**Recitation no 4**

**TASK 1: GraphX and PageRank**

GraphX comes with static and dynamic implementations of PageRank as methods on the PageRank object. Static PageRank runs for a fixed number of iterations, while dynamic PageRank runs until the ranks converge (i.e., stop changing by more than a specified tolerance). GraphOps allows calling these algorithms directly as methods on Graph. GraphX also includes an example social network dataset that we can run PageRank on. A set of users is given in graphx/data/users.txt, and a set of relationships between users is given in graphx/data/followers.txt.

Open the following files:

Edit built.sbt as needed.

Edit PageRank.scala as needed.

What does the code in PageRank.scala do?

What is the convergence criteria for PageRank in the example?

Also:

--- Compile the PageRank example from GraphX by downloading the below file:

<https://github.com/apache/spark/blob/master/graphx/src/main/scala/org/apache/spark/graphx/lib/PageRank.scala>

--- What does the PageRank code do?

--- What is the convergence criteria?

**TASK 2:**

**Part 1: Getting started with Spark SQL**

Like Apache Spark in general, Spark SQL in particular is all about distributed in-memory computations on massive scale. The primary difference between Spark SQL’s and the “bare” Spark Core’s RDD computation models is the framework for loading, querying and persisting structured and semi structured data using structured queries that can be expressed using good ol’ SQL and HiveQL. More information about Spark SQL and its processes can be found here:

<https://amplab.cs.berkeley.edu/wp-content/uploads/2015/03/SparkSQLSigmod2015.pdf>

For this exercise we will use KDD Cup 1999 dataset, found here:

<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

The dataset contains a wide variety of intrusions simulated in a military network environment. This dataset will be later used to develop a machine learning based intrusion detection system.

But first let’s manipulate and transform the data with Spark SQL.

Make the following folder with the Task2 problems that you download from sakai: recitation4/problems

wget http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data\_10\_percent.gz

gzip -d kddcup.data\_10\_percent.gz

jupyter notebook sparkSQL.ipynb

Follow the instructions from this point on and please save/store printscreens of all conducted instructions.

**Part 2: Summary Statistics with Spark SQL**

In descriptive statistics, summary statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible.

jupyter notebook summary.ipynb

**References:**

Parts of the recitation are adapted from Content taken from Spark Notebooks by Jose A Dianes and other third parties.

Learning Spark by Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia.