|  |
| --- |
| **佛 山 科 学 技 术 学 院**  **《大数据技术架构大作业》实训报告**  专 业 数据科学与大数据技术 姓 名 陈荣辉 成 绩  班 级 17大数据3 学 号 20170060316 日 期 2020.5.8 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 一、实验目的  对数据“上牌数--商用车销量数据样例”中的20000-50000行数据进行统计分析。  二、实验平台  操作系统：Ubuntu16.04； Hadoop 版本：2.7.1；MySQL版本：5.1.40；HBase版本：1.1.5；Hive版本：1.2.1；SqoopR版本：1.4.6；Eclipse:3.8。   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Eclipse | Sqoop | HBase | Hive | | R | | Hadoop | | MySQL | | | Linux系统 | | | | | |   图2-1 本案例所涉及的系统及软件  三、实验内容和要求  1、上传数据集Carszu.csv    2、创建运行本案例的目录bigdatacase      3、查看bigdatacase目录下的Carszu.txt的前5行记录    4、数据集预处理，删除文件第一行记录（即字段名称）    5、新建脚本文件pre\_deal.sh    6、执行脚本文件pre\_deal.sh，对Carszu.txt进行预处理      7、启动HDFS    8、把user\_table.txt上传到HDFS中    9、启动MySQL数据库    10、启动Hive    11、创建外部表bigdata\_user    12、查看bigdata\_user表的各种属性    13、查看bigdata\_user表的简单结构    14、Hive语句查询数据集的前十行记录    15、Hive语句查询数据集中品牌的前十行记录    16、Hive语句查询数据集中省份的前十行记录    17、Hive语句查询数据集中省份和区县的前二十行记录    18、Hive语句数据集中有多少条记录（我上传的数据是30000条，但此处确是60018）    19、Hive语句查询有多少个不重复的省份    20、Hive语句查询数据集中有多少个制造商品牌    21、Hive语句查询数据集中1月份到6月份的记录    22、Hive语句当月份为6月时，根据pinpa进行分组统计    23、Hive语句查询符合年份为2013，功率为45，轴距为2500的有多少辆      24、Hive语句查询当区县为万荣县的时候，根据区县进行分组    25、Hive语句根据品牌、制造商进行分组      26、Hive语句根据区县、制造商进行分组      27、Hive语句查询品牌为东风和五菱的数量，然后除以总量即可求得对应品牌的市场占有量      28、Hive语句根据品牌、使用性质、燃料类型进行分组      29、Hive语句根据月份、制造商进行分组      30、Hive、MySQL、HBase数据互导，创建临时表user\_action    31、启动HDFS，查看user\_action文件是否被创建      32、将bigdata\_user表中的数据插入到user\_action，并查询    33、使用Sqoop将数据从Hive导入MySQL，启动MySQL数据库      34、创建数据库dblab，并查看编码    35、在数据库dblab中创建一个新表user\_action，并设置其编码为utf-8    36、执行数据库导入操作      37、查询user\_action表中的数据      38、使用Sqoop将数据从MySQL导入HBase，启动Hadoop集群、MySQL服务、HBase服务，并创建user\_action,并导入数据            39、查看HBase中user\_action表数据      40、安装R      41、可视化分析，使用Echarts |