学习报告

一、Hadoop

1、Hadoop 简介

Hadoop 可运行与一般的商用机器上，具有高容错，高可靠性，高扩展等特点，特别适合写一次，读多次的场景

2、适用场景

A、大规模数据

B、流式数据（写一次，读多次）

C、商用硬件（一般硬件）

3、Hadoop 架构

A、HDFS：分布式文件存储

B、YARN：分布式资源管理

C、MapReduce：分布式计算

D、Others：利用YARN的资源管理功能实现其他的数据处理方式

内部各个节点基本都是采用 Master-Worker 架构

二、HDFS

1.hdfs的基本概念：

HDFS：Hadoop Distributed File System ，Hadoop分布式文件系统。主要用来解决海量数据的存储问题。

2.hdfs的写文件过程

将文件存到HDFS集群中

3.hdfs读文件过程

将集群中的数据读出

4.hdfs的优缺点：

优点：

①高容错性

②适合批处理

③适合大数据处理

④流式文件访问

⑤可构建在廉价机器上

缺点：

①低延时数据访问

②小文件储存

③并发写入，文件随机修改

三、MapReduce

1、简介

一种分布式的计算方式，指定一个 Map（映射） 函数，用来把一组键值对映射成一组新的键值对，指定并发的 Reduce（归约） 函数，用来保证所有映射的键值对中的每一个共享相同的键组。

Map 输出格式和 Reduce 输入格式一定是相同的。

2、基本流程

读取文件数据

进行 Map 处理

进行 Reduce 处理

把处理结果写到文件中

四、D3.js

D3.js是一个JavaScript库，它可以通过数据来操作文档。D3可以通过使用HTML、SVG和CSS把数据鲜活形象地展现出来。同时，它提供了强大的可视化组件，可以让使用者以数据驱动的方式去操作DOM。