

大数据环境用的最多的消息中间件就是kafka了，kafka的特点就是吞吐量很大，这种消息队列本身就是Hadoop生态系统中的一员，kafka这种消息中间件是基于zookeeper的，主要利用的就是zookeeper的一致配置等一致性特性，也就是说,kafka是一个分布式的集群，这种集群本身就自带着Zookeeper高可用的光环，在面对海量数据的业务压力时，能够从容应对。

很多场合，都是flume+kafka这种架构模式，由flume进行数据采集，kafka承担消费者。

在离线处理这部分，就是使用经久不衰的Hadoop解决方案，HDFS+YARN+spark这种，然后，就可以对数据进行一个离线的数据挖掘。

在流式数据处理中，斗鱼的技术栈将Storm和Spark streaming都用上了，可能也是跟自己的业务场景比较复杂有关，对于一般的公司，业务场景也比较简单，Storm还是Spark streaming很容易就可以选择出来了，其实，在很多情境下，大部分的中小型公司这两个方案差别不是很大。

**一. 数据存储**

数据是个很泛的概念，但是我们脑海里第一反应的就是关系型数据库和EXCEL这种二维表是数据。

而现在数据各种各样特色，有文档、有图片、有流式的音频视频、有日志数据、有IM消息数据、有索引数据、有社交应用的网状关系数据、有地图数据。他们对存储访问都有不同的要求，因而NoSQL兴起了。

KV型：先后出现了Memcached、Redis。

文档型：出现了CouchDB、MongoDB。

日志数据：出现了Facebook Scribe、Flume、Logstash。

即使在传统关系型数据领域，由于大数据规模也出现了真正的分布式关系型数据库，如GreenPlum、TiDB、OceanBase

为了多维分析，也出现了专门的列式数据库，如HBase。

**二. 数据仓库**

大数据仓库建设，必要的工作就是ETL（抽取、转换、导入）。抽取，这步就又细分为：数据抽取、数据清洗、数据校验。在转换这步，我们也需要关注数据安全脱敏，也就是说，进入大数据仓库的数据需要分级。

数据复制分发，我们可以使用消息队列和调度服务来工作。消息队列如Kafka、ZeroMQ、ActiveMQ、RabbitMQ。调度服务如ZooKeeper。

数据抽取传输开源项目，我能看到的Sqoop on Hadoop。可能很多数据都是直接被Spark、Storm、Presto、Hbase处理了。

**三. 大数据仓库基础架构**

我们第一个就要提出的就是Hadoop。目前它已经成为了一个生态。Hadoop最核心是两块：分布式文件系统HDFS、MapReduce。

在MapReduce中资源管理和作业管理均是由JobTracker实现的，集两个功能于一身，而在新的Hadoop2.0架构中，MapReduce作业管理由ApplicationMaster实现，而资源管理由新增系统YARN完成。

Hadoop Corona是facebook开源的下一代MapReduce框架。其基本设计动机和Apache的YARN一致。就是因为大家都在解决同一个问题，只不过不同时间段开源出来不同的解决方案项目，所以使大家大呼到底要用哪个啊。

Mesos是更宽泛的资源管理框架，不仅可以管理偏重于半结构化的大数据框架，也可以管理非结构化的框架。人们也在对比Mesos和Kubernetes，其实Kubernetes只是目前聚焦于容器的管理，而Mesos野心更大，当然也抽象的更通用化，留下各种抽象接口，以管理各种资源。

YARN成为了大数据处理资源管理框架，而Mesos更成为了有史以来更大集成者更通用的资源管理框架了。

Zookeeper已经成为最基层最具体最不可或缺的真正的协调执行者了。

**四. 大数据计算框架**

Spark、Storm、Flink。

Spark以前的领域在MapReduce工作的领域，不过Spark大量把MAP中间结果放到内存中，所以显得性能特别快。现在Spark也在往生态走，希望能够上下游通吃，一套技术栈解决大家多种需求，所以大家又渐渐看不清楚Spark聚焦的领域了。Spark Shark，是为了VS hadoop Hive，Spark Streaming是为了VS Storm。

Storm擅长处理实时流式。比如日志，比如网站购物的点击流，是源源不断、按顺序的、没有终结的，所以通过Kafka等消息队列来了数据后，Storm就一边开始工作。Storm自己不收集数据也不存储数据，随来随处理随输出结果。

它们都有各自的特色，但他们都无法给你处理具体的业务应用，它们只是大规模分布式计算的通用框架，所以具体用他们的特性干什么，还得你自己写。这就是大数据业务领域分析系统的事了。

不过，下面我会介绍一些技术，在特定的业务领域，能够帮你更简单的达成业务应用需要。

**五. 大数据应用技术**

1、大数据搜索：Lucene、Solr、ElasticSearch。ElasticSearch是新推出的比Solr在大规模数据情况下更好的开源解决方案。

2、大数据查询：这里有Hive/Impala，Hive的作用是你可以把结构化数据导入到Hadoop中然后用简单SQL来做查询。你可以把Impala看做是性能更快的Hive，因为Impala不强依赖MapReduce。而Facebook开源的Presto更是能查询多种数据源，而且一条Presto查询可以将多个数据源的数据进行合并。

3、大数据分析：咱们要提到去年新晋顶级Apache项目的Kylin。它创始于ebay，2014年进入apache孵化项目。Kylin不仅仅能做SQL查询，而且能做Cube多维分析。

4、大数据挖掘：这个领域包含精准推荐、机器学习/深度学习/神经网络、人工智能。自从AlphaGo火了以后，机器学习再度火热。Google开源了最新机器学习系统TensorFlow，微软亚洲研究院开源了分布式机器学习工具包-DMTK，雅虎也开源了Caffe On Spark 深度学习。Mahout是Apache的一个开源项目，提供一些机器学习领域经典算法的实现，包括聚类、分类、推荐过滤、频繁子项挖掘。

**六. 数据可视化工具**

这块有大量的可视化开源组件，但成系统的开源的确实出色的不多。这里面大有可玩。

Airbnb近日开源了数据探查与可视化平台Caravel。另外，百度推出的Echarts组件也不做，可以基于此做些系统性的工作。

**七. 数据平台监控运维**

这里只看到一套完整的适用于海量日志处理的工具：Facebook Scribe、Flume、Logstash、Kibana。但我觉得，如此复杂的大数据平台技术栈，一套成熟的监控运维系统，是必须要出现的。

**八. 最后总结：大数据平台建设总蓝图**

作为咱们日常业务，最直接的需要就是一套能分析具体业务的应用系统，但这套应用系统需要很多基础技术架构和服务才能达成。为了让大家有所预期，所以我把这些必备前提储备列出来。

技术平台建设：

1、大数据基础架构：推荐Hadoop、HDFS、YARN；大数据计算框架：推荐Spark；大数据日志收集推荐Flume+Logstash+Kibana。我们需要部署依赖中间件Zookeeper。

2、大数据仓库平台建设：分布式关系型TiDB、KV式Redis、文档型MongoDB、列式Hbase

3、大数据搜索，推荐选择Lucene、ElasticSearch；大数据查询，推荐选择Presto；大数据多维分析，推荐Kylin；大数据挖掘，推荐挖掘开源算法包MashOut。

大数据整理服务：

1、主数据管理：主数据标准制定、主数据清洗与校验、主数据转换（拆分合并）、主数据复制分发、主数据访问OpenAPI。

2、ETL：数据抽取、数据清洗、数据校验、数据安全脱敏

大数据分析系统建设：

1、大数据展示平台建设

2、大数据商业应用模型建模

3、大数据应用分析系统设计与开发

大数据分析服务：

1、大数据分析解读报告