IMPLEMENTACIÓN DE UN PROTOTIPO DE INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS PARA LA GESTIÓN DE APLICACIONES VIRTUALIZADAS EN COLPENSIONES

WILSON JAIR PEÑA MARTINEZ

CLARA HELENA ROJAS ORTIZ

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

BOGOTÁ D. C

2020

IMPLEMENTACIÓN DE UN PROTOTIPO DE INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS PARA LA GESTIÓN DE APLICACIONES VIRTUALIZADAS EN COLPENSIONES

WILSON JAIR PEÑA MARTINEZ

CLARA HELENA ROJAS ORTIZ

Trabajo de grado para optar al título de magister en tecnologías de la información y comunicación

Asesor

JORGE MARIO LONDOÑO PELAEZ

PhD.

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

BOGOTÁ D. C

2020

AGRADECIMIENTOS

Gracias a mis padres Alfonso y Gema, por ser los principales promotores de mis sueños, por confiar y creer en mis expectativas, por los consejos, valores y principios que me han inculcado.

- Clara Helena Rojas Ortiz

A Dios, ante todo por darme las capacidades suficientes para afrontar nuevos retos como este. A mis padres por incentivar la formación en mí; a mis hijos por ser mi motivo y mi motor sin importar las circunstancias. Y mi agradecimiento y reconocimiento especial para mi amor por su apoyo, paciencia y compresión ya que sin ella no hubiese sido posible.

- Wilson Jair Peña Martínez

Agradecemos a nuestros docentes de la Universidad Pontifica Bolivariana, por haber compartido sus conocimientos a lo largo de la preparación de nuestra maestría, de manera especial, al Msc. Jorge Mario Londoño tutor de nuestra tesis de grado quien ha guiado con su paciencia, y su rectitud como docente.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

GLOSARIO

BIG DATA: El término Big Data hace referencia al conjunto de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles y que ameritan, por tales razones, ser manipuladas con herramientas de gran capacidad específicamente creadas para trabajar con este volumen (PowerData, 2019).

BODEGA DE DATOS: Para visualizar únicamente los datos de interés es necesario crear una bodega de datos o *data warehouse*. En esta se almacenan datos recopilados e integrados desde múltiples fuentes ya que son usados por las organizaciones para crear análisis o reportes deben estar en un formato coherente y de fácil acceso (sas, 2019). Los *data warehouse* no solo permiten la visualización de los datos, sino que también esta información puede ser utilizada para aplicar modelos y técnicas de minería y ciencia de datos con algunos tratamientos adicionales dependiendo de la herramienta a utilizar y lo que se busque conocer.

CIENCIA DE DATOS: Es el estudio y análisis de la información que genera valor agregado para la organización y que se convierte en un recurso valioso para la definición de objetivos estratégicos, el propósito de este concepto es tratar de predecir comportamientos futuros usando datos de diferentes fuentes, ya sea estructurados o no estructurados.

INTELIGENCIA DE NEGOCIOS: La inteligencia de negocios (B.I. por sus siglas en inglés) concierne el tratamiento de las tecnologías, procesos, plataformas, aplicaciones, estrategias y herramientas facilitan la obtención rápida y sencilla de datos provenientes de los sistemas de gestión empresarial para su análisis e interpretación de manera que puedan ser aprovechados para la toma de decisiones por parte de la dirección del negocio, los datos usados aquí son históricos y estructurados.

METODOLOGIA CRISP-DM: Para el lado de la ciencia de datos, la metodología más utilizada es la CRISP-DM, en la que el conocimiento del negocio y los datos junto con la preparación de los datos toman la mayor relevancia, ya que para realizar análisis de este tipo es importante contar con un muy buen conocimiento del negocio y con datos limpios, útiles y que no tengan redundancias.

METODOLOGIA KIMBALL: Se basa en lo que el autor denomina como Ciclo de Vida Dimensional del Negocio, este ciclo se basa en 4 principios básicos, el primero es la identificación de los requerimientos del negocio, el segundo es diseñar una base de información única, integrada y fácil de usar en la que se reflejen todos los datos necesarios, el tercero es realizar entregas en incrementos significativos en plazos de 6 a 12 meses y el ultimo es ofrecer una solución completa en la que se entregan todos los elementos necesarios para generar valor a los usuarios (Rivadera, 2010).

MINERIA DE DATOS: Al igual que la ciencia de datos, la minería de datos o *data mining*, comprende las técnicas que permiten explorar grandes volúmenes de datos (Big Data) de manera automática con el objetivo de encontrar modelos o patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

PROCESOS DE ETL: (*Extract*, *Transform*, *Load*) implican la extracción, transformación y carga de los datos que finalmente generan los insumos de la bodega. Cada uno de estos procesos se da en su propia fase y se requiere de la ejecución en su propio orden. La primera fase es la Extracción (E), toma los datos existentes en los sistemas transaccionales conocidos también como OLTP (*On-Line Transaction Processing*). La fase de Transformación (T) es un proceso en el cual se aplican las reglas del negocio o funciones necesarias sobre los datos extraídos previamente para convertirlos en información útil que será cargada en la bodega. El último proceso es el de Carga (L) que finalmente interactúa con la base de datos destino depositando la data de acuerdo con las necesidades del negocio.

TABLEROS DE CONTROL O DASHBOARD: Son una composición de indicadores y gráficos que responden principalmente a la pregunta ¿Qué pasó? Basados en datos pasados es posible realizar acciones para mejorar distintos procesos (Curto, 2010). Son usados para explicar el pasado, estos datos y conclusiones son la base para predecir futuros valores (Valchanov, 2018).

VISUALIZACIÓN DE DATOS: Da respuesta a la necesidad de monitorear estos datos de manera rápida y eficaz. El concepto de "Visualización" hace referencia a la estética, el buen diseño y la claridad. El concepto general es poder convertir los datos en una herramienta de interpretación de los hechos con la ayuda del trabajo de mentes creativas y expertos con gran poder de análisis (Olivares, 2015).

RESUMEN

Se implementó un prototipo de inteligencia de negocios y ciencia de datos que monitorea el comportamiento de las aplicaciones virtualizadas en COLPENSIONES, incluyendo modelos de predicción y técnicas de análisis de minería de datos que permitan tener una información más valiosa y completa que se visualiza por medio de un tablero de control entendible a personas de diferentes áreas de la empresa.

PALABRAS CLAVE

Virtualización de Aplicaciones, Minería de Datos, Bodega de Datos, Ciencia de Datos, Inteligencia de Negocios.

ABSTRACT

Implementation of a business intelligence and data science prototype that aims to monitor the behavior of virtualized applications in COLPENSIONES, including prediction models and data mining analysis techniques that allow for more valuable and complete information, to visualize through of a dashboard that is understandable to people from different areas of the company.

KEY WORDS

App Virtualization, Data Mining, Data Warehouse, Data Science, Business Intelligence.

1. INTRODUCCIÓN

Colpensiones como Entidad Financiera de carácter especial centra su atención en el servicio que presta a los ciudadanos en materia de protección a la vejez, para esto hace uso de diferentes herramientas tecnológicas, por ejemplo, un entorno de escritorios y aplicaciones virtualizadas que se centralizan en servidores. Hace dos años la administración de este entorno se puso en manos de un tercero llamado Agilitix. Debido a que Colpensiones no poseía el control sobre sus aplicaciones y lo que conocía únicamente eran informes mensuales que no brindaban gran detalle a cerca del comportamiento de los escritorios y las aplicaciones, dichos informes no facilitaban la toma decisiones acertadas que permitieran la correcta administración de los recursos de infraestructura dispuestos para la prestación de los servicios.

Por esta razón la Entidad decidió tomar el control de sus aplicaciones, por lo que se hizo necesario contar con una herramienta de gestión de virtualización con administración propia que permita conocer su comportamiento en tiempo real y con información que permita tomar decisiones anticipadas y no sobre la marcha. El propósito de este proyecto es aplicar los conocimientos adquiridos durante la maestría para implementar un prototipo de inteligencia de negocios y ciencia de datos que permita monitorear y analizar la información. Adicionalmente se integrarán conceptos de la ciencia de datos para brindar predicciones útiles y así, formular acciones preventivas.

Los datos que se analizarán son generados por la virtualización de escritorios de usuario y aplicaciones en un entorno en el que no se había hecho previamente por la organización. Se realizó un modelo de *data warehouse* que contiene la información necesaria para visualizar el comportamiento de las cargas transaccionales de los servidores de la granja de virtualización, de los usuarios y de las aplicaciones, además de esto se realizaron modelos de clusterización para segmentar los usuarios según el uso que le dan a las aplicaciones y series de tiempo que muestran un posible comportamiento a futuro.

En este documento se describen las etapas de implementación de un prototipo de sistema de gestión basado en herramientas de Inteligencia de Negocios y Ciencia de Datos a partir de las metodologías existentes. En el planteamiento del problema abordado en el capítulo 1, se describe la situación actual que da oportunidad para la mejora.

La justificación descrita en el capítulo 2 explica el por qué se ejecuta el proyecto y los beneficios de su implementación. También se describen a detalle el objetivo general y los específicos en el capítulo 3. El capítulo 4 describe el marco referencial el cual incluye el marco conceptual, contextual y el estado del arte. En el capítulo 5 se encuentra la metodología utilizada para el desarrollo del proyecto. El capítulo 6 muestra el análisis de los resultados y los objetivos alcanzados.

2. PLANTEAMIENTO DEL PROBLEMA

2.1 Problema

Para la Dirección de Infraestructura Tecnológica de Colpensiones es incierto decidir sobre los recursos que utilizan las aplicaciones virtualizadas ya que no se tiene un sistema de gestión que permita tener ese control. Esto genera que la toma de decisiones adecuadas y oportunas concernientes a la administración y gestión de la plataforma no esté basada en análisis de datos ni conocimiento de la realidad sino en la percepción de la situación del momento. Este escenario no permite realizar análisis, consultas ni seguimientos al comportamiento de los usuarios y los hábitos de consumo de las aplicaciones generando incertidumbre en la ejecución de los procesos actuales de la organización que dependen de esas aplicaciones.

Es importante para la dirección conocer los tiempos de uso de las aplicaciones, los orígenes de las conexiones, los usuarios que se conectan y sus hábitos de trabajo entre otras, además de poder predecir estos comportamientos para tomar acciones tempranas. La Organización carece de gobierno sobre el entorno de aplicaciones virtualizadas de los usuarios ya que se confió esta labor a un tercero 2 años atrás mediante la modalidad de contratación. Actualmente la plataforma de aplicaciones y servicios virtualizados se encuentra a cargo de la empresa colombiana Agilitix. Amparado en su experticia, Colpensiones ha puesto en las manos de esta empresa la administración del entorno virtualizado debido a la limitada gestión del conocimiento interno en este aspecto. A pesar de la existencia de herramientas especializadas para analítica, inteligencia de negocios, minería de datos y demás, el procesamiento de los datos entregados por el contratista concernientes a los consumos de recursos, rendimiento de la plataforma, estado de los usuarios, sus hábitos y sus consumos, etc., se realiza de forma manual, con una alta probabilidad de error debido entre otras, al alto volumen, ocasionando subutilización de recursos tecnológicos y humanos (Funeme, 2019).

Otra situación presente es el desconocimiento de los componentes del entorno lo cual hace que no se identifiquen posibles puntos de falla en la operación ante incidentes reportados por los mismos usuarios que experimentan bajo desempeño en las aplicaciones Core del negocio. Esto finalmente se ve reflejado en el incumplimiento de los logros estratégicos, generación de sobrecostos en la operación y reducción de utilidad económica que se reporta periódicamente a los entes de control del país. En la Figura 1 se encuentra representado el árbol de problemas de manera más gráfica:

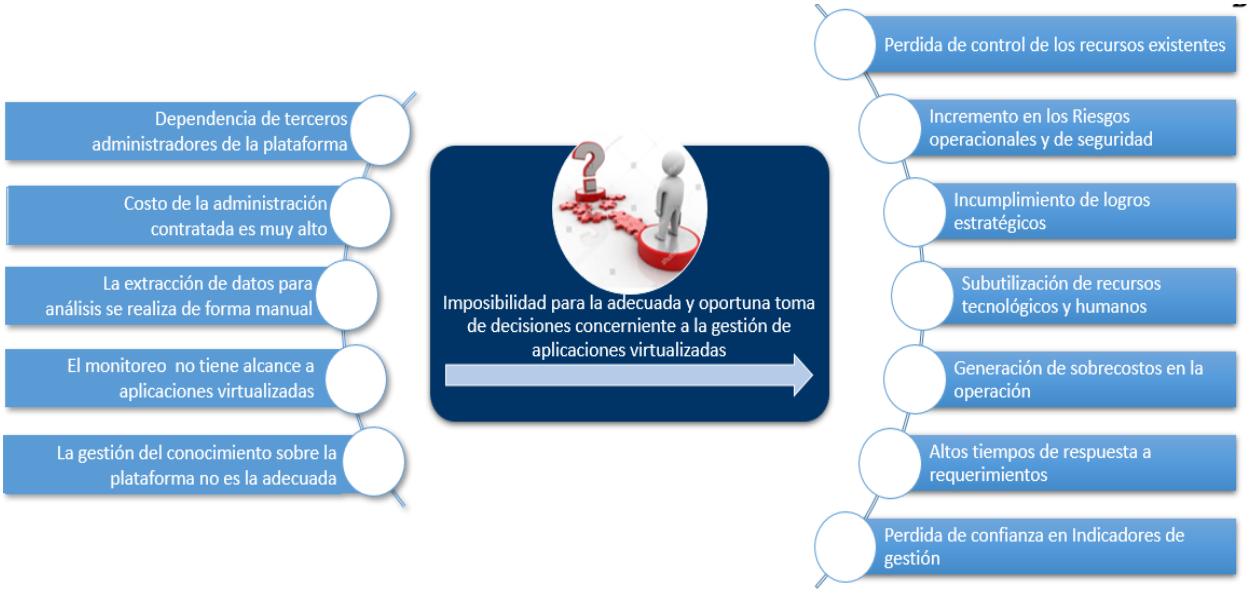

Figura 1 Árbol de problemas – Fuente: Autores

Por decisión de las directivas, a partir de enero de 2020 la Organización empezó la gestión para tener el gobierno total de la plataforma virtualizada de las aplicaciones y por ende de la información que se obtenga a partir de su uso. Esto implica que el gran volumen de datos que se genere no será útil para la Entidad si no es debidamente analizado.

2.2 Justificación

Actualmente la organización tiene inconvenientes con la administración y control del entorno virtualizado de aplicaciones ya que esta labor se encuentra tercerizada y requiere contar con un sistema de gestión con herramientas de inteligencia de negocios y ciencia de datos que le permita tomar decisiones de manera oportuna, basadas en los datos generados a partir de la implementación de una plataforma tecnológica administrada desde la DIT (Funeme, 2019).

Con este prototipo se pretende conocer los datos y establecer modelos de análisis de datos descriptivos y predictivos que ayuden a la Entidad a tomar decisiones sobre los recursos tecnológicos con mayor seguridad y rapidez de la que se tiene actualmente con el contratista, además de que se podrán tomar acciones preventivas y correctivas de manera anticipada. Con el diseño de estas estrategias de Inteligencia de Negocios y Ciencia de Datos se le puede otorgar a Colpensiones una mayor ventaja competitiva ya que el conocimiento y entendimiento de los datos permite mantener el control eficiente sobre los recursos tecnológicos existentes, los procesos y los usuarios disminuyendo así los riesgos operacionales y de seguridad. Este prototipo permitirá también utilizar de mejor manera los recursos tecnológicos y humanos existentes conllevando a la optimización de los costos de operación y al incremento de las utilidades para la organización (Logicalis, 2015), ya que teniendo la administración y control propio de la plataforma virtualizada se puede obviar la contratación con el tercero que asciende a los 8.000 millones de pesos anuales. Aunque la plataforma soporta casi 3.000 usuarios es importante tener en cuenta que el costo que se paga actualmente a Agilitix es elevado considerando que este valor corresponde casi al 2% del presupuesto de operación anual de la Entidad (Colpensiones, 2019).

Una administración propia y no tercerizada puede aumentar la confianza en los indicadores de gestión mejorando la toma de decisiones y disminuyendo los tiempos de respuesta de análisis de datos lo cual se refleja en el cumplimiento de los objetivos estratégicos de la organización. En la Figura 2 se enumeran los beneficios que se podrían obtener con la implementación de proyecto.

Aunque en la organización existen herramientas de gestión y monitoreo de la plataforma tecnológica, no existe una que específicamente se enfoque en el monitoreo del uso y desempeño de las aplicaciones virtualizadas del negocio. El monitoreo actual se encarga de presentar en tiempo real el rendimiento de los servidores, la disponibilidad de los servicios, el estado de los canales de comunicación, la respuesta de servicios web, el desempeño de bases de datos, etc.

Sin embargo y más allá de una simple gestión de monitoreo, no existe una herramienta que permita ver el desempeño en tiempo real, o que permita tomar decisiones a partir del comportamiento de las aplicaciones, ni tampoco que permita predecir su comportamiento futuro. El costo de adquisición de una herramienta comercial para tal gestión puede llegar a los US$60.000 anuales y no está contemplado dentro del presupuesto de la Entidad.

Sumado a esto, las herramientas comerciales son rígidas, de fábrica y responden a la atención de requerimientos generales de las empresas de todos los sectores y sin posibilidad de personalizarlas a las necesidades propias de un negocio lo cual hace que la Entidad deba adecuarse a ellas y no las herramientas a la Entidad generando más inconvenientes que soluciones.

El prototipo de visualización que se plantea en este proyecto permite analizar y predecir, desde diferentes dimensiones, las situaciones que se presentan con las aplicaciones virtualizadas, su estado y uso de acuerdo con las necesidades propias de la organización y orientada tanto a perfiles Directivos como Técnicos. En la figura 2 se muestran las mejoras que se pueden llevar a cabo con esta proyecto y a lo que esto conllevaría en caso de ser implementado por la empresa.

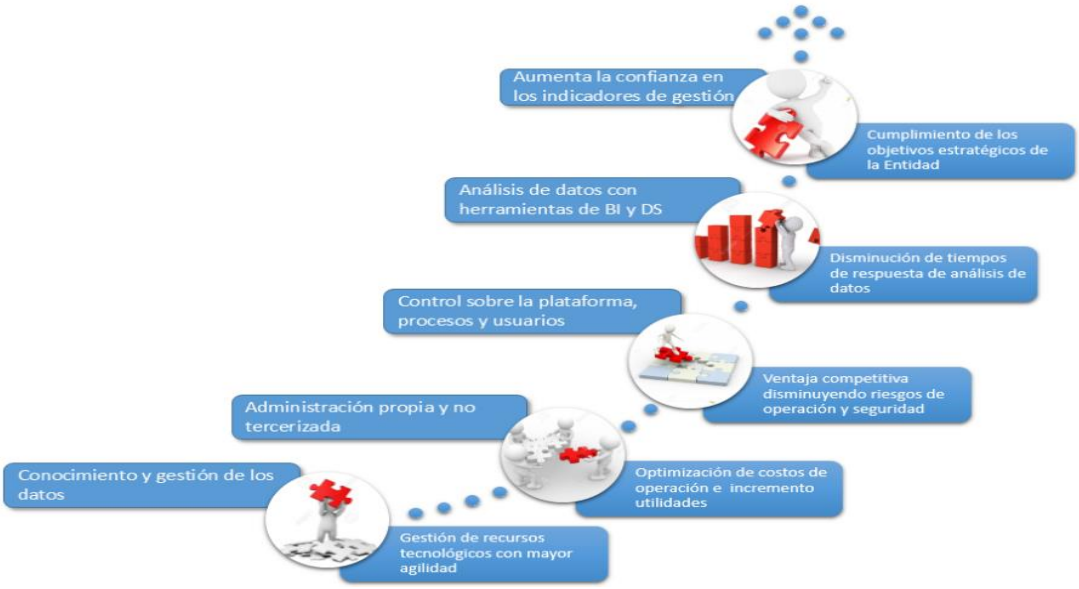

Figura 2 Justificación – Fuente: Autores

La justificación del proyecto está sustentada en el marco estratégico 2019-2022 de la Entidad la cual pretende “Promover la transformación digital en la gestión institucional para hacer más eficientes los procesos, trámites y servicios” (Colpensiones 2020). Desde la perspectiva de TICs E Infraestructura ,se considera igualmente “Disponer de una óptima arquitectura de tecnología y aplicaciones que esté alineada con la estrategia del negocio” y “Proporcionar y mantener oportunamente la infraestructura requerida” (Colpensiones, 2016). Como se observa en la Figura 3.

Figura 3 Mapa Estratégico Colpensiones 2019 – 2022 Fuente: https://www.colpensiones.gov.co/Publicaciones/nuestra_entidad_colpensiones/marco_estrategico

En el mapa de procesos de la Organización que se muestra en la Figura 4, la Gestión de Servicios de Tecnología e Información se encuentra como un área transversal que apoya los procesos y busca ofrecer la disponibilidad de todos los servicios necesarios para el funcionamiento de la Entidad a través de una eficiente gestión de los recursos y componentes tecnológicos.

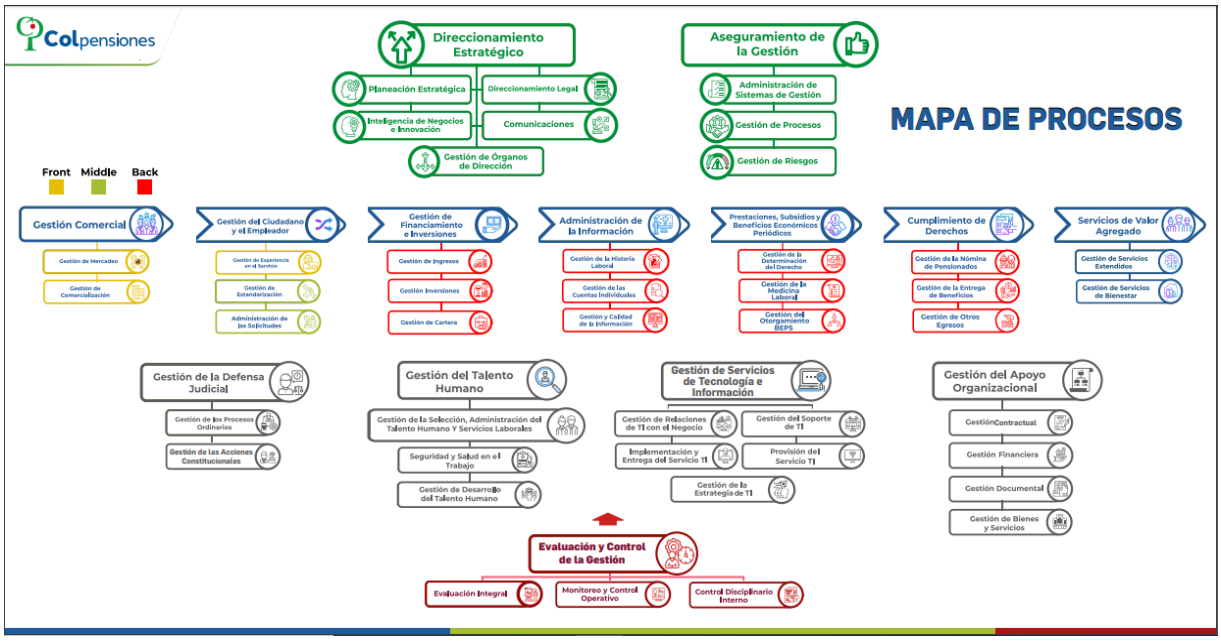

Figura 4 Mapa de procesos – Fuente: https://www.colpensiones.gov.co/Publicaciones/nuestra_entidad_colpensiones/marco_estrategico

3. OBJETIVOS

3.1 Objetivo General

Implementar un prototipo de sistema de gestión y control de aplicaciones virtualizadas basada en herramientas de inteligencia de negocios y ciencia de datos para la identificación de factores que afecten el desempeño de la plataforma tecnológica de la Administradora Colombiana de Pensiones Colpensiones.

3.2 Objetivos Específicos

- Realizar el levantamiento de información y requerimientos del negocio necesarios para la elaboración del prototipo de sistema de gestión del entorno virtualizado.
- Implementar el modelo del *Data Warehouse* y los procesos de extracción, transformación y carga (ETL's) de la data requerida que permita luego procesarla y analizarla desde las diferentes dimensiones y con óptimos tiempos de respuesta.
- Realizar un análisis descriptivo y predictivo que refleje el estado de los escritorios y las aplicaciones virtualizadas a partir de la data recolectada para facilitar la toma de decisiones concernientes a la gestión de la infraestructura tecnológica que soporta los servicios.
- Implementar un modelo de *Dashboard* que incorpore los análisis realizados y la información más relevante del entorno virtualizado por medio de representaciones gráficas, permitiendo la optimización de los recursos existentes a partir de las decisiones tomadas por los responsables.

4. MARCO REFERENCIAL

4.1 Marco contextual

A partir de la creación de la Administradora Colombiana de Pensiones Colpensiones en 2007 y su entrada en funcionamiento en 2012, se consideró que el activo más valioso para la organización iba a ser los datos apoyados en una eficiente gestión tecnológica. Como reposa en el Auto 110 de 2013, la liquidación del antiguo Instituto de Seguros Sociales (ISS) se basó en la existencia de una serie de dificultades administrativas y operativas que no le permitían atender

oportunamente las peticiones de sus afiliados y vinculados. Algunas de las dificultades presentadas más notables eran la planta de personal insuficiente, la carencia de un sistema integrado de información tecnológica, el represamiento de expedientes sin fallar en los centros de decisión, la desactualización de las historias laborales y el incremento de solicitudes de corrección de estas.

Sumado a esto se contaba con una muy pobre y casi nula calidad de los datos registrados en sus precarios sistemas de información y por supuesto, el gobierno de datos no existía. Las dificultades del ISS se gestaron a lo largo de varios años, pero finalmente no fue posible superarlas, originando una situación de desbalance entre la demanda de servicios y la capacidad institucional de la entidad para atenderlos.

La organización ha sido consciente desde su inicio de que debe implementar y mantener la ejecución del Plan Estratégico de Tecnologías de la Información enfocado en una estrategia digital con mecanismos de control que permitan no solo salvaguardar la información del negocio en bases de datos y sistemas asegurados, sino administrarlos bajo buenas prácticas de gobierno de datos. La Vicepresidencia de Planeación y Tecnologías de la Información tiene a su cargo los procesos de planeación, incluido el seguimiento, el sostenimiento del sistema integrado de planeación y gestión, y la operación tecnológica de la entidad. Así mismo, es el área encargada de definir los lineamientos, políticas y procedimientos que enmarcan el tratamiento de los sistemas de información y de la infraestructura tecnológica de la organización. La adopción de nuevas tecnologías ha permitido a la organización mantenerse a la vanguardia brindando servicios ágiles y seguros soportando la operación del negocio en procesos, personas y tecnologías. Aunque es claro que la razón de ser de Colpensiones son las pensiones, la contratación de terceros especializados en tecnologías es de gran importancia debido a que son estos los que tienen un conocimiento más profundo y la experticia para su implementación y administración. Sin embargo, para Colpensiones, la obtención rápida y sencilla de datos provenientes de los sistemas controlados por terceros para su análisis e interpretación, de manera que puedan ser aprovechados para la toma de decisiones, algunas veces se torna compleja. Esta situación crea dependencia de terceros que soportan las plataformas tecnológicas generando altos costos por la administración delegada y con riesgo de pérdida o fuga de información lo que se traduce finalmente en un nivel de gobernanza bajo lo que conlleva a que no se disponga de herramientas apropiadas para implementar los mecanismos de gobierno y gestión de la información.

Al no tener control sobre la información, ya sea en tiempo real, en tableros de control o en informes ejecutivos, la ejecución de acciones preventivas se convierte en acciones correctivas haciendo compleja la identificación de incidentes y por supuesto, la toma de decisiones acertadas por parte de la gerencia, que complementando algo mencionado previamente, conlleva a que no se disponen de las herramientas apropiadas para implementar los mecanismos de gobierno y gestión de la información. Esto ocasiona claramente la pérdida de control sobre los recursos existentes y el incremento de riesgos operacionales y de seguridad.

4.2 Marco Conceptual

El desarrollo de este proyecto se basó en datos generados por la virtualización de aplicaciones, este es el concepto que permite a sus usuarios acceder de manera remota desde sus escritorios de trabajo a las diferentes aplicaciones, en estos se almacenan los datos y ejecutan los procesos en un solo servidor central. Esto facilita las labores de muchas personas, ya que pueden acceder a sus escritorios y aplicaciones de manera remota desde cualquier dispositivo conectado a internet.

Algunos de los principales beneficios de su implementación son una mayor seguridad tanto en los escritorios como en los datos almacenados y los costos de soporte se reducen al igual que los costos generales en hardware (SIAG, 2018).

El concepto base que engloba todo el contenido y procesos que se desarrollan en este proyecto es el de Inteligencia de Negocios y Ciencia de Datos. La inteligencia de negocios (B.I. por sus siglas en inglés) concierne el tratamiento de las tecnologías, procesos, plataformas, aplicaciones, estrategias y herramientas facilitan la obtención rápida y sencilla de datos provenientes de los sistemas de gestión empresarial para su análisis e interpretación de manera que puedan ser aprovechados para la toma de decisiones por parte de la dirección del negocio, los datos usados aquí son históricos y estructurados. Muy comúnmente, la inteligencia de negocios se complementa con la ciencia de datos que es el estudio y análisis de la información que genera valor agregado para la organización y que se convierte en un recurso valioso para la definición de objetivos estratégicos, el propósito de este concepto es tratar de predecir comportamientos futuros usando datos de diferentes fuentes, ya sea estructurados o no estructurados.

Aunque son conceptos diferentes, si existe una estrecha relación entre la ciencia de datos y la inteligencia de negocios como se explica en la Figura 5:

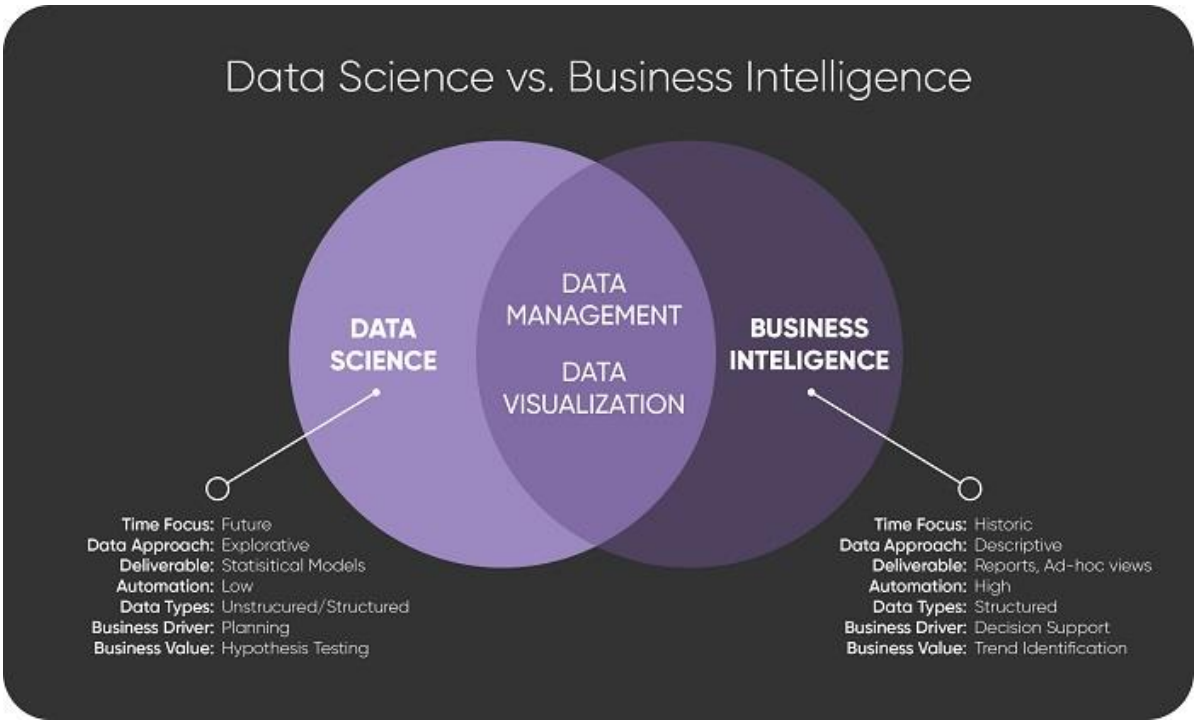

Figura 5 DS vs BI - Fuente: https://tdwi.org/articles/2017/12/05/bi-all-understanding-differences-data-science-and-bi.aspx

Se hizo uso de los datos provenientes de la virtualización combinando los conceptos y herramientas usadas tanto para inteligencia de negocios, como para la ciencia de datos. Dentro de la inteligencia de negocios se usan herramientas de tipo OLAP que permiten tener modelos multidimensionales que contienen información de diferentes fuentes, dentro de estas herramientas se encuentran por ejemplo Tableau, Click, Cognos y MicroStrategy, entre muchas otras. Por parte de la ciencia de datos, se utilizan algoritmos que mezclan conceptos de estadística, aprendizaje de máquina y computación, dentro de las herramientas usadas en este campo, se encuentran lenguajes de programación como Python y R o frameworks como RapidMiner. Al integrar ambos conceptos es posible responder preguntas como: ¿Qué pasó? ¿Qué va a pasar? o ¿Qué se puede hacer para cambiar el futuro? El responder a todas estas preguntas dentro de la visualización de los datos da un valor agregado que puede ser de gran provecho para la organización. Es posible encontrar la manera de visualizar por medio de tablas, gráficas o indicadores lo que se encontró durante el análisis predictivo o prescriptivo (Kotu, 2017).

Al igual que la ciencia de datos, la minería de datos o *data mining*, comprende las técnicas que permiten explorar grandes volúmenes de datos (Big Data) de manera automática con el objetivo de encontrar modelos o patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto. El término Big Data hace referencia al conjunto de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles y que ameritan, por tales razones, ser manipuladas con herramientas de gran capacidad específicamente creadas para trabajar con este volumen (PowerData, 2019).

La implementación de un sistema de este tipo genera grandes cantidades de datos, un uso adecuado de estos puede brindar información valiosa que permita tomar decisiones importantes. Durante los últimos años la cantidad de información almacenada ha crecido de manera exponencial, a causa de esto la capacidad para analizarla y hacer de esta algo útil también se ha venido desarrollando (Few & Edge, 2012). La visualización de datos da respuesta a la necesidad de monitorear estos datos de manera rápida y eficaz. El concepto de "Visualización" hace referencia a la estética, el buen diseño y la claridad. El concepto general es poder convertir los datos en una herramienta de interpretación de los hechos con la ayuda del trabajo de mentes creativas y expertos con gran poder de análisis (Olivares, 2015).

Una de las maneras que se han creado para visualizar estos datos son los tableros de control o *dashboards*, con el correcto uso de estos las empresas son capaces de monitorear sus procesos en tiempo real y de manera eficaz. Se debe tener un conocimiento de los procesos y generar indicadores usando los datos de principal importancia. Los *dashboard* son una composición de indicadores y gráficos que responden principalmente a la pregunta ¿Qué pasó? Basados en datos pasados es posible realizar acciones para mejorar distintos procesos (Curto, 2010). Como se mencionó, los *dashboard* son usados para explicar el pasado, estos datos y conclusiones son la base para predecir futuros valores. Existen diferentes técnicas para poder lograrlo y cada una de ellas puede brindar una mayor o menor aproximación según lo que se esté analizando. Se buscará saber el futuro comportamiento de los servidores para prever posibles fallas y evitarlas antes de que ocurran (Valchanov, 2018).

Los reportes, ya sean impresos o digitales, son otra de las herramientas válidas para la visualización de la información. Deben ser claros, precisos y concisos y ser capaces de decir en forma directa, lo que se quiere contar. Un informe extenso se convierte en carga operacional.

En la Figura 6 se puede observar un tablero de control que se actualiza en tiempo real y aporta información muy relevante para el monitoreo, en este caso específico, del movimiento en redes sociales de cierta empresa:

Figura 6 Dashboard o tablero de control - Fuente: https://www.klipfolio.com/blog/6-dashboards-i-use-daily

Para conocer estas predicciones es necesario utilizar conceptos de la minería y la ciencia de datos, donde la primera busca encontrar patrones o reglas dentro de los datos que no se ven a simple vista, de esta forma se amplía la visión y el conocimiento sobre el negocio, el segundo concepto, es decir, la ciencia de datos busca predecir comportamientos futuros a partir de los datos pasados usando diferentes técnicas. Existen herramientas que permiten la unión de ambos mundos, principalmente por medio de diferentes lenguajes de programación como R o Python (Kotu, 2017). Para visualizar únicamente los datos de interés es necesario crear una bodega de datos o *data warehouse*. En esta se almacenan datos recopilados e integrados desde múltiples fuentes ya que son usados por las organizaciones para crear análisis o reportes deben estar en un formato coherente y de fácil acceso (sas, 2019). Existen múltiples herramientas para este propósito, se explorarán las que mejor se acomoden a la necesidad, de acuerdo con el tamaño y formato de los datos.

Cabe mencionar que los *data warehouse* no solo permiten la conservación de los datos, sino que también esta información puede ser utilizada para aplicar modelos y técnicas de minería y ciencia de datos con algunos tratamientos adicionales dependiendo de la herramienta a utilizar y lo que se busque conocer. En la Figura 7 se muestra cómo un *data warehouse* toma datos de diferentes fuentes de la organización y agrupa los más relevantes haciendo análisis sobre ellos, visualizándolos con herramientas especializadas como Power BI, Tableau o Qlik.

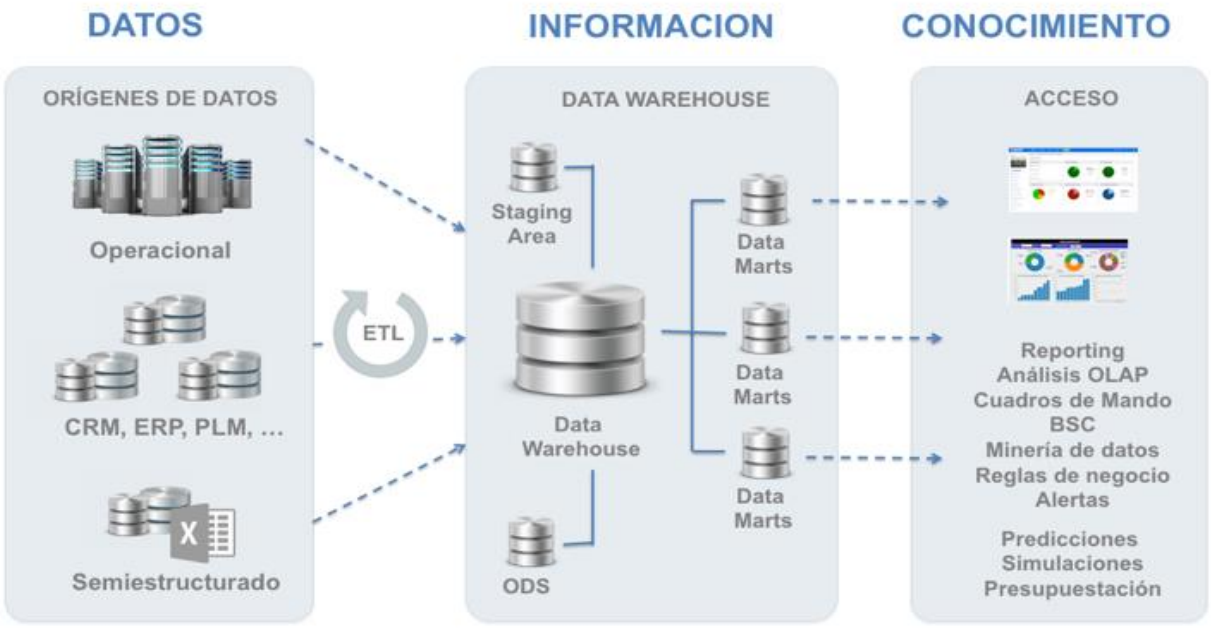

Figura 7 Arquitectura de un data warehouse. Fuente: http://www.diegocalvo.es/data-warehouse/

Para obtener una bodega de datos valiosa y adecuada es importante tener una buena calidad de datos, una característica esencial que determina la fiabilidad de los datos en la toma de decisiones (IBM, n.d.). Los datos son un activo valioso en toda empresa por lo tanto deben ser manejados dentro de un gobierno de datos por parte de la compañía.

El gobierno de datos es la disciplina encargada de la orquestación de las personas, los procesos y la tecnología que permita que la información sea un recurso de valor empresarial (Jimmy Martínez, 2012). Ampliando lo anterior, dentro de todo *data warehouse* es vital contar con datos limpios y completos, se debe realizar una preparación previa de los datos, a fin de remover ruido e inconsistencias, revisando que los datos estén completos, normalizados, coherentes y libres de posibles variables redundantes (Durán & Costaguta, 2007).

Con un excelente gobierno de datos se pueden establecer políticas y controles desde los mismos sistemas transaccionales que finalmente reflejan un alto porcentaje en la calidad de datos que lleguen a la *data warehouse*.

Parte esencial de la implementación de un *data warehouse* son los procesos de ETL (*Extract, Transform, Load*) que implican la extracción, transformación y carga de los datos que finalmente generan los insumos de la bodega. Cada uno de estos procesos se da en su propia fase y se requiere de la ejecución en su propio orden. La primera fase es la Extracción (E), que toma los datos existentes en los sistemas transaccionales conocidos también como OLTP (On-Line *Transaction Processing*).

La calidad de los datos debería controlarse desde la misma captura en los sistemas de información transaccionales y deben ser convertidos a los formatos adecuados para iniciar el proceso de transformación. La fase de Transformación (T) es un proceso en el cual se aplican las reglas del negocio o funciones necesarias sobre los datos extraídos previamente para convertirlos en información útil que será cargada en la bodega. El último proceso es el de Carga (L) que finalmente interactúa con la base de datos destino depositando la data de acuerdo con las necesidades del negocio previamente planeadas durante el análisis de los requerimientos disponiéndolo como un sistema de procesamiento analítico u OLAP (On-Line Analytical Processing). Los procesos de ETL son útiles para las aplicaciones de bases de datos, migraciones, sincronización de data o interfaces con otros sistemas.

Para llevar a cabo el proyecto, se deberán tener en cuenta metodologías tanto para la inteligencia de negocios, como para la ciencia de datos. Para el lado de la inteligencia de negocios se usa principalmente la metodología de Kimball, que se basa en lo que el autor denomina como Ciclo de Vida Dimensional del Negocio, este ciclo se basa en 4 principios básicos, el primero es la identificación de los requerimientos del negocio, el segundo es diseñar una base de información única, integrada y fácil de usar en la que se reflejen todos los datos necesarios, el tercero es realizar entregas en incrementos significativos en plazos de 6 a 12 meses y el ultimo es ofrecer una solución completa en la que se entregan todos los elementos necesarios para generar valor a los usuarios (Rivadera, 2010).

En la Figura 8 se ven representadas estos principios y etapas:

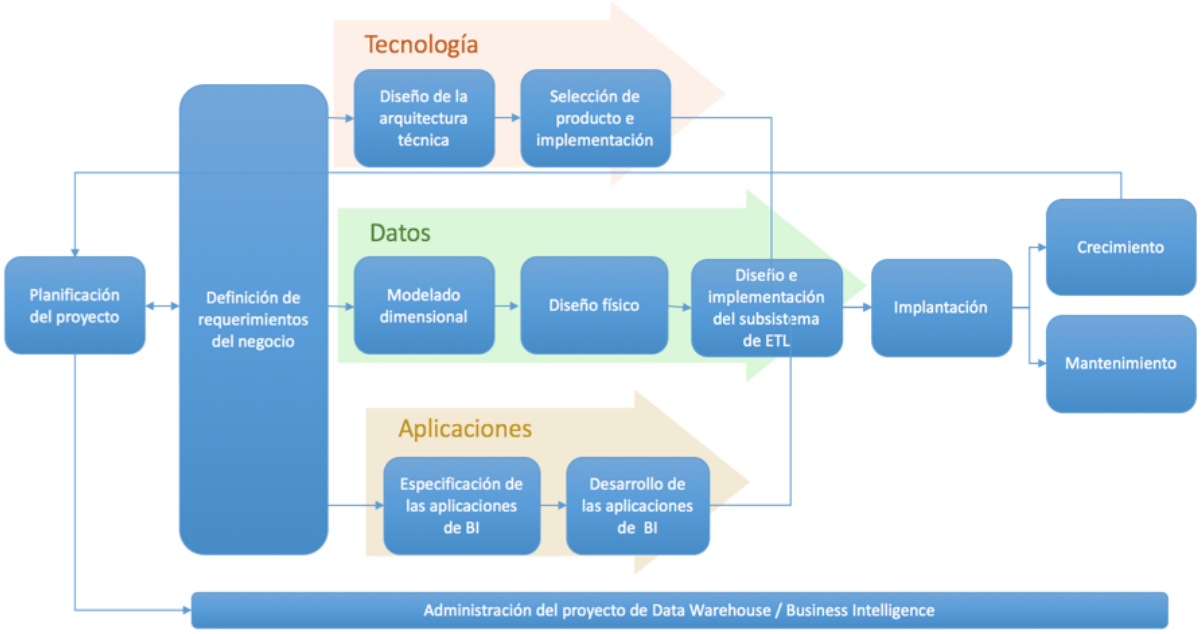

Figura 8 Ciclo de vida de proyecto de BI - Fuente: http://ww.diegocalvo.es/inteligencia-de-negocio

Para el lado de la ciencia de datos, la metodología más utilizada es la CRISP-DM, en la que el conocimiento del negocio y los datos junto con su preparación toman la mayor relevancia, ya que para realizar análisis de este tipo es importante contar con un muy buen conocimiento del negocio y con datos limpios, útiles y que no tengan redundancias.

Desglosando cada una de las fases, se tiene que la primera es la fase de compresión del negocio o problema, en la que se determinan los objetivos del negocio y del análisis a realizar, en la segunda fase se debe realizar la comprensión de los datos, es decir, entender qué significa cada una de las variables a utilizar, realizar análisis exploratorios y verificar la calidad de los datos obtenidos, durante la tercera fase se deben preparar los datos, en la que por medio de una exploración profunda de los datos se seleccionan los datos más relevantes para el análisis buscado. La cuarta fase se trata del modelado, en esta fase se seleccionan las técnicas más adecuadas para la solución del problema y se aplican a los datos, para identificar si las técnicas utilizadas son efectivas se debe realizar la quinta fase que evalúa los modelos aplicados basada en diferentes métricas. Por último se encuentra la fase de implementación, en la que se despliegan los modelos creados y se entregan los resultados al negocio para generar valor (EPB 603, n.d.).

En el capítulo 5 del presente trabajo se explica en forma detalla la unión de las metodologías y la integración propuesta para el desarrollo del proyecto. En la Figura 9 se representa cada etapa del ciclo de vida de un proyecto de minería:

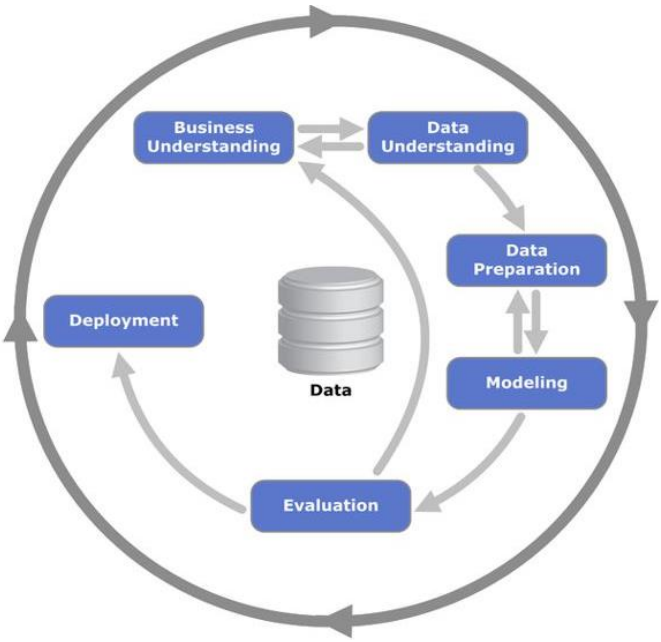

Figura 9 Diagrama de proceso CRISP - Fuente: //riis.com/blog/machine-learning-data-pre-processing/

Los modelos escogidos para el desarrollo del proyecto son el *clustering*, usando el algoritmo de K-means, redes neuronales MLP y la serie de tiempo. El *clustering* es una técnica de aprendizaje de máquina no supervisada que involucra el agrupamiento de puntos de datos. En teoría, los puntos de datos que se encuentran en el mismo grupo poseen propiedades o características similares (Priy, n.d.). Existen diferentes algoritmos para este modelo, el que probablemente es el más utilizado es el K-means, para aplicar este algoritmo, en primer lugar se debe seleccionar la cantidad de grupos a crear, en el proceso, se buscan diferentes puntos al azar (que se convierten en el centro del clúster) y se mide la distancia con respecto a los demás puntos, se realizan varias iteraciones hasta encontrar el centro que una distancia promedio menor a los demás puntos (Seif, 2018).

Las redes neuronales MLP (Multi-layer Perceptron) tratan de emular la manera de aprender del cerebro humano, están formadas por un conjunto de nodos conocidos como neuronas artificiales que se encuentran conectadas y transmiten señales entre sí, estas señales se transmiten desde la entrada hasta generar una salida. El objetivo de este modelo es aprender modificándose a sí mismo de forma que puede llegar a realizar tareas complejas.

Las redes reciben una serie de valores de entrada y cada una de estas entradas llega a uno o varios nodos llamados neurona, las neuronas se encuentran agrupadas por capas para formar la red y cada neurona posee un peso, es decir, un valor numérico que modifica la entrada recibida, lo nuevos valores continúan su camino por la red y una vez que se alcanza el final de la red se obtiene una salida que será la predicción calculada por la red (Innovation, 2019).

Las series de tiempo por su parte son secuencias en las que una métrica es medida en intervalos regulares de tiempo, por ejemplo, diariamente, basados en esta información es posible intentar predecir lo que sucederá en periodos siguientes con la variable medida, las series se pueden dividir en dos tipos: univariadas y multivariadas. Uno de los modelos más utilizados en esta técnica, es el ARIMA (Auto Regressive Integrated Moving Average), este algoritmo explica una serie temporal dada en función de sus valores pasados, es decir, sus propios retrasos y errores de pronóstico, ya que es un modelo de regresión lineal. Si la serie que se intenta predecir tiene patrones estacionarios, se deben agregar términos estacionarios y convertir el modelo en SARIMA (Seasonal SARIMA) (Prabhakaran, 2019).

Existe un conjunto de coeficientes (p, d, q) que se utilizan para parametrizar este tipo de modelos. Por esta razón se denotan como ARIMA (p, d, q). Juntos, estos tres parámetros explican la estacionalidad, tendencia y ruido en los conjuntos de datos:

- p es la parte autorregresiva del modelo. Permite incorporar el efecto de valores pasados al modelo. Intuitivamente, esto sería similar a afirmar que es probable que haga calor mañana si ha estado caliente los últimos 3 días.
- d es la parte integrada del modelo. Esto incluye términos en el modelo que incorporan la cantidad de diferenciación (es decir, el número de puntos de tiempo pasados para restar del valor actual) para aplicar a la serie de tiempo. Intuitivamente, esto sería similar a afirmar que es probable que sea la misma temperatura mañana si la diferencia de temperatura en los últimos tres días ha sido muy pequeña.
- q es la parte del promedio móvil del modelo. Esto permite establecer el error del modelo como una combinación lineal de los valores de error observados en puntos de tiempo anteriores en el pasado (Vincent, 2017).

4.3 Marco Legal

Parte fundamental del marco legal de este proyecto es la Constitución Política de Colombia (Const., 1991). “Con el fin de fortalecer la unidad de la Nación y asegurar a sus integrantes la vida, la convivencia, el trabajo, la justicia, la igualdad, el conocimiento, la libertad y la paz, dentro de un marco jurídico, democrático y participativo que garantice un orden político, económico y social justo, y comprometido a impulsar la integración de la comunidad latinoamericana”.

Bajo este preámbulo se establece que la Constitución es el mecanismo para garantizar los derechos de los ciudadanos y la protección que el Estado debe brindar. Por medio del Decreto número 2011 del 28 de septiembre de 2012, se determina y reglamenta la entrada en operación de la Administradora Colombiana de Pensiones - Colpensiones y se dictan otras disposiciones (Trabajo, 2012) y se complementa con los Decretos 2012 del 28 de septiembre de 2012, por el cual se suprimen unas dependencias de la estructura del Instituto de Seguros Sociales -ISS- (Trabajo, 2012), y con el Decreto 2013 del 28 de septiembre de 2012, por el cual se suprime el Instituto de Seguros Sociales -ISS-, se ordena su liquidación y se dictan otras disposiciones (Trabajo, 2012).

Durante los últimos 8 años de servicio de la Entidad se han generado muchos actos administrativos y quizá uno de los más importantes es el Auto 110 de 2013, Ref. Expediente acumulado T-3287521 por Acción de tutela instaurada por Raúl1 y otros, en forma separada contra el Instituto de Seguros Sociales y Colpensiones. Este auto publicado en Internet (Auto 110/13, 2013) comprueba la existencia de una situación constitucionalmente relevante durante el proceso de transición del ISS a Colpensiones.

A partir de la expedición de la Ley 100 de 1993 que trata sobre el Sistema General de Seguridad Social, en el tema de pensiones se establece que el sistema está compuesto por dos regímenes distintos que tiene como objetivo cubrir los riesgos de invalidez, vejez y muerte para sus afiliados, nuevos o provenientes del antiguo Seguro Social (Corte, 1993).

Al interior de la Organización también se han expedido normas reglamentarias. El Decreto 309 por el cual se adopta la Reestructuración de la Entidad, con un alcance de tener una entidad fortalecida, centrada en el ciudadano y volcada al cumplimiento efectivo, integral y oportuno de los derechos de los ciudadanos (Colpensiones, 2017).

Otra norma es la Circular Externa 007 del 5 de junio de 2018 emitida por la Superintendencia Financiera de Colombia, numeral 3.12 que establece criterios para Gestionar la seguridad de la información y la ciberseguridad en los proyectos que impliquen la adopción de nuevas tecnologías, complementando así los requerimientos mínimos para la gestión de este riesgo en las entidades vigiladas (Superfinanciera, 2018).

Recientemente ha tomado mucho auge la LEY 1581 de 2012 por la cual se dictan disposiciones generales para la protección de datos personales (Presidencia, 2012) y la Organización ha sido consciente de la importancia de la custodia y buen manejo de la información de todos sus afiliados.

4.4 Estado del Arte

Con el paso de los años la demanda de almacenamiento y computación ha impulsado el crecimiento de los centros de datos, creando grandes granjas de servidores. Un centro de datos puede comprender muchos miles de servidores y puede utilizar tanta energía como una ciudad pequeña. Se investigaron artículos y patentes relacionados con la inteligencia de negocios y la ciencia de datos, en la primera parte se presentan trabajos relacionados directamente con la inteligencia de negocios para mostrar en la segunda parte cómo la utilización de la ciencia de datos puede complementar de manera sustancial la información de las aplicaciones virtualizadas.

En el año 2009 se publicó un artículo relacionado con la utilización de los recursos en los servidores utilización para la virtualización de escritorios, para optimizar el rendimiento algunas herramientas implementan el uso compartido de páginas, este identifica las páginas de memoria de máquinas virtuales con un contenido idéntico y las consolida en una única página compartida. Estas páginas pueden estar ubicadas en diferentes segmentos por lo que los autores propusieron un sistema llamado “Memory Buddies”, este sistema monitorea los recursos en tiempo real de las máquinas virtuales y los distribuye para así mejorar el rendimiento y aumentar la capacidad de los centros de datos en un 17% (Wood et al., 2009). Posteriormente se publicó una patente en el año 2012 en la que se implementó un sistema de visualización en tiempo real para monitorear los procesos y recursos utilizados por los escritorios (Barber, Friedlander, Hagan, & Kaminsky, 2012), por medio del análisis de estos datos es posible tomar decisiones y así evitar posibles problemas.

En la Escuela Politécnica del Ejército en Ecuador se propuso una solución a la subutilización del hardware, software y almacenamiento además de la sobrecarga de trabajo del personal informático en los laboratorios de computación en el año 2011. Esta solución consistió en proponer la aplicación de técnicas de consolidación de servidores y virtualización de aplicaciones, permitiendo la instalación de software especializado de manera centralizada. Para esto se realizaron distintas pruebas en tiempo real, evaluando el rendimiento de la red, además del consumo de CPU y memoria (Díaz, Vilac, & Gallo, n.d.). Cuando se implementa la virtualización de escritorios, suele haber subutilización o aprovisionamiento excesivo de recursos en las máquinas virtuales, es difícil predecir las necesidades de cada usuario, una solución a estos inconvenientes es usar los conceptos de la ciencia de datos desde diferentes puntos de vista.

Para dar solución a la problemática anteriormente nombrada, durante el año 2008 fue publicada una investigación usando como referencia las necesidades de cada usuario para estudiar problemas relacionados con la implementación de aplicaciones dentro de un centro de datos virtualizado, en este artículo se demuestra como los modelos de virtualización pueden ser utilizados para predecir las necesidades de las aplicaciones y como el hecho de compartir memoria de manera dinámica entre las máquinas virtuales puede mejorar el rendimiento (Wood, 2008). En el 2011 se publicó una patente de un modelo de rendimiento de sesión de escritorio remoto que intenta predecir la capacidad que va a necesitar el usuario y se la asigna de manera dinámica (Talwar, Basu, & Kumar, 2011). Esto se implementó para antes de iniciar las sesiones de escritorio remoto. Posteriormente en 2012 se publicó un artículo que, utilizando el procesamiento en múltiples nodos, propuso modelizar, estudiar su escalabilidad, analizar y predecir el performance de aplicaciones paralelas. Adicionalmente se intentó clasificar todos los escritorios de máximos recursos para mejorar y garantizar un máximo rendimiento eficaz (Quisbert, 2012). En el año 2014 se presentó una patente dirigida a la gestión, predicción y visualización de la capacidad, asignación y utilización de recursos.

La solución presentada por (Quisbert, 2012) es un sistema para calcular, detectar, predecir y mostrar la asignación de recursos para eliminar los cuellos de botella en las redes de los sistemas virtualizados. La aplicación posee un método para predecir la futura utilización de los recursos para cada objeto de la red en un periodo determinado de tiempo basada en la información histórica.

En el mismo año (Tan, Nguyen, Shen, Gu, & Venkatramani, 2012) presentaron un sistema llamado PREPARE (Predictive Performance Anomaly Prevention) que provee prevención automática de anomalías para infraestructuras virtualizadas en la nube. Este sistema integra predicción de anomalías, inferencia de la causa basada en aprendizaje y actúa de manera preventiva para minimizar los posibles problemas de rendimiento sin intervención humana.

En el trabajo de (Arroba, Zapater, & Ayala, 2014) presentado el mismo año no se tiene en cuenta sólo el rendimiento y aplicaciones utilizadas, sino también, la energía utilizada por estos. En este se realizó un estudio de la energía consumida por los diferentes centros de datos, por medio de datos históricos se pudo llegar a la conclusión de que todos los servidores no se usan en todo momento, por lo tanto, es posible apagar algunos mientras no se esté trabajando con ellos y así ahorrar energía, que trae beneficios tanto ecológicos como monetarios. Además de la energía, otro factor externo importante es la temperatura óptima a la que trabajan los centros de datos. En esta propuesta se crea un perfil térmico de los servidores con todos los datos guardados por sensores para tener un conocimiento de la temperatura que permite un mejor rendimiento, por medio de un modelamiento de predicción es posible controlar el ambiente en el que se encuentran los servidores (Chaudhry, Chong, Ling, Rasheed, & Kim, 2016).

También en el año 2014 el trabajo de (Wang, Qiu, & Guo, 2014) se publicó un trabajo sobre máquinas virtuales para el servicio de salud remota en la nube, en el que se propone un protocolo de coherencia de datos que logró medir la demanda de ancho de banda, posteriormente se diseñó un algoritmo predictivo HMM (Hidden Markov Model) que ayudó a manejar los recursos de la red de manera efectiva, este enfoque tiene similitud con el utilizado en Colpensiones, ya que lo empleados utilizan máquinas virtuales que se encuentran centralizadas en servidores, en el caso de este trabajo estos están ubicados en la nube, sin embargo para la entidad son propios.

Por otra parte en el año 2019 (Aldossary, Djemame, & Alzamil, 2019) presentaron un sistema que predice el costo y energía utilizada de un sistema de máquinas virtualizadas en la nube que ayudan a crear sistemas con un uso de energía eficiente y a un menor costo, para la creación del modelo de predicción utilizaron el algoritmo ARIMA de series de tiempo y obtuvieron resultados precisos.

Durante la investigación se encontró que el tener un monitoreo en tiempo real de las aplicaciones virtualizadas es bastante útil, ya que se puede obtener información de muchas fuentes como, por ejemplo, las aplicaciones más utilizadas y el rendimiento que necesita cada usuario, es posible tener claros los recursos de cada servidor, la temperatura, horarios y energía utilizada.

Para ver la real importancia de esta información es necesario contar con una adecuada visualización y análisis, además de poder crear modelos de predicción que permitan afrontar futuros problemas antes de que sucedan y llegar así a una mejora continua en la implementación. Adicionalmente, se encontró que el procesamiento en paralelo mejora en gran medida el manejo de grandes cantidades de datos, la implementación del sistema de virtualización va a generar datos constantemente, por lo que tener en cuenta esto será de vital importancia. En cuanto a las posibles técnicas para realizar los análisis predictivos se deben tener en cuenta que los datos en su gran mayoría serán numéricos, por lo que técnicas como las redes neuronales o las máquinas de soporte vectorial serán relevantes opciones, para visualizar este tipo de datos que cambian y se actualizan con el tiempo es necesario contar con gráficas entendibles y resumidas en un único tablero.

5. METODOLOGÍA

Este es un proyecto de investigación aplicada que pretende resolver un problema específico dentro de la organización de Colpensiones. En primera instancia se hizo un estudio exploratorio para encontrar el problema a resolver y de esta manera poder describirlo, encontrando sus causas, consecuencias y las variables que inciden en él. Luego se realizó una investigación bibliográfica para conocer el estado del trabajo de personas que han realizado trabajos similares a este, para finalmente iniciar con la ejecución del proyecto. Para unir los conceptos de las dos ramas que participaron en la realización de este proyecto (inteligencia de negocios y ciencia de datos) se utilizaron las metodologías de Kimball para el diseño e implementación de la *data warehouse* y CRISP-DM para la creación de los modelos predictivos, que cuentan con algunos pasos en común y otras tareas en paralelo.

El primero de ellos es el entendimiento del negocio, sus objetivos y del problema, así como de los datos a analizar, en esta etapa se analizaron a profundidad todas las variables que influyen dentro del proceso y se seleccionaron cuáles de ellas sirven para realizar los análisis tanto descriptivos como predictivos, este es uno de los pasos más importantes que además se comparte

en ambas perspectivas, así como es común la preparación de los datos, ya que una vez entendidos, es necesario estructurarlos dentro de un *data warehouse* que se diseñó e implementó en el proceso revisando que tuvieran una buena calidad, adicionalmente se clasificaron como datos categóricos o numéricos. En esta etapa se seleccionaron las variables a utilizar dentro de los diferentes modelos y se hicieron las transformaciones necesarias previas a los análisis, tanto descriptivo como predictivo, para crear los conjuntos de datos necesarios. Desde esta parte se empiezan a dividir las tareas, ya que en la ciencia de datos es necesario hacer algunas transformaciones en los tipos de datos que se obtengan según las técnicas a utilizar.

Mientras que en la parte de inteligencia de negocios se diseñaron los entregables de las diferentes visualizaciones de los datos analizados que se extraen desde el *data warehouse* previamente implementado, en la ciencia de datos se aplicaron, evaluaron y desplegaron los diferentes modelos, se utilizaron series de tiempo para visualizar a futuro el comportamiento de las aplicaciones según sus costos y su uso, adicionalmente se utilizaron técnicas de minería de datos para encontrar patrones que no son visibles fácilmente, finalmente ambas ramas se volvieron a encontrar buscando la manera adecuada de hacer visibles los resultados de los modelos aplicados.

Es importante tener en cuenta que el desarrollo de esta metodología permitió realizar análisis descriptivos con base en los hechos ocurridos que son los registrados en los sistemas transaccionales y permitieron responder a las preguntas que buscan conocer los hábitos y comportamientos de los usuarios, aplicaciones y servidores. A partir de estos hechos, se procedió a realizar los análisis predictivos buscando el perfilamiento de lo que podrá suceder en el tiempo futuro para los mismos factores. En la figura 10 se muestran de manera gráfica los pasos a seguir según la metodología implementada para desarrollar el proyecto.

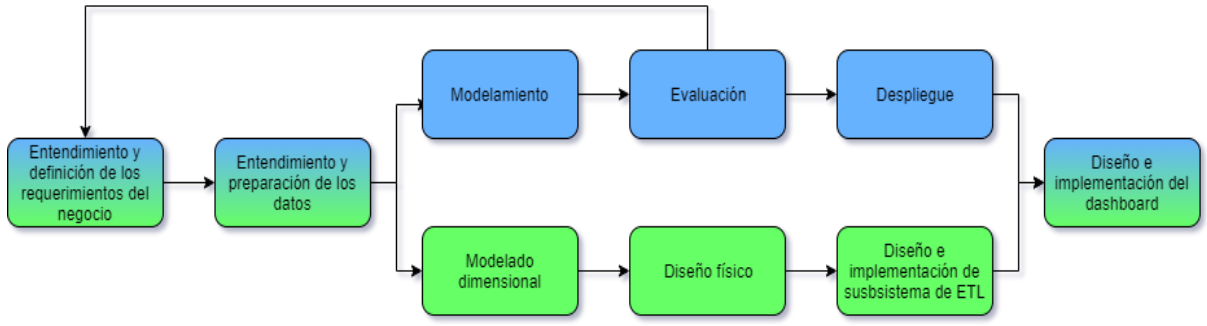

Figura 10 Metodología – Fuente: Autores

En las tablas 1 a 4 se muestra el cronograma establecido para el cumplimiento de los objetivos específicos de acuerdo con la metodología implementada en el desarrollo el proyecto. Las tablas muestran las actividades correspondientes por cada mes de trabajo las cuales conllevan a la entrega de productos que finalmente, dan cumplimiento a los objetivos específicos planteados.

OBJETIVO ESPECIFICO No. 1							
Realizar el levantamiento de información y requerimientos del negocio necesarios para la elaboración del prototipo de sistema de gestión del entorno virtualizado.							
Productos							
Documentación sobre el levantamiento de información.							
Actividades							
No	**Descripción**	**Cronograma**					
		M1	**M2**	**M3**	**M4**	**M5**	**M6**
1	Comprensión del negocio y sus datos: • Identificar los objetivos de la empresa y en área encargada de los datos. • Realizar entrevistas a las personas involucradas. • Realizar el levantamiento de los requerimientos.	X					
2	Creación del diccionario de datos: • Identificar las variables de interés en el análisis. • Conocer el significado y sentido de cada una de las variables.		X				
3	Consolidación de la información recopilada en las actividades anteriores • Crear documento de consolidación del análisis y requerimientos del negocio		X				
4	Definición de procesos de obtención e ingestión de datos: • Establecer las variables para los análisis y eliminar las redundantes o las que no aporten información relevante. • Realizar la estadística descriptiva de cada variable, clasificándolas como categóricas o numéricas. • Generar indicadores a partir de las variables establecidas.		X	X			

Tabla 1 Cronograma objetivo específico No.1 – Fuente: Autores

OBJETIVO ESPECIFICO No. 2							
Implementar el modelo del *Data Warehouse* y los procesos de extracción, transformación y carga (ETL's) de la data requerida que permita luego procesarla y analizarla desde las diferentes dimensiones y con óptimos tiempos de respuesta.							
Productos							
Data Warehouse implementado (incluyendo data necesaria para el análisis predictivo)							
Actividades							
No	**Descripción**	**Cronograma**					
		M1	**M2**	**M3**	**M4**	**M5**	**M6**
1	Análisis de datos necesarios para la implementación • Seleccionar el modelo de data Warehouse a utilizar.			X			
2	Diseño de la estructura del Data Warehouse • Diseñar el diagrama de data Warehouse.			X			
3	Creación de las consultas necesarias para la implementación del modelo: • Implementación de los modelos de extracción, transformación y carga de los datos.				X		
	• Extracción de las variables especificas a utilizar en los modelos predictivos				X		

Tabla 2 Cronograma objetivo específico No.2 – Fuente: Autores

OBJETIVO ESPECIFICO No. 3							
Realizar un análisis descriptivo y predictivo que refleje el estado de los escritorios y las aplicaciones virtualizadas a partir de la data recolectada para facilitar la toma de decisiones concernientes a la gestión de la infraestructura tecnológica que soporta los servicios de la Entidad							
Productos							
Resultado de los análisis realizados con un análisis correspondiente							
Actividades							
No	**Descripción**	**Cronograma**					
		M1	**M2**	**M3**	**M4**	**M5**	**M6**
1	Análisis descriptivo de las variables • Aplicación de técnicas de minería de datos para encontrar patrones en los datos. • Análisis y reporte de los hallazgos				X		
2	Análisis y selección de los modelos predictivos a usar: • Prueba y evaluación de diferentes modelos según la o las variables que se busque predecir.				X	X	
3	Implementación e integración de los modelos elegidos: • Despliegue de los modelos realizados. • Diseño de visualización para los resultados obtenidos.					X X	

Tabla 3 Cronograma objetivo específico No.3 – Fuente: Autores

OBJETIVO ESPECIFICO No. 4							
Implementar un modelo de *Dashboard* que incorpore los análisis realizados y la información más relevante del entorno virtualizado por medio de representaciones gráficas, permitiendo la optimización de los recursos existentes a partir de las decisiones tomadas por los responsables.							
Productos							
Prototipo de Dashboard que incluya el análisis descriptivo y predictivo.							
Actividades							
No	**Descripción**	**Cronograma**					
		M1	**M2**	**M3**	**M4**	**M5**	**M6**
1	Diseño del Dashboard • Diseñar la estructura a mostrar.					X	
2	Creación de gráficas y KPIs a mostrar: • Selección y creación de gráficas. • Selección y creación de KPIs.					X	
3	Realización de pruebas del dashboard • Visualización con datos controlados, • Interpretación y evaluación por parte de los usuarios.						X X

Tabla 4 Cronograma objetivo específico No.4 – Fuente: Autores

El presupuesto para la ejecución del proyecto se basó en 4 pilares principales qué fueron la mano de obra, el hardware, el software y otros componentes qué sumados nos dan un costo total del proyecto de $11.515.793 pesos ejecutados durante 6 meses. La siguiente tabla muestra la relación distribuida por ítem del presupuesto:

DATA WAREHOUSE	ENE	FEB	MAR	ABR	MAY	JUN	CANT.	COSTO UNIT. ($)	COSTO TOTAL ($)
MANO DE OBRA – Horas									
Levantamiento y obtención de datos	40	-	-	-	-	-	40	30.000	1.200.000

Creación de diccionario de datos	-	20	-	-	-	-	20	30.000	600.000
Análisis de calidad	-	20	-	-	-	-	20	30.000	600.000
Preparación de datos	-	-	20	-	-	-	20	30.000	600.000
Creación de modelo multidimencional - StarNet	-	-	20	-	-	-	20	30.000	600.000
Creación de modelos predictivos	-	-	-	40	-	-	40	30.000	1.200.000
Creación de la BD - DWH	-	-	-	40	40	-	80	30.000	2.400.000
Creación de procesos ETL	-	-	-	-	40	40	80	30.000	2.400.000
Diseño e implementación de tablero de control y reportes	-	-	-	-	-	40	40	30.000	1.200.000
Subtotal	**40**	**40**	**40**	**80**	**80**	**80**	**360**		**8.400.000**
HARDWARE – Horas									
Depreciación Acumulada Equipo de Computo (Horas)	40	40	40	80	80	80	360	50	18.000
Subtotal	**40**	**40**	**40**	**80**	**80**	**80**	**360**		**18.000**
SOFTWARE – Horas									
Sistema Operativo MS Windows 10	40	40	40	80	80	80	2	669.999	1.339.998
Ofimática (Procesador de Texto (MS Word) y Hoja de Cálculo (MS Excel))	40	20	40	-	-	-	2	249.999	499.998
Gestor de Bases de Datos (SQL	-	-	-	40	80	40	1	-	-

2017 Developer Edition SSMS)									
Visualizador de Datos (Power BI)[1]	-	-	-	-	-	40	1	-	-
Software para Análisis de Datos (Weka) [1]	-	-	-	20	-	-	1	-	-
Software para Desarrollo (Python) [1]	-	-	-	20	-	-	1	-	-
Software para Calidad de Datos (DQ-Analyzer)[1]	-	20	-	-	-	-	1	-	-
Programas de comunicaciones (Messenger, Whats-App WEB)	10	10	10	10	10	10	1	-	-
Subtotal	**90**	**90**	**90**	**170**	**170**	**170**	**11**	**0**	**1.839.996**
OTROS COSTOS – Unidades									
Energía Eléctrica (Kw/H)	25,20	25,20	25,20	50,40	50,40	50,40	226,80	440,80	99.973,80
Telefonía (Min)	30,00	30,00	30,00	30,00	30,00	30,00	180,00	100,00	18.000,00
Banda Ancha / Internet (Kb)	3.925,93	3.925,93	3.925,93	7.851,85	7.851,85	7.851,85	35.333,33	1,00	35.333,33
Transportes (Und)	4,00	4,00	4,00	4,00	4,00	4,00	24,00	2.400,00	57.600,00
Subtotal									**210.907**
							Subtotal		10.468.903
							10% Imprevistos		1.046.890
							TOTAL ($)		**11.515.793**

Tabla 5 Presupuesto del proyecto - Fuente: Autores

[1] Software Libre

6. DESARROLLO, PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

6.1 Entendimiento y definición de los requerimientos

6.1.1 Comprensión del Negocio

Para abordar el proyecto se hizo una investigación inicial basada en la misión, visión y objetivos estratégicos de la compañía, que permitiera ubicarse en un contexto de entendimiento de las necesidades, esta investigación se podrá encontrar en el Anexo 1 del documento. A partir de la información recolectada se logró identificar que los objetivos estratégicos planteados por la organización para los años 2019 a 2022 y que se relacionan con el desarrollo de este proyecto se basan en la implementación de mejores prácticas de gestión con enfoque en procesos y riesgos para responder a los cambios en el entorno y las expectativas de las partes interesadas, así como promover la transformación digital en la gestión institucional para hacer más eficientes los procesos, trámites y servicios.

Igualmente se logró identificar que la Dirección de Infraestructura Tecnológica busca mejorar la administración controlada de los recursos disponibles para la realización de las operaciones de la Entidad y específicamente para el control de las aplicaciones virtualizadas y los escritorios, conociendo el comportamiento y uso de estas y el comportamiento de los usuarios que las utilizan de una manera fácil y accesible permitiendo tomar acciones que faciliten el trabajo de las personas que hacen parte de la organización. (Colpensiones, 2019)

El modelo inicialmente implementado se basó en la utilización de los recursos tecnológicos existentes y buscó la reducción de costos operacionales y la optimización de la seguridad, confiabilidad y disponibilidad de la información contribuyendo al logro de los objetivos estratégicos de la Entidad.

Es importante aclarar que actualmente la gestión de la plataforma virtualizada, a cargo de un tercero no tiene en cuenta la oportunidad de realizar análisis de los datos con modelos de ciencia de datos e inteligencia de negocios que sirvan para la acertada toma de decisiones. El objeto contractual establecido con el tercero establece la administración funcional de la plataforma virtual de escritorios y aplicaciones de la Entidad para el periodo 2019-2020 con soporte y capacidad de operación para 3.000 usuarios concurrentes.

6.1.2 Entrevistas

Con base en los objetivos del proyecto para la primera etapa del entendimiento se realizan entrevistas con las personas involucradas en la administración y monitoreo del entorno virtualizado, que permiten tener una visión más amplia sobre los procesos realizados y las necesidades actuales de la organización teniendo en cuenta los diferentes niveles jerárquicos y sus propias necesidades.

El día 15 de enero de 2020 en las oficinas de la Dirección de Infraestructura Tecnológica de Colpensiones se realiza una reunión de contextualización con el analista del Grupo Capa Media en donde se expone el alcance del proyecto.

El funcionario procede a realizar una presentación de la implementación de la herramienta TERMINAL SERVER (TS) que se está llevando a cabo por parte de la Entidad para la publicación de escritorios y aplicaciones virtualizadas de los usuarios.

Como resultado de esta reunión se logra entender el alcance del proyecto que la Dirección de Infraestructura Tecnológica está ejecutando debido a la inminente finalización de obligaciones contractuales del proveedor Agilitix quien actualmente es el socio estratégico de COLPENSIONES para brindar la experiencia de todo su equipo, con el objetivo de asegurar la calidad del análisis, arquitectura, implementación y gestión del proyecto escritorios virtuales y virtualización de aplicaciones gracias al conocimiento de la plataforma y al acompañamiento realizado en la operación del negocio. Adicionalmente, Agilitix es el socio de valor agregado que implementó la plataforma inicialmente y que en la actualidad administra y da soporte a la plataforma del centro de datos principal y de contingencia. Este proyecto contempla el cambio de plataforma de virtualización tercerizada a uno con gestión propia basado en los recursos tecnológicos actuales y contratados dentro de la línea base del socio de tecnología IBM. Se pretende que a través de una herramienta licenciada llamada Microsoft Terminal Server, se permita a los equipos cliente conectarse a un equipo remoto y utilizar programas instalados en él.

Según el funcionario, los programas o aplicaciones remotas facilitan la administración del sistema porque no hay sólo una copia de un programa para actualizar o mantener en lugar de muchas copias instaladas en equipos individuales para cada usuario.

Teniendo en cuenta el catálogo de servicios, el inventario de aplicaciones y los roles y perfiles de los usuarios se asignan bajo grupos de seguridad las aplicaciones y los recursos necesarios para la ejecución de las actividades de cada uno.

El día 29 de enero de 2020 en las oficinas del NOC (Network Operation Center) de Colpensiones se realiza una reunión de contextualización con el analista del área en donde se expuso el alcance del proyecto. El analista procedió a realizar una explicación de la herramienta de monitoreo y el alcance sobre la plataforma de virtualización actual en donde se identifica que solamente se tiene visualización de la cantidad de usuarios conectados a la granja de servidores de Agilitix y el porcentaje de uso de la memoria y CPU de los servidores. El analista refiere las ventajas y oportunidades al tener un mejor sistema de visualización que el actual. No se evidencia la existencia de reportes, tableros de control o utilitarios para el monitoreo de conexiones de usuarios a las aplicaciones o escritorios. De acuerdo con lo expuesto, el área NOC monitorea y solamente reacciona ante posibles irregularidades o alertamientos que se generen y son escalados por los operadores de turno a los responsables quienes finalmente deben realizar cualquier acción para la mitigación de las alarmas reportadas. En este caso, las acciones se limitan a tomar acciones correctivas ya que no se tienen las herramientas necesarias que permitan realizar análisis predictivos sobre el comportamiento de los usuarios, de las aplicaciones o de los escritorios.

El día 13 de febrero se realizó una segunda reunión con el analista del grupo Capa Media en donde informó que se dispondrá de más de 70 aplicaciones virtualizadas para los usuarios. Igualmente refirió que no se tiene implementado en este momento ningún reporte sobre el uso de las aplicaciones ni de los escritorios. Tampoco se evidenció la existencia de tableros de control ni utilitarios que muestren información concerniente. Solamente refiere que se están guardando en una base de datos SQL los registros generados por la autenticación de los usuarios a las sesiones del Terminal Server.

El día 21 de febrero de 2020 en las oficinas de la Dirección de Infraestructura Tecnológica se realiza una reunión de contextualización con el PMP del socio de tecnología para el proyecto Colpensiones. Actualmente están ejecutando pruebas sobre la implementación de la herramienta IBM APM (Application Performance Management) para realizar el monitoreo, la visibilidad y el control completos de la infraestructura y del entorno de las aplicaciones.

Según lo indicado por el PMP, con IBM APM se podrá identificar los cuellos de botella y encontrar rápidamente la causa de los problemas de la aplicación. Este proyecto se encuentra en fase de pruebas de concepto para Colpensiones. El alcance del proyecto no contempla análisis de uso de las aplicaciones ni los escritorios por parte de los usuarios, únicamente el rendimiento de las aplicaciones.

6.1.3 Requerimientos

Las necesidades de la Organización se enfocan en tres grandes aspectos que considera importantes y que se abordarán en el desarrollo de este proyecto para la solución del problema planteado:

- Infraestructura
- Usuarios
- Aplicaciones

Desde cada una de estas aristas se buscó analizar el comportamiento por las diferentes variables del negocio teniendo en cuenta las fuentes de información disponibles como el catálogo de servicios, los sistemas transaccionales (OLTP), la actual plataforma de virtualización, el directorio activo (DA), el Web-Site, el inventario de los componentes de infraestructura y la CMDB para finalmente, obtener un sistema OLAP que permita desde múltiples dimensiones, obtener información que aporte valor a las decisiones de la Entidad.

El análisis de la infraestructura se efectuó únicamente para conocer la carga transaccional proveniente de las conexiones de los usuarios y las peticiones realizadas para el uso de las aplicaciones.

No fue del alcance de este análisis el comportamiento de los recursos físicos (CPU, Memoria, Disco) de los servidores de la granja de virtualización administrados por el tercero contratado actualmente.

El análisis buscó dar respuesta a las siguientes preguntas de la Entidad:

- ¿Cuáles son los servidores de la granja con mayor carga transaccional por cantidad de conexiones?
- ¿Cuál ha sido el comportamiento de los usuarios referente al uso de las aplicaciones y los escritorios y cómo podría ser su comportamiento futuro?
- Caracterización de los usuarios: ¿Quién es el que más se conecta? ¿Quién es el que menos se conecta?
- ¿Cuáles son las fechas de mayor afluencia de usuarios?
- ¿Cuál es la hora pico de conexiones?
- Caracterización de las sedes de trabajo: ¿Cuáles son las sedes que más se conectan?
- ¿Cuáles son las sedes que menos se conectan?
- ¿Cuáles son las aplicaciones más usadas?
- ¿En qué periodos de tiempo (mes, día, hora) se usan más?
- ¿De dónde provienen las conexiones y cuál es su distribución?
- ¿Cuál es el tiempo de conexión de los usuarios?
- ¿Cuál es el tiempo de uso de las aplicaciones?
- ¿Cuál es el tiempo de inicio de sesiones?
- ¿Cómo se relaciona el tiempo de conexión de los usuarios con las aplicaciones utilizadas?
- ¿Qué aplicaciones son las más usadas en los diferentes momentos del mes/semana?
- ¿Qué relación existe entre las diferentes áreas de la empresa con las aplicaciones usadas y el tiempo de conexión de los usuarios?
- ¿Según el uso de las diferentes aplicaciones por usuario es posible reducir las licencias utilizadas y así reducir costos?
- ¿Cuál es el costo por cada una de las aplicaciones usadas?
- ¿Cuál es la proporción de conexiones internas y externas a las aplicaciones y escritorios de la entidad?

Teniendo en cuenta estas preguntas, se generaron los requerimientos para la solución del proyecto:

Nombre	RF01-El sistema deberá permitir el análisis de la carga transaccional de los servidores.
Resumen	Se debe realizar el análisis de uso de los servidores que soportan la plataforma de los escritorios y las aplicaciones virtualizadas. Se debe conocer, en un periodo de tiempo específico, la cantidad de usuarios conectados por cada servidor, el promedio de la duración de las conexiones y la distribución de estas por cada plataforma virtualizada.
Entradas	
Información de identificación del servidor, plataforma virtual y fechas de conexión.	
Resultados	
Análisis de la carga transaccional por cantidad de conexiones a servidores por periodo de tiempo.	

Tabla 6 Requerimiento Infraestructura – Fuente: Autores

Realizar análisis comparativos por periodos de tiempo sobre los componentes de la infraestructura busca identificar la tasa de crecimiento o de reducción de los recursos lo cual permitirá la optimización y la mejora continua de la infraestructura.

Nombre	RF02-El sistema deberá permitir el análisis del comportamiento y tendencias de los usuarios.
Resumen	Se requiere analizar los hábitos, comportamientos y tendencias de conexiones por parte de los usuarios. Se requiere conocer las sedes con mayor cantidad de usuarios conectados y el uso de aplicaciones, origen de las conexiones, tiempos de conexión, usuarios con más y menos número de conexiones, duración total de las conexiones y promedio diario, fechas y horas de inicio y fin de las conexiones por usuario y las aplicaciones usadas.
Entradas	
Información de identificación del servidor, hora de conexión, día, mes, identificación de usuario,	
Resultados	
Análisis del comportamiento de los usuarios referente a las conexiones y modelo predictivo de su comportamiento.	

Tabla 7 Requerimiento Usuarios – Fuente: Autores

Analizar el comportamiento de las aplicaciones por periodos de tiempo y por áreas funcionales para gestionar su licenciamiento, determinar su uso, rendimiento, prestigio, experiencia del usuario, funcionalidad y optimización de la plataforma sobre la que se ejecutan.

Nombre	RF03-El sistema deberá permitir el análisis del comportamiento y tendencias del uso de las aplicaciones.
Resumen	Se requiere generar un análisis del uso de las aplicaciones y los escritorios virtuales. Analizar los hábitos, comportamientos y tendencias de uso de las aplicaciones por parte de los usuarios. Es necesario también conocer qué aplicaciones tienen mayor frecuencia de uso, mayor cantidad de usuarios, mayor tiempo de utilización, origen de las conexiones, tiempos de permanencia de los usuarios en las aplicaciones, tiempos de respuesta de las aplicaciones para el inicio de las sesiones, picos y valles de las aplicaciones.
Entradas	
Información de identificación del servidor, hora de conexión, día, mes, identificación de usuario, aplicación, origen, tiempo de conexión.	
Resultados	
Análisis del comportamiento de los usuarios referente al uso de las aplicaciones y los escritorios y modelo predictivo de su comportamiento.	

Tabla 8 Requerimiento Aplicaciones – Fuente: Autores

6.2 Entendimiento y preparación de los datos

6.2.1 Información recolectada

Se recolectó información de los meses de octubre, noviembre y diciembre de 2019 desde las diferentes fuentes de información mencionadas anteriormente:

- Servidores
- Usuarios
- Segmentos
- Sedes

- Plataformas
- Aplicaciones
- Perfiles
- Vicepresidencias
- Gerencias
- Direcciones
- Conexiones

Esta información de obtuvo de las fuentes de información disponibles en la Entidad, por ejemplo, la información de los usuarios se obtuvo del Directorio Activo, el inventario de aplicaciones so logró obtener a partir del catálogo de servicios. La siguiente tabla muestra el origen de los datos obtenidos.

TABLA	FUENTE DE DATOS
Servidores	Plataforma de Virtualización
Usuarios	Directorio Activo
Segmentos	Firewall de Red
Sedes	Sitio web / Organigrama Institucional
Plataformas	Plataforma de Virtualización
Aplicaciones	Catálogo de Servicios
Perfiles	Plataforma de Virtualización
Vicepresidencias	Sitio web / Organigrama Institucional / Directorio Activo
Gerencias	Sitio web / Organigrama Institucional / Directorio Activo
Direcciones	Sitio web / Organigrama Institucional / Directorio Activo
Conexiones	Sistema Transaccional / Plataforma de Virtualización

Tabla 9 Fuente de datos – Fuente: Autores

La información extraída se convirtió en tablas estructuradas indexadas y se creó un diccionario que contiene la información de cada uno de los campos.

6.2.2 Análisis de calidad de datos

Como parte esencial del proyecto se considera necesario realizar el análisis de la calidad de datos obtenidos durante el proceso de levantamiento de información. A pesar de que la mayoría de los datos provienen de fuentes y sistemas ya normalizados, se procede a realizar el análisis con la herramienta *DQ Analyzer* para garantizar un alto porcentaje de calidad (75-95%). Al final del ejercicio se obtiene un 99,69% de calidad en los datos, considerado optimo para la realización de los modelos del presente proyecto. El resultado del análisis realizado se detalla en el Anexo 2.

6.2.3 Análisis estadístico de datos para la etapa de modelado

6.2.3.1 Preparacion de los datos para el modelo de clustering

Se realizó una exploración estadística inicial de los datos, eliminando variables que no son necesarias para el análisis, identificando los tipos de datos, y graficando sus distribuciones. Se eliminaron variables como:

- fechainicio
- inicio
- fin
- cliente
- ip
- plataforma
- ID_Usuario
- Nombre
- Inicial
- Apellido
- Nombreparamostrar
- Descripcion
- Telefono
- Direccion y Ciudad
- Servidor y usuario
- segmento

Que no contienen información relevante para el análisis. Adicionalmente se eliminaron variables que no poseían información o cuyo valor era único:

- accountlocketout
- passwordlastset
- cannotchangepassword
- passwordneverexpired
- EstadoProvincia
- CodigoPostal
- Pais
- WebPage
- Compañia
- Email
- Tipo
- tipo_conexion
- Estado

Finalmente se filtraron únicamente los registros que se relacionan con el inicio de aplicaciopnes, ya que los datos contiene también información acerca del inicio de sesión.

6.2.3.1.1 Variables numéricas

En este análisis la única variable continua es “Duración” que muestra la cantidad de minutos en la que se usó determinada aplicación.

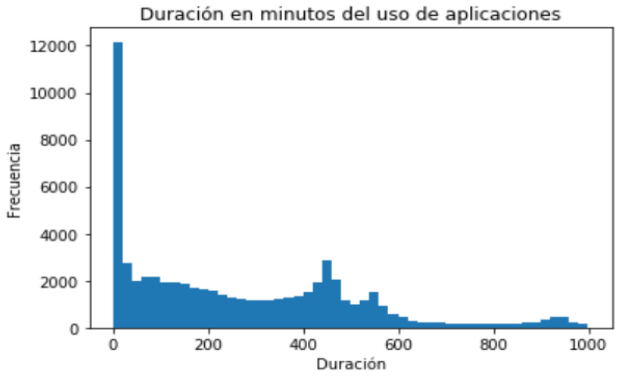

Figura 11 Duración en minutos del uso de las aplicaciones – Fuente: Autores

Se observa en la figura 11 que sus valores van desde 0 a 999 y que hay una gran cantidad de aplicaciones que se usan durante menos de un minuto, lo que se debe a caídas de las aplicaciones antes de que sea posible usarlas.

6.2.3.1.2 Variables categóricas

A continuación, en la figura 12, se muestra el análisis estadístico de cada variable categórica utilizada en el análisis:

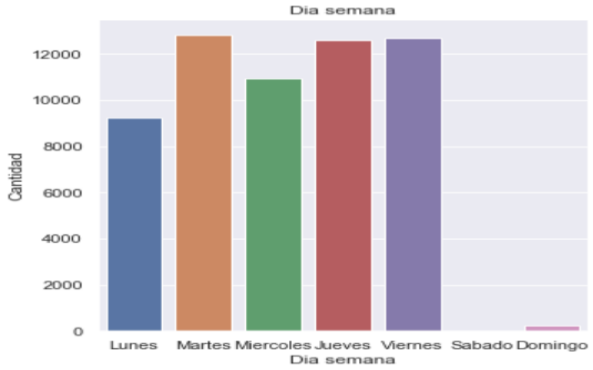

Figura 12 Variables categóricas día-semana – Fuente: Autores

Se observa que los días de más actividad en el uso de las aplicaciones son los martes, jueves y viernes, en la ventana de tiempo analizada hay bastantes lunes festivos, por lo que esto se ve reflejado en la actividad del lunes. Curiosamente el domingo hay más actividad que el sábado, esto obedece a las campañas comerciales que realiza la entidad para el acercamiento a los ciudadanos y que se realizan en sitios públicos como los centros comerciales.

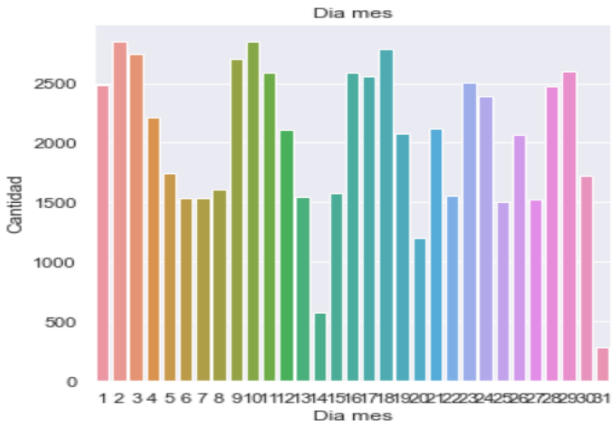

Figura 13 Variables categóricas día-mes – Fuente: Autores

En la figura 13 se muestra que hay unos picos de los días del en las que los usuarios se conectan más a las aplicaciones, como los primeros días, días previos y posteriores a la mitad del mes y a final de mes. La figura 14 muestra las horas en las que se usan las aplicaciones, las horas pico de su uso son de 6 am a 8 am y de 2 pm a 3 pm.

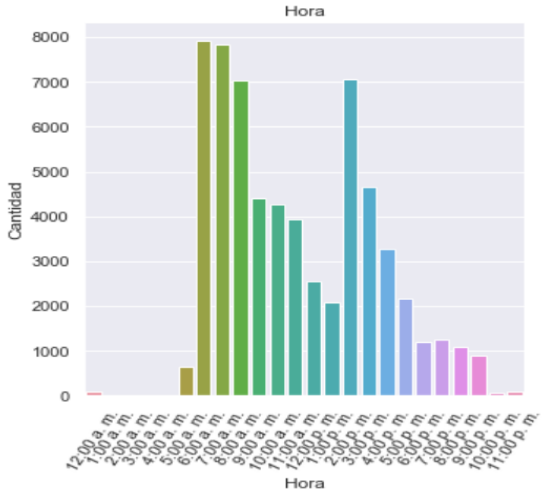

Figura 14 Variables categóricas hora – Fuente: Autores

Aplicación	Cantidad
IBM Consulta Pagos	16188
IBM Historia Laboral Tradicional	8470
IBM CORRDNC	5863
Historia Laboral Imputacion 157	4175
SoapUI 715	3340
DevolucionIngresosAFP	2153
IBM_Consgre_ISS	2013
Notepad++ 715	1983
Excel Avanzado 715 v8	1967
Interactive SQL 715	1759
SyBase Central 715	1727
IBM_Historico Subsidiados	1599
Filezilla 715	1513
IBM_CONSDNC	1337
AutoAudit_715	999

IBM_Modican_ISS	744
IBM_Devoluciones y Egresos 15	681
Putty 715	583
SQL Management 2017	573
SQL Server Management Studio 715	507
IBM_Devoluciones y Egresos 11	402
IBM_Consulta Pagos Inconsistencia	363
Whisper 715	356
Nomina Antigua	308
WinSCP 715	243
Debido Cobrar	223
HashMyFiles 715	195
IBM_Consgre	171
Verigre	170
Enterprise Architect Full	160

Tabla 10 Top de uso de aplicaciones – Fuente: Autores

En la tabla de aplicaciones se observa que la más usada es IBM Consulta Pagos, existen múltiples aplicaciones que tienen un único uso en los datos analizados por lo que no se muestran en la tabla.

Vicepresidencia	Cantidad
Vicepresidencia de operaciones del régimen de prima media	20817
Vicepresidencia de gestión corporativa	13919
Vicepresidencia comercial y de servicio al ciudadano	12740
Vicepresidencia de planeación y tecnologías de la información	9281
Vicepresidencia de seguridad y riesgos empresariales	3475
Vicepresidencia de beneficios económicos periódicos	2399

Tabla 11 Cantidad de conexiones por vicepresidencia - Fuente: Autores

La vicepresidencia que posee un mayor uso de las aplicaciones es la de operaciones del régimen de prima media, mientras que la que les da un menor uso es la de beneficios económicos periódicos.

Gerencia	Cantidad
Gerencia administrativa	10864
Gerencia de servicio y atención al ciudadano	9606
Gerencia de financiamiento e inversiones	8506
Gerencia de tecnologías de la información	5965
Gerencia de defensa judicial	5171
Gerencia de administración de la información	3851
Gerencia de planeación institucional	3316
Gerencia de determinación de derechos	3289
Gerencia comercial	3134
Gerencia de talento humano y relaciones laborales	3055
Gerencia de prevención del fraude	2008
Gerencia de riesgos y seguridad de la información	1467
Gerencia de administración de cuentas individuales	1213
Gerencia de redes e incentivos	1186

Tabla 12 Uso de aplicaciones por gerencia – Fuente: Autores

La gerencia que posee un mayor uso de las aplicaciones es la administrativa, mientras que la que les da un menor uso es la de redes e incentivos.

Dirección	Cantidad
Direcciones regionales	3485
Dirección de acciones constitucionales	3400
Dirección de bienes y servicios	2956

Dirección de relacionamiento con el negocio	2843
Dirección de contribuciones pensionales y egresos	2591
Dirección de administración de solicitudes y PQRS	2416
Dirección de inversiones	2322
Dirección de afiliaciones	2314
Dirección de Tesorería	2185
Dirección financiera	2181
Dirección de estandarización	2158
Dirección de prevención del fraude	2008
Dirección de comercialización y acompañamiento empresarial	1912
Dirección documental	1906
Dirección de sistemas de información	1894
Dirección de cartera	1893
Dirección de prospectiva y estudios	1807
Dirección de procesos judiciales	1771
Dirección de ingresos por aportes	1700
Dirección de desarrollo del talento humano	1657
Dirección contractual	1636
Dirección de atención y servicio al ciudadano	1547
Dirección de historia laboral	1537
Dirección de planeación y proyectos	1509
Dirección de riesgos y seguridad de la información	1467
Dirección de gestión del talento humano	1398
Dirección de infraestructura tecnológica	1228
Dirección de mercadeo	1222

Dirección de administración de cuentas individuales	1213
Dirección de redes e incentivos	1186
Dirección de prestaciones económicas	1137
Dirección de nómina de pensionados	1116
Dirección de medicina laboral	1036

Tabla 13 Uso de aplicaciones por dirección – Fuente: Autores

La dirección que posee un mayor uso de las aplicaciones es la regional, mientras que la que les da un menor uso es la de medicina laboral.

6.2.3.2 Preparacion de los datos para el modelo de serie de tiempo

En el caso de la serie de tiempo fue necesario agrupar y sumar los costos diarios por tipo de conexión según el escritorio utilizado, por lo que en primer lugar se filtraron solo los datos de inicio de sesión y posteriormente se hizo la consulta para obtener un *data frame* que contuviera los valores de costo totalizados por día, el resultado obtenido se muestra en la figura 15.

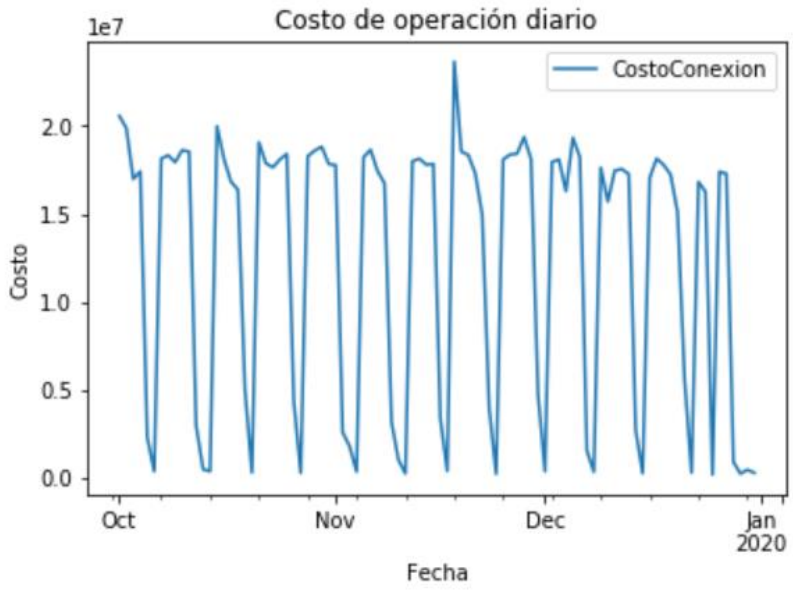

Figura 15 Serie de tiempo costo diario de operación – Fuente: Autores

6.3 Modelado Multi-Dimensional

6.3.1 Propuesta de Esquema de Datawarehouse

En la figura 16 se muestra el diagrama lógico propuesto para la implementación del modelo de B.I en donde se puede observar cada una de las ramas de interés para la obtención y análisis de información.

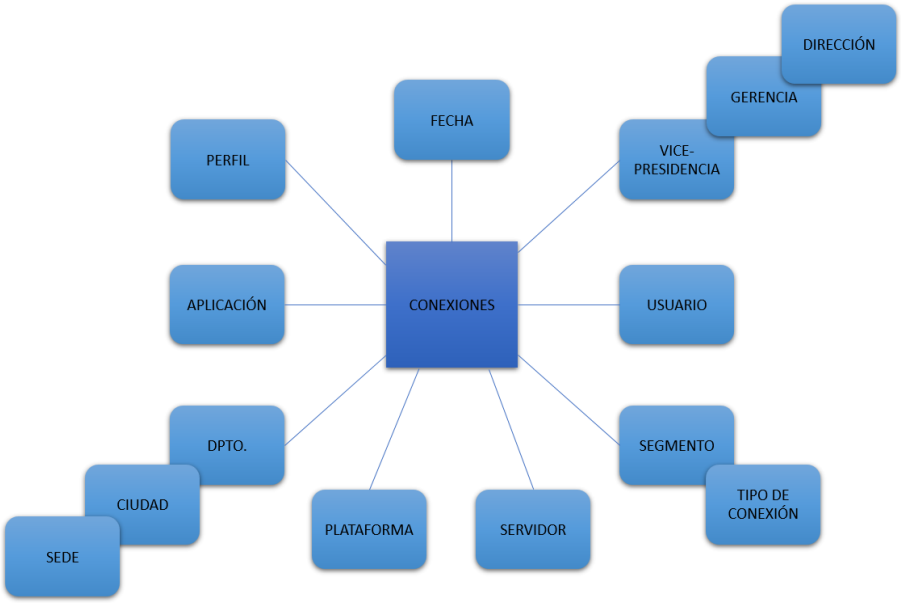

Figura 16 Modelo lógico – Fuente: Autores

De acuerdo con lo establecido en el proceso de levantamiento de la información, se requirió tener en cuenta en el desarrollo de la bodega de datos la información que permita conocer de forma granular, la cantidad de conexiones que provengan desde las diferentes ciudades y sedes de la Entidad, así como tener granularidad según la organización jerárquica. También se requiere conocer el perfil de conexión de los usuarios y el costo generado por cada conexión. El modelo lógico planteado permitió cumplir con el requerimiento de los interesados.

La arquitectura dispuesta para la implementación del modelo de *data warehouse* contempla el desarrollo de un único *datamart* ya que las necesidades del negocio y el alcance del proyecto delimitan el interés de la información solamente a la Gerencia de Tecnologías de la Información. Para lograr el objetivo, se tuvo en cuenta la obtención de los datos desde las siguientes fuentes de información:

- **Directorio Activo:** Se obtuvo la información pertinente a los usuarios, la vicepresidencia, gerencia y dirección a la que pertenece (Unidades Organizativas).
- **Catálogo de Servicios:** Contiene el inventario de las aplicaciones del negocio y la función de cada una.
- **Sistema Transaccional:** Contiene los registros de inicio y fin de sesión de los usuarios y los correspondientes a las aplicaciones o peticiones de los usuarios registrados. El sistema transaccional almacena la fecha y hora exacta de autenticación y de terminación de la sesión.
- **Plataforma Virtual:** Se obtuvo la información concerniente a los perfiles establecidos, los segmentos de red involucrados y las plataformas usadas de acuerdo con las transacciones de los usuarios.
- **Estructura Organizacional:** Para la obtención del organigrama de la entidad y la dependencia entre vicepresidencias, gerencias y direcciones.

La figura 17 ilustra el modelo de arquitectura utilizado y sus diferentes componentes necesarios para la realización de los análisis descriptivos y predictivos planteados en este proyecto. Se observa las fuentes de datos, el componente ETL, el sistema administrador de las bases de datos, las herramientas analíticas y de visualización:

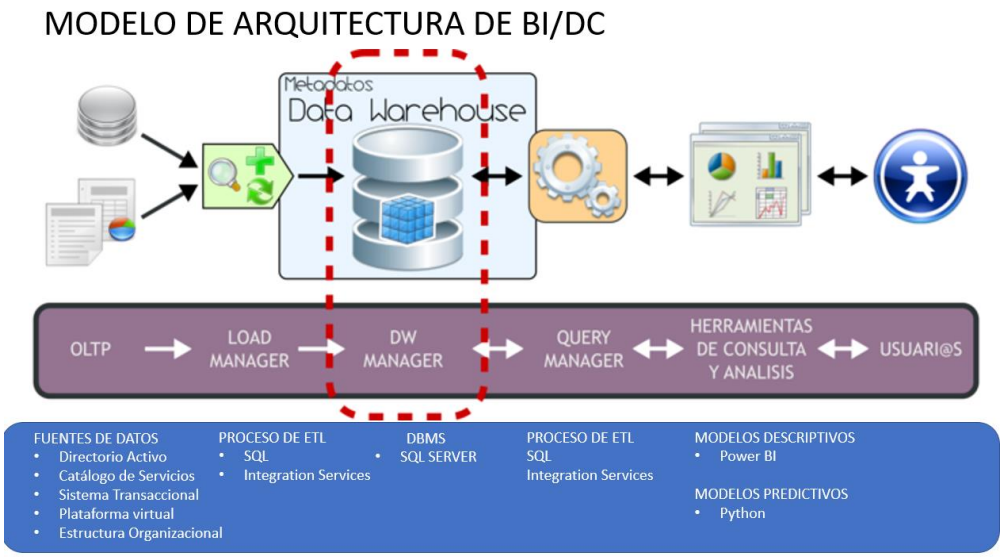

Figura 17 Modelo de arquitectura de DWH - Fuente: Autores

En la figura 18 se ilustra el modelo *StarNet* que muestra los campos de la tabla de hechos relacionados con cada una de sus dimensiones. Este admitió el análisis de la información obtenida para dar respuesta desde las diferentes perspectivas de interés de los usuarios. Por ejemplo, permitió conocer el porcentaje de uso de las aplicaciones por cada usuario o el costo de las conexiones en un periodo de tiempo por cada vicepresidencia, gerencia o dirección de la Entidad. La figura 18 enseña el modelo multidimensional propuesto para el *datamart*.

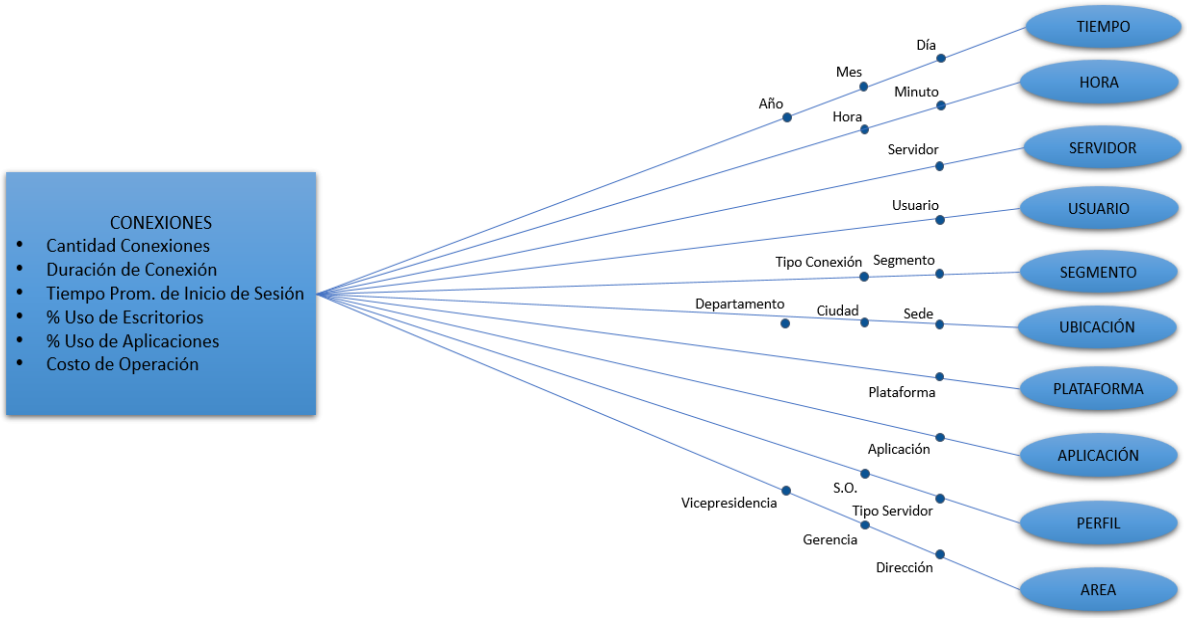

Figura 18 Star Net – Fuente: Autores

El modelo del *datamart* exhibido plasma el diseño de la *StarNet* realizado y permitió realizar el análisis del uso de aplicaciones, la duración, el costos de conexión, y el tiempo promedio de su uso desde las dimensiones del usuario, área, servidor, etc. teniendo en cuenta variables de tiempo y espacio como el filtrado por un periodo de tiempo inicial y final, hora de inicio y fin de la conexión y desde una ubicación geográfica definida.

La dimensión Servidores permitió la realización del análisis descriptivo por los componentes de infraestructura dispuestos en la granja de virtualización. Cada servidor recibe un número de peticiones de usuarios para el cargue de su perfil y de las aplicaciones necesarias para la realización de sus labores. Este análisis permite conocer la carga transaccional de cada uno de los servidores dispuestos y propone una oportunidad de mejora en el sistema de balanceo de cargas. La dimensión áreas contiene las diferentes vicepresidencias, gerencias y direcciones de la Entidad

y su correspondiente relación permitiendo la granularidad del análisis y respondiendo a los intereses de la organización. En la dimensión Segmentos se puede determinar los orígenes de la conexión de los usuarios, que pueden ser conexiones internas o externas a la red institucional. Para conocer los orígenes físicos de las conexiones se incorporó la dimensión Ubicación que permitió identificar los departamentos, ciudades y sedes desde donde se realizaron las conexiones. La dimensión Plataforma y Perfil permiten analizar las métricas desde cada una de las estructuras tecnológicas dispuestas para el entorno virtualizado y los recursos asignados de acuerdo con el nivel del usuario y sus funciones. La dimensión Aplicaciones examina el comportamiento de cada una de las herramientas de trabajo asignadas a los usuarios y su relación con sus funciones. Esta dimensión permitió conocer los hábitos de trabajo sobre las aplicaciones y su rendimiento.

En el modelo del *datamart* se incluyen dimensiones iniciales de tiempo, hora y minuto para permitir realizar análisis filtrados por criterios de fecha y hora específicos. Se incluye también dentro de la dimensión Hora_Ini y Hora_Fin el atributo Franja para realizar los análisis de acuerdo con los turnos de trabajo establecidos.

La tabla de hechos que finalmente contiene la información y los valores de las medidas requeridas para el *data warehouse*, está compuesta por los indicadores de las dimensiones y se establece como la tabla central del modelo. Esta tabla contiene las llaves principales de las dimensiones que posteriormente se transforman en llaves foráneas permitiendo así la integración y la relación en modelo multidimensional.

Principalmente, la tabla de hechos contiene los valores numéricos que se desean medir, para este caso, contiene cada registro de conexión realizado por los usuarios, la duración, el tiempo de inicio de cada sesión, la aplicación que utilizó un usuario, el perfil, fecha y hora exacta de inicio y fin de cada conexión, el área y la ciudad o sede origen de la conexión.

En la figura 19 se ilustra el modelo de *datamart* propuesto con sus dimensiones y tabla de hechos mencionados anteriormente:

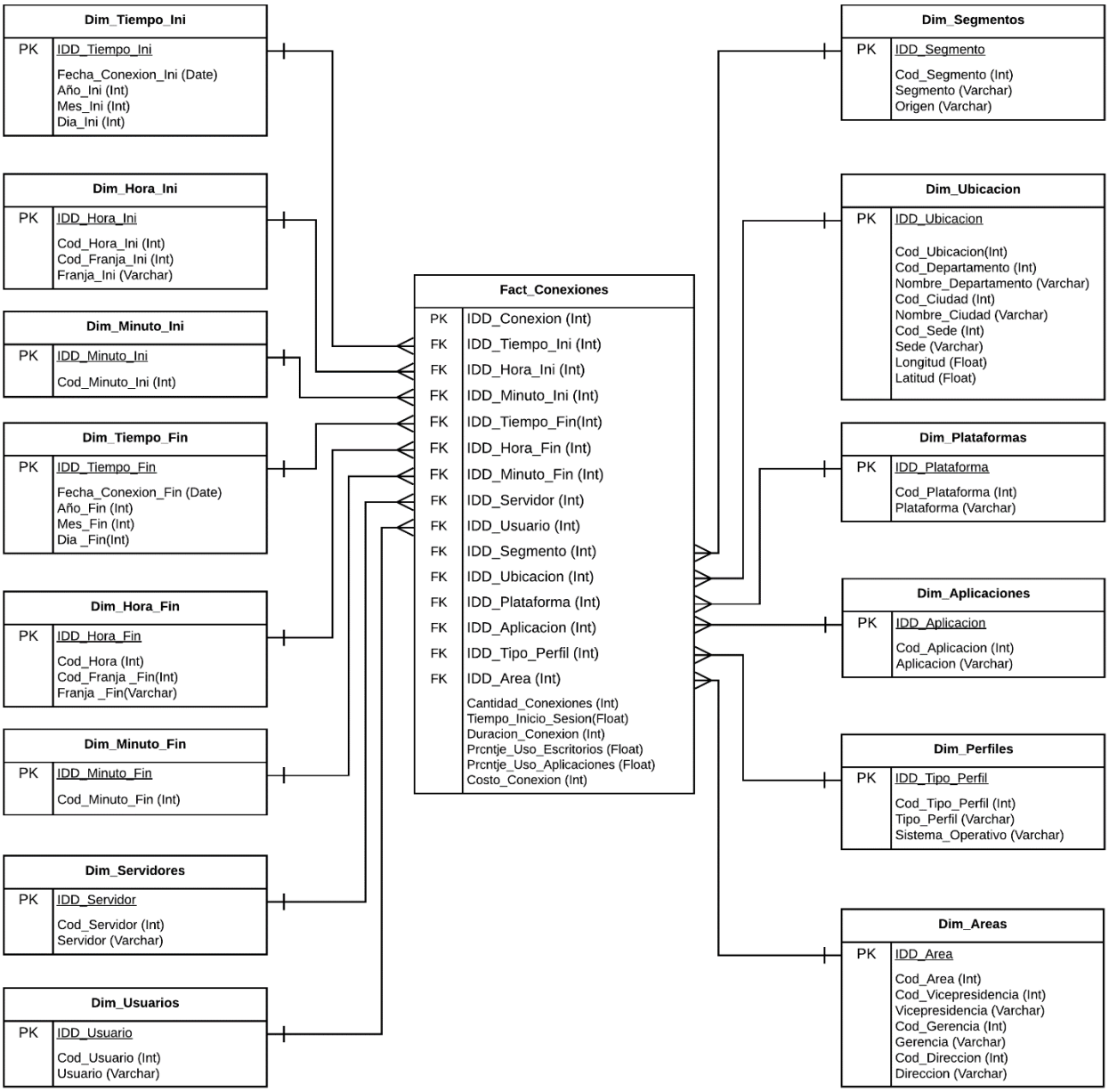

Figura 19 Modelo datamart propuesto – Fuente: Autores

6.3.2 Diseño del Modelo de Datos Transaccional

A continuación, en la Figura 20, se muestra el modelo implementado para la base de datos transaccional, en el cual se observan las tablas maestras y sus relaciones con la tabla de conexiones que registran la actividad de los usuarios. Se observa la relación jerárquica entre las vicepresidencias, las gerencias y las direcciones a la cual pertenece un usuario. Para el modelo

multidimensional se tuvo en cuenta la dependencia y la posible pérdida de data histórica si un usuario llega a ser trasladado de dirección realizando una ETL que transformara y cargara la data.

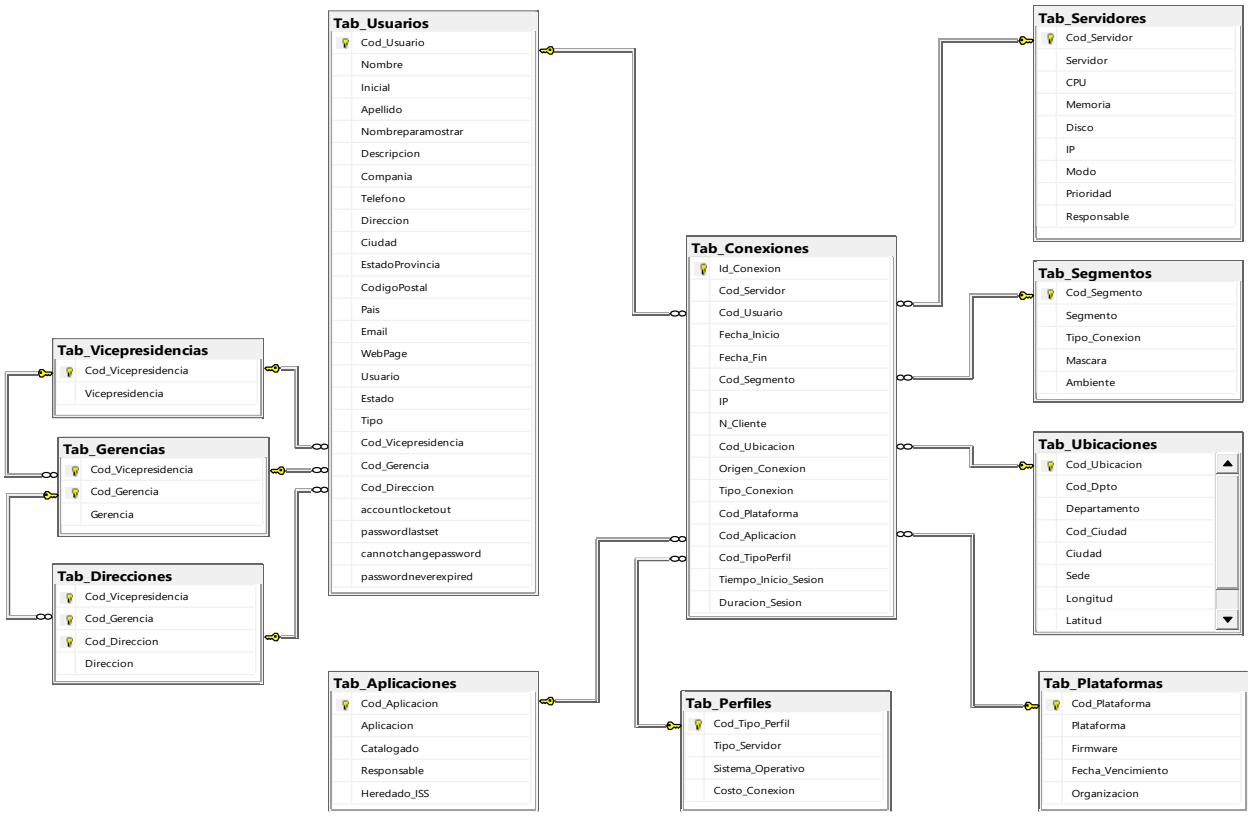

Figura 20 Diseño lógico del sistema OLTP – Fuente: Autores

6.3.3 Implementación de los modelos de extracción, transformación y carga

Siguiendo la metodología planteada para el desarrollo del proyecto, a partir de la base de datos transaccional (OLTP) y de las fuentes de datos, se construyeron las ETL's necesarias para la alimentación del área de almacenamiento y procesamiento intermedia llamada *staging area*. Los datos organizados en esta sección se convierten en la fuente de información que posteriormente alimentará las dimensiones del modelo y reposarán en las bodegas de datos del *datamart*.

Se crearon dos procesos de ETL: el primero para facilitar el proceso de extracción y transformación de la información del sistema transaccional y cargarlo en la *staging area*, y el segundo, para transformar la información recolectada y subirla en el *data warehouse*. En las figuras 21 y 22 se puede apreciar los modelos de ETL's anteriormente mencionados y que se ejecutaron en Visual Studio con Analysis Services para la obtención de los resultados:

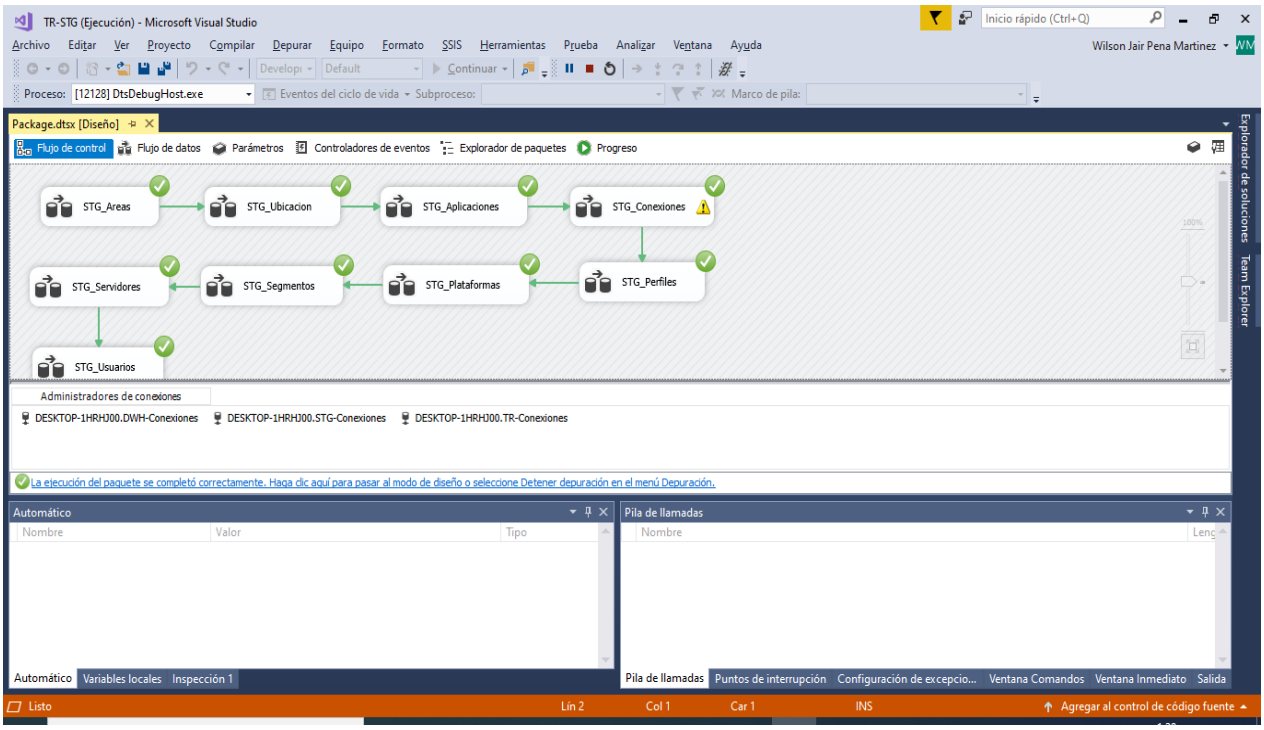

Figura 21 Modelo de ETL TR – STG Area – Fuente: Autores

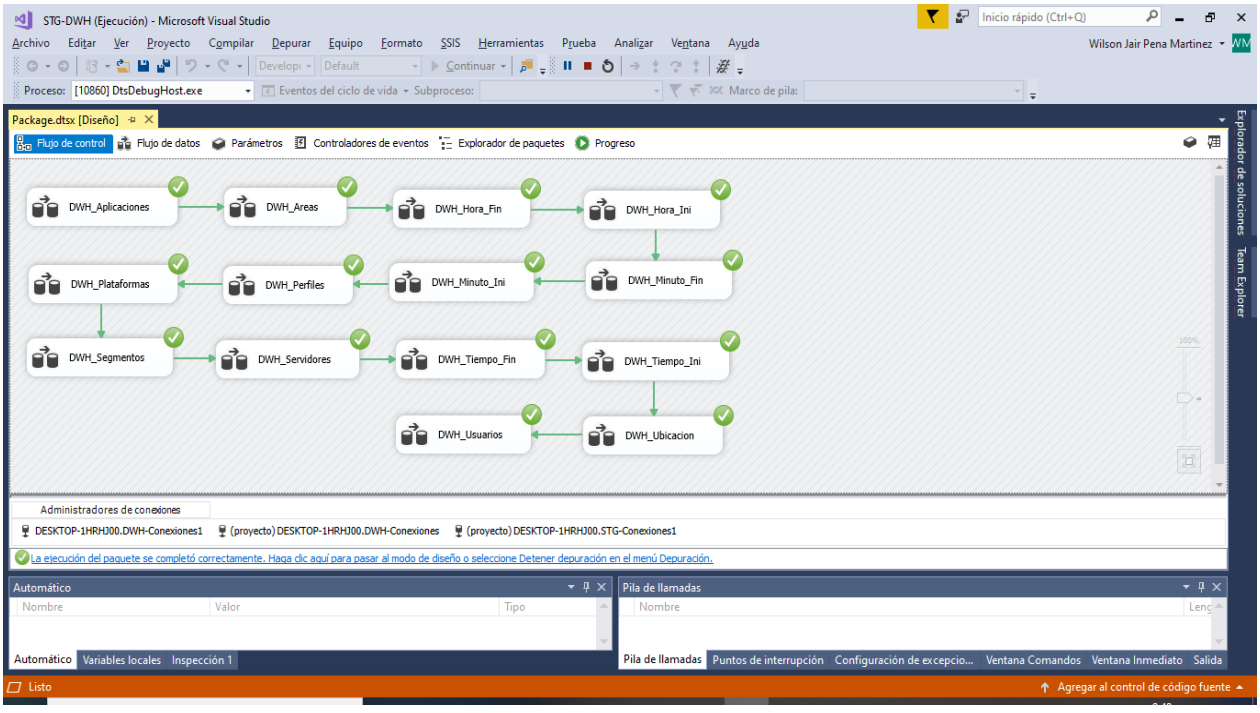

Figura 22 Modelo de ETL STG-DWH - Fuente: Autores

El detalle de la definición de los procesos ETL se pueden observar en el anexo 4 del presente proyecto.

La figura 23 ilustra las tablas dimensionales y la tabla de hechos cargadas en el *datamart* a partir de la ejecución de las ETL mencionadas. La tabla “Fact_Conexiones” contiene 150.489 registros correspondientes a la transaccionalidad efectuada durante 90 días.

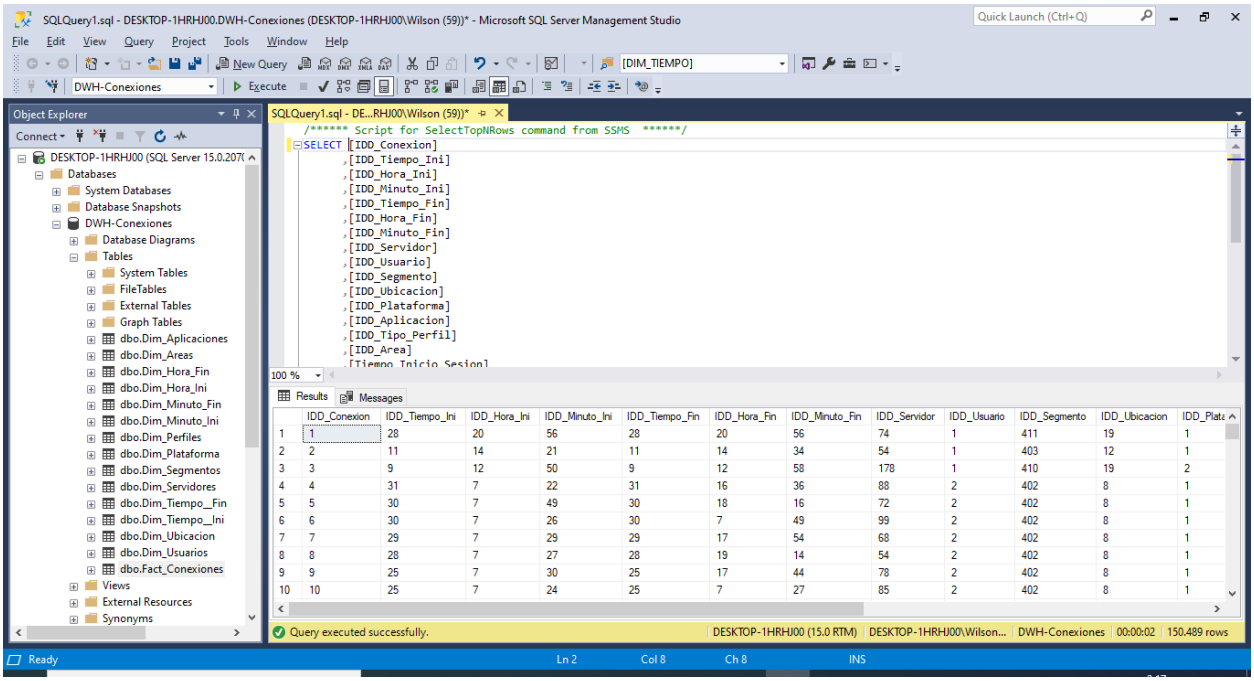

Figura 23 Tabla de hechos data warehouse - Fuente: Autores

Una vez teniendo la data cargada en las bases de datos del *datamart* y siguiendo la metodología propuesta, se dispone de la estructura necesaria para realizar las propuestas de visualización que den respuesta a las necesidades planteadas, modelos predictivos, análisis y evaluación.

6.4 Modelado y evaluación de la etapa predictiva

Se utilizaron 3 técnicas de minería de datos para dar una visión más amplia sobre la operación de las aplicaciones, en la primera se crearon seis modelos de clusterización, uno por cada vicepresidencia de la compañía, esto con el fin de agrupar los usuarios con comportamientos similares, lo anterior teniendo en cuenta variables como, la hora, el día de la semana y del mes, las aplicaciones que usan, por cuanto tiempo las usan y las gerencias a las que pertenecen, lo que permite ver las necesidades de los usuarios e implementar posibles mejoras.

En caso de que lleguen nuevos usuarios a la entidad, para que el sistema sea capaz de predecir a qué grupo pertenecen se utilizó una segunda técnica, un modelo predictivo de redes neuronales, que tiene como entrada las mismas variables y clasifica a los usuarios nuevos según su comportamiento. Finalmente, para visualizar el comportamiento de cada grupo se creó una pestaña en el dashboard que muestra gráficamente cómo están compuestos los grupos según las diferentes variables que los componen.

La tercera técnica consiste en una serie de tiempo que predice los costos de la operación, al tener la operación de las aplicaciones virtualizadas como propia la empresa necesita tener un control sobre los costos que esta le genera, además de poder anticiparse. Con el tiempo la empresa irá recogiendo más datos que permitirán que el modelo mejore y muestre los costos con una mayor precisión.

6.4.1 Modelo de cluster K-means

Se planteó un modelo de clusterización con la técnica de K-means en el que se segmentaron las conexiones de los usuarios a las aplicaciones discriminados por la vicepresidencia a la que pertenecen, se utilizó el método del codo para identificar la cantidad de clústeres adecuada para cada vicepresidencia.

La mayoría de las variables elegidas para crear el modelo son categóricas como: el día de la semana, el día del mes, la hora, la aplicación usada y la dirección a la que pertenece el usuario.

La única variable numérica en el análisis es la duración del uso de la aplicación. La técnica de *clustering* necesita de variables numéricas ya que lo que hace es medir distancias entre los puntos de un plano, por lo que se realizó la conversión de variables categóricas a numéricas usando variables "*dummy*", es decir, creando una nueva columna por cada categoría de la variable y denotar cada ocurrencia con un "1" o en caso contrario con un "0", esto para evitar que existiera un orden o jerarquía en las categorías.

Cada modelo se evaluó con las medidas de: inercia que mide la coherencia interna de los clústeres y cuyo valor entre más cercano sea a "0" indica un mejor modelo, esta medida no tiene una escala definida por lo que es difícil saber qué valor indica mejores grupos, índice Davies and Bouldin que indica la "similitud" promedio entre grupos, donde la similitud es una medida que

compara la distancia entre grupos con el tamaño de estos, en este caso también valores cercanos a "0" indican una mejor separación entre los grupos y el índice de silueta donde un valor alto significa un mejor modelo, ya que este índice mide la definición de los clústeres, teniendo en cuenta la distancia desde una muestra a todos los otros puntos en su mismo grupo y de la misma muestra a los puntos en el clúster más cercano (Scikit-learn, 2011). En cuanto a la vicepresidencia de operaciones del régimen de prima media, en la figura 24 se muestra el método del codo que se usó para determinar la cantidad adecuada de grupos, el codo se evidencia entre los números 4 y 5, se creó este modelo usando 5 clústeres.

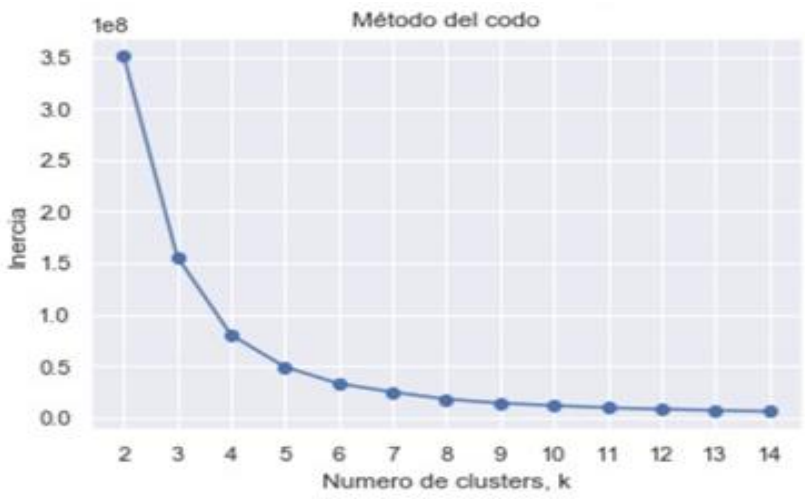

Figura 24 Método del codo – Fuente: Autores

Las métricas obtenidas en este caso fueron:

$$Inertia = 49108600{,}45$$
$$Davies\ and\ Bouldin = 0{,}47$$
$$Silhouette = 0{,}61$$

En este caso el índice de Davies and Bouldin indica una buena separación entre los grupos, la inercia presenta un valor que parece alto, sin embargo, evaluando los grupos obtenidos se considera que hay coherencia con el uso de las aplicaciones en la compañía y como se mencionó anteriormente, este no es un valor escalado.

Se realizó el mismo procedimiento para los modelos de cada una de las 6 vicepresidencias y se obtuvieron los siguientes resultados:

Vicepresidencia	Inercia	Davies and Bouldin	Silhouette	Codo
Vicepresidencia de gestión corporativa	33194692,73	0,47	0,60	5
Vicepresidencia de operaciones del régimen de prima media	49108600,45	0,47	0,61	5
Vicepresidencia comercial y de servicio al ciudadano	48003493,45	0,45	0,62	4
Vicepresidencia de planeación y tecnologías de la información	22647386,23	0,49	0,58	5
Vicepresidencia de seguridad y riesgos empresariales	12792119,15	0,44	0,63	4
Vicepresidencia de beneficios económicos periódicos	7622423,59	0,44	0,65	4

Tabla 14 Evaluación del modelo de clustering – Fuente: Autores

Se encontró que la inercia más baja entre los modelos es la de la vicepresidencia de beneficios económicos periódicos, el índice de Davies and Bouldin más pequeño también pertenece a esta vicepresidencia, junto con el de Silhouette más alto, se utilizó el método del codo para identificar la cantidad de clústeres de cada modelo y los resultados obtenidos junto con la caracterización de cada grupo se encuentran en el anexo 3. En la siguiente tabla se muestran los resultados obtenidos para el modelo de la vicepresidencia nombrada:

CLÚSTER	DÍA SEMANA	DÍA MES	HORA	APLICACIÓN	DURACIÓN	DIRECCIÓN
Validación de operaciones	Viernes, martes	24,2,10	9 a 10	IBM consulta pagos, IBM historia laboral tradicional	125-300	Dirección de acciones constitucionales, Dirección de inversiones
Consulta tutelas	Jueves, miércoles	28	7 a 9	Sybase central 715, soapui 715	750-1000	Dirección de acciones constitucionales, Dirección de cartera
Validación de derechos	Viernes, martes	9,18,11	6 a 7	IBM consulta pagos, IBM historia laboral tradicional	300-475	Dirección de acciones constitucionales, Dirección de contribuciones

						pensionales y egresos
Validación de contribuciones	Martes, jueves	1,3,18	6 a 8	IBM consulta pagos	0-120	Dirección de contribuciones pensionales y egresos, Dirección de acciones constitucionales
Validación de acciones constitucionales	Jueves, martes	18,23,29	14,7	IBM consulta pagos, IBM historia laboral tradicional	500-700	Dirección de acciones constitucionales

Tabla 15 Resultado del modelo de clustering - Fuente: Autores

Se nombraron los grupos de acuerdo con la operación y necesidades de la compañía, por ejemplo, el clúster de validación de operaciones contiene conexiones realizadas a las aplicaciones de IBM Consulta Pagos e IBM Historia Laboral Tradicional, realizadas por las direcciones de acciones constitucionales y de inversiones los viernes y martes, a mediados o finales de mes, en las horas de 9 a 10 de la mañana y con una duración de 125 a 300 minutos. Las conexiones de este clúster realizan registros y control de las operaciones concernientes de las acciones constitucionales. En el caso del grupo de consulta tutelas las personas usan las aplicaciones de SyBase Central 715 y SoapUI 715 los jueves y miércoles a finales de mes de 7 a 9 de la mañana, con una duración de 750 a 1000, la mayor entre los grupos identificados. Lo anterior basado en la cantidad de registros similares en cada clúster, es decir, se buscaron los atributos más comunes en cada grupo para ser capaces de capaces de caracterizarlo.

El código detallado de la implementación de esta técnica se encuentra en el anexo 5, junto con una breve descripción, ya que el análisis más relevante se encuentra en este documento.

6.4.2 Modelo de redes neuronales (MLP)

Para ser capaces de identificar a qué clúster de los anteriormente creados pertenece un nuevo registro, se creó un modelo supervisado de clasificación de red neuronal perceptron multicapa, siendo el nombre del clúster la variable objetivo que se utilizará para clasificar las nuevas conexiones. Para la preparación de los datos de este modelo se tomaron las mismas modificaciones hechas en el *clustering*, es decir, los datos categóricos de las variables se convirtieron a numéricos usando variables "*dummy*" que no tengan una jerarquía.

Los parámetros de la red creada son 3 capas ocultas de 2, 3 y 5 neuronas respectivamente, esta distribución fue seleccionada luego de múltiples iteraciones y teniendo en cuenta que la salida contiene 5 posible opciones, la función de activación es ReLU (Rectified Lineal Unit):

$$f(x) = \max(0, x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases}$$

Esta función transforma los valores introducidos anulando los valores negativos y dejando los positivos tal y como entran, en este caso específico ninguno de los datos de entrada es negativo, por esta razón se eligió dicha función de activación (Calvo, 2018).

Para la evaluación del modelo se utilizó una división 70/30. Luego de aplicarlo a los datos, se obtuvo una precisión de 0.97, es decir, el modelo es bastante acertado a la hora de predecir a qué clúster pertenecen los nuevos registros, en la siguiente tabla se muestra la matriz de confusión obtenida en donde se evidencian los pocos errores obtenidos con respecto a los datos reales:

6161	0	0	0	0
0	3927	78	0	0
17	0	3583	0	0
0	0	0	385	64
0	105	0	0	1883

Tabla 16 Matriz de confusión - Fuente: Autores

El código detallado de la implementación de este modelo se encuentra en el anexo 5, junto con una breve descripción, ya que el análisis más relevante se encuentra en este documento.

6.4.3 Modelo de series de tiempo para los costos diarios

Se realizó un modelo de series de tiempo para predecir el costo diario de los inicios de sesión, en primer lugar, se hizo el cruce en la base para obtener únicamente los valores de costo relacionados a las fechas, para este diseño se contó únicamente con información de 91 días, por lo que posee un valor de error alto, sin embargo, se realizaron pruebas con varios algoritmos para llegar a un mejor resultado.

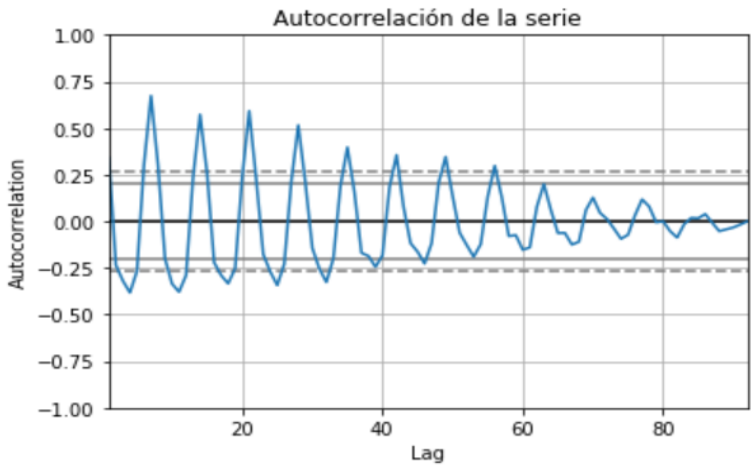

Figura 25 Autocorrelación – Fuente: Autores

Para realizar el análisis de la serie se realizó la gráfica de auto correlación que muestra que la serie se repite aproximadamente cada 7 días, es decir, semanalmente, como se muestra en la figura 25.

Para la predicción se utilizó el modelo ARIMA (Autoregressive Integrated Moving Average Model), que es ampliamente utilizado para estos análisis por sus buenos resultados y simplicidad, se hizo una partición para entrenamiento y prueba con el 70% y 30% de los datos, para la predicción a futuro se tomó una semana del mes de enero.

En cuanto a la hiperparametrización se utilizó un ciclo *for* que fuera cambiando los parámetros del modelo, mientras se calculaba el índice AIC (Akaine Information Criteria) que cuantifica que tan bueno es el modelo y su simplicidad, un menor valor significa un mejor modelo, por lo que, con estas bases, se escogieron los parámetros y se procedió a ejecutar el modelo.

En la figura 26 se muestran los resultados de la evaluación del modelo, las conclusiones de estas métricas son que en primer lugar, los errores residuales son cercanos a cero y no parecen tener ninguna correlación, en segundo lugar, el diagrama de densidad parece tener una distribución normal con media cercana a 0, sin embargo, se observa que hay diferentes fluctuaciones debidas a la limitada cantidad de datos, en tercer lugar, los puntos deberían estar lo más cercano posible de la línea, ya que una desviación alta significaría que la serie se encuentra sesgada, finalmente la última gráfica muestra la correlación entre los errores residuales, estos no deberían repetirse periódicamente.

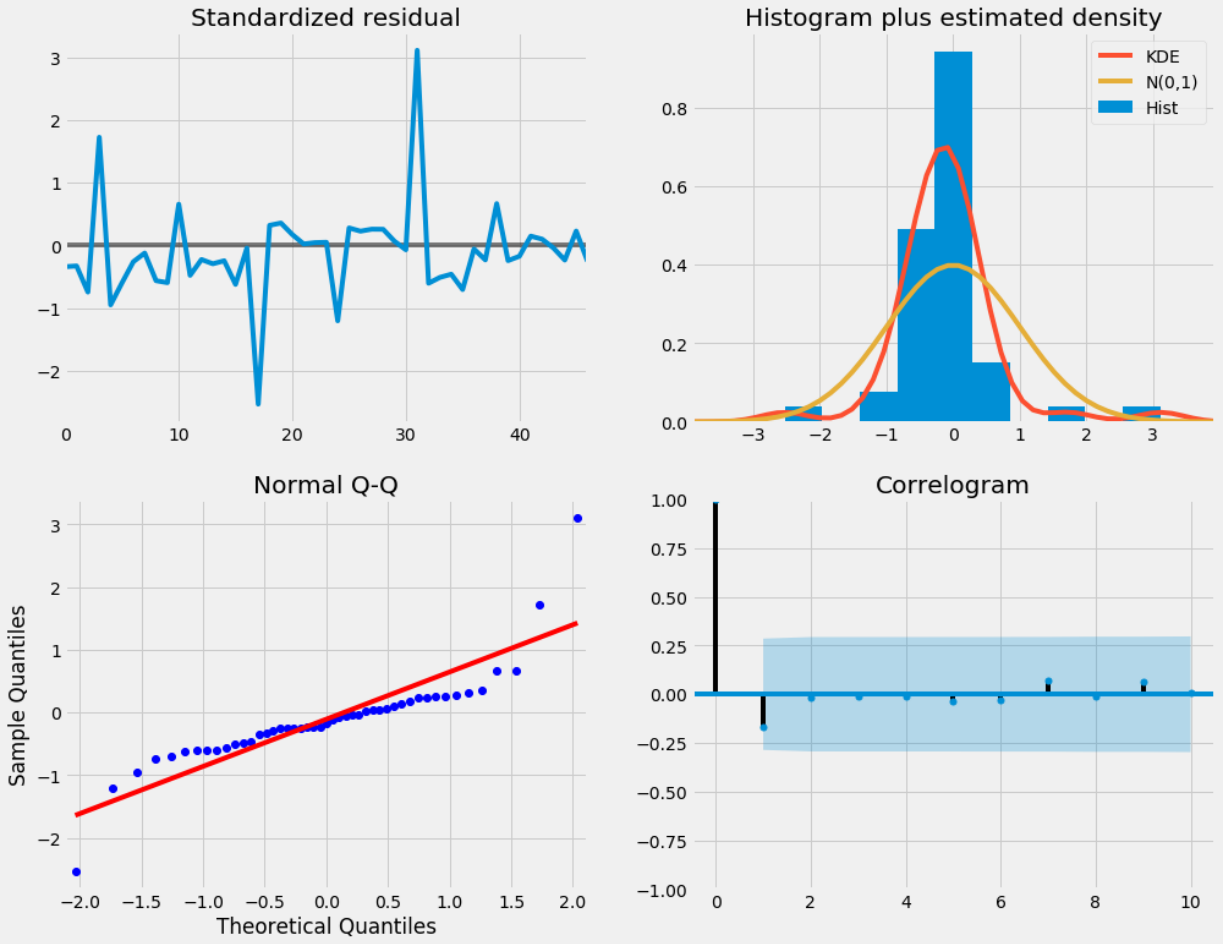

Figura 26 Diagnósticos del modelo – Fuente: Autores

Finalmente se realizó una representación gráfica (Figura 27) en donde se muestra una primera predicción de la primera semana del mes siguiente, además de la predicción realizada para los datos de prueba, aunque el costo parece no ser muy exacto se considera que con una mayor cantidad de datos este mejorará.

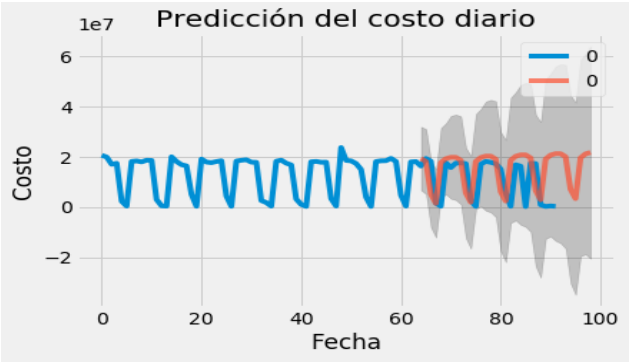

Figura 27 Predicción del costo – Fuente: Autores

El código detallado de la implementación de este modelo se encuentra en el anexo 6, junto con una breve descripción, ya que el análisis más relevante se encuentra en este documento.

6.5 Propuesta de Prototipos de Visualización

En las Figuras 28 a 32 se muestran diferentes graficas propuestas para la visualización de la información que exponen el uso de las aplicaciones en el tiempo, según el tipo de escritorio desde el que se usan, la cantidad de conexiones mensuales y la duración de estas, la relación de costos por usuario, perfil y aplicación y las conexiones por sede alrededor del país. El objetivo de estos prototipos es lograr el acercamiento al cumplimiento de las necesidades de los usuarios y la respuesta a las preguntas planteadas. En la figura 28 se propone la visualización de la utilización de las aplicaciones en los meses analizados.

Figura 28 Prototipo visualización: uso de las aplicaciones en el tiempo – Fuente: Autores

En la figura 29 se observa el uso de los escritorios a través del tiempo, se encuentra que el tipo de escritorio más utilizado es el liviano, seguido por el mediano y el pesado, teniendo los 3 una mayor utilización en el mes de octubre. En cuanto a los escritorios NOC, OBSERVEIT y VIP se observa que su uso fue muy similar durante los 3 meses analizados.

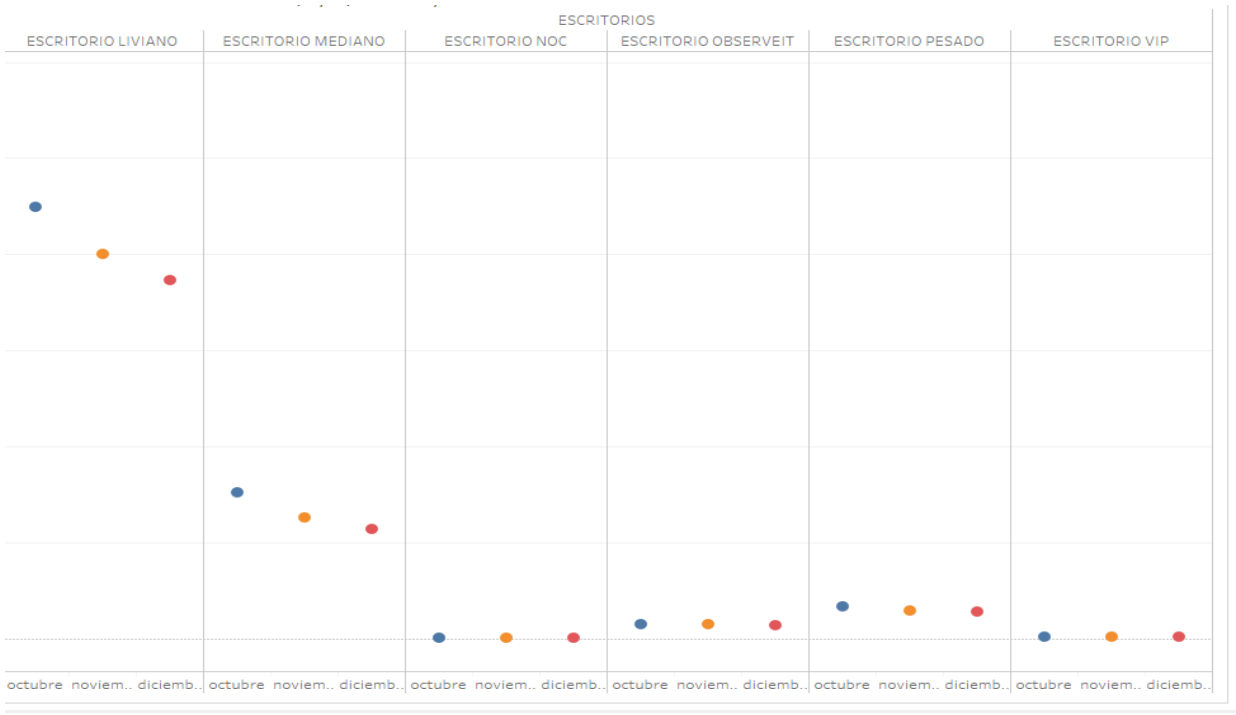

Figura 29 Prototipo visualización: uso de los escritorios en el tiempo – Fuente: Autores

En la figura 30 se muestra la cantidad de conexiones mensuales por aplicación, en este caso de IBM Consulta Pagos, se observa que en noviembre hubo un menor uso de la aplicación con respecto a los meses analizados.

Figura 30 Prototipo visualización: cantidad de conexiones mensuales por aplicación – Fuente: Autores

Se puede observar en la figura 31 la cantidad de conexiones por servidor, los servidores más utilizados son los IBMCXXAA003 y IBMCXXAA004, con un promedio de conexiones de 4400 a 4500, se evidencia igualmente que el uso de estos 2 servidores es mucho mayor en comparación con los demás, por lo que podría ser posible nivelar cargas y mejorar el rendimiento en general.

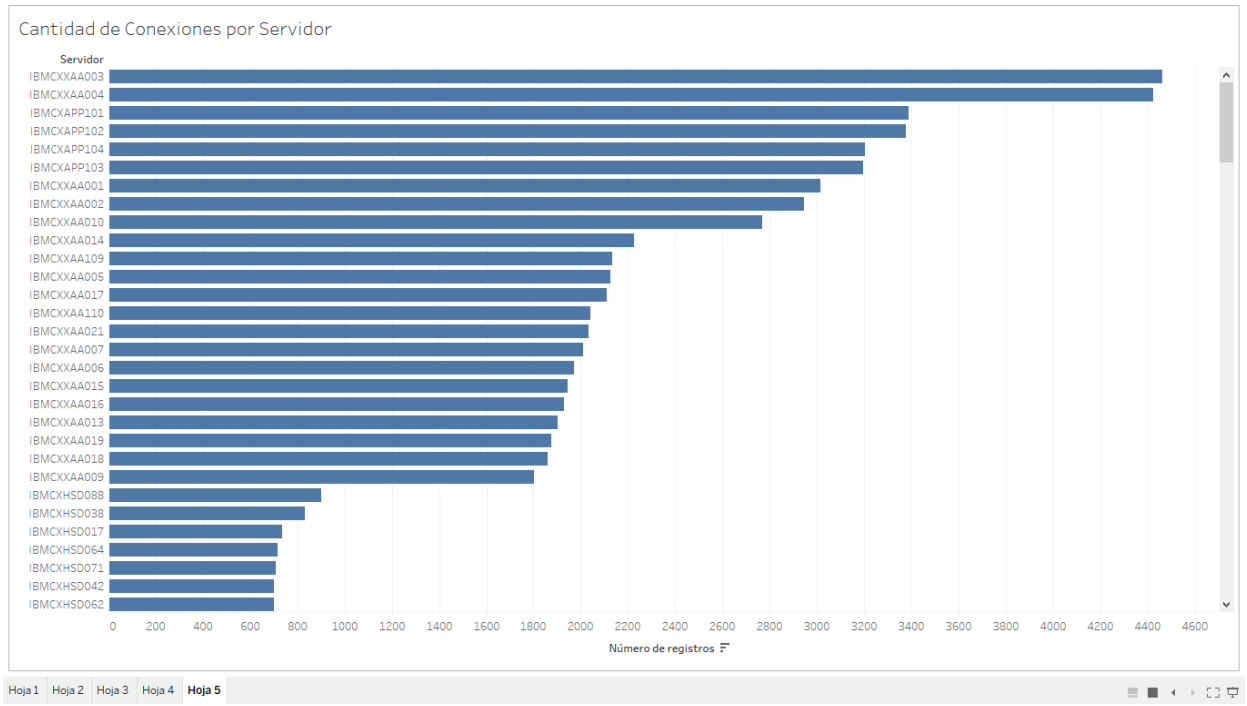

Figura 31 Prototipo visualización: cantidad de conexiones por servidor – Fuente: Autores

En la figura 32 se expone la distribución de la cantidad de conexiones en un momento del tiempo provenientes del departamento de Antioquia, la etiqueta muestra el municipio y el número de conexiones de este.

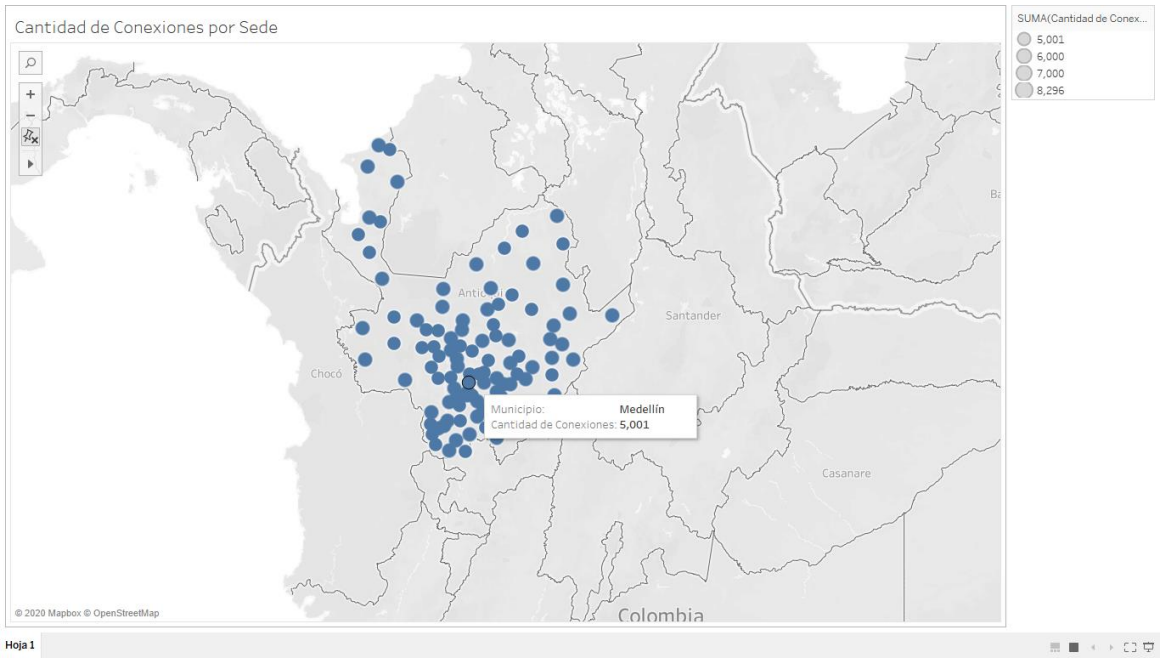

Figura 32 Prototipo visualización: cantidad de conexiones por sede – Fuente: Autores

Para la visualización de los análisis realizados se dispuso del uso de la herramienta Microsoft Power BI ya que por su flexibilidad y facilidad permite el análisis de la información desde las diferentes dimensiones de forma más fluida.

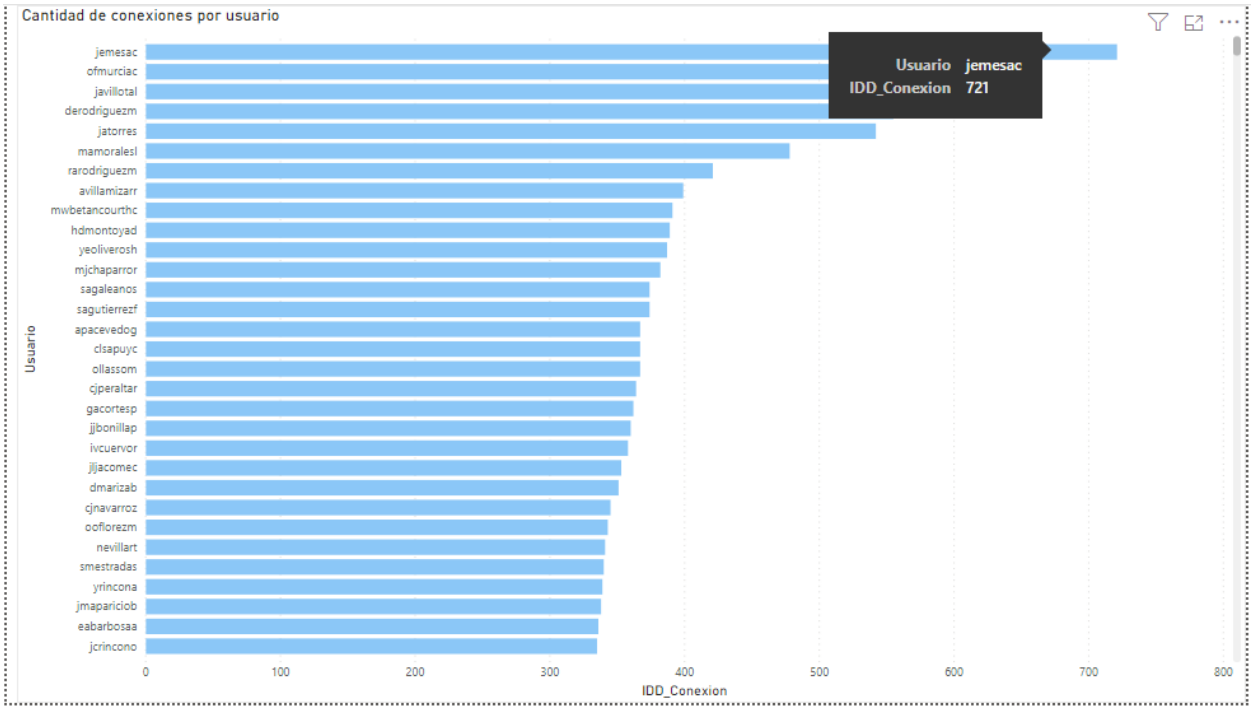

Figura 33 Cantidad de conexiones por usuario – Fuente: Autores

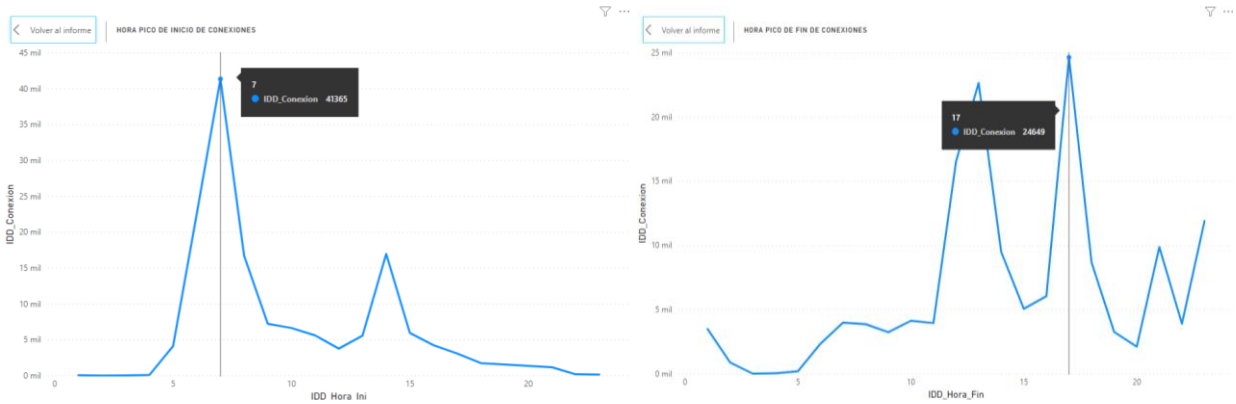

Figura 34 Hora pico de conexión y desconexión – Fuente: Autores

Igualmente, la herramienta de visualización permite la generación de tableros de control que la información sintetizada, dinámica, concluyente y fácil de manipular por parte de los actores interesados.

La figura 35 ilustra un tablero de control propuesto que permite conocer por sede y en un periodo de tiempo deseado, la cantidad de conexiones, la duración y el valor de cada una de acuerdo con la aplicación ejecutada.

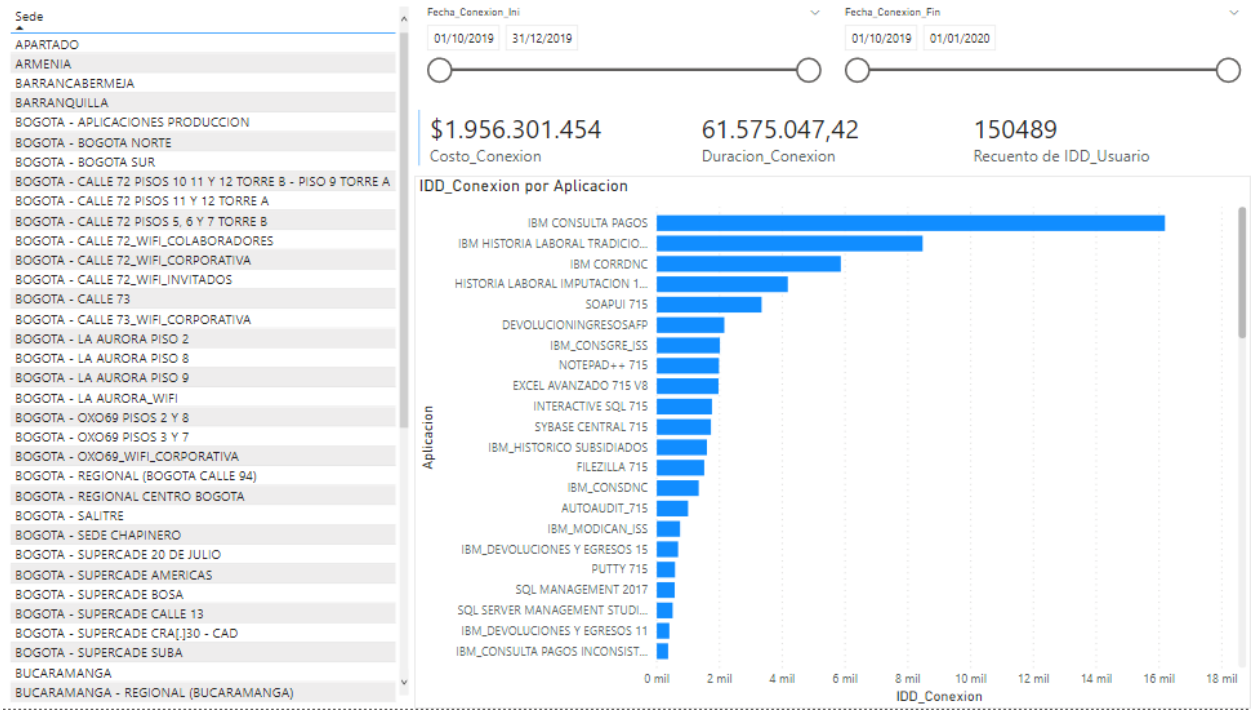

Figura 35 Dashboard Conexiones por sede – Fuente: Autores

6.6 Pruebas e Implementación de Dashboard

Como resultado final del ejercicio y una vez realizada la socialización y evaluación de los prototipos diseñados con los usuarios interesados, se estableció que la visualización más adecuada para responder a las preguntas del negocio era un tablero de control que mostrara la gestión de la infraestructura, otro las aplicaciones y un tercer panel con información concerniente a los usuarios para análisis descriptivos. Igualmente, se definió un *dashboard* que muestre el análisis de *clustering* realizado y otro que permita predecir el costo de la operación a través de una serie de tiempos. Una vez recogida la información se determinó que por su facilidad de uso e interactividad los *dashboard* se implementarían en la herramienta Power BI.

A continuación, se muestran los tableros de control finales de acuerdo con lo validado con los usuarios:

6.6.1 Dashboard descriptivo de infraestructura

En la figura 36 se observa el *dashboard* diseñado para la gestión de la infraestructura dispuesta para la virtualización de escritorios y aplicaciones. El objetivo de este tablero es mostrar, controlar y tomar decisiones concernientes al comportamiento de los servidores de la granja de virtualización y su comportamiento transaccional. De acuerdo con los requerimientos planteados, este tablero permite conocer, en un periodo de tiempo específico, la cantidad de usuarios conectados por cada servidor, el promedio de la duración de las conexiones y la distribución de estas por cada plataforma virtualizada:

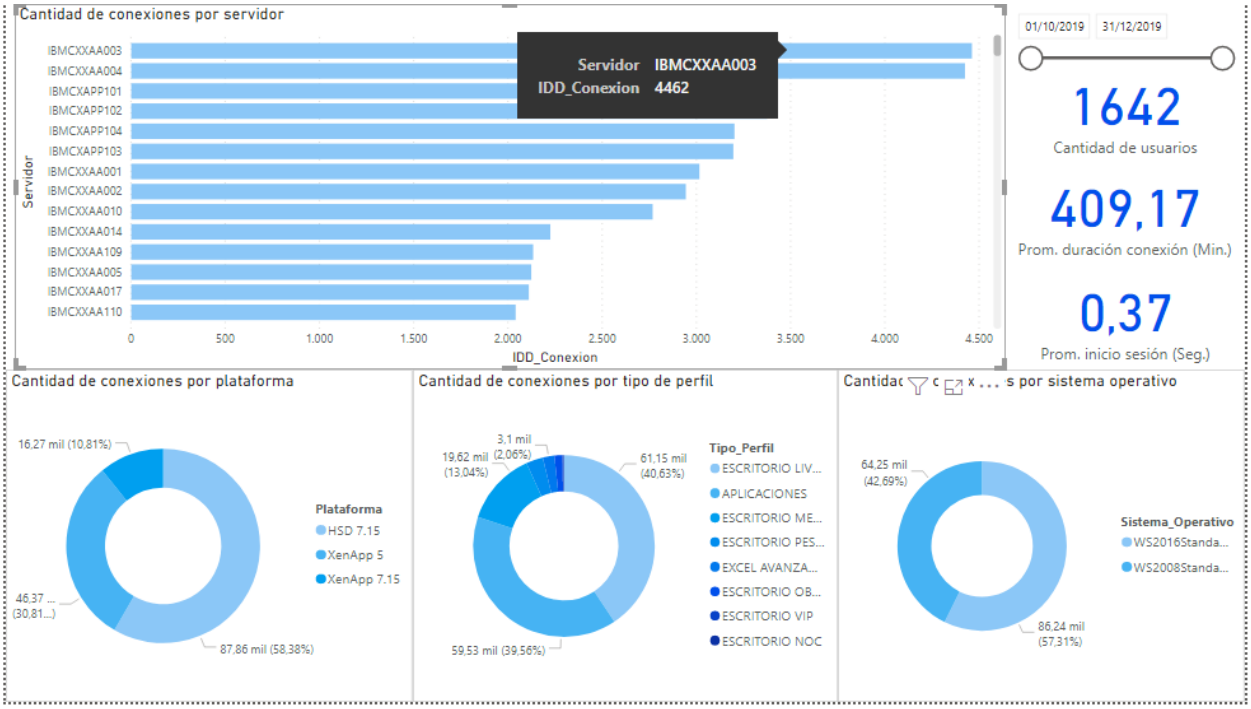

Figura 36 Dashboard infraestructura - Fuente: Autores

Se puede observar, por ejemplo, que los servidores IBMCXXAA003 e IBMCXXAA004 son los que mayor carga transaccional tuvieron durante los 90 días de data recolectada con un total de 4.462 y 4.460 conexiones respectivamente entre el 01-oct-2019 y 31-dic-2019. Se observa también que el tiempo promedio diario de las conexiones de los usuarios a los servidores fue de 409.17 minutos, es decir, de 6.81 horas de labor y un tiempo promedio de inicio de sesión de 0.37 segundos, considerado optimo a pesar de la gran carga transaccional de estos 2 servidores. Con base en este análisis se sugiere a la Dirección de Infraestructura Tecnológica realizar un chequeo del balanceo de las cargas de los servidores con el ánimo de optimizar los recursos disponibles.

6.6.2 Dashboard descriptivo de usuarios

En este tablero (figura 37) y de acuerdo con lo solicitado, se pueden analizar los hábitos, comportamientos y tendencias de conexiones por parte de los usuarios. Se conocen las sedes desde donde se conectan, las aplicaciones que usan, duración total de las conexiones y promedio diario. También se muestran las fechas y horas de inicio y fin de las conexiones en un determinado periodo de tiempo y la ubicación organizacional de los mismos. Se puede realizar búsqueda del usuario en la caja superior izquierda.

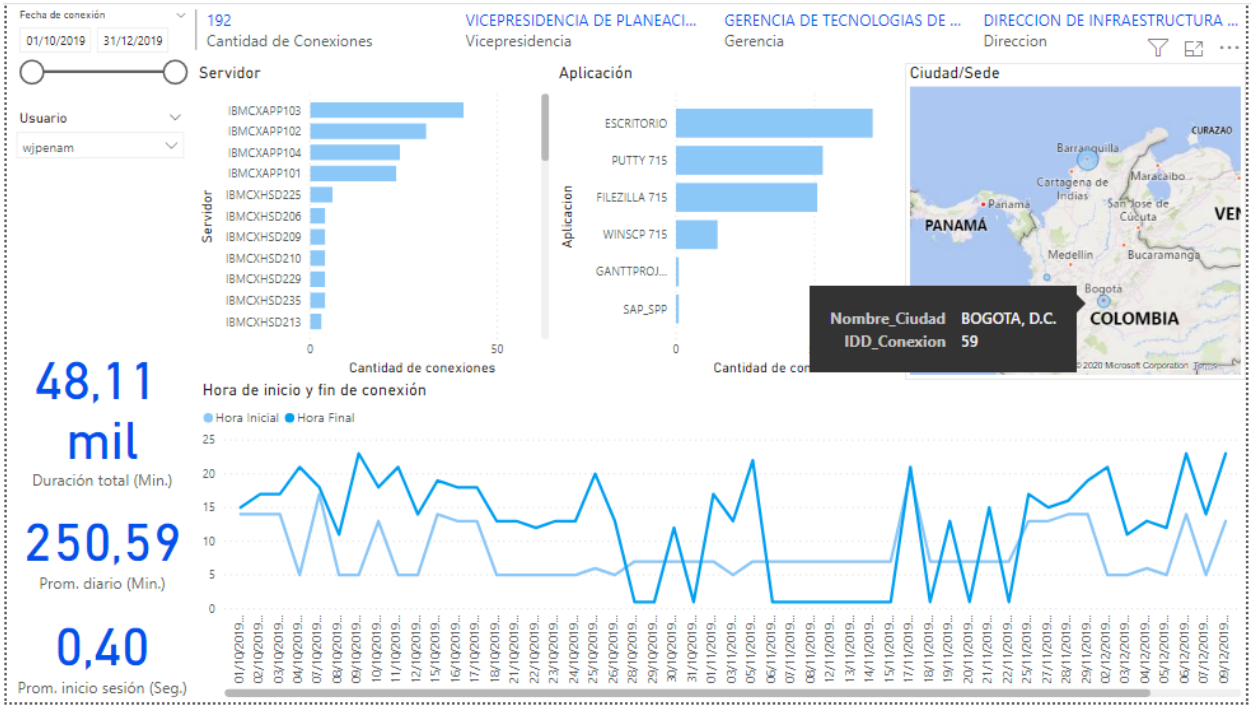

Figura 37 Dashboard usuarios - Fuente: Autores

Se observa que el usuario "*wjpenam"* realizó 192 conexiones entre el 01-oct-2019 y 31-dic-2019, 59 de las cuales se realizaron desde la ciudad de Bogotá y las demás distribuidas entre Barranquilla y Quibdó. Se observa también que la mayor distribución de sus conexiones fue hacia el escritorio virtual y la aplicación PUTTY 715 y la menos usada fue SAP_SPP con una sola conexión. Se observan algunos *outliers* en las horas de conexión y desconexión del usuario ya que la hora de inicio es mayor a la de terminación. Probablemente se deba a inicios de sesión en jornadas nocturnas. El tiempo total de las conexiones del usuario fue de 48.110 minutos para un promedio diario de 250.59 minutos y un promedio de 0.40 segundos de inicio de sesión. Se sugiere a los administradores de la plataforma revisar los tiempos en los días indicados en la visualización.

6.6.3 Dashboard descriptivo de aplicaciones

El siguiente *dashboard* muestra el comportamiento de las aplicaciones. Permite analizar el uso de las aplicaciones y los escritorios virtuales. Analizar los hábitos, comportamientos y tendencias de uso de cada aplicación del catálogo por parte de los usuarios. Permite también conocer las aplicaciones que tienen mayor frecuencia de uso, mayor cantidad de usuarios, mayor tiempo de utilización, origen de las conexiones, tiempos de permanencia de los usuarios en las aplicaciones, tiempos de respuesta de las aplicaciones para el inicio de las sesiones, durante un tiempo determinado por el usuario. Da respuesta también a conocer el uso de las aplicaciones por cada vicepresidencia, gerencia o dirección y calcular el costo generado por estas. La figura 38 muestra el tablero indicado:

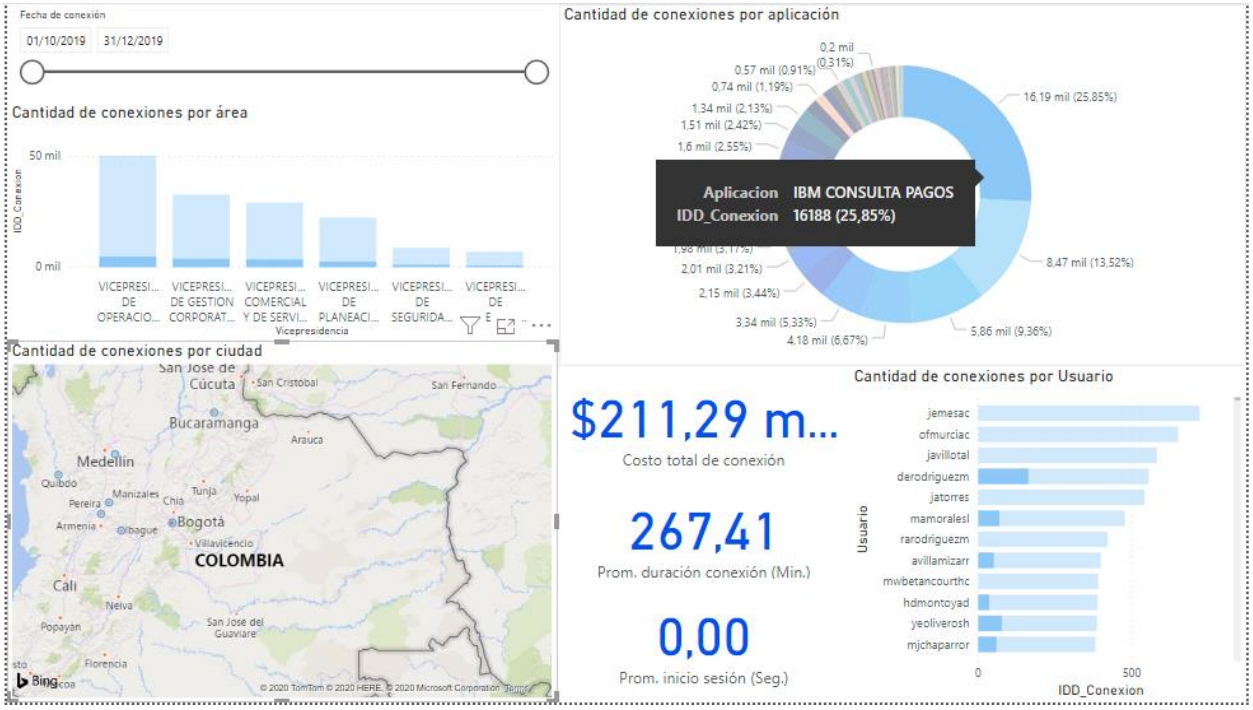

Figura 38 Dashboard aplicaciones - Fuente: Autores

Se observa que la aplicación IBM CONSULTA PAGOS es la más usada en la VICEPRESIDENCIA DE OPERACIONES generando un costo de utilización de $211.29 millones y que su uso se distribuye entre varias sedes de la organización como Bogotá, Medellín, Bucaramanga y otras. El uso de esta aplicación constituye el 25.85% del total con 16.188 conexiones realizadas por los usuarios entre el 01-oct-2019 y 31-dic-2019. Este indicador sugiere oportunidades de mejora con la revisión de los usuarios que acceden a esta aplicación.

6.6.4 Dashboard del modelo de clustering

En cuanto al análisis realizado, se diseñaron dos tableros que explican de manera gráfica y entendible al usuario los modelos creados.

En el primero de ellos (Figura 41) se muestra el *clustering*, en el que por medio del filtro ubicado en la parte superior izquierda es posible distinguir todas las variables que lo componen, en este caso específico se observa el clúster de validación de derechos de la vicepresidencia de régimen de prima media, la gerencia más activa es la de financiamiento e inversiones, usando las aplicaciones de IBM consulta pagos e IBM historia laboral tradicional principalmente, en los primeros y últimos días del mes, en las horas de la mañana y con una duración promedio de uso de 300 a 490 minutos. Es posible aplicar distintos filtros de los datos que muestran información valiosa a los usuarios.

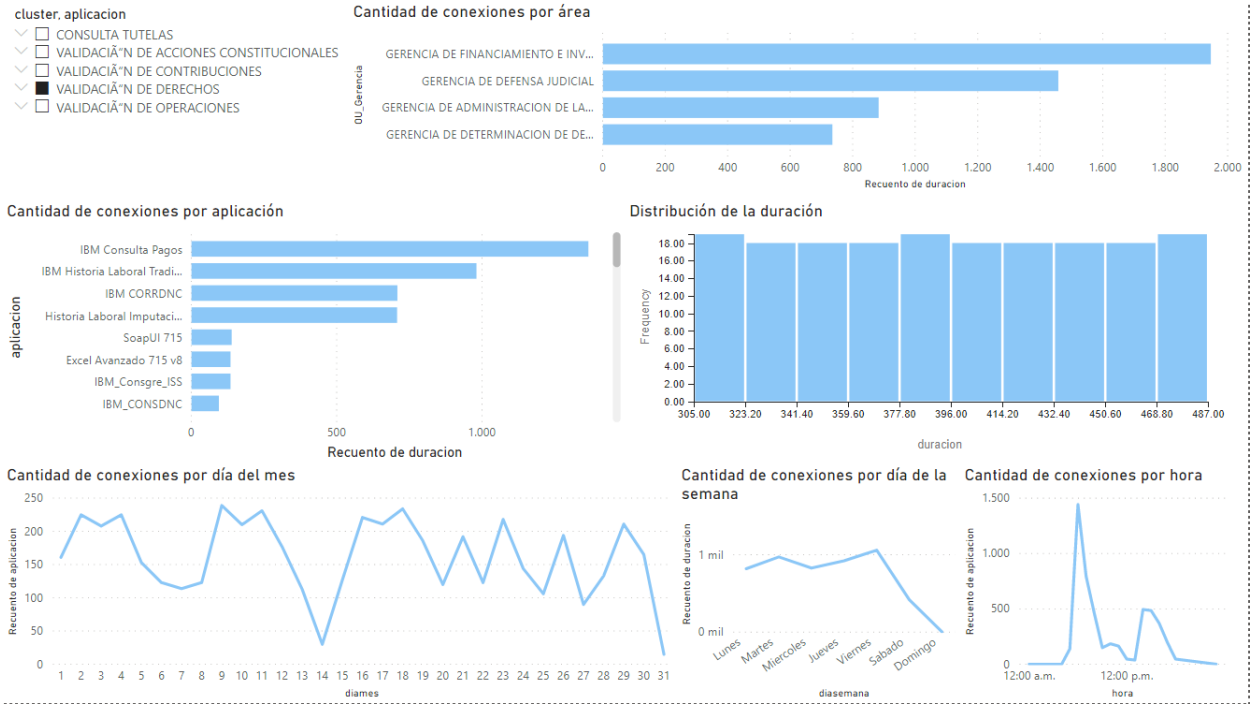

Figura 39 Dashboard clustering - Fuente: Autores

6.6.5 Dashboard predictivo de serie de tiempo

Para la visualización de la serie de tiempo que predice el costo diario de la operación de los inicios de sesión, se creó el panel de la figura 42, en el que se muestra el costo total de los 3 meses analizados, con la posibilidad de filtrar en una ventana más corta de tiempo, también es posible conocer el costo discriminado por los perfiles de operación y finalmente se visualiza la serie de tiempo, en la que la línea más larga muestra los valores reales, y la corta muestra la predicción del modelo para finales de diciembre, permitiendo ver la comparación de los valores reales versus lo que predijo el modelo, y principios de enero cuyos valores son desconocidos en el set de datos trabajado.

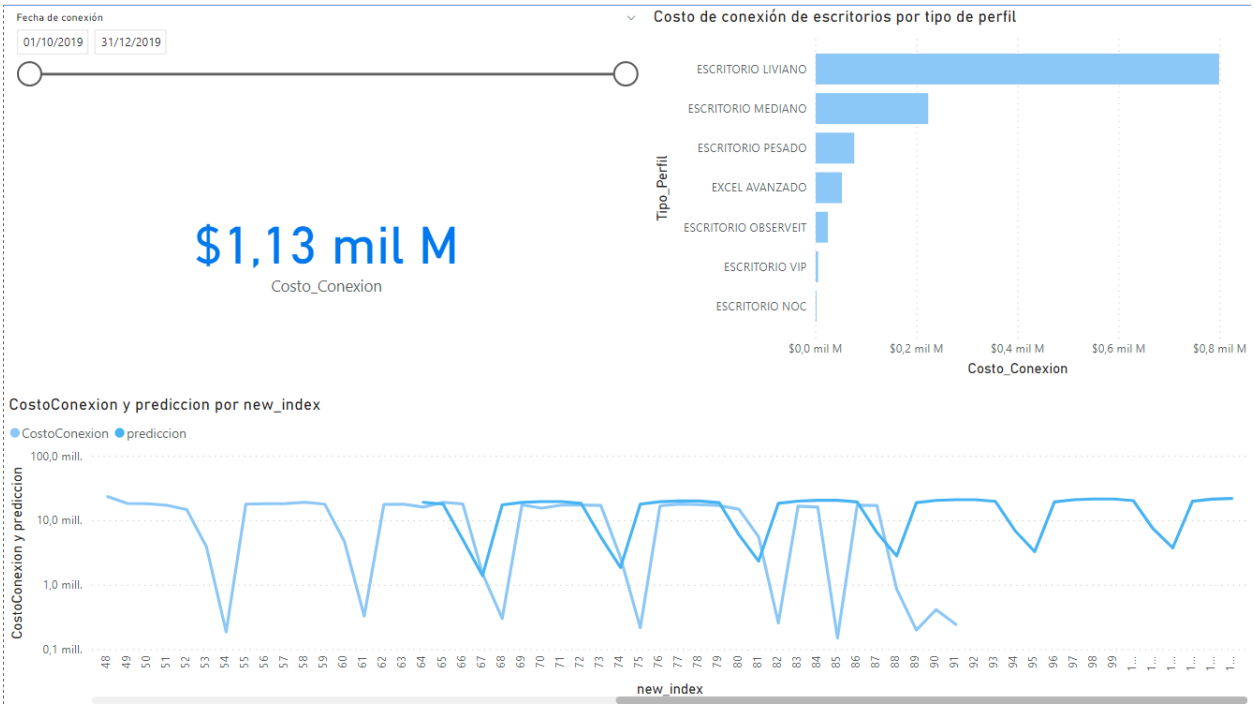

Figura 40 Dashboard serie de tiempo - Fuente: Autores

En el siguiente código QR se pueden consultar las visualizaciones interactivas desarrolladas como complemento del proyecto.

Power BI

7. CONCLUSIONES

En la etapa de entendimiento del negocio se logró comprender la importancia de la administración de los recursos y cómo repercuten en la obtención de los logros estratégicos de la entidad. Una asignación equivocada de recursos de infraestructura, basada en instintos y no en datos reales, puede afectar considerablemente la ejecución de procesos y los tiempos de respuesta de estos, ocasionando incluso el incremento de costos operativos, reducción de utilidades y aumento de riesgos operacionales.

Es necesario contar con una óptima calidad de datos a partir de campos controlados en sistemas transaccionales, ya que permiten tener un mayor entendimiento y certeza de la operación y facilitan la preparación del entorno para para la definición de los modelos descriptivos y predictivos que se puedan desarrollar a partir de la información registrada. A través del conocimiento del negocio y de sus datos, se logró identificar que el área de TI es consciente de la importancia de la gestión y administración de los recursos para el cumplimiento de los logros estratégicos de la organización. Por lo tanto, los modelos aquí propuestos responden a las necesidades transversales del negocio promoviendo e impulsando una transformación digital basada en tecnologías de punta que sean capaces de brindar la confiabilidad, integridad y disponibilidad necesaria para satisfacer las necesidades de todos los clientes internos. Dar a los usuarios interesados en esta gestión la posibilidad de contar con herramientas ágiles, seguras y con total disponibilidad basada en la Inteligencia de Negocios y la Ciencia de Datos, facilita la toma de decisiones a partir de la administración de la información.

La implementación de bodegas de datos a partir de un modelamiento multidimensional y la proyección de modelos predictivos junto con el aprovechamiento de todos sus beneficios, se convierten en ventajas competitivas para las organizaciones que buscan producir al máximo a partir del potencial de sus datos. Colpensiones ha heredado la información que durante años poseía el antiguo Instituto de Seguros Sociales y sumada a su propia gestión, ha instaurado una óptima administración de su información, con políticas, gobierno de datos y con herramientas que permitan, no solo almacenar los datos sino sacarles provecho para obtener mejores resultados operativos y financieros, que finalmente se ven reflejados en la satisfacción y el bienestar de los ciudadanos.

A partir del análisis descriptivo y predictivo realizado en este proyecto, se propone la generación de valor desde los distintos puntos de vista involucrados. La capacidad analítica desarrollada se puso a consideración y evaluación de los interesados y permitió cuestionar los datos generados a partir de los hechos permitiendo que la información compilada diera las respuestas a las preguntas del negocio para entender lo que está sucediendo, por qué, y lo que puede suceder permitiendo tomas las decisiones pertinentes.

Con la elaboración e implementación de los *dashboards* no solamente se logró generar una representación visual de los datos en un punto del tiempo. La implementación de esta herramienta incide en la consecución de los objetivos estratégicos de la entidad ya que permite visualizar el problema y facilita la toma de decisiones mitigando posibles errores que se puedan presentar ya que su objetivo es transformar los datos en información gráfica útil para orientar las decisiones estratégicas hacia la obtención de los objetivos de la organización.

En el proceso de formación de esta Maestría y extendiendo las expectativas de aprendizaje a partir de la investigación formativa, se logró asimilar y aplicar principios, conceptos, metodologías y herramientas para la implementación de sistemas de gestión basados en Inteligencia de Negocios y Ciencia de Datos que permitieron integrar modelos de descripción y predicción, es decir, permitieron unir ambos conceptos usando herramientas de bases de datos y lenguajes de programación, reflejando los resultados en diferentes entornos de visualización y aportando valor a la organización para la toma acertada de decisiones. Estos resultados se enfocaron en el análisis de los hechos ocurridos y a partir de esto, generar la prospectiva de su comportamiento permitiendo un análisis exploratorio basado en modelos estadísticos y probabilísticos.

El mundo está cambiando y no es posible ser ajenos a ese cambio. Se debe estar atentos a lo que el entorno dice. Es importante transformarse en agentes con visión y con capacidad de innovación, con capacidad de trasmitir lo que está pasando, pero también con capacidad analítica para tomar decisiones necesarias para el progreso de las empresas del país basado en las fortalezas que se pueden obtener a partir de un acertado conocimiento de los datos y el aprovechamiento pleno de la información.

8. TRABAJOS FUTUROS

A partir de los análisis realizados y los resultados obtenidos se pueden desarrollar actividades que no fueron del alcance de este proyecto, e incluso proyectos completos y que se ven como una oportunidad de mejora continua teniendo en cuenta la aplicación de la inteligencia de negocios, de la ciencia de datos, *big data*, la minería de datos y la inteligencia artificial entre otras.

En lo concerniente a la administración de la infraestructura tecnológica que soporta la prestación de los servicios virtualizados y que se evidenció en el análisis de la carga transaccional de los servidores, se tiene la oportunidad de aplicar técnicas predictivas que determinen los picos de conexiones y el performance de los servidores. A través de modelos analíticos se puede realizar un balance efectivo de dichas cargas que permita a los administradores de la plataforma realizar los ajustes correspondientes para optimizar el uso de los recursos disponibles, la adecuación de los ya existentes o la decomisión de los que no son usados.

Un estudio de costos basado en el análisis del uso de las aplicaciones permite realizar presupuestos que se ajusten a la realidad de la organización ya que se puede determinar qué se está usando y de qué manera por parte de los usuarios y qué aplicaciones no son necesarias. De esta forma los planes presupuestales serán más acertados.

El análisis de conexiones permite realizar estudios sobre la carga laboral de los funcionarios para determinar planes de acción enfocados al bienestar laboral y la asignación de recursos humanos a las distintas áreas de la Entidad.

Analizar el comportamiento de los usuarios puede ser determinante para encontrar posibles vulnerabilidades de seguridad informática ya que al realizar estudios que determinen, por ejemplo, el origen de las conexiones o conexiones en fechas u horas fuera de lo común, permiten cerrar brechas de seguridad y posibles fugas de información que afecten las operaciones de la entidad.

Los procesos de las áreas pueden ser analizados de forma descriptiva y prescriptiva identificando los picos, frecuencias y tendencias a través de líneas de tiempo que permitan determinar un conjunto de condiciones óptimas para su ejecución.

9. REFERENCIAS

Aldossary, M., Djemame, K., & Alzamil, I. (2019). Energy-aware cost prediction and pricing of virtual machines in cloud computing environments. *Future Generation Computer Systems*, *93*, 442–459. https://doi.org/10.1016/j.future.2018.10.027

Arroba, P., Zapater, M., & Ayala, J. L. (2014). Hacia la conciencia social del consumo energético en el centro de datos. *Novática*, *227*(enero-marzo), 45–50.

Barber, K., Friedlander, J., Hagan, R., & Kaminsky, D. L. (2012). *United States Patent: VISUALIZATION AND CONSOLIDATION OF VIRTUAL MACHINES IN A VIRTUALIZED DATA CENTER*.

Calvo, D. (2018). Función de activación – Redes neuronales. Retrieved from https://www.diegocalvo.es/funcion-de-activacion-redes-neuronales/

Chaudhry, M. T., Chong, C. Y., Ling, T. C., Rasheed, S., & Kim, J. (2016). Thermal prediction models for virtualized data center servers by using thermal-profiles. *Malaysian Journal of Computer Science*, *29*(1), 1–14. https://doi.org/10.22452/mjcs.vol29no1.1

Colpensiones. (2016). *Direccionamiento estratégico*.

Colpensiones. (2019). *Colpensiones mapa estratégico 2019-2022*.

Const. (1991). *Constitución política de Colombia*. Retrieved from http://wsp.presidencia.gov.co/Normativa/Documents/Constitucion-Politica-Colombia.pdf

Curto, J. (2010). Introducción a Business Inteligence. In *Introducción a Business Inteligence* (Editorial, pp. 17–19).

Díaz, W., Vilac, J., & Gallo, D. (n.d.). Laboratorios de computación multiplataforma aplicando tecnologías de virtualización. *Congreso.Investiga.Fca.Unam.Mx*. Retrieved from http://congreso.investiga.fca.unam.mx/docs/anteriores/xvi/docs/13B.pdf

Durán, E., & Costaguta, R. (2007). *Minería de datos para descubrir estilos de aprendizaje*. (1988).

EPB 603. (n.d.). Metodología para el desarrollo de proyectos en minería de datos CRISP-DM. *Sistemas Del Conocimiento*.

Few, S., & Edge, P. (2012). Information dashboard design. *O'Reilly*.

Funeme, J. (2019). *Situación actual de las aplicaciones*.

IBM. (n.d.). ¿Qué es la calidad de datos? Retrieved from https://www.ibm.com/co-es/analytics/data-quality

Innovation, A. (2019). Qué son las redes neuronales y sus funciones. Retrieved from https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

Jimmy Martínez. (2012). Seis pasos para el Gobierno de Datos ¿ Qué es y cómo se implementa un programa de Gobierno de Datos ? *DeveloperWorks*, *1*(Gobierno de Datos), 1–5.

Kotu, V. (2017). *Business Intelligence & Data Science*. Retrieved from https://www.youtube.com/watch?v=mRUSooe3cPM&feature=youtu.be

Logicalis. (2015). *Mayor ventaja competitiva con herramientas Big Data Analytics*. Retrieved from https://blog.es.logicalis.com/analytics/mayor-ventaja-competitiva-con-herramientas-big-data-analytics

Olivares, E. (2015). ¿Qué es la visualización de datos? Conoce todos los detalles. Retrieved from https://ernestoolivares.es/historias-visuales-visualizacion-de-datos/

PowerData. (2019). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Retrieved from https://www.powerdata.es/big-data

Prabhakaran, S. (2019). ARIMA Model – Complete Guide to Time Series Forecasting in Python. Retrieved from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Priy, S. (n.d.). Clustering in Machine Learning. Retrieved from https://www.geeksforgeeks.org/clustering-in-machine-learning/

Quisbert, V. (2012). Multiprocesador Distribuidas. *REVISTA DE INFORMACIÓN TECONOLOGÍA Y SOCIEDAD*, 100–101.

Rivadera, G. R. (2010). La metodología de Kimball para el diseño de almacenes de datos (Data

warehouses). *Cuadernos de La Facultad n. 5*, 56–71.

sas. (2019). Data Warehouse Qué es y por qué es importante. Retrieved from https://www.sas.com/es_co/insights/data-management/data-warehouse.html

Scikit-learn. (2011). Clustering. Retrieved from https://scikit-learn.org/stable/modules/clustering.html

Seif, G. (2018). The 5 Clustering Algorithms Data Scientists Need to Know. Retrieved from https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68#:~:text=Clustering is a Machine Learning,the grouping of data points.&text=In theory%2C data points that,dissimilar properties and%2For features.

SIAG, C. (2018). Virtualización de escritorios. ¿Qué es y cómo lo aplico a mi empresa? Retrieved from https://siagconsulting.es/virtualizacion-de-escritorios/

Talwar, V., Basu, S., & Kumar, R. (2011). *United States Patent: REMOTE DESKTOP PERFORMANCE MODEL FOR ASSIGNING RESOURCES*.

Tan, Y., Nguyen, H., Shen, Z., Gu, X., & Venkatramani, C. (2012). PREPARE: Predictive Performance Anomaly Prevention for Virtualized Cloud Systems. *2012 IEEE 32nd International Conference on Distributed Computing Systems*, 285–294.

Trabajo, M. *Decreto 2011 de 2012*. , Pub. L. No. 2011 (2012).

Valchanov, I. (2018). Data Science Predicting The Future. Retrieved from https://www.kdnuggets.com/2018/06/data-science-predicting-future.html

Vincent, T. (2017). A Guide to Time Series Forecasting with ARIMA in Python 3. Retrieved from https://www.digitalocean.com/community/tutorials/a-guide-to-time-series-forecasting-with-arima-in-python-3

Wang, J., Qiu, M., & Guo, B. (2014). High reliable real-time bandwidth scheduling for virtual machines with hidden Markov predicting in telehealth platform. *Future Generation Computer Systems*, *49*, 68–76.

Wood, T. (2008). Improving Data Center Resource Management, Deployment, And Availability

With Virtualization slide. *Communication of the ACM*, *51*(7), 9–10. https://doi.org/10.1145/1364782.1364786

Wood, T., Tarasuk-Levin, G., Shenoy, P., Desnoyers, P., Cecchet, E., & Corner, M. D. (2009). Memory buddies: Exploiting Page Sharing for Smart Colocation in Virtualized Data Centers. *ACM SIGOPS Operating Systems Review*, *43*(3), 27. https://doi.org/10.1145/1618525.1618529