Business Intelligence for Big Data

MapReduce assignment

By analyzing the historical usage data of the Barcelona bike sharing system, identify the most critical stations and moments, i.e., when and where the usage demand saturates the offer.

Input data

The data is coming from sensors placed inside each bike-sharing station. Bike-sharing stations consist of groups of parking stalls. Every 10 minutes the sensors detect the presence or absence of parked bikes in the stalls, which can be free (bike absent) or used (bike present).
Each input document from the database has the following data:

station_id: 234 time_hour: 12 time_minute: 34 date: 21-01-2017 day_week: Saturday **free_stalls: 2** **used_stalls: 8**	station_id: 234 time_hour: 23 time_minute: 45 date: 21-01-2017 day_week: Saturday **free_stalls: 1** **used_stalls: 9**	station_id: 654 time_hour: 23 time_minute: 45 date: 21-01-2017 day_week: Saturday **free_stalls: 6** **used_stalls: 5**	station_id: 654 time_hour: 12 time_minute: 59 date: 29-05-2017 day_week: Monday **free_stalls: 7** **used_stalls: 4**

Assignment

Execute the following 3 exercises by writing the key-value parameters of the requested Map Reduce jobs, considering the available attributes in the original documents of the database and applying the required computations.

Exercise 1

Write a MapReduce to identify the most critical places by computing the average usage for each bike sharing station.
The average usage of a station over all the measurements is defined as follows:

total number of used stalls / (total number of free+used stalls)

Map

Key = …
Value = …

Reduce

Key = …
Value = …

Exercise 2

Write a MapReduce to identify the most critical moments by computing the average usage for each timeslot, over all stations.
The timeslot is defined as the hour of the day-of-the-week, e.g., Monday at 13, Saturday at 12 (which means all measurements of all Saturdays of all stations from 12:00 to 12:59).
The average usage of the system over all the stations is defined as follows:

total number of used stalls / (total number of free+used stalls)

Map

Key = …
Value = …

Reduce

Key = …
Value = …

Exercise 3

Write a MapReduce to identify the most critical places and moments together, by computing the average usage for each timeslot and for each station.
The timeslot is defined as the hour of the day-of-the-week, e.g., Monday at 13, Saturday at 12.
For instance, if the timeslot is Saturday at 12, the average usage must be computed over all measurements of all Saturdays from 12:00 to 12:59, separately for each station.
The average usage is defined as follows:

total number of used stalls / (total number of free+used stalls)

Map

Key = …
Value = …

Reduce

Key = …
Value = …

Exercise 4

Using the *restaurant* collection of MongoDB described in Lab Practice #7, write the Map Reduce to count the number of restaurants whose review is higher than 4.5 and that can contain more than 5 people.