1、hadoop 的介绍以及发展历史

1. Hadoop **最早起源于 Nutch**。Nutch 的设计目标是构建一个大型的全网搜索引擎，包括网页抓取、索引、查询等功能，但随着抓取网页数量的增加，**遇到了严重的可扩展性问题——**如何解决数十亿网页的存储和索引问题。

2. 2003 年、2004 年**谷歌发表的两篇论文为该问题提供了可行的解决方案**。

——分布式文件系统（GFS），可用于处理海量网页的**存储**

——分布式计算框架 MAPREDUCE，可用于处理海量网页的**索引计算**问题。

3. Nutch 的开发人员完成了相应的**开源实现 HDFS 和 MAPREDUCE**，并从 Nutch 中剥离成为独立项目 HADOOP，到 2008 年 1 月，HADOOP 成为 Apache 顶级项目(同年，cloudera 公司成立)，迎来了它的快速发展期。

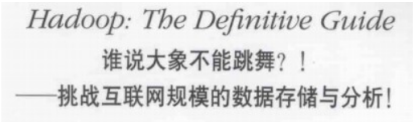
狭义上来说，hadoop 就是单独指代 hadoop 这个软件，

广义上来说，hadoop 指代大数据的一个生态圈，包括很多其他的软件

Hadoop的概念、版本、发展史

Hadoop是什么？
Hadoop的起源
Hadoop发展史
Hadoop的四大特性（优点）
Hadoop的版本
如何选择Hadoop版本

Hadoop是什么？



Hadoop： **适合大数据的分布式存储和计算平台**
Hadoop不是指具体一个框架或者组件，它是Apache软件基金会下用Java语言开发的一个开源分布式计算平台。实现在大量计算机组成的集群中对海量数据进行分布式计算。适合大数据的分布式存储和计算平台。
Hadoop1.x中包括两个核心组件：MapReduce和Hadoop Distributed File System(HDFS)
其中HDFS负责将海量数据进行分布式存储，而MapReduce负责提供对数据的计算结果的汇总

Hadoop的起源

2003-2004年，Google公布了部分GFS和MapReduce思想的细节，受此启发的Doug Cutting等人用2年的业余时间实现了DFS和MapReduce机制，使Nutch性能飙升。然后Yahoo招安Doug Gutting及其项目。
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2月被分离出来，成为一套完整独立的软件，起名为Hadoop
Hadoop名字不是一个缩写，而是一个生造出来的词。是Hadoop之父Doug Cutting儿子毛绒玩具象命名的。
Hadoop的成长过程
Lucene–>Nutch—>Hadoop

总结起来，Hadoop起源于Google的三大论文
GFS：Google的分布式文件系统Google File System
MapReduce：Google的MapReduce开源分布式并行计算框架
BigTable：一个大型的分布式数据库

演变关系
GFS—->HDFS
Google MapReduce—->Hadoop MapReduce
BigTable—->HBase

Hadoop发展史

Hadoop大事记
2004年— 最初的版本(现在称为HDFS和MapReduce)由Doug Cutting和Mike Cafarella开始实施。
2005年12月— Nutch移植到新的框架，Hadoop在20个节点上稳定运行。
2006年1月— Doug Cutting加入雅虎。
2006年2月— Apache Hadoop项目正式启动以支持MapReduce和HDFS的独立发展。
2006年2月— 雅虎的网格计算团队采用Hadoop。
2006年4月— 标准排序(10 GB每个节点)在188个节点上运行47.9个小时。
2006年5月— 雅虎建立了一个300个节点的Hadoop研究集群。
2006年5月— 标准排序在500个节点上运行42个小时(硬件配置比4月的更好)。
2006年11月— 研究集群增加到600个节点。
2006年12月— 标准排序在20个节点上运行1.8个小时，100个节点3.3小时，500个节点5.2小时，900个节点7.8个小时。
2007年1月— 研究集群到达900个节点。
2007年4月— 研究集群达到两个1000个节点的集群。
2008年4月— 赢得世界最快1TB数据排序在900个节点上用时209秒。
2008年7月— 雅虎测试节点增加到4000个
2008年9月— **Hive成为Hadoop的子项目**
2008年11月— Google宣布其MapReduce用68秒对1TB的程序进行排序
2008年10月— 研究集群每天装载10TB的数据。
2008年— **淘宝开始投入研究基于Hadoop的系统–云梯**。云梯总容量约9.3PB，共有1100台机器，每天处理18000道作业，扫描500TB数据。
2009年3月— 17个集群总共24 000台机器。
2009年3月— **Cloudera推出CDH（Cloudera's Dsitribution Including Apache Hadoop）**
2009年4月— 赢得每分钟排序，雅虎59秒内排序500 GB(在1400个节点上)和173分钟内排序100 TB数据(在3400个节点上)。
2009年5月— Yahoo的团队使用Hadoop对1 TB的数据进行排序只花了62秒时间。
2009年7月— **Hadoop Core项目更名为Hadoop Common;**
2009年7月— **MapReduce 和 Hadoop Distributed File System (HDFS) 成为Hadoop项目的独立子项目。**
2009年7月— **Avro 和 Chukwa 成为Hadoop新的子项目。**
2009年9月— 亚联BI团队开始跟踪研究Hadoop
2009年12月—亚联提出橘云战略，开始研究Hadoop
2010年5月— Avro脱离Hadoop项目，成为Apache顶级项目。
2010年5月— HBase脱离Hadoop项目，成为Apache顶级项目。
2010年5月— IBM提供了基于Hadoop 的大数据分析软件——InfoSphere BigInsights，包括基础版和企业版。
2010年9月— Hive(Facebook) 脱离Hadoop，成为Apache顶级项目。
2010年9月— Pig脱离Hadoop，成为Apache顶级项目。
2011年1月— **ZooKeeper 脱离Hadoop，成为Apache顶级项目。**
2011年3月— Apache Hadoop获得Media Guardian Innovation Awards 。
2011年3月— Platform Computing 宣布在它的Symphony软件中支持Hadoop MapReduce API。
2011年5月— Mapr Technologies公司推出分布式文件系统和MapReduce引擎——MapR Distribution for Apache Hadoop。
2011年5月— HCatalog 1.0发布。该项目由Hortonworks 在2010年3月份提出，HCatalog主要用于解决数据存储、元数据的问题，主要解决HDFS的瓶颈，它提供了一个地方来存储数据的状态信息，这使得 数据清理和归档工具可以很容易的进行处理。
2011年4月— SGI(Silicon Graphics International)基于SGI Rackable和CloudRack服务器产品线提供Hadoop优化的解决方案。