Universität Freiburg
Institut für Informatik

Georges-Köhler Allee, Geb. 51
D-79110 Freiburg

Fang Wei-Kleiner

fwei@informatik.uni-freiburg.de

Advanced Databases and Information Systems
Summerterm 2019
Discussion on

7. Sheet: Distributed Processing II

Exercise 1 (Dealing with data in Spark)

Spark offers different concepts (or abstractions) to deal with data, namely *RDDs*, *DataFrames* and *Datasets*. The latter were introduced in different versions of Spark and are frequently adapted for more efficient and transparent use for end-users. Finally, dealing with data differs based on the programming language you chose (i.e. Scala, Java, Python or R).

a) Summarize the central properties of the mentioned approaches for data usage (take into account how Spark evolved with respect to different versions).

b) Discuss their commonalities and differences.

c) What are their generell advantages and disadvantages?

d) Which restrictions to available programming languages impose on them?

Exercise 2 (Spark environment setup and dataset)

Given the Docker image with Hadoop and Spark installation you were introduced to during the exercise, you will process the dataset "citeulike". In ILIAS you will find the file *citeulike_adbis.zip* which contains the dataset and a *README.txt* file. Read it in order to get a better understanding of the content of each file.

Exercise 3 (Loading the dataset into an RDD)

Create a ***pair*** RDD (i.e. a collection as key-value pairs) for the citeulike dataset.

a) *userRatingsRDD*: create a pair RDD from *user_libraries.txt* using the user hash as the key and the liked paper(s) (*paper_id*) as the value(s).

b) *paperTermsRDD* : create a pair RDD from *papers.csv* using the *paper_id* as the key and the words contained in the abstract as the value(s).

Note: read the data directly into RDD, i.e. do not build DataFrames or any other data models and then convert them to RDDs.

Exercise 4 (Basic analytics with RDDs)

Using Spark's capabilities, retrieve the following information:

a) Number of (distinct) users, number of (distinct) items, and number of ratings

b) Min number of ratings a user has given

c) Max number of ratings a user has given

d) Average number of ratings of users

e) Standard deviation for ratings of users

f) Min number of ratings an item has received

g) Max number of ratings an item has received

h) Average number of ratings of items

i) Standard deviation for ratings of items

Exercise 5 (Loading datasets into Data Frames)
In contrast to RDDs, DataFrames allow one to handle structured, distributed data in a table-like representation with named and typed columns. DataFrames are therefore applicable in any instance that requires manipulation of structured data.

Load the dataset into DataFrames leveraging the structure of the data. Choose a reasonable schema.

Exercise 6 (Tasks on top of DataFrames)
Solve exercise 4 again using DataFrames instead of RDDs. Use only DataFrame's functionalities, i.e. avoid converting your model to RDD and then using RDD's functionalities.