Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Diplomado Big Data

Ecosistema Hadoop: Fundamentos

Laboratorio 1: Análisis de texto con Hadoop

En este laboratorio utilizarán múltiples herramientas del ecosistema Hadoop, para analizar datos provenientes de la base de datos de la empresa DualCore, que incluye productos (`products`), clientes (`customers`) y evaluaciones (`ratings`). En base a este análisis, podrá detectar problemas en la base de datos y proponer potenciales soluciones.

1 Preparación de los datos

Como paso previo al trabajo de laboratorio, deberá preparar los datos a utilizar y cargarlos en sus respectivas cuentas del cluster y/o máquina virtual. Los datos se encuentran en archivos de texto disponibles en el sitio del curso. Los pasos a realizar son los siguientes:

1. Utilizando la interfaz de Hue, cree en el MetaStore de Hive una base de datos llamada *grupoN*, donde N es su número de grupo. **NO** almacene la base de datos en la ubicación por defecto, créela en la carpeta `/user/grupoN/warehouse/dualcore`.

2. Utilizando la interfaz de Hue, cree en el MetaStore de Hive tablas llamadas `products`, `customers` y `ratings`, que cumplan con los siguientes esquemas:

name	type
prod_id	int
brand	string
name	string
price	int
cost	int
shipping_wt	smallint

(a) `products`

name	type
cust_id	int
fname	string
lname	string
address	string
city	string
state	string
zipcode	string

(b) `customers`

name	type
posted	timestamp
cust_id	int
prod_id	int
rating	tinyint
message	string

(c) `ratings`

   Al igual que en el caso anterior, **NO** use la ubicación por defecto, utilice la carpeta `/user/grupoN/warehouse/dualcore/table_name`, donde `table_name` indica el nombre de la tabla. Revise los archivos para utilizar el carácter de finalización de campos correcto.

3. Utilizando la interfaz de Hue, llene las tablas con los datos contenidos en los archivos. La manera más sencilla de hacerlo es copiar en las carpetas correspondientes, los archivos con los datos de las respectivas tablas.

2 Análisis de las evaluaciones de los productos

Como primera tarea, deberá encontrar aquellos productos que obtienen las mejores y peores calificaciones, por parte de los clientes de DualCore. Para esto, escriba un script en *Pig Latin*, que a partir de los datos almacenados en el *MetaStore*, retorne el nombre y promedio de puntaje de los 10 productos mejor evaluados. A continuación, modifique el script para que también retorne la cantidad de evaluaciones que estos productos tienen. Analice y comente sobre el efecto que la cantidad de evaluaciones juega sobre el ranking de los productos mejor evaluación, y en base a esto, defina un criterio para aceptar un puntaje promedio como válido. Es importante notar que debe existir una justificación razonable para la selección del criterio. Finalmente, repita el proceso anterior, incluyendo la selección del criterio, para encontrar los productos peor evaluados. Es perfectamente posible que el criterio seleccionado no sea el mismo que para lo productos mejor evaluados.

Actividades interesantes para el informe: Repita el proceso anterior usando ahora *Impala*. Comente sobre las diferencias en los tiempos de ejecución. **Nota**: Si no aparecen las tabla, ejecutar la consulta `INVALIDATE METADATA` en *Impala*.

3 Análisis de los comentarios

A continuación, deberá analizar los comentarios realizados sobre el producto peor evaluado en base a **n-gramas**. Formalmente, un n-grama representa una subsecuencia de n elementos de una secuencia dada. En este caso, la secuencia está dada por un comentario y los elementos son sus palabras. El análisis de n-gramas permite encontrar expresiones populares dentro de grandes volúmenes de texto.

Desde el punto de vista práctico, para este análisis utilizará las funciones de análisis de texto de Hive, en particular, las funciones `SENTENCES`, `NGRAMS` y `EXPLODE`. Por ejemplo, para obtener los **k** n-gramas con más apariciones, basta con ejecutar la consulta `SELECT EXPLODE(NGRAMS(SENTENCES(LOWER(message)), n, k)) FROM ratings`, que generará los k n-gramas más populares encontrados en los comentarios de todos los productos.

En base a esto, obtenga y analice, al menos, los bigramas y trigramas más populares del peor producto, y comente su posible significado en términos de la mala evaluación por parte de lo usuarios.

4 Análisis de los comentarios parte 2

Finalmente, en este última parte del laboratorio, utilizando la información de los n-gramas más populares y significativos encontrados en la sección anterior, procese los comentarios usando Pig, Impala o Hive, con el fin de identificar aquellos que entreguen fuerte evidencia de las interpretaciones dadas en la sección anterior. En base a estos comentarios, identifique la causa de los malos comentarios e indique como puede ser esto corregido en la base de datos.

Actividades interesantes para el informe: Si utilizó Pig en el análisis anterior, repita el proceso Hive o Impala, y si usó Hive o Impala, repítalo usando Pig. Comente sobre las diferencias en facilidad de uso y tiempo de ejecución

Bonus

Automatice todo el proceso realizado, utilizando *Oozie*.