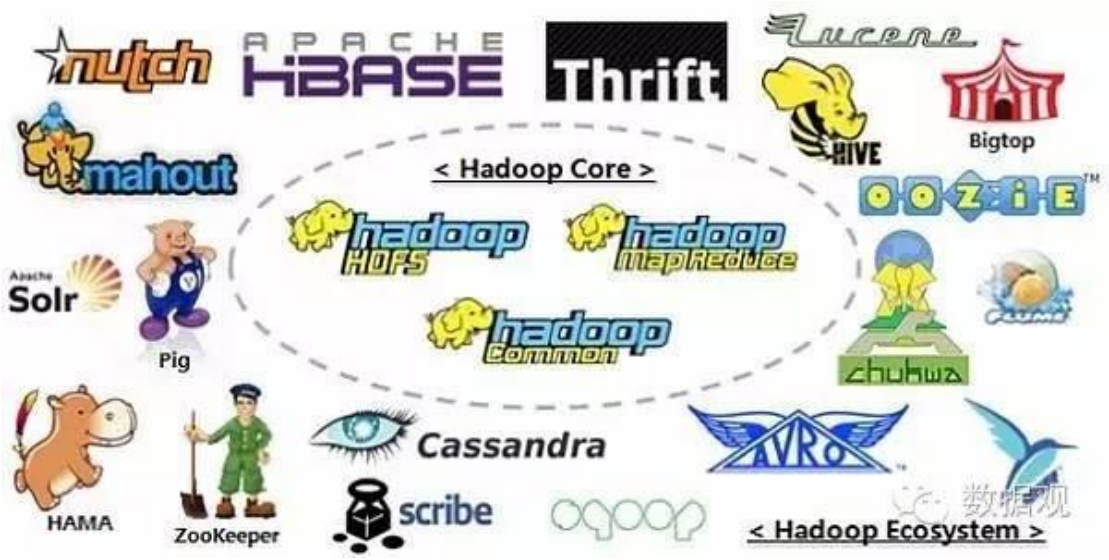
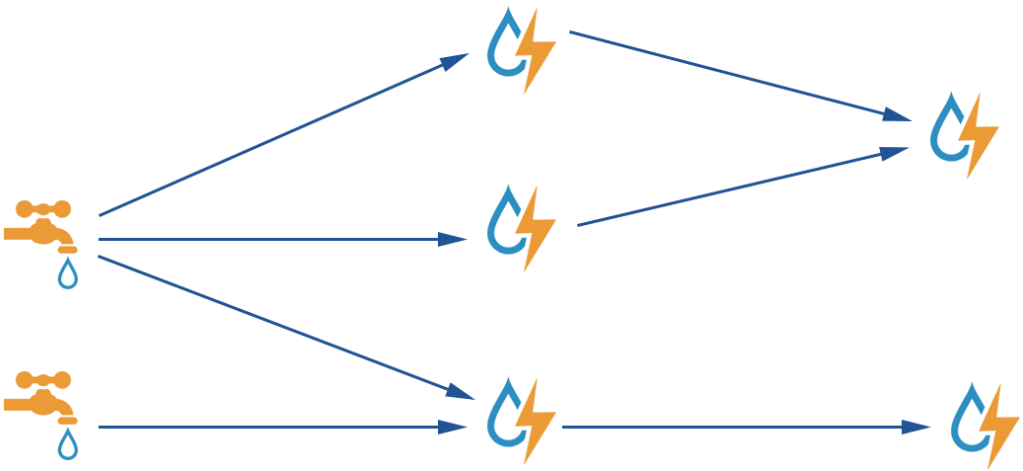
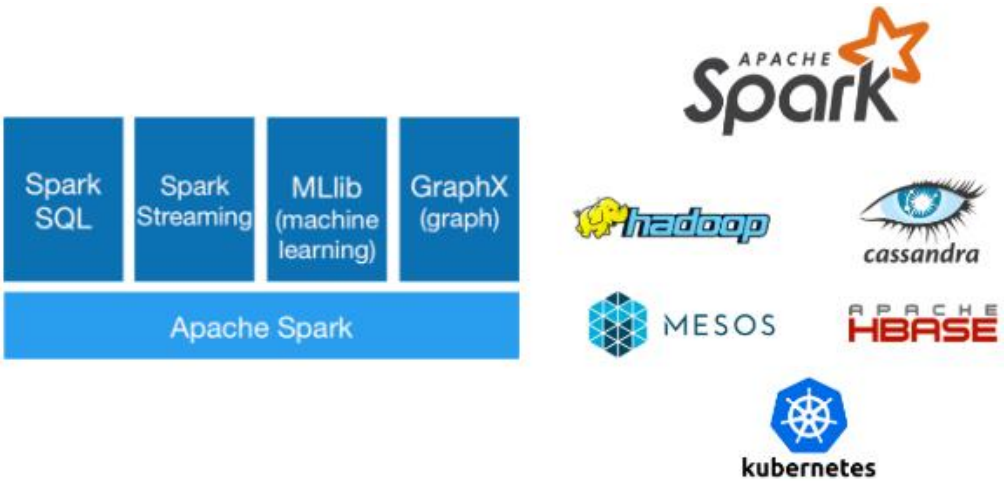
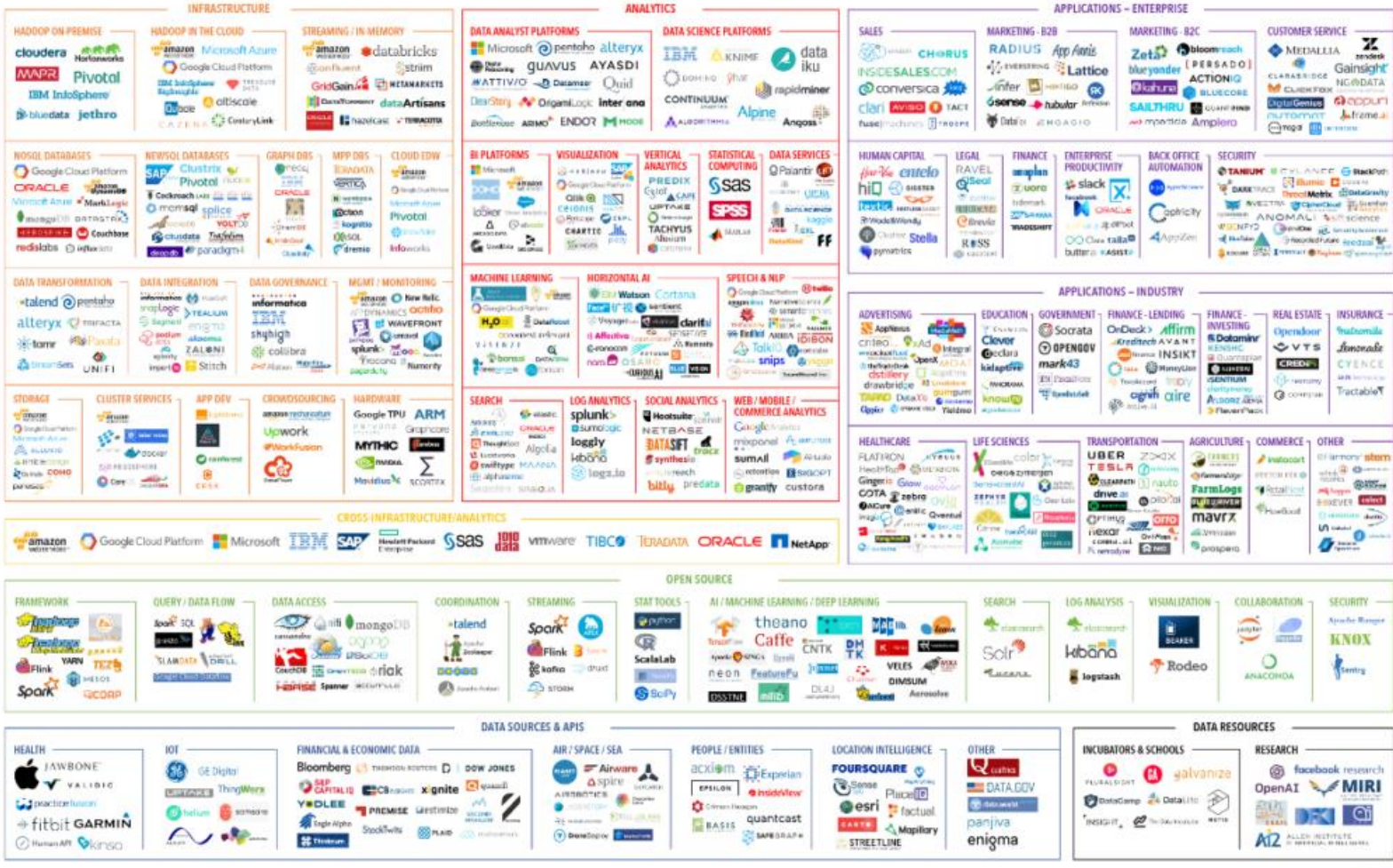
Hadoop 生态圈


nutch
APACHE HBASE
Thrift
Lucene
HIVE
Bigtop
mahout
< Hadoop Core >
OOZIE
hadoop HDFS
hadoop MapReduce
Solr
Pig
hadoop Common
chukwa
HAMA
ZooKeeper
Cassandra
AVRO
scribe
< Hadoop Ecosystem >




APACHE Spark
Spark SQL
Spark Streaming
MLlib (machine learning)
GraphX (graph)
Apache Spark
hadoop
cassandra
MESOS
APACHE HBASE
kubernetes

BIG DATA LANDSCAPE 2017



V2 – Last updated 5/3/2017

 mattturck.com/bigdata2017

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

目录

1.分布式基础

1. 1. 计算机科学与技术

掌握本质。以不变应万变。

- 计算机组成原理
- 计算机操作系统
- 编译原理
- 程序设计语言（C、C++、Java、C#、PHP、JavaScript、Python、Scala）

集合，NIO，多线程，类加载，GC，内存管理

- 计算机网络

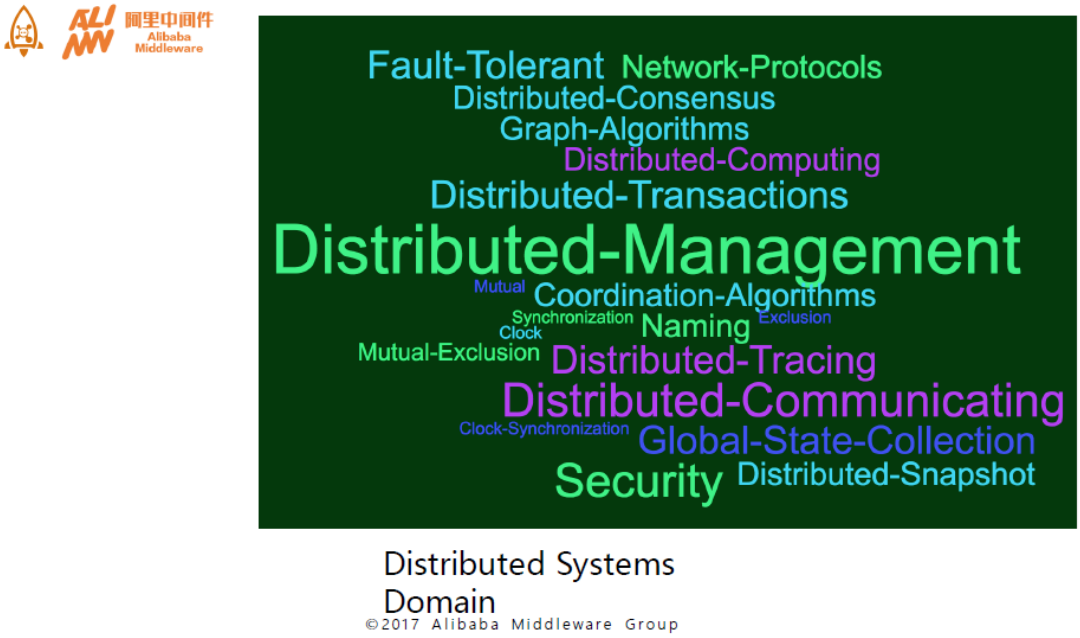
MQTT，HTTP，FTP

- 数据结构与算法
- 数据库原理
- 软件工程

需求，设计，开发，测试，上线，维护

- More：数学（线性代数、概率与统计、离散数学）&英语

1.2. 大数据（分布式系统）



从技术的角度讲，大数据就是指分布式相关的技术。

1.2.1. 4V

Value, Volume, Variety, Velocity 价值、容量、多样、速度

生活中大数据的应用：电商/资讯推荐、反电信欺诈、个人信用画像、共享单车跟踪

1.2.2. 数据

数据源：

按行业，互联网、金融、电信等；

按系统，上网日志、业务应用日志、交易记录等。

数据格式：

文本，二进制。文本有 JSON、XML 系、CSV 等。

数据编码：

Unicode、GBK、UTF-8 等。

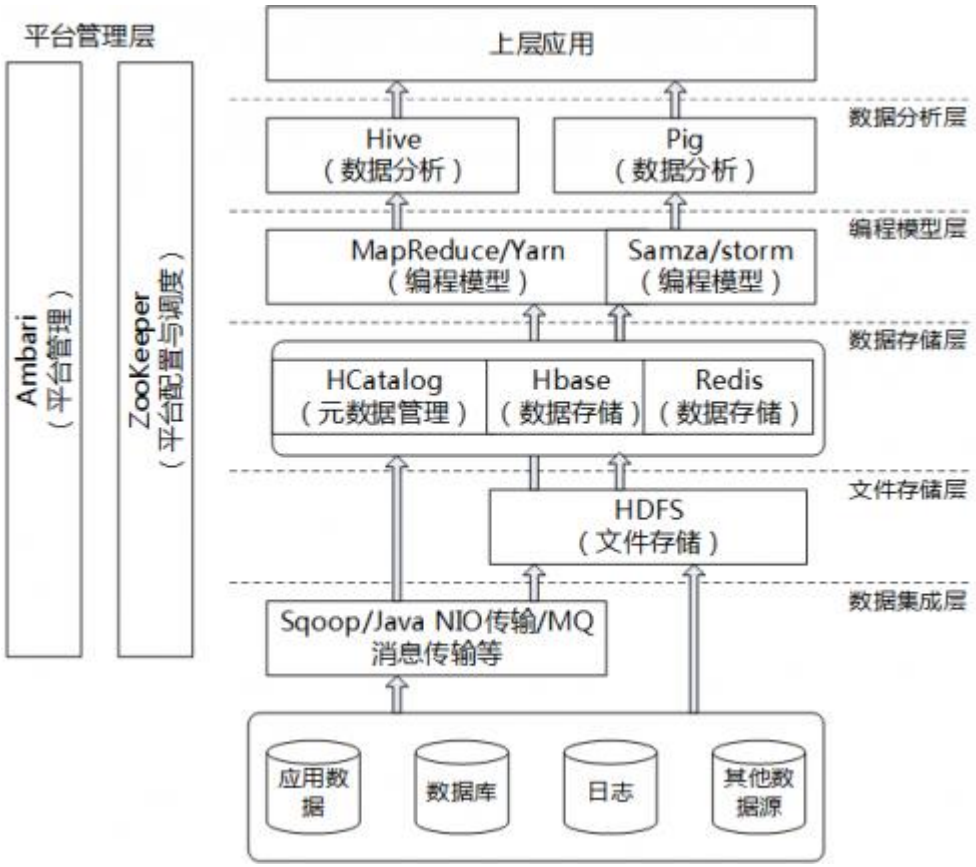
数据压缩：

Zip、Snappy 等。

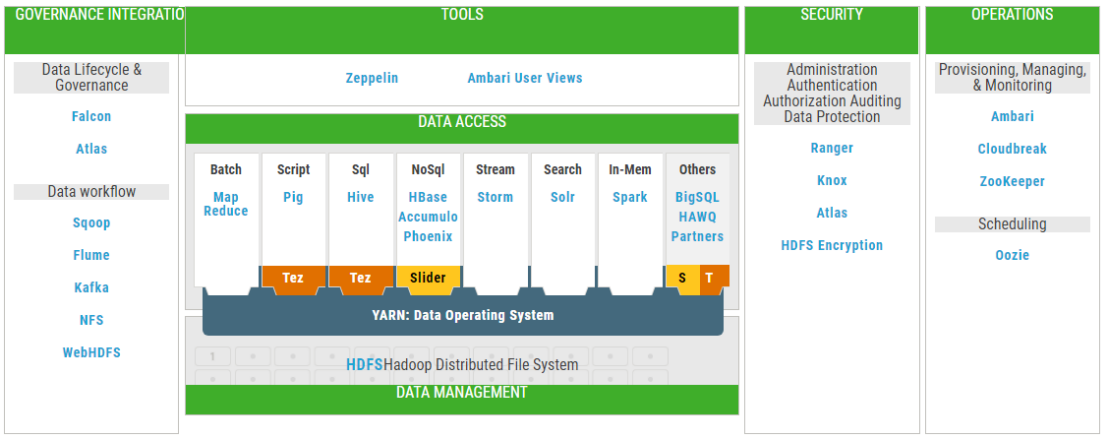
1.2.3. 技术架构

1.2.3.1. 大数据平台

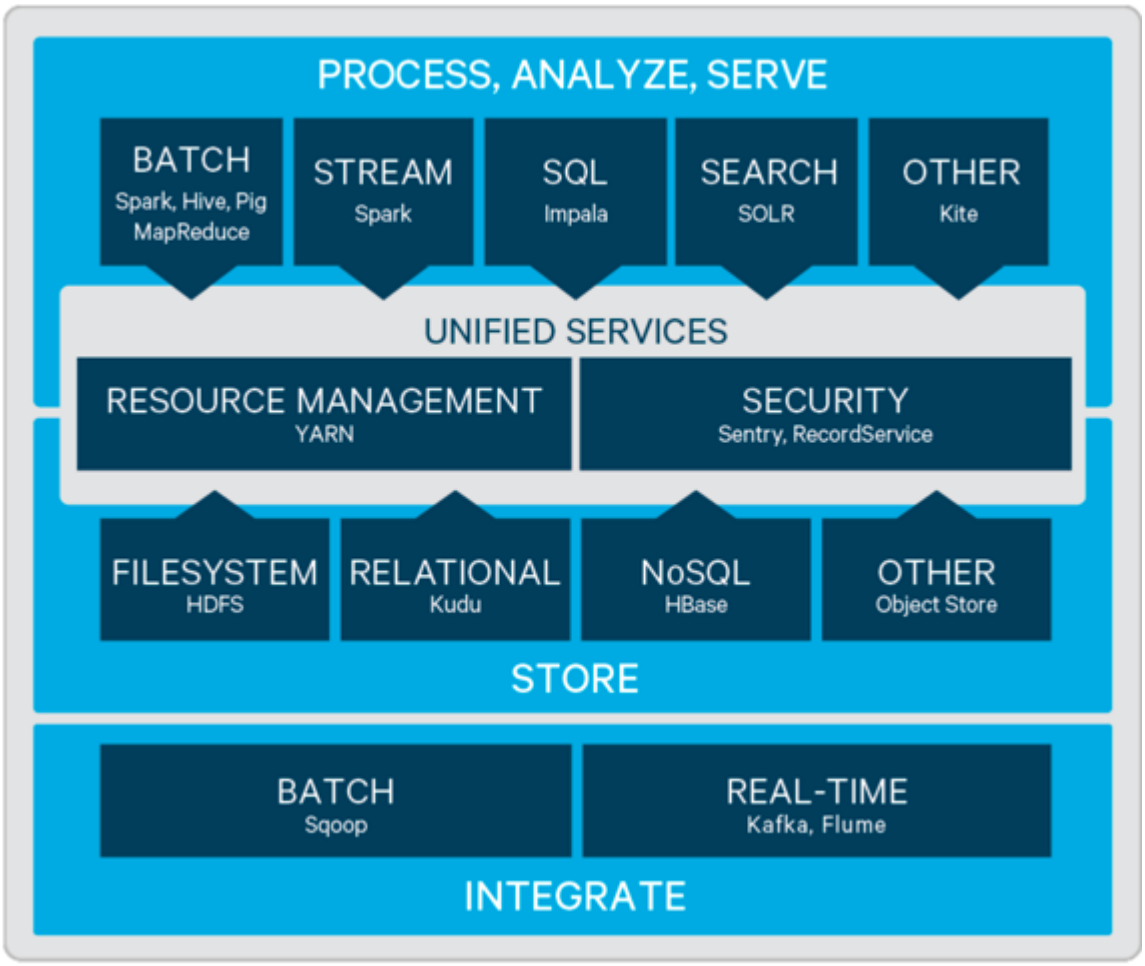
数据采集、数据处理/标准化、数据存储、数据分析/挖掘、数据应用/服务



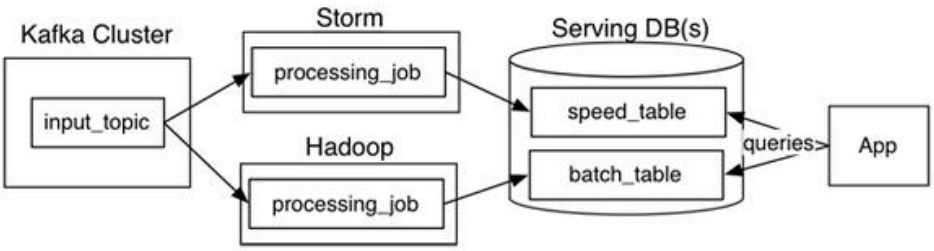
Hortonworks:


GOVERNANCE INTEGRATIO
TOOLS
SECURITY
OPERATIONS
Data Lifecycle & Governance
Falcon
Atlas
Data workflow
Sqoop
Flume
Kafka
NFS
WebHDFS
Zeppelin
Ambari User Views
DATA ACCESS
Batch
Map Reduce
Script
Pig
Sql
Hive
NoSql
HBase
Accumulo
Phoenix
Stream
Storm
Search
Solr
In-Mem
Spark
Others
BigSQL
HAWQ
Partners
Tez
Tez
Slider
S
T
YARN: Data Operating System
HDFSHadoop Distributed File System
DATA MANAGEMENT
Administration
Authentication
Authorization
Auditing
Data Protection
Ranger
Knox
Atlas
HDFS Encryption
Provisioning, Managing, & Monitoring
Ambari
Cloudbreak
ZooKeeper
Scheduling
Oozie

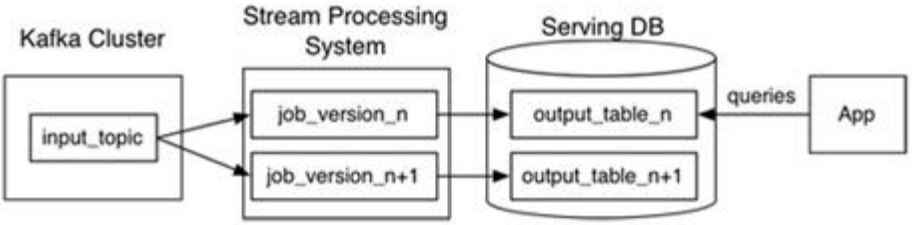
Cloudera:


PROCESS, ANALYZE, SERVE
BATCH
Spark, Hive, Pig
MapReduce
STREAM
Spark
SQL
Impala
SEARCH
SOLR
OTHER
Kite
UNIFIED SERVICES
RESOURCE MANAGEMENT
YARN
SECURITY
Sentry, RecordService
FILESYSTEM
HDFS
RELATIONAL
Kudu
NoSQL
HBase
OTHER
Object Store
STORE
BATCH
Sqoop
REAL-TIME
Kafka, Flume
INTEGRATE

1.2.3.2. Lambda 架构

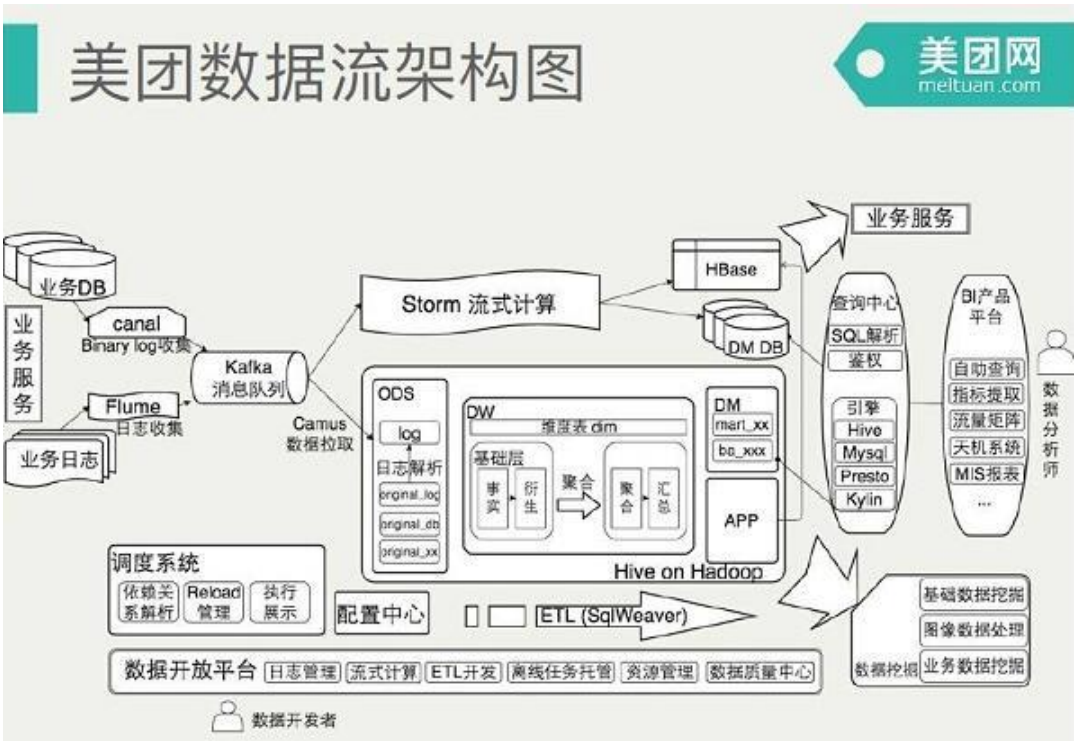


1.2.3.3. Kappa 架构

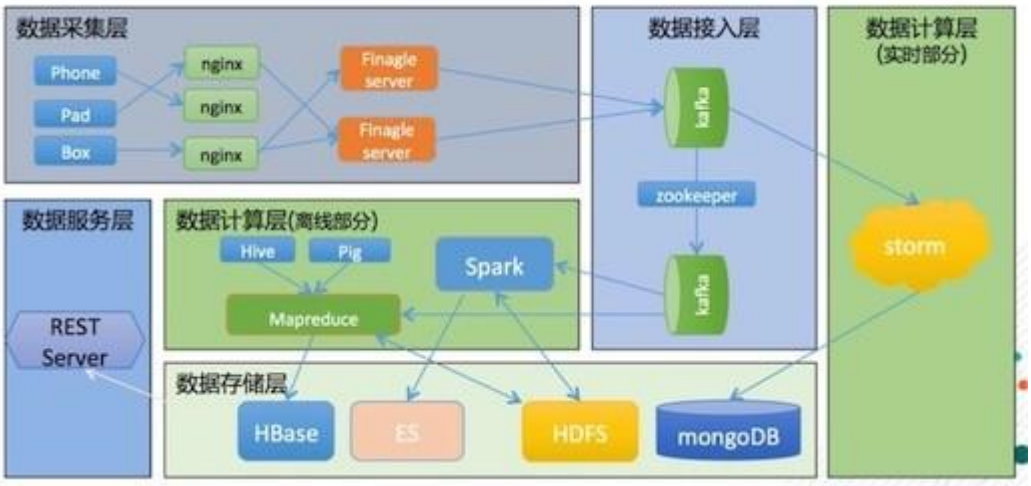


1.2.3.4. 部分互联网公司大数据架构

美团：



友盟：



1.3. 开源&Apache 软件基金会

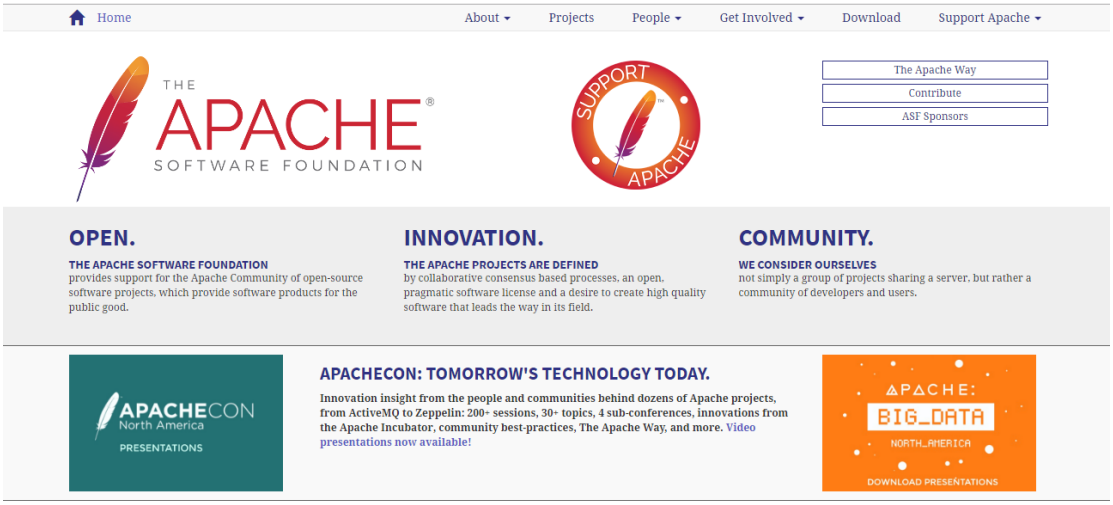
1.3.1. 开源

Apache 基金会（http://www.apache.org/）

Linux 基金会（https://www.linuxfoundation.org/）

GitHub（https://github.com/）

1.3.2. Apache 基金会



http://www.apache.org/index.html#projects-list

User -> Contributor -> Committer -> PMC member -> ASF member

http://www.apache.org/foundation/how-it-works.html#roles

http://www.apache.org/licenses/#2.0

1.3.3. 怎样学习一个 Apache 开源项目？

1.3.3.1. What ? How ? Why ?

知其然，知其所以然。

官网：http://storm.apache.org/

文档：http://storm.apache.org/releases/current/index.html

Quikstart/Example :

https://github.com/apache/storm/tree/master/examples/storm-starter

代码：https://github.com/apache/storm

Issue：https://issues.apache.org/jira/browse/STORM/

Wiki：https://cwiki.apache.org/confluence/display/STORM/Storm+Home

Mailing list：user@storm.apache.org　　dev@storm.apache.org

1.3.3.2. 出现问题？

自己解决（官网 FAQ、邮件、Issue、Google） --（超过 1 天）--> 寻求帮助（邮件）

1.4. 分布式系统原理

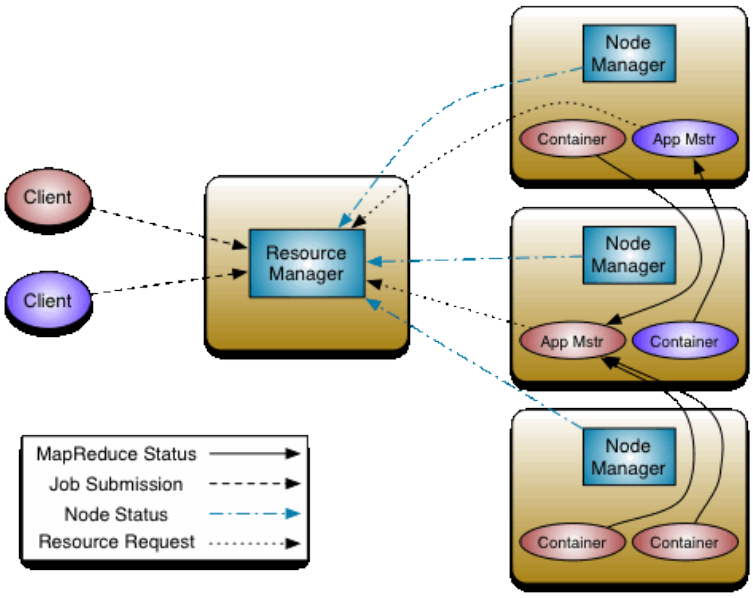
1.4.1. 核心思想：分而治之

分什么？ 资源、任务/数据

一台计算机，从单线程到多线程；一个集群，从单节点到多节点。

通过 Mater-Slave 模式对集群进行管理。Master 负责集群的资源管理与任务管理。

Hadoop YARN 架构：



1.4.2. 数据分区

常用分区算法：均匀（随机、轮询），hash，一致性 Hash

Hash 分区：hash(key) % n

1.4.3. 容错

分布式计算：重新调度+消息重发

分布式存储：多副本

容错：检查点，心跳，租约 Lease

1.4.4. 缩容扩容

分布式计算：缩容重新调度，扩容对原有任务无影响

分布式存储：缩容无影响，扩容对原有数据无影响