Tid Chile

Analizando Terabytes de datos de movilidad anonimizada

Telefónica
Investigación y Desarrollo Chile

Agenda

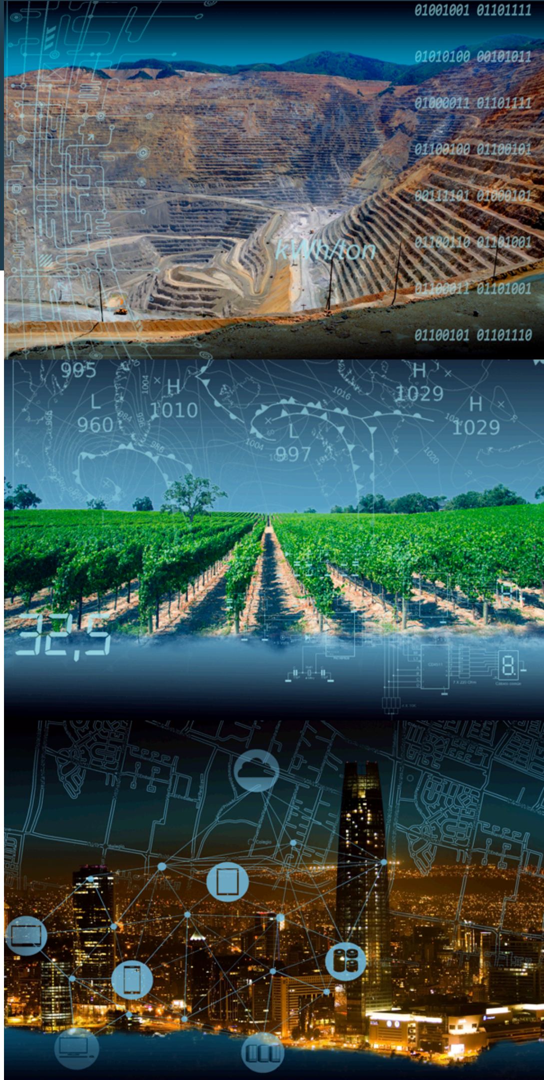
- About.
- El problema.
- Los datos y el procesamiento.
- Los papers y desafíos para los devs.
- Performance.
- Lecciones aprendidas.

About TI+D

MUVIO
movilidad urbana

OptimalCrop®

GEO
for mining



El equipo

- UX/UI
- PO
- Researchers
- Devs/QA

Problema

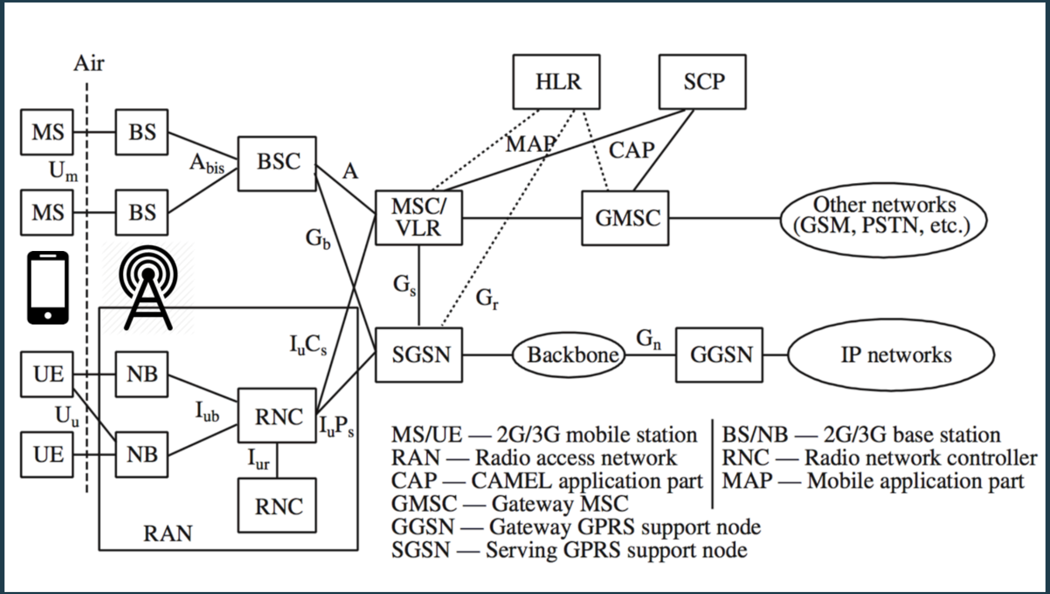
¿Se pueden usar datos de telefonía para desarrollar aplicaciones útiles para transportistas y planificadores urbanos?

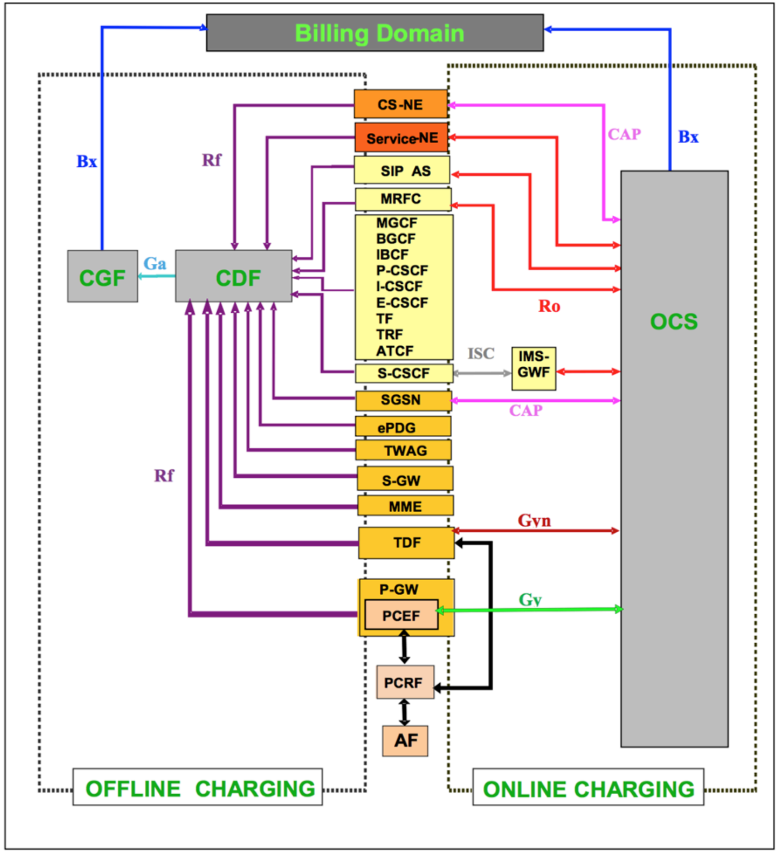
Si esto es posible, ¿qué desafíos técnicos encontraremos?

Datos

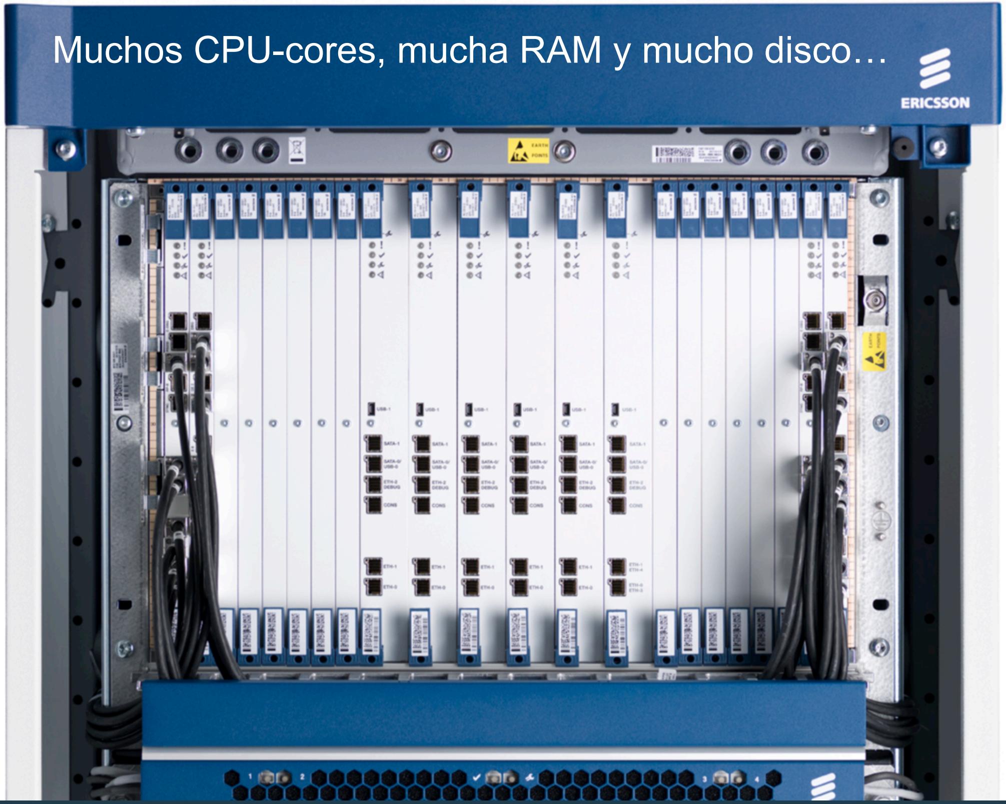
- CDR – Registros de facturación anónimos (SHA-256, métodos probabilísticos y agregación)
- Sectores o antenas 2G, 3G y 4G
- Datos públicos: censo, uso de suelo, matriz y zonas origen-destino

CDR – Registros de facturación anónimos



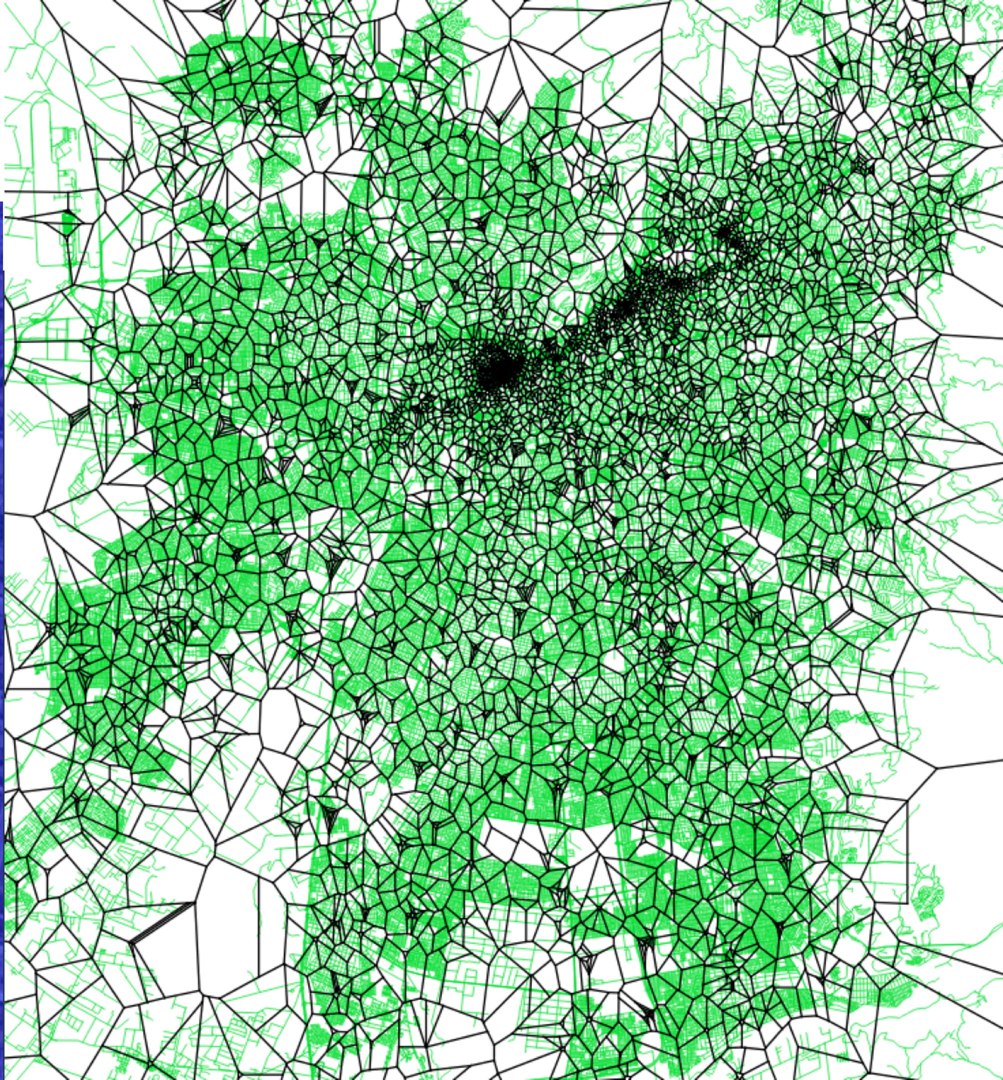


CDR – Registros de facturación anónimos



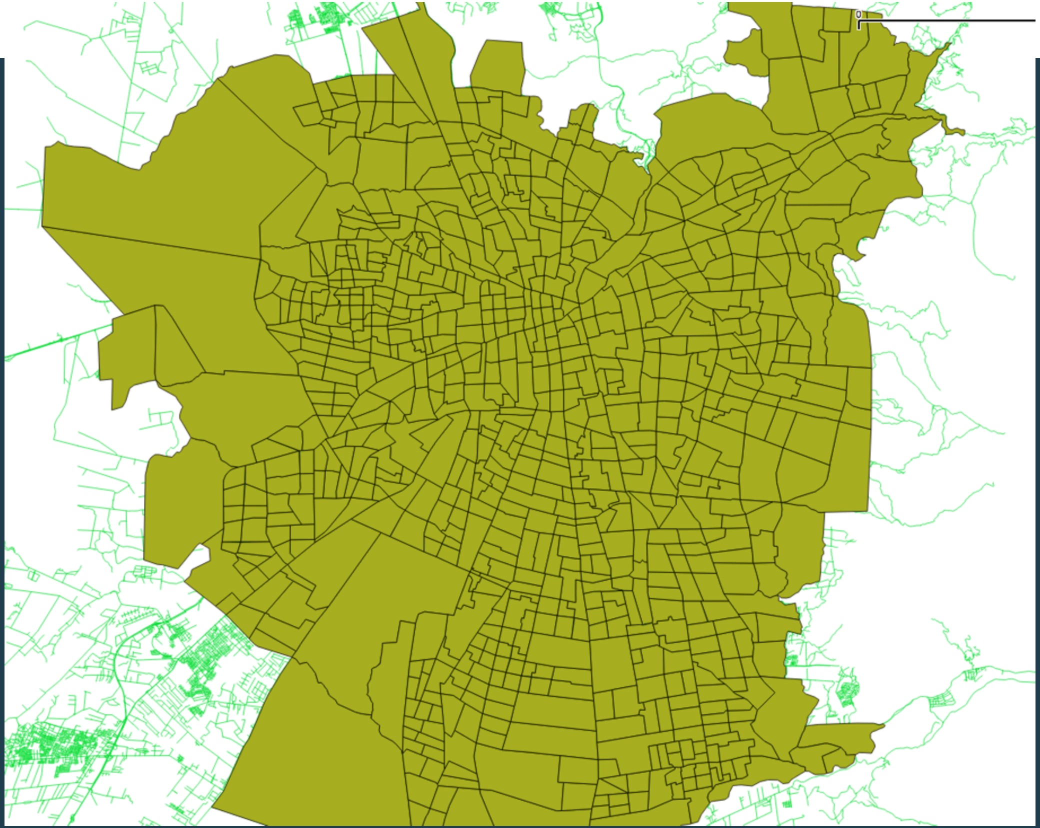
Antenas 2G/3G/4G

Antenas 2G/3G/4G



Censo / Uso de suelo

Matriz y zonas OD



Arquitectura de datos

```
                                                      +----------+ Censo
                                                      |Datos     | Uso de suelo
                                                      |Públicos  | EOD
  Terabytes                                           +----------+
                                                           |          Gigabytes
                                                           |
+----------+      +----------+     +-----------+      +----V-----+
|Datos     |      |  ETL     |     | Group by  |      |Zorzal    |
|Telco     +----->|          +---->| anon ID   +----> |(viajes)  |
+----------+      +----------+     +-----------+      +----------+
CDR                Java/Spark        Hadoop(Pig)        Python Multiprocessing
Antenas            Anonimización     Spark
                   Storage HDFS

Telefonica         Azure WASB        Azure HDInsight,
Open Cloud         y Amazon S3       Amazon EC2
y servidores                         y servidores
in-house                             in-house
```

Tid Chile

Las publicaciones.

A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow

Eduardo Graells-Garrido
Telefónica I+D
Santiago, Chile

Diego Saez-Trumper
EURECAT
Barcelona, Spain

ıtion about urban mo-

Our main contribution is a method to detect these disaggregated daily journeys using *Call Detail Records* (CDRs). Our approach

Probabilistic Mobile Phone Loca

Claudio Albornoz
Telefónica I+D
Santiago, Chile
c.albornoz.flores@gmail.com

***Abstract*—A probabilistic method to infer mobil location is presented. By using street maps and tir vance values as prior information, probability map: constructed in the coverage areas of the cell sectors (ai The probability maps are used to generate simulated of each mobile phone inside the coverage area wit precision. The models can be improved by adding (Strength Signal measurements in the area. A case stu event-driven mobile phone data sets is presented, par using Charging Data Records and Internet usage (mid-size and large-size cities of Chile.**

***Keywords*-GPS simulation; Anonimization; Event-dr bile phone data;**

I. Introduction

Telecommunication cell sectors, commonly kr antennas, are widely deployed in almost every cit largest portion of the world. Seven billion people (95 global population) live in an area that is covered by a network with GSM or modern technologies [1]. Sin Telecom operators are starting to use their subscrib as a proxy of urban mobility, increasing interest ha

sensors MDPI

Article

Sensing Urban Patterns with Antenna Mappings: The Case of Santiago, Chile †

Eduardo Graells-Garrido [1,2,*], Oscar Peredo [2] and José García [2]

1 Data Science Institute; Faculty of Engineering, Universidad del Desarrollo, Las Condes 7610658, Chile
2 Telefónica I+D; Av. Manuel Montt 1404, Third Floor, Providencia, Providencia 7501105, Chile; oscar.peredoa@telefonica.com (O.P.); joseantonio.garcia@telefonica.com (J.G.)
* Correspondence: egraells@udd.cl; Tel.: +56-2-232-791-10

† This paper is an extended version of a paper published in "Graells-Garrido, E.; García, J. Visual Exploration of Urban Dynamics Using Mobile Data. In *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information*, Proceedings of the 9th International Conference (UCAmI 2015), Puerto Varas, Chile, 1–4 December 2015; Springer International Publishing: Cham, Switzerland, 2015; pp. 480–491".

Academic Editors: Vladimir Villarreal and Carmelo R. García
Received: 5 May 2016; Accepted: 4 July 2016; Published: 15 July 2016

Abstract: Mobile data has allowed us to sense urban dynamics at scales and granularities not known before, helping urban planners to cope with urban growth. A frequently used kind of dataset are Call Detail Records (CDR), used by telecommunication operators for billing purposes. Being an already extracted and processed dataset, it is inexpensive and reliable. A common assumption with respect to geography when working with CDR data is that the position of a device is the same as the Base Transceiver Station (BTS) it is connected to. Because the city is divided into a square grid, or by coverage zones approximated by Voronoi tessellations, CDR network events are assigned to corresponding areas according to BTS position. This geolocation may suffer from non negligible error in almost all cases. In this paper we propose "Antenna Virtual Placement" (AVP), a method to geolocate mobile devices according to their connections to BTS, based on decoupling antennas

Urban Dynamics Estimation using Mobile Phone logs and Locally Varying Anisotropy

Oscar F. Peredo[1], José A. García[1], Ricardo Stuven[1] and Julián M. Ortiz[2]

Abstract In telecommunications, the billing data of each telephone, denoted Call Detail Records (CDR), are a large and rich database with information that can be geo-located. By analyzing the events logged in each antenna, a set of time series can be constructed measuring the number of voice and data events in each time of the day. One question that can be addressed using these data involves estimating the movement or flow of people in the city, which can be used for prediction and monitoring in transportation or urban planning. In this work, geostatistical estimation techniques such as kriging and inverse distance weighting are used to numerically estimate the flow of people. In order to improve the accuracy of the model, secondary information is included in the estimation. This information represents the locally varying anisotropy (LVA) field associated with the major streets and roads in the city. By using this technique, the flow estimation can be obtained with a better quantitative and qualitatively interpretation. In terms of storage and computing power, the volume of raw information is extremely large, for that reason Big Data technologies are mandatory to query the database. Additionally, if high-resolution grids are used in the estimation, High Performance Computing techniques are necessary to speed up the numerical computations using LVA codes. Case studies are shown, using voice/data records from more than 1 million clients of Telefónica Movistar in Santiago, capital of Chile.

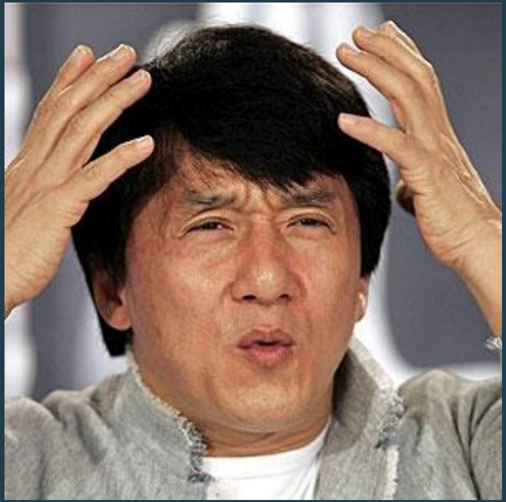
Introduction

The usage and ubiquity of mobile phones in all countries and social groups are generating an unprecedented amount of behavioral data. According to the International Telecommunications Union [1], by the end of 2015 there are more than 7 billion mobile cellular subscriptions, corresponding to a penetration rate of 97%,


I HAVE NO IDEA
WHAT I'M DOING

Expectativas vs Realidad

- **Data sample vs data real.**
- **Aplicaciones vs notebooks de Jupyter.**
- **¿Cuándo tenemos el MVP? (Product Owner).**



Arquitectura de datos

```
                                                        +----------+ Censo
                                                        |Datos     | Uso de suelo
                                                        |Públicos  | EOD
 Terabytes                                              +----------+
                                                             |
                                                             |          Gigabytes
   +----------+        +----------+       +-----------+     +----V-----+
   |Datos     |        |  ETL     |       | Group by  |     |Zorzal    |
   |Telco     +------> |          +---->  | anon ID   +---->|(viajes)  |
   +----------+        +----------+       +-----------+     +----------+
   CDR                  Java/Spark          Hadoop(Pig)       Python Multiprocessing
   Antenas              Anonimización       Spark
                        Storage HDFS

   Telefonica           Azure WASB          Azure HDInsight,
   Open Cloud           y Amazon S3         Amazon EC2
   y servidores                             y servidores
   in-house                                 in-house
```

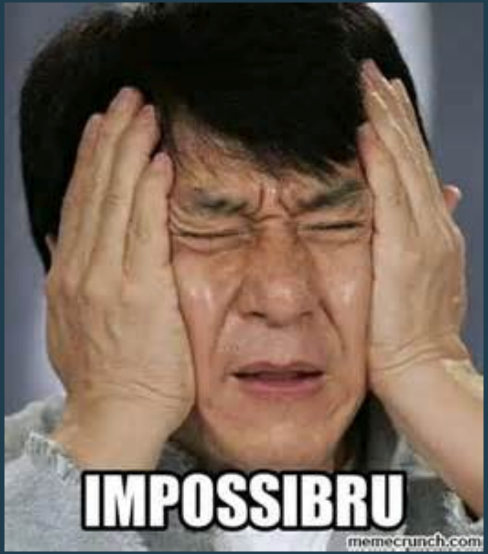
Tid Chile

Arquitectura v.1 Muvio

```
+---------+          +---------+            +-------+             +-----------+
|         |          |         |            |       |             |           |
| Zorzal  +--------->| Liebre  +----------->| API   +------------>| Frontend  |
|         |          |         |            |       |             |           |
+---------+          +---------+            +-------+             +-----------+
Batch ETL            Microservice           Backend API
(Python Pandas)      (Python Tornado)       Django
                                                ^
                                                |
                                        +---------------+
                                        |               |
                                        |               |
                                     +--v---+       +---v----+
                                     |      |       |        |
                                     | DB   |       | Cache  |
                                     |      |       |        |
                                     +------+       +--------+
```

Nuevo dataset - Santiago

- **Tiempos respuesta > 2 min. (400M registros).**
- **Nuevas agregaciones y filtros.**



Arquitectura v.2 Muvio

```
+---------+          +-------------------------+          +-------+           +-----------+
|         |          |                         |          |       |           |           |
| Zorzal  +--------->+          Rocket         +--------->| API   +---------->| Frontend  |
|         |          |                         |          |       |           |           |
+---------+          +-------------------------+          +-------+           +-----------+
Batch ETL             API Gateway +                       Backend API
(Python Numpy         Lambda +                            Django
 Multiprocessing)     Athena (Presto) -> ORC                  ^
                                                              |
                                                      +---------------+
                                                      |               |
                                                      |               |
                                                   +--v--+         +---v---+
                                                   |     |         |       |
                                                   | DB  |         | Cache |
                                                   |     |         |       |
                                                   +-----+         +-------+
```

MUVIO

MIS SETS DE DATOS | HISTORIAL DE CONSULTAS | METODOLOGÍA | PLANES

Centro de ayuda | Cerrar sesión

Limpiar todos los filtros

Crear consulta

Consulta sin título

Lugar

Elegir origen y destino

Fechas

03/10/2016 | 09/10/2016

Días laborales

☑ Lun ☑ Mar ☑ Mie ☑ Jue ☑ Vie

Fin de semana

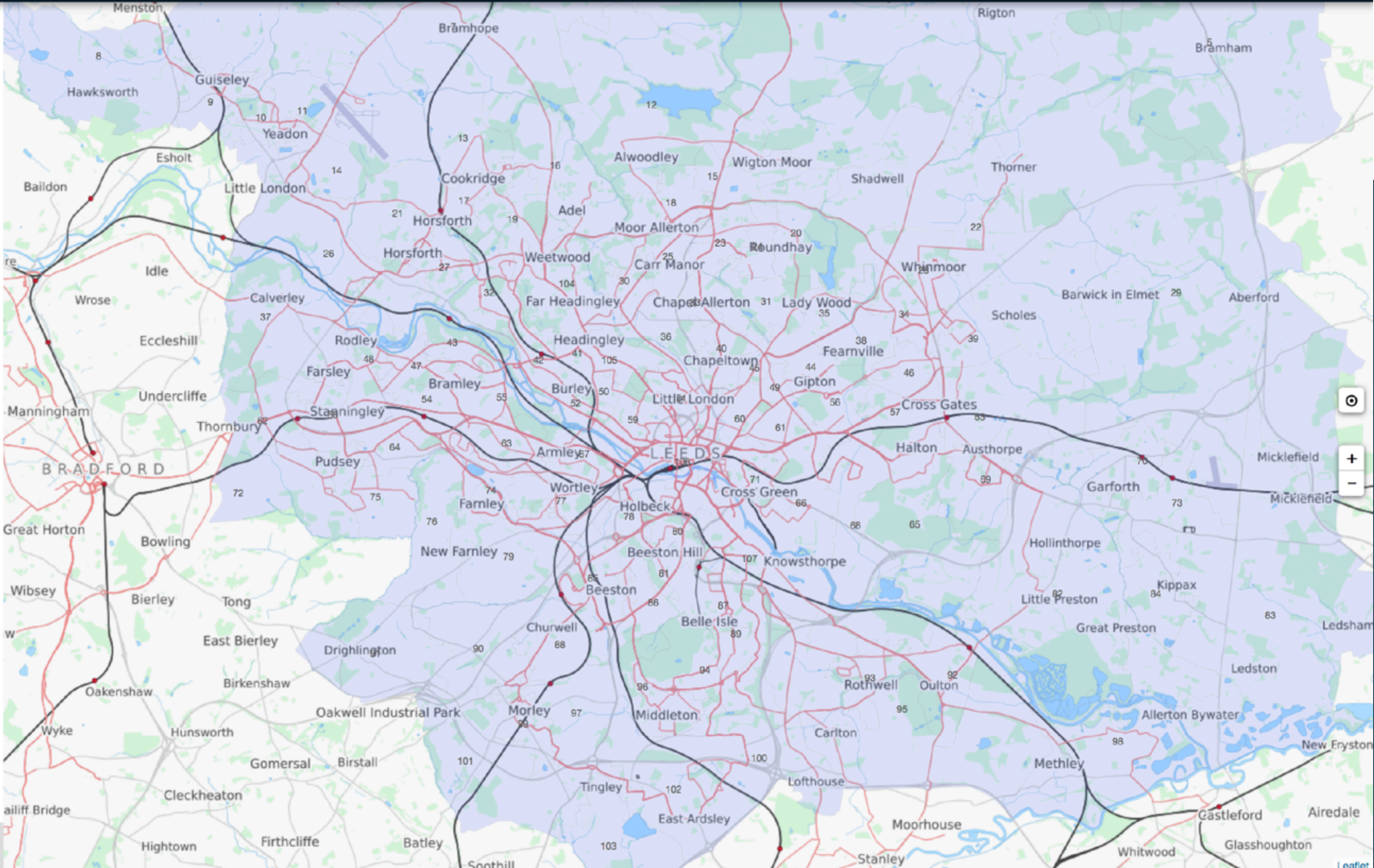
☑ Sáb ☑ Dom

Horario

Inicio 00:00

Fin 23:59

Realizar consulta



Tid Chile

Lecciones Aprendidas (Devs)

- ¿Optimizo antes, durante o después?
- Pandas != High performance
- Formatos de archivos en storage.

Tid Chile

DEMO

Gracias por tu tiempo!
Tienes Preguntas?

https://muvio.cl/