|  |
| --- |
| 【实验名称】网站流量分析 (MapReduce+Hive综合实验） |
| 【实验目的】 |
| 综合应用MapReduce+Hive |
| 【实验原理】 |
| 使用MapReduce做数据过滤，使用Hive做数据分析。 |
| 【实验环境】 |
| 开发环境:   1. IDEA/Eclipse 2. 按前面的实验《基于Maven的JAVA开发环境》配置和创建工程   运行环境:  （1）OS: Ubuntu16.04  （2）已经部署好的Hadoop环境 |
|  |
| 【需求描述】 |
| http.log日志文件，是电信运营商记录用户手机上网访问某些网站行为的日志记录数据，其中上行流量+下行流量 = 总流量；  phone.txt是手机号段规则，是手机号码对应地区城市和运营商的数据。  数据格式说明：  1、http.log日志 数据格式：  手机号码，请求网站的URL，上行流量（20字节），下行流量（5000字节）  例如：18611132889 http://v.baidu.com/tv 20 5000  2、phone.txt 数据格式：  手机号前缀，手机号段，手机号码对应的省份，城市，运营商，邮编，区号，行政划分代码  例如：133 1332170 广西 南宁 电信 530000 0771 450100  （一条数据中多个字段用空格或制表符分隔）  要求：   1. 根据给的用户上网日志记录数据，计算出总流量最高的网站Top3(网站例如：v.baidu.com，weibo.com) 2. 根据给的用户上网日志记录数据，计算出总流量最高的手机号Top3 3. 根据给的手机号段归属地规则，计算出总流量最高的省份Top3 4. 根据给的手机号段运营商规则，计算出总流量最高的运营商Top2 5. 根据给的手机号段归属地规则，计算出总流量最高的城市Top3 |
| 【实验步骤】 |
| （1）将数据文件http.log、phone.txt上传到hdfs，保存的目录以个人学号区分，如001为学号：  （2）数据清洗  编写MapReduce程序实现数据清洗，将不完整的数据过滤掉，保证输出数据以 ','分割，输出数据保存在两个文件中。  （3）创建hive表(要求表名以姓名缩写首字母+学号)   1. 将清洗后的数据导入Hive 2. 使用SQL查询满足条件的数据（等连接的多表查询）   1、根据给的用户上网日志记录数据，计算出总流量最高的网站Top3(网站例如：v.baidu.com，weibo.com)  2、根据给的用户上网日志记录数据，计算出总流量最高的手机号Top3  3、根据给的手机号段归属地规则，计算出总流量最高的省份Top3  3、根据给的手机号段运营商规则，计算出总流量最高的运营商Top2  4、根据给的手机号段归属地规则，计算出总流量最高的城市Top3   1. 删除hive数据库（以姓名缩写首字母+学号创建的数据库） |