

# Hive Project

## Hive Incremental Data Load

**Objective:** Write a shell script for the daily updating of a table that contains current wage data of employees.

**Given Data files:** emp\_data\_20230901.txt, emp\_data\_20230902.txt, emp\_data\_20230903.txt,

**Created execution files:**

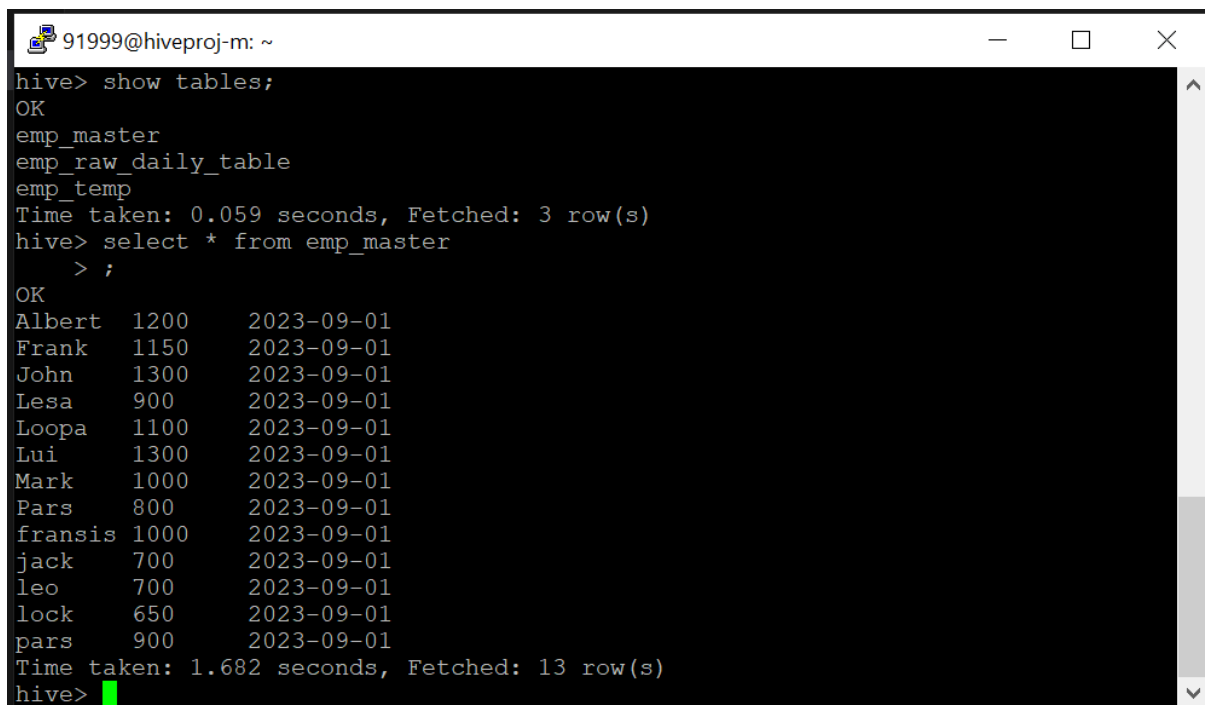
- **scd.hql** [consists of hive ql statements]
- **daily\_script.sh** [shell script required to execute the hql file]

**Steps:**

1. Loaded the 3 data files and scd.hql file from the local machine to the GCP Bucket.
2. Created bin and data folders inside the cluster VM which is a Linux machine.
3. Used Vim to write daily\_script.sh into the bin folder and copied scd.hql from GCP bucket to GCP cluster VM.
4. Copied the 3 data files from the GCP Bucket into the GCP Cluster VM.
5. Used the command **chmod -x daily\_script.sh** to give execution permissions.
6. Used the command **sh daily\_script.sh 20230901** to perform the data load.

**Results:**

### 1. sh daily\_script.sh 20230901



```
91999@hiveproj-m: ~
hive> show tables;
OK
emp_master
emp_raw_daily_table
emp_temp
Time taken: 0.059 seconds, Fetched: 3 row(s)
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: 1.682 seconds, Fetched: 13 row(s)
hive>
```

## 2. sh daily\_script.sh 20230902

```
91999@hiveproj-m: ~  
jack      700      2023-09-01  
leo       700      2023-09-01  
lock      650      2023-09-01  
pars      900      2023-09-01  
Time taken: 1.682 seconds, Fetched: 13 row(s)  
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: 0.296 seconds, Fetched: 15 row(s)  
hive> █
```

## 3. sh daily\_script.sh 20230903

```
91999@hiveproj-m: ~  
Mark      1000      2023-09-02  
Pars      800      2023-09-02  
leo       700      2023-09-02  
lock      650      2023-09-02  
Time taken: 0.296 seconds, Fetched: 15 row(s)  
hive> select * from emp_master;  
OK  
fransis 1000      2023-09-01  
jack     700      2023-09-01  
pars     900      2023-09-01  
Frank    1150      2023-09-02  
John     1500      2023-09-02  
Lesa     900      2023-09-02  
Loopa    1100      2023-09-02  
Lui      1300      2023-09-02  
Mark     1000      2023-09-02  
Pars     800      2023-09-02  
Albert   22100      2023-09-03  
Bhut     1800      2023-09-03  
Lio      1500      2023-09-03  
leo      700      2023-09-03  
lock     1650      2023-09-03  
Time taken: 0.159 seconds, Fetched: 15 row(s)  
hive> █
```

#### 4. describe formatted emp\_master;

91999@hiveproj-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:           91999
CreateTime:      Sun Nov 26 07:08:03 UTC 2023
LastAccessTime:  UNKNOWN
Retention:       0
Location:        hdfs://hiveproj-m/user/hive/warehouse/incr_load.db/emp_m
aster
Table Type:      EXTERNAL_TABLE
Table Parameters:
    COLUMN_STATS_ACCURATE  {"BASIC_STATS": "true"}
    EXTERNAL               TRUE
    bucketing_version      2
    numFiles                3
    numPartitions          3
    numRows                15
    rawDataSize            135
    totalSize              150
    transient_lastDdlTime  1700982483

# Storage Information
SerDe Library:    org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:      org.apache.hadoop.mapred.TextInputFormat
OutputFormat:     org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputForm
at
Compressed:       No
Num Buckets:      -1
Bucket Columns:   []
Sort Columns:     []
Storage Desc Params:
    field.delim        ,
    line.delim         \n
    serialization.format
```