UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MAESTRÍA EN ANÁLISIS Y VISUALIZACIÓN DE DATOS

Ingeniería para el Procesado Masivo de Datos

**Análisis exploratorio con Apache Hive sobre HDFS**

Presenta

**Carmen Sayuri Maldonado Pinto**

**Eduardo García García**

**Juan Carlos de Jesús Garcés Carrillo**

**Kevin David Farinango Cinilin**

Agosto 2022

Contenido

[Introducción 1](#_Toc110794888)

[Objetivo 1](#_Toc110794889)

[Desarrollo 1](#_Toc110794890)

[Fuentes de Información por utilizar 2](#_Toc110794891)

[Información del dataset 2](#_Toc110794892)

[Pasos de la ingesta de datos 3](#_Toc110794893)

[Proceso en el Notebook 4](#_Toc110794894)

[Ejercicio 14](#_Toc110794895)

[Conclusiones 17](#_Toc110794896)

[Hoja de control 18](#_Toc110794897)

Introducción

En el presente análisis, se trabajó con la información proporcionada en los csv adjuntos (features, stores, sales). La practica la realizamos principalmente con la herramienta databricks, enfocándonos en utilizar el lenguaje de SQL con Spark.

Objetivo

Realizar un análisis utilizando las herramientas y teoría vistas en la materia. Realizar un proyecto de colaboración y aplicación como científicos de datos.

Desarrollo

Se utiliza el repositorio de Github para compartir el código con los miembros del equipo como se puede ver en Ilustración 1

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 1Codigo y archivos en Github

Esto nos permitió trabajar de manera conjunta y ordenada con el fin de que cada integrante pudiera descargar los archivos en su disco local. Los archivos correspondientes a la información por analizar fueron cargados en la herramienta de Databricks y se creo un notebook en el cual se verán reflejados cada uno de los pasos solicitados, así como de los resultados obtenidos.

Fuentes de Información por utilizar

Como se menciono en las secciones anteriores nuestras fuentes de información son los archivos features, stores y sales.

Información del dataset

El archivo de features se compone de las variables de:

* Store, Date, Temperature, Fuel\_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment, IsHoliday.

El archivo de sales se compone de las variables de:

* Store, Dept, Date, Weekly\_Sales, IsHoliday.

El archivo de stores se compone de las variables de:

* Store, Type, Size

Como se puede observar la información que se puede extraer de features son características adicionales de la tienda y la información de la región donde se ubica como lo es la temperatura media, el precio de combustible, índice de precio al consumo, tasa de desempleo, así también información sobre la semana en que se registraron dicha información. En la tabla de sales es la información semanal de las ventas por tienda, departamento. Por último, esta la tabla de stores que contiene información general de la tiendo como el tipo y el tamaño.

Pasos de la ingesta de datos

Como se mencionó anteriormente la mayoría del trabajo se llevó a cabo en Databricks, sin embargo, para la muestra del funcionamiento con HDFS se trabajó directamente en consola, usando vagrant. Por lo cual, dicho proceso se ve reflejado en la siguiente Ilustración 2

|  |  |
| --- | --- |
| Se suben los ficheros de datos al sistema de ficheros local |  |
| Se crea la carpeta con el nombre del Equipo\_7 |  |
| Se suben los ficheros a la carpeta creada |  |

Ilustración 2 Captura de pantalla de la consola con el ejemplo del funcionamiento de HDFS

Con databricks se hizo un proceso muy similar, pero utilizando las herramientas disponibles. Primeramente, se cargó la información de los archivos csv a databricks dicho proceso se muestra en la siguiente imagen.

|  |  |
| --- | --- |
|  |  |

Ilustración 3 Carga de la información en databricks

Cabe resaltar que estos archivos quedaron almacenados en la ruta “/FileStore/Equipo\_7” para efectos de la actividad demostrativa y ya en la ejecución en “/FileStore/tables”.

Después se creó un cluster y donde, se generó el Notebook “actividad\_3” el cual se utilizo para poder plasmar nuestros pasos y mostrar la información solicitada en cada uno de los incisos de la actividad.

Proceso en el Notebook

Estando en el notebook primeramente se decidió crear una función que pudiera realizar la lectura de los archivos csv y después crear las tablas correspondientes con el esquema adecuado, estos pasos se muestran en la Ilustración 4

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 4 Proceso de lectura de csv y creación de tablas de cada archivo

Mostrar las cinco primeras filas de cada tabla cargada. Para este caso se realizo la selección de 5 registros por tabla con el query

***“Select \* from features limit 5***

***Select \* from sales limit 5***

***Select \* from stores limit 5”***

El detalle de estas tablas se muestra en las siguientes imágenes (Ilustración 5 y Ilustración 6)

Tabla

Descripción generada automáticamente

Ilustración 5 Las cinco primeras filas de la tabla features

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ilustración 6 Las cinco primeras filas de la tabla sales y stores

Contar el número de filas de cada tabla. Indicar este número en el informe. Para este caso se realizó el conteo por tabla con el query

***“SELECT COUNT(\*) as feature FROM feature***

***SELECT COUNT(\*) as sales FROM sales***

***SELECT COUNT(\*) as stores FROM stores”***

El detalle de estas consultas se muestra en la imagen (Ilustración 7) obteniendo como resultado que la tabla features contiene 8,190 filas, la tabla sales tiene 421,570 filas y la tabla stores cuenta con 45 filas.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 7 Contabilización de filas por tabla

Estudiar el rango (máximo y mínimo) de cada variable numérica. Indicar estos datos en el informe. Para este caso se utilizó nuevamente una query de SQL para obtener el máximo y mínimo de cada variable numérica sin embargo algunas variables contenían datos como NA y se modifico con el fin de que la variable pudiera medirse correctamente por lo cual la query de sql es la siguiente.

***SELECT MIN(Store) as Min\_Store, MAX(case when Store='NA' then null else cast( Store as double) end ) as Max\_Store FROM feature***

Para este caso solo se mostrará una consulta ya que existen en total 10 variables numéricas entre las 3 tablas. Adicionalmente y con el propósito de realizar la consulta de manera más automática se creo una función para detectar de manera automática cuales eran las variables numéricas y de esa misma manera poder realizar una presentación de todas las variables. Los pasos se pueden ver en las siguientes imágenes.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 8 Función para detectar las variables numéricas.

Aplicación

Descripción generada automáticamente con confianza baja

Ilustración 9 Presentación de las 3 primeras variables numéricas con sus mínimos y máximos

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración 10 Presentación de la cuarta, quinta y sexta variables numéricas con sus mínimos y máximos

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración 11 Presentación de las ultimas 3 variables numéricas con sus mínimos y máximos

Estudiar las diferentes categorías de las principales variables categóricas y el número de filas correspondientes a cada categoría. En este punto se utilizaron las variables categóricas las cuales de acuerdo al análisis realizado fueron, isHoliday para features y sales; para stores fue la variable Type. La consulta de esta informaciòn se realizo con la query:

***“SELECT IsHoliday, COUNT(\*) as CountValues FROM feature GROUP BY IsHoliday”***

***“SELECT IsHoliday, COUNT(\*) as CountValues FROM sales GROUP BY IsHoliday”***

***“SELECT Type, COUNT(\*) as CountValues FROM stores GROUP BY Type”***

Para este caso la información se muestra en la Ilustración 12 e Ilustración 13.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 12 Consulta de variable categórica de la tabla features

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 13 Consulta de variable categorica de la tabla stores y sales

Buscar valores inexistentes o anómalos. Primeramente se analizaron todas las varibles con el fin de observar si tenían algun valor nulo o como NA. Y se contabilizaron los datos inexistentes por variable. A continuación se muestra un ejemplo de la consulta de SQL que se utilizo para cada variable:

***“SELECT COUNT(Store) as Filas\_Totales, SUM(CASE WHEN Store IS NULL then 1 else 0 end) as NullCount, SUM(CASE WHEN Store = 'NA' then 1 else 0 end) as NACount FROM feature”***

En las siguientes imágenes se muestran algunas de las consultas sobre los datos inexistentes.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente con confianza media

Como se puede ver en las imágenes hay algunas variables que cuentan con estos datos nulos. Las variables correspondientes son:

Tabla features:

* Markdown1 con 4,158 datos como NA
* Markdown2 con 5,269 datos como NA
* Markdown3 con 4,577 datos como NA
* Markdown4 con 4,726 datos como NA
* Markdown5 con 4,140 datos como NA
* CPI con 585 datos de NA
* Unemployment con 585 datos de NA

Tabla sales y stores no continen datos inexistentes.

Por otro lado, se hizo un análisis de las variables encontrando detalles en las de Temperature, Weekly\_Sales.

Para la temperatura se realiza el supuesto que muestra los grados Fahrenheit por lo cual se consideraron como anómalos aquellos que fueran menor a 10 grados. La consulta de SQL es la siguiente:

***“ SELECT COUNT(\*) as Cantidad\_Datos\_Anomalos FROM feature WHERE Temperature < 10”***

Como se muestra en la siguiente imagenes

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 14 Anomalía de la variable Temperature

Para la variable Weekly Sales se detectaron como datos anómalos aquellos que fueran menor a 0. Dado a que las ventas no pueden ser negativas. La consulta de SQL es la siguiente:

***“SELECT COUNT(\*) as Cantidad\_Datos\_Anomalos FROM sales WHERE Weekly\_Sales < 0”***

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Ilustración 15 Anomalía de la variable Weekly Sales

Realizar alguna operación join entre las tablas creadas. En este caso se decidió unir las 3 tablas features y sales con las variables store y Date, y estas se unieron con la tabla Stores. La consulta de SQL utilizada fue.

***“SELECT***

***stores.Type,***

***feature.IsHoliday,***

***sales.Date,***

***sales.Dept,***

***sales.Weekly\_Sales***

***FROM stores***

***INNER JOIN feature***

***ON stores.Store = feature.Store***

***INNER JOIN sales***

***ON feature.Store = sales.Store***

***and feature.Date =sales.Date***

***WHERE Weekly\_Sales > 0***”

En la siguiente imagen se muestra la consulta y la presentación de los datos resultantes

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración 16 Resultado de utilizar un join con las tablas ejemplo

Ejercicio

A la compañía en cuestión le gustaría, tras un primer análisis exploratorio sencillo, deducir alguna información interesante y que les pueda dar pistas sobre qué tal funcionan las ventas en cada tienda y en los departamentos de estas, cómo evolucionan las ventas a lo largo del año.

Algunas de las preguntas de negocio que estaremos contestando a continuación son las siguientes:

**Preguntas de negocio**

¿Cómo son las tendencias de ventas de los últimos 3 años?

¿Qué tipo de tienda registra las mayores ventas?

¿Qué periodos del año se tiene mas ventas?

Estas preguntas que típicamente un negocio pueda tener, serán respondidas con base a la información generada con la extracción ya análisis de los datos de nuestro datawarehouse o lago de datos que es como podría llegar a definirse a un cluster basado en hdfs.

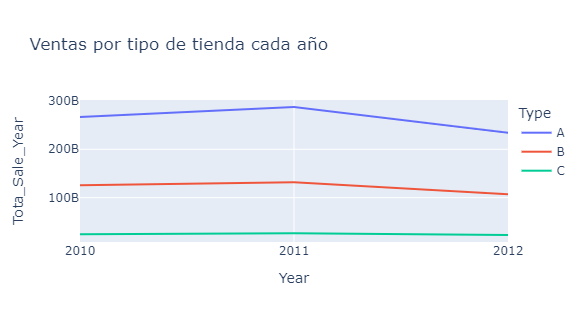


Ilustración 12 Ventas por tipo de tienda cada año

Podemos determinar que la tienda tipo A durante los últimos 3 años ha sido una de las que se ha mantenido en primer lugar de ventas.

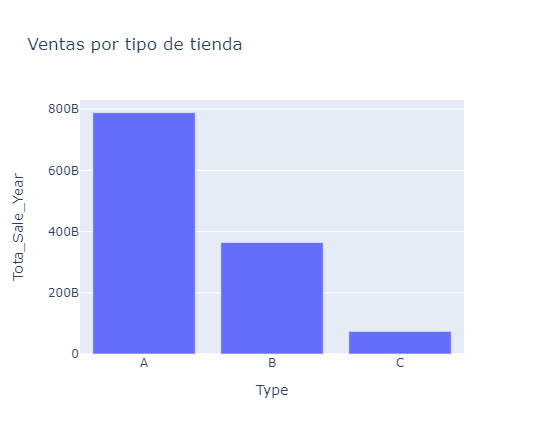


Ilustración 13

Podemos ver que los ingresos de la tienda B representa el 50% de ventas de la tienda A, y la tienda C tiene un nivel de ingresos muy bajo.

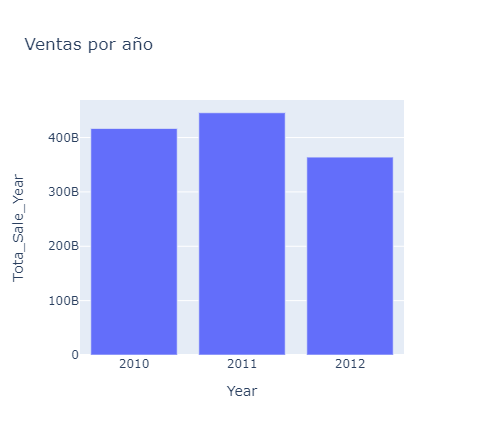


Ilustración 14

Aquí podemos ver que el año 2011 es el que mas ventas tuvo en comparación con el año anterior y el posterior.

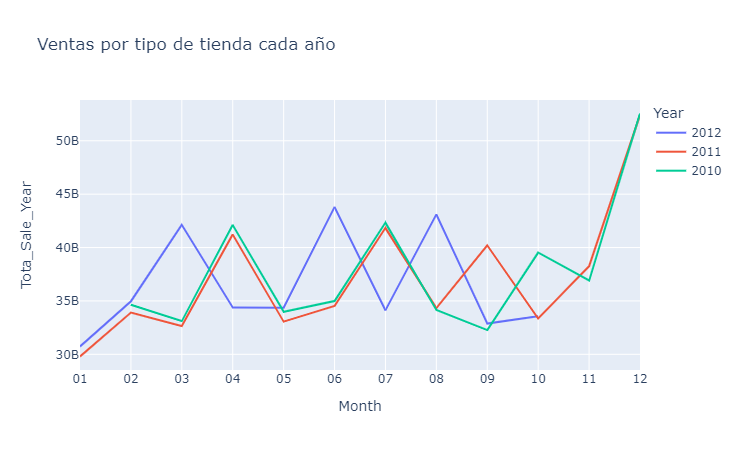


Ilustración 15

Podemos ver la tendencia de las ventas en relación a los meses del año, donde siempre en las épocas decembrinas

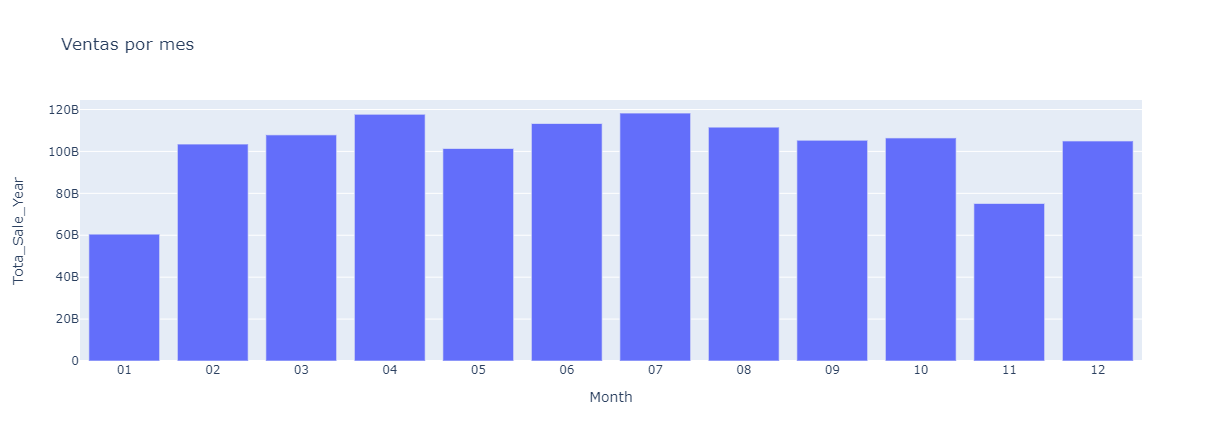


Ilustración 16

Con esta vista podemos ver los ingresos mensuales por tienda

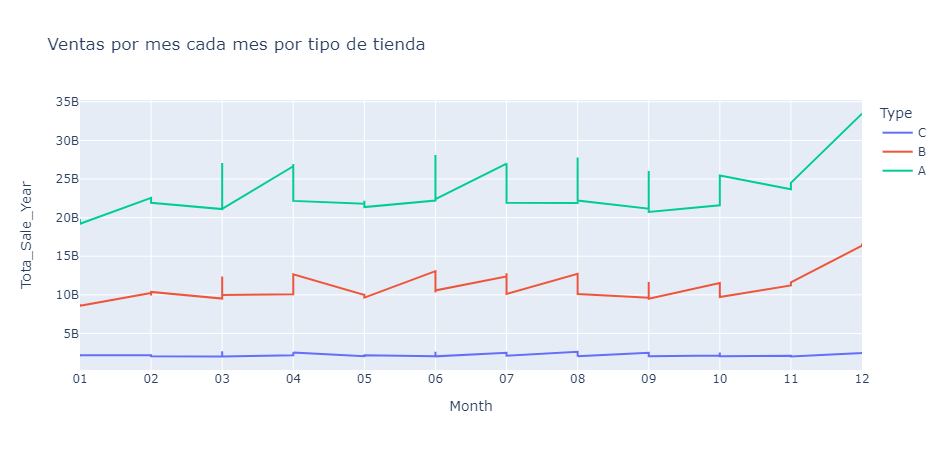


Ilustración 17

Con esta vista hacemos la comparativa de ventas por tipo de tienda durante el año con la referencia de mes a mes.

Conclusiones

Como hemos visto, la exploración de los datos que son cargados a un cluster diseñado en HDFS y en el caso de Data bricks que es DBFS, usando consultas simples de SQL, permite a los científicos de datos, obtener la información necesaria sin uso de complejas líneas de código, se puede visualizar en segunda información de diferentes conjuntos de datos y crear relaciones entre ellos y generar informes y análisis de los mismos.

Con el tipo de consultas generadas, pudimos conocer el numero de registros, sus valores mínimos y máximos, así como el tipo de información que tiene cada conjunto de datos, lo que nos permitió pode generar de forma visual información que permita ya aun usuario de negocio a evaluar la información y poder tomar decisiones.

Hoja de control

La siguiente tabla presenta en lo que cada uno de los miembros del equipo ayudaron.

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Documentación** | **Desarrollo** | **Investigación** |
| Carmen Sayuri Maldonado Pinto | Si | Si | No |
| Eduardo García García | Sí | Sí | Sí |
| Juan Carlos de Jesús Garcés Carrillo | No | No | No |
| Kevin David Farinango Cinilin | Sí | Sí | Sí |

Se recomienda únicamente colocar la calificación a los estudiantes:

Eduardo García García, Kevin David Farinango Cinilin y Carmen Sayuri Maldonado Pinto quienes fueron los únicos que aportaron con la creación del documento y el proyecto.