Hive入门

一、实验介绍

1.1 实验内容

本节课程主要介绍理论 ：

- Hive 的定义
- Hive 的体系结构
- Hive 与关系数据库的区别
- Hive 的应用场景
- Hive 的存储

1.2 实验知识点

- Hive QL
- 数据 ETL
- 元数据存储

1.3 实验环境

- Hive V2.0.0
- hadoop2.4.1
- Xfce终端

1.4 适合人群

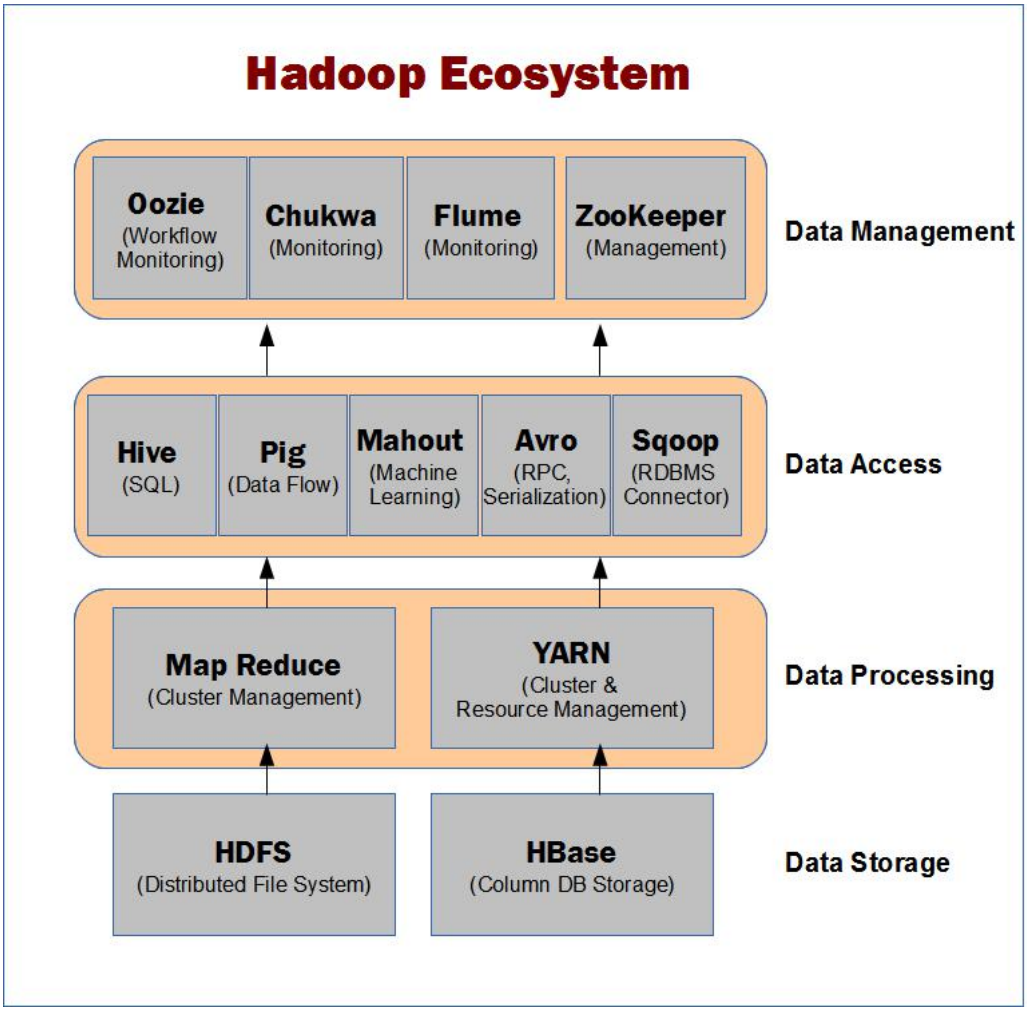
本课程难度为一般，属于初级级别课程，适合具有hadoop基础的用户，熟悉linux基础知识

一、什么是 Hive ?

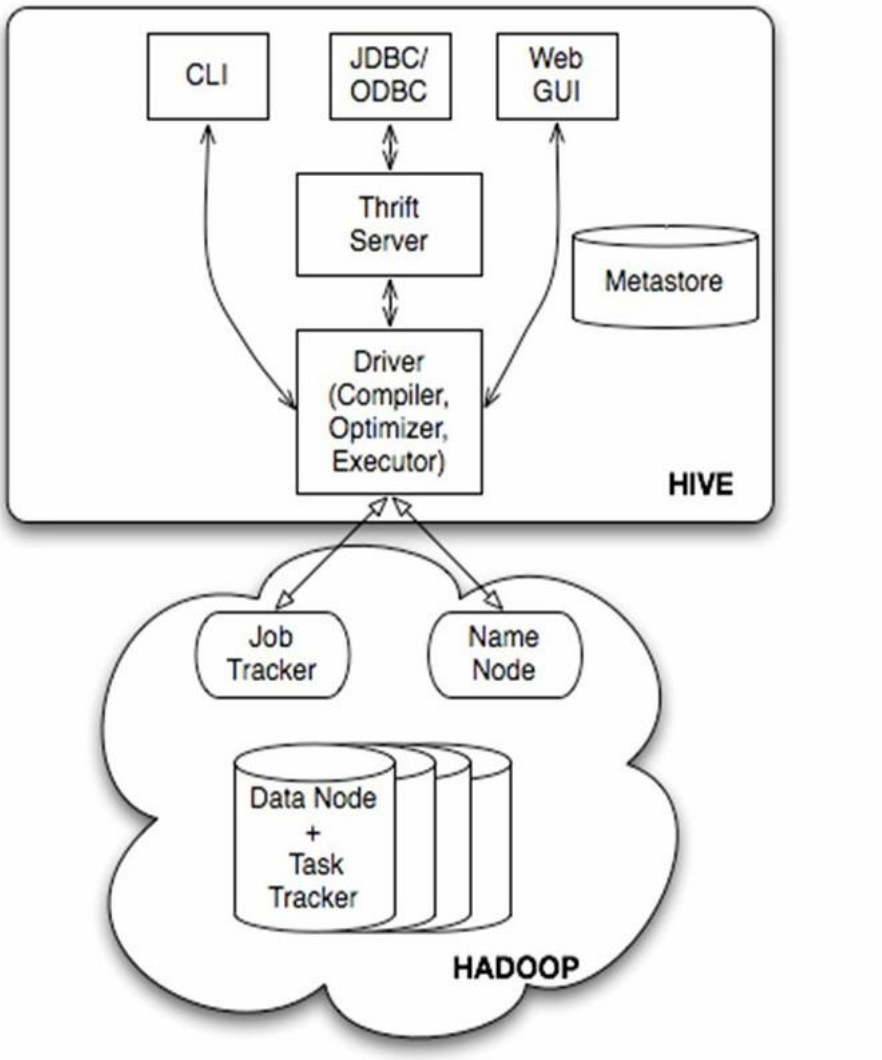
Hive 是一个基于 Hadoop 文件系统之上的数据仓库架构。它为数据仓库的管理提供了许多功能：数据 ETL（抽取、转换和加载）工具、数据存储管理和大型数据集的查询和分析能力。同时 Hive 还定义了类 SQL的语言 -- Hive QL. Hive QL 允许用户进行和 SQL 相似的操作，它可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能。还允许开发人员方便地使用 Mapper 和 Reducer 操作，可以将 SQL 语句转换为 MapReduce 任务运行，这对 MapReduce 框架来说是一个强有力的支持。

二、Hive 体系结构

Hive 是 Hadoop 中的一个重要子项目，从下图我们就可以大致了解 Hive 在 Hadoop 中的位置和关系。



上图描述 Hadoop EcoSystem 中的各层系统。而 Hive 本身的体系结构如下：



从图中我们可以看出 Hive 其基本组成可以分为：

- 用户接口，包括 CLI, JDBC/ODBC, WebUI
- 元数据存储，通常是存储在关系数据库如 MySQL, Derby 中
- 解释器、编译器、优化器、执行器
- Hadoop, 用 HDFS 进行存储，利用 MapReduce 进行计算

三、Hive 与关系数据库的区别

Hive 在很多方面与传统关系数据库类似（例如支持 SQL 接口），但是其底层对 HDFS 和 MapReduce 的依赖意味着它的体系结构有别于传统关系数据库，而这些区别又影响着 Hive 所支持的特性，进而影响着 Hive 的使用。

我们可以列举一些简单区别：

- Hive 和关系数据库存储文件的系统不同，Hive 使用的是 Hadoop 的HDFS（Hadoop的分布式文件系统），关系数据库则是服务器本地的文件系统；
- Hive 使用的计算模型是 MapReduce，而关系数据库则是自己设计的计算模型；
- 关系数据库都是为实时查询的业务进行设计的，而 Hive 则是为海量数据做数据挖掘设计的，实时性很差；实时性的区别导致 Hive 的应用场景和关系数据库有很大的不同；
- Hive 很容易扩展自己的存储能力和计算能力，这个是继承 Hadoop 的，而关系数据库在这个方面要差很多。

四、Hive 应用场景

通过对 Hive 与传统关系数据库的比较之后，其实我们不难得出 Hive 可以应用于哪些场景。

动手实践是学习 IT 技术最有效的方式！　　开始实验

Hive 构建在基于静态批处理的 Hadoop 之上，Hadoop 通常都有较高的延迟并且在作业提交和调度的时候需要大量的开销。因此，Hive 不适合在大规模数据集上实现低延迟快速的查询。

Hive 并不适合那些需要低延迟的应用，例如，联机事务处理（OLTP）。Hive 查询操作过程严格遵守 Hadoop MapReduce 的作业执行模型，Hive 将用户的 HiveQL 语句通过解释器转换为 MapReduce 作业提交到 Hadoop 集群上，Hadoop 监控作业执行过程，然后返回作业执行结果给用户。Hive 并非为联机事务处理而设计，Hive 并不提供实时的查询和基于行级的数据更新操作。

Hive 的最佳使用场合是大数据集的批处理作业，例如，网络日志分析。

五、Hive 的数据存储

Hive 的存储是建立在 Hadoop 文件系统之上的。Hive 本身没有专门的数据存储格式，也不能为数据建立索引，因此用户可以非常自由地组织 Hive 中的表，只需要在创建表的时候告诉 Hive 数据中的列分隔符就可以解析数据了。

Hive 中主要包括 4 种数据模型： 表（`Table`）、 外部表（`External Table`）、 分区（`Partition`）以及 桶（`Bucket`）。

Hive 的表和数据库中的表在概念上没有什么本质区别，在 Hive 中每个表都有一个对应的存储目录。而外部表指向已经在 HDFS 中存在的数据，也可以创建分区。Hive 中的每个分区都对应数据库中相应分区列的一个索引，但是其对分区的组织方式和传统关系数据库不同。桶在指定列进行 Hash 计算时，会根据哈希值切分数据，使每个桶对应一个文件。

六、Hive 的元数据存储

由于 Hive 的元数据可能要面临不断地更新、修改和读取操作，所以它显然不适合使用 Hadoop 文件系统进行存储。目前 Hive 把元数据存储在 RDBMS 中，比如存储在 MySQL, Derby 中。这点我们在上面介绍的 Hive 的体系结构图中，也可以看出。

七、实验总结

> 本次实验是 Hive 的简介。介绍了什么是 Hive, 它与传统关系数据库的区别，以及 Hive 的体系结构和使用场景等等。

参考文档

- 《Hadoop实战 第2版》陆嘉恒，机械工业出版社；
- 《Hadoop 权威指南》Tom White, 清华大学出版社；
- Hive 数据库仓库工具 (http://baike.baidu.com/subview/699292/10164173.htm#8)；
- Hive 体系结构 (http://sishuok.com/forum/blogPost/list/0/6231.html)；

下一节 (/courses/38/labs/766/document)