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

* C’est un outil de datawarehousing dans HDFS
* Couche intermédiaire entre Mapreduce et SQL
* Cela consiste à écrire des requêtes SQL et Hive génère le job MapReduce ou Tez
* Conçu pour les utilisateurs SQL qui ne connaissent pas de langage de programmation
* En mode batch

# Apache Hive: un peu d’histoire

* Développé par Facebook pour des usages internes
* C’est après qu’il est devenu un projet Apache Open source

# Terminologie

### Database:

* metastore : chargé du stockage des métadonnées pour chaque table
* Il enregistre par exemple les schémas ainsi que les localisations en incluant les métadonnées de partitionnement afin d'aider le driver à suivre la distribution des bases de données au sein du cluster.
* En pratique, les données sont stockées à la manière d'un SGBD relationnel

### Table :

### Partition:

### Bucket/Cluster: TDB

# Types de données

### Primitifs : TDB

### Complexes : TDB

TDB

# HiveQL, SQL pour Hive

HiveQL supporte les langages et les fonctions suivants :

### DDL (Langage de Définition de Données) :

Create, alter, drop

### DML (Langage de Manipulation de Données) :

Load, insert, select

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

# Table interne vs table externe :

TDB

### Les tables internes :

Pour les tables internes, Hive déplace les données dans le répertoire de son entrepôt.

* Si la table est supprimée, les métadonnées de la table et les données seront supprimées.
* Les données sont temporaires
* Hive gère complètement le cycle de vie de la table et des données.

### Tables externes :

Pour les tables externes, Hive ne déplace pas les données dans son répertoire d’entrepôt.

* Si la table externe est supprimée, les métadonnées de la table sont supprimées mais pas les données.
* Les données sont également utilisées en dehors de Hive. Par exemple, les fichiers de données sont lus et traités par un programme existant qui ne verrouille pas les fichiers.
* Les données doivent rester dans l’emplacement sous-jacent, même après un DROP TABLE.

# 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