HIVE CHEAT SHEET

Hive Basics

Apache Hive

It is a data warehouse infrastructure based on Hadoop framework which is perfectly suitable for data summarization, analysis and querying. It uses an SQL like language called HQL (Hive query Language)

HQL: It is a query language used to write the custom map reduce framework in Hive to perform more sophisticated analysis of the data

Table: Table in hive is a table which contains logically stored data

Hive Interfaces:

- Hive interfaces includes WEB UI
- Hive command line
- HD insight (windows server)

Components of Hive

Meta store: Meta store is where the schemas of the Hive tables are stored, it stores the information about the tables and partitions that are in the warehouse.

SerDe: Serializer, Deserializer which gives instructions to hive on how to process records

Thrift

A thrift service is used to provide remote access from other processors

Meta Store

This is a service which stores the metadata information such as table schemas

Indexes

Indexes are created to the speedy access to columns in the database

Syntax: Create index <INDEX_NAME> on table <TABLE_NAME>

Hive Function Meta Commands

Show functions: Lists Hive functions and operators

Describe function [function name]: Displays short description of the particular function

Describe function extended [function name]: Displays extended description of the particular function

Hive Functions

- UDF(User defined Functions): It is a function that fetches one or more columns from a row as arguments and returns a single value
- UDTF(User defined Tabular Functions): This function is used to produce multiple columns or rows of output by taking zero or more inputs
- Macros: It is a function that uses other Hive functions
- User defined aggregate functions: A user defined function that takes multiple rows or columns and returns the aggregation of the data
- User defined table generating functions: A function which takes a column from single record and splitting it into multiple rows

Hive SELECT Command

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
FROM table_reference
[WHERE where_condition]
[GROUP BY col_list]
[HAVING having_condition]
[CLUSTER BY col_list | [DISTRIBUTE BY col_list] [SORT BY col_list]]
[LIMIT number]
;
```

- Select: Select is a projection operator in HiveQL, which scans the table specified by the FROM clause
- Where: Where is a condition which specifies what to filter
- Group by: It uses the list of columns, which specifies how to aggregate the records
- Cluster by, Distribute by, Sort by: Specifies the algorithm to sort, distribute and create cluster, and the order for sorting
- Limit: This specifies how many records to be retrieved

Hive Data Types

Integral data types:

- Tinyint
- Smallint
- Int
- Bigint

String types:

- VARCHAR-Length(1 to 65355)
- CHAR-Length(255)

Union type: It is a collection of heterogenous data types.

- Syntax: UNIONTYPE<int, double, array<string>, struct<a:int,b:string>>

Timestamp: It supports the traditional Unix timestamp with optional nanosecond precision

- Dates
- Decimals

Complex types:

- Arrays: Syntax-ARRAY<data_type>
- Maps: Syntax- MAP<primitive_type, data_type>
- Structs: STRUCT<col_name : data_type [COMMENT col_comment], ...>

Bucketing

It is a technique to decompose the datasets into more manageable parts

Partitioner

Partitioner controls the partitioning of keys of the intermediate map outputs, typically by a hash function which is same as the number of reduce tasks for a job

- Partitioning: It is used for distributing load horizontally. It is a way of dividing the tables into related parts based on values such as date, city, departments etc.

Hcatalog

It is a metadata and table management system for Hadoop platform which enables storage of data in any format.

Hive commands in HQL

Data Definition Language(DDL): It is used to build or modify tables and objects stored in a database. Some of the DDL commands are as follows:

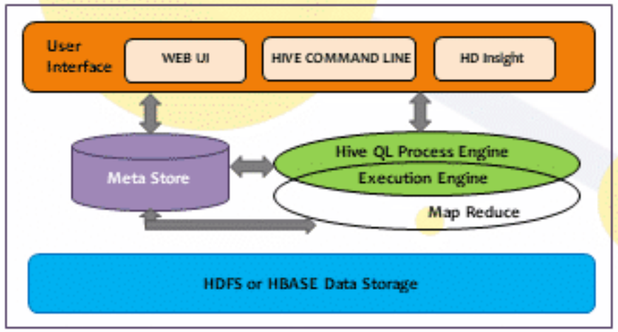
- To create database in Hive: create database<data base name>
- To list out the databases created in a Hive warehouse: show databases
- To use the database created: USE <data base name>
- To describe the associated database in metadata: describe<data base name>
- To alter the database created: alter<data base name>

Data Manipulation Language(DML): These statements are used to retrieve, store, modify, delete, insert and update data in a database

- Inserting data in a database: The Load function is used to move the data into a particular Hive table.

  LOAD data <LOCAL> inpath <file path> into table [tablename]
- Drop table: The drop table statements deletes the data and metadata from the table: drop table<table name>
- Aggregation: It is used to count different categories from the table :

  Select count (DISTINCT category) from tablename;
- Grouping: Group command is used to group the result set, where the result of one table is stored in the other: Select <category>, sum(amount) from <txt records> group by <category>
- To exit from the Hive shell: Use the command quit



Operations - Performed on Hive

Function	HQL Query
To retrieve information	SELECT from_columns FROM table WHERE conditions;
To select all values	SELECT * FROM table;
To select a particular category values	SELECT * FROM table WHERE rec_name = "value";
To select for multiple criteria	SELECT * FROM TABLE WHERE rec1 = "value1" AND rec2 = "value2";
For selecting specific columns	SELECT column_name FROM table;
To retrieve unique output records	SELECT DISTINCT column_name FROM table;
For sorting	SELECT col1, col2 FROM table ORDER BY col2;
For sorting backwards	SELECT col1, col2 FROM table ORDER BY col2 DESC;
For counting rows from the table	SELECT COUNT(*) FROM table;
For grouping along with counting	SELECT owner, COUNT(*) FROM table GROUP BY owner;
For selecting maximum values	SELECT owner, COUNT(*) FROM table GROUP BY owner;
Selecting from multiple tables and joining	SELECT pet.name, comment FROM pet JOIN event ON (pet.name = event.name);

Command Line Statements

Function	Hive Commands
To run the query	hive -e 'select a.col from tab1 a'
To run a query in a silent mode	hive -S -e 'select a.col from tab1 a'
To select hive configuration variables	hive -e 'select a.col from tab1 a' –hiveconf hive.root.logger=DEBUG,console
To use the initialization script	hive -i initialize.sql
To run the non-interactive script	hive -f script.sql
To run script inside the shell	source file_name
To run the list command	dfs –ls /user
To run ls (bash command) from the shell	!ls
To set configuration variables	set mapred.reduce.tasks=32
Tab auto completion	set hive.<TAB>
To display all variables starting with hive	set
To revert all variables	reset
To add jar files to distributed cache	add jar jar_path
To display all the jars in the distributed cache	list jars
To delete jars from the distributed cache	delete jar jar_name

Metadata Functions and Query

Function	Hive Commands
Selecting a database	USE database;
Listing databases	SHOW DATABASES;
listing table in a database	SHOW TABLES;
Describing format of a table	DESCRIBE (FORMATTED|EXTENDED) table;
Creating a database	CREATE DATABASE db_name;
Dropping a database	DROP DATABASE db_name (CASCADE);