初识Hadoop 2.x

谭唐华

课程大纲

- Hadoop 2.x部分
- Hadoop 2.x生态系统部分
- Spark 1.x部分
- 实战商城离线分析项目

Hadoop部分

- 初识Hadoop
- 深入Hadoop
- 高级Hadoop
- Hadoop实战应用

Hadoop生态系统部分

- 数据仓库Hive
- 数据库Hbase
- 数据转换Sqoop
- 日志收集Flume
- 任务调度Oozie
- 图表显示Hub

Spark部分

初识Hadoop 2.x 大纲

- 大数据发展及背景
- Hadoop2.x 由来概述
- Hadoop2.x 生态圈
- Hadoop2.x 环境搭建

前言

马云留言

- 过去7年我们从互联网创业到互联网产业，很快进入互联网经济，而且正在从IT走向DT时代，也许昨天称为IT领袖峰会，未来要称DT领袖峰会，DT不仅仅是技术提升，而是思想观念的提升。DT和IT时代区别，IT以我为中心，DT以别人为中心，DT要让企业越来越强大，让你员工强大。DT越来越讲究开放、透明。我们所有企业都要思考什么样的文化、什么样的组织、什么样的人才才能适应未来DT时代，相信整个DT时代到来，在海外这被称为D经济。

大数据基本特征

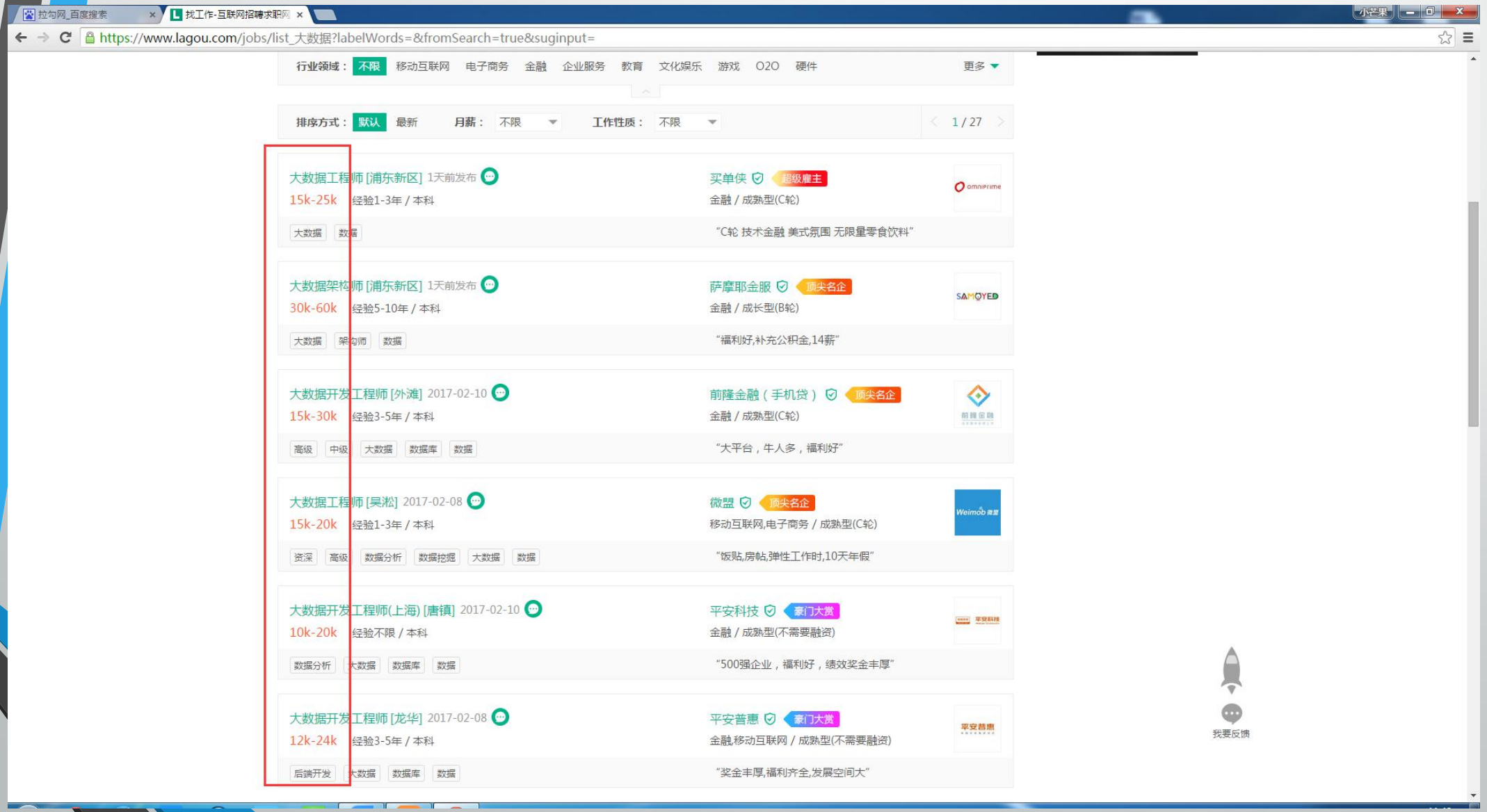
- 大容量
- 多样性
- 快速度
- 真实性

大数据的应用及发展前景

- 大数据时代带给我们的思考
  - 大数据计算提高数据处理效率，增加人类认知盈余
  - 大数据通过全局的数据让人类了解事物背后的真相
  - 大数据有助于了解事物发展的客观规律，利于科学决策
  - 大数据提供了同事物的连接，客观了解人类行为
  - 大数据改变过去的经验思维，帮助人们建立数据思维
- 大数据的企业应用场景
  - 1）医疗行业
  - 2）生物技术
  - 3）金融行业
  - 4）零售行业
  - 5）电商
  - 6）农牧业

  大数据的政府应用

大数据工资情况


拉勾网_百度搜索
找工作-互联网招聘求职网
https://www.lagou.com/jobs/list_大数据?labelWords=&fromSearch=true&suginput=
行业领域： 不限 移动互联网 电子商务 金融 企业服务 教育 文化娱乐 游戏 O2O 硬件 更多
排序方式： 默认 最新 月薪： 不限 工作性质： 不限 1 / 27
大数据工程师 [浦东新区] 1天前发布
15k-25k 经验1-3年 / 本科
买单侠 超级雇主
金融 / 成熟型(C轮)
大数据 数据
"C轮 技术金融 美式氛围 无限量零食饮料"
大数据架构师 [浦东新区] 1天前发布
30k-60k 经验5-10年 / 本科
萨摩耶金服 顶尖名企
金融 / 成长型(B轮)
大数据 架构师 数据
"福利好,补充公积金,14薪"
大数据开发工程师 [外滩] 2017-02-10
15k-30k 经验3-5年 / 本科
前隆金融（手机贷） 顶尖名企
金融 / 成熟型(C轮)
高级 中级 大数据 数据库 数据
"大平台，牛人多，福利好"
大数据工程师 [吴淞] 2017-02-08
15k-20k 经验1-3年 / 本科
微盟 顶尖名企
移动互联网,电子商务 / 成熟型(C轮)
资深 高级 数据分析 数据挖掘 大数据 数据
"饭贴,房帖,弹性工作时,10天年假"
大数据开发工程师(上海) [唐镇] 2017-02-10
10k-20k 经验不限 / 本科
平安科技 豪门大赏
金融 / 成熟型(不需要融资)
数据分析 大数据 数据库 数据
"500强企业，福利好，绩效奖金丰厚"
大数据开发工程师 [龙华] 2017-02-08
12k-24k 经验3-5年 / 本科
平安普惠 豪门大赏
金融,移动互联网 / 成熟型(不需要融资)
后端开发 大数据 数据库 数据
"奖金丰厚,福利齐全,发展空间大"
我要反馈

职位需求


https://www.lagou.com/jobs/2462445.html
拉勾
首页
公司
一拍
言职
大鲲
买单侠招聘
大数据工程师
超级雇主
15k-25k /上海 / 经验1-3年 / 本科及以上 / 全职
大数据
数据
1天前 发布于拉勾网
收藏
投个简历
完善在线简历
上传附件简历
职位诱惑：
C轮 技术金融 美式氛围 无限量零食饮料
职位描述：
工作职责： 1） 参与Hadoop大数据平台的设计与开发，解决海量数据面临的挑战； 2） 管理、优化并维护Hadoop集群，保证集群规模持续、稳定； 3） 负责Hadoop/Spark的功能扩展和性能优化，解决并实现业务需求； 4）协助建立数据模型，对数据挖掘脚本进行优化； 5）协助完成ETL开发工作。 职位要求： 1） 1-3年Hadoop/Spark工作经验，本科或以上学历，计算机相关专业； 2） 精通JAVA语言，熟悉Linux开发环境，具有实际系统开发经验； 3） 熟悉/Hive/Hbase/Spark/flume/storm/sqoop,具有实际开发经验； 4） 具备Java/Scala等开发经验； 5） 具有很强的学习能力、钻研精神、较强的沟通能力以及团队精神。
工作地址
上海 - 浦东新区 - 东方路1215弄陆家嘴软件园4号楼4层
查看地图
职位发布者:
Sherry
HRBP
聊天意愿
很弱
回复率 -- 用时1分钟
简历处理
很慢
处理率5% 用时7天
活跃时段
全天
早11点最活跃
面试评价
omniprime
买单侠
金融
C轮
红杉，策源(A轮)，真格基金(天使轮)，顺为资本、京东金融、晨兴创投、人人公司等(C轮)，人人网，红杉，策源(B轮)
500-2000人
http://www.fenqi.im
相似职位
大数据工程师
15k-25k
买单侠 [上海·东方路]
大数据工程师
我要反馈
在拉勾
发现新的职业机会
161,173 公司
2,740,495 职位
登录
极速注册 →

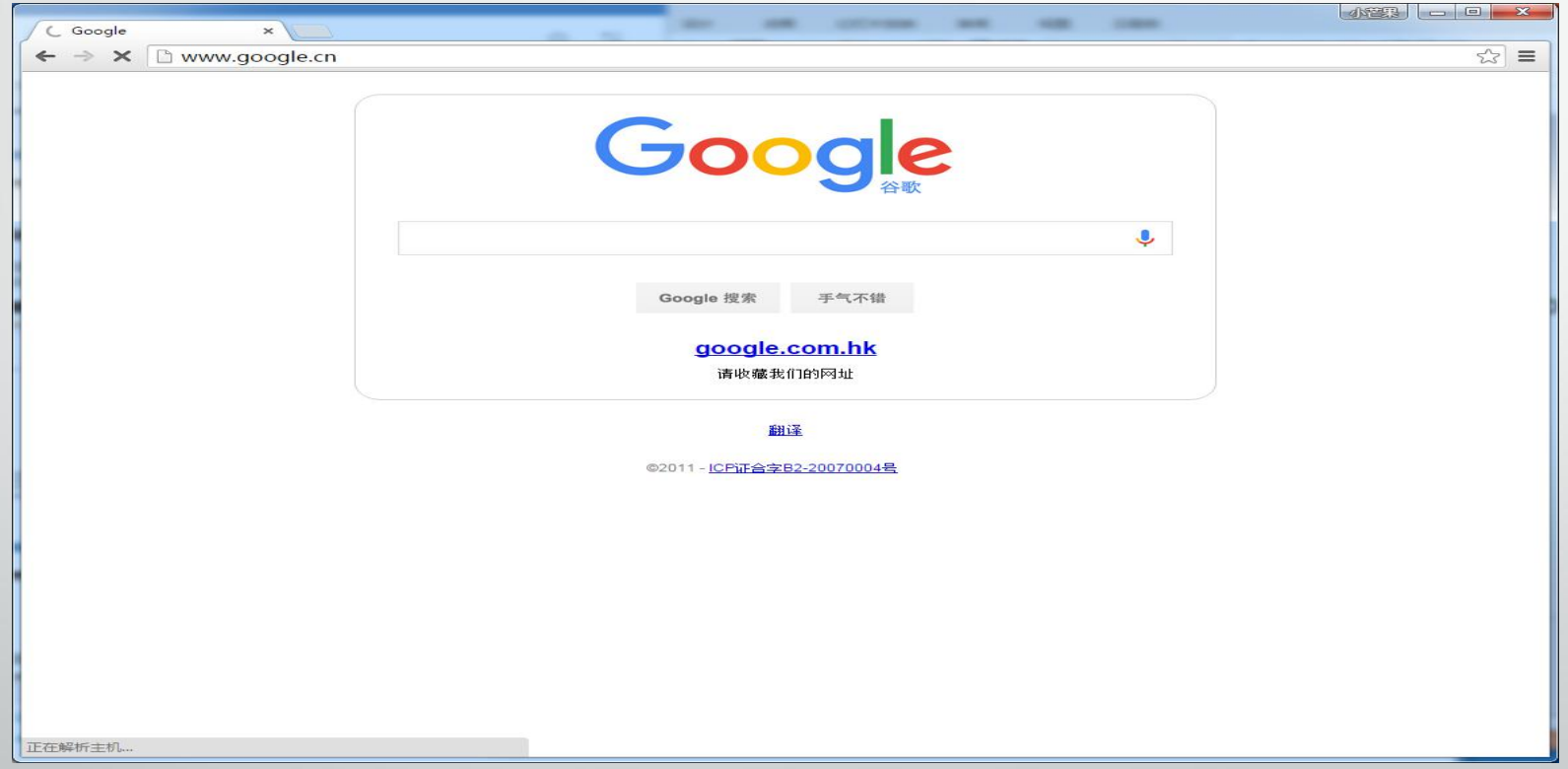
初识Hadoop 2.x 大纲

- 大数据发展及背景
- Hadoop2.x 由来概述
- Hadoop2.x 生态圈
- Hadoop2.x 环境搭建

大数据之Hadoop

- Hadoop是一个开源的、可靠的、可扩展的分布式并行计算框架
- 主要组成：分布式文件系统HDFS和MapReduce计算模型
- 作者：Doug Cutting
- 语言：Java，支持多种编程语言，如Python、C++

Hadoop思想起源：Google



Google的低成本之道

- 不使用超级计算机，不使用存储（淘宝的去i，去e，去o之路）
- 大量使用普通的PC服务器（去掉机箱，外设，硬盘），提供有冗余的集群服务
- 全世界多个数据中心，有些附带发电厂
- 运营商向Google倒付费

Google面对的数据和计算难题

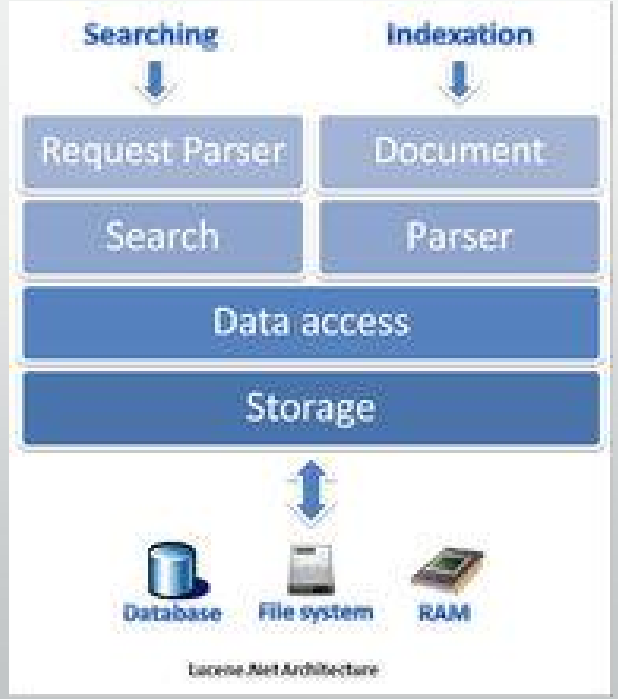
- 大量的网页如何存储？
- 搜索算法
- Page-Rank计算问题

Google大数据框架

- GFS
- Map-Reduce
- BigTable

Hadoop的起源--Lucene

- Lucene最初是由Doug Cutting开发的，在SourceForge的网站上提供下载。在2001年9月做为高质量的开源Java产品加入到Apache软件基金会的Jakarta家族中。随着每个版本的发布，这个项目得到明显的增强，也吸引了更多的用户和开发人员。



Lucene到nutch，nutch到hadoop

- 2003-2004年，Google公开了部分GFS和Mapreduce思想的细节，以此为基础Doug Cutting等人用了2年业余时间实现了DFS和Mapreduce机制，使Nutch性能飙升。
- Yahoo招安Doug Cutting及其项目Hadoop 于 2005 年秋天作为 Lucene的子项目 Nutch的 一部分正式引入Apache基金会。
- 2006 年 3 月份，Map-Reduce 和 Nutch Distributed File System (NDFS) 分别被纳入称为 Hadoop 的项目中，名字来源于Doug Cutting儿子的玩具大象。

hadoop达到高度

- 实现云计算的事实标准开源软件。
- 包含数十个具有强大生命力的子项目。
- 已经能在数千节点上运行，处理数据量和排序时间不断打破世界纪录。

Hadoop与Google对标

Hadoop	Google
HDFS	GFS
MapReduce	Map-Reduce
Hbase	BigTable

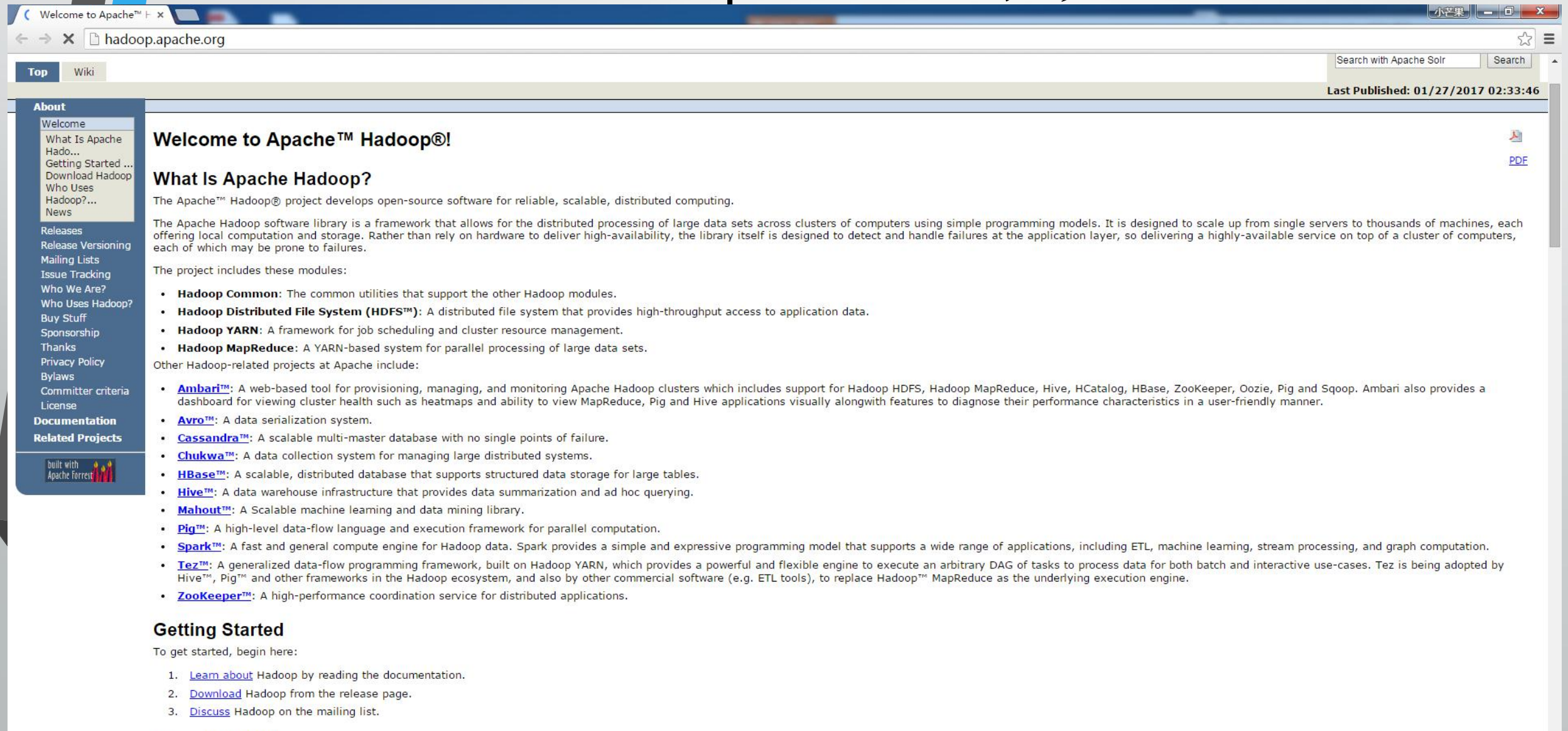
Hadoop的何去何从

- Nutch->Hadoop1.x->Hadoop2.x->Spark1.x->Spark2.x->Hadoop生态圈
- 轻量化->重量化
- 单台化->集群化
- 硬盘化->内存化

初识Hadoop 2.x 大纲

- 大数据发展及背景
- Hadoop2.x 由来概述
- Hadoop2.x 生态圈
- Hadoop2.x 环境搭建

Hadoop2.x生态系统



什么是Hadoop

- Hadoop项目主要包括以下四个模块
  - **Hadoop Common**:

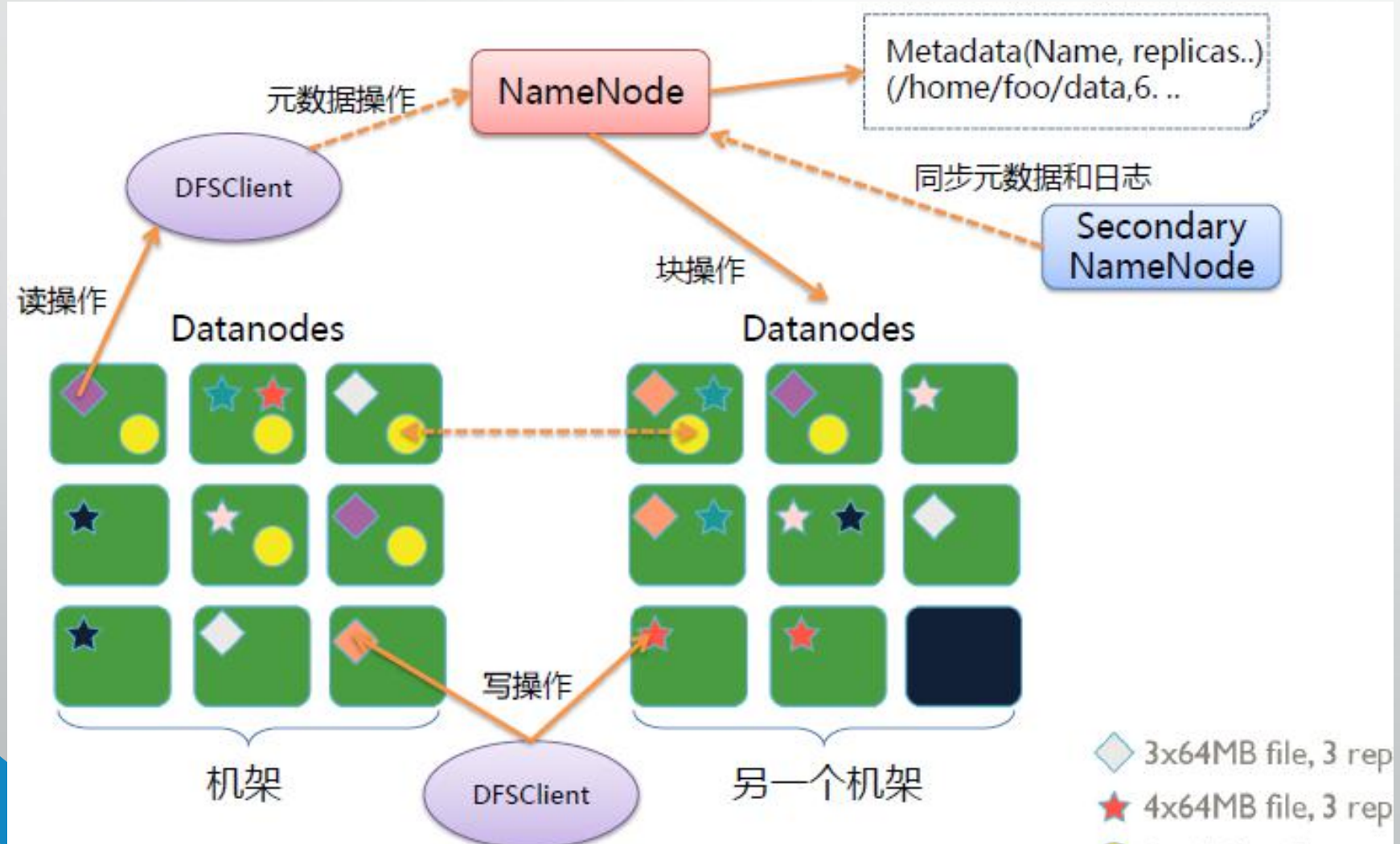
    为其他Hadoop模块提供基础设施。
  - **Hadoop HDFS**:

    一个高可靠、高吞吐量的分布式文件系统
  - **Hadoop MapReduce**:

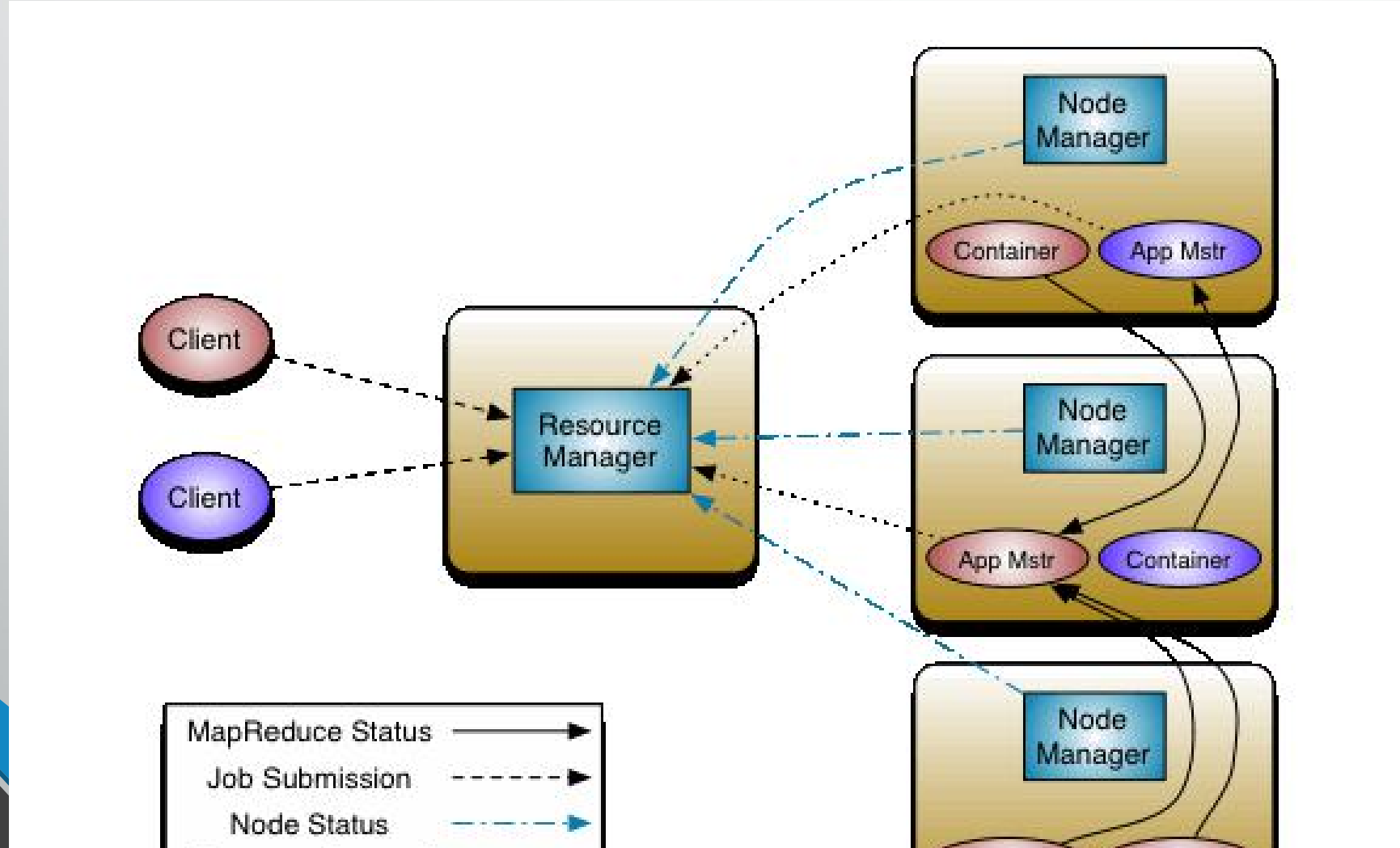
    一个分布式的离线并行计算框架
  - **Hadoop YARN**:

    一个新的MapReduce框架，任务调度与资源管理

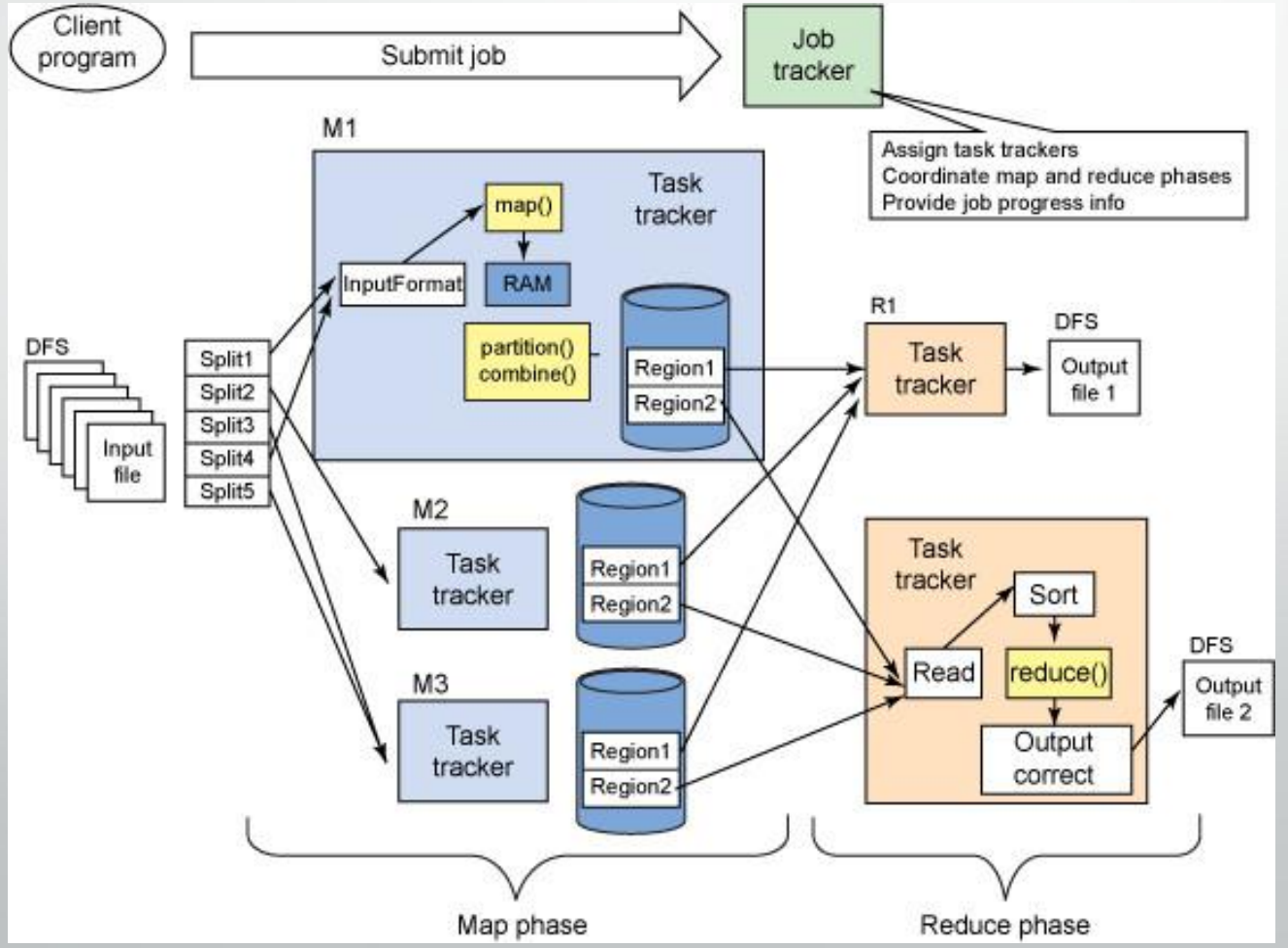
HDFS架构图


Metadata(Name, replicas..)
(/home/foo/data,6. ..
元数据操作
NameNode
DFSClient
同步元数据和日志
Secondary
NameNode
读操作
块操作
Datanodes
Datanodes
写操作
机架
另一个机架
DFSClient
3x64MB file, 3 rep
4x64MB file, 3 rep

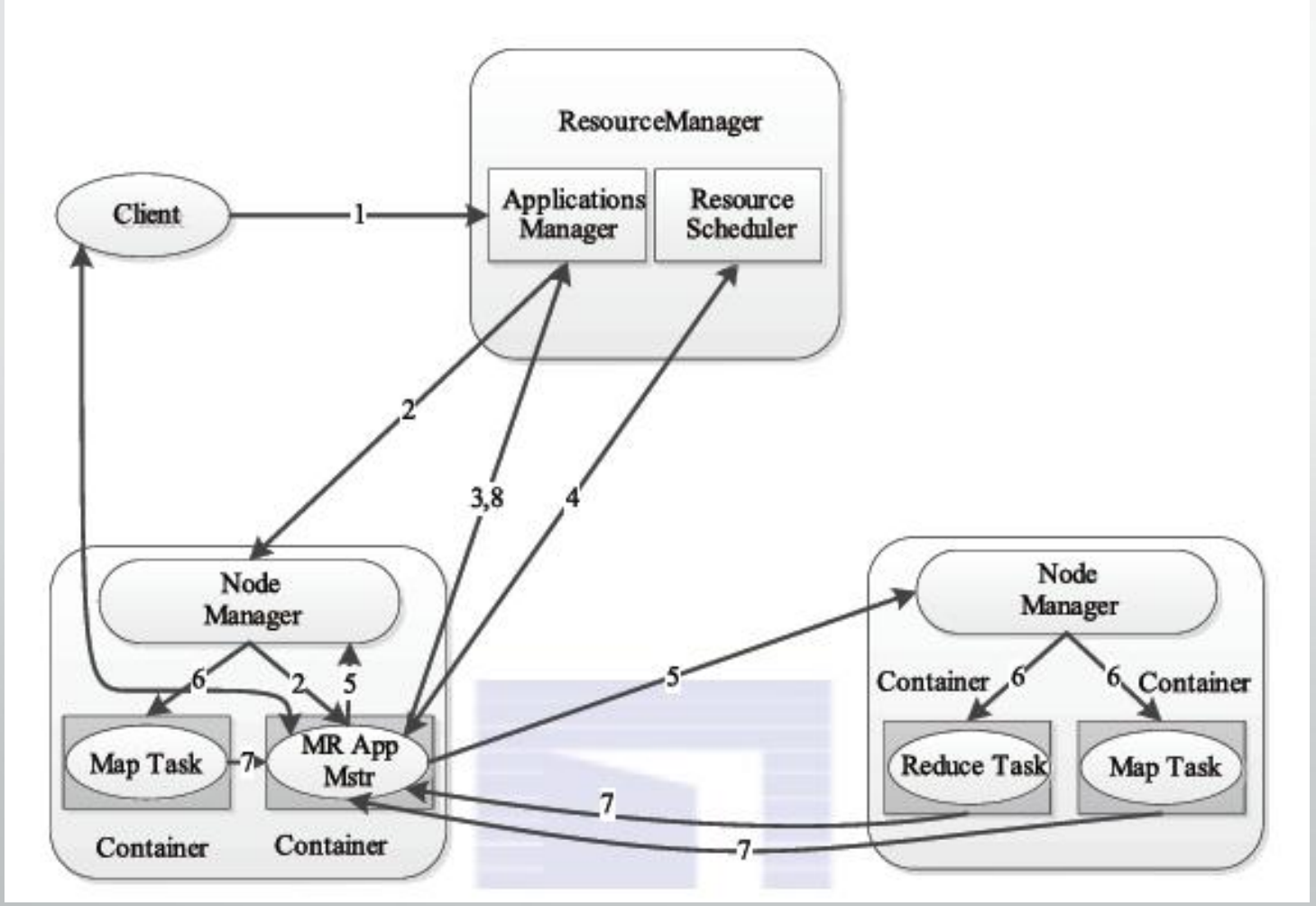
YARN架构图


Node Manager
Container
App Mstr
Client
Client
Resource Manager
Node Manager
App Mstr
Container
Node Manager
MapReduce Status
Job Submission
Node Status

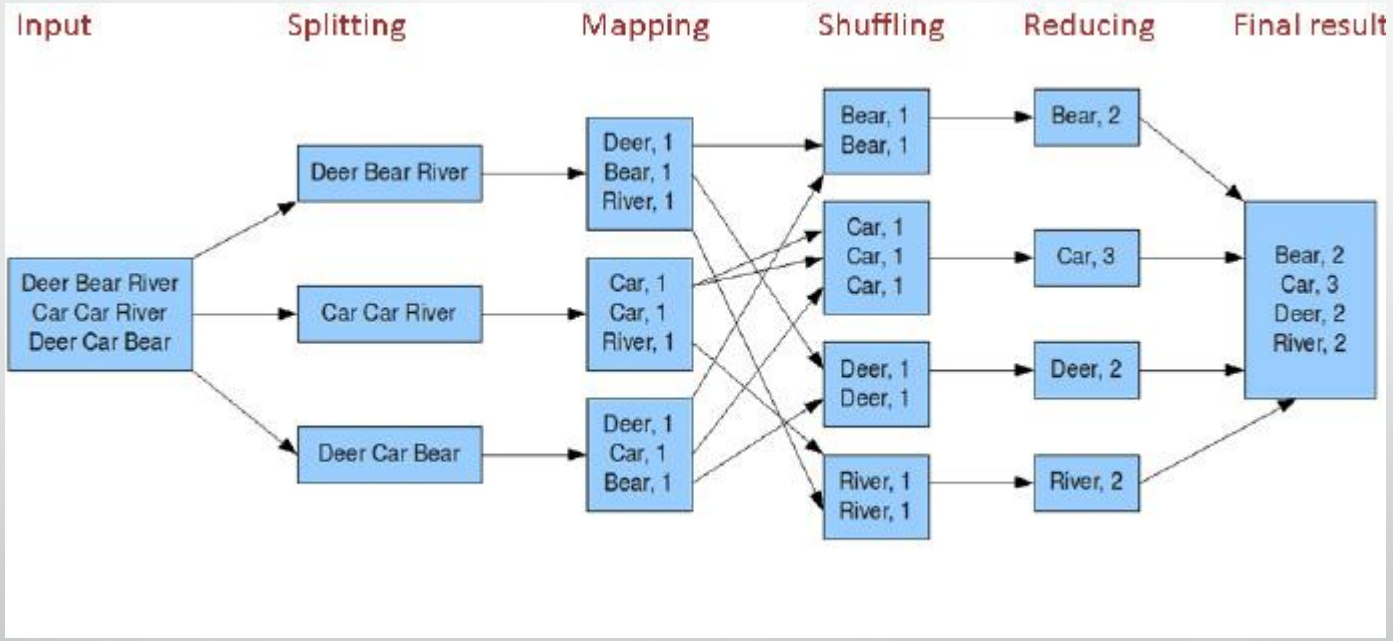
MapReduce架构图



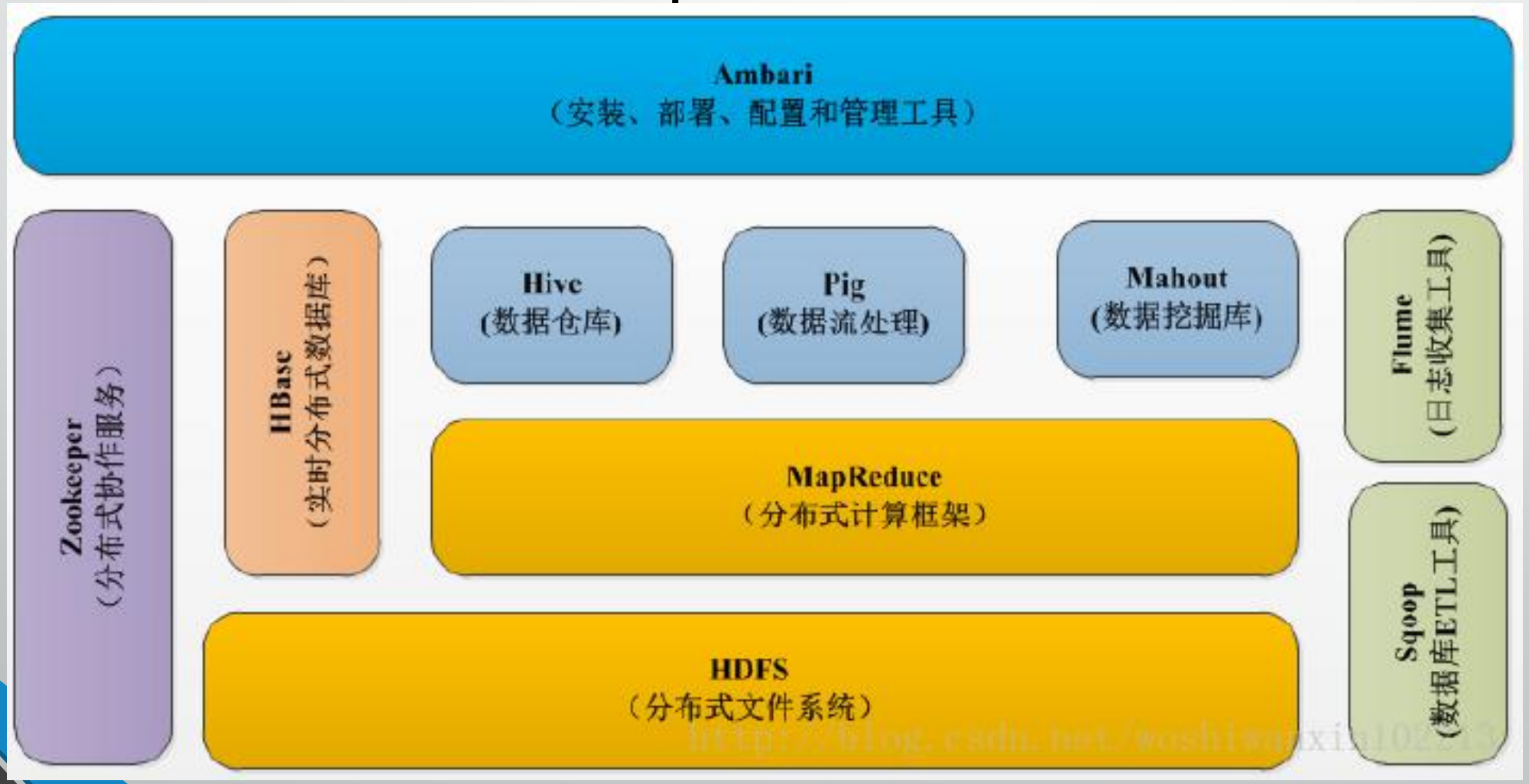
MapReduce On YARN



MapReduce工作流程



Hadoop2.x生态系统



初识Hadoop 2.x 大纲

- 大数据发展及背景
- Hadoop2.x 由来概述
- Hadoop2.x 生态圈
- Hadoop2.x 环境搭建

Hadoop2.x系统安装方式

- 单机模式
- 单机伪分布模式
- 集群分布模式

Hadoop部署前准备

- jdk1.7
- centos6.5
- hadoop2.x
- ssh

单机模式安装基本步骤

- 准备操作系统
- Centos6.4（64bit）
- 主机名、IP地址
- 普通用户
- 防火墙、selinux
- 主机名映射
- 安装JDK
- 修改Hadoop的相关配置文件
- hadoop-env.sh、core-site.xml、hdfs-site.xml
- yarn-env.sh、mapred-env.sh
- yarn-site.xml、mapred-site.xml

单机模式部署

- 安装jdk
- 安装hadoop（不用hdfs）
- 安装hadoop（用hdfs）
- 安装hadoop（运行在yarn）

日志聚集部署

- 历史服务器
- 查看已经运行完成的MapReduce作业记录，比如用了多少个Map、用了多少个Reduce、作业提交时间、作业启动时间、作业完成时间等信息。
- 默认情况下，历史服务器是没有启动的。

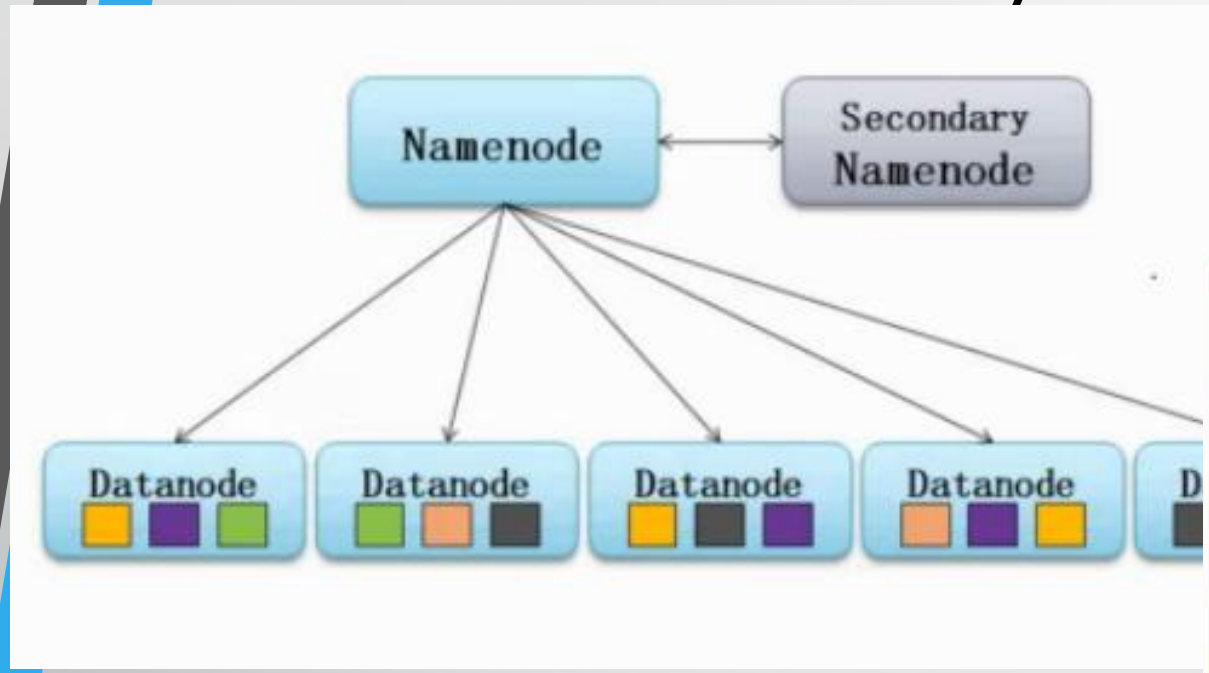
hadoop配置文件介绍

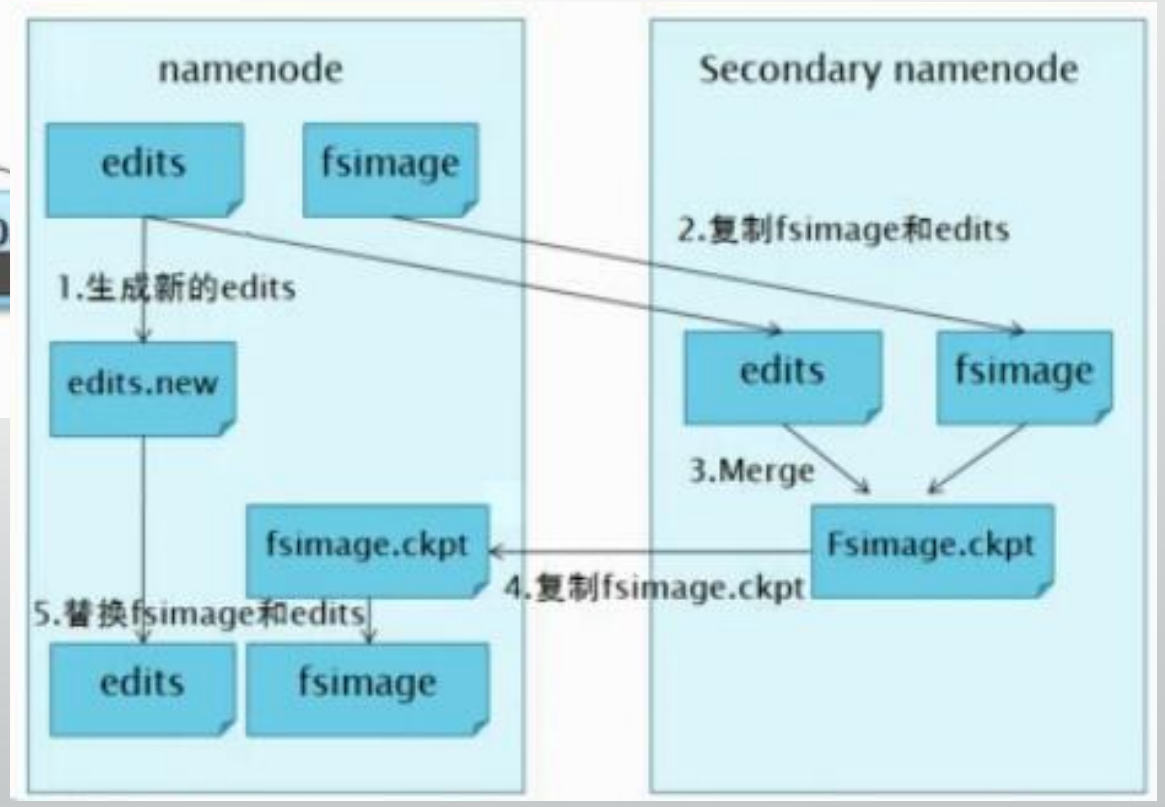
- 默认配置文件

  4个*-default.xml
- 自定义配置文件

  4个*-site.xml

SecondaryNameNode架构图





SecondaryNameNode配置

- 辅助节点配置
- 历史日志服务器地址配置

总结-HDFS模式启动

- 格式化NameNode
- 启动NameNode
- 启动DataNode
- HDFS监控web页面
- 启动SecondaryNameNode
- SecondaryNameNode监控web页面

总结-YARN模式启动

- 启动ResourceManager
- 启动NodeManager
- 查看启动守护进程
- 查看日志
- yarn监控web页面

启动HDFS和YARN的方式

- 方式一：逐一启动

  hadoop-daemon.sh, yarn-daemon.sh

- 方式二：分开启动

  start-dfs.sh, start-yarn.sh

- 方式三：全部启动

  start-all.sh

hadoop2.x 本地编译

- * 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)
- * Zlib devel (if compiling native code)
- * openssl devel (if compiling native hadoop-pipes)
- * Internet connection for first build (to fetch all Maven and Hadoop dependencies)