Hadoop简介

模块一至四:19760111-安航正

模块五至九:19760116-章舒婕

语音:19760113-李晚月

## 一 大数据的特点

与传统数据比较：

传统：使用MB,GB 等单位，是集中式存储 结构化 稳定的数据模型 紧密的内部关系

大数据：GB,TB,PB（海量）分布式存储 结构化，半结构化，无结构化 拥有平面模型 松散的数据关系

## 二 大数据存储的挑战

大数据是海量的（数据存储中的容量增长是没有止境的）

**要求**：不断扩张存储空间，简化用户对数据的管理与应用

应用要求：海量存储技术已经不是单台的存储设备，因此，要求统一平台的数据管理产品能够整合不同平台的存储设备在一个单一的控制界面上，结合虚拟化软件对存储资源进行管理

**问题存在**：

1. 降低成本
2. 易于扩展（爆炸性增长且难以预计）
3. 易使用，统一访问控制（日益复杂环境使存储数据无法管理）

## 三 分布式存储

1. 分布式文件系统

相对于本地文件系统而言，通常指c/s架构或网络文件系统，用户数据没有直接连接到本地主机上，而是存储在远程存储服务器上

1. 集群存储

* 多台存储设备中的存储空间集合成一个能给应用服务器提供统一访问接口和管理界面的存储池，应用可以通过访问接口透明的访问和利用所有存储设备上的磁盘，可以充分发挥存储设备的性能和磁盘利用率
* 采用开放式架构，包括存储节点，前端网络，后端网络，每个元素都可以非常容易地进行扩展而不用改变集群存储的架构

（侧重点不同分为）

高性能集群HPC

高可用集群HAC

负载均衡集群LBC

1. P2p存储

无中心结构，节点之间对称，通过相互合作完成用户任务

用户通过该平台自主寻找其他节点进行数据备份和存储空间交换，为用户构建了大规模存储交换的系统平台

P2p存储用于构建更大规模的分布式存储系统，可以跨多个大型数据中心或高性能计算中心使用

## 四 典型的分布式文件系统

1. **NFS**

它被移植到了几乎所有主流的操作系统，为运行于unix和其他系统上的客户程序提供了访问远程文件的透明性，系统中每台计算机需安装NFS客户模块和服务器模块，客户和服务器是对等的，每台服务器可以同时扮演客户和服务器两种角色

NFS的设计独立于操作系统，但被几乎所有操作系统支持

1. **AFS**

CUM设计，反映了通过减少客户-服务器通信量来支持大范围内的共享信息，被用来满足更多用户需要，其可伸缩性的主要策略是在客户节点上缓存整个文件

主要用来管理分部在不同网络节点上的文件，AFS采用安全认证和灵活的访问控制提供一种分布式的文件和授权服务，该服务可以扩展到多个客户端

（以上两个是早期的，其目的是提供标准接口和远程文件访问，更多的关注访问的性能和数据的可靠性，都是c/s文件系统）

1. **GPFS**

应用范围较广，可以支持大容量的文件系统和大文件，通过采用分片存储，较大的文件系统块，数据预读等方法获得了较高的数据吞吐量

通过文件分布在多个节点和硬盘上，GPFS可以超越单一节点和单一文件系统的性能极限

1. **GFS**

谷歌研制，由一个主服务器节点和大量块服务器节点构成，并由大量客户访问

把文件分成64MB，减少了元数据的大小，使主服务器节点能够非常方便的将元数据放置在内存中以提升访问效率

数据块分布在集群的机器上，使用linux的文件系统存放，同时每块文件至少有3份以上的冗余

1. **HDFS**

高容错性，设计用来部署在低廉的硬件上，实现了异构软硬件平台间的可移植性

硬件故障是常态而不是异常，自动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计划任务，实现了故障的检测和自动快速回复

## 五 大数据处理的挑战

1.数据处理的发展：

集中式数据处理：把所有数据，任务都集中到一个计算机上进行处理；

分布式数据处理:是指由多个相互连接的计算机组成的一个整体，这些计算机在系统软件协调下，合作执行一个共同的或不同的任务，最少依赖于集中的控制过程、数据和硬件。

2.新数据存储模式对数据处理框架的要求。  
数据处理与数据存储的强关联：  
 传统的并行计算方法主要从体系结构和编程语言的层面定义了一些较为底层的并行计算抽象和模型，但由于大数据处理问题具有很多高层的数据特征和计算特征，因此大数据处理需要更多地结合这些高层特征考虑更为高层的计算模式。

3.计算类型

面向大数据处理的数据查询、统计、分析、挖掘等需求，促生了大数据计算的不同计算模式，整体上我们把大数据计算分为三种：

1. 离线批处理计算：批量（batch）获取数据、批量传输数据、周期性批量计算数据、数据展示；
2. 实时交互计算：立刻得到反馈信息的交互；
3. 流计算：数据实时产生、数据实时传输、数据实时计算、实时展示。

## 六 大数据存储和处理的框架——Hadoop

ApacheTM Hadoop ® project：一个开源软件项目，致力于实现可靠的，可扩展的，分布式的计算。Hadoop:是一个开源的，基于Java的开源编程框架，该框架支持在分布式环境中实现超大规模数据的存储和处理。

Apache Hadoop software library:是一个框架(framework)，利用一种简单的编程模型，支持在计算机集群中实现大数据的分布式处理。

## 七 Hadoop的发展

2002：起源于Apache Nutch项目（一个开源的网络搜索引擎）

2003——2004：谷歌发表介绍了Google分布式文件系统（GFS）和分布式计算框架（MapReduce）的论文

2004：Nutch项目模仿GFS开发了自己的分布式文件系统NDFS（Nutch Distributed File System），以及分布式计算框架（MapReduce）

2006：Nutch中的NDFS和MapReduce开始独立出来，成为Lucene项目的一个子项目，称为Hadoop

Hadoop1.0 VS Hadoop2.0

Hadoop1.0（Single Use System：Batch APPS）:MapReduce，HDFS；

Hadoop2.0(Multi Use Data Platform：Batch，Interactive，Online，Streaming，…)：MapReduce，Tez，Others，YARN，HDFS2

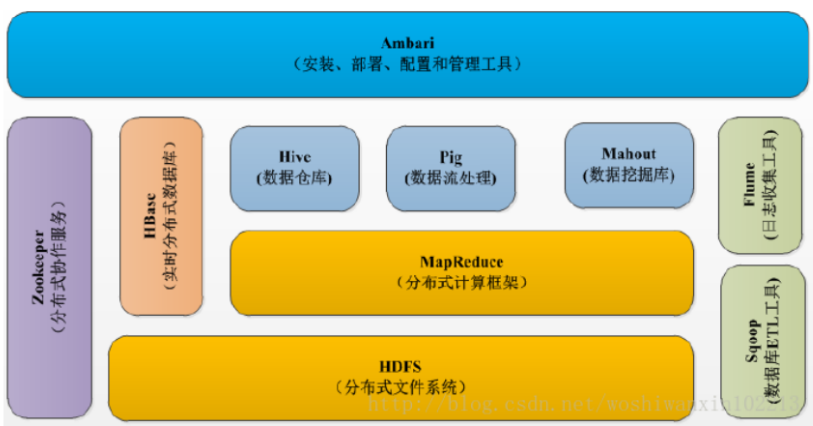
## 八 Hadoop与大数据

大数据：数据采集，数据存储，数据处理，数据应用展示

Hadoop核心

Hadoop：是一种利用集群的方式进行高效数据存储、处理的一种分布式框架。

## 九 Hadoop的生态



Hadoop生态的特点:  
1.提供了完整的大数据管理框架(数据存储和数据处理);  
2.在这一框架上有丰富的扩展支持;

3. HDFS:高效、高容错;  
4.MapReduce计算框架:高效率;