hive是建立在hadoop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，半结构化文件映射成一张数据库表，基于表提供了一种类似于SQL的查询模型，将sql语句翻译成mapreduce程序运行，处理hdfs中结构化数据

为什么使用HIVE:

直接使用mapreduce需要掌握JAVA语言，学习成本高，实现复杂查询逻辑开发难度大

背靠HADOOP，擅长存储分析海量数据集

若翻译成mapreduce程序，有如下问题：

数据文件在哪里

使用什么符号作为列的分隔符

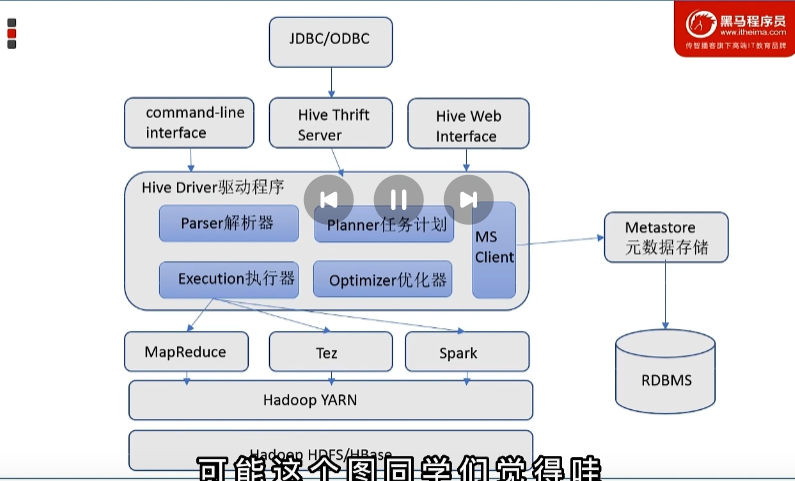
那些列可以作为city使用

city列是什么类型的数据

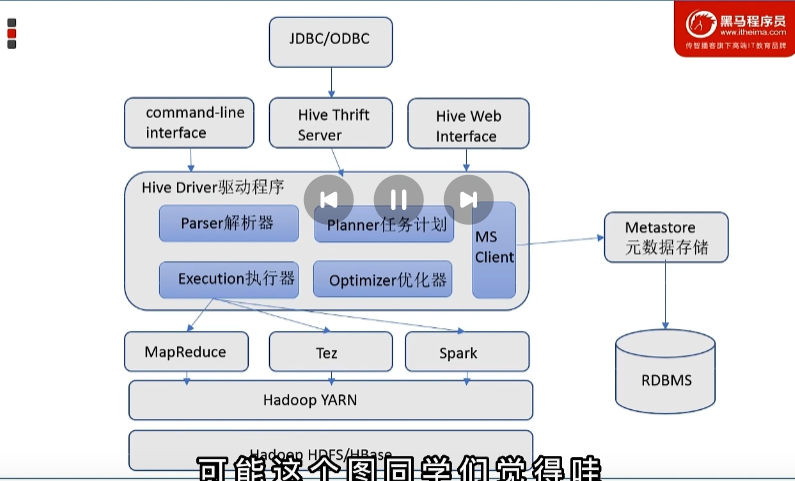
功能1元数据管理：数据位置，数据结构，对数据进行描述

功能2sql解析器：sql分析，sql到mapreduce程序的转换，提交mapreduce程序运行并收集执行结果

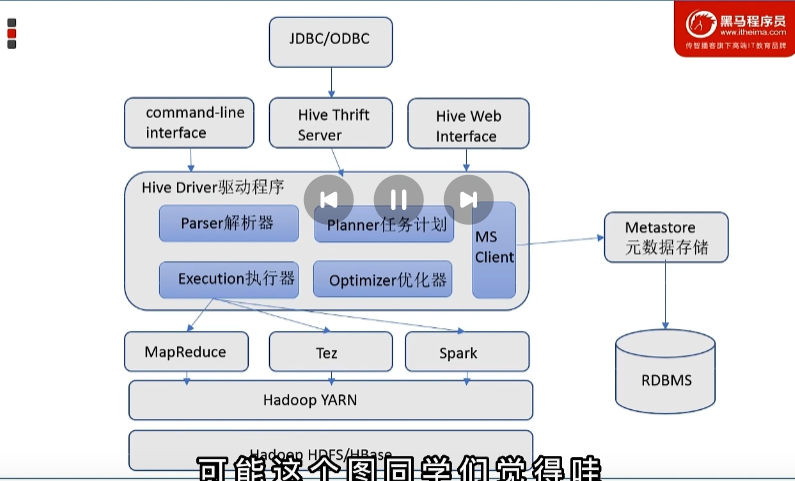
基础架构



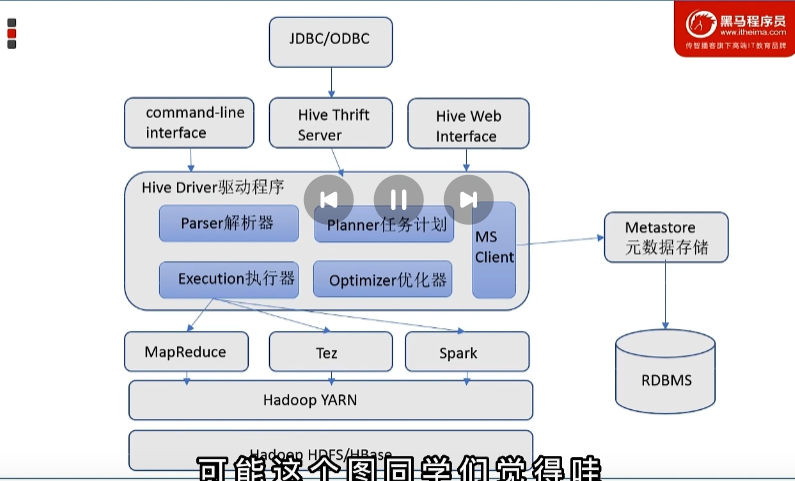
元数据存储



sql解析



用户接口



hive本身是单机工具，只需部署在一台机器上即可，它可以提交分布式程序

mysql安装配置



部署hive

1. 部署mysql数据库，并配置root账户密码，因为hive有元数据管理的功能，元数据的存储依赖于外部数据库
2. 下载hive并上传解压，设置软连接
3. 下载mysql驱动jar包放入hive的lib目录内

wget -b https://repo1.maven.org/maven2/mysql/mysql-connector-java/5.1.34/mysql-connector-java-5.1.34.jar

1. 修改配置文件，环境变量和hive-site.xml,hive-env.sh
2. 初始化元数据库

/export/server/hive/bin/schematool –initSchema –dbType mysql -verbos

1. 修改文件所属权限

chown –R hadoop:hadoop /export/server/apache-hive-3.1.3-bin /export/server/hive

1. 启动hive的metastore服务

nohup bin/hive --service metastore >> logs/metastore.log 2>&1 &

启动hiveserver2服务

nohup bin/hive –service hiveserver2 >> logs/hiveserver2.log 2>&1 &

1. 启动hive命令行

bin/hive

看起来hive处理的是表，实际上处理的是hdfs中的文件，处理结果以表格的形式返回给用户。

启动hiveserver2服务，让第三方工具能连接到hive数据库

数据库本质上就是hdfs上的文件夹

默认分隔符为\001，也可以建表的时候指定分隔符

row format delimited fields terminated by ‘\t’