# PA1910-2-ADACOP

## ADACOP

Portal Web para la Consulta y Análisis de Datos Abiertos Relacionados con la Contratación Pública

Ing. Jaime Andrés Mendoza Mendoza
Ing. Daniel Alejandro Calambás Marín
Ing. Julián Alexander Malaver Moreno
Ing. Leidy Andrea Ruiz Rodríguez

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2019

PA1910-2-ADACOP

# ADACOP

Portal Web para la Consulta y Análisis de Datos Abiertos Relacionados con la Contratación Pública

**Autor:**

Ing. Jaime Andrés Mendoza Mendoza

Ing. Daniel Alejandro Calambás Marín

Ing. Julián Alexander Malaver Moreno

Ing. Leidy Andrea Ruiz Rodríguez

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TÍTULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

**Director**

Ing. Rafael Andrés Gonzalez Rivera PhD.

**Co-Director**

Ing. José Francisco Molano Pulido
Ld. Analítica - CAOBA

**Comité de Evaluación del Trabajo de Grado**

< - >

< - >

**Página web del Trabajo de Grado**

http://pegasus.javeriana.edu.co/~PA1910-2-ADACOP

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
Mayo, 2019

**PONTIFICIA UNIVERSIDAD JAVERIANA**
**FACULTAD DE INGENIERIA**
**MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**

**Rector Magnífico**

Jorge Humberto Peláez, S.J.

**Decano Facultad de Ingeniería**

Ingeniero Lope Hugo Barrero Solano, ScD.

**Director Maestría en Ingeniería de Sistemas y Computación**

Ingeniera Angela Carrillo Ramos PhD.

**Director Departamento de Ingeniería de Sistemas**

Ingeniero Efraín Ortíz Pabón PhD.

## Artículo 23 de la Resolución No. 1 de Junio de 1946

*"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"*

# AGRADECIMIENTOS

Este trabajo de grado es un esfuerzo que ha permitido aprovechar las competencias y experiencias, no solo de nosotros como estudiantes, sino también de personas interesadas, dando su opinión y aporte en las tomas de decisiones que permitieron lograr el objetivo de este.

Agradecemos en primer lugar a nuestro director Rafael Andrés Gonzales Rivera, por su dirección, su orientación en la investigación y sus aportes que se convirtieron en bases sólidas para justificar el desarrollo del proyecto.

A José Francisco Molano, Líder de analítica de la Alianza CAOBA, por su guía, acompañamiento y disponibilidad durante las fases de desarrollo.

Agradecemos a la Alianza CAOBA, por su apoyo, experiencia y por permitirnos desarrollar este proyecto en sus instalaciones y bajo su infraestructura.

A Luis Carlos Reyes Hernández, director del Observatorio Fiscal y docente del Departamento de Economía PUJ, Adriana Salinas, líder de gasto público del Observatorio Fiscal y profesora de cátedra del Departamento de Ciencias Políticas PUJ, quienes nos orientaron en definiciones y vieron una oportunidad interesante de trabajo conjunto.

Al profesor Jorge Alvarado, director de la Maestría de Analítica para la Inteligencia de Negocios, por resaltar la generación de valor del proyecto.

Y por supuesto queremos agradecer a nuestras familias, por el acompañamiento, la paciencia y el sentido de responsabilidad inculcado durante todas las etapas de nuestras vidas.

Finalmente, a la Pontificia Universidad Javeriana y especialmente al Departamento de Ingeniería de Sistemas por permitirnos realizar nuestro proceso de formación en la maestría y por promover el desarrollo de profesionales éticos.

# Contenido

# Lista de figuras

# Lista de tablas

## ABSTRACT

The present project was carried out within the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA), in order to develop a Big Data architecture for extracting, storing, processing and analytical exploration of the Colombian government open data, initially including public contracts; the goal is to improve the transparency and usefulness that citizens can extract from such data, which implies work on the consistency, data standardization and highlighting certain properties that can generate value and contribute to decision making. This document describes the implementation of the architecture, the resulting prototype and provides the results of TAM (technology acceptance model) of acceptance with potential expert users.



## RESUMEN

El presente proyecto, realizado de la mano con el Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA), se focaliza en el desarrollo de una arquitectura de *Big Data* que soporte la extracción, almacenamiento, procesamiento y exploración analítica sobre los datos abiertos del gobierno colombiano, inicialmente relacionados con la contratación estatal; con el fin de mejorar los procesos de transparencia y brindar una mayor cercanía hacia el ciudadano, ya que se debe trabajar en la consistencia, estandarización de los datos y en particular en las propiedades que generen valor y aporten a la toma de decisiones. Este documento describe la implementación de la arquitectura, el prototipo resultante y provee los resultados de las pruebas TAM (*technology acceptance model*) de aceptación a usuarios interesados.

# RESUMEN EJECUTIVO

Para el aprovechamiento de los datos abiertos de Colombia, el Gobierno nacional dispone de una plataforma web de datos abiertos [1] que almacena datos y publica los mismo bajo licenciamiento abierto disponible para cualquier usuario. Mediante la Carta Internacional de Datos Abiertos [2] , en 2016, Colombia adoptó seis principios base para el uso y buenas prácticas sobre los datos abiertos [3] . Dentro de los datos expuestos y para el desarrollo del presente proyecto, el interés se centra en las contrataciones públicas, aquellos datos dispuestos por la agencia de contratación Colombia Compra Eficiente, donde se encuentra la información relacionada con las compras y gastos del estado [4].

A pesar de que Colombia está bien posicionada en datos abiertos de gobierno según indicadores de la OCDE [5], aún no hay herramientas o mecanismos efectivos y de libre acceso que soporten el procesamiento, el análisis de la información o visualizaciones que aporten a la toma de decisiones para usuarios interesados. Por otra parte, los procesos de descarga, las incertidumbres en la actualización de los datos y los problemas de calidad identificados en los mismos, hacen que el acceso y aprovechamiento real de estos datos abiertos no logre la efectividad y apropiación requerida ni en usuarios del sector público (ej. alcaldes y gobernadores) ni en la ciudadanía. De esta necesidad parte el desarrollo del presente proyecto que, en el marco de la Alianza CAOBA, realiza la implementación de una herramienta con tecnologías de *Big Data* (*Open Source*) que consolida y enriquece el contenido de los datos abiertos (por medio de fuentes externas), procesa la información contenida en los datos, genera métricas de calidad para la generación datos limpios y permite desplegar modelos analíticos y visualizaciones descriptivas con el fin de facilitar el entendimiento de los datos.

A partir de la anterior necesidad, se realizó una tarea exhaustiva de levantamiento de información que se dividió en dos enfoques, el primero buscaba plantear una arquitectura de *Big Data* escalable que soportara la extracción, el almacenamiento, procesamiento, administración y visualización de los datos. En este ejercicio se evaluaron herramientas de uso libre y mantenidas por la comunidad, como resultado se tiene que Apache y Hortonworks contienen soluciones para el almacenamiento, procesamiento y administración de la información; el segundo

enfoque buscaba crear y soportar técnicas analíticas avanzadas para integrar información adicional (fuentes externas) sobre los datos abiertos, específicamente sobre procesos y contratos públicos. Con el lenguaje de programación *Python*, se realizaron las tareas de entendimiento, descripción, limpieza, aplicación de criterios de calidad de datos y la generación de *fact tables* que posteriormente serían usadas para realizar la visualización de los tableros de control o *dashboards*. De igual manera, y con el fin de realizar un análisis de sentimiento, en este primer prototipo se integraron los resultados extraídos de contratación con fuentes externas (Twitter), teniendo en cuenta la polaridad (en términos de sentimiento positivo o negativo) de las palabras, bajo el contexto de la contratación pública.

Para el desarrollo del proyecto se adoptó una metodología de desarrollo de *software* ágil. *Scrum* fue seleccionada teniendo en cuenta dos puntos: primero, que la Alianza CAOBA implementa la metodología dentro de sus desarrollos y segundo, que en su flujo de trabajo se realizan entregas iterativas por *sprint*, lo cual permitió que el entendimiento de los requerimientos, enlazados a las actividades y tareas de cada uno de los miembros del equipo, fuesen controlados y refinados por medio de historias de usuarios.

# 1 Introducción

La Alianza CAOBA es el Centro de Excelencia y Apropiación en *Big Data* y *Data Analytics* el cual fue creado en el año 2015 como una iniciativa de Colciencias y el Ministerio de las Tecnologías de la Información y las Telecomunicaciones (MinTIC) y conglomera empresas del sector público, privado y la academia. CAOBA tiene como misión la generación de nuevo conocimiento relacionado con el *Big Data* y *Data Analytics* por medio de la investigación aplicada y el desarrollo de soluciones tecnológicas [6].

Al ser patrocinado por el Gobierno Nacional, CAOBA está comprometido con el apoyo de las iniciativas tecnológicas estatales y tiene definido un portafolio de proyectos estratégicos alineados a las necesidades del Estado y la industria. Uno de estos proyectos pretende fortalecer el uso de los datos abiertos del Gobierno con el fin de proporcionar mayor transparencia y soportar la toma de decisiones por parte del Gobierno y la ciudadanía, aplicando procesos analíticos sobre información correspondiente a la contratación estatal. El desarrollo de este proyecto se encuentra alineado con uno de los pilares del MinTIC que parte de los datos abiertos de entidades públicas como elementos clave para promover la transparencia, la competitividad, el desarrollo económico y la generación de impacto social en el contexto de apropiación de las TIC [1].

El Gobierno de Colombia ha reconocido que la compra y contratación pública es un asunto estratégico. Por este motivo, se creó la organización Colombia Compra Eficiente en el año 2011 la cual crea políticas unificadas que sirvan de guía para los administradores de compras del estado y que permitan monitorear y evaluar el desempeño del sistema generando mayor transparencia [4]. Con la creación de esta organización, se adoptó el sistema de información SECOP promoviendo la contratación abierta y el uso de la información de los procesos de compra y contratación para fomentar la colaboración, la innovación y la transformación de la entrega de bienes, obras y servicios a los ciudadanos [7].

En el contexto mundial se han realizado una gran cantidad de esfuerzos para promover la publicación y uso de datos abiertos gubernamentales. El primer país en implementar una estrategia de datos abiertos fue el gobierno de los Estados Unidos en el año 2009 basado en tres

pilares: colaboración, partición y transparencia [8]. De igual manera, la Unión Europea ha introducido el concepto de gobernanza inteligente, que tiene como eje central el uso de las grandes cantidades de datos que la administración pública genera, en el desarrollo de sus actividades y en sus relaciones con la ciudadanía y las empresas, para apoyar los procesos de toma de decisiones estatales y aumentar la transparencia [9]. También resalta que el rol de las agencias del gobierno no debe basarse únicamente en liberar los datos, sino en crear estrategias para atraer entidades externas y *stakeholders* que generen innovación a partir de los datos abiertos [7]. En una encuesta realizada por Deloitte - *Shaping the Future of Open Data An assessment of the open data* - en 2016, se preguntó ¿Qué tipo de información del sector público o conjuntos de datos específicos le interesaría o cree que debería ponerse a disposición de los ciudadanos? [10], el 38% de los encuestados desean conocer más sobre decisiones de gobierno y un 6% sobre contratación pública.

Bajo el anterior contexto, este proyecto se desarrolló en varias fases, el diseño de la arquitectura de Big Data, la implementación de dicha arquitectura con tecnologías libres, la integración de los componentes de software y el diseño y despliegue del portal web para la visualización de la información y la validación de calidad de los datos disponibles.

## 1.1 Problemática y oportunidad

A pesar de las iniciativas relacionadas con la publicación de datos abiertos de contratación pública, aún existe una carencia de mecanismos para soportar la toma de decisiones apoyadas en este tipo de información.

La Alianza CAOBA, apoyó esta primera fase del proyecto, para la creación de una herramienta que consolida y enriquece el contenido de los datos abiertos con fuentes externas (p. ej. Redes Sociales) y permite visualizarlos mediante *dashboards* creados a partir de modelos analíticos descriptivos. Los *dashboards* o tableros de control, son herramientas de visualización que permiten conocer, interactuar y analizar información de manera visual [11] por medio de diferentes gráficos [12] con el fin de facilitar su entendimiento.

Se desarrolló una plataforma tecnológica llamada ADACOP, que utiliza componentes tecnológicos de *Big Data* para almacenar, procesar y visualizar la información recolectada, con el

fin de disponer su uso para la creación de modelos analíticos por parte de la Alianza CAOBA. Una vez desarrollada la herramienta, se utilizaron metodologías de validación (por medio del modelo TAM, para medir la utilidad de ADACOP), y verificación de los requerimientos (funcionales y no funcionales) en cada uno de los cuatro componentes definidos en la arquitectura de la solución: recolectar, almacenar, procesar y visualizar. Lo anterior, se realizó siguiendo la metodología propuesta en el trabajo referenciado [13], que consiste en definir los requerimientos de la solución en términos de las 4 Vs del *Big Data* (Volumen, Velocidad, Variedad, Variabilidad). En cuanto a la interfaz, el portal web contiene módulos de selección, carga, y visualización de los gráficos descriptivos generados, y *dashboards* o tableros de control específicos para el entendimiento de SECOP.

La recolección, reproducción y procesamiento de los datos utilizados en esta solución, están cubiertos bajo la legislación colombiana, mediante la Ley 1712 de 2014, la cual permite el uso de datos abiertos para su aprovechamiento y/o transformación de forma libre y sin restricciones, para hacer aplicaciones por parte de terceros y contenidos de su propia creación [14].

La solución construida es propiedad intelectual de la Alianza CAOBA mientras que la propiedad moral del mismo es de los integrantes que hicieron parte del desarrollo de este proyecto.

# 2 Descripción del Proyecto

## 2.1 Objetivo general

Diseñar e implementar una arquitectura con tecnologías de *Big Data* para recolectar, almacenar, procesar y visualizar información relacionada a datos abiertos de contratación pública estatal del gobierno de Colombia, que será utilizada por la Alianza CAOBA para el desarrollo de modelos analíticos.

## 2.2 Objetivos específicos

1. Identificar fuentes primarias en el portal de datos abiertos (www.datos.gov.co) que contengan información relacionada con contratación pública.
2. Identificar fuentes externas complementarias que permitan enriquecer el contenido de las fuentes primarias.
3. Diseñar la arquitectura de la plataforma para la recolección, análisis y visualización de información relacionada con la contratación estatal.
4. Implementar un repositorio para las fuentes identificadas considerando los requerimientos de volumen, velocidad, variedad y velocidad de los datos.
5. Desarrollar los componentes de software que permitan la carga de las fuentes identificadas hacía el repositorio.
6. Desarrollar un componente de software para el análisis de estadística descriptiva de las fuentes cargadas en el repositorio, que contribuya a la validación de la calidad de datos.
7. Desarrollar un componente web, para la visualización de *dashboards* con la información almacenada en el repositorio de datos.

## 2.3 Delimitación

### 2.3.1 Alcance

El proyecto culmina con el diseño, implementación y producción de ADACOP sobre un ambiente *Big Data* estable, teniendo en cuenta que las herramientas y tecnologías serán analizadas y definidas dentro del proyecto. Las herramientas utilizadas en el desarrollo son de licencia abierta y son descritas en la Sección 4.1 de este documento.

La evaluación del proyecto se realizará mediante un juicio de expertos, dentro del ambiente productivo el cual es desarrollado con el modelo de aceptación tecnológica, TAM y SUS, que se describen en la Sección 5 de este documento.

Los datos a procesar están relacionados contratación pública, procesos de compra y contratación a nivel Colombia los cuales proceden de los conjuntos de datos del Sistema Electrónico de Contratación Estatal (SECOP) [15] disponibles en el portal de datos abiertos de Colombia [12].

El inventario de fuentes debe contener únicamente información disponible en portales de datos abiertos y fuentes alternas asequibles de forma libre, como lo es el portal de datos abiertos y la red social Twitter accesibles por medio de sus respetivas APIs.

## 2.4 Metodología

El Centro de Excelencia de excelencia en *Big Data* y *Data Analytics* desarrolla todos sus proyectos bajo la metodología de desarrollo ágil *SCRUM*. ADACOP al ser desarrollado como parte del portafolio de proyectos de CAOBA fue gestionado y desarrollado por medio de esta metodología. *SCRUM* permite implementar cambios y nuevos requerimientos en el desarrollo a lo largo del mismo, lo cual, es conveniente para ADACOP debido a que durante las validaciones e investigación surgen nuevos requerimientos y necesidades para implementar como parte del sistema.

El proceso de *SCRUM* requiere definir los roles de los integrantes e interesados del proyecto para así asignar las respectivas responsabilidades. Principalmente el equipo consta de cinco investigadores, el líder de analítica de CAOBA y el director del presente trabajo de grado. Con un total de siete personas, los roles fueron distribuidos como se describe en la tabla 2.1.

| *Rol* | *Integrante* | *Responsabilidades* |
|---|---|---|
| *Product Owner* | Francisco Molano | Gestionar los requerimientos de ADACOP.<br><br>Priorizar los requerimientos para el desarrollo.<br><br>Definir las metas y alcance en cada sprint de desarrollo.<br><br>Fijar criterios de aceptación de los entregables desarrollados. |
| *Scrum Master* | Jaime Mendoza | Facilitar la comunicación entre el equipo de trabajo y el product Owner.<br><br>Organizar el desarrollo del sprint de acuerdo a las metas definidas por el product Owner.<br><br>Asegurar el cumplimiento de los roles y responsabilidades. |
| *Scrum Team* | Daniel Calambás<br><br>Julian Malaver<br><br>Jaime Mendoza<br><br>Angelica Pacheco<br><br>Andrea Ruiz | Desarrollar el producto final basados en la planeación de cada sprint. |
| *Director Alianza CAOBA* | Rafael Gonzalez | Director del trabajo de grado ADACOP. |

Tabla 2.1 - Roles en proceso *SCRUM*.

El grupo interesados en el funcionamiento de ADACOP y principalmente SECOP, está compuesto por Luis Carlos Reyes Hernández, director del Observatorio Fiscal y docente del Departamento de Economía PUJ, Adriana Salinas, líder de gasto público del Observatorio Fiscal y profesora de cátedra del Departamento de Ciencias Políticas PUJ.

### 2.4.1 Proceso *SCRUM*

*SCRUM* es la metodología de desarrollo de *software* ágil de mayor auge en los últimos años [16]. Como toda metodología de trabajo, esta tiene definido un proceso que, a diferencia de metodologías tradicionales, es de naturaleza iterativa como se describe en la figura 5. Es importante resaltar que la implementación de esta metodología depende de la organización y proyecto donde sea adoptada y en esta sección se pretende describir esta adopción en el proyecto ADACOP liderado por la Alianza CAOBA.

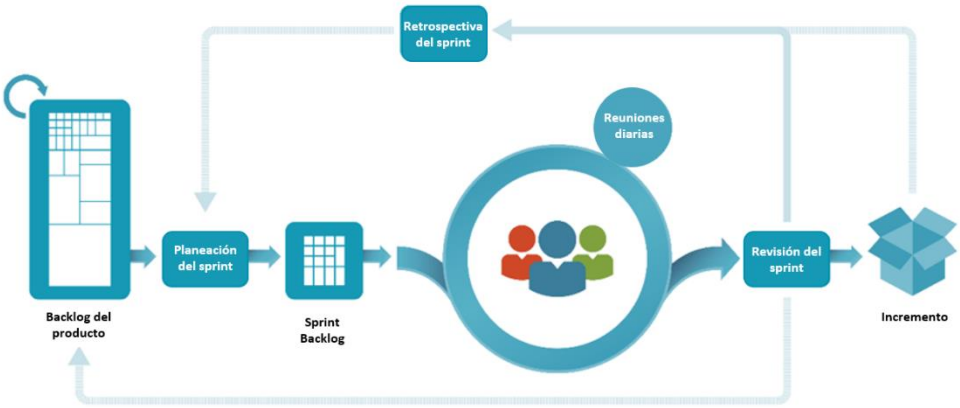


Figura 1 - El proceso de SCRUM.
Fuente: https://www.scrum.org/resources/blog/que-es-scrum

Definidos los roles de trabajo dentro de la metodología, la primera actividad fue realizar la primera iteración en la creación del *Product Backlog*. El *Product Backlog* es el listado de los requerimientos del sistema, el cual es administrado y priorizado por el *Product Owner*. Este listado debe ser definido de acuerdo con los objetivos y alcances definidos para el proyecto. En el caso de ADACOP, basado en los objetivos planteados en la sección 2 de este documento. La

gestión y administración del *Product Backlog* debe ser iterativa, a medida que el proyecto avanza, teniendo en consideración los nuevos requerimientos y necesidades de los interesados.

El *Product Owner* y el *SCRUM Master* tienen la responsabilidad de definir las actividades y metas de los *sprints* de desarrollo para la asignación de trabajo dentro del *SCRUM Team*. Un *sprint* es un periodo de tiempo en el cual se desarrolla una nueva funcionalidad del producto. Para cada *sprint* se define su respectivo *sprint backlog* con las actividades a desarrollar. Para ello, es necesario conocer el tiempo que los diferentes integrantes del *SCRUM Team* disponen para dedicar a las actividades de desarrollo del proyecto. Este tiempo se definió de acuerdo a la tabla 2.2 considerando las horas que el equipo debe dedicar siendo miembros de CAOBA y las horas a dedicar como parte de los trabajos de grado.

| ***Integrante*** | ***Horas CAOBA*** | ***Horas Trabajo de Grado Maestría Sistemas*** | ***Horas Trabajo de Grado Maestría Analítica*** | ***Total horas día hábil*** |
|---|---|---|---|---|
| *Daniel Calambás* | 4 | 2 | 2 | 8 |
| *Julian Malaver* | 4 | 2 | 0 | 6 |
| *Jaime Mendoza* | 4 | 2 | 2 | 8 |
| *Angelica Pacheco* | 4 | 0 | 2 | 6 |
| *Andrea Ruiz* | 0 | 2 | 2 | 4 |

Tabla 2.2 - Horas de trabajo *SCRUM* Team ADACOP.

La planeación de cada *sprint* se realiza en una reunión conocida como el *Sprint Planning* donde se reúnen el *Product Owner*, el *Scrum Master* y el *Scrum Team* para socializar las actividades y alcance del *sprint*. Esta reunión es de suma importancia dado que todo el equipo de trabajo estima el tiempo que podría llegar a tardar en completarse cada actividad basada en la experiencia de todos los integrantes.

Basados en el tiempo disponible estipulado en la tabla 2.2 y en la estimación de cada actividad, se establecía la duración del *sprint* de tal forma que no supere las dos semanas de trabajo ya que el objetivo es realizar *sprints* cortos con el fin de agregar o redefinir funcionalidades de acuerdo con las necesidades. De igual manera, asignaba un responsable a cada actividad equilibrando las cargas de trabajo entre los miembros del equipo.

El seguimiento constante del desarrollo de las actividades le permitió al *Product Owner* y al *Scrum Master* detectar posibles problemas o retrasos en el desarrollo de las actividades debido a la ocurrencia de eventos inesperados. Por esto, lo que *Scrum* propone es realizar reuniones cortas diarias llamadas *Daily Scrum*; para el desarrollo de ADACOP se definió llevar a cabo estas reuniones, en un tiempo no mayor a quince minutos, donde debían estar presentes todos los miembros del equipo y compartir los avances desarrollados desde el día anterior, el trabajo que se encuentra realizando y los inconvenientes que pudo haber tenido. El *Scrum Master* contaba con la responsabilidad de agendar los *Daily Scrum* y de ayudar a solucionar los problemas que se presentaran o reportaran considerándolos para el siguiente *Scrum Planning*.

Al concluirse el tiempo estipulado para cada *sprint*, el equipo se reunía nuevamente con el fin de que cada integrante presentara sus resultado al *Product Owner*, quién aprobaba o no los resultados (estas reuniones son conocidas como *Sprint Review*), en caso de no hacerlo o de que una tarea no haya sido completada por una estimación desacertada, estas actividades entraba a ser parte del *Sprint Backlog* de la siguiente iteración. Adicionalmente, *Scrum* propone que al finalizar cada sprint se realice una reunión de retrospectiva (*Sprint Retrospective*) en la cual se discuta que se hizo bien y que se hizo mal durante la iteración para. ya sea continuar haciéndolo, o corregirlo para el siguiente *sprint*, en el caso de ADACOP, este proceso se realizó como parte del *Sprint Review* y no como una reunión separada.

Para ADACOP el proceso de *Scrum* se llevó a cabo tal como se describe anteriormente en un periodo de 16 semanas y 8 *sprints* con una duración promedio de 6 días hábiles por cada uno.

# 3 Estado del arte

En el año 1942 el sociólogo estadounidense Robert King Merton explica la importancia de hacer de libre acceso los resultados de las investigaciones, de tal manera que cada investigador aporte al conocimiento común y así poder avanzar [17]. El concepto de Datos Abiertos fue mencionado por primera vez en un artículo científico en 1995 donde los autores promueven un intercambio completo y abierto de información científica entre diferentes países como prerrequisito del análisis y entendimiento de la atmosfera, los océanos y la biosfera. *"Our atmosphere, oceans and biosphere form an integrated whole that transcends borders."*[17]. De acuerdo con la definición dada por Open Data Charter [2], los datos abiertos son datos digitales que son dispuestos públicamente, con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar. En el contexto colombiano, el gobierno ha desarrollado estrategias para promover los datos abiertos como un activo, este esfuerzo se fundamenta en seis principios de los datos abiertos que se mencionan a continuación:

1. **Abiertos por defecto**: La información debe estar disponible para su consulta.
2. **Oportunos y completos**: Aportan valor al usuario.
3. **Comparables e interoperables**: Comparados desde distintos sectores y periodos de tiempo.
4. **Mejorar la gobernanza y la participación ciudadana**: Datos para fortalecer la transparencia.
5. **Accesibles y usables**: Datos gratuitos y bajo licencia abierta (lectura de maquina).
6. **Apoyen el desarrollo y la innovación**: El uso de datos abiertos permite la construcción de nuevo conocimiento.

Teniendo en cuenta el cuarto principio, la ley 1712 de 2014 de transparencia y derecho de acceso a la información pública, articulo 6, define que los datos deben estar a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados. De acuerdo con el argumento anterior, Colombia dispone del portal de datos abiertos, el cual cuenta con un proceso cíclico de comunicación, flexibilidad, retroalimentación e interacción, alineado, no solo a la preparación y publicación de los datos

por parte de las entidades, sino también a estrategias concretas para promover el uso de los datos publicados, realizar seguimiento e identificar el valor agregado que se está generando con su utilización [18].

Para asegurar la transparencia y el acceso de estos datos, es necesario realizar tareas la recolección, mediciones de calidad exhaustiva (precisión), análisis y visualizaciones, esto soportado en una arquitectura robusta y escalable, donde las colecciones de datos sean almacenadas y procesados de forma masiva, de aquí parte la solución con herramientas *Open Source* de *Big Data*. Técnicamente, Gartner define *Big Data* como "un activo de información de gran volumen, alta velocidad y / o gran variedad que exige formas de procesamiento de información innovadoras y rentables que permitan una visión mejorada, toma de decisiones y automatización de procesos" [19]; el concepto anterior mencionado, se desarrollado en base a las 4 Vs [20]:

- **Volumen**. Cantidad de información generada o almacenada.
- **Velocidad**. Frecuencia de actualizaciones de los datos, relacionado con el cambio de estado y la fluidez de estos.
- **Variedad**. Los datos pueden ser de tipo Estructurado, semiestructurado o no estructurado.
- **Variabilidad**. Los valores de los datos tienen diferentes tipos o formatos.

El desarrollo de soluciones de *Big Data* parte de un objetivo, la toma de decisiones basada en los datos, para soportar este objetivo, el desarrollo parte desde la implementación de tecnologías superpuestas que primero, gestionen los diferentes tipos de datos (estructurados, no estructurados y semiestructurados), segundo, partiendo de la premisa de eficiencia, se genere el procesamiento de grandes volumen de datos y tercero permita visualizar el comportamiento de dichos datos en tiempo real y de forma entendible al usuarios [21]. Un pipeline de una arquitectura típica de *Big Data* se basa en 6 capas (ver figura 2):


Figura 2. Pipeline *Big Data* **[21]**.

1. Extracción o adquisición de las fuentes de datos.
2. Transformación y estandarización de datos. Aplicando normalización y un nivel aceptable de calidad en los datos a procesar.
3. Almacenamiento de datos.
4. Procesamiento de información (algoritmos).
5. Análisis. Aplicación de técnicas y modelos analíticos.
6. Visualización y presentación al usuario.

A partir de este estudio inicial y junto con la experiencia de la Alianza CAOBA se identificaron los siguientes referentes para realizar la definición de la arquitectura de ADACOP e implementar la solución con tecnologías Big Data; incluyendo dentro de cada capa herramientas Open Soure. Para el procesamiento de los datos, se definió Hadoop [22] como ecosistema, que trabaja con Yarn [22] (API) para la gestión de los recursos del clúster, en la capa de almacenamiento del clúster se definió HDFS [23] y Hive[24] como herramienta de consulta. En el ejercicio de definición de las herramientas a implementar dentro de la arquitectura para ADACOP, se trabajó de forma iterativa, lo que concluyo con la definición de nuevas soluciones dentro de cada capa (la arquitectura adoptada para ADACOP se define en la sección 4.3.2).

El desarrollo de ADACOP se llevó a cabo de forma iterativa apoyado en la metodología de desarrollo de software *Scrum,* aplicando un proceso de investigación aplicada a este desarrollo. Como lo proponen Gonzalez et al. [25], esto implicó un desarrollo iterativo basado el diseño donde la relevancia, el diseño y el rigor son los ejes principales del proceso. La relevancia se ve reflejada en una correcta definición de los requerimientos que son implementados durante el proceso *Scrum* de ADACOP, considerando el entorno de aquellas personas que interactuarán con la herramienta y quienes tienen unas necesidades específicas que fueron satisfechas por medio de la plataforma. Estos requerimientos se contemplaron en los ciclos de diseño de forma tal que la evolución y cambio en los mismos se vieron reflejados en la herramienta, no sin evaluarlos dentro del proceso de desarrollo, verificando el correcto funcionamiento de estos. Cada requerimiento desarrollado fue sustentado, no solo en las necesidades de los interesados, si no en una base de conocimiento aplicable basado en los desarrollos y experiencias de la Alianza CAOBA y de la comunidad científica en general. Para ello los ciclos de

rigor en esta investigación, permitieron encontrar métodos, teorías, experiencias y arquitecturas de referencia aplicables a el proceso de desarrollo y a la solución como tal.

## 3.1 Metodología de investigación.

En el proceso de entendimiento del contexto de las políticas de datos abiertos, contratación pública y el desarrollo de soluciones *Big Data*, se realizó una tarea de revisión de literatura, que apunta a conocer y verificar el objeto de estudio soportado por dos componentes, el primero teórico y el segundo orientado hacia recursos literarios publicados de casos de éxito con tecnologías aplicadas, investigaciones previas de análisis de datos de portales internacional y el interés social del individuo.

Para realizar la investigación, se estableció un marco previo de búsqueda orientada a examinar los enunciados relacionados con el campo objetivo, como lo son las palabras clave; esto, con el fin de tener un acercamiento a actividades o implementaciones similares, relacionadas con los objetivos del presente proyecto. En base a este marco se estableció el proceso de revisión de sistemática de literatura [26].

### 3.1.1 Revisión sistemática.

Como trabajo previo y contextualización, se realizó una revisión sistemática de literatura, con el fin de sintetizar la información científica disponible e identificar trabajos relacionados bajo tres variables: analítica, arquitecturas de Big Data y negocio. En el ejercicio de revisión, se creó un proceso iterativo (Ver Figura 3), cuyo cuerpo está compuesto de 3 fases:

1. Exploración de la información. Búsqueda de literatura de acuerdo con palabras clave especificas relacionadas con el objetivo del proyecto.
2. Selección. Se establecen criterios de selección de las bases científicas, estos criterios los miden la credibilidad de la información.
3. Teorización. Modelar el conocimiento en bases conceptuales.

Vale la pena aclarar, que la figura 3 muestra el proceso iterativo, que identifico en la primera fase 58 artículos de referencia; bajo los criterios de selección, se reducen a 18, estos dan

justificación a la conceptualización, al diseño de la solución ADACOP y que complementan el componente de negocio.

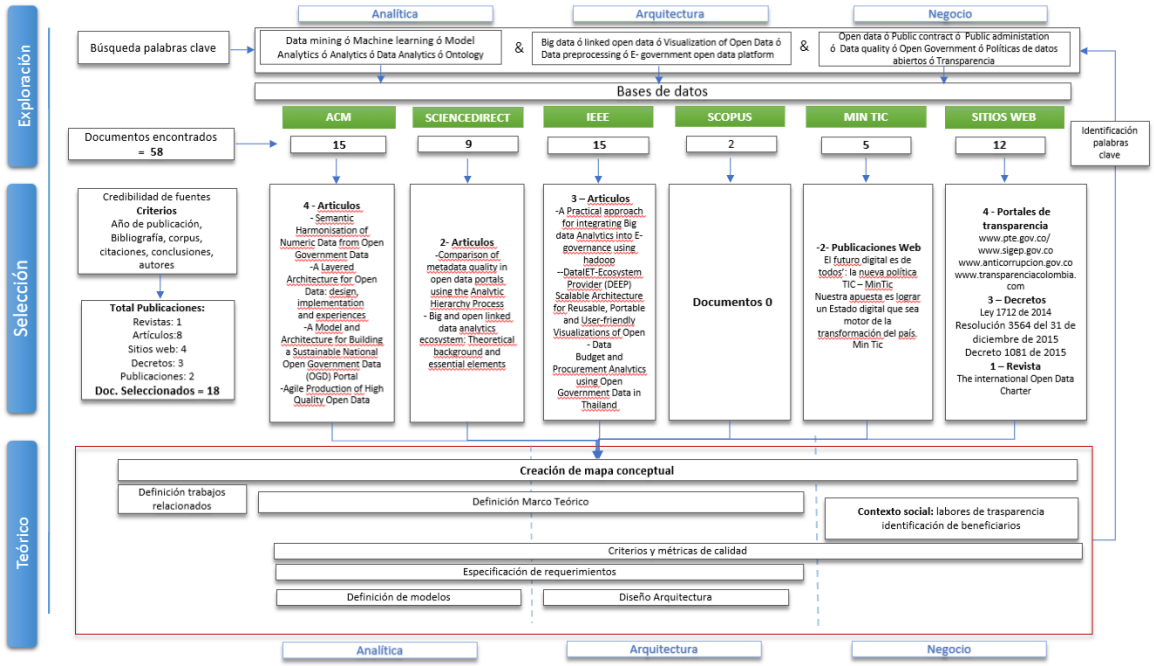


Figura 3. Revisión sistemática. Fuente: Elaboración propia.

## 3.2 Trabajos relacionados

Siguiendo la metodología de investigación descrita en la sección 5.1, se encontraron distintos trabajos relacionados con el uso de datos abiertos, datos de contratación pública y el uso de tecnologías Big Data para el desarrollo de este tipo de soluciones, con el fin de determinar aquellos más relevantes con conocimiento aplicable para el desarrollo de ADACOP.

R. De Donato *et al.*[27] desarrollaron una metodología de trabajo para que las entidades públicas generen de manera colaborativa conjuntos de datos de forma iterativa para poner a la disposición de la ciudadanía información de la mejor calidad posible y empoderar a los mismos en el uso proactivo de esta información. Los autores resaltan la necesidad de mejorar la calidad

en los datos públicos y que de acuerdo con informe del *Open Data Barometer* [28] aún los datos abiertos carecen de un buen índice de calidad. Colombia no es la excepción ya que según la política nacional de explotación de datos emitida por el Departamento Nacional de Planeación [29] resaltan que en general los datos abiertos de Colombia carecen de calidad y actualización. El trabajo de De Donato *et al.*[27] permitió determinar que para entender y estructurar correctamente los datos abiertos es necesario tener acceso al menos a los metadatos de los mismos y además tener en mente los problemas de calidad que se encontraron dentro de los conjuntos de datos.

Bhushan et al. proponen una implementación de una solución *Big Data* basado en una metodología para la integración de tecnologías analíticas con propósitos de gobernanza digital [30]. Los autores utilizan el ecosistema de *Big Data* Hadoop con el fin de construir un pipeline analítico para procesar información gubernamental y generar *insights* para soportar procesos de toma de decisiones. Como resultado los autores resaltan el uso de Hadoop ya que al ser un ecosistema *open source* permite ser personalizado de acuerdo a los requerimientos del problema y adicionalmente integrarlo para propósitos gubernamentales y de gobernanza electrónica es fácilmente alcanzable. Sin embargo, este trabajo no tiene en consideración el empoderamiento de la ciudadanía en este proceso y dentro de su arquitectura no se contempla un medio por medio del cual se brinde mayor acceso y transparencia en la información y mucho menos relacionada a la contratación pública dado que se aborda como una solución general para el estado. Es importante resaltar que el uso del ecosistema *Big Data* Hadoop en este trabajo generó un punto de partida para la definición arquitectónica de ADACOP.

Bakshi et al. [31] resaltan el crecimiento exponencial constante de los datos en la industria en los últimos años y algunas consideraciones a tener en cuenta a la hora de implementar sistemas de procesamiento y almacenamiento de grandes volúmenes de datos. A su vez, este crecimiento ha creado la necesidad de nuevas tecnologías como los sistemas de bases de datos no relacionales, que son altamente utilizados para el almacenamiento de metadatos, y sistemas de almacenamiento de datos distribuidos, para soportar grandes volúmenes de datos, sobre clústeres computacionales. El uso de estas dos tecnologías es imprescindible para ADACOP debido a la naturaleza de los datos a ser procesados. También es importante resaltar de este trabajo las recomendaciones de procesamiento para tener en cuenta como lo son la cantidad de nodos del

clúster que soporta la solución *Big Data* Hadoop, la concurrencia de usuarios y los requisitos de red necesarios entre los nodos del clúster.

Vergilio et al. [32] resaltan la importancia de una buena definición de los requerimientos no funcionales a la hora de diseñar e implementar soluciones *Big Data*. Este trabajo fue realizado tomando como referentes tres gigantes de la tecnología: Facebook, Twitter y Netflix. De esta investigación es importante resaltar que para ADACOP se tomaron en cuenta parte de los requerimientos no funcionales más importantes detallados por los autores.

Un caso de éxito de análisis de información de contratación estatal fue llevado a cabo por Surasvadi el al., [33] quienes analizaron datos abiertos de contratación en Tailandia con el fin de encontrar patrones en el gasto público de las diferentes entidades gubernamentales. Los autores definieron un *pipeline* analítico por medio del cual recolectaron, limpiaron y procesaron la información para ejecutar un proceso de clasificación y encontrar patrones de interés con el fin de visualizarlos en un tablero de control (*dashboard*). Los resultados de este caso demuestran que si es posible extraer *insights* de los datos de contratación pública para soportar procesos de tomas de decisiones. Sim embargo, en este trabajo no se tiene en cuenta una plataforma sobre la cual poner en producción los modelos analíticos desarrollados y mucho menos teniendo en cuenta que el gran volumen de información y su variabilidad hacen de este un problema a abordar como *Big Data*.

El trabajo de Agrahari et al. [34] resalta la necesidad de disponer la información en forma de visualizaciones para entender el comportamiento de los datos y hallar *insights* que permitan apalancar una mejor toma de decisiones y más aún en lo que respecta a datos generados por entidades gubernamentales. De igual manera resalta la necesidad de tecnologías aptas para no solo soportar los procesos de contratación estatal sino además de analizar este tipo de información. En el caso colombiano existen dos plataformas para la gestión de contratación estatal en las cuales las entidades del gobierno hacen públicos sus procesos. Por otra parte, también existe un portal de datos abiertos donde el gobierno publica los datos de las plataformas de contratación. Sim embargo existe la carencia de mecanismos y plataformas que permitan a la ciudadanía la consulta y análisis de esta información que también pueda ser utilizada para los gobernantes para soportar la toma de decisiones en materia presupuestal y gasto público.

# 4 Desarrollo de ADACOP

## 4.1 Selección de herramientas de desarrollo

La selección de las herramientas a utilizadas para el desarrollo de ADACOP dependen de diferentes factores, por una parte, se tienen las restricciones tecnológicas del entorno CAOBA las cuales afectan directamente el desarrollo de la solución. Por otra parte, el gran volumen de datos a extraer y procesar hace necesario el uso de tecnologías Big Data, área de conocimiento en la cual ADACOP es de gran contribución para la academia. Finalmente, se consideran las tecnologías como tal a utilizar de acuerdo con los requerimientos, necesidades y restricciones de los interesados. Actualmente existen muchos lenguajes de programación y *frameworks* de desarrollo para este tipo de proyectos y por ello fue necesario seleccionar aquellas que son las más adecuadas para la construcción de soluciones relacionadas con *Big Data*.

### 4.1.1 Evaluación de herramientas

Para el desarrollo de ADACOP fue necesario seleccionar las plataformas y leguajes de programación adecuados teniendo en cuenta las restricciones tecnológicas dentro de la Alianza CAOBA y las tecnologías existentes en *Big Data* que pueden soportar a ADACOP.

### 4.1.2 Restricciones tecnológicas de la Alianza CAOBA

La Alianza CAOBA cuenta con un servidor físico, cuyas características básicas se describen en la tabla 4.1, sobre el cual se ha realizado virtualización con el fin de poner a disposición máquinas virtuales que puedan ser usadas para configurar clústeres para Big Data. Utilizando soluciones tecnológicas como las ofrecidas por Cloudera y Hortonworks ya se han configurado ecosistemas de Hadoop para el desarrollo de herramientas en la Alianza CAOBA. Dado el alcance definido para ADACOP, se requiere de un clúster *Big Data* debidamente configurado para desarrollar los diferentes componentes y almacenar grandes volúmenes de información, por lo tanto, ADACOP es desarrollada sobre los recursos tecnológicos existentes dentro de CAOBA.

| | |
|---|---|
| Memoria RAM | **503 GB** |
| Disco Duro | 2.5 TB |
| Procesador | Intel(R) Xeon(R) CPU E5-2658 v4 @ 2.30GHz, 56 cores |

Tabla 4.1 - Características del servidor de CAOBA.

Utilizando la infraestructura tecnológica de CAOBA, se configuraron ocho máquinas virtuales para el despliegue del ecosistema Big Data. Cada una de estas máquinas cuentan con las características descritas en la tabla 4.2.

| | |
|---|---|
| Memoria RAM | 32 GB |
| Disco Duro | 50 GB |
| Cores de procesador | 4 cores |

Tabla 4.2 - Características máquinas virtuales.

### 4.1.3 Tecnologías para el desarrollo de ADACOP

Para el desarrollo de ADACOP fue necesario analizar, por una parte, las soluciones de Big Data que pueden ser implementadas para soportar el ecosistema Big Data y por otro lado los lenguajes de programación para el desarrollo de la herramienta.

### 4.1.4 Ecosistema de Big Data

El ecosistema de Big Data está conformado por un conjunto de plataformas desarrolladas y mantenidas por la comunidad de Apache y conocido como Hadoop [22]. *Big Players* tecno-

lógicos como Cloudera y Hortonworks ofrecen soluciones que facilitan la instalación y configuración de este ecosistema. Para ADACOP, fue necesario tener como mínimo disponibles las siguientes tecnologías *Big Data* del ecosistema Hadoop los cuales serán descritos en la sección 6.3.1 de este documento.

Adicionalmente, se considera el uso de tecnologías Big Data, principalmente, por tres factores. El primero de ellos va relacionado directamente al volumen de los datos pues inicialmente será de aproximadamente 8GB diarios, mensualmente aumentaría más o menos 1,2%, para que al final del primer año se tengan más de 3.1 TB contando solo las bases de datos de SECOP. El segundo corresponde a la experiencia previa de la Alianza CAOBA en el desarrollo de estas soluciones lo cual fue definido como requerimiento por parte del líder de analítica de CAOBA. El tercero, porque CAOBA desarrolló una herramienta de Anonimización de datos no estructurados [35], herramienta la cual fue desarrollada de la mano con el banco más grande del país, y donde se diseñó una arquitectura escalable para el procesamiento de información almacenada dentro de un ambiente distribuido.

- **HDFS**: mecanismo de almacenamiento.
- **Yarn**: administrador de recursos.
- **Spark**: mecanismo de procesamiento.
- **Hive**: mecanismo de consulta y acceso a la información.

### 4.1.5 Otros elementos considerados en la Selección de herramientas

Como se menciona anteriormente, ADACOP requiere la definición de la solución para la instalación y configuración del ecosistema de Big Data y de los lenguajes de programación que se requieren para la creación de los componentes de la herramienta.

La Alianza CAOBA ha venido desarrollado capacidades en cuanto al desarrollo de soluciones que abordan problemáticas de Big Data bajo el ecosistema de Hadoop. Por lo tanto, dentro de la Alianza ya se ha trabajado con la instalación y configuración de ecosistemas de

este tipo, principalmente con las soluciones tecnológicas ofrecidas por las empresas tecnológicas Cloudera y Hortonwors. Según la experiencia adquirida por CAOBA, la solución Big Data con mejores beneficios es la suite de Hortonworks llamada HDP (*Hortonworks Data Platform*), el principal de ellos su licenciamiento gratuito el cual permite tener a disposición el componente de software que ayuda a administrar las diferentes herramientas que hacen parte del ecosistema.

Los lenguajes de programación seleccionados para el desarrollo de ADACOP fueron definidos de acuerdo con las capacidades de desarrollo de CAOBA y en las tendencias de la industria. Según Siddiqui et al. [36], Python es un lenguaje famoso para desarrollar algoritmos para analítica de datos ya que proporciona una gran cantidad de librerías enfocadas en el procesamiento de datos y aprendizaje de máquina, además, los componentes *back-end* de los productos de software *Big Data* desarrollados por CAOBA se han realizado en este lenguaje, por lo cual se cuenta con esta capacidad dentro de la alianza. Adicionalmente ADACOP requiere de un componente de visualización por medio del cual el usuario interactuará con el sistema para la extracción de información y consulta de los gráficos generados por este. Para ello se seleccionó como lenguaje de programación Angular versión 7, ya que este permite fácil integración con el componente de *back-end* por medio de servicios web y es uno de los lenguajes que más se usan en la actualidad para el desarrollo de componentes de *front-end*.

Por último, ADACOP también requiere el desarrollo de un componente Big Data el cual permita la consulta y procesamiento de la información almacenada dentro del HDFS del ecosistema. Para esto se hace uso del componente de procesamiento del ecosistema Hadoop conocido como Apache Spark. Spark proporciona un lenguaje de programación llamado PySpark el cual permite escribir programas para manipular y procesar los datos existentes dentro del HDFS.

| *Componente* | *Tecnología* |
|---|---|
| *ADACOP Front End* | Angular 7 |
| *ADACOP Back End* | Python 3.7 |

| | |
|---|---|
| *ADACOP Big Data* | Python 2.7 / PySpark |

Tabla 4.3 - Tecnologías de ADACOP.

## 4.2 Identificación de requerimientos

ADACOP es una solución que presenta una serie de retos especialmente en la incertidumbre de sus requerimientos. De otra parte, el carácter innovador del proyecto, la integración de nuevas herramientas y tecnologías, podrían llegar a limitar las funcionalidades de la plataforma. Ésta es una de las razones por las que se opta por utilizar la metodología *scrum* (sección 4.4), ya que se permite un manejo mucho más consciente del cambio, y hace de éste un proceso evolutivo, que busca siempre la mejora continua.

Mediante reuniones, talleres y entrevistas con los clientes y usuarios, definidos en la sección sección 2.4, *scrum* nos permitió realizar una definición inicial de épicas (casos de usos) o funcionalidades de alto nivel, y derivarlas en una serie de historias de usuario (requerimientos) mucho más simples, cortas y fáciles de implementar. De esta manera, nos encontramos primero con las personas que van a tener un acercamiento y uso inicial de la plataforma (cliente CAOBA), y, en segundo lugar, con un grupo de expertos en el área que nos indicaron temáticas y características relevantes que debería tener un sistema como éste.

- CAOBA:
  - José Francisco Molano: Líder de analítica – Alianza CAOBA
  - Rafael Gonzalez: Director – Alianza CAOBA
- Expertos Observatorio Fiscal:
  - Luis Carlos Reyes Hernández: Director - Observatorio Fiscal PUJ; Profesor de cátedra - Departamento de Economía PUJ.
  - Adriana Francisca Salinas Esteban: Líder gasto público – Observatorio Fiscal PUJ; Profesora de cátedra - Departamento de Ciencia Política PUJ.

En la entrevista con el cliente CAOBA se identificaron 5 funcionalidades esenciales que se debería desarrollar en el sistema ADACOP: extracción, almacenamiento, procesamiento,

administración y visualización de datos (ver sección 4.2). Para los efectos de este proyecto/plan piloto, el tema de visualización y análisis se centrará únicamente en las tablas relacionadas con el tema de contratación estatal SECOP.

Por otro lado, surgieron requerimientos adicionales sugeridos por el grupo de expertos, especialmente en el tema de trazabilidad de la información, esto es, estar en la capacidad de observar el cambio a lo largo del tiempo entre versiones de una misma tabla, por ejemplo, la variación de los contratos en SECOP para analizar posibles incongruencias en la información. Además, surge el tema de la evaluación de la calidad de datos como un diferencial de nuestra plataforma frente a las soluciones actuales del mercado.

### 4.2.1 Prototipos

Con base en la metodología *Scrum*, un prototipo es un personaje ficticio altamente detallado que representa a la mayoría de los usuarios y *stakeholders* que pueden llegar a utilizar el producto final, dentro de ADACOP se identificaron 3 prototipos o perfiles de usuario fundamentales:

- Perfil 1: Alejandro es un joven de 27 años, con algunos conocimientos en manejo de datos, pero con un fuerte enfoque hacia los negocios y análisis de micro-data, encuentra mucho más enriquecedor un gráfico y/o visualización sencilla, y resultados concretos y definitivos que mucha información poco cohesionada. Además, lleva un estilo de vida bastante acelerado, por lo que gusta y necesita de respuestas bastante rápidas y acertadas para la toma de decisiones.
- Perfil 2: David es un profesional senior de 42 años de edad, experto en estadística y analítica de datos. Posee grandes conocimientos en aplicaciones, metodologías y técnicas en el manejo de datos, por lo que desea poder realizar análisis mucho más avanzados, y espera ante todo alto rendimiento y confiabilidad en los resultados e información entregada.
- Perfil 3: Andrea es una mujer de 35 años, profesional en áreas no relacionadas con tecnologías de la información, con un gran interés en explorar las bases de datos abiertos proporcionadas por el Gobierno y relacionadas con sus áreas de trabajo (sectores

ambientales, de la salud, política, entre otros). Desea una interfaz sencilla de manejar, adecuada para sus conocimientos básicos en estadística y manejo de datos.

### 4.2.2 Épicas

- Extracción de datos: como usuario deseo extraer diferentes tablas y conjuntos de datos desde la plataforma de datos abiertos del gobierno para su posterior análisis, enriquecimiento y comparación con fuentes de datos adicionales.
- Almacenamiento de datos: como usuario quiero poder almacenar la información para acceder a ella rápidamente, cuando quiera y desde donde quiera.
- Procesamiento de datos: como usuario necesito procesar, modificar y analizar los datos para encontrar información y comportamientos relevantes.
- Administración de datos: como usuario quiero administrar las tablas que he descargado para observar sus estadísticas, propiedades y metadatos.
- Visualización de datos – SECOP: como usuario necesito visualizar de manera gráfica los análisis y estadísticos de las tablas para comprender los resultados del procesamiento de la información.

### 4.2.3 Requerimientos funcionales

Una vez determinadas las épicas del proyecto, se procedió a descomponerlas en historias de usuario o requerimientos atómicos, no ambiguos y verificables.

#### 4.2.3.1 Extracción de datos

En la tabla 4.4 se describen los requerimientos de extracción definidos e implementados dentro del desarrollo de ADACOP.

| ID | Requerimiento | V | Estado |
|---|---|---|---|
| REQ_EXT_001 | El sistema debe extraer un conjunto de datos de la plataforma de datos abiertos del gobierno. | Volumen<br>Variedad | Implementado |

| | | | |
|---|---|---|---|
| REQ_EXT_002 | El sistema debe extraer los metadatos de un conjunto de datos de la plataforma de datos abiertos del gobierno. | Variedad | Implementado |
| REQ_EXT_003 | El sistema debe extraer información de Twitter. | Volumen | Implementado |
| REQ_EXT_004 | El sistema debe revisar periódicamente cambios en los metadatos de un conjunto de datos en la plataforma de datos abiertos del gobierno. | Variabilidad | Implementado |

Tabla 4.4 - Requerimientos de extracción.

#### 4.2.3.2 Almacenamiento de datos

En la tabla 4.5 se describen los requerimientos de almacenamiento definidos e implementados dentro del desarrollo de ADACOP.

| ID | Requerimiento | V | Estado |
|---|---|---|---|
| REQ_STG_001 | El sistema debe establecer una conexión con la base de datos de Mongo | | Implementado |
| REQ_STG_002 | El sistema debe establecer una conexión con el sistema de archivos distribuido HDFS. | | Implementado |
| REQ_STG_003 | El sistema debe almacenar archivos dentro del HDFS. | Volumen | Implementado |
| REQ_STG_004 | El sistema debe eliminar archivos dentro del HDFS. | Volumen | Implementado |
| REQ_STG_005 | El sistema debe almacenar los metadatos de una tabla en Mongo. | Variabilidad | Implementado |
| REQ_STG_006 | El sistema debe actualizar los metadatos de una tabla en Mongo. | Variabilidad | Implementado |

| | | | |
|---|---|---|---|
| REQ_STG_007 | El sistema debe eliminar los metadatos de una tabla en Mongo. | Variabilidad | Implementado |
| REQ_STG_008 | El sistema debe obtener los metadatos de una tabla en Mongo. | Variabilidad | Implementado |
| REQ_STG_009 | El sistema debe administrar el versionado de archivos en los metadatos. | Variabilidad | Implementado |
| REQ_STG_010 | El sistema debe administrar el versionado de tablas en los metadatos. | Variabilidad | Implementado |
| REQ_STG_011 | El sistema debe obtener el esquema de una tabla en los metadatos de Mongo. | Variabilidad | Implementado |
| REQ_STG_012 | El sistema debe guardar estadísticas del procesamiento de datos. | Velocidad | Implementado |

Tabla 4.5 - Requerimientos de Almacenamiento.

#### 4.2.3.3 Procesamiento de datos

En la tabla 4.6 se describen los requerimientos de procesamiento definidos e implementados dentro del desarrollo de ADACOP.

| ID | Requerimiento | V | Estado |
|---|---|---|---|
| REQ_PRC_001 | El sistema debe permitir la creación de una tabla desde un archivo del HDFS. | Volumen | Implementado |
| REQ_PRC_002 | El sistema debe permitir eliminar una tabla. | Volumen | Implementado |
| REQ_PRC_003 | El sistema debe generar la muestra de una tabla. | Velocidad | Implementado |
| REQ_PRC_004 | El sistema debe generar el histograma de una columna. | Velocidad | Implementado |

Tabla 4.6 - Requerimientos de Procesamiento.

#### 4.2.3.4 Administración de datos

En la tabla 4.4 se describen los requerimientos de administración de datos definidos e implementados dentro del desarrollo de ADACOP.

| ID | Requerimiento | V | Estado |
|---|---|---|---|
| REQ_ADM_001 | El sistema debe mostrar la lista de las tablas almacenadas. | Variedad | Implementado |
| REQ_ADM_002 | El sistema debe mostrar los metadatos de una tabla. | Variabilidad | Implementado |
| REQ_ADM_003 | El sistema debe presentar la muestra de una tabla. | Velocidad | Implementado |
| REQ_ADM_004 | El sistema debe mostrar el histograma para la columna de una tabla. | Velocidad | Implementado |
| REQ_ADM_005 | El sistema debe permitir agregar una nueva tabla con la URL del api en la plataforma de datos abiertos. | Variabilidad | Implementado |
| REQ_ADM_006 | El sistema debe permitir agregar una nueva tabla con su código en la plataforma de datos abiertos. | Volumen | Implementado |
| REQ_ADM_007 | El sistema debe permitir eliminar una tabla. | | Implementado |

Tabla 4.7 - Requerimientos de Almacenamiento.

#### 4.2.3.5 Visualización de datos - SECOP

| ID | Requerimiento | Estado |
|---|---|---|

| | | |
|---|---|---|
| REQ_VIS_001 | El sistema debe presentar las tablas relacionadas con SECOP. | Implementado |
| REQ_VIS_002 | El sistema debe mostrar un gráfico de agregación (mes, año, año-mes) por cantidad de contratos para una columna tipo fecha. | Implementado |
| REQ_VIS_003 | El sistema debe mostrar un gráfico para analizar las relaciones entre proveedor y contratista. | Implementado |
| REQ_VIS_004 | El sistema debe mostrar el índice de calidad de datos de una tabla. | Implementado |

Tabla 4.8 - Requerimientos de Validación SECOP.

### 4.2.4 Requerimientos no funcionales

- **Modularidad**: la arquitectura y funcionamiento de la aplicación es modular, y consta de un grupo de funcionalidades genéricas transversales a múltiples componentes, y algunas otras bastante especializadas, pero siempre garantizando bajo acoplamiento entre los mismos y alta cohesión. Además, es posible la implementación e integración de nuevos módulos de visualización específicos para realizar análisis más avanzados de datos, para el alcance de este proyecto solo se desarrolló un módulo de visualización enfocado a SECOP pero a futuro podrá expandirse a set de datos diferentes de la plataforma de datos abiertos [37]. Esta característica implementada en ADACOP permite que puedan ser integrados nuevos módulos de procesamiento de datos para enriquecer las funcionalidades del sistema.
- **Reusabilidad**: capacidad y facilidad para reutilizar componentes, módulos y elementos dentro del código en futuras implementaciones [37]. Se ve reflejado en los componentes de acceso y procesamiento a información que fueron desarrollados como parte de ADACOP y que permiten la reutilización de funcionalidades por medio de *end points*.
- **Robustez y tolerancia a fallos**: el sistema debe estar en la capacidad de continuar operando, aunque uno o más nodos del sistema fallen [32]. Teniendo en cuenta que los

sistemas distribuidos implican retos en consistencia, duplicidad de datos y concurrencia [38]. Esto puede ser medido por medio de la cantidad de nodos del clúster que pueden fallar pero que sus esquemas de replicación permiten el correcto funcionamiento de la solución.

- **Escalabilidad**: se requiere mantener el rendimiento de la aplicación al incrementar los datos o la carga y añadiendo más recursos al sistema. Se busca especialmente escalabilidad de tipo horizontal, es decir, agregando más maquinas al *stack* del sistema [38]. La escalabilidad de ADACOP puede ser medida por la cantidad de usuarios concurrentes en el sistema, la cantidad de datos almacenados y procesados y la cantidad de nodos que pueden ser agregados como parte del clúster donde se despliega la solución.
- **Mantenimiento**: se debe procurar disminuir la complejidad de implementación de los componentes del sistema de ADACOP, con el fin de facilitar el mantenimiento del código en caso de una actualización o fallos [38]. La arquitectura modular de ADACOP permite que el mantenimiento pueda ser llevado a cabo solo en aquellos de componentes que lo requieran.

| ID | Requerimiento |
|---|---|
| REQ_NOF_001 | Modularidad |
| REQ_NOF_002 | Reusabilidad |
| REQ_NOF_003 | Robustez y tolerancia a fallos |
| REQ_NOF_004 | Escalabilidad |
| REQ_NOF_005 | Mantenimiento |

Tabla 4.9 - Requerimientos de No funcionales ADACOP.

## 4.3 Arquitectura de ADACOP

ADACOP es un portal web para la consulta y análisis de datos abiertos relacionados con la contratación estatal del gobierno de Colombia y es desarrollado como parte de los trabajos

de grado de las Maestrías en Ingeniería de Sistemas y Computación y Analítica para la Inteligencia de Negocios cursadas por los autores en la Pontificia Universidad Javeriana. Es desarrollado como parte del portafolio de proyectos definido por el centro de excelencia y apropiación en *Big Data* y *Data Analytics*, Alianza CAOBA. ADACOP tiene como objetivo poner a disposición un portal web por medio del cual los *stakeholders* pueden descargar, analizar y visualizar información proveniente del portal de datos abiertos del gobierno de Colombia relacionados con la contratación pública con el fin de soportar la transparencia en estos procesos y mejorar el acceso a esta información y así empoderar a los interesados en los procesos de contratación pública. Este proyecto se diseñó como una solución *Big Data* dado el gran volumen de información que es procesada y almacenada por el sistema. Además, se pretende generar nuevo conocimiento relacionado con *Big Data* y *Data Analytics* como parte de la misión de CAOBA.

### 4.3.1 Ecosistema *Big Data*

ADACOP en una plataforma que almacena grandes volúmenes de datos para procesarlos y generar información de interés, según fue definido como parte de sus requerimientos. Para abordar el diseño e implementación de un software que cumpla con estos requisitos, se hizo uso de tecnologías relacionadas con *Big Data*. En esta sección se describe el ecosistema *Big Data* utilizado para la construcción de ADACOP y los componentes que lo conforman.

Hadoop es actualmente el ecosistema de *Big Data* por excelencia. Este ecosistema de *Big Data* está conformado por un conjunto de programas y procedimientos de uso libre desarrolladas y mantenidas por la comunidad de Apache [39]. *Big players* tecnológicos como Cloudera [40] y Hortonworks [41] ofrecen soluciones que facilitan la instalación y configuración de este ecosistema; sobre se almacena y procesa información, además permite el despliegue de componentes tecnológicos para tareas específicas de manipulación de datos. Hadoop es instalado sobre un clúster de computadores interconectados entre sí, permitiendo de esta manera escalar horizontalmente agregando más nodos de computación cuando sea necesario. Ya se han propuesto soluciones en ecosistemas *Big Data* para el procesamiento y almacenamiento de in-

formación estatal como en el caso de Bhushan et al. [30], quienes implementaron una arquitectura *Big Data* para ejecutar procesos analíticos sobre información de gobierno digital utilizando Hadoop.

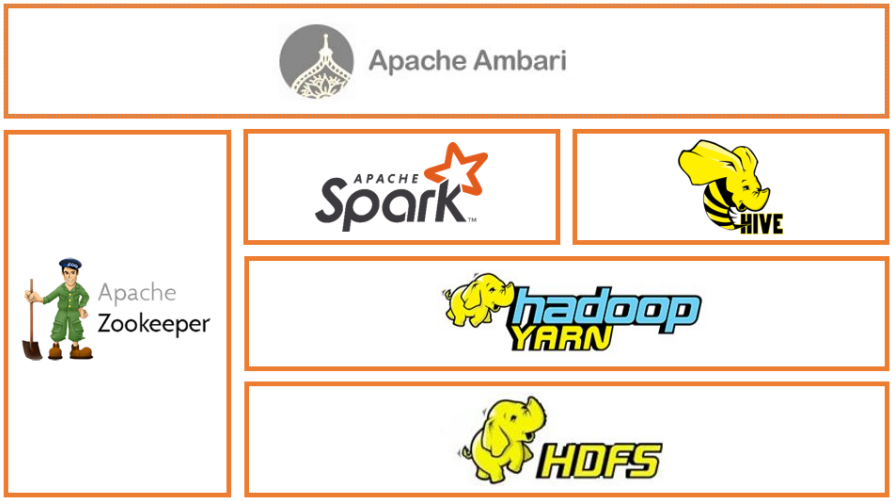


Figura 4 - Ecosistema *Big Data*. Fuente: Elaboración propia.

En la figura 4 se encuentran los componentes tecnólogos que hacen parte del ecosistema, los cuales son necesarios para el desarrollo de ADACOP teniendo en cuenta los requerimientos definidos. Sin embargo, el ecosistema puede tener muchos más componentes los cuales no se detallarán en este documento dado que no son necesarios para la construcción y puesta en marcha de ADACOP. A continuación, se describen los componentes del ecosistema y su funcionalidad.

**HDFS**: el HDFS es un sistema de almacenamiento distribuido el cual permite persistir grandes cantidades de datos sobre un clúster de *Big Data* de tal forma que los conjuntos de datos son particionados y almacenados a lo largo de todos los nodos del clúster [23]. Para ADACOP, este componente será utilizado con el fin de almacenar toda la información proveniente del portal de datos abiertos y los conjuntos de datos resultantes de las tareas de procesamiento.

**Yarn**: yarn es el componente que permite una administración centralizada de todos los nodos que hacen parte del clúster de *Big Data* y es el encargado de gestionar sus recursos [42].

Además, por medio de este componente es posible añadir y eliminar nodos al clúster mitigando el riesgo de pérdida de información.

**Spark**: spark es el motor de procesamiento que permite ejecutar tareas sobre los datos distribuidos a lo largo de los diferentes nodos del clúster [43]. Sobre este componente se desarrolla el módulo de ADACOP encargado de la manipulación, procesamiento y limpieza de datos.

**Hive**: hive permite consultar la información existente dentro del clúster por medio de lenguaje SQL, dado que permite asociar un esquema de datos a los archivos planos que contienen la información como tal y que están distribuidos a lo largo del clúster [24].

**Zookeeper**: este componente tiene como funcionalidad centralizar la información de configuración de los distintos nodos del clúster, permitiendo administrar de forma fácil y ágil los diferentes nodos [44].

**Ambari**: este componente permite una fácil administración e instalación de los diferentes componentes dentro de los nodos del clúster. Este fue utilizado inicialmente para toda la configuración y despliegue del ecosistema *Big Data* sobre el cual es desarrollado ADACOP [45].

Como se menciona anteriormente, las dos empresas pioneras en el desarrollo de este tipo de ecosistemas con Hortonworks [41] y Cloudera [40]. En CAOBA ya se han realizado diferentes proyectos relacionados con *Big Data* y se cuenta con experiencia en este tipo de soluciones y ecosistemas. Teniendo esto en cuenta, para ADACOP se utiliza la suite de Hortonworks la cual tiene una licencia de libre uso sin soporte y permite configurar el clúster más rápidamente.

### 4.3.2 Modelo de arquitectura propuesto

Basado en las restricciones tecnológicas de la Alianza CAOBA descritas en la sección 4.1.2 y en los requerimientos definidos para ADACOP descritos en la sección 4.2, se propone un arquitectura de solución *Big Data* basado en capas motivado en la arquitectura de referencia

propuesta por Pääkkönen et al. [46]. En la figura 5 se relacionan las diferentes capas de la arquitectura y los componentes de ADACOP que la soportan.

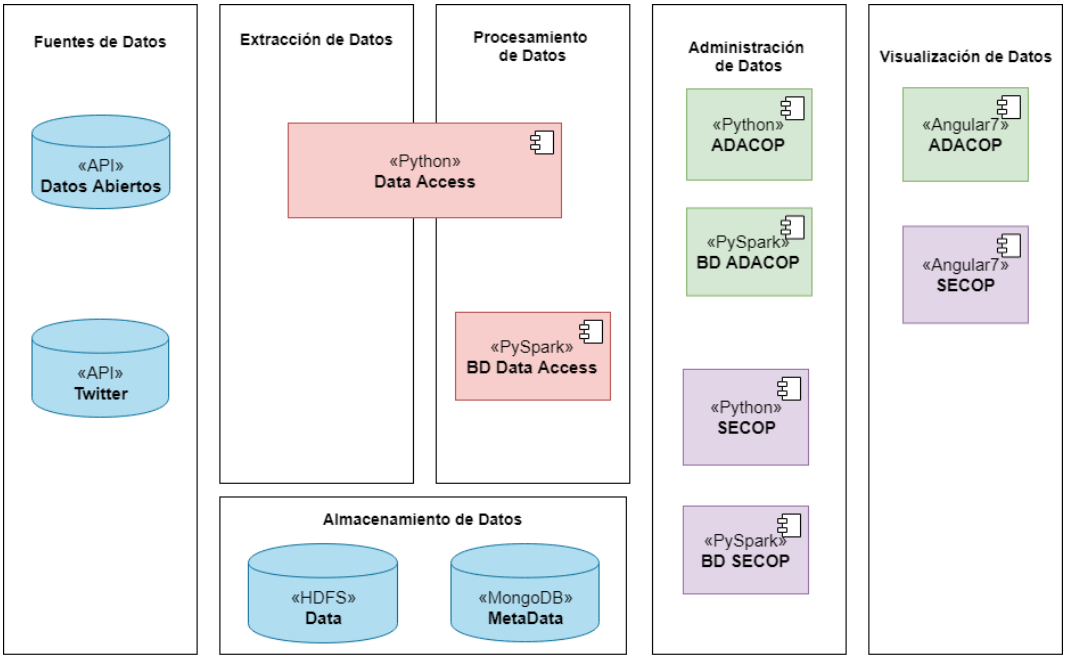


Figura 5 - Arquitectura de Capas de ADACOP. Fuente: Elaboración propia.

Al ser una herramienta que realiza tareas de obtención, procesamiento y almacenamiento de datos, se construyen componentes motivados en brindar servicios por capas, con el fin de facilitar la abstracción del flujo de datos correspondiente, ya que en cada capa se realizan tareas específicas de acuerdo con sus responsabilidades. Los componentes que soportaran la solución se pueden ver en la figura 6.

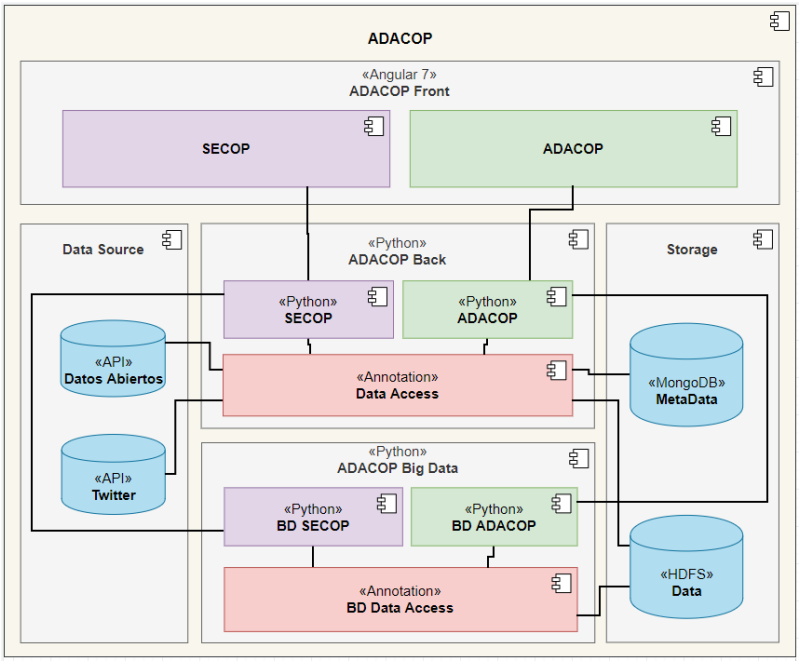


Figura 6 - Arquitectura de Componentes de ADACOP. Fuente: Elaboración propia.

En la figura 6 se presenta el diagrama correspondiente a la arquitectura lógica de ADACOP con los diferentes componentes que soportan esta arquitectura basada en capas. A continuación, se describen las diferentes capas de la arquitectura y como los diferentes componentes desarrollados soportan las responsabilidades de estas.

#### 4.3.2.1 Fuentes de datos

El alcance de ADACOP contempla la extracción de dos fuentes principales identificadas, por una parte, el portal de datos abiertos de Colombia en el cual diferentes entes gubernamentales ponen a disposición de la ciudadanía conjuntos de datos con información pública [18] y por otro lado información proveniente de la red social twitter.

El portal de datos abiertos de Colombia [18] ofrece, en su mayoría, conjuntos de datos tabulares los cuales pueden ser accedidos y descargados por medio del API disponible para tal fin. Esta API está construida con una tecnología de acceso libre llamada *socrata* [47] la cual expone un servicio para extraer los conjuntos de datos y los metadatos asociados al mismo.

La red social twitter [48] de igual manera ofrece un API para acceder a la información que es publicada por sus usuarios [49]. Esta API permite extraer tweets en tiempo real de acuerdo a un conjunto de palabras clave que le sean dadas y, adicionalmente, permite consultar un historial de estas palabras.

#### 4.3.2.2 Extracción de Datos

ADACOP contiene un módulo llamado ADACOP Back (como se muestra en la figura 6) el cual contiene un módulo llamado Data Access que contiene credenciales de acceso a los diferentes API. Este componente extrae la información proveniente de datos abiertos y de twitter según lo requiera el usuario desde la interfaz gráfica de la herramienta. Este proceso de extracción es orquestado por medio del módulo ADACOP quien a su vez expone por medio de servicios REST diferentes servicios a la interfaz gráfica de la solución.

El componente de extracción recibe la URL correspondiente al conjunto de datos que el usuario desea descargar. Con esta URL el componente de extracción crea una tarea por medio de un programador (*scheduler*). Este programador, que hace parte de ADACOP Back, realiza el proceso de extracción del conjunto de datos y metadatos cada vez que detecta una actualización en el portal de datos abiertos realizando una revisión cada veinticuatro horas. ADACOP permite descargar cualquier conjunto de datos proveniente del portal por medio de este componente con sus respectivos metadatos.

Para la extracción de twitter el componente de extracción recibe un archivo con las palabras clave a descargar de acuerdo con lo definido por el usuario. Dado que el alcance inicial de ADACOP es descargar y procesar información relacionada con la contratación pública, el sistema contiene un archivo con las palabras correspondientes a los contratistas más representativos (palabras seleccionadas por expertos) y el sistema extrae esta información en línea. Sin embargo, este componente permite extraer tweets de cualquier tema según las palabras clave que sean ingresadas.

*4.3.2.2.1 Manejo de versiones*

La trazabilidad en la información es una de las funcionalidades clave de ADACOP, ya que aparte de permitirnos manejar diferentes versiones de una misma tabla y gestionarlas dentro de los metadatos del proyecto, permite evaluar diferentes métricas de calidad que relacionan las diferencias entre una versión y otra.

Este módulo parte de un *scheduler* parametrizable, que puede lanzarse a una hora especifica del día y con una periodicidad definida (por defecto se ejecuta a las 00:00 am todos los días), es decir, realiza procesamiento en *batch* o por lotes. Este módulo se encarga de comparar la *metadata* original extraída de datos abiertos y la compara con la almacenada en los metadatos de ADACOP, si encuentra diferencias entre ambos registros procede a realizar la descarga de la nueva versión.

Dentro de este proceso de actualización surgieron algunos retos, especialmente cuando se encontraron diferencias significativas entre una versión a otra, como, por ejemplo, cambios en los nombres de las columnas, o incluso la presencia de nuevas variables, ya que como se había nombrado con anterioridad existen una serie de métricas de trazabilidad que se verían afectadas en su cálculo. En estos casos, es posible analizar dichas variaciones únicamente entre columnas con el mismo nombre, si se quisiese hacer un manejo mucho más profundo de dichos cambios podría necesitarse un operador que realizará las homologaciones de manera manual. Pero esto podría contravenir la simplicidad y usabilidad por parte del usuario (sea avanzado o no) y facilidad de mantenimiento del sistema. Se toma la decisión de definir algunas métricas de calidad adicionales que permitan observar estos cambios estructurales en los metadatos de la tabla.

#### 4.3.2.3 Procesamiento de Datos

Como se menciona anteriormente, los componentes desarrollados de ADACOP permiten la extracción de cualquier conjunto de datos proveniente del portal de datos abiertos. Sin embargo, para el procesamiento de los datos, se implementaron componentes los cuales procesan específicamente las bases de datos de contratación estatal SECOP. Esto conlleva a que ADACOP no solo puede procesar información de contratación pública, sino que, además, el

modularidad de su arquitectura permite la integración de nuevos componentes de procesamiento de acuerdo con las necesidades de CAOBA.

El procesamiento de los datos de SECOP es realizado por sus respectivos componentes y está dividido en dos partes. Por una parte, se encuentra el proceso de tomar los datos que son extraídos de las fuentes de datos, persistirlos dentro del HDFS y crear la tabla en Hive asociándola a su esquema y haciéndola asequible para todos los componentes del ecosistema *Big Data*. Esta primera parte del procesamiento es realizada exclusivamente por el componente de ADACOP y puede hacerse para cualquier conjunto de datos del portal de datos abierto. Por otra parte, una vez creada la tabla en Hive, se ejecutan los procesos de creación de las diferentes estadísticas de calidad según lo definido dentro de los requerimientos para SECOP. Para ello, los componentes de SECOP toman las respectivas tablas para generar las tablas de estadísticas y de limpieza que de igual manera quedan persistidas dentro de Hive.

#### 4.3.2.4 Almacenamiento de Datos

ADACOP consta de dos componentes de almacenamiento como se muestra en la figura 6. Los metadatos generados y extraídos de las diferentes fuentes son almacenados en una base de datos NoSQL, MongoDB, en archivos de formato Json.

Los datos son almacenados dentro del sistema de archivos distribuido del ecosistema *Big Data* y son asociados a los metadatos almacenados para poder tener acceso de forma fácil a estos.

#### 4.3.2.5 Administración de Datos

La figura 5 muestra los componentes que hacen parte de esta capa y la figura 6 muestra cómo están distribuidos dentro de la arquitectura de ADACOP. El objetivo principal de estos componentes es orquestar los procesos de extracción, procesamiento y consulta de los resultados de estos procesos los cuales son solicitados desde los componentes de visualización.

#### 4.3.2.6 Visualización de Datos

Al igual que los componentes del *back-end* de ADACOP, el componente de visualización consta de dos grandes módulos correspondientes a SECOP y ADACOP (figura62). ADACOP permite al usuario administrar el contenido de información almacenada en el sistema

además de realizar un análisis exploratorio básico por medio de la generación de histogramas según lo solicite el usuario. SECOP permite visualizar al usuario los *dashboards* (tableros de control) integrados como parte de este componente. El sistema cuenta con cinco *dashboards*.

1. ADACOP
2. Dashboards SECOP I
3. Dashboard SECOP II
4. Dashboard de Calidad

### 4.3.3 Flujo de Datos

Los componentes que conforman la arquitectura de ADACOP tienen como fin soportar el pipeline analítico desde la extracción de los datos, pasando por la creación de las estadísticas generadas hasta la visualización de los *dashboards*. Como se menciona en los requerimientos de ADACOP (sección 6.2), el sistema está en la capacidad de, primeramente, extraer cualquier conjunto de datos proveniente del portal de datos abiertos de Colombia y de procesar los datos correspondientes a las bases de datos de contratación estatal disponible en el portal. Es por esta razón, que dentro de la arquitectura descrita en la figura 6, se observan componentes independientes de ADACOP y SECOP. Los componentes ADACOP permiten la administración del sistema y la extracción general de información y los de SECOP permiten crear y ejecutar procesos analíticos específicos de estos datos. La principal ventaja de esta desagregación de responsabilidades por medio de módulos es la posibilidad de integrar nuevos componentes analíticos para nuevas fuentes de interés. Teniendo en consideración esta separación de componentes y responsabilidades dentro de la arquitectura, se desarrollaron los siguientes flujos de datos:

#### 4.3.3.1 Extracción de información

En la figura 7, se describe el flujo del proceso de extracción de datos provenientes del portal de datos abiertos desde el momento en que el usuario ingresa la URL correspondiente hasta que la información y sus respectivos metadatos son almacenados y estructurados dentro de hive (recuadro de extracción).

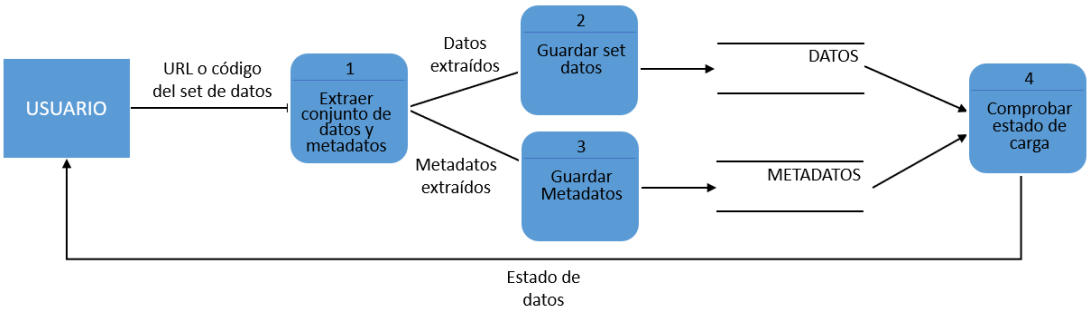


Figura 7 . Flujo de datos de extracción en ADACOP. Fuente: Elaboración propia.

### 4.3.3.2 Procesamiento de datos

Los componentes de procesamiento de datos corresponden a los relacionados con SECOP y tienen dos funcionalidades principales, la creación de diferentes estadísticas de calidad de la información de contratación pública y la limpieza de estos. Dentro del proceso de limpieza de igual manera se crean las tablas de hechos (*fact tables*) que posteriormente serán utilizadas para soportar los respectivos *dashboards* en el *front-end* de ADACOP. Esta información generada de igual manera es persistida dentro del componente de almacenamiento y relacionada por medio de los metadatos correspondientes como se describe en el flujo de la figura 8 (recuadro de procesamiento). Este proceso se ejecuta una vez se extrae el conjunto de datos correspondiente.

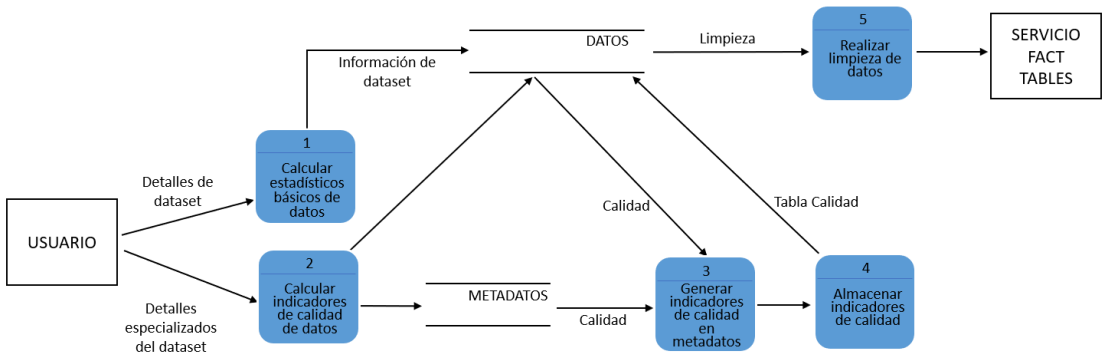


Figura 8. Flujo de datos de procesamiento en ADACOP. Fuente: Elaboración propia

#### 4.3.3.3 Administración de datos

El componente de ADACOP permite la administración de datos por medio de la interfaz gráfica del sistema donde el usuario puede visualizar la información que tiene disponible (ver figura 9). Adicionalmente el componente encargado de la administración es el que crea los procesos de descarga automática por medio de un *scheduler* el cual obtiene una nueva versión de la información cada vez que esta es actualizada dentro del portal de datos abiertos. Cuando el sistema revisa cada 24 horas los datos y encuentra una actualización, esta versión se descarga por medio del flujo de extracción, guardando así diferentes versiones de los conjuntos de datos y estadísticas generadas.

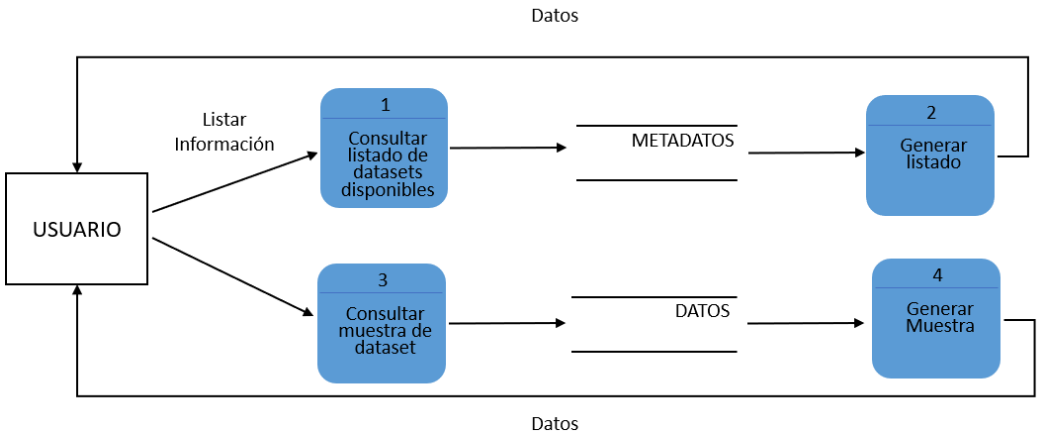


Figura 9 . Flujo de datos de administración ADACOP. Fuente: Elaboración propia.

#### 4.3.3.4 Visualización de datos – SECOP

Tanto el componente de ADACOP como el de SECOP brindan la posibilidad al usuario de visualizar e interactuar con diferentes gráficos para poder analizar descriptivamente la información, como se muestra en la figura 10.

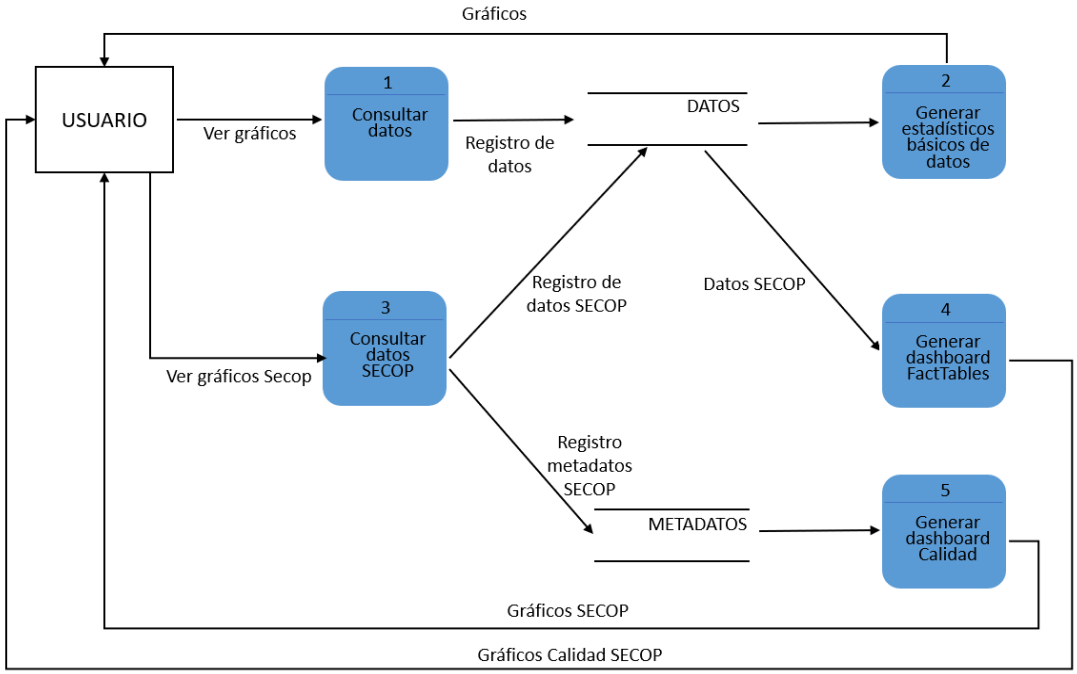


Figura 10. Flujo de datos de Visualización ADACOP. Fuente: Elaboración propia

#### 4.3.3.5 Prototipo resultante

A continuación, se muestran las interfaces de usuario resultantes:

a. Pantalla inicial, menú principal.

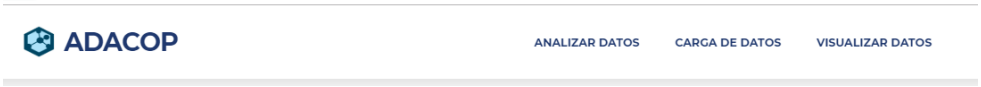


Figura 11. Menú principal ADACOP Fuente: Elaboración propia.

b. Análisis de datos. Módulo de datos descriptivos de fuentes de SECOP y de otras fuentes cargadas.


Figura 12. Módulo 1. Análisis de datos ADACOP. Fuente: Elaboración propia.

- Análisis descriptivos SECOP. En esta sección se muestran los dashboard provenientes de SECOP I y SECOP II.


Figura 13. Sección análisis de datos ADACOP. Fuente: Elaboración propia.

- SECOP

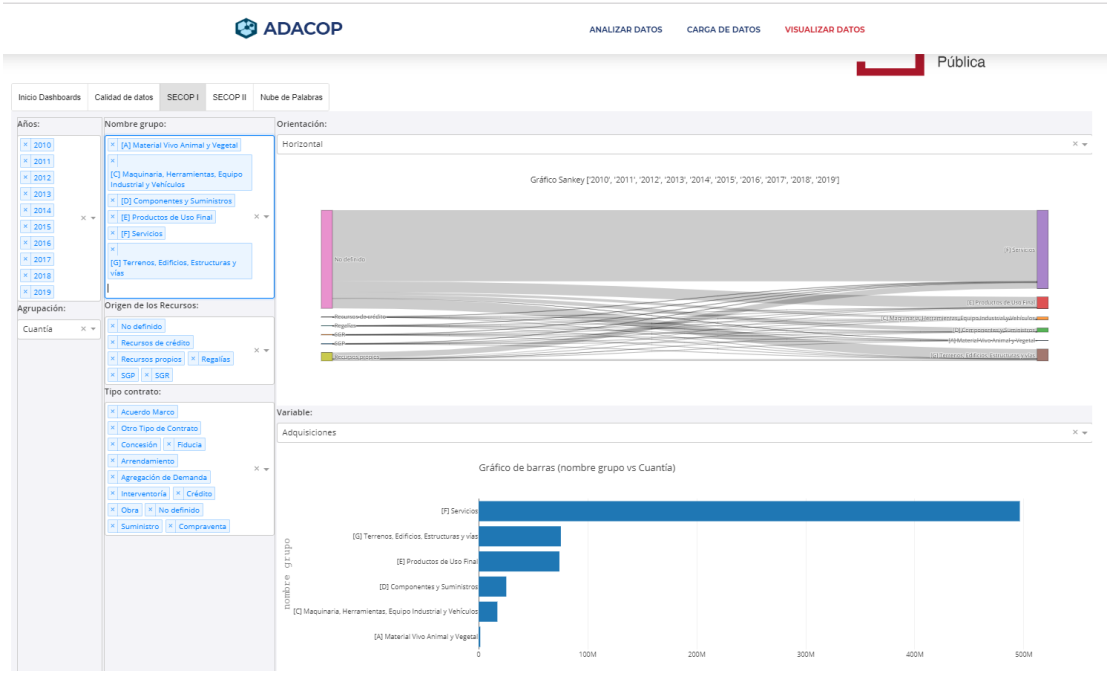


Figura 14. Sección análisis de datos SECOP. Fuente: Elaboración propia.

- Calidad de datos. Comparación de trazabilidad de las versiones cargadas de SECOP.

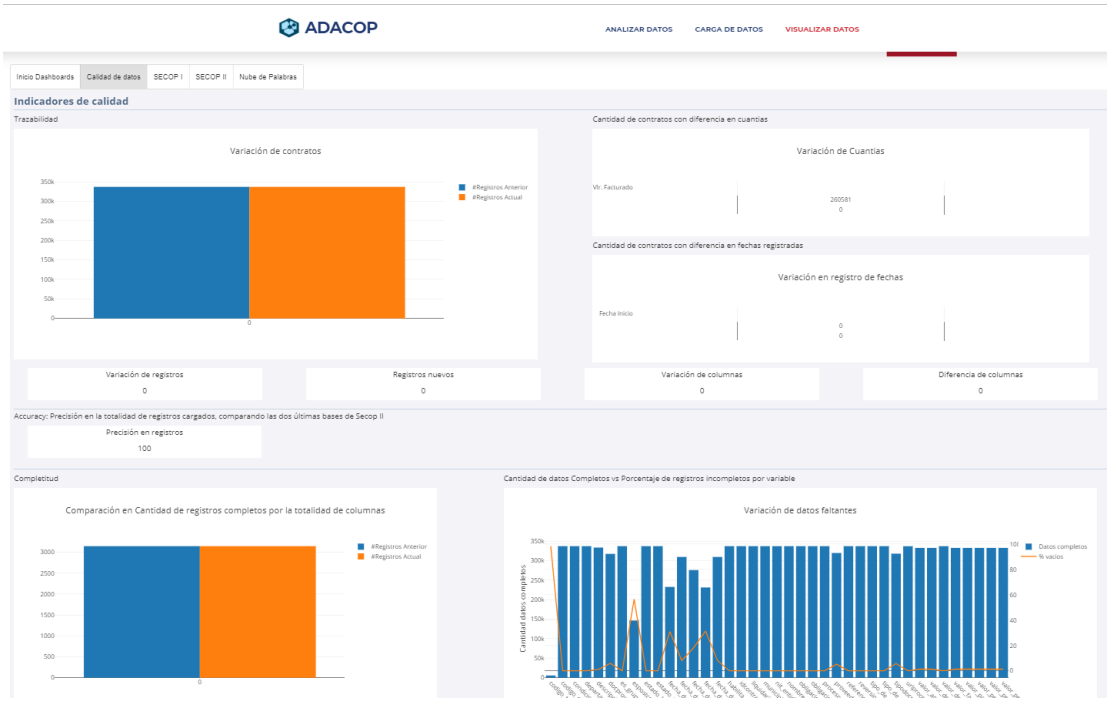

Figura 15. Calidad de datos SECOP. Fuente: Elaboración propia.

- Nube de palabras (Twitter). Palabras frecuentes extraídas de Twitter


Figura 16. Gráfico fuentes externas SECOP. Fuente: Elaboración propia.

c. Carga de datos, fuentes de datos. Antes de realizar la interacción con este módulo, se debe seleccionar la fuente de datos a integrar proveniente de datos abiertos, una vez sea seleccionada se debe pegar la URL del navegador dentro del campo Cargar Datos.

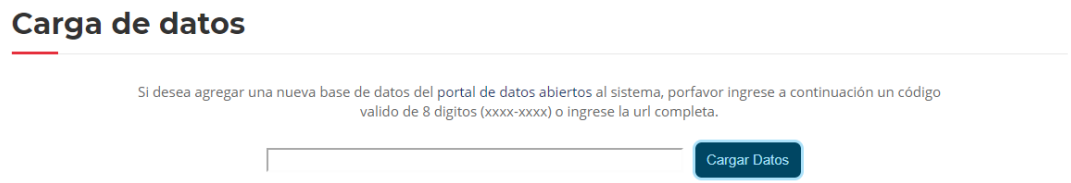


Figura 17. Carga de datos ADACOP. Fuente: Elaboración propia.

d. Visualización de datos. Lisa de bases cargadas y listas para usar.

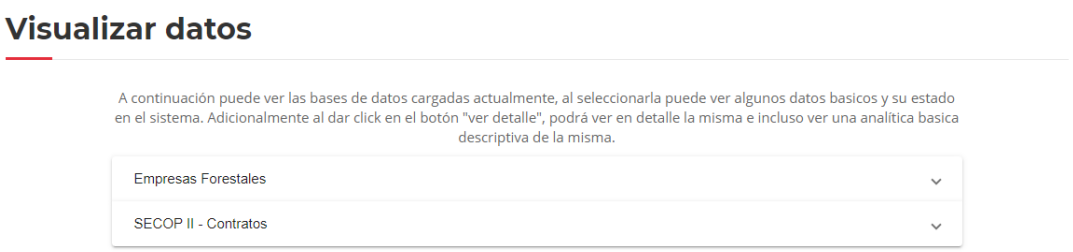


Figura 18. Visualización de datos ADACOP. Fuente: Elaboración propia.

- En la lista se muestran los detalles de los datos extraídos, el estado para su uso y la fecha de la última actualización en datos abiertos.

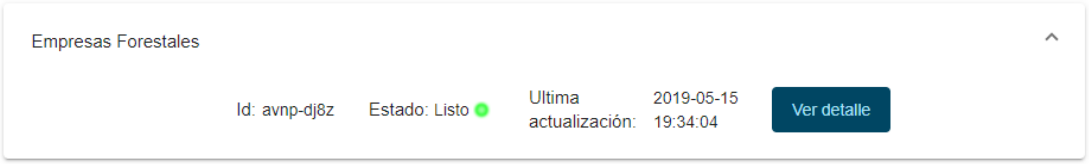


Figura 19. Sección lista de fuentes cargadas. Fuente: Elaboración propia.

- Ver detalles.

  En el encabezado se muestra el detalle de los metadatos, las variables y el contenido del conjunto de datos.


Figura 20. Detalle conjunto de datos. Fuente: Elaboración propia.

- Analizar datos.

  Se realiza una gráfica de distribución sencilla, para el ejemplo, se muestra el conteo de los registros clasificados por tipo.

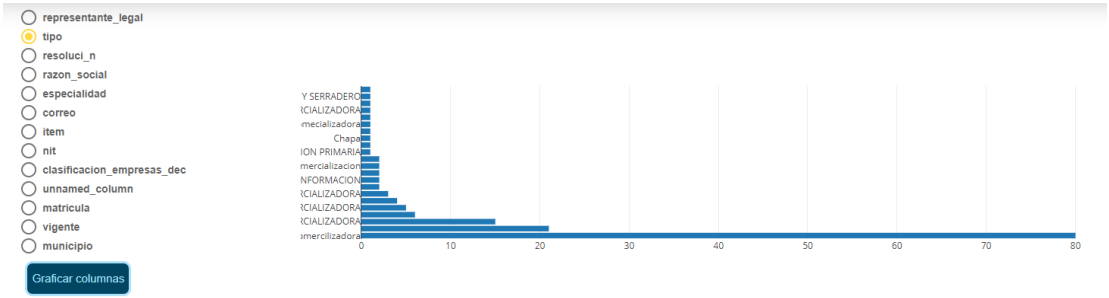


Figura 21. Descriptivos de los conjuntos de datos. Fuente: Elaboración propia.

# 5 Resultados y Validación

## 5.1.1 Resultados de Verificación

Consiste en definir las "V" del *Big Data* en las cuales serán agrupados los requerimientos, las "V" para tener en cuenta son Volumen, Velocidad, Variedad y Variabilidad. Posteriormente se procede a evaluar cada requerimiento con cual componente está relacionado y soportado/medirlo para su cumplimiento. Esta tarea se realizará tanto para requerimientos funcionales como para no funcionales (Ver 22).

| | Requerimientos | Volumen | Variedad | Variabilidad | Velocidad |
|---|---|---|---|---|---|
| FUNCIONALES | REQ_EXT_001 | ✓ | ✓ | | |
| | REQ_EXT_002 | ✓ | ✓ | | |
| | REQ_EXT_003 | ✓ | | | |
| | REQ_EXT_004 | ✓ | | ✓ | ✓ |
| | REQ_STG_003 | ✓ | | | ✓ |
| | REQ_STG_004 | ✓ | | | |
| | REQ_STG_005 | ✓ | | ✓ | ✓ |
| | REQ_STG_006 | ✓ | | ✓ | ✓ |
| | REQ_STG_007 | | | ✓ | ✓ |
| | REQ_STG_008 | | | ✓ | ✓ |
| | REQ_STG_009 | ✓ | | ✓ | |
| | REQ_STG_010 | ✓ | | ✓ | |
| | REQ_STG_011 | | | ✓ | |
| | REQ_STG_012 | | | | ✓ |
| | REQ_PRC_001 | ✓ | | | ✓ |
| | REQ_PRC_002 | ✓ | | ✓ | |
| | REQ_PRC_003 | | | | ✓ |
| | REQ_PRC_004 | | | | ✓ |
| | REQ_ADM_001 | | ✓ | | |
| | REQ_ADM_002 | ✓ | | ✓ | |
| | REQ_ADM_003 | | | ✓ | ✓ |
| | REQ_ADM_004 | | | | ✓ |
| | REQ_ADM_005 | ✓ | | ✓ | |
| | REQ_ADM_006 | ✓ | | | |

| | | | | | |
|---|---|---|---|---|---|
| NO FUNCIONALES | Modularidad | | | | |
| | Reusabilidad | | | | |
| | Robustez | | | | |
| | Escalabilidad | | | | |
| | Mantenimiento | | | | |

Figura 22. Componentes de *Big Data* vs Requerimientos Definidos. Fuente: Elaboración propia.

De la anterior evaluación, se realiza la asignación de la V correspondiente a cada requerimiento, ver sección 4.2.3.

### 5.1.2 Validaciones

Como metodología de validación, teniendo en cuenta que desde hace varios años se han desarrollado una serie de modelos para validar la aceptación de cualquier tecnología, pero el Modelo de Aceptación Tecnológica (TAM) desarrollado por Fred D. Davis Jr. en 1985 destaca desde sus inicios, pues ha sido comprobado y reconocido por ser un modelo efectivo, convirtiéndose en uno, si no el más utilizado. Este modelo busca evaluar las variables utilidad percibida, facilidad de uso percibida y la intención hacia el uso en su versión inicial, más adelante se le agregaron algunas variables externas como lo son la influencia social y las condiciones que facilitan el uso de la tecnología[50]. Para los propósitos de esta validación, de las pruebas TAM, nos interesan los siguientes ejes:

- La Utilidad Percibida (PU) se refiere al grado en que una persona cree, que ADACOP en particular, mejorará su desempeño en el trabajo.
- Intención hacia el Uso (BI): Grado en el que una persona ha formulado planes conscientes para desarrollar (o no) alguna conducta futura utilizando ADACOP.

Adicionalmente, otro de los modelos de aceptación de sistema, este enfocado exclusivamente en la usabilidad percibida, se encuentra el modelo de Escala de Usabilidad del Sistema (SUS), desarrollado en 1986 por John Brooke. Como se mencionó anteriormente este modelo se centra en medir la usabilidad del sistema, y para esto cuenta con 10 preguntas en las que se

encuentran algunas positivas y otras negativas, que se tendrán en cuenta de manera inversa para el cálculo del resultado final [51]. Y este es uno de los más, sino el más utilizado para la evaluación de aceptación de dashboards.

Para el sistema ADACOP, las variables que son relevantes son la Utilidad y la Usabilidad, por este motivo y teniendo en cuenta que lo que el usuario final tendrá un mayor enfoque a analizar el dashboard desplegado en el sistema, se seleccionó el modelo SUS para medir la usabilidad. Por otro lado, el modelo TAM se seleccionó para validar la utilidad de ADACOP teniendo en cuenta las variables de utilidad percibida y la intención hacia el uso.

#### 5.1.2.1 Encuesta adaptada para ADACOP

Para predecir el uso de ADACOP por parte de la población de usuarios para la cual fue desarrollada, el modelo TAM se adapta para responder a las variables más relevantes para el sistema, apoyado con el modelo SUS para evaluar la usabilidad de este. En la tabla 1 se presentan los ítems utilizados para determinar cada variable.

| Modelo | Factor | Item | Código |
|---|---|---|---|
| Utilidad (TAM) | Utilidad percibida | El uso de ADACOP para analizar datos abiertos del sector público me beneficia. | PU1 |
| | | El uso de ADACOP en datos abiertos del sector público me permitirá realizar mi investigación más rápidamente. | PU2 |
| | | Usar ADACOP en datos abiertos del sector público aumentará mi productividad. | PU3 |
| | | El uso de ADACOP en datos abiertos del sector público mejora mi desempeño en mi trabajo. | PU4 |
| | Intención hacia el uso | Tengo la intención de utilizar la herramienta ADACOP en el futuro para analizar información de datos abiertos. | BI1 |
| | | Planeo usar la herramienta ADACOP en el futuro para analizar datos abiertos. | BI2 |

| | | | |
|---|---|---|---|
| | | Encuentro ADACOP útil en mi trabajo. | BI3 |
| Usabilidad (SUS) | Usabilidad | Creo que me gustaría usar el sistema ADACOP con frecuencia. | SUS1 |
| | | Encontré el sistema ADACOP innecesariamente complejo. | SUS2 |
| | | Pensé que el sistema ADACOP era fácil de usar. | SUS3 |
| | | Creo que necesitaría el apoyo de un técnico para poder utilizar ADACOP. | SUS4 |
| | | Encontré que las diversas funciones en ADACOP estaban bien integradas. | SUS5 |
| | | Pensé que había demasiada inconsistencia en el sistema ADACOP. | SUS6 |
| | | Me imagino que la mayoría de las personas aprenderían a usar ADACOP muy rápidamente. | SUS7 |
| | | Encontré el sistema ADACOP muy incómodo de usar. | SUS8 |
| | | Me sentí muy seguro usando el sistema ADACOP. | SUS9 |
| | | Necesitaba aprender muchas cosas antes de poder utilizar ADACOP. | SUS10 |

Tabla 10 - Instrumento de validación de tecnología para ADACOP

#### 5.1.2.2 Procedimiento

Para validar la aceptación de ADACOP por parte del usuario final, se siguieron los siguientes pasos:

- Selección de los usuarios sobre los cuales se realizarán las pruebas en representación de todos los posibles usuarios, para esto se seleccionó personal del Observatorio Fiscal de la Universidad Javeriana.

- Investigar diferentes modelos de validación de tecnologías, enfocado a Big Data, datos abiertos y Dashboards, llegando a la selección de TAM y SUS como se describe en la sección anterior.
- Elaboración del instrumento de medición: para esto se tuvieron en cuenta los elementos básicos del modelo TAM y el modelo SUS y se adaptaron específicamente para ADACOP, sin modificar su esencia tal y como se puede apreciar en la Tabla 1.
- Validación del instrumento de medición: el desarrollo del instrumento fue iterativo y estuvo constantemente en validación tanto con el equipo de desarrollo del portal web, como con el cliente CAOBA, para tener la seguridad de que no quedara pendiente la evaluación de ninguno de los aspectos relevantes y para que además fuera claro y sencillo para aquellas personas que fueran a realizar las pruebas.
- Aplicación del instrumento de medición: el instrumento aprobado fue planeado para ser evaluado por Luis Carlos Reyes Hernández, director del Observatorio Fiscal y docente del Departamento de Economía PUJ y Adriana Salinas, líder de gasto público del Observatorio Fiscal y profesora de cátedra del Departamento de Ciencias Políticas PUJ, personas que han demostrado un gran interés en el sistema y potenciales usuarios del portal web. Esto se llevó a cabo durante en el mes de mayo de 2019 en una reunión con estas personas del Observatorio Fiscal de la Universidad Javeriana, luego de una breve demostración de las diferentes opciones y navegación por sistema ADACOP, realizada por algunos de los desarrolladores de este.

#### 5.1.2.3 Resultados y discusión

Los resultados fueron contundentes y demostraron una gran aceptación por parte de los dos evaluadores que tanto para el modelo TAM como para el modelo SUS, tal y como se puede apreciar a continuación en la tabla 2, los resultados:

| Modelo | Factor | ITEM | FRECUENCIA | | | | | MEDIANA | ACEPTACIÓN POR ÍTEM | ACEPTACIÓN POR VARIABLE |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | | | |
| Utilidad (TAM) | Utilidad percibida | PU1 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | 100% |
| | | PU2 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | |
| | | PU3 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | |
| | | PU4 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | |
| | Intención | BI1 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | 100% |
| | | BI2 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | |
| | | BI3 | 0 | 0 | 0 | 0 | 2 | 5 | 100% | |
| | Total, aceptación utilidad | | | | | | | | | 100% |
| **Modelo** | **Factor** | **ITEM** | **1** | **2** | **3** | **4** | **5** | **VALOR REAL** | **ACEPTACIÓN # POR ÍTEM** | **ACEPTACIÓN % POR ÍTEM** |
| Usabilidad (SUS) | Usabilidad | SUS1 | 0 | 0 | 0 | 0 | 2 | 4 | 10 | 100% |
| | | SUS2 | 2 | 0 | 0 | 0 | 0 | 4 | 10 | 100% |
| | | SUS3 | 0 | 0 | 1 | 0 | 1 | 3 | 7.5 | 70% |
| | | SUS4 | 2 | 0 | 0 | 0 | 0 | 4 | 10 | 100% |
| | | SUS5 | 0 | 0 | 0 | 0 | 2 | 4 | 10 | 100% |
| | | SUS6 | 2 | 0 | 0 | 0 | 0 | 4 | 10 | 100% |
| | | SUS7 | 0 | 0 | 0 | 0 | 2 | 4 | 10 | 100% |
| | | SUS8 | 2 | 0 | 0 | 0 | 0 | 4 | 10 | 100% |
| | | SUS9 | 0 | 0 | 0 | 0 | 2 | 4 | 10 | 100% |
| | | SUS10 | 2 | 0 | 0 | 0 | 0 | 4 | 10 | 100% |
| | Total, aceptación usabilidad | | | | | | | 40 | 97.5 | 100% |

Tabla 11 - Resultados pruebas de validación

Para el modelo TAM se consolidó el resultado, calculando inicialmente la mediana de las evaluaciones, y calculando con esta la aceptación por cada uno de los ítems, luego se consolidó para cada factor y finalmente totalizando para obtener la aceptación de utilidad del sistema.

El modelo SUS tiene un modo de evaluación diferente pues para este modelo hay que tener en cuenta que los ítems impares se valoran como el valor seleccionado menos 1, mientras que a los pares se les calcula restándoles a 5 el valor seleccionado. Para cada ítem se obtiene el valor real teniendo en cuenta estas reglas y sacando el promedio de todos los valores, para posteriormente obtener el total sumando cada resultado. Finalmente se multiplica por 2.5 y se obtiene la aceptación total para la usabilidad. Este resultado para el modelo sus no es equivalente a un porcentaje directamente, hay que basarse en la siguiente grafica para poder asociar el resultado obtenido a su respectivo porcentaje de aceptación tal y como se ve en la figura 23. [45]

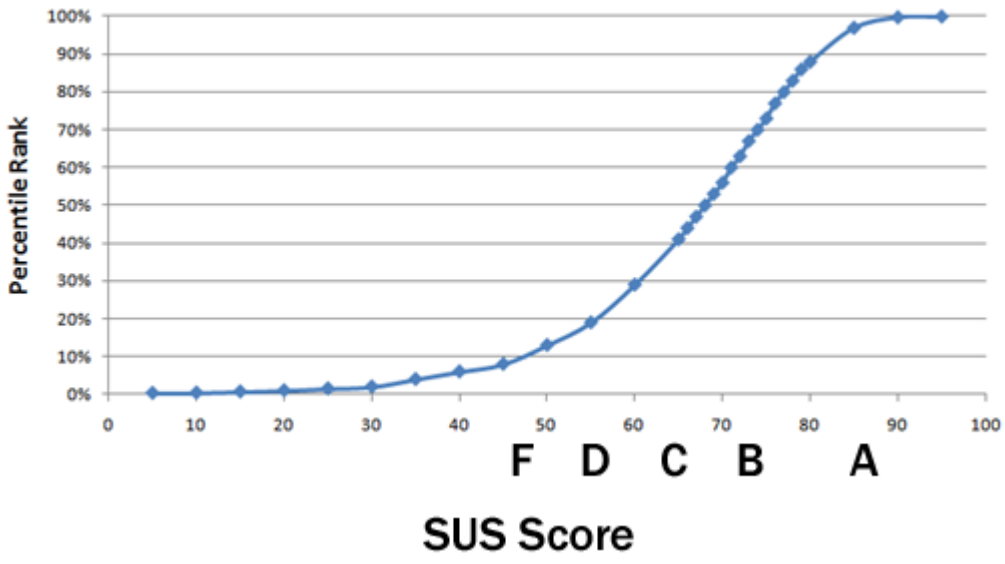


Figura 23 - Relación SUS con porcentaje de aceptación. Fuente: "MeasuringU: Measuring Usability with the System Usability Scale (SUS)"**[52]**

# 6 Conclusiones y trabajo futuro

## 6.1 Conclusiones

Desde el inicio y durante el desarrollo del proyecto, se dedicó un esfuerzo importante orientado a la investigación y entendimiento de modelos de arquitectura de Big data y analítica avanzada, aplicados a datos abiertos estatales; en este ejercicio, la definición y adopción de un proceso de revisión sistemática, orientó la búsqueda de trabajos relacionados, permitiendo establecer criterios de selección para las herramientas *Open Source* integradas, aplicabilidad de algoritmos avanzados en modelos analíticos y replicabilidad científica, por medio de los hallazgos tecnológicos encontrados en cada fase del desarrollo.

Para la contratación pública, las bases de SECOP presentan información estructurada, pero dicha información no es integral, presenta una gran cantidad de datos faltantes y no es de buena calidad, por ello, para generar valor se incluyó dentro del flujo de procesamiento cálculo de estadísticas y criterios de calidad. Dado que la base de SECOP es grande, el proceso de identificación de las cuentas se realizó por medio agrupación de contratistas de acuerdo a sus cuantías, identificando cuatro grupos: *top*, *high*, *medium* y *low*, de cada grupo se seleccionaron el top 25 de contratistas. Se requirió usar el API de Twitter para la extracción de información. Dado que en algunos casos se repetían las cuentas, se incluyó también la ubicación / localización y de esta manera obtener un listado más acotado. Con conclusión, se desarrolló el componente de extracción de Twitter, pero por la calidad de los datos de SECOP, es complejo realizar el enriquecimiento de estas fuentes.

En cuanto a la arquitectura, para satisfacer la recolección, análisis y visualización de la información, se diseñó una arquitectura de Big data por capas modular que permite integrar componentes de análisis para cualquier conjunto de datos, estos nuevos modelos se integran desde la capa de procesamiento de datos (tomando como referencia SECOP, ver figura 5).

En el proceso de diseño, se realizó la especificación de requerimientos de acuerdo a la definición de volumen, variabilidad, velocidad y variedad, esta definición, junto con la tarea de investigación, facilitó el diseño arquitectónico y selección de tecnologías *Open Sorce* (Python principalmente) como fuentes de orquestación de los servicios de la aplicación; en la fase de

desarrollo y la implementación, se crean componentes funcionales para cargar, almacenar y administra los datos extraídos, integrándolos dentro de un único flujo,

Para la implementación del repositorio, se generan servicios en Python, que se consumen desde la interfaz e interactúan con el HDFS, adicionalmente, por la lógica de negocio, se incluye dentro del repositorio el almacenamiento de los metadatos en Mongo.

En cuando a las tareas análisis y exploración de los datos, en el componente Back de ADACOP, se incluye el flujo de medición de criterios de calidad de datos, estas tareas dan como resultado datos limpios y normalizados para posteriormente realizar representaciones descriptivas; en esta integración, se presentan inconvenientes relacionados con los formatos de los datos (UTF), generando un trabajo adicional, para la solución se realizaron funciones de *encode* por dato y así permitir la creación de las *Fact Tables* definitivas.

Finalmente, dada variabilidad de los diferentes tipos de fuentes provenientes de datos abiertos y que se pueden cargar en ADACOP, en el componente web Front de ADACOP donde se visualiza la descriptiva de los datos, se optó por hacer graficas de distribución por variable, esto ya que las labores de entendimiento y selección de datos son netamente analíticas y no las cubre el alcance de este proyecto.

## 6.2 Trabajo futuro

El desarrollo de este proyecto levantó una gran cantidad de dudas en los miembros del observatorio fiscal quienes se interesaron en gran manera en continuar con su desarrollo en una siguiente etapa. Al ser un proyecto liderado por la Alianza CAOBA y el cual pretende ser parte del portafolio de proyectos de este centro de excelencia, a continuación, se presentan una serie de mejoras y nuevas implementaciones que pueden llegar a ser parte de la herramienta:

1. El foco de este trabajo fue el diseño e implementación de una arquitectura modular para el análisis de datos abiertos de contratación estatal. Sin embargo, el diseño gráfico de la solución puede ser trabajado más a profundidad teniendo en cuenta aspectos como la navegabilidad del sistema y *user experiencie*.

2. Para el observatorio fiscal es de gran interés poder enriquecer los datos de SECOP con fuentes alternas provenientes de diferentes entidades gubernamentales. Trabajando de la mano con el observatorio se podrían integrar estas fuentes
3. La arquitectura modular de ADACOP permite integrar de forma fácil nuevos componentes para el análisis de diferentes conjuntos de datos del portal de datos abiertos de Colombia. Sería de gran valor realizar el ejercicio de integrar el análisis de otro conjunto de datos realizando el ejercicio de entendimiento de negocio para aplicar modelos analíticos y hallar información de valor. De igual manera esto fortalecería ADACOP brindando mayor valor para futuros interesados.

## 6.3 Reflexiones

El uso de *Scrum* como metodología ágil de desarrollo de ADACOP permitió diseñar e implementar una arquitectura robusta aun cuando los requerimientos de la solución no eran del todo claros. Fue gratificante para el equipo los resultados obtenidos al validar la solución con el observatorio fiscal ya que el gran factor diferenciador de este trabajo es el hecho de tener la Alianza CAOBA como cliente y el observatorio fiscal como usuario potencial.

Jaime Mendoza

La campaña de datos abiertos a nivel mundial es muy importante y tiene mucho potencial, aun así, aunque trata de generar un gobierno más transparente, se queda corta teniendo en cuenta que no se tienen en cuenta barreras técnicas y tecnológicas que vuelven prácticamente inservible toda información ahí publicada. Para cualquier persona que quiera enfrentarse a la labor de revisar los datos, requerirá de una infraestructura tecnológica grande con la que la mayoría no tiene al alcance y además debe tener algunos conocimientos medios para poder analizar de manera adecuada y eficiente todo el volumen de la información. Estas primeras barreras ya descartan a la gran mayoría de la población y adicionalmente hay que tener en cuenta que se debe tener un conocimiento de negocio para poder tener claridad sobre el significado de uno u otro valor para cada variable del *dataset*. Todo esto hace que la iniciativa de datos abiertos sea muy bonita en el papel, pero poco practica en la realidad.

Daniel Calambás

El ejercicio de investigación es el punto de partida para tener un acercamiento a información relacionada, en este proyecto, el trabajo de investigación fue constante y en el pude evidenciar que hay mucho enfoques y trabajo por hacer relacionado con el manejo de datos abiertos, específicamente en como la calidad de los datos puede satisfacer el objetivo de la transparencia. Por otro lado, técnicamente, afiance mis habilidades análisis y uso de los leguajes aquí implementados, específicamente en Python.

Andrea Ruiz

Este proyecto de profundización ayuda a mostrar las grandes ventajas de la integración entre la academia y la industria, La Pontifica Universidad Javeriana junto al Centro de Apropiación en Big Data y Analítica se integraron para generar un producto con impacto real, y gracias a la realimentación constante de las partes interesadas fue posible generar un producto valioso, que integra el desarrollo y aplicación sobre nuevas tecnologías a situaciones y necesidades reales. Se espera que este proyecto pueda ser un acercamiento hacía las políticas estatales de transparencia y de datos abiertos, y que pueda ser un ejemplo de aprovechamiento de la información libre del gobierno.

Julián Malaver

# 7 Referencias

[1] «Datos Abiertos - Ministerio de Tecnologías de la Información y las Comunicaciones». [En línea]. Disponible en: https://www.mintic.gov.co/portal/604/w3-article-62310.html. [Accedido: 03-may-2019].

[2] «The International Open Data Charter», *International Open Data Charter*. [En línea]. Disponible en: https://opendatacharter.net/. [Accedido: 10-may-2019].

[3] «Guia de Datos Abiertos», n.º 01, p. 36.

[4] «Colombia Compra Eficiente | Colombia Compra Eficiente». [En línea]. Disponible en: https://www.colombiacompra.gov.co/colombia-compra/colombia-compra-eficiente. [Accedido: 03-may-2019].

[5] «Open Government Data - OECD». [En línea]. Disponible en: http://www.oecd.org/gov/digital-government/open-government-data.htm. [Accedido: 17-may-2019].

[6] «Alianza CAOBA Centro de Excelencia big data y Data Analytics Colombia, tic.» [En línea]. Disponible en: http://alianzacaoba.co/. [Accedido: 13-may-2019].

[7] «Datos abiertos», *Colombia Compra Eficiente*, 24-sep-2015. [En línea]. Disponible en: https://www.colombiacompra.gov.co/transparencia/gestion-documental/datos-abiertos. [Accedido: 03-may-2019].

[8] «Information Strategies for Open Government: Challenges and Prospects for Deriving Public Value from Government Transparency», *ResearchGate*. [En línea]. Disponible en: https://www.researchgate.net/publication/221561300_Information_Strategies_for_Open_Government_Challenges_and_Prospects_for_Deriving_Public_Value_from_Government_Transparency. [Accedido: 03-may-2019].

[9] A. Cerrillo-Martínez, «Datos masivos y datos abiertos para una gobernanza inteligente», *El Prof. Inf.*, vol. 27, n.º 5, pp. 1128-1135, sep. 2018.

[10] «European Commission | Search». [En línea]. Disponible en: https://ec.europa.eu/search/?query_source=ISA&QueryText=Report+on+high-value+data-sets+from+EU+institutions+value+of+a+dataset&op=Search&swlang=en&form_build_id=form-RpZHpbi9Uauw7pE5kIDBMAZgY5cihUMgIjPUo0esL8E&form_id=nexteuropa_europa_search_search_form#. [Accedido: 03-may-2019].

[11] «SBOK Guide - SCRUMstudy.com». [En línea]. Disponible en: https://www.scrumstudy.com/sbokguide. [Accedido: 03-may-2019].

[12] «TÉRMINOS Y CONDICIONES DE USO PORTAL DE DATOS ABIERTOS», *Datos Abiertos Colombia*, 19-sep-2016. [En línea]. Disponible en: https://herramientas.datos.gov.co/es/terms-and-conditions-es. [Accedido: 03-may-2019].

[13] P. Nesi, G. Pantaleo, y G. Sanesi, «A hadoop based platform for natural language processing of web pages and documents», *J. Vis. Lang. Comput.*, vol. 31, pp. 130-138, dic. 2015.

[14] S. Nadal *et al.*, «A software reference architecture for semantic-aware Big Data systems», *Inf. Softw. Technol.*, vol. 90, pp. 75-92, oct. 2017.

[15] «CONSULTE EN EL SECOP I», *Colombia Compra Eficiente*, 17-nov-2016. [En línea]. Disponible en: https://www.colombiacompra.gov.co/secop/consulte-en-el-secop-i. [Accedido: 13-may-2019].

[16] «Why Agile Is Eating The World». [En línea]. Disponible en: https://www.forbes.com/sites/stevedenning/2018/01/02/why-agile-is-eating-the-world%E2%80%8B%E2%80%8B/#55c7d85f4a5b. [Accedido: 20-may-2019].

[17] «A Brief History of Open Data and GIS», *GovLoop*, 26-ago-2014. [En línea]. Disponible en: https://www.govloop.com/a-brief-history-of-open-data-and-gis/. [Accedido: 06-feb-2019].

[18] «Datos Abiertos Colombia | Datos Abiertos Colombia». [En línea]. Disponible en: https://www.datos.gov.co/. [Accedido: 07-may-2019].

[19] «What Is Big Data? - Gartner IT Glossary - Big Data». [En línea]. Disponible en: https://www.gartner.com/it-glossary/big-data/. [Accedido: 10-may-2019].

[20] N. Khan, M. Alsaqer, H. Shah, G. Badsha, A. A. Abbasi, y S. Salehian, «The 10 Vs, Issues and Challenges of Big Data», presentado en Proceedings of the 2018 International Conference on Big Data and Education, 2018, pp. 52-56.

[21] R. Schmidt y M. Möhring, «Strategic Alignment of Cloud-Based Architectures for Big Data», en *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops*, 2013, pp. 136-143.

[22] T. White, *Hadoop: the definitive guide*, Fourth edition. Beijing: O'Reilly, 2015.

[23] «HDFS Architecture Guide». [En línea]. Disponible en: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html. [Accedido: 07-may-2019].

[24] «Apache Hive TM». [En línea]. Disponible en: https://hive.apache.org/. [Accedido: 07-may-2019].

[25] R. A. Gonzalez y A. Pomares, «La investigación científica basada en el diseño como eje de proyectos de investigación en ingeniería», en *ResearchGate*.

[26] F. J. García-Peñalvo, «Revisión sistemática de literatura para artículos», ene. 2017.

[27] R. De Donato *et al.*, «Agile production of high quality open data», presentado en Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, 2018, p. 84.

[28] «Informe Global | Open Data Barometer». [En línea]. Disponible en: https://opendatabarometer.org/4thedition/report/?lang=es. [Accedido: 28-ene-2019].

[29] J. M. S. Calderón, Ó. A. N. Trujillo, G. A. R. Flórez, M. C. Santamaría, L. C. V. Echeverri, y A. G. Uribe, «CONSEJO NACIONAL DE POLÍTICA ECONÓMICA Y SOCIAL CONPES», p. 116.

[30] B. Jadhav, A. B. Patankar, y S. B. Jadhav, «A Practical approach for integrating Big data Analytics into E-governance using hadoop», en *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1952-1958.

[31] K. Bakshi, «Considerations for big data: Architecture and approach», en *2012 IEEE Aerospace Conference*, 2012, pp. 1-7.

[32] V. Thalita Vergilio y R. Muthu, «Non-Functional Requirements for Real World Big Data Systems - An Investigation of Big Data Architectures at Facebook, Twitter and Netflix».

[33] N. Surasvadi, C. Saiprasert, y S. Thajchayapong, «Budget and procurement analytics using open government data in Thailand», en *2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)*, 2017, pp. 1-6.

[34] A. Agrahari y S. K. Srivastava, «A data visualization tool to benchmark government tendering process: Insights from two public enterprises», 2019.

[35] A. Pomares-Quimbaya, A. Sierra-Múnera, J. Mendoza-Mendoza, J. Malaver-Moreno, H. Carvajal, y V. Moncayo, «Anonylitics: From a Small Data to a Big Data Anonymization System for Analytical Projects», presentado en 21st International Conference on Enterprise Information Systems, 2019, pp. 61-71.

[36] T. Siddiqui, M. Alkadri, y N. A. Khan, «Review of Programming Languages and Tools for Big Data Analytics», *Int. J. Adv. Res. Comput. Sci.*, vol. 8, n.º 5, p. 1112, may 2017.

[37] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, y M. Niranjan, «Trendminer: An Architecture for Real Time Analysis of Social Media Text», en *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[38] N. Marz y J. Warren, *Big data: principles and best practices of scalable real-time data systems*. Shelter Island, NY: Manning, 2015.

[39] «Welcome to The Apache Software Foundation!» [En línea]. Disponible en: https://www.apache.org/. [Accedido: 07-may-2019].

[40] © 2019 Cloudera, I. A. rights reserved Terms, C. | P. Policy, D. P. A. Hadoop, associated open source project names are trademarks of the A. S. F. F. a complete list of trademarks, y C. Here, «Cloudera | The enterprise data cloud company», *Cloudera*. [En línea]. Disponible en: https://www.cloudera.com/. [Accedido: 07-may-2019].

[41] «Plataforma de gestión de datos y soluciones de análisis de macrodatos | Hortonworks». [En línea]. Disponible en: https://es.hortonworks.com/. [Accedido: 07-may-2019].

[42] «Apache Hadoop 2.9.2 – Apache Hadoop YARN». [En línea]. Disponible en: https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html. [Accedido: 07-may-2019].

[43] «Apache Spark™ - Unified Analytics Engine for Big Data». [En línea]. Disponible en: https://spark.apache.org/. [Accedido: 07-may-2019].

[44] «Apache ZooKeeper». [En línea]. Disponible en: https://zookeeper.apache.org/. [Accedido: 07-may-2019].

[45] «Ambari -». [En línea]. Disponible en: https://ambari.apache.org/. [Accedido: 07-may-2019].

[46] P. Pääkkönen y D. Pakkala, «Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems», *Big Data Res.*, vol. 2, n.º 4, pp. 166-186, 01 2015.

[47] «Socrata Developers | Socrata». [En línea]. Disponible en: https://dev.socrata.com/. [Accedido: 07-may-2019].

[48] «About». [En línea]. Disponible en: https://about.twitter.com/. [Accedido: 07-may-2019].

[49] «Docs». [En línea]. Disponible en: https://developer.twitter.com/en/docs.html. [Accedido: 07-may-2019].

[50] F. D. Davis, Jr., «A technology acceptance model for empirically testing new end-user information systems: Theory and results», *1985*, p. 291.

[51] R. Lewis y J. Sauro, «Revisiting the Factor Structure of the System Usability Scale», vol. 12, n.º 4, p. 10, 2017.

[52] «MeasuringU: Measuring Usability with the System Usability Scale (SUS)». .