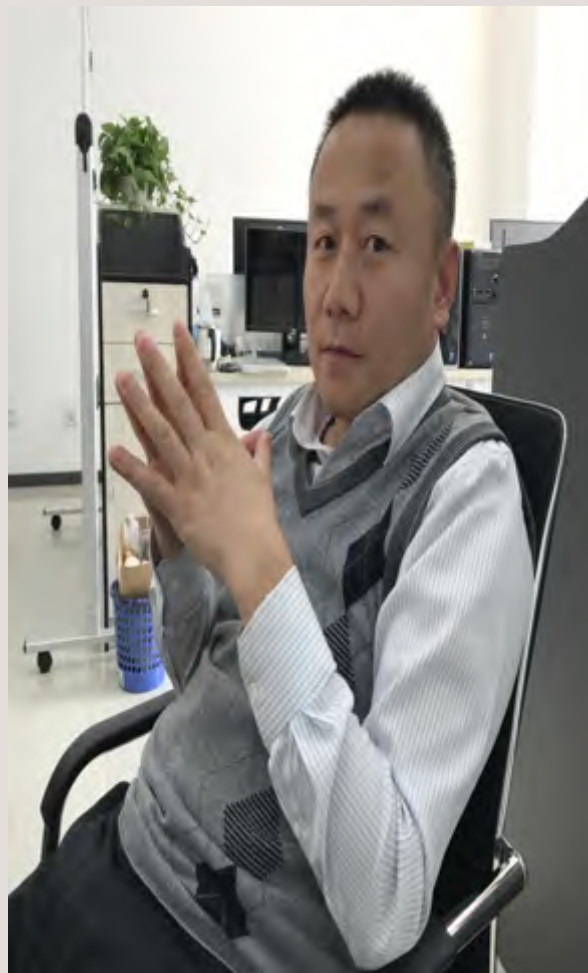




用户行为分析大数据平台的选型之路

吴江林

个人简介



吴江林，北京交通大学计算机专业硕士、北京师范大学工商管理硕士，曾就职于雅虎、阿里巴巴、百合网等大型互联网公司，拥有近20年互联网行业从业经历，在数据挖掘、数据建模、虚拟化及云计算领域拥有丰富的研发经验。目前就职于北京极数科技有限公司，担任CTO一职。



使用AVRO序列化数据

- Hadoop的子项目
- 基于二进制传输
- 高性能中间件
- 支持数据密集型应用

AVRO特点

- 依赖模式实现数据结构定义
- 支持二进制编码和JSON编码
- 支持压缩功能
- 支持数据排序功能
- 序列化时速度快且结果内容少





使用Parquet列式存储

Parquet是面向分析型业务的列式存储格式，由Twitter和Cloudera合作开发，2015年5月升级为Apache顶级项目。

Parquet列式存储特点

- 可以跳过不符合条件的数据，只读取需要的数据，降低I/O吞吐量
- 压缩编码可以降低磁盘存储空间
- 支持向量运算，能够获取更好的扫描性能
- 查询性能提升30倍，存储空间节约75%



列式存储 VS 行式存储

- 传统的RDBMS提供ACID原子操作和事务，在大多数列式存储数据库上是不具备的，而大多数列式存储数据库对分布式架构支持的更好。
- 需要了解的是列式存储以及行式存储在查询和存储上的本质，在技术选型上可以多点筹码。





使用 Apache Sqoop

- ETL转换工具
- RDBMS \Leftrightarrow HDFS
- 对某些NoSQL数据库亦提供支持

使用Apache Kafka

Kafka是一种高吞吐量分布式发布订阅消息系统，它可以处理消费级规模网站中的所有动作流数据。



Kafka特点



- 实时数据流 (real-time data streaming)
- 水平可扩展 (horizontally scalable)
- 容错机制 (fault tolerance)
- 支持分区 (partition)
- 出奇的快 (wicked fast)

使用Apache Spark

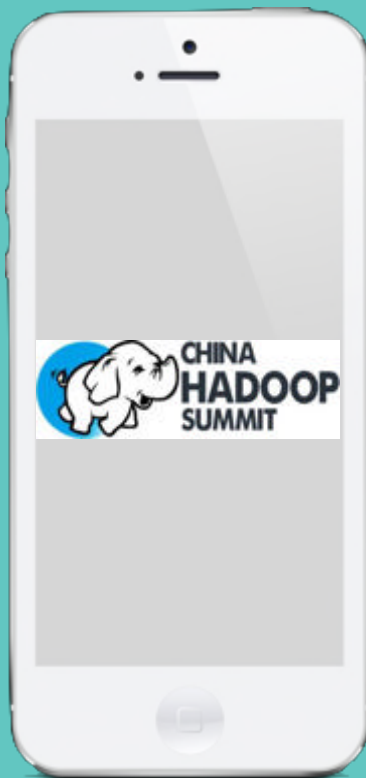
- 高性能，执行速度100倍于Hadoop
- 支持Java、Scala、Python、R语言
- 支持SQL、机器学习、图运算、流
- 支持Hadoop、Cassandra、HBase
- 支持独立部署运行



Spark基于这样的理念：当数据庞大时，把计算过程传递给数据要比把数据传递给计算过程要更有效率。每个节点存储它的数据集，然后任务被提交给节点。

LOREM
IPSUM

使用Druid



Druid是一个面向列式存储
的高性能分布式数据仓库。

Druid特点



- 高速聚合，亚秒级OLAP查询
- 流数据注入，实时多维度查询
- 集群动态扩展，PB级数据量，每秒百万级事件注入
- 低成本部署，与现有数据系统无缝集成



使用Impala进行离线查询

数极客



Impala是Cloudera公司主导开发的新型查询系统，它提供SQL语义，能查询存储在Hadoop的HDFS和HBase中的PB级大数据。



Impala的特点

- 中间结果放在内存中，节约大量I/O开销
- 省掉了MapReduce作业启动的开销
- 采用数据仓库MPP设计思想
- C++实现，做了很多有针对性的硬件优化
- 支持Data locality的I/O调度机制

使用Apache HIVE

Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。



HIVE特点

- 支持索引，加快数据查询
- 不同的存储类型
- 将元数据保存在关系数据库中，大大减少了在查询过程中执行语义检查的时间。
- 可以直接使用存储在Hadoop 文件系统中的数据。
- 支持UDF函数来完成内置函数无法实现的操作。
- 类SQL 的查询方式，将SQL 查询转换为MapReduce 的job 在Hadoop集群上执行。





使用Apache HDFS

- Hadoop分布式文件系统(HDFS)被设计成适合运行在廉价硬件上的分布式文件系统。
- HDFS在最开始是作为Apache Nutch搜索引擎项目的基础架构而开发的，目前是Apache Hadoop Core项目的一部分。



HDFS特点

- 高容错性 (fault-tolerant)
- 流式访问 (streaming access)
- 支持千万级别文件 (tens of millions)
- 异构平台可移植性 (heteron portablity)



大数据技术平台逻辑架构



数极客



Thanks!

