Hadoop 简介

什么是大数据（big data）

大数据（big data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。 在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》 中大数据指不用随机分析法（抽样调查）这样捷径，而采用所有数据进行分析处理。大数据的 5V 特点（IBM 提出）：Volume（大量）、Velocity（高速）、Variety（多样）、Value（低价值密度）、Veracity（真实性）。--百度百科

意思就是：反正很多，多到你用常规办法干不了那么多。

什么是 hadoop

hadoop 就是用来解决大数据问题的工具。

Hadoop 是一个由 Apache 开发的分布式系统基础架构。编程人员可以不需要了解分布式底层的情况下，开发分布式程序。充分利用集群来进行高效快速运算以及存储。

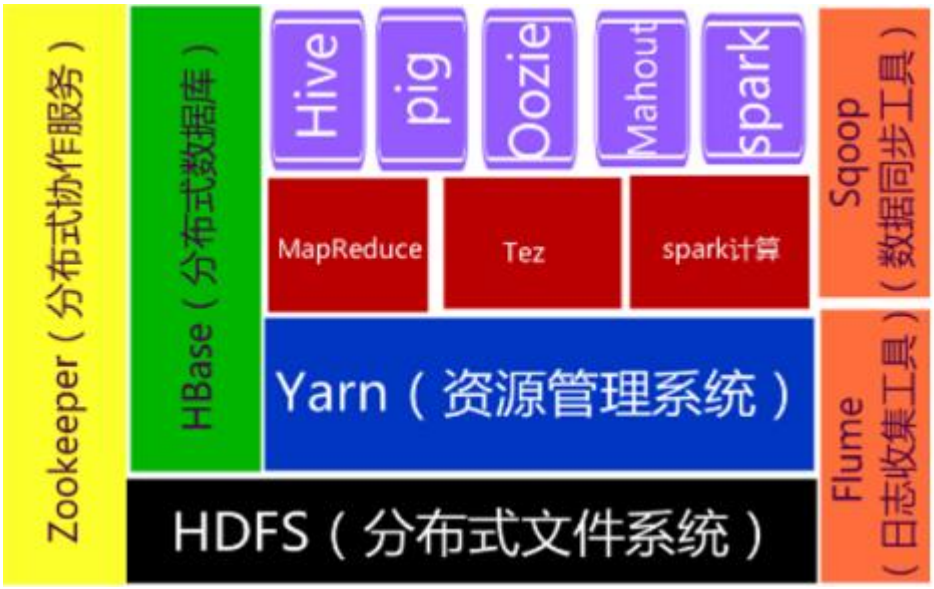
Hadoop 实现了一个分布式文件储存系统（Hadoop Distributed File System），简称 HDFS。

Hadoop 的框架的核心的设计是：HDFS 和 MapReduce。HDFS 用来存储数据(大量的数据)，而 MapReduce 是一个分布式计算框架，为大数据提供计算工作。

HDFS：Hadoop Distributed File System。用来存储大数据。它牛逼的地方是：把 N 台机器搞到一个文件系统管理，访问网络数据跟访问本地的体验一致。客户端丢给 hdfs 的文件它会自动分片放到集群之中。通过配置还可以自动备份。

MapReduce(MR)：它是一个分布式计算框架。MR 的主体思想就是分而治之。这个算法主要经历 3 步：拆分大问题—解决所有小问题—合并结果。MR 的主要工作其实就是把上诉过程自动化了。

hadoop 生态系统



- Hdfs：用来做存储的
- Yarn： Hadoop 集群的资源管理系统
- MapReduce:已经提过，负责计算任务，MapReduce 处理时有较高时延问题。
- TeZ:TeZ 是一个运行在 Yarn 之上的项目，主要是对 MapReduce 进行更细的拆分 ，然后以有向图（DAG）的方式重新组织来完成计算任务。通过 TeZ 可将有依赖关系的作业转化为一个任务，减少 hdfs 的写操作和一些中间环节，提高作业的效率。

- Spark:也是分布式计算框架，比 MapReducer 抽像化程度更高，可以简单理解为 MapReduce 为手工作坊式工作，Spark 为流水线式的现代化作业。两者各有自己的使用场景。
- Hbase：是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。
- Hive：基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，通过类 SQL 语句快速实现简单的 MapReduce 统计，不必单独写 mr 代码。
- Pig：编写 MR 需要用 java 代码实现 map 和 reduce 方法。Pig，就是用另一种类似的 sql 的语言，写代码，这个代码会自动转换成 mapreduce。
- Oozie：是一个工作流引擎服务器, 用于管理和协调运行在 Hadoop 平台上 MR 任务。一般情况下，MR 的任务是在多个节点上并行运行的，是没有顺序的。比如想做：先抓取数据到 HDFSàMR 处理 à 存数据库，这样有顺序的事情，这个就可以用了。
- Mahout：一个数据挖掘工具，实现了一些机器学习的算法。
- Flume：采集日志的工具。
- Sqoop：hdfs 和关系型数据之间转换数据的工具。
- Zookeeper（ZK）：是一个开源的分布式协调服务。分布式应用程序可以基于 ZooKeeper 实现诸如数据发布/订阅、负载均衡、命名服务、分布式协调/通知、集群管理、Master 选举、分布式锁和分布式队列等功能。Hadoop 只有一个 namenode，当这个 namenode 挂了之后，需要手动启动 Secondary NameNode 去替代，通过 Zookeeper 可以搞几个 namenode，一个是活动的，其余的 standby。当主 namenode 挂了，立即另选一个 namenode 激活。

其它

- Ambari：创建、管理、监视 Hadoop 的集群的工具
- Hue：也是监控集群的，可以在界面上直接写 MR

hadoop 发展历史

- Hadoop 由 Apache Software Foundation 公司在 2005 年秋天作为 Lucene 的子项目 Nutch 的一部分正式引入。
- 2006 年 3 月份，Map/Reduce 和 NDFS 被纳入 Hadoop 的项目中。
- Cloudera 是美国的一家软件公司，该公司在 2008 年开始提供基于 Hadoop 的软件以及服务。
- GoGrid 是一家云计算基础设施公司，在 2012 年，GoGrid 与 Cloudera 合作基于 Hadoop 的软件以及服务，加速了 Hadoop 应用的步伐。

Google 三篇重要论文:

2003 年发布关于 GFS（google filesystem）的论文，基于此实现了 hdfs.

2004 年发布 mapreduce 的论文，基于此实现了 mapreduce 框架

2006 年发布了 big table 的论文，基于此实现了 hbase 数据库

公有云平台提供的 hadoop 服务

- AWS：Amazon EMR 框架产品,可以快速部署 hadoop 集群，进行 hadoop 集群的托管。

- 微软 azure 云 ：Azure HDInsight ,它是 Hadoop 组件的云分发版，可以通过 Azure HDInsight 轻松、快速且经济有效地处理大量数据。 可以使用 Hadoop、Spark、Hive、LLAP、Kafka、Storm、R 等最常用的开源框架。可以通过这些框架启用各种各样的方案，例如提取、转换和加载 (ETL)；数据仓库操作；机器学习；IoT。
- 阿里云:MaxCompute,它能提供快速、完全托管的 PB 级数据仓库解决方案,可以通过 MaxCompute Studio 与 IntelliJ IDEA 等开发工具进行整合使用。

hadoop 能做什么

1.通过 HDFS 实现海量数据存储

2. 海量日志分析

3. 批处理

这里的批处理是指通过提取、转换和加载，将非结构化或结构化数据从异类数据源中提取出来， 转换成某种结构化格式，然后加载到数据存储中的过程。可以将转换的数据用于数据科学或数据仓库。

2. 作为数据仓库

可以使用 hadoop 对任何格式的结构化或非结构化数据执行 PB 规模的交互式查询。比如配合 hive 实现数据仓库。

hadoop 应用领域

1) 移动数据

2) 能源开采

3) 在线旅游

4) 图像处理

5) IT 安全

6) 电子商务

7) 节能

8) 医疗保健

9) 诈骗检测

10) 基础架构管理

Hadoop 版本

1.x 比较老了，不建议使用

2.x 建议使用 2.9 以后版本

3.x 比较新，暂时不建议在生产环境中使用

3.x 有许多改进，最大一点就是采用 hdfs 存储采用厅偶校验 容错，可以更加节约空间。