**Hadoop 2.x 深入浅出基础课程笔记之Lesson 01**

***内容提纲：***

1. What is Apache Hadoop？
2. Hadoop 生态系统简介
3. 编译Hadoop 2.2.0
4. Hadoop 2.X 目录结构
5. 导入源码至Eclipse

**1、什么是Hadoop？**

**“谁说大象不能跳舞？！**

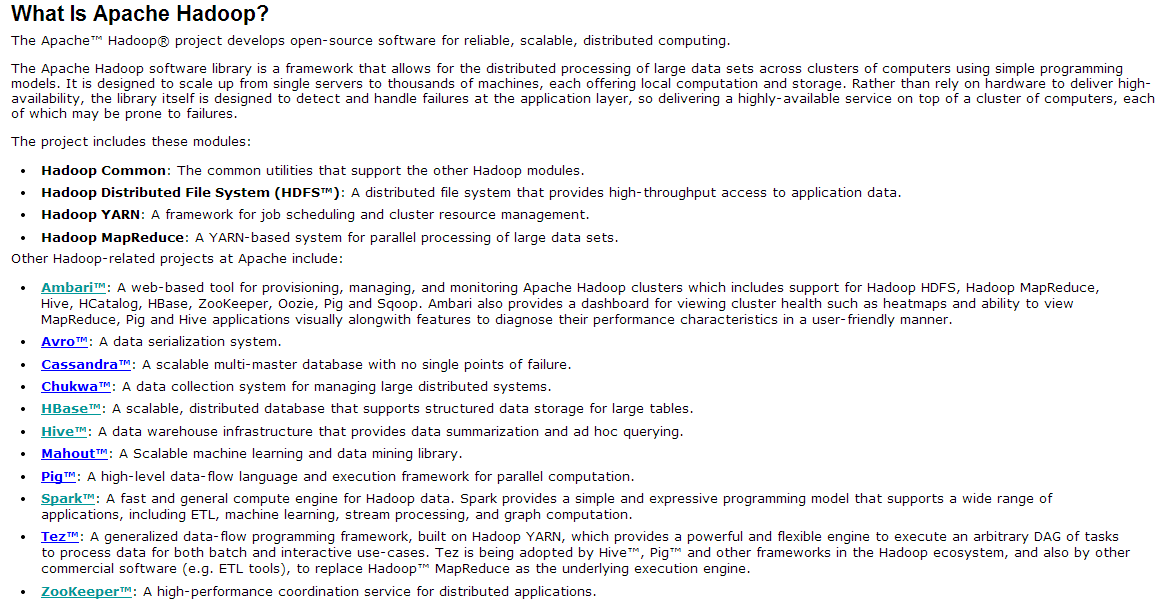
**Hadoop——轻松应对海量数据存储和分析所带来的挑战！“**

1. 海量数据**存储**，HDFS 分布式文件系统

2）海量数据**分析**，MapReduce 并行的离线计算框架

Apache Hadoop:

**http://hadoop.apache.org/**



**2、Hadoop 项目的四大模块**

* **Hadoop Common**: The common utilities that support the other Hadoop modules.
* **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
* **Hadoop YARN**: A framework for job scheduling and cluster resource management.
* **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

**2.1、YARN：”云的操作系统” -- Hortonworks （Hadoop 商业版本的）**

给部署在YARN上的应用，分配资源。

管理资源

JOB/APPLICATION 调度

Win7 ，XP：

QQ、YY、播放器等等，分配一些资源（CPU、内存、网络IO、磁盘IO等）

**2.2、技能**

1. **云计算，Hadoop 2.x**
2. **服务总线，SOA/OSB，Dubble**
3. **全文检索，Lucunce、Solr、Nutch**

**2.3、Hadoop 版本：**

1. 以0.20.x和1.x.x 版本为主
2. 0.23.x 版本为主
3. **2.x.x 版本为主**

**3、编译Hadoop 2.x 源码**

**3.1、环境：**

1）Linux 64 位操作系统，CentOS 6.4 版本，VMWare 搭建的虚拟机

2）虚拟机可以联网

**3.2、官方编译说明：**

Build instructions for Hadoop

----------------------------------------------------------------------------------

Requirements:

\* **Unix System**

**\* JDK 1.6+**

**\* Maven 3.0 or later**

**\* Findbugs 1.3.9 (if running findbugs)**

**\* ProtocolBuffer 2.5.0**

**\* CMake 2.6 or newer (if compiling native code)**

**\* Internet connection for first build (to fetch all Maven and Hadoop dependencies)**

----------------------------------------------------------------------------------

Maven main modules:

hadoop (Main Hadoop project)

- hadoop-project (Parent POM for all Hadoop Maven modules. )

(All plugins & dependencies versions are defined here.)

- hadoop-project-dist (Parent POM for modules that generate distributions.)

- hadoop-annotations (Generates the Hadoop doclet used to generated the Javadocs)

- hadoop-assemblies (Maven assemblies used by the different modules)

- hadoop-common-project (Hadoop Common)

- hadoop-hdfs-project (Hadoop HDFS)

- hadoop-mapreduce-project (Hadoop MapReduce)

- hadoop-tools (Hadoop tools like Streaming, Distcp, etc.)

- hadoop-dist (Hadoop distribution assembler)

**3.3、怎么讲Hadoop Project 导入到Eclipse**

**Importing projects to eclipse**

When you import the project to eclipse, install hadoop-maven-plugins at first.

$ cd hadoop-maven-plugins

$ mvn install

Then, generate eclipse project files.

$ mvn eclipse:eclipse -DskipTests

At last, import to eclipse by specifying the root directory of the project via

[File] > [Import] > [Existing Projects into Workspace].

**3.4、如何编译**

Building distributions:

Create binary distribution without native code and without documentation:

$ mvn package -Pdist -DskipTests -Dtar

Create binary distribution with native code and with documentation:

$ mvn package -Pdist,native,docs -DskipTests -Dtar

Create source distribution:

$ mvn package -Psrc -DskipTests

Create source and binary distributions with native code and documentation:

**$ mvn package -Pdist,native,docs,src -DskipTests -Dtar**

Create a local staging version of the website (in /tmp/hadoop-site)

$ mvn clean site; mvn site:stage -DstagingDirectory=/tmp/hadoop-site

**3.4、MAVEN国内镜像配置** 1、进入安装目录 /opt/modules/apache-maven-3.0.5/conf，编辑 settings.xml 文件

\* 修改<mirrors>内容：

<mirror>

<id>nexus-osc</id>

<mirrorOf>\*</mirrorOf>

<name>Nexus osc</name>

<url>http://maven.oschina.net/content/groups/public/</url>

</mirror>

\* 修改<profiles>内容：

<profile>

<id>jdk-1.6</id>

<activation>

<jdk>1.6</jdk>

</activation>

<repositories>

<repository>

<id>nexus</id>

<name>local private nexus</name>

<url>http://maven.oschina.net/content/groups/public/</url>

<releases>

<enabled>true</enabled>

</releases>

<snapshots>

<enabled>false</enabled>

</snapshots>

</repository>

</repositories>

<pluginRepositories>

<pluginRepository>

<id>nexus</id>

<name>local private nexus</name>

<url>http://maven.oschina.net/content/groups/public/</url>

<releases>

<enabled>true</enabled>

</releases>

<snapshots>

<enabled>false</enabled>

</snapshots>

</pluginRepository>

</pluginRepositories>

</profile>

2、复制配置

将该配置文件复制到用户目录，使得每次对maven创建时，都采用该配置

\* 查看用户目录【/home/hadoop】是否存在【.m2】文件夹，如没有，则创建

$ cd /home/hadoop

$ mkdir .m2

\* 复制文件

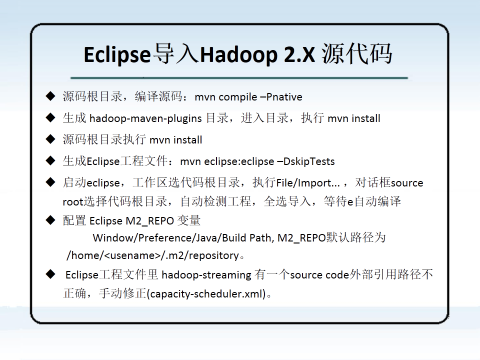
$ cp /opt/modules/apache-maven-3.0.5/conf/settings.xml ~/.m2/

**3.6、配置DNS**

修改： vi /etc/resolv.conf

nameserver 8.8.8.8

nameserver 8.8.4.4



额外的包：

