Práctica - Map Reduce

Map Reduce Teórico.

Supongamos que se quiere procesar el resultado del **Censo Nacional de Población, Hogares y Viviendas 2010: Censo del Bicentenario,** utilizando Map Reduce. Suponiendo que como entrada (una por habitante del hogar), se recibe como clave un identificador de entrada y como valor la información del censo que contiene:

- **edad:** Cantidad de años cumplidos a la fecha de referencia del Censo
- **alfabetismo**: Sabe leer y escribir
    - 0: Sin datos
    - 1: Si
    - 2: No
- **tipoVivienda:** Tipo de vivienda particular donde habita
    - 0: Sin datos
    - 1: Casa
    - 2: Rancho
    - 3: Casilla
    - 4: Departamento
    - 5: Pieza en inquilinato
    - 6: Pieza en hotel familiar o pensión
    - 7: Local no construido para habitación
    - 8: Vivienda móvil
    - 9: Persona/s viviendo en la calle
- **nombreDepto**: Nombre del departamento donde habita.
- **nombreProv:** Nombre de la provincia donde habita.
- **hogarId:** Identificador del hogar donde habita.

Indicar para cada una de las preguntas:

- La clave y el valor de salida del mapper (indicando si hubo filtrado o transformación).
- La operación a realizar en el proceso de reduce.
- La clave y valor de salida para el proceso de reduce.

Se requiere resolver:

1. Cantidad de habitantes total del país agrupados de acuerdo a su edad en tres grupos :
    a. 0 - 14 años
    b. 15 - 64 años
    c. 65 años y más
2. El promedio de habitantes por vivienda para cada tipo de vivienda.
3. Los n departamentos con mayor índice de analfabetismo, donde el índice se calcula por el número total de habitantes analfabetos del departamento sobre el total de población del departamento, donde n provee el usuario.

4. Los departamentos de la provincia `prov` con una cantidad de habitantes menor a `tope`, donde `prov` y `tope` lo provee el usuario.
5. Los pares de departamentos que tienen la misma cantidad de cientos de habitantes.

Práctica en Computadora

Para realizar esta práctica se utilizará la VM de quickstart provista por cloudera. La versión de la misma se encuentra instalada en las laptops, por lo que solo se necesita iniciar el virtual box y lanzar la VM llamada hdb-2018-02.
La VM tiene un CentOS instalado por cloudera con las herramientas del ecosistema Hadoop (Hadoop, Pig, Hive, HBASE). Inicia con un Firefox abierto que permite acceder a un tutorial armado por cloudera y además tiene bookmarks para el resto de las aplicaciones web provistas por las herramientas.

Datasets & HDFS

Bajar el archivo **datasets.tar.gz** y descomprimirlo en la el directorio /home/cloudera, de manera que quede un directorio **datasets**.
El contenido tiene dos series de datasets:

- datasets-small: Con 1 o varios archivos que representan un volumen pequeño para poder hacer un testeo rápido de querys y aplicaciones
- datasets-med: Con archivos con más volumen para pruebas más “reales”.

Dentro de cada tipo de dataset se encuentran los siguientes directorios:

- **books**: Son libros pasados a txt por voluntarios que se pueden bajar de este link. El formato es simple un encabezado por licencia y el libro en ASCII, sin estilos ni imágenes.
- **temp-hourly**: Son mediciones de los sensores de temperaturas tomadas compilados por el NCEI (Centro de Información de Clima de Estados Unidos). Cada archivo es un csv que contiene las medidas de tiempo sumarizada a la hora de un mes. El formato es CSV con un encabezado y una línea por cada registro.
- **imdb**: Información de películas en formato json obtenidas del sitio imdb a través de su API.
- **censo:** Información del censo obtenido de data.ar, curado para que tenga el formato dado al principio de este TP.

para realizar el resto de los ejercicios suba a hdfs el contenido del directorio de manera que quede en la siguiente uri: **hdfs:/datasets**.

1. Genere la carpeta datasets-small dentro de **hdfs:/datasets** con sus subcarpetas, subiendo cada archivo a la subcarpeta correspondiente.
2. Utilizando el comando **put** suba todo el contenido datasets-med en un solo comando.

Map Reduce en Java.

1. Bajar el archivo **word-count.tar.gz** y descomprimirlo
2. Desde la consola acceder al directorio y correr el comando **mvn eclipse:eclipse**
3. Abrir el eclipse (icono en el escritorio) e importar el proyecto file -> import -> existing projects into workspace y buscar la carpeta que se creó en el paso 1.
4. Con el código importado agregar al mapper y al reducer para realizar el word-count
5. Desde la clase Driver hace botón derecho -> run as -> run configurations y en el tab arguments. En program arguments agregar el path hasta la carpeta del datasets que contiene a los libros: **/home/cloudera/datasets/datasets-small/book** y un path de salida **/home/cloudera/output.** Recordad que cada vez que se corre hay que primero borrar el segundo directorio.
6. Correr con run.
7. En caso de querer correr en modo "cluster".
    a. Hay que ir por consola hasta el directorio descomprimido y correr **mvn install**.
    b. dentro de la carpeta target se genera un archivo **word-count-1.0-SNAPSHOT.jar.** (supongamos en /home/cloudera/word-count/word-count-1.0-SNAPSHOT.jar)
    c. Correr el comando

    hadoop jar /home/cloudera/word-count/word-count-1.0-SNAPSHOT.jar ar.com.met.hadoop.examples.wordcount.Word/countDriver /datasets/datasets-small/books /results/hadoop/wc-small

8. [Opcional] Realizar otro proyecto donde con la información de temperaturas sacar el promedio por día de las muestras tomadas
9. [Opcional] Pasar a código alguna de las consultas realizadas para el censo en este TP.