PEC1 – Infr. Big Data, tratamiento batch

Entrega

- El práctica se entregará en formato pdf, el nombre del fichero debe tener el siguiente formato <apellidos>_PEC1.pdf
- Las imágenes deben ser nítidas de tamaño suficiente para facilitar su lectura
- Los comandos o programas utilizados se incluirán dentro del pdf en formato texto, por delante de la imagen donde se ve la ejecución (si procede).

Antecedentes

Durante las clases hemos visto varias herramientas y arquitecturas utilizadas para resolver la necesidad del tratamiento de datos masivos.

En este ejercicio vamos a poner en práctica esos conocimientos en un entorno controlado y con una cantidad de datos mucho menor a la que nos encontramos en la realidad, pero que servirá de muestra y, sobre todo, permitirá que ubiquéis los conocimientos en el marco de una aplicación Big Data.

Requerimientos de la aplicación

Tenemos la necesidad de realizar el tratamiento de los datos de calidad del aire que generan las estaciones de medida de la Comunidad de Madrid.

Debemos descargar los datos de 2020 de su página web, en este caso nos interesan los datos por hora recogidos durante 2020, en la propia página encontraréis información sobre su contenido.

Se pide:

1. Almacenamiento en HDFS (20% de la nota)

Los ficheros deben estar descargados en la carpeta /home/template/work-bigdata/pec1/data, en adelante la llamaremos “data_local”.

a) ¿Qué formato de fichero es el más adecuado para el tratamiento posterior de los datos mediante Hive?

b) Copiar los ficheros de cada mes de la carpeta data_local y depositarlos en la carpeta /user/template/pec1/data de hdfs, en adelante nos referiremos a ella como “data_hdfs”

c) Visualizar el espacio ocupado por el directorio data_hdfs procurando que el espacio se muestre en unidades de almacenamiento legibles, no en bytes.

d) Mostrar el número de bloques y la cantidad de réplicas de cada fichero. ¿Es un buen caso de almacenamiento de información en hdfs? ¿Explica por qué?

2. Tratamiento con Hive (30% de la nota)

Vamos a utilizar esta herramienta para realizar el tratamiento batch de los datos.

a) Crear una tabla externa (nombre tabla medidasAire)que nos permita mapear los datos de todos los ficheros de la carpeta data_hdfs

b) Realizar una query que nos muestre mensualmente la media de dióxido de azufre medida por estación a las 12h. A modo de ejemplo el resultado debería mostrar:

Año	Mes	Día	SO2 (media de las 12h.)
2020	01	01	7.56
...			

c) Crear una tabla interna particionada por año-mes (nombre tabla medidasAirePart) y realizar una carga (con selección de la partición de forma dinámica), de todos los datos contenidos en la tabla medidasAire

d) Exporta a un fichero CSV el resultado de una query hive que presente la media por año-mes-dia-hora tomada de cada magnitud

Año	Mes	Día	Magnitud	h01	h02	...
2020	01	01	1	7	8	
...						

3. Tratamiento con Mongo DB (30% de la nota)

En los procesos anteriores hemos realizado el tratamiento batch de los datos de calidad del aire, ahora queremos llevar esos datos a Mongo para que se puedan consultar.

a) ¿Cuál crees que sería la estructura de documento adecuada para almacenar los datos del fichero csv del punto 2.a) de Hive?, represéntala mediante un documento de ejemplo

b) Crea un proceso apoyándote en la librería pymongo que recoja los datos del fichero csv del punto 2.a) y los almacene en una colección de mongo.

c) Sobre la colección creada en el punto 3.b), realiza una query presente el siguiente contenido (el formato es sólo ilustrativo):

Año	Mes	Día	Magnitud	Media
2020	01	01	SO2	7
...				

Si no conseguiste realizar el punto 3.b utiliza la utilidad mongoimport indicando el tipo de fichero csv.

4. Arquitectura Big Data (20% de la nota)

En este punto trataremos de definir un esquema de arquitectura que se adecúe a las necesidades cubiertas en los apartados anteriores.

1. ¿Qué tipo de arquitectura big data es mejor utilizar para el conjunto de apartados resueltos anteriormente?. Realiza una **breve** explicación de la elección.
2. Realiza un esquema que represente la arquitectura elegida, ubicando en él las piezas construidas en los apartados anteriores. En el esquema se espera ver la ubicación de los productos: hdfs, hive, mongo, … y también los flujos de información.