聚贸大数据技术架构

根据目前公司的业务情况和已有技术选型，以及昨天会议讨论意见，以及今天下午与架构组交流。聚贸大数据技术架构总结如下：

1. 数据源， 分为内部数据（日志， 数据库），外部数据（互联网，第三方）
2. 采集层， 系统日志收集Flume + Kafka， 数据库ETL工具Kettle， 互联网数据用爬虫来抓取， 第三方数据通过合作购买获得，相应对接技术如Webservice，Mina socket，FTP。

备注：Flume + Kafka+ Elastic Search这是目前公司架构组技术选型，大数据平台延续这一思路可以无缝的和聚贸各业务平台对接，获取相关日志信息

1. 计算层， 分为实时业务和批处理离线业务， 实时数据根据业务响应时间要求采用Storm， Spark Stream， 批处理业务采用MR， Hive，Pig进行数据清洗，转换
2. 数据层，选用HBase， Redis， Cassandra， Gemfire
3. 服务层， 将相关的业务逻辑封装成服务。采用Webservice Dubbo/DRPC
4. 应用层， web数据展示，报表统计等统一用公司内部框架Dove
5. 分布式一致性Zookeeper
6. 平台监控 Ambari，Ganglia，Nagios。 Ambari负责大数据集群， Ganglia，Nagios负责服务器预警监控

平台监控

Nagios

应用层

Web应用层 Dove框架

分布式

一致性

Zookeeper

服务层

数据服务层 DubboX/

数据持久层 HBase

数据层

MR 离线业务分析处理

Storm实时业务分析处理

计算层

接口

Flume+Kafka

爬虫

ETL（Kettle）

采集层

内部数据

外部数据

第三方

**互联网**

日志

DB

B

数据源

备注：以上高亮模块为目前技术团队已经用到的技术栈，关于监控参考博客 <http://www.cnblogs.com/scotoma/archive/2013/05/18/3085248.html>