Hadoop这一节的主要内容包括， Hadoop发展历程，Hadoop的特性以及Hadoop的应用现状

首先我们来看一下Hadoop的来历

• Hadoop是APACHE软件基金下面的一个顶级项目，是一个开源分布式的计算平台，这种分布式计算平台之所以具有这么旺盛的生命力，一方面是因为免费，一方面是因为极大的降低了使用的复杂性，Hadoop对于我们普通用户，屏蔽了所有大数据底层实现的细节，只要按照它提供的更高层的接口，做一些傻瓜式的编程操作，后台的所有的工作都是由系统自己实现。完全不像以前学习的高性能计算编程方式那么复杂，正是因为有了Hadoop这个平台，我们做分布式编程更加简便，在Hadoop平台上开发应用可以使用多种语言C ,C++或者Python。Hadoop是用Java语言编写的，因此具有很好的跨平台性。

• Hadoop这个名字不是一个缩写，而是一个虚构的名字。该项目的创建者，Doug Cutting解释Hadoop的得名：“这个名字是我孩子给一个棕黄色的大象命名的。我的命名标准就是简短，容易发音和拼写，没有太多的意义，并且不会被用于别处。小孩子恰恰是这方面的高手，Hadoop最初是由Apache Lucene项目的创始人Doug Cutting开发的文本搜索库。 在2004年，Nutch项目也模仿GFS开发了自己的分布式文件系统NDFS（Nutch Distributed File System），也就是HDFS的前身

2004年，谷歌公司又发表了另一篇具有深远影响的论文，阐述了MapReduce分布式编程思想，Hadoop由 Apache Software Foundation 公司于 2005 年秋天作为Lucene的子项目Nutch的一部分正式引入。它受到由 Google Lab 开发的 Map/Reduce 和 Google File System(GFS) 的启发 。 2005年，Nutch开源实现了谷歌的MapReduce。所以可以这么讲，大数据的主要技术其实都是由谷歌公司先研发出来，而Hadoop是将这些技术开源实现了。

• 到了2006年2月，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，同时，Doug Cutting加盟雅虎

• 2008年1月，Hadoop正式成为Apache顶级项目，Hadoop也逐渐开始被雅虎之外的其他公司使用

• 2008年4月，Hadoop打破世界纪录，成为最快排序1TB数据的系统，它采用一个由910个节点构成的集群进行运算，排序时间只用了209秒

• 在2009年5月，Hadoop更是把1TB数据排序时间缩短到62秒。Hadoop从此名声大震，迅速发展成为大数据时代最具影响力的开源分布式开发平台，并成为事实上的大数据处理标准

• 我们通常说到的hadoop包括两部分，一是Hadoop核心技术（或者说狭义上的hadoop），对应为apache开源社区的一个项目，主要包括三部分内容：hdfs，mapreduce，yarn。其中hdfs用来存储海量数据，mapreduce用来对海量数据进行计算，yarn是一个通用的资源调度框架（是在hadoop2.0中产生的）。

• 另一部分指广义的，广义上指一个生态圈，泛指大数据技术相关的开源组件或产品，如hbase、hive、spark、pig、zookeeper、kafka、flume、phoenix、sqoop等。

• 生态圈中的这些组件或产品相互之间会有依赖，但又各自独立。比如habse和kafka会依赖zookeeper，hive会依赖mapreduce。

• 下面图给出了Hadoop技术生态圈的一个大致组件分布图

• 第一个组件是Hdfs

• Hdfs是一种分布式文件系统，是Hadoop体系中数据存储管理的基础。它是一个高度容错的系统，能检测和应对硬件故障，用于在低成本的通用硬件上运行。Hdfs简化了文件的一致性模型，通过流式数据访问，提供高吞吐量应用程序数据访问功能，适合带有大型数据集的应用程序。

• 二是Mapreduce

• MapReduce分为第一代（称为 MapReduce 1.0或者MRv1，对应hadoop第1代）和第二代（称为MapReduce 2.0或者MRv2，对应hadoop第2代）。第一代MapReduce计算框架，它由两部分组成：编程模型（programming model）和运行时环境（runtime environment）。它的基本编程模型是将问题抽象成Map和Reduce两个阶段，其中Map阶段将输入数据解析成key/value，迭代调用map()函数处理后，再以key/value的形式输出到本地目录，而Reduce阶段则将key相同的value进行规约处理，并将最终结果写到HDFS上。它的运行时环境由两类服务组成：JobTracker和TaskTracker，其中，JobTracker负责资源管理和所有作业的控制，而TaskTracker负责接收来自JobTracker的命令并执行它。

•

当然这里只是介绍了一部分，其他组件还在后面会详细介绍

•

• Hadoop是Apache软件基金会旗下的一个开源分布式计算平台，为用户提供了系统底层细节透明的分布式基础架构

• Hadoop是基于Java语言开发的，具有很好的跨平台特性，并且可以部署在廉价的计算机集群中

• Hadoop的核心是分布式文件系统HDFS（Hadoop Distributed File System）和MapReduce

• Hadoop被公认为行业大数据标准开源软件，在分布式环境下提供了海量数据的处理能力

它具有以下几个方面的特性：

• 高可靠性:整个Hadoop平台采用冗余副本机制，可以实现非常好的可靠性。一旦发生故障，冗余的机器就可以提供服务。

• 第二：高效性：因为它是利用集群做运算，可以把成百上千台服务器集中起来，做一个分布式并行处理，所以它可以非常高效的处理海量分布式数据集。

• 高可扩展性：你可以10个节点也可以20 个节点，可以不断往里面增加机器，可以加到几千个节点，一个集群由几个到几千个节点，可扩展性非常好

• 高容错性：它采用多副本机制，即使一部分副本发生问题，其他的副本也可以保证能正常使用。

• 成本低：Hadoop集群不像以前使用的HPC（高性能计算机，很多企业都是用比较贵的小型机，Hadoop不需要这种昂贵的机器，可以节省很多成本。

• 运行在Linux平台上

• 支持多种编程语言，它可以采用多种编程语言支持开发。

• Hadoop凭借其突出的优势，已经在各个领域得到了广泛的应用，而互联网领域是其应用的主阵地

• 2007年，雅虎在Sunnyvale总部建立了M45——一个包含了4000个处理器和1.5PB容量的Hadoop集群系统。Hadoop在很多大型公司都有相关应用，

• 尤其是Facebook， Facebook作为全球知名的社交网站，Hadoop是非常理想的选择，Facebook主要将Hadoop平台用于日志处理、推荐系统和数据仓库等方面

• 国内采用Hadoop的公司主要有百度、淘宝、网易、华为、中国移动等，其中，淘宝的Hadoop集群比较大。中国移动也是专门用Hadoop做大数据分析。

• 那么Hadoop在企业当中到底是怎么用的呢？我们一起来看一下。在企业当中的应用架构可以用这张图来表示，在企业中需要把大量数据源抓过来进行分析，它需要进行的分析包括几类：一个是数据分析、一个是数据实时查询、还有一个是数据挖掘，在企业应用当中，最典型的就是这三种应用。从底层数据源获取数据以后，为了支撑上层的这三种应用，Hadoop的相关技术是如何来做到的呢？我们来看一下中间这一层：大数据层。大数据层采用的就是相关的大数据技术，这里面很多都是Hadoop平台软件框架中的技术，不同的Hadoop组件可以帮助实现不同的企业分析，最底层可以使用Hadoop平台的HDFS分布式文件存储来满足企业中大量数据存储的需求，存储问题解决了，接下来就是要进行数据分析。第一类数据分析是什么呢：离线分析，我们把很多数据拿出来后进行批量处理，Hadoop中的mapreduce最擅长的就是做批处理，它的长处就是可以对批量的数据进行离线分析，进行批处理，所以这个MR就是指MapReduce，就是它的简称。除了MapReduce之外，我们还可以用Hadoop数据平台上的Hive和Pig来帮助我们进行离线数据分析。对于实时数据查询，我们可以用HBase数据库，HBASE是支持几十亿行数据的非常好的分布式数据库。如果是数据挖掘应用，可以使用Hadoop平台上的Mahout，它把各种数据挖掘、机器学习、商务智能的算法都用MapReduce实现。它也是开源的，它里面包含非常多算法的MapReduce实现。如果没有这些套件的话，开发人员就需要自行开发一些算法 ，现在不需要了，你直接就可以拿过来用了，当然在企业中的应用不止这么多，我们这里是介绍了主要的几个。我们的课程是入门级课程，更多的内容还需要大家继续深入去学习。