北风网 IBEIFENG.COM

大数据Hadoop高薪直通车课程

Hive 初识入门

讲师：轩宇（北风网版权所有)

课程大纲

1. Hive 体系结构
2. Hive 环境搭建
3. Linux下MySQL安装
4. Hive 元数据配置
5. Hive 基本操作

课程大纲

1. Hive 体系结构
2. Hive 环境搭建
3. Linux下MySQL安装
4. Hive 元数据配置
5. Hive 基本操作

大数据招聘需求之【Hive】

Hive [北京] 2015-08-24

12k-20k 经验3-5年 / 硕士

"数据挖掘、机器学习、人工智能"

EmoKit 海妖情绪识别引擎

移动互联网 · 医疗健康 / 初创型(天使轮)

股票期权 | 弹性工作 | 扁平管理 | 岗位晋升

Hive [北京] 2015-07-16

15k-25k 经验3-5年 / 本科

"正A轮"

UCAR用车

移动互联网 · O2O / 成长型(A轮)

技能培训 | 年底双薪 | 节日礼物 | 绩效奖金

Hive [北京] 2015-07-16

15k-30k 经验3-5年 / 大专

"扁平化管理 晋升空间大 五险一金"

UCAR用车

移动互联网 · O2O / 成长型(A轮)

技能培训 | 年底双薪 | 节日礼物 | 绩效奖金

Hive [上海] 2015-08-28

17k-23k 经验3-5年 / 本科

"轻松自由的办公文化"

新实数码

移动互联网 · 生活服务 / 成长型(B轮)

节日礼物 | 年底双薪 | 免费班车 | 股票期权

Hive [杭州] 2015-08-22

15k-25k 经验3-5年 / 本科

"免费午餐，免费晚餐，免费夜宵"

金升奇

电子商务 · 旅游 / 成长型(A轮)

年底双薪 | 专项奖金 | 绩效奖金 | 年终分红

MapReduce编程的不便性

- MapReduce is **hard** to program
  - 【八股文】格式编程，三大部分

- **No Schema**, lack of query lanaguages, eg. SQL
  - 数据分析，针对DBA、SQL语句，如何对数据分析
  - MapReduce编程成本高
  - FaceBook 实现并开源Hive

Apache Hive

Apache Hive

The **Apache Hive**™ data warehouse software facilitates querying and managing large datasets residing in distributed storage. Built on top of **Apache Hadoop**™, it provides

- Tools to enable easy data extract/transform/load (ETL)
- A mechanism to impose structure on a variety of data formats
- Access to files stored either directly in **Apache HDFS**™ or in other data storage systems such as **Apache HBase**™
- Query execution via MapReduce

什么是Hive?

✓ 由Facebook开源用于解决海量**结构化**日志的数据统计；

✓ 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；

✓ **构建在Hadoop之上的数据仓库**；

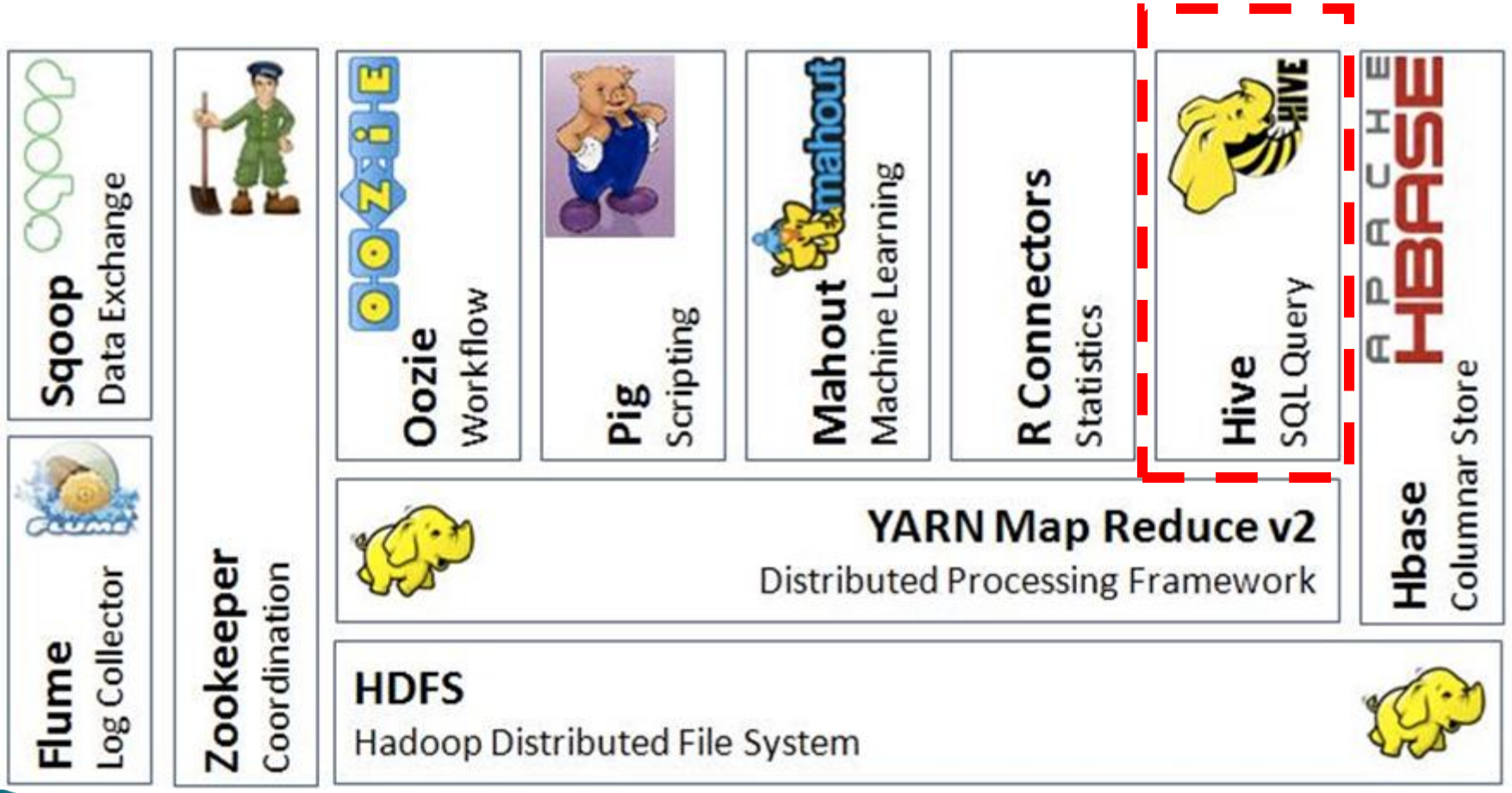
- 使用HQL作为查询接口；
- 使用HDFS存储；
- 使用MapReduce计算；

✓ **本质是：将HQL转化成MapReduce程序**

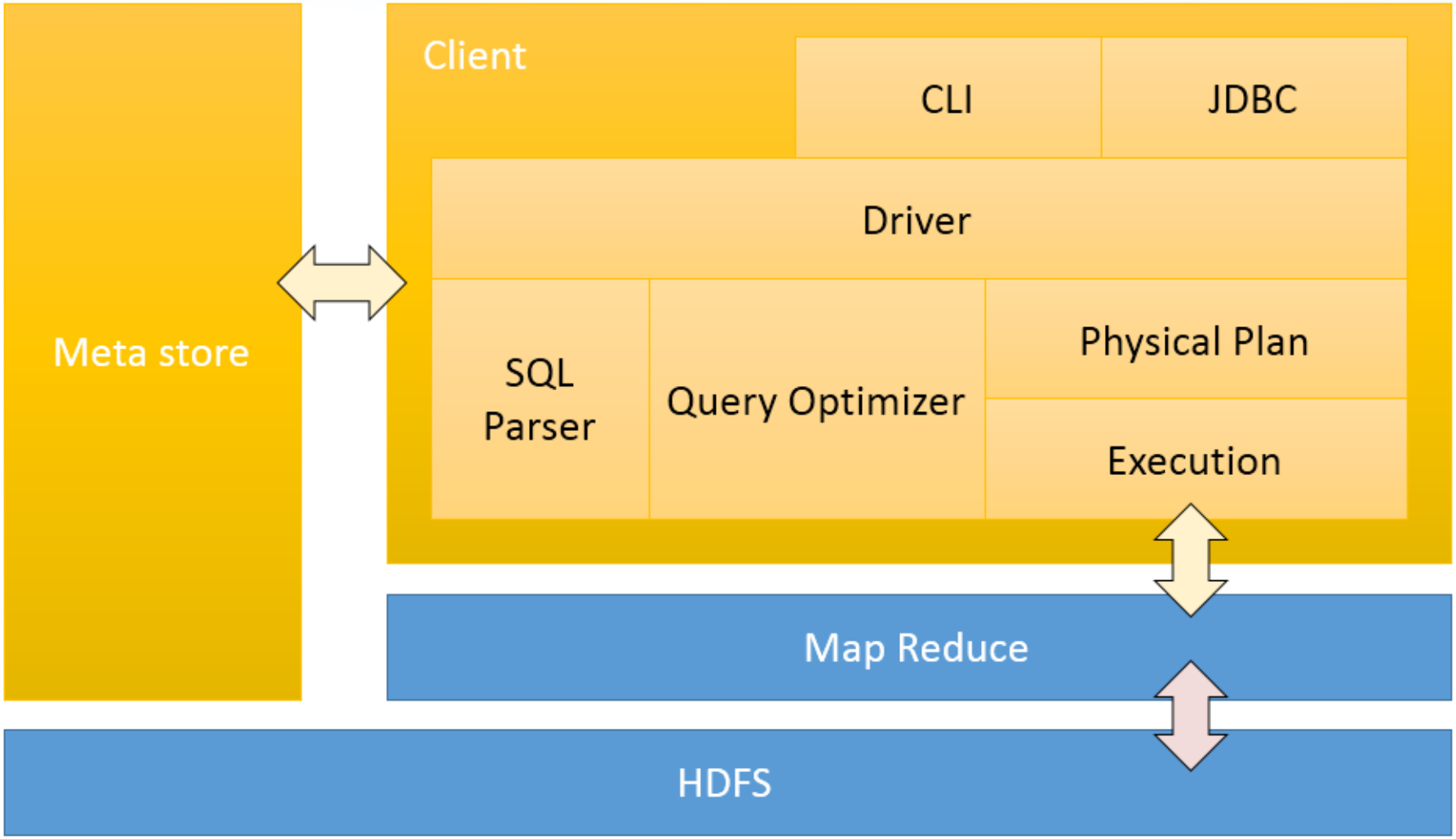
✓ 灵活性和扩展性比较好：支持UDF，自定义存储格式等；

✓ 适合**离线数据处理**；

Hive在生态系统的位置



Hive 架构


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

Hive 架构

- **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存储Metastore；

- **Hadoop**

使用HDFS进行存储，使用MapReduce进行计算；

Hive 架构

➢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

✓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

1. Hive 体系结构
2. Hive 环境搭建
3. Linux下MySQL安装
4. Hive 元数据配置
5. Hive 基本操作

Hive 相关文档

◆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/

Requirements

Requirements

- Java 1.7
  Note: Hive versions 1.2 onward require Java 1.7 or newer. Hive versions 0.14 to 1.1 work with Java 1.6 as well. Users are strongly advised to start moving to Java 1.8 (see HIVE-8607).
- Hadoop 2.x (preferred), 1.x.
  Hive versions up to 0.13 also supported Hadoop 0.20.x, 0.23.x.
- Hive is commonly used in production Linux and Windows environment. Mac is a commonly used development environment. The instructions in this document are applicable to Linux and Mac. Using it on Windows would require slightly different steps.

Installing Hive

Start by downloading the most recent stable release of Hive from one of the Apache download mirrors (see Hive Releases).

Next you need to unpack the tarball. This will result in the creation of a subdirectory named `hive-x.y.z` (where `x.y.z` is the release number):

```
$ tar -xzvf hive-x.y.z.tar.gz
```

Set the environment variable `HIVE_HOME` to point to the installation directory:

```
$ cd hive-x.y.z
$ export HIVE_HOME={{pwd}}
```

Finally, add `$HIVE_HOME/bin` to your `PATH`:

```
$ export PATH=$HIVE_HOME/bin:$PATH
```

Installing Hive

It will create the subdirectory build/dist with the following contents:

- README.txt: readme file.
- bin/: directory containing all the shell scripts
- lib/: directory containing all required jar files
- conf/: directory with configuration files
- examples/: directory with sample input and query files

Subdirectory build/dist should contain all the files necessary to run Hive. You can run it from there or copy it to a different location, if you prefer.

In order to run Hive, you must have Hadoop in your path or have defined the environment variable HADOOP_HOME with the Hadoop installation directory.

Moreover, we strongly advise users to create the HDFS directories /tmp and /user/hive/warehouse (also known as hive.metastore.warehouse.dir) and set them chmod g+w before tables are created in Hive.

Running Hive

Hive uses Hadoop, so:

- you must have Hadoop in your path OR
- `export HADOOP_HOME=<hadoop-install-dir>`

In addition, you must create `/tmp` and `/user/hive/warehouse` (aka `hive.metastore.warehouse.dir`) and set them `chmod g+w` in HDFS before you can create a table in Hive.

Commands to perform this setup:

```
$ $HADOOP_HOME/bin/hadoop fs -mkdir       /tmp
$ $HADOOP_HOME/bin/hadoop fs -mkdir       /user/hive/warehouse
$ $HADOOP_HOME/bin/hadoop fs -chmod g+w   /tmp
$ $HADOOP_HOME/bin/hadoop fs -chmod g+w   /user/hive/warehouse
```

Running Hive

You may find it useful, though it's not necessary, to set HIVE_HOME:

```
$ export HIVE_HOME=<hive-install-dir>
```

Running Hive CLI

To use the Hive command line interface (CLI) from the shell:

```
$ $HIVE_HOME/bin/hive
```

DDL Operations

create table user

```
create table user2(id int,name string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' ;
```

load data

```
load data local inpath '/opt/datas/user2.tsv' overwrite into table user2 ;
```

query data

```
select * from user ;
```

user.tsv

```
1	Zhangsan
2	Lisi
3	wangwu
```

课程大纲

1. Hive 体系结构
2. Hive 环境搭建
3. Linux下MySQL安装
4. Hive 元数据配置
5. Hive 基本操作

Supported Backend Databases for Metastore

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`

MySQL安装

采用yum安装方式安装

```
yum install mysql-server
```

判断MySQL是否已经安装好：

```
chkconfig --list|grep mysql
```

启动mysql服务：

```
service mysqld start
```

或者

```
/etc/init.d/mysqld start
```

检查是否启动mysql服务：

```
/etc/init.d/mysqld status
```

启动mysql服务：

```
service mysqld start
```

或者

```
/etc/init.d/mysqld start
```

检查是否启动mysql服务：

```
/etc/init.d/mysqld status
```

MySQL安装

设置MySQL开机启动：

```
1.  chkconfig mysqld on
```

检查设置MySQL开机启动是否配置成功：显示2 3 4 5为on

```
1.  chkconfig --list|grep mysql
```

创建root管理员:

```
1.  mysqladmin -uroot password root
```

登录

```
1.  mysql -uroot -proot
```

设置用户连接权限

```
$ sudo cat /root/.mysql_secret
```

修改 root 用户密码为 123456

```
> SET PASSWORD = PASSWORD('123456');
```

```
# 查询用户信息
mysql> select User,Host,Password  from user;
# 更新用户信息
mysql> update user set Host='%' where User = 'root' and Host='localhost' ;
# 删除用户信息
mysql> delete from user where user='root' and host='127.0.0.1';
# 刷新信息
mysql> flush privileges;
```

课程大纲

1. Hive 体系结构
2. Hive 环境搭建
3. Linux下MySQL安装
4. Hive 元数据配置
5. Hive 基本操作

Configuration Parameters

https://cwiki.apache.org/confluence/display/Hive/AdminManual+MetastoreAdmin

Config Param	Config Value	Comment
javax.jdo.option.ConnectionURL	jdbc:mysql://<host name>/<database name>?createDatabaseIfNotExist=true	metadata is stored in a MySQL server
javax.jdo.option.ConnectionDriverName	com.mysql.jdbc.Driver	MySQL JDBC driver class
javax.jdo.option.ConnectionUserName	<user name>	user name for connecting to MySQL server
javax.jdo.option.ConnectionPassword	<password>	password for connecting to MySQL server
hive.metastore.warehouse.dir	<base hdfs path>	default location for Hive tables.

课程大纲

1. Hive 体系结构
2. Hive 环境搭建
3. Linux下MySQL安装
4. Hive 元数据配置
5. Hive 基本操作

Hive Basic Operations

```
show databases;

use hive;

show tables;

create table helloworld(id int, name string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

load data local inpath '/home/hadoop/data/helloworld.txt' overwrite into table helloworld;

create table helloworld2(id int, name string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';

SHOW TABLES 'hello*';

desc helloworld;

desc extended helloworld;

desc formatted helloworld;

show functions;

desc function upper;

DESCRIBE FUNCTION EXTENDED upper;
```

```
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask.
MetaException(message:javax.jdo.JDODataStoreException: An exception was thrown while
adding/validating class(es) : Specified key was too long; max key length is 767 bytes
com.mysql.jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: Specified key was too long; max
key length is 767 bytes
        at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
        at sun.reflect.NativeConstructorAccessorImpl.newInstance(
        NativeConstructorAccessorImpl.java:57)
        at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(
        DelegatingConstructorAccessorImpl.java:45)
        at java.lang.reflect.Constructor.newInstance(Constructor.java:526)
        at com.mysql.jdbc.Util.handleNewInstance(Util.java:411)
        at com.mysql.jdbc.Util.getInstance(Util.java:386)
        at com.mysql.jdbc.SQLError.createSQLException(SQLError.java:1054)
```

http://blog.csdn.net/wind520/article/details/39890967

MySQL的varchar主键只支持不超过768个字节 或者 768/2=384个双字节 或者 768/3=256个三字节的字段
而 GBK是双字节的，UTF-8是三字节的。

Hive常用属性配置

- Hive数据仓库位置配置
- Hive运行日志信息位置
- 指定hive运行时显示的log日志的级别
- 在cli命令行上显示当前数据库，以及查询表的行头信息
- 在启动hive时设置配置属性信息
- 查看当前所有的配置信息

Hive Shell常用操作

Hive常用命令行操作：

- -e SQL from command line
- -f SQL from files
- -v,--verbose Verbose mode (echo executed SQL to the console)
- -S,--silent Silent mode in interactive shell

Hive Shell常用操作

Hive交互式命令行操作：

- quit/exit
- set key=value
- set
- set -v
- !
- dfs
- query string

查看操作历史命令

```
$HOME/.hivehistory
```

本课程版权归北风网所有

欢迎访问我们的官方网站

www.ibeifeng.com