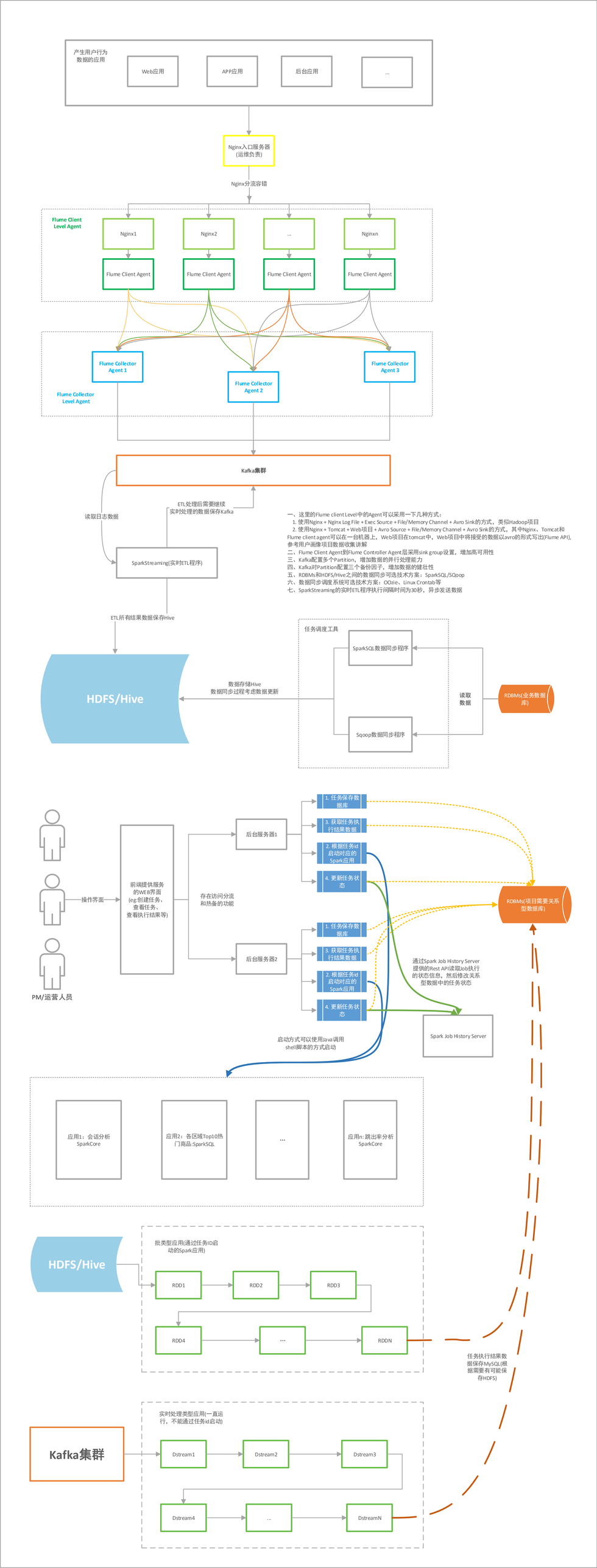
产生用户行为数据的应用
Web应用
APP应用
后台应用
...
Nginx入口服务器(运维负责)
Nginx分流容错
Flume Client Level Agent
Nginx1
Nginx2
...
Nginxn
Flume Client Agent
Flume Client Agent
Flume Client Agent
Flume Client Agent
Flume Collector Agent 1
Flume Collector Agent 2
Flume Collector Agent 3
Flume Collector Level Agent
Kafka集群
读取日志数据
ETL处理后需要继续实时处理的数据保存Kafka
SparkStreaming(实时ETL程序)
ETL所有结果数据保存Hive
一、这里的Flume client Level中的Agent可以采用一下几种方式：
1. 使用Nginx + Nginx Log File + Exec Source + File/Memory Channel + Avro Sink的方式，类似Hadoop项目
2. 使用Nginx + Tomcat + Web项目 + Avro Source + File/Memory Channel + Avro Sink的方式，其中Nginx、Tomcat和Flume client agent可以在一台机器上，Web项目在tomcat中，Web项目中将接受的数据以avro的形式写出(Flume API)，参考用户画像项目数据收集讲解
二、Flume Client Agent到Flume Controller Agent层采用sink group设置，增加高可用性
三、Kafka配置多个Partition，增加数据的并行处理能力
四、Kafka对Partition配置三个备份因子，增加数据的健壮性
五、RDBMs和HDFS/Hive之间的数据同步可选技术方案：SparkSQL/SQoop
六、数据同步调度系统可选技术方案：OOzie、Linux Crontab等
七、SparkStreaming的实时ETL程序执行间隔时间为30秒，异步发送数据
任务调度工具
SparkSQL数据同步程序
HDFS/Hive
数据存储Hive
数据同步过程考虑数据更新
读取数据
RDBMs(业务数据库)
Sqoop数据同步程序
PM/运营人员
操作界面
前端提供服务的WEB界面(eg:创建任务、查看任务、查看执行结果等)
存在访问分流和热备的功能
后台服务器1
后台服务器2
1. 任务保存数据库
3. 获取任务执行结果数据
2. 根据任务id启动对应的Spark应用
4. 更新任务状态
RDBMs(项目需要关系型数据库)
通过Spark Job History Server提供的Rest API读取Job执行的状态信息，然后修改关系型数据中的任务状态
Spark Job History Server
启动方式可以使用Java调用shell脚本的方式启动
应用1：会话分析SparkCore
应用2：各区域Top10热门商品:SparkSQL
...
应用n: 跳出率分析SparkCore
HDFS/Hive
批类型应用(通过任务ID启动的Spark应用)
RDD1
RDD2
RDD3
RDD4
...
RDDN
任务执行结果数据保存MySQL(根据需要有可能保存HDFS)
实时处理类型应用(一直运行，不能通过任务id启动)
Kafka集群
Dstream1
Dstream2
Dstream3
Dstream4
...
DstreamN