Big Data Analytics
Exercise Sheet 9

Prof. Dr. Dr. Lars Schmidt-Thieme, Mohsan Jameel
Information Systems and Machine Learning Lab
University of Hildesheim

Uploaded June 4th, 2019,
Deadline on June 11th, 2019 at 08:00am,
(Should be submitted as a single unzipped PDF file on learn-web course "SoSe 2019: 3104 Big Data Analytics"

Instructions

Please following these instructions for solving and submitting the exercise sheet.

1. Student should clearly write his/her name, matriculation number and tutorial group number (i.e. “Group 1: Tuesday Tutorial”, “Group 2: Thursday Tutorial” and “Group 3: Wednesday Tutorial”).
2. The submission should be made before the deadline, only through learnweb to your group submission link.
3. Should be submitted as a single unzipped PDF file on learn-web course "SoSe 2019: 3104 Big Data Analytics".
4. Each student must submit an individual solution in-order to be eligible for bonus points.
5. Group submission are acceptable but will not contribute towards bonus points.

Exercise 1: Apache Spark Essentials (6 points)

Please provide brief answers to the following questions.

a. **(2 points)** The replication in a distributed file system provides fault-tolerance by storing replicas of the original data on different server nodes. Lets suppose there is a set of data manipulations performed on the original data to get some useful results. In case of a node failure (or a job failure), we can retrieve original data from one of the replicas. However, the set of data manipulations is lost with the node failure. Which technique is used to preserve the set of data manipulations so we can retrieve them in the case of a node failure? Also list its components (Only list).

b. **(2 points)** Apache Spark has Resilient Distributed Datasets (RDD), which works on lazy evaluation mechanism. Explain what is lazy evaluation. Can we avoid lazy evaluation in Apache Spark?

c. **(2 points)** How many types of transformations exist in Apache Spark? What are the affects of each type of transformation on partitioned data? [Hint: You are not asked about the names of the transformation functions but their types]

Exercise 2: Problem solving using Apache Spark (14 points)

For the following questions you are given a csv file, where each row represents a flight record and each cell is separated by a ‘,’. The csv file is given below (visualized in the form of a table). The delay time is positive if a flight depart or arrive late and negative if a flight depart or arrive early. In the questions below

FL_NUM	ORIGIN	DEST	DEP_TIME	DEP_DELAY	ARR_TIME	ARR_DELAY
"307"	"DEN"	"PHX"	"1135"	-10.00	"1328"	-17.00
"307"	"PHX"	"DEN"	"1502"	-8.00	"1653"	-8.00
"309"	"DCA"	"MIA"	"0646"	-13.00	"0930"	-14.00
"310"	"MIA"	"LGA"	"1402"	-3.00	"1646"	-14.00
"310"	"FLL"	"MIA"	"1540"	175.00	"1740"	165.00
"311"	"PHX"	"DEN"	"2310"	16.00	"0107"	28.00
"311"	"SEA"	"PHX"	"1800"	1.00	"2141"	-2.00
"309"	"FLL"	"LGA"	"1818"	18.00	"2016"	-4.00

you have to find an average delay in arrival time for each destination (DEST).

Part a: pseudocode (7 points)

In this part you are required to provide a pseudocode that consists of a set of Apache Spark transformations and actions that also includes loading the data into RDDs.

Part b: Working example (dry run 1) (7 points)

With the help of a flowchart/block diagram, show what happens at each step of the program in (Part a). Consider you are running with 4 executors. Your diagram should reflect step-by-step transformations and actions on RDDs. [Hint: You can explain your steps from the time your Apache Spark system gets ready for execution, loading data, transformations and actions on RDDs. Take a look at the lecture slide “C.2. Resilient Distributed Datasets: Apache Spark” (slides number 28-29)]

Annexture

1. Dry run `http://cpmadeeasy.blogspot.com/2013/03/dry-run.html`