# 大数据简介(Hadoop)

## Hadoop是什么

1）Hadoop是一个由Apache基金会所开发的分布式系统基础架构。

2）主要解决，海量数据的存储和海量数据的分析计算问题。

3）广义上来说，HADOOP通常是指一个更广泛的概念——HADOOP生态圈。

## Hadoop发展历史

1）Lucene是Doug Cutting开创的开源软件，用java书写代码，实现与Google类似的全文搜索功能，它提供了全文检索引擎的架构，包括完整的查询引擎和索引引擎

2）2001年年底Lucene成为Apache基金会的一个子项目

3）对于大数量的场景，Lucene面对与Google同样的困难

4）学习和模仿Google解决这些问题的办法 ：微型版Nutch

5）可以说Google是hadoop的思想之源(Google在大数据方面的三篇论文)

GFS --->HDFS

Map-Reduce --->MR

BigTable --->Hbase

6）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

7）2005 年Hadoop 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

8）名字来源于Doug Cutting儿子的玩具大象，如图2-20

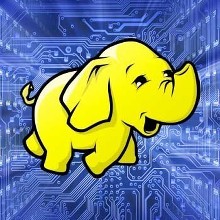


图2-20 Hadoop的logo

9）Hadoop就此诞生并迅速发展，标志着云计算时代来临

## Hadoop的优势

1）高可靠性：Hadoop底层维护多个数据副本，所以即使Hadoop某个计算元素或存储出现故障，也不会导致数据的丢失。

2）高扩展性：在集群间分配任务数据，可方便的扩展数以千计的节点。

3）高效性：在MapReduce的思想下，Hadoop是并行工作的，以加快任务处理速度。

4）高容错性：能够自动将失败的任务重新分配。

## Hadoop组成

在Hadoop1.x时代，Hadoop中的MapReduce同时处理业务逻辑运算和资源的调度，耦合性较大，在Hadoop2.x时代，增加了Yarn。Yarn只负责资源的调度，MapReduce只负责运算，如图2-21所示。

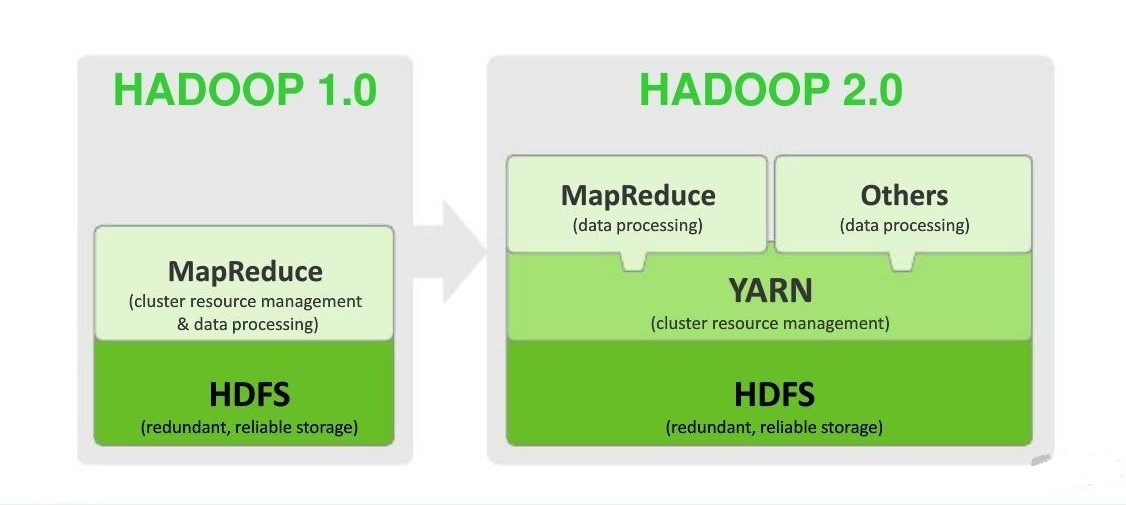
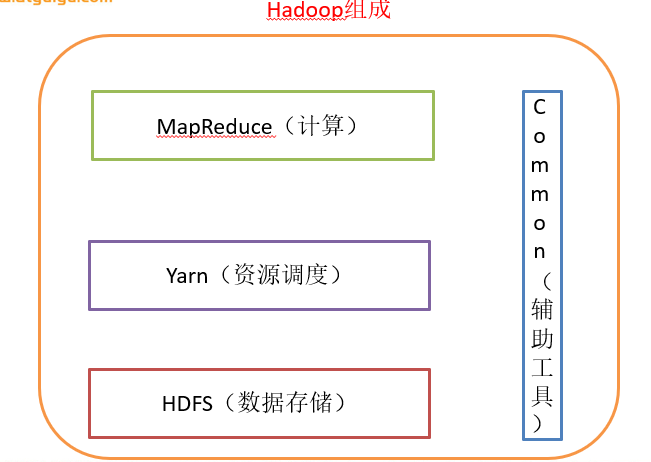
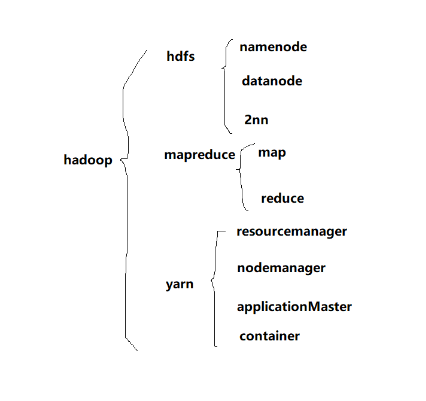


图2-21 Hadoop1.x与Hadoop2.x的区别





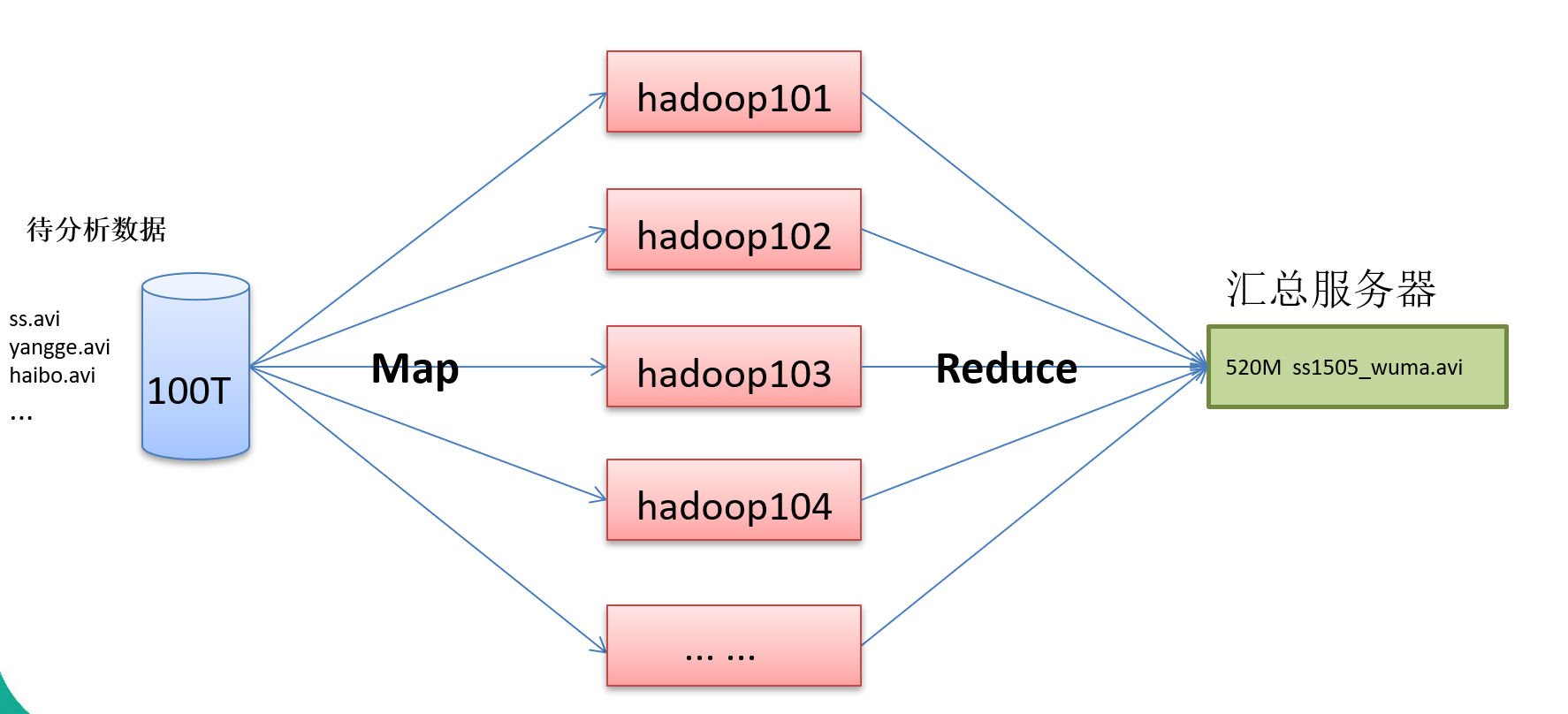
### HDFS架构概述

1. NameNode（nn）：存储文件的元数据，如文件名，文件目录结构，文件属性（生成时间、副本数、文件权限），以及每个文件的块列表和块所在的DataNode等。
2. DataNode(dn)：在本地文件系统存储文件块数据，以及块数据的校验和。
3. Secondary NameNode(2nn)：用来监控HDFS状态的辅助后台程序，每隔一段时间获取HDFS元数据的快照。

### MapReduce架构概述

1）Map阶段并行处理输入数据

2）Reduce阶段对Map结果进行汇总



### YARN架构概述

1）ResourceManager(rm):处理客户端请求,启动/监控applicationMaster,监控NodeManager,资源的调度与分配.

2）NodeManager(nm):单个节点上的资源管理,处理来自ResourceManager的命令,处理来自applicationMaster的命令

3) ApplicationMaster:数据切分,为应用程序申请资源,并分配给内部任务,任务监控与容错

4) Container:对任务运行环境的抽象,封装了CPU,内存等多维资源以及环境变量,启动命令等任务运行的相关信息

大数据技术生态体系

