基于地铁出行行程的大数据实时分析预测系统设计

团队编号:2002066　二仙桥老大爷

目录

1.引言

1.1 需求

在新一轮的科技革命和产业变革的浪潮推动下,近些年我国的城轨交通行业信息化建设步入了快速的发展阶段,信息化建设的成果初具规模,改变了传统的建设模式、服务手段和经营方式。为各个相关部门提供科学的数据,并且能够有效的分配资源和人力,提高整个交通系统的安全性、舒适性和经济效益。能够为有关部门处理紧急突发事件提供有效的数据支持和决策依据,尤其是在组织大型活动时、客流量的预测能够帮助轨道交通运营单位做好相应乘客运输能力的调整匹配,既能够保证活动的顺利进行也能够减少对其他居民的影响。

1.2 现状

随着国内各个城市轨道交通持续的高速发展,轨交乘客数量不断增长,在缓解城市整体交通拥堵的同时,轨道交通本身也面临这较大的客流管理压力,在数据层面主要普遍存在着以下几个问题:

(1)客流监测能力不足,当前的轨交 AFC 系统只能提供乘客进出站信息,而乘客的出行过程信息是缺失的,导致行业方无法全面的掌握路网内的客流分布和动态,对存在的站点、线路、车厢拥堵感知滞后。

(2)缺乏客流精准管控方法,没有对大客流条件下的高风险客流聚集点的精准管理和控制方法。

(3)缺乏对突发客流的提前评估和预测,各种预案的实施较为被动,

无法及时并且预见性的缓解可能出现的突发客流事件,确保正常客运不受影响。

1.3 目标

提供友好的用户交互方式，通过调整模型的各种相关因子，对指定时间、指定线路或者站点的客流进行预测和预警并且通过图形化的方式直观展现。

1.4 计划

该系统的网页端采用 Echarts 和 Bootstrap。数据处理层采用 Spark 和 Flink，为系统实时计算提供保障。数据存储层采用 Kafka、MySQL 和 Hive，Kafka 是基于 zookeeper 协调的分布式消息系统，它的最大的特性就是可以实时的处理大量数据以满足各种需求场景：比如基于 hadoop 的批处理系统、低延迟的实时系统、storm/Spark 流式处理引擎，web/nginx 日志、访问日志，消息服务等等。数据采集层采用 Flume，Flume 是一种分布式，可靠且可用的服务，用于高效地收集，汇总和移动大量日志数据，它具有可靠的可靠性机制以及许多故障转移和恢复机制，具有强大的容错性和容错能力。

1.5 创新

该系统能准确预测地铁人流量，方便管理员对地铁及其相关人员进行调度。增强客流量监控，提供乘客出行过程信息。

1.6 优势

功能强大：该系统功能包括单月整体的客流波动分析、工作人和周末的客流分析、单站的点出入站客流量分析、用户年龄结构分析、

早晚高峰客流站点发布分析、站点 OD 客流量分析、路线断面流量分析和天气预测。

模型精度高：训练神经网络的数据集庞大，模型精度在 90%以上。

响应政策：该系统响应《中国城市轨道交通智慧城轨发展纲要》，符合刚要中提出的重点建设方向。

2. 系统综述

2.1 系统结构

基于地铁出行行程的大数据实时分析预测系统 v1.0 由数据存储系统、spark streaming 实时计算系统、LSTM_Prophet 组合预测系统以及智能展示终端组成，基于 Internet 网络进行连接。

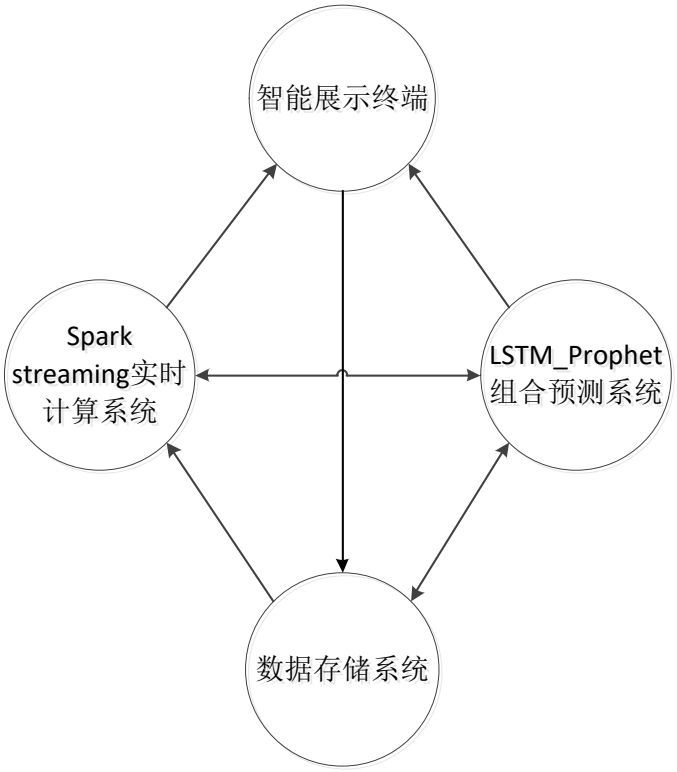


图 1 系统总体结构图

2.2 系统功能简介

数据存储系统将地铁乘客刷卡进出各地铁站的信息、乘客个人信息、

节假日信息及气象信息存入Hbase非关系型数据库中，将地铁工作人员信息存入Mysql关系型数据库中，spark streaming实时计算系统获取数据库存储系统中现有的乘客信息以及气象信息进行统计分析并与智能展示终端交互进行可视化展示，同时，LSTM_Porphet组合预测系统获取数据库系统中的时序数据，在经过训练预测后，将预测结果直接传输给智能展示终端进行可视化展示的同时将预测结果存入数据存储系统中去，智能展示终端在可视化展示时，并进行实时预警，将预警信息存入数据存储系统中去，当预测数据量过多时，也可将预测数据交由spark streaming实时计算系统，再由其传输给智能展示终端。此后，预测系统每隔一定时间就会获取spark streaming实时计算系统统计好的实时时序数据进行误差分析，发现预测数据的误差较大时，预测系统会再次获取数据库系统中的时序数据进行模型训练，更新权重，确保系统预测误差在可许范围内。

2.3 技术特点

（1）基于Hadoop、Spark streaming高可靠分布式集群实现地铁乘客信息及气象信息的实时数据分析和历史数据分析。

（2）采用LSTM_Prophet组合模型既支持短期预测也支持长期预测，在考虑节假日、气象信息等因素时也能很好的达到预测效果。

（3）利用 springboot 后端框架对数据进行存取，数据库采用关系型数据库MySQL和非关系型数据库Hbase。

（4）前端采用Echarts框架实现数据可视化。

3.运行环境

3.1 硬件设备

客户终端:

CPU 性能: ≥Pentium 2.0GHZ

内存: ≥512M

硬盘: ≥40G

服务器（5 台）:

CPU: ≥ 16CPU*4 核

内存: ≥ 32G DDR4

硬盘: ≥ Sata 8*4T

网卡: ≥ 1G

3.2 支持软件

终端: Windows98/XP/Win7/VISATA/Win10/Android/ios

服务器:

操作系统: centos7/ubuntu

数据库: Habse1.2.3 和 MySQL5.7

3.3 数据结构

本平台采用的 关系型数据库 Hbase1.2.3 和关系型数据库 MySQL5.7, 存储文件支持指定云存储系统。

4.系统操作说明

4.1 系统搭建

4.1.3 k8s 与 Docker 容器

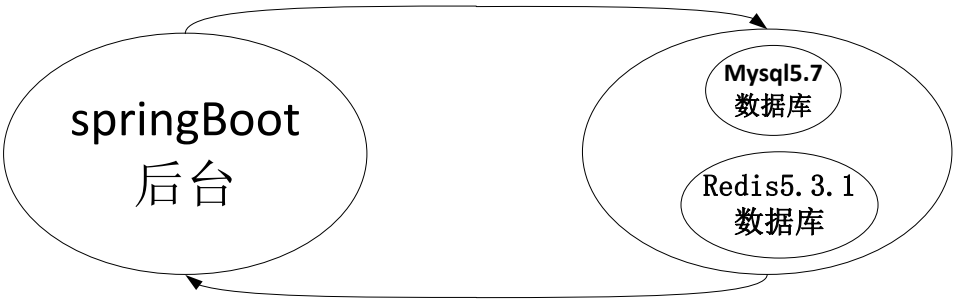
系统采用 kubernetes(以下简称 k8s)对 Docker 容器进行管理操作。对于 k8s 是一个开源的容器集群管理系统，可以实现容器集群的自动化部署、自动扩缩容、维护等功能。可以大大节省开发维护的时间，提高开发效率。k8s 可以虚拟出很多独立的虚拟机，通过这些虚拟机与 docker 打包成一个独立环境来运行程序，节省很多存储陈本和兼容调配成本。

由于 Docker 在系统中只是一个进程，只需要将应用以及相关的组件打包，在运行时占用很少的资源，在一定程度上可以节省宝贵的服务器资源消耗。并且，团队在进行相关的组件部署时可能会遇到千奇百怪的环境上的问题，使用 Docker 容器让各个环境隔离开，可以解决这一问题，大大提交开发效率

4.1.2 springBoot 数据存储系统

这种映射机制是双向的,当向数据库存入数据时,是将 java 对象映射为数据库对象,而从数据库取出数据时,却将数据库中的数据还原为 java 对象。关系型数据库都使用了 JPA 的一套执行标准,他结合 使用 Mybatis 实现了实体的持久化.后续的数据库管理设计都遵循 了 JPA 这一个标准规范,提供了相同的访问数据库的 API。

Java 对象映射为数据库对象



数据库对象还原为 Java 对象

4.1.3 LSTM_Prophet 组合预测模型

本模型时序数据由某城市地铁乘客行程数据约 80 万条、乘客信息数据约 12 万条、站点名称数据以及节假日数据组成，并将这些数据划分为训练集和测试集，为了提高模型的预测精度并加快收敛速度，需要对训练集和测试集进行归一化处理和反归一化处理。使用 LSTM 模型对数据集进行拟合，设定 batch_size 和 epochs 的可能取值，并由 Grid_Search 对其进行选择，使用均方误差作为损失函数，其值收敛至几乎无变化时，表示使用该组参数的模型最优。使用 Prophet 模型对数据集进行拟合，设置模型的 changepoint_prio_scale 值为 0.2，seasonality_mode 为乘法模型，n_change points 值为 30，yearly——seasonality 为 outo，设置这些参数以后可以提高模型拟合的灵活性，增加曲线的转折点数量，进而提高了模型拟合精度，然后将数据集导入 Prophet 模型中进行拟合，从而得到待预测日期数据的预测结果和训练集的预测结果。最后使用 BP 神经网络，将训练集的 LSTM 和 Prophet 模型的预测结果作为输入，真实值作为输出，对 BP 神经网络模型进行训练拟合，通过模型训练，由 BP 神经网络确定 2 个模型的预测值在组合中的权重，然后将测试集的 LSTM 和 Prophet 的预测结

果作为输入，并加入气象因素，输出通过 BP 神经网络非线性组合后的预测值，并与真实值进行比较，从而判断组合模型的预测效果。

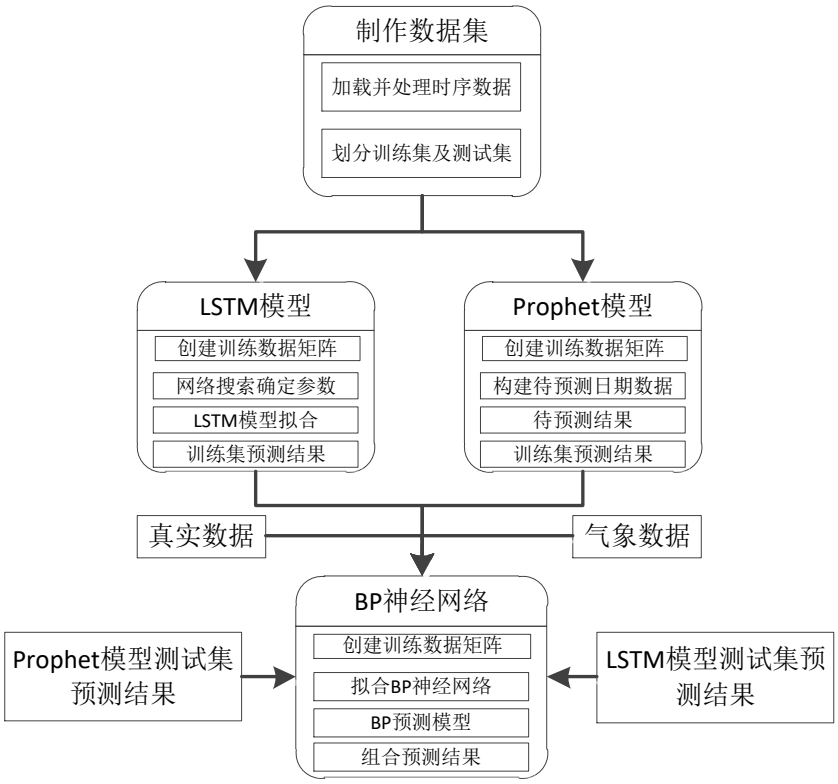


图 2 LSTM_Prophet 组合预测流程

4.1.4 spark 与 Hadoop 集群框架

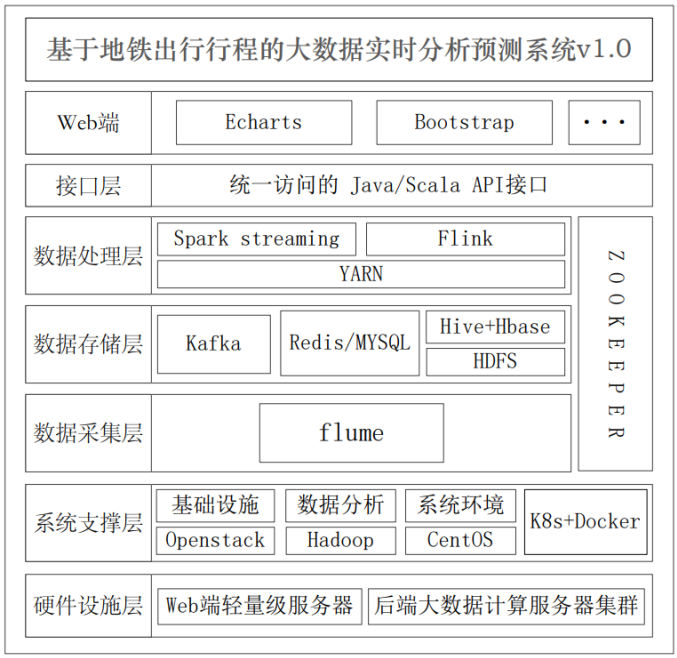
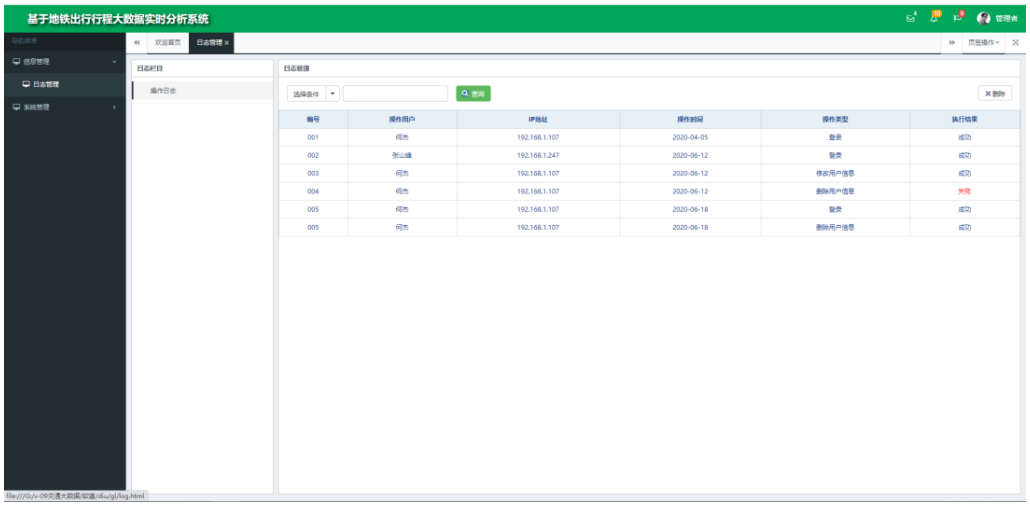


图 3 大数据平台总体架构

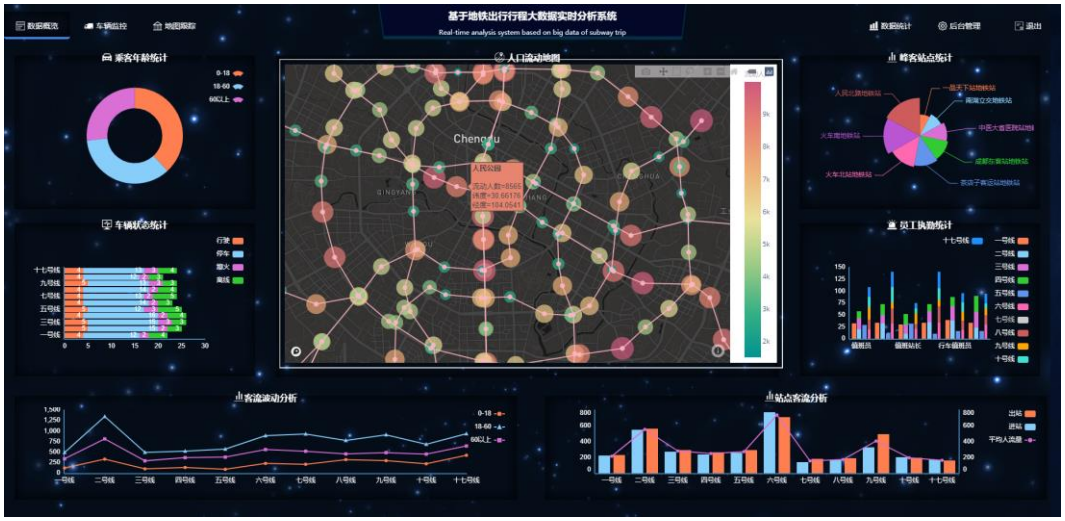
4.2 管理员登录

进入系统后台会有一个登录界面，我们用管理员账号登录进入 系统，然后可以对里面的数据进行操作。

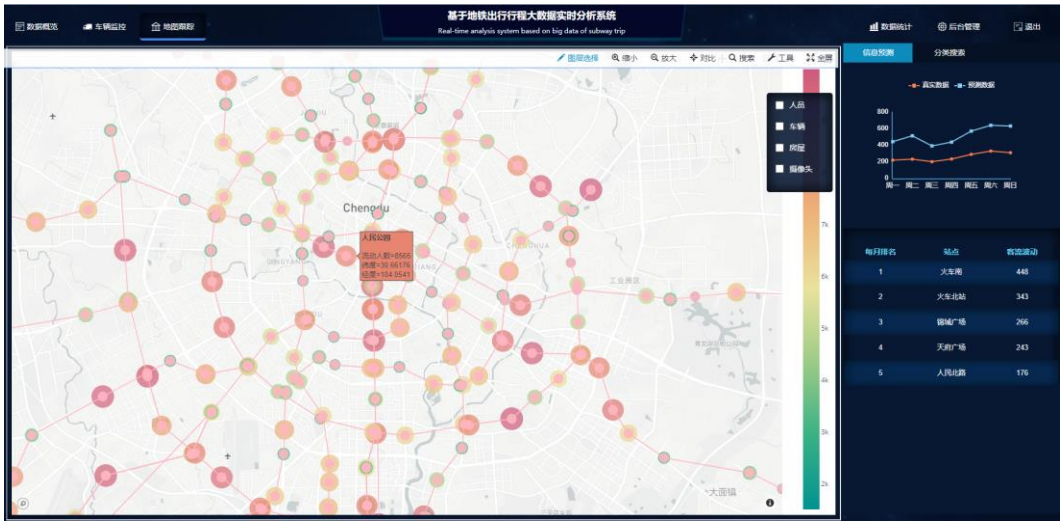


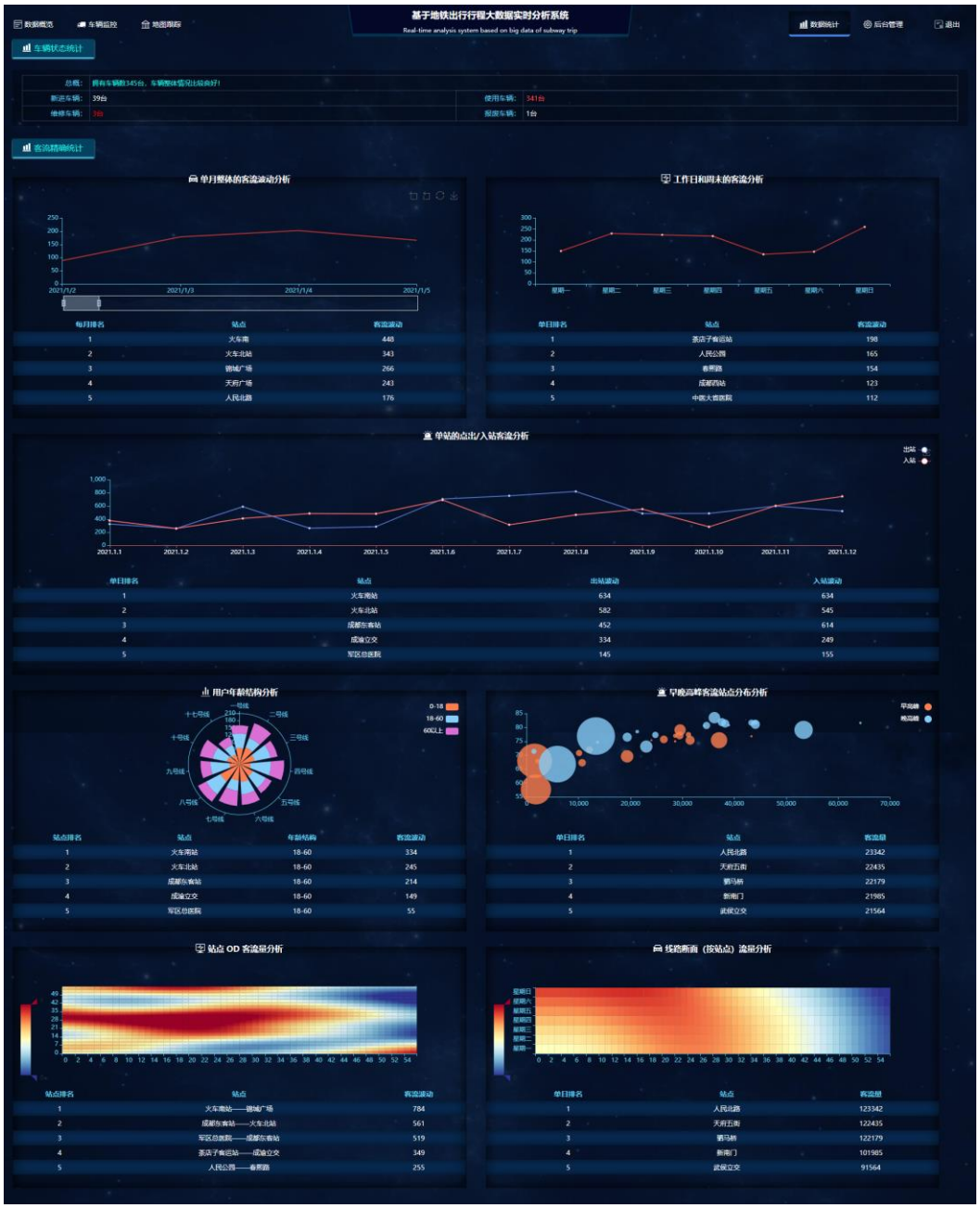
4.3 平台管理

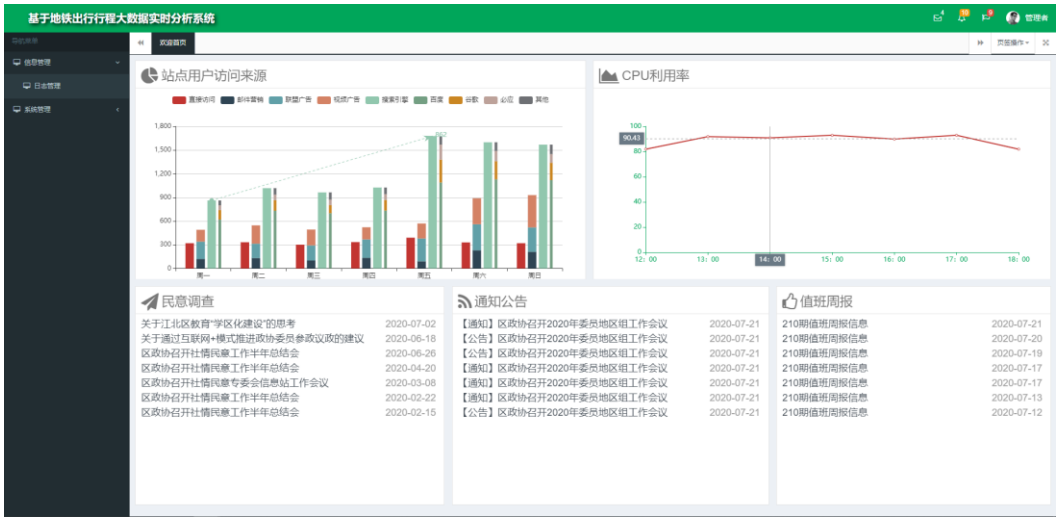
4.3.1 系统的运行

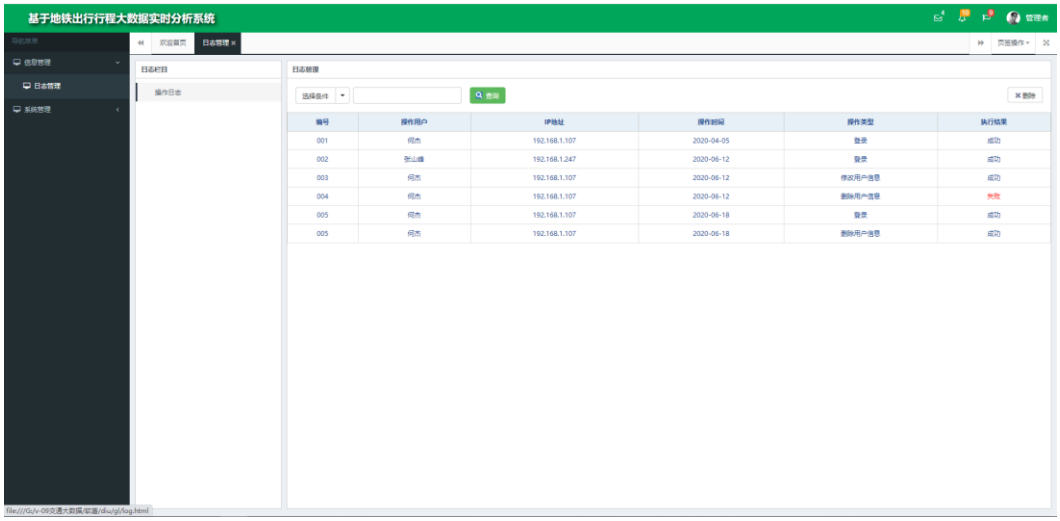

基于地铁出行行程大数据实时分析系统
Real-time analysis system based on big data of subway trip
乘客年龄统计
人口流动地图
峰客站点统计
车辆状态统计
员工轨迹统计
客流波动分析
站点客流分析

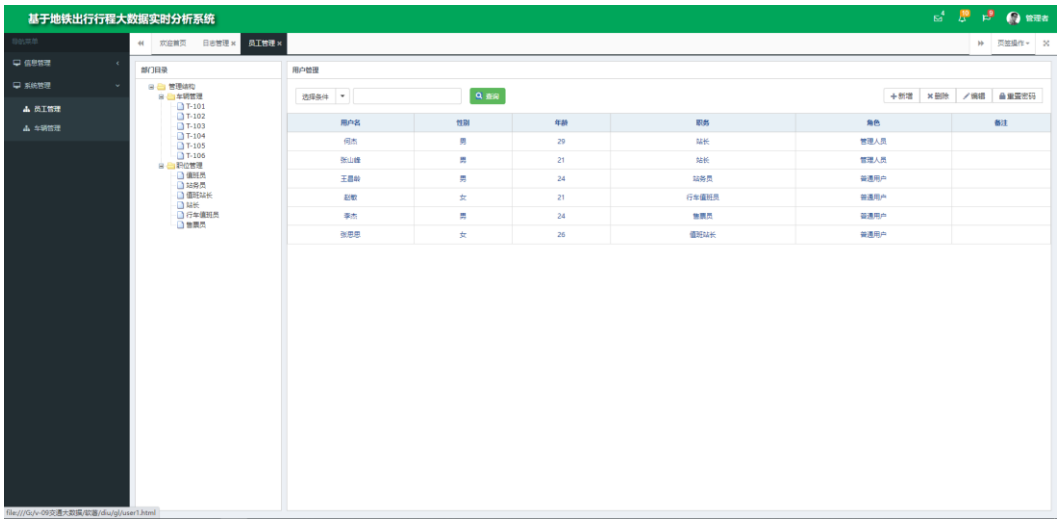

基于地铁出行行程大数据实时分析系统
Real-time analysis system based on big data of subway trip
温度统计:
基本信息
车牌号：B-431
驾驶人姓名：张章忠
行驶里程：27965
状态：形驶
所在位置：成都火车南-地铁站
热烈庆祝监控系统进入调试阶段
实时天气
温度：22 ℃
降雨：0 h

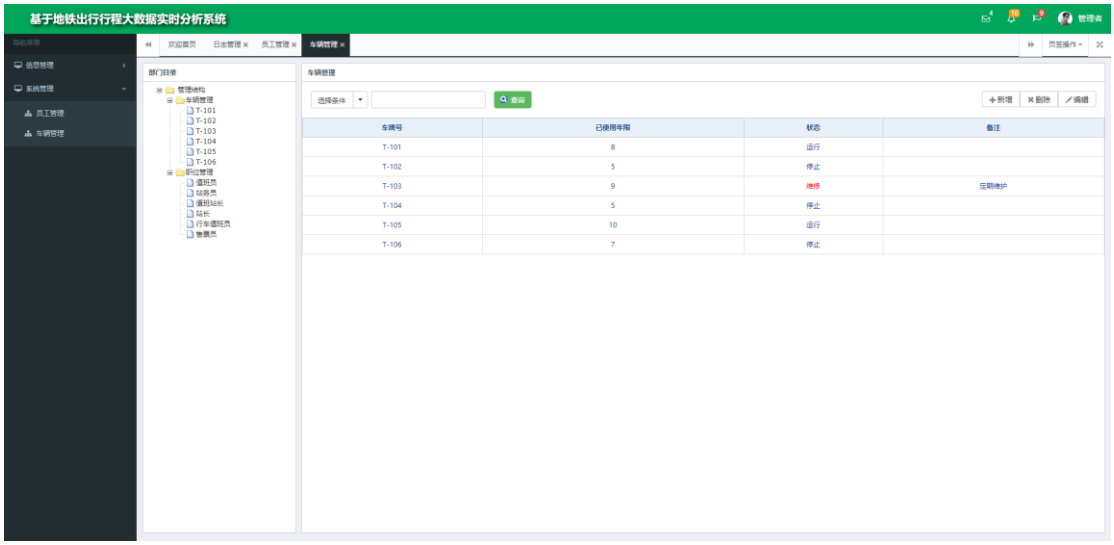

基于地铁出行行程大数据实时分析系统
Real-time analysis system based on big data of subway trip
人员
车辆
房屋
摄像头


基于地铁出行行程大数据实时分析系统
Real-time analysis system based on big data of subway trip
数据概览
车辆监控
地图模拟
数据统计
后台管理
退出
车辆状态统计
总概：拥有车辆数345台，车辆整体情况比较良好！
新运车辆：39台
使用车辆：341台
维修车辆：3台
报废车辆：1台
客流精确统计
单月整体的客流波动分析
每月排名 站点 客流波动
1 火车南 448
2 火车北站 343
3 锦城广场 266
4 天府广场 243
5 人民北路 176
工作日和周末的客流分析
单日排名 站点 客流波动
1 圣洁子食运站 198
2 人民公园 165
3 春熙路 154
4 成都西站 123
5 中医大省医院 112
单站的点出/入站客流分析
出站
入站
单日排名 站点 出站波动 入站波动
1 火车南站 634 634
2 火车北站 582 545
3 成都东客站 452 614
4 成渝立交 334 249
5 军区总医院 145 155
用户年龄结构分析
0-18
18-60
60以上
站点排名 站点 年龄结构 客流波动
1 火车南站 18-60 334
2 火车北站 18-60 245
3 成都东客站 18-60 214
4 成渝立交 18-60 149
5 军区总医院 18-60 55
早晚高峰客流站点分布分析
早高峰
晚高峰
单日排名 站点 客流量
1 人民北路 23342
2 天府五街 22435
3 锦马路 22179
4 新南门 21985
5 武侯立交 21564
站点 OD 客流量分析
站点排名 站点 客流波动
1 火车南站——锦城广场 784
2 成都东客站——火车北站 561
3 军区总医院——成都东客站 519
4 圣洁子食运站——成渝立交 349
5 人民公园——春熙路 255
线路断面（按站点）流量分析
单日排名 站点 客流量
1 人民北路 123342
2 天府五街 122435
3 锦马路 122179
4 新南门 101985
5 武侯立交 91564


基于地铁出行行程大数据实时分析系统
站点用户访问来源
CPU利用率
民意调查
通知公告
值班周报

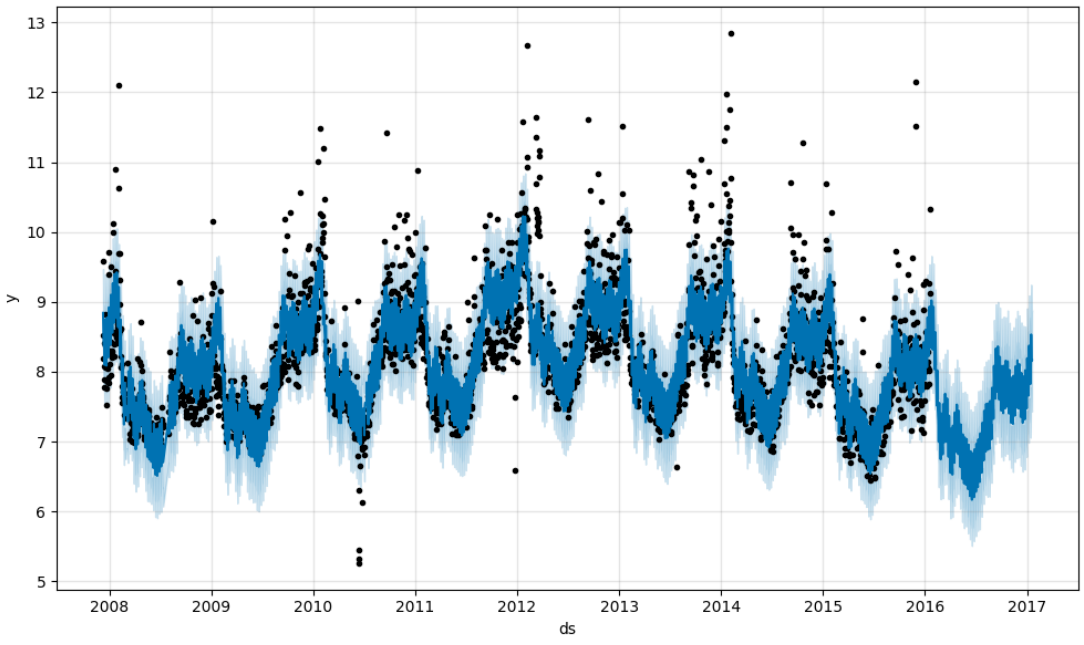

基于地铁出行行程大数据实时分析系统

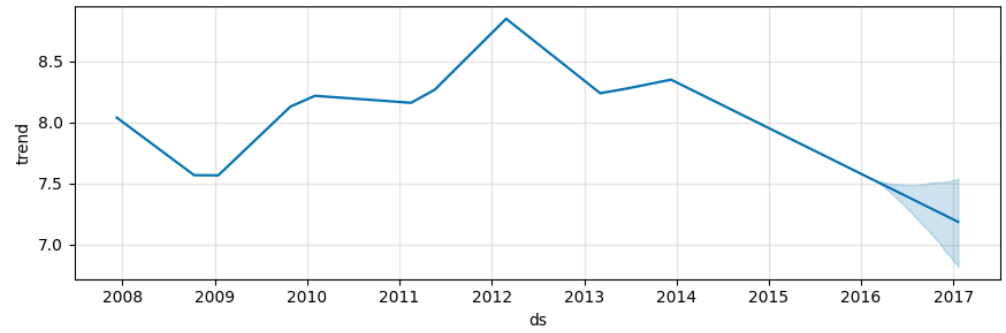

基于地铁出行行程大数据实时分析系统

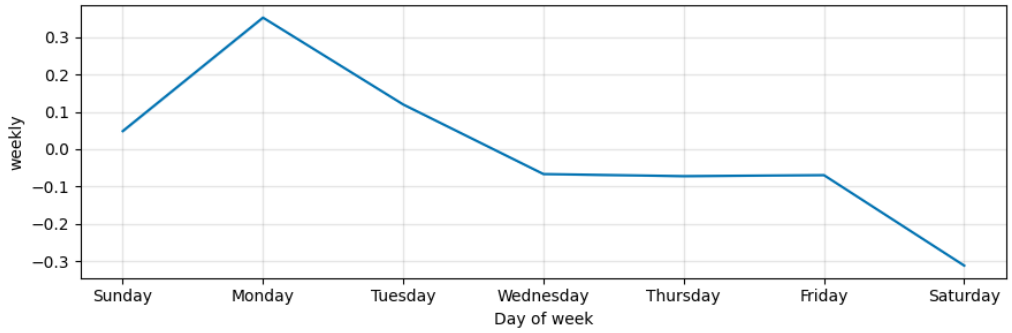


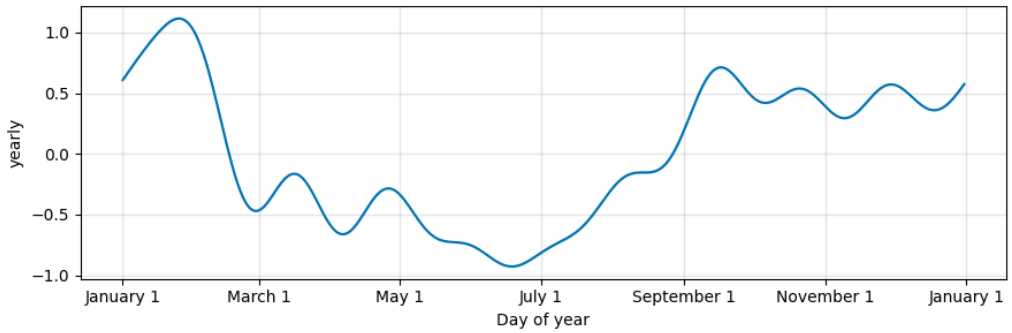
4.3.2 预测结果及分析

理想情况：

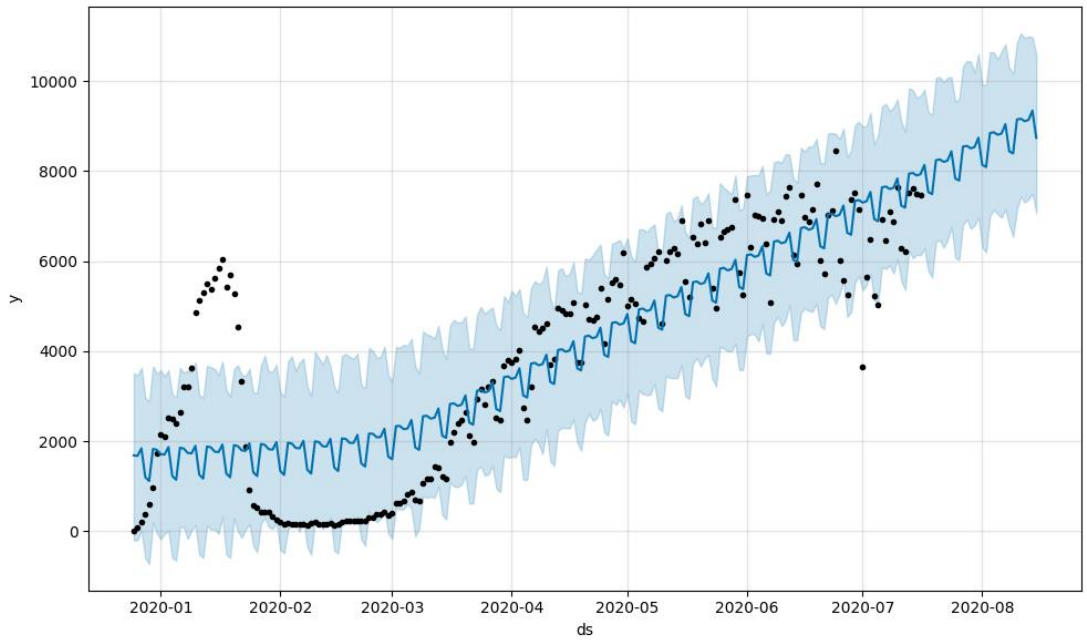


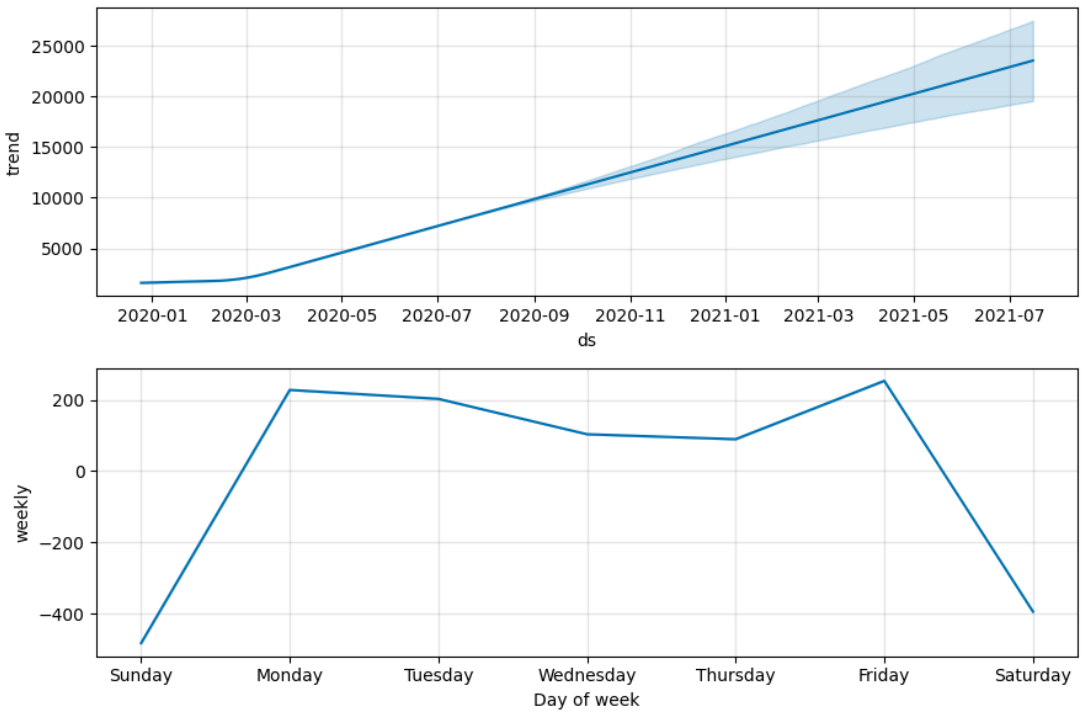

trend
8.5
8.0
7.5
7.0
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
ds


weekly
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week


yearly
1.0
0.5
0.0
−0.5
−1.0
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year

实际情况：


y
10000
8000
6000
4000
2000
0
2020-01
2020-02
2020-03
2020-04
2020-05
2020-06
2020-07
2020-08
ds



4.3.3 系统的维护与风险

全面监控，事前预警：实现全面的数据采集和整理，全面 支持网络设备，ip 地址空间，业务支撑组件，对平台进行实 时的监控和分析，并提供故障的安事前预警机制，帮助系统管 理人员进行监控，提早发现故障隐患，评估威胁，制定优化方 案，采取有效手段，预防故障发生带来的危害。

故障原因定位分析：当资源出现故障时，运维监控系统管 理系统应能够迅速提供一种方案，通过业务监控视图进行初步 的故障分析，快速定位故障点，解决故障。

网络管理：为管理员提供真实的信息，直观地反映了平台 的负载状况和设备属性，以及线路的事时流量；通过颜色显示 负载的和流量的压力，提醒 IT 的运维人员需要关注的信息动 态，随时告知可能存在的隐患。