1. 知道如何利用Hadoop以及其他相关技术设计分布式系统
2. 知道如何利用Pig和Spark创造能够在Hadoop集群上运算数据的脚本
3. 知道如何利用HBase Cassandra和MongoDB来分析非关系型数据库
4. 知道什么才是在不同状况下合适的数据存储技术
5. 知道如何利用Kafka Sqoop和Flume等技术将数据发布到Hadoop集群上
6. 知道如何利用HDFS和MapReduce来存储分析大规模的数据
7. 知道如何利用Hive和MySQL分析关系型数据
8. 知道如何利用Drill Phoenix和Presto查询数据
9. 理解YARN Tex Mesos Zookeeper Zepplin Hue和Oozie是如何管理Hadoop集群的
10. 利用Spark Streaming Flink和Storm消耗流数据

Hadoop是一个能够对大量数据进行分布式处理的软件框架。 Hadoop 以一种可靠、高效、可伸缩的方式进行数据处理。

Hadoop 是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。

Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度。

Hadoop 还是可伸缩的，能够处理 PB 级数据。

此外，Hadoop 依赖于社区服务，因此它的成本比较低，任何人都可以使用。

Hadoop是一个能够让用户轻松架构和使用的分布式计算平台。用户可以轻松地在Hadoop上开发和运行处理海量数据的应用程序。它主要有以下几个优点：

高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖。

高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

高效性。Hadoop能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

高容错性。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

低成本。与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低。

Hadoop带有用Java语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。Hadoop 上的应用程序也可以使用其他语言编写，比如 C++。

hadoop大数据处理的意义

Hadoop得以在大数据处理应用中广泛应用得益于其自身在数据提取、变形和加载(ETL)方面上的天然优势。Hadoop的分布式架构，将大数据处理引擎尽可能的靠近存储，对例如像ETL这样的批处理操作相对合适，因为类似这样操作的批处理结果可以直接走向存储。Hadoop的MapReduce功能实现了将单个任务打碎，并将碎片任务(Map)发送到多个节点上，之后再以单个数据集的形式加载(Reduce)到数据仓库里。

hadoop的应用程序

Hadoop 的最常见用法之一是 Web 搜索。虽然它不是唯一的软件框架应用程序，但作为一个并行数据处理引擎，它的表现非常突出。Hadoop 最有趣的方面之一是 Map and Reduce 流程，它受到Google开发的启发。这个流程称为创建索引，它将 Web爬行器检索到的文本 Web 页面作为输入，并且将这些页面上的单词的频率报告作为结果。然后可以在整个 Web 搜索过程中使用这个结果从已定义的搜索参数中识别内容。

MapReduce

最简单的 MapReduce应用程序至少包含 3 个部分：一个 Map 函数、一个 Reduce 函数和一个 main 函数。main 函数将作业控制和文件输入/输出结合起来。在这点上，Hadoop 提供了大量的接口和抽象类，从而为 Hadoop应用程序开发人员提供许多工具，可用于调试和性能度量等。

MapReduce 本身就是用于并行处理大数据集的软件框架。MapReduce 的根源是函数性编程中的 map 和 reduce 函数。它由两个可能包含有许多实例(许多 Map 和 Reduce)的操作组成。Map 函数接受一组数据并将其转换为一个键/值对列表，输入域中的每个元素对应一个键/值对。Reduce 函数接受 Map 函数生成的列表，然后根据它们的键(为每个键生成一个键/值对)缩小键/值对列表。

这里提供一个示例，帮助您理解它。假设输入域是 one small step for man,one giant leap for mankind。在这个域上运行 Map 函数将得出以下的键/值对列表：

(one,1) (small,1) (step,1) (for,1) (man,1)

MapReduce 流程的概念流

MapReduce 流程的概念流

(one,1) (giant,1) (leap,1) (for,1) (mankind,1)

如果对这个键/值对列表应用 Reduce 函数，将得到以下一组键/值对：

(one,2) (small,1) (step,1) (for,2) (man,1)(giant,1) (leap,1) (mankind,1)

结果是对输入域中的单词进行计数，这无疑对处理索引十分有用。但是，假设有两个输入域，第一个是 one small step for man，第二个是 one giant leap for mankind。您可以在每个域上执行 Map 函数和 Reduce 函数，然后将这两个键/值对列表应用到另一个 Reduce 函数，这时得到与前面一样的结果。换句话说，可以在输入域并行使用相同的操作，得到的结果是一样的，但速度更快。这便是 MapReduce 的威力;它的并行功能可在任意数量的系统上使用。图 2 以区段和迭代的形式演示这种思想。

回到 Hadoop 上，

它是如何实现这个功能的?一个代表客户机在单个主系统上启动的 MapReduce应用程序称为 JobTracker。类似于 NameNode，它是 Hadoop 集群中惟一负责控制 MapReduce应用程序的系统。在应用程序提交之后，将提供包含在 HDFS 中的输入和输出目录。JobTracker 使用文件块信息(物理量和位置)确定如何创建其他 TaskTracker 从属任务。MapReduce应用程序被复制到每个出现输入文件块的节点。将为特定节点上的每个文件块创建一个惟一的从属任务。每个 TaskTracker 将状态和完成信息报告给 JobTracker。图 3 显示一个示例集群中的工作分布。

Hadoop 的这个特点非常重要，因为它并没有将存储移动到某个位置以供处理，而是将处理移动到存储。这通过根据集群中的节点数调节处理，因此支持高效的数据处理。

总之：Hadoop的优点有：

1、hadoop是可伸缩的，可以处理PB级的大数据

2、hadoop是高容错的，hadoop能够自动保存数据到多个副本，从而保证任务失败后将该任务进行重新分配

3、hadoop是高效的，能够将数据在不同节点之间进行分配和运行，保证各个节点上数据的动态平衡，因此处理速度非常快。

高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖。

高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

高效性。Hadoop能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

高容错性。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

低成本。与一体机、商用数据仓库以及QlikView、Yonghong Z-Suite等数据集市相比，hadoop是开源的，项目的软件成本因此会大大降低。