项目需求

1. 项目概述

我们项目-基于电商的大数据分析平台，项目抽取出来的离线数据分析部分。

综合所学的大数据框架，进一步巩固所学的知识，并掌握企业中大数据开发整体流程，增加大数据开发项目经验。

通过真实的企业数据分析项目，提高项目架构，实战开发能力，做到立即上手开发大数据项目。

本项目贯穿了Hadoop生态圈中的主要技术框架，并联合使用。

项目分为五大部分：

（1）企业项目业务设计：学习企业中常见的**业务需求**。

（2）数据采集

（3）数据清洗

（4）数据分析

（5）数据展示

1. 大数据的分析流程

大数据开发流程也是大致遵循软件开发流程，分为：

1. 项目调研
2. 需求分析
3. 方案设计
4. 编码实现单元测试
5. 测试（集成测试，用户测试，压力测试）
6. 部署上线
7. 运行维护几个阶段

**2.1.1.项目调研**

（1） 目的是了解项目的需求，结合当前技术进行可行性分析，确定是否能够达到目标。

（2） 这个阶段业务性较强，开发人员必须了解业务，项目的进行都是以业务为导向。

（3） 一般会有比较专业的需求及概念，比如金融兴业：贷款，存款，利率，本息等金融行业的专业术语。

**2.1.2需求分析**

（1）明确整个项目中的实际需求：

①．要做什么

②．做成什么样子

（2）需求分析是否做好很大的影响了项目的进度，如果需求分析做的不明确很可能会导致项目的工期延迟或者会出现返工。

（3）明确客户的需求：

①显示需求：客户明确提出的需求。

②隐式需求：根据客户明确提出的需求，发掘出的需求，这对明确客户的需求有很大的帮助。

**2.1.3 方案设计**

（1）概要设计：从整体上对项目进行基本的技术架构的设计实现，比如模块的划分，使用的技术架构等。

（2）详细设计：根据概要设计中模块划分进行具体的设计，比如使用Hbase,要设计表结构，Rowkey等等。

（3）方案设计的好坏决定了编码的效率，方案设计相当于伪代码，伪代码转化为代码实现开发逻辑。

（4）方案设计的要求：程序的扩展性，可指定性。

**2.1.4 编码实现**

（1）根据详细设计文档对模块进行代码开发。

**2.1.5 测试阶段**

（1）功能测试：测试模块功能是否达到目标。

（2）联通测试：多个模块进行联合测试。

（3）压力测试：高并发，多用户条件下，系统是否能够正常运行。

（4）用户测试：选取关键用户对新系统进行测试，根据建议反馈可以修改。

**2.1.6部署上线**

（1）运行维护：数据备份，日志分析等

三．数据流量及集群规划

**3.1.1数据量：决定了集群的规划**

（1）字段个数，比如30~40个字段，每条记录大小为300~600字节

（2）每天处理的记录，大型网站：千万级别数据访问记录；中小型网站；百万级别的访问量。

（3）处理数据的总量

每条记录大小\*每天的记录数\*2年

（4）保存的时间，一般是保存2~3年

（5）计算每台服务器存储数据量：

①中小型集群：20~50台机器磁盘大小是16T（8槽x2T磁条）；

②一般的磁盘利用率不超过80%，也就是说每台机器存储12.8T

（6）计算需要多少台机器进行存储（DataNode节点个数），

①中小型集群：20~50台机器；

②中型集群：50~100台机器；

③大型集群：100~台以上

**3.1.2 集群规划**

（1）DataNode/NodeManager:25台

（2）NameNode/ResourceManager:2台

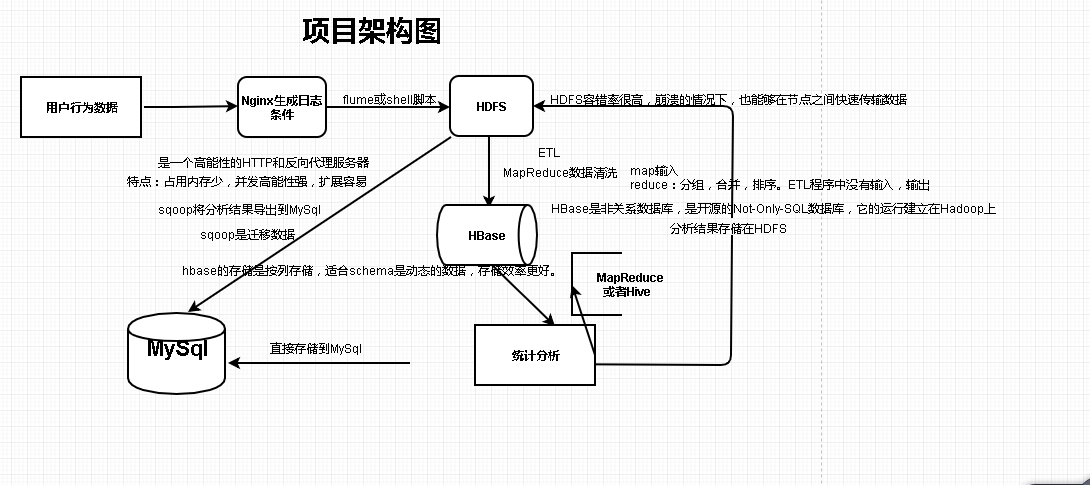
（3）Zookeeper:3台

（4）JournalNode:3台

（5）Hive/sqoop:1台

四．常见的电商大数据分析需求

数据分析工作是非常重要的，也是确实能够带来实际效果的。比如说利用数据分析做用户画像以进行精准化营销；利用数据分析来改进现有产品的结构，让用户有更好的购物体验；利用数据分析来管理用户的生命周期，提高用户的忠诚度，减少用户流失；根据用户的购买数据，挖掘用户的潜在需求，提供精准化服务，扩大影响力等等。



五．数据采集阶段

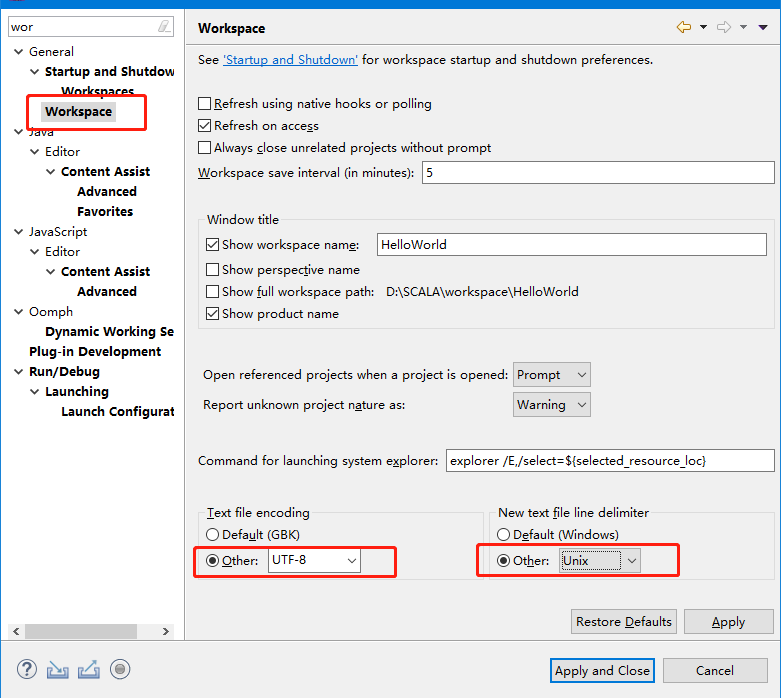
**（1）日志收集服务器Tengine：**

tengine日志分布到不同服务器上，首先使用日志服务来分别收集这些日志。再以logHub作为日志源，来进行监控。

**（2）JS SDK/Java SDK生成日志信息传入到Tengine**

一．JS SDK

1.修改配置workspace



2.导入代码步骤：1选中空白处右击—》Import—》Existing Projects into Workspace—》offline\_data\_analystics

**3.****Flume日志采集(source,channel,sink)**

## 一、Flume 组件

* Source 采集信息源
* Channel 运输数据/缓存数据
* Sink 下沉数据，负责和数据源目的地对接,并处理。

消息 Record，Source封装Event(事件)成为Record对象，并保存到Channel中，Sink拉取Record并保存到目标系统中。

Sink处理完成之后，会向Channel发送确认消息，提供消息处理的可靠性。

* 1. **日志数据导入HDFS**

1. 把日志数据上传到HDFS中进行处理，可以分为以下几种情况：
2. 如果是日志服务器数据较小，压力较小，可以直接使用shell命令把数据上传到HDFS中。
3. 如果是日志服务器数据较大，压力较小，使用NFS在另一台服务器上上传数据；
4. 如果日志服务器非常多，数据量大，使用flume进行数据处理；

**六．数据清洗**

**1. 数据清洗过程（ETL数据抽取、转换、加载）**

**2． MapReduce程序实现数据清洗**

（1）字段的提取过滤

（2）字段的格式化

（3）字段的补全

### 3. ETL（MapReduce）的流程

将日志解析存入hbase的表，没有排序/合并/聚合的需求，所以一般的ETL程序没有shuffle和reduce过程

MapReduce的具体过程

（1）输入-》map-》shuffle-》reduce-》输出

（2）map输入：

①从hdfs上读取昨天的日志

②数据类型

key：longwritable：行的偏移量

value：Text：日志文件一行的内容

③在hdfs上每天的日志是按照日期命名的单独的文件目录存放，所以要动态的去读取。

（3）进行日志的解析，调用LogParser类型进行字段的提取和补全

（4）返回的是map集合

（5）输出到Hbase

①输出数据类型

（map端）封装Put

rowkey：实现rowkey的生成

列簇

列标签

值

*（6）shuffle：ETL程序中没有*

*①分区*

*②排序*

*③合并*

*（7）reduce：分组，合并，排序：ETL程序中没有*

*①输入*

*②输出*

（8）驱动类

①初始化map和reduce

②tableMapReduceutil

（9）需要解决的问题

①动态的按照日期读取日志文件

②创建hbase表

如何设计hbase表

rowkey

列簇

**4. 为啥清洗数据结果放入Hbase**

前天大家商量大数据的架构,都认为应该要先清洗数据 然后把清洗好的数据入到hbase.说是这样节省磁盘空间。我认为这样做的话，没有发挥hdfs的实际作用。应该是先把原始数据直接入到hdfs，再通过hive和mapreduce做数据的清洗入到hbase。

**1.2、HBase的特点**

◆大表：一个表可以有上亿行，上百万列。

​◆面向列：面向列表（簇）的存储和权限控制，列（簇）独立检索。

​◆稀疏：对于为空（NULL）的列，并不占用存储空间，因此，表可以设计的非常稀疏。

​◆无模式：每一行都有一个可以排序的主键和任意多的列，列可以根据需要动态增加，同一张表中不同的行可以有截然不同的列。

​◆数据多版本：每个单元中的数据可以有多个版本，默认情况下，版本号自动分配，版本号就是单元格插入时的时间戳。

​◆数据类型单一：HBase中的数据都是字符串，没有类型。

**Hive 实现**

HIVE概念：hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进

行。

**七．数据分析**

**1.需求：**项目中：数据存储在hbase中

-》ETL-》hbase-》MapReduce（要自定义mysqloutputformat）-》MySQL

-》ETL-》hbase—》Hive-》HDFS-》SQOOP-》MySQL

**2.分析维度的组合**

维度 个数

2017年12月20日谷歌浏览器的1.40版本 50

2017年12月20日website平台 100

2017年12月20日IOS APP平台 1000

**3.Hbase数据前的过滤**

（1）过滤所有事件类型为launch事件的记录

①字段：en，值是e\_l

（2）过滤需要分析的字段

①uuid：访客id

②s\_time：时间

③pl：平台的类型

④version：平台的版本

⑤browsername：浏览器的名称

⑥browserversion：浏览器版本

所以，在开发的时候要有一个类（HbaseScanUtil.java）对hbase的数据进行过滤（1）（2），然后调用这个类返回一个scan对象给driver。