架构部 大数据
王刚

聚美优品

大数据应用
SQL和NoSQL融合实践

声明

本文所有内容均为个人主观理解

部分资料来源于网络

大纲

- SQL & NoSQL
- 曾经的 NoSQL 神话
- NoSQL 分类
- SQL on Hadoop
- SQL on Hadoop 聚美实践
- SQL and Hadoop

SQL & NoSQL

SQL

- ACID
- Join
- Schema
- ……

NoSQL

- CAP
- BASE
- Schema-less

✧ Not SQL
✧ No! SQL
✧ Not only SQL

SQL & NoSQL

- CAP
  - Consistency (一致性)
  - Availability (可用性)
  - Tolerance of network Partition (分区容忍性)
- Consistency
  - strong consistency (强一致性 - ACID)
  - weak consistency (弱一致性 - BASE)
    - Eventually consistent (最终一致性)

NoSQL 神话

1. NoSQL 将取代 SQL

1. NoSQL 比 SQL 更好/更坏

1. 语言+框架决定数据库选型

NoSQL 分类

分类	举例	典型应用场景	数据模型	优点	缺点
键值存储 (key-value)	Tokyo Cabinet Tyrant Redis Voldemort Oracle BDB	内容缓存 主要用于处理大量数据的高访问负载 也可用于一些日志系统等。	Key 指向 Value 的键值对 通常用hash table实现	查找速度快	数据无结构化 通常只被当作字符串或者二进制数据
列存储	Cassandra HBase Riak	按列存储数据 方便存储结构化和半结构化数据 对数据压缩友好 查询范围在一列或几列查询时有极大的IO 优势	以列簇式存储 将同一列数据存在一起	查找速度快，可扩展性强 容易进行分布式扩展	功能相对局限
文档存储	CouchDB MongoDb	一般用类似 json 的格式存储, 存储内容是文档型 可以对某些字段建立索引实现关系型数据库某些能力	Key-Value对应的键值对 Value为结构化数据	数据结构要求不严格，表结构可变 不需要像RDMS一样需要预定义表结构	查询性能不高 缺乏统一的查询语法
图存储 (Graph)	Neo4j InfoGrid Infinite Graph	社交网络，推荐系统等。专注于构建关系图谱	图结构	利用图结构相关算法	很多时候需要对整个图做计算 不太好做分布式的集群方案

- 易扩展
- 大数据量
- 数据模型灵活
- 高可用

NoSQL 分类

- **Column**: Accumulo, Cassandra, Druid, HBase, Vertica.
- **Document**: Apache CouchDB, ArangoDB, BaseX, Clusterpoint, Couchbase, Cosmos DB, IBM Domino, MarkLogic, MongoDB, OrientDB, Qizx, RethinkDB
- **Key-value**: Aerospike, Apache Ignite, ArangoDB, Couchbase, Dynamo, FairCom c-treeACE, FoundationDB, InfinityDB, MemcacheDB, MUMPS, Oracle NoSQL Database, OrientDB, Redis, Riak, Berkeley DB, SDBM/Flat File dbm, ZooKeeper
- **Graph**: AllegroGraph, ArangoDB, InfiniteGraph, Apache Giraph, MarkLogic, Neo4J, OrientDB, Virtuoso
- **Multi-model**: Apache Ignite, ArangoDB, Couchbase, FoundationDB, InfinityDB, MarkLogic, OrientDB, Cosmos DB

NoSQL 分类

Type	Notable examples of this type
Key-Value Cache	Apache Ignite, Coherence, eXtreme Scale, Hazelcast, Infinispan, Memcached, Velocity
Key-Value Store	ArangoDB, Aerospike
Key-Value Store (Eventually-Consistent)	Oracle NoSQL Database, Dynamo, Riak, Voldemort
Key-Value Store (Ordered)	FoundationDB, InfinityDB, LMDB, MemcacheDB
Data-Structures Server	Redis
Tuple Store	Apache River, GigaSpaces
Object Database	Objectivity/DB, Perst, ZopeDB
Document Store	ArangoDB, BaseX, Clusterpoint, Couchbase, CouchDB, DocumentDB, IBM Domino, MarkLogic, MongoDB, Qizx, RethinkDB
Wide Column Store	Amazon DynamoDB, Bigtable, Cassandra, Druid, HBase, Hypertable

SQL on Hadoop

- Batch SQL
  - 批量查询, 数据挖掘, 建模, 大规模 ETL
  - 单位: 分钟或小时
- Interactive SQL
  - 交互查询, 报表
  - 单位: 秒
- In-Memory SQL
  - 内存计算
  - 单位: 秒, 分钟
- Operational SQL
  - 侧重 OLTP, 单点查询, 超低延迟
  - 单位: 毫秒

SQL on Hadoop

- Batch SQL
  - Hive
- Interactive SQL
  - Impala, Drill, Presto
- In-Memory SQL
  - Spark
- Operational SQL
  - HBase

SQL on Hadoop

- MPP架构
  - 查询速度快, 毫秒或秒级
  - 粗粒度容错, 容错性差
  - 可横向扩展(一定范围内)
  - 并发不会随集群扩大而明显提高
  - 不适合大规模部署, 建议100节点以内
- 非 MPP 架构
  - 速度一般比 MPP 慢
  - 细粒度容错
  - 可扩展至上万节点
  - 适合大规模及超大规模部署

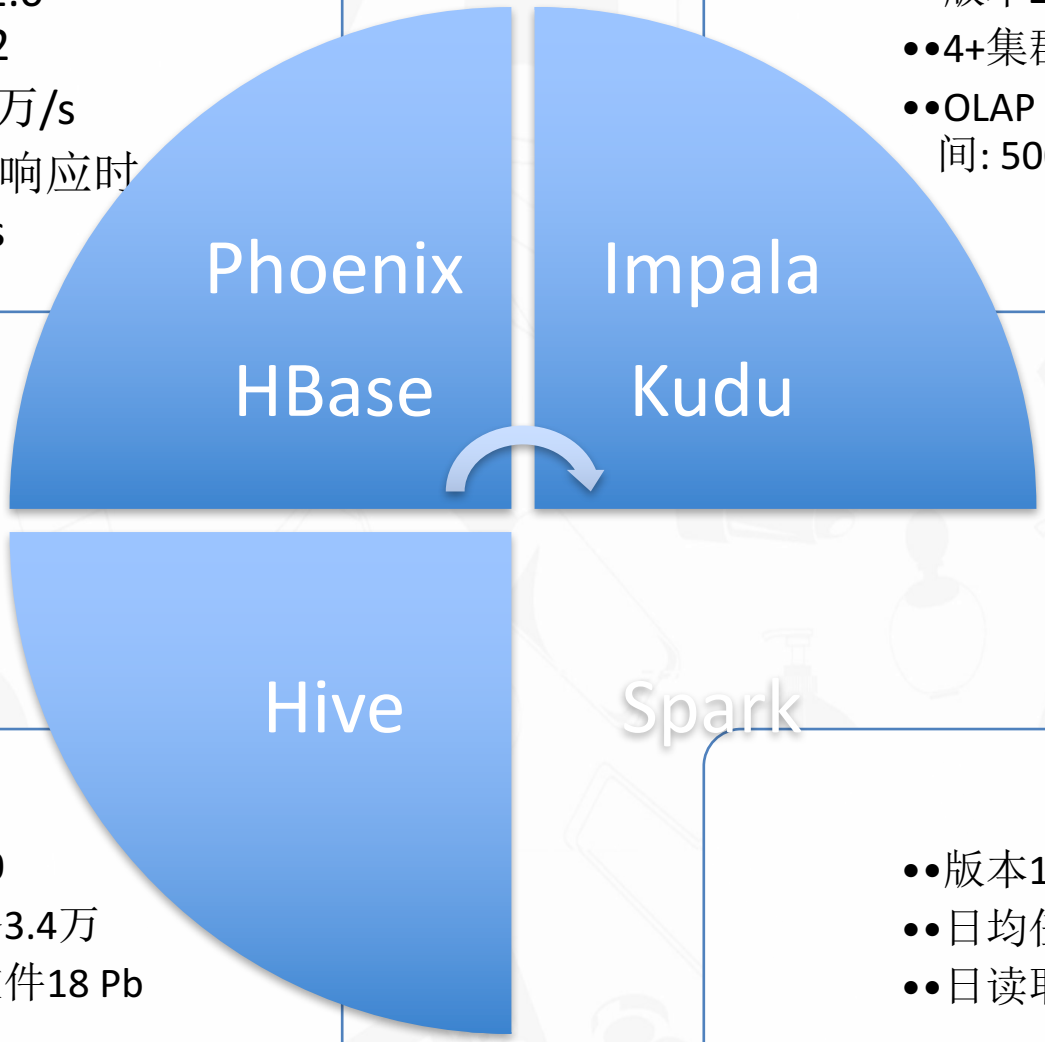
SQL on Hadoop

	scan读	随机读	写	删改	速度	稳定性	sql 支持
hive	极佳	不支持 可用 mapreduce	不支持 需写 hdfs	不支持	极慢	极佳	极佳
spark sql	极佳	不支持 可用 filter	较弱 写 hdfs 或 save api	不支持	快	一般	极佳
impala+kudu	极佳	其实是 scan	极佳	支持 仅通过 key 效率一般	快	优秀	极佳
Phoenix+HBase	一般	key和index极快 其他为scan较慢	极佳	支持 仅通过 key 效率极高	key和index极快 非index较慢	一般	一般
elasticsearch	优秀	极佳	极佳	支持 改效率一般	极快	优秀	较差 不支持join

SQL on Hadoop

- 优化器
  - 主要使用RBO, 少部分CBO
  - 优化核心为 shuffle 和 join
- 硬件相关策略
  - 磁盘优化
    - 本地化, 压缩, 列存, 分区, 块索引, disk-aware
  - CPU 优化
    - 向量化, 动态代码生成, 压缩算法
  - 内存 & CPU 缓存
    - 堆外存储, 缓存数据结构

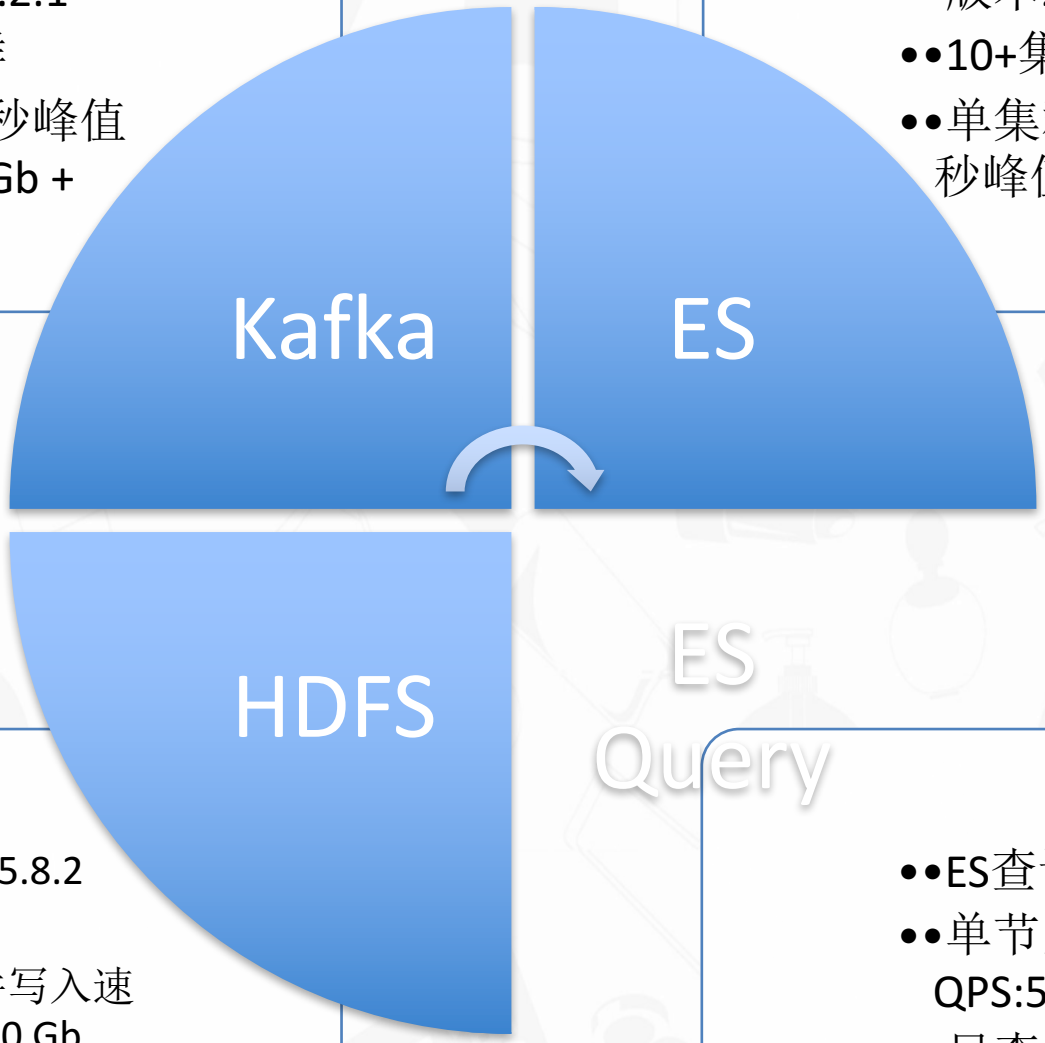
SQL on Hadoop



SQL on Hadoop 聚美实践

Phoenix HBase	••HBase 是列存储数据库, Phoenix 为SQL 封装 ••超高写入/读取性能(写70万+, 读45万+QPS) ••极度依赖 rowkey, 对 rowkey 设计要求较高
Impala Kudu	••Impala是MPP引擎, 极限压榨服务器CPU/内存/磁盘 ••Kudu 为针对 OLAP 场景优化设计的分布式存储引擎 ••Impala+Kudu 可以实现近实时的动态更新数据统计
Hive	••Hadoop生态中最稳定组件之一 ••磁盘密集型计算框架, 最终翻译成mapreduce执行 ••经过适当调优, 可应对 PB 级数据分析
Spark	••Hadoop生态最火组件之一 ••内存密集型计算框架, 可达到Hive性能的3-20倍 ••支持SQL, 流式处理, 机器学习, 图计算等

SQL on Hadoop 聚美实践



SQL on Hadoop 聚美实践

Kafka	••消息队列 ••适用于可少量丢失的消息传输(实际上没丢过) ••客户端支持 java，go支持尚可, php支持较弱
ES	••全文检索引擎 ••JSON 为主, 二次开发插件支持 SQL
ES Query	••Elasticsearch 查询接口 ••对 es 原生的深度包装, 支持 SQL 和 DSL
HDFS	••Hadoop 分布式文件系统 ••一次写入, 不可更改, 读取性能和副本数有关 ••目前 hadoop 生态中最稳定的底层组件

SQL and Hadoop

- Oracle Big Data SQL
  - 统一查询层, 跨 Oracle 和 Hadoop
  - Oracle 自定义元数据体系
  - 继承 Hive 元数据
  - 商业, 收费
- Presto
  - 多数据源, 含 MySQL, PostgreSQL 和 Hive 等
  - 兼容各个数据源元数据
  - 开源, 免费

谢 谢

聚美优品
JUMEI.COM