学习Hadoop的目的

1. 独立搭建Hadoop、hbase集群
2. 掌握Hadoop、Hbase、Hive等基础知识
3. 掌握shell、java api操作Hadoop、Hbase、Hive，能够上手工作。
4. hadoop
5. 适合大数据的分布式存储和计算平台。
6. 核心项目：HDFS（分布式文件系统）和MapReduce（并行计算框架）
7. HDFS的架构，主从结构，一个namenode和多个datanode。Namenode负责维护文件系统的结构，管理文件与block块之间的关系，block与datanode之间的关系。Datanode负责存储文件，文件被分成block存储在磁盘上，为了数据安全，会保留多个副本存在。
8. MapReduce的架构，主从结构，一个jobTracker和多个taskTracker。JobTracker负责接受客户提交的计算任务，把计算任务分给TaskTracker执行，监控TaskTracker的执行情况。TaskTracker负责执行JobTracker分配给的计算任务。
9. Hadoop的特点：扩容能力（处理pb级别数据）成本低（普通机器）高效率（并行处理）可靠性（多个数据副本）
10. 部署的三种方式：本地、伪分布、集群模式。伪分布模式的步骤，关闭防火墙，修改ip，修改hostname，设置ssh自动登录，安装jdk，安装hadoop。
11. Hadoop的目录包括哪些，分别解释一下。在hadoop的根目录下，有hadoop的核心jar包hadoop-core-1.1.2.jar；bin目录下存放hadoop的执行命令，如启动start-all.sh、关闭stop-all.sh、启动配置命令hadoop-config.sh、start-dfs.sh、stop-dfs.sh、start-mapred.sh、stop-mapred.sh、动态增加节点时候启动的命令hadoop-daemon.sh start datanode、hadoop-daemon.sh start tasktracker；conf下存放配置信息core-site.xml指定**fs.default.name**的值为hdfs的默认访问路径，即hdfs的根目录/，还指明了临时目录、hadoop-env.sh指定jdk路径等、hdfs-site.xml指明数据的副本数、mapred-site.xml指明mapred.job.tracker的值为JobTracker的IP地址和端口号、master主节点的主机名、slaves从节点的主机名；lib下存放hadoop所有的依赖jar包；docs下存放hadoop的api等文档信息；src下存放hadoop的源代码；logs下存放日志信息。另外hadoop启动的时候会生成tmp文件夹，里面存放hdfs中的文件。
12. 如何运行jar包中的hadoop程序？在eclipse中编写好程序，然后打包成jar文件，注意选择main方法，在hadoop分布式集群上运行，上传到集群的hdfs文件系统中，然后执行hadoop jar \*.jar就可以运行了。
13. Hdfs文件系统namenode包括文件有fsimage、edits、fstime。Datanode存储时候数据分成文件块存放。Secondarynamenode执行过程：从NameNode上下载元数据信息（fsimage,edits），然后把二者合并，生成新的fsimage，在本地保存，并将其推送到NameNode，同时重置NameNode的edits. Secondarynamenode一是镜像备份，二是日志与镜像的定期合并。
14. Hdfs通过javaapi操作的核心类是FileSystem,configuration通过new出来。ClientProtocol是客户端FileSystem与NameNode通信的接口。DatanodeProtocol是DataNode与NameNode通信的接口。NamenodeProtocol是SecondaryNameNode与NameNode通信的接口。DFSClient是直接调用NameNode接口的对象。用户代码是通过DistributedFileSystem调用DFSClient对象，才能与NameNode打交道。
15. Hdfs的高可靠如何实现？通过datanode节点的多个副本保证数据的可靠。
16. MapReduce是一种分布式计算模型，有google提出，主要用于搜索领域，解决海量数据的计算问题。
17. Mapreduce执行步骤：

1. map任务处理

1.1 读取输入文件内容，解析成key、value对。对输入文件的每一行，解析成key、value对。每一个键值对调用一次map函数。

1.2 写自己的逻辑，对输入的key、value处理，转换成新的key、value输出。

1.3 对输出的key、value进行分区。

1.4 对不同分区的数据，按照key进行排序、分组。相同key的value放到一个集合中。

1.5 (可选)分组后的数据进行归约。

2.reduce任务处理

2.1 对多个map的输出，按照不同的分区，通过网络copy到不同的reduce节点。

2.2 对多个map的输出进行合并、排序。写reduce函数自己的逻辑，对输入的key、value处理，转换成新的key、value输出。

2.3 把reduce的输出保存到文件中。

例子：实现WordCountApp

1. 源代码分析:JobTracker负责接收用户提交的作业，负责启动、跟踪任务执行。JobSubmissionProtocol是JobClient与JobTracker通信的接口。InterTrackerProtocol是TaskTracker与NameNode通信的接口。Tasktracker负责执行任务。JobClient是用户作业与JobTracker交互的主要接口。负责提交作业的，负责启动跟踪任务执行、访问任务状态和日志等。
2. Hadoop计数器：可以让开发人员以全局的视角来审查程序的运行情况以及各项指标，及时做出错误诊断并进行相应处理；combiner的作用就是在map端对输出先做一次合并，以减少传输到reducer的数据量，实现本地的规约处理；partitioner分区，默认为HashPartitioner；hadoop的压缩codec；MRUnit进行测试。Shuffler的过程是什么。
3. MapReduce常见算法：单词计数、数据去重、排序、top K、选择、投影、分组、多表连接、单表关联。
4. 面试题：现有1亿个整数均匀分布，如果要得到前1K个最大的数
5. Hadoop中使用zookeeper的事件处理确保整个集群只有一个NameNode，存储配置信息等。保证hadoop中的数据事务性一致。通过选举模式选举领导者。Zookeeper 作为 Hadoop 项目中的一个子项目，是 Hadoop 集群管理的一个必不可少的模块，它主要用来控制集群中的数据，如它管理 Hadoop 集群中的 NameNode，还有 Hbase 中 Master Election、Server 之间状态同步等。
6. Zoopkeeper 提供了一套很好的分布式集群管理的机制，就是它这种基于层次型的目录树的数据结构，并对树中的节点进行有效管理，从而可以设计出多种多样的分布式的数据管理模型
7. HBase – Hadoop Database，是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用HBase技术可在廉价PC Server上搭建起大规模结构化存储集群。HBase利用Hadoop HDFS作为其文件存储系统，利用Hadoop MapReduce来处理HBase中的海量数据，利用Zookeeper作为协调工具。行健，列族，cell，版本是其自然属性。
8. HBASE的结构，表是存放数据的，表有行和列组成数据模型。物理存储，Table在行的方向上分割为多个HRegion，每个region分散在不同的RegionServer中。可以实现数据的实时查询。
9. HBASE的常用命令hbase create/put/get/count/delete/drop/scan
10. 数据仓库与数据库的区别
11. Hive是建立在Hadoop上的数据仓库基础架构，它提供了一系列的工具，可以用来进行数据提取转换加载（ETL），这是一种可以存储、查询和分析存储在hadoop中的大规模数据的机制。Hive是SQL解析引擎，它将SQL语句转译成M/R Job然后在Hadoop执行。Hive的表其实就是HDFS的目录/文件，按表名把文件夹分开。如果是分区表，则分区值是子文件夹，可以直接在M/R Job里使用这些数据。
12. Hive的数据模型内部表、分区表、外部表。导入导出数据
13. 为什么选择Hive？基于Hadoop的大数据的计算/扩展能力，支持SQL like查询语言，统一的元数据管理，简单编程。
14. Sqoop的功能是实现关系型数据库与hadoop的hdfs之间的数据转换。