Apache Hive

Table des matières

[Introduction 3](#_Toc100650268)

[1. C’est quoi Apache Hive : 3](#_Toc100650269)

[2. Apache Hive: un peu d’histoire 3](#_Toc100650270)

[3. Terminologie 3](#_Toc100650271)

[a) Database: 3](#_Toc100650272)

[b) Table: TDB 3](#_Toc100650273)

[c) Partition: 3](#_Toc100650274)

[d) Bucket/Cluster: 3](#_Toc100650275)

[4. Types de données 3](#_Toc100650276)

[a) Primitifs : 3](#_Toc100650277)

[b) Complexes : 3](#_Toc100650278)

[5. HiveQL, SQL pour Hive 4](#_Toc100650279)

[• DDL : 4](#_Toc100650280)

[• DML : 4](#_Toc100650281)

[• Fonction utilisateurs (UDF) : (User Defined Function) : 4](#_Toc100650282)

[6. Table interne vs table externe : 4](#_Toc100650283)

[a) Les tables internes : 4](#_Toc100650284)

[b) Tables externes : 4](#_Toc100650285)

[7. Hive : partitionnement des tables : 4](#_Toc100650286)

[a) Création de tables : 5](#_Toc100650287)

[b) Ajouter une partition : 5](#_Toc100650288)

[c) Renommer une partition : TDB 5](#_Toc100650289)

[d) Supprimer une partition : TDB 5](#_Toc100650290)

[Avantage du partitionnement : 5](#_Toc100650291)

[Inconvenient du partitionnement: 5](#_Toc100650292)

[8. Hive: Bucketing c’est quoi ? 5](#_Toc100650293)

[a) La répartition des données sur les bucket se fait : TBD 5](#_Toc100650294)

[b) Création de tables buckettées : TBD 5](#_Toc100650295)

[c) Arborescence HDFS d’une table buckettée :TBD 5](#_Toc100650296)

[9. Apache Hive: Architecture 5](#_Toc100650297)

[10. Formats de fichier supportés par Hive ? 5](#_Toc100650298)

# Introduction

# C’est quoi Apache Hive :

Hive est une couche installée sur HDFS(C’est le data warehouse sur HDFS).

Son avantage est qu’on n’a pas à maîtriser un langage de programmation pour le manipuler (utiliser).

Hive n’est pas un SGBD

Hive on l’utilise uniquement en mode batch. Mais jamais en mode streaming (temps réel)

Hive génère des jobs Tez = map reduce (dans le fond)

Tez plus performant que Map Reduce car c’est lui le moteur d’exécution.

# Apache Hive: un peu d’histoire

Hive est utilisé la première fois par Facebook dans les années 2000. Par la suite a été développé et mis à la commercialisation fine des années 2000.

# Terminologie

### Database:

Les métastores (métadonnées des tables) et les tables (la donnée)

### Table: TDB

### Partition:

La partition consiste à créer des dossiers (répertoires) selon les différentes variées d’une colonne qu’on choisie

### Bucket/Cluster:

c’est une sorte de subdivision de fichier en plusieurs répertoire grâce à la fonction de hashage. (echantillonnage sur une clé)

# Types de données

### Primitifs :

integer, float, char, etc…

### Complexes :

map (clé, valeurs), structures, arraylist etc..

# HiveQL, SQL pour Hive

HiveQL supporte

### DDL :

Data Definition Language: create, alter, drop

### DML :

Langage de manipulation : load et insert

Dans Hive update & delete non supportés (principe même de HDFS)

NB : truncate est plus rapide que delete car truncate déclare directement la table

### Fonction utilisateurs (UDF) : (User Defined Function) :

On peut créer des fonctions qu’on veut utiliser

# Table interne vs table externe :

TDB

### Les tables internes :

Les tables internes sont entièrement mangées par Hive (métadonnées et les données) dans le répertoire : /apps/hive/warehouse. Donc la suppression de la table entraine la suppression de la donnée et de la métadonnée

### Tables externes :

Sont dans HDFS, elles ne sont pas gérées par hive donc leurs suppressions dans hive n’entrainent pas la suppression dans hdfs des données mais uniquement la suppression des métadonnées dans hive.

# Hive : partitionnement des tables :

### Création de tables :

C’est le mécanisme qui consiste à créer des tables à partir des fichiers de différentes extensions.

### Ajouter une partition :

Une partition se fait en choisissant une clé (attribut ou colonne de la table qu’on veut partitionner) de partition

Exemple : Souvent les colonnes date, année sont de bonnes clés pour partitionner, le choix de la clé qu’on utilise

### Renommer une partition : TDB

### Supprimer une partition : TDB

## Avantage du partitionnement :

Distribution des traitements.

Partition (réflexion)

## Inconvénient du partitionnement :

Group by à éviter dans Hive.car c’est long

# Hive: Bucketing c’est quoi ?

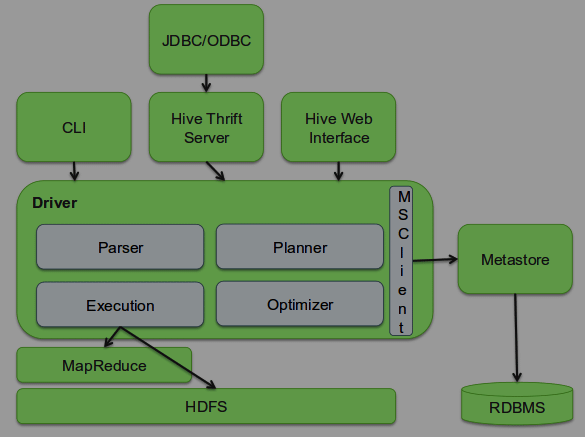
TBD

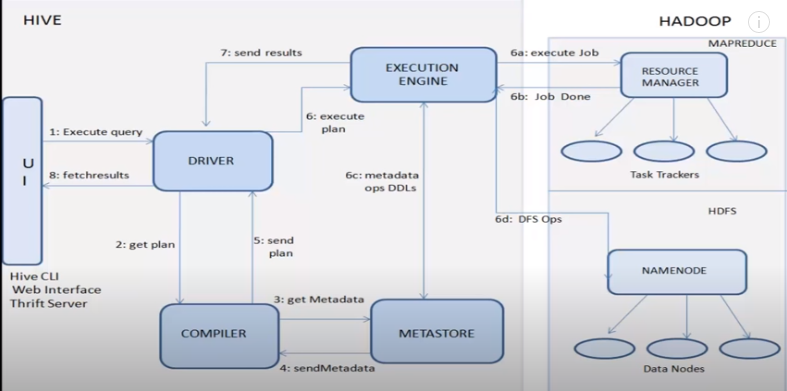
### La répartition des données sur les bucket se fait : TBD

### Création de tables buckettées : TBD

### Arborescence HDFS d’une table buckettée :TBD

# Apache Hive: Architecture





# Formats de fichier supportés par Hive ?

ORC File

Text file

RC file

Parquet file

Sequence file