Assignment 8: Introduction to Hadoop

Tasks:

1. Study Hadoop and MapReduce.
2. Write the record-reader routine, to read a csv file and extract information (as mentioned in the following sections).
3. Write the Mapper, Combiner and Reducer routines (those necessary) to implement the queries mentioned in the following sections.
4. Write the output-format routine to print results of queries to a *.txt* file.

   Write the code for these queries in R.

DataSet:

The dataset is a csv file which contain information about songs. The file contains following information:

- Name of artists who sung the song (artists separated by ;).
- Tags, list of tags for the song. Example: “New York”, “classic”, “rock” etc. The tags are based on both genre and lyrics of the song (tags separated by ;).
- Title, title of the song.
- Song-id, alpha-numeric format.

  To download the dataset, use this link.

Information to extract:

From the file extract names of the artists, title and tags of the song. Group all this information under the *song-id*. All this work is to be done in the Record-reader phase.

Queries:

The queries are to be implemented in Mapper, Combiner and Reducer phases. Some of them may be empty based on the query.

Query 1:

Print titles and ids of the songs which have more than *NUM_TAGS* tags.

Query 2:

Print names of the artists, along with song names and song ids, who have sung of more than *NUM_SONGS* songs. If a song is sung by multiple artists together, consider it separately for each artist.

Query 3:

It is an extension of query 2. In it implement the query 2 with one more condition. The condition is to consider only those songs which have more than NUM_TAGS tags. Thus first filter the songs which have NUM_TAGS tags, then implement the query 2 on this data.

Query 4:

In this query we develop an index for artist names. i.e. given an artist name you shall be able to retrieve names of all the songs sung by that artist. Example: on input *"Arjith Singh"*, print all the songs sung by *Arjith Singh*.

Important, you shall not search the whole dataset to find the songs sung by *Arjith Singh*.

Write code to construct an index for artist names. Corresponding to an artist, store all the songs sung by him/her. Since it is an index on artist names, it shall be sorted on artist names. On a query search in this index and print the results.

Also print the whole index to a separate file once. An index like this is called inverted index.

Input/Output:

Get NUM_TAGS, NUM_SONGS and *Artist name* as input. Construct a menu first to select the query, then to get required inputs.

Print output of each query and the index constructed to *.txt* files with appropriate formatting and names.

Deliverables:

R code for above functionality compressed as .tar.gz, named <YOUR_ROLL_No>.tar.gz.

****

LOGIC:

Hadoop has following phases:

1. record reader. // Reading and parsing the input to records.
2. map. // Execute an operation on each record.
3. combiner. // Do reductions local to a node.
4. partitioner. // Shuffling and sorting. Can't be altered except providing a Comparator.
5. reduce. // Combine the results from the combiner.
6. output format.

The assignment covers all the user editable phases along with three important uses of Hadoop i.e.

- **filtering** (finding artists with at least 3 vowels in their names)
- **numerical summarizations** (Counting number of songs of an artist)
- **indexing**.