处理和分析海量数据的神器——hadoop。

一、是什么

1、概念

Hadoop是一个开源的框架，可编写和运行分布式应用处理大规模数据，是专为离线和大规模数据分析而设计的，并不适合那种对几个记录随机读写的在线事务处理模式。

Hadoop 是以一种可靠、高效、可伸缩的方式进行处理的。

Hadoop 是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。

Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。

Hadoop 还是可伸缩的，能够处理 PB 级数据。

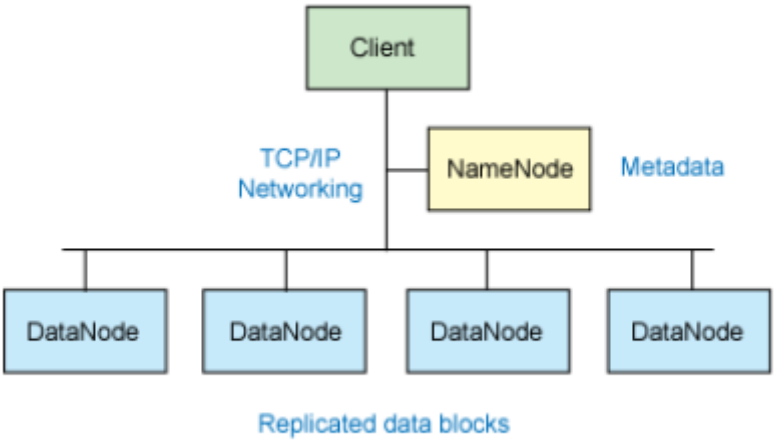
2、核心

Hadoop的核心就是HDFS和MapReduce，而两者只是理论基础，不是具体可使用的高级应用，Hadoop旗下有很多经典子项目，比如HBase、Hive等，这些都是基于HDFS和MapReduce发展出来的。

要想了解Hadoop，就必须知道HDFS和MapReduce是什么。

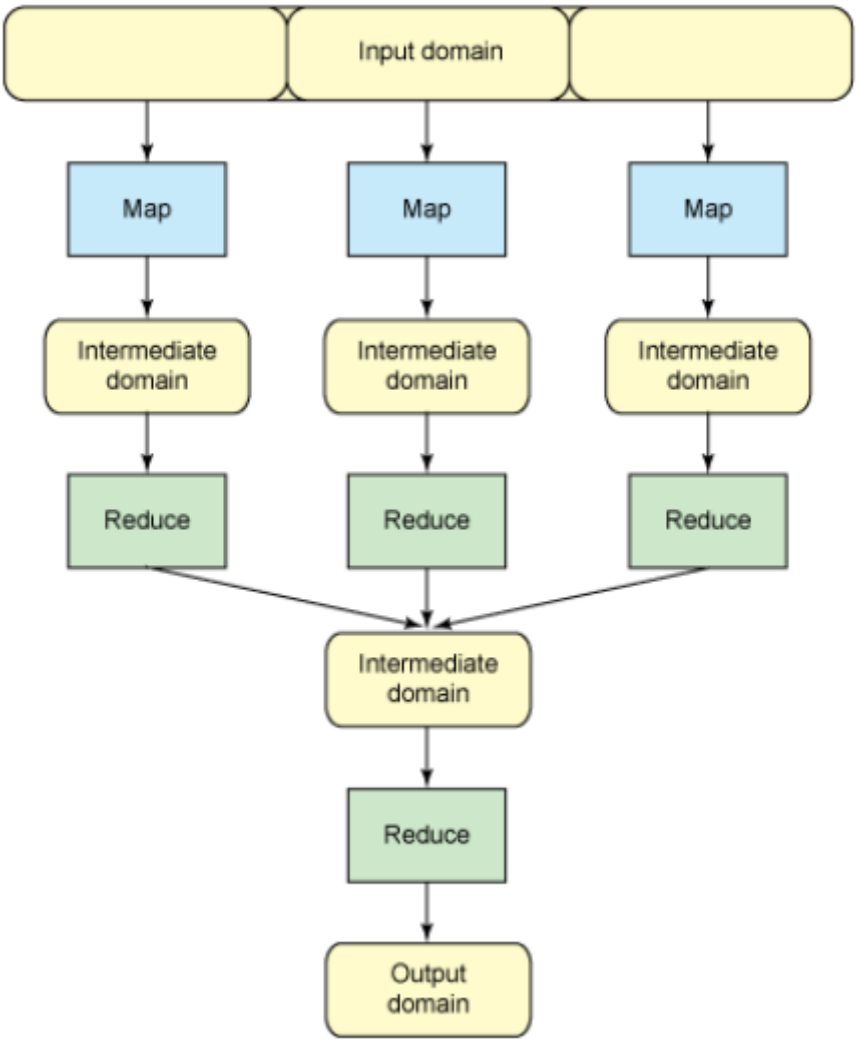
- **HDFS**

HDFS（Hadoop Distributed File System，Hadoop分布式文件系统），它是一个高度容错性的系统，适合部署在廉价的机器上。HDFS能提供高吞吐量的数据访问，适合那些有着超大数据集（large data set）的应用程序。



- **MapReduce**

Mapreduce是一个计算框架，一个处理分布式海量数据的软件框架及计算集群。



二、干什么

1、应用

1.搜索引擎（Doug Cutting 设计Hadoop的初衷，为了针对大规模的网页快速建立索引）。
2.大数据存储，利用Hadoop的分布式存储能力，例如数据备份、数据仓库等。
3.大数据处理，利用Hadoop的分布式处理能力，例如数据挖掘、数据分析等。
4.科学研究，Hadoop是一种分布式的开源框架，对于分布式计算有很大程度地参考价值。

2、优缺点

- **优点**

高可靠性。
Hadoop按位存储和处理数据的能力值得人们信赖。

高扩展性。
Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

高效性。
Hadoop能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

高容错性。
Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

低成本。
与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低。

- **缺点**

不适合低延迟数据访问。

无法高效存储大量小文件。

不支持多用户写入及任意修改文件。