Assignment 16.1

Step 1: Start Hadoop Daemons:

$ start-all.sh

$ jps

```
[acadgild@localhost dataset]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/08/28 16:35:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: namenode running as process 2830. Stop it first.
localhost: datanode running as process 2931. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: secondarynamenode running as process 3086. Stop it first.
17/08/28 16:35:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
resourcemanager running as process 3263. Stop it first.
localhost: nodemanager running as process 3364. Stop it first.
[acadgild@localhost dataset]$ jps
2931 DataNode
3364 NodeManager
2830 NameNode
3086 SecondaryNameNode
3263 ResourceManager
4111 Jps
[acadgild@localhost dataset]$
```

Step 2: create a dataset as below where it contains id, name, salary, unit:

$ cat employee.txt

```
[acadgild@localhost assignment16]$ cat employee.txt
1       Amit    100     DNA
2       Sumit   200     DNA
3       Yadav   300     DNA
4       Sunil   500     FCS
5       Kranti  100     FCS
6       Mahoor  200     FCS
8       Chandra 500     DNA
[acadgild@localhost assignment16]$
```

Step 3: start hive shell and load data:

$ sudo service mysqld start

$ hive

```
USE default;
CREATE TABLE employee
(
id INT,
name STRING,
salary INT,
unit STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';
LOAD DATA LOCAL INPATH '/home/acadgild/dataset/assignment16/employee.txt'
INTO TABLE employee;
SELECT * FROM employee;
```

```
[acadgild@localhost assignment16]$ sudo service mysqld start
[sudo] password for acadgild:
Starting mysqld:                                           [  OK  ]
[acadgild@localhost assignment16]$ hive

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-commo
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hive/lib/hive-jdbc-0.14.0-standalone.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hive> USE default;
OK
Time taken: 0.472 seconds
hive> CREATE TABLE employee
    > (
    > id INT,
    > name STRING,
    > salary INT,
    > unit STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t';
OK
Time taken: 0.325 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/dataset/assignment16/employee.txt'
    > INTO TABLE employee;
Loading data to table default.employee
Table default.employee stats: [numFiles=1, totalSize=115]
OK
Time taken: 0.684 seconds
hive> SELECT * FROM employee;
OK
1       Amit    100     DNA
2       Sumit   200     DNA
3       Yadav   300     DNA
4       Sunil   500     FCS
5       Kranti  100     FCS
6       Mahoor  200     FCS
8       Chandra 500     DNA
Time taken: 0.344 seconds, Fetched: 7 row(s)
hive>
```

Step 4: Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit:

Find the lead salary first:

SELECT id, name, salary, unit, LEAD(salary) OVER (PARTITION BY unit ORDER BY salary) AS lead_salary FROM employee;

```
hive> SELECT id, name, salary, unit, LEAD(salary) OVER (PARTITION BY unit ORDER BY salary) AS lead_salary
    > FROM employee;
Query ID = acadgild_20170828210000_9546a04b-6d73-45aa-a160-fe462d077881
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1503918153486_0020, Tracking URL = http://localhost:8088/proxy/application_1503918153486_
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1503918153486_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-08-28 21:00:23,634 Stage-1 map = 0%,  reduce = 0%
2017-08-28 21:00:28,852 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.71 sec
2017-08-28 21:00:35,156 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.04 sec
MapReduce Total cumulative CPU time: 2 seconds 40 msec
Ended Job = job_1503918153486_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.04 sec   HDFS Read: 332 HDFS Write: 141 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 40 msec
OK
1       Amit    100     DNA     200
2       Sumit   200     DNA     300
3       Yadav   300     DNA     500
8       Chandra 500     DNA     NULL
5       Kranti  100     FCS     200
6       Mahoor  200     FCS     500
4       Sunil   500     FCS     NULL
Time taken: 19.096 seconds, Fetched: 7 row(s)
hive>
```

Calculate the difference:

SELECT id, name, salary, unit, (lead_salary - salary) AS diff_salary FROM
(
SELECT id, name, salary, unit, LEAD(salary) OVER (PARTITION BY unit ORDER BY salary) AS lead_salary
FROM employee
) temp
WHERE lead_salary - salary > 100;

```
hive> SELECT id, name, salary, unit, (lead_salary - salary) AS diff_salary FROM
    > (
    > SELECT id, name, salary, unit, LEAD(salary) OVER (PARTITION BY unit ORDER BY salary) AS lead_salary
    > FROM employee
    > ) temp
    > WHERE lead_salary - salary > 100;
    > WHERE lead_salary - salary > 100;
Query ID = acadgild_20170828210101_9913de8d-1e1a-4d1d-bfdb-89f2b6918183
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1503918153486_0021, Tracking URL = http://localhost:8088/proxy/application_1503918153486_0021/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1503918153486_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-08-28 21:01:21,927 Stage-1 map = 0%,  reduce = 0%
2017-08-28 21:01:27,213 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.66 sec
2017-08-28 21:01:34,528 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.29 sec
MapReduce Total cumulative CPU time: 2 seconds 290 msec
Ended Job = job_1503918153486_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.29 sec   HDFS Read: 332 HDFS Write: 41 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 290 msec
OK
3       Yadav   300     DNA     200
6       Mahoor  200     FCS     300
Time taken: 19.429 seconds, Fetched: 2 row(s)
hive>
```

Step 5: List of all employees who draw higher salary than the average salary of that department:

SELECT id, name, salary, unit, avg_salary FROM
(
SELECT avg(salary) OVER (PARTITION BY unit) AS avg_salary, id, name, salary, unit
FROM employee
) temp
WHERE salary > avg_salary;

```
hive> SELECT id, name, salary, unit, avg_salary
    > FROM
    > (
    > SELECT avg(salary) OVER (PARTITION BY unit) AS avg_salary, id, name, salary, unit
    > FROM employee
    > ) temp
    > WHERE salary > avg_salary;
Query ID = acadgild_20170828210505_952a7c14-dbb2-4657-90ca-aad0f7808504
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1503918153486_0023, Tracking URL = http://localhost:8088/proxy/applicat
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1503918153486_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-08-28 21:05:32,532 Stage-1 map = 0%,  reduce = 0%
2017-08-28 21:05:38,793 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.64 sec
2017-08-28 21:05:45,056 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.16 sec
MapReduce Total cumulative CPU time: 2 seconds 160 msec
Ended Job = job_1503918153486_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.16 sec   HDFS Read: 332 HDFS Write: 8
Total MapReduce CPU Time Spent: 2 seconds 160 msec
OK
8       Chandra 500     DNA     275.0
3       Yadav   300     DNA     275.0
4       Sunil   500     FCS     266.6666666666667
Time taken: 19.234 seconds, Fetched: 3 row(s)
hive>
```