# 智慧城市决策大数据支撑平台

同济大学软件学院，智能城镇化协同中心

曹布阳

# CONTENTS


01 平台的建设目的

02 平台技术架构

03 平台技术

04 平台应用案例

# 研发团队

曹布阳 教授

吴志强 教授

刘岩 副教授

章恒

孔翎聿

高璟琦

饶伊文

徐小龙

计鹏玥

李贞良

杨博文

杨晨

01

# 平台的建设目的

- 发掘智能城镇或城镇群落关键数据的融合机制，实现生活、自然、社会、人文和环境等基础数据的融合和联动
- 研发涵盖城市内部动态数据，城市群落间联动实时数据，以及城市时空数据的分析框架
- 研发城镇群落指标评价的分析与评估模型， 提供实时可靠的数据分析引擎
- 为城市规划研究人员提供计算平台，提供所需的数据等

# 建设目的



## 协同特征

- 平台与城市研究专家合作，提供相关数据下载与共享服务接口
- 收集整理城市研究案例作为互动展示的案例库。

## 平台特点

- 面向长三角区域城市规划的时空大数据分析可视化平台
- 实现了实时数据采集、数据融合以及统一的时空数据框架
- 整合并储存了经济、社会、生态、交通与基础设施等五个领域51项近120亿条城市数据

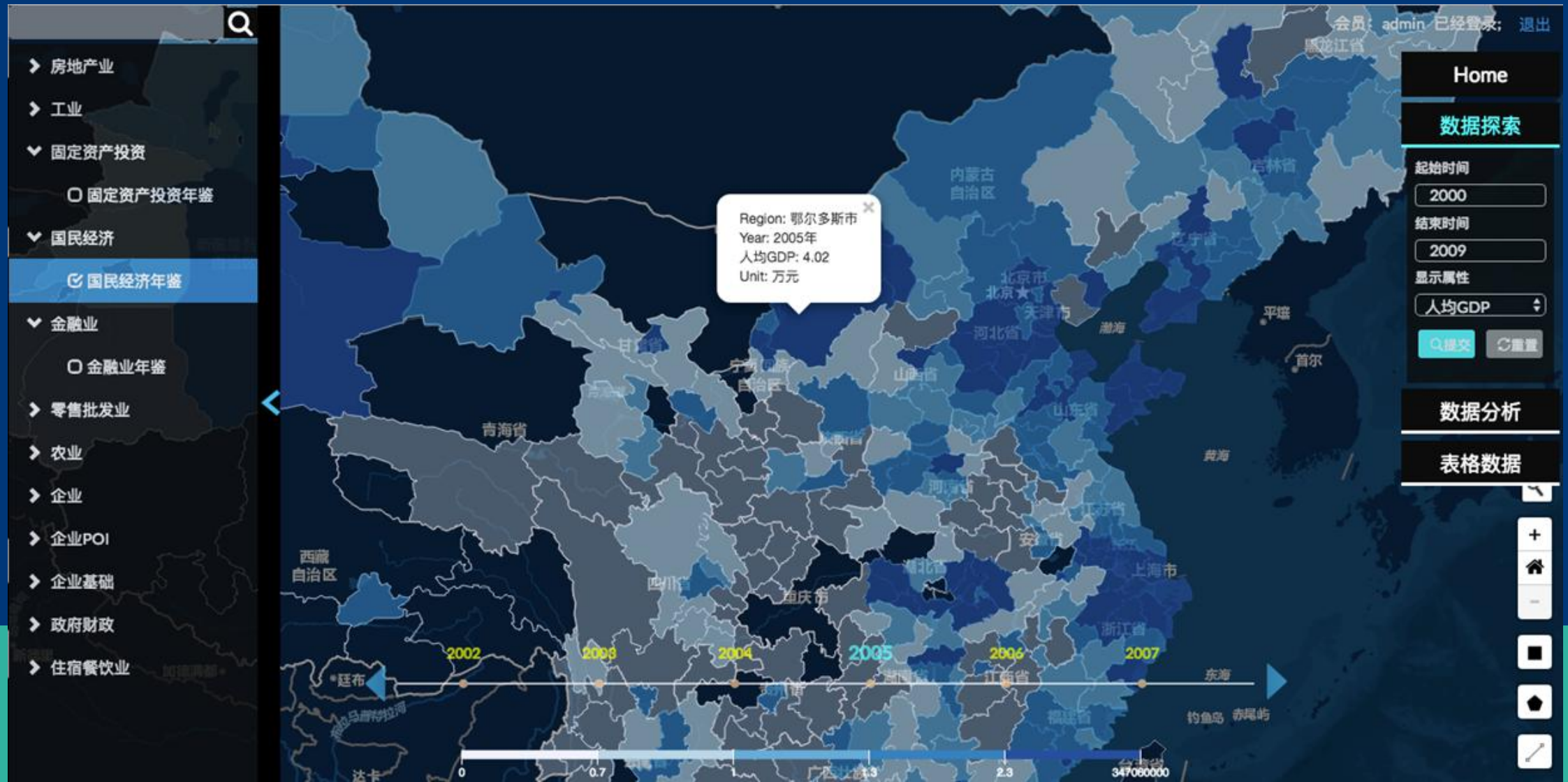
## 技术特征

- ArcGIS Server作为时空数据分析与地图可视化平台
- 结合Spark、ElasticSearch等大数据处理及应用等前沿技术进行实时数据处理与分析，提高时间数据分析效率

# 智慧城市大数据支撑平台



**http://119.23.22.130:3000/login**



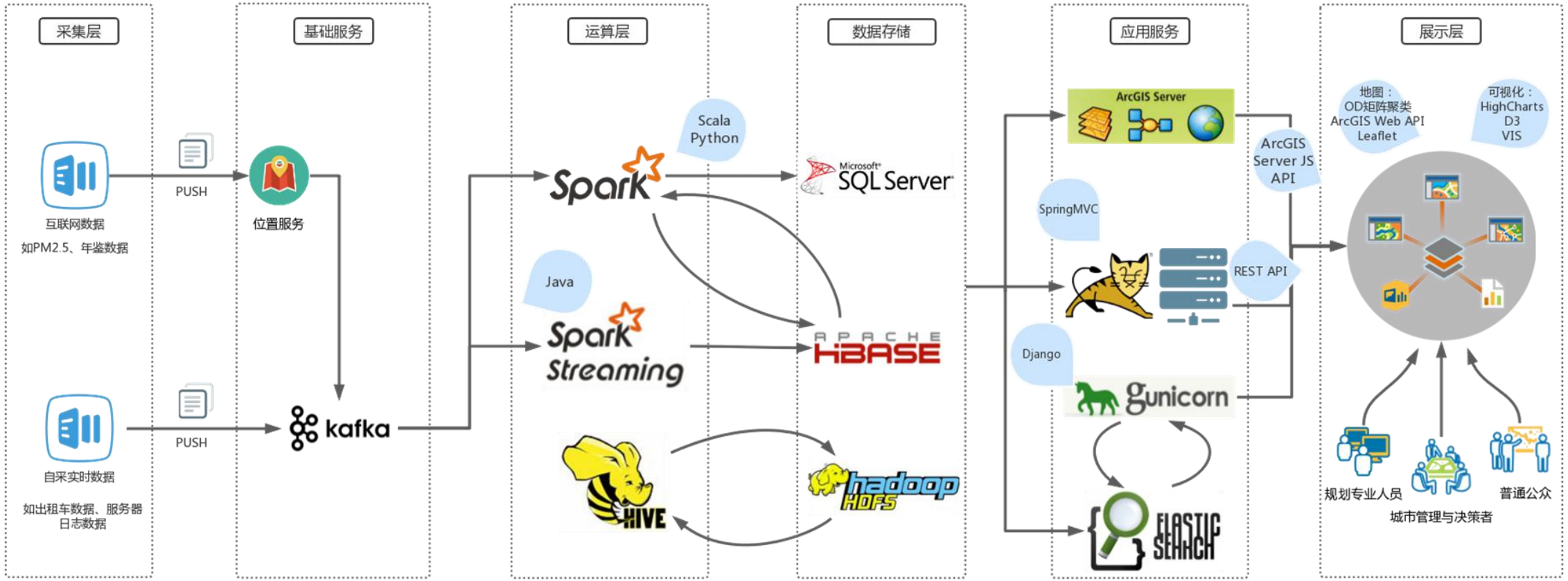
# 02

# 平台技术框架

- 平台总体框架
- 框架分层概要

# 总体框架




采集层
基础服务
运算层
数据存储
应用服务
展示层
互联网数据
如PM2.5、年鉴数据
PUSH
位置服务
自采实时数据
如出租车数据、服务器日志数据
PUSH
kafka
Scala
Python
Spark
Java
Spark Streaming
HIVE
Microsoft SQL Server
APACHE HBASE
hadoop HDFS
ArcGIS Server
SpringMVC
Django
gunicorn
ELASTIC SEARCH
ArcGIS Server JS API
REST API
地图：
OD矩阵聚类
ArcGIS Web API
Leaflet
可视化：
HighCharts
D3
VIS
规划专业人员
普通公众
城市管理与决策者

# 总体框架



**数据收集与基础服务**

- 实时数据的采集
- 数据的清洗与归一
- 空间化处理（地理编码）

1

2

**运算与数据存储**

- 对于数据做初步的整理，分析
- 根据预定的metadata和应用数据模型，把数据存入数据库或相应的HDFS。

3

**应用服务**

- 提供常用时空数据的分析模型和工具。
- 按需提供数据内容

4

**数据展示服务**

- 提供数据的展示。
- 时空要素的结合，比较，图表等

# 03

## 平台技术

- 数据收集与基础服务
- 运算与数据存储
- 应用服务
- 数据展示服务

# 数据收集

## 数据采集器设计目标

- 通过网络数据爬虫程序，从网络中获取经济、环境等数据。丰富系统的数据分析维度
- 通过爬虫获取实时数据保证数据的时效性。

## 数据采集器采集的各类数据源样例

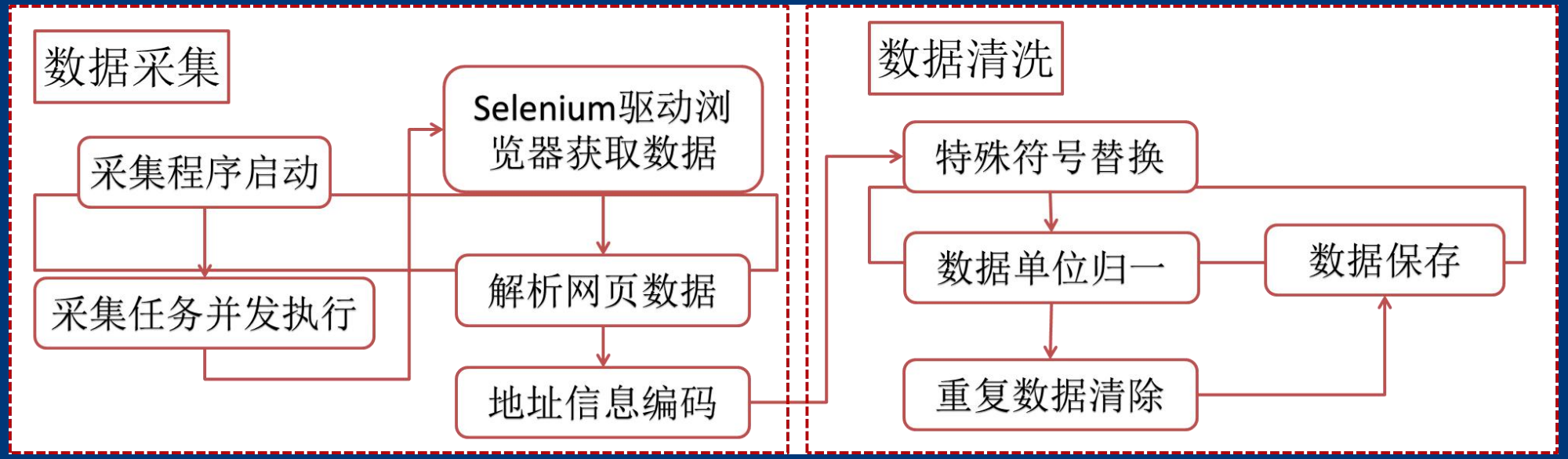
| 数据类型 | 数据 | 数据源 | 更新频率 |
|---|---|---|---|
| 经济 | 上海市注册企业信息 | 天眼查 | 每天 |
| | 地级市社会年鉴数据 | 各市统计局官网 | 每年 |
| | 上海市二手房交易信息 | 链家 | 每天 |
| 环境 | 全国主要流域水质周报 | | 每周 |
| | 全国主要城市PM2.5 | | 每小时 |
| 生活 | 上海市活动信息 | 豆瓣网 | 每天 |
| | 上海市餐馆及评论数据 | 大众点评 | 每天 |

## 数据采集层主要技术

- Selenium、PantomJS 模拟浏览器行为，获取网页实际数据
- Python并发编程技术

# 数据收集工作流程




数据采集
采集程序启动
采集任务并发执行
Selenium驱动浏览器获取数据
解析网页数据
地址信息编码
数据清洗
特殊符号替换
数据单位归一
重复数据清除
数据保存

# 基础服务



## 流数据地理编码服务

- 数据采集流入的本地表格数据需要被统一抽象为现实世界（空间）中的对象，经地理编码转换为WGS84经纬度坐标的地理对象。

## Kafka性能调优

- 合理的设置数据处理的间隔
- 启动多个数据接收实例
- 合理使用内存，并且定期清理不需要的缓存数据

## 分布式消息系统 Kafka的应用

- 提供一个统一的实时信息收集平台, 支撑较大的数据量,具备良好的容错能力
- 保证数据处理顺序，方便异步通信开发

# 运算服务



## 消息订阅

- 从基础服务中接收流数据，设置数据处理间隔，
- 合理分配内存，关注垃圾回收

1

## 数据预处理

2

- 数据的预处理以保证数据质量，
- 去除噪声和无关数据，
- 数据变换与数据集成。

## 数据运算

3

- 使用Spark SQL实现基本的统计与运算功能，
- 使用Spark MLib实现与机器学习相关功能。

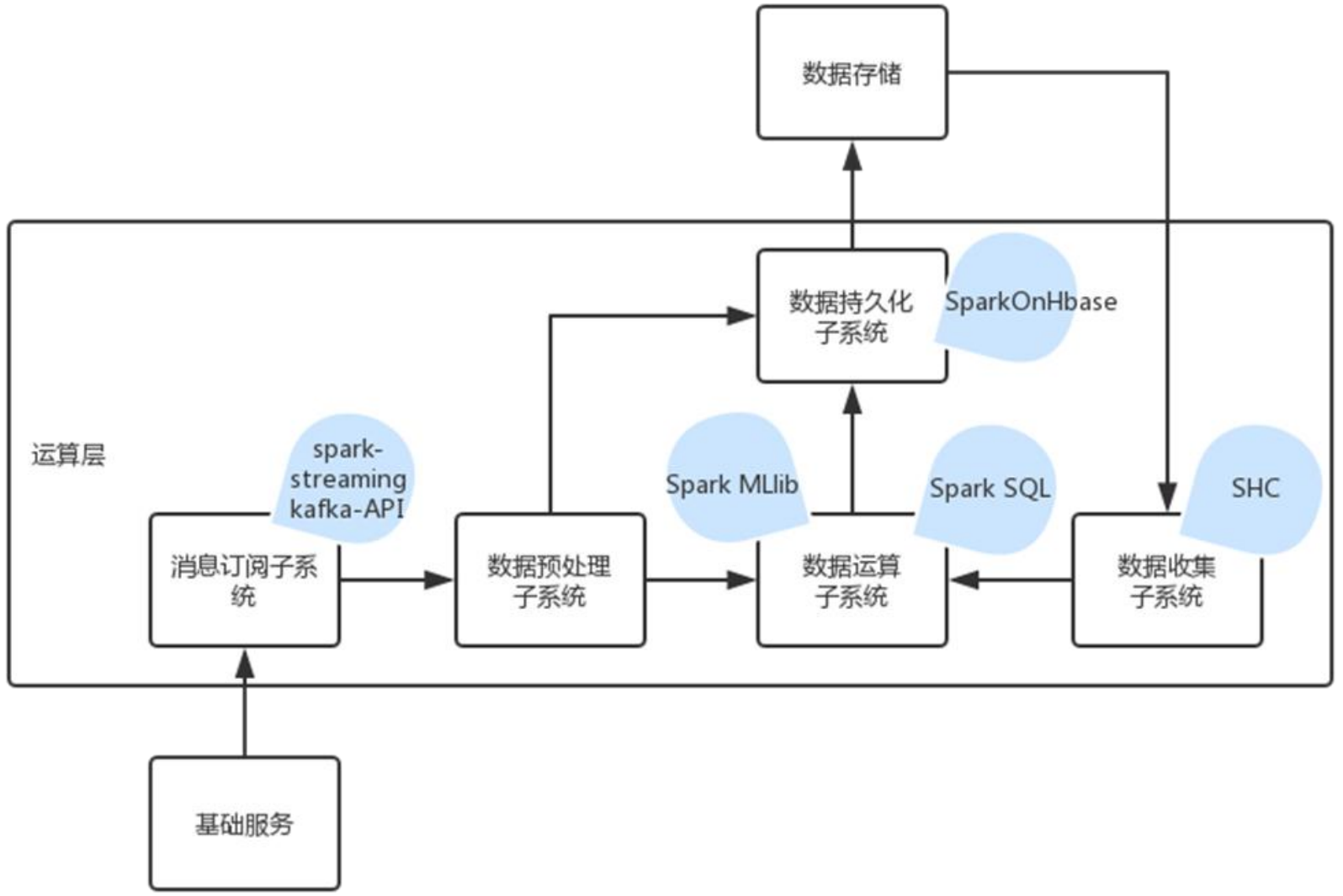
4

## 数据供给和持久化

- 针对不同类型数据源编写相应数据读取接口，实现基于HBase的大数据量下的快速读取。
- 针对不同类型数据源编写相应的持久化接口，将源数据和运算的结果进行持久化存储

# 运算服务框架




数据存储
数据持久化
子系统
SparkOnHbase
运算层
spark-
streaming
kafka-API
Spark MLlib
Spark SQL
SHC
消息订阅子系
统
数据预处理
子系统
数据运算
子系统
数据收集
子系统
基础服务

# 数据存储



## 设计理念

- 根据数据源本身的数据类型，结构化数据、空间数据、非结构化数据等分别对待，
- 采用SQL Server、GeoDatabase与HBase 等存储技术。

1

2

## SQL Server

- 数据储存量小
- 结构化数据，数据表间联系紧密
- 现有11个数据库，60张数据表，9126206条数据

3

## Geodatabase

- 矢量地图基础数据
- 栅格数据集
- 时空运算与分析的中间结果
- 可视化时空图层

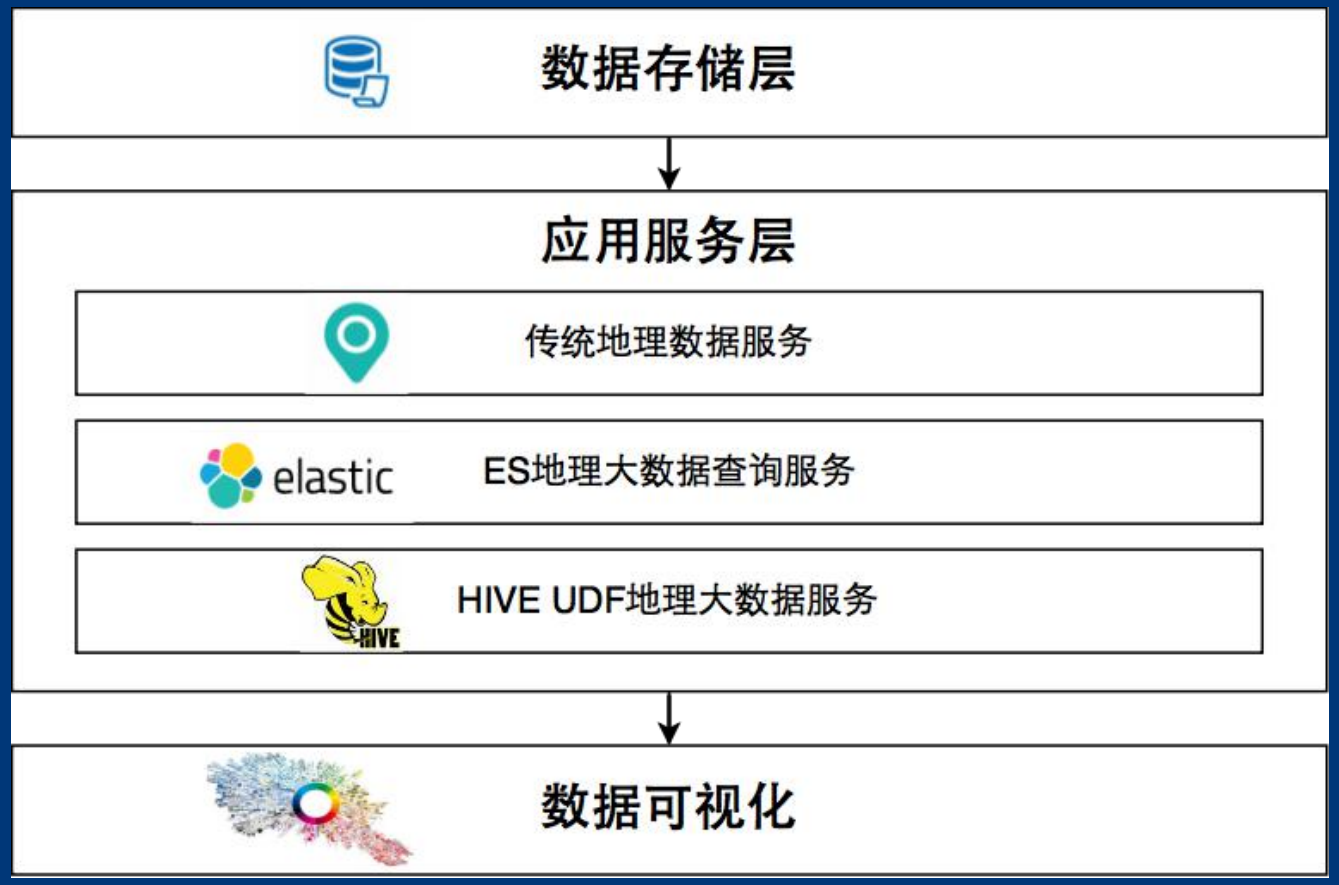
4

## Hbase

- 数据量大
- 可伸缩，面向列的数据
- 现有8个表，119亿条数。

# 应用服务






- 通过ArcGIS，提供通用的GIS数据管理服务，融合关系型数据库与地理数据并以图层形式推送到前端，支持地图图层可视化。
- ElasticSearch提供RESTful web接口，用于大规模时空数据的搜索
- 利用GIS Tools for Hadoop扩展基于Hive的UDF应用，通过Hive在时空数据上构建MapReduce应用程序序

从数据到内容。

# 应用服务 – 传统地理服务和大规模时空数据的服务

## 传统地理服务

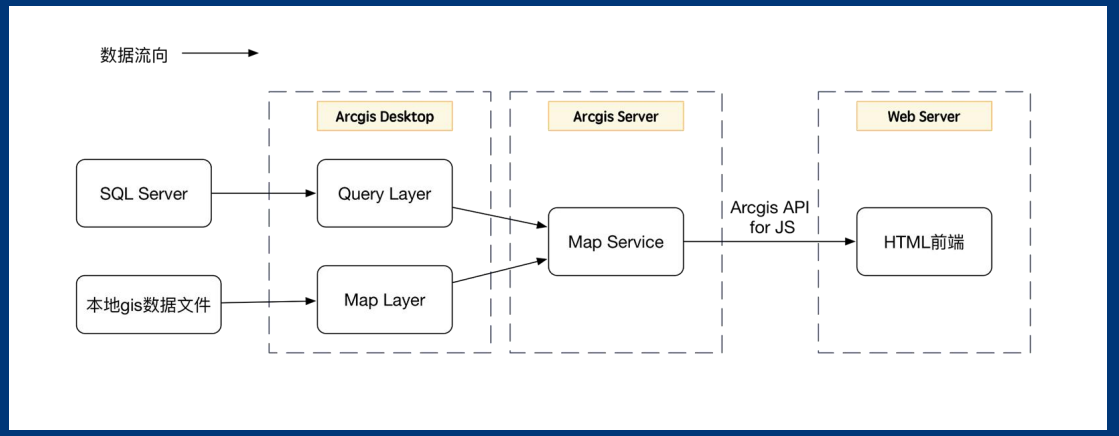
- 以Query Layer的形式读入客户端内，访问Layer属性获取相应数据。
- 利用Arcgis将制作好的地图数据打包发布成地图服务Map Service
- 利用JavaScript API接口进行前端开发，展示相关的时空数据

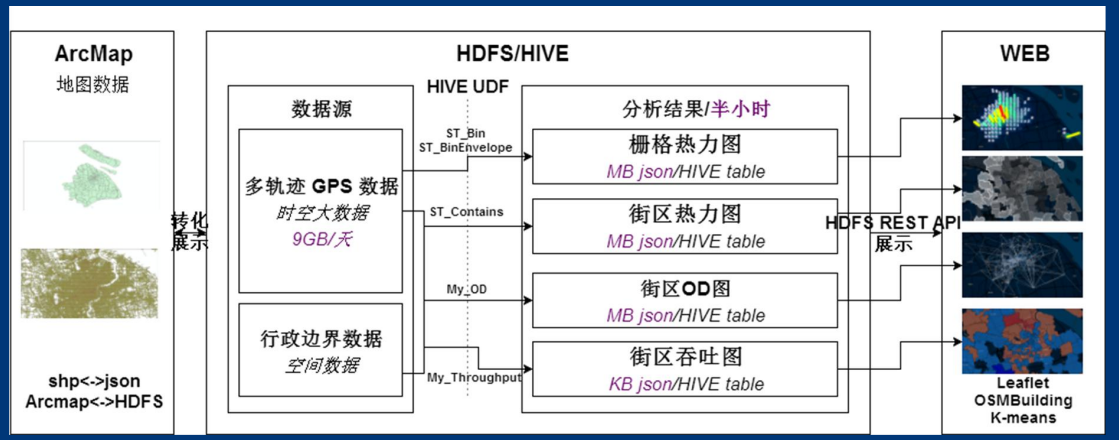
## 基于ES大规模时空数据服务

- 根据用户的需求，采用ES的Query DSL对数据进行过滤/搜索和聚合，并做必要的处理，返回给用户。

## UDF服务

- Hive UDF + Hive SQL
- 基于分布式处理方法对空间大数据使用空间分析函数统计分析。



# 数据展示服务



## 数据探索界面

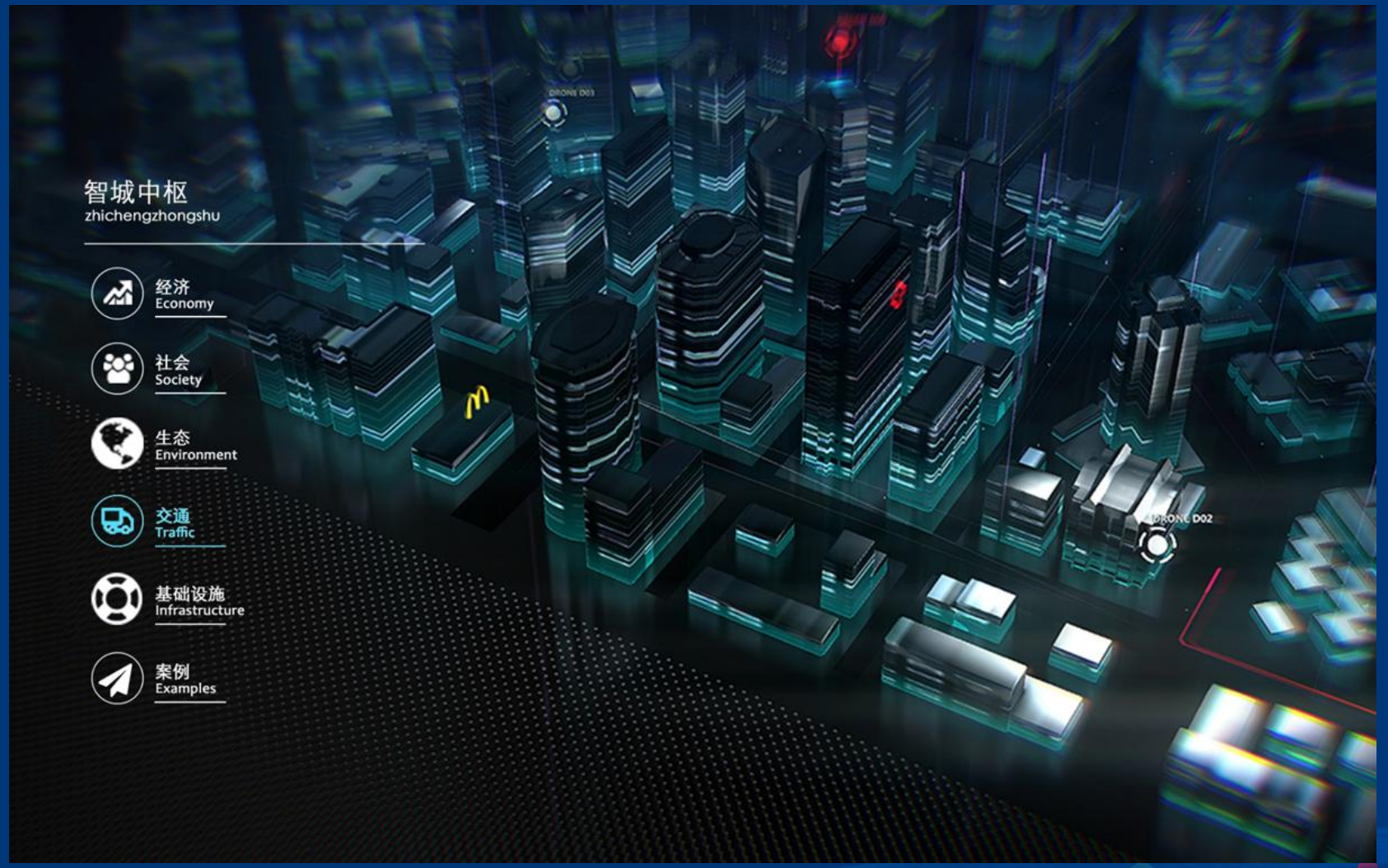
- 数据查询
- 数据分析
- 地理数据可视化

## 数据管理界面

- 服务器运行监控
- 数据库状态查询
- 数据结构管理
- 数据采集状态管理

## 城市案例库

- 城市案例更新
- 城市案例搜索
- 案例数据交互



# 数据探索



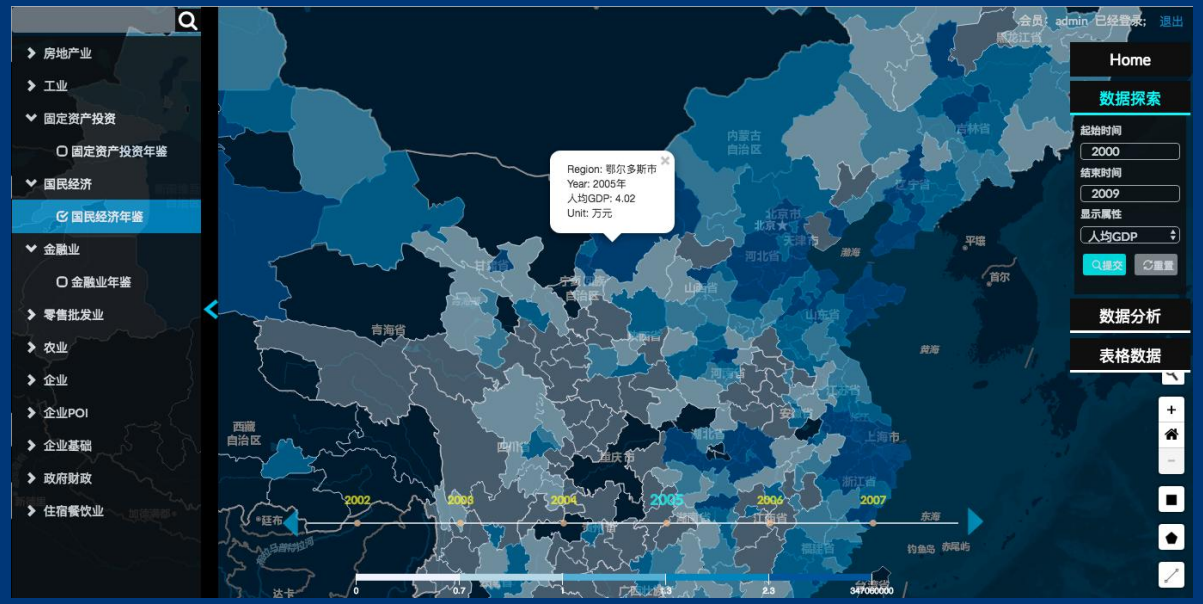
数据查询

数据分析

时空数据可视化

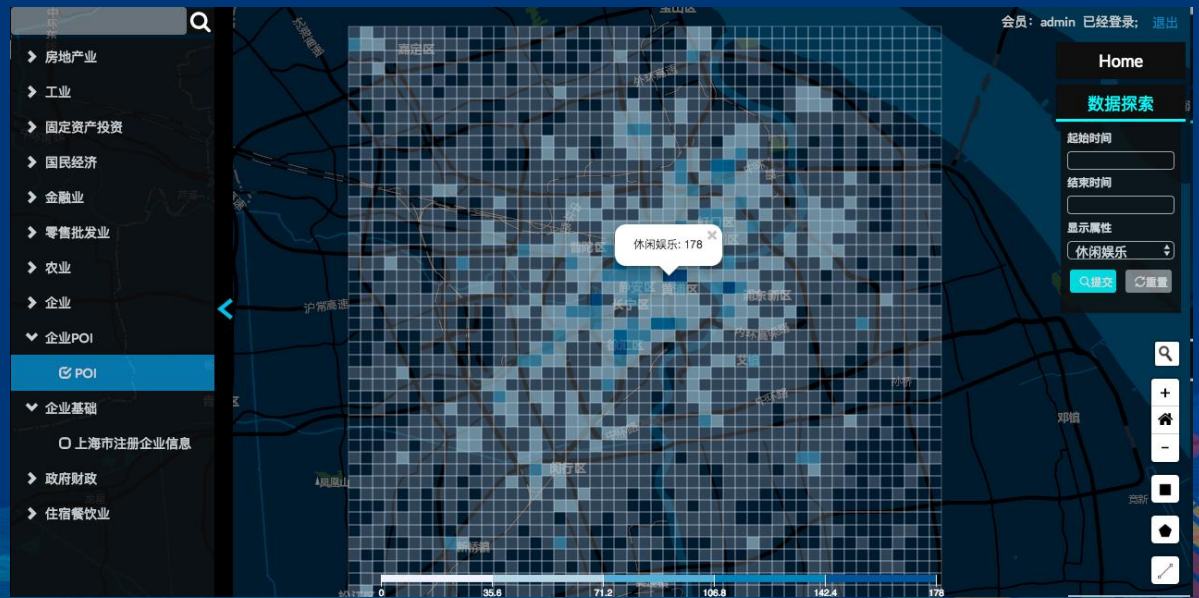
# 数据探索




房地产业
工业
固定资产投资
固定资产投资年鉴
国民经济
国民经济年鉴
金融业
金融业年鉴
零售批发业
农业
企业
企业POI
企业基础
政府财政
住宿餐饮业
会员：admin 已经登录; 退出
Home
数据探索
起始时间
2000
结束时间
2009
显示属性
人均GDP
提交
重置
数据分析
表格数据
Region: 鄂尔多斯市
Year: 2005年
人均GDP: 4.02
Unit: 万元
2002
2003
2004
2005
2006
2007

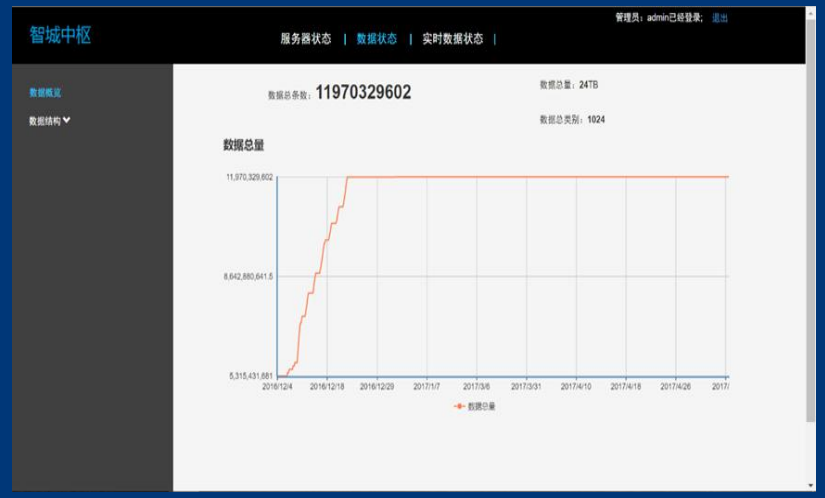
房地产业
工业
固定资产投资
固定资产投资年鉴
国民经济
国民经济年鉴
金融业
金融业年鉴
零售批发业
农业
企业
企业POI
企业基础
政府财政
住宿餐饮业
下载
id
year
region
code
flag
indicator
unit
value
536791 2005年 唐山市 xj0302 2 人均GDP 万元 0.00026699999580151
42107 2005年 定西市 XJ2811 2 人均GDP 万元 0.2393999993801117
574534 2005年 和田地区 XJ3111 2 人均GDP 万元 0.2712000012397766
1058281 2005年 陇南市 XJ2812 2 人均GDP 万元 0.2728999853134155
185534 2005年 临夏州 XJ2813 3 人均GDP 万元 0.2892999947071075
1429323 2005年 固原市 xj3004 2 人均GDP 万元 0.2991999983787536
Home
数据探索
起始时间
2000
结束时间
2009
显示属性
人均GDP
提交
重置
数据分析
表格数据
2002
2003
2004
2005
2006
2007

城市环境卫生
地理特征
空气质量
PM2.5
空气质量状况
监测站点
实时空气质量数据
空气数据说明
2016年空气质量状况
水文质量
会员：admin 已经登录; 退出
Home
数据探索
City: 上海
Position name: 青浦淀山湖
co: 0.5
Uint: 微克/立方米
Time: 2017-5-3 12:00
2017-5-3 9:00
2017-5-3 10:00
2017-5-3 11:00
2017-5-3 12:00
2017-5-3 13:00
2017-5-3 14:00


房地产业
工业
固定资产投资
国民经济
金融业
零售批发业
农业
企业
企业POI
POI
企业基础
上海市注册企业信息
政府财政
住宿餐饮业
会员：admin 已经登录; 退出
Home
数据探索
起始时间
结束时间
显示属性
休闲娱乐
提交
重置
休闲娱乐: 178

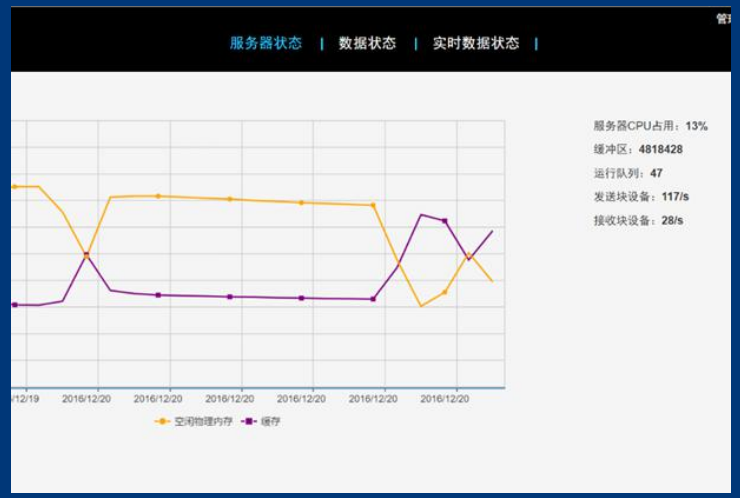
# 数据管理





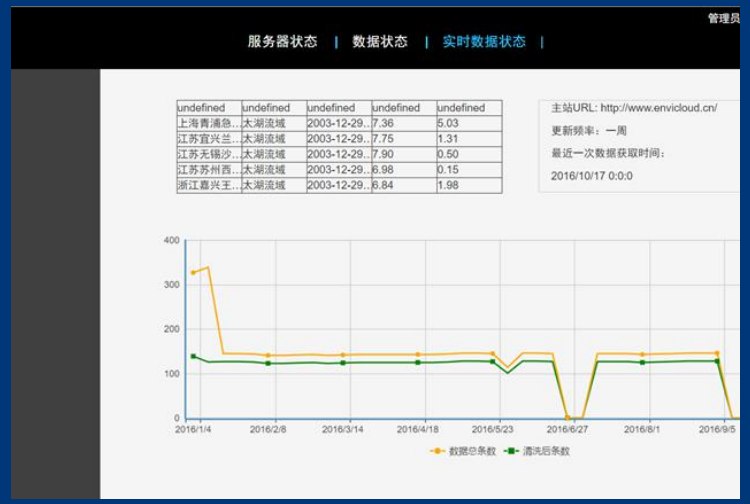

## 数据总量查询

•折线图表示一段时间内数据总量的变化趋势，




## 服务器状态监控

- 折线图展示服务器空间物理内存与缓存的变化趋势；
- 文字动态显示服务器其他实时状态指标。

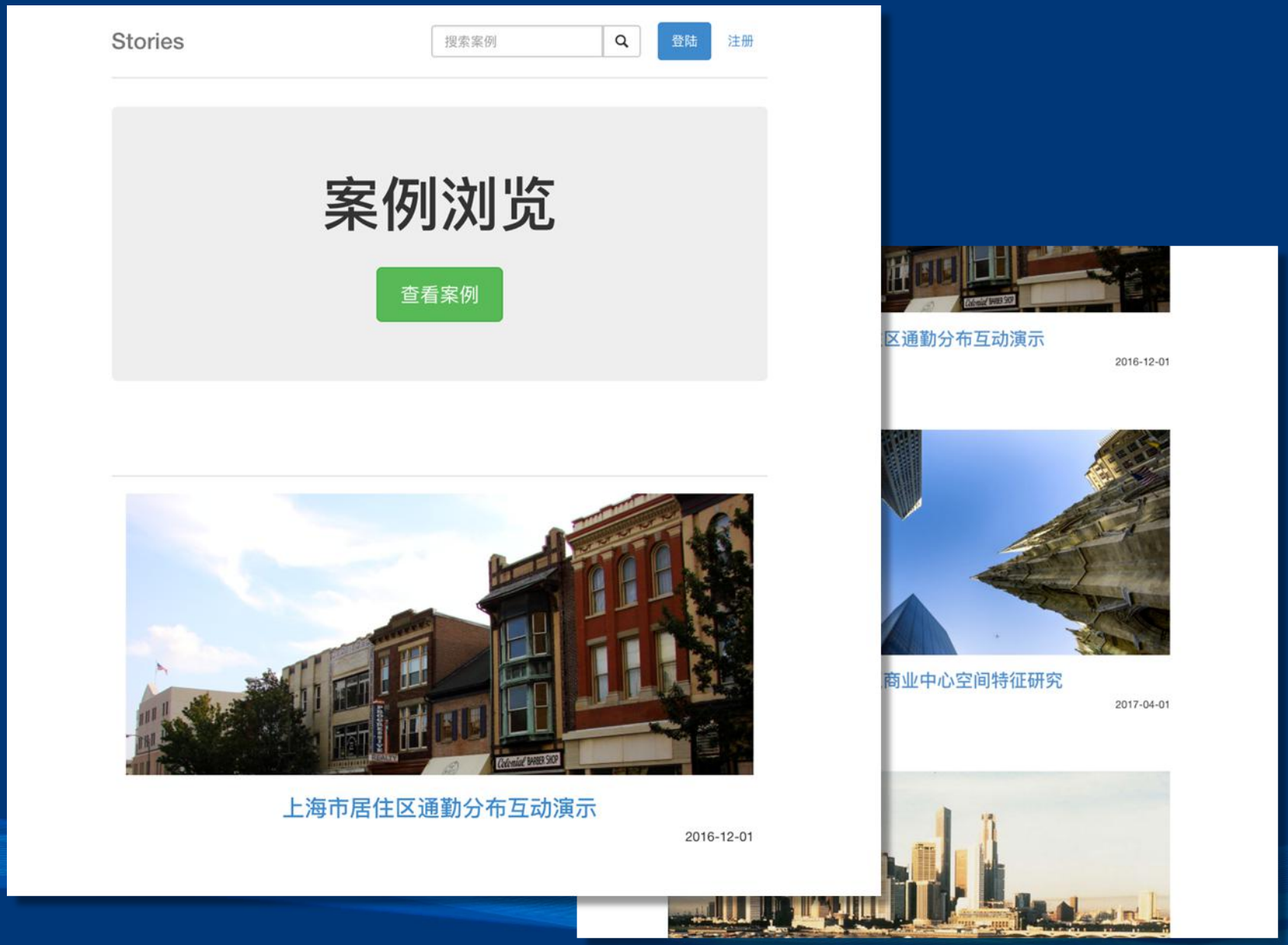



## 实时数据查询

- 表格显示实时数据格式，给出获取数据的最新时间。
- 折线图可以表现出数据获取的总条数的变化趋势，利用不同颜色线条表示总条数与清洗后条数。

# 案例库






- 以**Map Story**的形式建立了城市规划，状况分析案例库
- 可对收集的案例搜索和查看

# 案例库

- 关键词搜寻案例
- 案例地图具有交互功能

# 04

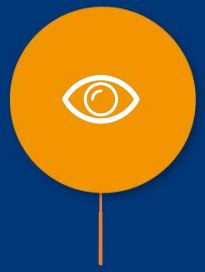
# 平台技术应用案例

- 出租车数据分析
- 交通卡数据分析与挖掘

# 交通数据分析 – 出租车为例

## 栅格结构的应用

根据栅格中某时段出租车的数据，构造栅格热力图



## 街区热力图

根据街区中某时段出租车的数据，构造街区热力图

## 街区OD分析

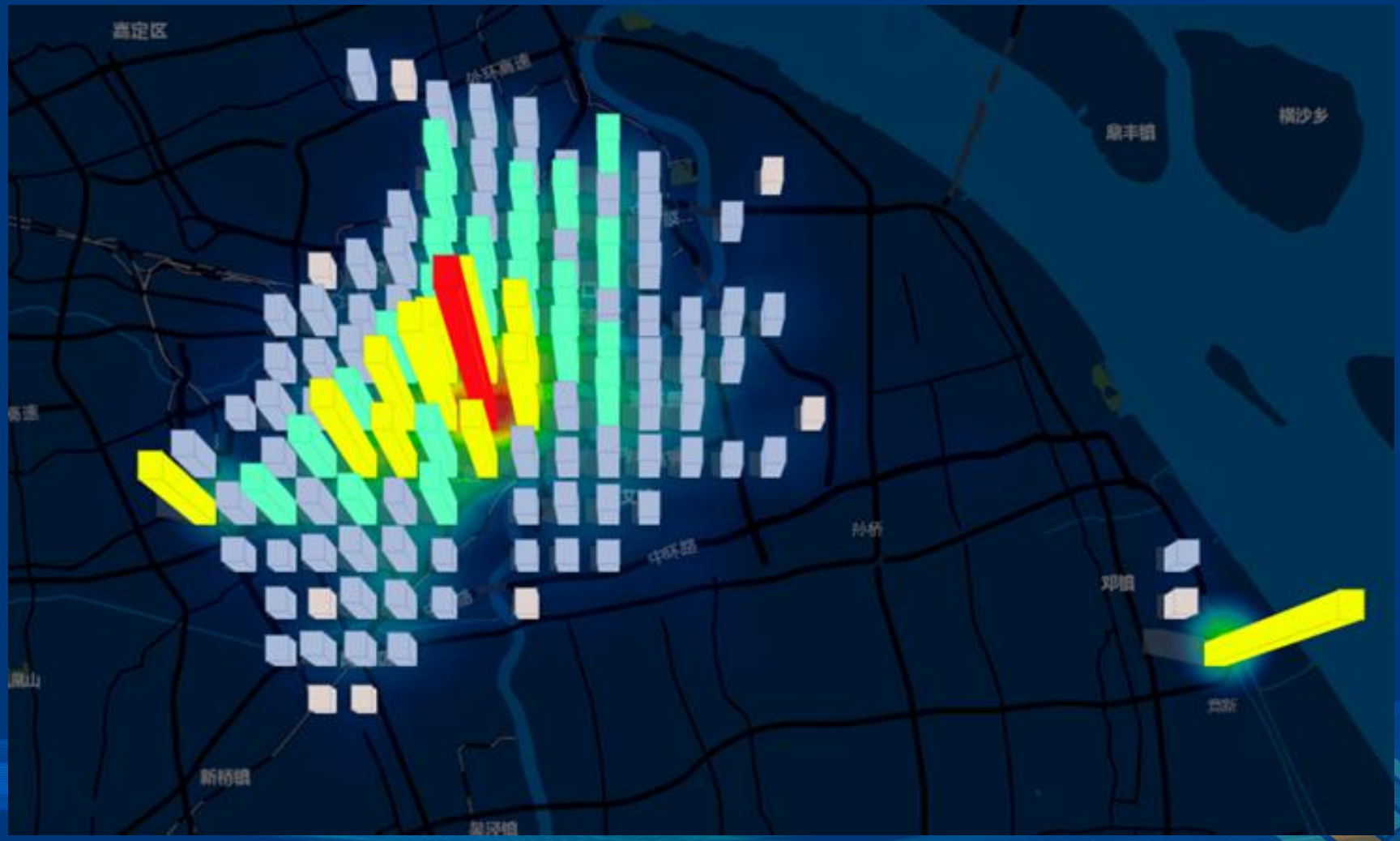
利用出租车轨迹数据，建立相应的街区OD关系。

## 街区吞吐分析

街区总输出和总输入的差值。

# 交通数据分析 – 出租车为例

## 栅格热力图 (半小时时间段 )



# 交通数据分析 – 出租车为例

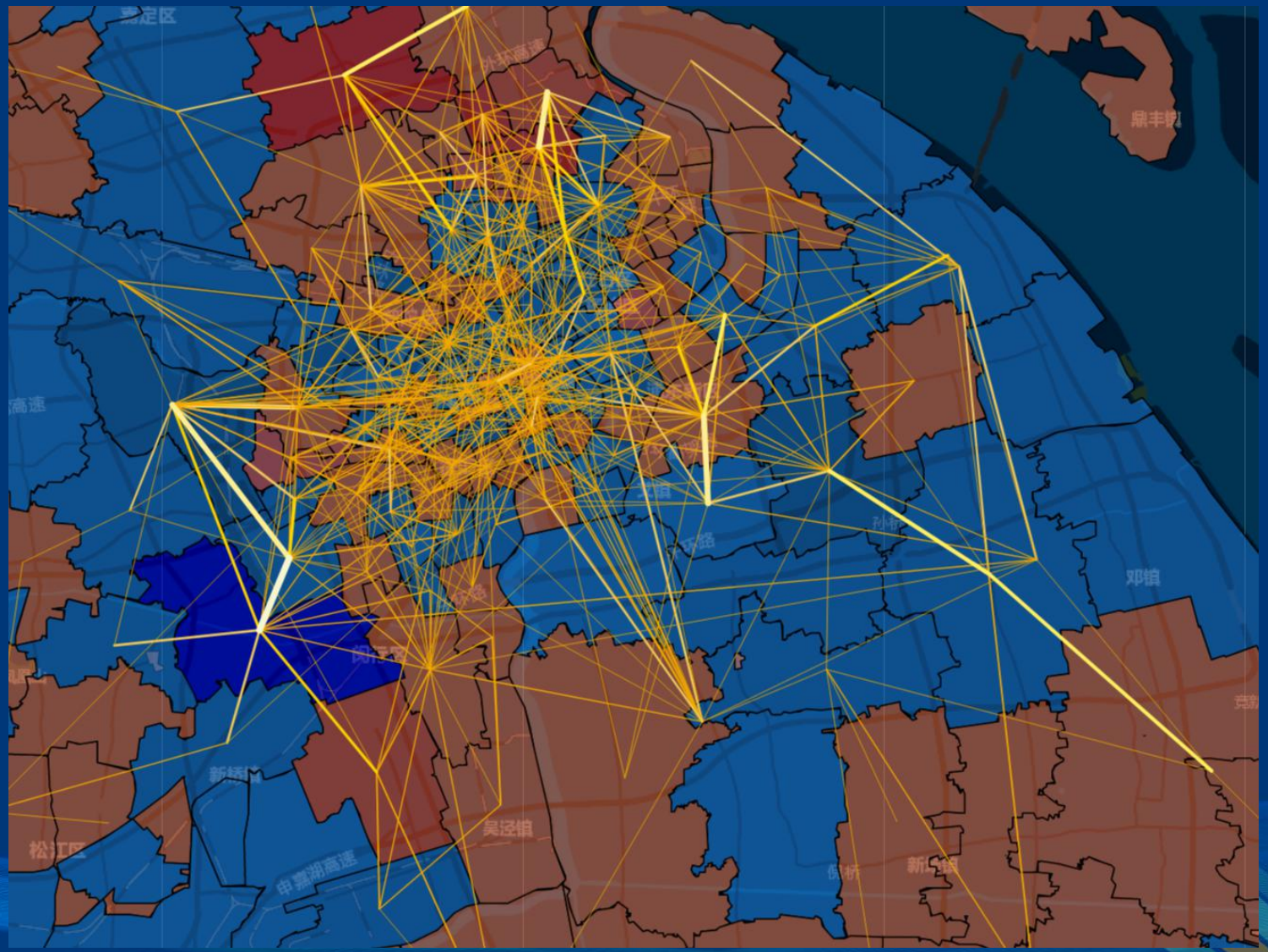


## 街区热力图（半小时时间段）

# 交通数据分析 – 出租车为例

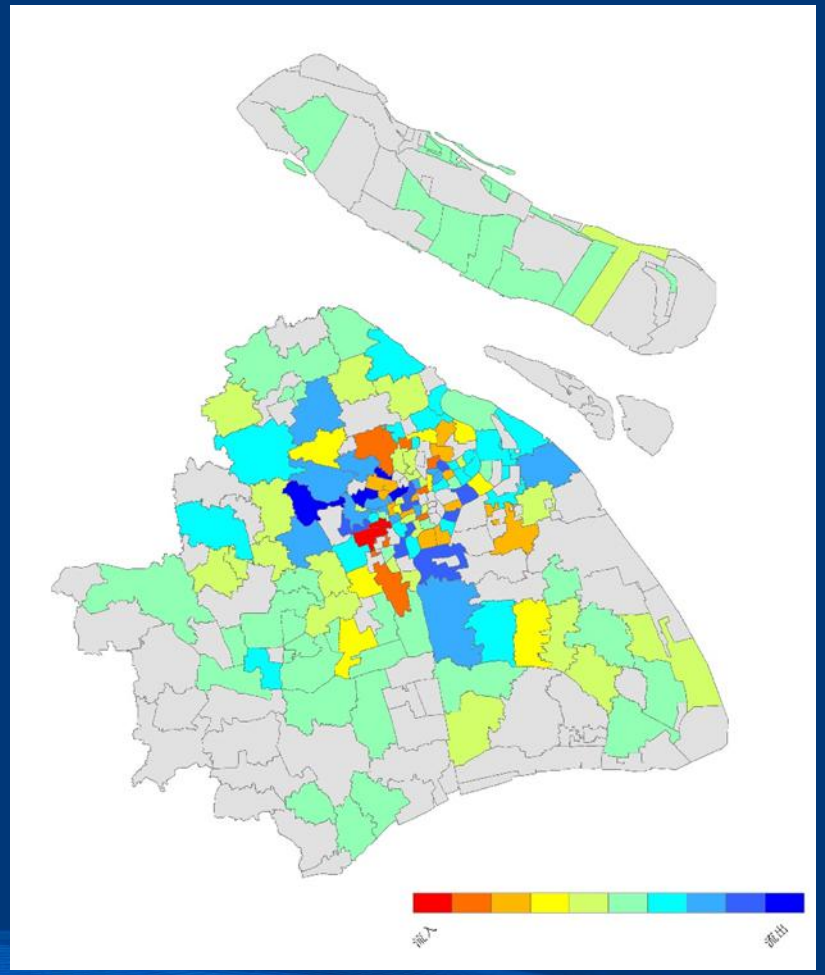


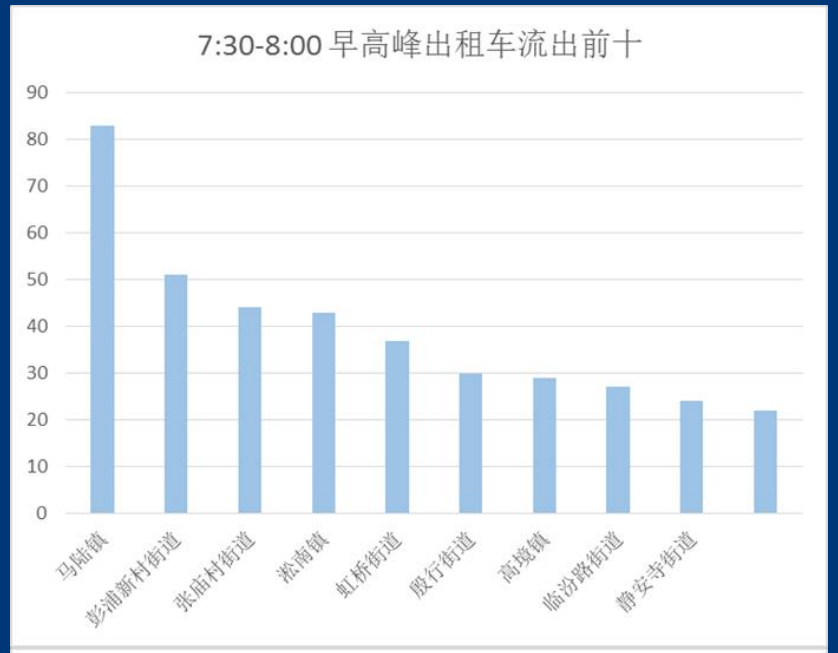
## 街区OD图（半小时时间段）

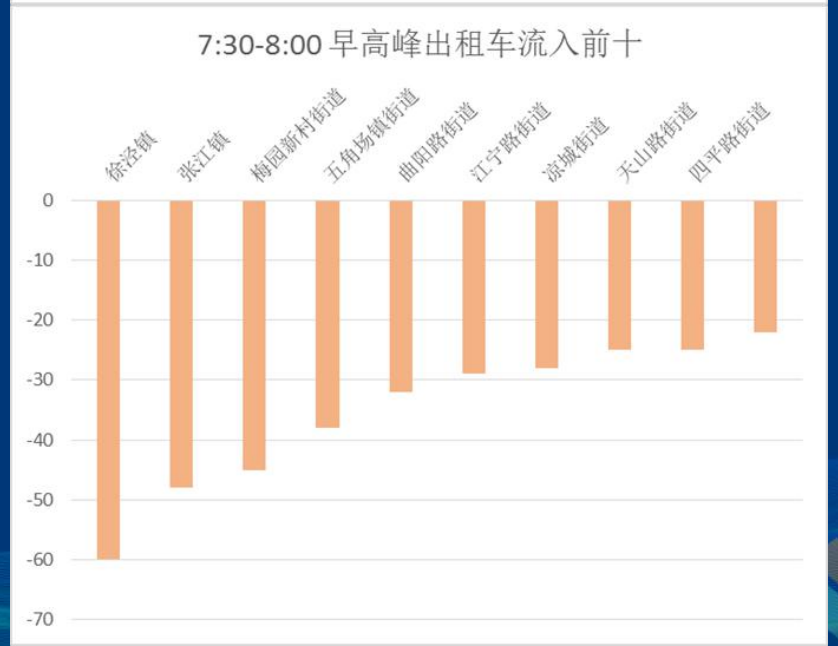


# 交通数据分析 – 出租车为例

## 早高峰吞吐量分析（7：30 至8：00）



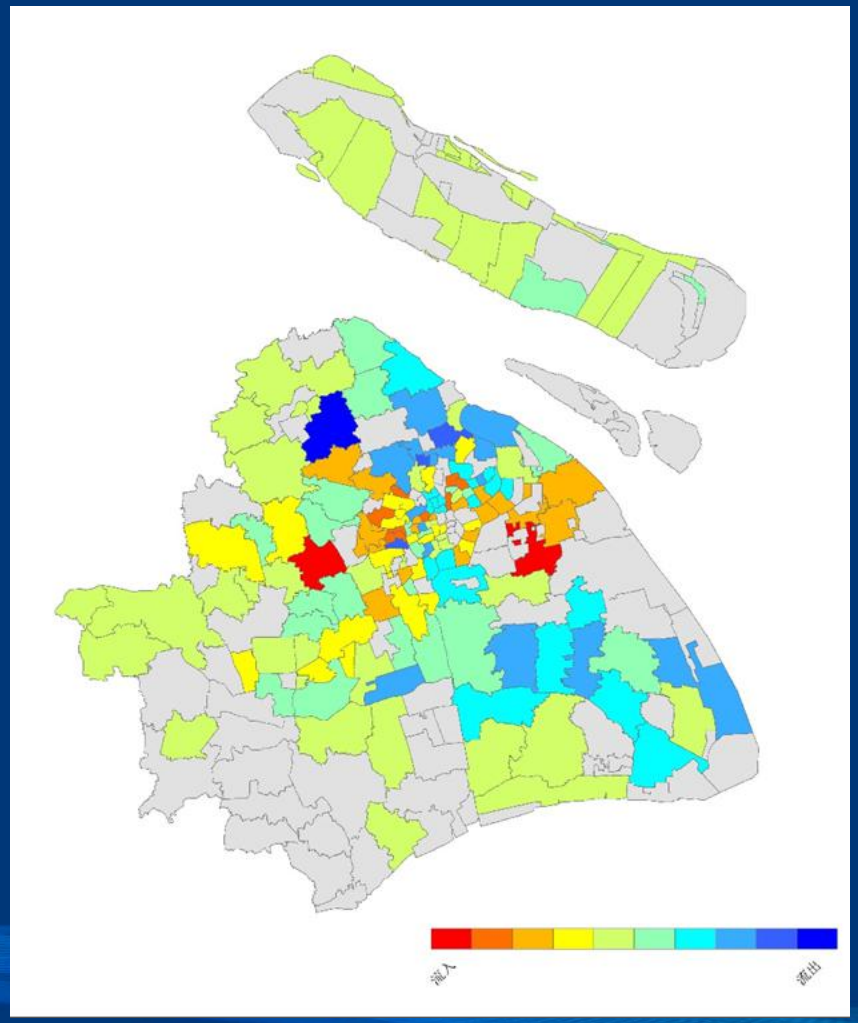



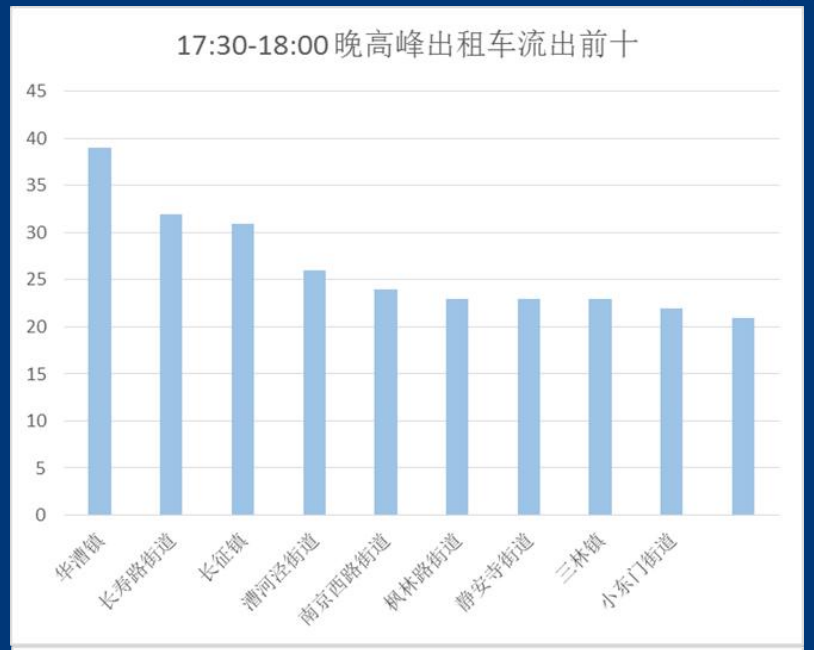



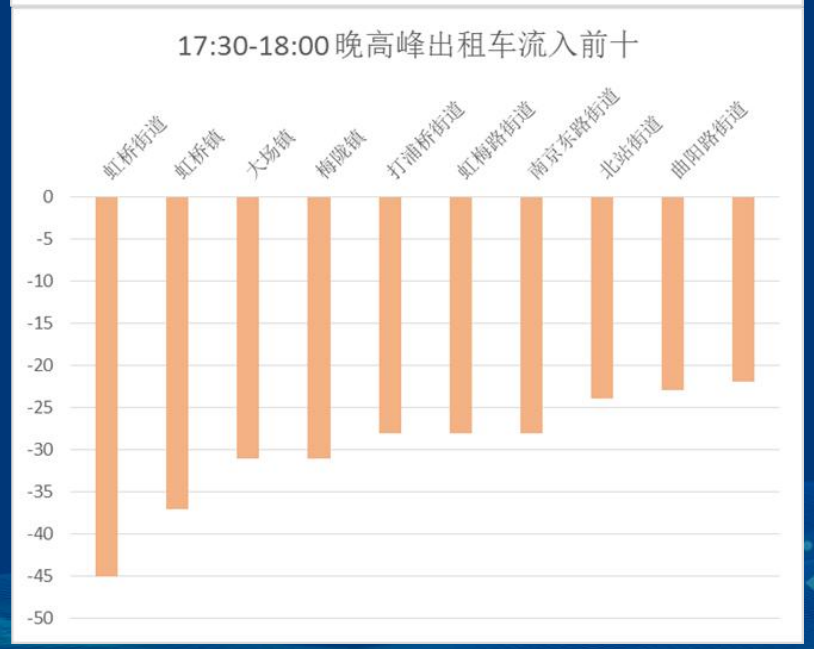
# 交通数据分析 – 出租车为例



## 晚高峰吞吐量分析（17：30 至18：00）



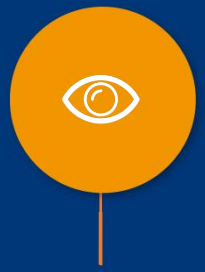



# 上班族与地铁

## 绿色出行

快捷，便利，绿色环保



## 地铁上班族

- 条件一：连续3天固定出行模式
- 条件二：到达时间早于中午12点
- 条件三：地铁通勤时间小于2个小时

## 数据来源

交通卡数据

## 数据量

- 4,160,000交通卡用户（一天）
- 约有1,200,000（28%）人坐地铁上班

# 上班族与地铁



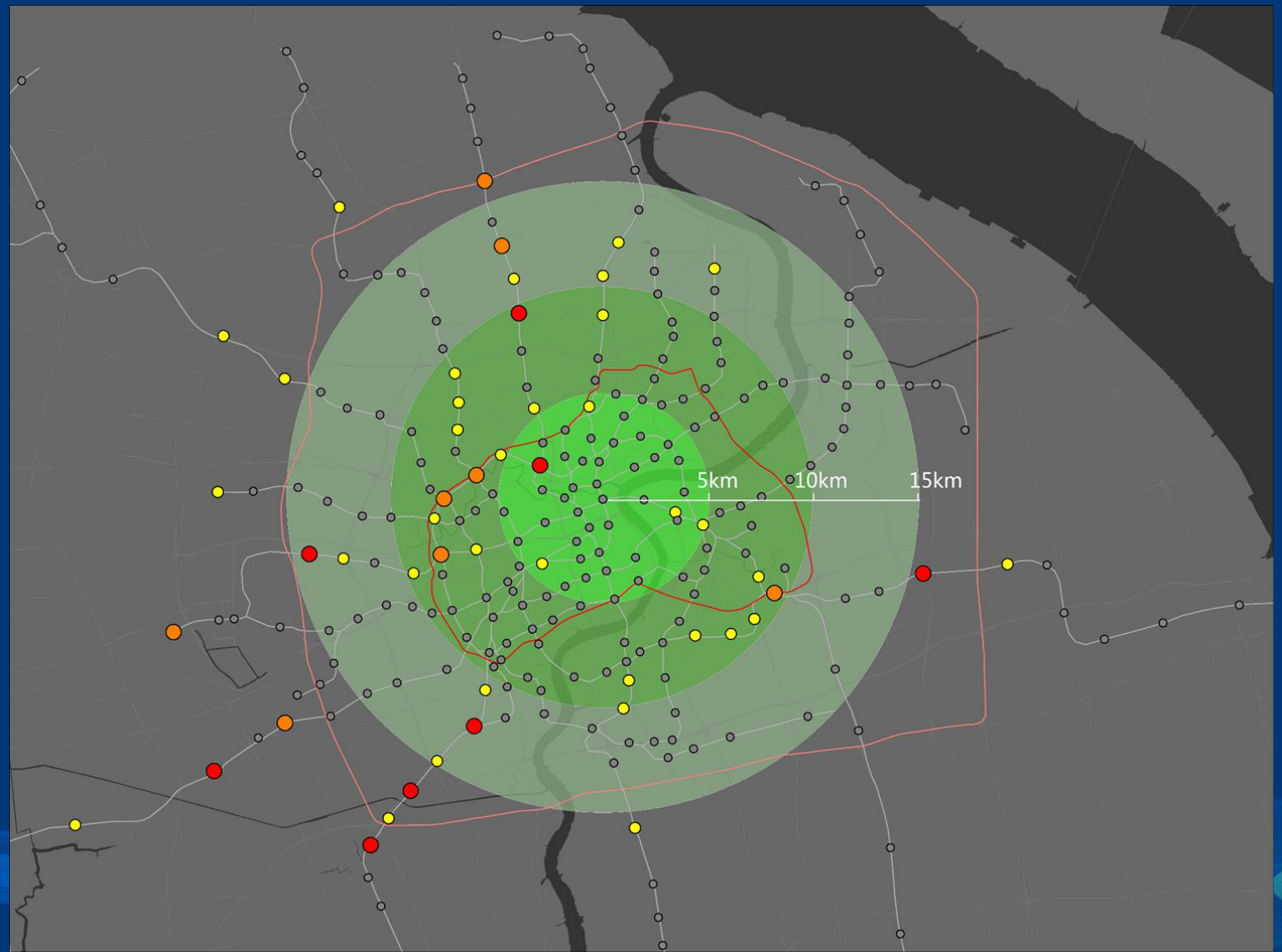
样例数据

| CardID | OStation | DStation | OTime | DTime | SpendTime |
|---|---|---|---|---|---|
| 79796 | 延长路 | 桂林路 | 7:52:05 | 8:36:42 | 44 |
| 84407 | 宋园路 | 四川北路 | 9:40:09 | 10:08:06 | 28 |
| 92416 | 北新泾 | 陆家嘴 | 7:44:46 | 8:14:38 | 30 |
| 92563 | 外环路 | 浦电路 | 7:47:56 | 8:30:50 | 43 |
| 93715 | 枫桥路 | 南京东路 | 7:52:26 | 8:19:09 | 27 |
| 100902 | 灵岩南路 | 源深体育中心 | 7:09:16 | 8:14:57 | 65 |
| 106780 | 巨峰路 | 东昌路 | 10:33:02 | 11:03:28 | 30 |
| 106979 | 隆昌路 | 复兴岛 | 7:02:24 | 7:09:22 | 7 |
| 106997 | 杨思 | 上海体育馆 | 8:28:09 | 8:56:54 | 28 |
| 107263 | 彭浦新村 | 人民广场 | 7:55:30 | 8:22:23 | 27 |
| 109455 | 隆昌路 | 金海路 | 7:33:12 | 7:58:07 | 25 |
| 109488 | 新村路 | 南陈路 | 6:25:42 | 6:40:52 | 15 |
| 117721 | 共康路 | 常熟路 | 5:56:29 | 6:31:22 | 35 |
| 118672 | 上海南站 | 静安寺 | 7:59:29 | 8:33:02 | 34 |

# 上班族与地铁



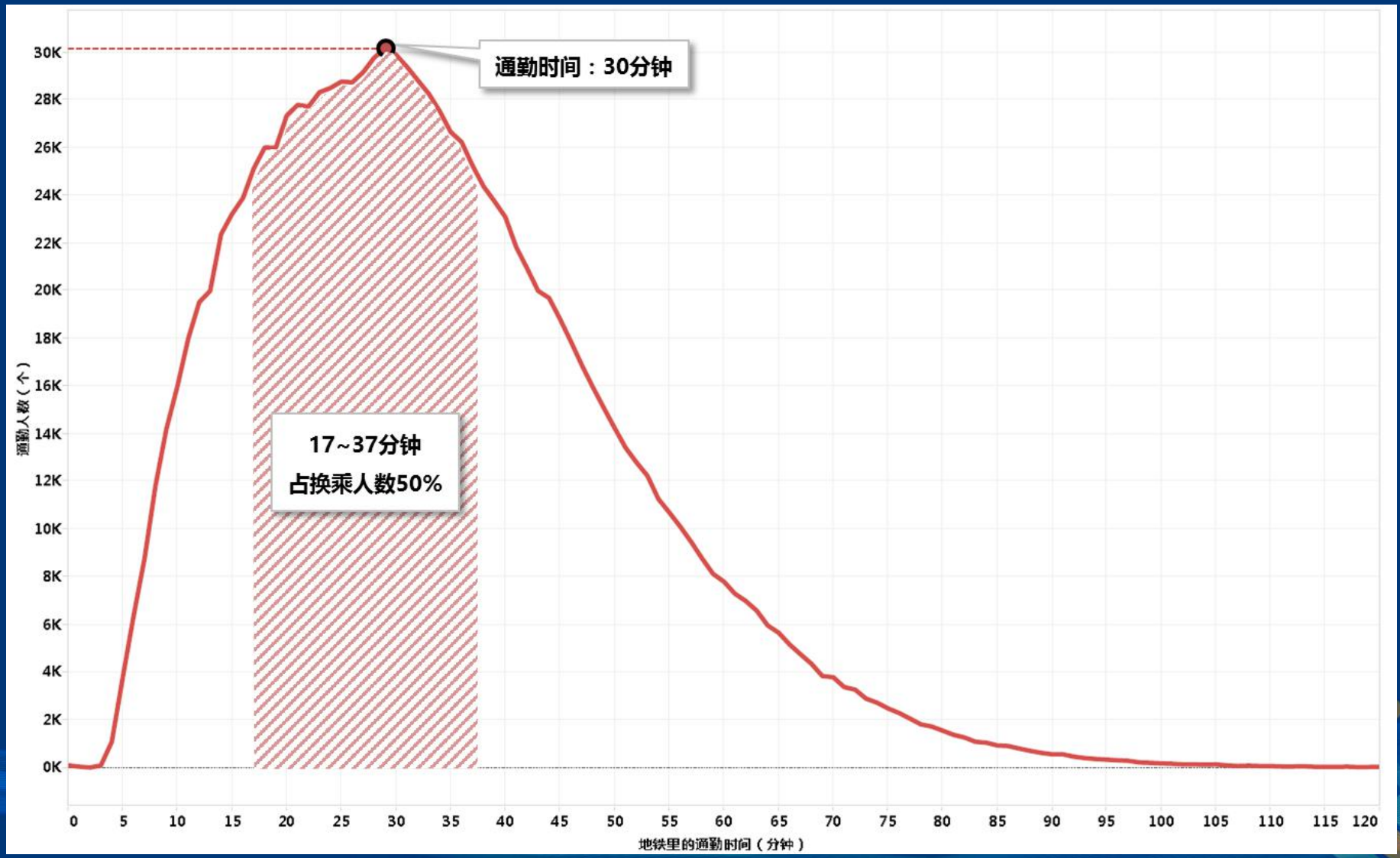
高房价高生活费把上班族赶到哪里去了？



# 上班族与地铁



上班族一生花了多少时间在地铁上（通勤时间）？



# 上班族与地铁

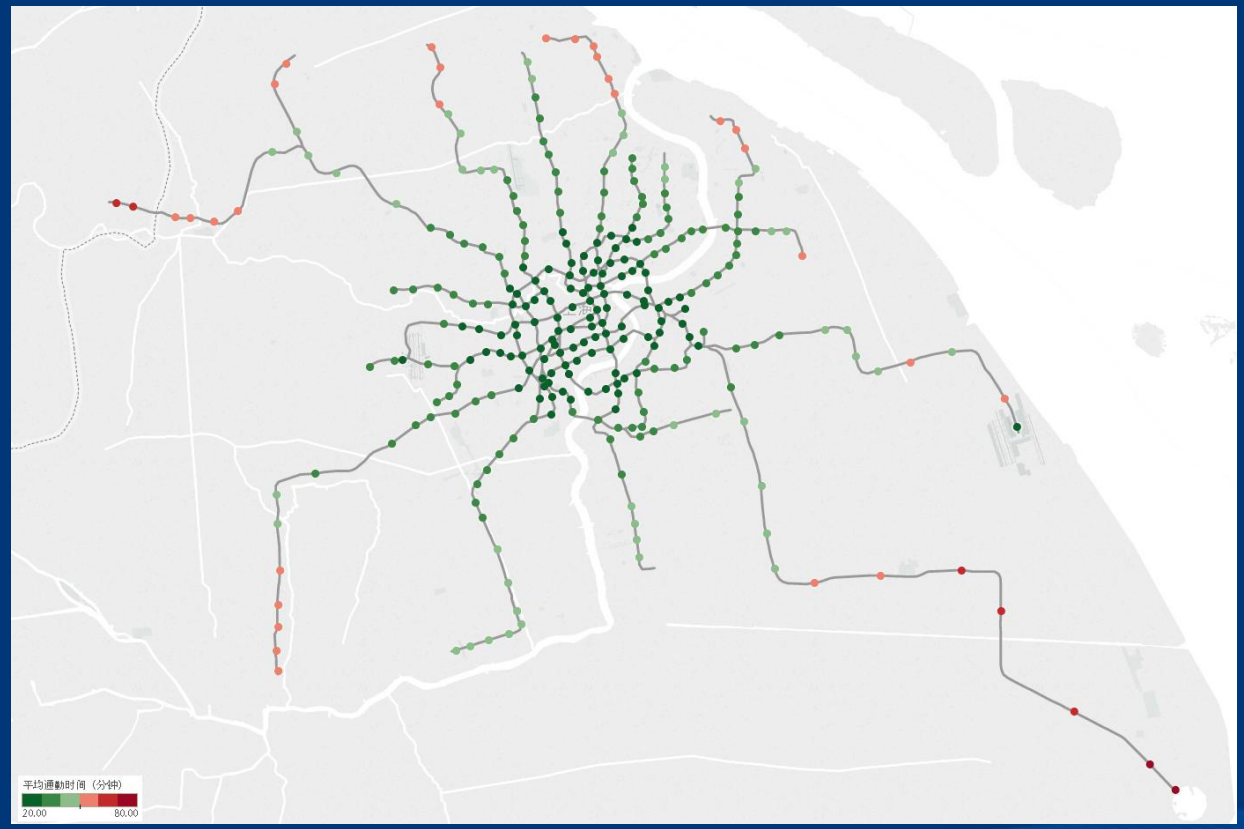


从各站点出发的平均通勤时间，word天，临港大道！

| 排序 | 站点名 | 平均时间（分钟） |
|---|---|---|
| **1** | **陆家嘴** | **20.57** |
| **2** | **衡山路** | **20.59** |
| **3** | **新天地** | **20.90** |
| **4** | **东昌路** | **21.02** |
| **5** | **陕西南路** | **21.06** |
| 6 | 南京西路 | 21.35 |
| 7 | 南京东路 | 21.61 |
| 8 | 上海图书馆 | 21.63 |
| 9 | 陆家浜路 | 21.65 |
| 10 | 上海科技馆 | 21.73 |

| 排序 | 站点名 | 平均时间（分钟） |
|---|---|---|
| **1** | **临港大道** | **75.92** |
| **2** | **滴水湖** | **71.71** |
| **3** | **光明路** | **65.76** |
| **4** | **书院** | **63.75** |
| **5** | **花桥** | **63.74** |
| 6 | 惠南东 | 62.67 |
| 7 | 惠南 | 61.55 |
| 8 | 嘉定北 | 59.85 |
| 9 | 兆丰路 | 58.46 |
| 10 | 上海汽车城 | 57.88 |

**地铁站点便利度**



# 上班族与地铁

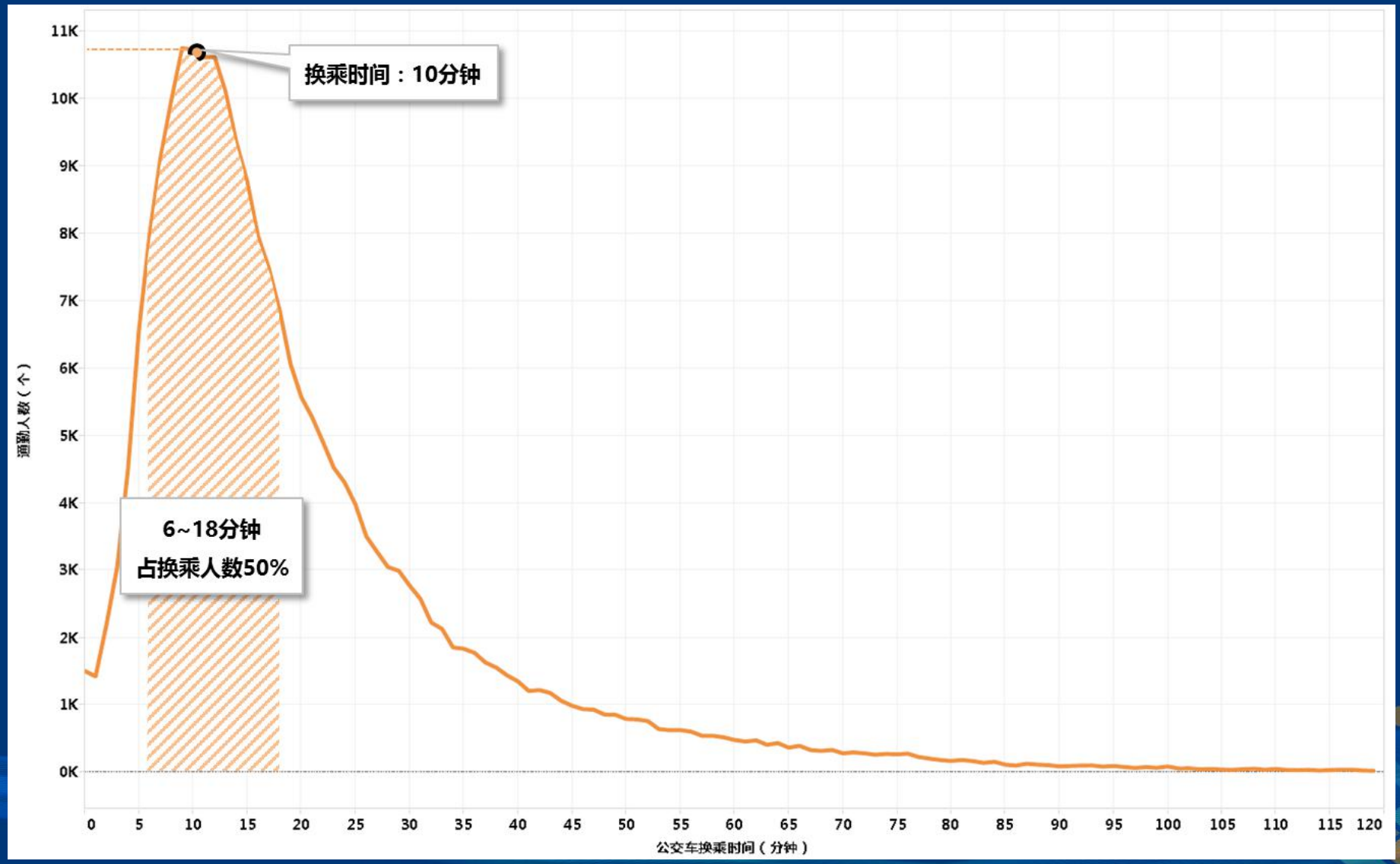


光有地铁是不够滴，公交车来支援，约有24万人坐公交车换地铁

| CardID | OTime | OStation | SpendTime |
|---|---|---|---|
| 100052870 | 7:37:07 | 4号线大连路 | 10.02 |
| 100056200 | 8:19:08 | 12号线东陆路 | 17.58 |
| 100075893 | 8:42:13 | 12号线国际客运中心 | 77.52 |
| 100077123 | 9:51:01 | 8号线江浦路 | 72.83 |
| 100085891 | 8:19:02 | 4号线浦东大道 | 75.23 |
| 100090160 | 8:12:55 | 7号线上海大学站 | 7.35 |
| 100096996 | 9:24:26 | 2号线中山公园 | 75.93 |
| 100101300 | 10:13:10 | 1号线徐家汇 | 49.02 |
| 100103682 | 7:54:25 | 1号线外环路 | 5.97 |
| 100106760 | 6:55:19 | 1号线彭浦新村 | 30.38 |
| 100116870 | 8:05:20 | 7号线美兰湖 | 38.02 |
| 100119023 | 7:24:25 | 11号线祁连山路 | 5.22 |
| 100135547 | 6:24:34 | 3号线友谊路 | 11.33 |
| 100139268 | 6:10:45 | 1号线莘庄 | 10.97 |

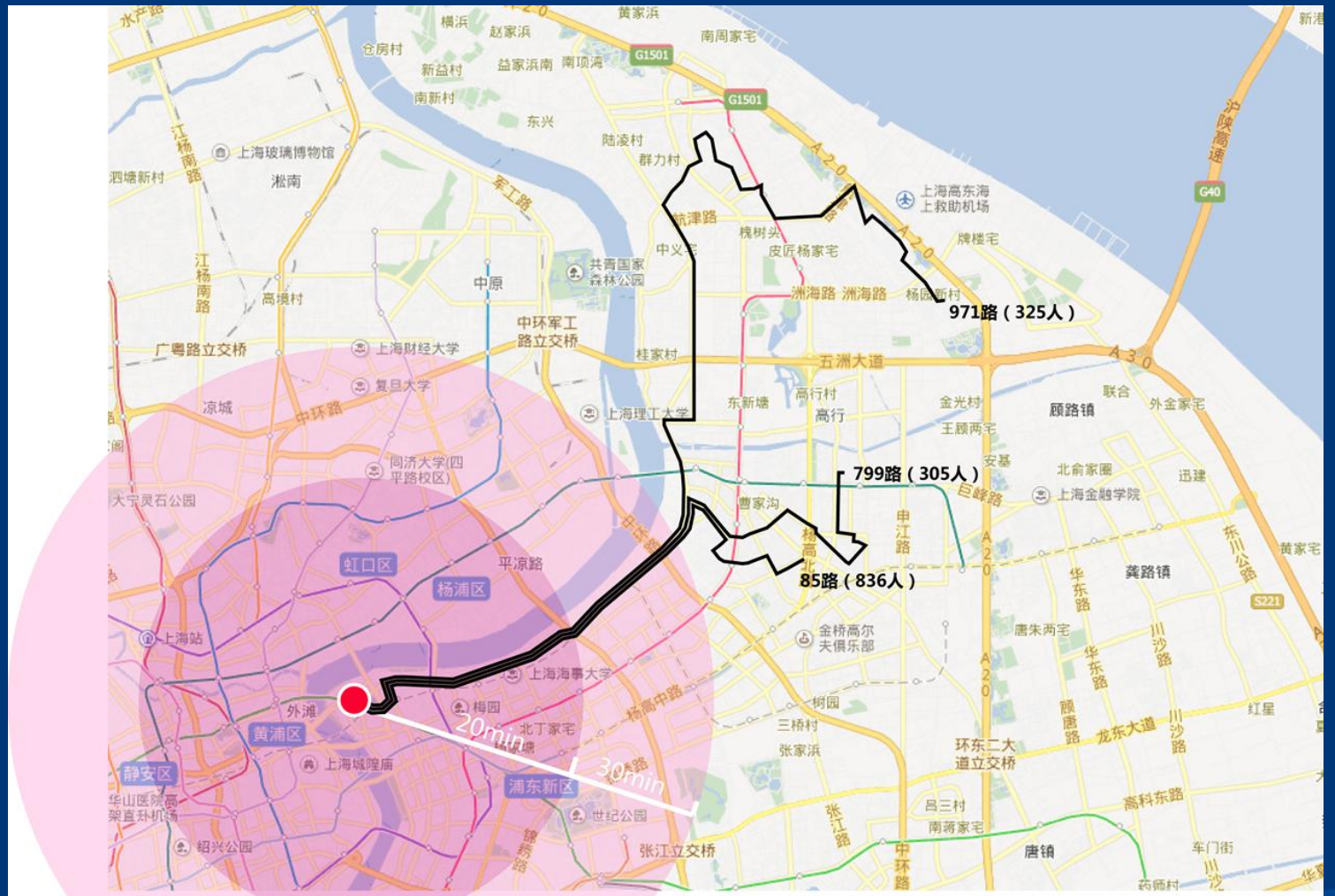
# 上班族与地铁

## 公交车上看风景时间（换乘时间）



# 上班族与地铁

宝宝不开心，我家附近没有地铁站！

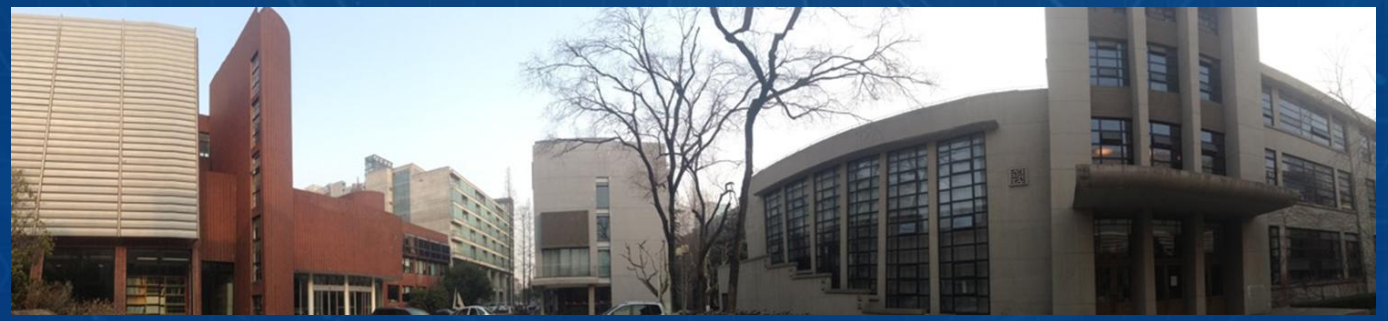


05

# 总结

- 实现多源数据，异构数据的整合
- 实时（流）数据的采集和利用
- 利用大数据计算框架实现大规模数据的分析，展示，挖掘
- 为城市现状的描述，诊断，预测，问题的解决提供了数据和相应的工具支撑

# 谢谢



曹布阳教授 15000215319 caobuyang@tongji.edu.cn 同济大学软件学院