# 大数据技术原理与应用网站教程

### 一、Hadoop下载安装

Ubuntu下载安装配置JDK：<https://jingyan.baidu.com/article/25648fc1797a649191fd00ca.html>

Java SE Download:

https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

Ubuntu下载安装配置Hadoop：

<http://dblab.xmu.edu.cn/blog/install-hadoop/>

Hadoop在线查看NameNode和DataNode的信息和HDFS中的文件：

<http://localhost:50070/dfshealth.html#tab-overview>

Hadoop中运行一个输出时，不可以和之前有重复的输出文件夹,可以通过增加程序代码块：

Configuration conf = new Configuration();

Job job = new Job(conf);

/\* 删除输出目录 \*/

Path outputPath = new Path(args[1]);

outputPath.getFileSystem(conf).delete(outputPath, true);

Hadoop启动的Shell命令：

./sbin/start-dfs.sh

Hadoop关闭的Shell命令：

./sbin/stop-dfs.sh

Ubuntu中配置环境变量的Shell命令：

sudo gedit /etc/environment #编辑用户环境变量

source /etc/environment #使的环境变量立即生效

sudo gedit /etc/profile #编辑全局环境变量

source /etc/profile #使的环境变量立即生效

## HDFS

HDFS学习指南：

<http://dblab.xmu.edu.cn/blog/290-2/>

Java API所在的jar包都在已经安装好的hadoop文件夹里：/usr/local/hadoop/share/hadoop

/hadoop/etc/hadoop目录下的core-site.xml和hdfs-site.xml放到当前工程项目下，即eclipse工作目录的bin文件夹下面，否则报错。

Hadoop API 文档：

<http://hadoop.apache.org/docs/stable/api/>

## HBase

HBase学习指南：

<http://dblab.xmu.edu.cn/blog/588-2/>

HBase下载地址：

<http://archive.apache.org/dist/hbase/>

## NoSQL

参考网络教程：

<http://www.runoob.com/mongodb/mongodb-tutorial.html>

MongDB安装和操作：

<http://dblab.xmu.edu.cn/blog/115/>

Linux MongDB下载地址：

<http://www.mongodb.org/downloads>

MongDB jar包下载地址：

<https://github.com/mongodb/mongo-java-driver/downloads>

## 云数据库

阿里云官网：

<http://www.aliyun.com/>

使用支付宝账户登录阿里云，账户登录成功后，点击“云数据库RDS“，即可进入云数据库RDS页面。点击“立即购买”，即可获得RDS服务。新用户可以免费体验半年的RDS服务。购买成功后，可以通过管理控制台对RDS实例进行使用

## MapReduce

使用 Hadoop 2.6.0 运行 WordCount 实例至少需要如下三个 jar:

$HADOOP\_HOME/share/hadoop/common/hadoop-common-2.6.0.jar

$HADOOP\_HOME/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.6.0.jar

$HADOOP\_HOME/share/hadoop/common/lib/commons-cli-1.2.jar

通过命令 hadoop classpath 可以得到运行 Hadoop 程序所需的全部 classpath信息

将 Hadoop 的 classhpath 信息添加到 CLASSPATH 变量中，在 ~/.bashrc 中增加如下几行：

export HADOOP\_HOME=/usr/local/hadoop export CLASSPATH=$($HADOOP\_HOME/bin/hadoop classpath):$CLASSPATH

MapReduce编程需要：

先运行 .java 的java源程序，生成 .class 文件

接着把 .class 文件打包成 jar，才能在 Hadoop 中运行 jar -cvf jar包名.jar 需要加入的class文件。Eg：jar -cvf WordCount.jar ./WordCount\*.class

最后运行hadoop运行jar包。Eg：/usr/local/hadoop/bin/hadoop jar WordCount.jar WordCount input output 解释：通过hadoop运行WordCount.jar的jar包中的WordCount类，输入文件为input，输出文件为output

MapReduce学习指南：

<http://dblab.xmu.edu.cn/blog/631-2/>

使用命令行编译打包运行自己的MapReduce程序学习指南：

<http://dblab.xmu.edu.cn/blog/hadoop-build-project-by-shell/>

使用Eclipse编译运行MapReduce程序学习指南：

<http://dblab.xmu.edu.cn/blog/hadoop-build-project-using-eclipse/>

## Hive数据仓库

学习指南：

<http://dblab.xmu.edu.cn/blog/1080-2/>

## Spark

Spark下载地址：

<http://spark.apache.org/downloads.html>

Spark1.6下载地址：

<http://www.apache.org/dyn/closer.lua/spark/spark-1.6.0/spark-1.6.0-bin-without-hadoop.tgz>

学习指南：

<http://dblab.xmu.edu.cn/blog/778-2/>

Spark安装与基本使用：

<http://dblab.xmu.edu.cn/blog/spark-quick-start-guide/>

Scala语言教程：

<http://dblab.xmu.edu.cn/blog/spark/>

Ubuntu中安装Scala语句：

sudo apt-get install scala

Spark DataFrame API：

<http://spark.apache.org/docs/latest/api/scala/index.html#org.apache.spark.sql.DataFrame>

sbt下载地址：

<https://www.scala-sbt.org/download.html>

sbt-lunch.jar下载地址：

<http://www.java2s.com/Code/Jar/s/Downloadsbtlaunchjar.htm>