1. DSP概述

DSP全称data service platform，是整合多种数据库以及多种系统框架可适应不同场景存储、计算和查询的数据服务平台。

1. DSP架构

DSP系统组成：collector kafka storm redis elasticsearch zookeeper hadoop hdfs hbase mongo等

* Collector：接收数据，数据入口
* Kafka：接收数据缓存
* Storm：实时统计框架
* Topology：实时统计脚本
* Hadoop：离线统计框架
* Zookeeper：配置维护，服务器间通讯
* Job：离线统计脚本
* Redis：存储collector接受的meta信息，以及做storm统计的缓存数据。其中redisDump用于统计数据备份，redisclean用于过期数据删除
* Hbase：存储数据统计结果，用于存储数据量大，查询速度无要求的数据结果，供用户查询
* Mongo：存储数据统计结果，用于存储数据量不大，需要快速查询或者复杂查询的数据结果，供用户查询
* ES：存储数据统计结果，用于存储实时性高、需要快速索引的数据结果，供用户查询
* HDFS：用于存储数据文件以及统计结果文件
* Portal：用于信息配置和展示
* Monitor：监控集群应用是否正常服务，异常发出告警
* Spark：准实时计算，介于Hadoop和storm之间，小型批处理，效率比storm高一些，处理数据量比Hadoop低一下。当前系统中用得较少
  1. **整体框架：**

Collector接收数据后，数据发送到kafka缓存。Camus从kafka获取数据并存储。Storm框架进行实时统计分析，从kafka获取数据。Hadoop框架进行离线统计分析，从Camus获取数据。



* 1. **数据接收存储流程：**

Collector维护type，接受clienttype数据分割后发送给kafka的topic。Camus从对应topic获取数据通过HDFS保存数据文件。



* 1. **数据统计流程：**

离线数据统计：

Hadoop框架对离线统计脚本（job）进行调度管理，从HDFS获取数据进行离线计算，统计结果数据可通过HDFS保存文件或者保存到HBASE供用户下载查询。



实时数据统计：

Storm框架对实时统计任务（topology）进去调度管理，从kafka获取数据进行实时计算，为了提高运算效率，将绑定一组redis缓存统计数据等，统计结果数据可通过HDFS保存文件或者保存到HBASE供用户下载查询。



1. DSP平台能力
   1. **接收存储数据**
      1. 数据接口：

collector提供API接口，通常可使用agent推送

* + 1. 接收模块：collector、kafka

Collector被动接收数据，发往kafka缓存，保证数据快速接收数据并且不丢失

* + 1. 存储模块：hdfs、hbase、mongo、elasticsearch

多个存储模块可灵活切换适用多种存储场景

* Hdfs冷门资源以文件方式存储，存储量大，用于无需查询的数据
* hbase存储量大，稳定性高，但是查询慢，不支持复杂查询。用于查询不频繁的场景
* mongo存储量小，支持比较大量的、复杂的查询。用于高频查询场景
* Elasticsearch可提供稳定、可靠、快速的搜索，用于数据量极大或者需要实时查询的场景
  1. **统计**
     1. 实时统计
* 支持对当前数据进行即时计算
* 对某些指标进行监控
* 对某些指标的近期数据计算汇总展示
  + 1. 离线统计

支持对历史数据或者及其庞大的数据量进行非即时的计算

* 1. **查询**

基于不同的存储方式，dsp支持多种查询方式

* 支持以文件方式下载
* 支持大量数据低频查询
* 支持高频查询
* 支持实时查询
  1. **平台可靠性**
* 所有服务均有负载均衡，服务异常自动切换（Collector通过lvs负载均衡，其余服务通过ha进行负载均衡）
* Monitor监控集群应用是否正常服务，异常时能够根据问题优先级发出电话、短信、邮件通知。
* 三副本存储
* 重要信息主备
* 统计中间数据备份

1. DSP当前任务（数据量、统计维度）
   1. **存储数据**

DSP当前已有5PB的存储量，平均日增2.1TB

* 1. **实时任务**

接入产品：

* XPOD（云宽带、WAP、KBA、CATM等）
* 云计算（IAAS、云存储、云监控）
* 流量经营平台
* 电视云

当前实时任务：

* 监控：数据统计，监控异常推送OMS进行告警
* 查询：数据统计，统计结果供其他系统（例如OPS）查询
  1. **离线任务**

Camus任务：数据转存任务

业务任务：当前已执行离线任务77w+，对32种产品进行相关数据统计分析，任务通过多种方式下发，可接口下发、定时调度、直接下发等。