Contenido

Introducción 1
Planteamiento de negocio 2
1.Realización Práctica – Ejercicio práctico 2
1.1 Ejercicio práctico #1 2
1.2 Ejercicio práctico #2 ..... **¡Error! Marcador no definido.**
1.3 Ejercicio práctico #3 2
1.4 Ayuda ejercicios prácticos 3
2.Dimensionamiento clúster Hadoop 3
3.Estimación arquitecturas Hadoop para distintos casos de uso 4
Datos de entrega de la práctica 4

Introducción

En la actualidad, la extracción de información y el análisis de datos son fundamentales para la toma de decisiones en los negocios. Sin embargo, la cantidad de datos generados cada día es enorme, lo que hace que el procesamiento y análisis de estos datos sea una tarea difícil y costosa. Hadoop es una tecnología que ha revolucionado la forma en que las empresas extraen información y analizan datos.

Hadoop es un framework de procesamiento distribuido de código abierto que permite el almacenamiento y procesamiento de grandes cantidades de datos en clusters de computadoras. Con Hadoop, es posible procesar grandes conjuntos de datos en paralelo, lo que reduce el tiempo de procesamiento y el costo. Además, Hadoop utiliza un sistema de archivos distribuidos que permite el almacenamiento de datos de manera eficiente.

Uno de los casos de uso más populares de Hadoop es el procesamiento de datos de usuario para implementar sistemas de recomendación en empresas como Netflix y Spotify. Con Hadoop, es posible procesar de manera eficiente los gustos de los usuarios, lo que permite a estas empresas proporcionar recomendaciones precisas y personalizadas a sus usuarios…

Planteamiento de negocio

Como titulado de Datahack, has sido recientemente contratado en MovieBuster, una startup de reciente fundación que pretende revolucionar el mundo del ocio digital, empezando por la manera en la que se oferta ocio audiovisual a los clientes.

El CEO de MovieBuster te ha encomendado el nada fácil desafío de crear el departamento de analítica desde cero, este departamento tendrá como función principal la extracción y análisis de los datos necesarios para generar insights que permitan a la mesa ejecutiva la definición de la estrategia de MovieBuster.

Como primera tarea, y de cara a decidir el Roadmap de proyectos que serán realizados durante los próximos meses, el CEO te encarga un análisis pormenorizado del mercado actual de películas, para, con los datos en la mano, poder decidir cuál va a ser la estrategia a seguir.

1.Realización Práctica – Ejercicio práctico -1 punto

Debido a que la startup aún no ha podido desplegar el cluster, tu objetivo es, a través de los datos contenidos de películas contenidos en un dataset público (https://github.com/dgarciaesc/sample_dataset)

1.1 Ejercicio práctico #1

De cara a definir por qué genero apostar, identificar los influencers que pueden potenciar el marketing de MovieBuster y definir una estrategia de publicidad, el CEO te pide averiguar los siguientes datos del momento de mercado actúal:

1. Cuál es la película con más opiniones?
2. Qué 10 usuarios son los más activos a la hora de puntuar películas?
3. Cuáles son las tres mejores películas según los scores? Y las tres peores?
4. Hay alguna profesión en la que deberíamos enfocar nuestros esfuerzos en publicidad? Por qué?
5. Se te ocurre algún otro insight valioso que pudiéramos extraer de los datos procesados? Cómo?

1.2 Ejercicio práctico #2

El CEO está preocupado con la eficiencia de las queries usadas para extraer los datos de los ejercicios prácticos #1 y #2 y exige poder ver estos resultados desde una web.

Implementa, a través de Sqoop, una BBDD relacional en MySQL que contenga al menos los datos de uno de los insights extraídos en el ejercicio práctico #1

1.3 Ayuda ejercicios prácticos

- El dataset disponible para la realización de la práctica se encuentra disponible en: https://github.com/dgarciaesc/sample_dataset

- Es necesario entender los campos del dataset para poder proceder a la realización de la práctica. Estos vienen convenientemente explicados en el Readme

- Es posible que los separarores del dataset no sean "comas" (","). Es responsabilidad del alumno solucionar este problema, bien mediante un pretratamiento de los ficheros de entrada o mediante la definición del delimitador en el comando que proceda de PIG o Hive

- Cualquier duda adicional puede ser planteada por correo a dgarciaesc@gmail.com

2.Dimensionamiento clúster Hadoop -1 punto

El CEO está muy contento con el trabajo realizado y quiere apostar aún más por tecnologías Big Data. Está pensando en montar otra infraestructura Hadoop que procese eventos de películas provenientes de distintas fuentes (cines, plataformas de streaming, etc..) y necesita estimación del tamaño plataforma teniendo en cuenta que:

Volumen de datos:

	Media Eventos	Tamaño por evento
Fuente 1	10.000 eventos/día	15 KB
Fuente 2	120.000 eventos/día	300 Bytes
	150.000 eventos/día	100 KB
	170.000 eventos/día	800 KB
	2000 eventos/día	1500 KB

Las características de las maquinas es que son capaces de tener hasta 22 discos de 2 Teras para almacenamiento cada una. Indicar el número de máquinas necesarias para poder almacenar todo el volumen de datos durante el próximo año, así como la justificación de porque se necesita dicha capacidad para un clúster Hadoop.

3.Estimación arquitecturas Hadoop para distintos casos de uso-1 punto

En la empresa están también pensando en conectar su plataforma Big Data con otras herramientas de la empresa y nos piden consejo sobre cómo podría integrarse/ejecutarse:

-Herramienta de BI (p.ej.: Microstrategy)
-Web de consultas sobre pedidos realizados
-Generación de informes SQL usando R que se ejecutan mensualmente
-Recopilación de información de redes sociales

Para cada una de estas tareas indica que posibles herramientas del ecosistema Hadoop aplicarían por requisitos de casuística teniendo en cuenta las ventajas e inconvenientes de cada una de ellas (por ejemplo, uso de Impala consume mucha RAM).

Datos de entrega de la práctica

Para la evaluación de la práctica el alumno deberá entregar:

-Memoria explicativa (formato a elegir: Word, PPT o web) donde se detallen las soluciones de las tres partes en las que se compone la práctica.

Los alumnos deberán subir en la carpeta práctica del aula virtual las memorias explicativas junto con los scripts utilizados en la parte 1 de la práctica

La evaluación de la práctica ser hará sobre un máximo de 3 puntos totales (1 por problema). Siendo 1.5 la puntuación necesaria para el apto.

Fecha límite: Mayo de 2024.