Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Diplomado Big Data

Ecosistema Hadoop: Fundamentos

Laboratorio 1: Análisis de datos usando el Ecosistema Hadoop

Parte 1: Análisis de texto con Hive e Impala

En esta parte del laboratorio, utilizarán herramientas exploratorias del ecosistema Hadoop, para analizar 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.

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.

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 *Hive*, 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 tablas, ejecutar la consulta `INVALIDATE METADATA` en *Impala*.

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.

Finalmente, utilizando la información de los n-gramas más populares y significativos recién encontrados, procese los comentarios usando 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 malos comentarios, identifique su posible causa e indique como puede ser esto corregido en la base de datos.

Parte 2: Procesamiento, corrección y generación de datos con Pig

La empresa Dualcore recientemente comenzó a utilizar anuncios en línea para atraer nuevos clientes a su sitio de comercio electrónico. Cada una de las dos redes publicitarias que utilizan proporciona datos sobre los anuncios que tienen publicados. Estos incluyen el sitio donde se colocó el anuncio, la fecha en que se colocó, las palabras clave que activaron su visualización, si el usuario hizo click en el anuncio, y el costo por clic.

Limpieza de datos

Los datos de ambas redes están en un formato diferente y también contienen algunos registros inválidos. Antes de que podamos analizar los datos, primero debemos corregir los problemas usando Pig. Específicamente, considere que los datos de las redes presentan el siguiente esquema:

Index	Field	Data Type	Description	Example
0	keyword	chararray	Keyword that triggered ad	`tablet`
1	campaign_id	chararray	Uniquely identifies the ad	`A3`
2	date	chararray	Date of ad display	`05/29/2013`
3	time	chararray	Time of ad display	`15:49:21`
4	display_site	chararray	Domain where ad shown	`www.example.com`
5	was_clicked	int	Whether ad was clicked	`1`
6	cpc	int	Cost per click, in cents	`106`
7	country	chararray	Name of country in which ad ran	`USA`
8	placement	chararray	Where on page was ad displayed	`TOP`

(a) Esquema de datos de la primera red (`ad_data1.txt`).

Index	Field	Data Type	Description	Example
0	campaign_id	chararray	Uniquely identifies the ad	`A3`
1	date	chararray	Date of ad display	`05/29/2013`
2	time	chararray	Time of ad display	`15:49:21`
3	display_site	chararray	Domain where ad shown	`www.example.com`
4	placement	chararray	Where on page was ad displayed	`TOP`
5	was_clicked	int	Whether ad was clicked	`Y`
6	cpc	int	Cost per click, in cents	`106`
7	keyword	chararray	Keyword that triggered ad	`tablet`

(b) Esquema de datos de la segunda red (`ad_data2.txt`).

Para la primera red, el proceso de limpieza considera eliminar cualquier registro que no contenga en el campo `country` la palabra `USA`. Luego, para cada registro, se debe transformar el contenido del campo `keyword`, cambiando todo a mayúsculas y eliminando los espacios en blanco previos y posteriores al contenido. Finalmente, se deben reordenar los campos de cada registro de acuerdo a lo indicado en la siguiente figura y posteriormente escribir el resultado en disco, utilizando tabulación como delimitador.

Para la segunda red, inicialmente se deben eliminar los registros duplicados. Luego, para cada registro, se debe transformar el contenido del campo `date`, cambiando los "-" por "/", y limpiar transformar el campo `keyword` como en la primera red. Finalmente, se deben reordenar los campos de cada registro de la misma manera que para la red anterior y posteriormente escribir el resultado en disco, utilizando tabulación como delimitador.

Index	Field	Description
0	campaign_id	Uniquely identifies the ad
1	date	Date of ad display
2	time	Time of ad display
3	keyword	Keyword that triggered ad
4	display_site	Domain where ad shown
5	placement	Where on page was ad displayed
6	was_clicked	Whether ad was clicked
7	cpc	Cost per click, in cents

Nuevo esquema de datos para ambas redes.

Análisis de las redes

Ambas redes publicitarias cobran una tarifa sólo cuando un usuario hace clic en el anuncio de Dualcore. Esto es ideal para Dualcore ya que su objetivo es atraer nuevos clientes a su sitio. Sin embargo, algunos sitios y palabras clave son más efectivas que otros para atraer a las personas, lo que resulta fundamental de considerar para cuando DualCore anuncie sus nuevos productos. Teniendo esto en consideración, complete los siguiente ejercicios.

1. Identifique qué sitios tienen el costo total más bajo. Utilizando los registros de ambas redes, elimine aquellos que **no** hayan sido clickeados y agrúpelos de acuerdo al campo `display site`. A continuación, calcule para cada sitio su costo total, ordénelos de manera ascendente en función del costo, e imprime en pantalla los tres primeros. Todo este proceso debe realizarse de manera simultánea para ambas redes, no de manera independiente.

2. Los términos que los usuarios escriben al realizar búsquedas pueden generar la aparición de anuncios de Dualcore. Dado que los anunciantes en línea compiten por el mismo conjunto de palabras clave, algunas cuestan más que otras. Utilizando Pig, y basándose en el ítem anterior, identifique las cinco palabras clave que han sido las más caras para Dualcore.

3. Calcule e imprima en pantalla la cantidad total de clicks que los anuncios han recibido. Luego, para cada anuncio, calcule la cantidad de clicks que ha recibido y la posición dentro del sitio que ha resultado más efectiva. Imprima en pantalla los resultados para los tres avisos más populares.

4. Calcule el costo total de una campaña publicitaria que genere 50.000 clicks. Para esto, realice dos estimaciones distintas, la primera utilizando el escenario más pesimista de costo por anuncio, y la segunda utilizando el costo promedio.

5. Para calcular la efectividad real del avisaje realizado, calcule para cada sitio el *click-trough rate*, que consiste en el porcentaje de avisos mostrados que fueron efectivamente clickeados por los usuarios. Imprima en pantalla los tres sitios más efectivos, y los tres menos efectivos. Compare estos resultados con los del primer item y comente.