Spark是

公开课就是Spark-Streaming 仅次于spark-sql

Spark学的好 找个工作不难

Spark逐渐成为大数据处理技术 ----- 人才市场受欢迎

----- Spark-streaming的介绍

以下五个方面

Ppt1

有很多数据 ----大数据是

京东 --- 搜索鞋子 记录哪个ip 记录一条搜索日志 ----- 评论有一个评论日志

点击一次 就会转换成一条日志 这是典型的流式数据

机器日志 ----- 数据保存到监控系统里 当机器非常有多 指标非常多 每台机器采集了1000个指标 每秒就是有10w\*1000条数据 ---- 这样 数据像流水一样产生

源源不断的采集到大数据系统里面 ---- 这种称为流式数据

流式数据的来源

移动终端 IOS 网站 物联网设备都会产生数据 ------ 使用那些应用来处理这些流式数据来产生价值呢？

流式数据的处理的应用场景？

\*\*实时分析用户发送的微博 微信数据 ---- 找出关键词 然后 看一下这个关键词的趋势

\*\*实时推荐 ---淘宝 推荐 --- 每个页面的右边或者下面 会根据你的行为 推荐一些内容 ---- 然后就是推荐各种各样的袜子 这就是实时推荐的过程

阿里成为了微博的一个股东 ----- 这样 淘宝给一个实时推荐的结果 推荐各种运动鞋 然后退出 登录到微博 还会有这个运动鞋 两个系统通了

\*\*网站指标的实时统计

------

典型的应用就是mapreduce spark

流式就是 来一条 我就处理一条 这个数据延迟非常低 典型的代表 就是storm

两者之间的 批处理 转化为流式处理

微批处理 --- 典型代表就是spark streaming

大的 就是批处理 和 流式处理

中间级别的引擎

Lambda architecture spark后续课程 会详细介绍

现在Spark非常火

Flick 和Apex都是比Spark晚一点

这个课程的特色是 有互动 一个小时 半小时讲演示 最后半小时是互动 ----- 并且基于Spark2.1 第三个是

流式处理 的四个模块

这个是非常抽象的知道思想

具体的架构如下

Android IOS、Site 数据写入到Kafka ---- 数据缓存到Kafka 然后 使用流式引擎来计算 最后 存储

数据收集不分 实时收集 架构一般是

公司机房 部署一个Kafka的集群 ---- 前面 有公司的各种各样的产品

在各种设备上看公司的产品

这些数据 有一个负载均衡器 把用户的观看行为 或者浏览行为发送到Http Server(Tomcat /Apache) ---- 简单汇总 就写入到Kafka

Kafka是分布式的消息队列 是一个生产者消费者模型 ----- 中建是一个分布式缓存的队列 多个broker

存储在Kafka中的数据 以 topic的形式进行组织

消息的队列 就是不同类型的数据 放到不同的主题下面

读到之后 数据被删掉 ---- 缓存数据可以有一定的过期时间

Spark streaming ----将

RDD是Spark中非常重要的一个模块

Spark streaming提供了非常丰富的算子 来复杂的计算 比Storm容易很多

存储的MySQL ------ 可以进一步进行数据分析