初识Hive

谭唐华

课程大纲

- Hive是什么
- Hive体系结构
- Hive环境搭建
- Linux下MySQL安装
- Hive元数据配置
- Hive基本操作

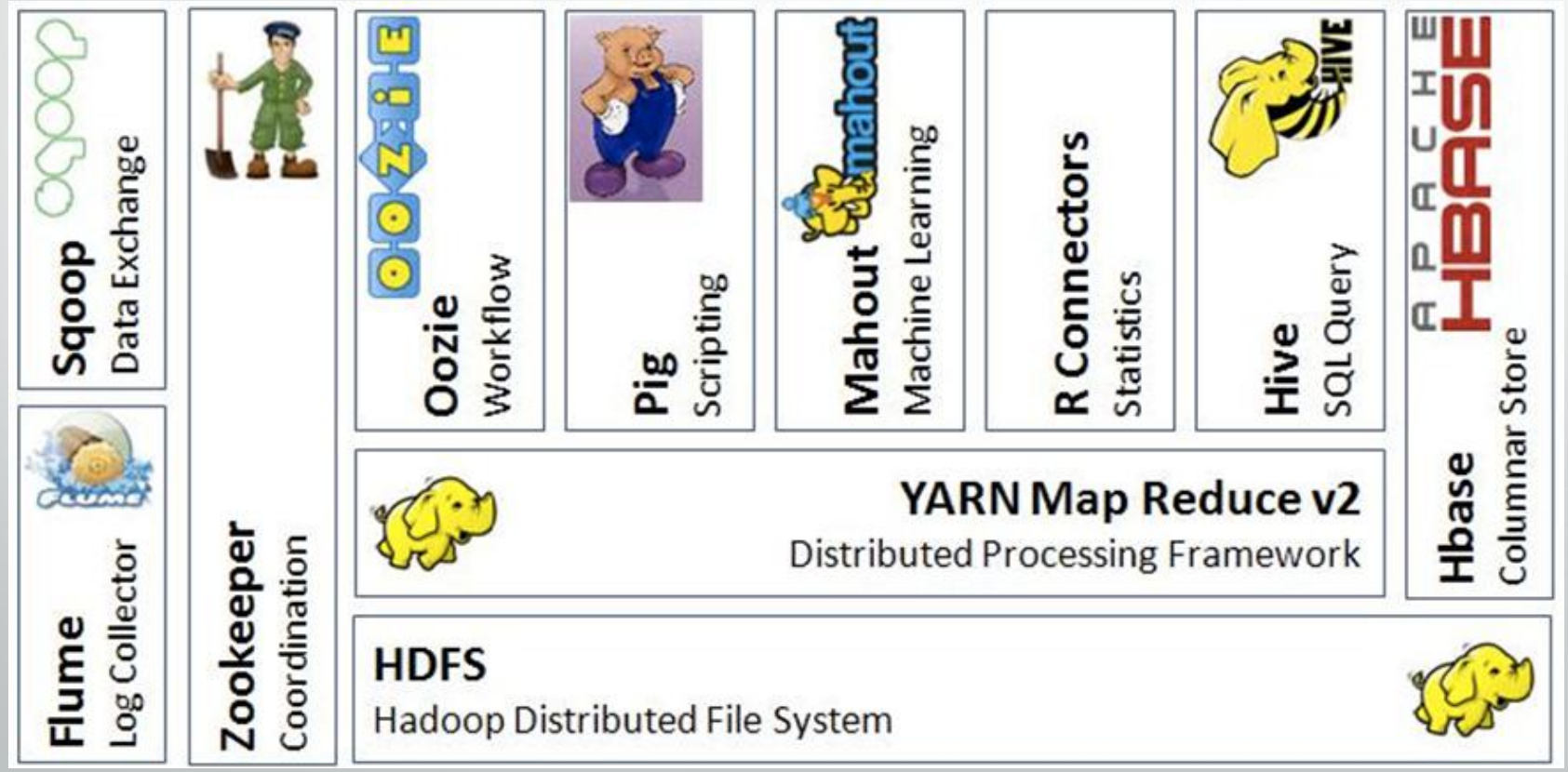
Hive是什么

- 由Facebook开源用于解决海量结构化日志的数据统计，后称为Apache Hive为一个开源项目
- Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射成一张表，并提供类SQL查询功能；

Hive是什么

- 构建在Hadoop之上的数据仓库；
  - 使用HQL作为查询接口；
  - 使用HDFS存储；
  - 使用MapReduce计算；
- 本质是：将HQL转化成MapReduce程序
- 灵活性和扩展性比较好：支持UDF，自定义存储格式等；
- 适合离线数据处理；

Hive生态系统位置


Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
APACHE HBASE
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

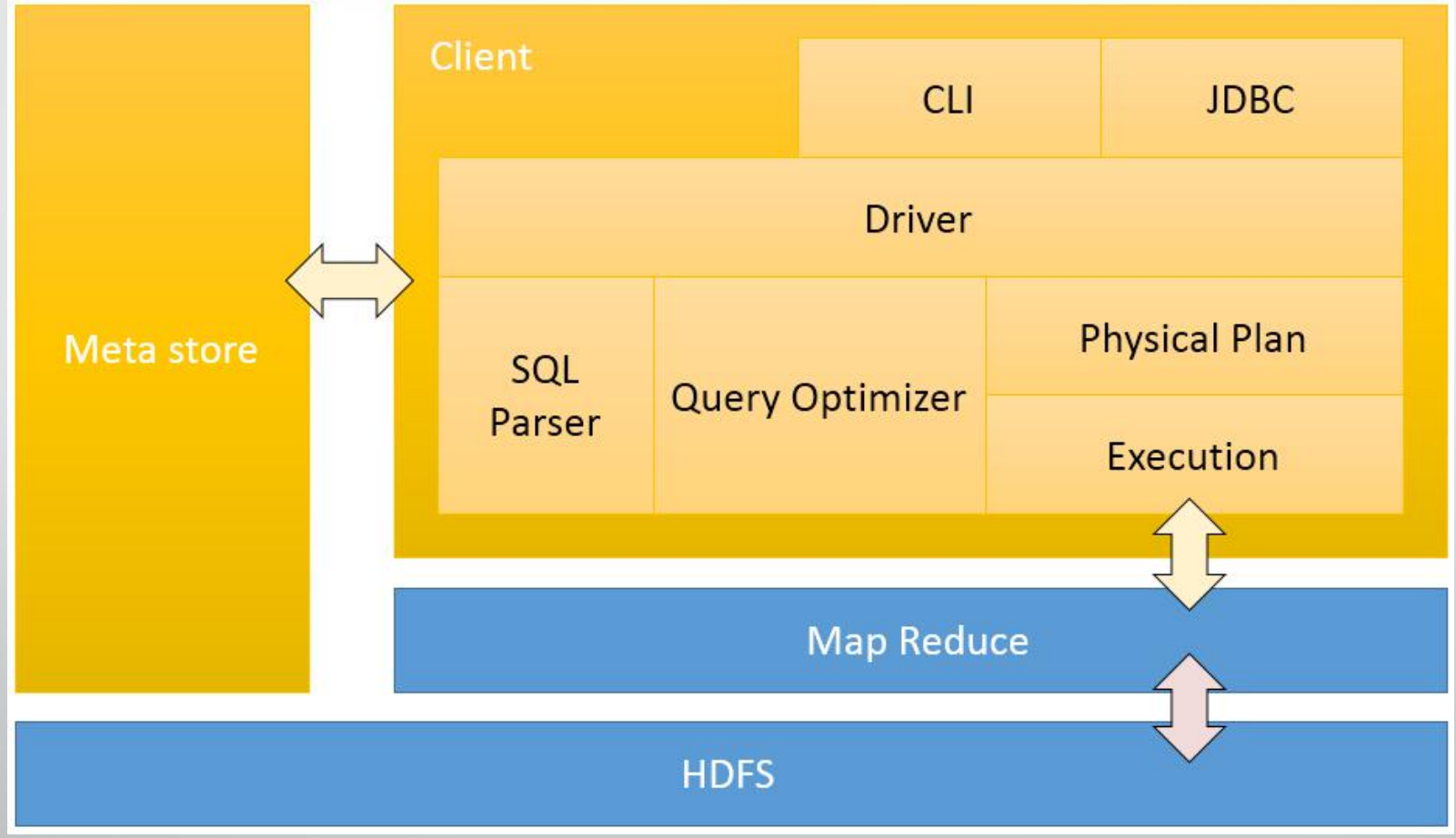
课程大纲

- Hive是什么
- Hive体系结构
- Hive环境搭建
- Linux下MySQL安装
- Hive元数据配置
- Hive基本操作

Hive体系结构

- Hive的数据存储基于Hadoop HDFS
- Hive没有专门的数据存储格式
- 存储结构主要包括：数据库、文件、表、视图、索引
- Hive默认可以直接加载文本文件（TextFile），还支持SequenceFile、RCFile
- 创建表时，指定Hive数据的列分隔符与行分隔符，Hive即可解析数据

Hive架构图


Client
CLI
JDBC
Driver
SQL Parser
Query Optimizer
Physical Plan
Execution
Meta store
Map Reduce
HDFS

Hive体系结构

- 用户接口: Client
  - CLI(hive shell)、JDBC/ODBC(java访问hive)，WEBUI(浏览器访问hive)
- 元数据: Metastore
  - 元数据包括：表名、表所属的数据库（默认是default）、表的拥有者、列/分区字段、表的类型（是否是外部表）、表的数据所在目录等；
  - 默认存储在自带的derby数据库中，推荐使用采用MySQL
- Hadoop
  - 使用HDFS进行存储，使用MapReduce进行计算；

驱动器: Driver

- 驱动器: Driver
  - 包含：解析器、编译器、优化器、执行器；
  - 解析器：将SQL字符串转换成抽象语法树AST，这一步一般都用第三方工具库完成，比如antlr；对AST进行语法分析，比如表是否存在、字段是否存在、SQL语义是否有误(比如select中被判定为聚合的字段在group by中是否有出现)；
  - 编译器：将AST编译生成逻辑执行计划；
  - 优化器：对逻辑执行计划进行优化；
  - 执行器：把逻辑执行计划转换成可以运行的物理计划。对于Hive来说，就是MR/TEZ/Spark；

Hive 优点与使用场景

- Hive 优点与使用场景
  - 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)；
  - 避免了去写MapReduce，减少开发人员的学习成本；
  - 统一的元数据管理，可与impala/spark等共享元数据；
  - 易扩展(HDFS+MapReduce：可以扩展集群规模；支持自定义函数)；
  - 数据的离线处理；比如：日志分析，海量结构化数据离线分析…
  - Hive的执行延迟比较高，因此hive常用于数据分析的，对实时性要求不高的场合；
  - Hive优势在于处理大数据，对于处理小数据没有优势，因为Hive的执行延迟比较高。

课程大纲

- Hive是什么
- Hive体系结构
- Hive环境搭建
- Linux下MySQL安装
- Hive元数据配置
- Hive基本操作

Hive环境搭建

- 官网
- http://hive.apache.org
- 文档
- https://cwiki.apache.org/confluence/display/Hive/GettingStarted
- https://cwiki.apache.org/confluence/display/Hive/Home
- 下载
- http://archive.apache.org/dist/hive/

Hive环境搭建

- 安装JDK和HADOOP环境
- 下载hive源文件
- 解压hive文件
  - tar zxf hive-0.13.1-cdh5.3.6.tar.gz -C /opt/
- 进入$HIVE_HOME/conf/修改文件
  - cp hive-env.sh.template hive-env.sh
  - cp hive-default.xml.template hive-site.xml
- 修改$HIVE_HOME/bin的hive-env.sh，增加以下三行
  - JAVA_HOME=/opt/jdk1.7.0_67
  - HADOOP_HOME=/opt/hadoop-2.5.0-cdh5.3.6
  - export HIVE_CONF_DIR=/opt/hive-0.13.1-cdh5.3.6/conf

课程大纲

- Hive是什么
- Hive体系结构
- Hive环境搭建
- Linux下MySQL安装
- Hive元数据配置
- Hive基本操作

Linux下MySQL安装

Supported Backend Databases for Metastore

Database	Minimum Supported Version	Name for Parameter Values
MySQL	5.6.17	mysql
Postgres	9.1.13	postgres
Oracle	11g	oracle
MS SQL Server	2008 R2	mssql

Linux下MySQL安装

- yum命令安装
  - yum install mysql mysql-devel mysql-server
- 启动服务
  - service mysqld start
- 开机启动启动
  - chkconfig mysqld on
- 创建root管理员密码
  - mysqladmin -uroot password ‘123456’
- 给用户和机器授权：
  - # mysql -uroot –p123456
  - mysql> grant all on *.* to root@ ' bigdata.beifeng.com' identified by '123456' ;
  - mysql> flush privileges;

课程大纲

- Hive是什么
- Hive体系结构
- Hive环境搭建
- Linux下MySQL安装
- Hive元数据配置
- Hive基本操作

Hive元数据配置

- 配置MySQL的metastore
  - 修改$HIVE_HOME/conf/hive-site.xml
  - 拷贝驱动包到$HIVE_HOME/lib

    cp mysql-connector-java-5.1.27-bin.jar /opt/apache-hive-0.13.1-bin/lib/

Hive仓库目录

- 创建hive目录
  - hive临时目录：$HADOOP_HOME/bin/hadoop fs -mkdir /tmp
  - Hive仓库目录：$HADOOP_HOME/bin/hadoop fs -mkdir -p /user/hive/warehouse
- 修改目录权限
  - 修改/tmp：$HADOOP_HOME/bin/hadoop fs -chmod g+w /tmp
  - 修改/warehouse：$HADOOP_HOME/bin/hadoop fs -chmod g+w /user/hive/warehouse

Hive的日志信息

- 重命名配置文件
- $ mv hive-log4j.properties.template hive-log4j.properties

- 修改log4j配置
- Hive下创建日志存放目录：$ mkdir logs
- 修改hive-log4j.properties：hive.log.dir=/opt/modules/apache-hive-0.13.1-bin/logs重命名配置文件
  - $ mv hive-log4j.properties.template hive-log4j.properties

- 修改log4j配置
  - Hive下创建日志存放目录：$ mkdir logs
  - 修改hive-log4j.properties：hive.log.dir=/opt/modules/apache-hive-0.13.1-bin/logs

课程大纲

- Hive是什么
- Hive体系结构
- Hive环境搭建
- Linux下MySQL安装
- Hive元数据配置
- Hive基本操作

Hive命令行模式

- 直接输入#/hive/bin/hive的执行程序，或者输入 #hive --service cli 启动

  hive>show tables;

  hive>create table test(id int,name string);

  hive>quit;

- 观察：#hadoop fs -ls /user/hive/warehouse/

  修改参数：hive.metastore.warehouse.dir表与目录的对应关系

Hive体系结构

- 与linux交互命令！
  - !ls
  - !pwd
- 与hdfs交互命令
  - dfs -ls /
  - dfs -mkdir /hive

DDL， DML操作

- 库
  - 创建库：create database t1 ;
  - 查看库：show databases;
  - 切换库：use t1 ;（修改hive.cli.print.header为true； hive.cli.print.current.db为true）
  - 删除库：drop database t1;
- 表
  - 创建表：create table table01(id int,name string) ;
  - 查看表：show tables;　　表概括：desc table01；表详情： desc formatted table01 ;
  - 查询表：select * from table01;
  - 删除表：drop table table01;

表操作

- 创建表
  - hive> create table student(id int,name string)
    > row format delimited fields terminated by '\t' ;
- 向表插入数据
  - load data local inpath 'student.txt' into table student ;
- 查询（验证是否加载mr）
  - select * from student ; select name from student ;

Hive脚本方式

- $> bin/ hive -e “hql"
- $> bin/ hive -e """>aaa
- $> bin/ hive -S -e """>aaa
- $> bin/ hive -f script.hql
- $> bin/ hive -i /home/my/hive-init.sql
- hive>source file

set命令使用

- hive控制台set命令:
  - set hive.cli.print.current.db=true;
  - set hive.metastore.warehouse.dir=/hive
- hive参数初始化配置set命令:
  - ~/.hiverc
- 补充：
  - hive历史操作命令集
  - ~/.hivehistory

Hive的JDBC模式

- JAVA API交互执行方式
- hive 远程服务 (端口号10000) 启动方式
- #bin/hive --service hiveserver2
- 在java代码中调用hive的JDBC建立连接

Hive与Hadoop的关系、mysql的关系

- Hive数据存储在HDFS
- hive.metastore.warehouse.dir
- /user/hive/warehouse
- 创建数据库
- 默认会到仓库目录下面去创建一个同名的目录,
- 这个目录就是用来保存该数据库所有的表的数据
- 创建表
- 在数据库目录下面生产一个同名的目录，用来保存该数据表的所有文件
- 载入数据
- load data ...
- 件上传到目录/user/hive/warehouse/t1.db/student
-
- 总结：　Hive数据就是存在HDFS之上
- Hive的数据并没有存入mysql,mysql只是存入了元数据

总结

- Hive数据仓库优点
- Hive与hadoop、MapReduce关系
- Hive环境部署
- Mysql数据库安装与元数据存储
- Hive基础操作命令