Analysis data insight from multi-players in a mobile social game

Luis Ángel Romero Gamarra
Grado de Ingeniería Informática

Humberto Andrés Sanz
Junio 2017

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FITXA DEL TREBALL FINAL

Títol del treball:	*Analysis data insight from multi-players in a mobile social game*
Nom de l'autor:	*Luis Ángel Romero Gamarra*
Nom del consultor:	*Humberto Andrés Sanz*
Data de lliurament (mm/aaaa):	*06/2017*
Àrea del Treball Final:	*Bussiness Intelligence*
Titulació:	*Grado de Ingeniería Informática*

Resum del Treball (màxim 250 paraules):

Estamos en la era de la información, donde cada año los datos que se encuentran en la red crecen exponencialmente, creando nuevas oportunidades de negocio a nivel empresarial.

A raíz de esto, han surgido distintos tipos de áreas que ayudan en el estudio de la información mediante nuevas tecnologías y herramientas creadas a partir de estos campos de estudio como son: La inteligencia de negocios y el Big Data.

Estos dos campos se unen para crear una nueva área interdisciplinar denominada Data Science, que busca mediante nuevas metodologías, extraer conocimientos de los datos.

Por tanto, este proyecto busca predecir, a través del análisis de los datos, el grado de jugabilidad de los usuarios, así como su grado de compromiso. Además de ello, se busca encontrar nuevas oportunidades de negocio a través de las ventas dentro de la aplicación.

Abstract (in English, 250 words or less):

We are in information era, where each year the data grows exponentially, making new opportunities of business.

As a result, different kinds of areas have emerged that aid in the data analysis through new technologies and tools created from other areas such as: Business Intelligence and Big Data.

This fields are joined to create a new area interdisciplinary called Data Science. This field seeks, through new methodologies, to extract knowledge or new data insight.

Therefore, this project intends to predict through data analysis, the interactivity and engagement of players in the mobile social game. In addition, it seeks to find new opportunities to increment the revenue from app purchases.

Paraules clau (entre 4 i 8):

Business Intelligence, Mobile social game, Data Analysis, Big Data, Data Insight.

Índex

Lista de figures

1. Introducción

El presente Trabajo de Final de Grado (TFG) se basará en el análisis de datos enfocado a los juegos móviles, estos análisis se llevarán a cabo en el ámbito de la Inteligencia de Negocio (Business Intelligence, por sus siglas en inglés, BI en adelante), en concreto, en las soluciones ofrecidas por las herramientas Big Data[1].

De esta manera, se mostrará, de forma breve, cuál es el papel del Big Data con relación al BI, la Ciencia de Datos (Data Science) y la gamificación.

Además, se utilizarán herramientas analíticas basadas en estas disciplinas con el fin de analizar los diferentes tipos de información mediante metodologías que se desarrollan en el campo de la Ciencia de Datos con el propósito de la correcta elaboración de cada fase hasta conseguir los objetivos del proyecto. Se describirán entonces, de manera clara y concisa, cuáles son estas etapas y como los datos son procesados en cada fase

Por otro lado, veremos también las distintas técnicas de análisis que nos ofrece la disciplina del Machine Learning[2] que permiten encontrar patrones de información con el fin de construir modelos de datos para ser evaluados.

Por último, presentaremos los resultados del análisis, donde se mostrarán los reportes que sirvan para la toma de decisiones a nivel empresarial y la búsqueda de nuevas oportunidades de negocio.

[1] Big Data es el término que se dan a los grandes conjuntos de datos.
[2] Machine Learning es una rama de la Inteligencia Artificial que tiene como objetivos desarrollar técnicas para el aprendizaje, descubrimiento de patrones y toma de decisiones a través de los datos.

1.1 Contexto y justificación del Trabajo

Hoy en día los datos representan un valor inmenso en el ámbito empresarial, ya que a través de los años estos datos se han ido incrementando de forma exponencial.

Un estudio elaborado en el año 2016 por la consultora McKinsey[3] indica que los volúmenes de datos se duplican cada tres años, donde el principal origen es la plataforma digital, aplicaciones de realidad virtual y miles de millones de teléfonos móviles.

En el contexto de las aplicaciones móviles y en concreto con la plataforma de juegos móviles, existe un desafío enorme ya que los datos producidos por parte de los videojuegos son muy diversos. Esta heterogeneidad implica grandes problemas para poder extraerlos, procesarlos y darles valor significativo para la mejora de sus servicios e ingresos. Por ejemplo, los datos basados en la geolocalización juegan un papel importante a la hora de crear nuevas oportunidades de negocio ya que a través de su geolocalización o interacción con la aplicación se puede ofrecer publicidad personalizada.

Por otro lado está la gamificación, este término que se da a las estrategias que se utilizan hoy en día para afianzar el compromiso o engagement de los jugadores hacia el juego móvil. Esta viene a ser una parte fundamental en el negocio, ya que es importante que los jugadores puedan crear un alto grado de compromiso.

Actualmente existen muchas plataformas especializadas con el ámbito de la inteligencia de negocios, algunas grandes empresas como Microsoft con su plataforma Power BI, Apache con Hadoop o Splunk, son algunas opciones que junto con otras plataformas externas como Amazon AWS o Spark, brindan soluciones que hoy en día existen en el mercado.

Todo esto junto a las nuevas tecnologías como la computación en la nube (Cloud Computing) permiten realizar todo tipo de tareas en cualquier momento y lugar, contribuyendo a que hoy en día exista el potencial de cómputo adecuado para poder analizar grandes volúmenes de datos de manera eficiente.

Es por ello que la empresa GameLab Inc[4]. Se encuentra en la necesidad de saber el grado de jugabilidad de los usuarios, así como su grado de compromiso, además de ello, se busca encontrar nuevas oportunidades de negocio a través de ventas dentro de la aplicación. Consecuentemente, este proyecto busca predecir, a través del análisis de los datos, las diferentes necesidades mencionadas anteriormente.

1.2 Objetivos del Trabajo

El objetivo principal del este proyecto es el análisis de datos producidos por la aplicación para predecir el compromiso de los jugadores (engagement), la interacción, ingresos económicos y tendencias que produce la interacción de los usuarios con la aplicación con el fin de crear nuevas oportunidades de negocio y mejora del servicio.

Para esto, tenemos que ejecutar los diferentes métodos para lograr transformar los datos brutos y procesarlos. Estos datos vienen a ser un conjunto de ficheros que simulan las distintas informaciones que produce una aplicación real que servirán de base para iniciar todo el proceso.

[3] Informe del Instituto Global Mckinsey especializado en Economía Internacional.
[4] GameLab Inc. Es el nombre de una empresa ficticia presentada solo para uso educativo.

En este sentido y a medida que se desarrolla el proyecto iremos definiendo los siguientes puntos claves:

- Construcción de la estrategia de Big Data
- Elección de la plataforma Big Data
- Adquisición y exploración del conjunto de datos para entenderlos
- Aplicación de la metodología de la ciencia de datos
- Entendimiento del negocio a través de las aplicaciones como son modelo de gamificación y las compras a través de la app.
- Análisis de la información
- El valor agregado resultante traducido en los reportes o “dashboards”.
- Líneas de actuación o recomendaciones

1.3 Enfocamiento y método a seguir

Como ya se ha indicado anteriormente, se utilizará la metodología que consta de distintas fases que se desarrollan en la ciencia de datos como son:

1. **Adquisición**, obtención de los datos a través de la aplicación.
2. **Preparación**, exploración y preproceso de los datos utilizando modelos gráficos y modelos estadísticos para la depuración de la información.
3. **Análisis**, donde se utilizan diferentes técnicas de análisis basadas en Machine Learning.
4. **Reportes**, se presentarán los reportes que ayuden a comunicar los resultados obtenidos.
5. **Actuación**, en esta etapa se deben de correlacionar los resultados con los objetivos planteados, es decir, aplicarlos al modelo de negocio de la empresa.

En consecuencia, cada uno de estos pasos se detallarán más adelante y se desarrollarán iterativamente con el objetivo de asegurar la correcta ejecución de la metodología y los objetivos planteados.

Es necesario hacer hincapié en iteración de cada fase, ya que, puede suceder que después del análisis se tenga que volver a la fase de adquisición en la búsqueda de nueva información, de esta forma, se valida de manera eficaz el resultado de cada proceso.

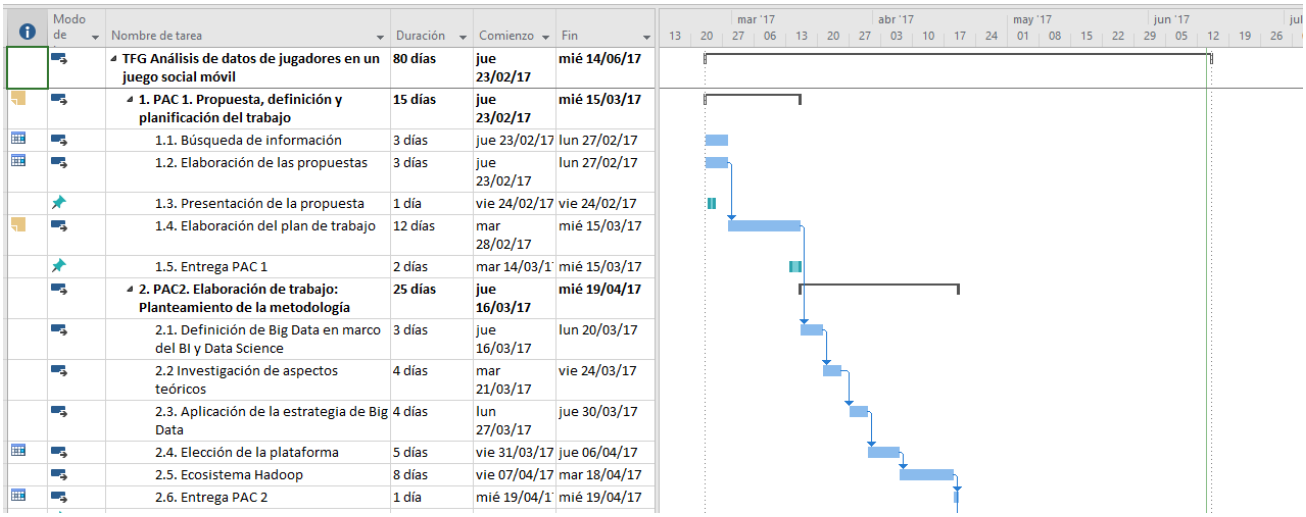
A nivel de infraestructura, se trabajará con una máquina virtual en la cual se utilizará la plataforma Cloudera[5] que contiene herramientas necesarias para poder trabajar con grandes volúmenes de datos.

[5] Cloudera es una compañía que proporciona software basado en Apache Hadoop, soporte y servicios, y formación para grandes clientes en el ámbito de BI.

1.4 Planificación del Trabajo

La planificación se realizará mediante un diagrama de *Gant* donde se fijan las tareas en cuestión de tiempo determinados, de esta manera realizar el seguimiento óptimo del proyecto.

La planificación viene a ser la siguiente:



Modo de	Nombre de tarea	Duración	Comienzo	Fin
	TFG Análisis de datos de jugadores en un juego social móvil	**80 días**	**jue 23/02/17**	**mié 14/06/17**
	1. PAC 1. Propuesta, definición y planificación del trabajo	**15 días**	**jue 23/02/17**	**mié 15/03/17**
	1.1. Búsqueda de información	3 días	jue 23/02/17	lun 27/02/17
	1.2. Elaboración de las propuestas	3 días	jue 23/02/17	lun 27/02/17
	1.3. Presentación de la propuesta	1 día	vie 24/02/17	vie 24/02/17
	1.4. Elaboración del plan de trabajo	12 días	mar 28/02/17	mié 15/03/17
	1.5. Entrega PAC 1	2 días	mar 14/03/1	mié 15/03/17
	2. PAC2. Elaboración de trabajo: Planteamiento de la metodología	**25 días**	**jue 16/03/17**	**mié 19/04/17**
	2.1. Definición de Big Data en marco del BI y Data Science	3 días	jue 16/03/17	lun 20/03/17
	2.2 Investigación de aspectos teóricos	4 días	mar 21/03/17	vie 24/03/17
	2.3. Aplicación de la estrategia de Big Data	4 días	lun 27/03/17	jue 30/03/17
	2.4. Elección de la plataforma	5 días	vie 31/03/17	jue 06/04/17
	2.5. Ecosistema Hadoop	8 días	vie 07/04/17	mar 18/04/17
	2.6. Entrega PAC 2	1 día	mié 19/04/1	mié 19/04/17

Ilustración 1. Planificación 1

	3. PAC 3. Primera parte del Desarrollo de la Metodología	**25 días**	**jue 20/04/17**	**mié 24/05/17**
	3.1. Instalación del ecosistema de trabajo	1 día	jue 20/04/17	jue 20/04/17
	3.2. Adquisición de datos	3 días	jue 20/04/17	lun 24/04/17
	3.3. Preparación de datos	**7 días**	**mar 25/04/1**	**mié 03/05/17**
	3.2.1. Exploración	3 días	mar 25/04/1	jue 27/04/17
	3.2.2. Preproceso	3 días	vie 28/04/17	mar 02/05/17
	3.2.2. Verificación del proceso iterativo	1 día	mié 03/05/17	mié 03/05/17
	3.4. Analisis de datos	**13 días**	**vie 05/05/17**	**mar 23/05/17**
	3.3.1. Investigación de información	5 días	vie 05/05/17	jue 11/05/17
	3.3.2. Elección de tecnicas Machine Learning	4 días	vie 12/05/17	mié 17/05/17
	3.3.3. Verificación del proceso iterativo	2 días	jue 18/05/17	vie 19/05/17
	3.5. Entrega PAC 3	1 día	mié 24/05/1	mié 24/05/17
	4. Pac 4 Segunda parte del desarrollo	**15 días**	**jue 25/05/17**	**mié 14/06/17**
	4.1. Reportes finales	**5 días**	**jue 25/05/17**	**mié 31/05/17**
	4.1.1. Instalación de plataforma	2 días	jue 25/05/17	vie 26/05/17
	4.1.2. Creación de informes	2 días	lun 29/05/17	mar 30/05/17
	4.1.3. Verificación del proceso	1 día	mié	mié 31/05/17

Ilustración 2. Planificación 2

	4.1. Reportes finales	**5 días**	**jue 25/05/17**	**mié 31/05/17**
	4.1.1. Instalación de plataforma	2 días	jue 25/05/17	vie 26/05/17
	4.1.2. Creación de informes	2 días	lun 29/05/17	mar 30/05/17
	4.1.3. Verificación del proceso iterativo	1 día	mié 31/05/17	mié 31/05/17
	4.2. Actuación	**6 días**	**jue 01/06/17**	**jue 08/06/17**
	4.2.1. Aplicación de los resultados	3 días	jue 01/06/17	lun 05/06/17
	4.2.2. Verificación de la consecución de los objetivos	2 días	mar 06/06/17	mié 07/06/17
	4.2.3. Rectificación y últimas valoraciones	1 día	jue 08/06/17	jue 08/06/17
	4.3. Revisión de la memoria	2 días	vie 09/06/17	lun 12/06/17
	4.4. Preparación y grabación del trabajo final	1 día	mar 13/06/17	mar 13/06/17
	4.5. Entrega PAC 4	1 día	mié 14/06/1	mié 14/06/17

Ilustración 3. Planificación 3

1.5 Breve sumario de productos obtenidos

Al final del trabajo se tendrán cada una de las entregas representando las soluciones de las PACs, además de esto se entregarán lo siguiente:

- Entrega de los recursos técnicos como son datos iniciales, ficheros en formato CSV.
- Ficheros técnicos con referente al resultado del proyecto: informes, diagramas y reportes técnicos.
- Entrega de la memoria del trabajo final de grado y presentación del mismo.

2. Definición de Big Data en el marco de la Inteligencia de negocios y la ciencia de datos

Para poder conocer las distintas relaciones entre estos diferentes conceptos es necesario conocerlo a fondo y saber cómo funcionan, de modo general, estas diferentes áreas.

Consecuentemente, en este apartado se definen los diferentes aspectos de la inteligencia de negocios (B.I.), se describen algunas características del Big Data y cuál es el rol de estas áreas con referente a la ciencia de datos o Data Science.

2.1 Bussines Intelligence

Business Intelligence es el término que se emplea cuando se trabaja con datos con el fin de darles valor y mejorar las tomas de decisiones a nivel empresarial, teniendo como objetivo el apoyo continuado y sostenible, facilitando la información necesaria en todo momento y de forma inmediata.

De esta manera, se puede citar la definición de BI, tomando el término introducido por la consultora Gartnet[6]:

"BI es un término que incluye aplicaciones, infraestructura y herramientas. Además de mejores prácticas que habilitan su acceso, así como el análisis de información para mejorar y optimizar decisiones y rendimiento."

(Gartner, 2017)

Se puede decir entonces que BI es un proceso iterativo para explorar y analizar información estructurada sobre un área, para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones.

[6] Extraído de la consultora tecnològica internacional Gartner, especializada en tecnologías de la información

La metodología que utiliza BI se basa en el principio de agrupar todos los datos empresariales en un servidor central llamado Data Warehouse, donde los datos se estructuran en una base de datos relacional, con conjuntos de índices y forma de acceso mediante tablas (cubos multidimensionales o cubos OLAP).

Componentes

Para lograr este cometido, este sistema de información utiliza diferentes herramientas y técnicas para el proceso de los datos.

Ilustración 4. Componentes BI

Herramientas ETL (Extraction, Transformation and Load)

El proceso de ETL extrae datos de diferentes fuentes de origen, después los valida, normaliza y transforma para almacenarlo en un entorno Data Warehouse (que se verá más adelante), para su posterior análisis.

- **Extraer:** como su nombre lo indica, extrae datos desde diferentes fuentes como ERP, CRM, y otro modelos de datos con distintos formatos (csv, XML, host).

- **Transformar:** Dependiendo el modelo de datos y el diseño lógico de almacenamiento, la transformación procesa estos datos de acuerdo a las reglas de negocio, donde se incluyen validaciones técnicas, normalización, homogeneización de códigos, cambios de formatos y procesos analíticos.

- **Carga (Load):** Representa la carga de los datos transformados en la etapa anterior, esta carga puede ser mediante ficheros batch[7] (por lotes, registro a registro o cargas tanto incrementales y totales):

[7] Archivo de texto que contiene secuencias de órdenes para ser ejecutadas en el sistema operativo.

Data Warehouse

Data Warehouse es un almacén de datos orientado a objetos, no volátil[8], integrado por colecciones de datos que pueden variar en el tiempo, creados para posteriormente transformarlos en información útil en la toma de decisiones organizacional.

DataMarts

Son subconjuntos de almacenes o Data Warehouses orientados a áreas específicas y diseñados con el fin de unir diferentes áreas departamentales dentro de una misma organización.

Cubos OLAP (On-Line Analytical Processing)

Son cubos de información que tienen un número indefinido de dimensiones, donde cada cubo contiene datos de una determinada variable que se desea analizar, proporcionando un vista lógica de los datos provistos por el sistema de información hacia el Data Warehouse.

Estas herramientas permiten a los usuarios finales realizar diferentes tipos de consultas, reportes o *dashboards*, con el objetivo de extraer conclusiones para la ayuda de la toma de decisiones empresarial.

2.2 Big Data

2.2.1 Introducción

El término Big Data se refiere al estudio de grandes torrentes de datos, y no sería posible gracias a la enorme evolución de internet, ya que es uno de las principales fuentes de datos donde cada año se incrementa a nivel exponencial.

Por otro lado están los diferentes dispositivos electrónicos, que interconectados hoy en día, son una fuente inagotable de información.

Si sumamos todo esto a las nuevas tendencias sociales a nivel digital como son las redes interconectadas de personas, podemos darnos cuenta de la gran era de los datos y como resultado aparecen ámbitos como el Big Data.

2.2.2 Definición

Big data es a menudo utilizado para referirse a un conjunto de datos que puede ser difícilmente gestionado usando sistemas de gestión de base de datos tradicionales. Por otro lado, también se puede decir que es una tecnología donde los datos están guardados en forma no estructurada.

[8] No volátil es un término que es referido cuando la información almacenada no se modifica ni elimina.

Si juntamos estas definiciones, podría decir que Big Data es un conjunto de tecnologías que permiten el almacenamiento, procesado y análisis de grandes cantidades de datos estructurados y no estructurados de forma escalable que permiten ganar nuevos conocimientos para la toma de decisiones.

2.2.3 Recursos que utiliza el Big Data

Existen tres fuentes principales de datos que nutren al Big Data:

Máquinas

Como máquinas se entiende por cada dispositivo inteligente, esto es para aquellos dispositivos que pueden conectarse a otros dispositivos o a internet (teléfonos inteligentes, Gps, relojes de actividad, etc.), así mismo, podemos decir que un dispositivo es inteligente si puede recolectar y analizar datos de forma autónoma (sensores térmicos, cámaras de tráfico, satélites, etc.)

Por último, un dispositivo inteligente también es aquel que provee un contexto ambiental, esto quiere decir definir un ambiente digitalizado e interconectado. Un ejemplo de este último contexto son aquellos relacionados con el Internet de las cosas (IoT[9]).

Personas

Las personas generan grandes cantidades de datos cada día mediante actividades en las redes sociales como son: Facebook, Twitter y LinkedIn. También utilizan otras plataformas como son Blogger, YouTube, Instagram o Google Drive.

Organizaciones

Los datos de las organizaciones están fuertemente relacionadas con el negocio de las mismas, vienen a ser datos empresariales tradicionales. Estos datos pueden ser diversos dependiendo del tipo de organización del que se habla. Por ejemplo, pueden ser datos gubernamentales, bancarios, transacciones comerciales, datos de investigación y desarrollo, etc.

Actualmente las organizaciones suelen guardar datos históricos para luego ser procesados, analizados y poder darles valor. Por ejemplo, los datos de transacciones pueden ser utilizados para detectar patrones relacionados con los usuarios, establecer patrones de demanda o capturar actividades fraudulentas.

[9] IoT, El internet de las cosas es un término definido para aquellos dispositivos interconectados entre sí que comparten datos de manera autómata.

Ilustración 5. Fuentes de datos Big Data

2.2.4 Características del Big Data

Según Gartner[10] Big Data utiliza tres características importantes:

- **Volumen:** Se refiere a la cantidad de datos que es generado en cada momento.
- **Velocidad:** Se refiere a la velocidad en la cual cada dato es generado, y su transferencia desde un lugar a otro.
- **Variedad:** Se refiere a la cantidad de formatos que pueden tener los datos.
- **Veracidad:** Se refiere a la calidad de los datos, el cual cambia constantemente.

A estas características pueden sumarse otras según el contexto del que se refiera:

- **Valencia:** Se refiere a la correlación de los datos entre sí.

Por último, según WOrDS[11], existe una sexta característica importante que debe ser mencionada:

- **Valor:** Se refiere al valor que Big Data debe brindar con los conocimientos obtenidos.

[10] Glosario de terminologías de la consultoría Gartner
[11] WOrDS, Centro de excelencia de flujo de ciencia de datos

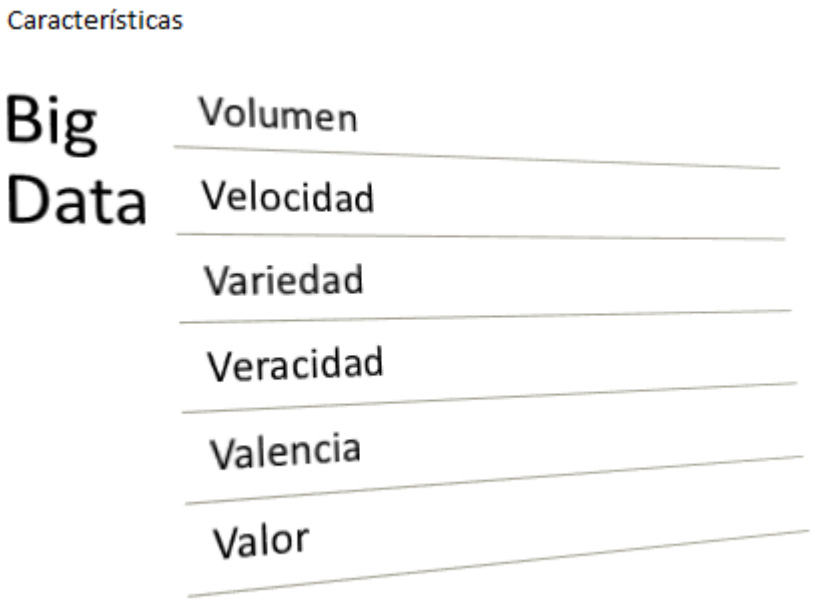


Ilustración 6. Características Big Data

2.3 Diferencias entre Business Intelligence y Big Data

Ahora como ya se ha visto cada una de estas tecnologías, en este apartado se compararán cuáles son las principales diferencias entre ellas.

- En un entorno Big Data, los datos se almacenan en un sistema de ficheros distribuidos, al contrario que en BI donde el almacenamiento se realiza en un servidor central (Data Warehouse). Por tanto, los entornos distribuidos son fáciles de escalar, no propensos a fallos, flexibles y de bajo costo.

- Big Data puede analizar diferentes modelos de datos, en diferentes formatos tanto estructurados como no estructurados, las soluciones Big Data solventan estos inconvenientes permitiendo un análisis global y centralizado desde diferentes fuentes de información.

- Big Data puede realizar procesos en tiempo real e histórico, a comparación de las soluciones BI donde solo se analizan datos históricamente guardados en el tiempo.

- Big Data ofrece una arquitectura escalable a nivel de cómputo, utiliza procesado paralelo para procesar la información, permitiendo las altas tasas de procesamiento. Por otro lado, esta arquitectura utiliza clústeres para el almacenamiento de datos, reduciendo el costo de almacenamiento. Al contrario que en la tecnología BI, donde el almacenamiento es centralizado, no permitiendo la escalabilidad.

De esta manera, podemos decir que Big Data ofrece soluciones para trabajos específicos, donde se requiere trabajar con grandes cantidades de datos y por ende, aumentar la escalabilidad sin problemas de costos. A nivel de cómputo también existen diferencias,

donde en Big Data es primordial los parámetros de procesamiento paralelo con lo cual es necesario una arquitectura diferente.
En este sentido, podemos concluir que Big Data no pretende sustituir las herramientas BI, sino que proporciona otras herramientas para determinados propósitos. Consecuentemente, estas tecnologías se utilizarán dependiendo las necesidades de la empresa.

2.4 Ciencia de Datos

La ciencia de datos o Data Science es un campo interdisciplinar que busca, mediante métodos científicos, extraer conocimiento de los datos. Esta disciplina se apoya en distintos campos como la Estadística, Data Mining, Machine Learning, Análisis predictivo y Ciencias de la computación.

Por tanto, la ciencia de datos utiliza información de diferentes fuentes (estructuradas y no-estructuradas), utiliza métodos científicos para analizarlos, obteniendo "data insights.[12]"

Estos métodos científicos o metodologías suelen ser iterativas, esto quiere decir, que existe una realimentación de información en cada fase. Más adelante veremos esta metodología con más detalles.

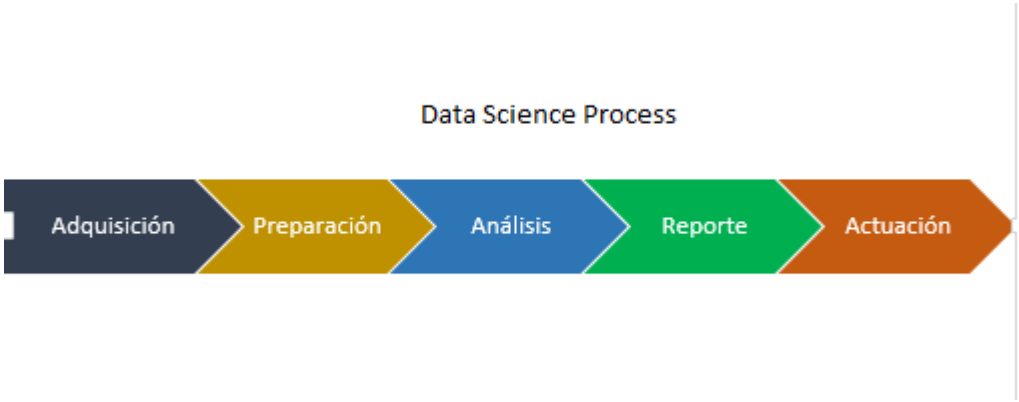


Ilustración 7. Proceso de la metología Data Science

[12] Data insght es el término que se refiere a los productos de datos en Data Science.

2.5 Ciencia de Datos, BI y Big Data

La ciencia de datos al igual que Big Data, suele trabajar con grandes cantidades de datos, los cuales suelen estar en diferentes modelos y formatos. Estos datos a menudo necesitan ser adquiridos ya que muchas veces están incompletos. Además, esta información suelen requerir un proceso de exploración y preparación antes de ser analizados.

Por otro lado, cuando hablamos de tecnologías BI, los conjuntos de datos suelen ser completos, extraídos desde archivos ya procesados, donde los conjuntos de datos son gestionables. Recordemos también que en Inteligencia de negocios, solo puede procesar datos estructurados, siendo esta una gran diferencia.

A nivel tecnológico, la ciencia de datos puede utilizar tanto herramientas Big Data como otras de diferentes campos, como son lenguajes de programación, herramientas predictivas del campo de Data Minning o Machine Learning, herramientas de reportes del campo Estadístico, incluso pueden utilizar herramientas de cuadros de mando de las tecnologías BI.

A nivel de metodologías, la ciencia de datos utiliza el método científico o método empírico, tal como habíamos comentado anteriormente. Por otro lado, BI utiliza el proceso ETL como metodología para el proceso de la información.

3. Estrategia Big Data enfocado a los objetivos del negocio

Este aparcado se desarrolla de acuerdo con los objetivos del trabajo, esto viene ser, definir las características del juego, definir ¿cuál es el problema?, ¿Qué cuestiones plantea el problema?, así como la construcción de la estrategia para alcanzar el objetivo inicial.

3.1 Características del Juego

Uno de los productos de la empresa GameLab Inc. Es un juego muy popular llamado CatchingPets[13]. El objetivo del juego es cazar diferentes tipos de mascotas en las diferentes misiones y niveles del juego. Estas misiones son en tiempo real y el juego provee un mapa interactivo donde se encuentran ocultos cada mascota.

Cada nivel del juego se ira desbloqueando con el fin de crear mayor compromiso con los jugadores, cada nivel será más complicado que el anterior y para pasar de nivel será necesario cierto habilidades y criterios realizados o mascotas obtenidas.

Cada mascota tiene distintas apariencias y habilidades, para atraparlas, es necesario darle clic en diferentes áreas de su cuerpo. Las puntuaciones estarán condicionadas a la

[13] El nombre CatchingPets es ficticio, que se usa para propósito educativo y con fines de desarrollar este trabajo de final de grado.

velocidad de conseguir las mascotas y a la exactitud del clic para atraparlas. Cada clic no satisfactorio será penalizado con puntos.

El juego es multiusuario, esto quiere decir que los usuarios iniciarán el juego de forma individual en el primer nivel, a medida que avanza el juego, podrán crear o unirse a diferentes equipos, cada equipo estará compuesto por un usuario o más. El primer nivel es de aprendizaje y fácil de jugar, pero a medida que suba de nivel, la complejidad de las misiones incrementarán.

Los jugadores se pueden comunicar entre ellos mediante el chat o mensajes asignados a cada equipo, además de esto, también pueden utilizar las redes sociales para comunicarse.

De modo general, existen las siguientes consideraciones:

- **Puntuación de usuarios:** Cada usuario será puntuado individualmente mediante la rapidez y exactitud a la hora de cazar una mascota. Las puntuaciones pueden ser vistas en tiempo real, mediante la app[14] del juego o desde la web del mismo. Dependiendo de los puntos del jugador, se desbloquearán distintas categorías en base al historial de puntos del jugador.
- **Puntación de equipos**: Cada equipo es puntuado públicamente, el equipo tendrá como máximo 30 miembros, cada equipo puede fichar otro jugados, así como cada jugador puede fichar por otro equipo. Las votaciones se realizarán entre cada miembro, donde tiene que haber un 80% de votos para aceptar a un nuevo jugador. Cuando los jugadores salen de un equipo, este se elimina automáticamente.
- **Compras en el juego:** Los jugadores pueden realizar compras en el juego, ya sea para incrementar sus opciones de cazar más mascotas, existen diferentes tipos de armas y herramientas que están disponibles según las misiones y niveles.
- **Finalización del juego:** El juego nunca acaba, el reto es encontrar nuevas estrategias para crear el mayor compromiso de los jugadores ya que los niveles seguirán aumentando. La empresa hará uso de herramientas analíticas Big Data para mantener a los usuarios enganchados en el juego.

En los siguientes apartados se analizarán a fondo las diversas fuentes de datos para empezar el proceso de análisis y obtención de resultados.

3.2 Construcción de la estrategia de Big Data

Por estrategia se entiende como un plan de acción diseñado para conseguir un objetivo trazado. Partiendo de esta definición, se procederá a trazar un plan de acción para llevar a cabo la realización de las metodologías en el proceso del desarrollo del proyecto.

Esta estrategia sirve de apoyo para la planificación realizada anteriormente. Debemos tener en cuenta que hay una gran diferencia entre la planificación del trabajo en general y el plan de acción del proceso de Big Data.

[14] App, es el término que se da a las aplicaciones móviles.

Por un lado la planificación nos ayuda a desarrollar todo el trabajo según las imposiciones y características desarrolladas en la planificación, esto quiere decir, el desarrollo total de las tareas para conseguir el acabado del proyecto final.

Por otro lado, el plan de acción de la estrategia de Big Data nos permite identificar los diferentes tipos de problemas con el fin de realizar analíticas enfocadas en el negocio de la empresa, y hallar así, nuevos objetivos que deben ser alcanzados.

Consecuentemente, se definen los pasos para poder describir las características del problema

3.2.1 Definir el Problema

En el apartado de objetivos generales se definieron, de modo general, los propósitos del presente trabajo. Recordemos que el principal objetivo era el análisis de datos para mejorar el servicio y generar nuevas oportunidades de negocio.

La empresa nos indica que estas propuestas vienen dado por la siguiente preocupación que respecta al ámbito de negocio:

- **Jugabilidad**, se refiere a la interacción del usuario con el juego, se mide el compromiso del usuario a través de su historial.
- **Ingresos económicos**, parte importante de la empresa son las ventas dentro de la aplicación, es necesario conocer a fondo a los usuarios quién son los que efectúan las compras.
- **Tendencias,** se refiere a la tendencia de jugabilidad sobre uno o más usuarios, además la tendencia en general del juego (incremento de usuarios, medida de dificultad de niveles, total de usuarios activos)
- **Mejora del servicio**, se refiere a la captura de nuevos usuarios y los indicadores de plataforma donde ellos juegan, es importante conocer que plataformas son más usadas, etc.

Mediciones:

De esta manera, se observa que es necesario enfocarnos en los ingresos, el compromiso de los usuarios y la mejora del servicio prestado que incluye la jugabilidad. Un ejemplo de los apartados que se intentarán medir son:

- El número de usuarios por determinado tiempo.
- Conocer que usuario son los que más compras ítems.
- Usuarios activos.
- Equipos más activos
- Categorías de usuarios
- Que plataformas son más usadas
- Usuarios con más mascotas
- Evolución de las ventas
- Ítems más vendidos

Análisis

Una vez que sabemos dónde tenemos que enfocarnos y realizado las preguntas anteriores para saber que nos dicen aquellos datos del negocio, es momento de pensar en el análisis.

Para el análisis tendremos cuestiones que no se pueden responder a simple vista, el procedimiento de analizar los datos se verán dentro del desarrollo de la metodología.

Algunos análisis importantes serían:

- Predecir que usuarios son los más probables que compren artículos con más costos que otros. Sería muy importante ya que las compras dentro de las aplicaciones son una fuente considerable de ingresos para la empresa.
- Predecir qué artículos o ítems serán más comprados.
- Conocer nuevas tendencias en el juego mediante el historial de los jugadores.
- Encontrar patrones entre diferentes variables del juego (por ejemplo, entre el usuario y las mascotas cazadas).
- Clasificar los usuarios más activos mediante el análisis de los chats, con el fin de realizar promociones personalizadas (marketing) para aquellos chat más activos.

3.2.2 Evaluar la situación

Para poder realizar tanto las mediciones y los análisis, es necesario evaluar la situación actual, esto quiere decir si se dispone de los medios adecuados para conseguir desarrollar la metodología adecuada y así conseguir los propósitos definidos.

Recursos

A nivel de fuente de datos, son importantes los recursos disponibles que brinda la empresa, estos datos sirven de punto de partida para empezar el proceso establecido. Además, estos están sujetos a determinados condiciones de confidencialidad.

Estos datos pueden ser de diferentes tipos: estructurados, semi-estructurados o no estructurados.

Los datos estructurados son aquellos definidos mediante modelos de datos y pueden ser almacenados en las tradicionales bases de datos relacionales.
Por otro lado, los datos **semi-estructurados** son datos que contienen cierta información no modelada, donde es necesaria homogeneizar la información.

Al contrario de estos, **los datos no estructurados** son aquellos que no están definidos y pueden tener diferentes modelos de datos y por lo tanto no pueden ser almacenados en una base de datos relacional, un ejemplo de estos datos son los contenidos de una red social como Facebook, donde no solo puede haber información de un solo modelo (por ejemplo textos), sino que pueden registrarse distintos tipos de datos como son: fotos, videos, imágenes, datos de geolocalización, enlaces, correos electrónicos, etc.

Estos datos se verán en el proceso de adquisición, donde se detallan que tipos de datos son y en que formato se encuentran. Además de las descripciones de cada una y como se implementarán en las plataformas tecnológicas.

A nivel de recursos tecnológicos, en este proyecto se utiliza una máquina virtual con los siguientes recursos:

A nivel de Hardware:

- Ordenador con procesador de 8 núcleos
- Memoria RAM de 16 GB y 500 GB de almacenamiento
- Sistema Operativo Windows 10 de 64 bits
- Conexión Ethernet 10 GB/s

A nivel de Software:
Se implementa la máquina virtual Cloudera[15] mediante la aplicación VirtualBox con las siguientes características:

- Procesador de 4 núcleos
- Memoria RAM de 8 GB y 250 GB de almacenamiento
- Sistema Operativo Linux Red Hat 64 bits
- Conexión Ethernet 1 GB/s
- Memoria de video 8 MB

4. Elección de la plataforma

Como se comentó anteriormente, se utilizará una máquina virtual, consecuentemente, se detalla ésta y las demás herramientas que se utilizarán en el proyecto.

4.1 Herramienta de Virtualización

VirtualBox

VirtualBox es una herramienta de virtualización desarrollada por Oracle Corporation que soporta diferentes arquitecturas (x86/amd64[16]), que sirve instalar sistemas “invitados” dentro de un sistema “anfitrión”.
Versión 5.1

4.2 Ecosistema Hadoop

[15] Versión del producto: cloudera-quickstart-vm-5.4.2-0-virtualbox
[16] x86/amd64, son denominaciones que se dan al distintito tipo de procesador, siendo x86 microprocesadores Intel, amd64 microprocesadores AMD.

Hadoop es un ecosistema Open Source[17] creado por Yahoo en el año 2005 y actualmente bajo desarrollo por la compañia Apache, tiene como principal característica englobar un conjunto de herramientas, aplicaciones y *frameworks* para el desarrollo de sistemas de cómputo distribuido y escalable.

Este ecosistema permite la utilización de aplicaciones para el procesamiento y análisis y almacenamiento de grandes cantidades de datos, ya que, se encuentran distribuidos en clústeres de uno o miles de nodos, con lo cual hace que la plataforma sea tolerable a fallos. Por otro lado, Hadoop esta optimizado para trabajar con variedades de datos y de diferentes fuentes, como datos en *Streaming*[18].

Una de las principales características de Hadoop es su sistema de archivos HDFS (Hadoop Distributed File System), donde se ejecutan la mayoría de aplicaciones.

Por otro lado, se encuentra MapReduce, el cual es un *framework* originalmente desarrollado por Google, que sirve para el desarrollo de aplicaciones y algoritmos.

Hoy en día existen más de 100 proyectos Open Source con referente a Big Data y el número continua creciendo.

Ilustración 8. Ecosistema Hadoop

Como se puede ver en la ilustración 5, el ecosistema Hadoop esa compuesto por estos principales y más populares *frameworks*. Este diagrama esta creado para entender que dichas herramientas forman una pila donde cada componente utiliza las características del componente superior (se entiende que el diagrama es de abajo hacia arriba), esta característica tiene como condicionante que un componente solo puede comunicarse con el siguiente más próximo, de esta manera, podemos observar que el componente HDFS no puede comunicarse con MapReduce.

A continuación, explicaremos brevemente algunos de los principales componentes del ecosistema Hadoop.

- **HDFS**, viene a ser el sistema de archivos en el que se basa Hadoop. Este sistema está basado en una arquitectura maestro-esclavo. Un nodo maestro es quién se encarga de coordinar a los Datanodes que guardarán la información.
  Los datos son replicados en diferentes Datanodes, asegurando la tolerancia a fallos.

[17] Open Source es un término utilizado en cuando se habla de aplicaciones de código abierto, no comerciales y sin costos.
[18] Streaming es relacionado con el flujo de datos en tiempo real.

- **YARN**, es un motor de gestión de recursos de procesos distribuidos, utiliza el componente MapReduce para la comunicación con el sistema de archivos. Con este componente se asegura la correcta distribución del trabajo, gestionando las ejecuciones de los diferentes componentes superiores.

- **MapReduce**, es un modelo de programación funcional que simplifica el cómputo paralelo. Está compuesto por dos funciones: Map quién aplica algoritmos a todos los elementos y Reduce que resume cada operación realizado en esos elementos.

- **Hive**, creado por Facebook es un modelo de programación que sirve de apoyo para las consultas SQL sobre el componente MapReduce, tomando como base modelos algebraicos.

- **Pig**, creado por Yahoo es un modelo de programación que utiliza modelos de flujo de datos usando MapReduce.

- **Giraph**, es un framework que se utiliza para analizar eficientemente grafos a gran escala, provenientes de redes sociales.

- **Spark, Storm y Flink**, están pensados para el proceso de Big Data en datos obtenidos en tiempo real. Estos datos son procesados en memoria, esto quiere decir, son procesados de forma instantánea con el fin de ejecutar aplicaciones Big Data lo más rápido posible.

- **HBase, Casandra y MongoDB**, son sistemas de almacenamiento NoSQL, donde los modelos datos son almacenados en diferentes formatos (XML, JSON, etc.)

- **Zookeeper**, es un sistema de gestión centralizado que monitorea la ejecución de cada componente en el ecosistema Hadoop. Se encarga de la sincronización, configuración y alta disponibilidad de las aplicaciones.

4.3 Distribuciones Hadoop

4.3.1 Cloudera

Cloudera es una compañía que proporciona software para trabajar con grandes cantidades de datos, donde se utiliza herramientas Big Data que están enfocadas al ámbito empresarial.

Actualmente esta empresa da soporte a numerosas tecnologías y en su mayoría al ecosistema Apache Hadoop. Cloudera lleva muchos años siendo patrocinador de la compañía Apache, ofreciendo certificaciones sobre la plataforma Hadoop a través de Cloudera university.

Cloudera Manager

Herramienta que sirve para administrar clústeres y distribuciones Hadoop, permitiendo a través de su interfaz la instalación del clúster desde cero, permitiendo la configuración, monitoreo y la gestión de servicios (añadir o quitar nodos).

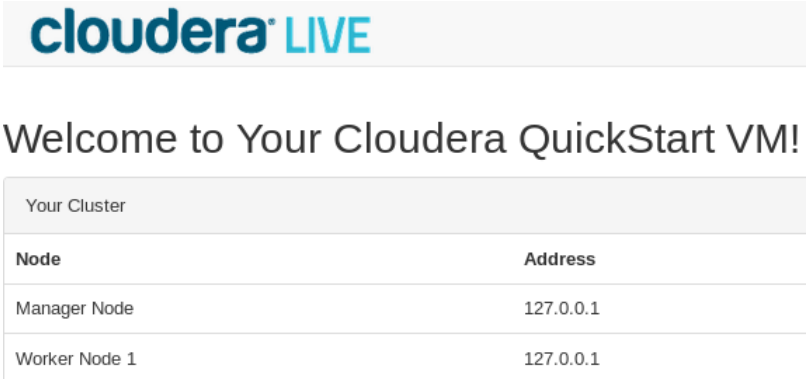


Ilustración 9. Cloudera Live

Esta herramienta solo está disponible en los productos de pago, ya que tienen licencia comercial.

4.3.2 Hortonsworks

Fundada en 2011, es otra de las distribuciones más usadas que ha emergido gracias a estar entre los más vendidos entre las distribuciones Hadoop.
Esta distribución provee una plataforma Open Source basada en Apache Hadoop para el análisis, almacenamiento y gestión del Big Data. Se puede decir que Hortonsworks es el único distribuidor que oferta su plataforma de forma completa siendo una plataforma no propietaria.

Esta distribución utiliza las últimas innovaciones en el ámbito Big Data, teniendo herramientas como YARN, o STORM.

4.3.3 Pivotal

Es una empresa que cuenta con muchos años de experiencia dentro del ámbito Big Data, es colaboradora de VMWare, y ofrece soluciones en la nube. Su plataforma Pivotal Enterprise trabaja mediante Hadoop 2.0.

Esta plataforma utiliza un centro de comando o *Command Center* que administra y monitoriza las distintas herramientas. Incluye aplicaciones para la consulta de datos y librerías de funciones para el análisis de datos.

4.3.4 Comparativa

A continuación se presenta una tabla comparativa con estas tres plataformas poniendo énfasis en las mejores opciones que tienen de acuerdo a las necesidades del proyecto.

	Cloudera	**Hortonsworks**	**Pivotal Enterprise**
Popularidad	Alta	Alta	Media
Licencia	Comercial y Open Source	Open Source	Comercial
Rendimiento	Alto	Alto	Alto
Tolerancia a fallos	Medio	Medio	Medio
Curva de aprendizaje	Baja	Media	Alta

Tabla 1. Breve comparativa de plataforma Big Data

¿Por qué se elige cloudera?

Se descarta Pivotal Enterprise puesto que tiene licencia comercial y sería necesario descargar una versión limitada del producto.

A pesar que ambas herramientas utilizan Apache hadoop como ecosistema, tiene distintos productos o distribuciones, detrás de ellas existen una gran comunidad que los respalda.

Por otro lado cloudera tiene software propietario llamado Cloudera Manager el cual facilita el uso de los productos integrados. Hortonsworks no tiene un gestor de aplicaciones.

Cloudera tiene una licencia comercial para sus productos empresariales, además tiene una versión para propósitos educacionales y para proyectos Open Source (la versión instalada) libre de costo. Hortonsworks tiene licencia Open Source y es completamente gratis.

Cloudera, al ser muy popular, cuenta con mucha información y sobre todo tutoriales donde la curva de aprendizaje no es tan elevada.

Por último, se elige Cloudera en el uso de la herramienta Apache Spark representa una curva de aprendizaje no alta, con respecto a sus competidores.

5. Instalación de la plataforma

De modo general, en este apartado detallamos las instalaciones de los distintos programas que utilizaremos en el trascurso del trabajo. Por otro lado, en el apartado *B. Procesos de instalación* del anexo, se encuentran los pasos detallados de cada instalación.

5.1. Herramienta de virtualización

Como ya se había comentado anteriormente, utilizamos la máquina virtual VirtualBox, con el objetivo de instalar el sistema operativo Linux que albergará las herramientas ofrecidas por Cloudera.

5.2. Cloudera

A través de Cloudera tendremos acceso a diferentes herramientas, en especial trabajaremos con PySpark para realizar el análisis correspondiente en fases posteriores.
Es necesario resaltar que, al ser un sistema ya pre-configurado, podremos hacer uso del lenguaje de programación Python[19], que mediante sus librerías ya incluidas, proporcionan funcionalidades que permiten gestionar los datos a analizar.

5.3. Splunk

Splunk es una aplicación de exploración, monitoreo, visualización, reportes y análisis de grandes conjuntos de datos (Big Data) provenientes de aplicaciones, servidores, páginas web y dispositivos de centros de datos (bases de datos), en la nube y dispositivos IoT o Internet de las cosas.

Esta aplicación recolecta los datos y los indexa a gran escala para facilitar la búsqueda, monitorización, análisis y visualización en tiempo real de los datos indexados. Estos datos pueden venir de otras aplicaciones o ecosistemas como Hadoop.

[19] Python es un lenguaje de programación interpretado, orientado a objetos e imperativo.

Ilustración 10. Diagrama Splunk

Esta herramienta tiene como objetivo hacer accesible grandes conjuntos de datos a las diferentes áreas de una organización, permitiendo las identificaciones de patrones, medidas estadísticas y evaluación de problemas, además de la provisión de BI corporativa.

Splunk puede ser instalado de forma local o mediante un dominio corporativo. Existen diferentes versiones de esta aplicación, por un lado está la versión *Enterprise* que es de pago o mediante un determinado tiempo de prueba. Existe la versión libre que es limitado a 500MB de datos al día.

Para la elección de esta herramienta se ha realizado una comparativa con otras similares, como son Tableu o Qlik.

Características	**Tableu**	**Qlik**	**Splunk**
Licencia	Limitada para estudiantes	Periodo de prueba	Libre
Soporte de datos Big Data	Solo BI	Solo BI	Si
Curva de aprendizaje	Media	Alta	Media
Soporte comunidad	Si	Si	Si
Editor de visualizaciones	Si	Si	Si
Tutoriales propios	Si	Si	Si
Tipo de instalación	Local, aplicación	Local, aplicación	Local y Remota, mediante navegador
Consumo de memoria	Alta	Media	Baja

Después de esta comparativa con diversas aplicaciones similares, optamos por Splunk puesto que es libre y la limitación no afecta al trabajo a realizar. Además, es importante el

soporte de la comunidad, tutoriales y sobre todo una curva de aprendizaje media. Otro punto a favor es que se trabaja mediante el navegador, teniendo un consumo mínimo de memoria.

La instalación de esta aplicación se encuentra en el *anexo B de procesos de instalación*.

5.4. KNIME

KNIME Analytics es una plataforma de código abierto para el análisis de datos, creación de reportes y visualización. Esta plataforma utiliza una interfaz gráfica basada en *drag and drop* o el arrastre de sus diferentes herramientas para facilitar la construcción de la estructura de la solución del análisis.

Ilustración 11. KNIME

KNIME utiliza cada componente como nodos que son conectados a otros nodos para construir cada fase de un flujo de trabajo. Cada nodo cumple una funcionalidad, por ejemplo el nodo *file reader* se encarga de importar la tabla que contendrá la fuente de datos.

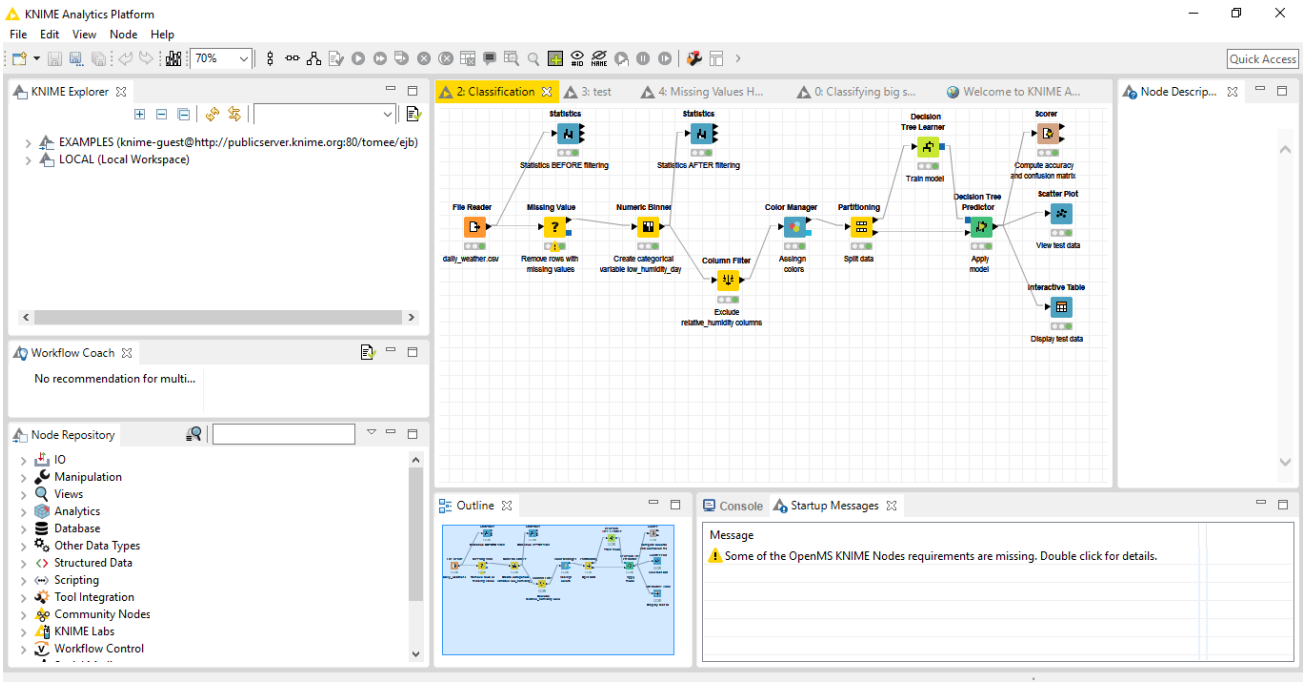


Ilustración 12. Diagrama KNIME

Como se muestra en la imagen anterior, cada nodo es seleccionado desde el repositorio de nodos (lateral izquiero, abajo). Cada nodo es configurado haciendo clic derecho sobre si, una vez completada la configuración, este nodo o nodos son ejecutados mediante la barra superior.

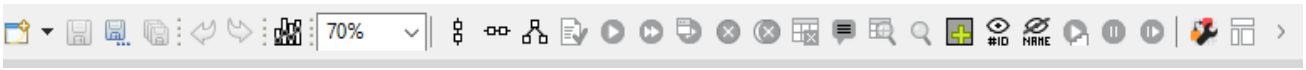


Ilustración 13. Barra herramientas KNIME

¿Por qué utilizamos KNIME?

A diferencia de otras plataformas, KNIME es de código abierto, a pesar que puede ser limitado para grandes conjuntos de datos, es más que suficiente para realizar nuestro trabajo. Por otro lado, es compatible con distintos sistemas operativos y sobre todo, su interfaz es muy intuitiva y fácil de usar, teniendo una curva de aprendizaje muy baja.
La instalación de esta plataforma se encuentra en el *anexo B. Procesos de instalación.*

5.5. Spark MLlib

Spark, perteneciente a Apache Software Foundation, es una herramienta útil y eficiente para realizar tareas de procesamiento masivo de datos, basada en MapReduce, permite realizar consultas, análisis iterativos y procesamientos de datos en tiempo real.

Esta herramienta utiliza computación en memoria con lo cual agiliza el procesamiento de algoritmos especialmente en grandes conjuntos de datos. Por otro lado, Spark provee soporte para las diferentes librerías de Python, Scala, Java y SQL.

Como podemos observar en la imagen siguiente, Spark trabaja por debajo de distintos componentes, cada uno de estos componentes tienen acceso a toda la infraestructura que implementa Spark (clústeres de datos).

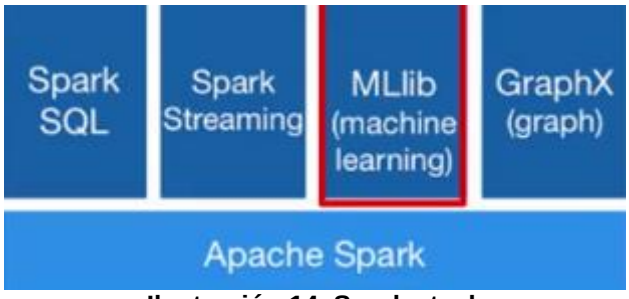


Ilustración 14. Spark stack

MLlib es una librería que permite trabajar con algoritmos Machine Learning para crear modelos de datos, esta librería no tiene una interfaz gráfica de usuario, consecuentemente, se utilizará Jupyter Notebook que es un entorno interactivo web de ejecución de código, estos son almacenados en la web y ser editados mediante cualquier navegador.

Spark viene incluido en el ecosistema de Cloudera, con lo cual no es necesaria su instalación. Para utilizar Jupyter Notebook necesitamos instalar Anaconda, que es un paquete que incluye librerías para trabajar con Python y se caracteriza porque trae soporte para el tratamiento de grandes volúmenes de datos en un solo paquete.

Para ver si tenemos instalados Anaconda, hay que ingresar desde el terminal en Cloudera y escribir conda info, como se muestra la imagen siguiente:

```
[cloudera@quickstart ~]$ conda info
Using Anaconda Cloud api site https://api.anaconda.org
Current conda install:

             platform : linux-64
        conda version : 4.0.5
  conda-build version : 1.20.0
       python version : 3.5.1.final.0
     requests version : 2.9.1
     root environment : /home/cloudera/anaconda3  (writable)
  default environment : /home/cloudera/anaconda3
     envs directories : /home/cloudera/anaconda3/envs
        package cache : /home/cloudera/anaconda3/pkgs
         channel URLs : https://repo.continuum.io/pkgs/free/linux-64/
                        https://repo.continuum.io/pkgs/free/noarch/
                        https://repo.continuum.io/pkgs/pro/linux-64/
                        https://repo.continuum.io/pkgs/pro/noarch/
          config file : None
    is foreign system : False
```

Ilustración 15. Anaconda

Podemos ver que tenemos la versión 3 instalada, con lo cual podemos empezar a trabajar.

5.6. Neo4j

Neo4j es una herramienta de análisis de grafos, entre las versiones actuales existen una comercial y otra libre para la comunidad y estudiantes. Esta aplicación es una base de datos orientada a grafos que ayuda a entender las relaciones entre datos y extraer su verdadero valor. Recordemos que un grafo se define por uno o varios nodos con una o varias aristas que se comunican entre ellos formando una relación.

Ilustración 16. Neo4j

Esta herramienta es capaz de analizar redes extremadamente complejas con millones de nodos y aristas o relaciones entre nodos. Por tanto, a diferencia de otras herramientas como por ejemplo Gephi o XGraph, Neo4j permite la creación de bases de datos enteras con grandes conjuntos de datos, siendo posibles consultas rápidas para el posterior análisis.

Hoy en día Neo4j se utiliza en grandes compañías como eBay, Walmart, Cisco o HP donde necesitan analizar grandes cantidades de información sin perder escalabilidad.

Actualmente, este tipo de herramientas se utiliza para la detección del fraude, recomendación en tiempo real a través de las redes sociales o la gestión de datos a gran escala.

Por último, el proceso de instalación de esta herramienta se encuentra en el *anexo B de procesos de instalación*. Más adelante se verá todo el potencial cuando se analicen los datos no-estructurados.

6. Desarrollo de la Metodología

6.1. Adquisición

En esta primera fase de desarrollo de la metodología, es importante conocer los datos que utilizaremos, identificar los datos con los cuales trabajaremos es el punto de partida para iniciar esta etapa.

Consecuentemente, se trabajara con los datos obtenidos de la empresa en cuestión, estos datos son semi-estructurados y no estructurados. Estos datos están en formato CSV y detallaremos a continuación.

6.1.1. Características de los datos

Los datos semi-estructurados vienen a ser los registros de la información durante el juego. Estos datos cuentan con 8 archivos, la tabla siguiente lista cada uno de estos archivos con una breve descripción y los campos que contienen:

Nombre del archivo	Descripción	Campos
ad-Clicks.csv	Este archivo contiene todos los datos del usuario cuando hace clic en un anuncio.	timestamp: cuando se registra el clic. txID: id identificador para cada clic userSessionid: Id de la sesión del usuario que ha realizado el clic. teamid: Id del Equipo actual que realize el clic en el anuncio. userid: Id del usuario quién realizó el clic en el anuncio. adID: Id del anuncio.

		adCategory: Categoría del anuncio
buy-clicks.csv	Este archive contiene todos los datos de cada compra realizada por cada jugador en el juego.	timestamp: Fecha cuando se realizó la compra. txId: Id de la compra. userSessionid: Id de la sesión del jugador quién realizó la compra. team: id del equipo que realizó la compra. userid: Id del usuario que realizó la compra. buyID: Id del producto comprado. price: Precio del producto.
users.csv	Este archive contiene los datos de cada jugador.	timestamp: Registro de la fecha que un usuario juega por primera vez. id: Id del jugador. nick: Alias del jugador. twitter: Identificador twitter del jugador. dob: Fecha de nacimiento del usuario. country: Identificador de dos letras del código del país del jugador.
team.csv	Archivo que contiene todos los datos de cada equipo creado o eliminado en el juego.	teamid: Id identificador para cada equipo. name: Nombre del equipo. teamCreationTime: Registro de creación del equipo.

		teamEndTime: Fecha cuando el último usuario dejo el equipo. Strength: Medida de jugabilidad de un equipo, se basa en el éxito del mismo. currentLevel: Nivel actual de cada equipo en el juego.
team-assignments.csv	Este archive contiene los datos de cada jugador al asignarse a un equipo determinado. Un jugador puede crear su propio equipo de un solo miembro.	time: Fecha de unión de un jugador a un equipo determinado. team: Identificador del equipo userid: identificador del usuario Assignmentid: identificador de asignación.
level-events.csv	Este archive contiene los datos del nivel del juego donde se encuentra cada jugador o equipo.	time: Registra cuando se inicia o acaba un nivel eventid: Identificador de la evento teamid: tidentificador de cada equipo. teamLevel: Nivel iniciado o completado. eventType: Tipo de evento, puede ser de inicio o final de cada nivel.
user-session.csv	Este archive contiene los datos de sesión del usuario, niveles y la plataforma que usa, entre otros datos.	timestamp: Registra la sesión userSessionid: identificador de sesión. userId: Id del jugador. teamId: Identificador del equipo. assignmentid: identificador que registra la unión de un usuario a un equipo.

		sessionType: Indica si la sesión inicia o termina. team_level: Nivel del equipo durante la sesión platformType: Tipo de plataforma durante la sesión.
game-clicks.csv	Registra cada clic que realiza el usuario o equipo durante los distintos niveles durante el juego.	timestamp: Registra la fecha cuando ocurre el clic clickid: Identificador del clic. userid: identificador del jugador. usersessionid: Identificador de la sesión. isHit: indica si el clic es efectivo (1) o no es efectivo (0) a la hora de completar una determinada tarea. teamId: Identificador del equipo. teamLevel: Actual nivel del equipo.

Por otro lado, tenemos los datos no-estructurados que constan de la información obtenido de los registros de conversaciones creadas en cada “chatroom” de cada equipo. Por razones de privacidad, no se incluyen los mensajes, sino los registros de cada conversación que nos servirán para el análisis de ciertos tipos de comportamientos que solo pueden ser observados mediante ciertos modelos de datos, en este caso, utilizaremos el análisis de grafos, estos análisis ser realizarán en los siguientes capítulos.

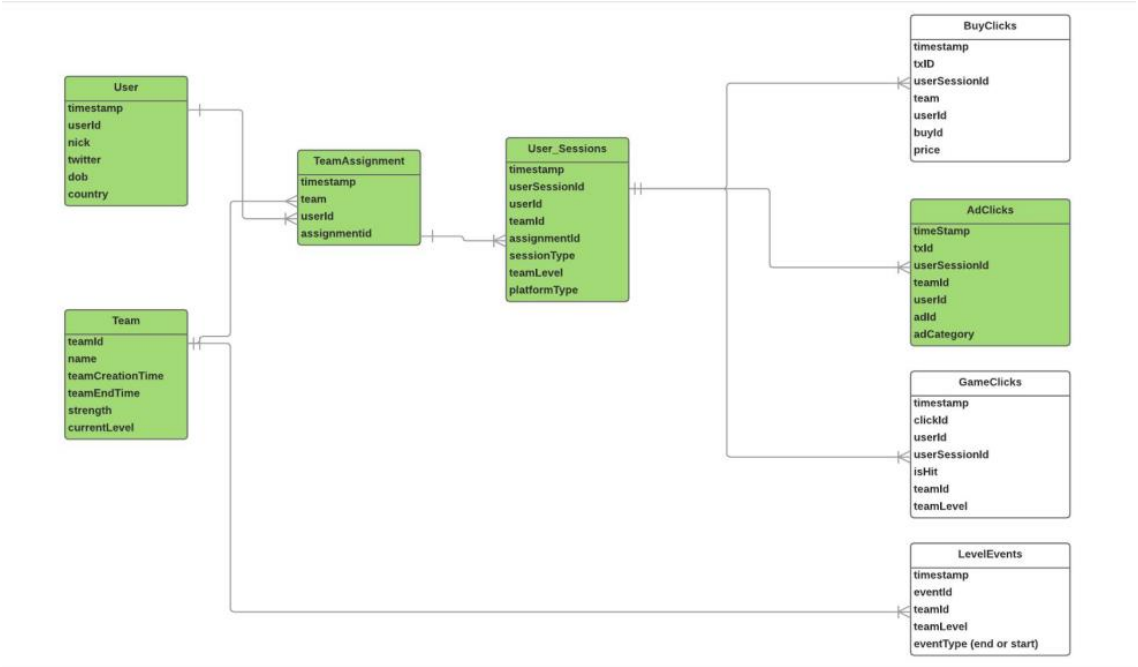
La información consta de 6 archivos en formato CSV. En la siguiente tabla se aprecia cada archivo con su descripción y el cada uno de los campos que contiene:

Nombre del archivo	Descripción	Campos
chat_create_team_chat.csv	Registra la creación de un chat por parte de un jugador.	userId: identificador del jugador. teamId: Identificador del equipo. teamChatSessionId: identificador de la sesión del

		chat. timestamp: Registra la fecha del evento
chat_item_team_chat.csv	Registra la creación de los nodos ChatItems y del arista “Partof”	userId: identificador del jugador. teamChatSessionId: identificador de la sesión del chat. chatItemId: identificador de cada nodo creado. timestamp: Registra la fecha del evento
chat_join_team_chat.csv	Registra la creación de las aristas “Joins” desde user hacia teamChatSession. Indica la unión de un jugador a un chat	userId: identificador del jugador. teamChatSessionId: identificador de la sesión del chat. timestamp: Registra la fecha del evento
chat_leave_team_chat	Registra la creación de la arista “Leaves” desde user hacia teamChatSession. Indica la salida de un jugador del chat.	userId: identificador del jugador. teamChatSessionId: identificador de la sesión del chat. timestamp: Registra la fecha del evento
chat_mention_team_chat	Registra la creación de la arista “Mentioned”, Indica las menciones de cada jugador o equipo en un chat.	chatItemId: identificador de cada nodo creado. userId: identificador del jugador. timestamp: Registra la fecha del evento
chat_respond_team_chat	Registra la respuesta que realiza cada jugador en un chat.	chatId1: Identificador del chat emisor. chatId2: Identificador del chat receptor. timestamp: Registra la fecha del evento

6.1.2. Modelo de datos

A continuación se muestra el diagrama entidad relación de los datos semi-estructurados:



Este diagrama muestra las tablas relacionales conectadas mediante llaves primarias de las distintas tablas a las que se compone. Por ejemplo, la tabla User tiene como llave primaria userId que sirve de identificador de cada usuario.

Por otro lado, se muestra el modelo de datos relacionados con los chat de cada jugador usando el modelo de grafos.

Nodos		
Nombre	**Propiedades**	**Descripción**
User	Id	Jugadores que interactúan en el chat.
Team	Id	Equipos de usuarios en cada sala de chat.
TeamChatSession	Id	Sesión creada por un equipo de jugadores
ChatItem	Id	Mensaje de cada chat representado por Id

Relaciones		
Nombre	**Propiedades**	**Descripción**
CreateSession	Timestamp	Arista "CreatesSession entre el nodo user y TeamChatSession
OwnerBy	Timestamp	Arista "OwnedBy" entre el nodo TeamChatSession y el nodo Team
Joins	Timestamp	Arista "Joins" desde User hacia TeamChatSession
Leaves	Timestamp	Arista "Leaves" desde User hacia TeamChatSession
Mentioned	Timestamp	Arista "Mentionedque va desde chatItem hacia User

PartOf	Timestamp	Arista "PartOf" desde el nodo ChatItem hacia el nodo TeamChatSession.
ResponseTo	Timestamp	Arista "ResponseTo" desde el nodo ChatItem hacia otro nodo ChatItem.
InteractsWith	Timestamp	Arista “InteractsWith " compuesto solo por users.

Gamificación

Rol de jugador dentro de un equipo

Cada usuario es un miembro de al menos un equipo. Cuando un usuario inicia el juego, por sí mismo pertenece a un equipo para el nivel de inicio, recordemos que este usuario puede unirse a cualquier equipo en el siguiente nivel.

Existen tres distintos grupos de jugadores divididos en dos categorías:

- Jugador dentro de un equipo
  - Jugando
  - No jugando
- Jugador fuera de un equipo
  - No asignado

Esta característica es importante para saber qué tipo de usuario es en cada nivel del juego, sí es un usuario activo, o este pertenece a un equipo.

Niveles de juego a nivel de equipos

Cuando un usuario está en un equipo, tiene una sesión de equipo con el cual se sabe cuándo se inicia el juego y cuando se acaba.

En cualquier otro caso, cuando un usuario pertenece a un equipo, estos tendrán un equipo asignado desde la primera vez que se han unido al mismo. Cada vez que un usuario se sale de un equipo, esta asignación es borrada.

Cada vez que un usuario sube de nivel, se registran los cambios; se guardan el nivel acabado y el nuevo. Al mismo tiempo, todos los jugadores que inician o terminan una sesión de juego, esta es registrada.

6.2. Preparación

Siguiendo con la fase metodológico, en esta etapa veremos la preparación de los datos que consta de dos partes: la exploración y el pre-proceso de los datos.

6.2.1. Exploración

En esta fase se realiza el primer análisis de la información con el fin de entender la naturaleza de los datos, esto quiere decir, qué significa la información adquirida, la calidad

y el tipo de formato con el cual está compuesta esta información. A este tipo de análisis preliminar le llamamos exploración.

Consecuentemente, buscamos aspectos claves como tendencias, correlaciones y otros aspectos estadísticos que puedan facilitar información útil para describir estos datos. Los resúmenes estadísticos son importantes ya que capturan distintas características a través de indicadores numéricos. Además, utilizando diagramas facilita el entendimiento de esta información.

Para realizar todos estos primeros análisis utilizaremos la aplicación Splunk, la instalación de la aplicación ficheros se encuentran en el apartado *B.3. Splunk del anexo de procesos de instalación.*

Para la carga de ficheros, nos tendremos que autenticar como admin en la aplicación y nos debemos dirigir a pestaña *settings* y luego a *add data*:

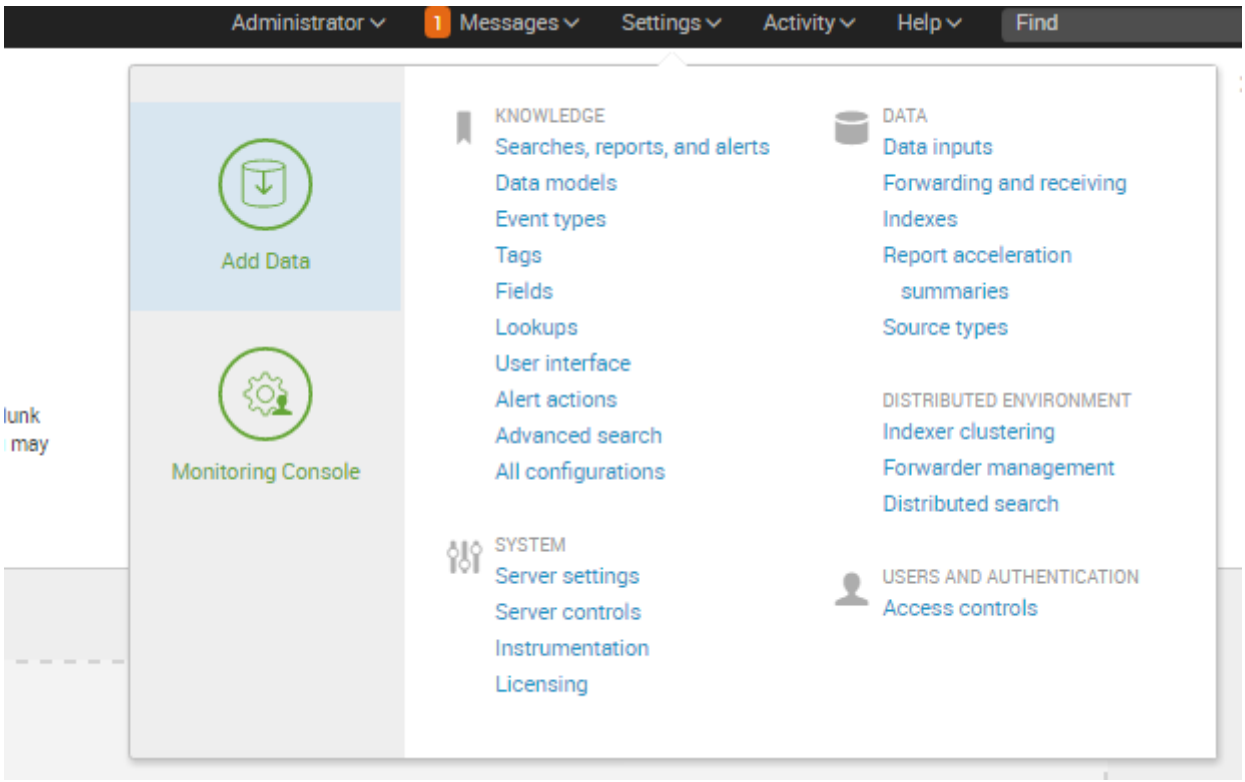


Ilustración 17. Opciones de carga de datos Splunk

Luego daremos clic en *upload* para luego, en la siguiente ventana, arrastrar el archivo en cuestión o podemos seleccionar el archivo para abrirlo desde la dirección correspondiente:

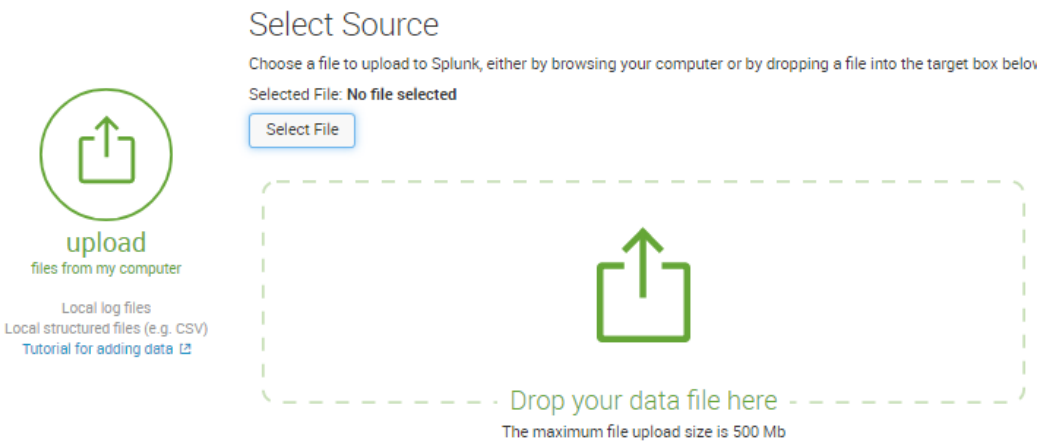


Ilustración 18. Carga de datos Splunk

En el siguiente paso, podremos configurar el formato del archivo origen, por defecto dejaremos como esta:

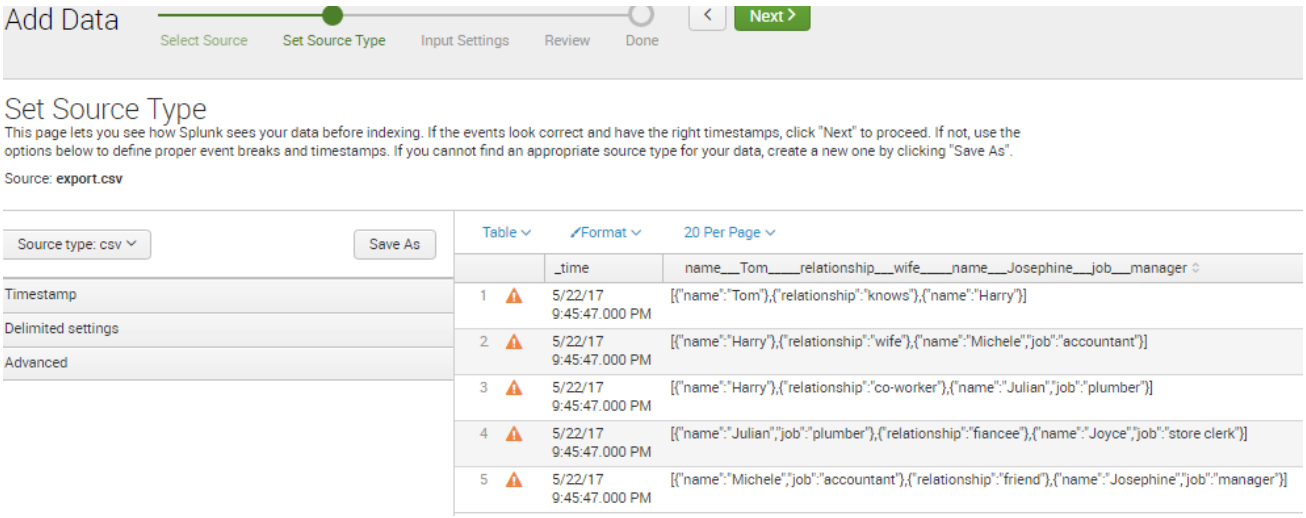


Ilustración 19. Configuración de datos Splunk

Después de esto nos mostrará un breve resumén de los datos importados. Este proceso se realizará para cada uno de los archivos.
Para ver los archivos añadidos, debemos ir a *Search and reporting* y luego podemos ver la opción *summary*:

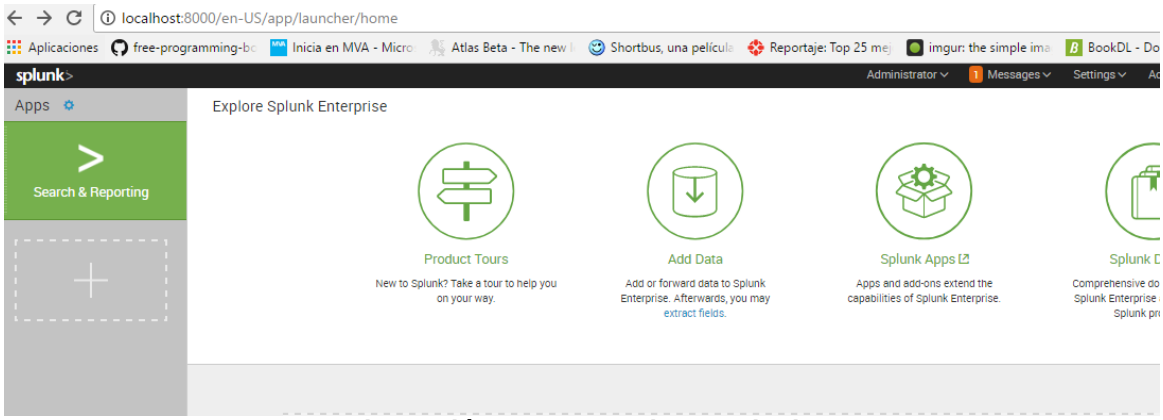


Ilustración 20. Busqueda en Splunk

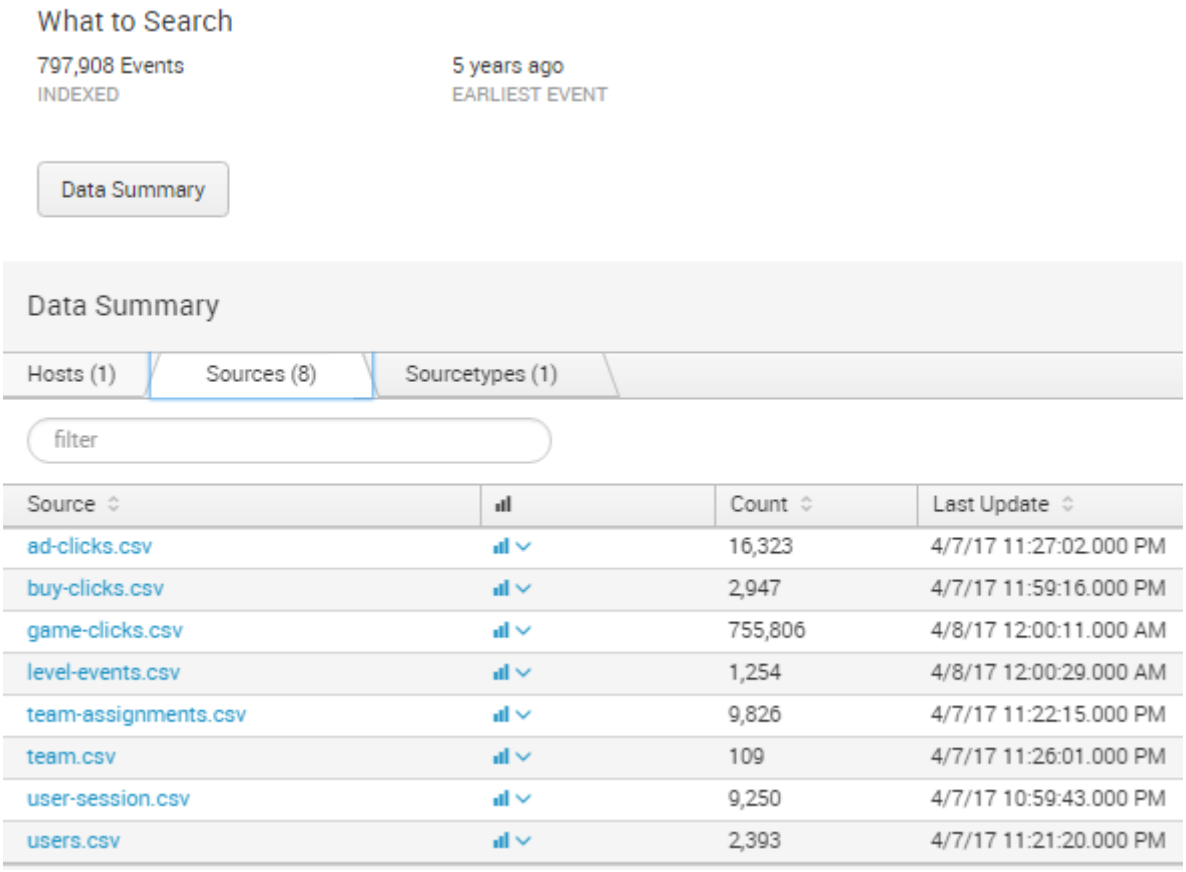


Ilustración 21. Archivos subidos Splunk

Una vez que tenemos los datos cargados, se procede a realizar la exploración de los datos. En este punto es importante revisar aquellos objetivos que nos planteamos al inicio del trabajo, recordando que uno de los puntos importantes es enfocarnos en mejorar los ingresos, el compromiso del usuario y la jugabilidad del mismo.

Viendo estas variables debemos preguntarnos cuál es la situación actual de los datos, es importante saber más acerca de aquellos ficheros críticos, como son los que contienen la información respecto a la cantidad de clic que hace el usuario, las compras que realiza, aquellas plataformas más usadas, etc.

Queremos saber las plataformas más usadas por cada jugador, para esto debemos seleccionar el archivo user-session que contiene cada una de las plataformas.

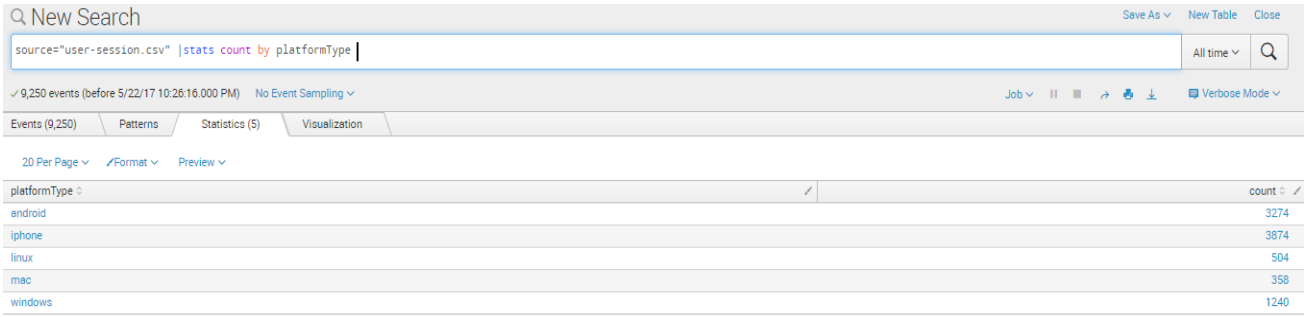


Ilustración 22. Plataformas más usadas

Observamos que la lista la encabezan las plataformas IPhone (3874) e Android (3274) seguidas de Windows (1240) quienes son las más utilizadas.

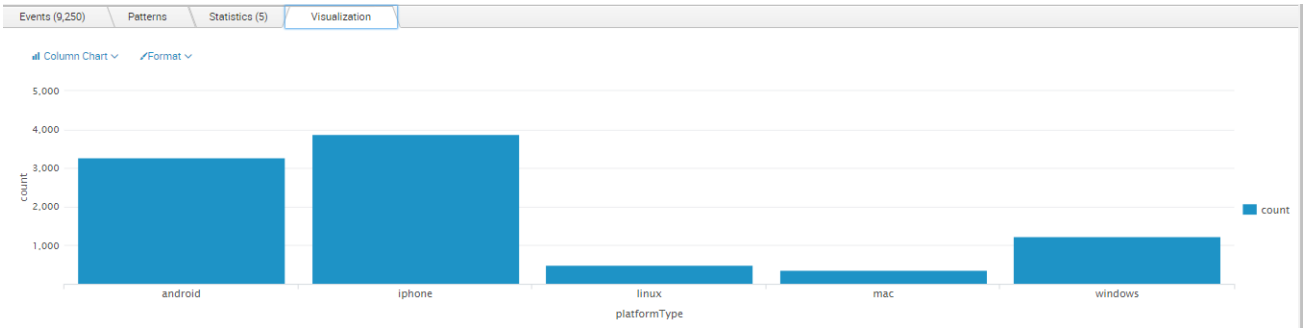


Ilustración 23. Diagramas de las plataformas más usadas

Queremos saber cuáles son las categorías de anuncios más clicadas o seleccionadas. Para esto debemos seleccionar como el archivo ad-clicks.csv que contiene las categorías de cada anuncio:

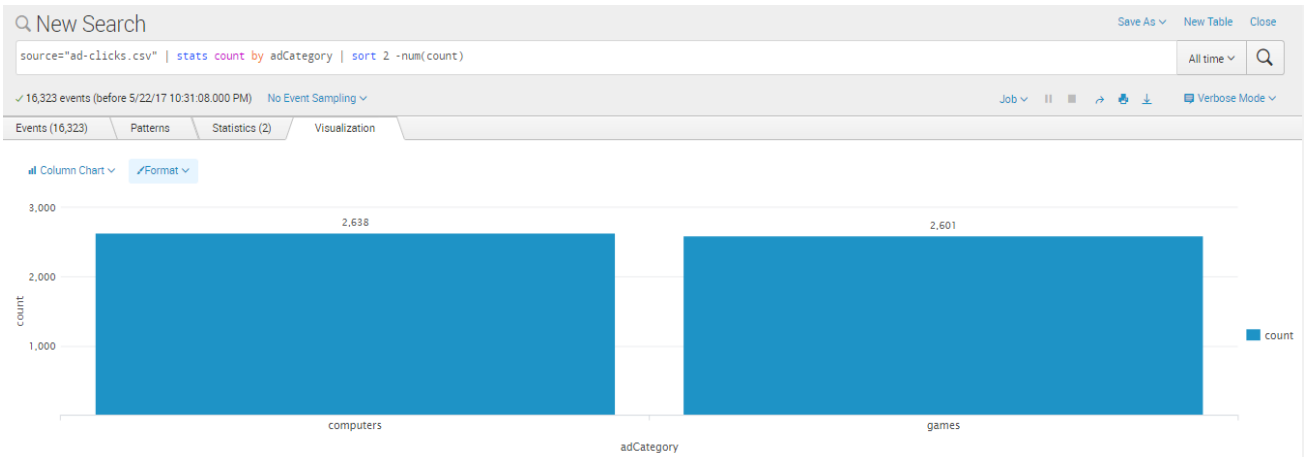


Ilustración 24. Categorías más vistas

Como podemos apreciar, existen dos mayores categorías, siendo *computers* (2638) la más clicada y en segundo lugar esta *games* (2601). Ahora se muestra por cada categoría:

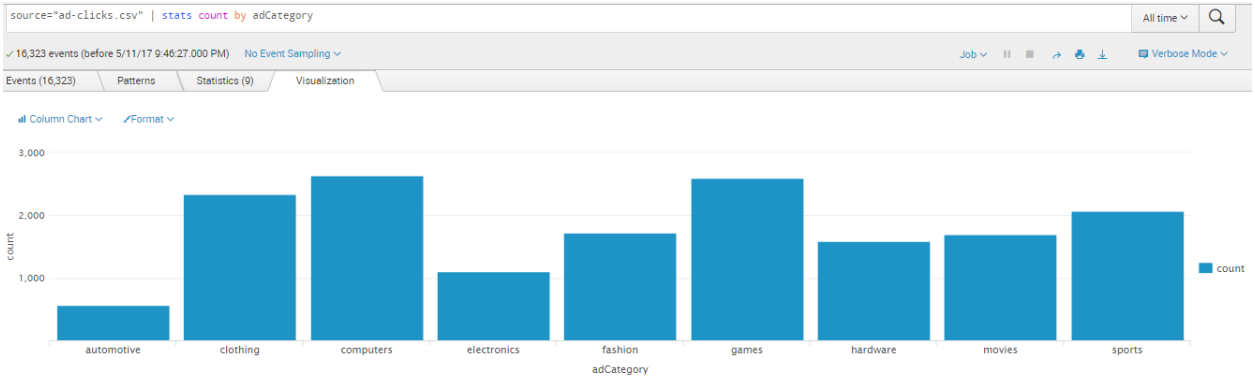


Ilustración 25. Categorías de anuncios

Según esto, podemos calcular los ingresos según el conjunto de valores de acuerdo a cada categoría. En el primer escenario, todas las categorías tienen igual valor, al contrario del segundo escenario.

Scenario #	Electronics	Fashion	Automotive	Total Revenue
1 - even	0.50	0.50	0.50	4928,25 €
2 - uneven	0.55	0.60	0.55	5184,10 €

Queremos saber cuáles son los productos más comprados, para esto debemos seleccionar la tabla o el archivo *buy-clicks.csv* que contiene los datos identificadores de los productos.

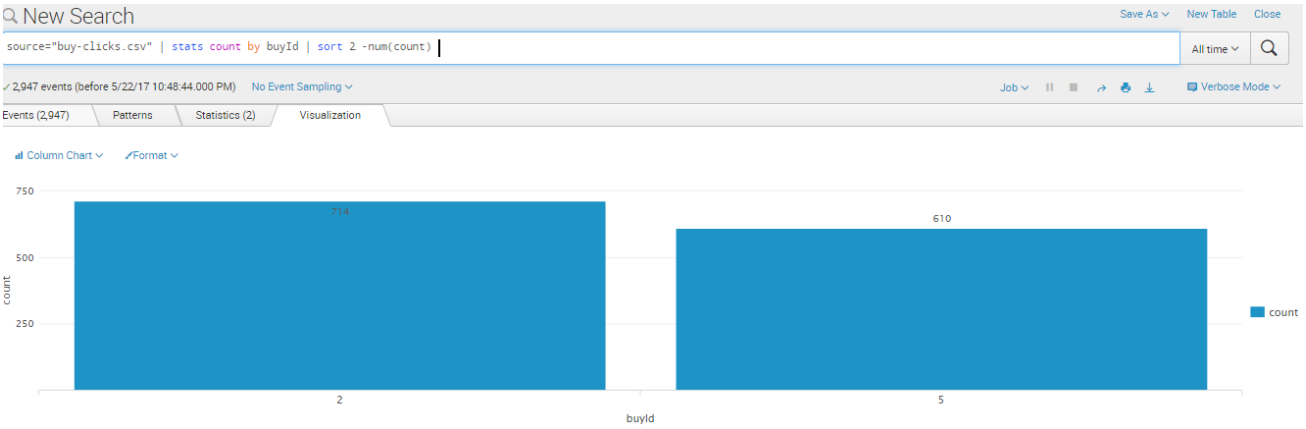


Ilustración 26. Items más comprados

Podemos observar que existen dos productos más comprados, siendo el ítem 2 con 714 veces comprado, seguido del ítem 5 con 610 veces.
Asimismo, podemos ver la cantidad de ítems que están disponibles para ser comprados. Podemos observar que existen seis ítems.

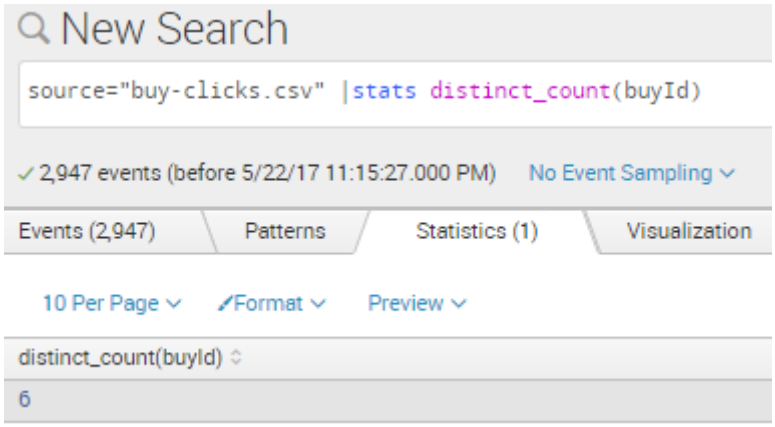


Ilustración 27. Cantidad de productos disponibles

Queremos saber el total de dinero gastado en la compra de items. Podemos observar que el total asciendo a 21407 euros.

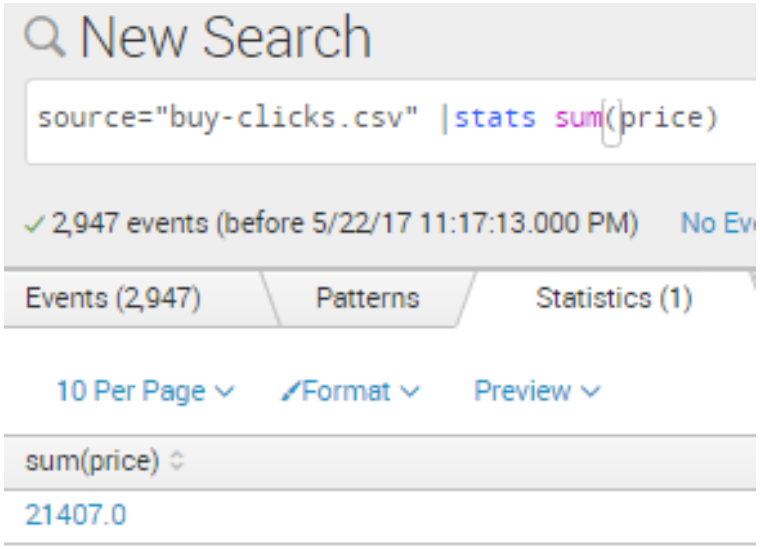


Ilustración 28. Total de dinero gastado

Queremos saber cuántas veces cada ítem es comprado.

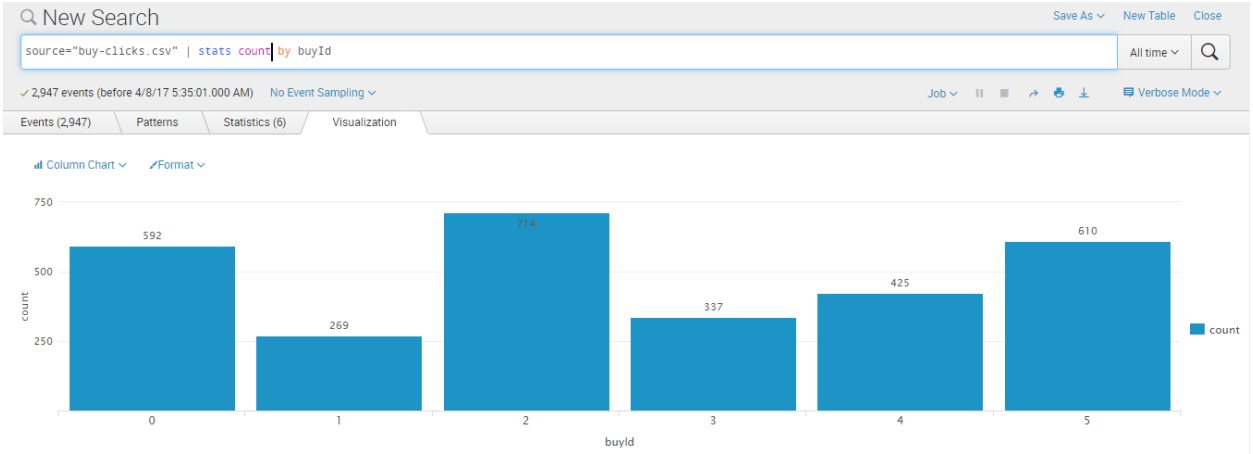


Ilustración 29. Veces de producto comprado

Ahora queremos saber cuánto dinero se ingresó por cada ítem. Observamos que el ítem con más ingresos es el que tiene el id 5, con 12200 euros.

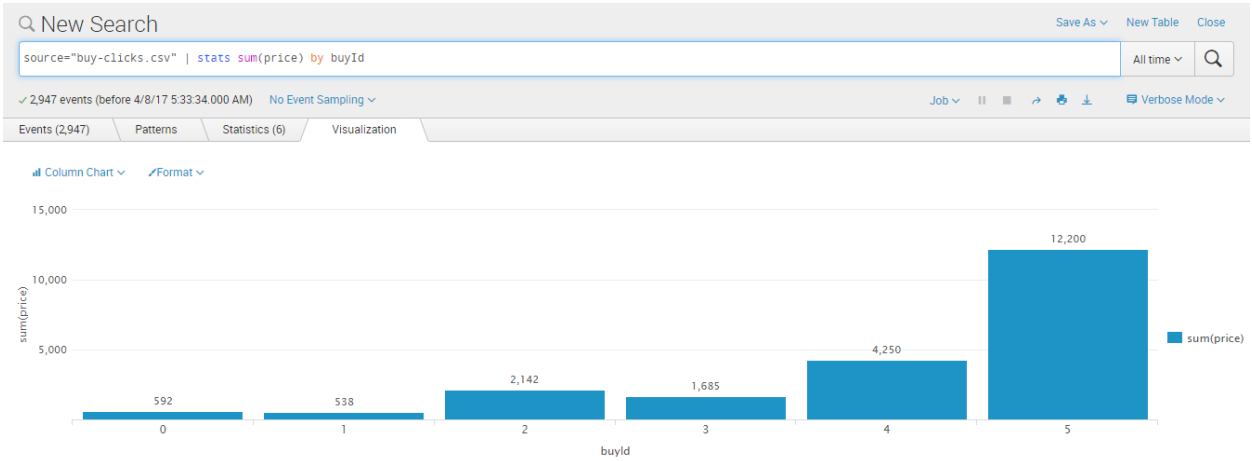


Ilustración 30. Total ingresos por productos

Queremos saber el total de dinero gastado, listando los 10 primeros usuarios. Observamos que el usuario 2229 encabeza el ranking con más de 200 euros gastados.

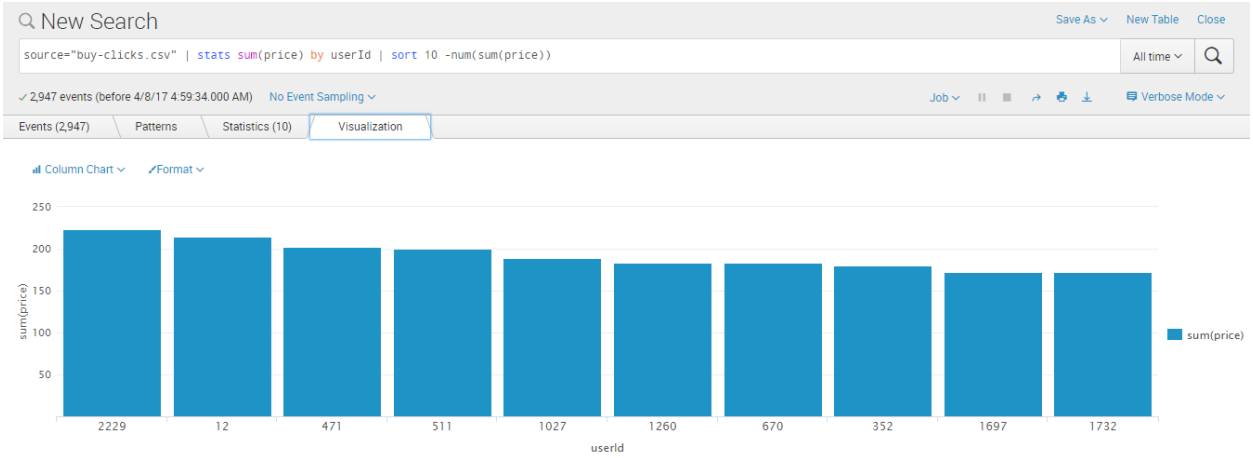


Ilustración 31. Gasto total de los diez top usuarios

En resumen, se saben datos claves para poder enfocar-nos en los siguientes análisis, así como podemos saber más sobre la información y el estado de ingresos de la compañía, el costo de cada producto, los usuarios más consumidores o aquellos que utilizan más una u otra plataforma. Esta información será esencial a la hora de agrupar datos para las siguientes etapas de la metodología.

6.2.2. Pre-Proceso

Una vez que se ha logrado entender los datos y se sabe que es lo que existe detrás de ellos, el siguiente paso es el pre-procesado de la información para los próximos análisis.

Este pre-proceso incluye "limpiar" los datos de valores nulos, duplicados, datos inconsistentes, ruidos en los datos etc. Además, es necesario filtrar la información, agregar nuevos subconjuntos de datos que puedan aportar nuevos conocimiento o darles un formato específico dependiendo el tipo de análisis que se realizará.

En este caso, se utilizará KNIME como plataforma para la limpieza, reportes, visualización y pre-análisis de los datos. La instalación de esta herramienta se describe en el apartado *B.4. KNIME del anexo de proceso de instalación*.

Para los análisis siguientes se utilizará el archivo combined_data.csv que es la unión de las tablas user-session.csv, buy-clicks.csv y game-click.csv, con el fin de procesar estos datos para el análisis de los ingresos producidos por el usuario mediante las compras producidas por los anuncios clicados durante el juego.

Estos datos se han visto en detalle en la fase anterior de adquisición. De modo resumen veremos que columnas tiene esta nueva tabla y una breve descripción:

Columna	Descripción
userId	Identificador de usuario
userSession	Identificador de sesión del usuario
team_level	Nivel de juego del equipo
platformtype	Plataforma usada por el usuario
count_gameclicks	Total de clics por sesión
count_hits	Total de golpes por sesión
count_buyid	Total de compras por sesión
avg_price	Promedio del precio de compra por sesión

Estos datos pueden presentar valores nulos que tienen que ser tratados para que en el análisis no exista inconsistencias. Para ellos es necesario importarlos a la plataforma y borrar aquellos campos vacíos.

Para importar los datos, insertaremos el nodo *File reader* desde el repositorios de nodos

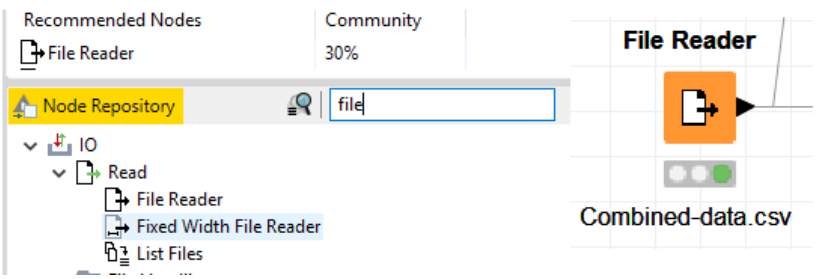


Ilustración 32. Importación de datos KNIME

Al darle doble clic veremos el archivo donde buscar los datos para importarlos.

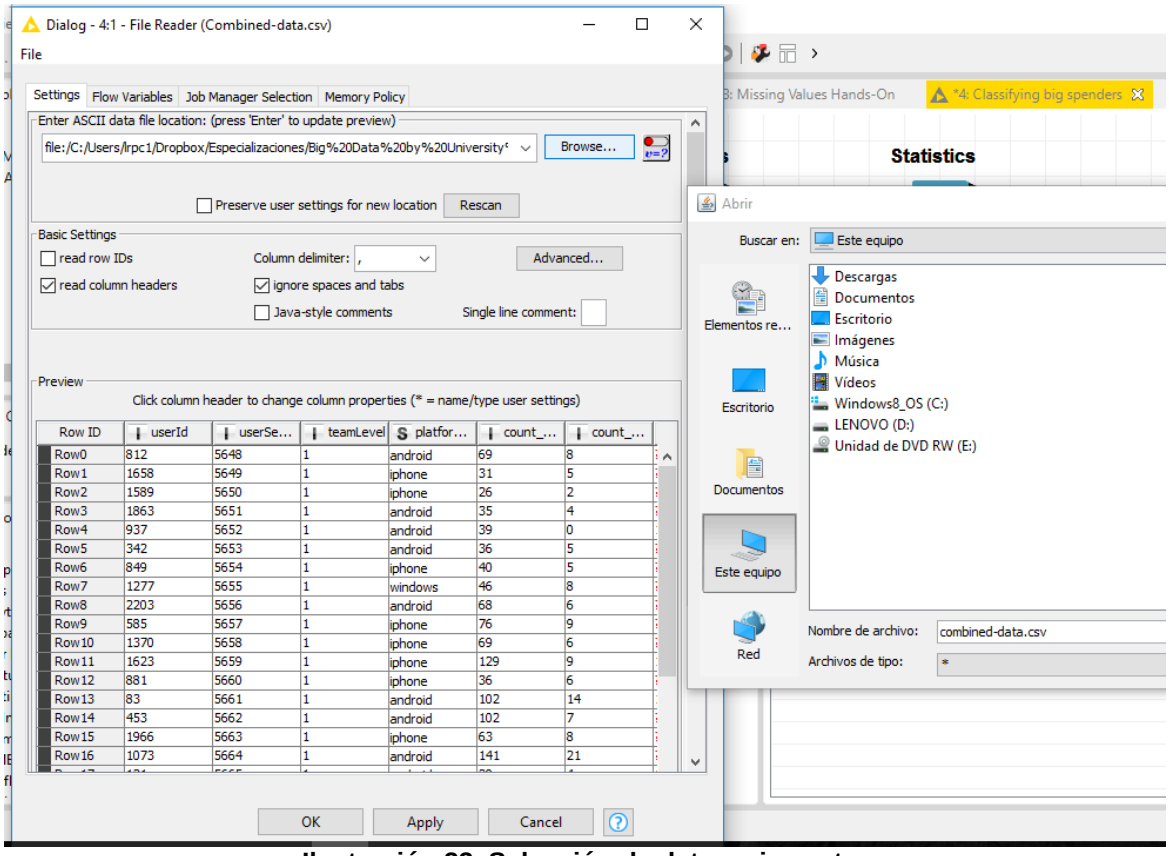


Ilustración 33. Selección de datos a importar

Para verificar si los datos tienen valores nulos, es necesario insertar un nodo de estadísticas y conectarlo al nodo anterior

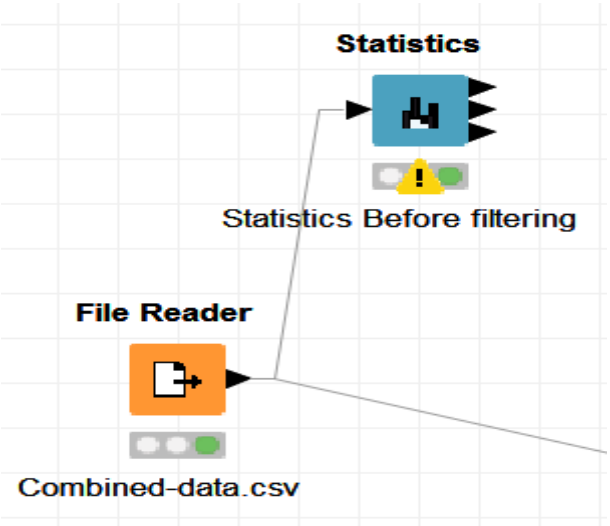


Ilustración 34. Nodo para ver valores nulos

Al darle clic derecho, veremos la opción para ver las estadísticas de la tabla. Como se puede apreciar en la imagen siguiente, existen valores nulos las columnas count_buyid y avg_price

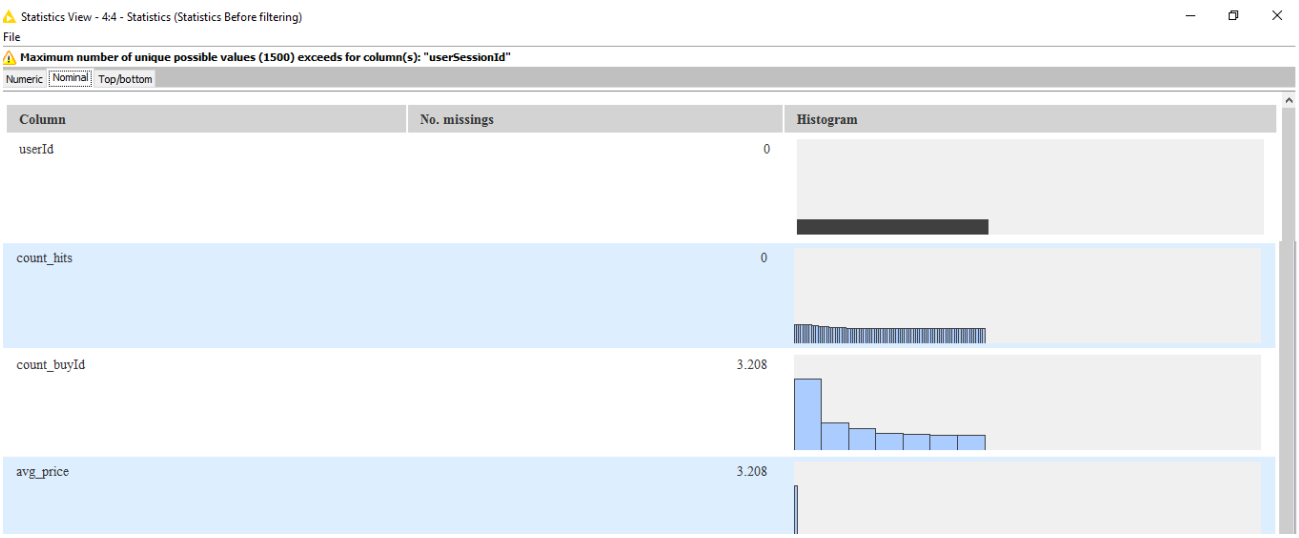


Ilustración 35. Estadisticas de datos en KNIME

Para eliminar estos valores es necesario insertar un nuevo nodo *Row Filter* y otro nodo estadístico para comprobar los datos nulos borrados. Cabe recordar que estos nodos se encuentran en el repositorio de nodos.

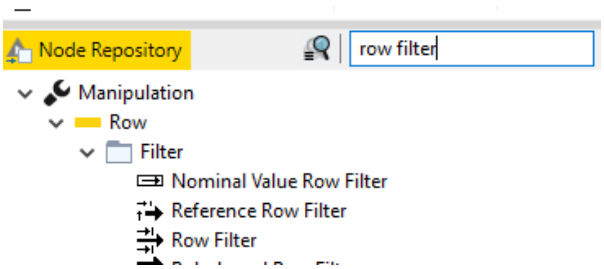


Ilustración 36. Repositorio de nodos

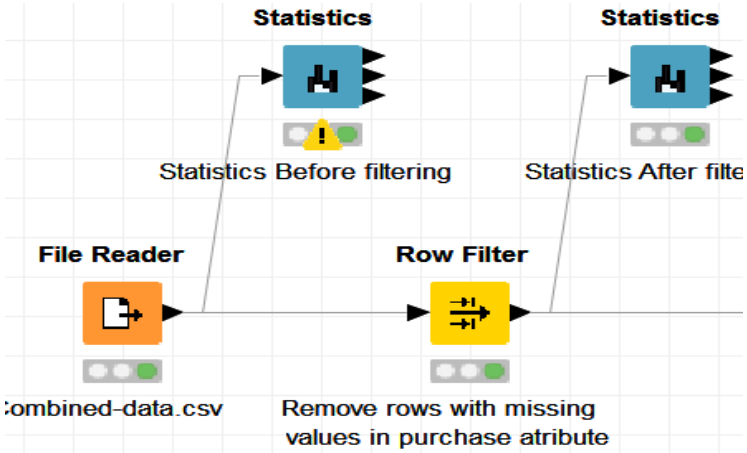


Ilustración 37. Inserción nodo row filter

Dar doble clic en el nodo *Row Filter* para configurarlo, Se debe asegurar de seleccionar aquella columna donde se desean eliminar los valores nulos y que sean necesarias para el análisis, en este caso sería avg_price. Se marca *Exclude rows by attribute value* para excluir las filas donde exista este valor nulo y se marca la opción *Only missing values match*, para eliminar solo los valores nulos. Por último se da a aplicar.

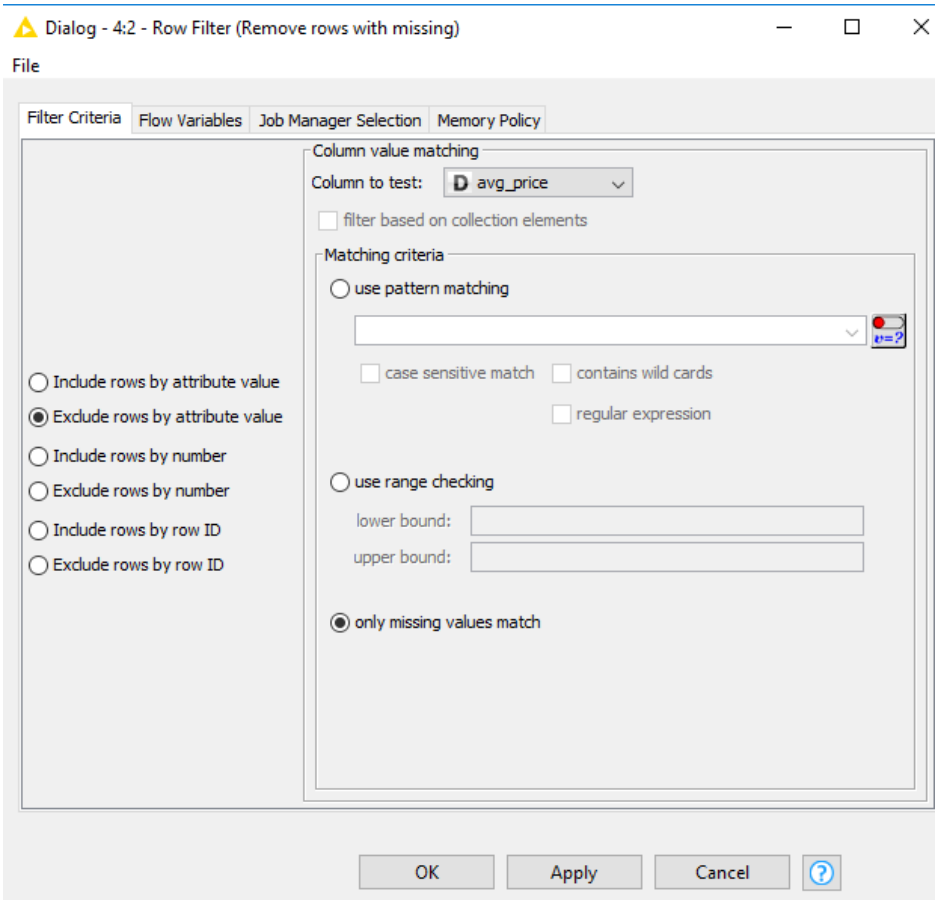


Ilustración 38. Configuración Row filter

Una vez realizado esta fase, conectamos este nodo con el nuevo nodo de estadística insertado. Se ve entonces que los valores nulos han sido eliminados.

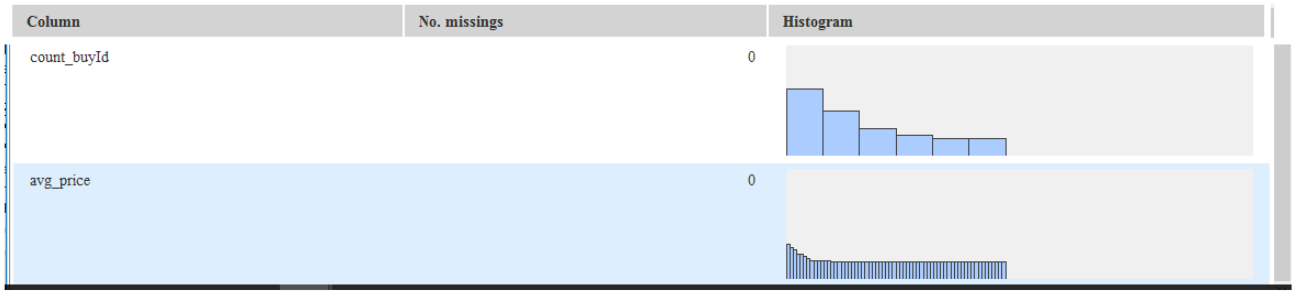


Ilustración 39. Verificación de la eliminación de valores nulos

6.3. Análisis de clasificación

En este apartado, se centra en el análisis, se seleccionan distintas técnicas que lleven a conseguir el objetivo trazado inicialmente, se construye modelos de datos y se analizan los resultados.

Se parte entonces de la fase anterior, donde se había quedado, antes de eso se define que es lo que se quiere conseguir con este análisis y qué técnica se utilizan para ello.

6.3.1. Definición del análisis

En este análisis se trata de predecir mediante la clasificación de los jugadores, quienes son aquellos que gastan más o menos dinero en función del costo del producto. Se busca entonces, segmentar a los usuarios mediante la condición de si más consumidores si las compras de producto son mayores a 5€ (a estos usuarios los llamaremos derrochadores o HighRollers) o aquellos compradores que compran ítems con un valor igual o inferior a 5€ (tacaños o PennyPinchers).

Se utiliza la plataforma de análisis KNIME para continuar el proceso anterior, de esta manera agilizar el análisis para así proveer mayor información acerca de los ingresos de la compañía.

Por otro lado, para este análisis es necesario hablar de las distintas técnicas que ofrecen la rama de Machine Learning como son, los modelos de clasificación mediante arboles de decisión.

6.3.1. Definición de la técnica a través del Machine Learning

Machine Learning es un campo de estudio que se centra en los sistemas de computación que pueden aprender de los datos, donde estos sistemas a menudo son llamados modelos que son capaces de ejecutar especificas tareas para analizar varios ejemplos de un determinado problema.

Entre las principales técnicas que se utilizan en Machine Learning están:

- **Clasificación**, sirve para predecir una categoría de los datos ingresados, por ejemplo podemos predecir el tipo de inclemencia climático partiendo de sus valores numéricos (lluvioso, soleado, nublado)
- **Regresión**, sirve para predecir un valor numérico como puede ser el precio de un producto.
- **Clúster Análisis**, tiene como objetivo organizar valores o similares dentro de un conjunto de valores, agrupándolos. Por ejemplo podemos segmentar al usuario dependiendo de sus compras (jóvenes, adultos o ancianos).
- **Asociación Análisis**, tiene como objetivo encontrar reglas que asocien diferentes tipos de productos o eventos, un ejemplo sería analizar el comportamiento de los clientes de acuerdo a sus compras.

Ahora nos centraremos en la clasificación de los usuarios dependiendo del grado de compra, utilizando estos valores numéricos para clasificarlos como derrochadores (HighRoller) o tacaños (PennyPincher).

Para realizar la clasificación, primero trabajaremos con los datos importados y procesados, se crea una nueva categoría mediante el nodo *numeric binner* con la condición establecida anteriormente, dividiendo a los jugadores en dos atributos según el monto de la compra.

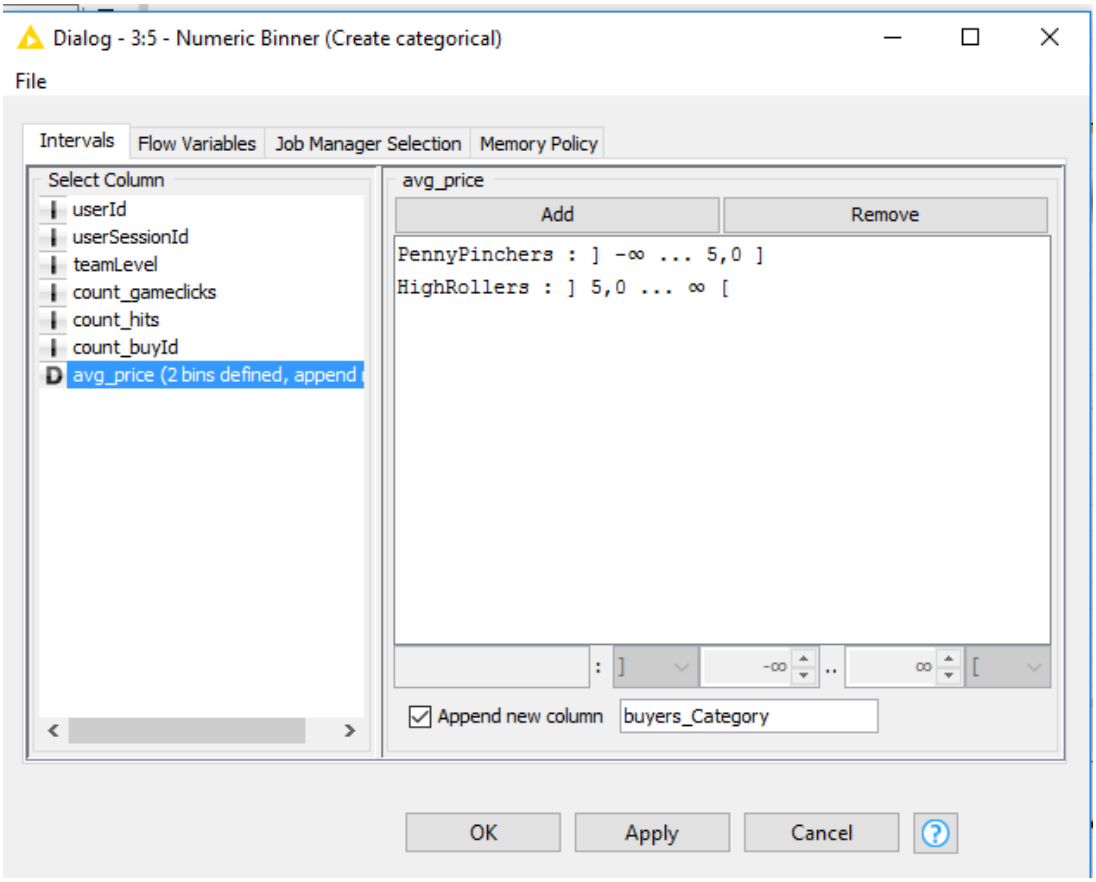


Ilustración 40. Nodo numeric binner

Los jugadores derrochadores (HighRollers) son compradores de ítems cuyo costo es mayor a 5€, al contrario, los jugadores tacaños (PennyPinchers) son compradores de ítems cuyo costo es menor o igual a 5€. La nueva categoría es llamada buyer_category y será la categoría objetivo para realizar la predicción
Una vez creado el nuevo atributo, es necesario filtrar aquellas columnas que no aportan al análisis o que ya han sido utilizadas para lograr la clasificación, por tanto detallamos en esta tabla los campos filtrados:

Atributo	**Causa de exclusión**
Avg_price	Atributo utilizado para la creación de la nueva categoría creada.
UserId	Atributo no relacionado con el atributo creado.
UserSessionId	No es necesario saber el id de sesión de cada usuario, por tanto es descartado.

Después de saber qué campos hay que excluir, se añade el nodo *column filter* para realizar este paso. Seleccionamos la siguiente configuración para excluir las columnas

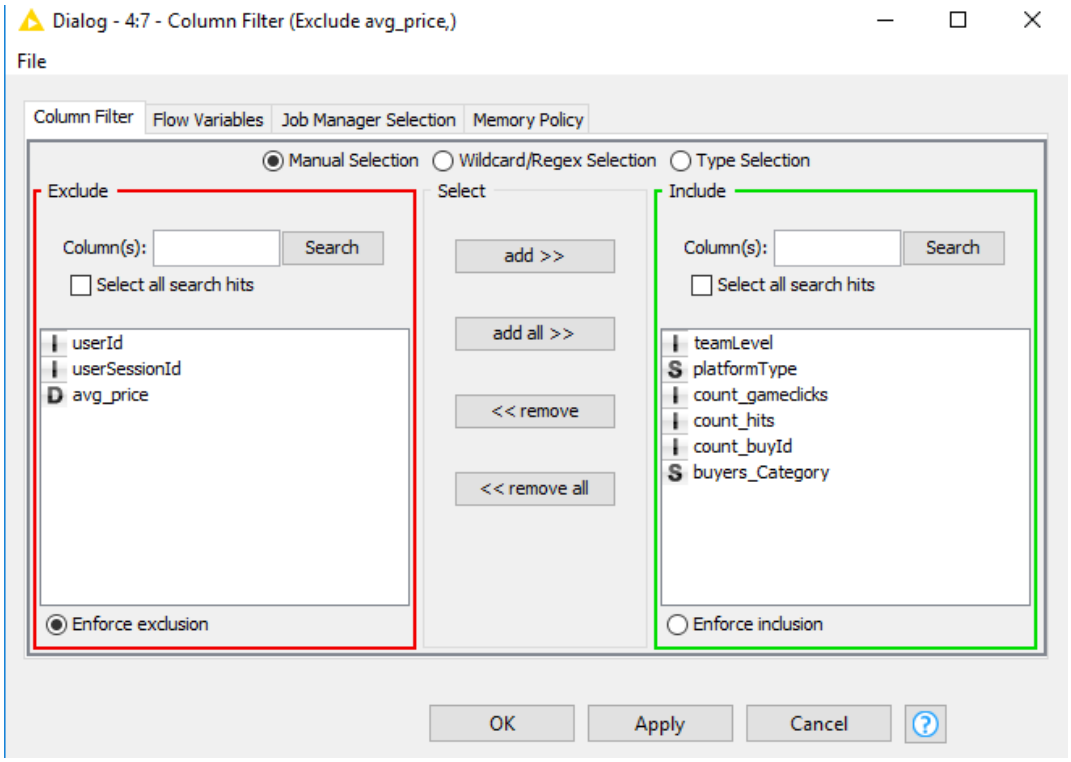


Ilustración 41. Exclusión de atributos

Se añade el nodo *color manager* para diferenciar ambos atributos de la nueva categoría creada.

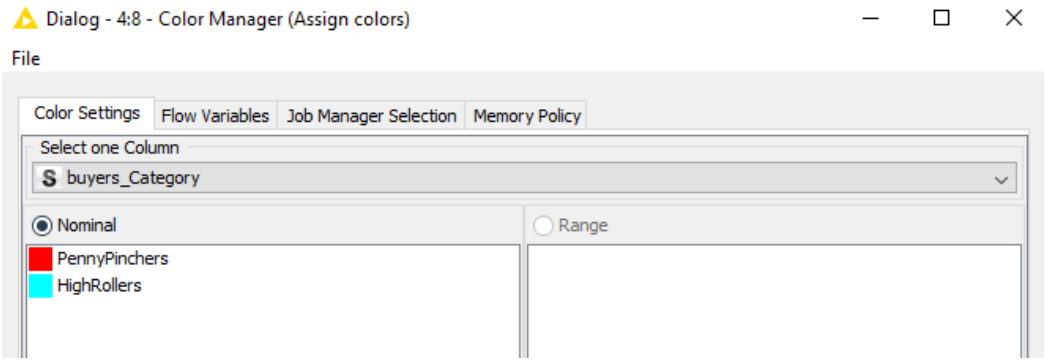


Ilustración 42. Nodo color manager

En el siguiente paso, Se necesitará crear el modelo de datos para el análisis. Este modelo de datos consta de dos fases. Por un lado, se tiene la fase de entrenamiento de los datos en el cual el modelo es construido ajustando parámetros mediante los datos entrenados, en otras palabras, los datos de entrenamiento es el conjunto de datos usados para entrenar o crear el modelo de datos.
Por otro lado, se tiene la fase de prueba, donde el modelo creado es aplicado a los datos de prueba.

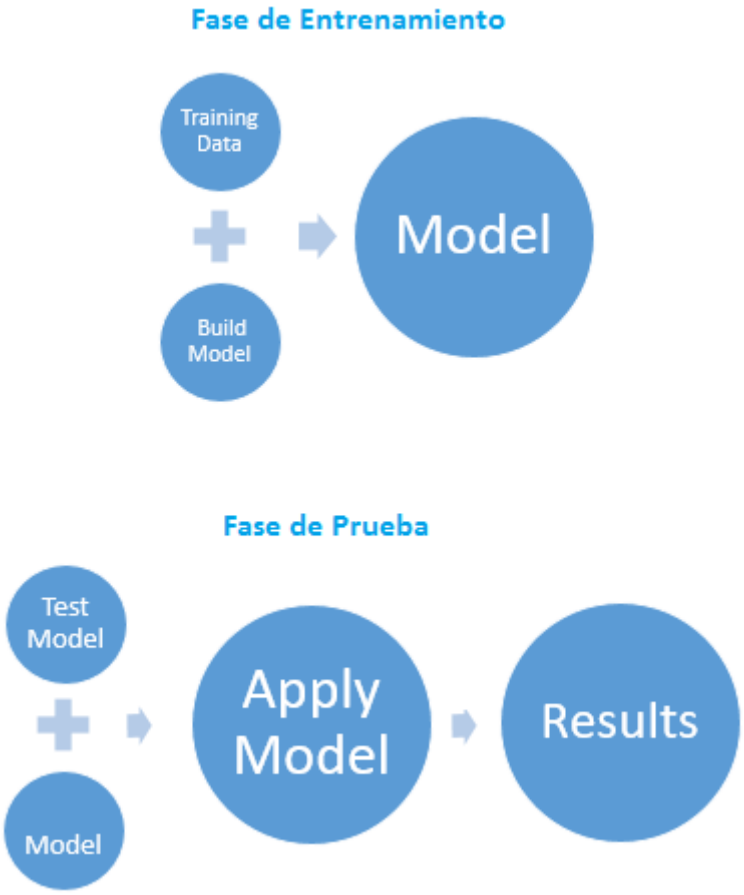


Ilustración 43. Aplicación del modelo de datos

Entonces, para particionar los datos creados hasta el momento se necesita insertar el nodo *partitioning* el cual dividirá en datos de entrenamiento y prueba.
El tamaño de la partición será de 60% para un conjunto de datos, y 40% para la otra. Como datos de muestra se escoge la categoría creada. El *random seed* es creado aleatoriamente y significa el patrón de resultados, cada *random* distinto daría valores distintos.

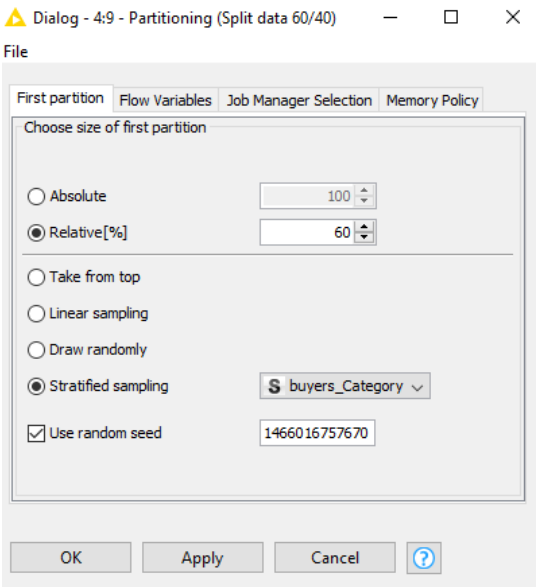


Ilustración 44. Nodo partitioning

El modelo de datos creado será el de árbol de decisión. Este modelo es un método popular usado en clasificación que divide los datos en pequeños conjuntos donde cada subconjunto pertenecerá a una sola categoría. Cabe destacar que estas pequeñas porciones divididas serán lo más puras posibles, quiere decir, lo más precisas posibles.

Un árbol de decisión es una estructura jerárquica con nodos y aristas, el nodo raíz se encuentra por encima de los otros nodos, y los últimos nodos son llamados nodos hojas. Cada uno de estos nodos tiene asociados etiquetas de subclases, entonces la técnica de clasificación se basa en etiquetar a diferentes nodos desde el nodo raíz hacia los nodos hojas dependiendo de su estructura condicionada por los valores asociados. Esto es importante a la hora de analizar los resultados.

Siguiendo con esta fase, es necesario añadir un nodo llamado *decisión tree learner* para los datos de entrenamiento y otro nodo llamado *decisión tree predictor* para aplicar el modelo. Se conecta la salida de los datos particionados hacia cada uno de ellos tal como muestra la imagen.

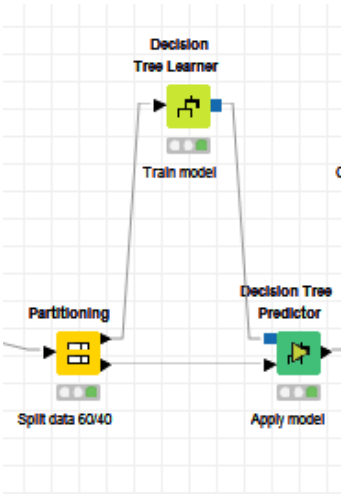


Ilustración 45. Nodo decision tree

Al final, se tiene el siguiente flujo de trabajo con todos los pasos realizados.

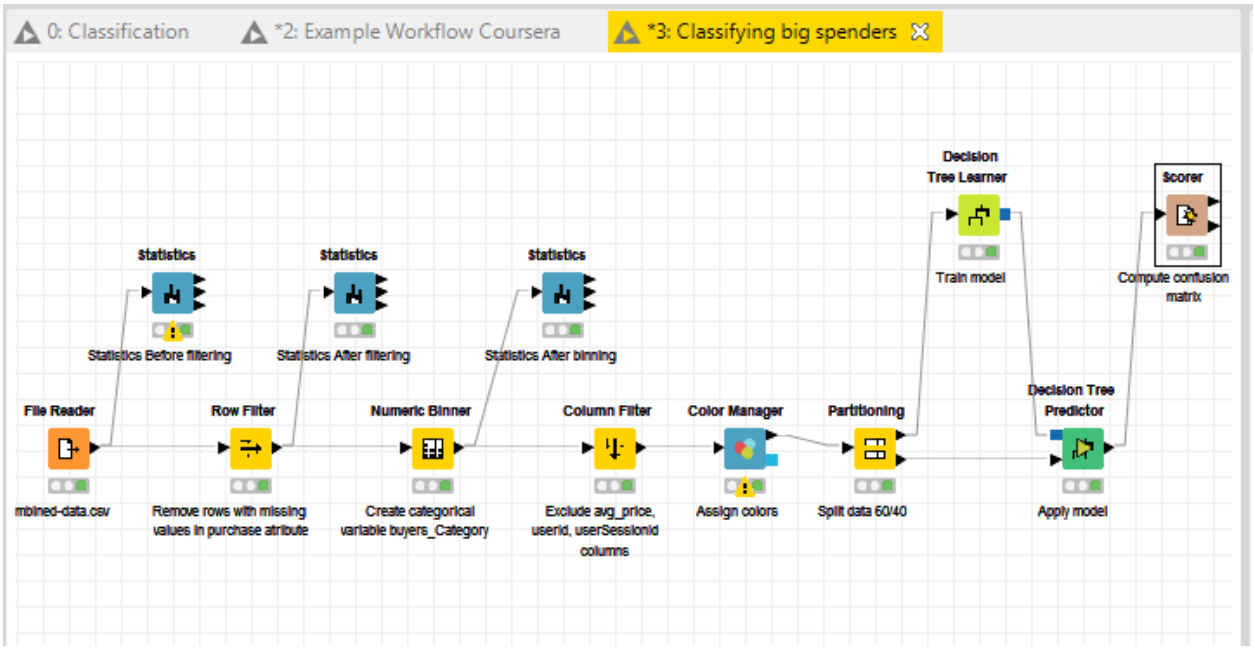


Ilustración 46. Diagrama final

Para ver los resultados del análisis se tiene que dar clic derecho al árbol de decisión predictivo y seleccionar *view decisión tree*.

Ilustración 47. Resultados del analisis de clasificación

6.4. Análisis de Clustering

Para este análisis se trabajará desde el entorno Cloudera, mediante MLlib y PySpark. Estas dos librerías serán importante a la hora de procesar los datos.

Este análisis tiene como objetivo organizar valores o similares dentro de un conjunto de valores, agrupándolos. De esta manera, se agrupan distintos grupos de usuarios basados en determinadas características acerca del comportamiento en el juego.

Nos centraremos en el comportamiento del usuario al realizar las compras de ítems, los clics que realiza en el juego para posteriormente saber la relación de estos parámetros con el ingreso que tiene la compañía.

Para agrupar a los jugadores, se necesita seleccionar los atributos que muestren los datos que se necesitan según el objetivo que se ha trazado anteriormente. El número de atributos dependerá de la información que se quiera recoger, teniendo especial cuidado en no escoger demasiados atributos para no malinterpretar el análisis con datos redundantes.

Los datos se crearán a partir de los ficheros fuente del proyecto, los atributos seleccionados serán:

Atributos	Descripción
totalBuyIds	Atributo que contiene los identificadores de los items comprados por cada usuario. Este atributo proviene de la tabla buy-clicks.csv, con el cual analizaremos las compras del usuario.
totalHits	Atributo que contiene el total de hits clicado por el usuario (clic correcto tiene valor 1, por el contrario es 0) en el juego. Este atributo proviene de la tabla game-clicks.csv.
revenue	Atributo que contiene el monto total gastado por cada usuario, esta información ayudará para relacionar los dos atributos anteriores y el ingreso aportado por el jugador.

Se procede a ingresar a la máquina virtual y ejecutar desde la línea de comandos el código pySpark para abrir Jupyter.

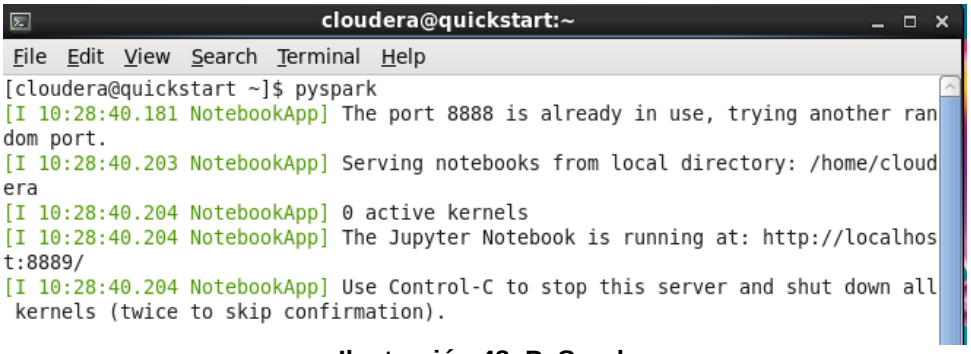


Ilustración 48. PySpark

Se abrirá entonces el navegador con el directorio de jupyter.

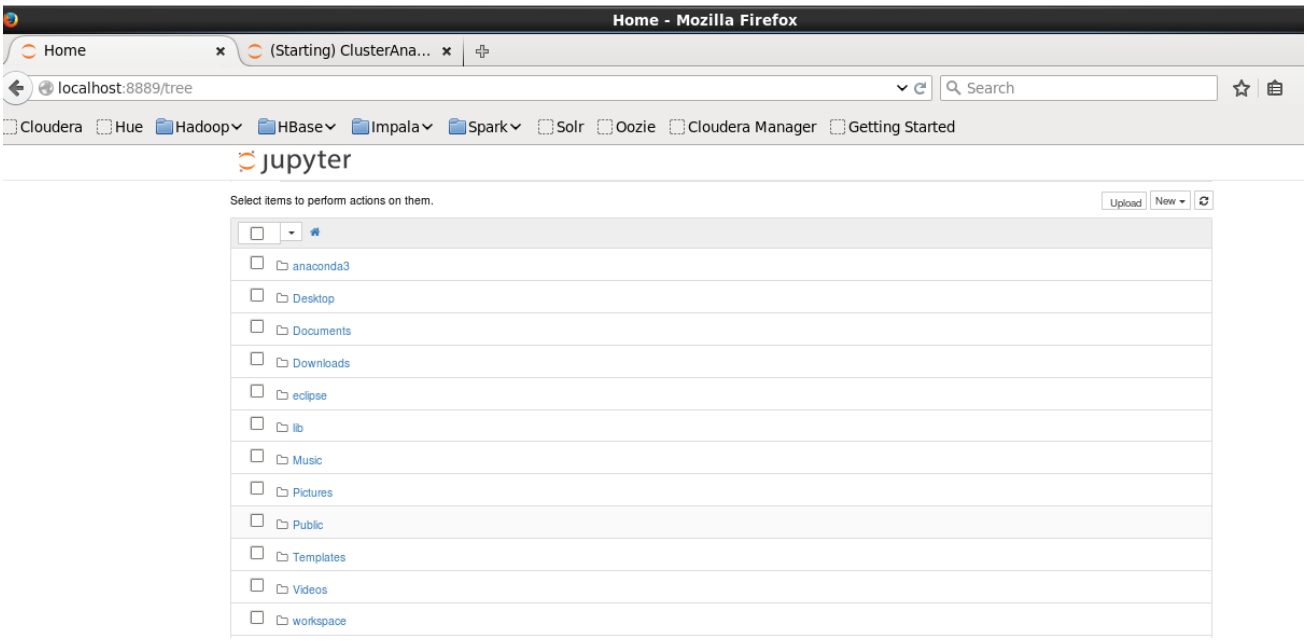


Ilustración 49. Directorio Jupyter

Desde la parte superior derecha se crea un nuevo Notebook y se introducirá un nombre.

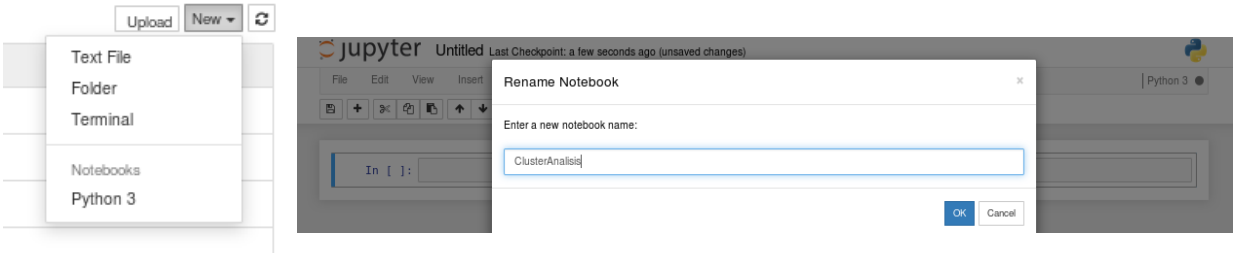


Ilustración 50. Cración de Notebook

Una vez realizado esto, se procede a importar librerías siguientes:

```
In [1]: from pyspark.sql import SQLContext
        from pyspark.ml.clustering import KMeans
        from pyspark.ml.feature import VectorAssembler
        from pyspark.ml.feature import StandardScaler
```

Ilustración 51. Importación de librerías

En la siguiente línea, se importará el archivo donde se encuentran los atributos, en este caso es W3_ClusterAnalysis.csv

```
In [2]: sqlContext = SQLContext(sc)
        data = sqlContext.read.load('file:///home/cloudera/Downloads/W3_ClustersAnalysis.csv',
                                    format='com.databricks.spark.csv',
                                    header='true',inferSchema='true')
```

Ilustración 52. Importación de datos

Se realiza el conteo de datos y se ve si se tienen las columnas que necesitamos:

```
In [4]: data.count()
Out[4]: 1193

In [5]: data
Out[5]: DataFrame[row ID: string, totalBuyIds: int, totalHits: int, revenue: int]
```

Ilustración 53. Conteo de datos

Se elimina la columna row ID, ya que no se necesita los identificadores de la tabla

```
In [8]: data = data.drop('row ID')

In [9]: data
Out[9]: DataFrame[totalBuyIds: int, totalHits: int, revenue: int]
```

Ilustración 54. Eliminación de columnas

Para ver las 5 primeras filas de cada columna, se ingresa:

```
In [31]: data.select('totalBuyIds', 'totalHits', 'revenue').show(5)
+-----------+---------+-------+
|totalBuyIds|totalHits|revenue|
+-----------+---------+-------+
|         12|      143|     21|
|          1|       96|     11|
|         31|      340|    113|
|         12|      403|     31|
|          1|       86|      2|
+-----------+---------+-------+
only showing top 5 rows
```

Ilustración 55. Ver columnas

Se necesita ver las estadísticas de los datos para ver si existen valores nulos, para esto importamos las librerías *numpy* y *pandas* que vienen por defecto dentro de Anaconda y que sirven para trabajar con grandes conjuntos de datos.

```
In [35]: import numpy as data

In [38]: import pandas as data

In [10]: data.describe().toPandas().transpose()
```

Out[10]:

	0	1	2	3	4
summary	count	mean	stddev	min	max
totalBuyIds	546	13.659340659340659	12.21509348806919	0	65
totalHits	1193	69.89354568315171	65.65942916167523	0	517
revenue	546	39.20695970695971	41.1546560299256	1	223

Ilustración 56. Estadisticas

Como se puede observar, el atributo totalHits tiene 1193 filas, siendo mayor que los otros atributos y representando valores nulos para aquellas filas donde los otros atributos tengan campos vacíos. Para eliminar los datos nulos ingresamos: *data = data.na.drop().* Con este comando se eliminan esos valores y se procede a verificar el número de filas para cada atributo.

```
In [11]: data = data.na.drop()

In [12]: data.count()
Out[12]: 546
```

Ilustración 57. Eliminación de valores nulos

Se procede a crear un vector que contendrá todas las columnas que se quieren combinar

```
In [14]: featuresUsed = ['totalBuyIds','totalHits','revenue']
         assembler = VectorAssembler(inputCols=featuresUsed, outputCol="features_unscaled")
         assembled = assembler.transform(data)
```

Ilustración 58. Creación del vector de datos

Ahora se tendrá que escalar los valores, siendo importantes los valores de desviación estándar (*withStd*) y la mediana (*withMean*), los cuales serán puntos de referencia para el escalado.

```
In [15]: scaler = StandardScaler(inputCol="features_unscaled", outputCol="features", withStd=True, withMean=True)
         scalerModel = scaler.fit(assembled)
         scaledData = scalerModel.transform(assembled)

In [16]: scaledData = scaledData.select("features")
         scaledData.persist()
Out[16]: DataFrame[features: vector]
```

Ilustración 59. Escalar los datos

Lo siguiente es utilizar el algoritmo K-Means el cual es usado en el análisis de clustering. Este algoritmo selecciona los K-primeros centroides o pequeños grupos “clusters”, estos son asignados a aquellos conjuntos de datos que están más próximos, calculando su distancia y asignando una muestra a dicho clúster. Luego se determina la mediana de cada clúster para determinar un nuevo centroide. Estos pasos son repetidos hasta que no se encuentre muestras más próximas.

Para realizar este procedimiento se procede a crear el K-Means con los valores de clústeres a crear, en este caso 3, una por cada columna. Se crea un modelo partiendo de los datos escalados para luego ser transformados.

```
In [17]: kmeans = KMeans(k=3, seed=1)
         model = kmeans.fit(scaledData)
         transformed = model.transform(scaledData)
```

Ilustración 60. K-Means

El siguiente paso es crear los centros de cada clúster para ser comparados, recordar que cada clúster viene a ser cada partición de datos.

```
In [18]: centers = model.clusterCenters()
         centers

Out[18]: [array([-0.64645597,  0.12239214, -0.61965848]),
          array([ 0.43584599, -0.25797935,  0.34764005]),
          array([ 2.24868783,  0.14530832,  2.38374319])]
```

Ilustración 61. Centros de cluster, resultados

La primera columna en el vector de centros viene a ser la versión escalada del atributo buyId la segunda columna corresponde a totalHits y la tercera corresponde al atributo revenue por cada usuario.

- Se debe comparar el primer valor de cada clúster para ver el comportamiento de cada jugador al comprar un ítem.
- Se debe comparar el segundo valor de cada clúster para diferenciar el comportamiento de cada jugador cuando realiza un hit en el juego.
- Se debe comparar el tercer valor de cada clúster para diferenciar los ingresos de cada jugador al comprar un ítem.

Esta comparativa se deberá incluir dentro del reporte final de este análisis, por tanto, se verá en el siguiente capítulo.

6.5. Análisis de Grafos

El objetivo del análisis es encontrar aquellas salas de chat que son más activas con el fin de ver que jugadores y qué equipos son más activos en las sesiones del juego.

Para este análisis se usarán los datos relacionados con las salas de chat de los jugadores. Esta información almacena los registros de cada jugador durante cada sesión del juego. Es necesario mencionar que estos datos han sido detallados en el apartado *Adquisición*.

Es necesario recordar que un grafo está compuesto varios nodos conectados por aristas, estas aristas viene a ser la relación que existe entre dos nodos. De esta manera, los nodos y aristas serán los siguientes:

Nodos		
Nombre	**Propiedades**	**Descripción**
User	Id	Jugadores que interactúan en el chat.
Team	Id	Equipos de usuarios en cada sala de chat.
TeamChatSession	Id	Sesión creada por un equipo de jugadores
ChatItem	Id	Mensaje de cada chat representado por Id

Aristas		
Nombre	**Propiedades**	**Descripción**
CreateSession	Timestamp	Arista "CreatesSession entre el nodo user y TeamChatSession
OwnerBy	Timestamp	Arista "OwnedBy" entre el nodo TeamChatSession y el nodo Team
Joins	Timestamp	Arista "Joins" desde User hacia TeamChatSession
Leaves	Timestamp	Arista "Leaves" desde User hacia TeamChatSession
Mentioned	Timestamp	Arista "Mentionedque va desde chatItem hacia User
PartOf	Timestamp	Arista "PartOf" desde el nodo ChatItem hacia el nodo TeamChatSession.
ResponseTo	Timestamp	Arista "ResponseTo" desde el nodo ChatItem hacia otro nodo ChatItem.
InteractsWith	Timestamp	Arista “InteractsWith " compuesto solo por users.

Para crear la base de datos dentro de Neo4j, primero es necesario crear el esquema de datos, este esquema contendrá cada tabla conteniendo información de las salas de chat.

- **Chat_create_team_chat.csv:** userid, teamid, TeamChatSessionID, timestamp
- **Chat_item_team_chat.csv:** userid, TeamChatSessionID, ChatItemID, timestamp
- **Chat_join_team_chat.csv:** userid, TeamChatSessionID, timestamp
- **Chat_leave_team_chat.csv**: userid, TeamChatSessionID, timestamp
- **Chat_mention_team_chat.csv**: ChatItem, userid, timestamp
- **Chat_respond_team_chat.csv:** ChatItem, ChatItem, timestamp

El proceso de subida de datos a Neo4j se detalla en el *anexo C. Código fuente*.

Se puede realizar la consulta del grafo creado, para esto es necesario escribir la consulta en la barra de codificación y ejecutarlo mediante el icono Play que se encuentra en la esquina superior derecha.

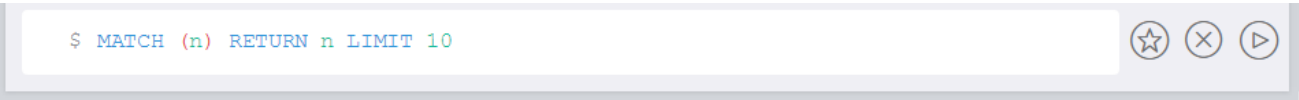


Ilustración 62. Barra de ejecución Neo4j

Este código indica que muestre los 10 primeros nodos donde se observa cada nodo unido mediante una arista que es quien relaciona uno o varios nodos entre sí.

En la imagen siguiente se observa el nodo del equipo con id 81 propietario del chat con id de sesión 6778, este último está relacionado con el usuario de id 740.

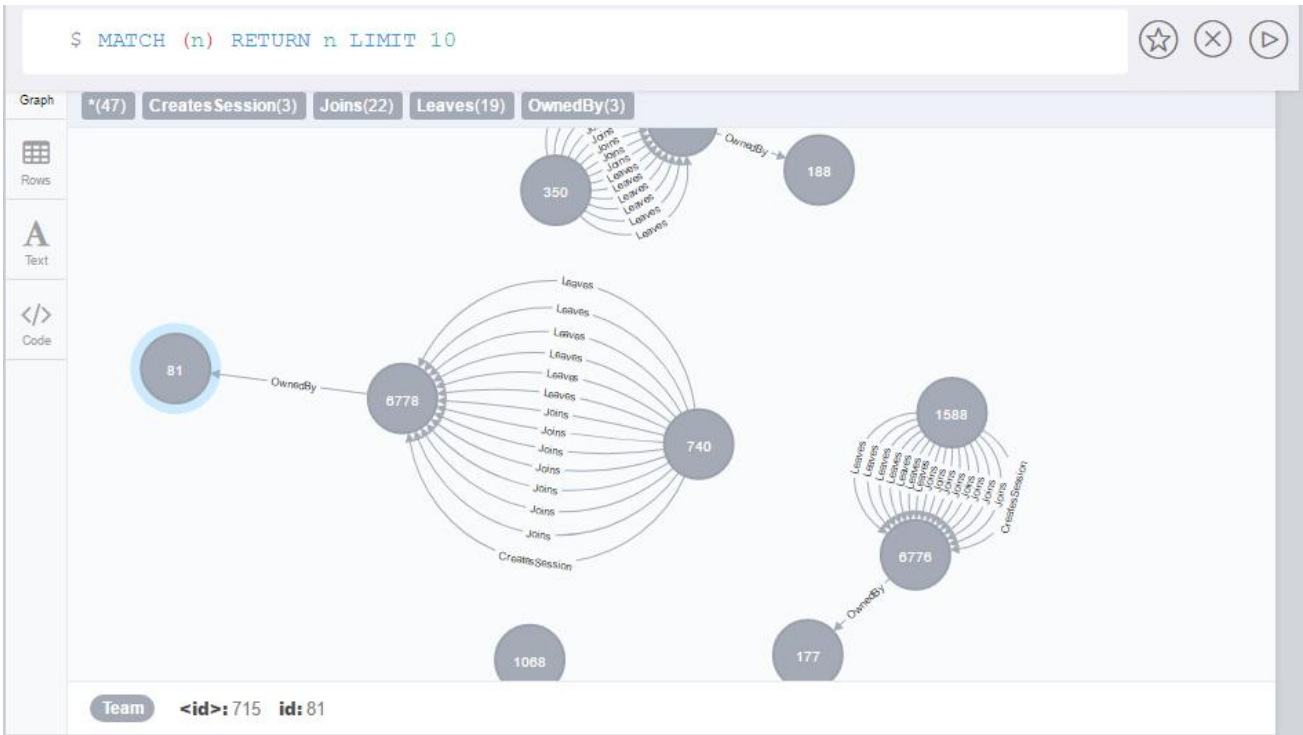


Ilustración 63. Top diez nodos

De esta manera, se puede realizar consultas con el fin de analizar el comportamiento de cada jugador y las relaciones entre jugadores en las diferentes salas de chat.

Número de participantes en conversaciones más largas

Primero se tiene que encontrar los chats con las conversaciones más largas, para esto se crea la siguiente consulta:

```
match p=(a)-[:ResponseTo*]->(c) return size(nodes(p)) order by size(nodes(p)) desc limit 1
```

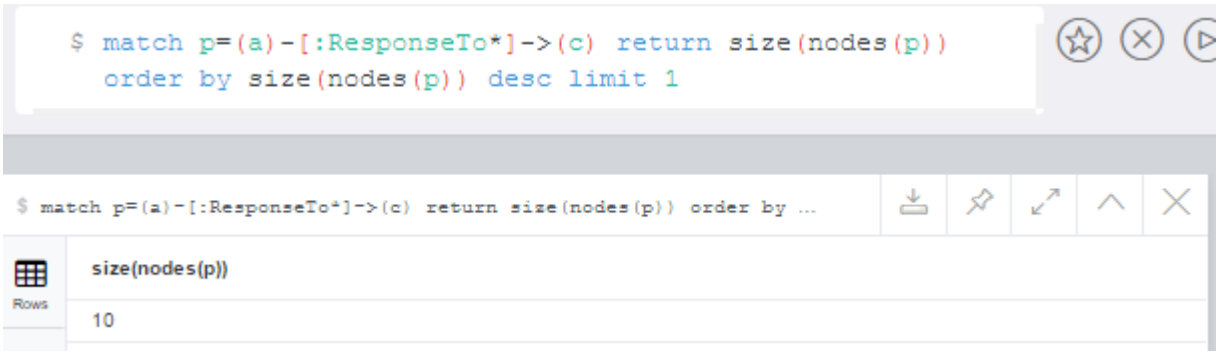


Ilustración 64. Consulta salas de chats

Se puede observar que existen 10 salas de chats que tienen las mayores respuestas o son más activas. Ahora es necesario ver qué usuarios son los que integran este chat. Para esto se realiza la siguiente consulta:

```
match p=(i:ChatItem)-[:ResponseTo*]->(j:ChatItem) where length(p)=9 with extract(n in nodes(p)|n.id) AS LongestPath match (u:User)-[:CreateChat]->(x:ChatItem) where x.id in LongestPath return count(distinct u) as NumUsers
```

Se puede apreciar que existen 5 jugadores, en la segunda imagen se aprecia el id de cada jugador:

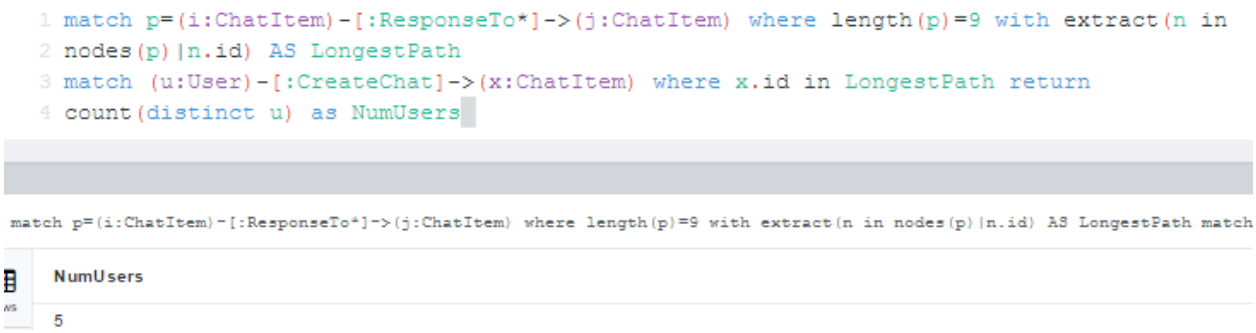


Ilustración 65. Número de jugadores

Jugadores por identificador:

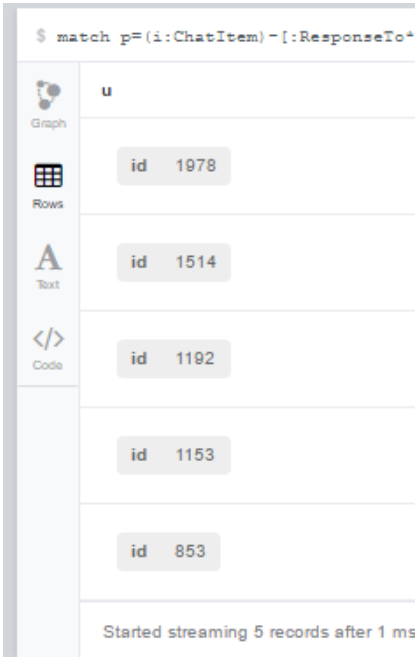


Ilustración 66. Identificador de jugadores

Para encontrar los usuarios más activos se necesita encontrar aquellos grupos de usuarios que interactúan entre sí. Para esto se necesita estimar cuan densos son sus vecinos, los vecinos vienen a ser aquellos nodos que están más próximos entre sí. Estos vecinos forman un grupo de vecino o vecindad.

Para realizar esta estimación, se necesita realizar una serie de pasos:

Primero se necesita agrupar los nodos de usuarios, para esto es necesario una condición que identifique aquellos usuarios elegidos. Las condiciones serán:

- Un usuario debe ser mencionado por otro o viceversa.
- Un jugador crea una respuesta a otro usuario mediante un chatItem (texto descriptivo).

Con estas condiciones se sabrá aquellos usuarios que interactúan entre sí, ya realizando una mención a otro jugador, o respondiendo a un usuario en concreto. Para realizar estos pasos, se necesita crear una consulta para cada condición.

Consulta para las menciones del usuario:

```
Match (u1:User)-[:CreateChat]->(i:ChatItem)-[:Mentioned]->(u2:User)
create (u1)-[:InteractsWith]->(u2)
```

Consulta para la creación de un chatItem en respuesta a otro usuario:

```
Match (u1:User)-[:CreateChat]->(i:ChatItem)-[:ResponseTo]->(i2:ChatItem)<-
[:CreateChat]-(u2:User) create (u1)-[:InteractsWith]->(u2)
```

Ya creadas las condiciones para cada usuario, necesitamos eliminar aquellas respuestas a cada chat que se realiza cada usuario, ejemplo: usuario A responde a sí mismo. Esto crearía un bucle que es necesario evitar ya que es irrelevante para el análisis. Para eliminar esto, se ejecuta el siguiente código:

```
Match (u1)-[r:InteractsWith]->(u1) delete r
```

El siguiente paso es crear un mecanismo de puntuación para encontrar aquellos usuarios que tienen más vecinos, en otras palabras, aquellos usuarios que interactúan más con otros usuarios, obteniendo una mayor densidad de usuarios (vecinos).

Este sistema de puntuación tendrá un rango desde 0 (los nodos o vecinos estan desconectados de otros nodos) a 1 (aquellos nodos o vecinos donde cada nodo es conectado con uno o más nodos). Este sistema es llamado coeficiente de agrupamiento, este sistema se define mediante varios valores que son:

- Total de aristas (edges)
- Número de vecinos (K)
- Pares posibles de aristas que salen de un vecino (*K* (K-1))*

La fórmula para calcular el coeficiente de agrupamiento es:

```
Coeficiente de agrupamiento (cc) = total de aristas (edges) / k*(k-1)
```

Para calcular este coeficiente de agrupamiento, se tiene que ejecutar la siguiente consulta:

```
match (u1:User)-[:InteractsWith]-(u2:User)
 where u1 <> u2
with u1, count(distinct u2) as k, collect(distinct u2.id) as vecinos
match (u2:User)-[:InteractsWith]-(u3:User)
where u2.id in vecinos and u3.id in vecinos
with distinct u1,u2,u3,k
with u1, sum(case when (u2)--(u3) then 1 else 0 end) as edge, k
return u1.id, edge, k*(k-1), 1.0*edge/k/(k-1) as cc
order by cc desc
```

Se obtiene el siguiente resultado donde se muestran los usuarios, sus aristas y el máximo número de posibles aristas (*K* (K-1))* y su coeficiente de agrupamiento:

```
$ match (u1:User)-[:InteractsWith]-(u2:User) where u1 <> u2 with u1, count(distinct u2) as k,...
```

Rows | Text | Code

u1.id	edge	k*(k-1)	cc
2064	20	20	1
1849	20	20	1
1639	6	6	1
2008	2	2	1
558	2	2	1
2028	30	30	1
697	6	6	1
1366	30	30	1
2098	30	30	1
1955	42	42	1
417	42	42	1
982	30	30	1
1894	12	12	1
1515	42	42	1
1310	2	2	1
1761	12	12	1

Ilustración 67. Coeficiente de agrupamiento

Para buscar aquellos tres usuarios más activos se debe realizar la siguiente consulta:

```
match p=(u:User)-[r:CreateChat]->(c:ChatItem)
return distinct ( u),count(p) order by count(p) desc limit 3
```

Como se observa, se obtiene aquellos usuarios más activos.

```
1 match p=(u:User)-[r:CreateChat]->(c:ChatItem)
2 return distinct ( u),count(p) order by count(p) desc limit 3
```

```
$ match p=(u:User)-[r:CreateChat]->(c:ChatItem) return distinct ( u),count(p) or
```

Graph | Rows | Text | </>

"u"	"count(p)"
{"id":"394"}	"115"
{"id":"2067"}	"111"
{"id":"209"}	"109"

Ilustración 68. Usuarios más activos

Ahora es necesario aplicar el coeficiente de agrupamiento a estos tres usuarios más activos, para esto se indica el id de cada uno en la siguiente consulta:

```
match (u1:User)-[:InteractsWith]-(u2:User)
where u1.id in [2075,254,996] and u1<>u2
with u1, count(distinct u2) as k, collect(distinct u2.id) as vecinos
match (u2:User)-[:InteractsWith]-(u3:User)
where u2.id in vecinos and u3.id in vecinos
with distinct u1,u2,u3,k
with u1, sum(case when (u2)--(u3) then 1 else 0 end) as edge, k
return u1.id, edge, k*(k-1), 1.0*edge/k/(k-1) as cc
order by cc desc
```

Obteniendo el siguiente resultado:

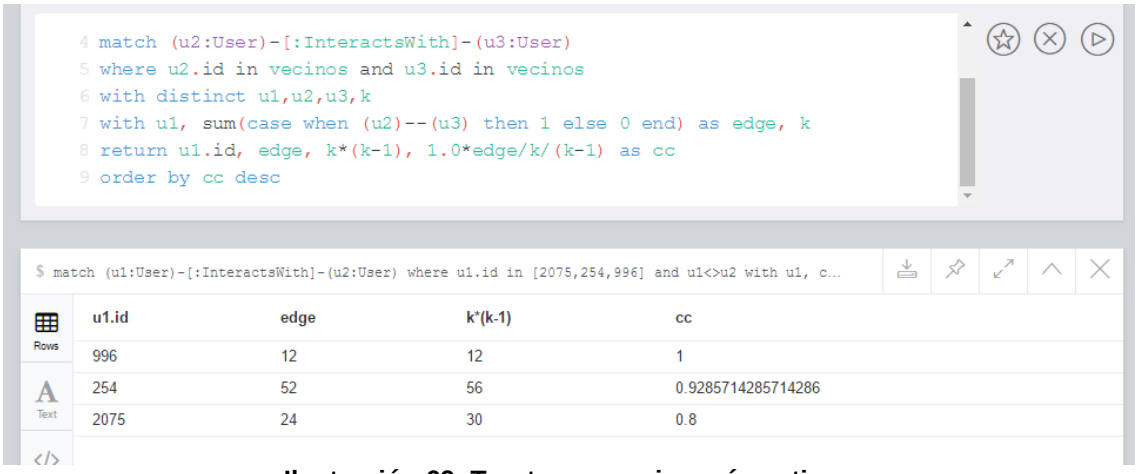


Ilustración 69. Top tres usuarios más activos

6.6. Reportes finales

6.6.1. Reportes de la exploración de datos

Plataformas más utilizadas

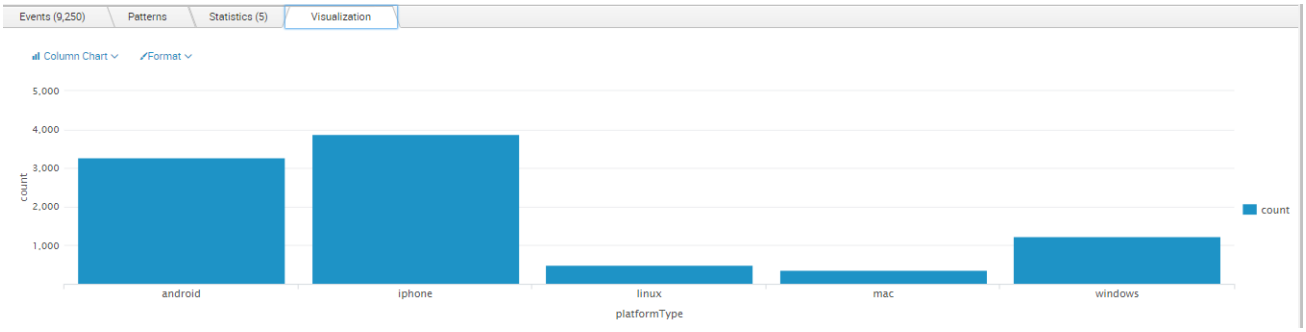


Ilustración 70. Reporte de las plataformas más utilizadas

Categorías más vistas mediante clics de los jugadores:

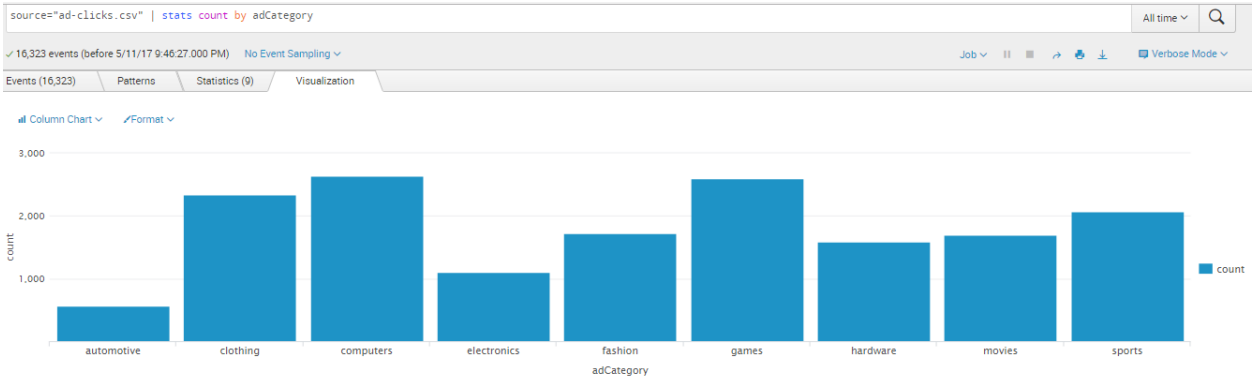


Ilustración 71. Reporte de las categorías mas vistas

Productos más comprados

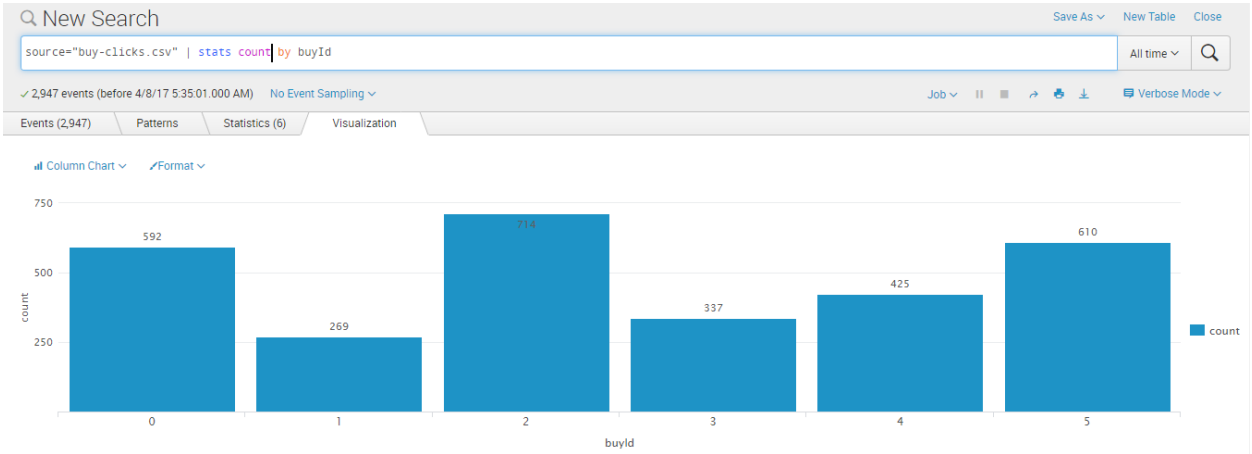


Ilustración 72. Reporte de los productos más comprados

Ítem con más ingresos

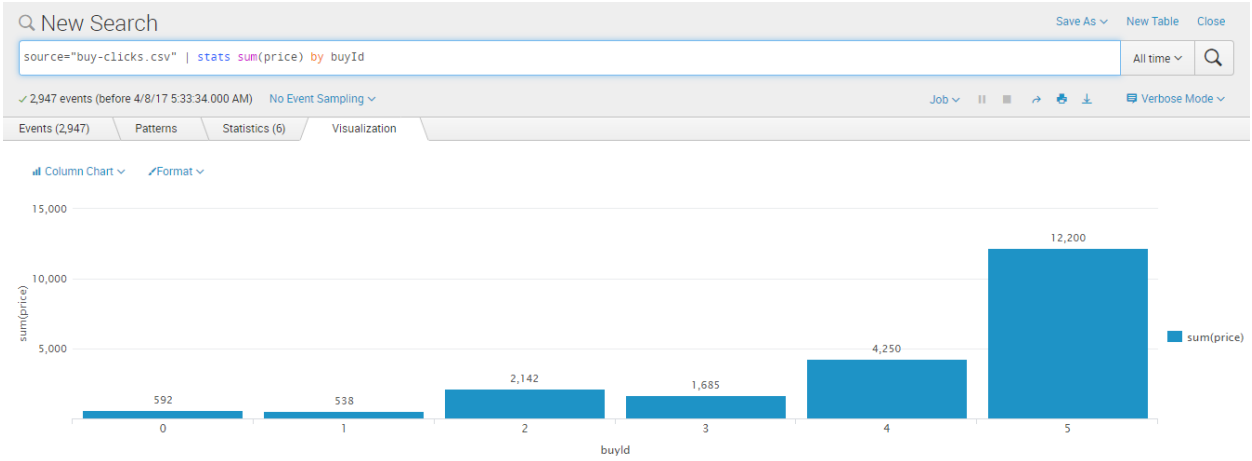


Ilustración 73. Reporte de los items con más ingresos

Resultados:

Los datos de exploración muestran los siguientes resultados:

- Las tres plataformas más utilizadas son: Iphone, Android y Windows
- Las tres categorías de anuncios más vistas son: Computers, Games y Clothings
- Existen seis productos de los cuales dos de ellos son más comprados (id 2 e id 5)
- El producto que más ingreso reporta es el ítem con id 5, con un total de 12.200€
- El total de ingresos por todas las compras dentro del juego tiene un valor de 21,407€
- El usuario id 2229 es el que más gasta (200€)

6.6.2. Reportes del análisis de clasificación

El resultado del análisis viene dado por la salida del resultado del árbol de decisión siguiente:

Ilustración 74. Reporte del análisis de clasificación

Resultados:

En este diagrama se puede apreciar que los jugadores que utilizan la plataforma Iphone tienen más probabilidades de ser clasificados como derrochadores (HighRollers) ver color de barra celeste. Por otro lado, los usuarios del resto de plataformas tienen más probabilidades de ser clasificados como tacaños (PennyPinchers), ya que gastan menos en las compras dentro de la *app*.

6.6.3. Reportes del análisis de clustering

El análisis de agrupamiento demuestra que existen tres grupos segmentados.

```
In [18]: centers = model.clusterCenters()
         centers

Out[18]: [array([-0.64645597,  0.12239214, -0.61965848]),
          array([ 0.43584599, -0.25797935,  0.34764005]),
          array([ 2.24868783,  0.14530832,  2.38374319])]
```

Ilustración 75. Reporte del análisis de clustering

Se puede observar, las siguientes diferencias:

Clúster 1 tiene los valores más bajos a comparación con el clúster 3, todos sus valores están por debajo de 0.7, por tanto demuestra que este grupo de usuarios han realizado pocas compras, son nuevos usuarios o no están interesado en los ítems ofertados.

Clúster 2 se diferencia teniendo los valores más altos que el clúster 1. Esto quiere decir que, aquellos jugadores con más hits tienden a comprar más ítems.

Clúster 3 tiene los valores más altos que los dos clústeres anteriores, corroborando que aquellos jugadores que tienden a hacer más hits, compran más ítems siendo los que más ingresos producen en el juego.

6.6.4. Reportes del análisis de grafos

El análisis de grafos muestra el siguiente resultado:

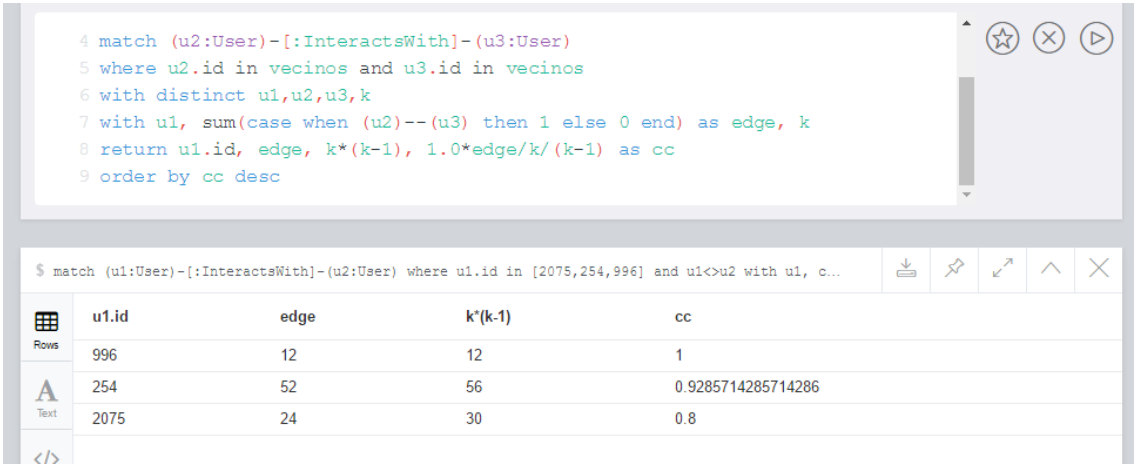


Ilustración 76. Reporte del análisis de grafos

Se puede observar los usuarios más activos:

- **User id 996** con 12 relaciones entre otros usuarios y con un máximo de 12 relaciones a crear, puntuándose con un coeficiente de agrupamiento de 1.
- **User id 254** con 52 relaciones con otros usuarios y con un máximo de 56, puntuado con un coeficiente de 0.9.
- **User id 2075** con 24 relaciones con otros usuarios y con un máximo de 30, puntuándose con un coeficiente de 0.8

Por tanto, estos usuarios son objetivos claros para nuevas campañas de ofertas y anuncios sobre nuevos productos ya que son más activos que otros jugadores.

6.7. Actuación

Recomendaciones finales

Después del exhaustivo análisis llevado a cabo en las diferentes etapas del trabajo, se concluye que, para el incremento de ventas, la empresa debe tomar atención en los siguientes puntos:

Los usuarios de la plataforma IPhone son aquellos que gastan más en las compras dentro del juego. Por tanto, es un gran objetivo para desplegar en ellos las futuras campañas de anuncios o nuevos productos. Así también, los jugadores y sus equipos más activos son otro objetivo para este *marketing* personalizado.

Enfocarse en aquellas categorías que son más vistas o clicadas por los usuarios. Las categorías como *Games* o *Computer*, son importantes para incrementar los ingresos publicitarios.

Para aquellos productos que están en stock, se podría realizar un estudio para ver la viabilidad de la subida de precio en aquellos que tienen más ingresos. Para aquellos productos que están por debajo de la media, es necesario estudiar por qué no son tendencia o de valor en las compras de los jugadores.

Los análisis de *clustering* y grafos nos indican que se debe tomar especial atención en los usuarios más activos ya que, tal como indica el análisis, aquellos jugadores que hacen más hits tienden a comprar más productos siendo los que más ingresos producen.

A nivel de gamificación, es necesario mejorar el compromiso o *engagement* con aquellos grupos de usuarios que registran menor actividad en el juego. Esto puede mejorar con campañas específicas o regalos por cada nivel finalizado, también sería posble crear nuevos ítems para estimular el compromiso hacia nuevos retos en el juego.

7. Conclusiones

El presente trabajo no solo trata de lograr cada objetivo planteado, también busca diferenciar los distintos campos que existen hoy en día para trabajar con diferentes tamaños de información. Por tanto, se ha querido remarcar las diferentes áreas que sirven para trabajar con conjuntos limitados de datos, como es la inteligencia de negocios a través de Data Warehouse,

Además de esto, se ha visto qué es lo que sucede cuando estos datos requieren ser manipulados en tiempo real o donde simplemente son grandes conjuntos de datos que necesitan ser trabajados a gran escala, es aquí donde entra en juego la solución Big Data.
También se ha hablado de la ciencia de datos, así como la metodología que presenta, el cual ha servido para ejecutar el plan de trabajo a lo largo del proyecto.

Ha sido muy importante la etapa de adquisición de la información, ya que además presentar los datos con los cuales se debe trabajar, se ha dado mucho importancia saber el modelo de los mismos. Este modelo representa la información obtenida con lo cual, se ha detallado de manera que pueda ser fácil de entender mediante un diagrama de datos relacional.

En la etapa de exploración ha servido para detallar todos los datos obtenidos mediante diagramas, utilizando Splunk como herramienta, se conocen distintos tipos de información que, sin ser explorados, sería imposible poder conocerlos.

La fase de pre-procesado de datos, sirve para eliminar aquellos datos que pueden distorsionar el análisis. Por tanto, esta fase es muy importante, siendo crítica si no se toman medidas para la “limpieza” de la información ya que puede afectar los resultados posteriores.
La parte fundamental de esta metodología ha sido el análisis elaborado utilizando diferentes herramientas. Se ha intentado aprender al máximo cada plataforma ya que, cada una de ellas, ofrece muchas posibilidades a la hora de procesar toda la información.
Al mismo tiempo, se han realizado varios tipos de análisis con el fin de enriquecer este trabajo y lograr los objetivos trazados al principio de este proyecto. Consecuentemente, se han reportado cada uno de ellos culminando con la etapa de actuación, donde se recomienda las líneas de actuación o medidas que la empresa debe llevar a cabo para satisfacer las necesidades inicialmente presentadas.

Por tanto, podemos corroborar que los objetivos iniciales han sido alcanzados, a nivel general se han detallado los puntos de actuación que la empresa debe realizar. Por otro lado, cada necesidad que al principio necesitaba saber la empresa, se han desarrollado en los análisis realizados con la conclusión que se indica en cada reporte presentado.

A nivel de planeamiento del trabajo, se ha seguido la planificación inicial, han aparecido problemas puntuales en las entregas pero estas han sido subsanadas, no presentando mayores problemas, es más, se han presentado diversos análisis que ayudan a enriquecer este proyecto. Por otro lado, la metodología utilizada ha sido totalmente acertada, siendo fácil de implementar e ayudando en el proceso de iteración de cada fase desarrollada.

A pesar de todo lo realizado, hubiera sido mejor desarrollar un poco más la parte teórica, con respecto a brindar más información acerca de las diferentes técnicas que hoy en día existen, pero no se ha realizado por tratar de no salir del contexto propio del proyecto.

A líneas de un futuro trabajo, sería posible implementar el desarrollo de nuevas técnicas, trabajar con datos en tiempo real de la aplicación móvil, así como la realización de procesado de los datos referente a las redes sociales de los jugadores para de esta forma, abrir más posibilidades a la empresa consiguiendo ingresos más ingresos publicitarios.

Para concluir, decir que este trabajo ha resultado muy beneficioso en el aprendizaje de nuevas herramientas, técnicas de análisis y sobre todo, a nivel teórico donde la ciencia de datos resulta algo más que una nueva área, sino un objetivo profesional para quién termina esta etapa estudiantil. Espero seguir recolectando este tipo de conocimientos a partir de esta valiosa experiencia.

8. Glosario

- **Anaconda**, es una plataforma Open Source desarrollado en Python, que ofrece soluciones para el ámbito de la Ciencia de Datos, en concreto para trabajar con grandes cantidades de información.
- **BI,** acrónimo de Business Intelligence, que describe a la inteligencia de negocios como el conjunto de estrategias que posibilitan el análisis de la información.
- **Big Data**, es un conjunto de tecnologías que permiten el almacenamiento, procesado y análisis de grandes cantidades de datos estructurados y no estructurados de forma escalable que permiten ganar nuevos conocimientos para la toma de decisiones.
- **Blogger**, es un servicio de blogs o páginas web orientadas a la creación de bitácoras a través de la web.
- **Clustering,** es el agrupamiento de conjuntos de datos similares no etiquetados, cada grupo es llamado *cluster*.
- **Csv,** acrónimo de *Comma Separated Values*, son un tipo de documentos en formato abierto que sirven para representar datos en forma de tabla, donde las columnas están separadas por comas y las filas por saltos de líneas.
- **Dashboard,** es una representación gráfica de los principales indicadores que intervienen en la consecución de objetivos empresarial orientado a la toma de desiciones.
- **Data Mart,** Son subconjuntos de almacenes o Data Warehouses orientados a áreas específicas y diseñados con el fin de unir diferentes áreas departamentales dentro de una misma organización.
- **Data Minning** o Minería de datos, es un campo que busca descubrir patrones en grandes volúmenes de datos.
- **Data Science**, acrónimo de la Ciencia de datos, es un campo interdisciplinar que busca, mediante métodos científicos, extraer conocimiento de los datos. Esta disciplina se apoya en distintos campos como la Estadística, Data Mining, Machine Learning, Análisis predictivo y Ciencias de la computación.
- **Data Warehouse,** es un almacén de datos orientado a objetos, no volátil[20], integrado por colecciones de datos que pueden variar en el tiempo, creados para posteriormente transformarlos en información útil en la toma de decisiones organizacional.
- **Drag and drop,** es el movimiento de arrastrar o desplazar con el puntero un objeto en la pantalla.
- **Facebook**, red social muy popular que conecta redes de personas.
- **Framework**, es un conjunto estandarizado de conceptos, prácticas y criterios para enfocar un tipo de problemas particular.
- **Hardware**, partes físicas o tangibles de un sistema informático.
- **Hortonsworks**, es una compañía de software especializada en Big Data, desarrolla y ofrece soporte a Apache Hadoop para el procesamiento de grandes conjuntos de datos.
- **Instagram,** es una red social y aplicación que permite subir fotos y videos desde cualquier punto de localización mediante una conexión a internet.

[20] No volátil es un término que es referido cuando la información almacenada no se modifica ni elimina.

- **IoT,** acrónimo de Internet de las Cosas, es utilizado para nombrar aquellos sistemas que están conectados entre ellos y a través de internet.
- **K-Mean,** es un método de agrupamiento que tiene como objetivo la partición de un conjunto en n observaciones en k grupos, en el que cada observación pertenece al grupo cuyo valor medio es más cercano.
- **Linkedin**, es una red social especializada en las relaciones laborales de sus usuarios y las empresas. Cada usuario tiene un perfil con sus datos curriculares con el fin de crear enlaces entre usuarios y empresas.
- **Log**, es un registro oficial de eventos durante un rango de tiempo determinado.
- **Machine Learning**, es un campo de estudio que se centra en los sistemas de computación que pueden aprender de los datos, donde estos sistemas a menudo son llamados modelos que son capaces de ejecutar especificas tareas para analizar varios ejemplos de un determinado problema.
- **OLAP (On-Line Analytical Processing),** Son cubos de información que tienen un número indefinido de dimensiones, donde cada cubo contiene datos de una determinada variable que se desea analizar, proporcionando un vista lógica de los datos provistos por el sistema de información hacia el Data Warehouse.
- **Open Source**, es la expresión con la que se conoce al programa distribuido y de libre desarrollo.
- **Oracle**, *Oracle Corporation* es una compañía de software que desarrolla base de datos y sistemas de gestión de bases de datos.
- **PySpark,** libería basada en Apache Spark, que mediante el lenguaje de programación Python.
- **Python,** es un lenguaje de programación interpretado, es orientado a objetos y funcional. Ofrece un tipado dinámico y se puede utilizar en distintos sistemas operativos ya que es multiplataforma.
- **Pivotal,** es una compañía de software localizada en San Francisco. Ofrece productos Big Data para el procesamiento de grandes conjuntos de datos.
- **Streaming,** se relaciona con el flujo de datos en tiempo real.
- **Software**, equipo o soporte lógico que hacen posible la realización de tareas específicas en contraposición de los componentes físicos llamados Hardware.
- **Twitter**, es un servicio de micro blog que permite enviar un conjunto limitado de palabras llamados tweet mediante texto plano, imágenes, enlaces, etc.
- **VirtualBox** es una herramienta de virtualización desarrollada por Oracle Corporation que soporta diferentes arquitecturas (x86/amd64[21]), que sirve instalar sistemas “invitados” dentro de un sistema “anfitrión”.
- **Youtube**, es un sitio web dedicado a compartir videos entre usuarios o grupos de usuarios.

[21] x86/amd64, son denominaciones que se dan al distintito tipo de procesador, siendo x86 microprocesadores Intel, amd64 microprocesadores AMD.

9. Bibliografía

i. **Apache Spark.** (2017) Python Programming Guide. [En línea]
ii. https://spark.apache.org/docs/0.9.1/python-programming-guide.html [Consulta: 03 de Abril del 2017]

iii. **Brown M.** (2011). In Gaming, Free to play can pay. [En línea] https://www.wired.com/2011/06/free-to-play/ [Consulta: 03 de Abril del 2017]

iv. **Benjamin E.** (2014). The average user does not exist in fermium gaming. [En línea] http://mobiledevmemo.com/the-average-user-doesnt-exist-in-freemium-gamin/ [Consulta: 03 de Abril del 2017]

v. **Cano J.L**. (2007). Business Intelligence: Competir con información.Llibre publicat per ESADE, Banesto. [En línea] http://itemsweb.esade.edu/biblioteca/archivo/Business_Intelligence_competir_con_informacion.pdf [Consulta: 20 de Abril del 2017]

vi. **Center for Machine Learning and Intelligence Systems,** (2017). Machine Learning Repository [En línea] http://archive.ics.uci.edu/ml/datasets.html [Consulta: 15 de Mayo del 2017]

vii. **Data Camp.** (2017). Data Analysis and Interpretation Tutorials. [En línea] https://www.datacamp.com/community/tutorials [Consulta: 20 de Mayo del 2017]

viii. **Dutcher J**. (2014). What is Big Data? University of Barkeley, California. [En línea]
ix. https://datascience.berkeley.edu/what-is-big-data/ [Consulta: 20 de Abril del 2017]

x. **Duvvuri S.; Singal B.** *Spark for Data Science.* Publicado por Packt Publishing. Birmingham, 2016.

xi. **Fundación Wikipedia**. (2016). Coeficiente de Agrupación. [En línea]
xii. https://es.wikipedia.org/wiki/Coeficiente_de_agrupamiento [Consulta: 05 de Mayo del 2017]

xiii. **Gartner** (2017). IT Glossary: Business Intelligence. [En línea]
xiv. http://www.gartner.com/it-glossary/business-intelligence-bi/ [Consulta: 20 de Abril del 2017]

xv. **Gartner** (2017). IT Glossary: Big Data. [En línea]
xvi. http://www.gartner.com/it-glossary/big-data/ [Consulta: 15 de Abril del 2017]

xvii. **Igual L.; Seguí S.** *Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications*. Publicado por Springer International Publishing, Switzerland 2017

xviii. **Kemper C**. *Beginning Neo4j*. 1ª Edición. Publicado por Apress, New York, 2015.

xix. **Kromer F.** (2008). Massive Scrape of Twitter Friend Graph. [En línea] http://blog.infochimps.com/2008/12/29/massive-scrape-of-twitters-friend-graph/ [Consulta: 13 de Mayo del 2017]

xx. **Mahesh L**. *Neo4j Graph Data Modeling*. Publicado por Packt Publishing. Birmingham, 2015

xxi. **Miller J**. *Mastering Splunk*. Publicado por Packt Publishing. Birmingham, 2015

xxii. **Manyika J.; Brown B.; Bughin J.; Dobbs R.; Roxburgh C.; Byers A.** (2011). Big Data: The next frontier for innovation, competition and productivity, McKinsey Global Institute [En línea] http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation [Consulta: 18 de Mayo del 2017]

xxiii. **Manyika J.; Chui M.; Farrell D.; Van Kuiken S.; Groves P.; Doshi E.** (2013). Open Data: Unlocking innovation and performance with liquid information. [En línea] http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information [Consulta: 18 de Mayo del 2017]

xxiv. **Navarro J**. (2016). Big Data y Retail: cómo mejorar los objetivos de negocio. [En línea] http://cleverdata.io/big-data-retail/ [Consulta: 13 de Mayo del 2017]

xxv. **Navarro J**. (2016). Diferencias entre Business Intelligence y Machine Learning. [En línea] http://cleverdata.io/diferencias-bi-machine-learning [Consulta: 13 de Mayo del 2017]

xxvi. **Robinson I.; Webber J.; Eifrem E**. *Graph Databases*. Publicado por O'Reilly Media, California, 2015.

xxvii. **Salvador F.** *Big Data: ¿La ruta? O ¿El destino?* publicado por IE Foundation, Oracle. California. 2014

xxviii. **SAS** (2017). Big Data, What is Big Data. [En línea]
xxix. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html [Consulta: 20 de Abril del 2017]

xxx. **Sitto K.; Presser M.** *Field Guide to Hadoop*. Publicado por O'Reilly Media, California, 2015.

xxxi. **Smith K**. *Splunk Developers Guide*. Publicado por Packt Publishing. Birmingham, 2015

xxxii. **Strong D.; Baker P.** (2017). The Best Self-Service Business Intelligence (BI) Tools of 2017. [En línea] http://www.pcmag.com/article2/0,2817,2491954,00.asp [Consulta: 20 de Abril del 2017]

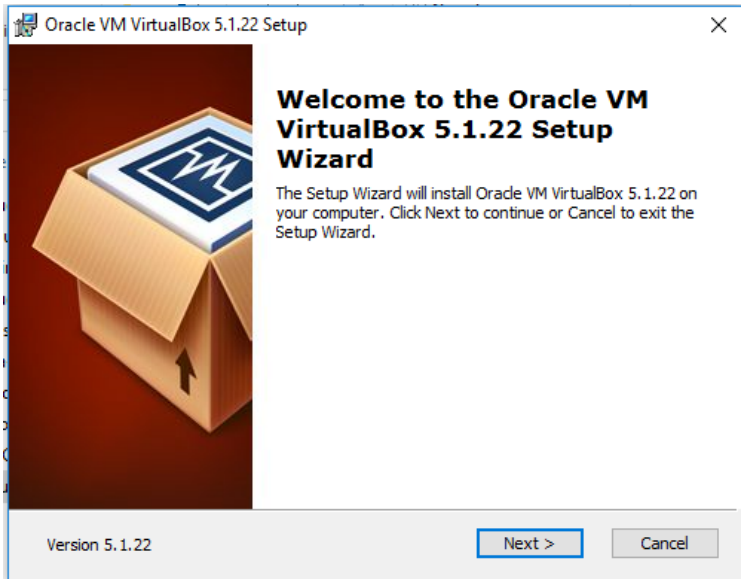
xxxiii. **Walz A.** (2015). The Data Behind Customer Acquisition and Retention For F2P Mobile Games. [En línea] https://www.apptentive.com/blog/2015/04/09/the-data-behind-customer-acquisition-and-retention-for-f2p-mobile-games/ [Consulta: 03 de Abril del 2017]

10. Anexos

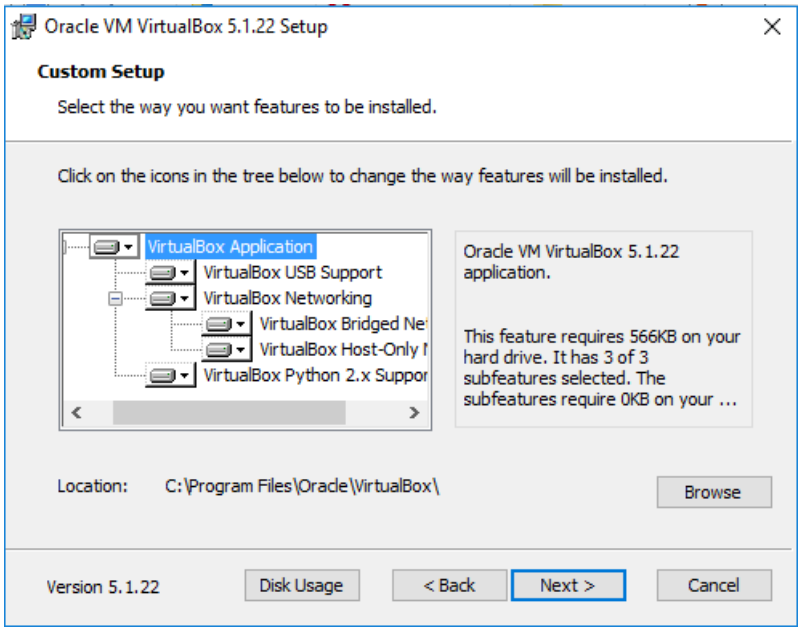
A. Planificación
B. Procesos de Instalación
B.1. VirtualBox

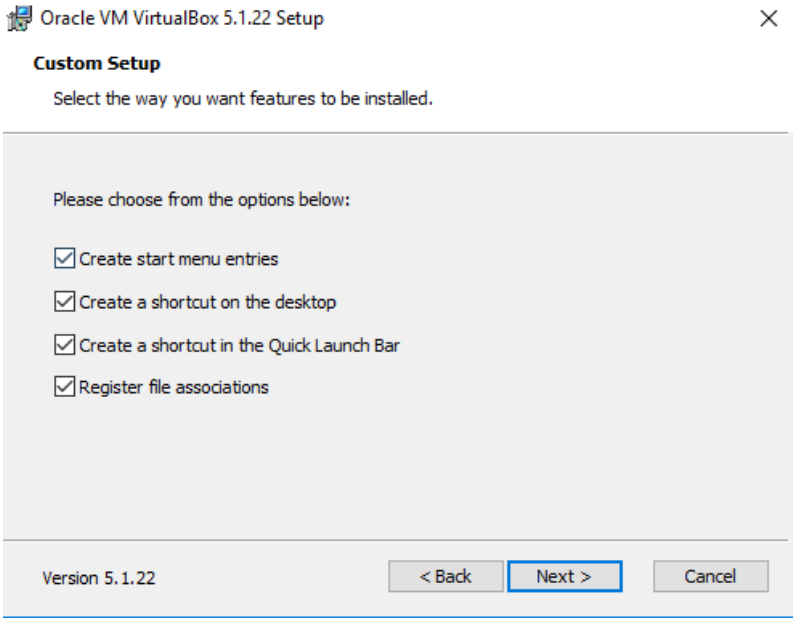
Para instalar la herramienta es necesario dirigirse a la web: https://www.virtualbox.org/wiki/Downloads, una vez allí, es necesario descargar la versión más actual dependiendo del sistema operativo, en este caso VirtualBox-5.1.22-115126-Win para Windows hosts

Una vez descargado el programa, procedemos a ejecutarlo.



En el siguiente paso, Se escogen los parámetros por defecto, para luego más adelante especificar los detalles técnicos para la máquina virtual. Se procede seleccionar la creación de accesos directos al escritorio o desde el menú inicio.
Por último, se dará clic en finalizar para terminar con la instalación.

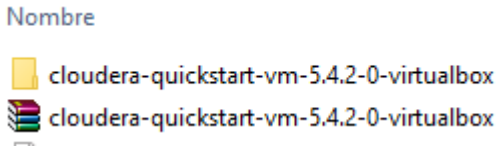




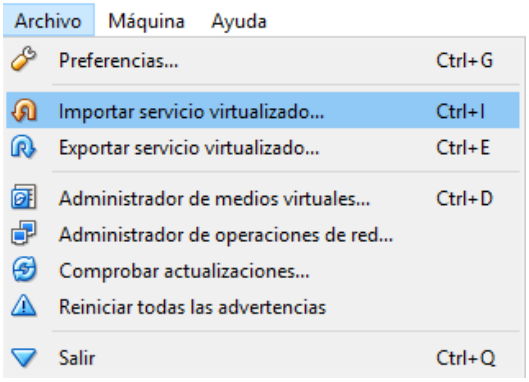
B.2. Cloudera

Para instalar Cloudera, se debe descargar la máquina virtual pre-configurada desde el siguiente enlace: https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.4.2-0-virtualbox.zip. Puede tardar en la descarga dependiendo la velocidad de conexión a internet, ya que el archivo es mayor a 4GB.

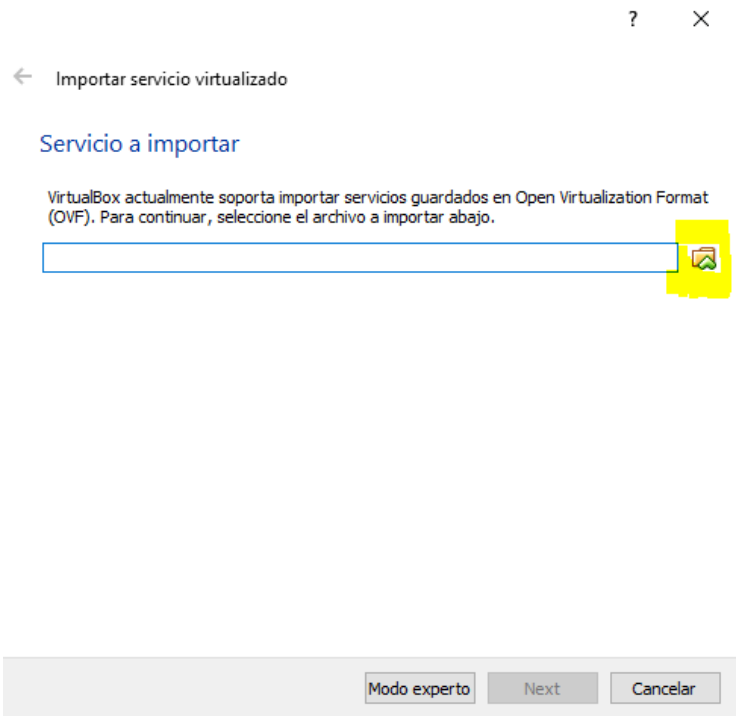
Una vez descargado el archivo, es necesario descomprimirlo, obteniendo la carpeta siguiente:



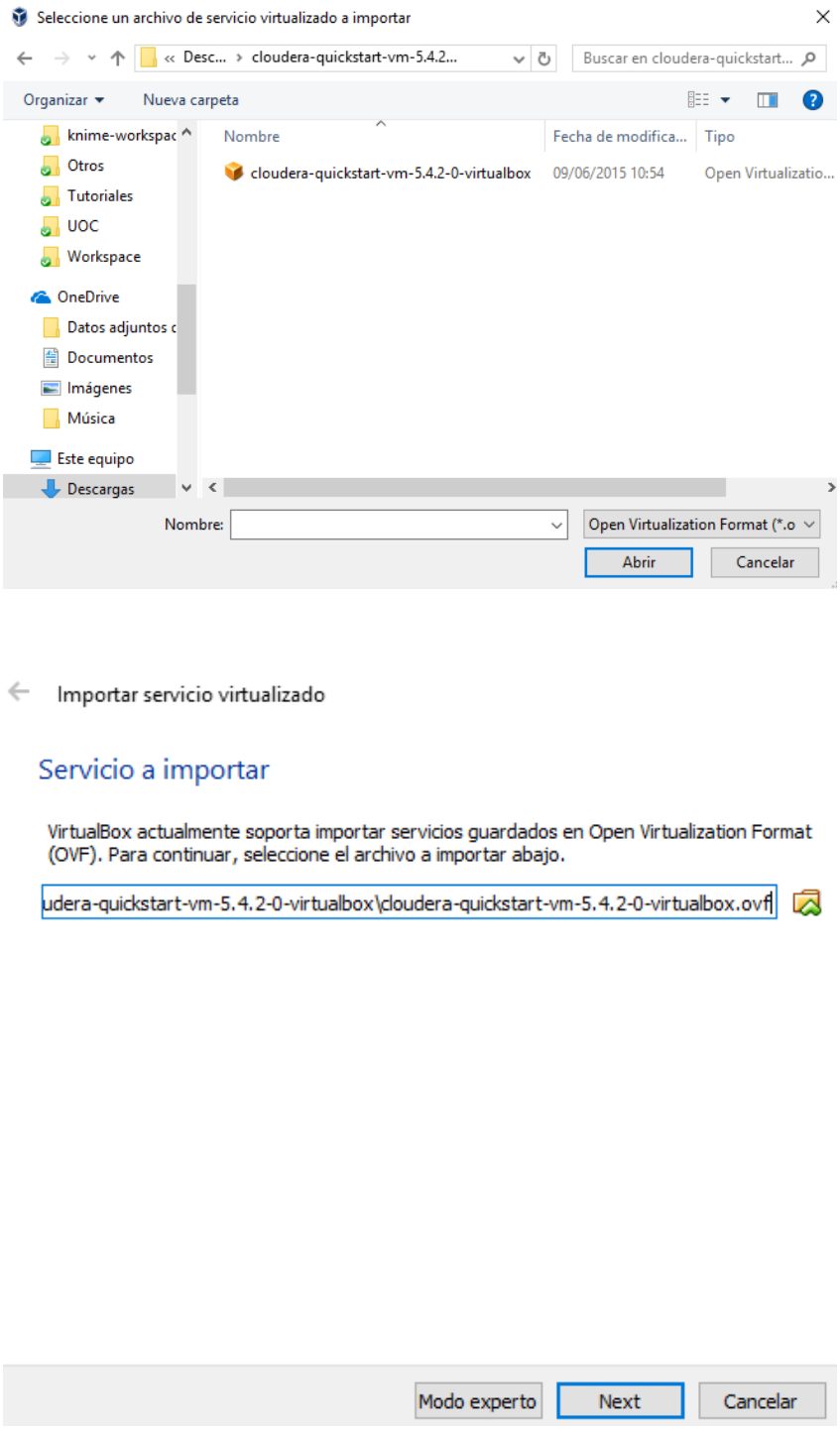
Ahora se procede a importar la máquina virtual en VirtualBox. Para esto se abre la aplicación y se selecciona archivo → Importar servicio virtualizado.



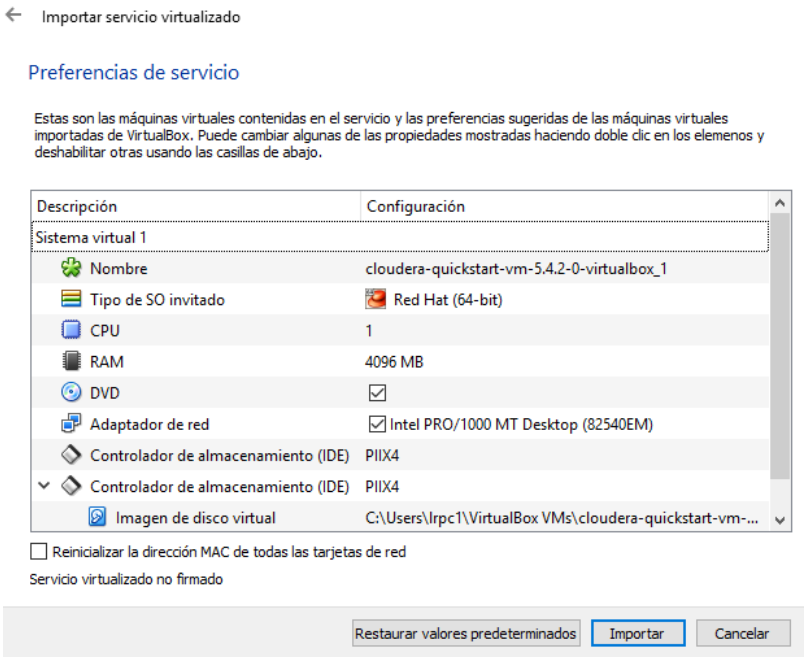
Después de esto, se dará clic en icono para buscar el archivo descargado anteriormente.



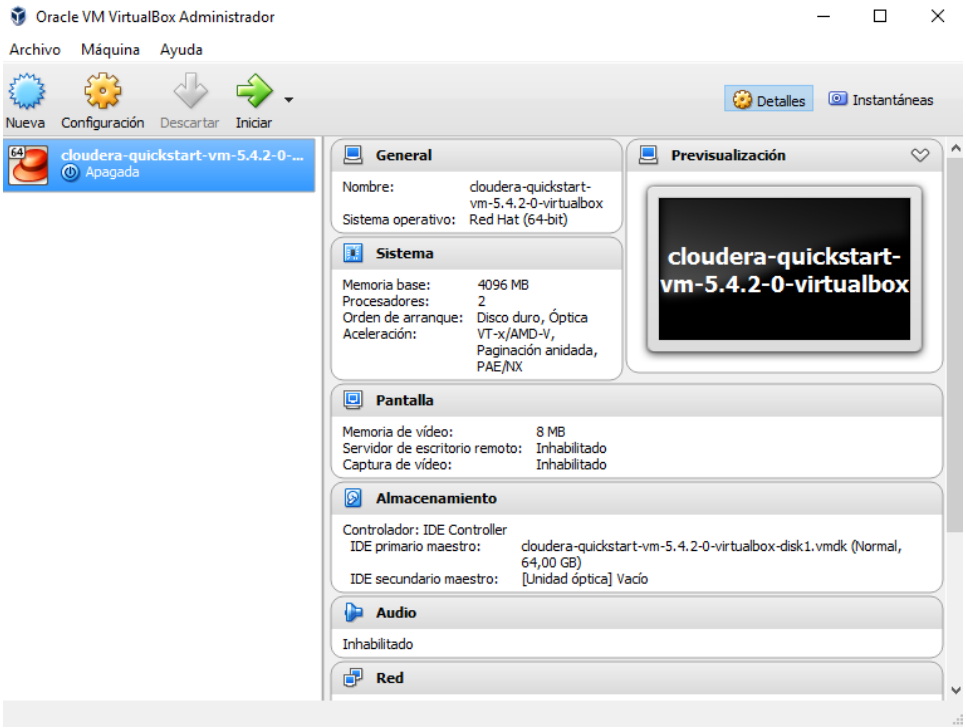
En el siguiente paso, se seleccionará el archivo que hemos descargado.



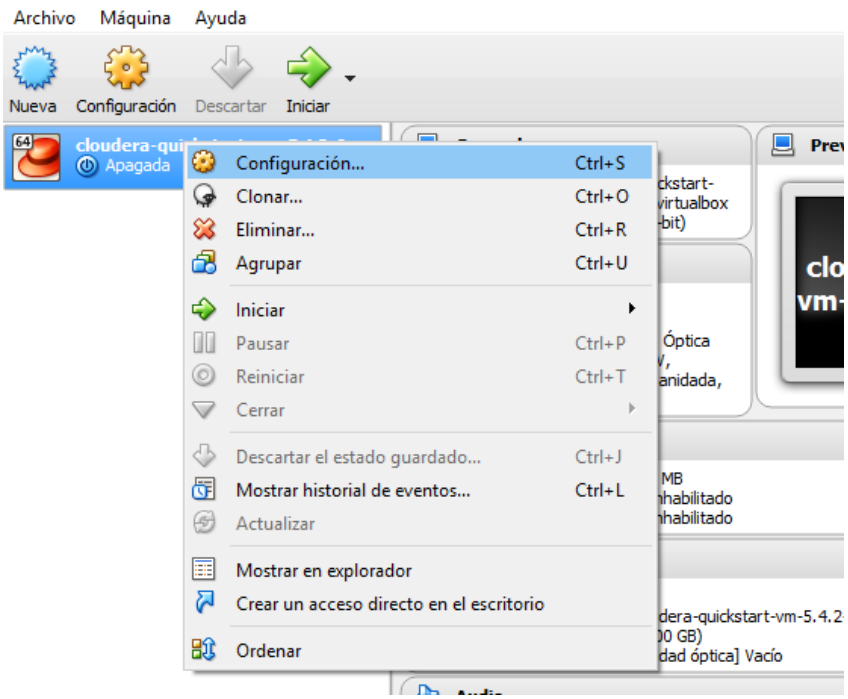
En la siguiente ventana se dejarán los valores predeterminados para luego cambiarlos en la fase posterior. Se da clic en importar para que empiece el proceso, esto puede tardar unos minutos dependiendo de las características del sistema donde se emplee.



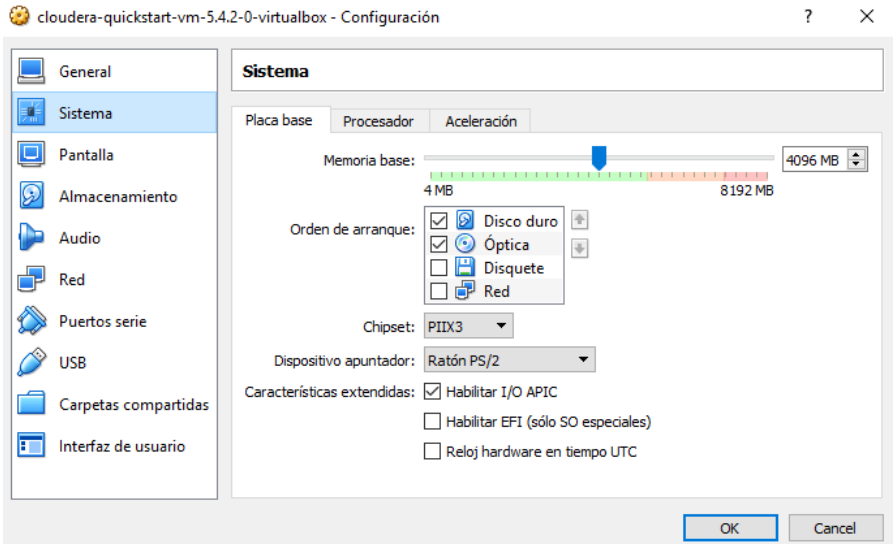
Al final de la importación, se observa que ya tenemos la maquina disponible.

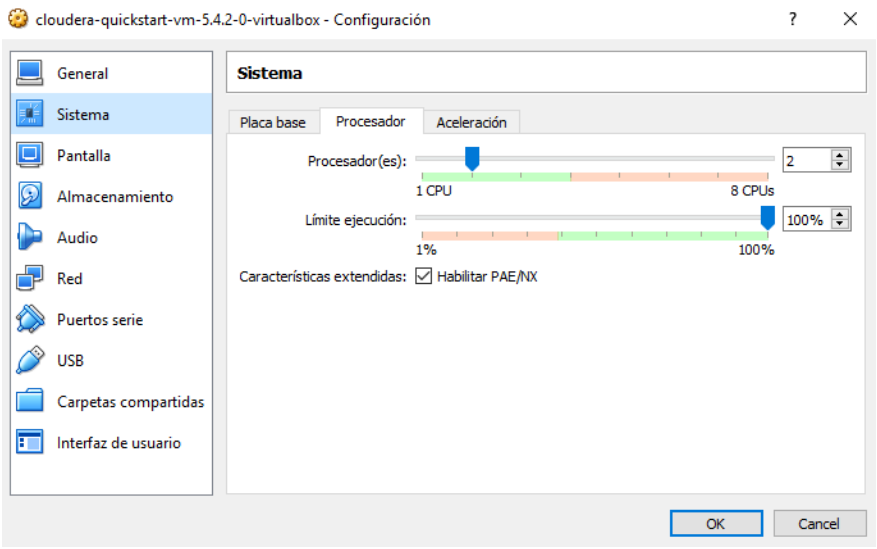


Para cambiar la configuración de la máquina virtual, se debe seleccionarla y dar clic derecho para ingresar en configuración.

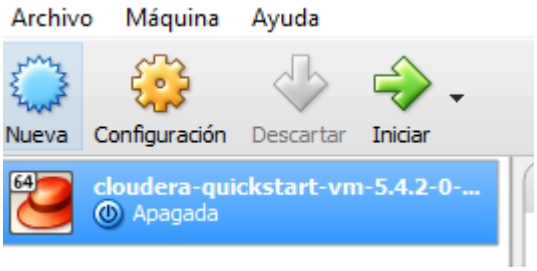


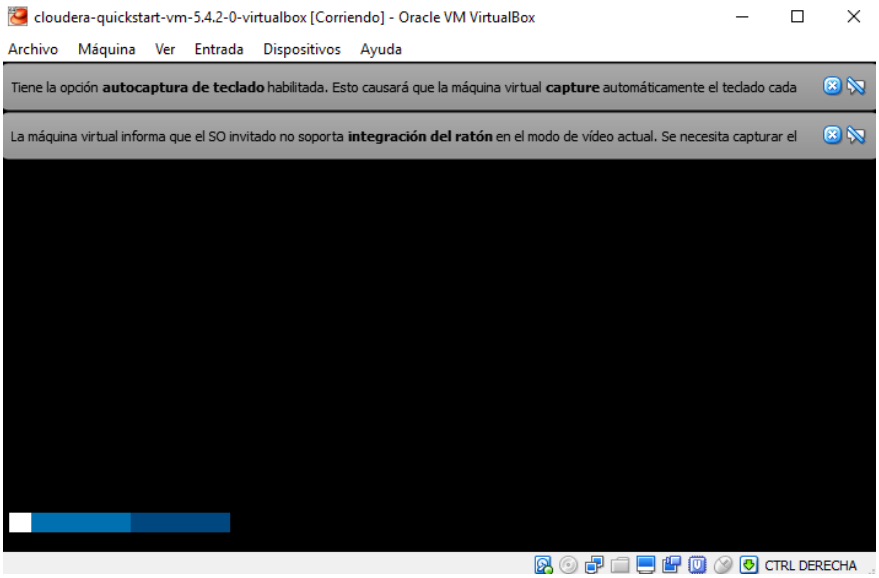
Nos interesa el apartado *Sistema*, el cual tiene la configuración de memoria, procesador y los periféricos. Como se puede apreciar, se puede cambiar el tamaño de la memoria RAM y el número de procesadores.



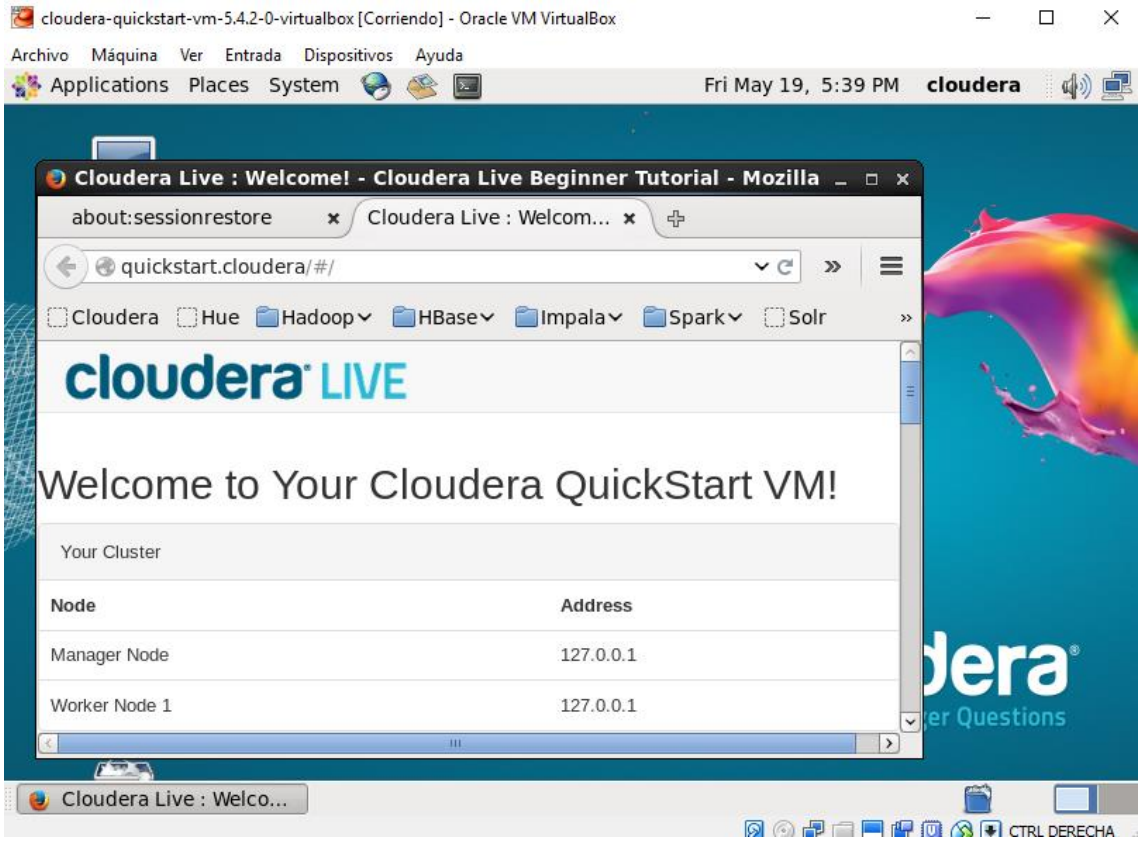


Una vez configurada, se procede iniciar desde el botón superior. Esto puede tardar varios minutos para que inicie y esté completamente activa.





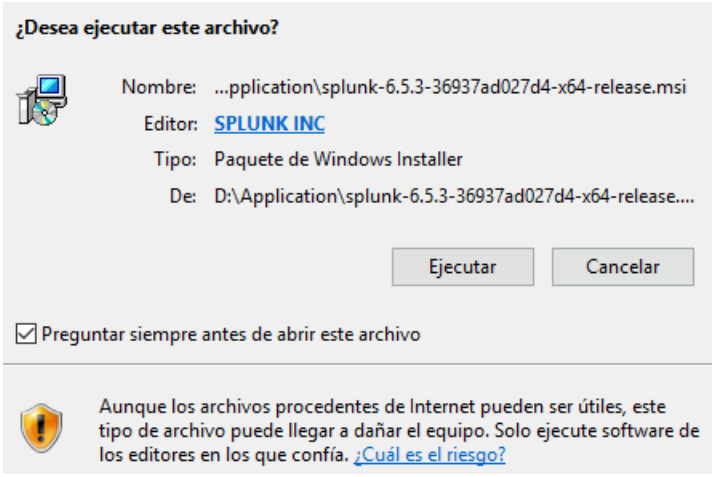
Una vez iniciada, aparecerá la pantalla de inicio de Cloudera junto con la ventana abierta del navegador.

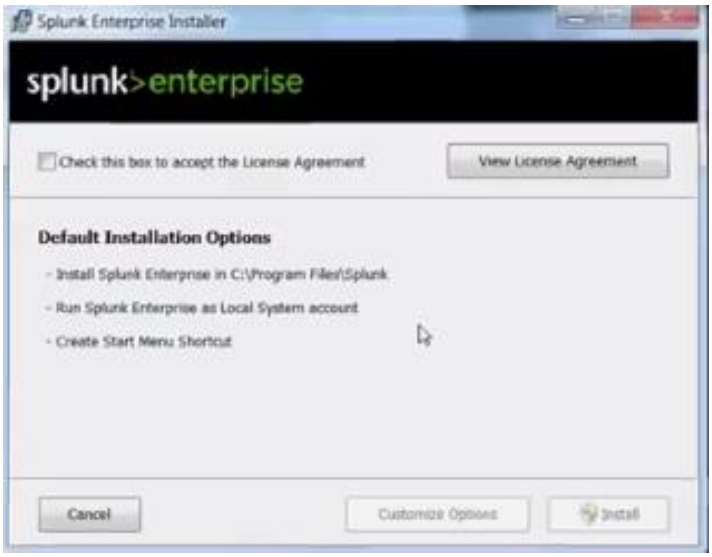


B.3. Splunk

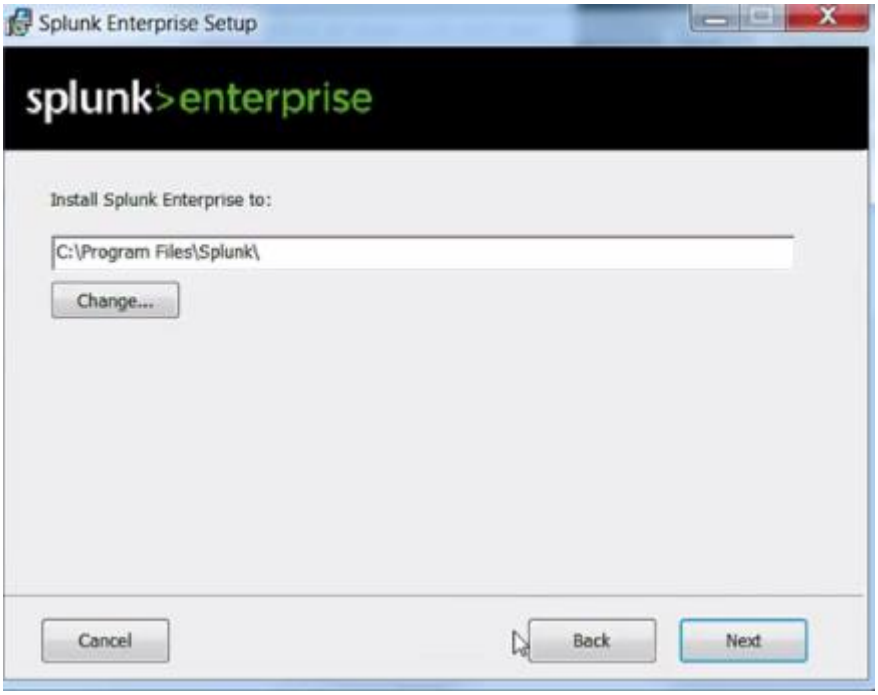
En la instalación de este programa se necesita descargar el siguiente archivo splunk-6.5.3-36937ad027d4-x64-release desde el siguiente enlace: https://www.splunk.com/page/previous_releases#x86_64windows

Una vez descargado se procede a instalarlo, al iniciar nos preguntara si se desea ejecutarlo, seleccionamos ejecutar y procedemos a aceptar la licencia de uso y se procede a continuar





En la siguiente pantalla se podrá cambiar la dirección donde se instalará la aplicación. Se selecciona en siguiente para seguir con la instalación.



En la siguiente ventana se selecciona la instalación de forma local, ya que no se necesita una cuenta de dominio, puesto que los datos están dentro de un único sistema. Se continuará la instalación seleccionando el icono siguiente.

Por último, se selecciona si es necesario un acceso directo al menú de inicio del sistema.

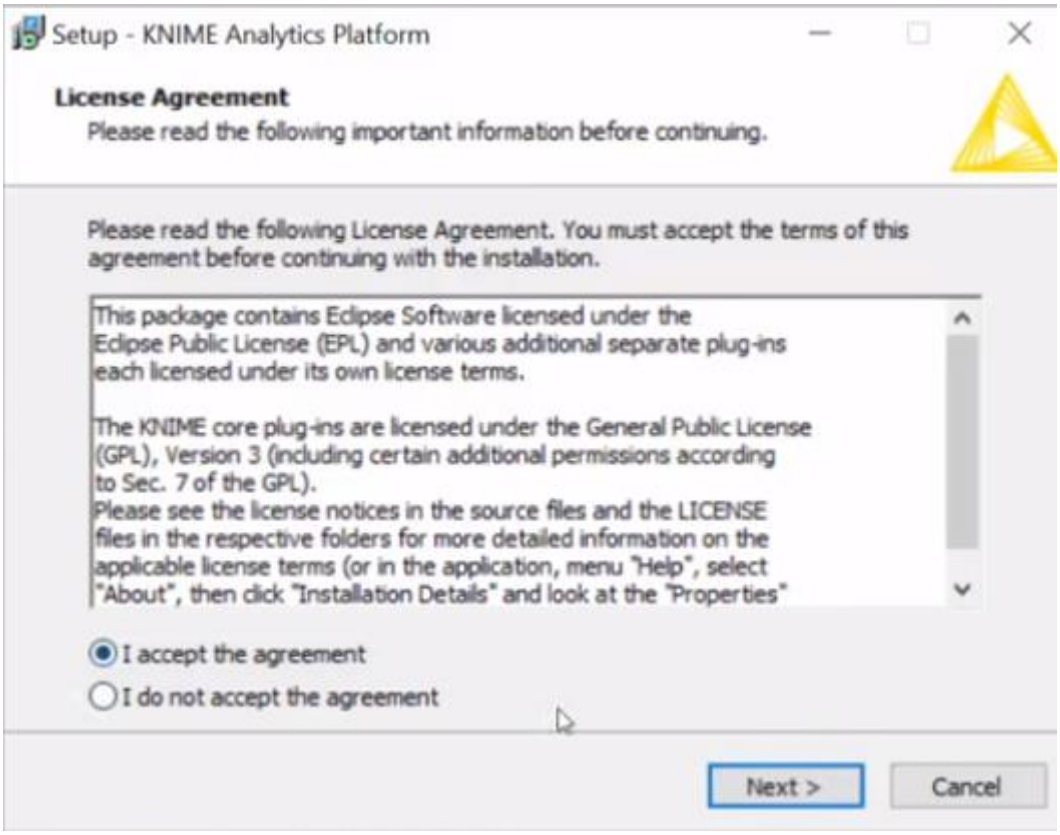
Una vez finalizada la instalación, al abrir la aplicación se presentará una pantalla donde es necesario logearnos para acceder. Se podrá ingresar con estos datos por primera vez ya que luego se podrá cambiar la contraseña desde el menú configuración.

B.4. KNIME

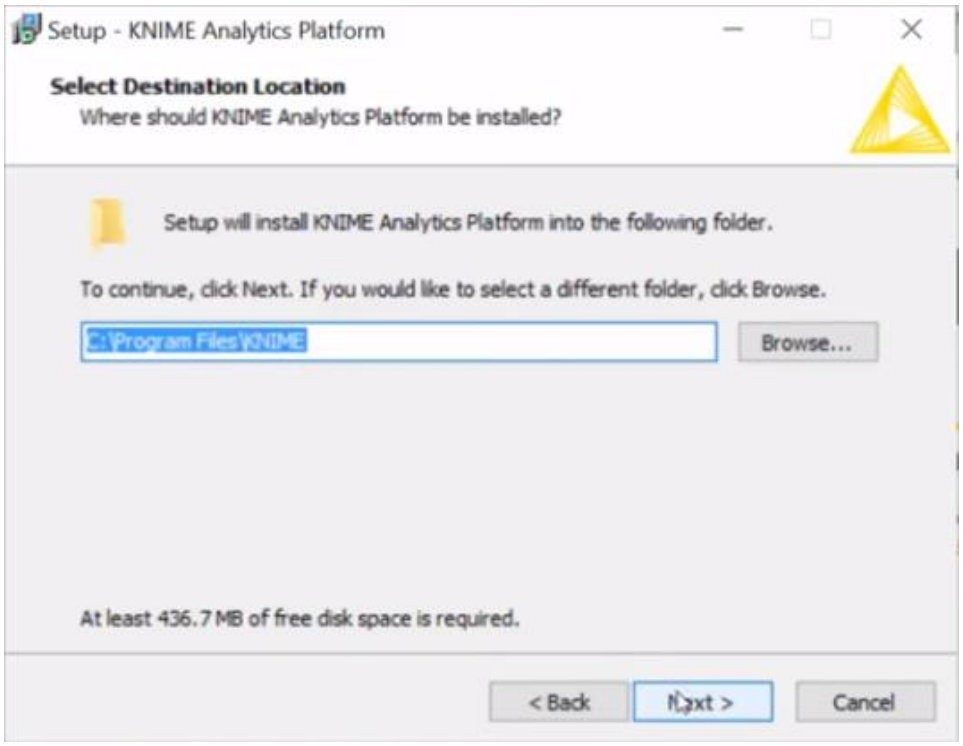
Este apartado describe la instalación de la aplicación KNIME para el pre-proceso y pre-análisis de los datos.

Antes de instalar, es necesario descargarse el instalador desde el siguiente enlace:
https://www.knime.org/downloads/overview?quicktabs_knimed=1#quicktabs-knimed
En nuestro caso, se descargará el instalador para Windows de 64 bits.

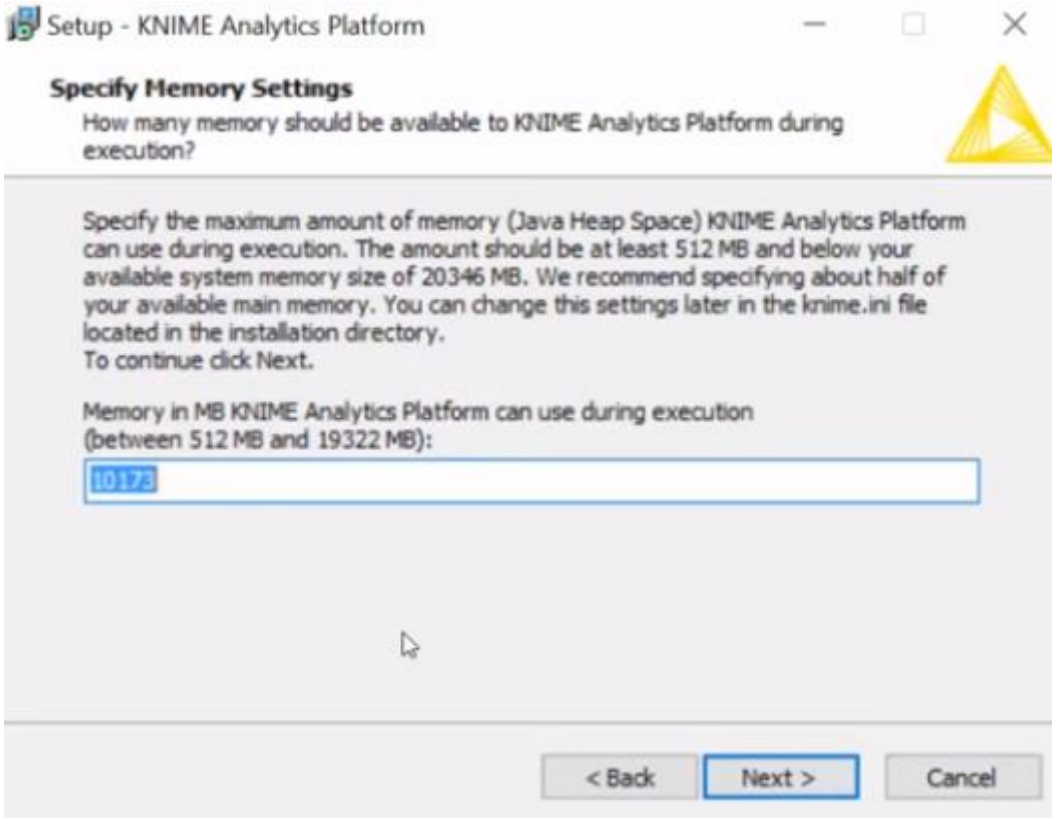
Una vez descargado, se debe dar clic en el instalador y aparecerá la ventana del acuerdo de licencia, el cual se deberá aceptar y se seguirá con la instalación.



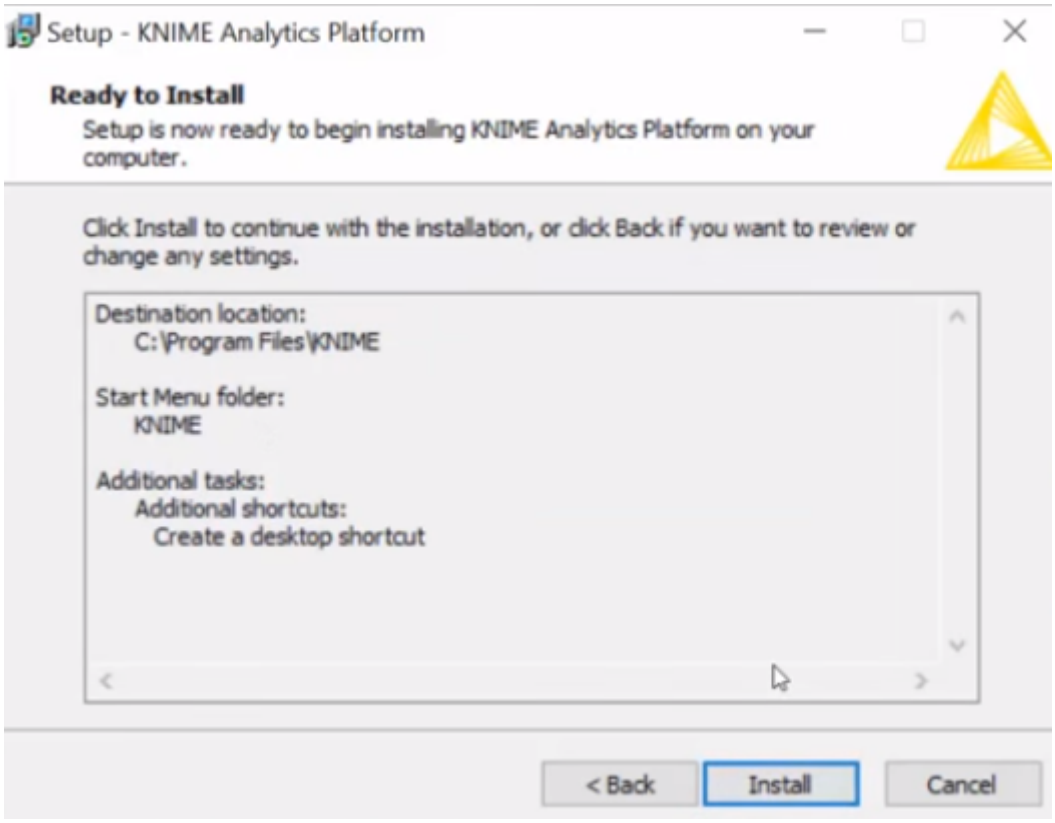
En el siguiente paso, se selecciona la ruta donde se guardará los archivos a instalar, se selecciona por defecto y se da clic en siguiente.



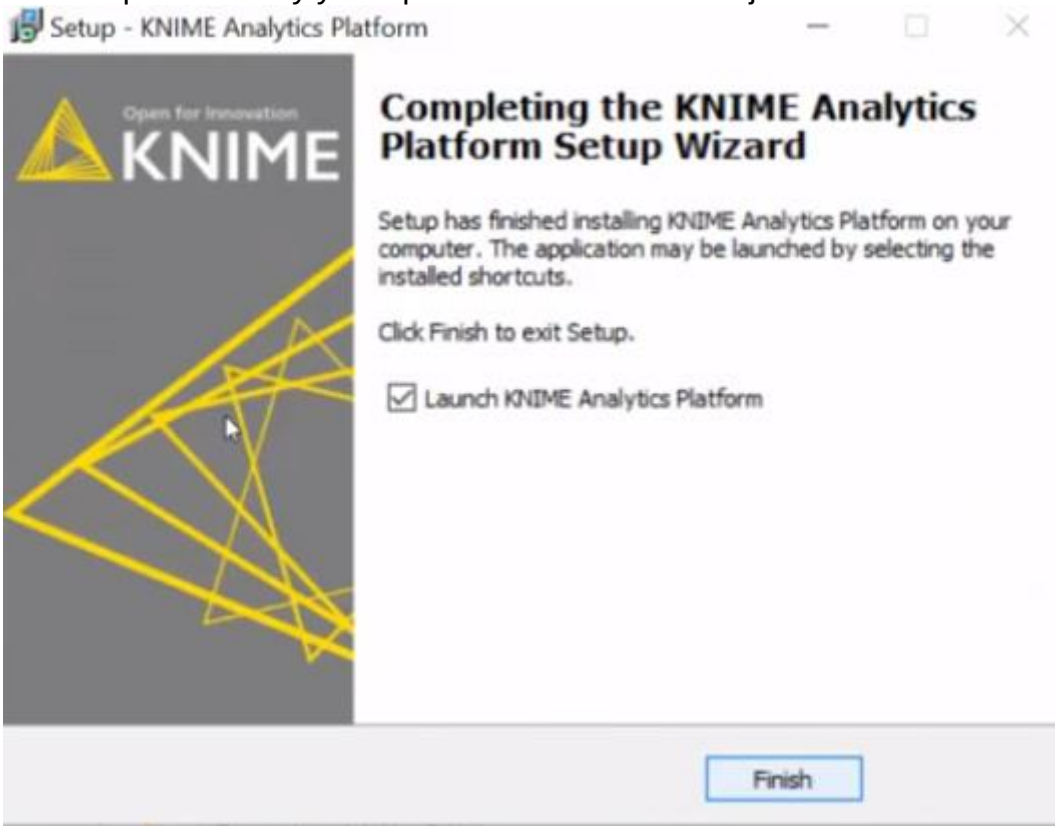
En los siguientes pasos nos muestra opciones para crear accesos directos, tanto en el menu inicio y el escritorio, se marcará por defecto y se dará clic en continuar. Después de esto preguntará cuanta memoria se debe dejar para el consumo de la aplicación, se dejará en 4048 MB

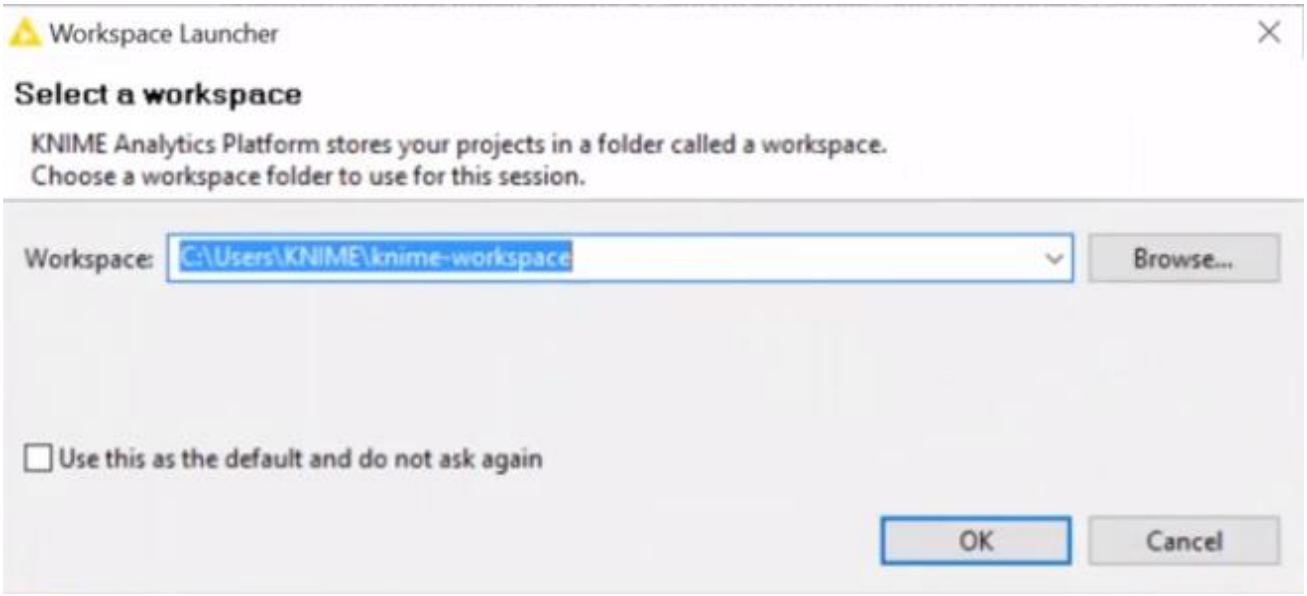


En el siguiente paso, se procede a instalar la plataforma.



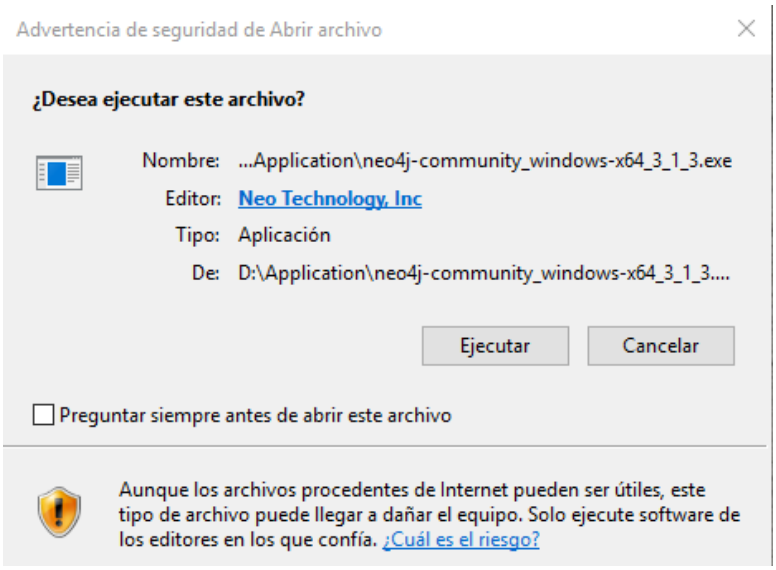
Por último, se procede a lanzar la aplicación que nos pedirá la ruta para guardar los trabajos que se vaya realizando, esta carpeta será llamará *Knime-workspace.*
Se dejará la ruta por defecto y ya se puede comenzar a trabajar.



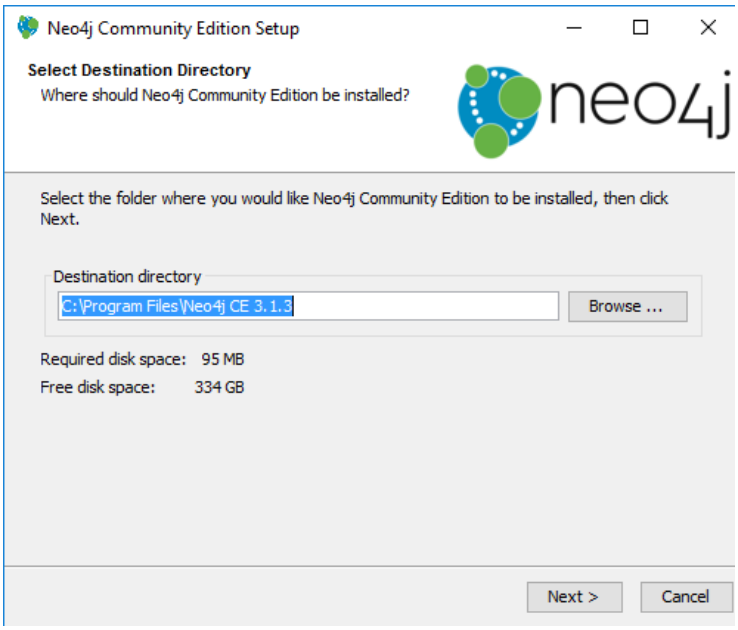


B.5. Neo4j

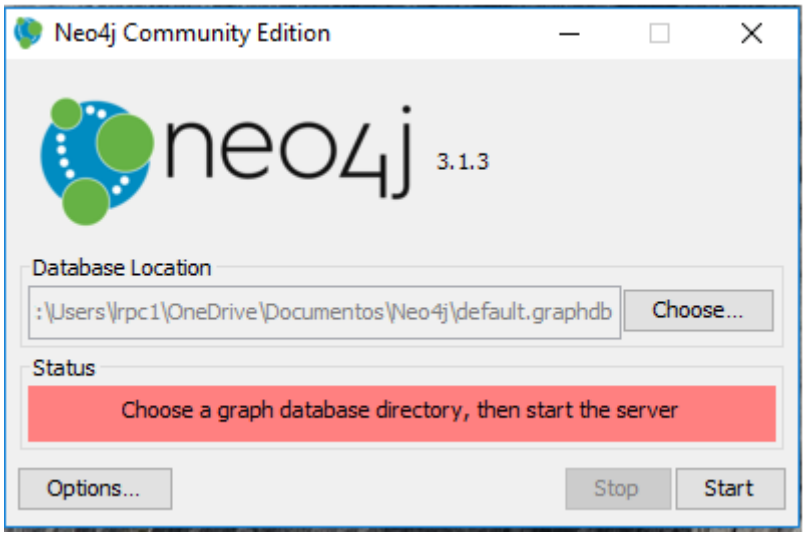
Para instalar es necesario descargar la última versión del instalador para Windows 10 desde la siguiente dirección: https://neo4j.com/download/
Una vez descargado, se procede a la instalación seleccionando el botón ejecutar como se aprecia en la siguiente imagen:



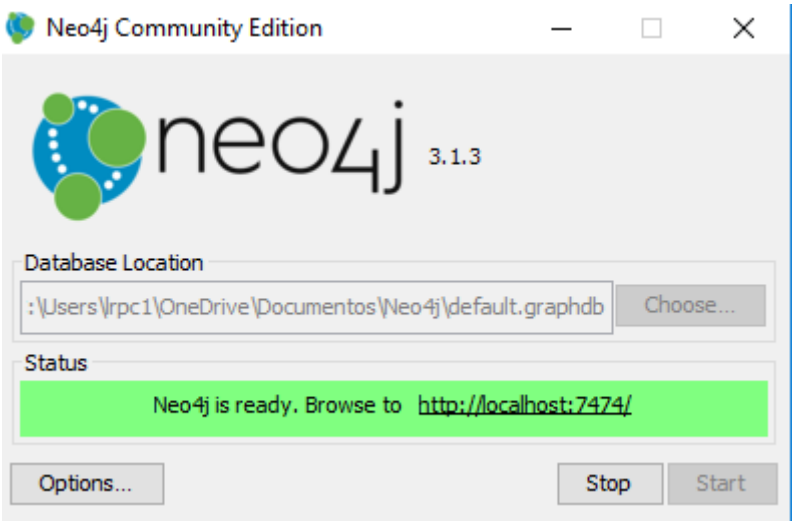
Seleccionar el directorio donde se almacenarán los datos de la aplicación, se dejará la ruta por defecto y seleccionar el boton Next.



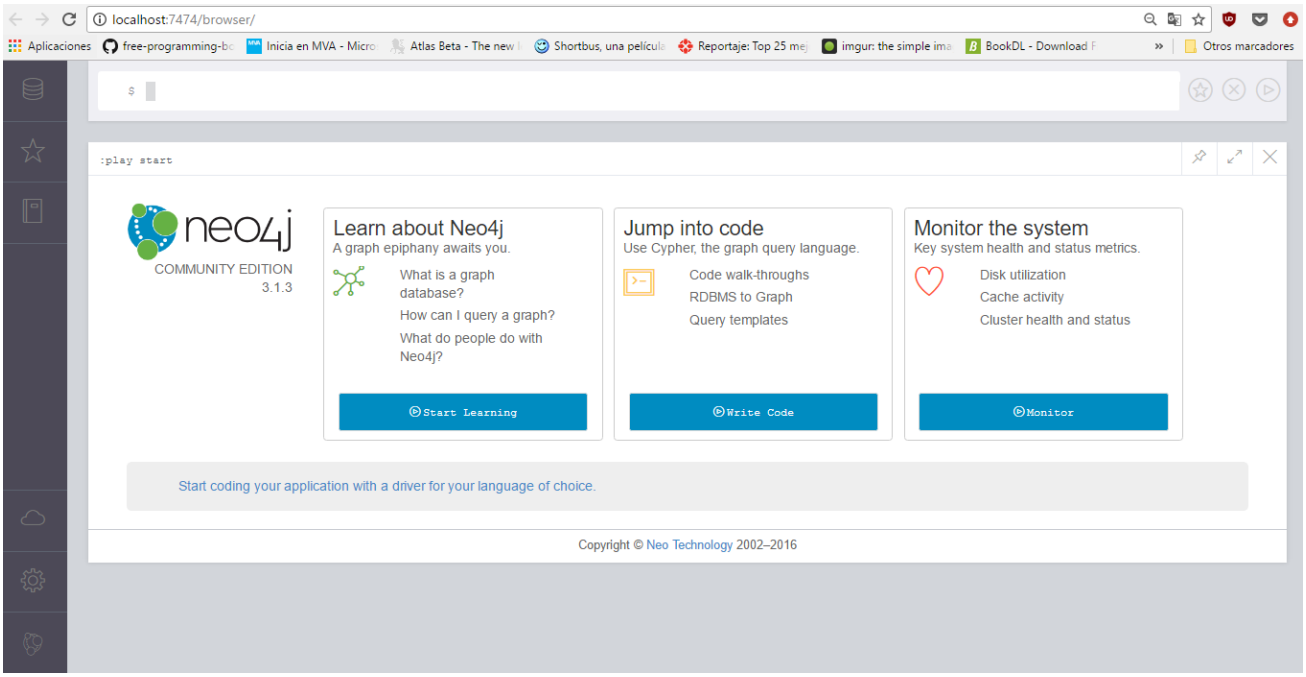
Al terminar la instalación, se debe ejecutar el instalador, aparecerá la siguiente ventana donde se debe seleccionar la ruta donde se encuentra la base de datos. Se eligirá la localización por defecto, seguidamente se selecciona en Start para iniciar la aplicación.



El siguiente paso es abrir en el navgador con el enlace que se indica, que abrirá e puerto 7474 mediante el localhost, como se puede apreciar, se puede parar en cualuier momento la aplicación mediante la opción Stop.



Se abrirá el navegador con la siguiente ventana de inicio donde ya podrá trabajar.

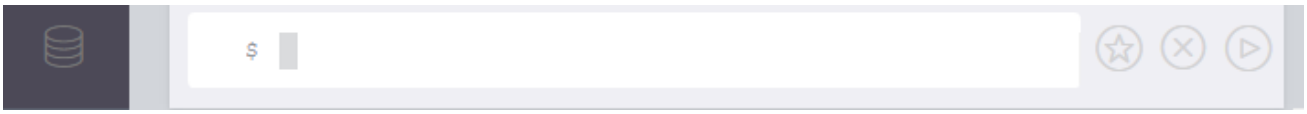


C. Código fuente

C.1. Carga datos a Neo4j

Código fuente utilizado para carga de datos de las seis tablas a la aplicación Neo4j. Se detallan los pasos a realizar:

- Crear un directorio dentro del directorio de Neo4j donde se guardarán las tablas, este directorio tendría el siguiente esquema:
- *"C:\Users**Alias**\OneDrive\Documentos\Neo4j\default.graphdb\import"* Donde Alias es el nombre de usuario de Windows.
- Se ejecutará los siguientes códigos desde la barra de codificación en Neo4j, uno a uno para cada una de las seis tablas con formato csv.




```
LOAD CSV FROM "file:///chat_create_team_chat.csv" AS row
MERGE (u:User {id: toInteger(row[0])})
MERGE (t:Team {id: toInteger(row[1])})
MERGE (c:TeamChatSession {id: toInteger(row[2])})
MERGE (u)-[:CreatesSession{timeStamp: row[3]}]->(c)
MERGE (c)-[:OwnedBy{timeStamp: row[3]}]->(t)

LOAD CSV FROM "file:///chat_join_team_chat.csv" AS row
MERGE (u:User {id: toInt(row[0])})
MERGE (c:TeamChatSession {id: toInt(row[1])})
MERGE (u)-[:Joins {timeStamp: row[2]}]->(c)

LOAD CSV FROM "file:///chat_leave_team_chat.csv" AS row
MERGE (u:User {id: toInt(row[0])})
MERGE (c:TeamChatSession {id: toInt(row[1])})
MERGE (u)-[:Leaves {timeStamp: row[2]}]->(c)

LOAD CSV FROM "file:///chat_item_team_chat.csv" AS row
MERGE (u:User {id: toInt(row[0])})
MERGE (c:TeamChatSession {id: toInt(row[1])})
MERGE (i:ChatItem {id: toInt(row[2])})
MERGE (u)-[:CreateChat {timeStamp: row[3]}]->(i)
MERGE (i)-[:PartOf {timeStamp: row[3]}]->(c)

LOAD CSV FROM "file:///chat_mention_team_chat.csv" AS row
MERGE (i:ChatItem {id: toInt(row[0])})
MERGE (u:User {id: toInt(row[1])})
MERGE (i)-[:Mentioned {timeStamp: row[2]}]->(u)

LOAD CSV FROM "file:///chat_respond_team_chat.csv" AS row
MERGE (a:ChatItem {id: toInt(row[0])})
MERGE (b:ChatItem {id: toInt(row[1])})
MERGE (a)-[:ResponseTo {timeStamp: row[2]}]->(b)
```

Podemos ver desde el icono *Database Information*, cada uno de los nodos y relaciones (aristas) cargados: