Apache Hive

Table des matières

[Introduction 3](#_Toc100593631)

[1. C’est quoi Apache Hive 3](#_Toc100593632)

[2. Apache Hive: un peu d’histoire 3](#_Toc100593633)

[3. Terminologie 3](#_Toc100593634)

[a) Database: 3](#_Toc100593635)

[b) Table: TDB 3](#_Toc100593636)

[c) Partition: TDB 3](#_Toc100593637)

[d) Bucket/Cluster: TDB 3](#_Toc100593638)

[4. Types de données 3](#_Toc100593639)

[a) Primitifs : TDB 3](#_Toc100593640)

[b) Complexes : TDB 3](#_Toc100593641)

[5. HiveQL, SQL pour Hive 3](#_Toc100593642)

[• DDL : TDB 3](#_Toc100593643)

[• DML : TDB 3](#_Toc100593644)

[• Fonction utilisateurs (UDF) : TDB 3](#_Toc100593645)

[6. Table interne vs table externe : 4](#_Toc100593646)

[a) Les tables internes : TDB 4](#_Toc100593647)

[b) Tables externes : TDB 4](#_Toc100593648)

[7. Hive: partitionnement des tables : 4](#_Toc100593649)

[a) Création de tables : TDB 4](#_Toc100593650)

[b) Ajouter une partition : TDB 4](#_Toc100593651)

[c) Renommer une partition : TDB 4](#_Toc100593652)

[d) Supprimer une partition : TDB 4](#_Toc100593653)

[Avantage du partitionnement : 4](#_Toc100593654)

[Inconvenient du partitionnement: 4](#_Toc100593655)

[8. Hive: Bucketing c’est quoi ? 4](#_Toc100593656)

[a) La répartition des données sur les bucket se fait : TBD 4](#_Toc100593657)

[b) Création de tables buckettées : TBD 4](#_Toc100593658)

[c) Arborescence HDFS d’une table buckettée :TBD 4](#_Toc100593659)

[9. Apache Hive: Architecture 4](#_Toc100593660)

[10. Formats de fichier supportés par Hive ? 4](#_Toc100593661)

# 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:

### Table: TDB

### Partition: TDB

### Bucket/Cluster: TDB

# Types de données

### Primitifs : TDB

### Complexes : TDB

TDB

# HiveQL, SQL pour Hive

TDB

HiveQL supporte :

### DDL : TDB

### DML : TDB

### Fonction utilisateurs (UDF) : TDB

# Table interne vs table externe :

TDB

### Les tables internes : TDB

### Tables externes : TDB

# Hive: partitionnement des tables :

TDB

### Création de tables : TDB

### Ajouter une partition : TDB

### Renommer une partition : TDB

### Supprimer une partition : TDB

## Avantage du partitionnement :

TDB

## Inconvenient du partitionnement:

TDB

# 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

TBD

# Formats de fichier supportés par Hive ?

TBD