División de Posgrado

Trabajo de Titulación

Propuesta de Arquitectura de Big Data para la Auditoria posterior de los procesos de Importación del Servicio de Aduana

Maestría de Sistemas de Información con Mención en Inteligencia de Negocios

Autor: Wilson Samaniego Vélez

Tutora: Doctora Alejandra Colina Vargas

Guayaquil - 2019

Dedicatoria

El presente trabajo se lo dedico en primer lugar a Dios por su protección y bendiciones, a mis padres por todo el amor que me han dado y los sacrificios realizados y a mi esposa e hijos que me apoyaron y lograron que este trabajo realice con éxito.

Ing. Wilson Samaniego Vélez

Agradecimientos

Agradezco a Dios por darme la sabiduría necesaria para cumplir esta meta, a mi familia por estar siempre a mi lado y apoyarme en este duro camino y en especial a mi tutora la Doctora Alejandra Colina Vargas que fue un apoyo para la realización de este proyecto.

Ing. Wilson Samaniego Vélez

Universidad Ecotec

TRABAJO DE TITULACIÓN

CERTIFICACION DE REVISION FINAL

QUE EL PRESENTE PROYECTO DE DESARROLLO TITULADO: "PROPUESTA DE ARQUITECTURA DE BIG DATA PARA LA AUDITORIA POSTERIOR DE LOS PROCESOS DE IMPORTACIÓN DEL SERVICIO DE ADUANA", FUE REVISADO, SIENDO SU CONTENIDO ORIGINAL EN SU TOTALIDAD, ASÍ COMO EL CUMPLIMIENTO DE LOS REQUERIMIENTOS QUE SE DICTAN EN LA GUÍA DE ELABORACIÓN DEL TRABAJO DE TITULACIÓN, POR LO QUE SE AUTORIZA A: ALEJANDRA MERCEDES COLINA VARGAS, QUE PROCEDA A SU PRESENTACION.

Samborondón, 24 de Julio de 2019

Alejandra Colina V.

TUTOR: ALEJANDRA MERCEDES COLINA VARGAS

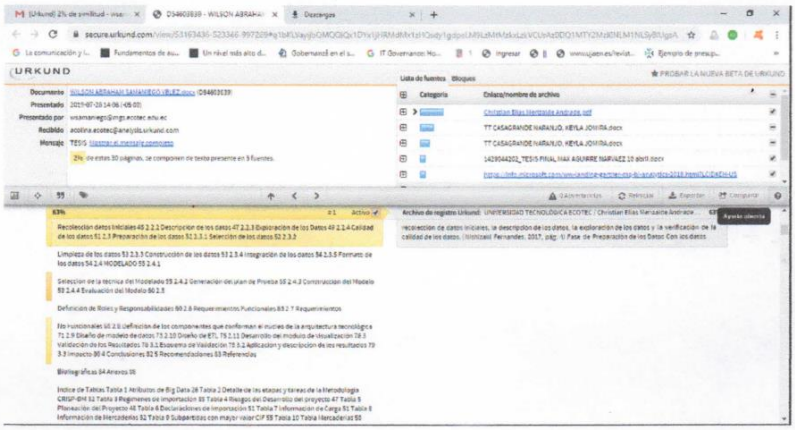
Universidad Ecotec

DIVISIÓN DE POSGRADO
PROCESO DE TITULACIÓN
CERTIFICADO DEL PORCENTAJE DE COINCIDENCIAS

Habiendo sido nombrado ALEJANDRA MERCEDES COLINA VARGAS, tutor del trabajo de titulación "PROPUESTA DE ARQUITECTURA DE BIG DATA PARA LA AUDITORIA POSTERIOR DE LOS PROCESOS DE IMPORTACIÓN DEL SERVICIO DE ADUANA" elaborado por WILSON ABRAHAM SAMANIEGO VÉLEZ, con mi respectiva supervisión como requerimiento parcial para la obtención del título de MAGISTER EN SISTEMAS DE INFORMACIÓN.

Se informa que el mismo ha resultado tener un porcentaje de coincidencias 2% mismo que se puede verificar en el siguiente link: secure.urkund.com/view/53163436- 523346-997269

Adicional se adjunta print de pantalla de dicho resultado.



Alejandra Colina V.

Doctora Alejandra Mercedes Colina Vargas
PROFESORA TUTOR DE LA CIENCIA

Resumen

En la actualidad la Aduana cuenta con gran cantidad de datos generados por la incorporación de nuevas fuentes de información y la implementación de nuevos sistemas transaccionales, trayendo consigo el crecimiento rápido de los datos, teniendo que recurrir a nuevas arquitecturas, técnicas de análisis y algoritmos para gestionar y extraer conocimiento de esa información. Esta investigación tiene como objetivo desarrollar una arquitectura de Big Data para la auditoria posterior de los tramites de importación y poder realizar la detección de posibles casos de evasión de impuestos los que pueden ser identificados implementando algoritmos de análisis de datos donde se permita determinar valores de declaración de aduanas que se encuentren por debajo de los precios (subvalorados). En la presente investigación se utilizó la metodología CRISP-DM para la implementación de una arquitectura de Big Data y la aplicación de un algoritmo para poder lograr determinar posibles casos de evasión de impuestos analizando las declaraciones aduaneras en el periodo establecido por ley. La arquitectura propuesta tiene las características de ser escalable, con la utilización herramientas de software libre, entre las que podemos mencionar Mysql como Base de Datos intermedia, Hadoop como base de datos NoSql, Sqoop se utilizó como herramienta para transferencia de datos.

Entre los principales hallazgos encontrados está la mejora en los tiempos en la auditoria posterior por medio del procesamiento en lote de los trámites de importación mediante la aplicación del algoritmo de agrupamiento de los datos y la detección de los posibles casos de evasión de impuestos bajo la modalidad de subvaloración.

Términos Claves: Subvaloración, Big data, contravenciones aduaneras.

Abstract

Nowadays customs counts with a huge number of data generated by the incorporation of new sources of information and the implementation of new transactional systems which is the reason data increases rapidly having to resort to new architectures, analysis techniques and algorithms with the goal of arranging and extracting knowledge of that information. This research has the purpose of proposing the implementation of a Big Data architecture for after the audit of import procedures and for the detection of taxes evasion possible cases which can be identified by implementing data analysis algorithms that can determine customs values declaration below the normal prices (undervalued). In this research the CRISP-DM methodology was used for the implementation of a big data architecture and the application of an algorithm to determine possible taxes evasion cases by analyzing the customs declaration in the stablished period by law. The architecture has the characteristic of being scalable with the use of free software without license tools, like Mysql as a base of intermediate database, Hadoop as a NoSql database and Sqoop was used as a tool for data transfer.

Among the main findings the principal is the improvement in time in the subsequent audit after the batch processing of import procedures through the application of a data grouping algorithm and the detection of taxes evasion possible cases under the undervalued modality.

Key Terms: Undervaluation, Big Data, customs contravention.

Índice de Contenido

1 Marco Teórico 20
1.1 Inteligencia de Negocios 20
1.1.1 Arquitectura de Inteligencia de Negocios 22
1.2 Metodología CRISP-DM 29
1.3 Servicio de Aduana 34
1.4 Marco Legal 42
2 Solución Propuesta 43
2.1 Comprensión del Negocio 43
2.1.1 Riesgos 47
2.1.2 Beneficios del Proyecto 48
2.1.3 Objetivos de la Minería de Datos 48
2.1.4 Plan de Proyecto 49
2.2 Comprensión de los Datos 50
2.2.1 Recolección datos Iniciales 50
2.2.2 Descripción de los datos 52
2.2.3 Exploración de los Datos 54
2.2.4 Calidad de los datos 56
2.3 Preparación de los datos 57
2.3.1 Selección de los datos 57
2.3.2 Limpieza de los datos 58
2.3.3 Construcción de los datos 58
2.3.4 Integración de los datos 59
2.3.5 Formato de los datos 59
2.4 MODELADO 60
2.4.1 Selección de la técnica del Modelado 60

2.4.2 Generación del plan de Prueba 60
2.4.3 Construcción del Modelo 64
2.4.4 Evaluación del Modelo 65
2.5 Definición de Roles y Responsabilidades 66
2.6 Requerimientos Funcionales 68
2.7 Requerimientos No Funcionales 73
2.8 Definición de los componentes que conforman el núcleo de la arquitectura tecnológica 77
2.9 Diseño de modelo de datos 78
2.10 Diseño de ETL 81
2.11 Desarrollo del módulo de visualización 88
3 Validación de los Resultados 91
3.1 Esquema de Validación 91
3.2 Aplicación y descripción de los resultados 92
3.3 Impacto 93
4 Conclusiones 95
5 Recomendaciones 97
Referencias Bibliográficas 98
Anexos 100

Índice de Tablas

Tabla 1 Atributos de Big Data 25
Tabla 2 Detalle de las etapas y tareas de la Metodología CRISP-DM 31
Tabla 3 Regímenes de Importación 35
Tabla 4 Riesgos del Desarrollo del proyecto 48
Tabla 5 Planeación del Proyecto 49
Tabla 6 Declaraciones de Importación 52
Tabla 7 Información de Carga 52
Tabla 8 Información de Mercaderías 53
Tabla 9 Subpartidas con mayor valor CIF 56
Tabla 10 Tabla Mercaderías 59
Tabla 11 Tabla de datos formateados 60
Tabla 12 Requerimiento Funcional 01 69
Tabla 13 Requerimiento Funcional 02 70
Tabla 14 Requerimiento Funcional 03 71
Tabla 15 Requerimiento Funcional 04 72
Tabla 16 Requerimiento Funcional 05 72
Tabla 17 Requerimiento Funcional 06 73
Tabla 18 Servidor de Base de Datos Transaccional 73
Tabla 19 Servidor Base NoSql 74
Tabla 20 Almacenamiento NAS 75
Tabla 21 Equipo de comunicación 75
Tabla 22 Requerimiento de Software 75
Tabla 23 Expertos encuestados 92
Tabla 24 Preguntas a expertos 92

Índice de Figuras

Figura 1: Importaciones en el Ecuador. Fuente: BCE 15
Figura 2 Transformación Datos – Información Conocimiento. Fuente: Business Intelligence BI & Analytics 21
Figura 3 : Arquitectura de Inteligencia de Negocios, Fuente: Business Intelligence BI & Analytics 22
Figura 4: Las 4V de Big Data. Fuente: McKinsey Global Institute Big data: The next frontier for innovation, competition, and productivity 24
Figura 5: Tipos de datos de Big Data. Fuente: IBM Developer Works..... 25
Figura 6: Arquitectura Hadoop. Fuente : www.apache.org 28
Figura 7: Metodología CRISP_DM. Fuente: IBM Manual de CRISP-DM SPSS Modeler 30
Figura 8: Etapas de la Metodología CRISP-DM 30
Figura 9: Estructura Organizacional Aduana..... 38
Figura 10: Proceso de subvaloración aduanera. Fuente: Elaboración propia 46
Figura 11: Comparación CIF por Subpartidas. Fuente: Elaboración propia 54
Figura 12: Estadística de Importaciones por Valor CIF Enero 2019. Fuente: Elaboración propia 55
Figura 13: Estadística de Importaciones por Valor CIF Febrero 2019. Fuente: Elaboración propia 55
Figura 14: Estadística de Importaciones por Valor CIF Marzo 2019. Fuente: Elaboración propia 56
Figura 15: Método Zscore..... 61
Figura 16: Método Outlier Detection Z-score : Fuente : Elaboración propia..... 62
Figura 17: Método de cálculo de distancia entre valores..... 63
Figura 18: Algoritmo K-means: Fuente Elaboración Propia 64
Figura 19: Construcción del Modelo de análisis. Fuente: Elaboración propia 65
Figura 20: Equipo del Proyecto. Fuente: Elaboración Propia 66
Figura 21: Componente de Hardware. Fuente : Elaboración Propia 77
Figura 22: Componentes de Software. Fuente: Elaboración Propia 78
Figura 23: Modelo de Datos Proyecto Aduana. Fuente: Elaboración propia 79
Figura 24: Extracción de la tabla de Mercaderías..... 83
Figura 25: Tabla de Declaraciones 84

Figura 26: Tabla de Caga 85
Figura 27: Tabla de Mercaderías Mysql 86
Figura 28: Tabla de Declaraciones Mysql 87
Figura 29: Tabla de Carga Mysql 88
Figura 30: Detección de casos de evasión de Impuestos por Subvaloración 89
Figura 31: Detección de Valores Atípicos 90
Figura 32 Tabulación Juicio de Expertos 93

PROPUESTA DE ARQUITECTURA DE BIG DATA PARA LAS AUDITORIAS DE LOS TRAMITES DE IMPORTACIÓN EN EL SERVICIO DE ADUANA

INTRODUCCIÓN

En las organizaciones existen grandes volúmenes de datos que se van acumulando en el transcurso del tiempo complejizando el análisis de la información al utilizar herramientas y bases de datos tradicionales, por tal motivo es preciso contar con tecnologías actualizadas para poder gestionar enormes cantidades de información. (Business Software, 2013)

Es por ello que, las empresas deben incorporar nuevas técnicas, herramientas y tecnologías informáticas para poder gestionar grandes volúmenes de datos que generan las operaciones del negocio.

Las instituciones y empresas están combinando los datos almacenados por las bases de datos convencionales y los datos almacenados en Big Data, sean estos comentarios y preferencias en las redes sociales, proporcionándoles nuevas oportunidades de negocios (Joyanes, 2013)

Con la introducción de Big Data las entidades están preparadas para obtener y aprovechar los datos más importantes y poder utilizarlos en la toma de decisiones del negocio. Aquellas que se encuentren actualizadas con la implementación de las nuevas tecnologías podrán ubicarse con una ventaja competitiva, mejorando en eficiencia y disminución de gastos operativos, optimizando la toma de decisiones, aumentando la fidelidad y vinculación con sus clientes y una mejor gestión con sus proveedores. (Pazmiño & Lopez, 2017)

En las empresas toda información generada por sistemas propios o por fuentes externas son utilizadas como base para la creación de estrategias y para la toma de decisiones convenientes a la empresa (Trujillo & Rojas, 2015).

En otro orden de idea, una de las principales actividades económicas de los seres humanos a lo largo del tiempo ha sido el comercio. Cuando se produce un intercambio de bienes y servicios entre un país con otros países se conoce como comercio exterior. Estos bienes de intercambio pueden ser dan de diferentes tipos siendo productos elaborados, materias primas o mercaderías varias. El comercio se da por el intercambio de mercaderías por dinero, donde el productor entrega la mercadería y el importador debe pagar el dinero acordado por la transacción (Ballesteros, 2015).

El intercambio comercial beneficia a los países a exponer sus productos a nuevos mercados, la venta de las mercaderías contribuye al aumento de sus ingresos y al incremento de la mano de obra directa e indirecta y en esa medida fortalece las arcas fiscales de los países. La dependencia encargada de controlar el ingreso y salida de mercaderías y la facilitación del comercio es el Servicio de Aduana (Plan estratégico aduana, 2015).

El Servicio de Aduana se encarga de la regulación y la gestión de los trámites aduaneros para la importación, exportación y tránsito de mercaderías en el territorio nacional. Así como también, del registro de grandes cantidades de transacciones que se originan por él envió de documentos electrónicos de los operadores de comercio exterior, información de los sistemas transaccionales, siendo guardada en base de datos relacionales para ser procesadas y analizadas posteriormente.

La información en el Servicio de Aduana se viene incrementando con la incorporación de nuevos sistemas y con el creciente aumento de las importaciones como también la incorporación de nuevos operadores al comercio exterior.

El detalle de los valores declarados en los trámites aduaneros de importación se muestra en la Figura 1, en ella se detallan los valores CIF que corresponde al Valor de la adquisición de la mercadería sumado el valor del seguro y del flete:

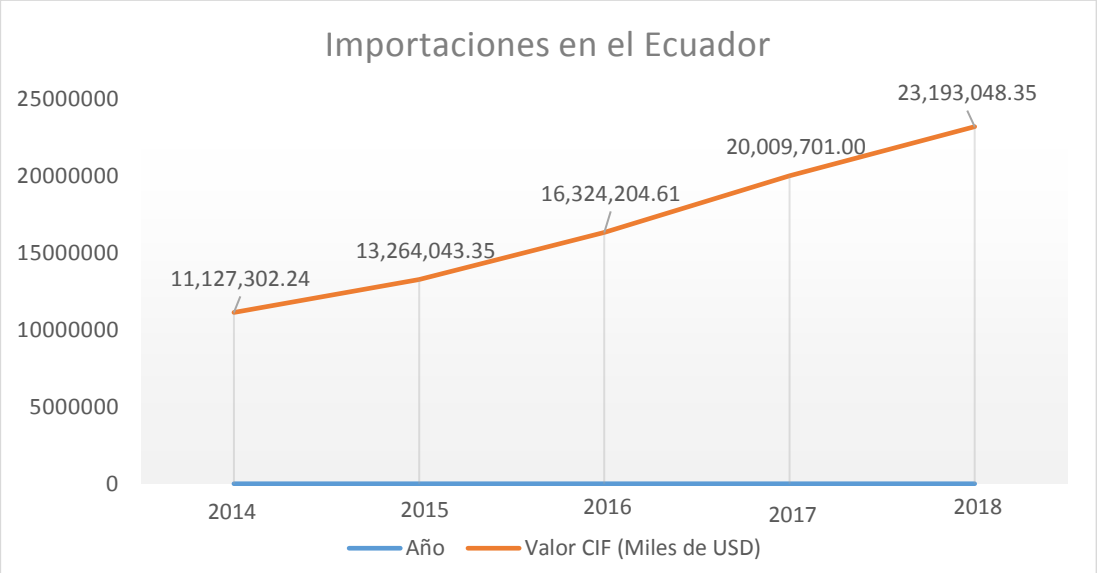


Figura 1: Importaciones en el Ecuador. Fuente: BCE

Para ejercer un mejor control sobre los trámites aduaneros de importación el Servicio de Aduana realiza el proceso de intervención en tres momentos: Control previo o anterior – Antes de la Llega de las mercaderías al territorio nacional, Control concurrente o durante el despacho de las mercaderías y Control posterior – Luego del pago y levante de las garantías.

El proceso de control posterior se lo efectúa sobre ciertos sectores económicos específicos, para esto la aduana viene implementando nuevos controles basados en el análisis de riesgo, mediante la utilización de perfiles de riesgo. Este análisis lo efectúa sobre declaraciones de importación, facturas comerciales, documentos bancarios, contables, manifiestos de carga con una vigencia a cinco años a la fecha del levante de las mercaderías.

La presente investigación se basa en una propuesta para diseñar una arquitectura de Big Data para el análisis posterior de trámites aduaneros de importación, la arquitectura propuesta se implementará con la integración de componentes de software libre y desarrollar un algoritmo para la identificación de posibles casos de evasión de impuestos bajo la modalidad de subvaloración aduanera utilizando métodos de agrupamiento para la detección de valores atípicos (outliers).

Pregunta Científica e idea a defender

¿Cómo mejorar el proceso de auditoría de trámites aduaneros de importación en el servicio de aduana bajo la modalidad de subvaloración?

Para ello, la idea a defender en esta investigación consiste en el mejoramiento de las auditorias posteriores de los trámites aduaneros de importación en el Servicio de Aduana. Esta institución maneja gran cantidad de registros provenientes de diversas fuentes y de distintos formatos razón por la cual se debe implementar nuevas tecnologías que procesen la información de manera más ágil.

Para mejorar el proceso se propone la implementación de una arquitectura de Big Data con su respectiva aplicación de algoritmo de agrupamiento (kmeans) sobre los valores declarados de las mercaderías permitiendo mejorar el proceso de auditoría de los trámites aduaneros y poder identificar posibles casos de evasión de impuestos.

Objetivos

Objetivo General

Implementar una arquitectura de Big data que permita mejorar el proceso de auditoría de trámites aduaneros de importación optimizando la identificación de

posibles casos de evasión de impuestos bajo la modalidad de subvaloración en el servicio de aduana.

Objetivos Específicos

- Identificar referentes teóricos relacionados con las tecnologías y herramientas necesarias para el desarrollo de una arquitectura de Big data.
- Analizar el proceso actual de auditoria de los trámites aduaneros para la identificación de posibles casos de evasión de impuestos por la modalidad de subvaloración.
- Definir los componentes que integran la arquitectura de big data a desarrollar en el servicio de aduana.
- Visualizar mediante una herramienta de inteligencia de negocios la identificación de los posibles casos de evasión de impuesto bajo la modalidad de subvaloración aduanera.

Justificación

La recaudación aduanera por concepto de tributos de importación es un rubro importante que debe ser controlado y auditado, esta representa un aporte importante en el desarrollo económico del país. Por tal motivo, el Servicio de Aduana debe mejoras los controles para garantizar que no existan casos de evasión de impuestos y esto provoque una disminución en las recaudaciones por concepto de tributos. Debe implementar nuevas estrategias de control, estas estrategias deben ser apoyadas por herramientas tecnologías.

Actualmente, el análisis de posibles casos de evasión de impuestos no se lo realiza a grandes volúmenes de datos, el análisis se lo realiza de manera segmentada ya que no cuenta con las herramientas informáticas que les permita realizar este tipo de

análisis. El análisis se lo realiza por medio de un perfilador de riesgo, es decir a ciertos grupos de operadores de comercio que ya han sido identificados como sujetos de control físico.

Se propone el desarrollo de una arquitectura de Big Data que permita reducir los recursos y fortalecer los procesos de auditoria para la obtención de evidencias de posibles evasiones de impuestos y poder garantizar una recaudación efectiva.

Procedimiento Metodológico

La presente investigación es una propuesta de implementación de una arquitectura de Big data que permita mejorar el proceso de auditoría de trámites aduaneros de importación optimizando la identificación de posibles casos de evasión de impuestos bajo la modalidad de subvaloración en el servicio de aduana; para resolver la problemática planteada se aplicó inteligencia de negocios implementando herramientas de Big Data y de visualización. Para el desarrollo de la propuesta se realizaron las siguientes gestiones:

- Reuniones con los técnicos de la Dirección de Gestión de Riesgos para evaluar la problemática a resolver y levantar los requerimientos para el desarrollo de la solución.
- Revisar el proceso actual y la plataforma informática que realiza el proceso de identificación de posibles casos de subvaloración aduanera.
- Diseño y construcción de un modelo de datos para realizar el proceso de minería.
- Creación del ETL para obtener los datos de las bases transaccionales y almacenarlos en la base de datos no relacional
- Diseño de los gráficos de visualización de la detección de posibles casos de evasión de impuestos a través de la herramienta Power BI.

Metodología

La metodología de investigación aplicada a la presente investigación fue de manera exploratoria y descriptiva, ya que el Servicio de Aduana actualmente no cuenta con herramientas para el manejo de base de datos no estructuradas para la detección temprana valores de las mercaderías que se desvían de su comportamiento normal, es decir casos de evasión de impuestos bajo la modalidad de subvaloración aduanera. Se tomaron varias muestras de datos para probar el modelo de minería de datos y se validaron los resultados obtenidos de las pruebas.

Para la implementación de la solución propuesta se utilizó la Metodología Cross Industry Standard Process for Data Mining (CRISP-DM), metodología utilizada en la mayoría de proyectos de Inteligencia de negocios. Cada fase de la metodología indica las tareas que se deben ejecutar garantizando una implementación efectiva del proyecto.

CAPITULO 1.

1 Marco Teórico

En este capítulo se detallaran los conceptos y teorías que se utilizan para el desarrollo de la propuesta para la implementación de una solución de Inteligencia de Negocios, la integración de componentes en una arquitectura de Big Data, el modelo aplicado en la minería de datos y el marco legal que rige el proceso de subvaloración en el Servicio de Aduana.

1.1 Inteligencia de Negocios

Para poder entender mejor la definición de inteligencia de negocios se debe tener diferenciado los conceptos de dato e información. La definición de dato de acuerdo al diccionario de la real academia de la lengua española es: "Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho", así también la definición de Información de acuerdo al diccionario de la lengua española es: "Comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada" (RAE, 2019)

La transformación y el análisis de los datos se convierten en inteligencia para la toma de decisiones en las organizaciones, esta secuencia de procesar los datos se puede visualizar mediante la Figura 2.



Figura 2 Transformación Datos – Información Conocimiento. Fuente: Business Intelligence BI & Analytics

De acuerdo a la definición de Ramos la inteligencia de negocios se conoce como el conjunto de estrategias, tecnologías y metodologías que permiten la conversión de los datos en información de calidad, para obtener conocimiento que facilite la toma de decisiones oportuna y contribuya a mejorar la competitividad (Ramos , 2016)

> La inteligencia de negocios emplea los procesos, tecnologías y herramientas para la transformación de los datos en información, esta convertirla en conocimiento y este conocimiento en la construcción de planes estratégicos que mejoren el rumbo de las organizaciones, así mismo integra el almacenamiento de datos con las herramientas de análisis y gestión de contenidos y gestión de conocimiento (Loshin, 2003)

Inteligencia de negocios logra posicionar mejor a las empresas a través de la combinación de la infraestructura tecnología, herramientas de análisis y aplicando las mejores prácticas, facilitando información relevante para la toma de las decisiones por parte de la alta gerencia (Gartner, 2013)

Marian Cristecu afirma sobre Inteligencia de Negocios es el proceso de recoger datos transaccionales y convertir los datos en inteligencia para la toma de mejores decisiones respecto al plan de la empresa. (Cristecu, 2017).

En atención a lo presentado, la inteligencia de negocios permite a las empresas dirigirse de acuerdo al plan estratégico de la empresa, determinar el cumplimiento de

los objetivos de la organización y finalmente mostrar el estado actual de la empresa por medio de indicadores de gestión.

1.1.1 Arquitectura de Inteligencia de Negocios

La arquitectura de Inteligencia de negocios está conformada por cuatro componentes, siendo estos: Origen de datos, Bodega de datos, Explotación y Visualización, mostrados en la Figura 3.

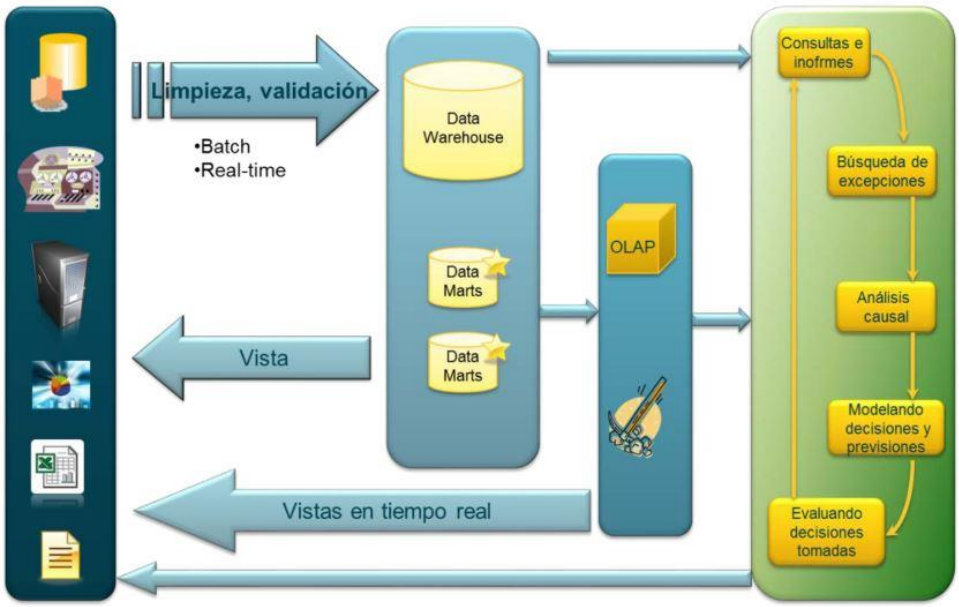


Figura 3 : Arquitectura de Inteligencia de Negocios, Fuente: Business Intelligence BI & Analytics

- **Origen de Datos**

El proceso de Inteligencia de negocios debe empezar con la selección de los datos que se utilizaran para la toma de decisiones. Los datos pueden ser escogidos de diferentes fuentes: archivos texto, archivos en Excel, redes sociales, desde base de datos relacionales Mysql, DB2, SQL Server, Oracle, archivos con formato Xml y Json. Los datos una vez extraídos deben ser transformados y cargados a un repositorio para luego ser analizado. (Oracle, s.f)

- **Etapa de Extracción**

En esta etapa se identifican los datos que deben ser extraídos, transformados desde el origen y luego almacenados en una bodega de datos, este almacén de datos puede ser en un Data Warehouse o Data Marts. En este proceso se debe también realizar la limpieza de los datos, es decir eliminar aquellos datos que no serán utilizados para los procesos de análisis. Esta definición está basada en lo indicado por (Reyes & Nuñez, 2015).

- **Etapa de Explotación**

En esta etapa los datos ya se encuentran representados en tablas y en dimensiones, se encuentran en modelos multidimensionales siendo estos los Cubos Multidimensionales OLAP o los DataMarts. Estos componentes poseen una gran capacidad de análisis de las relaciones y tiempos de respuestas óptimos.

- **Data Warehouse**

Repositorio donde se almacenan los datos luego del proceso de extracción. Transformación y limpieza, esta información ya depurada proviene de varias fuentes de información que al ser analizados servirán para la toma de decisiones en las organizaciones. Data Warehouse no es solamente una combinación de hardware y software si no también generar conocimiento estratégica para la toma de decisiones en las organizaciones y mejorar su rendimiento (Noriega, y otros, 2015).

Dentro de la implementación de la propuesta de investigación incluye la auditoria de los trámites aduaneros, este proceso se lo implementó mediante la aplicación de inteligencia de negocios. La implementación inicio con la selección de los datos que van a ser utilizados para el análisis, posteriormente se realizó la limpieza y la integración, seguido se efectuó la explotación de los datos mediante la aplicación del modelo de clusterización y finalmente se realizó el proceso de visualización.

Big Data

El concepto de Big Data, se define como integración de herramientas de hardware y software para gestionar y analizar los datos en diferentes formatos y procesarlos de manera rápida y transformarla en información. (Joyanes, 2013)

La empresa IBM realizó un estudio conjuntamente con la Said Business School entrevistando a profesionales en Informática para encontrar una mejor definición de Big Data, obteniéndose como resultados nuevos atributos como la variedad, volumen, velocidad, veracidad; y no, el almacenamiento de la información y el procesamiento de un masivo número datos. La Figura 4 muestra las cuatro dimensiones de los datos.

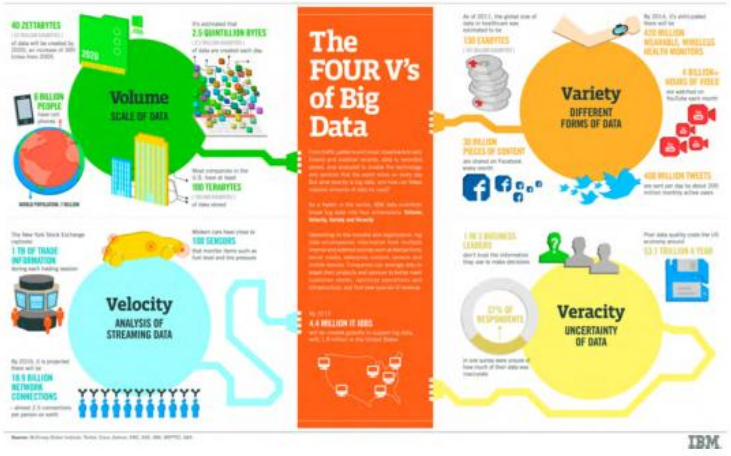


Figura 4: Las 4V de Big Data. Fuente: McKinsey Global Institute Big data: The next frontier for innovation, competition, and productivity

Big Data, se puede definir como los sistemas de información que gestionan grandes volúmenes de datos, gran capacidad para procesar en alta velocidad, veracidad y variedad para proporcionar información relevante para la toma de decisiones. (Gartner, 2013).

Los atributos de Big Data se muestran en la Tabla 1, estos son analizados por las empresas para la implementación de esta arquitectura dependiendo de los requerimientos del negocio:

Tabla 1
Atributos de Big Data

Volumen	Velocidad	Veracidad	Variedad
Registros	Tiempo cercano	Origen y Reputación	Dato No estructurado
Tablas y Archivos	Procesos	Responsabilidad	Probabilística
Transacciones	En tiempo Real	Disponibilidad	Multi-factor
Almacenamiento	Por lotes	Integridad	Estructurado

Fuente: UNIVERSITY OF AMSTERDAM. Defining the Big Data Architecture Framework [en línea]. Ámsterdam: Yuri Demchenko. Disponible en internet:http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf >

La plataforma Big Data cuenta con diferentes tipos de datos, según IBM los clasifica según la Figura 5:

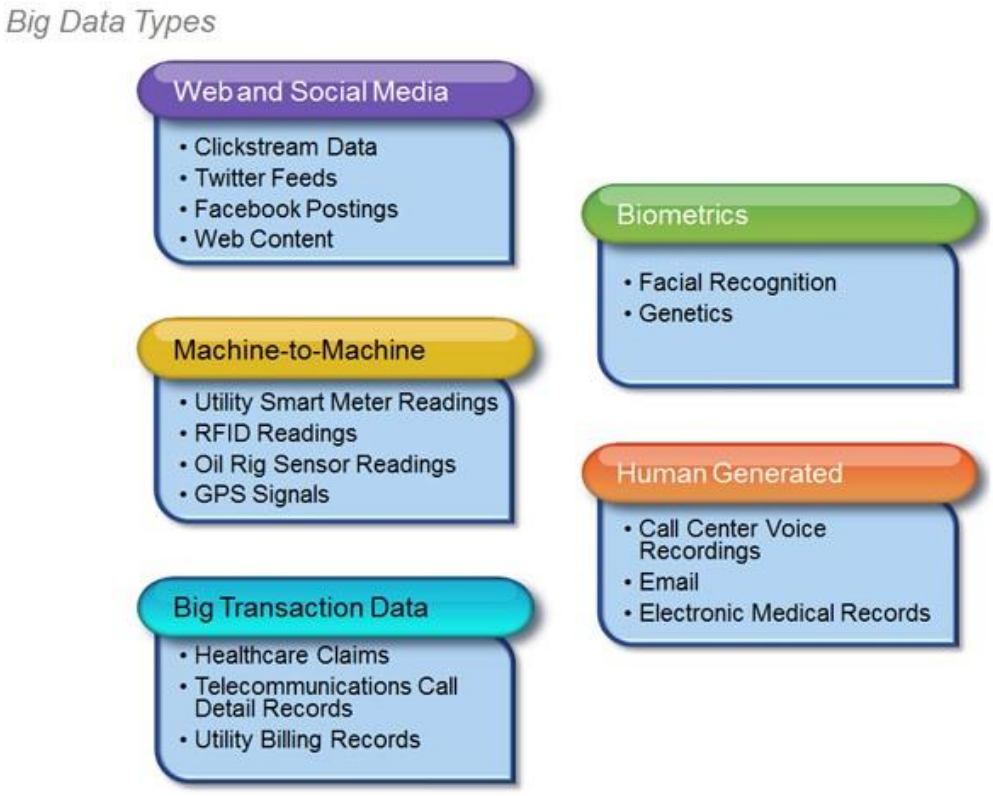


Figura 5: Tipos de datos de Big Data. Fuente: IBM Developer Works

Los datos se originan de diferentes formas, se detallan cada una de ellas:

Web and Social Media: Contenido web e información que es obtenida de las redes sociales. Ejemplos: www, Facebook, Twitter, LinkedIn, blogs

Machine to-Machine (M2M): Tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún

acontecimiento en particular. Se transmiten a través de redes alámbricas, inalámbricas o híbridas. Ejemplos: Velocidad, temperatura, presión, variables meteorológicas, variables químicas.

Big Transaction Data: Incluye datos procedentes de transacciones masivas de los centros de atención telefónica, de banca, finanzas, atención a clientes, etc. Ejemplos: Incluye registros de facturación, en telecomunicaciones los llamados registros detallados de las llamadas (Call Detail Record o CDR), etc.

Biometrics: Información biométrica. En el área de seguridad e inteligencia, los datos biométricos son sumamente importantes para los gobiernos, seguridad privada, servicios de inteligencia, policía. Ejemplos: Huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc.

Human Generated: Datos digitales generados por las personas, en sentido genérico. Ejemplos: Notas de voz, correos electrónicos, documentos electrónicos, resultados de estudios médicos, multas, etc.

- **Alcance del Big Data**

El Big Data es una nueva tecnología de análisis de información que las instituciones incorporan para mejorar el proceso de toma de decisiones, las cuales con tecnologías tradicionales no pueden procesar grandes volúmenes de datos.

La información generada por los diferentes procesos en las instituciones son almacenadas en base de datos relacionales, en estas bases la información es organizada en tablas y permiten controlar que esta no se duplique y mantenga su integridad referencial. Estas validaciones hacen que este tipo de tecnología se vuelva lenta al procesar grandes cantidades de datos y solo sean utilizadas para la lectura de datos estructurados.

Las nuevas tecnologías de base de datos no estructuradas cubren las limitaciones de las bases de datos relacionadas, mejorando en la capacidad de procesamiento y gestión de grandes volúmenes de datos.

- **Utilidad del Big Data**

Luego de mencionar el alcance de la implementación de la tecnología de Big data y los limitantes de las bases de datos relacionales corresponde mencionar las utilidades de este tipo de tecnología.

Existen varias utilidades en la implementación de una arquitectura de Big Data, entre las más importantes se pueden detallar las siguientes:

- Procesamiento paralelo de información en grandes cantidades
- Computación en la nube
- Gestión de datos estructurados, semiestructurados y no estructurados
- Identificación de patrón de comportamiento de los datos y poder realizar predicciones
- Mejora en la gestión de toma de decisiones
- Procesamiento en tiempo real de grandes volúmenes de datos
- Detectar tendencias

- **Hadoop (Highly Archived Object Oriented Programming)**

Es un framework desarrollado en Java, utilizado como repositorio de datos, gestiona grandes volúmenes de datos, procesamiento compartido de aplicaciones.

Diseñado para ejecutarse en un solo nodo o en clúster de servidores. Permite compartir recursos (Memoria, procesamiento, almacenamiento), consta de varios módulos mostrados en la Figura 6:

- **Hadoop Common**: Integra los diferentes módulos de Hadoop. Contiene el código fuente de los demás módulos que conforman esta plataforma.

- **HDFS - Hadoop Distributed File System:** Es el Sistema de archivo distribuido empleado en esta plataforma. Sus principales características son: escalable, confiable que puede almacenar gran cantidad de datos, no consume demasiados recursos de hardware, altamente tolerante a fallos, diseñado para ser portable de una plataforma a otra.

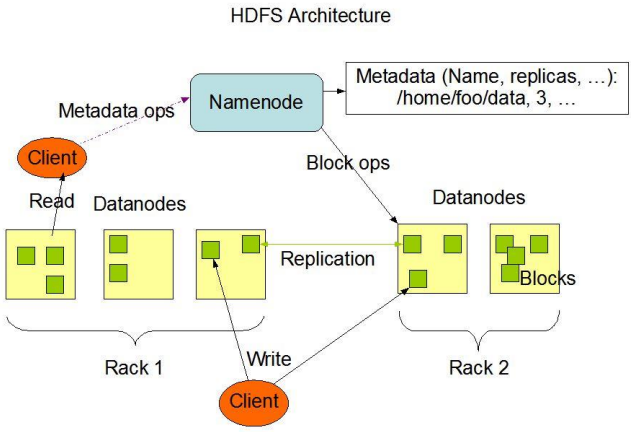


Figura 6: Arquitectura Hadoop. Fuente : www.apache.org

Esta plataforma almacena la información en bloques (chunks) de 64 MB.

HDFS está conformado por un Namenode, uno o más Datanodes y un Secundary Namenode.

El Namenode es el encargado de gestionar la metadata de los archivos y directorios del HDFS. Conecta los Datanodes que registran la información que sale e ingresa del clúster (Apache Software fundation, 2019).

- **Etapa de Visualización**

En esta etapa se utilizan herramientas que permitan interactuar con los datos analizados con los usuarios finales. Los datos son presentados a los usuarios de manera gráfica para que sean de fácil interpretación.

- **Power BI**

Es una herramienta de visualización de datos catalogada entre las más utilizadas de acuerdo al cuadrante de Gartner. (Microsoft Corporation, 2019). Esta solución se integra con diferentes orígenes de datos y contiene diversidad de gráficos para el análisis y presentación de los mismos.

Microsoft empresa propietaria del producto realiza actualizaciones de la herramienta de manera mensual, adicionando nuevas opciones de análisis o actualizando los ya gráficos existentes (Microsoft Power BI, 2016).

La propuesta de investigación incluye la implementación de la arquitectura de Big Data, en la solución se aplicó esta tecnología con la instalación de Hadoop. Para la auditoria de los trámites de importación se debe utilizar una herramienta que permita el almacenamiento y la gestión masiva de los datos.

1.2 Metodología CRISP-DM

CRISP-DM (Cross Industry Standard Processs for Data Minig), metodología utilizada para ejecutar el proceso de Minería de Datos. En esta metodología se describen los niveles que se deben cumplir para poder implementar un proyecto de minería de datos con éxito. Está organizada 4 niveles, el primer nivel son las fases, en el segundo nivel las tareas genéricas de cada fase, en el tercer nivel las tareas específicas y en el cuarto nivel las instancias del proceso.

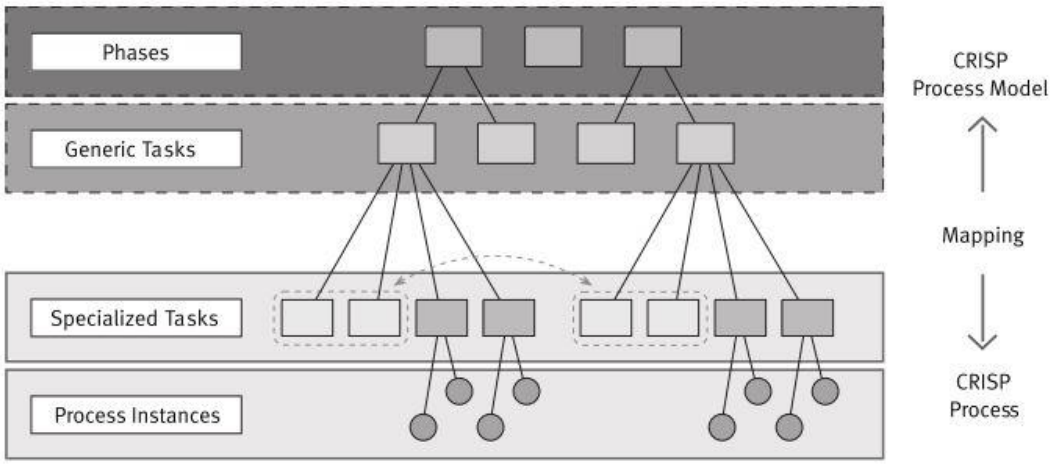


Figura 7: Metodología CRISP_DM. Fuente: IBM Manual de CRISP-DM SPSS Modeler

Fases de CRISP-DM

La metodología CRISP-DM explica detalladamente las tareas a realizar en cada fase, el cumplimiento de cada una de estas aseguran una implementación eficiente de un proyecto de minería de datos.

Las fases de CRISP-DM se muestran en la figura 8 las mismas que inician con el entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación del modelo y despliegue.

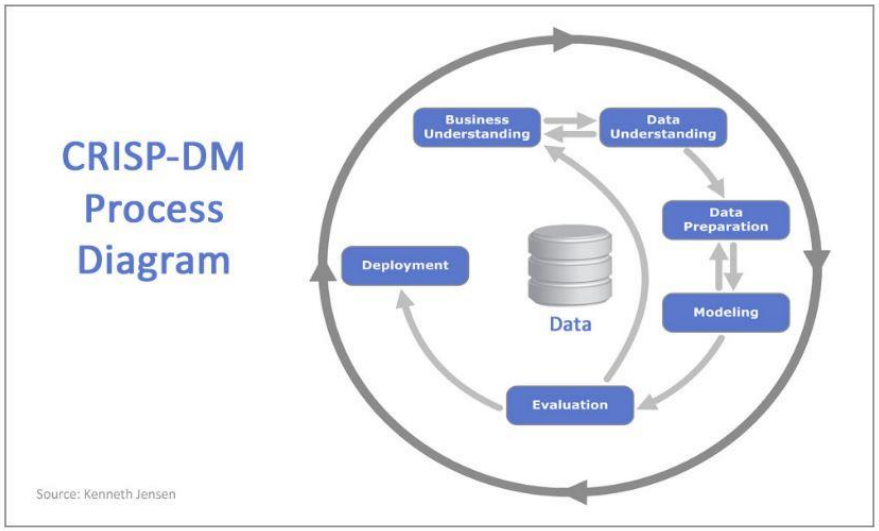


Figura 8: Etapas de la Metodología CRISP-DM

En la Tabla 2 se describen las tareas que deben ser realizadas para la implementación de un proyecto de inteligencia de Negocios:

Tabla 2
Detalle de las etapas y tareas de la Metodología CRISP-DM

No	Tarea
1	**Comprensión del Negocio**
1.1	Riesgos del proyecto
1.2	Beneficios del Proyecto
1.3	Objetivos de la Minería de Datos
1.4	Plan de Proyecto
2	**Comprensión de los datos**
2.1	Recopilación de los datos
2.2	Descripción de los datos
2.3	Exploración de los datos
2.4	Calidad de los datos
3	**Preparación de los datos**
3.1	Selección de los datos
3.2	Limpieza de los datos
3.3	Construcción de los nuevos datos
3.4	Integración de los datos
3.5	Formato de los datos
4	**Modelado de Datos**
4.1	Selección del modelo
4.2	Generación del Plan de Prueba
4.3	Construcción del Modelo
4.4	Evaluación del Modelo

Fuente: Manual CRISP-DM de IBM SPSS Modeler

- **Entendimiento del Negocio**

En esta fase se debe entender los objetivos y establecer los requerimientos del negocio y convertir el problema del negocio en una solución con la minería de datos. En esta fase de debe elaborar un plan donde se definan las tareas a cumplir y poder conseguir los objetivos esperados.

- **Entendimiento de los Datos**

En esta fase se inicia con la ingesta de los datos que se van a analizar, luego conocer en más detalle los datos, identificar los problemas de calidad de los datos y detectar subconjunto de datos para formarse una hipótesis sobre la información oculta.

- **Preparación de los datos**

En esta fase se construye el set de datos final que se utilizaran para la minería de datos. Se ejecutan además las tareas de transformación y limpieza de los datos para utilizarlos en el modelado de los datos.

- **Modelado de los datos**

En esta fase se deben aplicar las técnicas de modelado de datos. Las técnicas a ser aplicadas dependen de la preparación de los datos, algunos modelos dependen de cómo se encuentren preparados los datos, sí los datos no se encuentran bien preparados se deberá regresar a esa fase.

- **Evaluación**

En esta fase se deberá revisar exhaustivamente los pasos que se realizaron para llegar al modelo. En esta fase se debe revisar si los resultados obtenidos son los esperados, esto se lo debe realizar antes de poder desplegar los resultados

- **Despliegue**

Luego del análisis de los datos y obtenido el conocimiento este debe ser presentado para que el cliente pueda utilizarlo. Previamente debe ser revisado por parte del cliente sobre los resultados obtenidos en el análisis de los datos para ser publicados, cuando los resultados deben ser en tiempo real, en las etapas previas se debe realizar la revisión correspondiente.

Metodología KDD (Knowledge Discovery in Databases)

Metodología utilizada para encontrar información en un gran conjunto de datos y luego generar conocimiento. Esta metodología cuenta con varias etapas consecutivas.

Las etapas de esta metodología son:

- **Identificación del problema:** En esta etapa se debe identificar el problema a resolver, y definir el alcance de la solución propuesta.
- **Selección e integración de los datos:** En esta etapa se obtienen los datos de los sistemas transaccionales, estos datos pueden estar en diferentes

formatos y pueden contener valores nulos por lo que es importante que estos sean depurados.

- **Preparación de los datos:** En esta etapa los datos son depurados ya que en la mayoría de los casos estos contienen valores con errores o con valores nulos. En esta etapa se debe unificar formatos.
- **Transformación y almacenamiento de los datos:** En esta etapa se pueden agrupar o reducir los datos en características que sean de utilidad para su análisis, luego son almacenados estos datos en un Data warehouse o en un datamart,
- **Selección y Aplicación de algoritmos de Data Mining:** De acuerdo al tipo de análisis que se desea realizar se debe seleccionar y aplicar el algoritmo de minería de datos, El objetivo en esta etapa es por medio de la aplicación del algoritmo encontrar patrones de comportamiento de los datos.
- **Interpretación y evaluación de patrones:** En esta etapa se deben interpretar los patrones para convertirlo en conocimiento, este proceso debe ser apoyado por expertos en el negocio para la evaluación de los resultados obtenidos. Para interpretarlos es necesario poder identificarlos mediante la utilización de herramientas de visualización.

Metodología SEMMA

Esta metodología se la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de comportamiento en los datos. Esta metodología consta de cinco fases:

- **Selección de la muestra:** En esta etapa se debe seleccionar la muestra de los datos para su análisis. La muestra seleccionada debe ser representativa ya que de esto depende la obtención de los resultados.
- **Exploración de los datos:** Luego de la selección de los datos estos deben ser explorados con el fin de simplificar en lo posible el problema y optimizar la aplicación del modelo. Este proceso se lo debe realizar mediante la utilización de herramientas de visualización o con la aplicación de técnicas estadísticas para identificar la relación entre las variables.

- **Manipulación de los datos:** Esta etapa consiste en la manipulación de los datos, esto implica el aplicar formato adecuado a los datos que serán utilizados en el modelo.
- **Modelado de los datos:** En esta etapa se debe establecer una relación entre las variables explicativas y las variables de objeto de estudio. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos y algoritmos de minería de datos.
- **Valoración del Modelo:** En esta etapa consiste en la valoración de resultados mediante el análisis de los modelos utilizados o con la aplicación de otros métodos estadísticos o con la selección de nuevos datos de muestra.

En la propuesta de investigación se analizaron tres metodologías de minería de datos, las metodologías analizadas fueron CRISP-DM, KDD y SEMMA. (Rodriguez, Alvarez, Mesa, & González,s.f)

La metodología CRISP-DM es la más utilizada en el desarrollo de proyectos de minería de datos, se caracteriza por la descripción de las tareas y actividades que se deben ejecutar en cada fase. Este modelo se diferencia por el nivel de profundización y el detalle de las tareas y actividades.

Luego del análisis se determinó la utilización de la metodología CRISP-DM para la implementación del proyecto, la selección de basó por el nivel de detalle que se especifica para las actividades y las tareas que se deben desarrollar en cada fase.

1.3 Servicio de Aduana

El Servicio de Aduana es una institución gubernamental, manejada de manera autónoma, orientada a la atención y servicio de los operadores del comercio exterior. Los elementos que conforman parte de la identidad estratégica de la organización: e

Misión: Servir a los operadores de comercio exterior de manera integral y comprometida, cimentado en valores; que incentive la libre competencia en el país con un equilibrio entre el control efectivo y la facilitación del comercio (SENAE, 2019).

De acuerdo a lo establecido en el código orgánico de la producción comercio e inversiones facilitación del comercio es: promover la simplicidad y agilidad en los procesos aduaneros y el intercambio de la información sea de manera electrónica. (COPCI, 2013)

De acuerdo al COPCI el Servicio de Aduana debe ejecutar controles eficientes a las operaciones de comercio exterior por medio de procesos de gestión de riesgo, respetando las leyes, reglamentos y normativas, así también el buen uso de los recursos de las arcas fiscales (COPCI, 2013).

Para la importación de mercaderías al país la aduana tiene establecido regímenes aduaneros de importación, estos se encuentran están diferenciados de acuerdo al pago de los tributos, el detalle se los muestra en la Tabla 2.

Tabla 3
Regímenes de Importación

Régimen Aduanero	**Articulo - COPCI**
Importación para el consumo	147
Régimen de admisión temporal para perfeccionamiento activo	148
Reposición de mercancías con franquicia arancelaria	149
Transformación bajo control aduanero	150
Depósito aduanero.	151
Reimportación en el mismo estado.	152

Fuente: Regímenes Aduaneros de Importación Código Orgánico Producción Comercio e Inversiones COPCI. Elaboración propia

Los tributos aduaneros establecidos por el Servicio de Aduana y obligados a ser pagados por los operadores de comercio son los que se detallan a continuación:

1. Impuestos por las importaciones o exportaciones,
2. Derechos arancelarios;
3. Tasas por servicios otorgados por el Servicio de aduana. (COPCI, 2013).

Para el pago de los tributos se lo realiza en base al valor comercial de las mercancías que ingresan al país. Este valor es representado como valor en aduana, el cual se encuentra conformado por el valor comercial de la mercancía más los costos del transporte y del seguro. (COPCI 2013).

En el caso de que los operadores de comercio exterior llegarse a incurrir en alguna contravención el Servicio de Aduana tiene reglamentado las respectivas sanciones, estas se encuentran reglamentadas en el Código Orgánico de la Producción, Comercio e inversiones

Las contravenciones establecidas por el Servicio de Aduana y estipuladas en el COPCI, son las siguientes:

a) Ingreso no autorizado de personas a las zonas primarias aduaneras
b) Transmisión de las declaraciones aduaneras Total o parcial de manera electrónica fuera del plazo establecido
c) Entregar las Mercaderías a ser descargadas fuera de los tiempos establecidos.
d) Entregar información confidencial de la institución por parte de los servidores públicos sin autorización.
e) La no entrega del listado de pasajeros a la aduana, por parte del transportista, a la llegada o salida del medio de transporte.
f) La no entrega de las mercadurías por parte del transportista a la aduana registrada en el manifiesto de carga.
g) No permitir acciones de control por parte de la administración aduanera
h) No cumplir con los tiempos establecidos para los reembarques o trasbordos
i) La no presentación de documentos de soporte al mismo tiempo con la declaración aduanera.

j) La sobrevaloración o subvaloración de las mercancías en el proceso de control posterior. Encontrar mercancías no declaradas cuando se esté realizando el aforo físico de las mercadurías.

En el caso se incurrir en el método de subvaloración o sobrevaloración la sanción será:

- Con multa equivalente de trescientos por ciento (300%) del valor en aduana de las mercancías no declaradas o de la diferencia entre el valor declarado y el valor en aduana de las mercancías según corresponda; (COPCI, 2013)

La administración del Servicio de Aduana está conformada por un Director General y varios directores nacionales. La Figura 9 presenta su estructura organizacional:

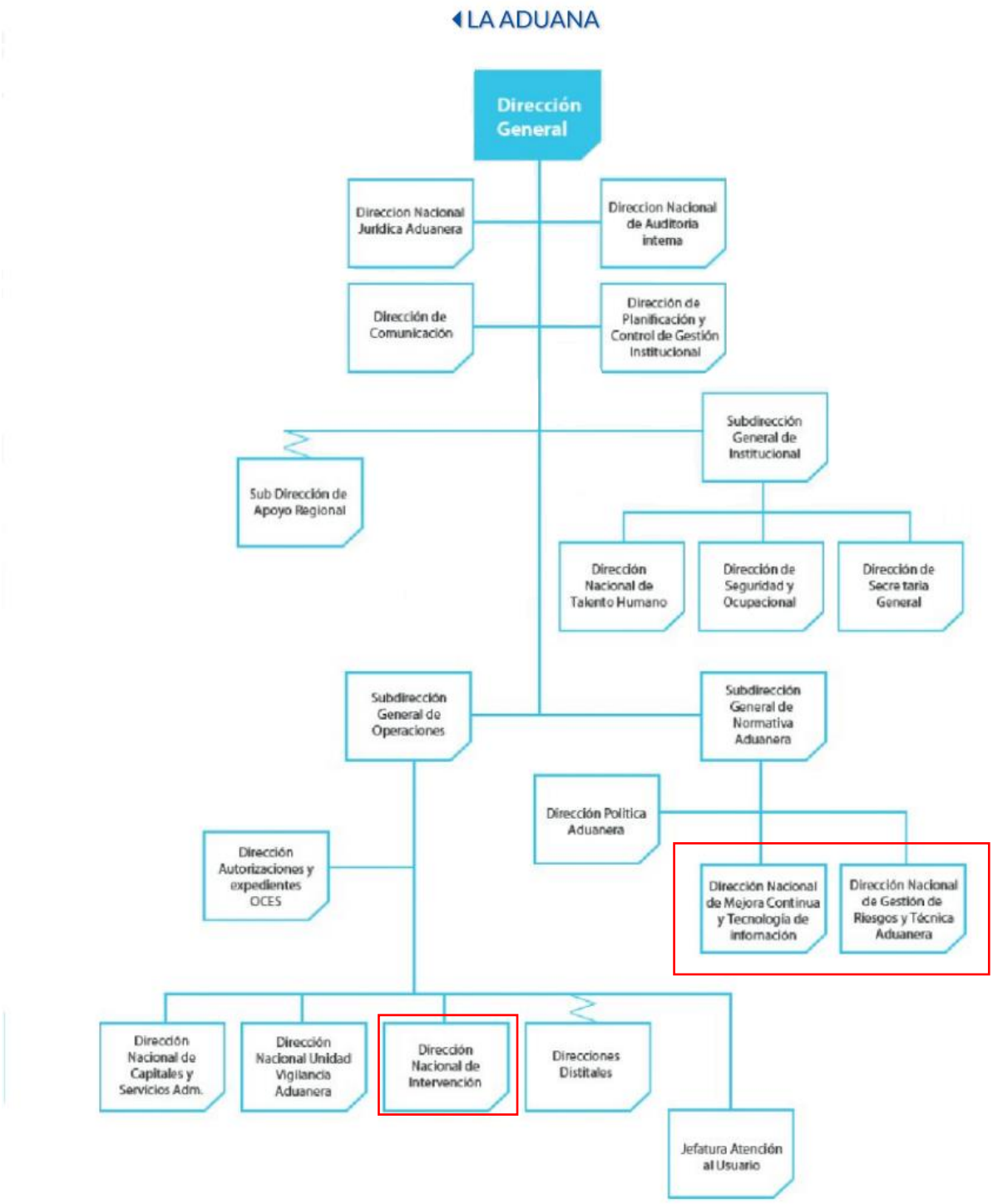


Figura 9: Estructura Organizacional Aduana

Las Direcciones involucradas para la implementación de la propuesta de solución fueron: Dirección Nacional de Gestión de Riesgos y Técnica Aduanera , la Dirección Nacional de Intervención y la Dirección Nacional de Mejora Continua y Tecnología

de la información. Las funciones y responsabilidades de las direcciones intervinientes se describen a continuación:

Dirección Nacional de Riegos y Técnica Aduanera

La misión de esta dirección es gestionar el riesgo de los operadores de comercio exterior, asegurar que se cumplan las obligaciones aduaneras, buscar que el despacho de las mercadurías se las realice de manera expedita. Dentro de las atribuciones y responsabilidades se van a detallar las siguientes:

- Identificar grupos de operadores de comercio y sectores económicos sujetos de control por medio de perfilamiento de riesgo.
- Análisis de fuentes de información interna y externa, para identificar los operadores de comercio exterior a controlar.

Dirección Nacional de Intervención

La misión de esta dirección es realizar los operativos de control a los operadores de comercio exterior y dentro de sus responsabilidades se tienen las siguientes:

- Establece mecanismos de selección de los controles a desarrollarse de acuerdo al Plan de Intervención.
- Retroalimentar a la Dirección Nacional de Gestión de Riesgos y demás direcciones los hallazgos encontrados en los controles posteriores

Dirección Nacional de Mejora Continua y Tecnología de la Información

Esta dirección tiene la misión de la elaboración y ejecución de proyectos con la utilización de nuevas tecnologías. Entre sus responsabilidades se tienen las siguientes:

- Gestionar y/o efectuar el análisis, diseño, desarrollo y despliegues de los sistemas informáticos, cumpliendo los estándares y con la calidad establecida

- Elaborar y evaluar tecnologías nuevas para la implementación en sus sistemas informáticos.

Control Posterior

De acuerdo al Régimen Andino de Control Aduanero, Control Posterior se lo realiza: "Posterior a la autorización del pago levante o levante de las garantías aduaneras.

El control posterior se lo puede realizar a:

- Facturas o documentos relacionados a la importación o exportación de las mercadurías;
- Las operaciones de perfeccionamiento de las mercadurías declaradas bajo este régimen.
- A las mercancías declaradas bajo importación temporal y demás destinos aduaneros.

Valoración Aduanera

En base al código orgánico de la producción, comercio e inversiones el cálculo de los tributos aduaneros se lo realiza sobre la Base imponible, a continuación el detalle El valor en aduana de las mercancías será el valor de transacción de las mismas más los costos del transporte y seguro.

Cuando la base imponible no se pueda establecer por medio del valor en aduana, se establecerá por medio de los métodos secundarios de valoración previstos en las normas.

La presente investigación fue implementada en el Departamento de Intervención en el Servicio de Aduana, este departamento se encarga de realizar los controles posteriores de los trámites de importación para verificar si estos han cumplido los requisitos de Ley.

Aplicación de Big Data en Investigaciones similares

De acuerdo a lo indicado por el Ing. Fernando Manso en su tesis para la obtención del título de Master en Gestión de Servicios Tecnológicos y de Telecomunicaciones, "Análisis de Modelo de Negocios basados en Big Data para operadores móviles", indica:

En la actualidad las empresas de telecomunicaciones se encuentran frente al gran desafío en revertir su situación en los próximos años realizando un cambio estratégico para no perder rentabilidad a largo plazo, este cambio se está realizando con la implementación de soluciones de Big Data para la gestión de grandes volúmenes de información. La aplicación de esta tecnología permite mejorar el modelo del negocio actual en lo que corresponde a la satisfacción del cliente, reducción del churn y la apertura de nuevos servicios en este negocio.

En la tesis realizada por el Ing. Sergio Espinoza sobre "Generar un marco de referencia para las implementaciones de Big Data en empresa de telecomunicaciones CNT", en la misma indica:

- La arquitectura Hadoop resulto ser la mejor tecnología que se adapta y encaja muy bien con Big Data ya que proporciona características como escabilidad , tolerancia a fallos, y soporta diferentes cambios en la arquitectura tanto en nivel de diseño o funcionalidades como a nivel de infraestructura, con un buen rendimiento en el trato de datos no estructurados y con una alta flexibilidad para añadir nuevas herramientas a una solución e incluso acoplarse con ya existentes para tener un efecto menor sobre las infraestructuras ya desplegadas.

1.4 Marco Legal

El marco legal que sustenta esta investigación son las leyes que rigen el comercio internacional y el comercio exterior en el país, además se referencia al Plan Nacional del Desarrollo de la nación teniendo el cual establece como meta adoptar medidas para el incremento de implementación de proyectos informáticos.

Comunidad Andina (1969)

Decisión 571

Normativa al tratamiento de la valoración aduanera de las mercancías: Normativa que regula los mecanismos para la determinación del valor en aduana utilizado para el cálculo de los trámites aduaneros.

Decisión 778

Normativa sobre el control aduanero: Normativa que regula los mecanismos el control que debe ejercer la aduana sobre las operaciones de comercio exterior

Convenio de Kyoto (1973-1974)

Directivas sobre el control aduanero: Convenio donde estipula las directrices que deben seguir los países para la simplificación de los trámites y facilitación del comercio internacional.

Código Orgánico de la Producción, Comercio e Inversiones COPCI. (2010)

Normativa Legal que se regirá por los principios que permitan una articulación internacional estratégica, a través de la política comercial, incluyendo sus instrumentos de aplicación y aquellos que facilitan el comercio exterior, a través de un régimen aduanero moderno transparente y eficiente.

Plan de Desarrollo del Ecuador (2017-2021)

El gobierno del Ecuador dentro de sus metas para el año 2021 incrementar de 4.6 al 5.6 el Índice de Desarrollo de Tecnologías de la Información y Comunicación

CAPITULO 2

2 Solución Propuesta

Para la implementación de la solución propuesta se utilizó la Metodología CRISP-DM, se siguieron cada una de las tareas de cada etapa descritas en este método. El seguir esta metodología facilito el desarrollo del proyecto.

2.1 Comprensión del Negocio

En esta etapa se menciona la misión de la empresa, su estructura organizacional, responsabilidades de las áreas involucradas, y así mismo se describirá el objetivo del negocio, realizar una valoración de la situación, determinar los objetivos de la minería de datos, y se describirá el plan de proyecto de minería de datos.

Para la comprensión del negocio se realizaron reuniones de trabajo con los técnicos de las Direcciones de Mejora Continua y Tecnología de la información, Dirección de Control Posterior y la Dirección de Riegos y Técnica Aduanera. Así mismo, se realizó un compromiso de confidencialidad debido a que los procesos involucrados contienen información sensible.

- **Beneficios para la institución**

Con la implementación de la propuesta podría alcanzar algunos beneficios para la institución, los cuales serían:

1. Incrementar la detección de posibles casos de evasión de impuestos por subvaloración a través de la identificación de trámites aduaneros de importación que hayan declarado un valor menor a lo permitido.
2. Identificar a los agentes o importadores aduaneros que utilizan la modalidad de subvaloración de las mercancías y categorizarlos como de mayor riesgo.

3. Poder determinar los países de origen de las mercaderías que se presentan con mayores casos de subvaloración.

- **Criterios de éxito del Negocio**

El poder determinar nuevos casos de evasión de impuestos por el método de subvaloración aduanera y poder identificar patrones de comportamiento permitirá incorporar nuevas reglas de perfilamientos en base a los hallazgos encontrados y categorizarlos como de alto riesgo a los operadores de comercio que hayan incurrido en esta modalidad y levantar una mejor logística para las revisiones físicas a las mercaderías.

- **Descripción de la Problemática**

Con el aumento de las importaciones en el país el Servicio de Aduana le resulta imposible poder verificar físicamente todas y cada una las declaraciones aduaneras presentadas, de darse la revisión se aumentaría sustancialmente el tiempo de salida de las mercaderías de aduana, razón por lo cual se debe implementar métodos de selección de operadores de comercio por medio de análisis de riesgos.

El análisis de riesgos debe poder identificar operadores de comercio exterior que podrían considerarse de alto o mediano riesgo tomando, uno de los parámetros a tomar como referencia son el país de procedencia de las mercaderías o montos en los valores de los tributos aduaneros, aquellas considerados de un menor riesgo serán sujetas a revisión documental o administrativa.

Las revisiones que se ejecutan por parte de la aduana se realizan de acuerdo a lo estipulado en por el convenio de Kioto, por lo que se verifica:

- Clasificación arancelaria, depende de las mercaderías a importarse.
- Lo declarado electrónicamente debe estar de acuerdo a los documentos de soporte.

- El régimen aduanero declarado debe corresponder a lo que realmente se está importando.
- Los valores de las mercaderías estén de acuerdo a su valor original de compra, soportado por medio de las facturas emitidas en el país donde fueron adquiridas.

La aduana luego de ejecutar el proceso de perfilamiento de riesgo y si existe algún delito aduanero deberá proceder a categorizar a los agentes y a los importadores de acuerdo al grado del delito, se los categoriza como operadores de comercio de alto riesgo, aquellos que no infringen con la ley aduanera se los categoriza como de menor riesgo. Las revisiones físicas se realizan a los operadores de comercio de alto riesgo y a los de menor riesgo solo revisión documental.

Con el fin de optimizar el control y la recaudación aduanera, la Dirección Nacional de Riesgos pretende mejorar la identificación de posibles actos de evasión aduanera, uno de los métodos más utilizados para la evasión de impuestos es por medio de la subvaloración de las mercancías, este método consiste en declarar un valor menor al estipulado en la aduana para beneficiarse con un pago menor en los tributos aduaneros.

Para poder detectar los casos de evasión por la modalidad de subvaloración se debe clasificar las mercaderías por subpartida arancelaria y por país de procedencia determinando aquellas transacciones cuyos valores tienen una desviación atípica sobre los que se consideran normales.

- **Proceso Actual**

El proceso de determinación de casos de evasión de impuestos se lo realiza por medio de un control posterior y es aplicado a las declaraciones aduaneras que pertenecen a operadores de comercio con un nivel de riesgo alto, aplicado a

mercaderías sensibles como electrodomésticos, textiles, cigarrillos, a continuación se describe con mayor detalle el proceso en la Figura 10:

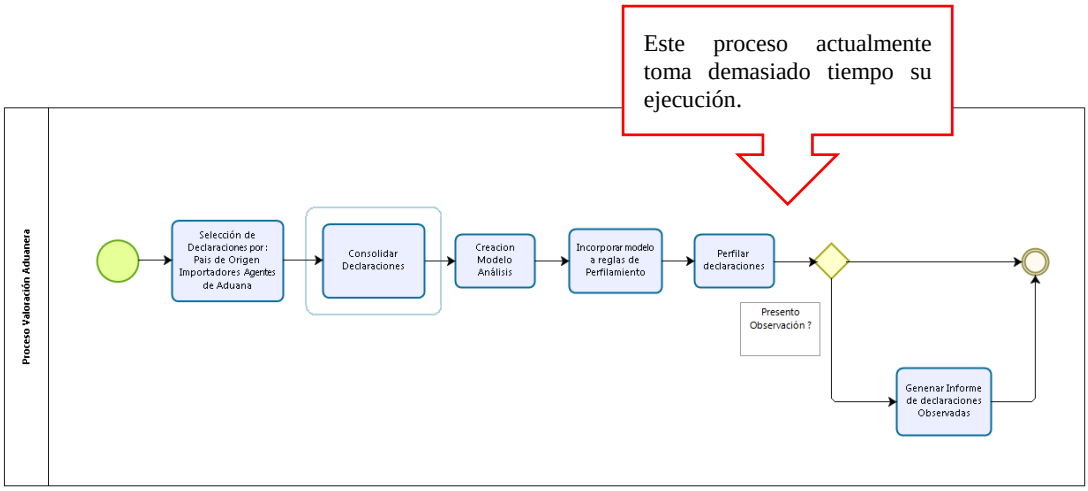


Figura 10: Proceso de subvaloración aduanera. Fuente: Elaboración propia

1. **Seleccionar las declaraciones aduaneras por País de Origen y por Agentes o Importadores:** Los precios de las mercaderías son diferentes en algunos países por lo que se debe realizar el análisis por un país específico. Además se puede filtrar por Agente de aduana que realzó el proceso de importación.
2. **Consolidación de declaraciones:** Las declaraciones que son seleccionadas para el análisis son extraídas de las bases de datos transaccionales y almacenados en un Data Warehouse para la ejecución del proceso de subvaloración.
3. **Creación del Modelo de Análisis:** Los analistas de gestión de riesgos realizan la verificación del modelo de análisis a ejecutar para el proceso de subvaloración aduanera.
4. **Incorporar el modelo a las reglas de perfilamiento:** Luego de la revisión del modelo de análisis este debe ser incorporado al módulo de perfilamiento de riesgos para ejecutar el proceso.
5. **Perfilar declaraciones** : Revisar que el valor declarado este dentro del rango de lo estipulado en la aduana (Validar el precio de las mercaderías registrado en la Base de Valor)
6. **Generar informes con las declaraciones que hayan sido reportadas como subvaloradas**: Las declaraciones aduaneras que hayan sido

identificadas por el perfilador de riesgo como subvaloradas, estas deben ser reportadas para que los infractores presenten documentos de soporte.

Uno de los inconvenientes que se presenta en el proceso de perfilamiento de riesgo es que cada vez se torna más lento debido a que debe validar los precios de las mercaderías declaradas contra los precios registrados en la Base de Valor Aduana (Precios de referencia de las mercaderías). Cabe señalar que, los precios de las mercaderías de importación son actualizados continuamente por técnicos de Valoración Aduanera, esta base crece constantemente debido a que se incrementan las mercaderías importadas.

Actualmente, los datos se encuentran registrados en una Base de Datos Oracle 12g las declaraciones de importación desde el año 1995 hasta la presente. Los datos a ser analizados van a ser del año 2015 con un set de datos de 10.000 registros. El set de datos va a contener información de las declaraciones, carga y de los ítems de las mercaderías.

2.1.1 Riesgos

En la implementación de proyecto de Inteligencia de negocios, es importante poder identificar los riesgos con el fin de poder mitigarlos o tomar las medidas para poder evitarlos o controlar el impacto de producirse. Se han identificado algunos riesgos que pueden darse, a continuación el detalle de los riesgos identificados en el desarrollo se presenta en la Tabla 4.

Tabla 4
Riesgos del Desarrollo del proyecto

Riesgo	Prioridad	Probabilidad	Causa
Aumento del tiempo de las actividades	Media	Poco Probable	Retraso en la entrega del Proyecto
Falta de compromiso de los usuarios finales	Alta	Muy probable	Disminución de la calidad del producto
Cambio de funcionarios en cualquier etapa del proyecto	Media	Poco Probable	Aumento del tiempo de las pruebas del producto

Fuente: Elaboración propia

2.1.2 Beneficios del Proyecto

A continuación el detalle de los beneficios identificados con la implementación del proyecto.

- Mayor procesamiento de declaraciones de importación para la detección de casos de evasión de impuestos
- Reducción de los costos en la implementación de servidores
- Reducción de costos en el almacenamiento
- Reducción de costos de licenciamiento en base de datos Oracle
- Reducción de costos en soporte y mantenimiento
- Reducción del tiempo de procesamiento y análisis de los datos
- Aumento de la efectividad en la toma de decisiones.

2.1.3 Objetivos de la Minería de Datos

Con la implementación del proyecto la Dirección de Gestión de Riesgos y técnica Aduanera van a mejorar en el control en la evasión de los impuestos con la incorporación de una nueva técnica de análisis de datos a través de la minería de datos.

A continuación el detalle de las metas de la minería de Datos:

a. Procesar grandes volúmenes de declaraciones de importación a nivel de partida arancelaria
b. Crear agrupamientos
c. Identificar aquellos casos en los cuales existan una mayor desviación de los precios normales
d. Visualizar estos casos para su posterior investigación

2.1.4 Plan de Proyecto

Para la ejecución de la propuesta de solución se identificaron las actividades que deben cumplirse, iniciando con la implementación de la Infraestructura tecnológica, la instalación del software, creación del algoritmo. A continuación el detalle de las tareas se presenta en la Tabla 5.

Tabla 5
Planeación del Proyecto

EDT	Nombre de la Tarea	Duración	Comienzo	Fin	Predecesoras
1	**IMPLEMENTACIÓN INFRAESTRUCTURA**	**4 días**	**lun 06/05/19**	**jue 09/05/19**	
1.1	CONFIGURACIÓN DEL SERVIDOR	2 días	lun 06/05/19	mar 07/05/19	
1.2	INSTALACIÓN SISTEMA OPERATIVO	2 días	mié 08/05/19	jue 09/05/19	2
2	**IMPLEMENTACIÓN AMBIENTE BIG-DATA**	**3 días**	**vie 10/05/19**	**mar 14/05/19**	
2.1	CREACIÓN DE BASE DE DATOS	2 días	vie 10/05/19	lun 13/05/19	3
2.2	CREACIÓN DE TABLAS	1 día	mar 14/05/19	mar 14/05/19	5
3	**PROCESO ETL**	**5 días**	**mié 15/05/19**	**mar 21/05/19**	
3.1	PROCESO DE EXTRACCIÓN	3 días	mié 15/05/19	vie 17/05/19	6
3.2	PROCESO DE CARGA	2 días	lun 20/05/19	mar 21/05/19	8
4	**MODELO DE MINERIA**	**7 días**	**mié 22/05/19**	**jue 30/05/19**	
4.1	ELECCIÓN MODELO	2 días	mié 22/05/19	jue 23/05/19	9

4.2	DESARROLLO	5 días	vie 24/05/19	jue 30/05/19	11
4.3	PRUEBAS	2 días	Vie 31/05/19	Lun 03/05/19	
5	**ANÁLISIS**	**6 días**	**mar 04/06/19**	**vie 09/06/19**	
5.1	ANÁLISIS DE LOS RESULTADOS	3 días	mar 04/06/19	jue 06/06/19	12
5.2	AJUSTES	3 días	jue 06/06/19	lun 10/06/19	14
6	**VISUALIZACIÓN**	**5 días**	**mar 11/06/19**	**lun 17/06/19**	
6.1	CREACIÓN DE INFORMES	5 días	mar 11/06/19	lun 17/06/19	15

Fuente: Elaboración Propia

2.2 Comprensión de los Datos

De acuerdo a la Metodología CRISP-DM en esta fase se debe iniciar con la recolección de datos, se identifica los datos a utilizar para el proceso de minería, los mismos se deben seleccionar de manera correcta, ya que de esto dependerán los resultados esperados. Luego de la selección se deberá validar su calidad y relevancia agregando valor al proceso que se está optimizando.

2.2.1 Recolección datos Iniciales

Los datos iniciales empleados para la propuesta de solución son las declaraciones aduaneras de importación, estas pertenecen al [1]régimen aduanero de Importación para el consumo (Régimen 10). Los datos recolectados son históricos ya que el proceso de fiscalización se lo realiza dentro de la etapa de control posterior. En esta etapa se analizan los datos de hasta cinco (5) años.

En el negocio de comercio exterior la información debe ser manejada de manera discreta, la divulgación provocaría que las investigaciones y la fiscalización no se desarrolle de manera eficaz imposibilitándose la identificación de posibles casos de

[1] Régimen 10: Código Orgánico de la Producción, Comercio e Inversiones Art. 147.- Importación para el consumo.

evasión, por tal motivo los datos de las declaraciones de importación utilizados para esta investigación fueron modificados para mantener la confidencialidad debido a que esta información es muy sensible.

A continuación se detallan los datos seleccionados en esta etapa:

Declaraciones de Importación: En esta tabla se registran las todas las declaraciones de importación de diferentes regímenes aduaneros, para la minería de los datos a realizar solo se va a considerar declaraciones régimen [10] son aquellas mercaderías de importación a consumo. Las declaraciones de importación seleccionadas corresponden al año 2015 y pertenecen a ciertas subpartidas arancelarias y de países de donde existe el mayor número de importaciones.

Carga: En esta tabla se registra el detalle de la mercadería que se va a importar, como es el país de origen, valor total de la transacción y los ítems totales a importar.

Mercaderías: En esta tabla se encuentra el detalle de los ítems declarados, como es el código de la subpartida arancelaria, la cantidad de ítems declarados y el valor de cada mercadería.

Los atributos que se seleccionaron para el proceso de minería de datos son los siguientes:

1. Número de declaración aduanera
2. País de Procedencia
3. Aduana Destino
4. Subpartida Arancelaria
5. Cantidad
6. Valor Unitario
7. Valor CIF (Costo + Seguro + Flete)

2.2.2 Descripción de los datos

Los datos a ser descritos pertenecen a la información de declaraciones aduaneras, información de carga e información de las mercaderías declaradas, a continuación el detalle:

DECLARACION DE IMPORTACION

Tabla 6 Declaraciones de Importación

Declaración aduanera	Código asignado por aduana
Código del Declarante	Código del agente
Aduana	Aduana donde se procesa la salida de la mercadería
Código Régimen	Régimen Aduanero declarada la mercadería
Tipo de Despacho	Tipo de Despachos – Normal – Fluvial
Tipo de Pago	Contado – Anticipado
Nombre Importador	Persona que realiza el ingreso de la mercadería al país
Tipo de Documento	Cedula – Ruc – pasaporte
Ciudad	Donde reside el importador
Dirección	Dirección del importador
Teléfono	Teléfono del importador
CIIU	Actividad económica del importador
Código Agente	Código del agente de aduana
Tipo documento Agente	Ruc del agente de aduana
Nombre del agente	Nombre completo del Agente
Dirección del agente	Dirección del agente de aduana

Fuente: Elaboración propia

CARGA

Tabla 7 Información de Carga

Declaración aduanera	Código asignado por aduana
País de Procedencia	País de origen de la mercadería
Consignatario	Persona a quien se le destina la carga
Número de Carga	Manifiesto generado por el transportista
Documento de transporte	Expedido por la Línea Naviera o Aérea
Valor FOB	Precio Neto de la mercadería
Flete	Valor del Flete
Seguro	Valor del Seguro de la mercadería
Valor en Aduana	Suma de los Valores de FOB + Flete + Seguro
Ítems Declarados	Cantidad de Ítems declarados

Peso Neto	Peso Neto en Kilos de la mercadería
Cantidad de Unidades Físicas	Cantidad Total de Unidades Físicas
Cantidad de Unidades Comerciales	Cantidad Total de unidades comerciales

Fuente: Elaboración propia

MERCADERIAS

Tabla 8 Información de Mercaderías

Declaración aduanera	Asignado por el sistema de aduana
Número de Ítem	Secuencial de ingreso de la mercadería
Subpartida	Código que identifica la mercancía ingresada
Código suplementario	Código adicional para la identificación un producto específico
Descripción	Descripción general de la mercadería a ser importada
País de Origen	Código del país de origen de donde proviene la mercadería a importar
Estado	Estado de la mercancía importada
Tipo de tratamiento	Tipo de tratamiento que se le da a la mercadería Importada
Tipo de Mercancía	Tipo de la mercancía a declarar
Peso Neto	Peso Neto en Kilogramos de la mercadería a importar
Ubicación geográfica	Lugar de destino de las mercaderías
Clase	Características que distinguen a la mercadería a importar
Modelo	Modelo del producto
Marca Comercial	Marca comercial de la mercadería a importar
Año de Fabricación	Año de fabricación de la mercadería a importar
Bien de Capital	Identificación de la mercadería a importar si es un bien de capital o No
Cantidad de Unidades Físicas	Cantidad de unidades por ítem de la mercadería a importar
Valor Unitario	Valor Unitario de la mercadería
Valor Total	Valor Total de la mercadería

Fuente: Elaboración Propia

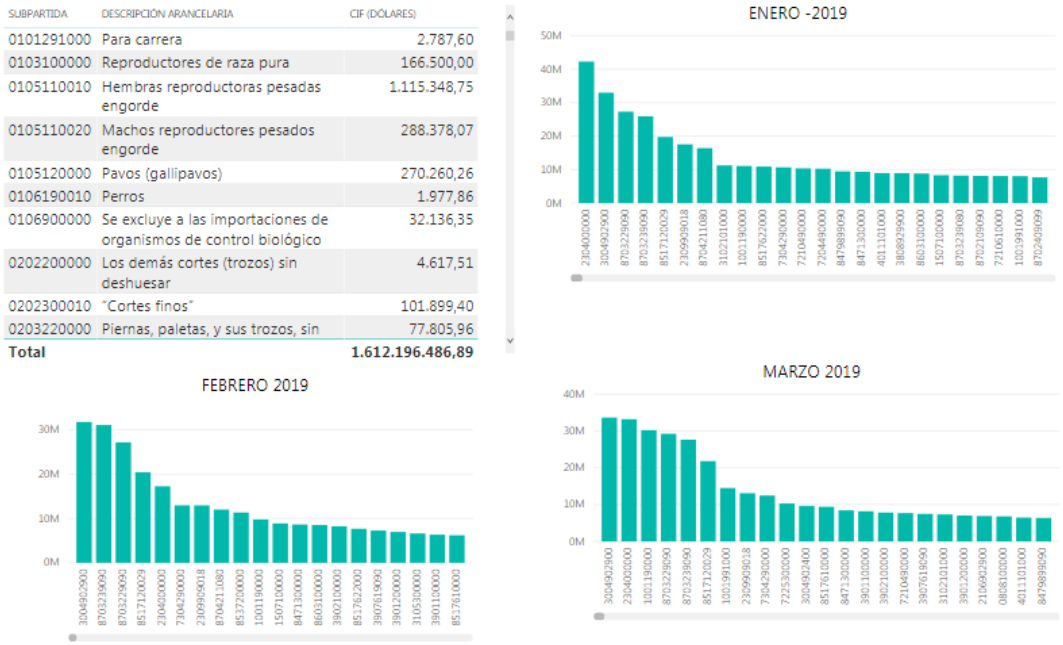
Herramienta de minería de datos

Para realizar la minería de datos de las declaraciones aduaneras de importación se utilizó la herramienta de Power BI donde se aplicó el algoritmo de Clustering

(Detection Outliers) Identificación de valores atípicos, estos corresponderían a los valores subvalorados.

2.2.3 Exploración de los Datos

Para la exploración de los datos se realizó análisis estadísticos para identificar las propiedades de los datos, Se tomaron ciertas subpartidas arancelarias del primer trimestre del año del 2019. La estadística obtenida se presenta en la Figura 11, Figura 12,Figura 13 y Figura 14.



SUBPARTIDA	DESCRIPCIÓN ARANCELARIA	CIF (DÓLARES)
0101291000	Para carrera	2.787,60
0103100000	Reproductores de raza pura	166.500,00
0105110010	Hembras reproductoras pesadas engorde	1.115.348,75
0105110020	Machos reproductores pesados engorde	288.378,07
0105120000	Pavos (gallipavos)	270.260,26
0106190010	Perros	1.977,86
0106900000	Se excluye a las importaciones de organismos de control biológico	32.136,35
0202200000	Los demás cortes (trozos) sin deshuesar	4.617,51
0202300010	"Cortes finos"	101.899,40
0203220000	Piernas, paletas, y sus trozos, sin	77.805,96
Total		**1.612.196.486,89**

Figura 11: Comparación CIF por Subpartidas. Fuente: Elaboración propia

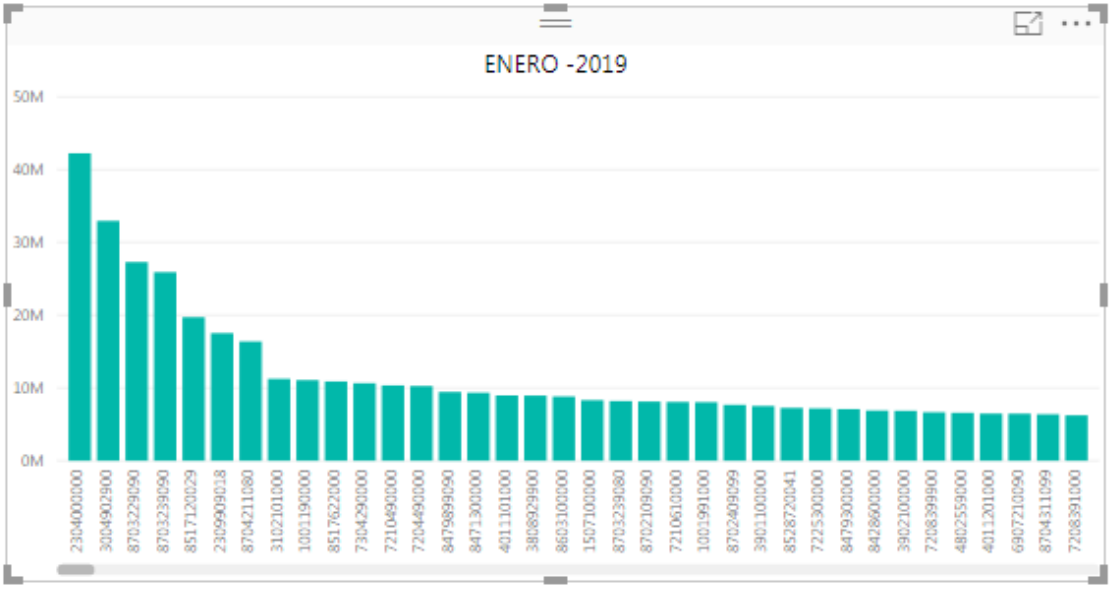


Figura 12: Estadística de Importaciones por Valor CIF Enero 2019. Fuente: Elaboración propia

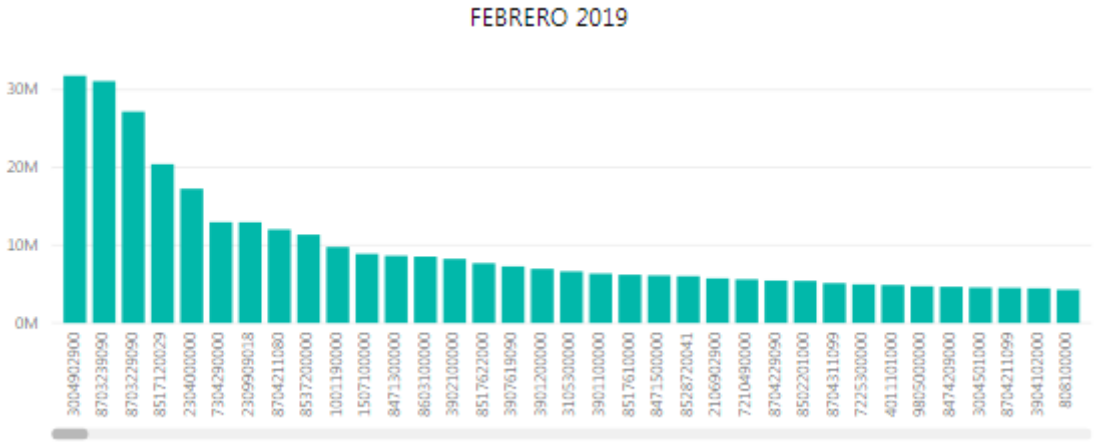


Figura 13: Estadística de Importaciones por Valor CIF Febrero 2019. Fuente: Elaboración propia

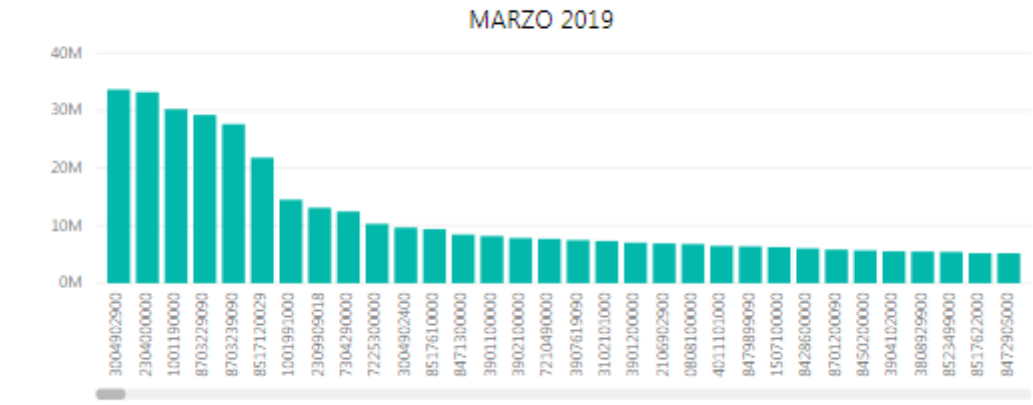


Figura 14: Estadística de Importaciones por Valor CIF Marzo 2019. Fuente: Elaboración propia

En el análisis estadístico detallado en la Tabla 9, se identificaron las siguientes subpartidas con los Valores CIF más altos:

Tabla 9
Subpartidas con mayor valor CIF

No	Subpartida
1	2304000000
2	3004902900
3	8703229090
4	8517120029
5	2309909018

Fuente: Elaboración propia

2.2.4 Calidad de los datos

La aduana tiene implementado el envío y recepción de la información a través de documentos electrónicos, estos edocs[2] tienen ya establecidos formatos y especificado la información que debe contener cada campo específicamente. En la recepción de las

[2] Edocs. Documentos en formato electrónico

declaraciones de importación la Aduana realiza un proceso de validación y verificación de la información, esta revisión consiste en comprobar que no existan campos con valores nulos, que se encuentre completa, que los códigos de los campos sean correctos y existan dentro de los catálogos, que la información de los valores de pesos, cantidades que se encuentra consolidada cuadre con la información del detalle.

Este proceso de validación de la información previo a la aceptación de las declaraciones por parte de la aduana garantiza que los datos a ser empleados en el proceso de minería de datos, y por lo tanto se podrá obtener los resultados esperados y de esta manera poder determinar posibles casos de evasión de impuestos por la modalidad de subvaloración aduanera.

2.3 Preparación de los datos

De acuerdo a la Metodología CRISP-DM en esta fase se deben preparar los datos para ser utilizados en el proceso de minería de datos, la preparación consiste en seleccionar los campos específicos para ser analizados en el modelo de minería.

2.3.1 Selección de los datos

Para ejecutar el proceso de minería de datos se han seleccionado los siguientes campos, a continuación el detalle:

- **Tabla Declaración aduanera**
  - Declaración aduanera
  - Aduana
- **Tabla Carga**
  - País de Procedencia
- **Tabla Ítems**
  - Subpartida
  - Cantidad de Unidades Físicas
  - Valor Unitario
  - Valor Total Aduana

Los datos seleccionados para la evaluación del modelo prototipo son los correspondientes al primer trimestre del año 2019.

2.3.2 Limpieza de los datos

Para el análisis de Los datos por medio de minería se debe realizar un proceso en el cual no se consideren registros en los cuales existan con valor en cero o que se encuentren con valores nulos.

Dentro de los datos se pudieron identificar registros cuyo país de procedencia tenían EC-Ecuador y como se trata de importaciones este valor debe ser diferente a EC.

Además no se consideraron los registros cuyos valores de las mercancías eran iguales a cero.

2.3.3 Construcción de los datos

Para realizar la minería de los datos se tuvieron que realizar las diferentes transformaciones:

Declaración Aduanera Importación DAI: La declaración aduanera está conformada de la siguiente manera:

Código de la Aduana – Año – Régimen Aduanero – Número secuencial de la DAI

055	2019	10	06793456

Este campo Texto se lo convertirá en número por lo que se debe eliminar los caracteres (-).

País de Procedencia: Este campo es carácter por lo que se debe transformar en campo numérico

AF – AFGANISTAN

DE – ALEMANIA

US- ESTADOS UNIDOS

…..

Se le asignara un valor numérico secuencial

1-AFGANISTAN

2-ALEMANIA

3-ESTADOS UNIDOS

…..

2.3.4 Integración de los datos

Para el proceso de minería de datos se creó la estructura que se muestra en la Tabla 10.

TB_MERCADERIAS

Tabla 10
Tabla Mercaderías

Campo	Tipo dato	Descripción
idDeclaracion	Char(30)	Número de Declaración aduanera
SubPartida	Numerico	Código de la partida arancelaria
idPuerto	Numérico	Código del país de origen de la mercadería
Cnatidad	Decimal(18,2)	Cantidad de la mercaduría declarada
Valor Unitario	Decimal(18,2)	Valor comercial de la mercadería
Valor CIF	Decimal(18,2)	Valor de la mercadería mas el flete y Seguro

Fuente: Elaboración Propia

2.3.5 Formato de los datos

Para realizar la minería de datos se debe transformar los datos tipo texto a datos numéricos. Los campos que se formatearon se describen en la tabla 11.

Tabla 11
Tabla de datos formateados

Campo	Tipo de Formato	Descripción
idPais	De Texto a Numérico	Código País de procedencia de la mercaduría a importar
idDeclaracion	De Texto a Numérico	Código de la declaración aduanera
idSubpartida	De Texto a Numérico	Código de la subpartida arancelaria

Fuente: Elaboración propia

2.4 MODELADO

En esta etapa se debe escoger el modelo que se va a utilizar para el proceso de minería de datos y realizar las pruebas de comprobación del modelo escogido.

2.4.1 Selección de la técnica del Modelado

Para determinar las declaraciones de importación que tienen valores anormales se va a utilizar K- means, dentro de este algoritmo se empleara la técnica de agrupamiento no supervisado. El agrupamiento se lo va a realizar seleccionando la subpartida arancelaria 8517120029 SMARTPHONE descrita en la tabla 6 Subpartidas con mayor valor CIF. Se selecciona esta subpartida ya que de acuerdo al histórico representa un producto donde se han presentado casos continuos de evasión de tributos.

El algoritmo se lo ejecutara relacionando las variables Cantidad, Valor Unitario de las mercaderías. Además se probara con las variables cantidad con Valor CIF.

2.4.2 Generación del plan de Prueba

A continuación se detallan los tres métodos por los cuales se identificaron los valores anormales sobre el precio unitario de las mercaderías

Outliers Detection (Zscore) : Este algoritmo identifica los valores que salen de lo normal. Este algoritmo utiliza la fórmula que permite la estandarización de una distribución. Esta fórmula se calcula restando al dato, la media de la distribución y

dividiendo el resultado por la desviación típica (distancia que tiene dicho dato respecto de la media)

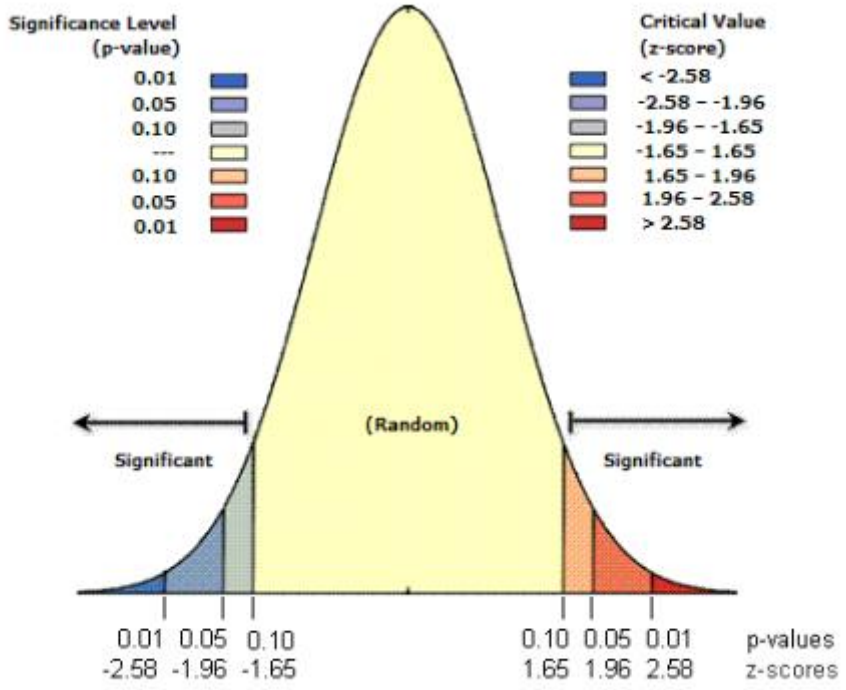


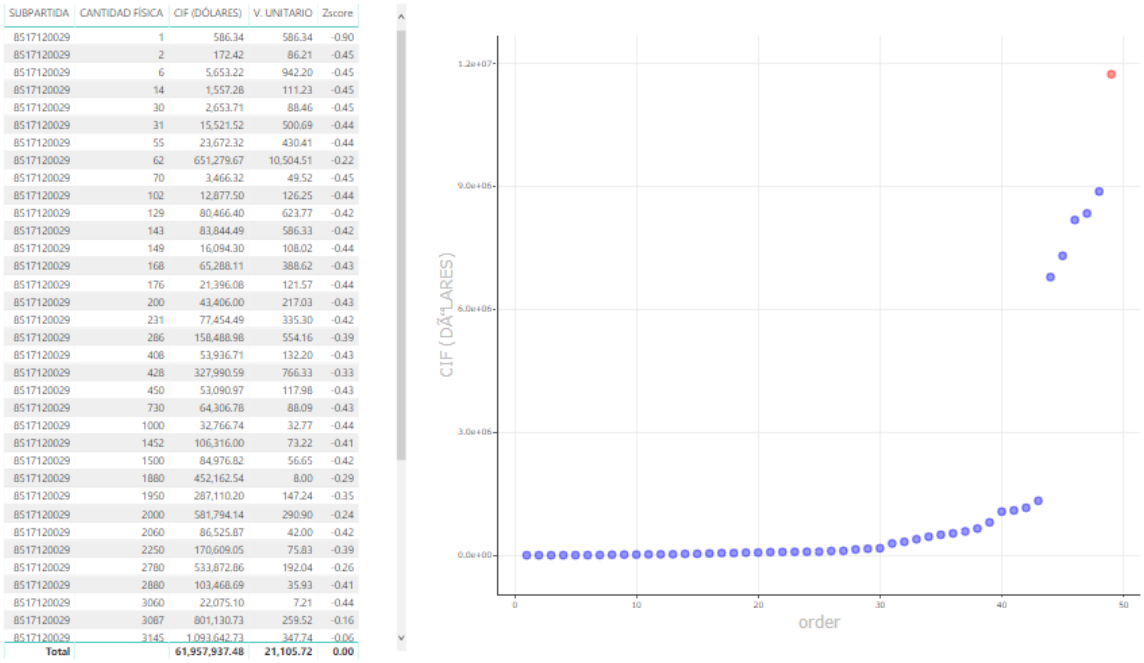
Figura 15: Método Zscore

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean

$\sigma =$ Standard Deviation

Este método detecta los valores atípicos, es decir aquellos valores que se encuentran muy alejados del valor de la media. En el grafico estos datos **z-scores** serían los que se encuentran en el lado derecho de la curva los mayores a 1.96 (encima de la media) y del lado izquierdo los menores a -1.96 (por debajo de la media).



SUBPARTIDA	CANTIDAD FÍSICA	CIF (DÓLARES)	V. UNITARIO	Zscore
8517120029	1	586.34	586.34	-0.90
8517120029	2	172.42	86.21	-0.45
8517120029	6	5,653.22	942.20	-0.45
8517120029	14	1,557.28	111.23	-0.45
8517120029	30	2,653.71	88.46	-0.45
8517120029	31	15,521.52	500.69	-0.44
8517120029	55	23,672.32	430.41	-0.44
8517120029	62	651,279.67	10,504.51	-0.22
8517120029	70	3,466.32	49.52	-0.45
8517120029	102	12,877.50	126.25	-0.44
8517120029	129	80,466.40	623.77	-0.42
8517120029	143	83,844.49	586.33	-0.42
8517120029	149	16,094.30	108.02	-0.44
8517120029	168	65,288.11	388.62	-0.43
8517120029	176	21,396.08	121.57	-0.44
8517120029	200	43,406.00	217.03	-0.43
8517120029	231	77,454.49	335.30	-0.42
8517120029	286	158,488.98	554.16	-0.39
8517120029	408	53,936.71	132.20	-0.43
8517120029	428	327,990.59	766.33	-0.33
8517120029	450	53,090.97	117.98	-0.43
8517120029	730	64,306.78	88.09	-0.43
8517120029	1000	32,766.74	32.77	-0.44
8517120029	1452	106,316.00	73.22	-0.41
8517120029	1500	84,976.82	56.65	-0.42
8517120029	1880	452,162.54	8.00	-0.29
8517120029	1950	287,110.20	147.24	-0.35
8517120029	2000	581,794.14	290.90	-0.24
8517120029	2060	86,525.87	42.00	-0.42
8517120029	2250	170,609.05	75.83	-0.39
8517120029	2780	533,872.86	192.04	-0.26
8517120029	2880	103,468.69	35.93	-0.41
8517120029	3060	22,075.10	7.21	-0.44
8517120029	3087	801,130.73	259.52	-0.16
8517120029	3145	1,093,642.73	347.74	-0.06
Total		**61,957,937.48**	**21,105.72**	**0.00**

Figura 16: Método Outlier Detection Z-score : Fuente : Elaboración propia

En la figura 16 se muestra un punto de color rojo, punto que se lo considera como un valor atípico que se encuentra por encima del valor media, es considerado atípico porque se aleja de la media de los demás valores. Para esta investigación se consideran los valores subvalorados, es decir que se encuentran por debajo del valor de la media.

Local Outliers Factor (LOF): El factor de valor atípico local se basa en un concepto de densidad local, donde la localidad viene dada por k vecinos más cercanos, cuya distancia se usa para estimar la densidad. Al comparar la densidad local de un objeto con las densidades locales de sus vecinos, uno puede identificar regiones de densidad similar y puntos que tienen una densidad sustancialmente menor que sus vecinos. Estos son considerados como valores atípicos.

La densidad local se estima por la distancia típica a la que se puede "alcanzar" un punto desde sus vecinos. La definición de "distancia de alcance" utilizada en la LOF es una medida adicional para producir resultados más estables dentro de los grupos.

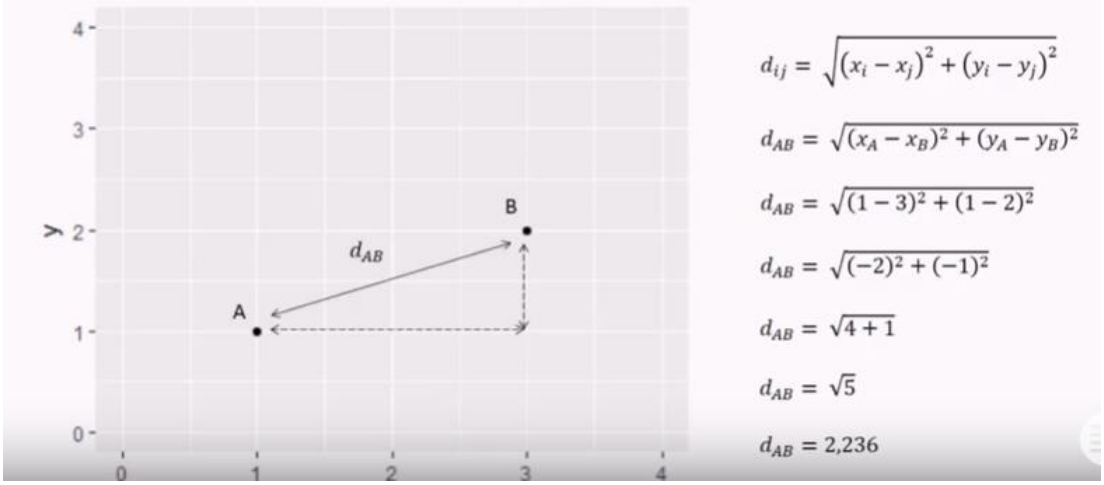


Figura 17: Método de cálculo de distancia entre valores

Este método utiliza la fórmula de distancia euclidiana (Teorema de Pitágoras), consiste en encontrar la distancia entre puntos y agrupar aquellos que se encuentren a menor distancia entre ellos. De esta forma se agrupan los datos formando varios clústeres.

Clustering Kmeans

Mediante este algoritmo agrupa los datos en grupos y en cada uno de ellos se encuentra un punto central denominado centroide. El objetivo en este proyecto es identificar valores anormales que se ubican fuera de cada uno de las agrupaciones.

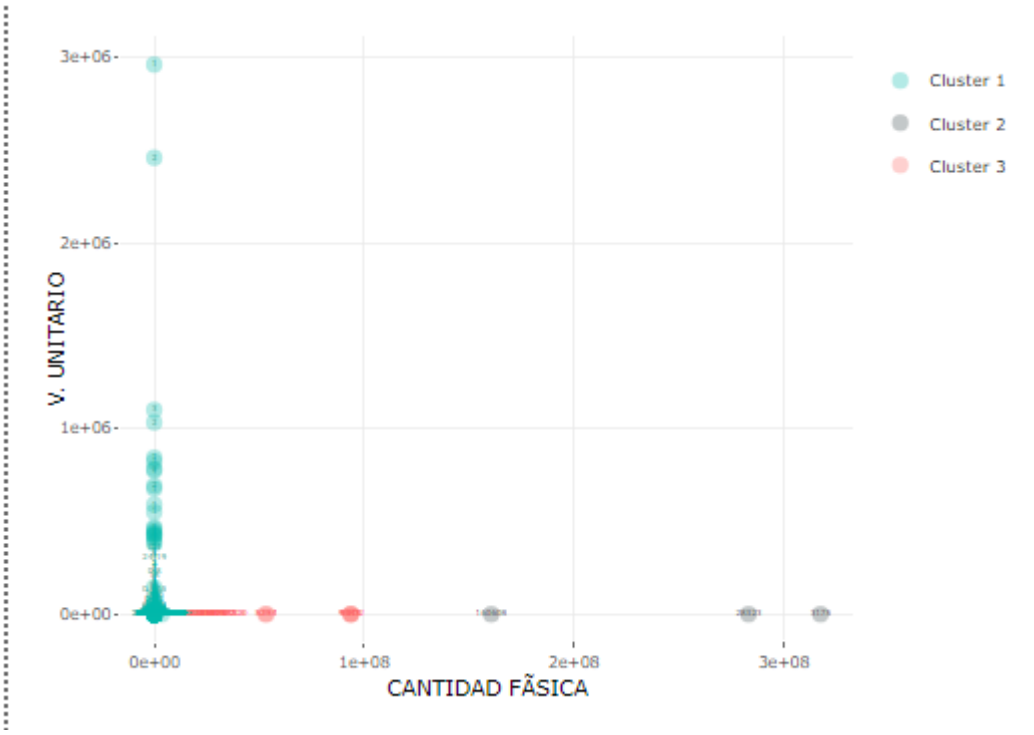


Figura 18: Algoritmo K-means: Fuente Elaboración Propia

En la figura 18 muestra los datos agrupados mediante la aplicación del algoritmo K-means, la agrupación consiste en encontrar grupos con características semejantes. En la gráfica se muestran tres grupos (Verde, Rojo y Plomo). Este tipo de algoritmo es no supervisado en donde no es necesario que los datos sean etiquetados y con su aplicación se determinan los grupos.

Este algoritmo permite identificar los valores atípicos en aquellas declaraciones que se hayan declarado con un valor menor a lo normal.

2.4.3 Construcción del Modelo

Objetivo: Identificar los valores de las mercancías que se encuentran muy debajo de los valores normales

El modelo a construir es en base al análisis de las cantidades de mercaderías versus el valor unitario. Las variables a ser consideradas en el análisis serán: País de procedencia, Valor CIF

Si en el modelo construido verificamos que existen valores unitarios demasiados bajos en comparación del cancroide entonces identificamos como posible caso de evasión de impuestos.

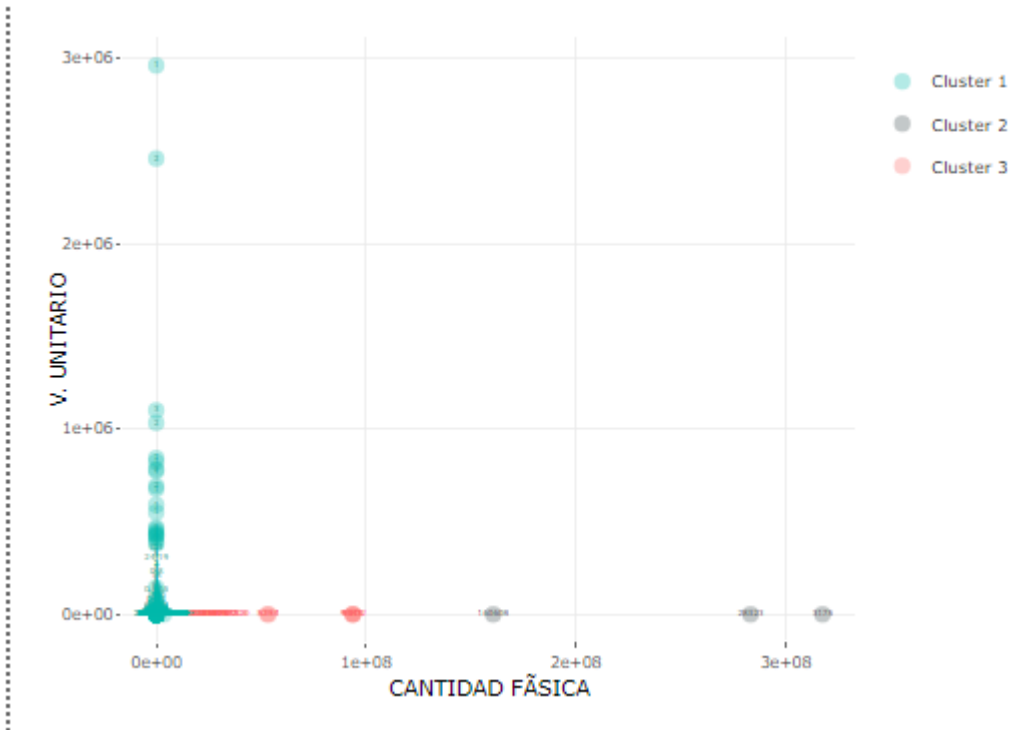


Figura 19: Construcción del Modelo de análisis. Fuente: Elaboración propia

2.4.4 Evaluación del Modelo

Se realizó un análisis utilizando los métodos Zscore, LOF y Clustering Kmeans , el modelo que mejores resultados se obtuvo fue utilizando Clustering con Kmeans.

Para la ejecución de los modelos se utilizó la Herramienta Power BI con algoritmos de detección de valores atípicos (outliers).

El objetivo del modelo es poder identificar valores anormales dentro de varias agrupaciones de datos.

Los datos de las declaraciones aduaneras fueron evaluados con distintas subpartidas arancelarias pero se enfocó en aquellas que poseían valores CIF más altos y se tomó una muestra de los Meses de Enero, Febrero y Marzo de 2019.

Se realizaron varias pruebas con varios números de clusters (5,4,3), la prueba con mejores resultado fue con 3 clusters. Se asignaron variables como país de procedencia de la mercancía, Valor CIF, Valor Unitario y Cantidad de Unidades Físicas.

2.5 Definición de Roles y Responsabilidades

En el desarrollo del Proyecto de Inteligencia de Negocios para el proceso de detección de casos posibles de evasión de impuestos bajo la modalidad de subvaloración aduanera se estableció un equipo de trabajo. En la Figura 19 se presenta la estructura del equipo de trabajo.

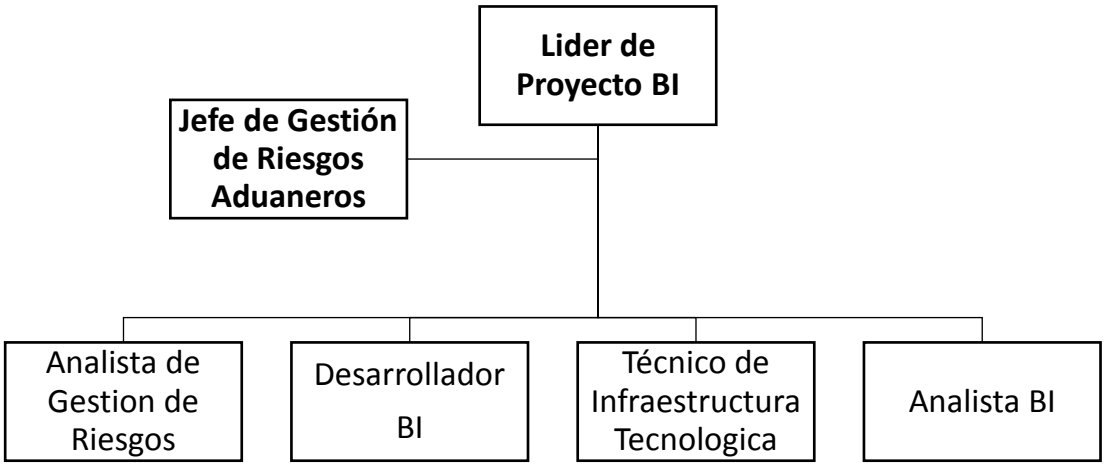


Figura 20: Equipo del Proyecto. Fuente: Elaboración Propia

Líder de Proyecto BI

Los roles y responsabilidades principales del Líder de BI son:

1. Dirigir y organizar al equipo del proyecto

2. Coordinar el cumplimiento del alcance y el tiempo establecido en el proyecto
3. Gestionar los riesgos del proyecto
4. Interacción directa con el Jefe de Riesgos del Servicio de Aduana

Jefe de Gestión de Riesgos Aduaneros

Los roles y responsabilidades principales del Jefe de Riesgos Aduaneros:

1. Dirigir y organizar a los técnicos de Valoración Aduanera
2. Aprobar los requerimientos y alcance del proyecto de Inteligencia de Negocios.
3. Confirmar los agentes y los importadores que se deberían sujetar al análisis del proceso de subvaloración.
4. Aprobar los informes de observaciones de valoración aduanera

Analista de Gestión de Riesgos

Los roles y responsabilidades principales del Técnico de Gestión de Riesgos Aduaneros:

1. Seleccionar los agentes y los importadores sujetos al análisis de valoración aduanera
2. Operar el sistema que ejecuta el proceso de análisis de valoración.
3. Revisar el sistema para confirmar si los requerimientos fueron levantados fueron acogidos
4. Realizar los informes de observación de valoración aduanera.
5. Aprobar los informes de observaciones de valoración aduanera

Técnico de Infraestructura Tecnológica

Los roles y responsabilidades principales del Especialista de Infraestructura Tecnología de Inteligencia de Negocios son:

1. Instalación de los equipos informáticos que conforman la infraestructura de Big Data
2. Instalación del software de Big Data.
3. Realizar la conectividad con la plataforma del Servicio de Aduana
4. Monitorear el rendimiento de los equipos informáticos instalados.

Analista BI

Los roles y responsabilidades principales del Analista de BI son:

1. Identificar los datos que servirán para el análisis
2. Diseño del modelo de datos a utilizar para el análisis
3. Realizar y coordinar las pruebas con el Analista de Gestión de riesgos
4. Coordinar con el Desarrollador las pruebas internas.

Desarrollador BI

Los roles y responsabilidades principales del Desarrollador BI son:

1. Aplicar del algoritmo de detección de valores atípicos en las declaraciones de importación
2. Diseño de la Visualización de los resultados del algoritmo.
3. Realizar las pruebas unitarias e integrales de lo desarrollado.
4. Documentar lo desarrollado y entregar al Líder del proyecto
5. Seguir los lineamientos del Líder de Proyecto.

2.6 Requerimientos Funcionales

Para el desarrollo de la solución propuesta se realizó el análisis de los requerimientos funcionales con el usuario final. Este proceso se realizó por medio de reuniones, en los cuales se detalló las funcionalidades que debe tener la solución para mejorar el proceso de detección de fraudes aduaneros utilizando la arquitectura de big data. A continuación el detalle de los requerimientos funcionales.

Requerimiento 001: **Segmentar Partidas Arancelarias por país de Origen**

Requerimiento que permitirá clasificar las subpartidas arancelarias por país de origen y determinar cuáles son utilizadas con mayor frecuencia en el proceso de importaciones. Mediante esta opción el usuario final podrá ingresar varios criterios de búsqueda siendo estos:

- Fechas de Inicio y Fin

- Distrito aduanero donde el agente realizó el trámite de salida de las mercaderías importadas
- Por agente aduanero que ejecutó el proceso de importación.

Tabla 12 Requerimiento Funcional 01

<table>
<tr><td colspan="2">Requerimiento 01 : Segmentar Partidas Arancelarias por país de Origen</td></tr>
<tr><td colspan="2">Como : Técnico de Control Posterior
Quiero : Clasificar las partidas arancelarias de importación por País de procedencia
Para poder : Identificar las subpartidas más frecuentes</td></tr>
<tr><td rowspan="4">Validación

• El analista puede filtrar el informe por fechas (inicio y fin).
• El analista puede filtrar por Distrito Aduanero de Destino (Guayaquil, Quito, Manta, Esmeraldas, Puerto Bolívar, Tulcán, Cuenca)
• El analista podrá filtrar el informe por Agente aduanero que gestionó la importación</td><td>Valor : 150</td></tr>
<tr><td>Prioridad : 1</td></tr>
<tr><td>Estimación: 16 h</td></tr>
<tr><td>Calidad : INVEST</td></tr>
</table>

Fuente: Elaboración Propia

Especificaciones Funcionales:

- Esta opción solo está habilitada para los técnicos que designó el Director de Gestión de Riesgos
- Los campos obligatorios para la emisión del informe son la fecha de Inicio y Fecha Fin, los demás campos son opcionales
- El informe tendrá los siguientes campos de salida
    - Subpartida Arancelaria
    - Descripción de la Subpartida arancelaria
    - País de Origen de la Mercadería
    - Agente de Aduana que realizó el trámite de importación
    - Número de declaraciones aduaneras por agente de aduana
    - Cantidad Física de Mercaderías Importadas
    - Valor CIF de las declaraciones

Requerimiento 02: Detección de Valores Atípicos en las declaraciones de Importación

Este requerimiento permite identificar aquellas declaraciones de importación cuyos valores unitarios de las mercaderías estén por debajo de su valor real.

La información a ser procesada debe estar clasificada por subpartida arancelaria y por puerto de origen, ya que el valor de la mercadería varia por país y además el arancel difiere por cada subpartida arancelaria

Tabla 13 Requerimiento Funcional 02

<table>
<tr><td colspan="2">Requerimiento de Usuario : Detectar Outliers – Valores Unitarios de Mercaderías Atípicos</td></tr>
<tr><td colspan="2">Como : Jefe de Control Posterior
Quiero : Identificar Valores de las mercancías que exceden mayormente su valor promedio
Para poder : Determinar las declaraciones aduaneras que pretenden declarar un valor menor del pago de los tributos</td></tr>
<tr><td rowspan="4">Validación
Se podrá filtrar el informe por fechas (inicio y fin).
Así también por País de Origen
Además se podrá filtrar por Partida Arancelaria</td><td>Valor : 200</td></tr>
<tr><td>Prioridad : 1</td></tr>
<tr><td>Estimación: 16 h</td></tr>
<tr><td>Calidad : INVEST</td></tr>
</table>

Fuente: Elaboración Propia

Requerimiento 003: Visualizar las declaraciones por subvaloración aduanera por país de Origen

Mediante esta opción el usuario final podrá visualizar las declaraciones aduaneras por país de origen que se identificaron con subvaloración aduanera y luego de esto realizar los informes y solicitar a los involucrados los respectivos documentos de soporte o la aplicación de la reliquidación aduanera y las multas correspondientes.

Tabla 14 Requerimiento Funcional 03

<table>
<tr><td colspan="2">Requerimiento 03 : Visualizar en detalle las declaraciones aduaneras con valores unitarios menores al promedio por país de Origen</td></tr>
<tr><td colspan="2">Como : Técnico de Valoración Aduanera
Quiero : Revisar información detallada de las declaraciones aduaneras subvaloradas
Para poder : Perfilar riesgos sobre las importaciones analizadas</td></tr>
<tr><td rowspan="4">Validación

En el detalle de las declaraciones se deberá mostrar:

información general:

• Nombre del Importador
• Dirección del Importador
• Teléfono
• Agente Aduanero
• Dirección del Agente
Información de Carga :

• País de procedencia
• Consignatario
• Valor CIF
• Cantidad Total de Unidades
Información Ítems :

• Partida Arancelaria
• Descripción de la Mercadería
• Cantidad
• Valor Unitario
• Clase
• Modelo</td><td>Valor : 150</td></tr>
<tr><td>Prioridad : 1</td></tr>
<tr><td>Estimación: 16 h</td></tr>
<tr><td>Calidad : INVEST</td></tr>
</table>

Requerimiento 04: Visualizar declaraciones aduaneras que han caído en el delito de subvaloración aduanera por Agente Aduanero

Esta opción permitirá visualizar el historial de las declaraciones procesados por un agente aduanero, con el fin de investigar en mayor detalle y confirmar si pudieran existir otros casos en los cuales se haya declarado con un valor menor al valor base.

Tabla 15 Requerimiento Funcional 04

<table>
<tr><td colspan="2">Requerimiento 04 : Visualizar en detalle las declaraciones aduaneras por Agente Aduanero</td></tr>
<tr><td colspan="2">Como : Técnico de Control Posterior
Quiero : Clasificar las partidas arancelarias de importación por Agente Aduanero
Para poder : Identificar si existen más casos de subvaloración por un mismo agente aduanero</td></tr>
<tr><td rowspan="4">Validación

• Agentes aduaneros activos
• Declaraciones aduaneras procesadas hace 5 años atrás</td><td>Valor : 100</td></tr>
<tr><td>Prioridad : 2</td></tr>
<tr><td>Estimación: 8 h</td></tr>
<tr><td>Calidad : INVEST</td></tr>
</table>

Fuente: Elaboración Propia

Requerimiento 05: Categorización de Agente Aduanero

Mediante esta opción el analista de Gestión de Riesgos podrá actualizar la categoría del agente Aduanero. Las categorías a ser asignadas son :

1. Alto Riesgo
2. Medio Riesgo
3. Bajo Riesgo

Tabla 16 Requerimiento Funcional 05

<table>
<tr><td colspan="2">Requerimiento de Usuario : Categorizar a los Agentes Aduaneros por declaraciones aduaneras Subvaloradas</td></tr>
<tr><td colspan="2">Como : Jefe de Control Posterior
Quiero : Visualizar los agentes aduaneros que incurrieron en casos de subvaloración
Para poder : Asignar una categoría y sus trámites se les asigne aforo físico</td></tr>
<tr><td rowspan="4">Validación

• Agentes aduaneros activos
• Declaraciones aduaneras procesadas hace 5 años atrás</td><td>Valor : 150</td></tr>
<tr><td>Prioridad : 1</td></tr>
<tr><td>Estimación: 8 h</td></tr>
<tr><td>Calidad : INVEST</td></tr>
</table>

Fuente: Elaboración Propia

Requerimiento 06: Reporte de Declaraciones a reliquidar

Esto opción permite visualizar las declaraciones aduaneras que hayan incurrido en el delito de subvaloración aduanera. El objetivo es emitir un informe mediante el cual se reliquiden aquellas declaraciones con los valores correctos y el importador tenga que pagar el tributo faltante.

Tabla 17 Requerimiento Funcional 06

<table>
<tr><td colspan="2">Requerimiento de Usuario : Reporte de Declaraciones a reliquidar Tributos y Establecer Multas</td></tr>
<tr><td colspan="2">Como : Analista Financiero
Quiero : Identificar las declaraciones que incurrieron en casos de Subvaloración
Para poder : Liquidar nuevamente las declaraciones y establecer las multas por tratar de ocasionar un perjuicio económico al estado</td></tr>
<tr><td rowspan="3">Validación

• Declaraciones aduaneras ingresadas hace 5 años atrás</td><td>Valor : 200</td></tr>
<tr><td>Prioridad : 1</td></tr>
<tr><td>Estimación: 16 h</td></tr>
<tr><td></td><td>Calidad : INVEST</td></tr>
</table>

Fuente: Elaboración Propia

2.7 Requerimientos No Funcionales

Los requerimientos no funcionales utilizados para el desarrollo de la propuesta se detallan en las Tablas 14,15,16,17, estos recursos corresponden a hardware y software.

HARDWARE

Tabla 18 Servidor de Base de Datos Transaccional

TIPO	**RACK**
FUNCIONALIDAD	BASE DE DATOS
	PROCESADOR
TIPO	RISC 64 BITS
NUMERO DE CORES	6

VELOCIDAD DEL PROCESADOR	3.3 GHZ
MEMORIA RAM	
INSTALADA	72 Gb
VELOCIDAD DE LA MEMORIA	1033 MHZ
ALMACENAMIENTO	
CAPACIDAD	584 GB
TIPO DE DISCO	SAS
VELOCIDAD	15,000 RPM
INSTALADOS	6

Fuente: Elaboración Propia

SERVIDOR HADOOP

Tabla 19 Servidor Base NoSql

TIPO	**RACK**
FUNCIONALIDAD	BASE DE DATOS HADOOP
PROCESADOR	
MARCA Y TIPO DE PROCESADOR	INTEL XEON 5600 QUAD CORE
SISTEMA OPERATIVO	LINUX 64 BITS
MEMORIA RAM	
CANTIDAD	16 GB
VELOCIDAD DE LA MEMORIA	1333MHZ
ALMACENAMIENTO	
CAPACIDAD TOTAL	292 GB
VELOCIDAD	10000 RPM

Fuente: Elaboración Propia

ALMACENAMIENTO NAS

Tabla 20 Almacenamiento NAS

Características	
Controlador	Dual (activo-activo)
Tipo y velocidad	Intel® 2.2 GHz Celeron
Procesadores	2
Memoria	4 Gb
Velocidad Puertos Fibra Canal	4 - 4 Gbps
Velocidad Puerto Ethernet	4 -1 Gbps
Almacenamiento	
Capacidad	512 TB
Discos internos	32

Fuente: Elaboración Propia

SWITCH LAYER 3

Tabla 21 Equipo de comunicación

Características	
ESCALABLE	CHASIS 7 SLOTS PARA PODER EXPANDIRSE
TARJETA SUPERVISORA	1 PRINCIPAL Y 1 REDUNDANTE.
MEMORIA	128 Megabytes
PUERTOS DE FIBRA Y COBRE	2 – 8

Fuente: Elaboración propia

SOFTWARE

Tabla 22 Requerimiento de Software

Item	**Descripción**
ORACLE	
Producto	Oracle DB Enterprise Edition
Versión	12c
Plataforma	Red Hat Linux
Detalle	Base de Datos Transaccional donde se encuentran registradas las declaraciones aduaneras
TOAD	
Producto	Toad para Oracle
Version	Compatible con la versión 12c de Oracle
Plataforma SO	Windows Server 2012
Detalle	Utilizado para la consulta de la Base de Datos de Oracle
SISTEMA OPERATIVO	
Producto	Sistema Operativo Ubuntu

Versión	16.04
Plataforma	Linux
Detalle	Sistema operativo donde se encuentran instalados los componentes de la arquitectura de Big Data el resto de componentes de software
MYSQL	
Producto	Base de datos Mysql
Versión	8.0
Plataforma	Linux Ubuntu
Detalle	Base de datos Relacional donde se almacenaran los datos de las declaraciones aduaneras ya transformadas para el análisis
SQOOP	
Producto	Sqoop
Versión	1.4.6
Plataforma	Linux Ubuntu
Detalle	Software para extraer los datos de Mysql a Hadoop
HADOOP	
Producto	Hadoop
Versión	2.9
Plataforma	Linux Ubuntu
Detalle	Base de datos NO SQL donde se almacenaran las declaraciones aduaneras para el análisis de subvaloración
Power BI	
Producto	Power BI
Versión	2.59
Plataforma	Windows
Detalle	Programa para el análisis estadístico y de minería de datos
BIZAGI	
Producto	Bizagi
Versión	3.1
Plataforma	Windows 64 bits
Detalle	Programa para el modelado de procesos
VISUAL STUDIO	
Producto	Visual Studio
Versión	2015
Plataforma	Windows
Detalle	Programa para construir el ETL
POWER DESIGNER	
Producto	Power Designer
Versión	16.
Plataforma	Windows
Detalle	Programa para construir el modelo de datos de la Solución

Fuente : Elaboración propia

2.8 Definición de los componentes que conforman el núcleo de la arquitectura tecnológica

Los componentes que conforman la arquitectura tecnológica se describen en la figura 20. El origen es de la Base de Datos Transaccional, base de datos en Oracle. Esta base se alimenta de las declaraciones que transmiten los operadores de comercio exterior a la aduana. Los datos de las declaraciones aduaneras se encuentran almacenados en un equipo de almacenamiento NAS.Por medio de un proceso de extracción se almacenan los datos en una Base de Datos NoSQL y por medio de Power BI se visualizan los datos.

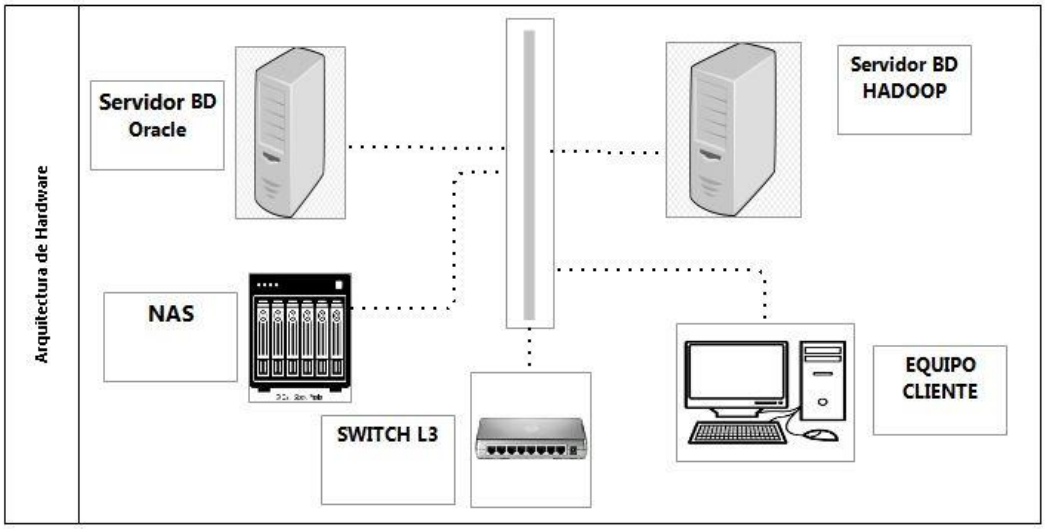


Figura 21: Componente de Hardware. Fuente : Elaboración Propia

La arquitectura de software utilizada para el desarrollo de la propuesta de arquitectura de Big Data para la auditoria de los trámites de importación en el Servicio de Aduana. En la Figura 21 se muestra el detalle

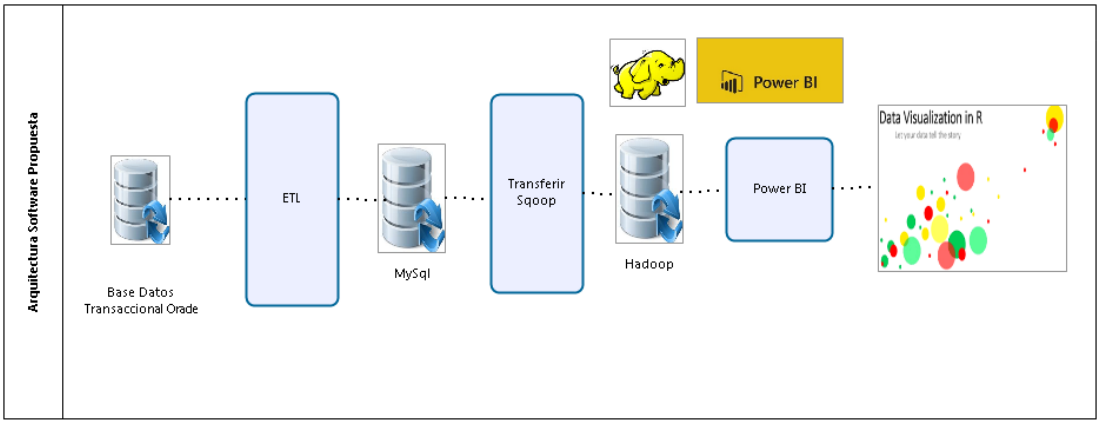


Figura 22: Componentes de Software. Fuente: Elaboración Propia

En los componentes de software podemos detallar:

1. **Base de Datos Oracle** : Fuente de datos transaccional
2. **Base de Datos Mysql**: Base donde se almacenaran los datos extraídos de Oracle
3. **Sqoop**: Software utilizado para transferir datos de Mysql a Hadoop
4. **Base de datos Hadoop**: Base de datos donde se realizara el proceso de análisis de los datos.
5. **Power BI:** Visualizador de los resultados del proceso de análisis

2.9 Diseño de modelo de datos

En este proceso se construye el modelo de datos para almacenar los satos extraídos de la Base de datos transaccionales y volcados a la Base de datos NoSql Hadoop. Sobre esta base se realizó el proceso de análisis y mediante el cual se determina los posibles casos de evasión de impuestos bajo la modalidad de subvaloración.

En la Figura 23 se muestran las tablas: TB_DECLARACIONES, TB_CARGA, TB_MERCADERIAS. Seguidamente se tiene el detalle de lo que contiene cada tabla y sus respectivos campos con su descripción.

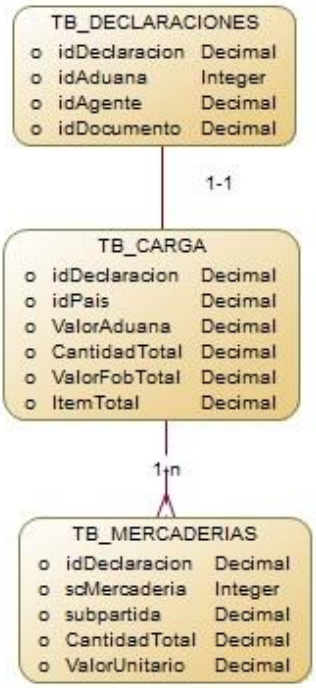


Figura 23: Modelo de Datos Proyecto Aduana. Fuente: Elaboración propia

TB_DECLARACIONES: En esta tabla se almacenan los datos referentes a las declaraciones aduaneras de importación. Su clave primaria es el idDeclaracion, este código es secuencial y asignado por la Aduana, la información enviada por los operadores de comercio son identificadas a través de este código y es asignado luego de ser validada la información.

IdDeclaracion: Tipo Numérico. Este campo identifica la declaración aduanera. Campo asignado por la aduana luego de ser validado y aceptado.

IdAduana: Tipo Numérico. Este campo identifica la aduana donde se realiza el trámite de desaduanizar las mercaderías de importación.

IdAgente: Tipo Numérico. Este campo identifica el agente aduanero que realiza el trámite de importación. Es el representante del importador ante la aduana.

IdDocumento: Tipo Numérico. Este campo identifica al agente de aduana por medio de la cedula de identidad / Pasaporte.

TB_CARGA. En esta tabla se almacenan los datos de la carga. La información registrada en esta tabla es totalizada del detalle de mercaderías que declara el importador.

IdDeclaracion: Tipo Numérico. Este campo identifica la declaración aduanera. Campo asignado por la aduana luego de ser validado y aceptado.

IdPais: Tipo Numérico. Campo que identifica el país de procedencia de la carga. Este campo es alfanumérico pero para el proceso de minería se lo transformo en código numérico.

ValorAduana. Tipo Decimal. Campo donde se registrar el valor [3]FOB de la mercancía agregado el valor el Flete y el Valor del Seguro. Sobre este valor se calculan los tributos aduaneros que deben declarar a la Aduana.

CantidaTotal: Tipo Decimal. En este campo se almacenan los valores totales de las cantidades de las mercancías que el importador declara a la Aduana.

Valor FOBTotal: Tipo Decimal. En este campo se registran los valores totales del precio de las mercancías a ser importadas.

ItemTotal: Tipo Decimal. Campo donde se almacena el número total de ítems de mercaderías que se encuentran en una declaración aduanera.

TB_MERCADERIAS. En esta tabla se almacena el detalle de las mercaderías declaradas para el proceso de importación. Las mercaderías se las identifica mediante un código, este código se lo conoce como subpartida arancelaria. A continuación el detalle de los campos que contiene esta tabla:

[3] FOB. Valor de la mercadería sin incluir el valor del flete

IdDeclaracion: Tipo Numérico. Este campo identifica la declaración aduanera. Campo asignado por la aduana luego de ser validado y aceptado.

scMercaderia. Tipo Numérico**.** Este campo es un secuencial de los ítems que constan en una declaración aduanera.

subPartida. Tipo Numérico. En este campo se identifica la mercadería. La codificación esta normada por la OMA[4]. En base a este código se calcula el valor del arancel a tributar por parte del Importador.

Cantidad. Tipo Decimal, Valor de la mercadería sin flete y seguro

2.10 Diseño de ETL

Para el análisis de la información previamente se ejecutó el proceso de extracción de datos de la Base Transaccional. La Base transaccional esta creada en Oracle. Se esta base se extrajeron los datos de las declaraciones aduaneras, información de carga y la información de las mercaderías. Así mismo se realizó el proceso de limpieza de los datos que fueron detallados en el capítulo 2 en la fase de preparación de los datos. Se continuó con la conversión de datos, algunos campos debieron ser convertidos de texto a numérico para poder realizar el proceso de análisis de agrupamiento.

[4] OMA: Organización Mundial de Adunas



Para el proceso de extracción se seleccionaron las tablas de Mercaderías, Declaraciones y Carga. En la figura 24 se muestra la tabla de Mercaderías, datos extraídos de la base de datos transaccional BD Oracle.

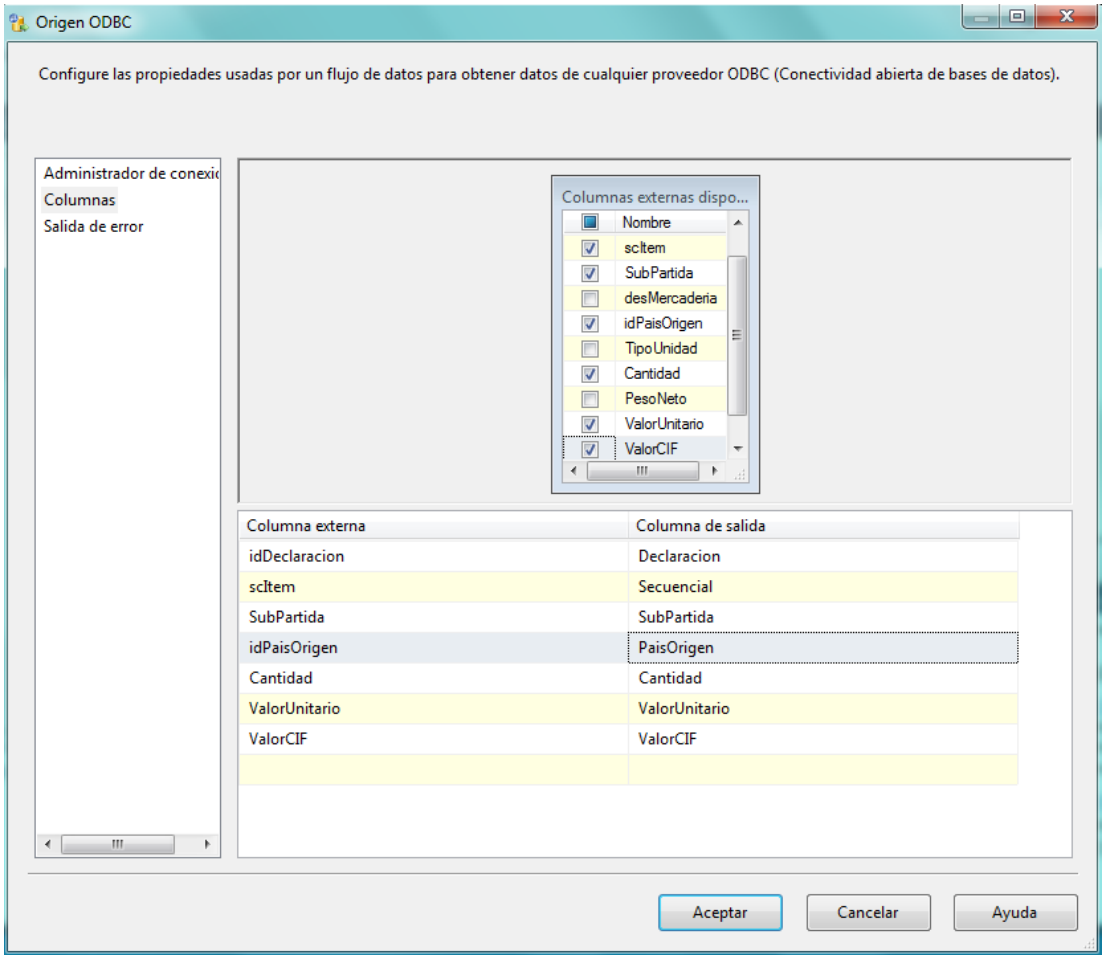


Figura 24: Extracción de la tabla de Mercaderías

En la figura 25 se muestra la tabla de Declaraciones, datos extraídos de la base de datos transaccional BD Oracle.

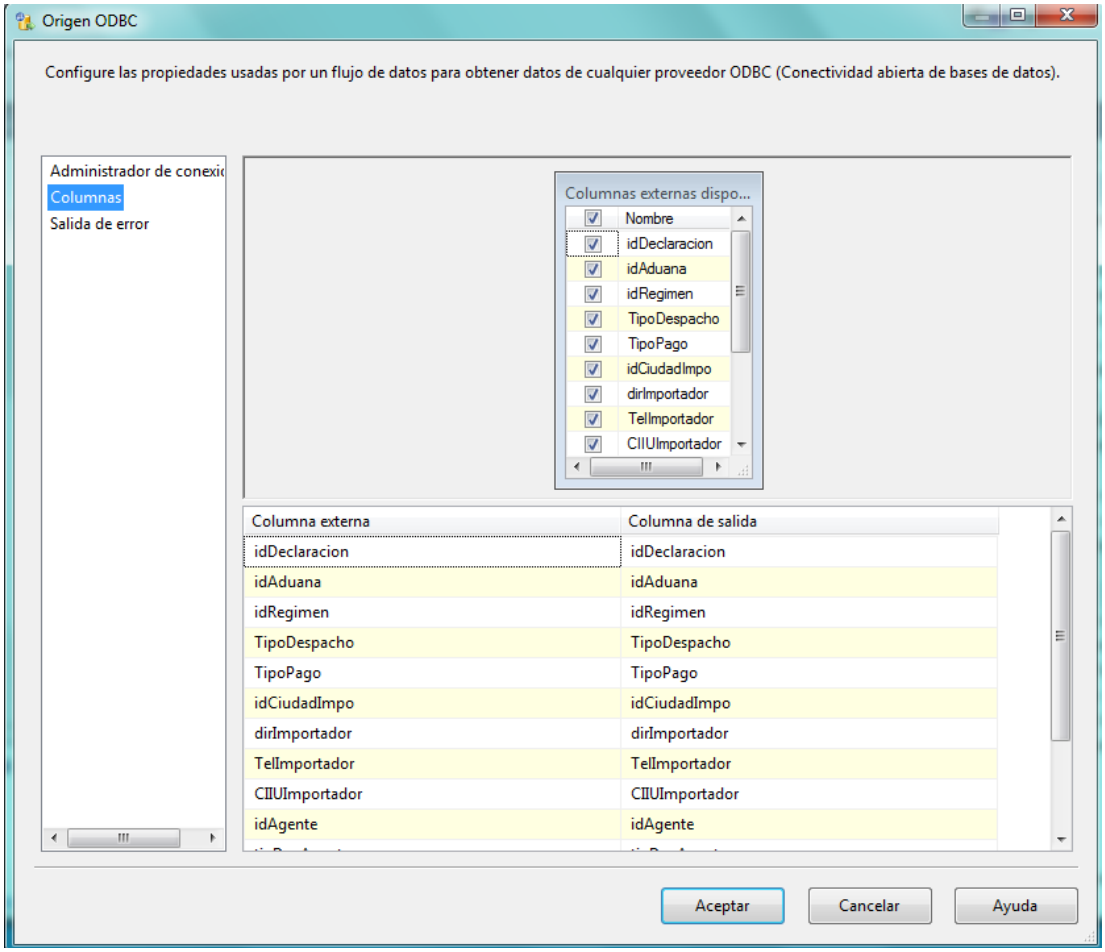


Figura 25: Tabla de Declaraciones

En la figura 26 se muestra la tabla de Carga, datos extraídos de la base de datos transaccional BD Oracle.

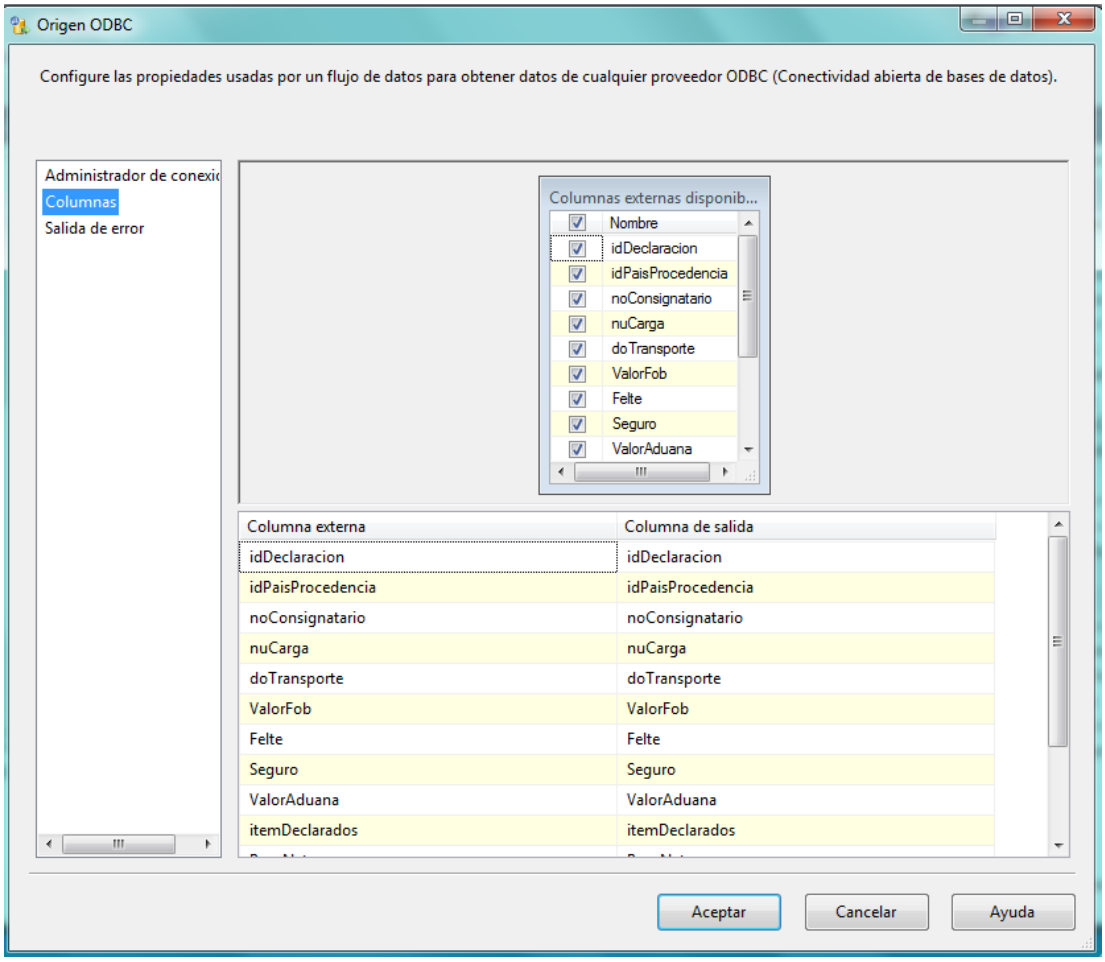


Figura 26: Tabla de Caga

Los datos fueron extraidos de BD Oracle y almacenados en BD Mysql, estee proceso fue realizado con la implementación de ETL desarrollado en Visual Studio 2015. A continuación se detalla el proceso de almacenamiento de las tablas en Mysql. En la figura 27 se muestra el proceso de extracción y almacenamiento de la tabla de Mercaderías.

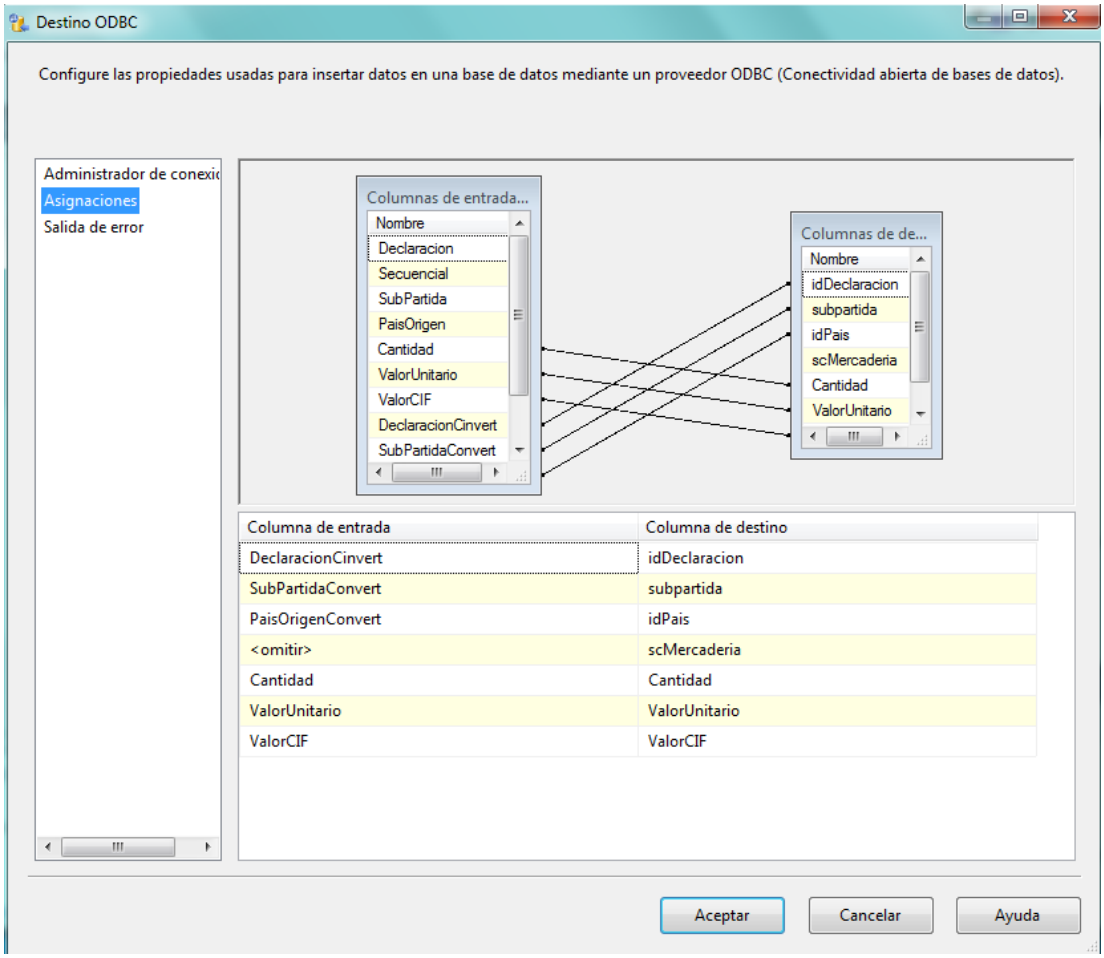


Figura 27: Tabla de Mercaderías Mysql

En la figura 28 se muestra el proceso de extracción y almacenamiento de la tabla de Declaraciones.

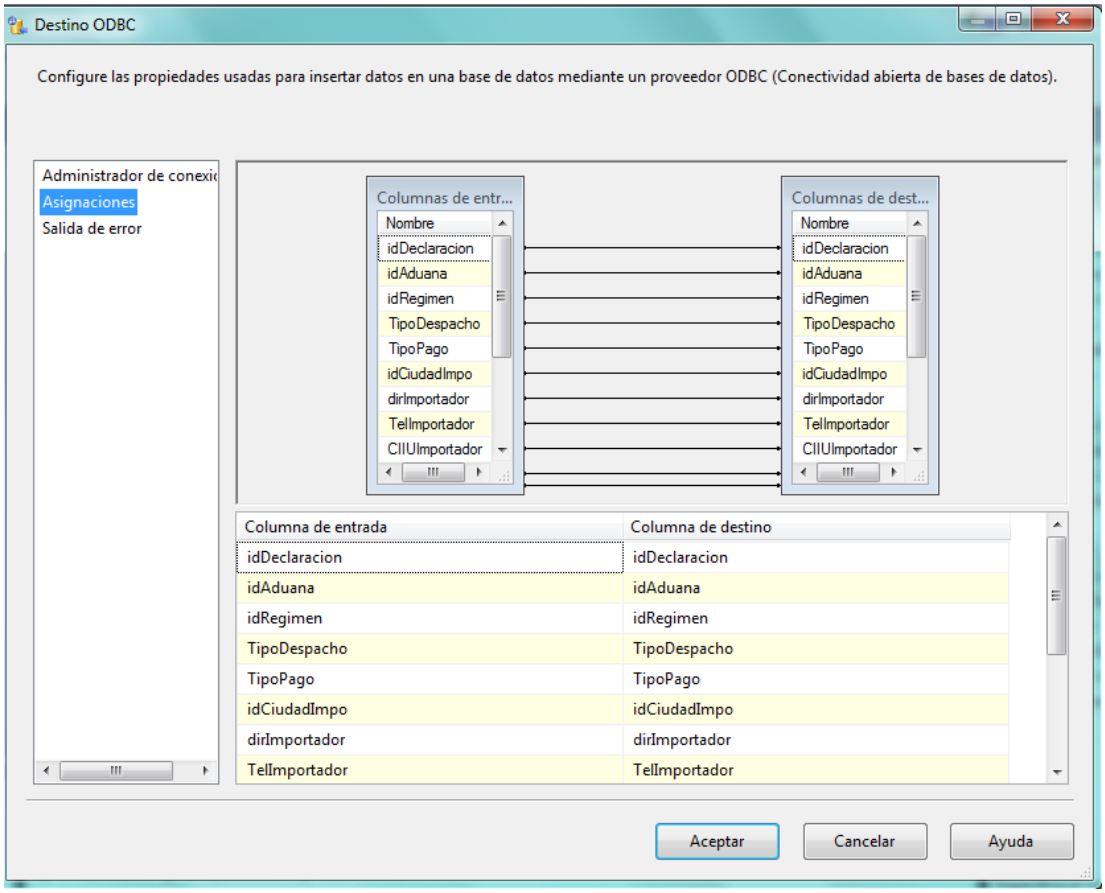


Figura 28: Tabla de Declaraciones Mysql

En la figura 29 se muestra el proceso de extracción y almacenamiento de la tabla de Carga.

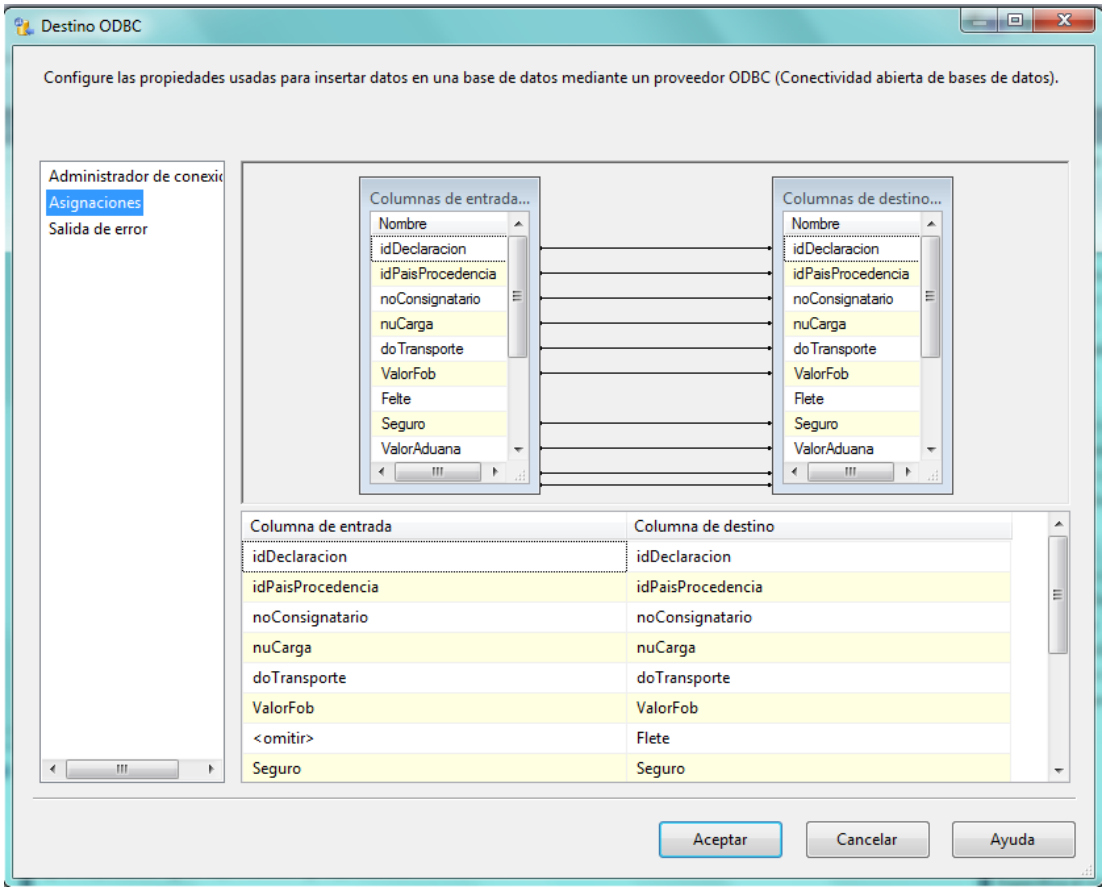


Figura 29: Tabla de Carga Mysql

2.11 Desarrollo del módulo de visualización

El proceso de visualización identifica los posibles casos de evasión de impuestos por la modalidad de subvaloración aduanera. El proceso de detección de posibles casos de evasión de impuestos es identificar aquellas declaraciones aduaneras que hayan sido declarados con valores muy por debajo de los valores normales.

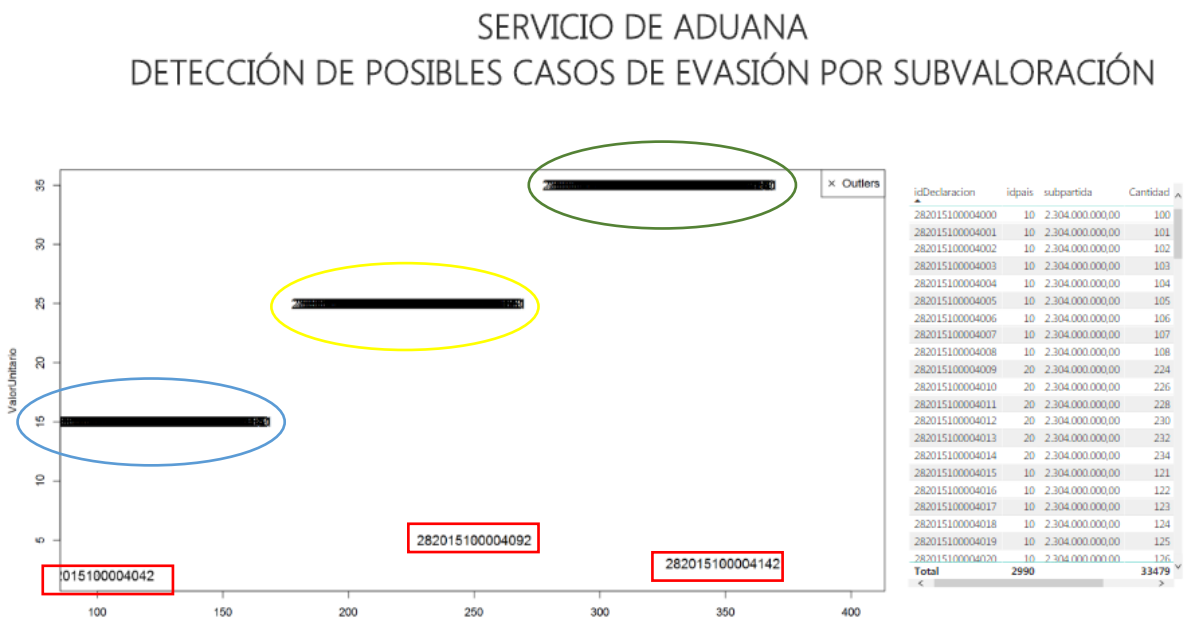


Figura 30: Detección de casos de evasión de Impuestos por Subvaloración

Se ejecutó el proceso de análisis y por medio de la aplicación del algoritmo de k-means, el mismo que identificó tres grupos. La agrupación la realizó por el Puerto de Origen de la mercadería. En la figura 23 muestra el análisis de los datos donde se detectaron tres declaraciones de importación con subvaloración aduanera (enmarcados en rojo). Luego de la identificación el técnico de intervención deberá realizar un informe de hallazgos para proceder con la reliquidación aduanera y la multa de comprobarse el delito.

Para la identificación de posibles casos de evasión de impuestos se utilizó Power BI con componente Detection Outliers, este componente agrupa los datos y analiza el comportamiento de los datos y genera un patrón, luego de la generación del patrón procede a la detección de valores atípicos, siendo estos para nuestra investigación los casos de subvaloración aduanera.

SERVICIO DE ADUANA
DETECCIÓN DE POSIBLES CASOS DE SUBVALORACION ADUANERA

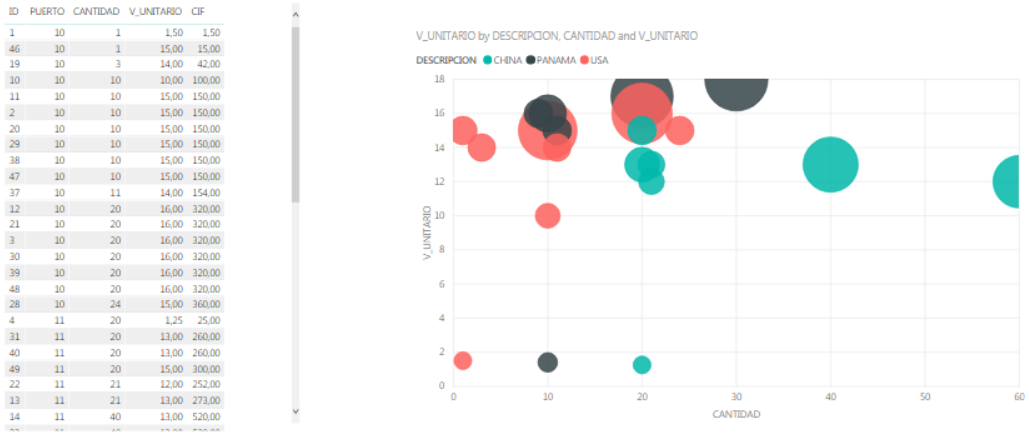


Figura 31: Detección de Valores Atípicos

En la figura 31 se muestra otra vista de los valores que no muestran comportamiento atípicos, valores que se encuentran muy separados de los valores normales.

CAPITULO 3.

3 Validación de los Resultados

En este capítulo se presentara la forma en que se verificó los resultados de la propuesta de implementación de una arquitectura de big data para las auditorias de los trámites de importación en el Servicio de Aduana.

Se realizó la validación de los resultados de lo desarrollado por el método de juicio de expertos. Luego de la validación se tabularon los resultados de la encuesta.

Esta validación permitirá identificar si la solución propuesta es viable de ser implementada.

3.1 Esquema de Validación

Para la validación de los resultados se siguió el siguiente esquema:

- Validación por Juicios de expertos

Validación por juicios de expertos

Para este método se realizó una encuesta a tres expertos en inteligencia de negocios. Para la consulta se entregaron cuestionarios con cinco preguntas sobre la propuesta de implementación de una arquitectura de big data para la auditoria de los trámites de importación en el Servicio de Aduana., los expertos analizaron y emitieron las respuestas correspondientes. Posteriormente fueron receptados los cuestionarios y las respuestas fueron tabuladas. En la tabla 23 se muestran los datos de los expertos que fueron encuestados.

Tabla 23 Expertos encuestados

No	Institución	Grado Académico	Cargo	Años Experiencia
1	Aduana	Master	Jefe de Mejora Continua	9
2	Universidad de Guayaquil	Master	Docente	10
3	IBM	Master	Jefe de Proyectos BI	15

Fuente: Elaboración propia

3.2 Aplicación y descripción de los resultados

Los resultados obtenidos fueron por medio de la entrega de cuestionarios a los expertos para su validación. Se formularon 5 preguntas, para las respuestas se generaron 5 opciones de grado de acuerdo. 1- muy en desacuerdo, 2-en desacuerdo ,3-en acuerdo,4-Medianamente de acuerdo,4-Muy de acuerdo.

En la tabla 24 se detallan las preguntas a ser consultadas a los expertos.

Tabla 24 Preguntas a expertos

		GRADO DE ACUERDO				
PREGUNTA	DESCRIPCION	1	2	3	4	5
1	ESTA USTED DE ACUERDO CON LA SOLUCIÓN PROPUESTA?					
2	LA INTERFASE PRESENTADA ES FACIL DE VISUALIZAR?					
3	ES FACITIBLE LA IMPLEMENTACIÓN DE ESTA SOLUCIÓN EN ADUANA?					
3	CON ESTA PROPUESTA SE IDENTIFICARAN MAS CASOS DE EVASIÓN DE IMPUESTOS?					
5	CONSIDERA USTED QUE SE PODRIAN AUMENTAR LAS RECAUDACIONES CON LA SOLUCIÓN PROPUESTA					

Fuente: Elaboración propia

En la figura 32 se muestra la tabulación de los resultados obtenidos a través de las respuestas realizadas por los expertos.



Figura 32 Tabulación Juicio de Expertos

De los datos obtenidos se puede concluir que los expertos están de acuerdo con la solución propuesta y recomendarían su implementación a mediano plazo.

3.3 Impacto

Con la implementación de la propuesta de arquitectura de big data para las auditorias en los trámites de importación en el servicio de aduana tendría el siguiente impacto:

1. **Operativo:** Los directivos del Departamento de control posterior van a poder identificar más casos de evasión de impuestos bajo la modalidad de subvaloración aduanera y poder establecer estrategias para minimizar este tipo de contravención.
2. **Económico**: Con la reducción de casos de evasión de impuestos la recaudación aduanera por concepto de pago de tributos se incrementara ya que los importadores deberán declarar de acuerdo a la ley.
3. **Social**: Al aumentar los valores en las recaudaciones existirá mayores recursos para ser redistribuidos en la construcción de viviendas, escuelas y hospitales beneficiando a la comunidad del país.

4. **Tecnológico**: La institución contraria con herramientas actualizadas para el proceso de análisis de grandes cantidades de datos que servirían para generar información efectiva para la toma de decisiones en todas las áreas del Servicio de Aduana.

4 Conclusiones

La revisión de las teorías y conceptos en la implementación de una solución de inteligencia de negocios permitieron identificar las herramientas y los componentes de hardware y de software para dar solución a la problemática presentada en el Servicio de Aduana.

Se realizó reuniones con los técnicos de la Dirección de Mejora Continua y Tecnología de la información y se revisó la arquitectura actual que el Servicio de Aduana cuenta para la detección de posibles casos de evasión bajo la modalidad de subvaloración aduanera.

La metodología CRISP-DM empleada en el desarrollo de la solución permitió estructurar y organizar las tareas a ejecutarse, mediante la identificación de riesgos, establecer los recursos a ser utilizados, levantar los requerimientos con la ayudad de los expertos del negocios y seleccionar, preparar, modelar y explotar los datos para proporcionar información para la toma de decisiones de la institución. El adoptar el esta metodología garantizó el éxito en la implementación del proyecto, comprobando porque esta metodología es recomendada por expertos para la minería de los datos.

Esta propuesta de solución difiere al proceso actual implementado el Servicio de aduana, esta solución procesa los datos por lotes, es decir que los agrupa e identifica los valores atípicos que en este caso son los casos de subvaloración por lo tanto el tiempo en el proceso es menor. El proceso actual procesa registro por registro y esto hace que a mayor cantidad de declaraciones evaluadas mayor sea el tiempo en el proceso.

Implementando la solución en Power BI se pudieron determinar casos de evasión de impuestos donde se agruparon los datos en cluster (3) y se obtuvieron valores por debajo de los valores normales de las mercaderías procesadas.

5 Recomendaciones

Para un mejor análisis de los casos de evasión de impuestos se recomendaría agrupar los datos en tres grupos (Kmeans = 3), luego de las pruebas realizadas este número resulto ser el más efectivos.

Así mismo se recomienda que de darse la implementación de la solución propuesta se modifique el rango de años que se puede realizare el análisis del control posterior, ya que al implementar una arquitectura de big data no hay la limitante del volumen de los datos a procesar, al aumentar la cantidad de datos a evaluar se podrá identificar más patrones de comportamiento y por lo tanto establecer estrategias para combatir ese tipo de contravenciones por parte de los operadores del comercio exterior.

Los componentes utilizados en la implementación de la solución no tienen costos de licenciamiento (Software Libre) por lo que la implementación del proyecto no representaría ningún costo a la institución, por lo que se recomendaría el evaluar su instalación en un mediano plazo,

El implementar esta solución podría asegurarse un aumento en las recaudaciones aduaneras y una disminución de los casos de evasión de impuestos.

Referencias Bibliográficas

Ballesteros, A. (2015). *Comercio Exterior: teoría y práctica.* Murcia, España: Universidad de Murcia.

Business Software. (2013). *Business Software*. Obtenido de http://businesssoftware.net/que es big data

Cristecu, M. P. (2017). Business Intelligence Integrated Solutions. *Land Forces Academy Review*, 1.

Gartner. (2013). *IT Glosary*. Obtenido de https://www.gartner.com/it-glossary/big-data

Group, IBM Software. (2007). *Introduction to Business Inteligence*. Obtenido de https://www-07.ibm.com/sg/events/blueprint/pdf/day1/Introduction_to_Business_Intelligence.pdf

Jimenez, M. Y. (2014). *Base de Datos relacionales y modelado de datos.* Málaga: IC Editorial.

Joyanes, L. (2013). *Big Data Analisis de grandes volumenes de datos en organizaciones.* México: Alfaomega Grupo Editor.

Loshin, D. (2003). *Business Inteligence The Savvy Manager Guide.* San Duego: Morgan Kaufmann Publishers.

Microsoft Corporation. (Febrero de 2019). *Gartner Magic Quadrant for Analytics and Business Intelligence Platforms*. Obtenido de https://info.microsoft.com/ww-landing-gartner-mq-bi-analytics-2019.html?LCID=EN-US

Noriega, R., Valdivia, G., Valenzuela, J., Tamer, M., Acosta, J., & López, R. (2015). Evolución de la Inteligencia de Negocios. *CULCyT*, 304.

Oracle. (s.f). *Inteligencia de Negocios.Oracle Business Intelligence*. Obtenido de https://www.oracle.com/ocom/groups/public/@otn/documents/.../317529_esa.pdf

Parr, O. (2000). *Data Mining cookbook*. Obtenido de http://books.google.com.co/books?id=L3w0loZrcU0C&printsec=front

Pazmiño, T., & Lopez, F. (2017). *Propuesta de un modelo de inteligencia organizacional como.*

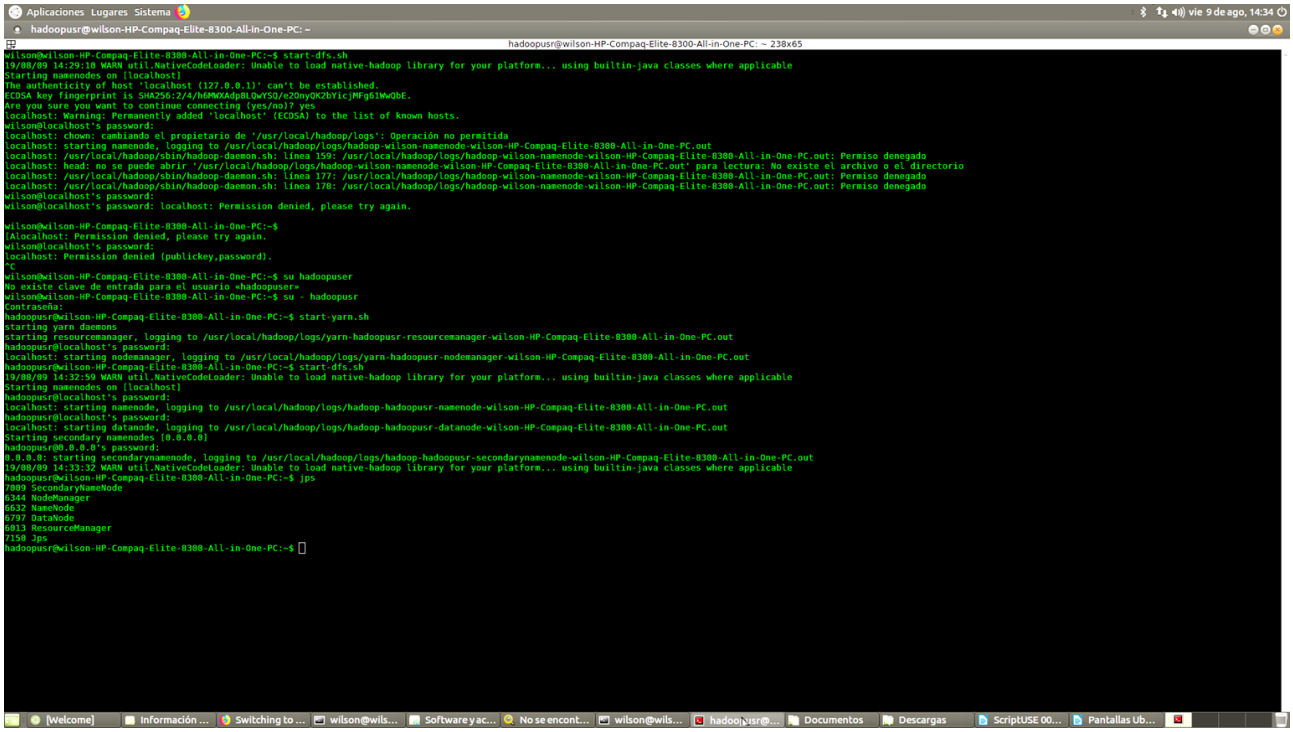
Ramos , S. (2016). *Business Intelligence (BI) & Analytics.* 2016: SolidQ Global S.A.

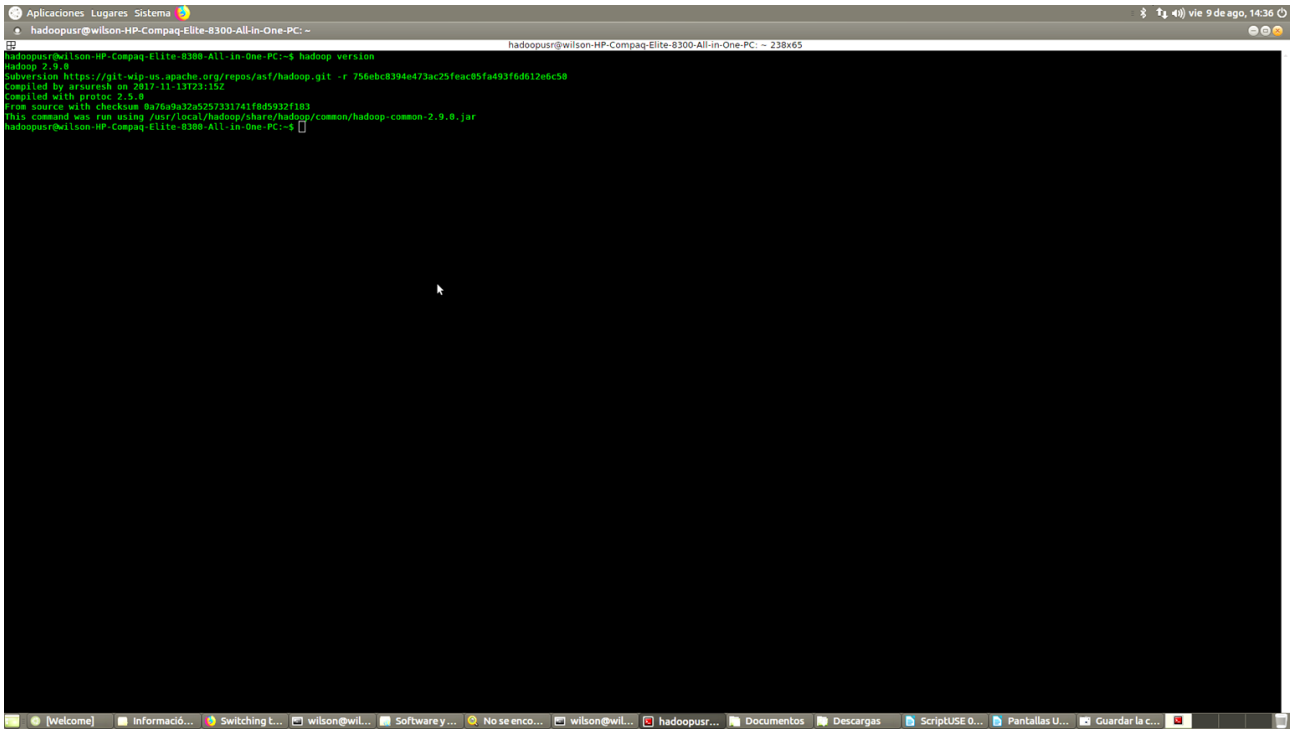
Reyes, D., & Nuñez, L. (2015). La inteligencia de Negocios como apoyo a la toma de decisiones en el ámbito académico. *Revista Internacional de Gestión del Conocimiento y la Tecnología* , 68.

Trujillo, J., & Rojas, P. (2015). *Modelo de Implementación de business Intelligence basado en Sap Hana para las empresas de Retail.*

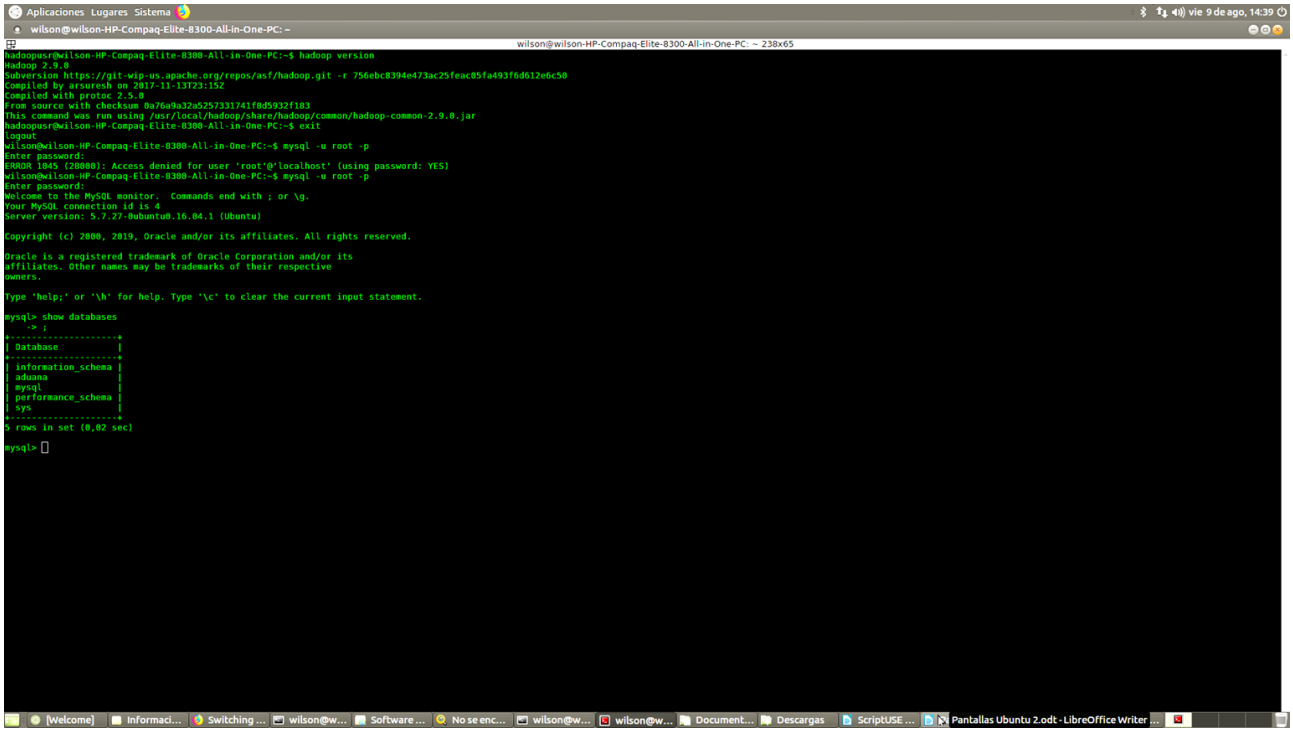
Anexos

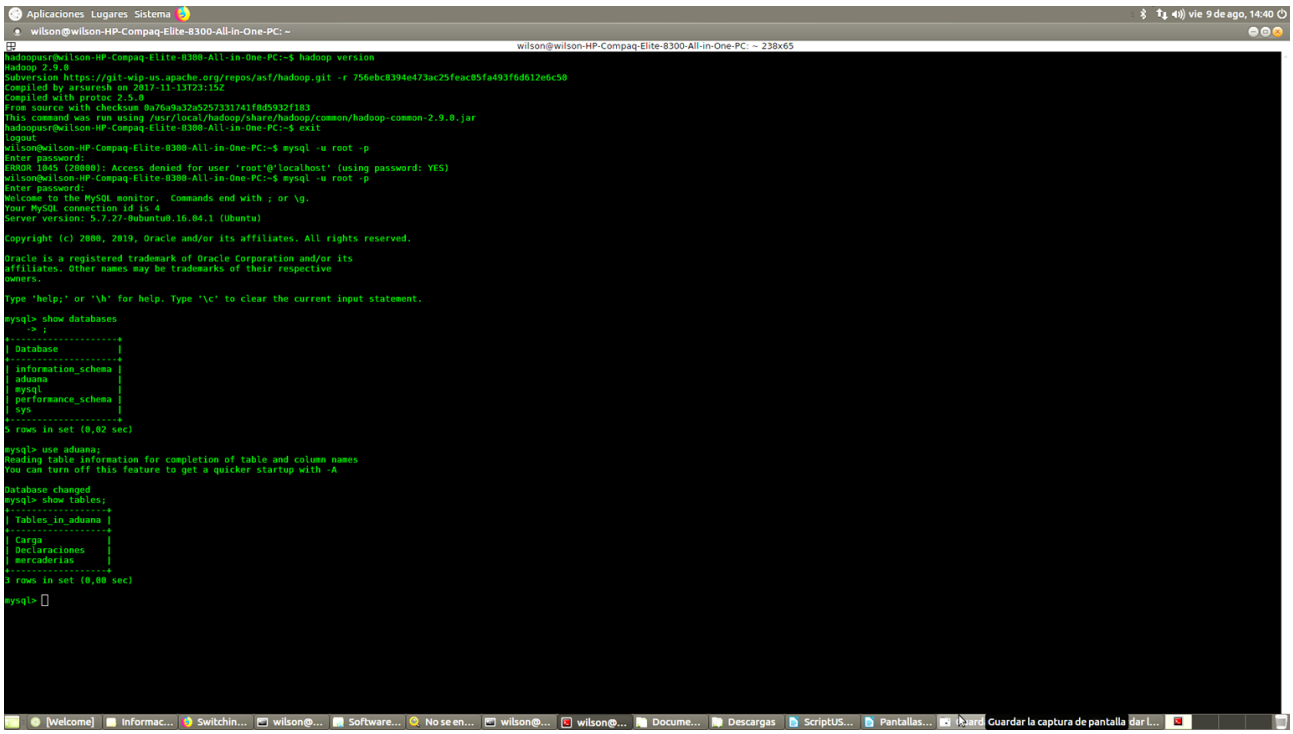
Anexo 1. Instalación Hadoop



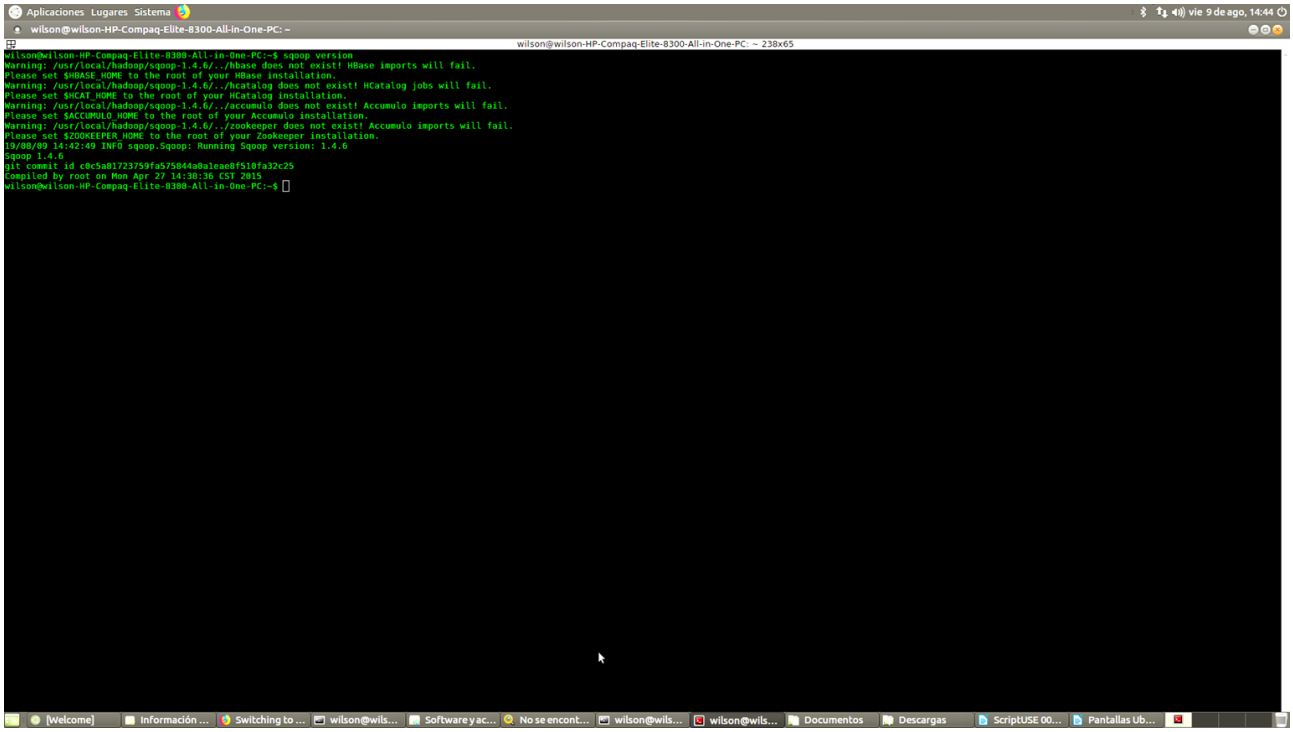


Anexo 2. Instalación Mysql





Anexo 3. Instalación Sqoop



Anexo 4. Respuestas a cuestionario expertos

EXPERTO 1 JEFE DE MEJORA CONTINUA

		GRADO DE ACUERDO				
PREGUNTA	DESCRIPCION	1	2	3	4	5
1	ESTA USTED DE ACUERDO CON LA SOLUCIÓN PROPUESTA ?					X
2	LA INTERFASE PRESENTADA ES FACIL DE VISUALIZAR ?				X	
3	ES FACITIBLE LA IMPLEMENTACIÓN DE ESTA SOLUCIÓN EN ADUANA?					X
3	CON ESTA PROPUESTA SE IDENTIFICARAN MAS CASOS DE EVASIÓN DE IMPUESTOS ?			X		
5	CONSIDERA USTED QUE SE PODRIAN AUMENTAR LAS RECAUDACIONES CON LA SOLUCIÓN PROPUESTA				X	
	TOTAL	**0**	**0**	**1**	**2**	**2**

EXPERTO 2 DOCENTE

		GRADO DE ACUERDO				
PREGUNTA	**DESCRIPCIÓN**	**1**	**2**	**3**	**4**	**5**
1	ESTA USTED DE ACUERDO CON LA SOLUCIÓN PROPUESTA ?				X	
2	LA INTERFASE PRESENTADA ES FACIL DE VISUALIZAR ?					X
3	ES FACITIBLE LA IMPLEMENTACIÓN DE ESTA SOLUCIÓN EN ADUANA?				X	
4	CON ESTA PROPUESTA SE IDENTIFICARAN MAS CASOS DE EVASIÓN DE IMPUESTOS ?			X		
5	CONSIDERA USTED QUE SE PODRIAN AUMENTAR LAS RECAUDACIONES CON LA SOLUCIÓN PROPUESTA				X	
	TOTAL	***0***	***0***	***1***	***3***	***1***

EXPERTO 3 JEFE DE PROYECTOS BI

		GRADO DE ACUERDO				
PREGUNTA	**DESCRIPCIÓN**	**1**	**2**	**3**	**4**	**5**
1	ESTA USTED DE ACUERDO CON LA SOLUCIÓN PROPUESTA ?			X		
2	LA INTERFASE PRESENTADA ES FACIL DE VISUALIZAR ?				X	
3	ES FACITIBLE LA IMPLEMENTACIÓN DE ESTA SOLUCIÓN EN ADUANA?				X	
4	CON ESTA PROPUESTA SE IDENTIFICARAN MAS CASOS DE EVASIÓN DE IMPUESTOS ?			X		
5	CONSIDERA USTED QUE SE PODRIAN AUMENTAR LAS RECAUDACIONES CON LA SOLUCIÓN PROPUESTA				X	
	TOTAL	**0**	**0**	**2**	**3**	**0**

Anexo 5. Tabulación de respuestas expertos

PREGUNTAS	GRADO DE ACUERDO
1	0
2	0
3	4
4	8
5	3
TOTAL	**15**

Anexo 6. Carta a Expertos

Guayaquil, Junio de 2019

Estimado (a) señor (a):

Ciudad

La presente es para solicitar de la manera más comedida se sirva colaborar en revisar y responder el cuestionario adjunto, el objetivo es poder validar los resultados del desarrollo de la tesis para la obtención de la maestría cuya propuesta es la "Implementación de una arquitectura de Big Data para la auditoria de los procesos de importación en el Servicio de Aduana"

Su aporte es de valiosa ayuda ya que con sus conocimientos y experiencia van aportar en gran medida al desarrollo de la tesis.

Agradecido por su colaboración y por su apoyo.

Atentamente

Ing, Wilson Samaniego
Maestrante