《教育大数据分析平台》技术方案

版本：草案

1. 项目描述
   1. 项目背景及现状

公司提出建设教育大数据分析平台，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。用户对课程的满意度是在线教育企业可持续发展基础，借助精细化运营手段，分析用户使用App的行为，推测用户的满意度，分析用户需求，向用户主动推送内容，提升用户活跃度，提高用户粘性。

整个项目分为离线统计和进线（实时）统计。对用户从多角度进线分析，提升用户对产品的满意度。

* 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；

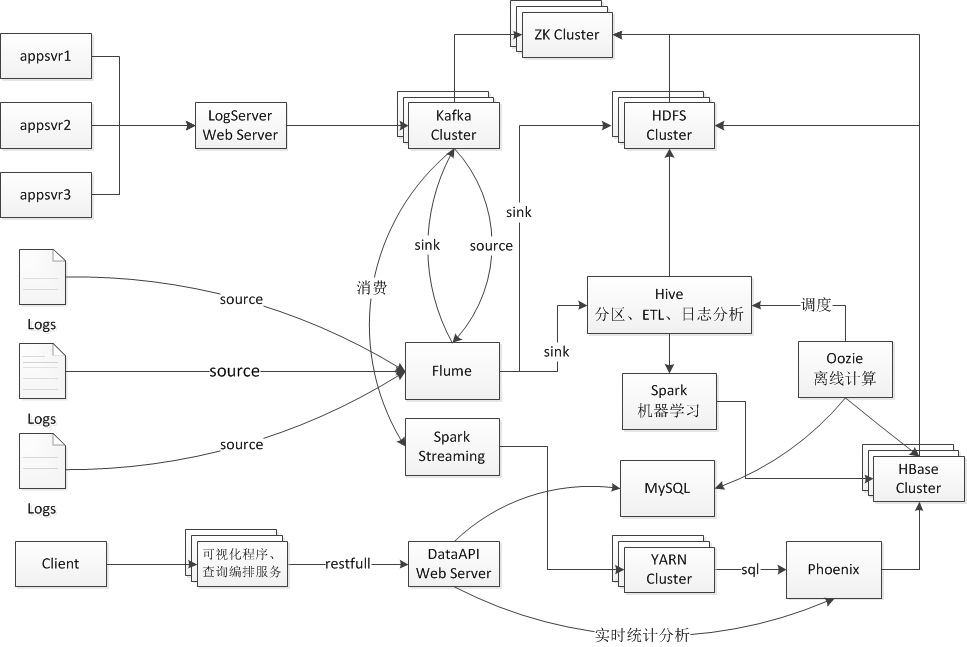
2.建立用户访问日志集中管控平台，整合所有操作日志数据；

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。

1. 业务功能---离线部分

离线业务功能分为3个模块：1、数据清洗；2、数据分析（用户基本信息分析、地域信息分析、用户浏览深度分析、订单分析、事件分析）；5、可视化展示（Grafana展示数据/Tableau图表展示）。

* 1. 数据清洗
     1. 数据清洗目的

将原生数据采集到HDFS，我们也称为不规整数据，即目前来说，该数据的格式还无法满足我们对数据处理的基本要求，需要对其进行预处理，转化为我们后面工作所需要的较为规整的数据，所以这里的数据清洗，其实指的就是对数据进行基本的预处理，以方便我们后面的统计分析，所以这一步并不是必须的，需要根据不同的业务需求来进行取舍，只是在我们的场景中需要对数据进行一定的处理。

* + 1. 数据清洗方案

我们采用HiveSQL和MapReduce来对数据进行预处理，预处理之后的结果，我们也是保存到HDFS中，即采用如下的架构：



* + 1. 数据清洗过程

**第一步：缺失值清洗**

缺失值是最常见的数据问题，处理缺失值也有很多方法，我建议按照以下四个步骤进行：

1. **确定缺失值范围：**对每个字段都计算其缺失值比例，然后按照缺失比例和字段重要性，分别制定策略。

**2、去除不需要的字段：**这一步很简单，直接删掉即可……但强烈建议清洗每做一步都备份一下，或者在小规模数据上试验成功再处理全量数据，不然删错了会追悔莫及(多说一句，写SQL的时候delete一定要配where!)。

**3、填充缺失内容：**某些缺失值可以进行填充，方法有以下三种：

**以业务知识或经验推测填充缺失值**

**以同一指标的计算结果(均值、中位数、众数等)填充缺失值**

**以不同指标的计算结果填充缺失值**

**4、重新取数：**如果某些指标非常重要又缺失率高，那就需要和取数人员或业务人员了解，是否有其他渠道可以取到相关数据。

以上，简单的梳理了缺失值清洗的步骤，但其中有一些内容远比我说的复杂，比如填充缺失值。

**第二步：格式内容清洗**

如果数据是由系统日志而来，那么通常在格式和内容方面，会与元数据的描述一致。而如果数据是由人工收集或用户填写而来，则有很大可能性在格式和内容上存在一些问题，简单来说，格式内容问题有以下几类：

**1、时间、日期、数值、全半角等显示格式不一致**

这种问题通常与输入端有关，在整合多来源数据时也有可能遇到，将其处理成一致的某种格式即可。

**2、内容中有不该存在的字符**

某些内容可能只包括一部分字符，比如身份证号是数字+字母，中国人姓名是汉字(赵C这种情况还是少数)。最典型的就是头、尾、中间的空格，也可能出现姓名中存在数字符号、身份证号中出现汉字等问题。这种情况下，需要以半自动校验半人工方式来找出可能存在的问题，并去除不需要的字符。

**3、内容与该字段应有内容不符**

姓名写了性别，身份证号写了手机号等等，均属这种问题。 但该问题特殊性在于：并不能简单的以删除来处理，因为成因有可能是人工填写错误，也有可能是前端没有校验，还有可能是导入数据时部分或全部存在列没有对齐的问题，因此要详细识别问题类型。

格式内容问题是比较细节的问题，但很多分析失误都是栽在这个坑上，比如跨表关联或VLOOKUP失败(多个空格导致工具认为“陈丹奕”和“陈 丹奕”不是一个人)、统计值不全(数字里掺个字母当然求和时结果有问题)、模型输出失败或效果不好(数据对错列了，把日期和年龄混了)。

**第三步：逻辑错误清洗**

这部分的工作是去掉一些使用简单逻辑推理就可以直接发现问题的数据，防止分析结果走偏。主要包含以下几个步骤：

**1、去重**

**2、去除不合理值**

一句话就能说清楚：举例，年龄200岁，年收入100000万(估计是没看见”万“字)，这种的就要么删掉，要么按缺失值处理。

**3、修正矛盾内容**

有些字段是可以互相验证的，举例：身份证号是1101031980XXXXXXXX，然后年龄填18岁。在这种时候，需要根据字段的数据来源，来判定哪个字段提供的信息更为可靠，去除或重构不可靠的字段。

逻辑错误除了以上列举的情况，还有很多未列举的情况，在实际操作中要酌情处理。另外，这一步骤在之后的数据分析建模过程中有可能重复，因为即使问题很简单，也并非所有问题都能够一次找出，我们能做的是使用工具和方法，尽量减少问题出现的可能性，使分析过程更为高效。

**第四步：关联性验证**

如果数据有多个来源，那么有必要进行关联性验证。例如，有汽车的线下购买信息，也有电话客服问卷信息，两者通过姓名和手机号关联，那么要看一下，同一个人线下登记的车辆信息和线上问卷问出来的车辆信息是不是同一辆，如果不是，那么需要调整或去除数据。

* + 1. 数据清洗格式

由于数据来源不同，可能涉及到几下几种数据格式，针对不同的格式需采取不同清洗方式。

涉及到的数据格式有：text、csv、pdf、parquet、和json等数据格式的文件。

* 1. 数据分析
     1. 用户基本信息分析

新增用户统计分析：分析新用户次日、次周留存情况，分析导致新用户流失原因，提升新用户留存。

活跃会员分析（只有产生看课、买课行为的用户才为活跃用户）：对比分析活跃用户、 累计用户的用户类型分布，如学生、教师、占比，了解平台内用户结构。

* + 1. 地域信息分析

对ip进行解析。分析出省市的看课百分比。

* + 1. 用户浏览深度分析（Hive）

用户浏览深度分析中，通过depth值来表示用户的浏览深度，分别从两个不同的角度来展示浏览深度，分别为会话和用户。会话是指，每个depth阶段对应的会话个数；用户就是指每个depth阶段的用户个数。Hive的最终数据保存到hdfs的指定目录中，通过sqoop将数据导入到mysql中。

计算规则：当前url的个数作为depth值。

* + 1. 订单分析（Hive）

包含订单数量、订单金额、成功支付的订单数量。

订单区间分析：不同价格课程的订单人数统计。

* + 1. 事件分析（Hive）

主要分析事件的触发次数，通过查看事件的触发次数可以得到事件转换率或者用户对此类事件的兴趣。

* + 1. 课程分析（Hive）

分析课程学习完成度、课程人均学习时长，课程学习频次等指标，掌握各课程受欢迎程度，优化课程提升核心竞争力；分析各教师受欢迎程度及学员对教师的评价等，提升教师教学能力，提高学员的用户满意度。

1、课程销售情况

实时获悉支付成功的总次数、总人数、实际支付总金额及支付成功的优惠券总金额，了解当前课程销售情况。

2、课程销售分析

分析课程每周、每月销量变化，对比各课程销售情况、课程预售页面浏览量，合理调整课程结构及定价策略，提升用户付费。

3、促销分析

分析试听引导成单率、课程详情页转化率等，结合用户复购比例、课程销售分布等指标，制定有效的促销策略，提高付费转化。

4、分享分析

统计分享了课程的学员的转化率。

* 1. 可视化展示

采用提供rest api的方式提供json格式的分析数据，主要方便前段程序来调用该接口来展示数据。整个可视化项目存放在kdc\_dataapi下。

教学仪表盘：可视化教学数据界面让教学过程尽在掌握，支持多视角切换、数据层次缩放，如课程进度监控、学生分组画像、教学内容分析。准确客观的教学过程数据及结果数据促进教学研究及教学模式创新。

1. 业务功能---近线部分

近线业务功能主要采用SparkStreaming的方式，通过消费kafka的数据实现以下功能：

实时计算每天各用户对各个课程的点击次数、计算每天各省的top3热门课程、对论坛网站动态行为的为维度分析。通过获得点击的课程信息，调用推荐算法模型，实时获得推荐的书籍并展示。

* 1. 每日热点搜索词统计

需求：解析访问日志数据，筛选出每天搜索uv排名前3的搜索词，按照每天的top3搜索词搜索总次数，倒序排序。

过程：采集实时产生的日志信息并交由SparkStreaming处理:由Flume实时采集日志信息，将信息Sink到kafka，KafKa将数据生产到SparkStreaming，最后由SparkStreaming消费数据，将结果保存到Hive中。

提交方式：将程序打成jar包，提交到YARN上执行。

* 1. 用户访客地区统计

需求：解析访问日志数据，根据访问的IP地址，转换成地市，然后根据课程地址进行访问的统计。

过程：采集实时产生的日志信息并交由SparkStreaming处理:由Flume实时采集日志信息，将信息Sink到kafka，KafKa将数据生产到SparkStreaming，最后由SparkStreaming消费数据，将结果保存到HBase。

提交方式：将程序打成jar包，提交到YARN上执行。

* 1. 最受欢迎的TopN课程

需求：解析访问日志数据，获得课程的访问地址列表，然后根据课程地址进行访问的统计。

过程：采集实时产生的日志信息并交由SparkStreaming处理:由Flume实时采集日志信息，将信息Sink到kafka，KafKa将数据生产到SparkStreaming，最后由SparkStreaming消费数据，将结果保存到HBase。

提交方式：将程序打成jar包，提交到YARN上执行。

1. 集群信息

两个账户：root，密码：123.com

hadoop,hadoop123

硬件环境：

|  |  |  |  |
| --- | --- | --- | --- |
| ip | 内存 | 硬盘 | 操作系统 |
| 192.168.9.20 | 24G | 500G | CentOS7.3.1611 |
| 192.168.9.80 | 32G | 800G | CentOS7.3.1611 |
| 192.168.9.220 | 16G | 500G | CentOS7.3.1611 |

软件环境：

|  |  |
| --- | --- |
| 软件 | 版本 |
| Linux相关软件 | CentOS7.3.1611 |
| mysql-5.6.40 |
| JDK1.8.0 |
|  |
| Hadoop相关软件  采用CDH5.14.2版本 | Hadoop-2.6.0-cdh5.14.2 |
| HBase-1.2.0-cdh5.14.2 |
| Oozie-4.1.0-cdh5.14.2 |
| ZooKeeper-3.4.5-cdh5.14.2 |
| ext-2.2 |
| hive-1.1.0-cdh5.14.2.tar.gz |
| Spark相关软件 |  |
|  |
|  |
|  |
| Hadoop本地开发环境 | ideaIU-2018.1.4 |
| apache-maven-3.5.3 |