### 大数据简介

#### 什么是大数据

海量数据，指的是在一定时间范围内无法用常规软件去管理和处理的数据集合

通俗的说：大数据要解决的问题就是大量数据存储和计算的问题

大数据：技术革新

#### 数据存储和数据计算

数据存储：分布式存储，数据（文件）一般都是分布到集群上的不同机器去存储

数据计算：

热门商品排行榜

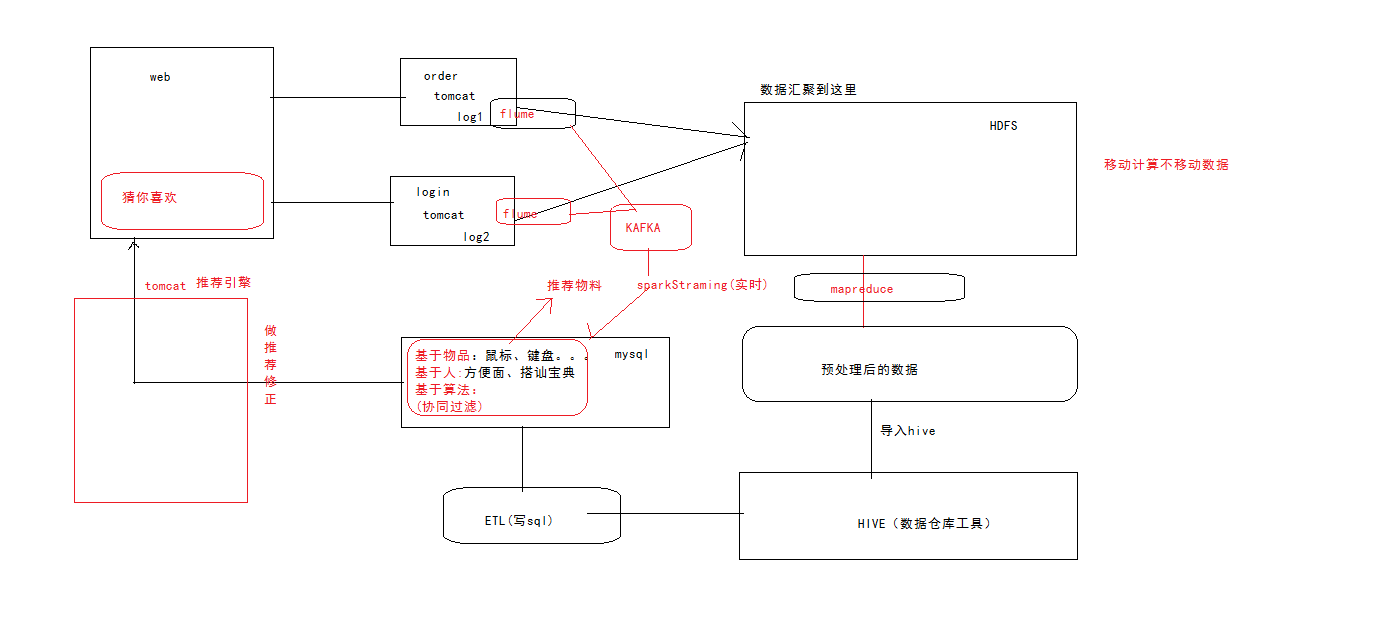
主播排行榜

礼物排行榜

每个人平均浏览的商品次数

浏览过A商品的人，还浏览过哪儿些产品

猜你喜欢（基于用户画像）



1. 大数据的思想

分而治之、移动计算不移动数据

1. 大数据来源
   1. 计算机产生的数据
   2. 移动运行商的数据
   3. 机器和传感器的数据
   4. 人为数据
   5. 开源数据
2. 数据处理流程
   1. 数据收集
   2. 数据导入和预处理
   3. 数据的统计和分析
   4. 数据的使用

### hadoop快速入门

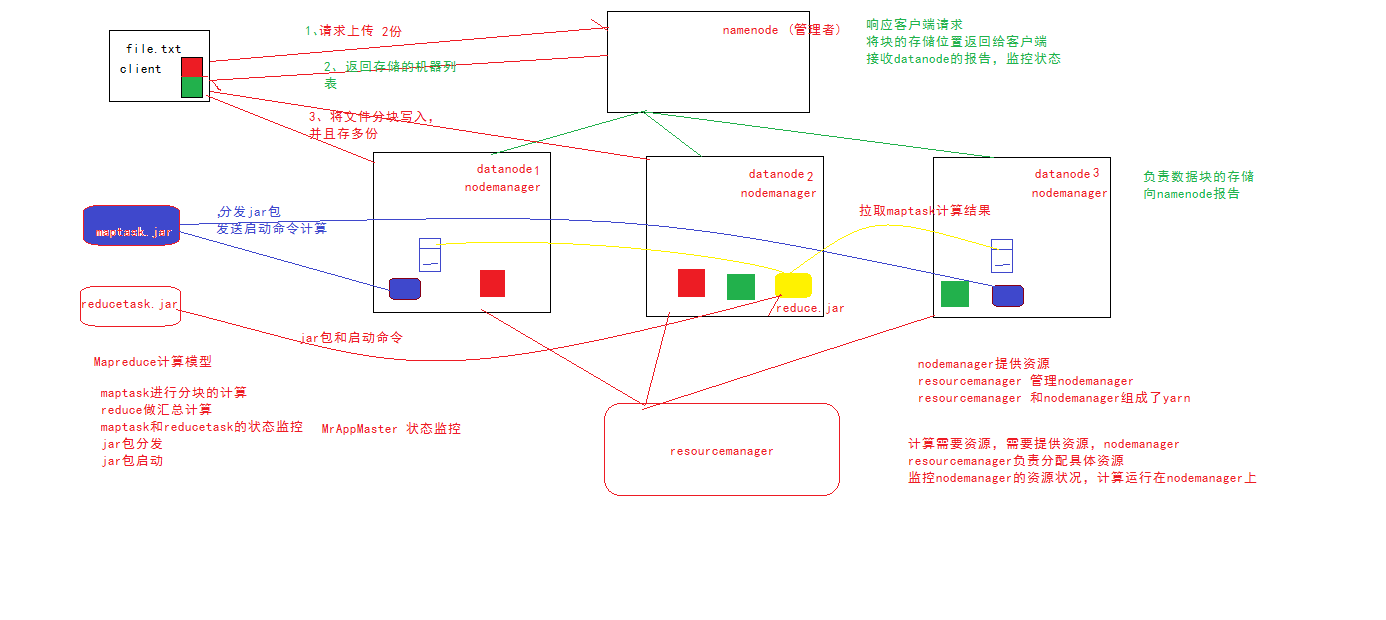
1. 什么是hadoop

是一个软件框架、允许使用简单的编程模型跨计算机集群的分布式处理大数据集合

被设计成从单一的服务扩展到数千台机器，每个机器提供本地的计算和存储

与依赖硬件高可用相比，hadoop框架设计的是故障的自检和失败转移，因为在一组计算机集群上提供服务，每一台计算都有可能发生硬件故障

1. 四大模块、三个核心
   1. Common：支撑hadoop其他模块
   2. HDFS：分布式文件系统，提供应用程序大数据的访问（高吞吐量）
   3. YARN：作业调度和集群资源管理
   4. MapReduce：一个基于yarn的并行的处理大数据集的系统
2. 自己完成一个分布式文件系统考虑问题：
   1. 每台机器负载情况
   2. 故障检测
   3. 故障转移
   4. 管理者（专职）
   5. 数据备份
   6. 数据还原（管理者）
3. 图解hadoop



1. Hadoop与云计算的关系
   1. 基础设施即服务
   2. 软件即服务
   3. 平台即服务
2. Hadoop生态圈及组成
   1. HDFS
   2. MAPREDUCE
   3. YARN
   4. HIVE
   5. HBASE
   6. ZOOKEEPER：分布式协调服务
   7. Mahout:机器学习算法库
   8. Oozie:工作流调度框架
   9. Flume:日志数据采集框架
   10. Spark:实时计算框架
   11. Storm：流式计算框架
   12. Sqoop：数据导入导出工具

### 分布式系统概述

* 1. 为什么要分布式系统

直白：承载量是分布式系统存在的原因

分布式系统的出现使用廉价的、普通的机器完成单个计算机无法完成的计算、存储任务

目的是利用更多的机器处理更多的数据，速度更快

* 1. 什么是分布式系统

一组计算机组成，通过网络通信，共同的协调完成任务

包括分布式存储和分布式计算

分而治之：即分片。对于计算，每个节点计算一部分，然后最终汇总，这就是MR的思想

（1）提升性能和并发，操作被分发到不同机器上，相互独立

（2）提升系统的可用性

* 1. 分布式系统的问题
     1. 不靠谱的网络
     2. 普遍的节点故障

### 离线数据分析流程

* 1. 数据采集
  2. 数据预处理
  3. 导入hive
  4. ETL
  5. 形成报表
  6. 导入mysql
  7. 数据应用

### 单机版hadoop的安装

1. 安装步骤
   1. 上传安装包，JDK、hadoop
   2. 首先安装JDK，环境变量配置 /etc/profile
   3. Source /etc/profile
   4. 安装hadoop,配置hadoop
   5. tar -zxvf /data/hadoop-2.7.1.tar.gz -C /apps/
2. 目录介绍
   1. Bin：普通用户操作hadoop的命令
   2. Etc:hadoop的配置文件的目录
   3. Sbin:启动或者停止hadoop服务的相关脚本
   4. Logs:日志
   5. Lib:hadoop的本地库
   6. Libexc:配置各个服务的shell配置文件的所在的目录
   7. Share:hadoop的jar包
3. 命令介绍
   1. Hadoop fs :fs涉及到一个通用的文件系统
   2. hadoop dfs ：翻译成hdfs dfs，也就是说这个命令仅对HDFS文件系统有效
   3. Hdfs dfs
   4. Yarn:运行yarn的相关服务
   5. Mapred：运行mr的相关服务
   6. Hdfs:调用hdfs文件系统的命令
4. 案例介绍
   1. hadoop jar /apps/hadoop-2.7.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar wordcount /word /out

Wordcount的案例

* 1. hadoop jar /apps/hadoop-2.7.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.7.1.jar pi 10 10

计算PI值

### HDFS集群的介绍安装

* 1. 设计思想：分而治之,将大文件、大批量文件，分布式存在大量服务器上，以便于采取分而治之的思想对海量数据进行运算
  2. 重点概念：文件切块、副本存放、元数据
  3. Hadoop的存储系统就是HDFS，一个分布式文件系统，对外部客户端而言，HDFS就像一个传统的分级文件系统，可以进行创建、删除、移动、重命名文件或文件夹的操作，与linux操作系统类似，但是HDFS的结构是基于一组特定的节点的构建，这些节点NameNode（HDFS元数据服务）、secondaryNameNode（辅助namenode管理元数据的节点）和DataNode（数据节点）
  4. 架构组成
  5. HDFS集群中的角色
     1. Namendoe：主节点，管理元数据
        1. 通常运行在集群上单独的机器上，负责管理文件系统名称空间（目录结构）

/data/a.txt /data/test/b.txt

* + - 1. 负责响应客户端的请求
      2. 负责元数据的管理
    1. Datanode：从节点，负责数据的存储
       1. 单独运行在一台计算机上的组件，hadoop集群包括至少一个namenode和多个datanode
       2. 负责与客户端进行通信进行读写
       3. 负责定期的向namenode汇报
    2. secondaryNameNode
       1. Namenode元数据管理的辅助节点，默认启动和namenode在同一台机器上

1. HDFS的安装
   1. 规划mini1\mini2\mini3\mini4
   2. 准备的内容：JDK,主机名，ip，域名映射，防火墙，时间同步
   3. 上传hadoop安装包进行配置
   4. 配置配置文件
      1. 修改hadoop-env.sh，修改JAVA\_HOME
      2. 修改core-site.xml
         1. fs.defaultFS:资源标识符

hdfs://mini1:9000

* + - 1. Io.file.buffer.size：缓冲区大小
      2. hadoop.tmp.dir:软件运行时临时文件存放位置
    1. 配置hdfs-site.xml
       1. dfs.replication
       2. 镜像文件检查目录
       3. 日志的目录
    2. 配置slaves（配置哪儿些机器是datanode）
    3. 分发hadoop的安装包
    4. 启动HDFS
       1. 初始化namendoe:hdfs namenode -format
       2. 启动namenode或者datanode

Hadoop-daemon.sh start/stop namenode/datanode

Start-dfs.sh 启动namenode和所有的datanode,启动secondarynamenode

1. 集群启动的错误
   1. 不能解析主机名

Could not resolve hostname mini11: Name or service not known

* 1. 配置文件出错



* 1. 集群重新格式化：使用start-dfs.sh时datanode无法启动，集群ID不一致
  2. 日志 $HADOOP\_HOME/logs/hadoop-{username}-{jobname}-{hostname}.log

1. 思考：自己实现一个HDFS，要考虑哪儿些问题
2. HDFS的概念和特性
   1. 首先，它是一个文件系统，用于存储文件，通过统一的命名空间-----目录树来定位文件
   2. 其次，它是分布式的，由很多服务器联合起来实现功能，集群中服务器都有各自的角色

重要特性

1. 物理上分块存储，块的大小可以通过参数配置（dfs.blocksize）,2.x版本默认大小128M，3.x版本256M
2. HDFS文件系统会给用户一个统一的抽象的目录树，客户端通过路径来访问文件

hdfs://namenode:port/data/a/b/c.txt

1. 目录结构及文件分块信息（元数据）的管理由namenode管理
2. 文件的各个block由datanode去存储，每个块会存多个副本（dfs.replication）
3. HDFS是设计成一次写入，多次读取的场景，且不支持文件的修改，但是支持文件的追加写入

### HDFS的基本操作

修改mini1中hdfs-site.xml文件中的

dfs.blocksize=64M

使用mini1上传文件到HDFS中，块的大小会是多大？

使用mini2机器上传文件到HDFS，块的大小是多大？

修改mini2 dfs.replacition 的值为2？

使用mini2上传文件查看副本数的数量？

使用mini1上传文件查看副本数量？

客户端决定