatoria

El presente proyecto lo dedicamos principalmente a Dios, por ser el inspirador y darnos fuerza para continuar en este proceso de obtener uno de los anhelos más deseados.

A nuestros padres, por su amor, trabajo y sacrificio en todos estos años, gracias a ustedes hemos logrado llegar hasta aquí y convertirnos en lo que somos.

Agradecimientos

Agradecemos a Dios por bendecirnos la vida, por guiarnos a lo largo de nuestra existencia, ser el apoyo y fortaleza en aquellos momentos de dificultad y de debilidad y a todas las personas que aportaron con sus conocimientos para la realización de este proyecto.

Resumen

En las últimas décadas, el desarrollo de las tecnologías se ha involucrado en todos los sectores y acciones de la vida humana, la inteligencia artificial, el metaverso, las super aplicaciones, la nube pública, las tecnologías Big Data, entre otras y todas estas son tan importante porque al aprovecharse puede combatir diferentes problemáticas que aquejan a la sociedad en general en cualquier ámbito, en temas relacionados a la salud, al sector financiero, al sector agropecuario, agroindustrial, ciudades inteligentes, a la gestión de recursos públicos y a una interminable lista de posibles casos donde se puede utilizar este tipo de tecnologías para resolver problemas cotidianos. Este proyecto se limita a una de las problemáticas que más impacto tiene a nivel mundial; la corrupción. Este mal que viene aquejando durante los últimos años y cada vez más, a la mayoría de los gobiernos que no tienen como controlar todos los recursos del estado, su contratación y compras, debido a la gran cantidad de contratos quese generan día a día.

Aquí es donde empieza a jugar un papel tan importante el Big Data, puesto que tiene el potencial de transformar las funciones gubernamentales y es ideal para solucionar problemáticas en diferentes casos de uso relacionados con la corrupción del sector público.

Un claro ejemplo sobre la problemática mencionada se presenta en Colombia, donde entidades como Transparencia Colombia y Monitor Ciudadano de la Corrupción en su informe (Radiografía de los hechos de corrupción en Colombia, 2019) del año 2019 indicó que el país perdió más de 17,9 billones de pesos. Además, se evidenció que el 73% del total de los hechos analizados respondieron a casos de corrupción administrativa; de este porcentaje más de la mitad compromete a la contratación pública, dividida en la adjudicación o celebración irregular de contratos, abuso de la figura de contratación directa, detrimento patrimonial por incumplimiento del objeto contratado y sobrecostos por irregularidades en la celebración de contratos son los que casos que más resaltan.

Es por eso, que se plantea el estudio de arquitecturas de referencia, herramientas y tendencias para procesar grandes cantidades de datos en tiempo real e históricos y el diseño e implementación de la arquitectura que sea capaz de procesar los datos disponibles, en el caso de uso de los contratos de gasto público en Colombia. De este modo, se desarrollaron procesamientos capaces de analizar la información histórica identificada e ingestada dentro de la arquitectura que permiten generar alertas tempranas de posibles irregularidades y sobrecostos en los contratos de gasto público y posterior a eso llegar a la visualización de toda la información en tableros de control.

La arquitectura planteada beneficiará a otras áreas de estudio, ya que servirá como punto de partida para que ahonden sus esfuerzos en la exploración y explotación de los datos y más no en el proceso de investigación y diseño de una arquitectura. Acá se genera una arquitectura lista para utilizar a partir de infraestructura como código y así se ahorrarán los esfuerzos de investigación. Esto servirá para futuras investigaciones sobre el uso de Big Data en la búsqueda de irregularidades en contratos de gasto público.

Abstract

In recent decades, the development of technologies has been involved in all sectors and actions of human life, artificial intelligence, the metaverse, super applications, the public cloud, Big Data technologies, among others, and all of these are so important because when taking advantage of it, it can combat different problems that afflict society in general in any field, in issues related to health, the financial sector, the agricultural and agro-industrial sector, smart cities, the management of public resources and an endless list of possible cases where this type of technology can be used to solve everyday problems. This project is limited to one of the problems that has the greatest impact worldwide, corruption. This evil has been afflicting in recent years and more and more, most governments that do not have a way to control all state resources, their hiring, and purchases, due to the large number of contracts that are generated every day.

This is where Big Data begins to play such an important role, since it has the potential to transform government functions and is ideal for solving problems in different use cases related to corruption in the public sector.

A clear example of the mentioned problem is presented in Colombia, where entities such as Transparency Colombia and the Citizen Monitor of Corruption in their report (Radiography of the facts of corruption in Colombia, 2019) of the year 2019 indicated that the country lost more than 17, 9 billion pesos. In addition, it was evidenced that 73% of the total events analyzed responded to cases of administrative corruption; Of this percentage, more than half commits to public procurement, divided into the awarding or irregular execution of contracts, abuse of the figure of direct contracting, patrimonial detriment due to non-compliance with the contracted object and cost overruns due to irregularities in the execution of contracts are those that cases that stand out the most.

That is why, the study of reference architectures, tools, and trends to process large amounts of data in real-time and historical and the design and implementation of the architecture that is capable of processing the available data, in the case of use, is proposed of public spending contracts in Colombia. In this way, processes were developed capable of analyzing the historical information identified and ingested within the architecture that allows generating early alerts of possible irregularities and cost overruns in public spending contracts and after that, reaching the visualization of all the information on dashboards of control.

The proposed architecture will benefit other areas of study since it will serve as a starting point for them to deepen their efforts in the exploration and exploitation of data and not in

the process of research and design of an architecture. Here, a ready-to-use architecture is generated from infrastructure as code; thus, research efforts will be saved. This will serve for future research on the use of Big Data in the search for irregularities in public spending contract.

Índice General

Índice de Figuras

Índice de Tablas

Anexos

Anexo 1. Modelo de datos orígenes

Excel que contiene todo el gobierno de datos realizado a las fuentes de datos identificadas. Se incluyen las transformaciones que se deben realizar para la limpieza de los datos y para la eliminaciónde las inconsistencias encontradas. (Anexo 1 - Modelo de datos orígenes.xlsx)

Anexo 2. Manuales y código fuente de la arquitectura

Se ha creado un repositorio público en GitHub donde se encuentra alojado todos los manuales, documentación y código fuente de la arquitectura, para que pueda ser accedido por cualquier interesado. Así mismo acá se podrá encontrar toda la construcción de los indicadores propuestos. (https://github.com/LauraMilenaRB/bigdata-corruption-indicators)

Anexo 3. Reporte de contratación pública con posibles irregularidades.

Se ha generado un reporte con la ejecución de todos los indicadores con los datos históricos de la contratación pública en Colombia, procesado en la plataforma construida y generados en los tableros de control construidos en QuickSight. (Reporte de contratos con alertas de posibles irregularidades.pdf)

1 Introducción

El desarrollo de las tecnologías se ha involucrado en todos los sectores y acciones de la vida humana, la inteligencia artificial, el metaverso, las super aplicaciones, la nube pública, las tecnologías Big Data, entre otras y todas estas son tan importante porque al aprovecharse puede contribuir a combatir diferentes problemas que aquejan a la sociedad en muchos ámbitos posibles.

Este proyecto se centra en Colombia y limita a una de las problemáticas que más impacto tiene a nivel mundial; la corrupción en el sector público, según transparencia internacional (Transparency International, 2021) y su reporte global donde se mide el Índice de Percepción de Corrupción en el Sector Público en una escala de 0 a 100, donde 0 indica el más alto nivel de corrupción y el número 100, indica el menor índice. Entre los países más afectados se encuentran: Sudán del Sur, Somalia, Siria, Yemen y Venezuela, en el caso de Colombia se ubica en la posición 92 de 180 países, por debajo del promedio mundial (Transparencia Colombia, 2021).

Para combatir esta problemática se ha propuesto utilizar el Big Data, puesto que tiene el potencial de transformar las funciones gubernamentales y es ideal para solucionar problemáticas en diferentes casos de uso relacionados con la corrupción del sector público, donde se maneja tanta información diferente.

Hay que tener en cuenta la complejidad de la corrupción y más aún en el sector público. Desafortunadamente esta no se puede generalizada, puesto que cada país tiene sus propias particularidades. En el caso de Colombia, se realizó una revisión de antecedentes los cualesindicaron que entre los años 2016 y el 2018 el país perdió más de 17,9 billones de pesos por causa de la corrupción y el 73% correspondía a contratación (Monitor Ciudadano de la corrupción, 2019).

Uno de los factores que contribuyen a estos altos índices es el encubrimiento de todos los intervinientes en estos hechos. Hasta el momento una de las principales formas en la cual sepuede descubrir la corrupción es gracias a la investigación de los medios de comunicación, puesto que a pesar de que los datos en temas de contratación con el estado son de carácter público (no todos, pero si una gran cantidad), no existe un control por parte de la ciudadanía, ni del estado, ni de los entes de control y por lo tanto este es uno de los puntos más álgidos en cuanto a la corrupción se refiere.

Como se menciona en los párrafos anteriores, los temas de contratación con el estado son de carácter público y en Colombia tenemos disponibles varias fuentes de datos como el SECOPI, SECOP II, PACO, Datos Abiertos Colombia, entre otros. Tal es el caso que durante el año 2021 se generaron más de 1'000.000 de contratos en las

diferentes modalidades de contratación (Acosta, 2021).

Tener esta cantidad de datos disponibles permite generar una base importante para comenzar a sacar provecho a las herramientas tecnológicas y de esta forma poder descubrir patrones que puedan alertar posibles irregularidades o anomalías en temas de contratación.

El Big Data ha venido tomando mucha fuerza durante los últimos años, aunque algunas definiciones datan del año 2001. En ese año se publicó el informe aMeta que hoy en día se conoce como Gartner (Laney, 2001), y es importante precisar que, aunque en dicho informe no menciona la palabra Big Data por ningún lado, esta definición ha nombrado los tres aspectos más importantes a tener en cuenta: el volumen, la velocidad y la variedad de los datos.

Ahora bien, ya teniendo una introducción del Big Data, empezamos a hablar del procesamiento de grandes volúmenes de datos, lo cual es fundamental para darle sentido a todos los datos y tomar decisiones lo más rápido posible.

Por ejemplo, en Colombia solamente en el año 2021 se gastaron más de 8 billones de pesos en los procesos adjudicados en infraestructura en las diversas modalidades según el informe anual de contratación, publicado por la Agencia Nacional de Infraestructura (ANI, 2022). Sitotalizamos no solo en infraestructura, sino en general, se gastan más de 100 billones de pesosal año en contratación (Monterrosa, 2018). Es por ello que una decisión en tiempo real en este tipo de situaciones ayudará a disminuir la probabilidad de perder dinero por contrataciónirregular.

Para construir una arquitectura que soporte el análisis de grandes cantidades de datos tanto para datos históricos como para alertas tempranas es fundamental conocer el estado del artede las arquitecturas de referencia que soportan estas necesidades. En la revisión bibliográficase descubrieron dos de las principales arquitecturas que abarcan estos temas, la arquitectura Lambda (Kreps, 2014) y la arquitectura Kappa (Marz, 2011), fuentes de inspiración para el diseño e implementación que se explicará más adelante.

Posterior a tener la arquitectura de referencia, es importante conocer aquellas herramientas o investigaciones donde se pueda combinar los dos términos fundamentales que comprometen este proyecto: la corrupción y Big Data; puesto que conocer lo que han invertido otros países, regiones o empresas en este ámbito podrá identificar casuísticas necesarias o servibles como punto de partida para el desarrollo de este proyecto.

En las revisiones bibliográficas encontramos proyectos como: (1) “Compras Públicas y Big Data: Investigación en Chile sobre índice de riesgo de corrupción” (Jorquera M. , 2019), allíse planea una idea similar utilizando Big Data para determinar potenciales riesgos de corrupción de las compras públicas, aplicado a la contratación de Chile, un sistema diferente al Colombiano. (2) “Hacia la transparencia 4.0, el uso de la inteligencia artificial y Big Data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales” (Hueso, 2020) se ha realizado una investigación sobre algunas herramientas con inteligencia artificial y tecnologías contra el fraude y la corrupción. (3) “The Next Generation of Anti- Corruption Tools: Big Data, Open Data & Artificial Intelligence” (Pasquarelli & Stirling, 2019) es importante resaltar que dicho reporte muestra a Colombia como uno de los países con mayor cantidad de datos públicos disponibles, pero también con altos niveles de corrupción percibida, esto hace que sea muy importante contar con herramientas como la propuesta en este proyecto de grado. (4) ProZorro, una plataforma donde se encuentran todos los contratos públicos de Ucrania, basados en una plataforma Big Data con un componente de inteligencia artificial (Antoniuk, Kuzyk, Zhurakovska, Sydorenko, & Sakhno, 2020). Estas herramientas no buscan definir una arquitectura pública, ni le dan la importancia a ella.

Colombia tiene mucho potencial por la cantidad de datos abiertos que maneja y por los altos niveles de corrupción que existe. Se ha implementado una herramienta que permite hacer seguimiento a la contratación pública, donde se puede generar alertas tempranas de los indicadores de posibles casos de corrupción, así como el análisis de los datos históricos para tener cifras sobre aquellos contratos que han podido tener algún tipo de irregularidad.

2 Objetivos

2.1 Objetivo general

Implementar un sistema Big Data mediante la revisión de tendencias que soporte el procesamiento en tiempo real y que permita clasificar los contratos de gasto público para identificar posibles irregularidades.

2.2 Objetivos específicos

2.2.1 Definir una arquitectura Big Data y seleccionar los componentes necesarios para soportar el procesamiento de datos en tiempo real, a partir de la comparación dearquitecturas de referencia.

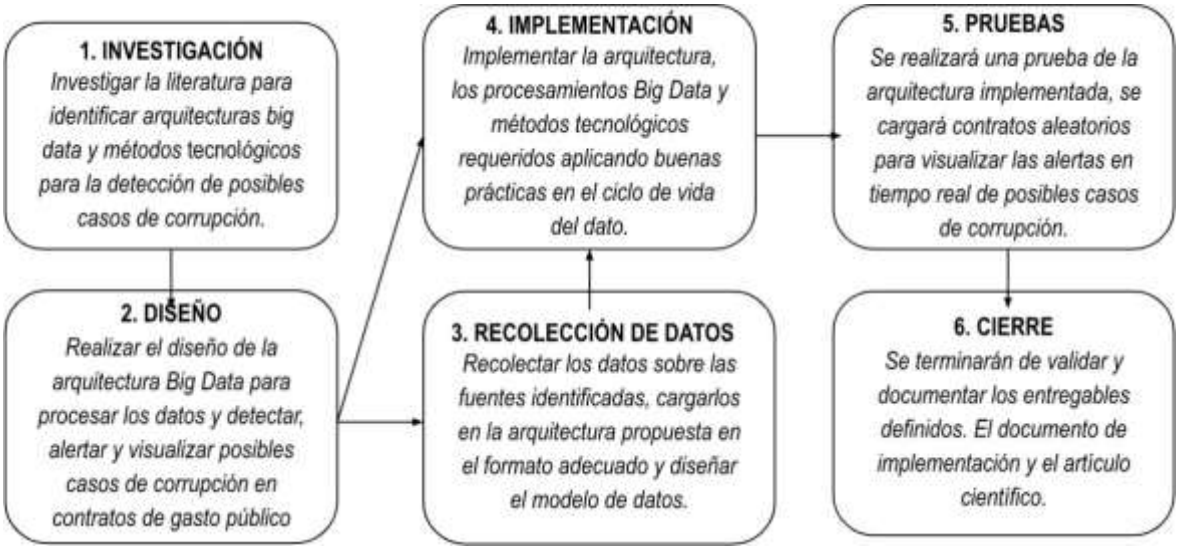
2.2.2 Identificar y desarrollar los indicadores y alertamientos de fraude en los contratos de gasto público para el contexto colombiano basados en los datos disponibles.

2.2.3 Establecer los tableros de control necesarios para la visualización de las alertas, indicadores y auditoría de los contratos de gasto público almacenados.

3 Metodología

El desarrollo del proyecto se dividirá en diferentes fases para cumplir con los objetivos específicos del proyecto, estas fases son investigación, diseño, recolección de datos, implementación, pruebas y cierre. El alcance del proyecto está previsto para desarrollarse por dos integrantes, ya que el proyecto no implica solo el diseño e implementación de una arquitectura, sino también el desarrollo de procesamientos para construir los indicadores de fraude en los contratos de gasto público que implican más actividades.

Ilustración 1. Metodología utilizada en el proyecto



Se ha realizado una investigación en sitios especializados, donde se resaltan IEEE, Google Scholar, entre otros, sobre proyectos similares al planteado en el caso de uso, tanto aplicado a Colombia como a todo el mundo, para conocer algunas similitudes y diferencias que permitieron el desarrollo de este proyecto.

Así mismo, se realizó la búsqueda de información periodística que permitió validar el tema asociado a los indicadores y posibles casos de corrupción en Colombia, inspiración para la definición de los indicadores o banderas rojas para el proyecto.

Al tener toda esta parte de investigación se llevó a cabo la consecución y recolección de toda la información de datos abiertos disponible, relacionada a la contratación

pública en Colombia, realizando ejercicios exploratorios con los datos disponibles para encontrar posibles irregularidades que permitieron generar las alertas correspondientes al caso de uso. Se hizo todo el gobierno de datos de toda la información recolectada de las diferentes fuentes de información para lograr la unificación de los datos y poder realizar un entendimiento de los datos. En otro hilo paralelo se avanzó en todo el proceso de construcción de la arquitectura en la nube, utilizando la infraestructura como código para que ésta pueda ser replicada por cualquier persona, este proceso se generó gracias a toda la investigación realizada en arquitecturas de referencias para el procesamiento de datos en tiempo real y por lotes, realizando una comparación entre servicios de la nube para encontrar la mejor alternativa a las necesidades. Seguido de ello, se realizó la construcción de todos los indicadores dentro de los servicios de la nube, permitiendo ingresar toda la información dentro de la arquitectura para generar los resultados correspondientes, a partir de esto se realizaron los reportes que permitieron consolidar los resultados de los indicadores propuestos tanto en tiempo real como los históricos.

4 Contexto y estado del arte

4.1 Contratación en Colombia

Es importante entender las características que tiene la contratación en Colombia, puesto que esto difiere de otros países y regiones. Es evidente que en nuestro país los contratos públicos son un tema que se ve afectado constantemente por la corrupción, tema que hablaremos en profundidad más adelante. Se definirán algunos conceptos claves sobre la contratación y cómo funciona el paso a paso en Colombia Compra Eficiente, el Sistema de compras públicas de Colombia.

4.1.1 Contrato

El primer concepto que definiremos es el contrato, *"es un acto por el cual una parte se obliga para con otra a dar, hacer o no hacer alguna cosa. Cada parte puede ser de una o de muchas personas"* (Código Civil Colombiano art. 1495, 1887), en otras palabras, es un acuerdo voluntario entre diferentes partes (personas naturales o jurídicas) donde se comprometen con una cosa u otra, siempre cumpliendo con obligatoriedad las condiciones pactadas. Con este acuerdo pactado y firmado, toma forma un documento para dar un respaldo y seguridad jurídica a las partes involucradas.

En el contexto de Colombia existen dos tipos de contratación, la pública y privada, sin embargo, para la síntesis del desarrollo del proyecto solo se abordará a detalle la contratación pública para la adquisición de los bienes y servicios requeridos por el estado.

4.1.2 Contratación pública

Cuando hablamos de contratación pública nos referimos a la compra de bienes o servicios por parte del gobierno a las empresas privadas o personas naturales que ofrecen los bienes y servicios requeridos por el estado. En este tipo de contratación encontramos el marco legal en Colombia, el cual se encuentra amparado bajo las leyes 80 de 1993 y 1150 de 2007, y el decreto 1510 de 2013, indicando que la contratación estatal se desarrollará con las reglas de interpretación de la contratación, los principios de transparencia, economía, responsabilidad y de conformidad con los postulados que rigen la función administrativa, todo esto para garantizar la alta calidad de la prestación del servicio y salvaguardar los interés públicos de los contribuyentes.

En la contratación pública encontramos cinco modalidades de selección previstas en la Ley 1150 de 2007, en donde las entidades estatales deben escoger a sus contratistas: (1) Licitación pública; (2) Selección abreviada; (3) Concurso de méritos; (4) Contratación directa; (5) Mínima cuantía.

4.1.3 Datos de la contratación pública en Colombia

Colombia es uno de los países que más datos abiertos genera, esta información está dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. Esto se debe gracias a la Ley 1712 de 2014, la cual define dichos datos como "*todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos*" (Congreso de la República de Colombia, 2014), gracias a esto la contratación pública ha generado grandes sistemas que permiten almacenar y tratar los datos relacionados a la contratación en SECOP y SECOP II, en Colombia existe una agencia nacional encargada de velar por todo el tema de la contratación pública llamada La Agencia Nacional de Contratación Pública - Colombia Compra Eficiente (ANCP - CCE).

4.1.3.1 La agencia Nacional de Contratación pública

En la labor de sus funciones el gobierno colombiano reconoce que la compra y contratación pública es un asunto estratégico de la nación, por ello con el decreto presidencial 4170 de noviembre 03 de 2011 se crea la Agencia Nacional de Contratación Pública – Colombia Compra Eficiente, como una entidad descentralizada de la Rama Ejecutiva del orden nacional, con personería jurídica, patrimonio propio y autonomía administrativa y financiera, adscrita al Departamento Nacional de Planeación. Actualmente es el rector del Sistema de Compra Pública en

el país y su objetivo consiste en desarrollar e impulsar políticas públicas claras, unificadas, buenas prácticas, herramientas para los procesos de compra y contratación estatal, con el fin de generar una mayor eficiencia, transparencia y optimización de los recursos del estado (Presidencia de la República de Colombia art. 2, 2011).

En la plataforma online de Colombia Compra eficiente, donde encontramos el sistema electrónico para la contratación pública *(www.colombiacompra.gov.co)*, en el apartado de 'Tienda Virtual' encontramos la Tienda virtual del estado colombiano, SECOP I, SECOP II, compra pública para la innovación, relatoría, manuales, guías, documentos tipo y mesa de servicio.

4.1.3.2 SECOP I

El SECOP I surge como respuesta a la adopción de medidas que garanticen los principios de eficiencia y transparencia en la contratación pública con miras a "*alcanzar dos objetivos: pulcritud en la selección de contratistas y condiciones de contratación más favorables parael Estado".*

Es el Sistema Electrónico para la Contratación Pública en la cual las Entidades Estatales deben publicar los Documentos del Proceso, desde la planeación del contrato hasta su liquidación. En la plataforma se puede consultar la información sobre los procesos de contratación estatal, permitiendo a las entidades estatales cumplir con las obligaciones de publicidad de los diferentes actos expedidos en los procesos contractuales. Además, permite consultar el estado de los procesos de contratación, a los interesados, proponentes, veedurías y a la ciudadanía en general. Esto permite a las Entidades Estatales y al sector privado tener una comunicación abierta sobre los procesos de contratación. A diferencia del SECOP II, esta no es una plataforma transaccional, por el contrario, se puede ver más como una herramienta para conocer los contratos que el estado está ofreciendo, como si se estuviera hablando de publicidad.

Ilustración 2. Esquema de integración del SECOP

Tomado de (República de Colombia, 2008)

4.1.3.3 SECOP II

Es la evolución a la versión anterior de la plataforma SECOP I, que le permite a entidades estatales y proveedores hacer todo el proceso de contratación en línea. SECOP II funciona como una plataforma transaccional donde las entidades crean, evalúan y adjudican procesos de contratación. A su vez los proveedores pueden hacer seguimiento en tiempo real al avance de los procesos de su interés, presentar ofertas y observaciones en línea.

Teniendo en cuenta a los ciudadanos como organismos de control, a los medios de comunicación o veedores interesados la plataforma les permite hacer búsquedas de procesos de contratación y planes anuales de adquisiciones de forma abierta, lo que permite mejorar la auditoría al sistema de compras, así mismo se tienen datos de sanciones disciplinarias, sanciones penales, responsabilidades fiscales, multas y sanciones contractuales y las colusiones en contratación pública. En la *Guía de búsqueda pública en el SECOP II* se puede encontrar el manual para usar la herramienta y cómo buscar procesos de contratación.

Dado que la información de SECOP I y II tiene como fin generar un control de la contratación pública en Colombia por medio de los datos abiertos, estos serán el insumo principal de nuestro proyecto, en especial la información histórica y actual de los procesos que terminen en la adjudicación un contrato de licitación pública (con los respectivos detalles). Para ello las plataformas permiten descargar la información de los procesos de compra pública que se han registrado en las plataformas mencionadas.

Cabe destacar que no es obligación de todas las entidades del estado registrar el proceso de contratación en estos portales, así que no estará disponible todo el histórico de toda la información, pero si existe una gran cantidad de datos que permitirá evaluar los posibles casos de contratación pública que sirvan de insumo para encontrar posibles irregularidades.

4.1.3.4 Portal Anticorrupción de Colombia – PACO

En *Portal Anticorrupción de Colombia – PACO | Bases de datos* encontramos los datos históricos de años anteriores con una frecuencia de actualización anual y en *Conjuntos de Datos abiertos* podemos encontrar los enlaces publicados en datos.gov.co donde se actualizan diferentes fuentes de información y que tomaremos como insumos para el proyecto. PACO tiene como objetivo a mediano plazo convertirse en el portal anticorrupción del Estado Colombiano y en la herramienta

central que las autoridades y líderes de la política pública de transparencia y lucha contra la corrupción usarán para monitorear este fenómeno (Portal Anticorrupción de Colombia – PACO | Bases de datos, s.f.).

4.2 Indicadores de corrupción

El caso de uso aplicado a este proyecto tiene relación con indicadores de la contratación pública en Colombia que permita detectar algunas posibles irregularidades en cuanto a los procesos de contratación pública se refiere, en las secciones siguientes, se podrá encontrar un poco acerca del contexto y estado del arte sobre indicadores de corrupción en Colombia y en la región, para un buen entendimiento, y así mismo para que sea un poco más sencillo la definición posterior de algunos indicadores de irregularidades que permita identificar posibles casos de corrupción en los contratos hallados en las bases de información de Colombia.

4.2.1 La corrupción

La corrupción es una de las principales problemáticas que más se presenta a nivel mundial y una de las más complejas de controlar puesto que su naturaleza es llevada de manera secreta y se busca evitar al máximo que salga a la luz pública, aquellos intervinientes muy pocas veces hacen público los acuerdos y la forma de llegar a ellos.

Cada país tiene sus propias particularidades que no permiten generalizar las causas, ni las formas en que se pueden intentar combatir, así mismo los niveles de corrupción son extremadamente complejos de medir, debido a que no existe un único indicador y ampliamente aceptado que permita evaluar la dimensión del fenómeno de la corrupción.

Aunque si bien es muy difícil llegar a medir los valores reales que se pierden en corrupción, existen algunas aproximaciones que permiten empezar a dimensionar de alguna forma la cantidad de dinero que se roban, algunos informes permiten recolectar información de casos que se han presentado y se han descubierto porque salen a la luz pública gracias, en su mayoría, al periodismo informativo que se ha tratado de enfocar en algunos contratos para buscar posible irregulares o carteles dedicados a robar mayormente el dinero público.

4.2.2 Medición de la corrupción

La principal forma de medir la corrupción se basa en algunas encuestas que intentan presentar cifras globalizadas para generar escalas de países con más corrupción o visto desde un punto de vista opuesto con menor transparencia, por nombrar algunos de

estos índices tenemos el International Country Risk Guide (ICRG), el cual se basa en encuestas de opinión de expertosy periodistas, el índice de Business International (BI) basando su información al igual que ICRG en una encuesta de opinión a nivel mundial, incluyendo algunos factores de riesgo comercial y político. El Global Competitiveness Report Index, basándose también en encuestas, pero involucrando más al sector privado pues se genera una clasificación que involucra a mandos medios y algunos directores de empresas alrededor del mundo, buscandola manera de reportar sobornos en negocios internacionales. Y por último y no menos importante y más utilizado el índice de percepción de la corrupción (IPC) a manos de Transparencia Internacional, una organización no gubernamental que promueve medidas contra crímenes corporativos y corrupción política en el ámbito internacional, este índice se basa en encuestas de opinión realizadas por diferentes organizaciones privadas, en entrevistascon analistas y expertos en el tema.

En la actualidad se viene cambiando un poco el enfoque, se están reemplazando las encuestas de ciudadanos y expertos en la materia, por la detección de algunos indicadores de riesgos que se han denominado banderas rojas, las cuales se establecen a partir del estudio exhaustivo de casos particulares de corrupción, reflejando los elementos más comunes detectados en estos casos, para generalizar estas casuísticas encontradas y generar indicadores que se puedan utilizar para detectar posibles irregularidades. Con el pasar de los tiempos, estas banderas rojas han empezado a ser un caso de investigación importante con el objetivo de sistematizar dichos patrones.

Algunos estudios importantes a la hora de conocer un marco referencial sobre este tema, inician en el año 2006, cuando Transparencia Internacional genera algunos indicadores dependiendo de cada fase de la contratación, puesto que de estas fases se pueden evaluar diversos riesgos que pueden desencadenar en banderas rojas (Transparencia Internacional,2006), al año siguiente, el Banco Mundial también genera otro importante proyecto donde seexplican los esquemas de corrupción que más se presentan: Kickback Brokers donde se paganlsobornos a funcionarios que ayudaron en la adjudicación de un contrato, el esquema denominado Bid Rigging donde se manipula la licitación para que sea adjudicada a un oferente específico y Front Companies la cual se basa en compañías fachadas para manipularla licitación y ejercer una influencia coercitiva sobre otros postores reales, a partir de ellos se generan algunas banderas rojas o indicadores que sirven para determinar posibles casos de corrupción (Campos & Pradhan, 2007).

Algunos otros artículos importantes para nombrar y que tratan esta misma problemática atacada a partir de banderas rojas son (Heggstad & Frøystad, 2011), (Coopers, 2013), (BancoMundial, 2014) y (Transparencia Internacional, 2014).

Todas estas alternativas imposibilitan crear patrones a nivel mundial para determinar posiblesíndices de corrupción, puesto que se deben realizar investigaciones y análisis de muchos documentos para detectar cada una de las posibles banderas rojas a partir de los casos detectados y centrados en cada país, pues la información disponible no está estandarizada a nivel mundial, lo que no permite globalizar dichas cifras a todo el mundo.

Estas banderas rojas buscan pasar de un criterio cualitativo de la corrupción a un criterio cuantitativo, precursores en el tema como Mihaly Fazekas (Fazekas, János, & King, 2013) busca que se evalúen la estructura y la magnitud de la problemática de corrupción para no generar juicios por culpa de la percepción de personas. Y este no ha sido el único estudio donde este autor busca demostrar la importancia de estos indicadores cuantitativos, en el trabajo de Fazekas y David-Barrett (David-Barrett & Fazekas, 2015) se analiza la aplicabilidad de la metodología del estudio anterior (2013) al conjunto de adquisiciones delReino Unido, a raíz de este trabajo se llegan a algunas conclusiones a nivel de política públicapara la mejora de dichos indicadores, partiendo del punto de invertir en la infraestructura de datos, desarrollar canales institucionalizados a través de los cuales se puedan informar estosanálisis y hallazgos clave de las adquisiciones públicas.

4.2.3 La corrupción en la contratación pública

La contratación pública es uno de los principales sectores donde se puede esconder la corrupción por las características que poseen, por un lado, el volumen de las transacciones y por el otro la magnitud de los montos que se manejan y más aún, gracias a estos dos ítems escasi imposible poder hacer seguimiento a cada uno de los contratos que se genera y a su vezse imposibilita el seguimiento de todos los recursos de la contratación pública. Es por ello que este proyecto se busca centrar en este ámbito, pues se posee una gran cantidad de información que permitirá monitorizar los contratos históricos y a su vez encontrar posibles banderas rojas que permita detectar posibles irregularidades.

4.2.4 Cifras de la corrupción en contratación pública en Colombia

Para tener algunas cifras claras y poder dimensionar el grave problema que afecta este país, en Colombia se gastan más de 150 billones de pesos al año en contratación estatal, lo que equivale a un aproximado del 15% del Producto Interno Bruto (PIB) de Colombia para el año2021 (Acosta, 2021). Entrando un poco más en profundidad y teniendo en cuenta sólo aquellos contratos que se han podido detectar con algún tipo de caso de corrupción, entre losaños 2016 y 2018 el país perdió más de 17,9 billones

de pesos en corrupción (Monitor Ciudadano de la corrupción, 2019), (ANI, 2021) cifras aproximadas, porque como se nombró anteriormente, solo se pueden dimensionar las cifras conocidas puesto que la corrupción y las irregularidades buscan ocultar la verdad; así mismo se evidenció que el 73% del total de los hechos analizados respondieron a casos de corrupción administrativa, de este monto más de la mitad compromete a la contratación pública, dividida en la adjudicación o celebración irregular de contratos, abuso de la figura de contratación directa, detrimento patrimonial por incumplimiento del objeto contratado y sobrecostos por irregularidades en la celebración de contratos son los que más resaltan.

4.2.5 Proyectos relacionados

Si bien existen algunos reportes e índices que dicen medir la corrupción en el mundo, es un tema imposible de controlar y conocer la realidad certera, pues si buscamos su definición, la corrupción busca ser oculta, anónima o secreta y en su mayoría se logra este objetivo, no se tienen cifras reales de cada país, estos índices generalmente se basan en encuestas a personas expertas en el tema o a los ciudadanos, por medio de entrevistas o algunas investigaciones que buscan estandarizar a nivel mundial la corrupción, cuando esto es un tema tan heterogéneo que es imposible de medir en realidad y a su vez genera un resultado que sufre de subjetividad y sesgo de cada participante.

Si bien es importante definir que esta investigación se centra en Colombia, es fundamental conocer algunos estudios, artículos, proyectos, herramientas y demás alternativas que permitan ser base para la detección de posibles patrones y formas de medir la corrupción, es por ello que se han buscado proyectos similares encontrando lo siguiente.

4.2.5.1 Compras Públicas y Big Data: Investigación en Chile sobre índice de riesgo de corrupción

Uno de los primeros proyecto que se debe nombrar, por presentar algunas similitudes en cuanto a los índices de corrupción y su medición se denomina *"Compras Públicas y Big Data: Investigación en Chile sobre índice de riesgo de corrupción"* (Jorquera M. , 2019), aunque no está situado en Colombia, es importante a nivel regional conocer el estado del arte y más aún cuando tiene partners importantes como el Banco Interamericano de Desarrollo, este proyecto si bien trata uno de los objetivos de este proyecto, no presenta la arquitectura que se utilizó para evaluar todos los datos, ni tampoco tiene como objetivo procesar datos en tiempo real para alertar la posible pérdida de dinero en el momento indicado. La idea principal de este proyecto es permitir determinar los potenciales riesgos de corrupción en las compras públicas,

posee un componente investigativo sobre la corrupción, basados en estudios internacionales de entidades como el Fondo Monetario Internacional, la OCDE, el Banco Mundial y Transparencia Internacional entre otros. Así mismo, el sistema de compraspúblicas de Chile, tiene grandes diferencias con el de Colombia, por lo tanto, son proyectos en contextos diferentes. La investigación de este proyecto ha partido de dos grandes pasos secuenciales fundamentales a la hora de identificar irregularidades en la contratación pública, el primero de ellos, la búsqueda de patrones de riesgo a partir de ejercicios exploratorios en los datos disponibles sobre compradores, proveedores y tipos de contratación y el segundo paso, la construcción de un índice de riesgo de corrupción basado en indicadores asociados a competencia, transparencia y anomalías o violaciones a la ley (Monitor Ciudadano de la corrupción, 2019). Este proyecto ha detectado los siguientes indicadores como alarmas que pueden llegar a detectar posibles irregularidades.

- El porcentaje de licitaciones públicas y privadas con un único oferente, es decir aquelloscontratos donde solo una persona o empresa esté interesada y cumpla las condiciones para licitar.
- El cambio en el promedio de participantes por procedimiento comparándolos con periodos anteriores para determinar posibles cambios abruptos que generen algún tipo dealarma.
- El número de empresas diferentes ganadoras entre todos los contratos, esto con el fin dedeterminar si la contratación está presentando posibles amaños para algunas empresas.
- El porcentaje de procedimientos que utilizaron adjudicación directa o licitación privada puesto que es la forma más sencilla en toda la contratación de generar corrupción porqueno existe grandes cantidades de proponentes.
- Porcentaje del monto por adjudicación directa o licitación privada con el fin de determinar que tanto está creciendo la adjudicación directa y si existen valores muy altosque estén generando algún tipo de alarma
- El cambio en la tendencia del porcentaje de adjudicaciones directas.
- El índice de Hirschman-Herfindahl, por número de contratos. Aclarando que dicho índice es una medida usada para determinar el nivel y los cambios de concentración en los mercados, a partir de modelos de competencia en cantidad con productos homogéneos.
- El índice de Hirschman-Herfindahl por monto.
- El índice de dominancia por número de contratos
- El índice de dominancia por monto.
- El índice de participación.
- El número de participantes es distinto entre todos los procedimientos.
- El índice de concentración de las 4 empresas con más procedimientos ganados.

- El índice de concentración de las 4 empresas con más monto adjudicado.
- El índice de concentración de las 4 empresas con ratio de participación y adjudicación enlicitaciones públicas y privadas.
- El porcentaje de licitaciones públicas y privadas sin ningún documento obligatoriopublicado.
- El promedio del porcentaje de incumplimiento en la publicación de documentos obligatorios.
- El porcentaje de licitaciones públicas y privadas con varias adjudicaciones.
- El porcentaje de las licitaciones mayores o iguales a 100 UTM (Unidad tributaria mensual chilena que consiste en un monto de dinero expresado en pesos que es determinado por la ley y es actualizada constantemente por el índice de precios al consumidor IPC) y menores a 1.000 UTM cuyo plazo de oferta fue acelerado.
- El porcentaje de las licitaciones mayores o iguales a 1.000 UTM y menores a 5.000 UTMcuyo plazo de oferta fue acelerado.
- El porcentaje de licitaciones sin bases de licitación.
- El porcentaje de licitaciones sin respuesta a todas las preguntas del foro.
- El porcentaje de procedimientos sin acta de evaluación.
- El porcentaje de procedimientos sin resolución de adjudicación.
- El porcentaje de procedimientos sin contrato.
- El cambio en el porcentaje de contratos publicados por procedimiento.
- El porcentaje de licitaciones que no cumplen con la normativa de publicación.
- El porcentaje de las licitaciones menores a 100 UTM cuyo plazo de oferta fue menor alplazo mínimo establecido por la ley de compras públicas.
- El porcentaje de las licitaciones mayores o iguales a 100 UTM y menores a 1.000 UTM cuyo plazo de oferta fue menor al plazo mínimo establecido por la ley de compras públicas.
- El porcentaje de las licitaciones mayores o iguales a 1.000 UTM y menores a 5.000 UTM cuyo plazo de oferta fue menor al plazo mínimo establecido por la ley de compras públicas
- El porcentaje de las licitaciones mayores o iguales a 5.000 UTM cuyo plazo de oferta fuemenor al plazo mínimo establecido por la ley de compras públicas.

Estos indicadores establecidos permitieron una identificación de posibles irregularidades adaptadas al mercado chileno y servirá como punto de partida para futuras investigaciones y adaptándose a la normativa vigente de cada país y en especial a los datos públicos que cada gobierno publique podrán ayudar a detectar algunos factores que permitan generar ciertas alertas.

4.2.5.2 Hacia la transparencia 4.0, el uso de la inteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales

El siguiente proyecto para referenciar se denomina *"Hacia la transparencia 4.0, el uso de lainteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales"* (Hueso, 2020) desde España y la universidad de Valencia se harealizado una investigación sobre algunas herramientas y tecnologías contra el fraude y la corrupción. Los sistemas automatizado con inteligencia artificial para la planificación de las entidades de control ayudan a localizar los focos de mayor riesgo de posible fraude aprendiendo de los grandes conjuntos de datos para identificar patrones que determinen algúntipo de sospecha de fraude, entre estos sistemas se nombran: el *"sistema holandés de tratamiento automatizado, profundo y predictivo de datos ilimitados SyRI"* es una de las herramientas para detectar diversas formas de fraude, incluidos beneficios sociales, subsidios y fraude fiscal. Esto funciona generando informes de riesgo, donde se determina si una persona es "digna" de ser investigada por posible fraude.

Por otro lado, nombra al sistema SALER valenciano, este es un sistema informático que analiza los datos generados por la Administración de la Generalitat Valenciana con el fin de obtener alertas tempranas. Esto ayudará a detectar cualquier posible irregularidad, negligencia o riesgo de fraude y corrupción de forma preventiva, lo que evita asuntos de corrupción. En este proyecto podemos encontrar algunas de las banderas rojas que maneja.

- Fraccionamiento con un mismo proveedor: el mismo proveedor aparece en varios contratos para el mismo proyecto, lo que implica el riesgo de que el licitador evite uncontrato más grande con procedimientos más estrictos.
- Fraccionamiento con distintos proveedores: varios proveedores para un único contrato implican el riesgo de que los licitadores eviten un único contrato principal con un proveedor o la cooperación ilegal entre los contratistas.
- Adjudicatario recurrente: las acciones comerciales repetidas en un régimen de monopolio suponen riesgo de que no haya competencia.
- Colusión en un procedimiento de licitación: las coincidencias de los mismos (dos o más) licitadores en diferentes convocatorias suponen riesgo de colusión entre los licitadores.
- Conflicto de intereses: el incumplimiento de las normas sobre incompatibilidad establecidas por ley supone riesgo de conflictos de intereses.
- Procedimiento no competitivo: la ausencia de justificación para el uso de un procedimiento no competitivo supone riesgo de corrupción.
- Prueba irregular de contrato: las ampliaciones y los cambios no justificados suponenriesgo de fraude.

4.2.5.3 Otros proyectos o reportes

Es importante nombrar el siguiente reporte, publicado en el sitio web de la Comisión Europea, que ha sido llevado a cabo por Oxford Insights y denominado *"The Next Generationof Anti-Corruption Tools: Big Data, Open Data & Artificial Intelligence"* (Pasquarelli & Stirling, 2019), acá se puede evidenciar que dicho reporte muestra a Colombia como uno delos países con mayor cantidad de datos públicos disponibles, pero también con altos niveles de corrupción percibida, esto hace que sea muy importante contar con herramientas como la desarrollada en este proyecto de grado. Según el reporte, la Unión Europea tiene un proyecto para identificar posibles instancias de corrupción, denominada *Digital Whistleblower*, una colaboración de 6 universidades europeas, instituciones anticorrupción y financiada con fondos europeos. Este proyecto ha dado como resultado la plataforma *Opentender*, donde se generan y visualizan algunas métricas respecto a identificación de riesgos de corrupción. Utiliza el aprendizaje automático para poder generar los resultados.

Ucrania también tiene una herramienta similar, denominada *ProZorro*, una plataforma donde se encuentran todos los contratos públicos de dicho país, basados en una plataforma Big Data con un componente de inteligencia artificial (Antoniuk, Kuzyk, Zhurakovska, Sydorenko, &Sakhno, 2020)

Todos estos estudios demuestran la importancia de tener un stock de datos abiertos y públicos capaces de identificar posibles irregularidades, esto se puede traducir en reportes y herramientas de monitoreo de corrupción y Colombia tiene mucho potencial por la cantidadde datos abiertos que maneja y más aún por los altos niveles de corrupción que tiene.

4.2.5.4 Banderas rojas en Colombia por periodistas y estudios.

Algunos estudios, de periodistas colombianos y parte del gobierno, han hecho el intento de identificar algunas banderas rojas aplicados al contexto colombiano, teniendo en cuenta las particularidades que contienen los contratos públicos en el país en virtud de la información disponible en las plataformas actuales y dependiendo de la fase del contrato, el tipo de contrato y el riesgo que genera. Una de estas definiciones la realizó Datasketch, donde ha definido los siguientes indicadores con el fin de detectar algunas posibles irregularidades enla contratación pública:

- El primer indicador es la ausencia de actores externos a la entidad que desean realizarla contratación para evaluar la apertura del contrato.
- Si hay evidencia de acuerdos informales para la apertura del contrato.
- Si el contratista tiene más de un contrato con el estado al mismo tiempo (puede tener un contrato como persona natural, pero además contratos como representante legal de alguna empresa u organización).
- Si el contratista tiene antecedentes de realizar actos ilegales en contrataciones.

- Si el contrato en algún momento de la ejecución aumenta su valor súbitamente.
- Si hay compra de información de especificaciones de los pliegos o condiciones delcontrato.
- Si el contratista celebra dos o más contratos con la misma entidad estatal en dos añosconsecutivos.
- Si se realizan adiciones presupuestales al contrato.
- Si existe alguna licitación con un solo proponente, o un número muy bajo.
- Si para la licitación de un contrato la similitud entre los pliegos o condiciones del procedimiento y los productos o servicios del contratista ganador.
- Si hay oferentes del contrato que se han quejado del proceso de contratación.
- Si las condiciones del procedimiento son muy específicas e inusuales en comparacióncon procesos similares anteriormente celebrados.
- Si los pliegos de condiciones de la adjudicación no son coherentes o no tienen una relación con las actividades que se requieren para un contrato parecido.
- Si no se publican los estudios previos que se tuvieron en cuenta para adjudicar elcontrato.
- Si la oferta ganadora de la adjudicación del contrato incluye elementos que no estabanen los pliegos o condiciones del procedimiento.
- Cuando hay una modificación sorpresiva en los pliegos o condiciones de la adjudicación del contrato.
- Cuando el periodo de licitación entre la oferta y la adjudicación del contrato es muycorto.
- Si se aceptan ofertas por fuera del plazo máximo de entrega.
- Si los miembros del comité de evaluación no tienen el conocimiento técnico necesariopara evaluar las ofertas presentadas y están dominados por un único individuo.
- Cuando se cambia el nombre y la personalidad jurídica de la empresa para laadjudicación del contrato.
- Si los procesos de interventoría del contrato han sido deficientes.
- Cuando se suscriben dos o más contratos con objetos idénticos en un periodo cortode tiempo y sin razón aparente.
- Si no se justifica adecuadamente los motivos para no realizar procesos competitivosque impliquen la escogencia del mejor candidato.
- Si hay ausencia de publicidad para la adjudicación del contrato.
- Cuando los criterios de evaluación del contrato no son anunciados.
- Cuando la información sobre el estado del contrato y el proceso de adjudicación noes pública.
- Si se elige un proceso de contratación directa en vez de uno más competitivo.

- Si se hace pública la apertura del contrato luego de la adjudicación del mismo.
- Si hay adendas extemporáneas: Hay una modificación sorpresiva en los pliegos ocondiciones de la adjudicación del contrato.
- Qué exista una relación social estrecha o cercana entre las dos partes del contrato,entre el contratista y el contratante.
- Cuando no hay una declaración de intereses del contratista y contratante.
- Si un empleado del órgano de contratación ha trabajado para una empresa ofertantede un contrato.
- Si un empleado del órgano de contratación insiste en obtener información sobre elprocedimiento de licitación de un contrato.
- Cuando los documentos necesarios para la adjudicación del contrato estánincompletos o errados.
- Si hay facturación falsa o pagos anticipados sin cumplimiento de requisitos.
- Cuando las facturas estén con montos redondeados.
- Si existen ofertas idénticas o similitudes sospechosas en el proceso de adjudicacióndel contrato.
- Si el funcionario puede llegar a consignar la información en el portal de contrataciónpública de manera inadecuada o incompleta (Datasketch, s. f.).

Estas banderas rojas pueden generar algunos indicios de posibles irregularidades, ya adaptadas a todo el proceso de contratación local, es importante tener en cuenta que muchasde ellas son complejas de monitorizar y automatizar con los datos disponibles y como algunos contratos no están completos o se presentan errores humanos que se puedan catalogar comofalsos positivos.

4.2.5.5 Banderas rojas en Colombia por el estado.

Desde el estado colombiano también se han tratado de definir algunos indicadores y métricas, por medio de la Secretaría de Transparencia quién ha realizado un manual para la identificación de banderas rojas en el proceso de contratación estatal dependiendo de la fase en que se encuentra el contrato a evaluar (Secretaria de Transparencia, s. f.).

4.2.5.6 Etapa precontractual

Para la etapa precontractual se han definido los siguientes indicadores:

- Dependiendo de la entidad contratante y la distancia entre la fecha de apertura delproceso y la fecha de elección presidencial o regional.
- Si existen incidencia de actores externos.
- Si existe una entidad o sector priorizado por el Gobierno nacional o por un órgano decontrol.
- Si el valor del contrato es de alta cuantía, dependiendo del plazo de ejecución yteniendo en cuenta el tipo de contrato puede generar cierta alarma.

- Dependiendo del tipo de contrato si hace parte de los sectores sensibles.
- Si el cronograma del proceso contiene alguna de las siguientes casuísticas:
    - Si el tiempo para la presentación de ofertas es considerado insuficiente o se han presentado reiteradas solicitudes de parte de los proponentes solicitando una extensión del plazo
    - Si el tiempo entre el inicio del contrato y la fecha de adjudicación es superiora 15 días.
- Si hay sospecha de pliegos de condiciones estructurados a la medida.
- Cuando el plazo para la presentación de ofertas es demasiado corto.

4.2.5.7 Etapa Contractual

Para la etapa contractual los siguientes indicadores pueden mostrar alguna posible irregularidad:

- El lapso entre la fecha de adjudicación y la fecha de inicio del contrato son muy extensos y puede generar algún tipo de sospecha por atentar contra el principio de transparencia y selección objetiva.
- Una adición o modificación del contrato.
- Si el contratista es sancionado por indebida ejecución del contrato.
- Si el contratista ha sido sancionado.
- Si el contratista que se ha ganado la adjudicación del contrato o la comprase encuentra en la "lista negra" de Confecámaras.
- Cuando el porcentaje de ejecución del contrato difiere en gran medida del porcentajereal entregado.
- Si hay adiciones o modificaciones en cierto porcentaje ejecutado y entregado.

4.2.5.8 Etapa postcontractual

Y la última etapa después de la culminación del contrato (postcontractual) puede generaralgún tipo de alerta si se presentan algunos de los siguientes indicadores:

- Cuando no se haya cumplido con el plazo definido y sin razón alguna.
- En caso de que el porcentaje ejecutado del valor del contrato no corresponda con el porcentaje entregado del bien, obra o servicio, se eleva una alerta inmediata.

Todas estas banderas rojas buscan la manera de generar ciertas alarmas o alertas para la contratación pública, es por ello que este proyecto quiere generar sus propias banderas rojas actualizadas a los datos que hoy en día se encuentres públicos y disponibles y generar dichoanálisis bajo una plataforma big data que permita analizar toda la información histórica disponible para generar aún más valor.

4.2.6 Contratos en Colombia que han sido identificados con corrupción

Para la definición de las banderas rojas aplicadas al contexto de contratación pública en Colombia es importante tener como referencia algunos casos de contratos que hayan presentado de alguna forma hechos de corrupción o por lo menos están en investigación, porque esto puede durar muchos años. En la actualidad existen varios ejemplos que permitirán encontrar patrones o ciertas casuísticas que ayudarán a definir estos indicadores de alerta para la contratación en curso. A continuación, se consolidará la información relevante de dichos contratos que durante los últimos años han salido a la luz pública, y quede alguna forma han generado pérdidas de dinero al estado colombiano, ya sea por incumplimientos en las condiciones de la contratación estipuladas, sobrecostos, peculado, celebración indebida de contratos, falsedad en documento público y concierto para delinquirentre otros.

Gracias a la investigación realizada por Transparencia por Colombia, Monitor Ciudadano de la corrupción y la Fundación Charles Leopold Mayer, se han podido identificar los siguientescasos relevantes; iniciando con un caso particular, presentado en el año 2011, para este año se ha elegido alcalde de Bucaramanga a Luis Francisco Bohórquez para el periodo 2012-2015, sus familiares empezaron a crear varias empresas para obtener numerosos contratos con sobrecostos incluidos. Así que personas relacionadas a un gobernante se vieronbeneficiados gracias a la contratación pública.

José Rubiel Páez, alcalde de Caldas en el departamento de Boyacá, contrató la remodelación de la estación de policía del municipio, usando una modalidad que no aplicaba por el valor del contrato, y no teniendo suficiente con esto, firmaron las actas de entrega sin que éste se hubiera complido, convirtiéndose en un elefante blanco más de este país. Así mismo, historia similar ocurrió con el alcalde del municipio de Sáchica en el departamento de Boyacá, Héctor Antonio Amado, se realizó el pago del contrato sin que se ejecutará las obras correspondientes, por eso es importante poder validar el cumplimiento de los entregables o avances de los contratos para poder seguir realizando los pagos correspondientes (Caracol Radio, 2022).

Otro caso importante se dio en la gobernación de Norte de Santander, allí se adjudicaron 4 contratos por más de 50 mil millones para el Programa de Alimentación Escolar (PAE), lo que generó un abuso en la contratación directa puesto que el mismo contratista o proveedor se quedó con dichos contratos.
Ahora pasamos al otro extremo del país, para el año 2014 la alcaldía de Leticia, esto en el departamento del Amazonas, contrató a la Asociación Zonal del Consejo

Municipal de Autoridades Indígenas de Tradición Autóctono - Azcaita para cumplir algunos compromisos por un valor cercano a los cien millones de pesos, en investigaciones de la Fiscalía General de la Nación se pudo determinar que algunos de los entregables del proyecto no se realizaron, esto demuestra un incumplimiento a los compromisos pactados y se vuelve parte fundamental poder controlar los entregables y avances de los contratos públicos.
En el año 2016, se celebró un contrato entre el hospital San Jerónimo de Montería (Contratación estatal), por más de dos mil millones de pesos, para la compra de algunos equipos médicos, en investigación realizada por la Fiscalía General de la Nación se pudo determinar que dicho contrato presentaba un sobrecostos por más del 50%, es decir más de mil millones de pesos, este otro tipo de caso, demuestra la importancia de comparar los precios entre las diferentes ofertas que se reciben por compras realizadas por entidades del estado (Monitor Ciudadano de la corrupción, 2019).

Y estos son solo algunos de los casos de corrupción en la contratación pública en Colombia, si quisiéramos hablar de más casos, tendríamos cientos de páginas por escribir, identificando cada caso partículas, pero no es el objetivo ni el alcance del proyecto y ni hablar de aquellos casos que han sido catalogados como los mayores desfalcos al país, como el carrusel de la contratación, el caso Odebrecht, el escándalo de Reficar, el carrusel de las alcaldías locales de Bogotá, entre otros.

Esta sección muestra la importancia de identificar posibles banderas rojas a la hora de la construcción del proyecto, es importante conocer el pasado para identificar posibles irregularidades que se puedan presentar en el futuro y alertar en el momento indicado para no estar condenados a repetir la historia de nuevo.

4.3 Big Data

Actualmente los datos son uno de los objetos más valiosos para las organizaciones, debido aque, gracias a ellos pueden obtener información importante sobre su negocio, el mercado, los clientes, la competencia y en sí cualquier tipo de información que les permite tomar decisiones con la ayuda de otras tecnologías importantes como lo es la inteligencia de negocio (BI), además de la inteligencia artificial (IA), está busca automatizar tareas de cómo percibir su entorno y resolver problemas de cálculo, aprendizaje, razonamiento, memoria y comprensión cómo lo hace el comportamiento humano, para lograr predecir cómo se van a comportar ciertas cosas del entorno, para adelantarse y ser pioneros basados en la información, así mismo el procesamiento de datos es parte fundamental del big data, pues es la principal herramienta para dar forma a los datos, para que se genere información, valor o conocimiento a partir de

ellos, datos hay cientos de miles de millones, pero darle un uso adecuado a esos datos no es tarea sencilla.

El término de big data se ha citado un número indeterminado de veces durante las últimas décadas, se volvió uno de los conceptos actuales más importantes y con implicaciones económicas en las organizaciones. El concepto de big data empezó a tomar fuerza en el año 2001 y de esa fecha a lo que es hoy se destacan las definiciones dadas por los grandes vendors de tecnología como Oracle, IBM o por Gartner; una de las más importantes empresas consultoras y de investigación de las tecnologías de la información, se siguen basando en los mismos términos para describir lo que es este paradigma.

Entonces, para entrar en razón, el término big data se puede entender como: *"activos de información de gran volumen, alta velocidad y/o gran variedad que exigen formas innovadoras y rentables de procesamiento de la información que permiten una mejor comprensión, toma de decisiones y automatización de procesos."* (Gartner, s.f.). En esta definición se están empleando tres conceptos importantes de los datos, también conocidos como las '*uves'* del Big Data: volumen, velocidad y variedad, sin embargo, en otras fuentes y empresas como BBVA. (BBVA, 2017) u Oracle (Oracle, s.f.), utilizan dos '*uves'* adicionales: veracidad y valor. A continuación, definimos estas.

- **Volumen:** Para considerar el Big Data debemos entender que la cantidad o el volumen de datos si importa y bastante, estamos hablando de procesar decenas de terabytes o hasta cientos de petabytes de datos, tengamos en cuenta que durante los últimos años son cada vez más los dispositivos que producen información, IoT, transacciones bancarias, los vehículos, fábricas, redes sociales, páginas web, celulares, contratación pública, control de obras, en fin, hoy en día cualquier cosa que se haga está recolectando datos, por eso este término es fundamental en la definiciónde big data.

- **Velocidad:** Para el negocio o las organizaciones la velocidad en muchos casos puede ser de mayor importancia que otras *'uves',* ya que, tener los datos disponibles en el momento justo es indispensable para tomar decisiones adecuadas que pueden dar beneficios ante la competencia del mercado, atado un poco a la cantidad de dispositivos que están generando los datos, esto crece de una manera inimaginable, la cantidad de información que circula por toda la red es impresionante, la velocidad con que dicha información se genera segundo a segundo es incalculable y con el pasar de los años será cada vez mayor.

- **Variedad:** En general cuando las organizaciones tienen un volumen de datos tan grande, es proporcional a tener muchas fuentes diferentes de información en diferentes tipos, estamos hablando tanto de bases de datos relacionales, no relacionales, como información de redes sociales, flujos de clics de páginas web, aplicaciones móviles, sensores, documentos, imágenes, videos, aplicaciones externas como CRM, ERP, entre otras, todo esto genera datos de diferente tipo.

- **Veracidad:** Está *'uve'* se refiere a la calidad de los datos que serán analizados en la organización, se debe tener la certeza de que los datos obtenidos de las diferentes fuentes son veraces y se puede confiar en ellos. Aunque no todas las definiciones de Big Data utilizan la veracidad como un atributo fundamental, es importante reconocer que no todos los datos que circulan por la red son reales, así que para la toma de decisiones de una empresa se debe contar con datos veraces, que sean de orígenes confiables y no generen falsos positivos.

- **Valor:** El valor se refiere al valor intrínseco que poseen los datos, sin embargo, este valor solo se puede validar cuando lo explotamos de manera correcta en el contexto del negocio. Al igual que el término anterior, no hace parte de la definición estándar de Big Data, ha sido un término que se viene utilizando para dar contexto a que los datos deben tener algún valor para ser considerados utilicen, no podemos decir que todos los datos almacenados generan valor.

4.4 Arquitecturas Big Data

Teniendo en cuenta el contexto del proyecto y los conceptos definidos de Big Data, se realizó una revisión de las arquitecturas Big Data para propósitos de diferentes tipos, como criptomonedas, salud, inteligencia de negocio y de modo general. En el ejercicio de comparar y revisar estas arquitecturas como referencia identificamos que la mayoría se componen principalmente de tres capas: **capa de almacenamiento**, **capa de procesamiento** y **capa de visualización**, sin embargo, en algunas de estas referencias también se han propuesto otras dos capas adicionales: **capa de gobierno de datos**, la cual nos ayuda con la administración, estandarización y mejora continua en la explotación de la información del negocio y la **capa de seguridad**.

Por ejemplo, en el estudio del estado del arte de sistemas para el análisis de datos en tiempo real de Big Data se presentan dos grandes arquitecturas, las arquitecturas de referencia *Kappa* y *Lambda,* estas buscan definir un estándar sobre el procesamiento de

datos, el manejo de las capas principales y ventajas y desventajas, este proyecto tiene sus bases en estas arquitecturas, pues son las más conocidas y estudiadas y tienen grandes ventajas que se verán más adelante en la construcción de la arquitectura. Ahora continuando con el estado del arte, se encuentra un artículo de revisión que busca consolidar todas estas arquitecturas aplicadas a un contexto de criptomoneda, este contexto está asociado a la toma de decisiones en tiempo real. A raíz de la investigación realizada, se hace una propuesta donde se incorporan capas.

adicionales como la de explotación con inteligencia artificial y una capa de seguridad**,** ya que en las otras arquitecturas revisadas por el autor no menciona nada al respecto de estos temas *(Arévalo, J., 2020).* Aquí se habla de la capa de almacenamiento, batch, streaming, servicio, predictiva y por último una capa transversal de seguridad.

Hay que tener en cuenta que las arquitecturas propuestas no solo están enfocadas al contexto de las criptomonedas, esto es solo un caso de uso de los muchos donde se puede aplicar dichas arquitecturas de referencia, de hecho, hay variedad de arquitecturas aplicadas a diferentes contextos como salud, telecomunicaciones, análisis de inteligencia de negocio (BI), entre otras.

Por ejemplo, Arias S, O. E. (2020) plantea una arquitectura con los componentes de fuentes de datos, ingesta, procesamiento, análisis y por último un componente transversal de gobierno de datos, sin embargo, no menciona una capa transversal de seguridad o monitoreo de costos e infraestructura. Al igual que en la arquitectura anterior, Tapia, M., y Marisol, S. (2021) han propuesto una arquitectura referencial enfocada a la gestión de las telecomunicaciones 5G basados en una arquitectura *Lambda*, en esta arquitectura mencionan la capa de fuentes de datos, ingesta o ETL y la capa de análisis de datos, en esta última engloba la visualización y el monitoreo de los datos, sin embargo, no se menciona el monitoreo de la infraestructura y seguridad. Como la anterior y otras más se puede evidenciar generalmente que las arquitecturas Big Data siempre tienen como componentes principales las fuentes de datos, ingesta de datos, almacenamiento, análisis y procesamiento *(Arévalo, J., 2020).*

Para plantear la arquitectura de referencia de nuestro proyecto, identificamos los componentes comunes y las diferencias puntuales entre las arquitecturas revisadas anteriormente. Para los componentes comunes identificamos la capa de **almacenamiento, procesamiento y visualización** que conforman las capas necesarias para gestionar una cantidad masiva de datos en cualquier proyecto Big Data. Las capas adicionales que se tuvieron en cuenta fueron el **gobierno de datos**, que nos ayudará a estandarizar la información para el análisis de los contratos, **seguridad de los datos e infraestructura** de nuestros componentes Big Data.

4.5 Infraestructura Cloud

Hoy en día existen diferentes soluciones tecnológicas de infraestructura para el mercado, *'OnPremise', 'Cloud Computing'* o infraestructuras '*Híbridas'*, estas tres clasificaciones han evolucionado con el pasar del tiempo, tanto así que en necesidades específicas se utilizan altiempo para soportar las operaciones del negocio y cada una tiene sus ventajas y desventajas frente a la otra con respecto al costo, seguridad, configuración, mantenimiento, entre otras características.

4.5.1 Modelos de despliegue

Existen varias formas de tener la infraestructura, ya sea en algún Data Center propio, en la nube de algún proveedor de servicios, una mezcla de las dos opciones anteriores o en una nube comunitaria. A continuación, se describe un poco más a detalle estos diferentes tipos de despliegue. Cada una tiene características diferentes y depende de la necesidad de cada organización o proyecto.

4.5.1.1 Nube privada o nube on premise

En esta infraestructura las mismas organizaciones se hacen responsables de sus propios servicios, de su mantenimiento, sus actualizaciones y de configurar su capacidad de hardwarede acuerdo con las necesidades del negocio. Mantienen sus Data Centers en una ubicación físicaen la organización y tienen un equipo dedicado a su administración. La organización debe controlar absolutamente toda la infraestructura, desde la seguridad física de los componentes hasta la actualización y compra de equipos necesarios para el óptimo funcionamiento.
En la mayoría de los casos esta infraestructura es utilizada por organizaciones que son reguladas donde deban tener absoluto control de la seguridad y administración de los datos de sus clientes.

Una de las principales desventajas de este tipo de despliegue son los costos iniciales requeridos para tener la infraestructura necesaria, puesto que se deben comprar todos los componentes físicos como servidores, almacenamiento, cómputo y todo lo requerido para tener la infraestructura necesaria, sumado a esto, el cálculo para saber cuántos recursos se requieren es un tema complejo de medir, puesto que muchas veces se termina haciendo subdimensionado los recursos, es decir, se utiliza tanto los componentes que la infraestructura se queda pequeña para la necesidad o el otro extremo, terminan sobredimensionados, lo que puede llegar a generar sobrecostos innecesarios, puesto que losrecursos no se utilizan completamente.

4.5.1.2 Nube pública o cloud computing

Hoy en día la computación en la nube es una de las principales alternativas para las nuevas empresas y consiste en los procesos para almacenar, administrar y acceder a los datos que son transferidos a través del internet. Existen diferentes proveedores de nube pública como por ejemplo Google Cloud, Amazon Web Services, Azure, Snow Fake, IBM, entre otros; estos proveedores externos tienen su propia infraestructura y plataforma, las cuales ofrece como servicio a las organizaciones para ejecutar sus procesos sin incurrir en grandesinversiones iniciales.

Esta solución permite a las organizaciones rentar espacios bajo demanda a cualquier proveedor que requieran o cumpla con sus necesidades donde su información esté segura y disponible en todo momento, además permite que las empresas paguen por el uso de servidores, almacenamientos, bases de datos, redes, software de manera elástica y en tiempo real. Acá prima las políticas de responsabilidad compartida, donde los proveedores de nubes se encargan de garantizar la disponibilidad de los recursos y a su vez de la seguridad física de los componentes. Por el contrario, los clientes o suscriptores a estos servicios son los responsables de configurar los servicios de manera segura. Esto empieza a suplir las desventajas del modelo de despliegue anterior, puesto que no se requiere una inversión inicialtan alta, puesto que, a medida que se van utilizando los servicios se pueden ir aprovisionando y a medida que se requieran más recursos con unos cuantos clics (y una buena configuración) se puede ir creciendo a medida que la necesidad lo indique y no solo el crecimiento, tambiénse puede decrecer a necesidad, así solo se paga realmente por lo que se utiliza y cuando se requiera. La nube pública tiene su principal aliado en los recursos compartidos para ser competitivo a nivel de precios, pero si se requiere servicios dedicados, también los ofrece. Así que este termina siendo una excelente opción para considerar a la hora de requerir servicios.

4.5.1.3 Nube híbrida

Por otro lado, existe una solución híbrida la cual combina lo mejor de las infraestructuras On Premise y Cloud. La infraestructura híbrida depende de algún proveedor de nube pública y una infraestructura On Premise alojada en las mismas instalaciones con una conectividad entre los diferentes entornos. Esto es muy usado generalmente cuando existen algunos temas regulatorios, donde, por ejemplo, se tienen que almacenar los datos en un lugar físico específico y las nubes públicas no tienen disponibilidad en esa ubicación, para ello se almacenan los datos en un Data Center privado, pero se realiza conexión a cómputo y demás servicios hacia la nube. Hoy en día se ofrecen herramientas potentes para que esta conexión nube pública-privada sea lo más transparente y rápida posible. Y como esté hay un sinnúmero de casos de uso, donde se requiere tener alguna parte de la arquitectura en sitio y el resto si en las nubes públicas. Así que, dependiendo de las necesidades de las

organizaciones y proyectos, esta opción puede ser interesante.

4.5.1.4 Nube comunitaria

Aunque la nube comunitaria no es una de las más utilizadas, vale la pena ponerla en contextoy nombrarla puesto que es una opción válida de modelo de despliegue. Tiene muchas similitudes con las nubes públicas, pero limitadas a un área o comunidad en específico. Esta infraestructura común es compartida por diversas organizaciones u empresas con objetivos similares y relacionadas de alguna forma. Estas nubes pueden ser gestionadas por las propias empresas o por terceros que estén en la capacidad de conseguirlo.

4.5.1.5 Comparación entre nube privada y nube pública.

Para realizar los temas comparativos se han seleccionado los dos modelos de despliegue másutilizados en la actualidad, tanto el Cloud Computing, como el modelo On premise. Y se han seleccionado 4 grandes características de comparación. La seguridad, los costos, la configuración y la capacidad, a partir de estas categorías se puede identificar, dependiendo la necesidad cuales son los principales beneficios/desventajas de cada opción.

Tabla 1. Comparación entre modelos de despliegue

Ítem de comparación	**On premise**	**Cloud Computing**
Seguridad	Absoluto control sobre la seguridad y accesibilidad a los datos. Tanto a nivel físico como a nivel lógico y de aplicaciones.	Depende del mismo nivel de seguridad del proveedor de la nube. Pero la seguridad física está 100% limitada al proveedor cloud. Para lo demás se basa en el modelo de responsabilidad compartida
Costo	Altos costos de capital (CapEx). Si no se realiza un balance previo se puede incurrir en gastos adicionales o quedarse sin presupuesto para aumentar las capacidades de infraestructura (Srinivasan, Ravi, & Raj, 2018).	Altos costos operacionales (OpEx). Aumentan los costos por los servicios que se usan y se debe monitorear frecuentemente. No hay límite de capacidades mientras se tenga el dinero para pagar su uso.

Ítem de comparación	On premise	Cloud Computing
Configuración	Se debe contar con un equipo de TI para configurar y montar las aplicaciones, además de realizar toda la configuración a nivel de hardware, software e infraestructura.	Una mínima configuración de los servidores, servicios que minimizan el tiempo de configuración para su uso.
Capacidad	Se debe tener un análisis previo del crecimiento de la infraestructura ya que es más costoso aumentar la capacidad	Es escalable, se puede aumentar y reducir las capacidades de acuerdo con la demanda de las aplicaciones automáticamente.

Para el proyecto se ha decidido utilizar una solución 'Cloud' luego de evaluar los beneficios, este modelo de despliegue tiene grandes utilidades respecto a los costos, agilidad, escalabilidad y seguridad en general, no requerimos de regulaciones como unas entidades bancarias, por lo que con los beneficios que mencionamos es suficiente. Además, podremos configurar las capacidades de la infraestructura en cuanto el almacenamiento y procesamiento de acuerdo con las necesidades del proyecto reduciendo los gastos de capital, los tiempos de implementación y el mínimo esfuerzo en la configuración del hardware. Así mismo al ser pensado este proyecto como una arquitectura de referencia donde se pueda aplicar a cualquier caso de uso, se podrá escalar de forma sencilla dependiendo de las necesidades de cada involucrado. Además, no se poseen todos los recursos necesarios para montar y administrar toda una infraestructura para soportar la arquitectura.

4.5.2 Modelos de servicio

Para entender un poco más cómo funcionan las soluciones en la nube debemos conocer los diferentes tipos de implementación que se pueden realizar en las soluciones en la nube, tanto para la nube privada, pública e híbrida como menciona *Yevge* (2022), con los requisitos actuales para este caso de uso y para la arquitectura de referencia que se implementó, solo utilizaremos la nube pública, aunque la arquitectura tiene ciertas capacidades agnósticas a la tecnología, que podrán terminar siendo implementadas en cualquier tipo de nube, creando servicios similares a los que

se implementaron acá. Continuando con el entendimiento de la nube, debemos saber que existen varios tipos de servicios que nos pueden ayudar con la infraestructura, plataformas o sistemas de software, aquí abordamos algunos de los principales: IaaS: Infraestructura como Servicio, PaaS: Plataforma como Servicio y SaaS: Software como Servicio (Bayrakdar & Nogara, 2019), aunque, con el pasar del tiempo, se vienen creando nuevas alternativas, dependiendo de las necesidades de las organizaciones, por ejemplo BaaS: Backend como Servicio, CaaS: Containers como Servicios, FaaS: Funciones como Servicios, entre otros, pero no serán explicadas en este documento, pero sí es importante nombrar que a día de hoy existen varios modelos y estos seguirán creciendo aún más dependiendo de las necesidades de cada usuario.

4.5.2.1 Infraestructura como servicio (IaaS)

En este modelo de servicio, el mismo proveedor de la nube aloja, administra y mantiene todo el hardware que soportará la operación del negocio de la organización, esto nos permite crear y ejecutar cargas de trabajo sin necesidad de tener el hardware de forma local ayudando a reducir los costos de capital, pagando únicamente por el uso del hardware requerido (Rajan, 2014). Para este modelo las capas que estarán a cargo del proveedor Cloud serán: La capa de Networking, Storage, Server y la virtualización, por otro lado, la responsabilidad que recae sobre el cliente será con las capas del: Sistema Operativo, Middleware, Runtime, Data, Applications. Cómo se puede evidenciar, hay varias capas donde se debe trabajar todo el tema de configuración y seguridad, la responsabilidad del cliente o usuario que consume el servicio será mayor a los demás modelos.

4.5.2.2 Plataforma como servicio (PaaS)

En el modelo de plataforma como servicio los proveedores ofrecen entornos estables y preconfigurados para desarrollar y alojar aplicaciones, de esta manera los usuarios pueden desarrollar aplicaciones en un entorno estable y eficaz permitiendo el ahorro de tiempo en las pruebas, ejecuciones, versionamiento, despliegues de las aplicaciones desarrolladas y configuraciones de los componentes (Gupta, Singh, Das, & Choudhary, 2022). Para este modelo de servicio, las capas bajo responsabilidad del proveedor cloud aumenta, facilitando así la configuración y mantenimiento de las aplicaciones por parte del usuario final, las capas que estarán a cargo del proveedor Cloud serán: La capa de Networking, Storage, Server, la virtualización, Sistema Operativo, Middleware y Runtime. Por el contrario, para el usuario final será Data y Applications. Comparado con el modelo IaaS, la responsabilidad del usuario disminuye notablemente, facilitando la construcción y mantenimiento de los servicios.

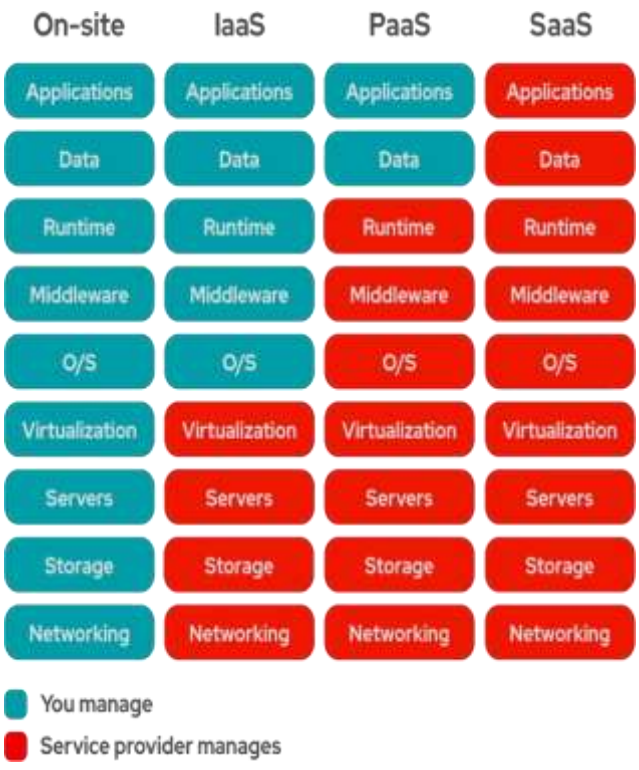
4.5.2.3 Software Como Servicio (SaaS)

En este modelo de software como servicio los proveedores de la nube ofrecen aplicaciones o software empresariales requeridos para los usuarios finales de las organizaciones y brinda a los usuarios la facilidad de acceder directamente desde cualquier navegador con acceso a internet (Gupta, Singh, Das, & Choudhary, 2022). Además, las licencias que se requieren para brindar el software a los usuarios se pueden pagar como la mayoría de los servicios en la nube; pago por uso (Rajan, 2014). Acá todas las capas estarán a cargo del proveedor cloud, puesto que lo único que deberá realizar el usuario final, será consumir el servicio directamente. Sin necesidad de realizar configuración a nivel de hardware, lo único que deberá configurar serán parámetros dentro del aplicativo a utilizar, así que la mayor responsabilidad recaerá sobre el proveedor.

4.5.2.4 Comparativa entre los diferentes modelos de servicios

En la siguiente figura se podrá encontrar un resumen de la responsabilidad tanto del proveedor Cloud como del usuario final por cada uno de los modelos, esto permitirá realizaruna comparativa entre todos los servicios.

Ilustración 3.Diferencias entre IaaS, PaaS y SaaS.



Tomado de (Redhat, 2022)

4.6 Proveedores de la nube

Luego de conocer los conceptos fundamentales de la nube, es importante poder identificar a los principales proveedores de computación en la nube en el mercado actual y los tipos de servicios que cada uno ofrece, centrando la información en las principales categorías de servicios que nos competen a la hora de la ejecución del proyecto como lo son el cómputo, almacenamiento, procesamiento de datos, bases de datos, herramientas de Big Data, análisis avanzado, monitoreo y costos, principales característica de las arquitecturas de referencias en el procesamiento de datos batch y en tiempo real.

Para identificar a los principales players de la nube consultamos el cuadrante mágico de Gartner. Es una representación gráfica que proporciona una visión panorámica de las posiciones relativas de un mercado específico sobre las empresas dedicadas a la tecnología anivel mundial (Carter, 2022)

El informe reportado por Gartner para servicios de infraestructura y plataforma en la nube nos expone a los proveedores líderes del mercado (Bala, Gill, Smith, Ji, & Wright, 2021), con un resumen de los puntos fuertes y precauciones por proveedor, adicional se encuentran señalados los líderes contundentes del cuadrante lo cuales son Amazon Web Services, Microsoft y en tercer lugar Google.

Teniendo en cuenta la visual anterior del cuadrante mágico tenemos la facultad para ceñir nuestra investigación exclusivamente en los líderes del mercado, que en este caso son los tres proveedores mencionados: Amazon web Services (AWS), Azure Microsoft y Google CloudPlatform (GCP).

A continuación, abordamos un poco más sobre estos proveedores dando un alcance amplio de las ventajas y desventajas de cada uno.

4.6.1 Amazon Web Services (AWS)

El proveedor de la nube Amazon web Services es uno de los jugadores más antiguos en el mercado, se inició en 2006 y por esto es uno de los líderes indiscutibles del mercado con más de 87 zonas de disponibilidad en 27 regiones geográficas de todo el mundo. Proporciona a sus usuarios servicios de almacenamiento, computación, bases de datos y muchos más (Yevge, Ghag, Solanki, & Mishra, 2022).

Tabla 2. Ventajas y desventajas de AWS

Ventajas	Desventajas

Ventajas	Desventajas
• Tiene una amplia disponibilidad de servicios.	• Costo alto frente a sus principalesAzure y GCP.
• Por su gran ventaja en el mercado tiene un crecimiento continuo de servicios. • Maneja grandes usuarios y recursos. • Alcance mundial. • Funcionalidades mejoradas de servicios.	• Tarifas adicionales para el soportetécnico. • Tarifas adicionales en serviciosesenciales. • Su plataforma no es familiar con lasmarcas de GCP y Azure.

Nota. Basado en (Kamal*, Raza, Alam, & Su'ud*, 2020) y (Gupta, Mittal, & Mufti, 2021).

4.6.2 Azure Microsoft

Es uno de los que ha tenido un crecimiento más rápido y está en el segundo puesto como líder del mercado. Azure ofrece servicios de fácil configuración y servicio rápido al aprovechar la estructura existente establecida por las ofertas de software y aplicaciones comerciales de Microsoft (Yevge, Ghag, Solanki, & Mishra, 2022).

Tabla 3. Ventajas y desventajas de Microsoft Azure

Ventajas	Desventajas
• Integración con herramientas y software Microsoft. • Compatibilidad con código abierto. • Precios bajos en comparación con AWS. • Nube pública y privada integrada.	• Problemas con el soporte técnico. • Diseño menos profesional para el usode servicios. • Requiere experiencia en plataformas.

Nota. Basado en (Kamal*, Raza, Alam, & Su'ud*, 2020) y (Gupta, Mittal, & Mufti, 2021).

4.6.3 Google Cloud Platform (GCP)

Google Cloud Platform se lanzó oficialmente en 2008. Es uno de los competidores activos de AWS y Azure, ofrece IaaS y PaaS con servicios de computación, almacenamiento, Big Data, Bases de Datos y otros más. Actualmente cuenta con 34 regiones geográficas en todo el mundo y 103 zonas de disponibilidad (Gupta, Mittal, & Mufti, 2021).

Tabla 4. Ventajas y desventajas de GCP

Ventajas	Desventaja
• Escalabilidad superior. • Fácil de usar y configurar. • Es compatible con lenguajes como Python y Java. • Precios ajustables	• No tiene funcionalidades mejoradas deservicios. • Menos oferta de servicios. • Menos diversidad de características. • Mejor seguridad que AWS y GCP.

Nota. Basado en (Kamal*, Raza, Alam, & Su'ud*, 2020) y (Gupta, Mittal, & Mufti, 2021).

4.6.4 Comparativo de nubes

Con esto definido y continuando con la investigación se presenta un cuadro comparativo por las categorías de servicios que se mencionaron en el capítulo de arquitecturas Big Data y que hacen semejanza a las arquitecturas de referencias, con ayuda del mapeo de nivel de servicioy características de *Ilyas* (s.f.) en *"A public cloud comparison*" y *"Compare AWS and azure services to"*, profundizamos por cada proveedor y la facilidad para desplegar componentes. Este cuadro busca definir las principales herramientas por cada categoría asociado a cada proveedor en la nube, esto teniendo como principal objetivo ampliar los posibles servicios a utilizar, puesto que la arquitectura de referencia propuesta será agnóstica al proveedor y al servicio elegido, así que si en algún momento se decide cambiar de proveedor en la nube se tenga claro qué alternativas de servicios se pueden utilizar.

Tabla 5. Comparación de servicios entre nubes

Categoría	Servicio	AWS	Google cloud platform	Azure
Almacenamiento y bases de datos	Object Storage	Amazon S3	Google cloud storage	Azure Blob
	Data Warehousing	Amazon Athena	BigQuery	Azure Synapse Analytics
Big data y Análisis	Big Data Query como servicio	Amazon Athena	BigQuery	Azure DataLake Analytics

Categoría	Servicio	AWS	Google cloud platform	Azure
avanzado	Cloud ETL	AWS Glue	AppFlow	Azure Dat aFactory
	Big Data Managed Cluster como servicio	Amazon EMR	Google Dataproc	Azure HDInsight
Monitoreo y costos	Logging & Monitoring	Amazon CloudWatch	Cloud Monitoring	Azure Monitor
	Cost explorer	AWS Cost explorer	Management Google	Azure Cost Management
Procesamiento en tiempo real	Streaming	Amazon Kinesis	Dataflow	Azure Stream Analytics
Gobierno de datos	Metadata	Glue Data Catalog	Data Catalog	Azure Purview
Seguridad	Web Applica tionFirewall	AWS WAF	Google Cloud Armor	WAF Azure

4.7 Infraestructura cómo código (IaC)

El concepto de infraestructura como código (IaC) es un enfoque que nos permite gestionar la infraestructura de TI en la Nube a través de archivos de configuración repetibles mediante código, en lugar de hacer el proceso de manera manual. Estos archivos de configuración facilitan la edición y distribución de las configuraciones ayudando a evitar cambios ad hoc yno documentados.

Este enfoque está directamente relacionado con la infraestructura en la nube, ya que parte de la práctica de DevOps y aplica la orquestación de la infraestructura con el control de versiones de repetibilidad que utilizan los desarrolladores para el código fuente de las aplicaciones (Redhat, 2022).

Esta práctica tiene beneficios tales como:

- Automatización.
- Eficiencia.
- Escalabilidad.
- Desarrollo ágil.
- Mayor uniformidad de la infraestructura en los entornos.

Hay dos maneras de abordar la infraestructura como código, mediante un enfoque declarativoo imperativo. El enfoque declarativo define un estado deseado del sistema sin indicar cómo conseguirlo e incluye los recursos necesarios y las propiedades de dichos sistemas, y la herramienta de infraestructura como código se encargará de configurarlo por usted.

Por otro lado, el enfoque imperativo define paso a paso los comandos específicos para lograrla configuración deseada pero también, la manera de conseguirlo. También se conoce comoel enfoque procedimental, esta emplea instrucciones explícitas y que no admiteactualizaciones.

4.7.1 Herramientas de la Infrastructure as Code

Actualmente en el mercado existen diversas herramientas para aprovisionar infraestructura como código que permiten asignar y configurar los recursos de una infraestructura en la Nube de forma que podamos gestionarla con versionamiento y documentación, ya sea por la línea decomandos, GUI, script (Dominguez-Quintero & Vargas-Lombardo, 2011).

Estas son algunas de las opciones más conocidas:

- Chef
- Puppet
- Saltstack
- Terraform
- AWS CloudFormation

5 Resultados y contribución

En esta sección presentaremos los resultados sobre el análisis y desarrollo de indicadores de alerta de fraude o presunta corrupción, eso sí, teniendo en cuenta que esto es un ejercicio académico y no emitimos juicios de si un contrato es corrupto o no, el objetivo de estos indicadores es alertar cualquier tipo de anomalía que se pueda presentar con la información obtenida de las plataformas de SECOP I, SECOP II y PACO, además presentaremos la arquitectura Big Data propuesta que soporta este casode uso

de presunta corrupción sobre la contratación pública.

Mediante estos resultados se espera que otros interesados sobre la temática de corrupción en la contratación pública puedan utilizar dichos indicadores para realizar nuevas alertas de fraude y que la arquitectura de referencia propuesta sea pertinente para usarla en otros casos de uso de manera ágil y automatizada ya que se entregarán el código fuente de su construcción, así como los manuales y documentación necesaria para un buen entendimiento.

5.1 Modelo y gobierno de datos

El modelo de datos es una representación visual o en un lenguaje estándar para poder hablar sobre los datos de una base de datos, empresa o proyecto. Es esencial para definir y estructurar los datos al contexto del proyecto, teniendo en cuenta que existen diferentes orígenes de información y cada uno maneja su propia nomenclatura, así que al ingresar todos al sistema se requiere unificar y así poder tomar decisiones coherentes con toda la información recolectada.
El modelo de datos propuesto definirá una nomenclatura de tablas y de campos estándar para todos los datos que utilizaremos para la construcción de nuestros indicadores y tableros informativos que presentaremos a continuación y en todo el ciclo de vida del dato.

5.1.1 Nomenclatura de tablas

Para crear los nuevos nombres de las tablas se definió una nomenclatura que está conformada por el siguiente estándar *t_yyyy_xxxxxxxxxxx_zzzzzzzzzzzz*, además, no deberá superar los 32 caracteres y deberá cumplir con las reglas presentadas a continuación:

- Toda tabla deberá iniciar por la letra t.
- Los 4 caracteres *yyyy* corresponden al origen de la información, para SECOP I se utilizará seci, para SECOP II será seii, para el origen en PACO será paco, para el SECOP Integrado se utilizará sein, para aquellos orígenes que sean de la tienda virtual del estado colombiano se utilizará tvec, para los demás orígenes se utilizará otro.
- Los 12 caracteres *x* corresponden a un identificador único de la tabla.
- Los 12 caracteres z corresponden al complemento de identificación de la tabla.

5.1.2 Nomenclatura de campos

Para crear los nuevos nombres de las columnas se definió una nomenclatura que está conformada por <indicativo>_<nombre_columna>, además, no deberá superar los 32 caracteres y deberá cumplir con las reglas a continuación:

- Para las columnas que tienen información en porcentajes se utilizará el indicativo ‘por’ al inicio del nombre de la columna y se debe castear a decimal de dos posiciones.Ej.: por_<nombre_columna>, decimal(7,4).
- Para las columnas que tienen información en pesos se utilizará el indicativo ‘monto’al inicio del nombre de la columna y se debe castear a decimal de dos posiciones. Ej.: monto_<nombre_columna>, decimal(30,3).
- Para las columnas que tienen información de fechas se utilizará el indicativo ‘fecha’ al inicio del nombre de la columna y se debe castear a date con el formato AAAA- MM-DD. Ej.: fecha_<nombre_columna>.
- Para las columnas que tienen información de identificadores únicos se utilizará el indicativo ‘id’ al inicio del nombre de la columna. Ej.: id_<nombre_columna>.
- Para las columnas que tienen información booleana o algún estado se utilizará el indicativo ‘tipo’ al inicio del nombre de la columna. Ej.: tipo_<nombre_columna>.
- Para las columnas que tienen información de nombres propios se utilizará el indicativo ‘nombre’ al inicio del nombre de la columna. Ej.: nombre_<nombre_columna>.
- Para las columnas que tienen información de descripciones o justificaciones largas seutilizará el indicativo ‘desc’ al inicio del nombre de la columna. Ej.: desc_<nombre_columna>.
- Para las columnas que tienen información de cantidades se utilizará el indicativo ‘cant’ al inicio del nombre de la columna. Ej.: cant_<nombre_columna>.

5.1.3 Datos orígenes identificados

El gobierno y modelo de datos mencionado anteriormente, las transformaciones y las inconsistencias mencionadas se implementaron para las fuentes identificadas a continuacióny se ven a detalle en el Anexo 1 **(Anexo modelo de datos orígenes)**:

- PACO - Colusiones en Contratación Pública - Superintendencia de Industria y Comercio.
- PACO - Responsabilidades fiscales
- PACO - Multas y sanciones contractuales
- PACO - Registro nacional de obras inconclusas

- PACO - Lista sancionados BID
- PACO - Lista Clinton - Lista sancionados OFAC
- SECOP Integrado
- SECOP I - Adiciones
- SECOP I - Modificaciones a Procesos
- SECOP I - Modificaciones a Adjudicaciones
- SECOP I - Multas y Sanciones SECOP I
- SECOP I - Origen de los Recursos
- SECOP I - PAA Detalle
- SECOP I - PAA Encabezado
- SECOP I - Procesos de Compra Pública
- SECOP II - Ubicaciones Adicionales
- SECOP II - Proveedores Registrados
- SECOP II - Rubros Presupuestales
- SECOP II - Garantías
- SECOP II - Adiciones
- SECOP II - Solicitudes CDPs
- SECOP II - Modificaciones a Procesos
- SECOP II - Ejecución Contratos
- SECOP II - Proponentes por Proceso
- SECOP II - Suspensiones de Contratos
- SECOP II - Facturas
- SECOP II - Procesos de Contratación
- SECOP II - PAA - Encabezado
- SECOP II - Objeto contiene PAE - Procesos de Contratación
- SECOP II - Compromisos Presupuestales
- SECOP II - Ofertas Por Proceso
- SECOP II - Contratos Cancelados
- SECOP II - Grupos de Proveedores
- SECOP II - Contratos Electrónicos
- SECOP II - Plan Anual De Adquisiciones Detalle
- SECOP II - Multas y Sanciones
- TVEC - Tienda Virtual del Estado Colombiano - Consolidado
- TVEC - Ítems
- Otros - Antecedentes de SIRI - Sanciones disciplinarias
- Otros - Conjunto servidores públicos
- Otros - Puestos Sensibles a la Corrupción
- Otros - Personas Expuestas Políticamente (PEP)
- Otros - Personas Naturales, Personas Jurídicas y Entidades Sin Ánimo de Lucro
- Otros - Base ingresos cuentas claras

5.1.4 Inconsistencias de datos encontradas

A la hora de analizar los datos, se pudo evidenciar una gran cantidad de inconsistencias en ellos. Esto generó diversos problemas a la hora de generar los indicadores, así que se necesitó identificar aquellos que podían generar resultados erróneos e incomprensibles para entrar a tomar algunas medidas de control de calidad de datos. Por otro lado, algunas herramientas utilizadas en la etapa de construcción de indicadores también necesitó realizar algunos ajustesa los datos para que no presentara fallos al ingresar al sistema. Todo esto quedó documentadoen la siguiente tabla, para que sean tenidos en cuenta. Cabe aclarar que estas no son las únicasinconsistencias que presentan los datos, acá se corrigieron todas aquellas que afectan a los indicadores propuestos o que generar algún inconveniente para poder ser registradas en la plataforma.

Tabla 6. Inconsistencias de los datos

Nombre del objeto en el data lake	Tipo inconsistencia	Nombre Columna	Descripción
t_paco_responsabilidad_fiscales	Inconsistencia de datos con respecto a las columnas	N/A	El archivo origen tenía 1127 registros que no corresponden al orden delas columnas.
t_paco_listas_sancionados_bid	Diferencias en el tipode formato de datos	fecha_inicio_sancion	Se encuentra el formato MMMM dd, yyyy sin embargo el Mes se registra en español como Enero, Febrero, Marzo, etc.
t_seci_pa_adquisicion_encabezado	Fechas inconsis tentes	fecha_mod_plan_a nual_adquisicion	Se encuentran años inferiores a 1582, lo que genera fallos en el casteode información. Senulean estas fechas.
t_seci_pa_adquisicion_detalle	Fecha Inicio	fecha_inicio_comp ra_item	Se encuentran años inferiores a 1582, lo que genera fallos en el casteode información. Senulean estas fechas.
t_seci_pa_adquisicion_detalle	Tipo de datos	monto_estimado_co mpra _futura	Se encuentran String en una columna que hace referencia a un tipo dedato Decimal.

Nombre del objeto en el data lake	Tipo inconsistencia	Nombre Columna	Descripción
t_seii_ejecucioncon_avancerevses	Fechas inconsistentes	fecha_entrega_real	Se encuentran años inferiores a 1582, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_seii_facturaplata_solopagassec	Fechas inconsistentes	fecha_factura, fecha_entrega, fecha_estimada	Se encuentran años inferiores a 1582, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_seii_rubros_presupuestales	Identificador Únicosin informar	id_unico_rubro	Se encuentran valores con 'No Definido'
t_otro_pernajuesadl_camarcomerci	Fechas inconsistentes	fecha_matricula	Se encuentran años inferiores a 1582, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_otro_pernajuesadl_camarcomerci	Fechas inconsistentes	fecha_renovacion	Se encuentran años inferiores a 1582, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_otro_pernajuesadl_camarcomerci	Fechas inconsistentes	fecha_vigencia	Se encuentran años inferiores a 1582, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_otro_pernajuesadl_camarcomerci	Fechas inconsistentes	fecha_cancelacion	Se encuentran años inferiores a 1582, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_otro_baseingresos_cuentasclara	Fechas inconsistentes	fecha_comprobante	Se encuentran años anteriores a 1900 para formatos timestamp, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_seii_garantias_publicadas	Fechas inconsistentes	fecha_creacion_poliza	Se encuentran años anteriores a 1900 para formatos timestamp, lo que genera fallos en el casteo de información. Se nulean estas fechas.

Nombre del objeto en el data lake	Tipo inconsistencia	Nombre Columna	Descripción
t_seii_garantias_publicadas	Fechas inconsistentes	fecha_envio_poliza	Se encuentran años anteriores a 1900 para formatos timestamp, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_seii_garantias_publicadas	Fechas inconsistentes	fecha_fin_poliza	Se encuentran años anteriores a 1900 para formatos timestamp, lo que genera fallos en el casteo de información. Se nulean estas fechas.
t_seii_garantias_publicadas	Fechas inconsistentes	fecha_registro_plataforma	Se encuentran años anteriores a 1900 para formatos timestamp, lo que genera fallos en el casteo de información. Se nulean estas fechas.

5.1.5 Nomenclatura - Resultados de los indicadores procesamiento en tiempo real

De acuerdo con el gobierno y modelo de datos realizado con los datos origen, se definió la estructura para los resultados de los indicadores de irregularidades propuestos cuando se evalúen los datos en tiempo real, que difiere con el análisis de los datos por lotes o históricos.La principal diferencia radica en que los indicadores en tiempo real validan la información anivel de contrato, es decir, por cada contrato que llega al sistema se valida cada uno de los indicadores definidos y se termina si el contrato tiene algún riesgo de generar presunta corrupción basado en estas banderas rojas.

Nombre del objeto en el data lake: *t_result_indicadores_stream*
Partición: fecha_ejecucion

Tabla 7. Nomenclatura utilizada en los resultados en tiempo real.

Nombre columna origen	Tipo dato	Descripción
id_contrato	string	Identificador único del contrato.
nombre_grupo_indicador	string	Esta información es una categorización del indicador que se está presentando. Ya sea indicador de inhabilidad, de incumplimiento u otros**.**
nombre_indicador	string	Es el nombre del indicador que se está reportando.

Nombre columna origen	Tipo dato	Descripción
tipo_alerta_irregularidad	string	Identifica si el indicador que se está analizando está generando algún tipo de inconsistencia. Puede tomar elvalor de Sí o No.
monto_contrato	decimal	Valor del contrato que se está analizando.
fecha_ejecucion	date	Fecha en que se ejecuta el análisis de los indicadores,generada automáticamente por el sistema.

5.1.6 Nomenclatura - Resultados de los indicadores de procesamiento por lotes

De acuerdo al gobierno y modelo de datos realizado con los datos origen, se definió la estructura para los resultados de los indicadores de irregularidades propuestos para elprocesamiento por lotes. Acá se busca analizar los datos históricos y determinar los totales que han presentado el indicador propuesto.

Nombre del objeto en el data lake: *t_result_indicadores_batch*
Partición: fecha_ejecucion

Tabla 8. Nomenclatura utilizada en los resultados por lotes.

Nombre columna origen	Tipo dato	Descripción
nombre_grupo_indicador	string	Esta información es una categorización del indicador que se está presentando. Ya sea indicador de inhabilidad, de incumplimiento u otros.
nombre_indicador	string	Es el nombre del indicador que se está reportando.
cantidad_irregularidades	bigint	Cantidad de irregularidades que se presentan en dicho indicador.
cantidad_contratos_irregularidades	bigint	Cantidad de contratos que se reportan que cumplen dicha irregularidad, a veces se presenta que el mismo contrato está más de una vez en cada irregularidad y por eso se determina la cantidad única de contratos.
monto_total_irregularidades	decimal	Representa la suma total del monto de los contratos que presentan dicha irregularidad.
cantidad_contratos_totales	bigint	Cantidad de contratos que son analizados en el indicador.
feca_ejecucion	date	Fecha en que se ejecuta el análisis de los indicadores, generada automáticamente por el sistema.

5.2 Indicadores de irregularidades propuestos

Luego de toda la investigación realizada, se han definido algunos indicadores para el caso de uso. Se ha dividido en 3 categorías, aquellos indicadores que generan algún tipo de alerta por inhabilidad de los proveedores o de aquellas personas que se les ha asignado el contrato, la segunda categoría aquellos indicadores por incumplimiento y por último una categoría genérica denominada Otros indicadores. Cabe aclarar que estos indicadores o banderas rojas se han generado por reglamentación, por contratos históricos que han presentado una casuística similar y ha generado pérdidas para el estado o basados en indicadores de otras investigaciones. Para este proyecto se han definido 10 indicadores en total, teniendo en cuenta que si bien, podrían existir cientos de indicadores de diferentes complejidades según toda la investigación, acá se limita mucho a los datos disponibles, este ha sido uno de los mayores inconvenientes a la hora de definirlos. A continuación, se detallarán todos estos y se dará una breve introducción a cada uno de ellos.

5.2.1 Indicadores por inhabilidad

La primera categoría definida, son aquellos indicadores que se presentan por inhabilidad de los proveedores asignados a los contratos, esto es una alerta que se debe tener en cuenta puesto que si los contratistas ya han tenido algunas multas, obras inconclusas o algún tema de responsabilidad fiscal existe una alta probabilidad que vuelva a suceder algo similar. Así que el objetivo de estos indicadores es alertar este tipo de situaciones para evitar posibles irregularidades por inhabilidad de los contratistas. A continuación, se definen 3 indicadores para esta categoría.

5.2.1.1 Inhabilitados por multas

Existen diversos motivos para ser multado en la contratación pública, pero uno de los más comunes es por incumplimiento de los términos y condiciones de los contratos, estos incumplimientos lo único que genera es impacto en los recursos económicos del estado y por consiguiente un posible detrimento patrimonial, es importante detectar aquellos contratos que han sido adjudicados a un contratista que ha tenido algún tipo de multa en la contratación pública, es por ello que se genera este indicador, pues será fundamental detectar este tipo de casos para alertar alguna posible inhabilidad del proveedor seleccionado. Es importante mencionar que si el contratista ya ha sido multado alguna vez, existe una alta posibilidad de que vuelva a suceder, así que es importante poder generar este tipo de alertas para hacer un mayor seguimiento y evitar posibles contratiempos en los contratos.

5.2.1.2 Inhabilitados por obras inconclusas

Es importante identificar que es una obra inconclusa, así que si validamos la definición, hace referencia a una obra que le falta algún elemento para que pueda ser

puesta en funcionamiento, ya sea a nivel físico, reglamentario, certificaciones, entre otros y se presenta cuando un año después de finalizado los términos del contrato dicha obra no entra en funcionamiento. Estas obras con el pasar del tiempo se empiezan a convertir en los llamados "elefantes blancos" puesto que se ha invertido dinero sin dar el beneficio esperado, muchas veces estas obras se convierten en un centro de corrupción puesto que se aprovechan los contratistas para dejar la obra sin terminar y sacando beneficio propio de los recursos del estado. Así que si algún contratista ha dejado alguna obra inconclusa no debería poder seguir contratando con el estado, puesto que, si ya lo hizo una vez, nada garantizará que no lo siga repitiendo así que estos contratistas deberán ser seguidos de forma especial para evitar posibles casos de corrupción.

5.2.1.3 Inhabilitados por responsabilidad fiscal

Según la Contraloría General de la Nación cuando se habla de la responsabilidad fiscal el principal objetivo es poder llegar a determinar si existe alguna irregularidad que pueda generar daños en el patrimonio público, esto quiere decir que aquellas personas que estén en esta lista han sustraídos al erario dineros de forma fraudulenta, así que el objetivo de este indicador es poder determinar si alguno de los contratos registrados en esta base tienen algún proveedor que ha sido catalogados como responsables fiscales dentro del país.

5.2.2 Indicadores de incumplimiento

Es importante poder identificar aquellos contratistas que han generado algún tipo de incumplimiento en la contratación pública en Colombia, estos indicadores tienen como objetivo poder alertar aquellos contratos que tiene asignado un contratista que ha tenido contratos cancelados (contratos que no tuvieron un buen término, es decir que no finalizaron como se esperaba) o aquellos contratos que han tenido uno o más incumplimientos en los entregables del proyecto, es importante hacer un seguimiento exhaustivo a estos contratos para evitar que se vuelva a repetir y se pierda dinero público.

5.2.2.1 Contratistas con contratos cancelados

Si un contratista ha tenido adjudicado contratos con el estado y por lo menos uno ha sido cancelado genera cierta incertidumbre de cumplimiento de este contratista, así que será importante reportar cuando este contratista vuelva a tener algún contrato con el estado colombiano, esto debido a que existe una alta probabilidad de generar el mismo incumplimiento, así que se deberá alertar cuando esto ocurra para poder realizar un mejor seguimiento y evitar posibles irregularidades en más contratos.

5.2.2.2 Contratos con incumplimiento de entregas

El SECOP II posee una fuente de datos que tiene toda la información relacionada con la ejecución estimada y real de los contratos firmados en dicha plataforma, esta fuente permiteidentificar aquellos contratos que están realizando entregas parciales de menor cantidad de elementos a recibir de lo esperado, es por ello que se busca identificar aquellos contratos, proveedores y montos asignados que presenten este tipo de casuísticas, puesto que asignar contratos a proveedores incumplidos es una señal de alerta que se debe tener en cuenta para evitar posibles casos de presunta corrupción y de elefantes blancos en la contratación estatal. Esta bandera roja permitirá saber a quién se debe controlar mejor en cuanto a seguimiento del contrato para evitar posibles casos de corrupción.

5.2.3 Otros indicadores

Existen otros indicadores que serán importante alertar, pues puede generar posibles irregularidades en la contratación pública. Se validaron casos como abuso de la contratación, selección de ofertas más costosas, aquellos contratos que tienen proveedores inactivos, contratos con proveedores de Personas Expuestas Políticamente y contratos con proveedores con puestos sensibles, a continuación, se dará una breve explicación de cada uno de ellos.

5.2.3.1 Abuso de la contratación

Si conocemos un poco de la historia de la contratación pública en Colombia e identificamos los principales casos de corrupción, podemos retroceder hacia el año 2016, allí en la gobernación de Norte de Santander se adjudicaron 4 contratos por más de 50 mil millones para el Programa de Alimentación Escolar (PAE) a un mismo proveedor, Corporación de Desarrollo Social Tanai Jawa, para el año 2018, la Procuraduría General de la Nación abrió un proceso disciplinario en contra del gobernador de esta región, incluyendo a la secretaría general y la secretaria de Educación, quienes fueron vinculadas al proceso (Radiografía de los hechos de corrupción en Colombia, 2019). De aquí nace el indicador de abuso decontratación, se valida que una misma entidad no tenga 3 o más contratos con un mismo proveedor dentro del mismo año, puesto que esto podría generar algunas irregularidades, así que será importante alertar cuando esto esté ocurriendo.

5.2.3.2 Ofertas costosas

La contratación pública tiene diferentes modelos o tipos de contratación, algunos de ellos permiten que diversos proveedores oferten valores de acuerdo con las condiciones que cada uno tenga para un producto o servicio, a partir de estas

ofertas, el estado selecciona la másacorde para el objetivo del contrato. Uno de estos atributos que se validan es el precio de oferta, es por ello que sabiendo que no necesariamente la más económica es la mejor oferta, se ha generado este indicador que puede evidenciar posibles irregularidades por no seleccionar la oferta más económica, teniendo en cuenta que no necesariamente un indicador sea garantía de un acto de corrupción, pero sí ayuda a identificar casuísticas que permita investigar más a fondo si existe algún acto de presunta corrupción. Si existe más de un oferente con el mismo valor ofertado, se debería seleccionar el primero en presentar la oferta. Cuando se selecciona una oferta más costosa se podría estar generando un detrimentopatrimonial de los recursos del estado.

5.2.3.3 Contratos con proveedores inactivos

Las empresas en Colombia deben estar debidamente legalizadas y estar adscritas a alguna Cámara de Comercio. Cada cierto tiempo, estas empresas deben renovar la documentación respectiva para estar activas y al día, si en algún punto estos proveedores tienen algún contrato público activo, pero su renovación en la Cámara de Comercio no se ha generado, deberá generar algún tipo de alertamiento por no estar legalmente funcionando. Esto puede ser una causal de posibles casos de corrupción que deberá ser investigado a mayorprofundidad.

5.2.3.4 Contratos con proveedores PEP

En Colombia existe una clasificación de Personas Expuestas Políticamente (PEP) registradas por las diferentes entidades y organismos del Estado en el Sistema de Información y Gestión del Empleo Público (SIGEP), esta clasificación permite identificar a aquellas personas que gracias a su posición destacada o influyente, es más susceptible a estar involucrada en soborno o corrupción, estas personas no deben tener ningún contrato con el estado, puesto que por su posición puede ser considerado irregular, o puede utilizar su poder para que le sean asignados contratos sin cumplir todos los requisitos para ello, poreso será fundamental poder identificar cuando esto ocurra.

5.2.3.5 Contratos con proveedores con puestos sensibles

En Colombia existe un listado de servidores públicos que tienen una vinculación activa en el Sistema de Información y Gestión del Empleo Público – SIGEP, estos empleados ejercen cargos de confianza y manejo presupuestal de nivel directivo y libre nombramiento y remoción, lo que permiten ser un punto de análisis por el alto riesgo de generar posibles casos de corrupción. Acá se buscará identificar aquellos contratos que tienen un proveedorasignado y que a su vez este proveedor tenga algún puesto sensible, estas personas no debentener ningún contrato con el estado, puesto que por su posición puede ser considerado irregular, o puede utilizar su poder para

que le sean asignados contratos sin cumplir todos los requisitos para ello, por eso será fundamental poder identificar cuando esto ocurra.

5.3 Resultado análisis de indicadores históricos por lotes

Se ha realizado la ejecución de los indicadores con los datos históricos descargados de las diferentes fuentes de información a corte del 28 de febrero del año 2023, estos datos contienen información histórica aproximadamente desde el año 2015, cabe aclarar que cada una de las fuentes de información tiene diferentes periodos históricos. Los resultados obtenidos se presentarán a continuación y se podrá evidenciar la cantidad de alertas que se hubieran podido reportar si estos análisis se hicieran de forma temprana a medida que llegaran los contratos. Además, se puede ver la cantidad de dinero que está en juego por no alertar estos contratos. Así que, estos resultados pueden evidenciar que estos indicadores podrán jugar un papel importante a la hora de buscar posibles irregularidades en la contratación pública soportada por la arquitectura construida. Para cada uno de los indicadores se presentará las fuentes de información involucradas, así como sus resultados.

5.3.1 Indicadores por inhabilidad

Acá se presentarán los resultados para la categoría de los indicadores por inhabilidad.

5.3.1.1 Inhabilitados por multas

El sistema de información del SECOP II, posee una fuente de datos que permite almacenar todos aquellos proveedores y sus respectivos procesos que han sido multados y sancionadospor cualquier motivo, a esta fuente se le conoce como *SECOP II - Multas y Sanciones*, y adicional, teniendo la fuente *SECOP II - Procesos de Contratación* donde se tienen todos loscontratos históricos de la plataforma, cruzamos estas dos fuentes de información para obtenertodos aquellos contratos que tienen un contratista con algún tipo de sanción o multa, para la ejecución a corte de 28 de febrero de 2023 existen 129 proveedores que han tenido alguna multa y que han tenido algún contrato adjudicado, estos proveedores tienen un total de 3339 contratos asignados, con un monto total de valor adjudicado de COP$1.394.242.081.966, más de 1 billón de pesos en contratos en más de 8 años de información.

Ilustración 4. Resultados obtenidos en un notebook para la ejecución del indicador inhabilitados por multas

5.3.1.2 Inhabilitados por obras inconclusas

En el sistema de información de PACO, se tiene el registro de todas las obras que no

fueron terminadas satisfactoriamente, lo cual se determinan como obras inconclusas *PACO - Registro nacional de obras inconclusas*, allí se tiene el histórico de dichas obras, dicha base fue creada por medio de la ley 2020 de 2020 y tiene como fin registrar las obras civiles inconclusas en el territorio colombiano, lo que se busca con este indicador es encontrar todos aquellos contratos y proveedores que han dejado alguna obra inconclusas y que tienen contratos con el estado cruzando contra la fuente *SECOP II - Procesos de Contratación* donde se tienen todos los contratos históricos de la plataforma. Así que los resultados evidencian lo siguiente, a corte del 28 de febrero de 2023, se tienen 38 proveedores que han tenido por lo menos un contrato inconcluso y que en la actualidad tienen algún contrato registrado en el SECOP II, en esos mismos resultados, se puede evidenciar que existen 666 contratos que están asignados a algún proveedor que ha tenido alguna obra inconclusas y todos estos contratos tienen un valor de COP$1.079.928.813.090, más de 1 billón de pesos en contratos en más de 8 años de información.

Ilustración 5. Resultados obtenidos en un notebook para la ejecución del indicado inhabilitados por obras inconclusas.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Indicadores por Inhabilidad	Inhabilitados por obras inconclusas	38	666	1079928813090.000	666	2023-02-28

5.3.1.3 Inhabilitados por responsabilidad fiscal

PACO posee una fuente de datos llamada *PACO - Responsabilidades fiscales*, esto gracias al trabajo de la Contraloría General de la República publica ya que trimestralmente actualiza según el Boletín de responsables Fiscales, que contiene un listado de las personas naturales o jurídicas que cuentan con procesos de responsabilidad fiscal en la Contraloría General de la República y las Contralorías departamentales, municipales o distritales. A partir de ello y cruzando contra *SECOP II - Procesos de Contratación,* el cual posee el registro histórico detodos los contratos registrados en la plataforma de SECOP II y a corte del 28 de febrero de 2023, existe un único proveedor que está en la lista de responsabilidad fiscal de la Contraloríay que tiene contratos registrados en el SECOP II, ese único proveedor tiene 4 contratos que tienen un valor total de COP$ 395.440.841, más de 390 millones de pesos en contratos en más de 8 años de información.

Ilustración 6. Resultados obtenidos en un notebook para la ejecución del indicador inhabilitados por responsabilidad fiscal.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Indicadores por Inhabilidad	Inhabilitados por responsabilidad fiscal	1	4	395440841.000	2548857	2023-02-28

5.3.2 Indicadores de incumplimiento

Acá se presentarán los resultados para la categoría de los indicadores por incumplimiento.

5.3.2.1 Contratistas con contratos cancelados

La fuente de *SECOP II - Contratos Cancelados* contiene todos los contratos dentro del sistema del SECOP II que han sido cancelados, de allí se toman todos aquellos proveedores que aparecen por lo menos una vez, posteriormente se cruza contra la fuente *SECOP II - Procesos de Contratación*, y poder determinar todos aquellos contratos que están activos dentro de la plataforma y que han sido asignados a alguno de estos proveedores con contratoscancelados. Para los datos a corte del 28 de febrero de 2023, existen 2.848 contratos que tienen asignado a alguno de estos proveedores, en total existen 360 proveedores diferentes, con un monto total de valor adjudicado de COP $80.996.426.975, más de 80 mil millones depesos en contratos en más de 8 años de información.

Ilustración 7. Resultados obtenidos en un notebook para la ejecución del indicador contratistas con contratos cancelados.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Indicadores incumplimiento	Contratistas con contratos cancelados	360	2848	80996426975.000	3487657	2023-02-28

5.3.2.2 Contratos con incumplimiento de entregas

Es importante poder detectar aquellos contratos que han tenido entregas incumplidas, en la plataforma del SECOP II se encuentra la fuente de *SECOP II - Ejecución Contratos* la cual posee toda información relacionada con la ejecución de los contratos firmados en el SECOPII, en términos de avance estimado y real. A partir de esta fuente y con datos a corte del 28 de febrero de 2023, se ha podido identificar 115 irregularidades, es decir casos donde la cantidad de elementos entregada es menor a la cantidad de elementos a recibir, eso traduce un incumplimiento en los entregables de 21 contratos, esto demuestra que cada contrato ha tenido más de un incumplimiento en los entregables, estos contratos suman un valoradjudicado de COP $11.762.172.817, más de 10 mil millones de pesos en contratos en más de 8 años de información.

Ilustración 8. Resultados obtenidos en un notebook para la ejecución del indicador contratos con incumplimiento de entregas.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Indicadores incumplimiento	Contratos con incumplimiento de entregas	115	21	11762172817.000	3649657	2023-02-28

5.3.2.3 Otros indicadores

Acá se presentarán los resultados de la última categoría, denominada otros indicadores.

5.3.2.4 Abuso de la contratación.

Al validar la fuente del *SECOP II - Procesos de Contratación*, donde se encuentran todos los procesos de contratación sin importar su estado de adjudicación, se pueden evidenciar que con datos a corte del 28 de febrero de 2023 existen *969* irregularidades en toda la base, esto quiere decir que hay 969 entidades que han contratado más de 3 veces al mismo proveedor dentro del mismo año, si se traduce eso en cantidad de contratos que presentan esta casuística tenemos un total de 4198, con una monto total de valor adjudicado de COP
$4.648.907.222.722, más de 4 billones de pesos en contratos en más de 8 años de información.

Ilustración 9. Resultados obtenidos en un notebook para la ejecución del indicador de abuso de la contratación.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Otros Indicadores	Abuso de la contratación	969	4198	4648907222722.000	2843657	2023-02-28

5.3.2.5 Ofertas costosas

Al validar la fuente del *SECOP II - Procesos de Contratación* contra la fuente *SECOP II - Ofertas Por Proceso,* donde se encuentran todas las ofertas que se han realizado para cada uno de los contratos (dependiendo del tipo de contrato aplica o no aplica esta modalidad de contratación con diversos oferentes), se pueden evidenciar que con datos a corte del 28 de febrero de 2023 existen *48840* contratos donde la oferta con el valor más bajo no fue seleccionada, teniendo en cuenta que si varias ofertas tienen el mismo valor, se debería tomar la primer oferta presentada por ese valor, como en este caso, la casuística está a nivel de contratos, la *cantidad_contratos_irregularidades* será el mismo valor, estos casos de no tomar la oferta más económica suman un total de COP $31.428.063.358.012, más de 31 billones de pesos en contratos en más de 8 años de información, esta validación se ha tomadode 1.548.624 contratos que tienen registros en las ofertas por proceso.

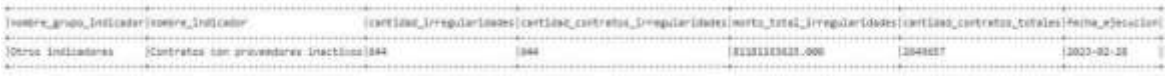
Ilustración 10. Resultados obtenidos en un notebook para la ejecución del indicador de ofertas costosa.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Otros Indicadores	Ofertas Costosas	48840	48840	31428063358012.000	1548624	2023-02-28

5.3.2.6 Contratos de proveedores inactivos

Al validar la fuente del *SECOP II - Procesos de Contratación* contra la fuente registrada porla Cámara de Comercio *Otros - Personas Naturales, Personas Jurídicas y Entidades Sin Ánimo de Lucro,* de allí se buscan todos los contratos activos que tengan el proveedor adjudicado con su registro en Cámara de Comercio en estado de Cancelado, según la información a corte del 28 de febrero de 2023, actualmente existen 644 proveedoresregistrados en la Cámara de Comercio que cruzan contra los contratos del SECOP II que tiene el registro Cancelado en la Cámara de comercio, cada proveedor tiene asociado un único contrato, puesto que la *cantidad_contratos_irregulares* es la misma, estos contratos suman un total de COP $81.181.163.625, más de 81 mil de pesos en contratos en más de 8 años de información.

Ilustración 11. Resultados obtenidos en un notebook para la ejecución del indicador de contratos de proveedores inactivos.



nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Otros indicadores	Contratos con proveedores inactivos	644	644	81181163625.000	2049657	2023-02-28

5.3.2.7 Contratos de proveedores PEP

Al validar la fuente del *SECOP II - Procesos de Contratación* contra la fuente que contiene el listado de las Personas Expuestas Políticamente (PEP) registradas por las diferentes entidades y organismos del Estado en el Sistema de Información y Gestión del Empleo Público (SIGEP) con datos a corte del 28 de febrero de 2023 se puede evidenciar que existen12 contratos asignado a 12 personas que han sido catalogadas como personas expuestas políticamente, esto afecta los principios de transparencia puesto que puede generar un tráficode influencias para ganar la contratación pública, estos 12 contratos suman un valor de COP $485.054.304, más de 485 millones de pesos en contratos en más de 8 años de información.

Ilustración 12. Resultados obtenidos en un notebook para la ejecución del indicador de contratos de proveedores PEP.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Otros indicadores	Contratos con proveedores PEP	12	12	485054304.000	2049657	2023-02-28

5.3.2.8 Contratos de proveedores con puestos sensibles

Al validar la fuente del *SECOP II - Procesos de Contratación* contra la fuente que contiene el listado de servidores públicos que tienen una vinculación activa en el Sistema de Información y Gestión del Empleo Público – SIGEP con datos a corte del 28 de febrero de 2023, se puede evidenciar que existen 143 contratos asignado a 115 personas que han sido catalogadas como personas con puestos sensibles, esto afecta los principios de transparenciapuesto que puede generar un tráfico de influencias para

ganar la contratación pública, estos 143 contratos suman un valor de COP $4.637.197.176, más de 4 mil millones de pesos en contratos en más de 8 años de información.

Ilustración 13. Resultados obtenidos en un notebook para la ejecución del indicador de contratos de proveedores con puestos sensibles.

nombre_grupo_indicador	nombre_indicador	cantidad_irregularidades	cantidad_contratos_irregularidades	monto_total_irregularidades	cantidad_contratos_totales	fecha_ejecucion
Otros Indicadores	Contratos con proveedores con puestos sensibles	115	143	4637197176.000	2646657	2023-02-28

5.4 Arquitectura propuesta

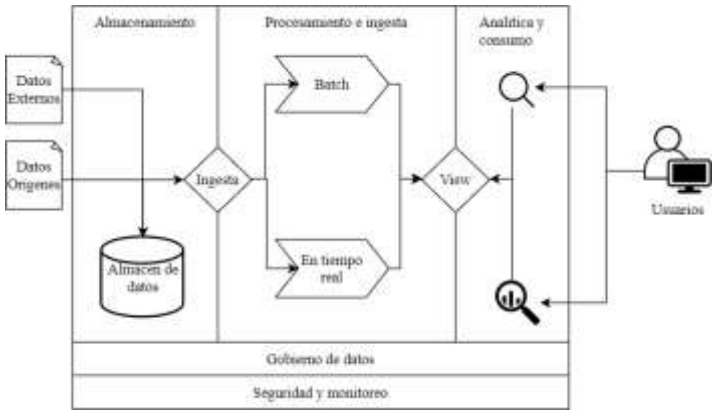
En esta sección se presentará la arquitectura propuesta para soportar el caso de uso. Se dará una breve explicación de cada uno de los componentes que se han utilizado para un mejor entendimiento.

5.4.1 Arquitectura de referencia

Basándose en el análisis de la revisión de las arquitecturas Big Data del marco teórico, planteamos la arquitectura de referencia de forma muy general para que pueda adaptarse de acuerdo al contexto del lector o interesado y pueda implementar diferentes herramientas y tecnologías On Premise o Cloud para su construcción. Este será el punto de partida, puesto que define las capas necesarias para soportar las necesidades, que en este caso son el procesamiento por lotes y el procesamiento en tiempo real, así como toda la parte de visualización de la información. La arquitectura está conformada por las capas de **datos fuente, almacenamiento, procesamiento, análisis y consumo, gobierno de datos, seguridad y monitoreo.**

A continuación, se explica en qué consiste cada capa, cómo se enfoca cada una en el presentecaso de uso y contexto del proyecto, y además, mencionamos las tecnologías que pueden serapropiadas para aprovisionar la infraestructura de cada una.

Ilustración 14. Arquitectura de referencia propuesta.



5.4.1.1 Datos fuente

En nuestro proyecto los datos fuente se encuentran en el sistema de compra pública SECOP I, SECOP II y TVEC (ANCP - CCE, 2019) y de datos abiertos. Los datos de estas plataformasse encuentran en distintos formatos como fotografías, PDF, archivos tipo Json, CSV, TXT, entre otros. Estas plataformas contienen todos los datos disponibles de compras públicas en Colombia. Los datos fuentes no son considerados una capa como tal de la arquitectura, pero es el insumo principal para el flujo de la arquitectura del proyecto. Así que todo caso de uso tendrá sus propias fuentes de datos, dependiendo de la información que se quiera cargar dentro del sistema.

5.4.1.2 Almacenamiento

El almacenamiento es una de las capas más importantes, aquí residirán los datos que vamos a utilizar en el proyecto, estos han sido ingestados desde los datos fuente origen. En la primera etapa se deberán almacenar los datos que obtengamos de las diferentes fuentes luego de realizar la ingesta o también conocido ETL, extracción, transformación y carga de datos. Se almacenan también todos los resultados procesados de los indicadores de riesgo de corrupción para la contratación pública en Colombia, estos procesamientos se desarrollarán de acuerdo a una lógica específica que analizará la información histórica de los datos fuente con tal de generar estos indicadores de posible corrupción.

Para la capa de almacenamiento existen soluciones Cloud y On Premise, pero una buena alternativa es el almacenamiento en la nube, ya que es una forma económica, de fácil acceso, mantenimiento y manejo. Existen diferentes tipos de almacenamiento: Bloque, objetos y archivos (Redhat, 2018), y por cada proveedor de la nube existen diferentes soluciones como *Amazon S3, Azure Blob Storage, Google cloud storage*, entre otros (Pletcher, 2022), (EdPrice-MSFT, s.f.).

También existen arquitecturas de almacenamiento de datos especializadas para Big Data como los *Data Warehouse y Data Lakes*. La nube nos permite almacenar datos en múltiplesservicios sin servidores, de forma rentable y altamente escalables, alguno de estos servicios son *BigQuery, Amazon Athena, Azure Synapse Analytics* (Google Cloud, s.f.).

5.4.1.3 Procesamiento

La capa de procesamiento es fundamental, ya que con ella se pueden hacer los análisis de datos correspondientes para las diferentes visualizaciones y las posibles alertas de fraude en los contratos ingestados de acuerdo a la historia existente de la contratación pública.

Esta capa tiene dos tipos de procesamiento de información: **procesamiento por lote o batch** y **procesamiento en tiempo real**. El componente de procesamiento en tiempo real es considerado en el proyecto para los insumos de datos que son entregados por los usuarios de la plataforma para generar visualizaciones, métricas e indicadores de fraude inmediatamente. Una de las herramientas más utilizadas en esta capa es el framework de *Spark*, es un motor multilenguaje para ejecutar ingeniería de datos, nos permite unificar el procesamiento batch y en tiempo real (Apache, s.f.), en la nube existen servicios de procesamiento de estadísticas y datos de código abierto que usan *Spark* o *Hadoop* con mayor eficiencia y seguridad, como *Azure HDInsight, Google Dataproc, Amazon Elastic MapReduce y AWS Batch*. Por el lado de las ingestas o ETL existen servicios como *Cloud Data Fusion, Amazon AppFlow, AWS Glue y Azure Data Factory* entre otros (Google Cloud, s.f.).

5.4.1.4 Análisis y consumo

En esta capa los usuarios de la plataforma pueden interactuar para visualizar métricas y alertas de posibles casos de fraudes con respecto a la información de la contratación pública almacenada en la arquitectura, esto mediante las salidas generadas por los procesamientos por lotes y en tiempo real, los cuales analizaron el histórico almacenado y otras características para identificar estas casuísticas considerables.

Para la visualización de la información existen muchas herramientas en el mercado tanto pagas como open source, entre ellas tenemos a *Tableau, MicroStrategy, Pentaho, Jaspersof, QuickSight* y muchas otras más, la selección puede depender de dos factores de comparación como el rendimiento y el costo (Elizondo & Joe, 2022).

Como algo adicional y que es importante mencionar, aunque no haga parte de los objetivos de nuestro proyecto es que en esta capa también podemos adicionar componentes de IA los cuales nos ayudan a tener entornos para entrenar modelos de aprendizaje automático a gran escala en la nube y hacer predicciones sobre datos nuevos, algunos de estos servicios son *Amazon SageMaker, Azure AI Platform y Vertex AI*. Adicionalmente existe un modelo de Big Data Query como servicio en la nube que nos ayuda a procesar y administrar consultas de petabytes de información mediante SQL, con servicios como BigQuery, Amazon Redshift Spectrum y Azure Synapse Analytics (Google Cloud, s.f.).

5.4.1.5 Gobiernos de datos

Esta capa no es una de las más mencionadas en las arquitecturas de referencia de Big Data, pero si bien es importante para la unificación de nombres dentro de la

arquitectura, esta capasirve *"para gestionar los datos durante su ciclo de vida, desde la adquisición hasta la eliminación, pasando por el uso"* (Google Cloud, s.f.). Podemos concluir que esto se basa en la implantación de estándares o políticas internas para regular el dato desde la recopilación, almacenamiento, procesamiento y eliminación de este si es el caso.

En la nube existen servicios que nos ayudan con la gestión de gobierno de datos como *AWS Glue Data Catalog*, *Data Catalog*, *Azure Purview*, *Azure Data Explorer* (Google Cloud, s.f.).

5.4.1.6 Seguridad y monitoreo

Esta capa, aunque no es una de las primordiales para una arquitectura Big Data es muy importante, ya que con ella podemos monitorear los diferentes componentes de la arquitectura, ver su comportamiento a través del tiempo, el uso de recursos y optimizar costoscuando hablamos de tecnología Cloud.

En cuanto a las herramientas de monitoreo y herramientas para supervisar, controlar y optimizar los costos en Cloud, podemos hablar de servicios como Management Google, AWS Cost explorer y Azure Cost Management, y para el monitoreo del rendimiento, disponibilidad y estado de la infraestructura encontramos servicios como Cloud Monitoring,Amazon CloudWatch y Azure Monitor (Google Cloud, s.f.).

5.4.1.7 Usuarios

Los usuarios no son una capa como tal de la arquitectura, pero son los stakeholders de los datos, en el proyecto los usuarios son los ciudadanos o personas, entidades públicas, periodistas o cualquier interesado en visualizar el comportamiento de la contratación públicadel país.

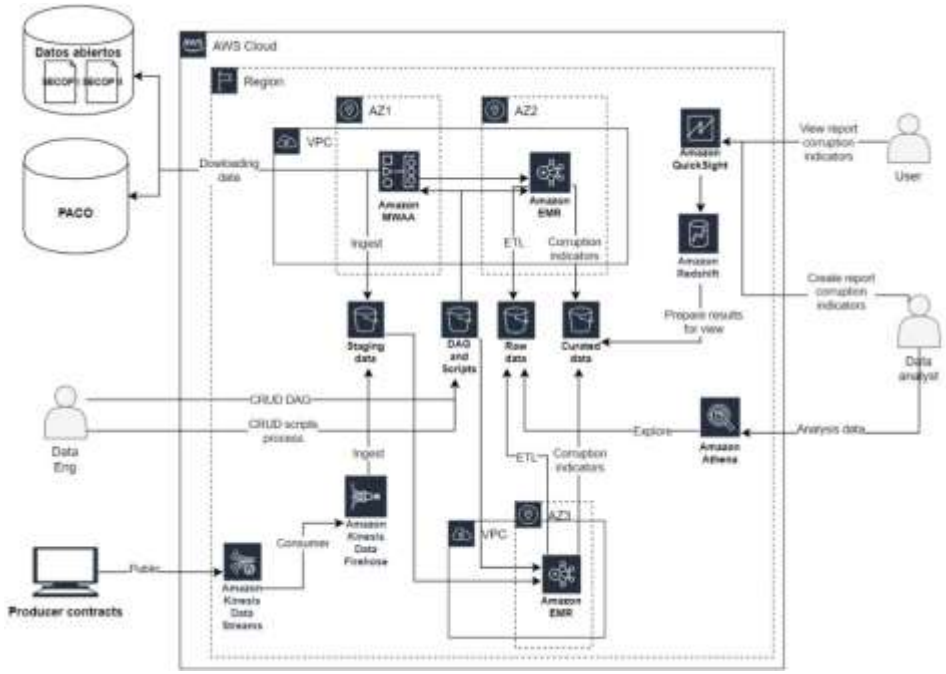
5.4.2 Arquitectura propuesta en AWS

Teniendo en cuenta la arquitectura de referencia propuesta, el análisis de cada capa y teniendodecidido que la implementación de la infraestructura se realizará en la nube por sus ventajasmencionadas en el marco teórico en el apartado Cloud Infrastructure, entonces, el siguiente paso será seleccionar el proveedor de la nube con el cual se desarrollara la infraestructura dela arquitectura y debido a que se cuentan con algunos conocimientos previos en AWS, sumado a que este proveedor es la compañía líder del mercado cloud (Carter, 2022) se ha decidido desplegar la arquitectura con este proveedor, que tiene mucha fuerza en el mercado y una inigualable cantidad de servicios que ayudarán a desarrollar con éxito el proyecto.

Esta arquitectura se definió sobre el proveedor de la nube de AWS, aquí se menciona

como funciona cada componente y a que capa corresponde frente a la arquitectura de referencia propuesta, a continuación, se presenta el diagrama de arquitectura con todos los componentes que interactúan para cumplir a cabalidad con el caso de uso propuesto:

Ilustración 15. Arquitectura big data construida en AWS para alertar las posibles irregularidades en la contratación del gasto público (Tiempo real y por lotes).



5.4.2.1 Almacenamiento

5.4.2.1.1 Amazon S3

Amazon Simple Storage Service (Amazon S3) es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento líderes en el sector, este ha sido el servicio elegido para el almacenamiento de todos nuestros datos.

5.4.2.1.2 Ingesta, procesamiento por lotes y tiempo real

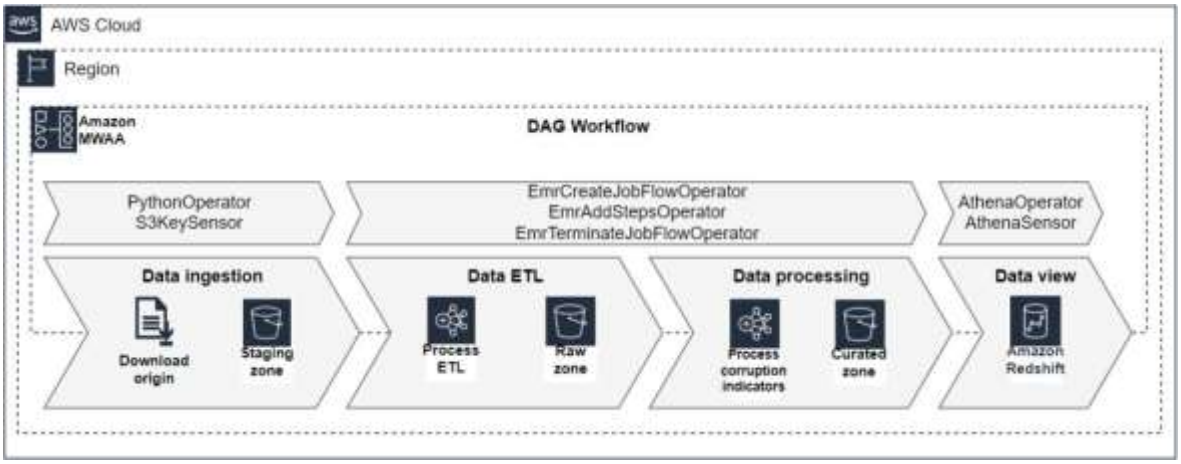
Esta sección es fundamental para conocer cómo funcionan las ingestas de la información (es decir cómo pasamos de los datos orígenes a la información en formato optimizado y con datos limpios), así mismo se explicarán las herramientas utilizadas para el procesamiento por lotes y en tiempo real de los indicadores de presunta corrupción.

5.4.2.1.3 Ingesta y procesamiento por lotes

En esta sección se encontrará los componentes utilizados para la ingesta de los datos y el procesamiento por lotes, hay que tener en cuenta que tanto el procesamiento en

tiempo real como el procesamiento por lote son necesidades diferentes, así que tiene herramientas diferentes. A continuación, se presenta el workflow de los datos.

Ilustración 16. Workflow de datos en Apache Airflow.



5.4.2.1.3.1 Amazon Managed Workflows Apache Airflow (MWAA)

Es un servicio que realiza la orquestación de flujos de trabajo administrados, seguros y de alta disponibilidad para Apache Airflow, organiza sus flujos de trabajo mediante gráficos acíclicos dirigidos (DAG) escritos en Python.

5.4.2.1.3.2 Apache Spark

Es un motor de análisis unificado para el procesamiento de datos a gran escala. Proporciona API de alto nivel en Java, Scala, Python y R, y un motor optimizado que admite gráficos de ejecución general. El ecosistema de Spark incluye cinco componentes clave: (1) Spark Core es un motor distribuido de uso general para procesar datos. (2) Spark SQL es el módulo de Spark que permite utilizar datos estructurados. (3) Spark Streaming facilita la creación de soluciones de streaming escalables y tolerantes a fallos. (4) MLlib es la biblioteca escalable de aprendizaje automático de Spark. (5) GraphX es la API de Spark para grafos y computación en paralelo de grafos.

5.4.2.1.3.3 Amazon EMR

Es la solución destinada al procesamiento de datos a escala de petabytes, análisis interactivo y aprendizaje automático mediante el uso de marcos de código abierto, como Apache Spark, Apache Hive y Presto. Este servicio es una plataforma de clúster administrada que simplifica la ejecución de los marcos de trabajo de Big Data.

5.4.2.1.4 Ingesta y procesamiento en tiempo real

En esta sección se encontrará los componentes utilizados para la ingesta de los datos y el procesamiento en tiempo real, esta parte es fundamental para este caso de uso, puesto que uno de los objetivos es poder emitir alertas tempranas para aquellos contratos que cumplan algunos indicadores de alerta generados.

5.4.2.1.4.1 Amazon Kinesis Data Streams

Es un servicio de datos de streaming sin servidor que hace que sea fácil capturar, procesar y almacenar flujos de datos a cualquier escala. Este servicio es igual que *Apache Kafka*, una solución de mensajería de publicación y suscripción (pub/sub). Sin embargo, se ofrece como un servicio administrado en la nube de *AWS* y, a diferencia de *Kafka*, no se puede ejecutar enlas instalaciones.

5.4.2.1.4.2 Amazon Kinesis Data Firehose

Es un servicio de extracción, transformación y carga (ETL) que captura, transforma y entrega de manera fiable datos de streaming en lagos y almacenes de datos y servicios de análisis.

5.4.2.1.4.3 Spark Streaming

Es una extensión de la API central de Spark que permite el procesamiento de flujos de datosen vivo escalable, de alto rendimiento y tolerante a fallas.

5.4.2.2 Analítica y consumo

Esta sección explicará las herramientas que se han utilizado para la parte de visualización de la información, esta parte es muy importante debido a que se trata de la zona donde se podrán ver aquellos contratos que están llegando en tiempo real y se podrá visualizar si tienen algún indicador que esté generando riesgos. Así mismo se podrá visualizar el análisis a los contratoshistóricos.

5.4.2.3 Amazon Athena

Es un servicio de análisis de consultas interactivo y sin servidor creado en marcos de código abierto, lo que lo hace compatible con formatos abiertos de archivos y tablas. *Amazon Athena* también facilita la ejecución interactiva de análisis de datos mediante *Apache Spark* sin tenerque planificar, configurar y administrar los recursos.

5.4.2.4 Amazon Redshift

Este servicio utiliza SQL para analizar datos estructurados y semiestructurados en almacenamientos de datos, bases de datos operativas y lagos de datos.

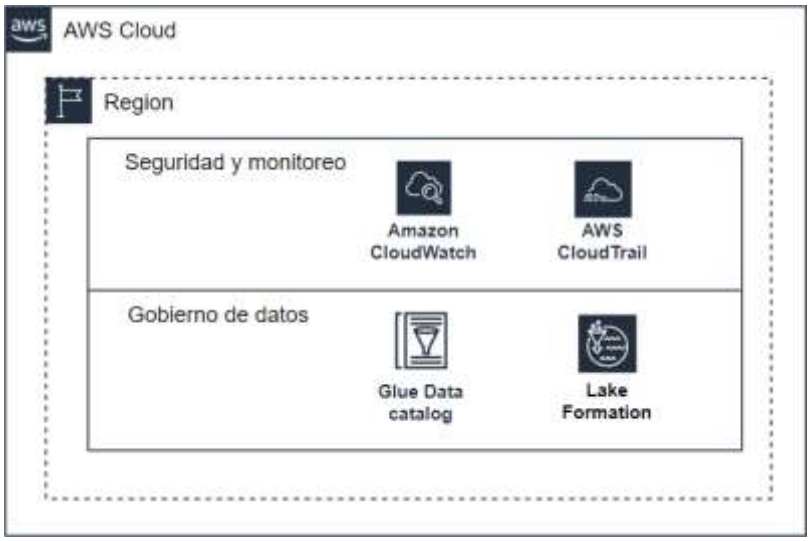
5.4.2.5 Amazon QuickSight

El servicio de inteligencia empresarial (BI) sin servidor permite que todos los miembros de su organización comprendan sus datos mediante preguntas en lenguaje natural, la exploración a través de paneles interactivos o la búsqueda automática de patrones y valores atípicos impulsada por Machine Learning.

5.4.2.6 Gobierno de datos

Como se ha nombrado en todo el documento una parte importante del caso de uso es el gobierno del dato, puesto que se unifican los nombres en todo el ciclo de vida del dato, lo que ayuda a tener un mejor entendimiento de toda la información. A continuación, se presentan las herramientas y servicios utilizados para la construcción de esta arquitectura.

Ilustración 17. Arquitectura de seguridad, monitoreo y gobierno de datos.



5.4.2.6.1 Lake Fomation

Es un servicio que facilita la configuración de un lago de datos seguro en cuestión de días. Un lago de datos es un repositorio centralizado, seleccionado y seguro que almacena todos sus datos, tanto en su forma original como preparados para análisis. Un lago de datos le permite desglosar los silos de datos y combinar diferentes tipos de análisis para obtener información y tomar mejores decisiones empresariales.

5.4.2.6.2 Glue Data Catalog

Es su almacén de metadatos técnicos persistentes en la nube de AWS. El catálogo de datos de *AWS Glue* proporciona un repositorio uniforme donde los sistemas dispares pueden almacenar y encontrar metadatos para realizar un seguimiento de los datos en los silos de datos.

5.4.2.7 Seguridad y monitoreo

En la Ilustración 17 se presentan las herramientas que se utilizarán en la capa de seguridad y monitoreo. A continuación, se presenta una breve explicación de los servicios utilizados en la implementación de esta arquitectura.

5.4.2.7.1 AWS CloudTrail

Monitorea y registra la actividad de la cuenta en toda la infraestructura de AWS, lo que le permite controlar las acciones de almacenamiento, análisis y reparación.

5.4.2.7.2 Amazon CloudWatch

Es un servicio de monitoreo y observabilidad creado por desarrolladores, ingenieros de fiabilidad de sitios (SRE), administradores de TI, propietarios de productos e ingenieros de DevOps.

5.4.2.7.3 AWS Identity and Access Management (IAM)

Administra de manera segura las identidades y el acceso a los recursos y servicios de AWS, con este servicio puede especificar quién o qué puede acceder a los servicios y recursos en AWS.

Estos dos servicios nos ayudan a monitorear el uso de los servicios de la arquitectura en general y de forma específica los servicios de IAM nos ayudan a controlar la seguridad de lacuenta y de los servicios desplegados.

5.4.3 Funcionamiento de la arquitectura propuesta en AWS

Iniciando el flujo de la arquitectura, se realiza la **ingesta de datos desde fuentes externas**. Primero, se toman las bases de datos de PACO y datos abiertos.gov.co, que contienen información sobre multas, puestos sensibles, obras inconclusas y datos de las plataformas SECOP I y SECOP II. Estos datos se cargan en la **capa Staging**, que representa un diseño de capas por buckets en el servicio de ***Amazon S3*** donde se almacenan sin ninguna transformación inicial.

Luego, los datos se trasladan a la **capa Raw**, otra capa de buckets en S3, donde se someten a algunas transformaciones y limpieza de datos. Aquí se realiza la **preparación inicial** para facilitar el posterior procesamiento como se muestra en Ilustración 15 y a mayor detalle en Ilustración 16.

Una vez preparados los datos, se inicia el **procesamiento en tiempo real**. Un "Producer" simula contratos públicos y envía continuamente datos a ***Kinesis Streams***, mientras que ***Kinesis Firehose*** actúa como el "Consumer" para procesar los datos en tiempo real y almacenarlos en la capa Staging, que corresponde a un buckets S3, tal como se muestra en Ilustración 15.

En paralelo, se gestiona el **procesamiento por lotes** utilizando ***Amazon MWAA***. Este servicio coordina y ejecuta en orden el despliegue de los servicios necesarios sólo si se cumplen las tareas predecesoras tal como se muestra en Ilustración 16, para gestionar el ciclo de vida del dato desde su origen hasta su uso posterior. El flujo de datos por lotes se programa para ejecutarse una vez por semana, todos los domingos (aunque es configurable según necesidades). Los datos almacenados en la **capa Staging** y **capa Curated** son utilizados como insumo de los clústeres de ***Amazon EMR*** para ejecutar los trabajos de Spark sobre los archivos que contienen la lógica de código de "Data ETL" y

"Data processing". Luego los resultados generados por el procesamiento de indicadores de corrupción se almacenan en la **capa Raw** para la limpieza y el trabajo de Data ETL, **capa Curated** para las agregaciones finales y el trabajo sobre Data processing.

Al finalizar, se crea una tabla final de resultados en ***Redshift*** para consultas a través de servicios de Business Intelligence como ***QuickSight*** o Power BI tal como se muestra en la Ilustración 15 y en el flujo de datos dela Ilustración 16 donde se aprovisiona el servicio de ***Redshift*** luego de tener los resultados del proceso de indicadores de presunta corrupción almacenados en la **capa *Curated***. La **capa Curated**, además de almacenar los resultados de los procesamientos de indicadores propuestos, también almacena el código requerido para ejecutar el workflow de datos (DAG), los ETL y los procesamientos finales.

Además, se lleva a cabo la **gestión de datos históricos**. Los datos almacenados en tiempo real se convierten en históricos si tienen una persistencia mayor a siete días desde la fecha en que se ejecuta el **procesamiento por lotes**.

En cuanto a la gobernanza y auditoría de datos referenciada en la Ilustración 17, el **catálogo de datos**, junto con ***CloudTrail*** y ***Lake Formation,*** servicios que no son necesarios desplegar de alguna forma, sino sólo utilizarlos directamente mapeando los orígenes de los buckets S3. Estos son esenciales, ya que son transversales a toda la arquitectura referenciada en la Ilustración 15 y brindan capacidades integrales de auditoría, gobernanza, así mismo rastreando cambios de esquema y controlando el acceso a los datos para evitar modificaciones inapropiadas o comparticiones sin autorización.

En conclusión, esta arquitectura garantiza la eficiente gestión y transformación de datos, desde su recolección en fuentes externas hasta su análisis y visualización. La combinación de servicios de AWS, como ***Kinesis***, ***EMR***, ***Redshift***, ***QuickSight*** y ***Athena***, proporciona un flujo de datos sólido y escalable para generar los resultados del caso de uso sobre los indicadores de corrupción y facilitar la toma de decisiones informadas y oportunas. Todo el proceso está respaldado por la seguridad, la gobernanza y la auditoría de datos para asegurar la integridad y confiabilidad de la información utilizada en el proceso.

5.5 Ciclo de desarrollo de la arquitectura propuesta

El ciclo de vida del desarrollo de software es una metodología que permite organizar y gestionar de manera eficiente todas las etapas involucradas en la creación y mantenimiento de un software, proporcionando una estructura para planificar, diseñar, implementar, probar y mantener el software de manera eficiente. En este proyecto se aplicaron estas etapas y se brindara una visión detallada de su implementación, destacando las buenas prácticas y herramientas utilizadas en cada

una de ellas.

Para la realización de este MVP desde el principio se seleccionaron dos metodologías de trabajo, (1) la programación por pares: una técnica en la que dos programadores trabajan juntos en el desarrollo de una funcionalidad, uno de ellos es el encargado de escribir el código, mientras que el otro integrante del equipo le corresponde revisar y sugerir posibles mejoras (Williams & Kessler, 2000). (2) El enfoque ágil: Es un marco de trabajo que promueve la entrega temprana y continua de software funcional, adaptándose a los cambios y priorizando la interacción en este caso con el director del proyecto de grado quién nos daba el aval y retroalimentación constante del entregable final (Beck, y otros, 2001).

Las metodologías mencionadas anteriormente, fomentan la colaboración efectiva, revisión continua, de calidad y retroalimentación constante del código tanto para las etapas de implementación y pruebas de la arquitectura desarrollada como para gestionar el desarrollo y el flujo del proyecto.

5.5.1 Etapas del ciclo de desarrollo de la arquitectura propuesta

De las etapas del ciclo de vida del desarrollo de software para el proyecto, en conjunto con las metodologías de programación por pares y el enfoque ágil, se aplicaron las siguientes fases del ciclo de vida del desarrollo de software.

5.5.1.1 Requisitos y análisis

En esta fase, se identificó y documentó los requisitos del software para el diseño de la arquitectura. Se realizó un análisis de los insumos de datos y documentación existente sobre otras arquitecturas realizadas en cloud y en A*WS Architecture Center.* De acuerdo con esto, se llevó a cabo la construcción diagrama de referencia de arquitectura general donde se especifica cada capa necesaria para cumplir con el entregable final, siendo agnóstico a cualquier tecnología.

Además, se identifica que uno de los requisitos y propósitos de este proyecto, es dejar una arquitectura de referencia para futuros trabajos, puesto que se ha diseñado, si bien pensando en este caso de uso, también paraque otros casos de uso puedan utilizarla y sea sencillo configurarla según los requerimientos.Así pueden enfocarse realmente en la exploración y explotación de grandes cantidades de datos y todo lo relacionado a este ámbito.

5.5.1.2 Diseño

Para el diseño, se define la arquitectura de la solución en la nube, utilizando los servicios y recursos disponibles en la plataforma seleccionada que en este caso fue con el cloud provider de AWS. En esta etapa es necesario utilizar patrones

arquitectónicos como el almacenamiento en nube, la computación *serverless*, temas de redundancia, escalabilidad horizontal y como parte importante la seguridad y el monitoreo de los servicios desplegados, tema fundamental para una arquitectura que soporte el caso de uso y es poco comentado en las arquitecturas investigadas.

Para realizar el diseño se revisaron diferentes herramientas de diagramación de arquitecturas en la nube, como AWS CloudFormation, Diagrams.net, Lucidchart y draw.io y por facilidad en el registro y uso de herramienta se seleccionó la última.

5.5.1.3 Implementación y programación en la nube

Luego del diseño, se inicia con la implementación de la arquitectura en la plataforma de nube de AWS. Utilizando siempre como buena práctica la documentación de los servicios y recursos de infraestructura proporcionados por el mismo proveedor para desarrollar y desplegar la solución planteada.
Adicionalmente, como buena práctica y facilidad para el desarrollo se utilizó *IntelliJ IDEA* como IDE, servicios de administración de configuración, como *AWS CloudFormation*, *SDK* como herramienta de *Infrastructure as Code (IaC)* para automatizar el aprovisionamiento y la configuración de recursos, *GitHub* como herramienta para la gestión de versiones, control de cambios y configuración del código.

5.5.1.4 Pruebas, revisión de código y validación en la nube

En esta etapa, se llevan a cabo pruebas para verificar el correcto funcionamiento de la arquitectura en la nube y del software. La revisión de código en pares permitió que se hicieran las validaciones y mejoras del código realizado y para el correcto funcionamiento de la arquitectura se contempló pruebas de código, seguridad y monitoreo en la nube, con los servicios como *AWS CloudWatch*.

5.5.1.5 Mantenimiento y mejora continua

En esta etapa, se realiza el seguimiento, la gestión de incidentes y las mejoras continuas en la arquitectura en la nube. Esto incluye la resolución de problemas, la aplicación de parches de seguridad, la optimización de recursos y la actualización de versiones.

Para el proyecto hubo una mejora y optimización de código y documentación continua, ya que a medida que se realizaban las pruebas surgían incidencias que debían solucionarse pronto para realizar el entregable final.

Cabe aclarar que el proyecto en mención se realiza en un contexto educativo por ende esta fase del ciclo de vida del desarrollo se termina una vez se ha cumplido con el entregable. Un nuevo alcance del proyecto para trabajos futuros podría contemplar la

implementación de actualizaciones de acuerdo con los servicios y parches de seguridad de AWS para las librerías utilizadas, también se podría recopilar y analizarlas con otros casos de uso para identificar oportunidades de mejora.

5.5.2 Códigos Fuente, manuales y documentación

Se ha buscado explicar y documentar al máximo la arquitectura para que pueda ser utilizaday ajustada de forma sencilla por cualquier persona con una necesidad similar, que sea un punto de partida avanzado para que se puedan dedicar a la explotación de datos y por qué no a continuar con este proyecto. Para hacer uso del código fuente, manuales para el uso e instalación y documentación por favor revisar al siguiente enlace, se ha creado un repositorio público que todos podrán descargar para que sea accedido desde cualquier parte y sea un punto de referencia para los demás interesados. https://github.com/LauraMilenaRB/bigdata-corruption-indicators

5.5.3 Costos asociados a la arquitectura

El manual de costos es una herramienta esencial en la planificación y gestión de arquitecturas en la nube, como las basadas en Amazon Web Services (AWS). Este manual proporciona una guía detallada y estructurada sobre los costos asociados con los diferentes componentes y servicios utilizados en una arquitectura propuesta, permitiendo a los equipos técnicos y de negocio comprender y estimar los gastos financieros asociados con la implementación, operación y escalabilidad de una arquitectura de AWS.

5.5.3.1 Recomendaciones para optimización de costos

Para empezar, hay que dar varias recomendaciones de cómo utilizar la arquitectura y que días ejecutar los procesos en otros casos de uso, el procesamiento batch se ejecuta únicamente cuando se programa el servicio de Apache Airlflow, en particular para este caso programamos el procesamiento por lotes cada domingo ya que los datos orígenes utilizados se actualizaban cada semana. Por esto es recomendable que se actualice según las necesidades del proyecto lo que optimiza los costos de la arquitectura.

También hay que tener en cuenta que la cantidad de nodos que debes ejecutar en un clúster de Spark para analizar una determinada cantidad de datos en GB depende de varios factores, como el tamaño total de los datos, la complejidad de las operaciones y los recursos disponibles en tu infraestructura, sin embargo, la ventaja del proyecto es que Amazon EMR, el servicio de AWS para ejecutar clústeres de Spark, proporciona opciones de escalado automático que te permiten ajustar dinámicamente

el tamaño del clúster en función de las necesidades de procesamiento y los datos analizados. Esto ayuda a optimizar los recursos y los costos utilizados.

Por otro lado, tenemos el procesamiento en tiempo real el cual debe ejecutarse indefinidamente, una recomendación es usar servicios en la nube para este fin que solo se utilice cuando se cumple y recibe los eventos externos de la arquitectura, con esto se garantiza que solo se ejecute cuando se requiere, adicionalmente utilizar servicios como lamba que pueden detener o desplegar los servicios en ejecución en tiempo real cuando se cumplen ciertas condiciones, con esto se pueden reducir los costos y es importante hacer estas validaciones ya que este procesamiento es el que más consume recursos y genera costes en la arquitectura.

5.5.3.2 Manual de costos

Luego de haber definido ciertas recomendaciones, compartiremos un cuadro de costos para cada servicio usados en nuestro caso de uso como referencia en USD con todos los servicios desplegados en la misma región de US East (N. Virginia). Además, cabe recordar que esto es ejecutado para un escenario académico.

Tabla 9. Costos de referencia.

Servicios	**Por mes**	**Configuración**
Amazon Managed Workflows for Apache Airflow	362.03	Seleccione un entorno (Pequeño), Mínimo de trabajadores (1), Máximo de trabajadores (2), Número de programadores (2), Tamaño de almacenamiento de datos (40 GB), Horas con el máximo de trabajadores (6 por mes)
Amazon Redshift	182.52	Nodos (1), Tipo de instancia (dc2.large), Utilización (solo bajo demanda) (730 horas/mes), Estrategia de precios (OnDemand), Almacenamiento de respaldo adicional (0 GB), Análisis de datos (0 TB), Almacenamiento administrado tamaño (1 GB), transferencia de datos a (0 GB)
Amazon QuickSight	26.01	Número de días laborables al mes (30), Capacidad de SPICE en gigabytes (GB)(10), Número de autores (1), Número de lectores (1)
Amazon EMR en tiempo real	181.77	Número de nodos EMR maestros (1), instancia EC2 (m6g.xlarge), Utilización (730 horas/mes) Número de nodos EMR principales (2), instancia EC2 (c4.2xlarge), Utilización (730 horas/mes)

Servicios	Por mes	Configuración
Amazon EC2 en tiempo real	337.26	Tenencia (instancias compartidas), sistema operativo (Linux), carga de trabajo (coherente, número de instancias: 3), instancia EC2 avanzada (m6g.xlarge), estrategia de precios (utilización bajo demanda: 730 horas/mes), habilitar el monitoreo (deshabilitado), DT entrante: no seleccionado (0 TB por mes), DT saliente: no seleccionado (0 TB por mes), DT intrarregional: (0 TB por mes)
Amazon EMR en batch	2.81	Número de nodos EMR maestros (1), instancia EC2 (m6g.xlarge), Utilización (24 horas/mes) Número de nodos EMR principales (2), instancia EC2 (m6g.xlarge), Utilización (24 horas/mes)
Amazon EC2 en batch	11.09	Tenencia (instancias compartidas), sistema operativo (Linux), carga de trabajo (coherente, número de instancias: 2), instancia EC2 avanzada (m6g.xlarge), estrategia de precios (utilización bajo demanda: 24 horas/mes), habilitar el monitoreo (deshabilitado), DT entrante: no seleccionado (0 TB por mes), DT saliente: no seleccionado (0 TB por mes), DT intrarregional: (0 TB por mes)
Amazon Kinesis Data Streams	10.96	Duración de la retención de datos (1 día), número de referencia de registros (5 por minuto), número de aplicaciones del consumidor (0), número máximo de registros (5 por minuto)
Amazon Kinesis Data Firehose	6.57	Tipo de fuente (Direct PUT o Kinesis Data Stream), unidades de registros de datos (número exacto), tamaño de registro (1000 KB), partición dinámica (complemento) (habilitado), conversión de formato de datos (opcional) (deshabilitado), proporción promedio de datos procesados a VPC vs datos ingeridos (1.3), Número de registros para la ingesta de datos (3 por minuto), Objetos de tamaño promedio entregados (64 MB), Procesamiento JQ (opcional) (Habilitado), Promedio de horas de procesamiento esperadas de JQ (6)
Amazon Simple Storage Service (S3)	28.80	Almacenamiento estándar de S3 (1000 GB por mes), solicitudes PUT, COPY, POST, LIST a S3 estándar (1000000), GET, SELECT y todas las demás solicitudes de S3 estándar (2000000), datos devueltos por S3 Select (0 GB por mes), datos escaneados por S3 Select (0 GB por mes)
Amazon Virtual Private Cloud	35.10	Días laborables al mes (22) Número de Gateways NAT (1)

Servicios	Por mes	Configuración
Amazon Athena	61.04	Número de DPU (24), Número total de consultas (1 por día), Cantidad de datos escaneados por consulta (2 GB), Número total de sesiones Spark (2 por semana), Duración (horas) de capacidad activa (0 horas por semana)
AWS CloudTrail	4.50	Unidades de eventos de administración (millones), registros de administración de escritura (1), registros de administración de lectura (1), unidades de eventos de datos (millones), registros de S3 (1), registros de Lambda (1), unidades de eventos de Insight (millones), registros con Insight (1), eventos de administración de escritura (10000 por mes), eventos de administración de lectura (10000 por mes), operaciones de S3 (1 por mes), eventos de datos Lambda (0 por mes), número de eventos de administración de escritura analizados (1 por mes)
Amazon CloudWatch	6.90	Cantidad de métricas (incluye métricas detalladas y personalizadas) (20), GetMetricData: cantidad de métricas solicitadas (1), GetMetricWidgetImage: cantidad de métricas solicitadas (1), cantidad de otras solicitudes de API (1), cantidad de paneles (1), Número de métricas de alarma de resolución estándar (1), número de métricas de alarma de alta resolución (1), número de alarmas compuestas (1)
AWS Lake Formation	6.42	Uso de almacenamiento en un mes (en millones) (1,5), Solicitudes en un mes (en millones) (2), Número de tablas (100), Número de archivos pequeños ingeridos por tabla por día (1), Tamaño de archivos pequeños (menos de 64 MB) (64 MB)
AWS Glue	6.00	Número de Objetos almacenados (0,5 millones por mes), Número de solicitudes de acceso (1 millón por mes)

Una conclusión importante que se puede extraer del cuadro de costos es la diferencia significativa en los precios entre Amazon EMR + Amazon EC2 en tiempo real y Amazon EMR + Amazon EC2 batch. La diferencia de precio entre EMR en tiempo real y EMR en batch se debe principalmente a los requisitos de rendimiento y recursos de cada servicio. Amazon EMR en tiempo real está diseñado para proporcionar una capacidad de respuesta rápida y en tiempo real, para esto necesita que las instancias de EC2 se mantenga en ejecución las 730 horas/mes, lo que requiere más recursos y, por lo tanto, tiene un costo mayor. Por otro lado, Amazon EMR en batch permite una programación más flexible y un procesamiento más lento, por lo que no necesita que las instancias de EC2 se mantengan en ejecución, solamente se ejecutara cuando se programes la tarea lo que reduce los costos en comparación con Amazon EMR + Amazon EC2 en tiempo real.

Otra conclusión sobre por qué el servicio de Apache Airflow en Amazon es costoso es debido a varias razones, 1. Administración y mantenimiento: se encarga de la administración y el mantenimiento de la infraestructura subyacente necesaria para ejecutar Apache Airflow, 2. Escalabilidad y disponibilidad: ofrece la capacidad de escalar automáticamente y mantener la disponibilidad del servicio, 3. Integración con otros servicios de AWS: proporciona integración y compatibilidad nativa con otros servicios de AWS, como Amazon S3, Amazon Redshift y Amazon EMR, 4. Servicio gestionado de AWS: Al ser un servicio gestionado por AWS, se espera que haya un costo adicional por la conveniencia y los beneficios proporcionados, como actualizaciones automáticas, parches de seguridad y soporte técnico, 5. Programación basada en cron: Apache Airflow utiliza la sintaxis de cron para programar la ejecución de tareas y flujos de trabajo. Puedes especificar la frecuencia de ejecución, ya sea por intervalo regular (por ejemplo, cada 5 minutos, cada hora) o en momentos específicos (por ejemplo, todos los días a las 8:00 a. m.).

Es importante tener en cuenta que, si bien el servicio de Apache Airflow en Amazon puede ser más costoso en comparación con una implementación local, ofrece ventajas significativas en términos de administración, escalabilidad y disponibilidad. La elección de utilizar este servicio dependerá de las necesidades y prioridades específicas de cada proyecto, teniendo en cuenta tanto los beneficios como los costos asociados.

Para terminar, hay que recordar que estas cifras son solo estimaciones y los precios pueden variar según la región y la configuración específica. Es importante consultar la página de precios de AWS para obtener información actualizada y precisa sobre los costos de los servicios para las condiciones y configuraciones específicas.

5.6 Resultados de los indicadores en la arquitectura

Ahora si vemos los resultados de estos indicadores en nuestra arquitectura, más específicamente en la capa de visualización de datos por medio de *Amazon QuickSight* tenemos los siguientes cuadros de control, tanto para los indicadores por lotes, como para losindicadores en tiempo real.

Iniciamos con los resultados a los indicadores calculados por lotes, es decir, la validación delos datos históricos que se tienen, según lo validado, se tienen registros desde el año 2015, hasta inicios del año 2023, así que se están validando más de 8 años de información.

Ilustración 18. Resultados de los indicadores históricos visualizados en QuickSight.

Resumen de contratos irregulares por indicador

Grupo de indicador	Nombre de indicador	Cantidad contratos irregulares	Monto total de irregularidades
indicadores incumplimiento	contratistas con contratos cancelados	2,841	81,238,635,735
	contratos con incumplimiento de entregas	21	11,762,172,817
	Subtotal	2,862	93,000,808,552
indicadores por inhabilidad	inhabilitados por multa	3,974	1,451,973,371,137
	inhabilitados por obras inconclusas	665	1,080,692,397,338
	inhabilitados por responsabilidad fiscal	4	395,440,841
	Subtotal	4,643	2,533,061,209,316
otros indicadores	abuso de la contratación	4,179	4,571,605,183,056
	contratos con proveedores con puestos sensibles	143	4,637,197,176
	contratos con proveedores inactivos	644	80,911,307,370
	ofertas Costosas	46,241	29,388,622,370,723
	Subtotal	51,207	34,045,776,058,325
Total		58,712	36,671,838,076,193

Como se puede validar en la Ilustración 18, se generó un tabla dinámica necesaria para observar la cantidad de contratos irregulares y el monto total comprometido de acuerdo al análisis de indicadores propuestos en todo el histórico de la contratación pública, todos los indicadores tienen por lo menos 1 contrato que presenta algún tipo de irregularidad, adicionalla sumatoria del valor de todos los contratos que presentan las irregularidades, esto permite visualizar de forma clara los resultados.

Así mismo, para facilitar la visualización se han creado algunas gráficas, en la Ilustración 19 se pueden visualizar el total de contratos agrupado por el grupo de indicador, la categoría creada para dividir los tipos de irregularidades. La tendencia de contratos con alertas de irregulares se encuentra en el grupo de otros indicadores, esto puede indicar que en trabajo futuro se puede hacer un análisis y clasificar a más detalle este grupo.

Ilustración 19. Total de contratos irregulares por grupo de indicador visualizados en QuickSight

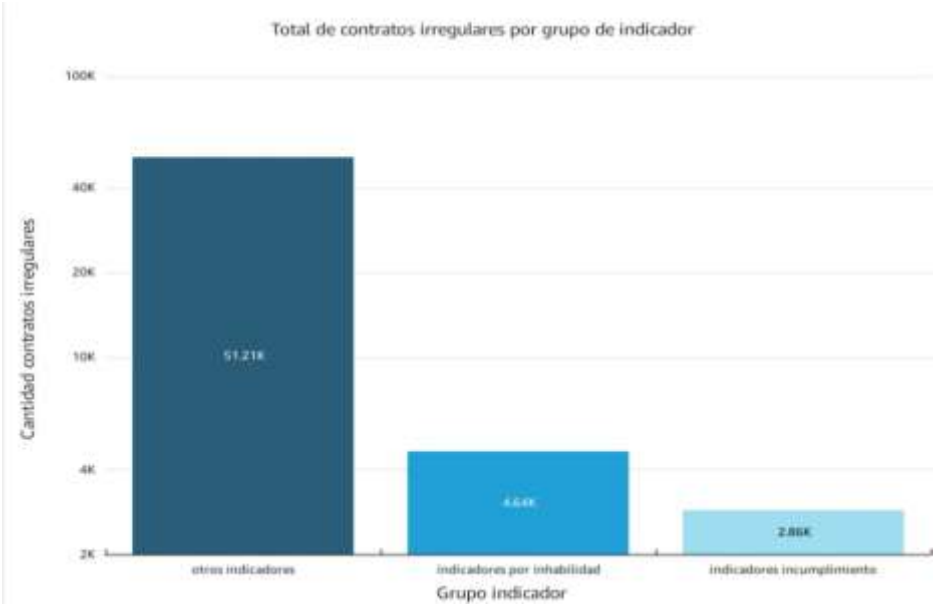
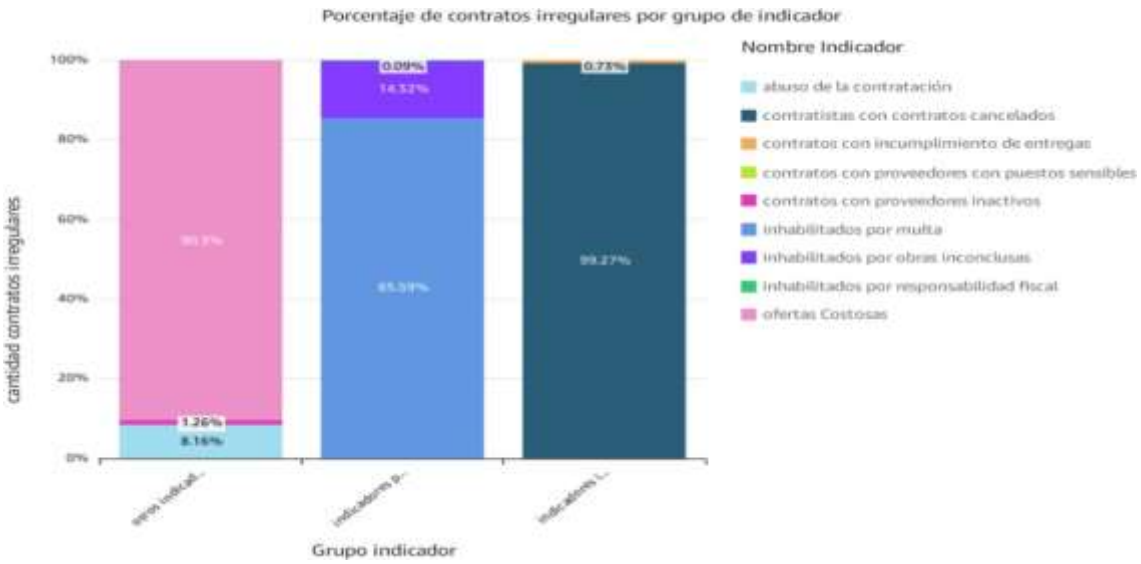


Ilustración 20. Porcentaje de contratos irregulares por indicador agrupado por el grupo de indicadores visualizados en QuickSight.



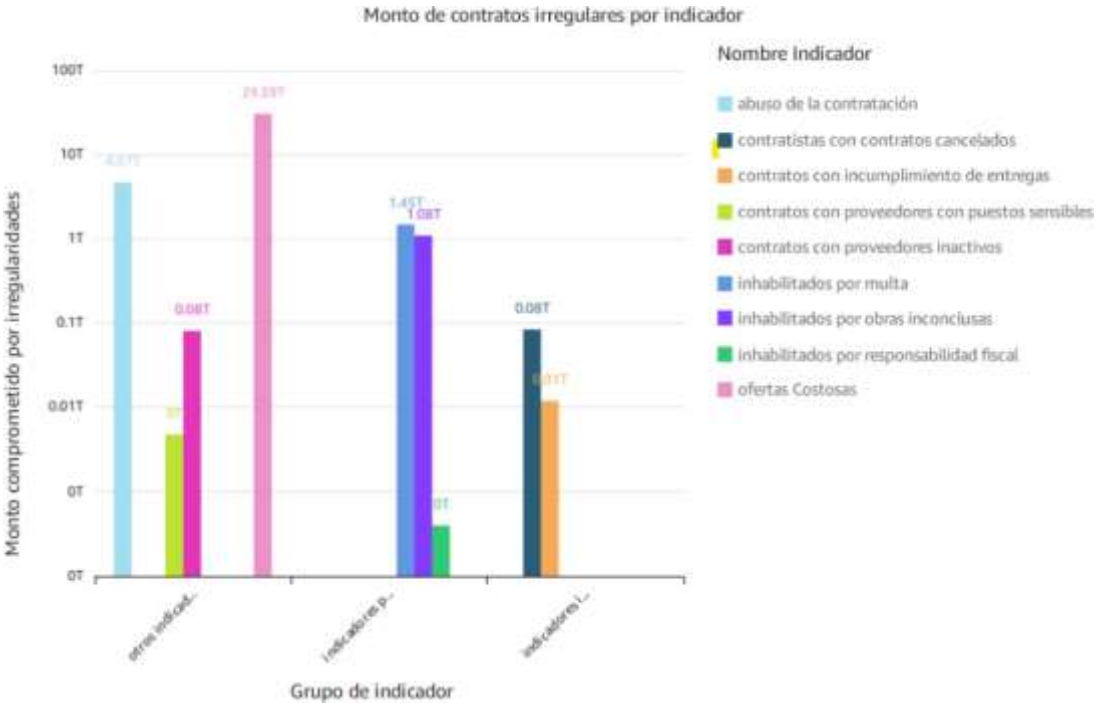
En la Ilustración 20, se tiene los porcentajes de cada indicador sobre su categorización de grupo, esto permite observar cuales son los indicadores que más se repiten. Para los datos analizados se tienen las siguientes conclusiones:

- El 90.3% de irregularidades en el grupo de otros indicadores se presenta con las ofertas costosas esto quiere decir que una gran cantidad de contratos asignados no tienen el precio más económico de todas las propuestas.

- El 85.59% de irregularidades en el grupo de indicadores por inhabilidad se presentan por aquellos proveedores que están inhabilitados por algún tipo de multa, pero así seestán asignando los contratos.

- El 99.27% de irregularidades en el grupo de indicadores por incumplimiento se presentan por contratistas que tienen o han tendido algún contrato cancelado, esto evidencia que así un contratista tenga algún tipo de contrato cancelado, se le están asignando nuevos contratos y puede seguir sucediendo lo mismo.

Ilustración 21. Monto total del valor adjudicado en los contratos irregulares por indicador visualizados en QuickSight.



En la gráfica reportada en la Ilustración 21 se evidencia los valores totales que se han adjudicado en cada indicador que presenta algún tipo de irregularidad detectado, esta gráfica es de gran importancia ya que podemos identificar el indicador con más dinero comprometido en la contratación pública, en este caso es el indicador de ofertas costosas en el grupo de otros indicadores, este presenta un valor cercano a los COP$ 30 billones de pesos. Este indicador en términos de dinero es uno de los más importantes y con mayor peso; y por esto debe ser prioridad en el monitoreo de alertas de irregularidades.

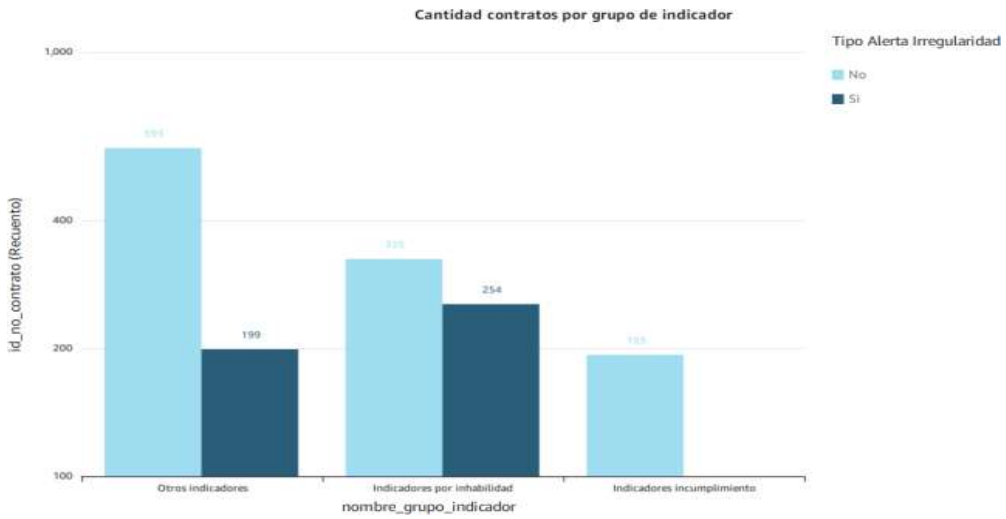
Para la parte de los indicadores en tiempo real, se han creado los siguientes cuadros de control, que permitirán validar en tiempo real si cada contrato que ingresa al sistema presenta algún posible caso de irregularidad y se debe realizar algún tipo de investigación adicional para validar si se trata de algún contrato con posible caso de corrupción.

Ilustración 22. Resultados de los indicadores en tiempo real visualizados en QuickSight.

No ...	Grupo indicador	Nombre de indicador	Alerta irregularidad
1	Indicadores incumplimiento	Contratistas con contratos cancelados	No
1	Indicadores por inhabilidad	Inhabilitados por multa	No
1	Indicadores por inhabilidad	Inhabilitados por obras inconclusas	Si
1	Indicadores por inhabilidad	Inhabilitados por responsabilidad fiscal	Si
1	Otros indicadores	Abuso de la contratación	No
1	Otros indicadores	Contratos con proveedores PEP	No
1	Otros indicadores	Contratos con proveedores con puestos sensibles	No
1	Otros indicadores	Contratos con proveedores inactivos	Si
105	Indicadores incumplimiento	Contratistas con contratos cancelados	No
105	Indicadores por inhabilidad	Inhabilitados por multa	No
105	Indicadores por inhabilidad	Inhabilitados por obras inconclusas	Si
105	Indicadores por inhabilidad	Inhabilitados por responsabilidad fiscal	No
105	Otros indicadores	Abuso de la contratación	No
105	Otros indicadores	Contratos con proveedores PEP	No
105	Otros indicadores	Contratos con proveedores con puestos sensibles	No
105	Otros indicadores	Contratos con proveedores inactivos	Si
109	Indicadores incumplimiento	Contratistas con contratos cancelados	No
109	Indicadores por inhabilidad	Inhabilitados por multa	Si

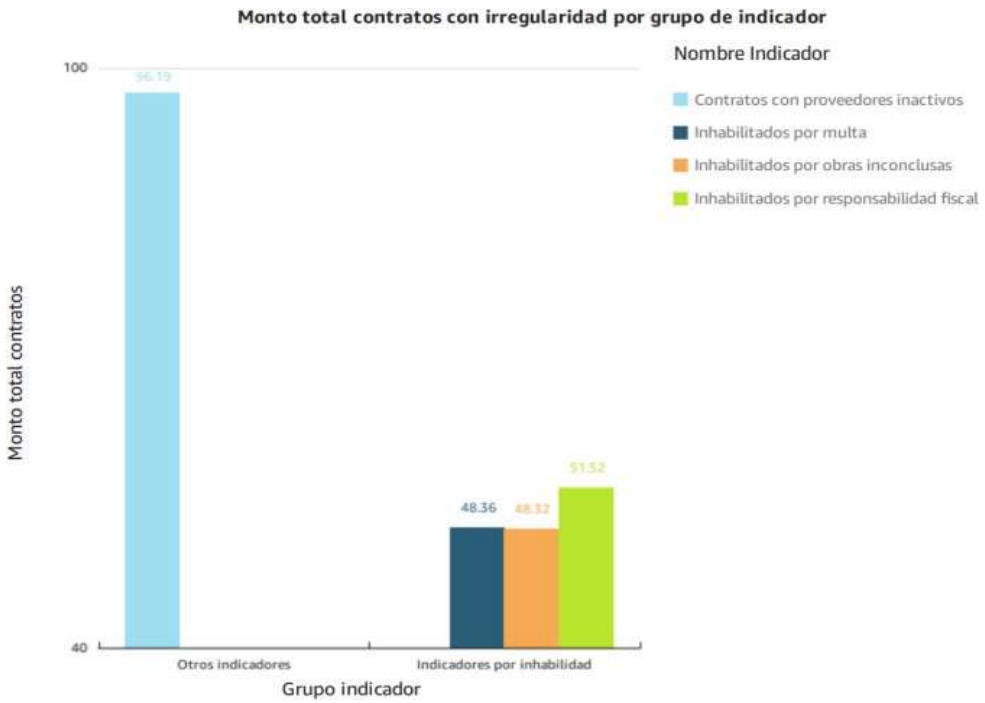
En la Ilustración 22 se presenta la tabla dinámica con la información en tiempo real, allí se puede observar si los contratos simulados (para nuestro caso, debido a que no se tiene unafuente de información que genere los contratos en tiempo real, así que se ha decidido hacer un componente encargado de simular dicha información, basado en contratos reales), tienenalgún tipo de irregularidad por grupo e indicador para la fecha en la que se simuló el contrato en tiempo real. Allí se podrá evidenciar el identificador del contrato, así mismo el resultadode la validación de cada uno de los indicadores, si presenta o no la irregularidad.

Ilustración 23. Número total de contratos irregulares por grupo de indicador analizados en tiempo real, visualizados en QuickSight.



La gráfica que se encuentra en la Ilustración 23, busca mostrar la cantidad de contratos analizados en tiempo real que presenta algún tipo de irregularidad en cada uno de los indicadores con el fin de identificar la tendencia de alertas de irregularidades de contratos recientes. Así mismo se puede visualizar la cantidad de contratos que no presentan ningún tipo de alerta sobre estos grupos de indicadores. Esta gráfica permite identificar aquellos indicadores que están siendo más incumplidos con los contratos que van llegando en tiempo real. Esto está agrupado por cada grupo de indicadores que se ha definido en el proyecto.

Ilustración 24. Monto total del valor adjudicado en los contratos irregulares por indicador analizados en tiempo real, visualizados en QuickSight.



En la Ilustración 24 se presenta la gráfica que muestra el monto total comprometido en los contratos con alerta de irregularidades por grupo de indicador y detallando cada indicador cuánto dinero se podría afectar por las irregularidades detectadas, esto con el fin de priorizar aquellos contratos e indicadores que mayor valor tenga y genere algún tipo de detrimento patrimonial al estado colombiano.

De acuerdo a los análisis realizados con las gráficas anteriores podemos concluir lo siguiente:

- Los indicadores con mayor monto comprometidos con alertas de irregularidades históricas son ofertas costosas con 29.39 billones, abuso de la contratación con 4.57 billones e inhabilitados por multa con 1.45 billones.

- En el grupo de otros indicadores se encuentran la mayoría de contratos con alertas deirregularidades tanto en los contratos históricos como en tiempo real.

- En cada grupo de indicadores existe un indicador que corresponde a más del 85% delos contratos con alertas de irregularidades históricas.
- Los indicadores con mayor monto comprometidos con alertas de irregularidades en tiempo real son contratos con proveedores inactivos 96.19 millones e inhabilitados por responsabilidad fiscal con 51.52 millones
- No se han alertado irregularidades en tiempo real para el grupo de indicadores de incumplimiento.

6 Trabajo futuro

Si bien este proyecto ha cumplido con los objetivos propuestos existen muchos ámbitos máspor los que se puede seguir avanzando y desarrollando el caso de uso. Si hablamos de la arquitectura, se puede seguir ampliando según las necesidades, se ha tratado de diseñar e implementar de una forma en que pueda crecer de forma independiente, por ejemplo, se puede añadir un componente de Machine Learning (ML), de análisis de texto, de inteligencia artificial, es decir cualquier tipo de componente que se requiera se podrá realizar sin ningúnproblema.
Existe mucha información que viene en documentos adjuntos como PDFs, imágenes y demásque no han sido utilizados para este caso de uso, así que este puede ser un punto en el cual se puede generar una mejora en el proceso de extracción de información para generar más indicadores. Así mismo, poder realizar modelos predictivos a partir de la información histórica para intentar predecir si un contrato puede tener mayor riesgo de generar algún tipo de presunta corrupción o irregularidad, también utilizar servicios adicionales de AWS para cargas de trabajo de IA/ML como pueden ser Amazon SageMaker for Business Analysts, AWS Textract o Amazon Comprehend.

Otra funcionalidad que se puede adicionar y desarrollar es para la verificación de costos de aquellos contratos que son de la modalidad de compras directas, estos contratos son compras de bienes materiales, la idea es poder verificar los costos, sobrecostos y diferenciar precios de este tipo de contratos en plataformas de comercio electrónico como Amazon, Mercado Libre, Google Shop o alguna otra tienda virtual, que permita identificar si algún producto que el estado quiera adquirir este en un precio mucho mayor al que está próximo a pagar.

Como se ha visto hay varios ámbitos donde se puede seguir avanzando con este proyecto, que alguien se interese y por qué no promocionar con las plataformas como SECOP II o con algún funcionario del gobierno u ONG que se interese en seguir avanzando para prevenir posibles casos de corrupción en Colombia y por qué no en

otros países.

7 Conclusiones y recomendaciones

Cuando se habla de Big Data una de las principales características es la veracidad de la información con que se trabaja y es acá donde se encuentra uno de los principales problemas que se afrontaron con este proyecto. Si bien existe mucha información disponible sobre la contratación pública en Colombia, mucha de esta información presenta algunos problemas, o los datos están incompletos, o datos erróneos o simplemente la información no concuerdacon lo que se está informando.

Pero esto puede tener diferentes explicaciones, la primera deella es errores humanos a la hora de registrar en plataformas como SECOP o SECOP II, comosabemos cuándo hay una interacción con un humano hay cierta probabilidad de que exista algún tipo de error involuntario y en este caso más al llenar cierta información requerida paralos contratos. La otra explicación que se puede dar es que todo esto se haga de forma premeditada para evitar identificar irregularidades en la contratación pública, puesto que al tener información 100% verídica posiblemente será más fácil encontrar culpables, responsables y casos de posibles irregularidades, pero es donde el estado debe tener un mayor control, la principal forma de evitar la corrupción en la contratación pública es tenerinformación verídica de cada uno de los contratos que el estado genera, cosa que tampoco se hace actualmente y que más adelante hablaremos del tema.

Todas las entidades que registreninformación en esta plataforma deberían seguir unos estándares altos para el manejo de información, tener catálogos para poder categorizar todo de forma más sencilla, puesto que se han encontrado casos donde escriben por ejemplo Bogotá, bogotá, bogota, Bogotá D. C. y todo para referirse a lo mismo, así que el manejo de catálogos para el registro de los datos será pieza fundamental para tener un mayor control.
Como lo comentamos en puntos atrás es fundamental poder controlar el 100% de los datos de la contratación pública en la plataforma como el SECOP II, que todas las organizacionesy entidades del estado sí o sí deban pasar por allí para poder realizar algún tipo de contratación sin importar el modo de contratación, con el objetivo de tener mayor control sobre los gastos del estado. Así mismo, no solo para la contratación pública, todo el dinero que salga y que entre del gobierno debería poder ser rastreado de cualquier manera por todos los ciudadanos, de esa forma tratamos de encontrar posibles desfalcos al estado y ponemos en cintura tanto estafador que anda suelto robando la plata de los impuestos de todos nosotros.

Para resumir el proyecto tuvo como objetivo implementar un sistema Big Data con

capacidad para procesar datos en tiempo real y clasificar los contratos de gasto público, con el propósito de identificar posibles irregularidades.
Luego de un arduo trabajo, se logró definir una arquitectura Big Data adecuada al contexto, gracias a la comparación exhaustiva de distintas arquitecturas de referencia. La selección de los componentes necesarios para el procesamiento en tiempo real fue un paso clave en el logro de este objetivo. Además, se desarrollaron indicadores y alertas de fraude específicamente adaptados al contexto colombiano. Estos indicadores se basaron en los datos disponibles y permiten generar alertas que podrán ser utilizadas por entidades gubernamentales, como la contraloría, la secretaría de transparencia y la fiscalía, para identificar potenciales irregularidades en los contratos de gasto público.

Otra contribución significativa del proyecto fue la creación de tableros de control que facilitan la visualización de las alertas e indicadores generados. Estos tableros proporcionan una visión clara y accesible para llevar a cabo auditorías y monitorear los contratos almacenados. Otra característica clave de la arquitectura implementada es su capacidad para permitir cambios en cualquier componente sin generar dependencias entre ellos. Esta flexibilidad fue diseñada con el objetivo general de convertirse en una arquitectura de referencia para otros proyectos. La posibilidad de introducir nuevos componentes de manera sencilla debido a su bajo acoplamiento es una ventaja significativa, lo que permite adaptarse a futuras necesidades y escalar según lo requiera el proyecto.

Asimismo, se enfatiza el esfuerzo realizado al documentar detalladamente toda la arquitectura. Esta documentación permite que cualquier interesado en el proyecto pueda configurarla fácilmente y ampliar su capacidad en pocos pasos para adaptarse a sus propias necesidades. Además, el hecho de ser una arquitectura de referencia ofrece la posibilidad de implementarla en diversos modelos de despliegue y en distintos proveedores cloud que ofrezcan servicios similares. La flexibilidad para mezclar diferentes proveedores brinda opciones adicionales para la implementación y puede aprovechar las ventajas específicas de cada uno.

La implementación de esta solución Big Data representa un paso importante en la lucha contra la corrupción y el mal uso de los recursos públicos. Si bien no determina directamente si un contrato es corrupto o no, sí provee un mecanismo eficaz para generar alertas y realizar un seguimiento más detallado por parte de las entidades pertinentes. De esta manera, se espera que este sistema pueda ser de gran ayuda en la identificación temprana de posibles irregularidades y, en última instancia, contribuir a fortalecer la transparencia y el buen uso de los recursos públicos en beneficio de toda la sociedad.

8 Bibliografía

Acosta, C. (9 de Diciembre de 2021). *"La contratación estatal asciende a $150 billones, lo cual constituye 15% del PIB".* Obtenido de asuntos:legales: https://www.asuntoslegales.com.co/actualidad/la-contratacion-estatal-asciende-a-150-billones-lo-cual-constituye-15-del-pib-3273780

Amazon. (s.f.). *AWS | Cloud Computing - Servicios de informática en la nube.* Recuperado el 6 de septiembre de 2022, de Amazon.com: https://aws.amazon.com/es/

ANCP - CCE. (29 de 06 de 2017). *Licitación Pública.* Obtenido de https://www.colombiacompra.gov.co/content/licitacion-publica

ANCP - CCE. (30 de octubre de 2019). *Conjuntos de Datos abiertos.* Obtenido de olombia Compra Eficiente | Agencia Nacional de Contratación Pública. : https://www.colombiacompra.gov.co/transparencia/conjuntos-de-datos-abiertos

ANCP - CCE. (27 de febrero de 2019). *SECOP.* Obtenido de https://www.colombiacompra.gov.co/secop/secop

ANCP - CCE. (s.f.). *Guía de búsqueda pública en el SECOP II.* Recuperado el 29 de septiembre de 2022, de https://www.gobiernobogota.gov.co/sites/gobiernobogota.gov.co/files/contratacion/guia_de_busqueda_publica_ciudadano.pdf

ANCP - CCE. (s.f.). *Órdenes de compra.* Recuperado el 11 de octubre de 2022, de https://www.colombiacompra.gov.co/tienda-virtual-del-estado-colombiano/ordenes-compra

ANI. (2021). *Estadísticas contratación.* Obtenido de https://www.ani.gov.co/estadisticas-contratacion

ANI. (31 de Enero de 2022). *Informe de Gestión Contratación Pública 2021.* Obtenido de https://www.ani.gov.co/estadisticas-contratacion

ANI. (s.f.). *Aniscopio.* Recuperado el 11 de octubre de 2022, de https://aniscopio.ani.gov.co/datos-abiertos

Antoniuk, O., Kuzyk, N., Zhurakovska, I., Sydorenko, R., & Sakhno, L. (2020). The role of ((BIG FOUR)) auditing firms in the public procurement market in Ukraine. *Independent Journal of Management & Production*, 2483-2495.

Apache. (s.f.). *Apache SparkTM - Unified Engine for large-scale data analytics.* Obtenido de Apache.org: https://spark.apache.org/

Arias, S., Garzón, S., H., A., Niño, G., L., V., & Rodriguez, k. (2020). *Diseño de solución informática de Big Data y BI para soportar la toma de decisiones del área de Desarrollo Comercial de Productos y Servicios de Valor Agregado en "SeNerg" con la finalidad de mejorar su rentabilidad.* Obtenido de Lorenanino.com: https://lorenanino.com/wp-content/uploads/2021/02/Entrega_TFM_G1_10150719.pdf

Bala, R., Gill, B., Smith, D., Ji, K., & Wright, D. (2021). *Magic Quadrant para servicios de infraestructura y plataforma en la nube. Gartner.com.* Obtenido de https://www.gartner.com/technology/media-products/reprints/AWS/1-271W1OTA-ESP.html

Banco Mundial. (25 de Abril de 2014). *Fraud and corruption awareness handbook : a handbook for civil servants involved in public procurement.* Obtenido de World Bank.: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/309511468156866119/fraud-and-corruption-awareness-handbook-a-handbook-for-civil-servants-involved-in-public-procurement
Bayrakdar, A., & Nogara, S. (2019). *SaaS vs. IaaS vs.* Obtenido de On-premise: https://gupea.ub.gu.se/handle/2077/62528
BBVA. (8 de mayo de 2017). *Las cinco uves del big data. BBVA.* Obtenido de https://www.bbva.com/es/las-cinco-uves-del-big-data/
Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., & ... & Kern, J. (2001). *Agile Manifesto.* Obtenido de Agile Alliance.
Campos, E., & Pradhan, S. (2007). *The many faces of corruption : tracking vulnerabilities at the sector level (English).*
Caracol Radio. (4 de enero de 2022). Condenados 2 exalcaldes de Boyacá por delitos asociados a corrupción. Obtenido de https://caracol.com.co/emisora/2022/01/05/tunja/1641339122_658123.html
Carter, R. (10 de noviembre de 2022). *Gartner Magic Quadrant for Cloud Infrastructure and Platform Services 2022.* Obtenido de Gartner: https://www.cxtoday.com/data-analytics/gartner-magic-quadrant-cloud-infrastructure-platform-services-2022/
Código Civil Colombiano art. 1495. (1887). *artículo 1495.* Obtenido de https://www.oas.org/dil/esp/codigo_civil_colombia.pdf
Congreso de la República Colombiana. (1993). *Ley 80 de 1993 atr. 30 parragrafo 1.*
Congreso de la República de Colombia. (2014). *Ley 1712 art 6 lit j.*
Coopers, P. (2013). *Identifying and reducing corruption in public procurement in the EU. Brussels: PricewaterhouseCoopers and Ecorys.*
Datasketch. (s. f.). *Banderas rojas en la contratación estatal.* Recuperado el 11 de octubre de 2022, de Datasketch.co: http://especiales.datasketch.co/contratos-colombia/banderas-rojas.html
Datos Abiertos Colombia. (s.f.). *la plataforma de datos abiertos del gobierno colombiano.* Recuperado el 9 de septiembre de 2022, de https://www.datos.gov.co/
David-Barrett, E., & Fazekas, M. (2015). *Corruption risks in UK public procurement and new anti-corruption tools.* Obtenido de Unpublished: https://doi.org/10.13140/RG.2.2.29841.02406
Dominguez-Quintero, L., & Vargas-Lombardo, M. (2011). *Herramientas de infraestructura como código: Ansible, Terrafom, Chef, Puppet. I+D Tecnológico.* Obtenido de https://doi.org/10.33412/idt.v17.2.3143
EdPrice-MSFT. (s.f.). *Compare AWS and Azure storage services - Azure Architecture Center.* Recuperado el 15 de septiembre de 2022, de Microsoft.com: https://docs.microsoft.com/en-us/azure/architecture/aws-professional/storage
Elizondo, R., & Joe, B. (2022). *Estudio comparativo de herramientas business intelligence de software libre y propietario para su aplicación y toma de decisiones en la inmobiliaria "Durancity". Babahoyo: UTB-FAFI.* Obtenido de 2022.
Fazekas, M., János, I., & King, L. (2013). *Corruption manual for beginners.* Obtenido de Crcb.eu: http://www.crcb.eu/wp-content/uploads/2013/12/Fazekas-Toth-King_Corruption-manual-for-beginners_v2_2013.pdf

Función Pública. (2018). *Curso de inducción a los gerentes públicos de la administración colombiana.* Recuperado el 5 de noviembre de 2022, de https://www.funcionpublica.gov.co/eva/gerentes/Modulo4/tema-2/1-modalidades.html

Gartner. (s.f.). *Definition of big data - Gartner information technology glossary.* Recuperado el 8 de septiembre de 2022, de https://www.gartner.com/en/information-technology/glossary/big-data

Google Cloud. (6 de septiembre de 2022). *Servicios de cloud computing.* Obtenido de https://cloud.google.com/?hl=es

Google Cloud. (s.f.). *¿Qué es el gobierno de datos?* Recuperado el 16 de septiembre de 2022, de https://cloud.google.com/learn/what-is-data-governance?hl=es

Google Cloud. (s.f.). *Compare AWS and azure services to.* Recuperado el 6 de septiembre de 2022, de https://cloud.google.com/free/docs/aws-azure-gcp-service-comparison

Gupta, B., Mittal, P., & Mufti, T. (2021). *A review on Amazon web service (AWS), Microsoft azure & Google cloud platform (GCP) services. Proceedings of the 2nd International Conference on ICT for Digital, Smart, and Sustainable Development, ICIDSSD 2020.* New Delhi, India.

Gupta, N., Singh, R., Das, S., & Choudhary, S. (2022). *AWS VS AZURE VS GCP: LEADERS OF THE CLOUD RACE.* Obtenido de International Research Journal of Modernization in Engineering Technology and Science: https://www.irjmets.com/uploadedfiles/paper//issue_7_july_2022/28711/final/fin_irjmets1658671480.pdf

Gutiérrez, I. (2020). *Diseño de una plataforma Big Data para predicción de patologías a partir de resultados médicos.*

Heggstad, K., & Frøystad, M. (Octubre de 2011). *The basics of integrity in procurement.* Obtenido de U4 Issue: https://www.cmi.no/publications/4211-the-basics-of-integrity-in-procurement

Hueso, L. (2020). Hacia la transparencia 4.0, el uso de la inteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales. *Repensando la administración digital y la innovación pública Publisher: Instituto Nacional de Administración Pública*.

Ilyas. (s.f.). *A public cloud comparison. Public Cloud Services Comparison.* Recuperado el 6 de septiembre de 2022, de https://comparecloud.in/

Jorquera, M. (2019). *Compras públicas y Big Data: Investigación en Chile sobre índice de riesgo de corrupción.* Obtenido de https://www.repository.fedesarrollo.org.co/handle/11445/3871

Jorquera, M. (2019). *Compras Públicas y Big Data: Investigación en Chile sobre índice de riesgo de corrupción. Santiago de Chile: Espacio Público.* Obtenido de BID: https://redflags.observatoriofiscal.cl/Content/documento/ModeloRedflagsInstitucion.pdf

Kamal*, M., Raza, H., Alam, M., & Su'ud*, M. (2020). Highlight the features of AWS, GCP and Microsoft Azure that have an impact when choosing a cloud service provider. International Journal of Recent Technology and Engineering. 4124–4132. Obtenido de https://doi.org/10.35940/ijrte.d8573.018520

Kreps, J. (2 de Julio de 2014). *Questioning the Lambda Architecture.* Obtenido de oreilly: https://www.oreilly.com/radar/questioning-the-lambda-architecture/
Laney, D. (2 de 2001). *META Group.* Obtenido de BibSonomy: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
Marz, N. (13 de 10 de 2011). *How to beat the CAP theorem.* Obtenido de http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html
Merchán, J. (2020). Arquitectura para el análisis de grandes cantidades de datos en tiempo real, aplicado a criptomonedas (Número 119). 65-75. Obtenido de Revista Escuela Colombiana de Ingeniería: https://www.escuelaing.edu.co/documents/866/Revista_119_pdf.pdf
Microsoft. (s.f.). *Servicios de informática en la nube.* Recuperado el 6 de septiembre de 2022, de Microsoft.com: https://azure.microsoft.com/es-es/
Ministerio de Comercio, Industria y Turismo. (2013). *Contratacion Pública | Cinco pasos sencillos que le permitirán acceder a una gran oportunidad de negocio.* Recuperado el 21 de septiembre de 2022, de https://www.aplicaciones-mcit.gov.co/cincopasos/c1.html
Molano, D. (2018). *La contratación pública y la contratación privada desde la realidad colombiana.* Escuela de Derecho y Ciencias Políticas.
Molano, D. (2018). *La contratación pública y la contratación privada desde la realidad colombiana. Escuela de Derecho y Ciencias Políticas.*
Monitor Ciudadano de la corrupción. (2019). *Radiografía de los hechos de corrupción en Colombia.* Obtenido de https://transparenciacolombia.org.co/Documentos/2019/Informe-Monitor-Ciudadano-Corrupcion-18.pdf
Monterrosa. (15 de Mayo de 2018). *"El Estado gasta $100 billones al año en contratación, según Colombia Compra Eficiente".* Obtenido de La republica: https://www.larepublica.co/economia/el-estado-se-gasta-100-billones-al-ano-en-contratacion-segun-colombia-compra-eficiente-2725948
Oracle. (s.f.). *¿Qué es el big data?* Recuperado el 8 de septiembre de 2022, de Oracle.com: https://www.oracle.com/co/big-data/what-is-big-data/
Pasquarelli, W., & Stirling, R. (2019). *The Next Generation of Anti-Corruption Tools: Big Data, Open Data & Artificial Intelligence.* OXFORD INSIGHTS.
Pletcher, S. (12 de enero de 2022). *Storage services compared: AWS vs Azure vs GCP.* Obtenido de A Cloud Guru: https://acloudguru.com/blog/engineering/storage-showdown-aws-vs-azure-vs-gcp-cloud-comparison
Portal Anticorrupción de Colombia – PACO | Bases de datos. (s.f.). *Gov.co.* Recuperado el 4 de octubre de 2022, de https://portal.paco.gov.co/index.php?pagina=descargarDato
Presidencia de la República de Colombia art. 2. (2011). *Decreto presidencial 4170 de 2011 art. 2.*
Rajan, R. (2014). *A comparative study of SaaS, PaaS and IaaS in cloud.* Obtenido de https://www.semanticscholar.org/paper/4a613de7b3a0ead4a08ed9a3caac79e553edbf46
Redhat. (8 de marzo de 2018). *Data storage: Dispositivos de almacenamiento de datos.* Obtenido de Redhat.com: https://www.redhat.com/es/topics/data-storage

Redhat. (11 de mayo de 2022). *¿Qué es la infraestructura como código - Infrastructure as Code?* Recuperado el 1 de noviembre de 2022, de https://www.redhat.com/es/topics/automation/what-is-infrastructure-as-code-iac
Redhat. (16 de agosto de 2022). *Diferencias entre IaaS, PaaS y SaaS.* Recuperado el 6 de septiembre de 2022, de https://www.redhat.com/: https://www.redhat.com/es/topics/cloud-computing/iaas-vs-paas-vs-saas
República de Colombia. (2008). *SISTEMA ELECTRÓNICO PARA LA CONTRATACIÓN PÚBLICA – SECOP.* Obtenido de https://www.idipron.gov.co/sites/default/files/docs/transparencia/normatividad/documentos/secop/secop.pdf
Srinivasan, V., Ravi, J., & Raj, J. (2018). *Google Cloud Platform for Architects: Design and manage powerful cloud solutions.* Obtenido de Packt Publishing.
Tapia, M., & Marisol, S. (2021). *Evaluación de una arquitectura de Big Data para la red móvil 5G a nivel de la capa ingestión utilizando aplicaciones de recolección de datos.*
Transparencia Colombia. (Enero de 2021). *Índice de Percepción de la Corrupción 2020.* Obtenido de https://transparenciacolombia.org.co/2021/01/28/indice-de-percepcion-de-la-corrupcion-2020/
Transparencia Internacional. (25 de febrero de 2006). *Handbook for curbing corruption in public procurement.* Obtenido de Transparency.org: https://www.transparency.org/whatwedo/publication/handbook_for_curbing_corruption_in_public_procurement
Transparencia Internacional. (24 de Julio de 2014). *Curbing corruption in public procurement: A practical guide.* Obtenido de Transparency.org. : https://www.transparency.org/whatwedo/publication/curbing_corruption_in_public_procurement_a_practical_guide
Transparency International. (28 de Enero de 2021). *CPI 2020: RESUMEN GLOBAL.* Obtenido de https://www.transparency.org/es/news/cpi-2020-global-highlights
Williams, L., & Kessler, R. R. (2000). *Pair Programming Illuminated. Addison-Wesley Professional.*
Yevge, A., Ghag, P., Solanki, C., & Mishra, A. (2022). *Review paper on cloud service provider - AWS, AZURE, GCP.* Obtenido de EasyChair.