CSE441: DATABASE SYSTEMS

ASSIGNMENT 4 (Spark)

DEADLINE: 9:00 pm, 1st April 2020

In this assignment, you have to simple problems using Spark architectures. Spark is a distributed general-purpose cluster-computing framework.

Programming Languages Allowed: Python

Prerequisites:

Spark Installation:

1. Spark architecture is provided in python under the library "PySpark". Install it using **pip**. Pyspark is similar to MapReduce. The main difference is in PySpark all operations are in memory.

Readings:

1. https://www.tutorialspoint.com/pyspark/index.htm
2. Examples: https://github.com/apache/spark/tree/master/examples/src/main/python

Dataset:

Dataset for the problem is a dataset on Airports which can be downloaded from moodle.

Problems:

1. Write a program to get the number of Airports by Country.
2. Write a program to find the Country having the highest number of airports.
3. Write a program to find airports whose latitude is between [10, 90] and longitude is between [-10, -90]. ([a,b] a,b both are included)

Note:

1. Take the number of CPUs as a command-line argument.

Input:

1. For all the problems take input from airport.csv file in the dataset.

Output:

1. Write outputs to a new file, take output filename as a command-line argument.

Deliverables

Create a folder with RollNuber_Assignment4 and put the following into it

1. All *pyspark_[problemnumber].py* files

Zip the folder and upload.

Strict action will be taken for copying in the Assignments.