Hadoop 体系介绍

目录

Hadoop 引言

思考问题：

1、Hadoop 是什么？Hadoop 是怎么产生的？
3、Hadoop 应用在哪里？Hadoop 能解决什么问题？
4、Hadoop 怎么使用？Hadoop 是如何工作的？
5、Hadoop 的核心设计思想和底层实现原理是什么？

学习目标：

第一天接触具体的大数据框架，总目标是让学习者建立起大数据和分布式的宏观概念
1、理解大数据的概念，理解 hadoop 是什么，应用在那些地方，解决什么问题
2、理解 hive 是什么，应用在那些地方，解决什么问题
3、理解 ZooKeeper 是什么，应用在那些地方，解决什么问题
4、理解 HBase 是什么，应用在那些地方，解决什么问题

讲课思路：
第一步：介绍这个东西，知道它是什么，用来解决什么问题
第二步：安装使用（安装，shell 操作， java API 操作）
第三步：讲该软件提供的功能的底层实现原理
第四步：讲该软件使用的高级应用和优化措施

1、Hadoop 快速入门

1.1、数据

数据(data)是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的未经加工的的原始素材。
数据可以是连续的值，比如声音、图像，称为**模拟数据**。也可以是离散的，如符号、文字，称为**数字数据**。
在计算机系统中，数据以二进制信息单元 0,1 的形式表示。

1.2、大数据

1.2.1、概念

指的是传统数据处理应用软件不足以处理（存储和计算）它们的大而复杂的数据集

最基本的衡量：大小
数据量最小的基本单位是 bit，按顺序给出所有单位：bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB

1 Byte =8 bit
1 KB = 1,024 Bytes = 8192 bit
1 MB = 1,024 KB = 1,048,576 Bytes（普通用户数据级别）
1 GB = 1,024 MB = 1,048,576 KB
1 TB = 1,024 GB = 1,048,576 MB
1 PB = 1,024 TB = 1,048,576 GB（企业级数据级别）
1 EB = 1,024 PB = 1,048,576 TB
1 ZB = 1,024 EB = 1,048,576 PB（全球数据总量级别）
1 YB = 1,024 ZB = 1,048,576 EB
1 BB = 1,024 YB = 1,048,576 ZB
1 NB = 1,024 BB = 1,048,576 YB
1 DB = 1,024 NB = 1,048,576 BB

据国际数据公司(IDC)统计，全球数据总量预计 2020 年达到 44ZB，中国数据量将达到 8060EB，占全球数据总量的 18%

1.2.2、大数据特点

容量大，种类多，速度快，价值高

容量（Volume）：数据的大小决定所考虑的数据的价值和潜在的信息
新浪微博，3 亿用户，每天上亿条微博
朋友圈，8 亿用户，每天亿级别朋友圈
种类（Variety）：数据类型的多样性，包括文本，图片，视频，音频
结构化数据：可以用二维数据库表来抽象，抽取数据规律
半结构化数据：介于结构化和非结构化之间，主要指 XML，HTML 等，也可称非结构化
非结构化数据：不可用二维表抽象，比如图片，图像，音频，视频等
速度（Velocity）：指获得数据的速度以及处理数据的速度
数据的产生呈指数式爆炸式增长
处理数据要求的延时越来越低
价值（Value）：合理运用大数据，以低成本创造高价值
综合价值大，隐含价值大
单条数据记录无价值，无用数据多

总结：
1、数据量大，处理难度大，但是蕴含价值也大
2、数据种类多样，更加个性化，针对不同数据源进行多样化的方式处理，结果更精确
3、要求对数据进行及时处理，追求更极致更完善的用户体验
4、数据成为新的资源，掌握数据就掌握了巨大的财富

大数据崛起的根本原因：
1、数据生成的速度呈指数式爆炸增长
2、数据的存储成本指数下降
3、流动数据增加，云端数据增加
4、企业可用数据资源增大

1.2.3、大数据价值

在总数据量相同的情况下，与个别分析独立的小型数据集（Data set）相比，将各个小型数据集合并后进行分析可得出许多额外的信息和数据关系性，可用来**政治经济国庆调控、察觉商业趋势、判定研究质量、避免疾病扩散、打击犯罪**或**测定即时交通路况**等，这样的用途正是大型数据集盛行的原因

1.3、Hadoop 的产生背景

- Hadoop 最早起源于 Nutch。Nutch 的设计目标是构建一个大型的全网搜索引擎，包括网页抓取、索引、查询等功能，但随着抓取网页数量的增加，遇到了严重的可扩展性问题——如何解决数十亿网页的存储和索引问题

- 2003 年、2004 年谷歌发表的两篇论文为该问题提供了可行的解决方案
  1、分布式文件系统 GFS，可用于处理海量网页的存储
  2、分布式计算框架 MapReduce，可用于处理海量网页的索引计算问题
  3、分布式数据库 BigTabl，每一张表可以存储上 billions 行和 millions 列
- Nutch 的开发人员完成了相应的开源实现 HDFS 和 MapReduce，并从 Nutch 中剥离成为独立项目 Hadoop，到 2008 年 1 月，Hadoop 成为 Apache 顶级项目，迎来了它的快速发展期

1.4、什么是 Hadoop？

1、Hadoop 是 Apache 旗下的一套开源软件平台
2、Hadoop 提供的功能：利用服务器集群，根据用户的自定义业务逻辑，对海量数据进行分布式处理
3、Hadoop 的核心组件有

A．Common（基础功能组件）（工具包，RPC 框架）JNDI 和 RPC
B．HDFS（Hadoop Distributed File System 分布式文件系统）
C．YARN（Yet Another Resources Negotiator 运算资源调度系统）
D．MapReduce（Map 和 Reduce 分布式运算编程框架）

4、广义上来说，Hadoop 通常是指一个更广泛的概念--Hadoop 生态圈
5、官网介绍：http://hadoop.apache.org/

Apache™ Hadoop®的项目开发开源软件可靠，可扩展，分布式计算。

Apache Hadoop 软件库是一个允许跨集群用简单的模型对于大数据的分布式处理的框架。它的目的是扩大从单一服务器到成千上万的机器，提供每个本地计算和存储。而不是依靠硬件来实现高可用性，库本身的是检测和处理在应用程序层的故障，所以提供高可用性服务除了计算机集群，每一种都可能导致故障。

该项目包括这些模块：

- Hadoop Common：基础功能类库支持其他 Hadoop 模块
- Hadoop Distributed File System：一个分布式文件系统，它提供了高通量访问应用程序数据
- Hadoop YARN：一个集群作业调度和资源管理的框架
- Hadoop MapReduce：YARN-based 系统并行处理大型数据集。（版本升级以 YARN 做资源管理器的 Hadoop）

其他在 Apache Hadoop 相关项目包括：

- Ambari™：一个基于 web 的工具配置，管理和监视 Apache Hadoop 集群，包括支持 Hadoop 的 Hadoop HDFS，Hadoop MapReduce，Hive，HCatalog，HBase，ZooKeeper，Oozie， Pig and Sqoop。Ambari 还提供了一个仪表板查看集群健康，如热图和能够有效看到 MapReduce，Pig 和 Hive 应用的特性来诊断性能特征以用户友好的方式。
- Avro™：数据序列化系统。
- Cassandra™：一个没有单点故障可伸缩的多主机数据库。

- Chukwa™：一个管理大型分布式系统的数据采集系统。
- HBase™：一个可扩展的分布式数据库，支持大型表的结构化数据存储。
- Hive™：一个数据仓库基础设施，提供了数据总结和特别查询。
- Mahout™：一个可扩展的机器学习和数据挖掘库。
- Pig™：一个高级数据流语言和并行计算的执行框架。
- Spark™：Hadoop 数据的快速、通用的计算引擎。Spark 提供了一个简单的和丰富的编程模型，支持广泛的应用程序，包括 ETL、机器学习、流处理和图计算。
- Tez™：一个广义数据流编程框架，基于 Hadoop 的 YARN，它提供了一个功能强大且灵活的引擎来执行任意 DAG（有向无环图）的任务来处理批处理和交互用例的数据。Tez 正在被 Hive™， Pig™ 和其他框架 Hadoop 生态系统，以及其他商业软件(例如 ETL 工具)，以取代 Hadoop MapReduce™作为底层执行引擎。
- ZooKeeper™：一个高性能的分布式应用程序的协调服务。

1.5、hadoop 在大数据和云计算当中的位置和关系

1、云计算是分布式计算、并行计算、网格计算、多核计算、网络存储、虚拟化、负载均衡等传统计算机技术和互联网技术融合发展的产物。借助 IaaS(基础设施即服务)、PaaS(平台即服务)、SaaS（软件即服务）等业务模式，把强大的计算能力提供给终端用户

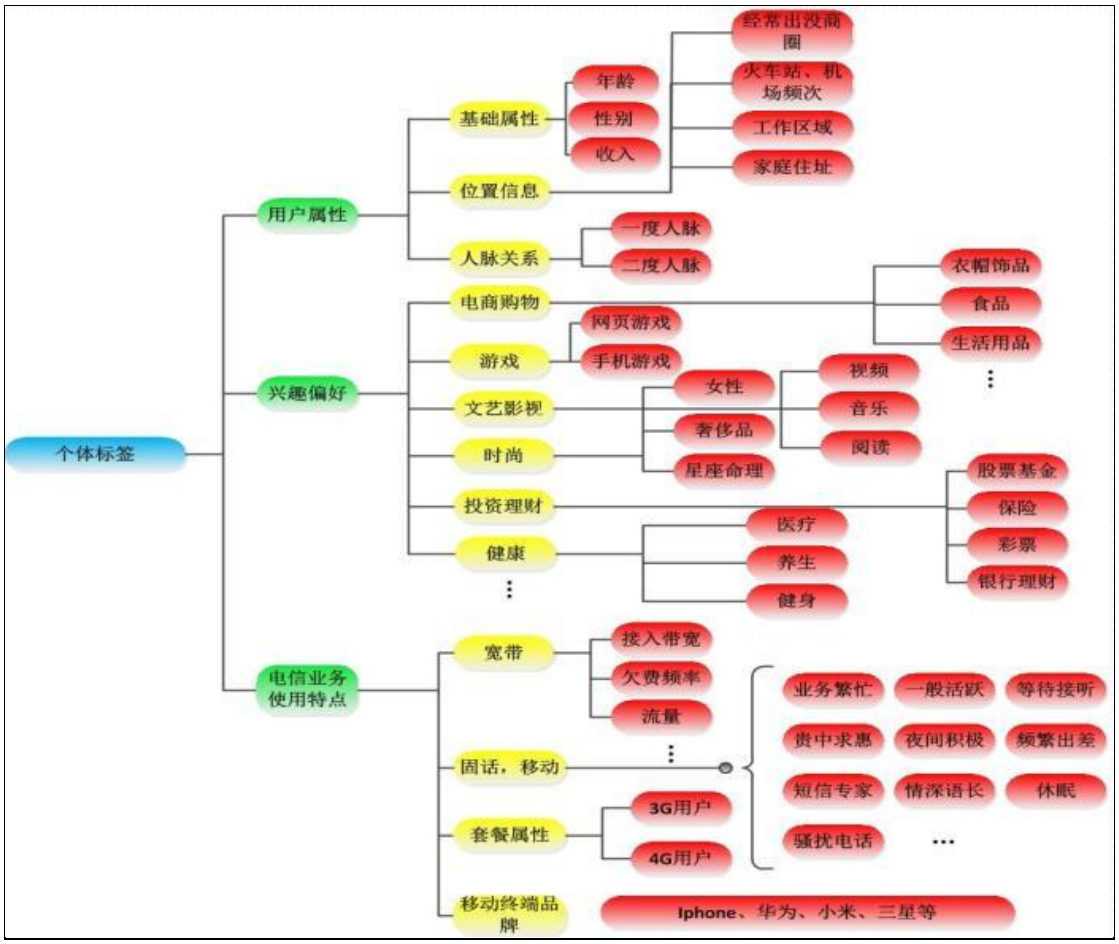
2、现阶段，云计算的两大底层支撑技术为“虚拟化”和“大数据技术”

3、而 Hadoop 则是云计算的 PaaS 层的解决方案之一，并不等同于 PaaS，更不等同于云计算本身

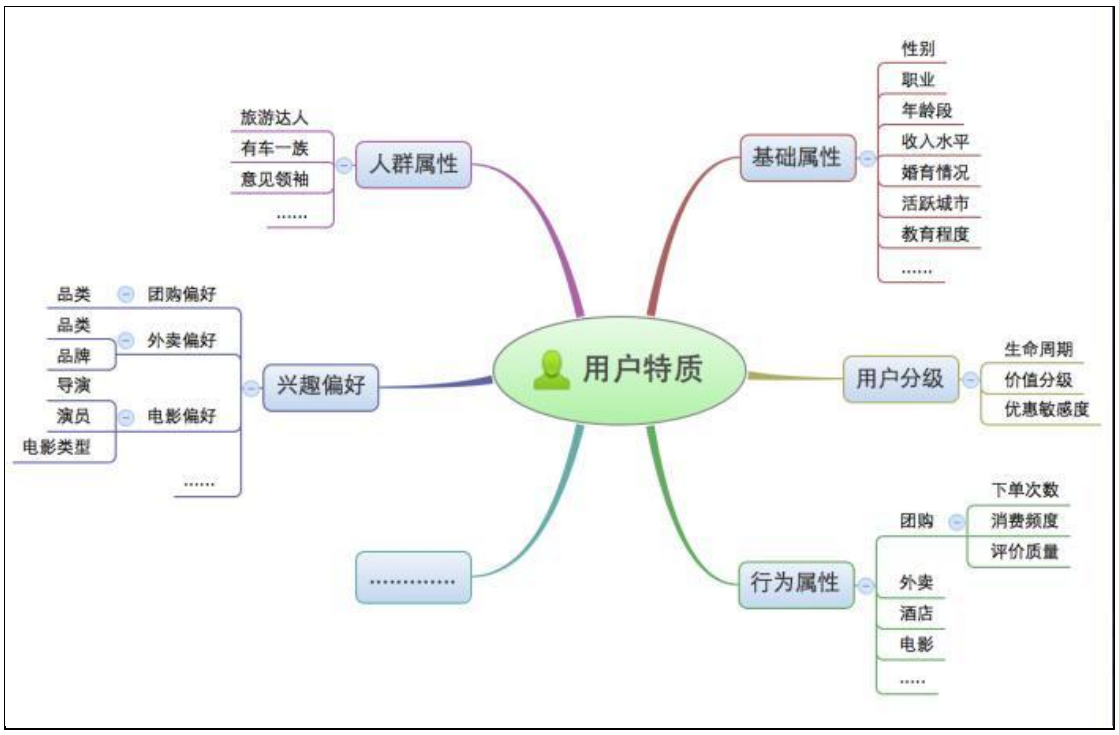
1.6、Hadoop 技术应用架构概览

1.6.1、Hadoop 应用于数据服务基础平台建设

1.6.2、Hadoop 用于用户画像

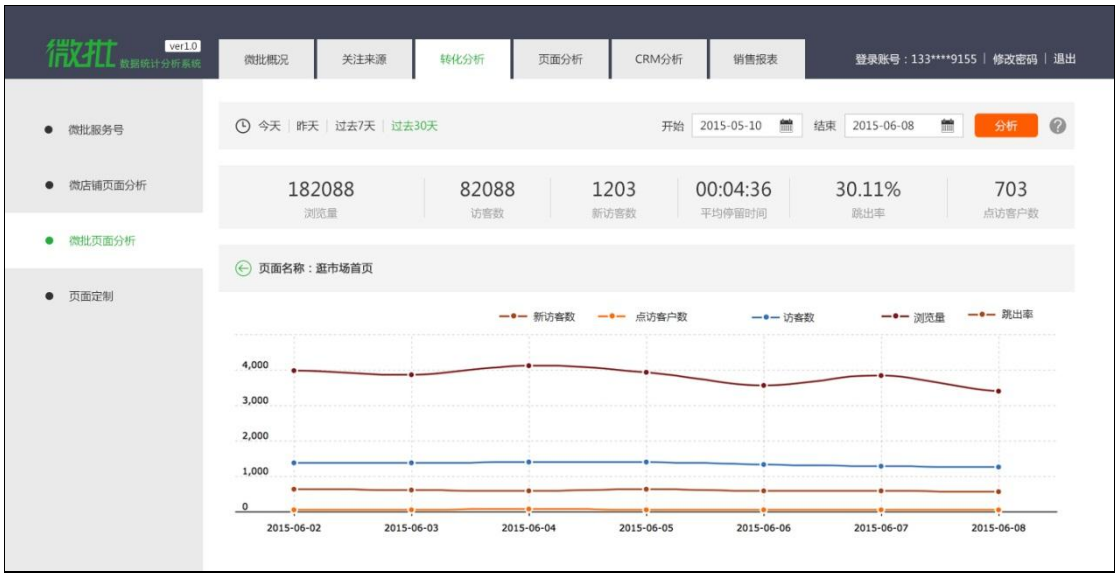


该图是中国电信的用户画像标签体系



该图是某团购网站的用户画像标签体系

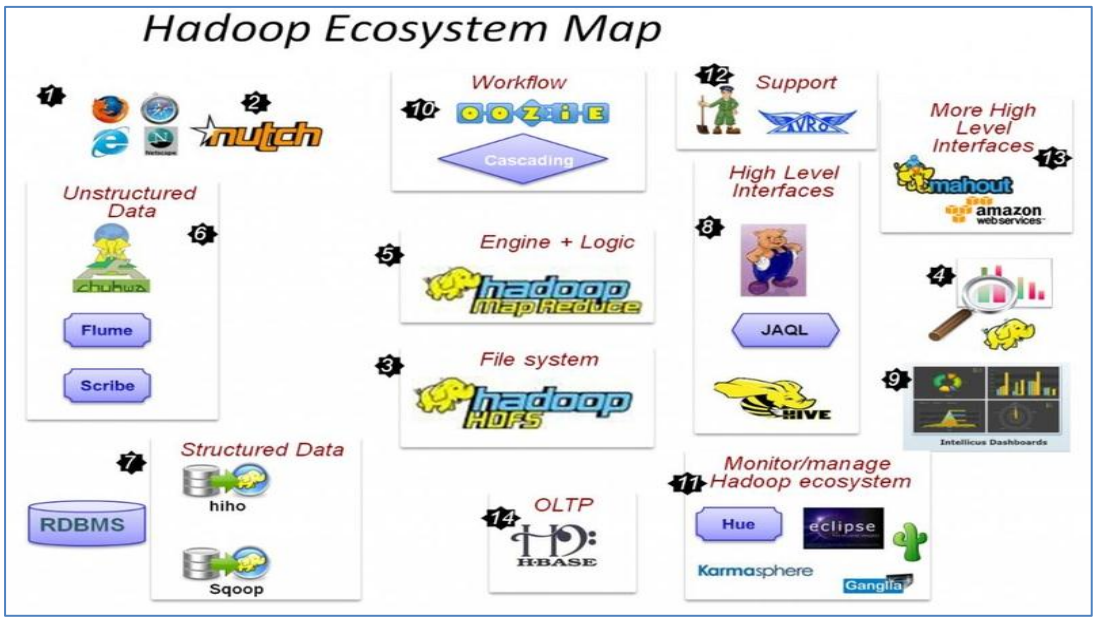
1.6.3、hadoop 用于网站点击流数据挖掘

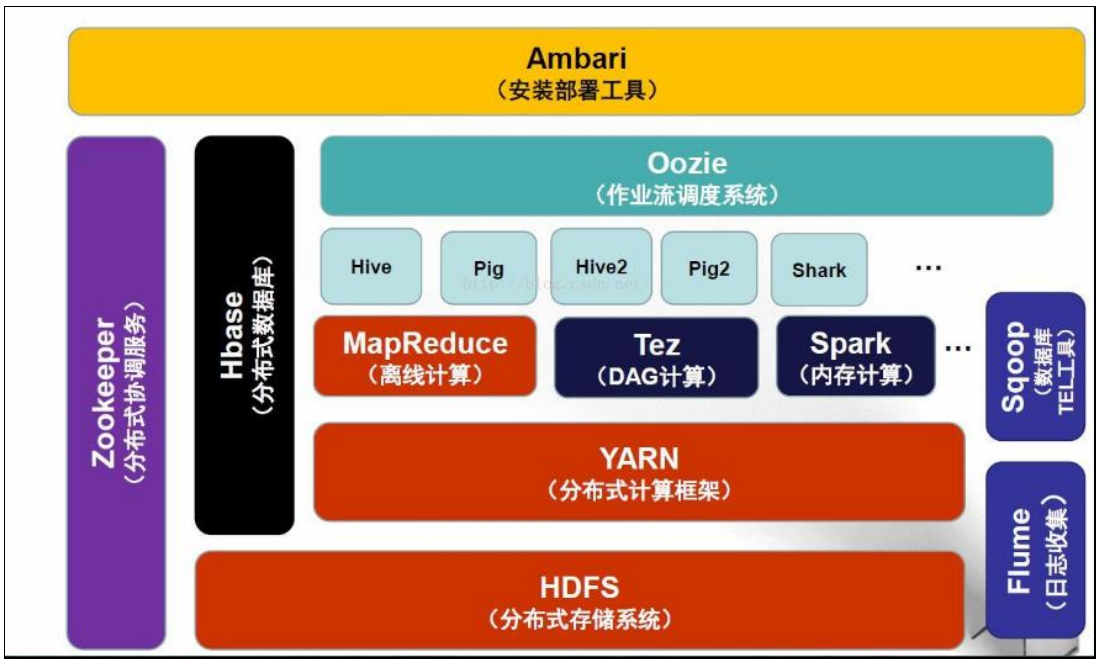


金融行业：个人征信分析
证券行业：投资模型分析
交通行业：车辆、路况监控分析
电信行业：用户上网行为分析
电商行业：用户浏览、购买行为分析
.....

最后总结：hadoop 并不会跟某个具体的行业或者某个具体的业务挂钩，它只是一种用来做海量数据分析处理的工具

1.7、hadoop 生态圈以及各组成部分的简介





重点组件：

HDFS：Hadoop 的分布式文件存储系统

MapReduce：Hadoop 的分布式程序运算框架，也可以叫做一种编程模型

Hive：基于 Hadoop 的类 SQL 数据仓库工具

HBase：基于 Hadoop 的列式分布式 NoSQL 数据库

ZooKeeper：分布式协调服务组件

Mahout：基于 MapReduce/Flink/Spark 等分布式运算框架的机器学习算法库

Oozie/Azkaban：工作流调度引擎
Sqoop：数据迁入迁出工具
Flume：日志采集工具

1.8、hadoop 就业情况及所需技能要求

1.8.1、hadoop 整体行业情况

A、大数据产业已纳入国家十三五规划
B、各大城市都在进行智慧城市项目建设，而智慧城市的根基就是大数据综合平台
C、互联网时代数据的种类，增长都呈现爆发式增长，各行业对数据的价值日益重视
D、相对于传统 JAVAEE 技术领域来说，大数据领域的人才相对稀缺
E、随着现代社会的发展，数据处理和数据挖掘的重要性只会增不会减，因此，大数据技术是一个尚在蓬勃发展且具有长远前景的领域

1.8.2、hadoop 就业职位要求

大数据是个复合专业，包括应用开发、软件平台、算法、数据挖掘等，因此，大数据技术领域的就业选择是多样的，但就 Hadoop 而言，通常都需要具备以下技能或知识

1、硬实力
A、Hadoop 分布式集群的平台搭建
B、Hadoop 分布式文件系统 HDFS 的原理理解及使用
C、Hadoop 分布式运算框架 MapReduce 的原理理解及编程
D、MySQL 数据库，Hive 数据仓库工具的熟练应用
E、Flume、Sqoop、Oozie/Azkaban 等辅助工具的熟练使用
F、Shell/Python 等脚本语言的开发能力

2、软实力
A、解决问题的能力（调试，阅读文档）
B、沟通协调能力（寻求帮助）
C、学习提升自己的能力（自我提高）
D、组织管控能力（管理能力）

1.8.3、hadoop 相关职位的薪资水平

大数据技术或具体到 HADOOP 的就业需求目前主要集中在北上广深一线城市，**薪资待遇普遍高于传统 JAVAEE 开发人员**，以北京为例：
北京 Hadoop：

职位名称	反馈率	公司名称	职位月薪	工作地点	发布日期
hadoop研发工程师		北京盛华合创科技有限公司	15001-20000	北京	10-31
Hadoop工程师		北京银丰新融科技开发有限公司	10001-15000	北京	10-31
BI工程师(Hadoop/大数据)	98%	北京易天新动网络科技有限公司	18000-25000	北京	10-31
hadoop开发工程师	99%	黑龙江邮政易通信息网络有限公司	10001-15000	北京 - 丰台区	10-31
北京分公司—架构师（HADOOP方向）	98%	哈尔滨乐辰科技有限责任公司	30001-50000	北京 - 朝阳区	10-31
Hadoop实习生	96%	上海蓬景数字营销策划有限公司	1000-2000	北京 - 朝阳区	10-31
Hadoop开发工程师	91%	上海蓬景数字营销策划有限公司	20000-40000	北京 - 朝阳区	10-31
Hadoop高级工程师	83%	北京科瑞明软件有限公司	15001-20000	北京	10-31
hadoop开发工程师 上市集团公司高薪 年薪50以上	80%	合优博奏管理咨询(北京)有限公司	30001-50000	北京	10-31
Hadoop系统设计师	79%	北京中亿华云科技有限公司	15001-20000	北京	10-31
大型互联网公司急聘Hadoop工程师	76%	合优博奏管理咨询(北京)有限公司	15001-20000	北京	10-31
高级hadoop开发	61%	北京中亿华云科技有限公司	15001-20000	北京	10-31
Hadoop工程师	60%	北京九美佐媒科技股份有限公司	10000-16000	北京 - 朝阳区	10-31

北京 Spark：

职位名称	反馈率	公司名称	职位月薪	工作地点	发布日期
OMG231-Spark高级研发工程师（北京）		深圳腾讯计算机系统有限公司 BEST	面议	北京	10-31
spark开发工程师	97%	北京奥维云网大数据科技股份有限公司	10000-20000	北京 - 朝阳区	10-31
spark开发工程师	83%	北京中亿华云科技有限公司	15001-20000	北京	10-31
Spark工程师	82%	北京数字新思科技有限公司	10001-15000	北京	10-31
Spark开发工程师（Java）		北京动力在线通信服务有限公司	6001-8000	北京	10-31
大数据开发工程师(spark)-103		北京金山云网络技术有限公司	15000-25000	北京	10-31
大数据开发工程师（hadoop/Spark）-QY		宜信公司	10000-20000	北京	10-31
大数据工程师(Spark)		北京理工新源信息科技有限公司	10000-20000	北京	10-31
Spark高级开发工程师		北京中油瑞飞信息技术有限责任公司	15001-20000	北京 - 昌平区	10-31
Hadoop/Spark大数据高级讲师（5险1金+双休+期权）		中科普开(北京)科技有限公司	15001-20000	北京 - 朝阳区	10-31
Spark开发工程师		北京中油瑞飞信息技术有限责任公司	10001-15000	北京 - 昌平区	10-31
大数据平台开发工程师(Spark)		北京东方金信科技有限公司	10001-15000	北京 - 海淀区	10-31

2、分布式系统概述

PS：由于大数据技术领域的各类技术框架基本上都是分布式系统，因此，理解 hadoop、storm、spark 等技术框架，都需要具备基本的分布式系统概念

概念讲解：

A．集群 + 负载均衡

B．分布式

1、 该软件系统会划分成多个子系统或模块，各自运行在不同的机器上，子系统或模块之间通过网络通信进行协作，实现最终的整体功能

2、 比如分布式操作系统、分布式程序设计语言及其编译(解释)系统、分布式文件系统和分布式数据库系统等。

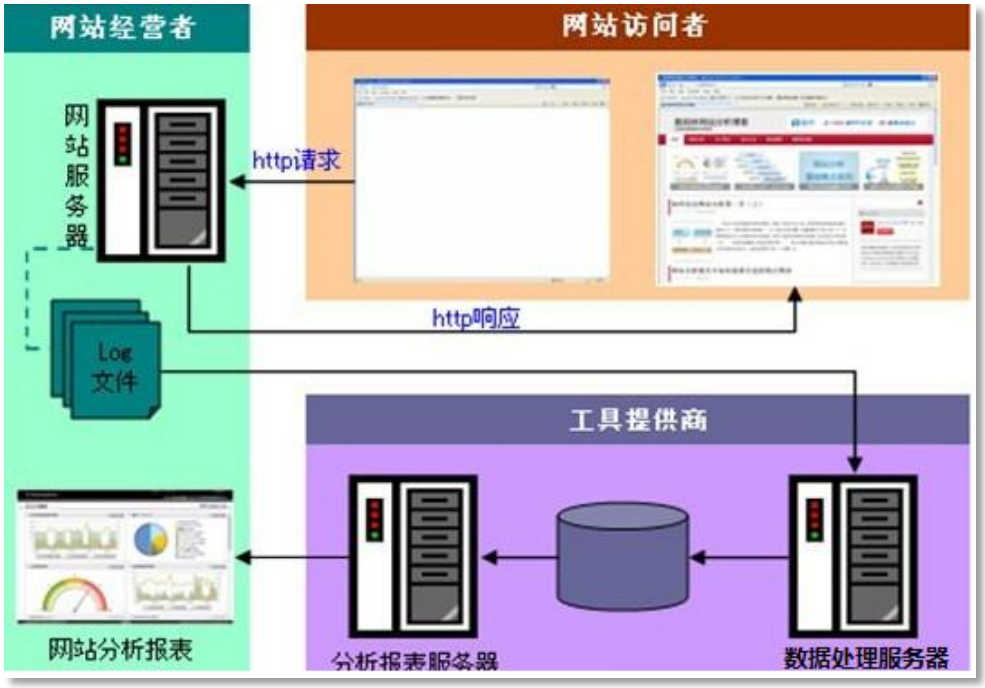
总结：利用多个节点共同协作完成一项或多项具体业务功能的系统就是分布式系统。

3、离线分析系统结构概述

PS：本环节主要感受数据分析系统的宏观概念及处理流程，初步理解 hadoop 等框架在其中的应用环节，不用过于关注具体实现细节

离线数据分析流程：

一个应用广泛的数据分析系统：**web 日志数据挖掘**



需求分析：

1、 案例名称：XX 网/XX app 点击流日志数据挖掘系统

2、 案例需求描述：“Web 点击流日志”包含着网站运营很重要的信息，通过日志分析，我们可以知道网站的访问量，哪个网页访问人数最多，哪个网页最有价值，广告转化率、访客的来源信息，访客的终端信息等。

3、 数据来源：本案例的数据主要由用户的点击行为记录

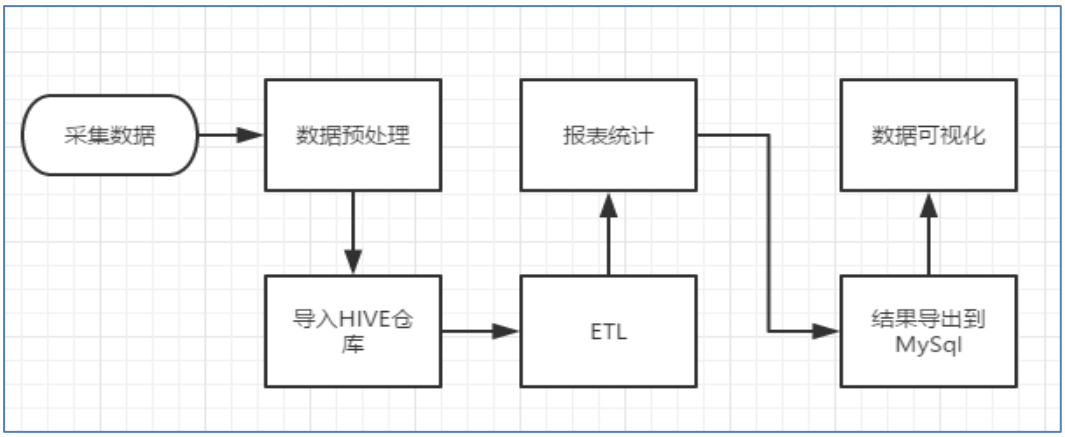
4、 获取方式：在页面预埋一段 js 程序，为页面上想要监听的标签绑定事件，只要用户点击

或移动到标签，即可触发 ajax 请求到后台 servlet 程序，用 log4j 记录下事件信息，从而在 web 服务器（nginx、tomcat 等）上形成不断增长的日志文件。

形如：

```
58.215.204.118 - - [18/Sep/2013:06:51:35 +0000] "GET /wp-includes/js/jquery/jquery.js?ver=1.10.2 HTTP/1.1" 304 0 "http://blog.fens.me/nodejs-socketio-chat/" "Mozilla/5.0 (Windows NT 5.1; rv:23.0) Gecko/20100101 Firefox/23.0"
```

5、 数据处理流程



A、数据采集：定制开发采集程序，或使用开源框架 Flume 或者 LogStash

B、数据预处理：定制开发 MapReduce 程序运行于 Hadoop 集群，或者专门数据收集工具也能进行数据预处理

C、数据仓库技术：基于 Hadoop 之上的 Hive

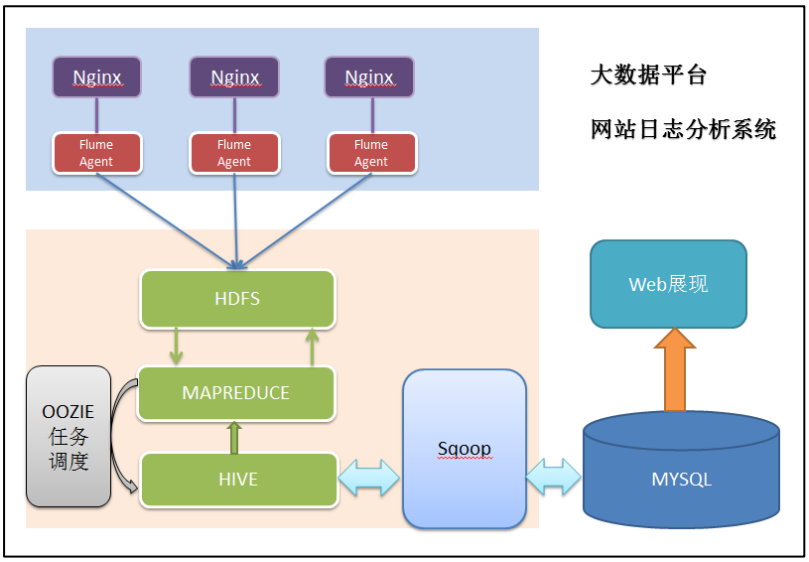
D、数据导出：基于 Hadoop 的 Sqoop 数据导入导出工具

E、数据可视化：定制开发 web 程序或使用 Kettle 等产品

F、数据统计分析：Hadoop 中的 MapReduce 或者基于 Hadoop 的 Hive，或者 Spark，Flink

G、整个过程的流程调度：Hadoop 生态圈中的 Oozie/Azkaban 工具或其他类似开源产品

6、 项目整体技术架构图



7、 项目相关截图

A、 MapReduce 运行：

```
[hadoop@hadoop02 ~]$ hadoop jar apps/hadoop-2.6.5/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.6.5.jar wordcount /w
c/input /wc/output
17/07/30 19:16:52 INFO input.FileInputFormat: Total input paths to process : 1
17/07/30 19:16:52 INFO mapreduce.JobSubmitter: number of splits:1
17/07/30 19:16:53 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1501413124905_0001
17/07/30 19:16:53 INFO impl.YarnClientImpl: Submitted application application_1501413124905_0001
17/07/30 19:16:53 INFO mapreduce.Job: The url to track the job: http://hadoop04:8088/proxy/application_1501413124905_0001
/
17/07/30 19:16:53 INFO mapreduce.Job: Running job: job_1501413124905_0001
17/07/30 19:17:07 INFO mapreduce.Job: Job job_1501413124905_0001 running in uber mode : false
17/07/30 19:17:07 INFO mapreduce.Job:  map 0% reduce 0%
17/07/30 19:17:21 INFO mapreduce.Job:  map 100% reduce 0%
17/07/30 19:17:36 INFO mapreduce.Job:  map 100% reduce 100%
17/07/30 19:17:36 INFO mapreduce.Job: Job job_1501413124905_0001 completed successfully
17/07/30 19:17:37 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=7183
                FILE: Number of bytes written=234237
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10180
                HDFS: Number of bytes written=6107
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
```

B、 在 Hive 数据仓库中查询数据：

```
hadoop02 | hadoop03 | hadoop04 | hadoop05
hive> select count(*) from weibo;
Query ID = hadoop_20170730191820_9a4b7cac-a612-4974-b8e7-7d695af194b0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1501413124905_0002, Tracking URL = http://hadoop04:8088/proxy/application_1501413124905_0002/
Kill Command = /home/hadoop/apps/hadoop-2.6.5/bin/hadoop job  -kill job_1501413124905_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-07-30 19:18:37,360 Stage-1 map = 0%,  reduce = 0%
2017-07-30 19:18:47,660 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.89 sec
2017-07-30 19:18:57,344 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.66 sec
MapReduce Total cumulative CPU time: 3 seconds 660 msec
Ended Job = job_1501413124905_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.66 sec   HDFS Read: 30966905 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 660 msec
OK
83026
Time taken: 39.11 seconds, Fetched: 1 row(s)
hive>
```

C、 Sqoop 运行演示

```
./sqoop export    \
```

```
--connect jdbc:mysql://localhost:3306/weblogdb    \
--username root    \
--password root    \
--table t_display_xx    \
--export-dir /user/hive/warehouse/uv/dt=2014-08-03
```

8、 项目最终效果