分布式系统实验报告

**实验三**

**分布式数据仓库Hive及其应用**

姓名： 樊启元

学号： 7203610712

院系： 未来技术学院

**2023年 5 月 19 日**

**1.实验内容**

Hive数据仓库是基于HDFS和MapReduce的分布式数据仓库，具备海量数据存储和处理能力，是大数据领域离线批量处理数据的常用工具，是将SQL语句解析成MapReduce作业的引擎，其本身不存储和计算数据，它完全依赖于HDFS和MapReduce。

实验题目：基站掉话率Top10统计。

实验要求：利用Hive，统计每个基站的掉话率，按降序排序，并找出掉话率比较高的前10个基站。

实验数据：jizhan\_information.scv文件。

jizhan\_information.scv文件说明

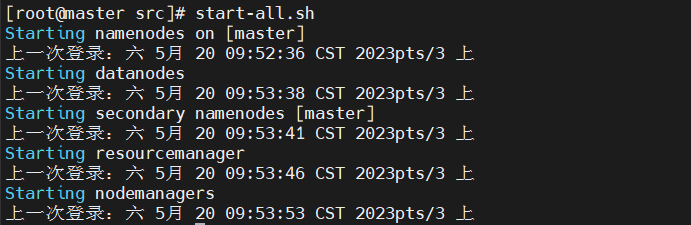
|  |  |  |
| --- | --- | --- |
| **字段名** | **说明** | **类型** |
| record\_time | 通话时间 | string |
| imei | 基站编号 | int |
| cell | 手机编号 | string |
| ph\_num | 接电话次数 | int |
| call\_num | 打电话次数 | int |
| drop\_num | 掉话的秒数 | int |
| duration | 通话持续总秒数 | int |
| drop\_rate | 掉话率 | double |
| net\_type | 通讯网络类型 | string |
| erl | 话务量的单位，话务量为呼叫次数与每次呼叫的平均占用时长的乘积 | int |

实验说明：基站掉话率的计算公式：



**2.实验步骤及结果**（截图说明）

启动 hadoop 服务



启动 mysql 服务

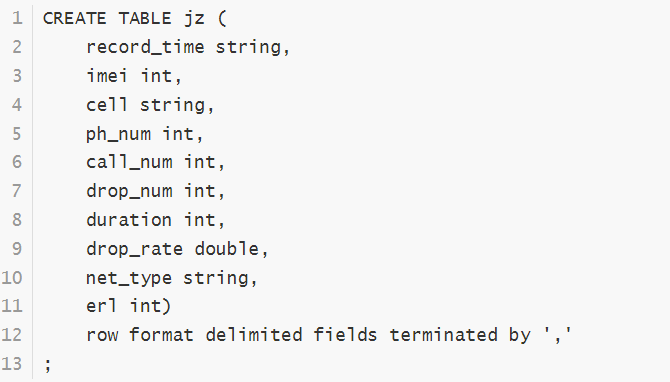


启动 hive 服务，并启动 hive

上传文件到 hdfs



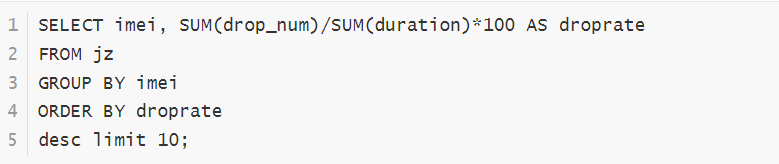
在 hive 的 card 数据库中创建表格：



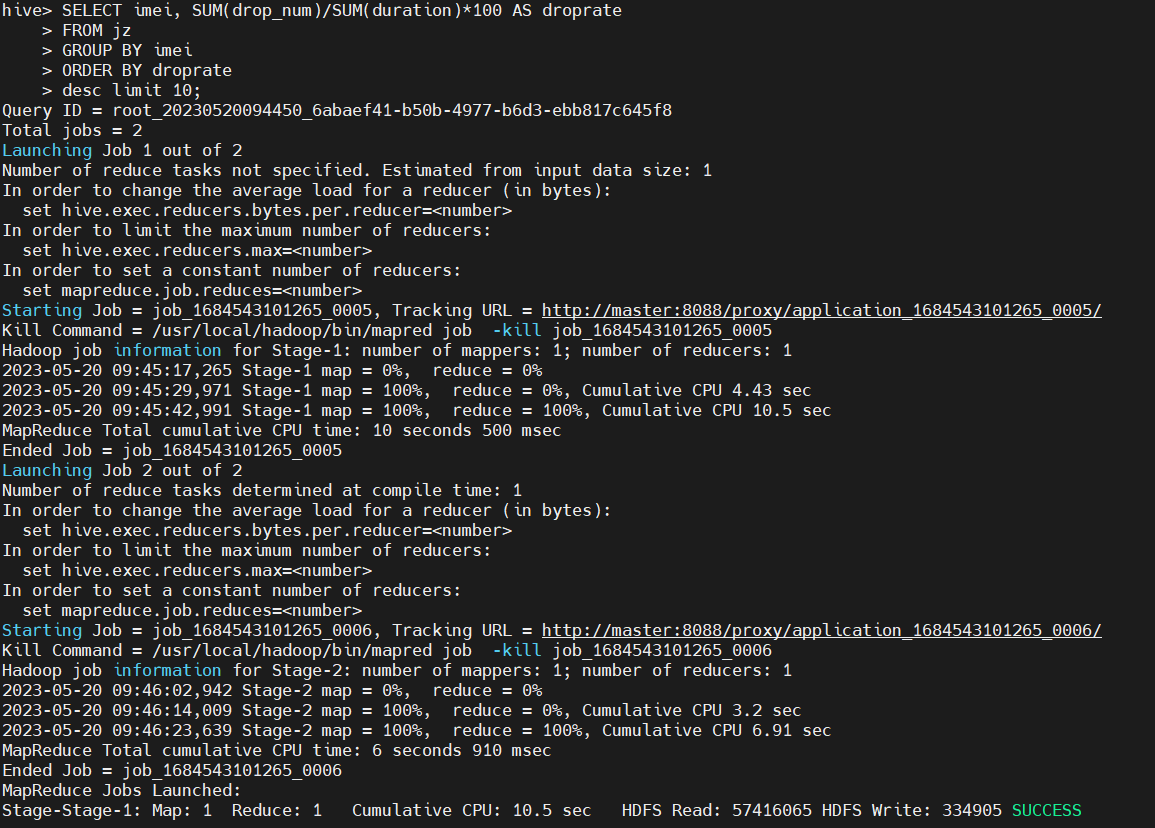
装载数据：

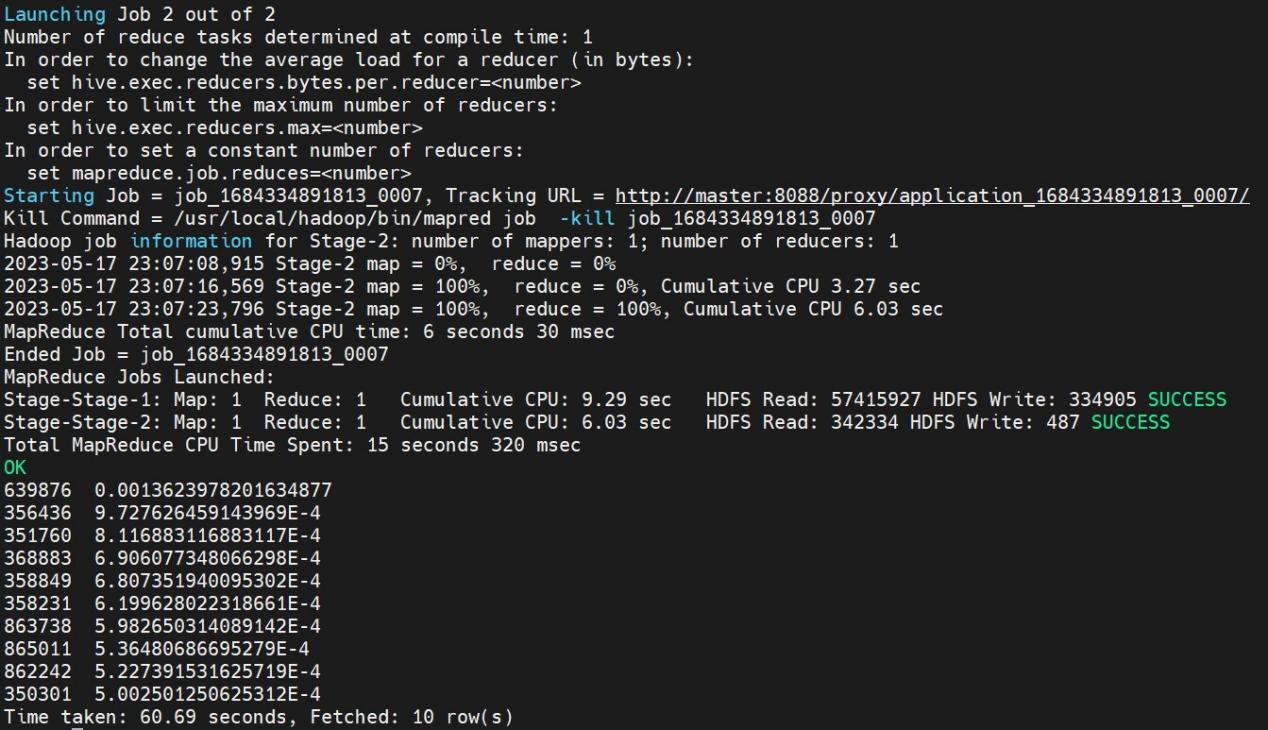


执行命令：



运行结果：





**3.个人建议及问题**

遇到的问题：

运行命令时：



解决方法：



再次执行命令即可。