

Project 1:

Problem description:

The wages of the person should be updated with the current wage on daily execution of the csv file of employees using Hive.

Datasets:

- 1) The dataset resides on the linux file system and the path is
/home/dell/data/hive_data/datasets/
 - a) Data 1: emp_data_20230901.txt
 - b) Data 2: emp_data_20230902.txt

Main files for execution:

- 1) Hive query language file & shell script file resides on /home/dell/bin/
 - a) Shell script file name (Main file): daily_script.sh - This is the file responsible to execute .hql file.
 - b) HQL file name: scd.hql

a) do a select * from emp_master; [data should match with the code base project output]

- 1) > sh daily_script.sh 20230901

Output:

```
hive> select * from emp_master;
OK
Albert  1200    2023-09-01
Frank   1150    2023-09-01
John    1300    2023-09-01
Lesa    900     2023-09-01
Loopa   1100    2023-09-01
Lui     1300    2023-09-01
Mark    1000    2023-09-01
Pars    800     2023-09-01
fransis 1000    2023-09-01
jack    700     2023-09-01
leo     700     2023-09-01
lock    650     2023-09-01
pars    900     2023-09-01
Time taken: 1.934 seconds, Fetched: 13 row(s)
hive>
```

Can see all the records of employees on 2023-09-01, where there is no change in wages.

2) >sh daily_script.sh 20230902

Output:

```
hive> select * from emp_master;
OK
fransis 1000      2023-09-01
jack     700      2023-09-01
pars     900      2023-09-01
Albert   1900      2023-09-02
Bhut     800      2023-09-02
Frank    1150      2023-09-02
John     1500      2023-09-02
Lesa     900      2023-09-02
Lio      500      2023-09-02
Loopa    1100      2023-09-02
Lui      1300      2023-09-02
Mark     1000      2023-09-02
Pars     800      2023-09-02
leo      700      2023-09-02
lock     650      2023-09-02
Time taken: 0.226 seconds, Fetched: 15 row(s)
hive>
```

Inference: Here, after executing the file “emp_data_20230902.txt”, the existing employee named Albert & John has increase in their wages; lock, leo, Mark, Lui, Loopa, Lesa, Frank & Pars has maintained their wages; where as fransis, jack & pars have not worked on 2023-09-02.

b) describe formatted emp_master; [this should display your user id of table in the owner]

3) hive >describe formatted emp_master;

Output:

```
dell@cluster-ce16-m: ~/bin x dell@cluster-ce16-m: ~
hive> describe formatted emp_master;
OK
# col_name      data_type      comment
emp_name        string
sal             int

# Partition Information
# col_name      data_type      comment
txn_data        date

# Detailed Table Information
Database:        incr_load
OwnerType:       USER
Owner:           dell
CreateTime:      Wed Sep 27 23:24:39 UTC 2023
LastAccessTime:  UNKNOWN
Retention:       0
Location:        hdfs://cluster-ce16-m/user/hive/warehouse/incr_load.db/emp_master
Table Type:      EXTERNAL_TABLE
Table Parameters:
  COLUMN_STATS_ACCURATE {\"BASIC_STATS\": \"true\"}
  EXTERNAL              TRUE
  bucketing_version     2
  numFiles              2
  numPartitions         2
  numRows              15
  rawDataSize          131
  totalSize             146
  transient_lastDdlTime 1695857079

# Storage Information
SerDe Library:    org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:      org.apache.hadoop.mapred.TextInputFormat
OutputFormat:     org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:       No
Num Buckets:      -1
```