Tutorial 8

This tutorial will be a demo session demonstrating the use of existing Hadoop Distributed File System (HDFS) infrastructure in Apache Spark.

The tutorial comprises of four steps:

1. Hadoop setup: Use Hadoop installation script, *03_hadoop_install.sh* (tested only on Ubuntu 14.04). You need a user *hduser* with root access using sudo and home directory as */home/hduser*. To setup user and root access via sudo, you can use scripts: *01_VM_start.sh* and *02_hadoop_add_user.sh*.

2. Start HDFS: Hadoop HDFS can be started by entering the command *start-dfs.sh*. To stop HDFS, enter the command *stop-dfs.sh*.

3. Copy the data to HDFS: Use command *hdfs dfs -put $LOCAL_FILE_PATH $HADOOP_DESTINATION_DIRECTORY*. Familiarize yourself with HDFS (see: `https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-common/FileSystemShell.html`)

4. Finally, we will make Spark SQL queries to HDFS data. For the demo, public transport data for Helsinki region will be used (available at `http://dev.hsl.fi/ajoaika_gps/`).

NOTE: The translation of column names from Finnish to English for HSL data is available at: `http://datasciencehackathon.aalto.fi/index.html?page_id=108`