1.什么是Hive

Hive是一款建立在Hadoop(分布式系统基础架构)之上的开源**数据仓库**系统，可以将存储在Hadoop文件中的结构化，半结构化数据文件映射为一张数据表，基于表提供了一种类似SQL的查询模块，称为Hive查询语言（HQL），用于访问和分析存储在Hadoop文件中的大型数据集。

Hive的核心是将HQL语言转换为MapReduce(面向大数据并行处理的计算模型、框架和平台)程序，然后将程序提交到 Hadoop群集执行。Hive是由FaceBook实现并开源。

2.为什么使用Hive

使用Hadoop MapReduce直接处理数据所面临的问题:

人员学习成本太高，需要掌握Java语言

MapReduce实现复杂查询逻辑开发难度太大

使用Hive处理数据的好处:

操作接口采用类SQL语法，提供快速开发的能力(简单，容易上手)

避免直接写MapReduce，减少开发人员学习成本

支持自定义函数，功能扩展很方便

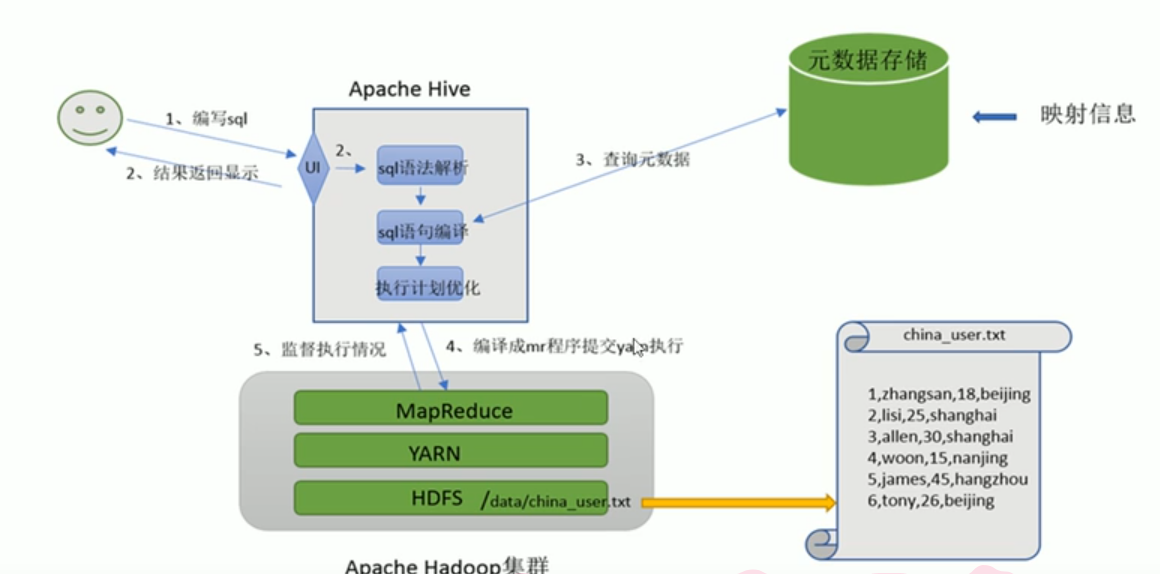
背靠Hadoop，擅长存储分析海量数据集

3.Hive和Hadoop的关系

数据仓库软件至少具备两种能力：(1)存储数据的能力，(2)分析数据的能力

Hive作为一款大数据时代的数据仓库软件，具备上述两种能力。只不过Hive并不是自己实现的，而是借助了Hadoop。Hive利用HDFS(HDFS是由Hadoop实现的分布式文件系统的一个组件)存储数据，利用MapReduce查询分析数据，Hive最大的优点就是使用户专注于编写HQL语句，由Hive将其转换为MapReduce程序完成对数据的分析

1. Hive的架构图



1. Hive的数据模型

数据模型：用来描述数据，组织数据和对数据进行操作，是对现实世界数据特征的描述

Hive的数据模型类似于关系数据库的库表结构，此外还有自己特有的模型，Hive中的数据可以在粒度级别上分为三类:table表，partition分区，bucket分桶。

Hive作为一个数据仓库，在结构上向传统数据库看齐，也分数据库，每个数据库下面有各自的表组成。默认数据库default。

Hive的数据是存储在HDFS上的，默认有一个根目录，在hive-site.xml中，由参数hive.metastore.warehouse.dir指定。默认值为**/user/hive/warehouse**。因此Hive中的数据库在HDFS上的存储路径为:

${hive.metastore.warehouse.dir}/databasename.db

例如，名为student的数据库存储路径为：

/user/hive/warehouse/student.db

Hive表与关系型数据库中的表相同。Hive中的表所对应的数据通常是存储在HDFS中，而表相关的元数据是存储在RDBMS中。

Hive中的表的数据在HDFS上的存储路径为：

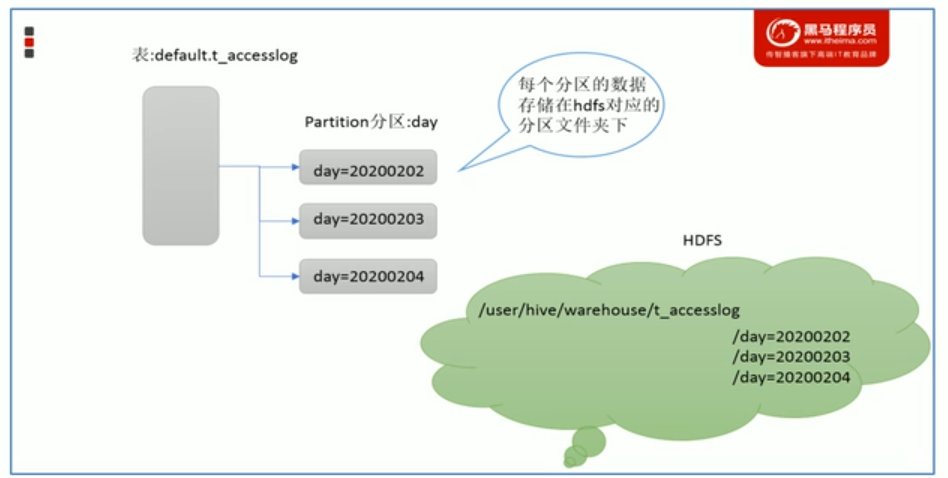
${hive.metastore.warehouse.dir}/databasename.db/tablename

例如，student数据库下的t\_user表存储路径为：

/user/hive/warehouse/student.db/t\_user

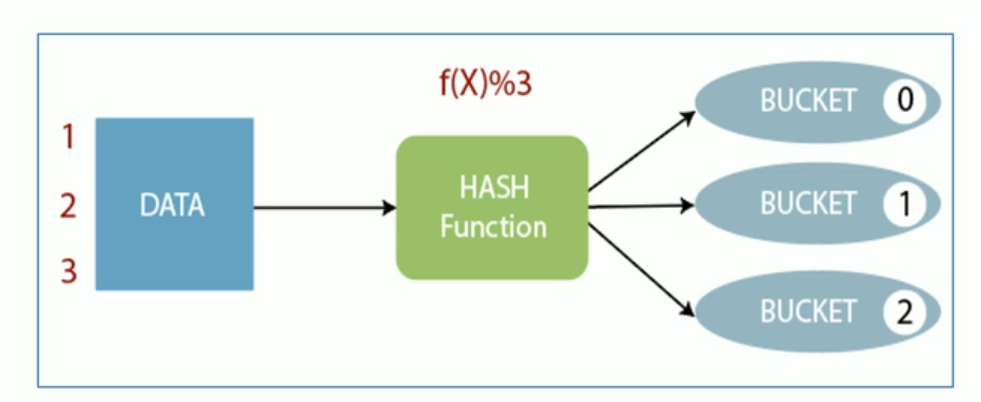
Partition分区是hive的一种优化表的手段。分区指根据分区列(例如“日期day”)的值将表划分为不同分区。这样可以更快地对指定分区数据进行查询。分区在存储层面上的表现是：table表目录下以子文件夹形式存在，一个文件夹表示一个分区，子文件命名标准：分区列=分区值。Hive还支持分区下继续建分区，

所谓的多重分区。



Bucket分桶表是优化表的手段。分桶是指根据表中的字段(例如“编号ID”)的值，经过hash计算规则将数据文件划分为指定的若干个小文件

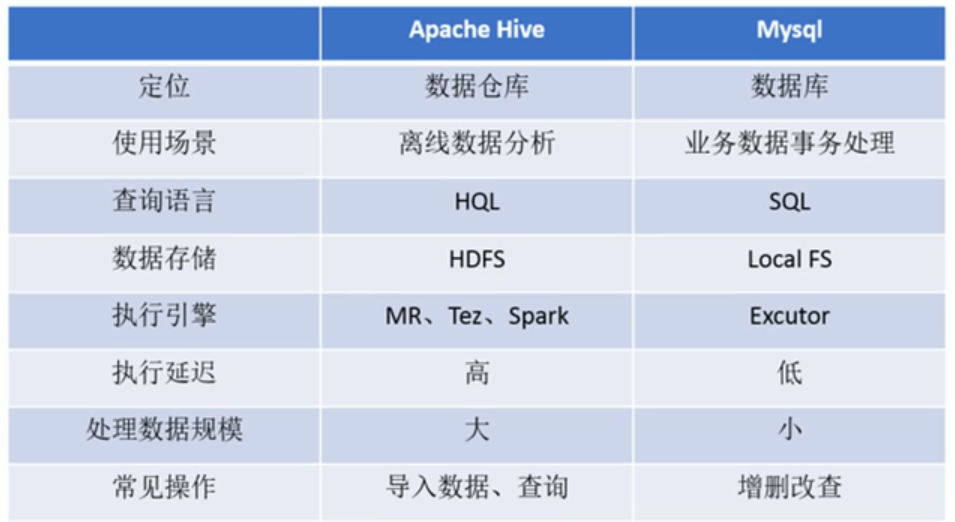
分桶规则：hashfunc(字段)%桶个数，余数相同的放在同一个文件



分桶的好处是可以优化join查询和方便抽样查询，Bucket分桶表在HDFS中的表现为同一个目录下数据根据hash散列之后变成多个文件。

6.Hive和MySQL

Hive虽然具有RDBMS数据库的外表，包括数据模型，SQL语法都十分相似，但应用场景却完全不同，Hive只适合用来做海量数据的离线分析。Hive的定位是数据仓库，是面向分析的OLAP系统，Hive和MySQL的区别

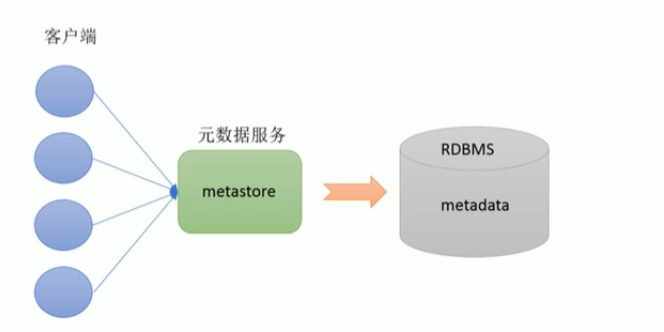


7.元数据

元数据(metadata)，又称为中介数据，中继数据，为描述数据的数据，主要是描述数据属性(property)的信息，用来支持如指示存储位置、历史数据、资源查找、文件记录等功能。

Hive元数据包含用Hive创建的database、table、表的位置、类型、属性、字段顺序类型等元信息。元数据存储在关系型数据库中。如Hive内置的Derby或者第三方MySQL等

metadatastore是元数据服务，元数据服务的作用是管理元数据，对外暴露服务地址，让各种客户端通过连metastore服务，由metastore再去连接MySQL数据库来存取元数据，有了metastore服务，就可以同时连接多个客户端，而且这些客户端不需要知道MySQL数据库的用户名和密码，只需要连接metastore服务即可，某种程度上来说保证了Hive元数据的安全。



8.metastore配置方式

metastore服务配置有三种模式：内嵌模式，本地模式和远程模式。

两个问题区分metastore的三种配置方式：

metastore服务是否需要单独配置、单独启动？

metastore是存储在内置的Derby中，还是第三方的RDBMS,比如MySQL

企业推荐模式--远程模式部署

