大数据处理综合实验课程（2025）

课程实验3 - HBase 和 Hive 操作实验

南京大学　　计算机学院

课程实验3：HBase 和 Hive 操作实验

- 实验目标：
  - 熟悉 HBase 的安装以及命令行基本操作；
  - 熟悉 MapReduce 的原理及其程序的编写，了解 Hive 外部表的使用；
  - 理解 Hive 分区与分桶表概念，熟悉 HQL 语句的编写。

- 实验环境（版本不强制要求一样，但要选择兼容版本）：
  - Hadoop-3.2.1（可采用单机伪分布式环境）；
  - Zookeeper-3.4.12；（HBase 依赖于 Zookeeper）
  - Hbase-2.4.0；
  - Hive-3.1.2；
  - 课程实验平台。

课程实验3：HBase 和 Hive 操作实验

- 兼容性检查

HBase Version	JDK 6	JDK 7	JDK 8	JDK 11	JDK 17
HBase 2.6	✗		✓	✓	✓
HBase 2.5	✗		✓	✓	!*
HBase 2.4	✗		✓	✓	✗
HBase 2.3	✗		✓	!*	✗
HBase 2.0-2.2	✗		✓	✗	
HBase 1.2+	✗	✓	✓	✗	
HBase 1.0-1.1	✗	✓	!	✗	
HBase 0.98	✓	✓	!	✗	
HBase 0.94	✓	✓	✗		

	HBase-2.4.x	HBase-2.5.x
Hadoop-2.10.[0-1]	✓	✗
Hadoop-2.10.2+	✓	✓
Hadoop-3.1.0	✗	✗
Hadoop-3.1.1+	✓	✗
Hadoop-3.2.[0-2]	✓	✗
Hadoop-3.2.3+	✓	✓
Hadoop-3.3.[0-1]	✓	✗
Hadoop-3.3.2+	✓	✓

	HBase-2.5.x	HBase-2.6.x
Hadoop-2.10.[0-1]	✗	✗
Hadoop-2.10.2+	✓	✓
Hadoop-3.1.0	✗	✗
Hadoop-3.1.1+	✗	✗
Hadoop-3.2.[0-2]	✗	✗
Hadoop-3.2.3+	✓	✗
Hadoop-3.3.[0-1]	✗	✗
Hadoop-3.3.[2-4]	✓	✗
Hadoop-3.3.5+	✓	✓

课程实验3：HBase 和 Hive 操作实验

- 任务1：HBase 安装及命令行操作；
- 任务2：MapReduce 编程与 Hive 外部表管理；
- 任务3：Hive 分区分桶表与 HQL 实践；

- 数据集（IOlog.trace）说明：
  - 概述：一个分布式存储系统的 I/O 操作日志，记录了每次对数据块（block）的操作详情；
  - 字段：block_id, io_offset, io_size, op_time, op_name, user_namespace, user_name, rs_shard_id, op_count, host_name；
  - 分隔：一行为一条记录，一行中每个字段以单个空格 ' ' 作为分隔符。

任务1- HBase 安装及命令行操作

- 步骤0：安装 Zookeeper，并完成配置；
- 步骤1：安装 HBase，并完成配置；
- 步骤2：进入 HBase shell 命令行，熟悉 HBase shell 基础操作；
- 步骤3：现有以下关系型数据库中的表和数据，要求以伪分布式方式运行 HBase，通过 Shell 将其转换为适合于 Hbase 存储的表并插入数据。
  1. 设计并创建合适的表；
  2. 查询选修 Big Data 的学生的成绩；
  3. 学生表增加了联系方式，修改相应的 HBase 表，并添加数据；
  4. 查询 Zhang Li 的联系方式；
  5. 删除创建的表。

任务1- HBase 安装及命令行操作

课程表

课程号（C_No）	课程名（C_Name）	学分（C_Credit）
123001	Math	4.0
123002	English	3.0
123003	Big Data	4.0

学生表

学号（S_No）	姓名（S_Name）	性别（S_Sex）	年龄（S_Age）
2025001	Li Lei	male	20
2025002	Han Meimei	female	21
2025003	Zhang Li	female	20
2025004	Li Ming	male	19

任务1- HBase 安装及命令行操作

选课表

学号（SC_Sno）	课程号（SC_Cno）	成绩（SC_Score）
2025001	123001	68
2025001	123002	90
2025001	123003	96
2025002	123001	85
2025002	123002	73
2025003	123001	82
2025003	123002	91

增加联系方式后的学生表

学号（S_No）	姓名（S_Name）	性别（S_Sex）	年龄（S_Age）	联系方式（S_Email）
2025001	Li Lei	male	20	lilei@qq.com
2025002	Han Meimei	female	21	hmm@qq.com
2025003	Zhang Li	female	20	zl@qq.com
2025004	Li Ming	male	19	lm@qq.com

任务2- MapReduce 编程与Hive外部表管理

程序功能：基于 HDFS 中的 IOlog.trace，编写一个 MapReduce 程序统计每个用户命名空间下的所有用户的写操作（op_name=2）次数之和，并输出<user_namespace, sum(op_count)>。在 Hive 中创建一张 **外部表** 来管理输出的数据，并查询该外部表的全部数据。

说明：本任务第一部分即用一个 MapReduce 程序实现如下 SQL 语句的功能。

```
Select user_namespace, sum(op_count)
From Iolog.trace
Where op_name = 2
Group by user_namespace;
```

任务3- Hive 分区分桶表与 HQL 实践

- 步骤1：在 Hive 中创建一张分区表 IOlog_part_<学号>（例如 IOlog_part_12345678），要求以命名空间为分区条件；
- 步骤2：在 Hive 中创建一张分桶表 IOlog_bucket_<学号>（例如 IOlog_bucket_12345678），要求以 block_id 为分桶条件（桶的数量为3）；
- 步骤3：任意导入5条数据到分区表，从 HDFS 将 IOlog.trace 导入分桶表中；
- 步骤4：HQL 查询
  1. 查询分区表中某个（任意）分区下的所有数据；
  2. 查询分桶表中每个用户有几个不同的主机地址（host_name）；
  3. 查询分桶表中每个命名空间下所有用户的写操作（op_name=2）的总次数。

注：在 Hive 开启允许所有分区都是动态的

```
SET hive.exec.dynamic.partition=true;
SET hive.exec.dynamic.partition.mode=nonstrict;
```

注：在 Hive 开启动态分区和分桶写入

```
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;
SET hive.enforce.bucketing = true;
```

实验报告要求

实验报告的内容至少包括：

1. 任务一中每个步骤的结果截图及相关说明；
2. 任务二中 MapReduce 程序的结果截图（输出结果和程序运行时 Yarn 的 Web 界面任务监控截图），以及 Hive 查询结果的截图；
3. 任务三中每个步骤的结果截图及相关说明；

其他提交内容：

实验源代码，Jar 包及其使用说明。

实验报告的格式要求：

- 报告格式应为PDF；
- 代码及其他资源（如果有）打包为 zip；
- 报告/压缩包文件名：学号_姓名_BG_3；
- 如果提交内容包含 Jar 包，应当附带 Jar 包的执行说明。

- DDL：2025年4月15日