大数据技术课程结课设计实验

1 实验目标

本学期常规实验主要有三个，实验一主要熟悉大数据技术常用框架的搭建部署，实验二目的在于熟悉 kafka 消息队列组件以及常用的存储架构，实验三的目的在于熟悉大数据处理技术的常用计算框架 MapReduce 和 Spark。在本次课程结课大实验中，主要目标是要求大家使用分布式处理与分布式存储技术对结构化的日志数据进行一些常用的统计分析处理，同时对处理结果进行 WEB 展示，从而使同学们熟悉大数据在在企业中的一般应用场景和工作流程。

具体实验目标包括：

(1) 利用 Kafka 拉取日志数据，并创建消费者将数据消费至分布式存储平台；

(2) 利用分布式处理技术对结构化日志数据进行聚合处理，并将处理结果存入 MySQL；

(3) 对数据处理结果进行 WEB 展示；

(4) 对数据进行统计指标计算，形成报表存储至 HDFS.

2 实验要求

2.1 数据集说明

本次实验所使用的数据集包含了 50W 条 GDS 订票请求日志数据。GDS 为全球分销系统，用来向全球各个国家销售国内航空公司的机票。国内外旅客通过 GDS 向中航信发送订票请求，产生一条 GDS 订票请求日志数据，日志数据记录了订票请求的请求方、响应方和所购机票的起飞机场、所属航空公司等信息。GDS 订票请求日志数据格式和详细信息如下：

数据样例

4722024&29730498|20190423010000|27201020|24H|ATH|1P|T|GR|1P|MF|ERC|999|No reply message|0|1

样例	说明
4722024&29730498	Rowkey：一条报文的标识符
20190423010000	Dt:时间，2019 年 04 月 23 日 01 时 00 分 00 秒
27201020	机票的属性一 MAC，GDS 处理标记一
24H	机票的属性二 PCM，GDS 处理标记二
ATH	机票的属性三 Airport，记录起飞机场的三字码
1P	机票的属性四 Airline，记录机票所属的航空公司
T	机票的属性五 Agent，记录销售机票的代理
GR	机票的属性六 Country，记录订票请求来自的国家
1P	Req:请求方，常见的还有 1A，1E 等
MF	Res:响应方，常见的还有 CA，MU 等
ERC	错误标识，错误报文为 ERC，正确为 S
999	错误码，定义了错误的类别
No reply message	错误类型，和错误码一一对应
0	成功数：若该条报文是成功报文，则这个字段记为 1，否则记为 0
1	失败数：若该条报文是失败报文，则这个字段记为 1，否则记为 0

2.2 要求介绍：

(1) 统计并展示一段时间内，请求方和请求订购机票的属性三 Airport 都相同的 GDS 订票请求数。查询维度包括：时间跨度（小时级别、天级别）、请求方。

例如某次查询统计结果如下所示，其中时间字段表示从这个时间开始的一个小时内：

时间	请求方	Airport	请求数
2019-04-23 12:00	1P	BKG	1000
2019-04-23 12:00	1P	ATH	800
…	…	…	…

(2) 统计并展示一段时间内，某个请求方在所有响应方上的 GDS 订票请求成功数、失败数、成功率。查询维度包含：时间跨度（小时级别、天级别）、GDS 请求方。（**选做，加 10 分**）

例如某次查询统计结果如下所示：

时间	请求方	响应方	请求成功数	请求失败数	成功率
2019-04-23 12:00	1P	MF	800	200	0.8
2019-04-23 12:00	1P	MU	300	200	0.6
…	…	…	…	…	…

(3) 计算并统计指标，每一天中请求总数最多的 5 个 GDS 请求方，成功数最多的 5 个请求方、失败数最多的 5 个响应方。

3 实验内容

- 分析日志数据，设计日志数据的 HBase 细节存储模型（**若选用 HDFS 作为分布式存储平台则无需设计存储模型**）。设计 MySQL ER 模型，要求能够存储聚合得到的小时粒度、天粒度数据。

- 编写生产者程序，将日志数据发送至 Kafka，并通过消费者（**流式计算框架 Storm、Spark Streaming、单机消费者三者任选其一**）将数据消费至分布式存储平台（**HDFS 或 HBase 任选其一**）。
- 利用分布式计算框架 （**Spark 或 MapReduce 任选其一**） 计算日志数据，并解析聚合成小时粒度以及天粒度数据，存储至 MySQL。
- 根据要求设计前台网页交互，推荐 Java Web，按照要求指定的展示粒度，从 MySQL 中读取数据并展示（**可以选用 Grafana 等开源数据可视化平台也可自行编写展示界面**）。
- 完成实验要求中的第三条，生成报表，并存储至 HDFS。

4 实验步骤

(1) 分析数据和需求，设计数据处理系统架构以及存储模型。包括，HBase 细节数据存储模型，MySQL 小时粒度天粒度数据存储模型；
(2) 编写 Kafka 生产者，读取文本日志数据并发送至 kafka 集群，编写消费者将数据消费至分布式存储平台；
(3) 利用分布式计算框架，解析日志数据，得到小时粒度的 GDS 订票请求数据，存储至 MySQL，并将小时粒度数据聚合为天粒度数据，存储至 MySQL；
(4) 设计 WEB 展示界面，对 MySQL 中的计算结果进行展示；
(5) 计算统计指标，生成报表并存储至 HDFS。

5 实验环境

(1) MySQL；
(2) Kafka 集群；
(3) 分布式文件系统：HDFS ；
(4) 数据仓库：Hive；
(5) 分布式数据库：HBase；

(6) 分布式近实时处理框架：Spark Streaming；或分布式流式处理框架：Storm；

(7) 分布式计算框架：Hadoop MapReduce 或 Spark 等；

(8) 编程语言：Java（推荐使用）或 Scala 或 C++等；

6. 实验报告

实验分小组开展，实验结果由实验报告和汇报 ppt 两部分组成。实验报告每个人需要完成各自的报告，除了总体的成果外，还需单独介绍个人在分组设计中的贡献和心得体会等内容。汇报 ppt 一个小组共同完成一份即可，通过课程结课汇报验收。

结课汇报时间：第 17 周周五下午，地点另行通知