Hadoop 介绍

侯在钱

目 录

1. Hadoop 简介

Hadoop 是一个运行在大规模廉价计算机上的分布式集群框架，它可以把几十台到几千台计算机连接在一起，形成一个巨大的计算机集群系统。所以，Hadoop 非常适用于大数据存储和处理，从大数据技术上来讲，当前大数据等于 Hadoop，Hadoop 等于大数据，学大数据技术必须学 Hadoop。

2. Hadoop 系统特点

- 易扩容（Scalable）：很容易增加存储节点和计算节点。
- 成本低（Economical）：可以通过普通计算机组成的服务器群来处理数据。这些服务器群总计可达数千个节点。
- 高效率（Efficient）：通过分发数据，Hadoop 可以在数据所在的节点上并行地（Parallel）处理它们，这使得处理非常的快速。移动计算的代价比之移动数据的代价低。一个应用请求的计算，离它操作的数据越近就越高效，这在数据达到海量级别的时候更是如此。将计算移动到数据附近，比之将数据移动到应用所在显然更好。

- 可靠性（Reliable）：Hadoop 能自动地维护数据的多份复制，并且在任务失败后能自动地重新部署（Redeploy）计算任务。
- 移植性：Hadoop 是使用 Java 实现的，在异构的软硬件平台间的可移植性。

3. Hadoop HDFS 适应性

适用条件

HDFS 是为以流式数据访问模式存储超大文件而设计的文件系统。所以适用如下应用：

- 超大文件。指的是几百 MB，几百 GB，几百 TB，甚至几百 PB。
- 流式数据访问，HDFS 建立的思想是：一次写入、多次读取的模式，是最高效的。
- 价格低廉的商用硬件。Hadoop 不需要运行在昂贵并且高可靠的硬件上。

不适用条件

HDFS 不适用以下的环境：

- 大量的小文件。
- 多用户写入，频繁修改。
- 低延迟数据访问。HDFS 是为了达到高数据吞吐量而优化的，这是以延迟为代价的，对于低延迟访问，可以用 HBase。

4. Hadoop 生态系统

Hadoop 是一个基础框架，为了满足大数据统计分析与挖掘的业务需求，则还需要和其他一些基于 Hadoop 的软件工具一起实现对业务的支撑。这些软件工具和 Hadoop 一起共同组成一个能支撑业务的系统环境，我们称之为 Hadoop 的生态系统。这些软件工具包括如下图：

- ***HDFS*** 可以支持千万级的大型分布式文件系统。
- ***HBase*** 是一个构建在 Hadoop 之上的，面向列的分布式数据库。
- ***MapReduce*** 是 Google 提出的一种分布式算法，用于超大型数据集的并行运算。
- ***Mahout*** 是一个可扩展的机器学习系统，旨在帮助开发人员更加方便快捷地创建智能应用程序。Mahout 包含许多机器学习算法实现，包括聚类、分类、推荐过滤、频繁子项挖掘等。通过使用 Hadoop，Mahout 可以有效地扩展到云中。
- ***R 语言***是一个统计分析、绘图的软件工具。使用一个开源项目 RHadoop 可以将 R 语言与 Hadoop 结合在一起，很好发挥了 R 语言特长，可以在大数据领域发挥更大作用。
- ***Hive*** 是一个数据仓库工具，适用于 ETL 方面的工作。
- ***Pig*** 是在 MapReduce 上构建的查询语言(SQL-like)，适用于大量并行计算。
- ***Sqoop*** 主要用于在 Hadoop 与传统关系型数据库间进行数据的传递，可以将一个关系型数据库（如 MySQL，Oracle，Postgres 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导进到关系型数据库中。
- ***Flume*** 是一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统。支持在日志系统中定制各类数据发送方，用于收集数据；同时还提供对数据进行简单处理，并写到各种数据接收方的能力。

- **Zookeeper** 分布式锁设施，提供分布式协作服务。功能包括：配置维护、名字服务、分布式同步、组服务等，用于分布式系统的可靠协调系统，保证数据的一致性。

5. Hadoop 必备知识

学习 Hadoop，需要预先掌握以下技术：

（1） 学习掌握 Linux 操作系统。因为 Hadoop 是安装 Linux 系统上的，不能安装在 Windows，有人说能安装在 Windows 上，那是在 Windows 创建了 Linux 虚拟机，还是安装在 Linux 上的。Linux 需要掌握的内容主要包括：（1）文件的操作，如文件创建、修改、删除、拷贝等；（2）如何在 Linux 上安装软件；（3）如何使用 SSH；（4）如何执行程序；（5）IP、主机名等常用配置。掌握 Linux 的程度只需一个入门级的程度就可以，可学习本网站内的《Linux 入门教程》。

（2） 学习掌握 Java 程序语言。在使用 MapReduce 分布式计算时需要使用 Java 语言来调用。需要掌握继承、方法的重写、输入输出（IO）流、内部类、数组、JAR 打包方法等。