一、数据清洗 nineninehome-analyze-transform

(数据清洗为MapReduce任务)

输入

上传到hdfs的接收到的原始数据

输出

清洗后的数据

执行频率

每小时执行一次

执行时间

2min左右

二、数据导入 LoadCleanedData

输入

清洗后的数据

输出

将清洗后的数据导入到hive表：

- startup_log_2015022700
- valid_log_2015022700

执行频率

每小时执行一次

执行时间

1min左右

三、数据处理 LogsProcess

输入

清洗后的数据：

- startup_log_2015022700
- valid_log_2015022700

输出

- startup_log_imei_20150227（每小时新增）
- valid_log_imei_20150227（每小时新增）
- startup_log_uid_20150227（每小时新增）
- valid_log_uid_20150227（每小时新增）
- new_imei（每小时新增）
- new_uid（每小时新增）

执行频率

每小时执行一次

执行时间

5min左右

依赖关系

- 依赖于数据清洗

四、备份

输入

- new_imei
- new_uid

输出

- new_imei_2015022723
- new_uid_2015022723

执行频率

每天执行一次

执行时间

2min左右

依赖关系

- 每天的数据处理结束后，下一天的数据处理开始前，进行备份

五、数据统计

1) 第一级结果表

1. 按小时统计 LoginLogStatHour

输入

- startup_log_imei_20150227
- valid_log_imei_20150227
- startup_log_uid_20150227
- valid_log_uid_20150227

输出

- login_log_stat_hour_2015022700

执行频率

每小时执行一次

执行时间

5min左右

依赖关系

- 某个小时的数据处理结束之后，才可以开始该小时的LoginLogStatHour统计

2. 按天统计 LoginLogStatDay

输入

- new_imei_2015022723
- new_uid_2015022723

输出

- login_log_stat_day_20150227

执行频率

每天执行一次

执行时间

3min左右

依赖关系

- 某一天的数据处理全部结束并做备份之后，开始该天的LoginLogStatDay统计

3. 次日留存统计 RetentionStat

输入

- login_log_stat_day_20150227
- login_log_stat_day_20150226（计算20150226的次日留存）

输出

- retention_20150226

执行频率

每天执行一次

执行时间

3min左右

依赖关系

- 次日留存为隔天统计，即今天统计昨天的次日留存
- 当天的**LoginLogStatDay**统计完毕之后，才可以进行前一天的RetentionStat的统计

2) 第二级结果表

1. DayOverview

输入

- login_log_stat_hour_20150227
- login_log_stat_day_20150227

输出

- day_overview_2015022700

执行频率

每小时执行一次

执行时间

1min左右

2. TimeAnalyze

输入

- login_log_stat_hour_20150227

输出

- time_analyze_2015022700

执行频率

每小时执行一次

执行时间

1min左右

3. ProductSummary

输入

- login_log_stat_hour_20150227
- login_log_stat_day_20150227

输出

- product_summary_20150227

执行频率

每天执行一次

执行时间

1min左右

4. ActiveOverview

输入

- login_log_stat_hour_20150227
- login_log_stat_day_20150227
- retention_20150227

输出

- active_overview_20150227

执行频率

每小时执行一次

执行时间

1min左右

5. ChannelRank

输入

- login_log_stat_hour_20150227

输出

- channel_rank_20150227

执行频率

每天执行一次

执行时间

1min左右

6. DataTrend

输入

- login_log_stat_hour_20150227
- login_log_stat_day_20150227

输出

- data_trend_20150227

执行频率

每天执行一次

执行时间

1min左右

7. UserStartupOverview

输入

- login_log_stat_hour_20150227
- login_log_stat_day_20150227

输出

- user_startup_overview_20150227

执行频率

每天执行一次

执行时间

1min左右

8. ActiveAnalyze

输入

- login_log_stat_hour_20150227
- login_log_stat_day_20150227

输出

- active_analyze_20150227

执行频率

每天执行一次

执行时间

1min左右

9. RegionAnalyze

输入

- startup_log_imei_20150227
- startup_log_uid_20150227

输出

- region_analyze_20150227

执行频率

每天执行一次

执行时间

2min左右

10. ChannelAnalyze

输入

- startup_log_imei_20150227
- startup_log_uid_20150227

输出

- channel_analyze_20150227

执行频率

每天执行一次

执行时间

2min左右

第二级结果表的依赖关系

- 第二级结果表之间都没有相互的依赖，可以同时进行统计
- 全部第二级结果表都必须要在第一级结果表统计完成之后才可以进行