日志

目录

[Overview 5](#_Toc492886802)

[2017 6](#_Toc492886803)

[Week\_37 6](#_Toc492886804)

[Plan 6](#_Toc492886805)

[Require 6](#_Toc492886806)

[2017-09-11 6](#_Toc492886807)

[2017-09-12 6](#_Toc492886808)

[2017-09-13 6](#_Toc492886809)

[2017-09-14 6](#_Toc492886810)

[2017-09-15 6](#_Toc492886811)

[2017-09-16 6](#_Toc492886812)

[2017-09-17 6](#_Toc492886813)

[Week\_36 7](#_Toc492886814)

[Plan 7](#_Toc492886815)

[Require 7](#_Toc492886816)

[2017-09-04 7](#_Toc492886817)

[2017-09-05 7](#_Toc492886818)

[2017-09-06 7](#_Toc492886819)

[2017-09-07 8](#_Toc492886820)

[2017-09-08 9](#_Toc492886821)

[2017-09-09 9](#_Toc492886822)

[2017-09-10 9](#_Toc492886823)

[Week\_37 10](#_Toc492886824)

[引用 12](#_Toc492886825)

[附录 13](#_Toc492886826)

[0000-年模板 14](#_Toc492886827)

[Week\_00-周模板 14](#_Toc492886828)

[Plan 14](#_Toc492886829)

[Require 14](#_Toc492886830)

[0000-01-01 14](#_Toc492886831)

[0000-01-02 14](#_Toc492886832)

[0000-01-03 14](#_Toc492886833)

[0000-01-04 14](#_Toc492886834)

[0000-01-05 14](#_Toc492886835)

[0000-01-06 14](#_Toc492886836)

[0000-01-07 15](#_Toc492886837)

# Overview

# 2017

## Week\_37

### Plan

|  |
| --- |
|  |

### Require

|  |
| --- |
| 1. 完成数据资产组0.7版本的需求和开发计划确定 2. 评估人脸和车辆系统接口可用性，形成技术方案 |

* V0.7需求及开发计划
  + 缺陷修复、代码优化
  + 基于spark-sql开发
    - Spark-sql数据仓库部署及测试
    - 采集、转换清洗、融合基于spark-sql dataframe开发
    - 与数据使用方老葛协商spark-sql使用
    - ETL链路日志重构（data\_version\_load、etl\_core）
    - 原有hdfs数据过渡到spark-sql方案设计与开发
    - 数据质量与大数据公共模块基于spark-sql dataframe开发
    - 上述模块设计及测试
  + 增量table数据采集重构及任务调度集成
    - 增量table数据采集功能重构设计
    - 增量table数据采集功能重构开发
    - 增量table数据采集功能与任务调度集成
    - 增量table数据采集功能与任务调度集成后功能测试
  + 5s人脸及车辆数据接入
* Face\traffic 技术方案
  + 评估人脸和车辆系统接口可用性

### 2017-09-11

1. 深入学习spark-sql技术
   1. 学习spark-sql function模块，了解spark-sql具体有哪些原生function
   2. 学习spark-sql2.2 Cost-Based Optimizer核心概要模型
      1. 通过计算table、column的条数、大小结合计算链路来计算cost

### 2017-09-12

1. 讨论数据资产组内v0.7的工作计划
   1. 具体需要做哪些功能
      1. 具体我和张思为先整理，然后还需要做的邱工补充
   2. 我与张思为分别做哪些功能
   3. 确定v0.7的开发计划任务

### 2017-09-13

1. Intellij 文件头注释模板统一确立
2. Scala 代码单元测试学习及简单实际应用
   * 1. 采用FunSuite风格

数据质量--识别源数据更新频率

data\_load\_pioneer 数据摸底表映射及操作实现

表结构设计

大数据公共模块：

Hdfs 增删封装--hdfs操作核心方法封装

### 2017-09-14

1. 和大家一起初步制定数据资产组v0.7版本的开发计划
   1. 具体任务、完成时间及开发人员
2. 学习spark技术
   1. 阅读spark 新增主推功能structure streaming官方文档

备注：

经与老葛沟通，数据放到spark-sql数据仓库，他那边没技术风险；但hdfs上的text数据v0.7暂时保留；他那边v0.7暂时不把spark-sql列入工作计划

王圣沟通算法那边使用车辆与人脸的数据，由王辉自己获取只是权宜之计；

车辆与人脸的数据获取从长期来看还得数据资产组实现

V0.7 的相应工作计划得排，但不安排测试

Question：人脸底层存储的是照片，没有人员信息主键的话？如何使用？

即：一旦抓拍库数据被采集到hbase（本来就是不准确的？）；我们自己如何通过一个人去拿一个人的所有抓拍轨迹数据？算法如何进行使用

数据质量--识别源数据更新频率；定时任务调度实现（采用quartz实现）并测试

### 2017-09-15

### 2017-09-16

### 2017-09-17

## Week\_36

### Plan

|  |
| --- |
|  |

### Require

|  |
| --- |
|  |

### 2017-09-04

### 2017-09-05

1. 部署solr集群
   1. 安装solr集群
   2. 更改solr相关配置并重启生效
   3. 上传solr配置文件并创建zsmap[百度转换的经纬度数据集]、mapzs[poi数据集] 的collection
   4. 将poi数据集写入mapzs collection
2. 数据资产组转换清洗及融合升级
   1. 备份元数据库、清除相关hdfs数据及元数据库表数据及更改相关配置
   2. 更新统一视图并执行转换清洗任务[第二天查看，运行成功(3.5h)]
3. 问题：
   1. 算法python导数据程序内存溢出异常
      1. 通过更改spark-submit参数予以解决

### 2017-09-06

1. 经纬度数据写入solr
   1. 将百度转换的经纬度数据集写入zsmap collection
      1. 问题：程序报 not Serialization Exception
         1. 排查分析：本地运行正常，线上异常；推测为相关maven jar依赖未解决
         2. 解决：笔记本电脑联网解决maven依赖并重新打包
2. 执行融合任务
   1. 执行成功，运行3.5小时
3. 转换清洗、融合结果数据写入spark-sql数据仓库
   1. 删除vulcanus\_core、vulcanus\_merge相关表
   2. 写入转换清洗相关表数据
   3. 写入融合结果相关表数据
4. 执行李金龙vulcanus\_load(贴源层) sql统计语句
5. 帮助葛挺导出vulcanus\_load部分表数据并加密发给葛挺
6. Yarn上自带job，杀不死并且越杀越多
   1. 通过咨询组件组提供解决予以解决[不确定机器重启之后该问题是否重现]

**出差总结：**

* 1. 案件地图poi、百度转换经纬度数据集均按设计导入solr搜索引擎
  2. 数据资产组转换清洗、融合按设计进行升级
  3. 上述出差目标均已达到，并按计划完成

### 2017-09-07

1. 整理出差记录文档
2. 协助算法解决215个卡口7亿多过车数据提取困难
   1. 向邢金彪索要卡口id列表文件
   2. 根据实际需求生成抽数据spark-sql语句
   3. 使用说明：
      1. 
      2. columns 全部替换为实际需要的 列名并逗号分隔
      3. conditions 全部替换为实际需要的where 条件语句 and连接
3. 协助测试理解采集增量实现
4. 表conf\_dict ：为配置字典，
   1. 记录：
   2. 即表示自增规则已配置
5. 表 conf\_dict\_elem ：为字典规则应用
   1. `conf\_dict\_elem\_id` int(11) NOT NULL AUTO\_INCREMENT,
   2. `conf\_dict\_id` 外键 REFERENCES `conf\_dict` (`conf\_dict\_id`)
   3. `elem\_key` 表id
   4. `elem\_value` 属性id
6. 增量实现：在表conf\_dict\_elem中插入记录即可；
   1. Eg：
      1. 解释：即为id=1的源表的attr\_id=1的属性为增量字段
7. 测试问题记录，表属性类型为rowid的采集失败
   1. 已分析(通过走读源码)并测试为rowid底层解析为long，理应解析为string
   2. 解决：通过重写该代码予以解决
   3. 

### 2017-09-08

专利文档书写

1. 修复增量采集数据类型支持bug
   1. 丰富增量字段数据类型支持数学、string、日期等
   2. 测试增量字段数据类型（数学、string、日期）实现等功能[测试通过]
2. 组内工作：书写自己负责模块文档
   1. 学习github wiki使用
   2. 书写 数据质量模块文档

### 2017-09-09

自己负责模块文档书写

专利文档书写

1. 继续书写自己负责模块文档
   1. 书写数据仓库资料文档
      1. 数据仓库部署
      2. 数据仓库访问
      3. 使用数据仓库遇到的问题及解决方法
   2. 案件地图地址转经纬度文档
   3. 案件地图solr搜索引擎部署文档

### 2017-09-10

## Week\_37

# 引用

# 附录

# 0000-年模板

## Week\_00-周模板

### Plan

|  |
| --- |
|  |

### Require

|  |
| --- |
|  |

### 0000-01-01

### 0000-01-02

### 0000-01-03

### 0000-01-04

### 0000-01-05

### 0000-01-06

### 0000-01-07