1.5大数据的关键技术

现在介绍大数据技术

在数据采集 数据存储与管理

数据处理与分析 数据隐私与安全

这五个层面都有相应的大数据技术存在

但是在最近十几年发展起来

最核心的大数据技术

是在数据存储与管理 数据处理与分析这两层

这两层代表了很多非常核心的大数据技术

这两层归结起来可以称为两大核心技术

也就是说大数据技术有两大核心

一个是分布式存储 一个是分布式处理

大数据技术解决的两大核心问题

就是怎么能够把这么多海量数据存储起来

一台机器存储不下

只能借助于整个集群网络去存储

叫做分布式存储

要对数据进行处理与分析

单台计算机无法满足实时需求

怎么办呢 也和刚才一样

构建一个集群

由这个集群来完成处理

整个课程要介绍两大核心

一个是分布式存储 一个是分布式处理

分布式存储主要以谷歌的技术为代表

当然分布式处理也是以谷歌技术为代表

这里面最核心的就是教材当中讲的

一个是分布式数据库

还有一个是分布式文件系统

以及分布式处理技术

分布式处理是谷歌提出来的MapReduce

讲了大数据技术以后再看一下

大数据计算模式的问题

为什么讲这个问题呢

因为现在有非常多的大数据相关技术产品存在

但是没有任何一种产品可以满足所有需求

所以我们要非常清楚地知道

大数据产品到底能够解决什么问题

它所服务的计算模式是不一样的

有些是用于批处理

有些用于实时计算

有些是用于交互式查询

每个产品运用的领域是不同的

所以我们要对计算模式进行区分

主要的计算模式分为四种

第一种是批处理

我们所熟悉的MapReduce

就是属于批处理模式

它是把一堆数据一起拿过来进行批量处理

它不适合于做实时计算

批处理显然不能满足时效性的要求

只要是批处理的东西

都不可能实现秒级响应

大数据计算模式

第一个非常典型的就是批处理

这里面代表性的产品就是MapReduce

还包括2015年异军突起的产品叫Spark

后面也会专门给大家讲Spark

它也是批处理

但它的实时性要比MapReduce更好

而且它解决了MapReduce当中的一些缺点

MapReduce无法高效地执行迭代计算

但是Spark可以高效地做迭代计算

许多应用当中需要做迭代计算

比如说数据挖掘

这个时候就不需要MapReduce 而是Spark

这就是批处理

第二种计算模式是流计算

流计算是专门针对流数据的实时计算

比如说日志流 用户点击流

这些都属于流式数据

它源源不断地到达系统

需要实时给出分析 马上给出响应

而不能一批一批进行处理

只要进行批处理 它马上失去实时商业价值

所以这时候不能进行批处理

像MapReduce那样去做批处理

只能用流计算框架去做

它是实现秒级的针对实时数据流的响应

第三种计算模式是图计算

很多产品像MapReduce可以处理图结构数据

但是效率不高

所以有些公司专门研发了

可以高效地处理图结构数据的产品

现实当中很多应用可以被转化为图结构

比如说社交网络数据

地理信息系统数据

这些数据都可以用图计算的软件去处理

最后一类是查询分析计算

就是刚刚讲的交互式查询

企业分析人员在分析系统中输入一个指令

要在一两秒钟得到查询结果

这个时候就需要查询分析软件

在大数据时代 这种大数据查询分析软件

就是用来满足交互式查询分析的需求

它有非常高的实时性

典型代表产品有Google公司的Dremel

还有在这里列出的Hive Cassandra

还有Impala 都是属于这类产品

所以这张表大家一定要理解

不同的产品是服务于不同的计算问题的

所以以后到了企业也是一样的

对不同的业务要有针对性地

去选择能够解决企业业务问题的产品