2025 年第 5 学期大数据方向项目设计要求

项目整体要求

1) 项目的主旨是使用 Hive 和 HBase，对存储在 HDFS 和 HBase 中的数据进行分析整理,获得各种统计结果。
2) 每人至少 5 种借助 Hive 的统计结果进行分析和至少 1 种借助 HBase 实现的分析。
3) 将分析所得数据结果进行数据可视化,可视化不能使用静态文件。
4) 由于不能参加项目评估导致的不及格，不能参加下学期期初的补考

项目实现的要求

1) 代码库：在 gitee.com（码云，其他代码托管平台也可以）创建自己项目的公开项目，将项目组代码统一提交到代码库
2) 项目组进度：进度统计文件，保证每周至少更新一次，每周五 18:00 之前，保证项目进度是最新的情况。
3) 作为可视化分析系统，需要 Web 访问的能力，需要将开发完成的分析展示功能，部署到 Linux 服务器上，并通过其他电脑访问。系统应该具备基本的登录、不同用户看到的菜单和页面不同的基本权限控制功能。
4) 数据文件：从互联网查找可用的大数据文件，可以优先选择从阿里的天池项目获取项目的数据，数据文件在 200M-500M 之间为宜，被处理的数据文件需要放在 Hive 表以及 HBase 表中。
5) 数据字典：对 Hive 和 HBase 数据表所有列设计数据字典，描述清楚每个字段的含义。参考模板《项目数据字典模板》。
6) Hive 数据表：根据数据字典创建 Hive 数据表，并编写 Hive SQL 语句将数据导入到表中。

   HBase 数据表:规划好 HBase 的表结构以及 RowKey,对于 RowKey 的设计需详细说明,如何避免数据热点和数据倾斜的。
7) 在将数据导入表前对数据进行适当的清洗和整理,清洗和整理的步骤和过程需要以文档记录的方式保存下来。可以使用 MapReduce、Linux shell、编写独立的应用程序等方法对数据进行清洗和整理。

8) 编写 Hive SQL 语句对表中的数据进行统计分析，自己的所有 HQL 中**必须出现**分组、子查询、分区、分桶、内置函数和自定义函数等方法。
9) 将统计分析结果导出到 MySQL 等数据库。再完成数据可视化展示。
10) 数据展示：使用列表或者图表方式展示统计分析的结果。图表方式尽量使用第三方图表类库完成。本学期推荐使用积木报表和帆软等免费和收费软件实现数据可视化（都可以免费使用）

    积木报表：http://report.jeecg.com/1423422

    帆软报表：https://www.finereport.com/
11) 相关限制说明：
    a) 不可以使用查询客户端的自动根据查询结果展示可视化报表的功能
    b) 同一个小组中，数据集不可以相同
    c) 每个 HQL 必须包含表连接、聚合等操作
    d) 两个 HQL 之间不能有逻辑上的相似性（比如：一个查最大值，一个查最小值）

项目讲评要求

1) 使用 NIIT 标准 PPT 模板，完成演讲 PPT 的制作
2) 项目展示分为以下步骤：
    [1] 项目组介绍
    [2] 项目功能介绍
    [3] 项目的亮点介绍
    [4] 项目整体演示
3) Hive SQL 脚本分析：每人重点讲解自己实现的 Hive SQL 脚本，并回答相应的问题。

项目提交要求：

1) 将自己的 PPT 以及代码项目文件和测试数据压缩（防止文件大小太大）后提交
2) 压缩包格式，（所有不按照格式提交的，无法统计分数）

    班级名称_小组名称_组长_组员 1_组员 2.zip

    例如：

    CLS01_0BUGGroup_张三_李四_王五.zip