datahack

Práctica ecosistema Hadoop

Tenemos una página web: http://datahack.mooo.com

En esta página se realizan los pagos por compras de productos, el director de la empresa quiere saber la efectividad de la forma de pago que actualmente está habilitada, los métodos de pago son:

- PagoCR: Pago contra reembolso
- Pago: Pago con tarjeta de débito

El director nos pide que generemos informes como tablas en HIVE con el resumen de cuantas de las visitas que llegan a nuestra web terminan realizando una compra (y como pagan) y cuantas no realizan el pago para evaluar si incluir otros métodos como PayPal que faciliten al usuario esta operación.

Para realizar este estudio tenemos nuestra plataforma Hadoop y acceso a 2 fuentes de datos:

Logs Apache de la web

Ubicados en la máquina brigde: **caronte** en **/var/log/httpd/access_log** y formato:

```
80.38.199.34 - - [24/Jun/2016:09:19:26 +0000] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36"
```

- Las entradas con método “GET” indican que se ha realizado una visita en un momento indicado en el campo fecha
- Las entradas con método “POST” indican que se ha realizado una comprar en el momento indicado en campo fecha

Nota: Ten en cuenta que la máquina brigde está fuera del clúster de Hadoop y tendrás que ver la forma de llevarte el fichero de log, del servidor web al clúster de Hadoop. Para este efecto el **comando scp**, puede serte de ayuda. Aquí te dejamos más información del mismo: https://www.garron.me/es/articulos/scp.html

BBDD con los registros de compras

Los registros de compras se almacenan en mysql ubicado en la máquina **cdm1**, dentro de la base de datos **web**. Para acceder a la misma, usar el usuario **datahack** y contraseña **datahack2017**:

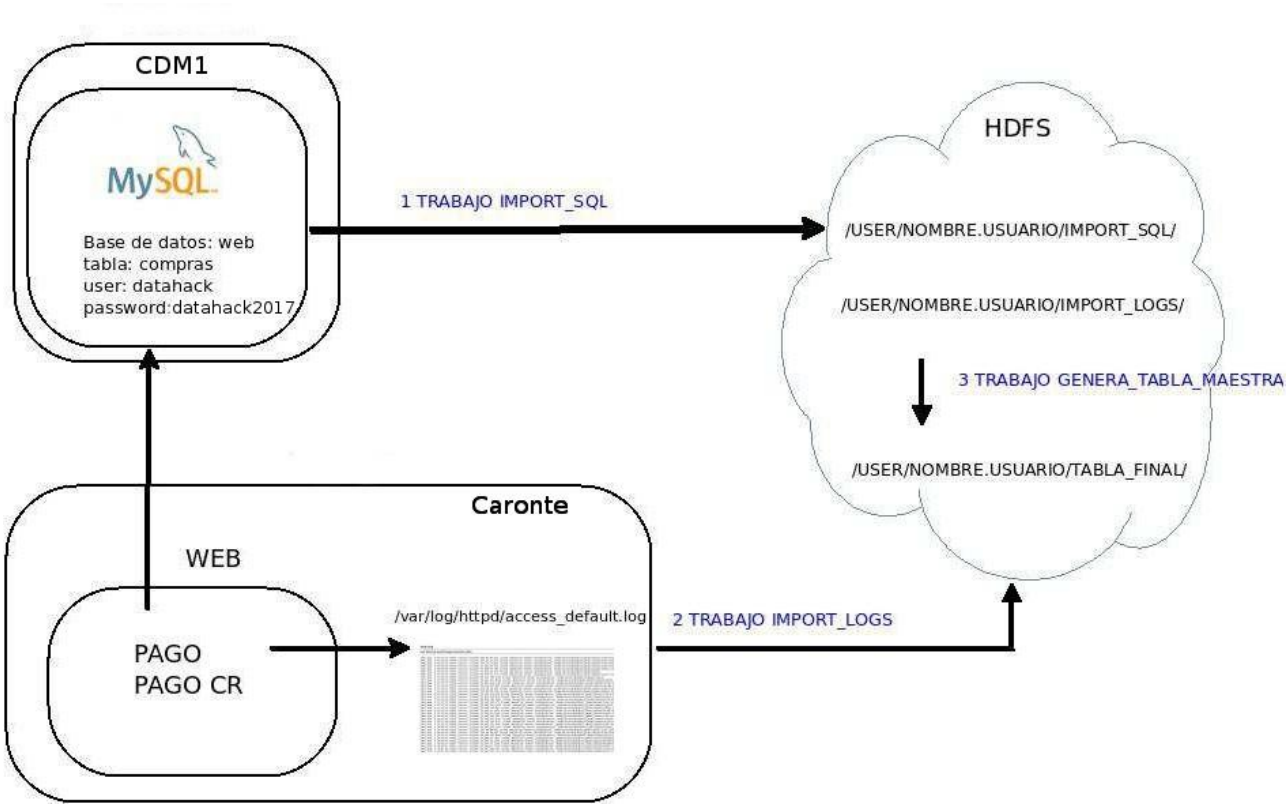
```
mysql> use web;
Database changed
mysql> show tables;
+---------------+
| Tables_in_web |
+---------------+
| compras       |
+---------------+
1 row in set (0,00 sec)

mysql> select * from compras limit 1;
+---------------------+-----------+
| hora                | tipo_pago |
+---------------------+-----------+
| 2016-06-24 09:15:16 | tarjeta   |
+---------------------+-----------+
1 row in set (0,00 sec)
```

En la tabla compras se almacena la fecha con hora y el tipo de pago realizado “tarjeta” o “reembolso”.

Apartado 1:

Obtener la tabla resultante del estudio en la que se muestre el número de compras (por tipo) y no compras por hora, del estilo:

Hora	PagoCR	Pago	NoCompra
10:00	12	11	2
11:00	30	15	9



Para realizar esta tarea vamos a involucrar diferentes componentes del ecosistema y los organizaremos en 3 tareas diferentes:

1- Trabajo import_sql:

Importaremos la base de datos con sqoop y dejaremos su contenido en HDFS, cada alumno deberá dejarlo bajo su home de HDFS como muestra el path: **/user/nombre.usuario/import_sql**.

Crearemos una tabla externa en HIVE con el nombre **bbdd** que apunte a este path para acceder posteriormente a ella mediante SQL.

2- Trabajo import_logs:

Montaremos flume para que mande todos los logs que se vayan generando a un directorio de HDFS, cada alumno deberá dejarlo bajo su home de HDFS como muestra el path: **/user/nombre.usuario/import_logs**.

Mediante PIG procesaremos los logs importados de HDFS para generar un dato estructurado con los datos que nos interesan de estos logs, en concreto queremos quedarnos con la IP, la fecha y si se trata de un campo GET o POST, por ejemplo, para la entrada:

80.38.199.34 - - [24/Jun/2016:09:19:26 +0000] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36"

Queremos generar una tabla:

IP	Fecha	METODO
80.38.199.34	24/Jun/2016:09:19:26 +0000	GET

Esta tabla con el resultado se almacenará en HIVE con el nombre de tabla “logs”, bien usando una tabla gestionada bajo **/user/hive/warehouse/nombreusuario.db/**, o externa en un path de HDFS dentro del home del usuario en **/user/nombre.usuario/tablas_hive/logs**

3- Trabajo generar la tabla maestra:

A partir de las dos tablas anteriores generadas en hive **logs** y **bbdd y** mediante transformaciones en sql genera la tabla maestra final en la que se muestre el numero de acciones realizadas por el usuario. Esta tabla debe contener las siguientes columnas:

1. Hora: Hora en la que se han contabilizado las acciones. Ejemplo: Hora=10 corresponde a las acciones realizadas entre las 10.00 y las 10.59
2. PagoCR: Número de peticiones de tipo pago contra reembolso de dicha hora
3. Pago: Número de peticiones de tipo pago tarjeta de crédito de dicha hora
4. No Pago: Número de usuarios que no han realizado ninguna acción en dicha hora

Apartado 2:
El director está muy contento con el trabajo realizado y quiere apostar aún más por tecnologías Big Data y nos pide consejo para las siguientes consultas:

Apartado 2.1:
Está pensando montar otra infraestructura Hadoop en Asia que contenga los datos de la empresa referentes a su negocio allí 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.

Apartado 2.2:

En la empresa están también pensando en conectar su plataforma Big Data con otras herramientas de la empresa y nos piden consejo sobre como 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 la entrega de la practica:

Para la evaluación de la práctica el alumno deberá entregar:

-Memoria explicativa del apartado 1 de las operaciones realizadas, junto con los scripts, así como la configuración o acciones realizadas para poder ejecutar dichos scripts comentados (por ejemplo, la configuración de FLUME)

Para este apartado los profesores consultarán si existen las tablas finales dentro de los HOMEs de HDFS de los usuarios con los nombres de tablas indicados, así como cualquier otra tabla auxiliar que se mencione en la memoria para el correcto funcionamiento de la práctica.

-Memoria explicativa del apartado 2 y justificación de las decisiones a tomar (no mas de 3 páginas)

Los alumnos enviarán .zip con dicho contenido a la dirección de email de la práctica, aquellos alumnos que no envíen nada no podrán ser evaluados.

Fecha límite: 10/03/2017

academica@datahack.es