Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)
Department of Computer Science and Engineering (Data Science)

Subject: Big Data Engineering (DJ19DSL604)

AY: 2022-23

Experiment 6

(Data Warehouse)

Name Sarvagya Singh
SAPID : 60009200030
BATCH : K1

Aim: Implement data warehousing using HIVE.

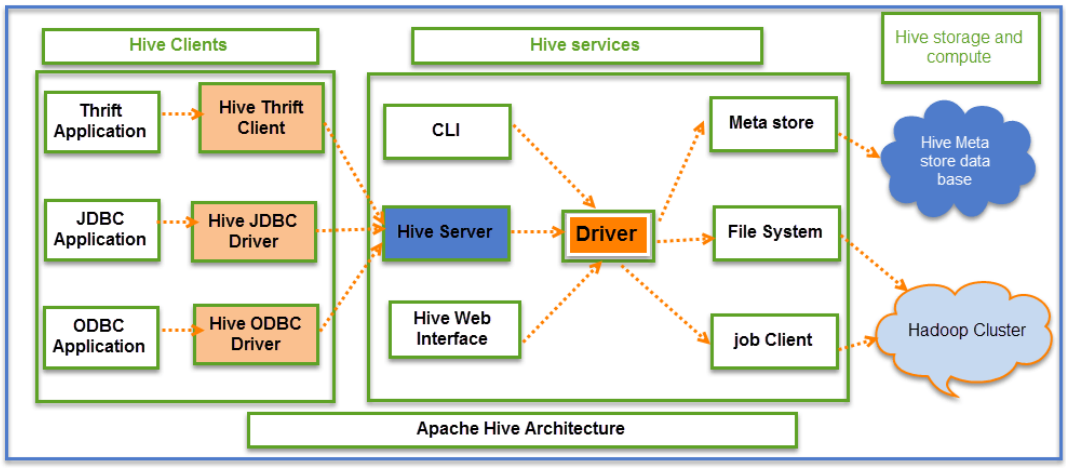
Theory:

Introduction to HIVE

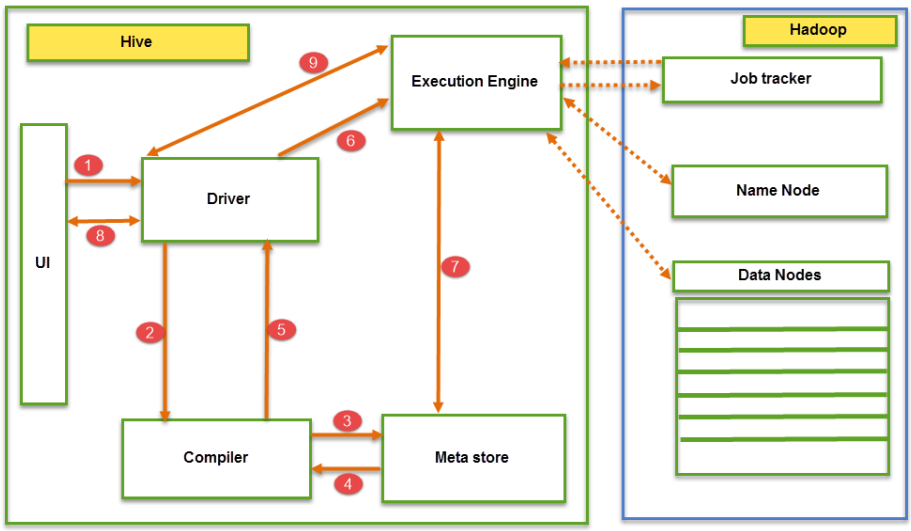
Hive as an ETL and data warehousing tool on top of Hadoop ecosystem provides functionalities like Data modeling, Data manipulation, Data processing and Data querying. Data Extraction in Hive means the creation of tables in Hive and loading structured and semi structured data as well as querying data based on the requirements.

For batch processing, we are going to write custom defined scripts using a custom map and reduce scripts using a scripting language. It provides SQL like environment and support for easy querying.

HIVE Architecture



Job execution flow:



Different modes of Hive:

Hive can operate in two modes depending on the size of data nodes in Hadoop.

These modes are,

- **Local mode**
- **Map reduce mode**

When to use Local mode:

- If the Hadoop installed under pseudo mode with having one data node we use Hive in this mode
- If the data size is smaller in term of limited to single local machine, we can use this mode
- Processing will be very fast on smaller data sets present in the local machine.

When to use Map reduce mode:

- If Hadoop is having multiple data nodes and data is distributed across different node we use Hive in this mode
- It will perform on large amount of data sets and query going to execute in parallel way
- Processing of large data sets with better performance can be achieved through this mode

Lab Assignment:

1. Installation of HIVE.
2. Implement the following SQL queries in HIVE on any database:
    a. Create Database
    b. Order by Query
    c. Group by Query
    d. Sort By
    e. Cluster By
    f. Distribute By
3. Working with HIVE ETL:
    a. Structured Data using Hive.
    b. Semi structured data using Hive (XML, JSON).

```
hadoop@vallabh-virtual-machine:~/apache-hive-3.1.2-bin$ cd bin
hadoop@vallabh-virtual-machine:~/apache-hive-3.1.2-bin/bin$ ls
beeline  hive            hiveserver2  init-hive-dfs.sh  schematool
ext      hive-config.sh  hplsql       metatool
hadoop@vallabh-virtual-machine:~/apache-hive-3.1.2-bin/bin$
```

```
hadoop@vallabh-virtual-machine:~$ tar -xzf apache-hive-3.1.2-bin.tar.gz
hadoop@vallabh-virtual-machine:~$ hdfs dfs -mkdir -p /user/hive/warehouse
hadoop@vallabh-virtual-machine:~$ hdfs dfs -mkdir /tmp
hadoop@vallabh-virtual-machine:~$ hdfs dfs -chmod g+w /user/hive/warehouse
hadoop@vallabh-virtual-machine:~$ hdfs dfs -chmod g+w /tmp
```

```
hive> create database test;
OK
Time taken: 0.2 seconds
hive> use test;
OK
Time taken: 0.141 seconds
hive> show tables;
OK
values__tmp__table__3
Time taken: 0.186 seconds, Fetched: 1 row(s)
```

```
hive> create table test.emp
    > (
    > sno int,
    > usr_name string,
    > city string)
    > ROW FORMAT delimited fields terminated by ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
OK
Time taken: 0.77 seconds
hive> show tables;
OK
emp
values__tmp__table__3
Time taken: 0.111 seconds, Fetched: 2 row(s)
hive> load data local inpath '/home/test/usr_data.txt' into table emp;
Loading data to table test.emp
OK
Time taken: 3.15 seconds
```

```
hive> select * from emp;
OK
1       gowtham chennai
2       saravana        chennai
3       ram     delhi
4       alex    mumbai
5       rahul   delhi
6       arun    goa
7       nila    chennai
8       nandini chennai
9       anita   delhi
10      jaya    delhi
Time taken: 0.943 seconds, Fetched: 10 row(s)
hive>
```

```
hive> insert into emp (sno,usr_name,city) values (1,'g','chennai');
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive
 1.X releases.
Query ID = test_20210712200241_a543353e-ddc1-47d0-a13b-4cbb134815e1
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1626090715853_0004, Tracking URL = http://localhost:8088/proxy/application_1626090715853_0004/
Kill Command = /home/test/hadoop-2.9.1/bin/hadoop job  -kill job_1626090715853_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2021-07-12 20:03:20,124 Stage-1 map = 0%,  reduce = 0%
2021-07-12 20:03:50,834 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 18.76 sec
MapReduce Total cumulative CPU time: 18 seconds 760 msec
Ended Job = job_1626090715853_0004
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:50000/user/hive/warehouse/test.db/emp/.hive-staging_hive_2021-07-12_20-02-41_029_2861965364404090998-1/-ext-10000
Loading data to table test.emp
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 18.76 sec   HDFS Read: 4362 HDFS Write: 76 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 760 msec
OK
Time taken: 75.824 seconds
```

```
hive> desc extended emp;
OK
sno                     int
usr_name                string
city                    string

Detailed Table Information      Table(tableName:emp, dbName:test, owner:test, createTime:1626099881, lastAccessTime:0, retention:0, sd:StorageDescriptor(cols:[FieldSche
ma(name:sno, type:int, comment:null), FieldSchema(name:usr_name, type:string, comment:null), FieldSchema(name:city, type:string, comment:null)], location:hdfs://localho
st:50000/user/hive/warehouse/test.db/emp, inputFormat:org.apache.hadoop.mapred.TextInputFormat, outputFormat:org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat,
 compressed:false, numBuckets:-1, serdeInfo:SerDeInfo(name:null, serializationLib:org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe, parameters:{serialization.format=,
, line.delim=
, field.delim=,}), bucketCols:[], sortCols:[], parameters:{}, skewedInfo:SkewedInfo(skewedColNames:[], skewedColValues:[], skewedColValueLocationMaps:{}), storedAsSubDi
rectories:false), partitionKeys:[], parameters:{transient_lastDdlTime=1626100436, totalSize=459, numRows=0, rawDataSize=0, numFiles=4}, viewOriginalText:null, viewExpan
dedText:null, tableType:MANAGED_TABLE, rewriteEnabled:false)
Time taken: 0.285 seconds, Fetched: 6 row(s)
```

```
hive> select * from emp where city in ('chennai');
OK
1       g       chennai
1       gowtham chennai
2       saravana        chennai
7       nila    chennai
8       nandini chennai
1       gowtham chennai
2       saravana        chennai
7       nila    chennai
8       nandini chennai
1       gowtham chennai
2       saravana        chennai
7       nila    chennai
8       nandini chennai
Time taken: 1.216 seconds, Fetched: 13 row(s)
```