1. **需求背景**

依据用户行为记录,统计出区域每日热门商品排名,并将排名结果推送至指定的MySQL数据库中,以供其它部门使用

本功能为一次模拟线上每日热门商品统计功能的若泽零基础班结业作业.用以检验,复习以及融汇在若泽零基础班学习结业后的成果.具体要求如下:

* 原始数据:每天的用户行为记录,已收集在HDFS.商品信息,区域信息在MySQL上
* 任务功能幂等性,整个统计过程可以反复执行并保证最终正确
* 可以更新统计算法并以新算法对历史记录以天为单位进行重新统计
* 原始数据有可能出现中文,内容混乱等情况
* 使用在零基础班的学习成果
* 尽可能的以实际生产项目情况来模拟完成

1. **概要设计**

该功能的整体设计思路为数据仓库的离线数据ETL.

以Hive作为整个数据仓库的基石,使用Sqoop来作为数据仓库(Hive)和其它异构数据存储(MySQL)的数据桥梁.

功能主要分为两个作业:

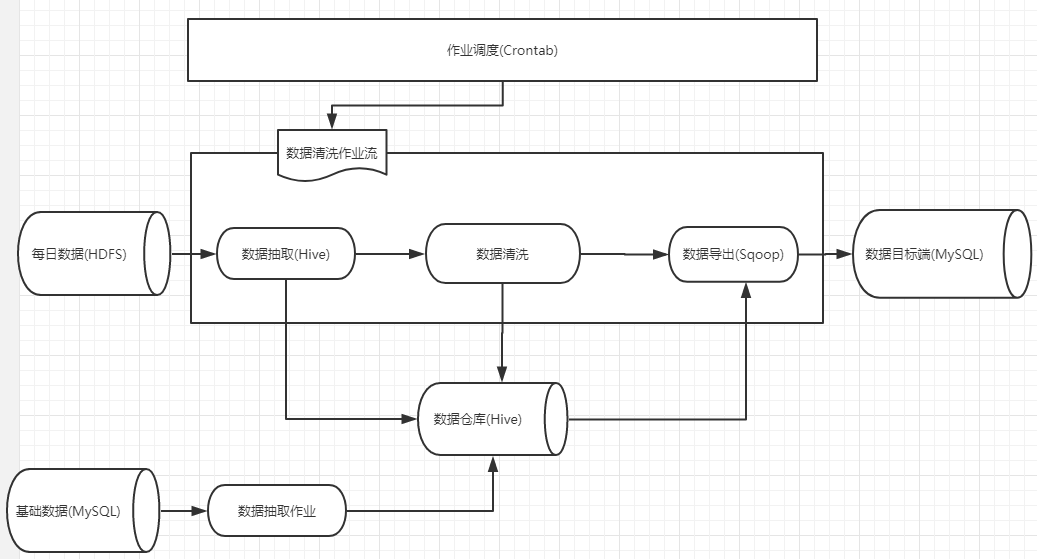
* **基础数据抽取作业**

将基础数据全量抽取到数据仓库中

* **每日数据清洗作业**

抽取每日数据到数据仓库中, 在数据仓库中统计生成每日热门商品排名,再将统计结果推送至目标存储系统以供页面展示

整个功能架构图如下:



1. **目录结构**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| bin | 本功能执行文件目录 | | | |
| init.sh | 数据仓库初始化脚本 | | |
| extact\_data.sh | 基础数据导入脚本 | | |
| trans\_day.sh | 数据清洗脚本  每日数据导入,清洗,导出至目的端 | | |
| shell\_lib | Shell基础库 | | |
| commmon.sh | | Shell中用到的通用函数 |
| conf | 配置文件目录 | | | |
| hot\_product.conf | | 每日区域热门商品统计未配置 | |
| others | 其它(数据准备等非功能目录) | | | |
| prepare\_all.sh | | 原始数据准备脚本 | |
| prepare\_data\_sql.sql | | 原始数据准备SQL | |
| user\_click.txt | | 用户行为记录20160505 | |
| clear\_all.sh | | 清理所有所有先关数据(包括原始数据) | |
| doc | 描述文档 | | | |

1. **实现过程**
   1. **数据仓库初始化脚本**

|  |  |
| --- | --- |
| 脚本 | /bin/init.sh |
| 概述 | Hive数据仓库的建立.并在数据仓库中建立四张承载表.   * 产品信息ruoze.product\_info * 城市信息ruoze.city\_info * 用户行为记录 ruoze.user\_click * 日分区热门商品排名结果 ruoze.hot\_product |
| 使用 | ./bin/init.sh |
| 参数 | 无 |
| 代码  预览 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\EGZ$35]LVP]L_9_3S1S0{YK.png |

* 1. **基础数据抽取脚本**

|  |  |
| --- | --- |
| 脚本 | /bin/init.sh |
| 概述 | 用Sqoop,从不同的MySQL库中抽取本功能需要的产品信息和城市信息.   * 城市信息来自MySQL.ruoze\_basedata库 * 产品信息来自MySQL.ruoze\_product库 |
| 使用 | ./bin/init.sh |
| 参数 | 无 |
| 脚本预览 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\O$S$IA`OLU`9K_PP3N4Q~F4.png |

* 1. **数据清洗脚本**

.

|  |  |
| --- | --- |
| 脚本 | /bin/trans\_day.sh |
| 概述 | * 按日导入用户行为记录. * 清理该日的目的端MySQL结果数据 * 统计该日区域商品排名结果 * 统计结果推送到目的端MySQL |
| 使用 | ./bin/trans\_day.sh –day 20160505  ./bin/trans\_day.sh –daysection 20160504 20160507 |
| 参数 | -day 统计日期(格式:yyyyMMdd)  -daysection 统计起始时间(格式:yyyyMMdd) 统计结束时间(格式:yyyyMMdd) |
| 代码预览 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\2}OKOQNEB%}D~BO7G2WZA4G.png  C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\WH1NG]5W$ZIEE96~~SXIAXI.png |

* 1. **其它脚本**
     1. **初始数据准备**

|  |  |
| --- | --- |
| 脚本 | /others/prepare\_all.sh |
| 概述 | 用于本功能之外的模拟数据准备,这些模拟的原始数据准备不视为本功能的一部分,在实际生产中,这些原始数据是由其它小组或其它部门直接给出的   * MySQL中建ruo\_product库,建表product\_info并写入一些模拟数据 * MySQL中建ruo\_basedata库,建表city\_info并写入一些模拟数据 * HDFS上放入用户20160505行为记录 user\_click.txt * Mysql中建ruoze库,建表hot\_product作为统计结果的目的端 |
| 使用 | 无 |
| 参数 | 无 |
| 脚本预览 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\G)7K2CG96DBY[0$17(CLHXD.png |

* + 1. **清理全部数据**

|  |  |
| --- | --- |
| 脚本 | /others/clear\_all.sh |
| 概述 | 为从头演示方便,删除本功能使用到的全部数据(包括原始数据)   * MySQL删除ruo\_product库, ruo\_basedata库, ruoze库 * HDFS删除user\_click.txt * Hive删除 ruoze库 |
| 使用 | 无 |
| 参数 | 无 |
| 脚本预览 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\PYFFI88{T`_2ULBW6T3ZR)R.png |

* + 1. **Shell基础库**

.

|  |  |
| --- | --- |
| 脚本 | /bin/shell\_lib/commmon.sh |
| 概述 | 封装Shell中用到一些基本函数   * 格式化提示消息   *print\_msg 消息内容 [-err | -waring | -info]*   * 读取配置文件信息   *read\_conf 配置文件地址 配置节名* |
| 使用 | 无 |
| 参数 | 无 |
| 脚本预览 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\XXP$M1@`Y9@F3DUB`$D~}B0.png |

1. **运行过程**
   * 1. **原始数据**

|  |  |
| --- | --- |
| 执行 | ./others/prepare\_all.sh |
| 产品信息 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\3]D2`@T$_DQJMUI06A(P16B.png |
| 城市信息 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\UDSDNMC%S2J8_64SZ8$8@NJ.png |
| 用户行为 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\YXO}W_WZA@W)SG3}EIS(]FF.png |

* + 1. **初始化数据仓库**

.

|  |  |
| --- | --- |
| 执行 | ./bin/init.sh |
| 执行过程 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\_MLOMM32Y{FKKC[E7G0~L]J.png  C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\N3I(}2YZN(4`)%BG%`KUQFN.png |
| 执行结果商品表 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\PSA0)[$8_J4_9GOR}U8@PSQ.png |
| 执行结果城市表 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\BB}FS`)$W~E@YWJE]OGM7%6.png |
| 执行结果用户行为表 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\[MT@COPASMR@XK_}MU~~A4I.png |
| 执行过节热门商品统计结果 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\]96{E$O2N@GPV~Q`8{)V5BF.png |

* + 1. **基础导入**

|  |  |
| --- | --- |
| 执行 | ./bin/extact\_data.sh |
| 执行过程 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\6UA{%FI96GOSYTH08%%TF9O.png |
| 执行结果商品表 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\CI{%8RRQ_3]9JJ8GS{G}Z2E.png |
| 执行结果城市表 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\WSP6TP$Y@V]_GUH)V83GK9C.png |

* + 1. **数据清洗-20160505**

|  |  |
| --- | --- |
| 执行 | ./bin/trans\_day.sh -day 20160505 |
| 执行过程 | C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\[]6L[Q`]@M05RHO52%K0X{C.png  C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\A3K{4M[Z(HNQE$0{22[EG_2.png  C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\V0O2@N2EPVZSEEKB1@GUL0G.png |

* + 1. **目的端结果**

|  |
| --- |
| C:\Users\Night\AppData\Roaming\Tencent\Users\38695281\QQ\WinTemp\RichOle\C5)WF4}B4[VAV5)MA)F0[V7.png |

数据共计21条.

发现没有NE区的用户行为记录,所以统计结果只有6个区,没有NE区的

但实际计算排名时,某些地区有并列的情况,所以某些区实际会出4条记录.