**一、背景**

现在随着互联网的发展，好多公司的主营业务在线上(app/网站)（外卖公司、各类 app，比如：下厨房；头条；安居客；内涵段子；斗鱼；每日优鲜>、页游、唯品会、拼多多、当当、苏宁 易购、国美在线、网易严选、携程、中华会计网校、多易教育、小象学院、慕课网、尚德教育、有钱花、上哪借 ……）

这类公司，都需要针对用户的线上访问行为、消费行为、业务操作行为进行统计分析，数据挖掘！以支撑公司的业务运营、精准画像营销、个性化推荐等，来提高业务转化率，改善公司运营效果！ 这些需求，都需要通过构建一个**综合数据处理系统**来支撑！

**二、名词解释**

ODS层：存放原始数据，它的作用就是**备份**。在我们这个第一版的系统里就是直接存放日志数据。

DWD层：对 ODS 层的数据进行数据清洗，同时对清洗后的数据生成**结构**与**粒度**相同的明细表。粒度举例：如果 ODS 层每条数据代表一小时的数据，那么 DWD 的每条数据就是一小时数据，如果 ODS 是一天的数据，DWD 就是一天的数据。 **结构不太理解**

DWS层：以 DWD 为基础，进行**轻度聚合**的表。举例：如果 DWD 的每条数据代表一小时的数据，那么 DWS 就可以是按照天来聚合的数据。

ADS层：以 DWS 或 DWT 为基础，为各种统计报表提供数据。统计报表就是要进行展示的表。

任务调度：任务管理就是指在具有依赖关系的多个任务里，先做哪个任务，后做哪个任务，当某个任务失败时怎么实现自动重启。

元数据管理：元数据管理就是对 hive 表中的数据进行血缘分析。作用是在表多的时候，如果某一个表出问题了，可以通过可视化的形式快速找出该表时怎么来的，从而解决问题，如果在表少的情况下则用处不太大。

**三、设计目标**

**3.1 实现功能**

实现的功能在需求文档里已经进行了详细说明，这里不再赘述。

**3.2 性能指标**

Ⅰ.数据精确度

严格验证每个数据的录入保证数据录入的无误，显示用户数据要和原始数据一样无差别。

Ⅱ.时间特性

在网络通常的情况下用户点击网页后5s内必须出结果

Ⅲ.适应性

网页支持主流浏览器的正常加载显示，控制可视化图表和报表显示正常。

**四、系统环境**

**4.1 相关软件和硬件**

软件：Hadoop、Spark、Flume、Kafka、Sqoop、Azkaban、Atlas、Echarts、SpringBoot

硬件：每台节点内存 8G、CPU 6 核、磁盘 300G，集群中共三台节点

**4.2 数据规模预估**

待定

**五、系统设计**

**5.1 基础介绍**

数据采集传输：Flume、Kafka、Sqoop

数据存储：Mysql、HDFS

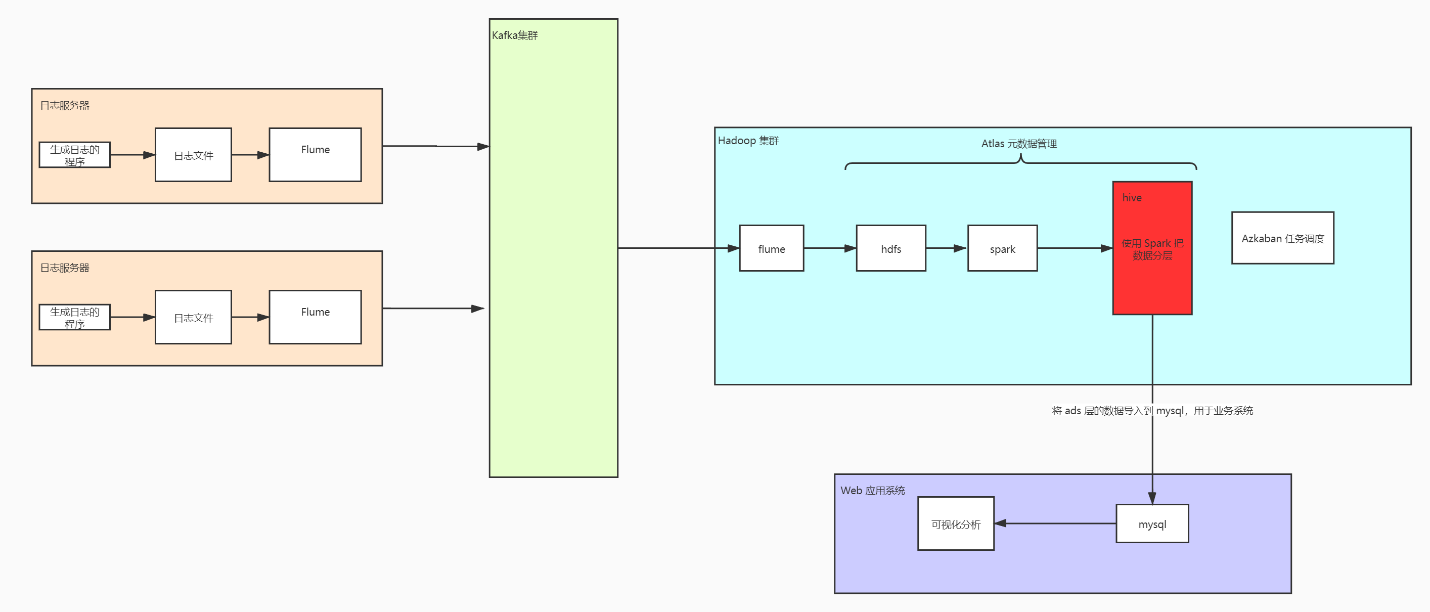
数据计算：Spark

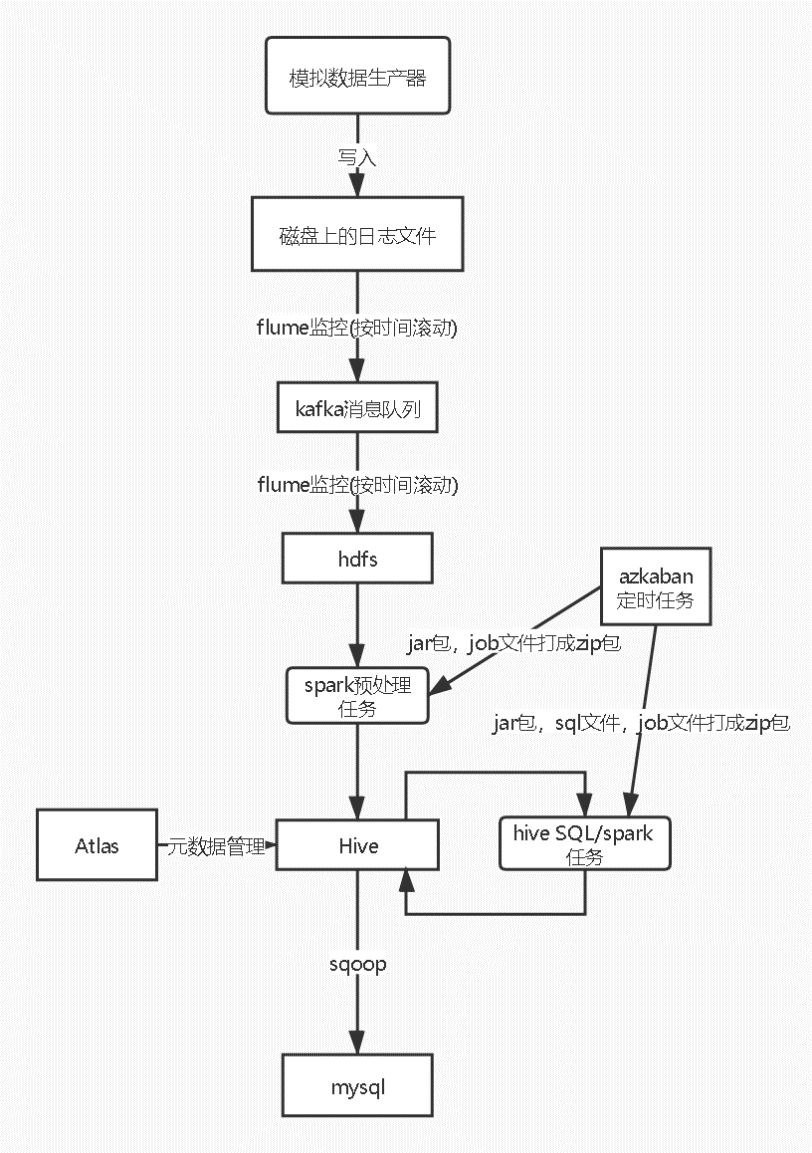
数据可视化：Echarts

任务调度：Azkaban

元数据管理：Atlas

**5.2 系统架构图**

**5.3 系统流程图**

****

**六、风险评估**

**6.1 已知的或可预知的风险**

因为这是第一次系统地开发大数据的项目，所以可能会遇到各种意料之外的问题，比如集群不稳定，web 应用承受不了高访问量。

**6.2 与其他系统可能的影响**

本系统为闭环系统，不涉及与其他系统的交互，所以这方面目前不存在问题。

**七、附件及参考资料**