TDDE31/732A54-Big Data Analytics Lab Compendium

Notice: *Please make sure you have read the whole lab compendium before you start to work on the server from NSC.*

April 3rd, 2025

1 Description and Aim

In the lab you will work on the Sigma[1] set up which is a HPC cluster from the National Supercomputer Centre (NSC). You are supposed to work with the historical meteorological data from the Swedish Meteorological and Hydrological Institute (SMHI). Specifically, you will work with air temperature readings and precipitation readings from 812 meteorological stations in Sweden[2]. In these exercises, you will work with Spark 3.5.1[3].

There are three lab topics which are *BDA1-Spark*, *BDA2-Spark SQL* and *BDA3-Machine Learning with Spark*. After completing the first two labs you will have basic knowledge of the programming environment, and PySpark API. You will work on exercises with Spark and Spark SQL and thus will be able to compare the differences between the two approaches. In the third lab, you are supposed to achieve a machine learning method with Spark. The overview of three labs is that you need to upload data to Hadoop Distributed File System (HDFS), read the data from HDFS in your code and then program with PySpark[4] to solve the assignments in each lab.

2 SMHI Data

The data includes air temperature and precipitation readings from 812 stations in Sweden. The stations include both currently active stations and historical stations that have been closed down. The latest readings available for active stations are from October 10, 2016. The air temperature and precipitation records are hourly readings, however some stations provide only one reading every three hours.

The provided files[5] are prepared csv files with removed headers. Values are separated with semicolons. Some files are too big to be read using some text editors. Therefore, please use

[1]Sigma at: `https://www.nsc.liu.se/systems/sigma/`
[2]If interested in other readings please check: `http://opendata-catalog.smhi.se/explore/`
[3]Spark 3.5.1 at: `https://spark.apache.org/docs/3.5.1/`
[4]PySpark 3.5.1 at: `https://spark.apache.org/docs/3.5.1/api/python/reference/index.html`
[5]Within the BDA_demo: `https://www.ida.liu.se/~732A54/lab/data.zip`

either python to read the files or bash commands such as *tail* and *more* to get an overview of a file's content. Provided files:

- temperature-readings.csv - ca 2 GB
- precipitation-readings.csv - ca 660 MB
- stations.csv
- stations-Ostergotland.csv

All these files are available at the following folder at Sigma:
/software/sse2/tetralith_el9/manual/spark/course-examples/BDA_demo/input_data

The headers of these files are shown in Table 1, Table 2 and Table 3. If you notice any mistakes in the dataset or have any comments please contact the lab assistants.

Station number	Date	Time	Air temperature (in °C)	Quality[a]

[a] G - controlled and confirmed values, Y - suspected or aggregated values

Table 1: Headers for temperature-readings.csv

Station number	Date	Time	Precipitation (in mm)	Quality[a]

[a] G - controlled and confirmed values, Y - suspected or aggregated values

Table 2: Headers for precipitation-readings.csv

Station number	Station name	Measurement height	Latitude	Longitude	Readings from (date and time)	Readings to (date and time)	Elevation

Table 3: Headers for stations.csv and stations-Ostergotland.csv

3 Running PySpark Program on Sigma

3.1 Working on Sigma

The Sigma server is available at ***sigma.nsc.liu.se*** (log in using your NSC accounts). You can use (**Option 1:**) ssh forwarding connection or (**Option 2:**) Thinlinc[6] connection to log in Sigma. You can also use regular ssh connection without forwarding option, in this case, you have to program locally and use *scp* command to upload your code to Sigma, and then check the history log from text file. **Note:** The forwarding function of ssh supports to run graphics applications (e.g., Atom editor) remotely.

- **Option 1:** when you are at a computer in an SU room at the university,

[6] `https://www.cendio.com/thinlinc/download`

- ***ssh -X username@sigma.nsc.liu.se*** where username is your NSC username (not the LiU one), -X indicates forwarding function of ssh which is used for running graphics applications remotely.
- ***[username@sigma ~] $ exit*** is used to logout Sigma. If it is hung on, please use ctrl-c to terminate the connection.
- You can use Emacs editor for coding by running following command first in a terminal:
  - ***emacs &***
- You can also use VS Code editor for coding by running following two commands first in a terminal:
  - ***/proj/liu-compute-2025-6/shared/VSCode-linux-x64/code --no-sandbox &***
- ***scp LOCAL_FILE_PATH username@sigma.nsc.liu.se:Documents*** is used for uploading files from your local machine to Sigma. (**Note:** you are supposed to run *scp* command before you log in Sigma when you want to upload files to Sigma.)
- You can always first connect to the university server and from there connect to Sigma. Information regarding remote access to LiU's system can be found here.

- **Option 2:** Sigma's Thinlinc server is available same as ***sigma.nsc.liu.se***. In this way, you can get a graphical environment on Sigma and given that you work directly on Sigma there is no need to use *ssh* or *scp*. **Note:** Please remember to log out when done working on the labs so that Sigma does not keep open Thinlinc sessions. Also, each pair of students, please use at most one Thinlinc connection during lab sessions, due to the limited number of Thinlinc licenses on Sigma.
  - When you are at a computer in an SU room at the university:
    - You need to open a terminal and run following two lines:
    - ***$ module load courses/732A54*** or ***$ module load courses/TDDE31***
    - ***$ tlclient***
  - If you use your own computers, you can just download Thinlinc first and then connect directly.
  - As mentioned above, on Sigma, you can use Emacs or VS Code editors for coding by running aforementioned commands in a terminal.

NSC reserves nodes for each lab session of the course, which means other jobs on Sigma will not use these nodes during our lab sessions. Table 4 shows reservation names for our scheduled lab sessions as well as a reservation, **devel**, which you can use at any time (e.g., outside of lab sessions).

Each time when you submit a job, the options '**-A liu-compute-2025-6**', and '**--reservation RESERVATION_NAME**' should be specified.

- ***[username@sigma ~] $ listreservations***, by running this command, you can see all the reservations as shown in Figure 1.
- ***[username@sigma ~] $ sbatch -A liu-compute-2025-6 --reservation liu-bda-2025-04-15 run_yarn_with_historyserver.q***

Table 4: Time and Reservation Name

RESERVATION_NAME	Time
devel	Anytime
liu-bda-2025-04-15	04-15, 13:15 to 17:15
liu-bda-2025-04-17	04-17, 08:15 to 10:15
liu-bda-2025-04-22	04-22, 15:15 to 17:15
liu-bda-2025-04-24	04-24, 08:15 to 10:15
liu-bda-2025-04-25	04-25, 15:15 to 17:15
liu-bda-2025-04-28	04-28, 10:15 to 12:15
liu-bda-2025-04-29	04-29, 13:15 to 17:15
liu-bda-2025-05-05	05-05, 10:15 to 12:15
liu-bda-2025-05-06	05-06, 13:15 to 19:15
liu-bda-2025-05-08	05-08, 08:15 to 10:15
liu-bda-2025-05-09	05-09, 15:15 to 17:15
liu-bda-2025-05-13	05-13, 15:15 to 17:15
liu-bda-2025-05-15	05-15, 08:15 to 10:15
liu-bda-2025-05-16	05-16, 15:15 to 17:15
liu-bda-2025-05-19	05-19, 10:15 to 12:15
liu-bda-2025-05-20	05-20, 13:25 to 17:15

```
[x_huali@sigma ~]$ listreservations
Reservations available to user:x_huali / project(s):liu-compute-2024-28,liu-compute-2025-6,liu-compute-2024-23
 devel  from 2025-04-03T08:00:00 to 2025-04-03T22:00:00  ALL USERS
 liu-bda-2025-04-15  from 2025-04-15T13:15:00 to 2025-04-15T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-04-17  from 2025-04-17T08:15:00 to 2025-04-17T10:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-04-22  from 2025-04-22T15:15:00 to 2025-04-22T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-04-24  from 2025-04-24T08:15:00 to 2025-04-24T10:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-04-25  from 2025-04-25T15:15:00 to 2025-04-25T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-04-28  from 2025-04-28T10:15:00 to 2025-04-28T12:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-04-29  from 2025-04-29T13:15:00 to 2025-04-29T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-05  from 2025-05-05T10:15:00 to 2025-05-05T12:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-06  from 2025-05-06T13:15:00 to 2025-05-06T19:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-08  from 2025-05-08T08:15:00 to 2025-05-08T10:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-09  from 2025-05-09T15:15:00 to 2025-05-09T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-13  from 2025-05-13T15:15:00 to 2025-05-13T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-15  from 2025-05-15T08:15:00 to 2025-05-15T10:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-16  from 2025-05-16T15:15:00 to 2025-05-16T17:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-19  from 2025-05-19T10:15:00 to 2025-05-19T12:15:00  (project:liu-compute-2025-6)
 liu-bda-2025-05-20  from 2025-05-20T13:15:00 to 2025-05-20T17:15:00  (project:liu-compute-2025-6)
```

Figure 1: List reservations.

3.2 Submitting Jobs to Hadoop Cluster and Sigma

To submit your pyspark code to the Hadoop cluster, you will use:

- ***spark-submit --deploy-mode cluster --master yarn --num-executors 9 --driver-memory 2g --executor-memory 2g --executor-cores 4 CODE.py*** where CODE.py is the python script in the current folder. In this command, Yarn is used for resource management and the cluster-deploy mode is used.

To run your pyspark code on Hadoop cluster, you also need to first submit a job to Sigma. Some scripts in the demo (in next section) are provided to you so you can make the call of

your *spark-submit* command. You will use the non-interactive way to submit a job on Sigma. Each time after you submit, the job will enter the scheduling queue. You can use **sbatch** command to submit the job, **squeue** command to monitor the submitted job and may use **scancel** command to cancel a job.

- ***[username@sigma ~] $ sbatch -A liu-compute-2025-6 --reservation liu-bda-2025-04-09 run_yarn_with_historyserver.q*** (Note: Please remember to use '-A liu-compute-2025-6' if you are in more than one project from NSC, it will guarantee to use the allocation reserved on Sigma for our course.)
- ***[username@sigma ~] $ squeue -u username***
- ***[username@sigma ~] $ scancel JOB_ID***

Sigma uses Slurm for scheduling. Once you submit a job, the job will be assigned an ID. After the job is finished, you will see a **slurm-JOB_ID.out** file returned, which includes the output information of the job script. The script **run_yarn_with_historyserver.q** is supposed to contain commands for constructing the Hadoop cluster for your account, commands for interacting with Hadoop Distributed File System (HDFS) and *spark-submit* command for running pyspark code. A detailed example (Figure 4) will be shown in following section which introduces the demo.

Another way to access the logs after the execution, you will need to set the *spark.eventLog.enabled* which is already contained in '*run_yarn_with_historyserver.q*'. After the job finishes, you can access the history server by running:

- ***spark_browse_job***

Then a firefox session that points at the history server web interface will be opened. In this way, it only lists the latest finished job on the history server.

3.3 Demo Code

You can find the demo from following folders after you log in Sigma.

- */software/sse2/tetralith_el9/manual/spark/course-examples/BDA_demo/*

The steps to run the *BDA_demo* are shown as below, and in Figure 2 and Figure 3.

- Step 1: Login Sigma with *'ssh -X'* connection or Thinlinc.
- Step 2: Copy the demo to your home folder on Sigma.
- Step 3: Use *sbatch* to submit the job. Then use *squeue* command to monitor your job. Once the job is finished, check the returned file named *slurm-ID.out*. Notice: when you want to run your own code, you need to make some changes in '*run_yarn_with_historyserver.q*' to make sure you create folders on HDFS and copy data to HDFS correctly (see Figure 4).
- Step 4: Check the history log from *slurm-ID.out* or run *spark_browse_job*, or check the output folder.

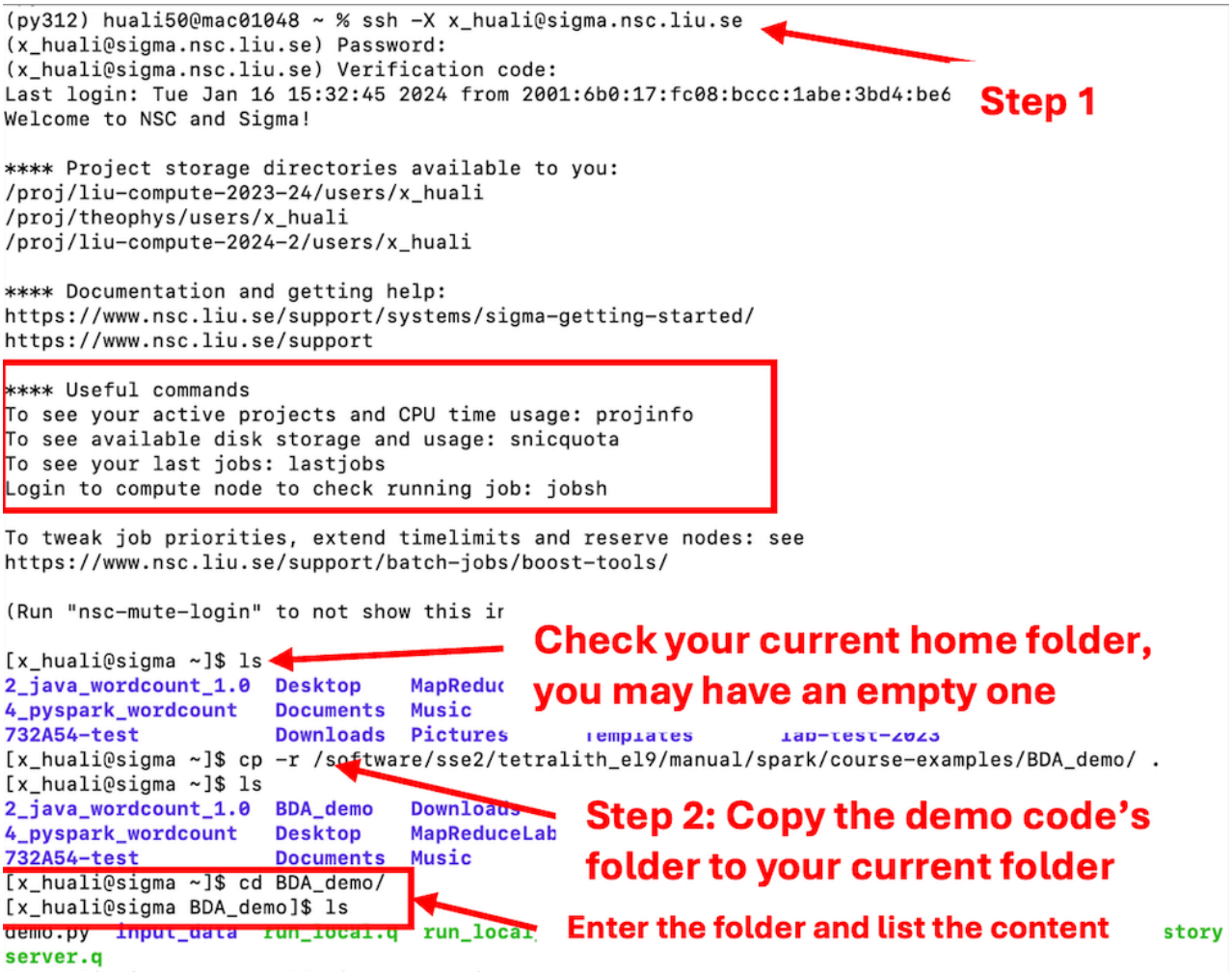

Figure 2: Login Sigma and Copy the demo code (Steps 1 and 2).

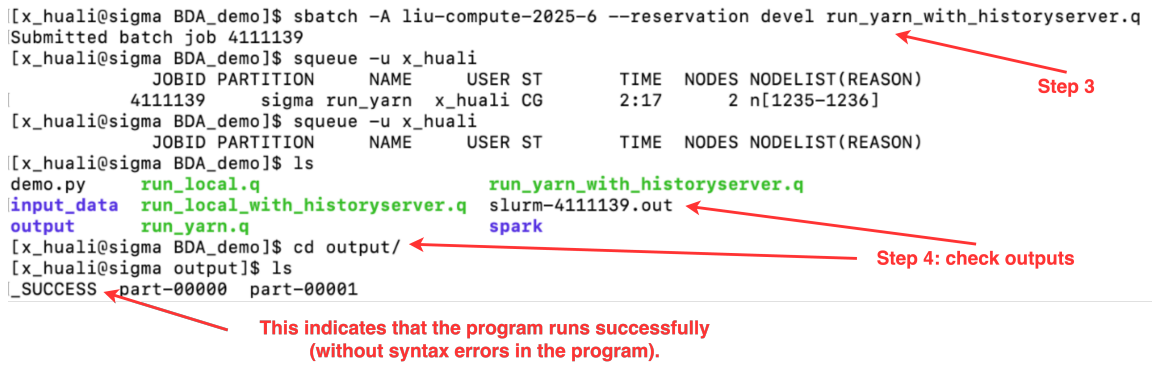

Figure 3: Submit a job and check the output (Steps 3 and 4).

In '*run_yarn_with_historyserver.q*' as shown in Figure 4, you can see a number of commands that are used to interact with HDFS.

- ***hadoop fs -mkdir <FOLDER_NAME>*** -make a folder on HDFS
- ***hadoop fs -mkdir -p <FOLDER_NAME> <FOLDER_NAME>*** -make multiple folders

- ***hadoop fs -test -d <FOLDER_NAME>*** -if the path is a directory, return 0
- ***hadoop fs -rm -r <FOLDER_NAME>*** -deletes the directory and any content under it recursively
- ***hadoop fs -cat <FOLDER_ON_HDFS> [local]*** -copy HDFS path to stdout
- ***hadoop fs -copyFromLocal <localsrc> ... <dst>*** -copy single src, or multiple srcs from local Sigma to HDFS
- ***hadoop fs -copyToLocal <dst> ... <localsrc>*** -copy single src, or multiple srcs from HDFS to local Sigma

```
#!/bin/bash
#SBATCH --time=10:00
#SBATCH --nodes=2
#SBATCH --exclusive

echo "START AT: $(date)"

module load spark/3.5.1-hadoop-3.3.6-hpc1-bdist

# Cleanup and start from scratch
rm -rf spark

# Startup hadoop filesystem and yarn
hadoop_setup

echo "Prepare output and input directories and files..."
# The following command will make folders on your home folder on HDFS, the input and output fold
ers should be corresponding to the parameter you give to textFile and saveAsTextFile functions i
n the code
hadoop fs -mkdir -p "BDA" "BDA/input"
hadoop fs -test -d "BDA/output"
if [ "$?" == "0" ]; then
    hadoop fs -rm -r "BDA/output"
fi

hadoop fs -copyFromLocal ./input_data/temperature-readings-small.csv "BDA/input/"
# Remove the comment when you need specifc file below
#hadoop fs -copyFromLocal ./input_data/temperature-readings.csv "BDA/input/"
#hadoop fs -copyFromLocal ./input_data/precipitation-readings.csv "BDA/input/"
#hadoop fs -copyFromLocal ./input_data/stations.csv "BDA/input/"
#hadoop fs -copyFromLocal ./input_data/stations-Ostergotland.csv "BDA/input/"

# Run your program
echo "Running Your program..."
exec 5>&1
APPLICATION_ID=$(spark-submit --conf spark.eventLog.enabled=true --deploy-mode cluster --master
yarn --num-executors 9 --driver-memory 2g --executor-memory 2g --executor-cores 4 demo.py 2>&1 |
 tee >(cat - >&5) | awk '!found && /INFO.*Yarn.*Submitted application/ {tmp=gensub(/^.*Submitted
 application (.*)$/,"\\1","g");print tmp; found=1}')

echo "================ FINAL OUTPUT ========================================="
hadoop fs -cat "BDA/output"/*
echo "================================================================================"
echo "Applicaton id: $APPLICATION_ID"
echo "================================================================================"
echo "=                                stderr                                      ="
echo "================================================================================"
yarn logs -applicationId "$APPLICATION_ID" | awk -F: '/^LogType/ {if($2=="stderr") {output=1} el
se {output=0}} output==1 {print}'
echo "================================================================================"
echo "=                                result                                      ="
echo "================================================================================"
yarn logs -applicationId "$APPLICATION_ID" | awk -F: '/^LogType/ {if($2=="stdout") {output=1} el
se {output=0}} output==1 {print}' | grep -v "WARN\|INFO"

rm -rf output
hadoop fs -copyToLocal 'BDA/output' ./
hadoop_stop

echo "END AT: $(date)"
```

Please read these comment lines, you need to change these hadoop commands.

Here you need to replace with your own code.

Figure 4: '*run_yarn_with_historyserver.q*' for running the code

4 Hand In

You are supposed to use GitLab[7] to submit your report and code. For each lab, please submit the code and a report that contains your results (a snippet of the results is enough if the results contain many rows) and answers to the questions. In cases where a plot of your results is asked, you can include the figure directly in the report. You can use a tool of your preference to produce the plots (R, Excel, matplotlib in Python, etc.). Comment each step in your code to provide a clear picture of your reasoning when solving the problem.

[7] `https://gitlab.liu.se/olaha93/bigdata`