# WQD7007 Big Data Management

## Hive

# Agenda

- Hive - easy data extraction, transformation and loading (ETL).

# Hive

- Hive is an **SQL like query language** that enables those analysts familiar with SQL to run queries on **large** volumes of data.
  - a dialect of SQL (Hive QL) that focuses on analytics and presents a rich set of SQL semantics e.g. OLAP functions, sub-queries, and common table expressions.
  - has three main functions:
    - data summarization,
    - query, and
    - analysis.
- Data analysts use Hive to explore, structure and analyze that data, then turn it into **business insights**

# Hive

- Hive also allows programmers familiar with the MapReduce framework to plug in their **custom mappers and reducers** to perform more sophisticated analysis that may not be supported by the built-in capabilities of the language.
- Hive users have a choice of 3 runtimes when executing SQL queries. Users can choose between Apache Hadoop MapReduce, Apache Tez or Apache Spark frameworks as their execution backend.

# Hive

| FEATURE | DESCRIPTION |
| --- | --- |
| **Familiar** | Query data with a SQL-based language |
| **Fast** | Interactive response times, even over huge datasets |
| **Scalable and Extensible** | As data variety and volume grows, more commodity machines can be added, without a corresponding reduction in performance |

# Hive

- **The tables in Hive are similar to tables in a relational database**, and data units are organized in a taxonomy from larger to more granular units.
- Databases are comprised of tables, which are made up of partitions.
  - Within a particular database, data in the tables is serialized and each table has a corresponding Hadoop Distributed File System (HDFS) directory.
  - Each table can be sub-divided into **partitions** that determine how data is distributed within sub-directories of the table directory. Data within partitions can be further broken down into buckets.

# Hive

- **HCatalog** is a component of Hive.
  - a **table and storage management layer** for Hadoop that enables users with different data processing tools (including Pig and MapReduce)
  - more easily read and write data on the grid.
  - It holds a set of files paths and metadata about data in a Hadoop cluster.
    - This allows scripts, MapReduce and Tez, jobs to be decoupled from data location and metadata like the schema.
    - Additionally, since HCatalog also supports tools like Hive and Pig, the location and metadata can be shared between tools.

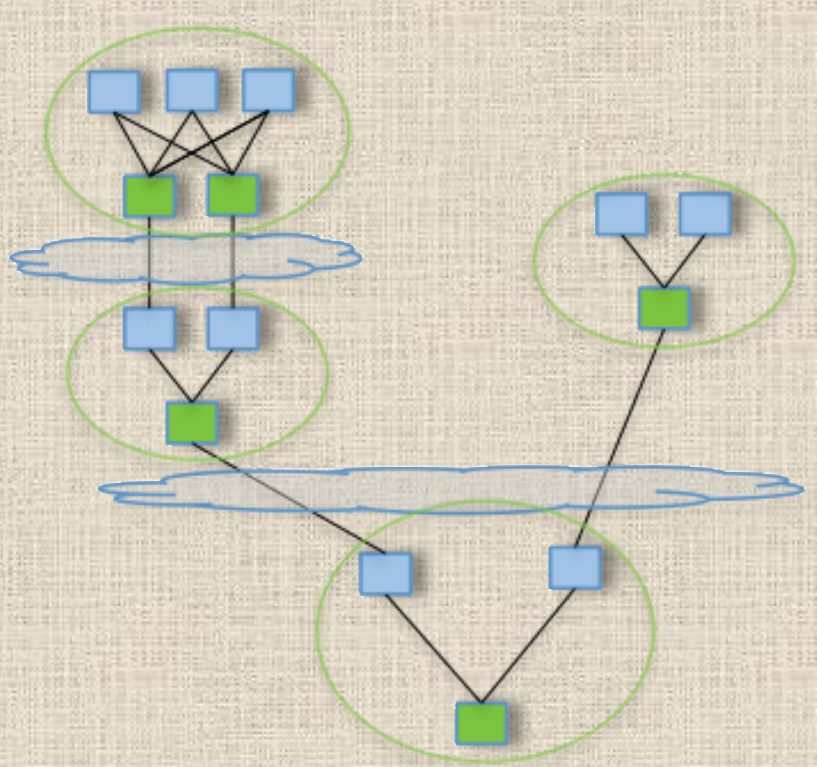
# Hive

- **WebHCat** provides a service that you can use to run Hadoop MapReduce (or YARN), Pig, Hive jobs or perform Hive metadata operations using an HTTP (REST style) interface.

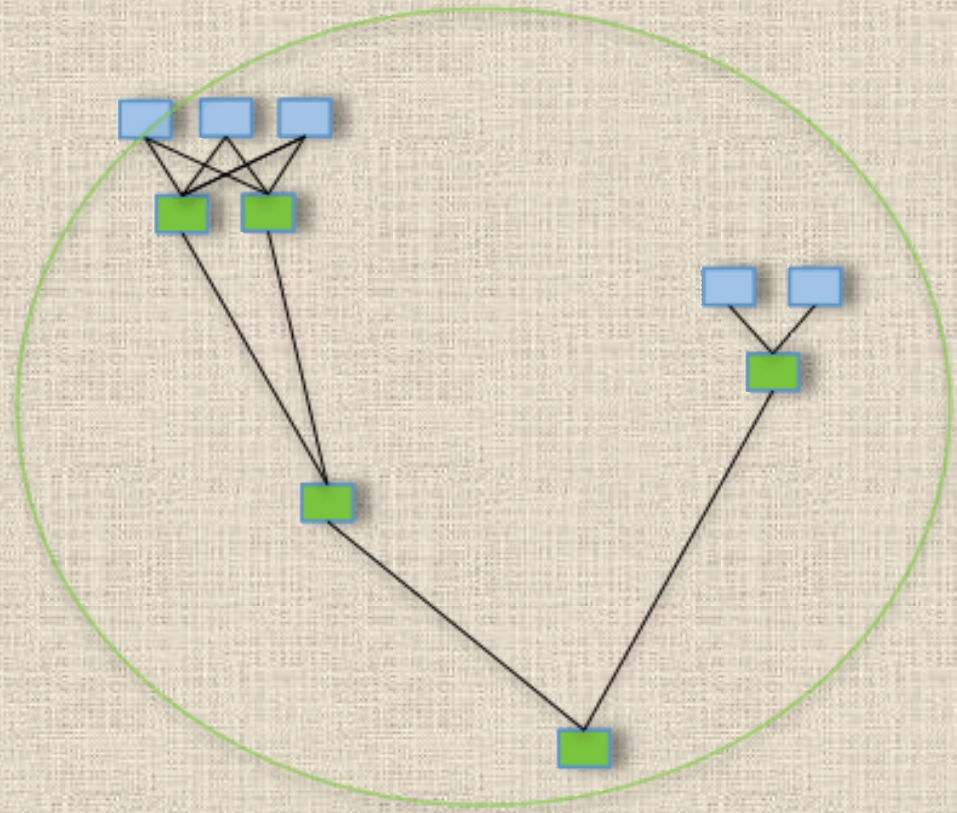
# Apache Tez

- Apache Tez is an **extensible framework** for building high performance batch and interactive data processing applications, coordinated by YARN in Apache Hadoop.
  - Tez **improves the MapReduce paradigm by dramatically improving its speed**, while maintaining MapReduce's ability to scale to petabytes of data.
  - Important Hadoop ecosystem projects like Apache Hive and Apache Pig use Apache Tez, as do a growing number of third party data access applications developed for the broader Hadoop ecosystem.

# Hive without and with Tez



Pig/Hive - MR



Pig/Hive - Tez

# Online reference

- http://www.javachain.com/hive-crud-operation
- http://www.informit.com/articles/article.aspx?p=2756471&seqNum=4

# Create table from csv file

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS Names_text(
    > EmployeeID INT,FirstName STRING, Title STRING,
    > State STRING, Laptop STRING)
    > COMMENT 'Employee Names'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > LOCATION '/user/hdfs/names';
OK
Time taken: 8.001 seconds
hive> Select * from Names_text limit 5;
OK
10      Andrew  Manager DE      PC
11      Arun    Manager NJ      PC
12      Harish  Sales   NJ      MAC
13      Robert  Manager PA      MAC
14      Laura   Engineer        PA      MAC
Time taken: 2.64 seconds, Fetched: 5 row(s)
```

# Create table manually

```
hive> CREATE TABLE STUDENT
    > (
    > STD_ID INT,
    > STD_NAME STRING,
    > AGE INT,
    > ADDRESS STRING
    > )
    > CLUSTERED BY (ADDRESS) into 3 buckets
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED as orc tblproperties('transactional'='true');
OK
Time taken: 1.349 seconds
hive> INSERT INTO TABLE STUDENT VALUES (101,'JAVACHAIN',30,'PAUL REVERE RD'),
    > (102,'ANTO',18,'29 NATHAN HALE'),
    > (103,'PRABU',23,'34 henry road'),
    > (104,'KUMAR',24,'gandhi road'),
    > (105,'jack',35,'Modi street');
Query ID = hive_20180417032150_92248d47-703a-4b62-8851-bb949749e2b3
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1523892949440_0007)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 21.88 s
--------------------------------------------------------------------------------
Loading data to table default.student
Table default.student stats: [numFiles=1, numRows=0, totalSize=1003, rawDataSize=0]
OK
Time taken: 57.745 seconds
hive>
```

# Select entry

```
hive> Select * from student;
OK
101     JAVACHAIN       30      PAUL REVERE RD
102     ANTO    18      29 NATHAN HALE
103     PRABU   23      34 henry road
104     KUMAR   24      gandhi road
105     jack    35      Modi street
Time taken: 0.279 seconds, Fetched: 5 row(s)
```

# Update entry

```
hive> update STUDENT
    > SET std_id = 110
    > WHERE std_id = 105;
Query ID = hive_20180417033441_cf9933ed-d560-4cc9-8412-15ed3410294b
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1523892949440_0007)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.11 s
--------------------------------------------------------------------------------
Loading data to table default.student
Table default.student stats: [numFiles=2, numRows=0, totalSize=1857, rawDataSize=0]
OK
Time taken: 14.657 seconds
hive> Select * from student;
OK
101     JAVACHAIN       30      PAUL REVERE RD
102     ANTO    18      29 NATHAN HALE
103     PRABU   23      34 henry road
104     KUMAR   24      gandhi road
110     jack    35      Modi street
Time taken: 0.208 seconds, Fetched: 5 row(s)
```

# Delete entry

```
hive> DELETE FROM STUDENT
    > where std_id=104;
Query ID = hive_20180417033734_93ea0bdb-f23f-46ea-8b40-f9a23564662d
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1523892949440_0007)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.49 s
--------------------------------------------------------------------------------
Loading data to table default.student
Table default.student stats: [numFiles=3, numRows=0, totalSize=2397, rawDataSize=0]
OK
Time taken: 13.619 seconds
hive> Select * from student;
OK
101     JAVACHAIN       30      PAUL REVERE RD
102     ANTO    18      29 NATHAN HALE
103     PRABU   23      34 henry road
110     jack    35      Modi street
Time taken: 0.194 seconds, Fetched: 4 row(s)
```

# Concluding Remarks

- Hive – SQL-like scripting language for fast processing