《日志集中管控平台》技术方案

版本：草案

1. 项目描述
   1. 项目背景及现状

公司提出建设用户访问日志集中管控平台，实现对用户访问行为的有效跟踪管控，为业务系统提供新的增长极。

* 1. 项目建设目标

本项目计划周期：2018.09.01~2018.10.25。

采集业务支撑系统部和网络部各系统的敏感信息操作日志，存储于总部日志集中管控平台。日志集中管控平台，实现业务支撑系统部和网络部日志数据的集中化采集、存储和可视化。

1.实现日志数据格式统一；

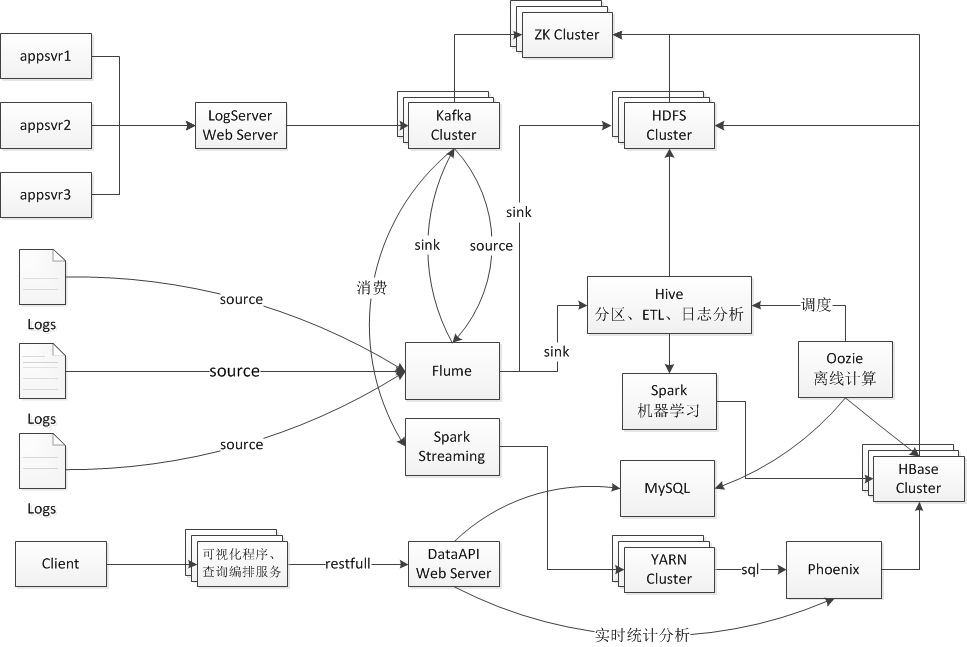
2.建立用户访问日志集中管控平台，整合所有操作日志数据；

3.实现用户访问操作日志条件查询；

4.提升平台能力，增加关联分析及审计功能，及时发现违规操作；

5.提供全文检索及报表视图；

1. 项目解决方案
   1. 技术架构



技术架构分为四层：数据采集层、存储计算层、数据服务层、应用层。

* 1. 数据采集层

负责从各类数据源中提取、导入数据，主要包括被动采集（LogServer）、主动采集Flume。

* + 1. LogServer

提供日志写入接口，所有已接入的应用程序发送日志至LogServer，在LogServer进行预处理后转发至消息队列Kafka，预留预处理逻辑的接口。

* + 1. Flume

数据源包括各类动态数据（如行为数据）、静态数据（如属性数据）、日志文件以及其它数据等，主要是半结构化的和非结构化的数据（txt、csv、json、xml……）。对于已结构化的数据（sql），使用sqoop方式从关系数据库导入。

* 1. 存储计算层

包括：分布式存储HDFS系统、数据仓库Hive、No-SQL数据库HBase、SparkStreaming流计算、YARN资源调度、协调系统ZooKeeper等。此外，还需用到关系数据库MySQL担任数据中转、元数据存储、供某些软件使用等用途。

* + 1. HDFS

备份原始日志数据。

* + 1. Hive

分区表：将日志按时间建立分区表，精确到天。

ETL处理：

1）识别流量攻击、网络爬虫、流量作弊；

2）数据缺项补正：对日志中的一些重要数据做取值归一，标准化处理或反向补正（基于新数据对老数据的回补修正）；

3）无效数据剔除：因业务变更或配置不当，在采集到的日志中会存在一些无意义、已经失效、或者冗余的数据，剔除它们减少存储空间与计算能力的消耗；

日志分析：新增用户统计、活跃会员分析、用户浏览深度、用户基本信息分析、操作系统分析、地域信息分析、用户浏览深度分析、外链数据分析、（后续分析维度视数据情况定……）

* + 1. HBase

存储机器学习数据模型、提取的数据特征集等。

* + 1. Spark streaming

流式计算，实时日志处理。

* + 1. YARN

为Hive与Spark的执行引擎提供统一资源调度。

* 1. 数据服务层

提供数据检索接口DataAPI。在ElasticSearch提供的数据搜索功能的基础上，针对业务需求进行定制和优化。

* 1. 应用层

日志检索：基于字段条件的精确查询以及基于字段内容的模糊查询。

可视化：基于Grafana的日志数据不同维度的统计分析指标显示。

推荐系统：基于用户访问行为的内容推荐。

3.项目模块