THE UNIVERSITY OF SYDNEY

School of Information Technologies
Dr Ying Zhou and A/Prof. Uwe Roehm

COMP5349: Cloud Computing **1.Sem./2015**

Assignment 2: Hadoop Hive and Spark

Group Work: 20% **12.05.2015**

Introduction

This second assignment is based on the same data set as Assignment 1. Your task is to perform a time series data analysis as specified under 'Problem Description' below, but with two different implementations. You are also required to compare the performance of the two implementations, and to write a short performance report.

Problem description

You are asked to find out how many times a user has visited a particular country, as well as the maximum, minimum, average, and total time he/she spent in this country.

In the assignment data set, a user is considered of having visited a place if he/she has taken at least one photo in that place. A user may have stayed in many different places in a country; as long as the user did not visit another country in between, these are considered as one visit of that country. A visit of a country may last for several days, weeks or even years.

The following sample data shows a particular user's photo records, sorted in chronological order. Each line represents a photo record. For simplicity, we only show user id as well as the time and place the photo is taken. From the sample data, we can tell that the user 10530649@N05 has visited four countries in total. He/she has visited Australia twice and all other countries once.

```
10530649@N05 2007-04-19 12:31:34 /Australia/QLD/Bundaberg/Kensington
10530649@N05 2007-05-23 08:08:27 /Australia/QLD/Cairns/Brinsmead
10530649@N05 2008-03-24 09:04:52 /Brazil/Maranhao/Primeira+Cruz
10530649@N05 2008-04-02 11:23:03 /Brazil/Espirito+Santo/Anchieta
10530649@N05 2008-04-10 16:33:25 /Brazil/Espirito+Santo/Guarapari
10530649@N05 2008-04-20 15:17:43 /United+Kingdom/Scotland/Spean+Bridge
10530649@N05 2008-05-06 13:24:03 /United+Kingdom/Scotland/Eshaness
10530649@N05 2009-10-23 11:00:16 /United+States/California/Cayucos
10530649@N05 2009-10-25 08:21:44 /United+States/California/Mill+Valley
10530649@N05 2009-12-10 20:53:56 /Australia/SA/Adelaide/Eden+Hills
10530649@N05 2009-12-14 10:11:31 /Australia/SA/Whyalla/Whyalla+Norrie
```

You are asked to compute the time spent in a country as shown in Table 1. The duration is computed as the difference in **days** between the first time a user took a photo in a country and the first time the user took a photo in next country. If the country is the last one or the only one a user visits, the duration is computed as the difference in **days** between the first and the last time a user took a photo in this country (see example of Australia #2). For countries with just one photo it is Ok to calculate the duration as 0.

Table 1: Computing the time spent in a country

Country	[start time, end time]	Duration (days)
Australia #1	[2007-04-19 12:31:34, 2008-03-24 09:04:52]	339.8
Brazil	[2008-03-24 09:04:52, 2008-04-20 15:17:43]	27.3
United Kingdom	[2008-04-20 15:17:43, 2009-10-23 11:00:16]	550.8
United States	[2009-10-23 11:00:16, 2009-12-10 20:53:56]	48.4
Australia #2	[2009-12-10 20:53:56, 2009-12-14 10:11:31]	3.6

For the above sample data, you are expected to produce the following result as a single record (line) for user 10530649@N05:

```
10530649@N05 Australia(2,339.8,3.6,171.7,343.4),
/ Brazil(1,27.3,27.3,27.3,27.3), United Kingdom(1,550.8,550.8,550.8,550.8),
/ United States(1,48.4,48.4,48.4,48.4)
```

There is no particular requirement on the order of countries in a user's record.

Implementation Requirements

You should provide **two implementations** of the above problem. Both implementations should produce the same results.

- The first implementation should use some HIVE queries to handle certain functions. It can be a mixture of HIVE queries and hand coded MapReduce jobs. There is no requirement on the number of HIVE queries you should use. You can decide how you mix HIVE queries and hand code MapReduce jobs to produce the final result. For instance , you may do everything in HIVE; or you may use HIVE to do some filtering/joining jobs and use hand coded MapReduce jobs for other parts. Such decisions are usually made based on your analysis of the problem and your familiarity with SQL and programming language. If you use a mixture of HIVE and MapReduce, you do not have to chain them automatically.
- The second implementation should be in Spark.

Performance Report Requirements

The report should contain the following three sections:

- **HIVE and/or MapReduce Design**
  In this section , describe the HIVE queries and the MapReduce jobs you use to produce the final result in implementation 1. Give a brief reason of choosing HIVE or hand coded MapReduce jobs to implement certain functions.

- **Spark Design**
  In this section, describe the Spark application you have designed. You should draw a lineage graph and briefly describe each operation you use.

- **Performance Evaluation**
  In this section, discuss the performance of your implementations. In your discussion, you should show at least the total execution time and explain the difference you observed. You are encouraged to cite other useful performance metrics such as the number jobs in implementation 1, and the number of stages in your Spark application and the size of the data shuffled.

Submission Guidelines

- Demo both implementations using a small data set in week 13's lab.

- Submit a hard copy of your report and a signed group assignment cover sheet in week 13's lab.

- Submit your **source code** and **hive query script** as a zip file to Blackboard Learn before **mid-day** 4^{th} of June, 2015. Please be advised that you need to submit your report before the demo for assignment 2.