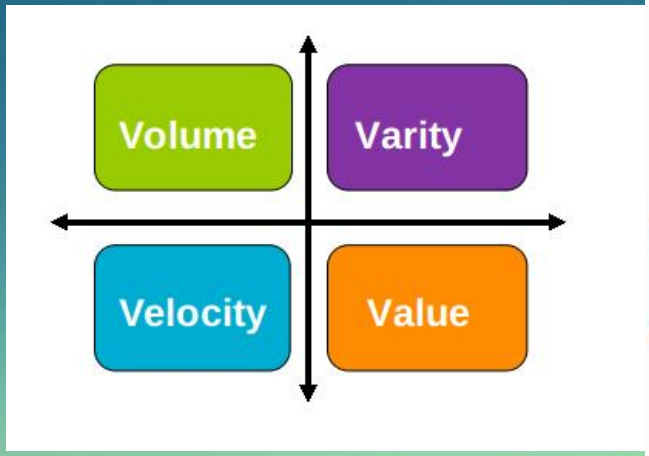
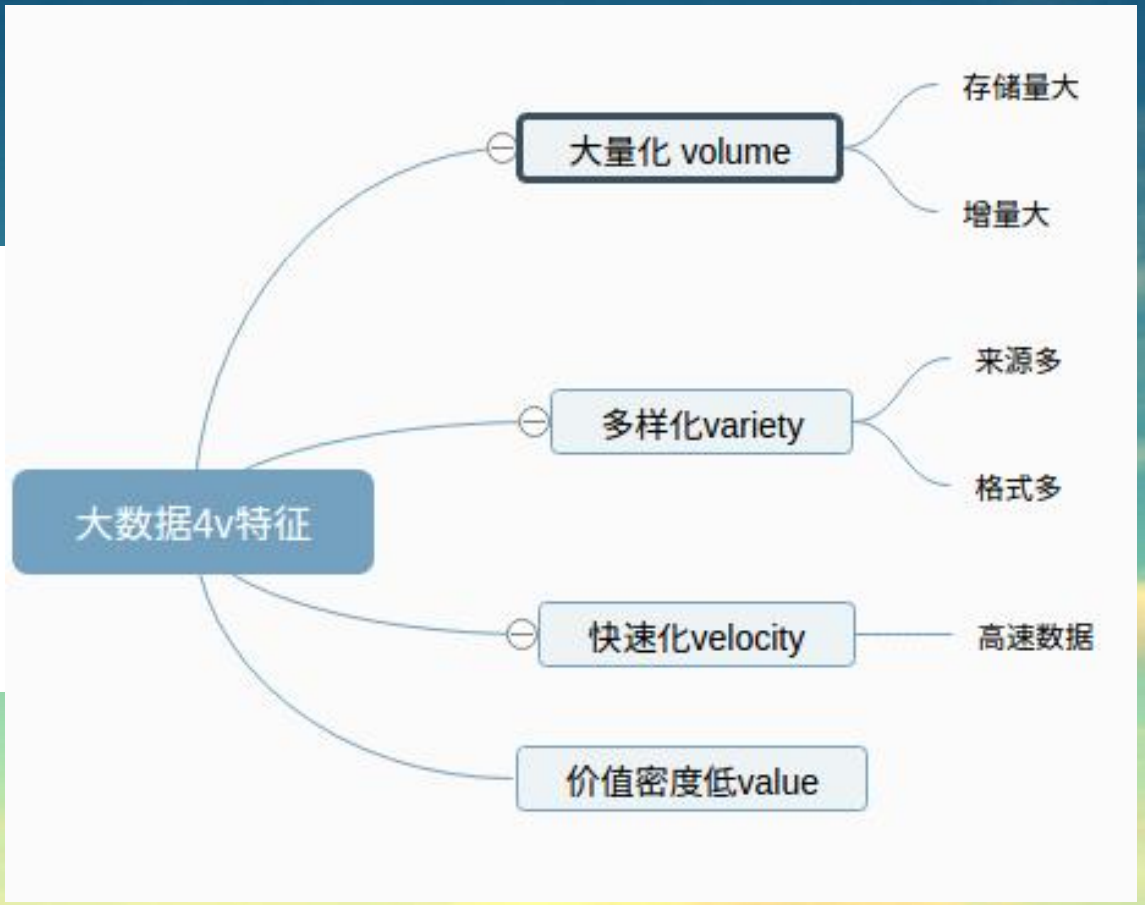
北京易腾时代科技有限公司

大数据

Robin Gao 2016-06-28

1.什么是大数据?


Volume
Varity
Velocity
Value


大数据4v特征
大量化 volume
存储量大
增量大
多样化variety
来源多
格式多
快速化velocity
高速数据
价值密度低value

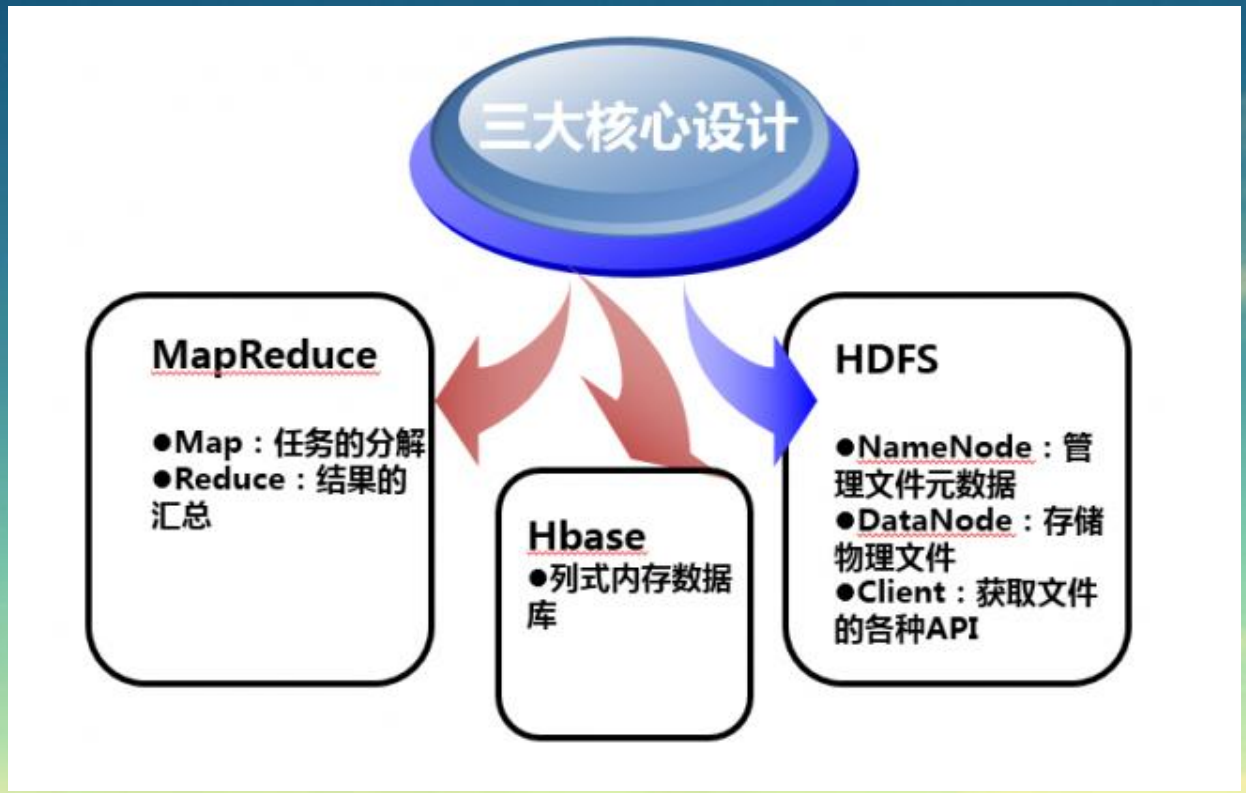
2.大数据技术起源

1.Google File System : 用来解决数据存储的问题

2.Map-Reduce : 函数式编程

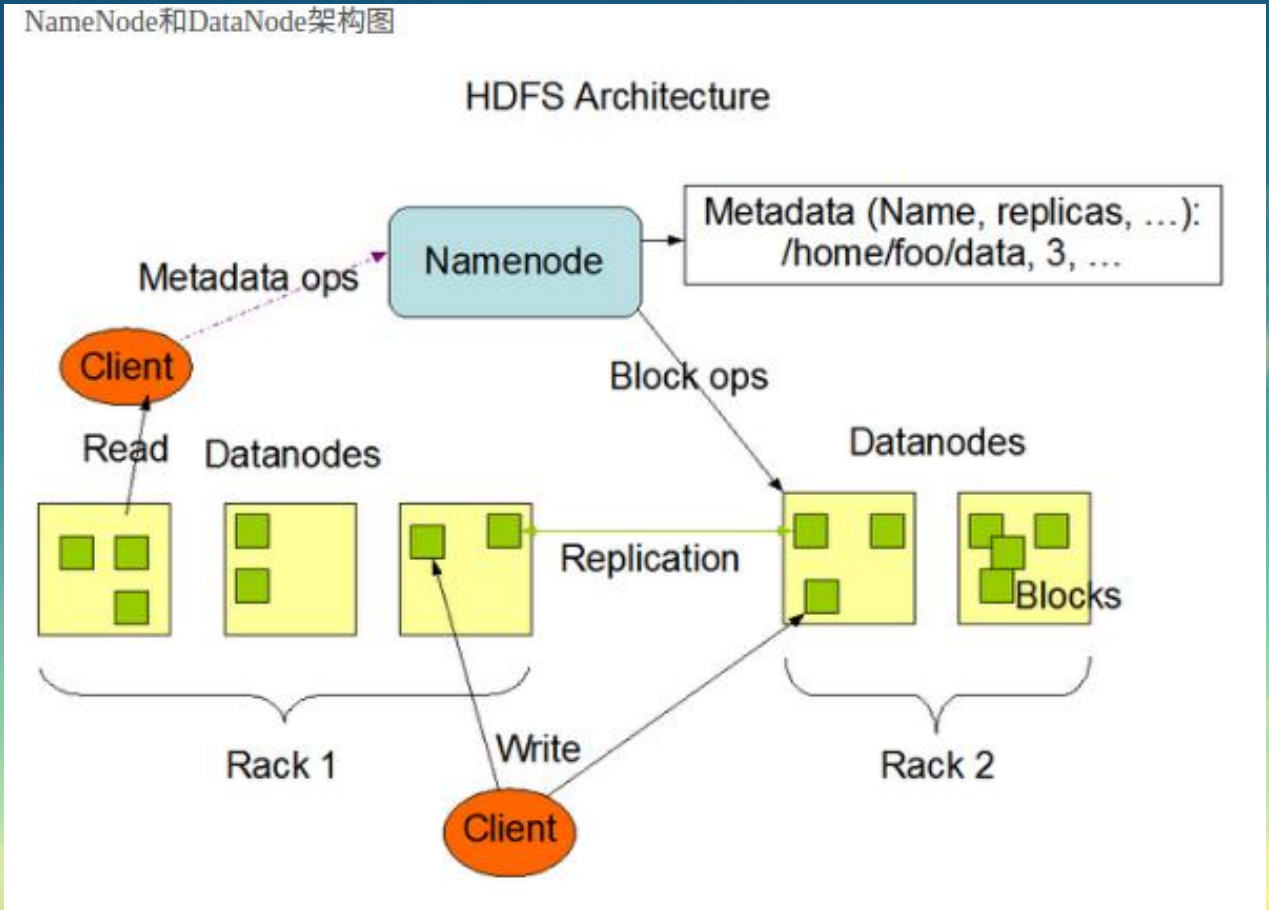
3.BigTable : 在分布式系统上存储结构化数据的
一个解决方案

3.最初hadoop核心设计

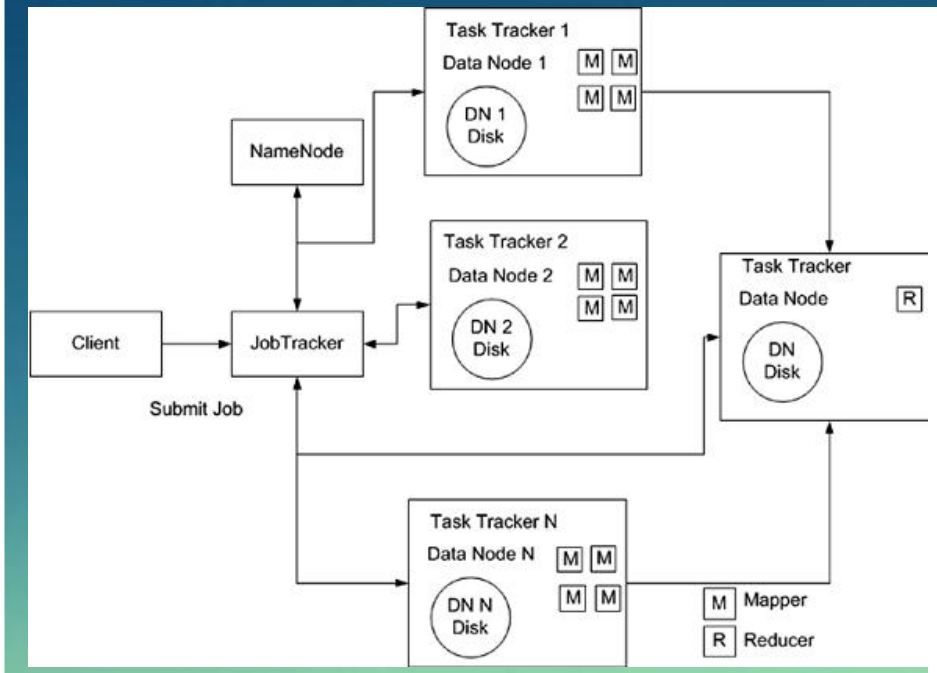


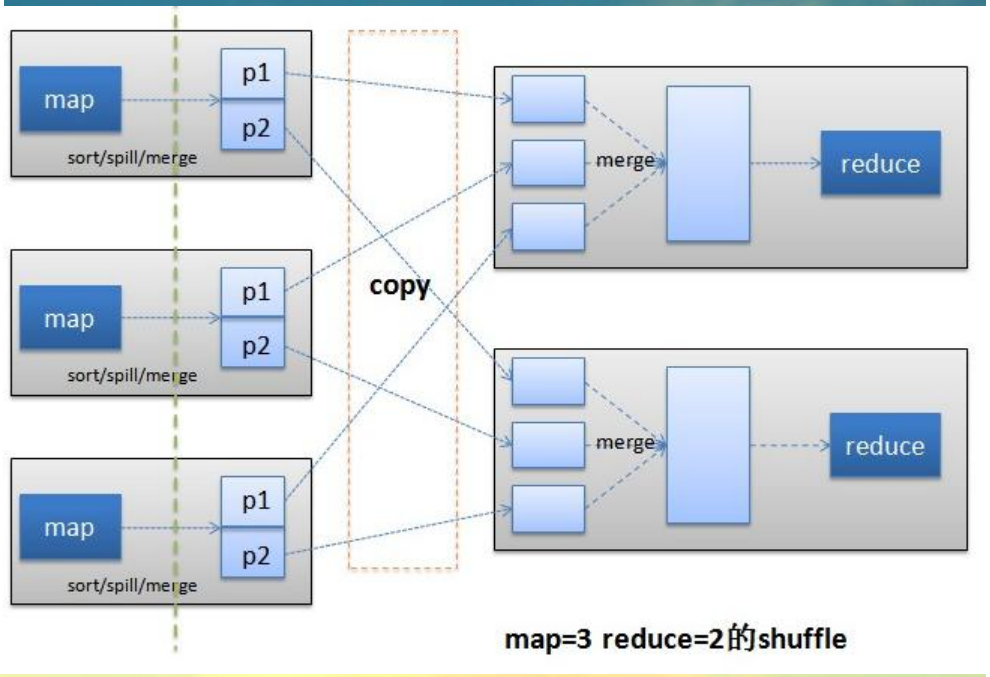
3.1 HDFS

NameNode和DataNode架构图

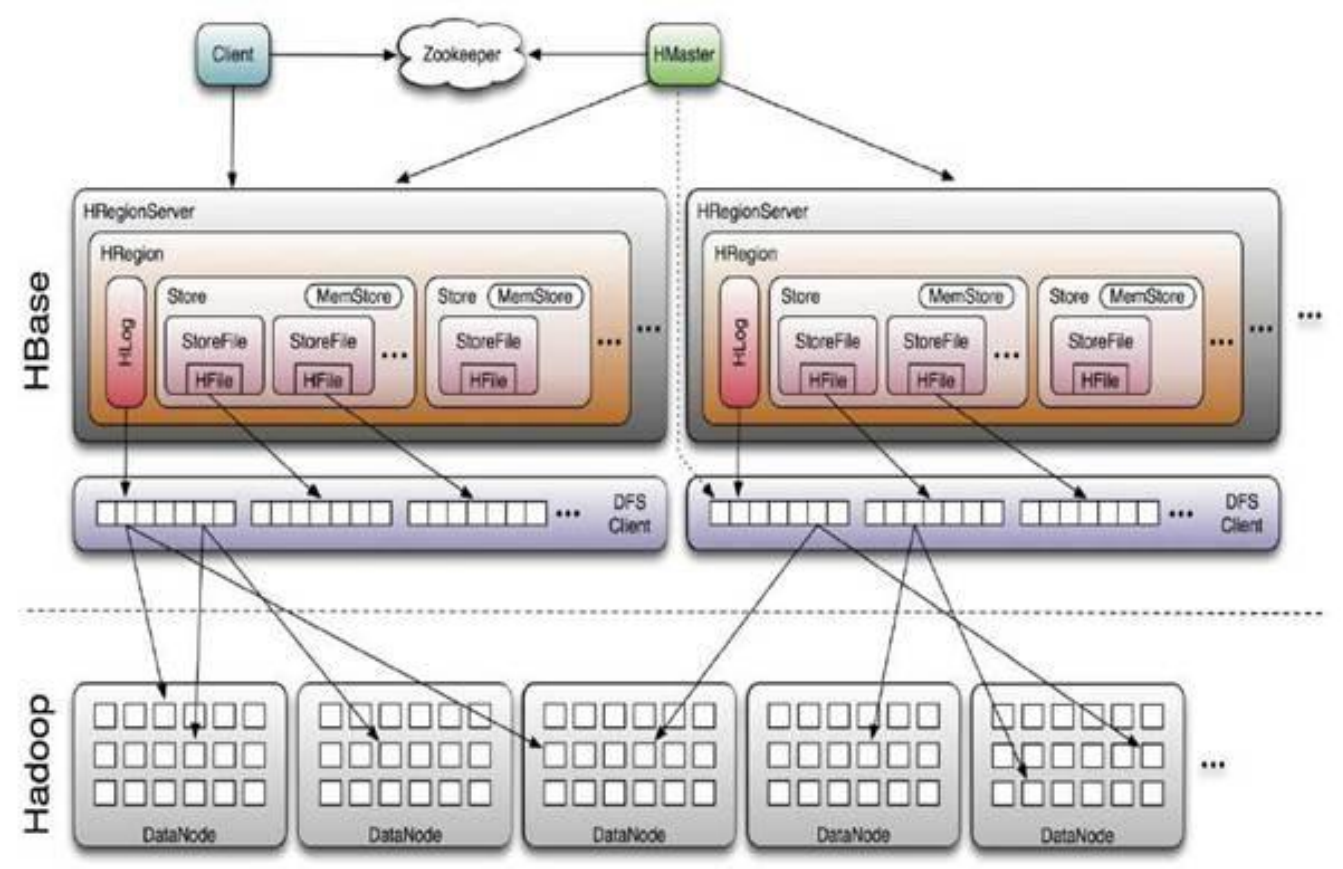


3.2 MapReduce

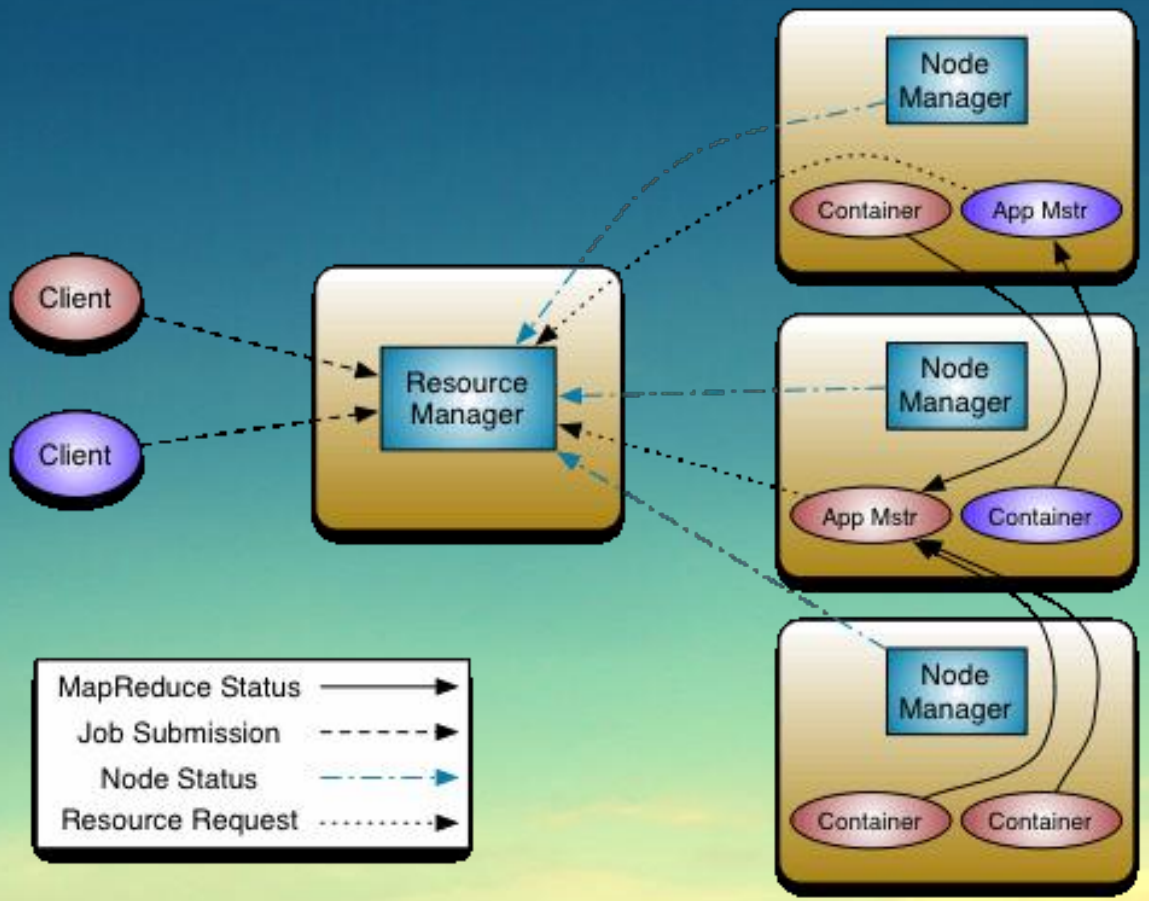




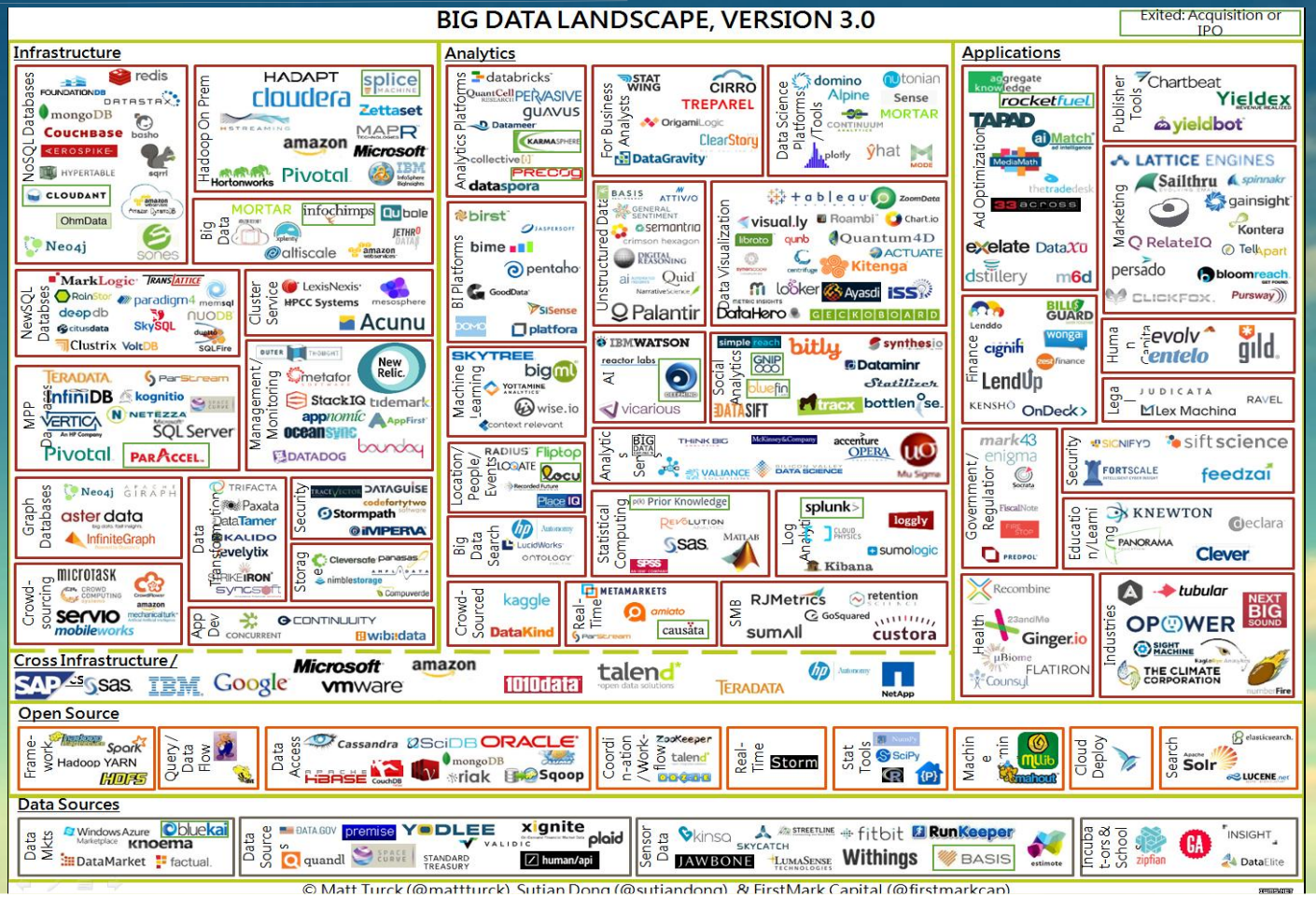
3.3 HBase


Client
Zookeeper
HMaster
HBase
HRegionServer
HRegion
HLog
Store
MemStore
StoreFile
HFile
DFS Client
Hadoop
DataNode

3.4 YARN


Client
Client
Resource Manager
Node Manager
Container
App Mstr
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request

4. 大数据生态图谱


BIG DATA LANDSCAPE, VERSION 3.0
Exited: Acquisition or IPO
Infrastructure
Analytics
Applications
Open Source
Data Sources
Cross Infrastructure/
© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

5. 我们想让做什么

1.拥有独立的大数据存储,
2.能够接收多源数据,
3.能接收结构化数据和非结构化数据,
4.提供实时或者近实时的分析,
5.能够提供线性的分析和非线性分析
6.预测分析
7.相关性分析
8.?

6. 总结一下

1.业务多维度分析
2.非关系型数据分析
3.数据挖掘,
4.?

7. 我们的数据?

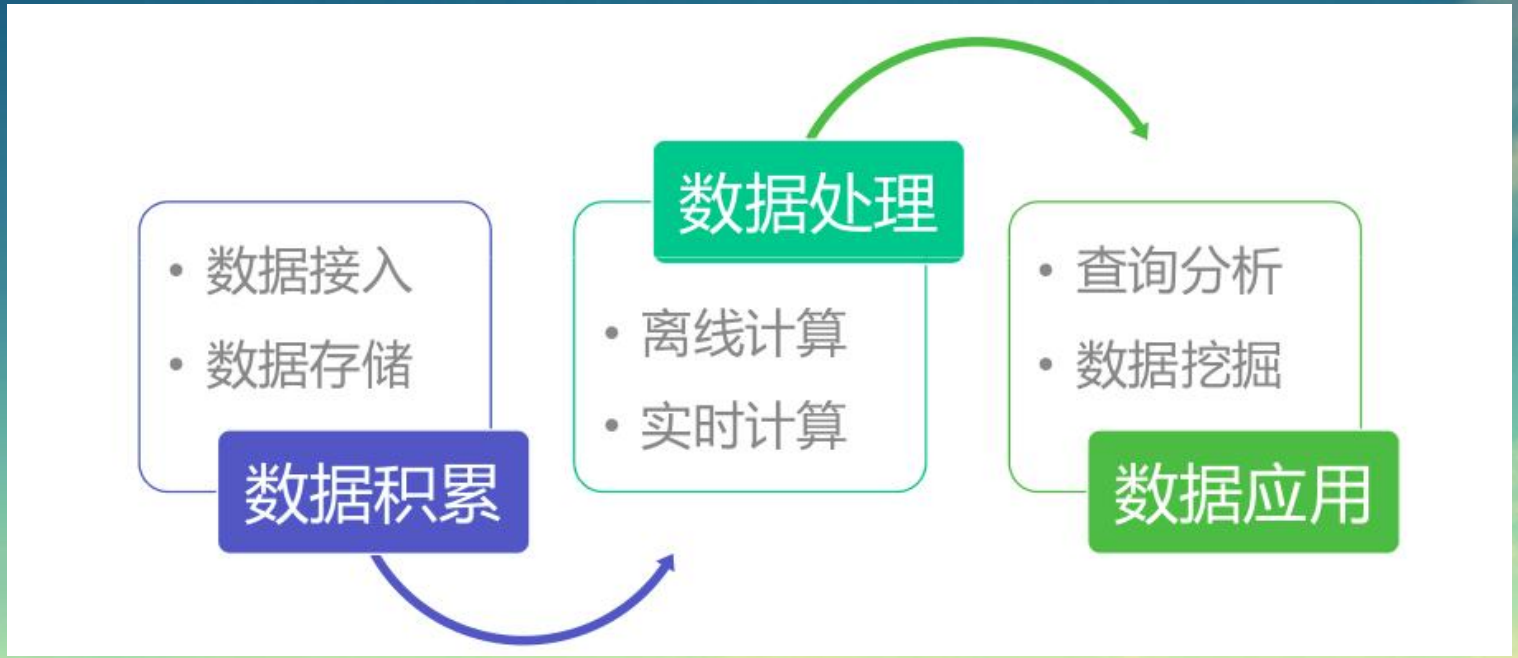
1. 业务数据 mysql
2. 操作日志 nginx ; tomcat
3. 销售政策,
4.?????

生产厂基本数据 ,原材料数据 ,生产过程数据
农业基础数据,
地域 ,农田面积,人口数据 ,农作物,品种,产量.
全国天气数据
粮食,食品,价格 , 物价
国家政策. 国际影响.
农户行为数据.
系统外销售数据.

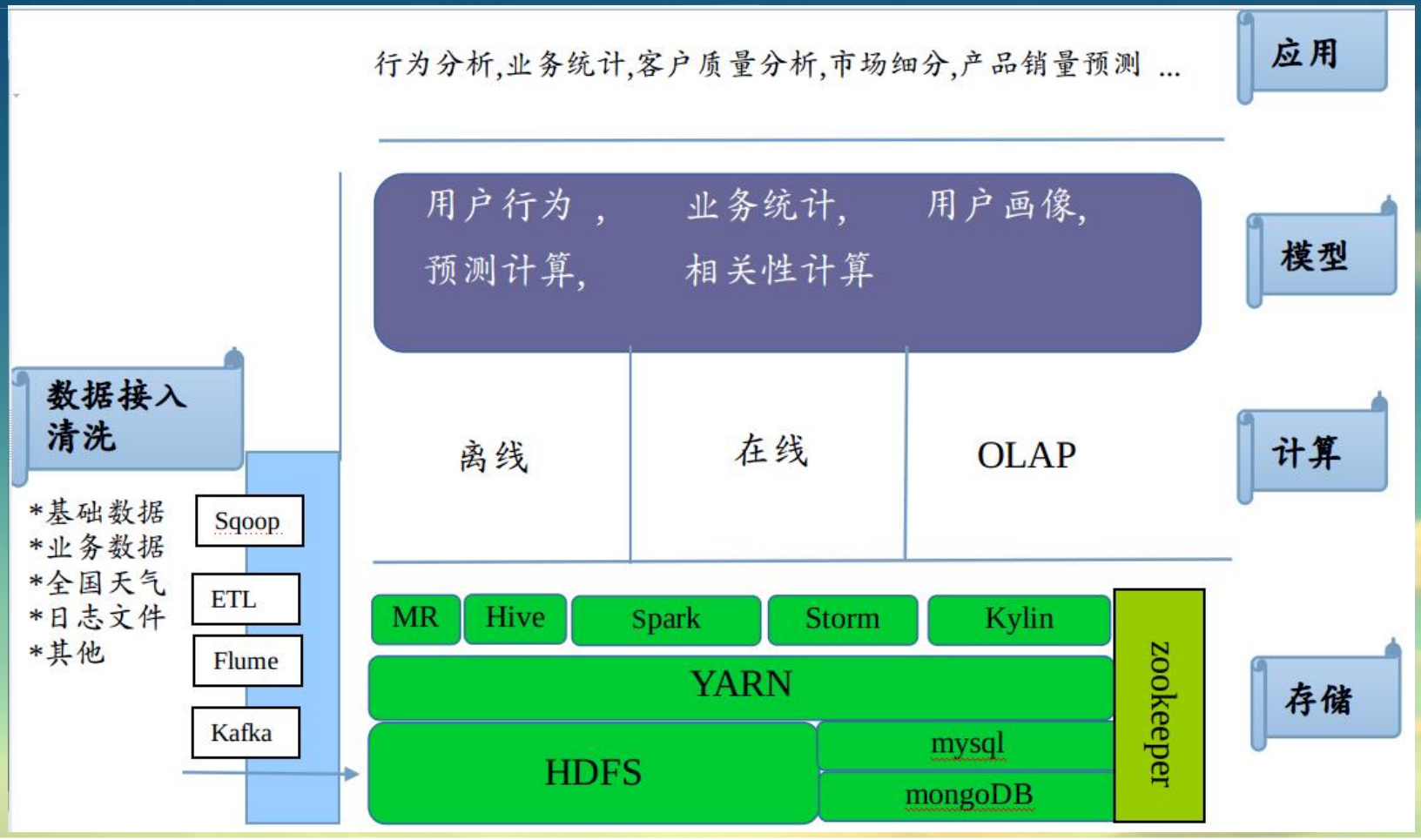
8. 具体做什么

1.多维度分析,
2.预测分析,
3.用户行为分析,
4.挖掘分析,
5.相关性分析
6.人工智能,机器学习,深度学习

9. How? 平台化基本思路



10. 架构图


行为分析,业务统计,客户质量分析,市场细分,产品销量预测 ...
应用
用户行为 , 业务统计, 用户画像,
预测计算, 相关性计算
模型
数据接入
清洗
*基础数据
*业务数据
*全国天气
*日志文件
*其他
Sqoop
ETL
Flume
Kafka
离线
在线
OLAP
计算
MR
Hive
Spark
Storm
Kylin
YARN
HDFS
mysql
mongoDB
zookeeper
存储

11. 处理步骤

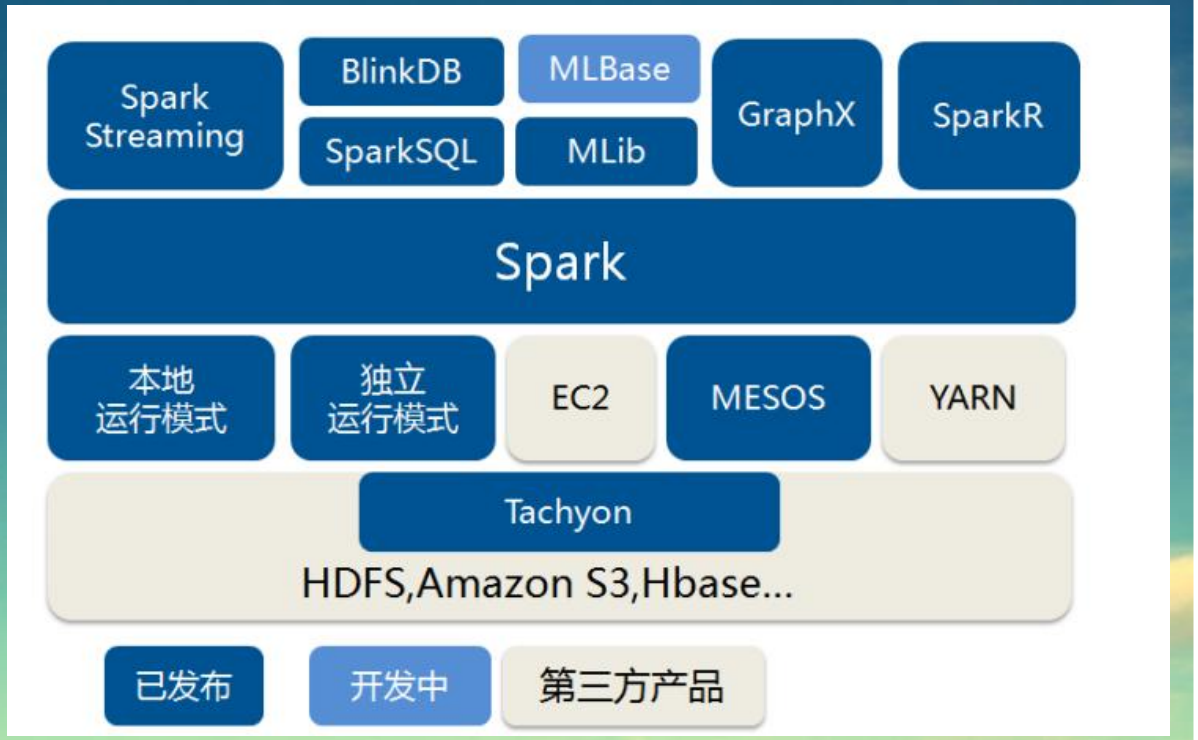
数据源->数据清洗->数据存储服务->数据分析->可视化展示

12. Spark

1.运行速度快

2.通用性强

3.随处运行



12. Storm

编程简单：开发人员只需要关注应用逻辑，而且跟Hadoop类似，
高性能，低延迟：可以应用于实时响应的场景。
分布式：可以轻松应对数据量大，单机搞不定的场景
可扩展： 随着业务发展，数据量和计算量越来越大，系统可水平扩展
容错：单个节点挂了不影响应用
消息不丢失：保证消息处理

使用Storm时你需要关注以下几点：
如果使用的是自己的消息队列，需要加入消息队列做数据的来源和产出的代码
需要考虑如何做故障处理：如何记录消息队列处理的进度，应对Storm重启，挂掉的场景
需要考虑如何做消息的回退：如果某些消息处理一直失败怎么办？

12. Apache kylin

- 可扩展超快OLAP引擎: 为减少在Hadoop上百亿规模数据查询延迟而设计
- Hadoop ANSI SQL 接口: 为Hadoop提供标准SQL支持大部分查询功能
- 交互式查询能力: 用户可以与Hadoop数据进行亚秒级交互
- 多维立方体: 为百亿以上数据集定义数据模型并构建立方体
- 与BI工具无缝整合: Kylin提供与BI工具，如Tableau，的整合能力
- 其他特性:
- Job管理与监控
- 压缩与编码
- 增量更新
- 利用HBase Coprocessor
- 基于HyperLogLog的Dinstinc Count近似算法
- 友好的web界面以管理，监控和使用立方体
- 项目及立方体级别的访问控制安全
- 支持LDAP

12. 时间规划

开发环境 ,
数据接入,ETL
任务调度 ,
日志抓取,分析, （两周,)

spark 数据挖掘(客户画像) ,
销量预测模型,
相关性分析模型, (一月/个)

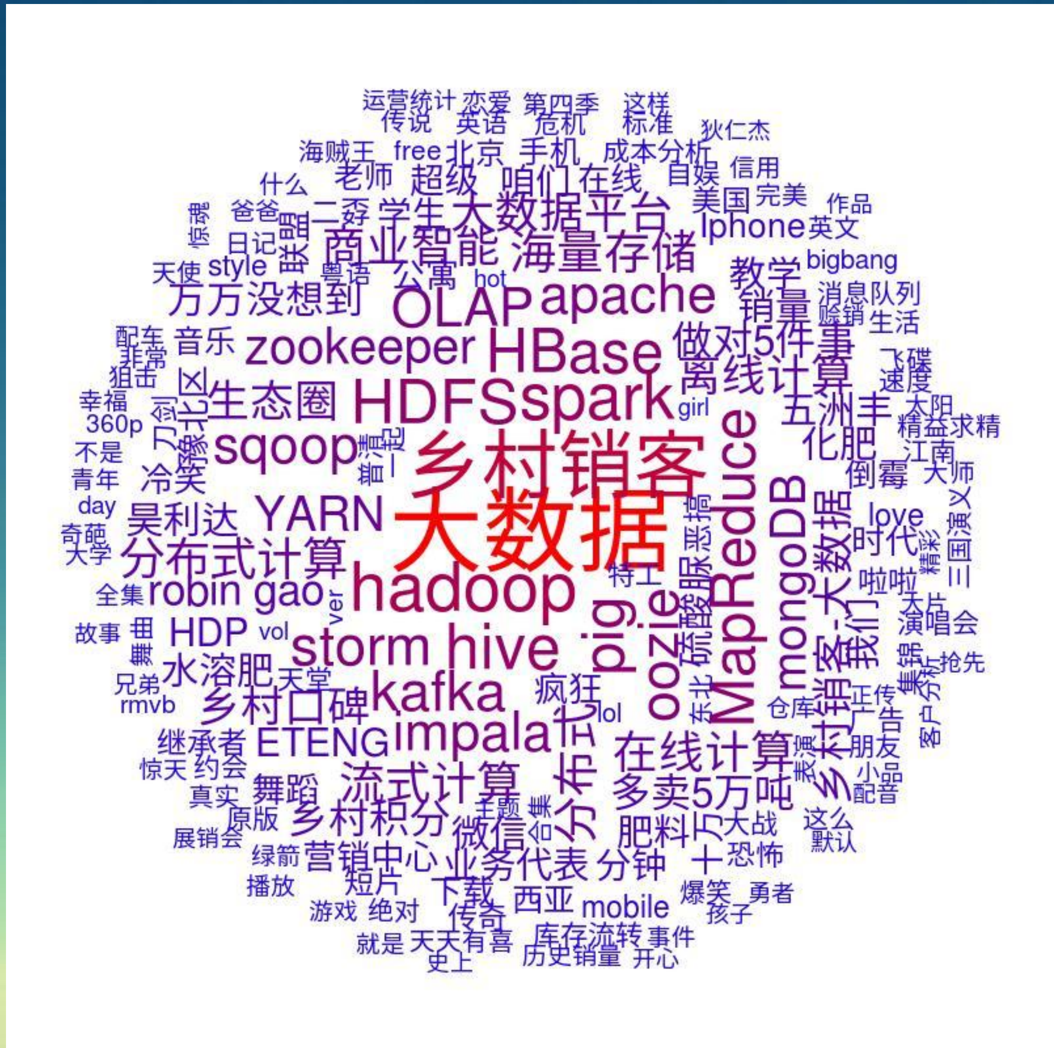
olap分析应用,(3周)

监控系统,

问题&难点

1. 阿里云结合
2. 模型算法
3. 运维

谢谢!



大数据
乡村销客
hadoop
HDFSspark
MapReduce
HBase
apache
OLAP
zookeeper
sqoop
YARN
storm
hive
kafka
impala
pig
oozie
mongoDB
学生大数据平台
商业智能
海量存储
分布式计算
生态圈
万万没想到
做对5件事
离线计算
在线计算
流式计算
乡村口碑
乡村积分
robin gao
ETENG
HDP
营销中心业务代表