Procesamiento y análisis de viajes en taxi

Trabajo fin de master - III Master de arquitectura Big Data

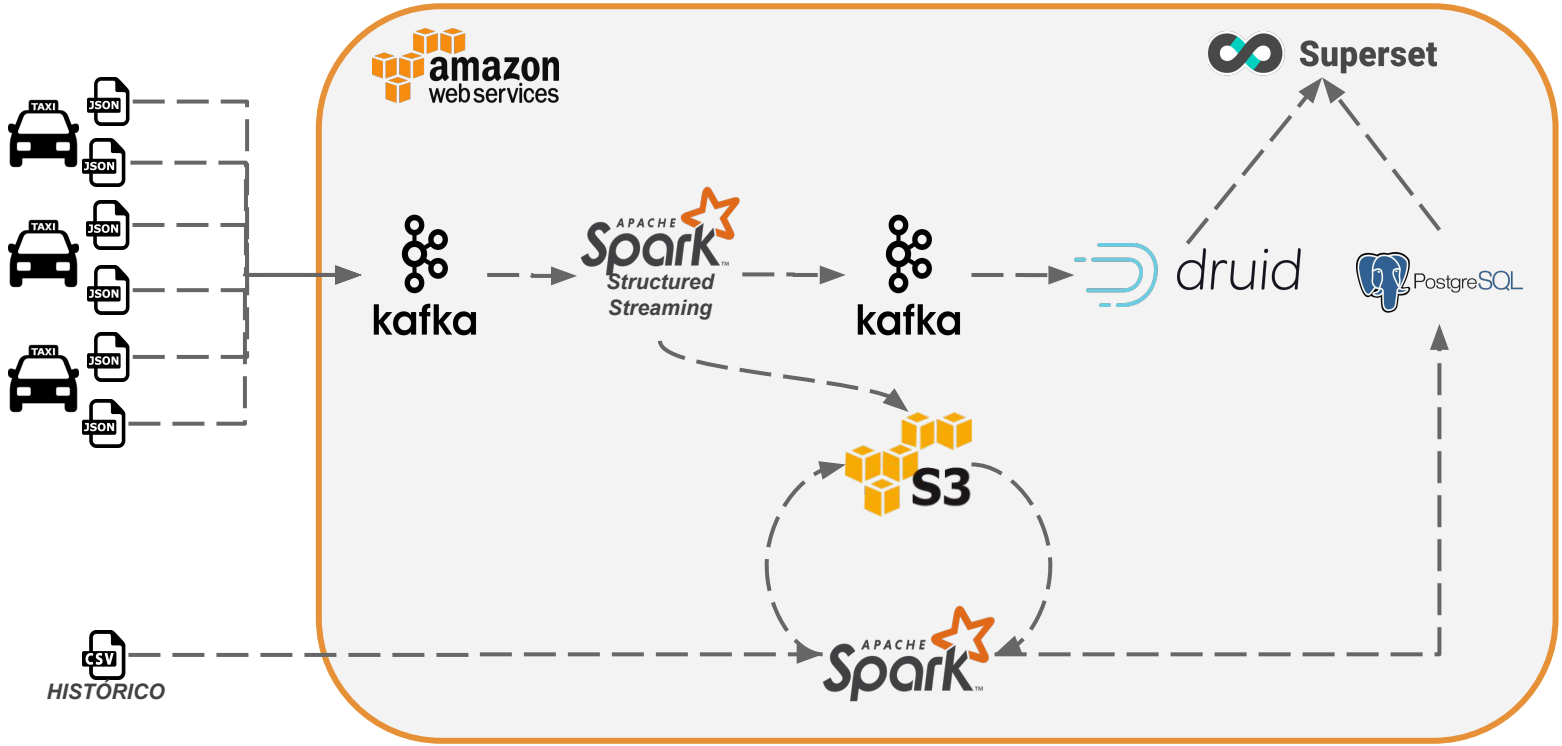
KSCHOOL

Requisitos

Se quiere disponer un sistema con el que poder:

- ❏ Procesar, en tiempo real, la información de cada viaje enviada por los taxis.
- ❏ Ejecutar de distintos tipos de consultas sobre la información en tiempo real. Por ejemplo, conocer el número de taxis activos total y por zonas o compañías, volumen total de viajes y por zonas o compañías en las últimas horas.
- ❏ Analizar el histórico completo de datos, para poder ejecutar consultas y obtener distintas estadísticas de actividad de los taxis. Por ejemplo, duración y coste medio de los viajes, empresas de taxis con mayor volumen de negocio, zonas donde más viajes empiezan, zonas en las que terminan más viajes.
- ❏ Consultar la información, tanto la tratada en tiempo real como la del histórico, desde una herramienta de visualización. Mostrando diferentes DashBoards.

Arquitectura



Spark - procesos

- StreamingTaxiTrips.py: tiene como objetivo procesar los mensajes enviados, en formato JSON por los taxis en tiempo real, enviarlos a kafka y almacenarlos en S3 para su posterior procesamiento con el resto del histórico.

- IngestHistoricTrips.py: tiene como objetivo almacenar en S3 los datos históricos de los viajes en taxi.

- TransformTaxiTrips.py: tiene como objetivo procesar los viajes en taxi almacenados en S3, tanto por el batch como por el streaming, para almacenarlos en PostgreSQL y que puedan ser consultados desde Superset.

- AreasLoc.py: tiene como objetivo generar el “maestro de áreas”, que será usado para enriquecer tanto los datos en tiempo real como los históricos.

Spark - procesos

Proceso	***Periodicidad***
StreamingTaxiTrips.py	Real-time
IngestHistoricTrips.py	Puntual
TransformTaxiTrips.py	Diaria
AreasLoc.py	Puntual

Druid - indexación

- kafkaIngestionTaxiTrips.json: tarea de indexación que lee, de un topic de kafka, los datos enviados por los taxis, previamente procesados en el proceso Spark StreamingTaxiTrips.py.

Configuración	***Valor***
segmentGranularity	fifteen_minute
queryGranularity	NONE
taskDuration	PT15M

Modelo de datos - Real time (DRUID)

taxi-trips	
Columna	**Descripción**
Dimensiones	
taxi_id	Id del taxi.
company	Empresa del taxi
pickup_community_area	Área de inicio del viaje.
pickup_community_area_name	Nombre del área de inicio del viaje.
pickup_centroid_latitude	Latitud del punto central del área de incio.
pickup_centroid_longitude	Longitud del punto central del área de incio.
dropoff_community_area	Area de fin del viaje.
dropoff_community_area_name	Nombre del área de inicio del viaje.
dropoff_centroid_latitude	Latitud del punto central del área de final.
dropoff_centroid_longitude	Longitud del punto central del área de final.

taxi-trips	
Columna	**Descripción**
Metricas	
trips	Número de viajes.
triptotal_sum	Suma del coste total del viaje.
trip_seconds_sum	Suma del tiempo de viaje.
trip_miles_sum	Suma de la distancia de viaje.
fare_sum	Sumas de las tarifa.
tips_sum	Suma de las propinas.
tolls_sum	Suma de los peajes.
extras_sum	Suma de los extras.

Modelo de datos - Histórico (PostgreSQL)

companies_pickup_area_view_[YEAR]	
Columna	**Descripción**
trip_start_date	Fecha de incio del viaje
company	Empresa de taxis.
pickup_community_area	Area de inicio del viaje.
pickup_community_area_name	Nombre del area de inicio del viaje.
pickup_centroid_latitude	Latitud del punto central del area de incio.
pickup_centroid_longitude	Longitud del punto central del area de incio.
fares	Sumas de las tarifa.
tips	Suma de las propinas.
tolls	Suma de los peajes.
extras	Suma de los extras.
trip_totals	Suma del coste total del viaje.
trips	Número de viajes.
taxis	Número de taxis activos.

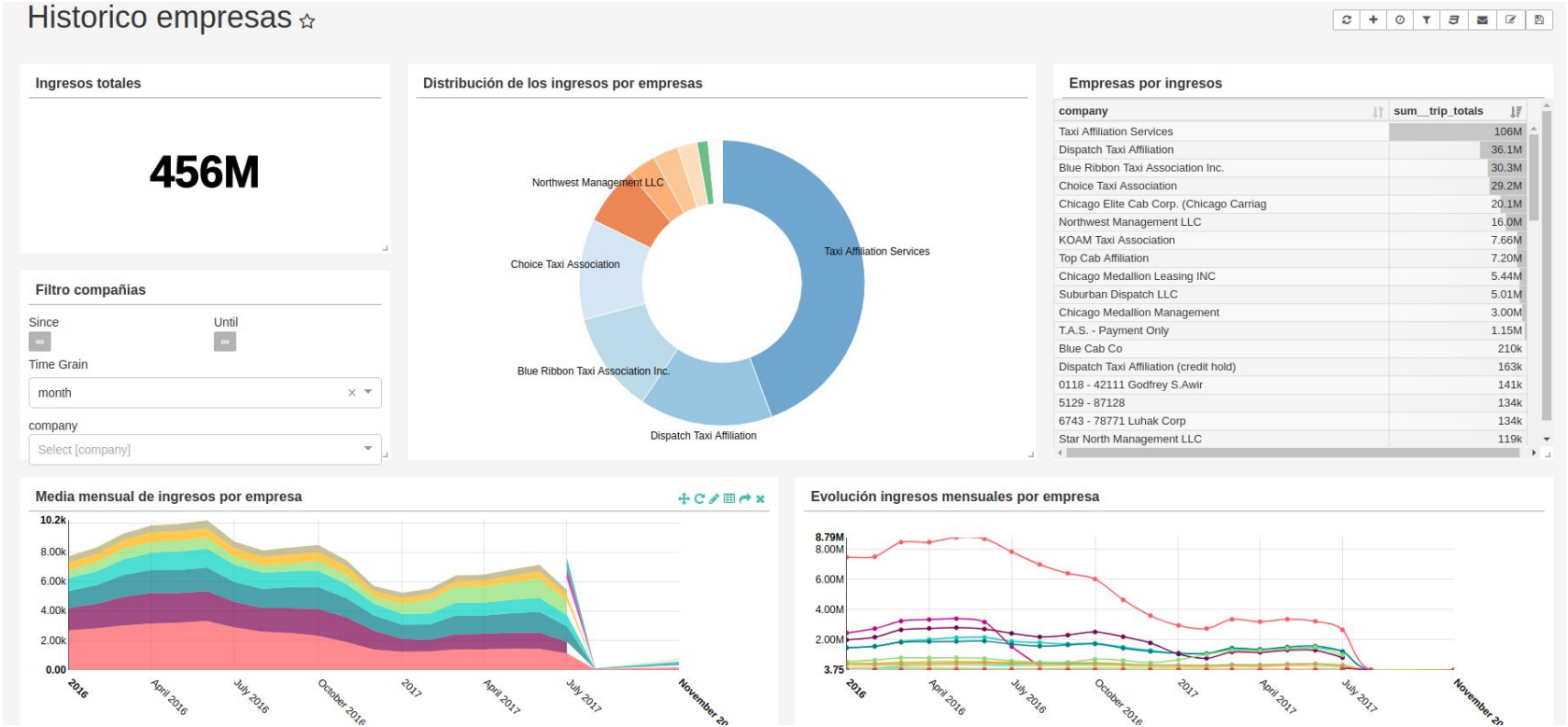
pickup_area_view_[YEAR]	
Columna	**Descripción**
trip_start_date	Fecha de incio del viaje
pickup_community_area	Area de inicio del viaje.
pickup_community_area_name	Nombre del area de inicio del viaje.
pickup_centroid_latitude	Latitud del punto central del area de incio.
pickup_centroid_longitude	Longitud del punto central del area de incio.
fares	Sumas de las tarifa.
tips	Suma de las propinas.
tolls	Suma de los peajes.
extras	Suma de los extras.
trip_totals	Suma del coste total del viaje.
trips	Número de viajes.
taxis	Número de taxis activos.

Modelo de datos - Histórico (PostgresQL)

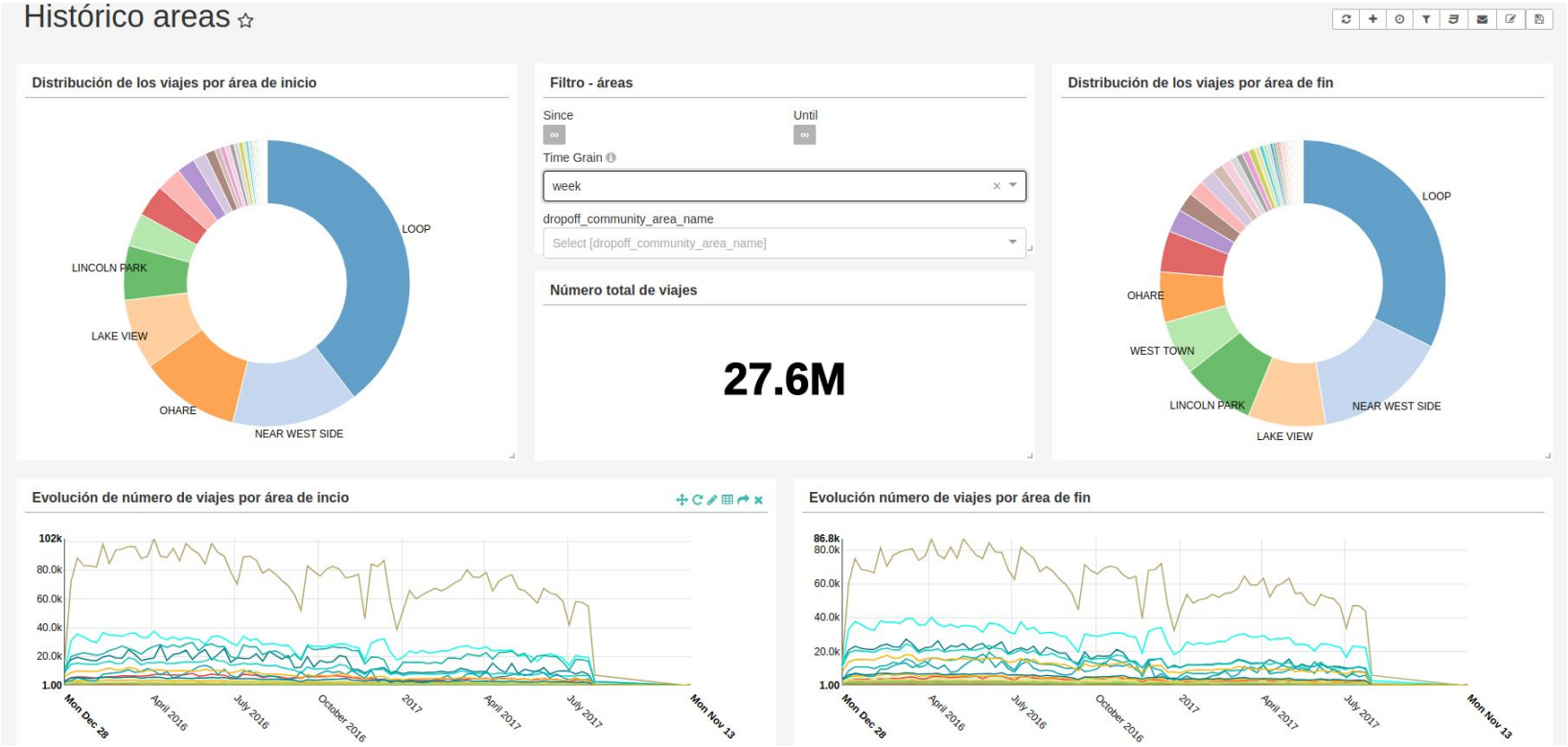
companies_dropoff_area_view_[YEAR]	
Columna	**Descripción**
trip_start_date	Fecha de incio del viaje
company	Empresa de taxis.
dropoff_community_area	Area de final del viaje.
dropoff_community_area_name	Nombre del area de final del viaje.
dropoff_centroid_latitude	Latitud del punto central del area de fin.
dropoff_centroid_longitude	Longitud del punto central del area de fin.
fares	Sumas de las tarifa.
tips	Suma de las propinas.
tolls	Suma de los peajes.
extras	Suma de los extras.
trip_totals	Suma del coste total del viaje.
trips	Número de viajes.
taxis	Número de taxis activos.

dropoff_area_view_[YEAR]	
Columna	**Descripción**
trip_start_date	Fecha de incio del viaje
dropoff_community_area	Area de inicio del viaje.
dropoff_community_area_name	Nombre del area de inicio del viaje.
dropoff_centroid_latitude	Latitud del punto central del area de incio.
dropoff_centroid_longitude	Longitud del punto central del area de incio.
fares	Sumas de las tarifa.
tips	Suma de las propinas.
tolls	Suma de los peajes.
extras	Suma de los extras.
trip_totals	Suma del coste total del viaje.
trips	Número de viajes.
taxis	Número de taxis activos.

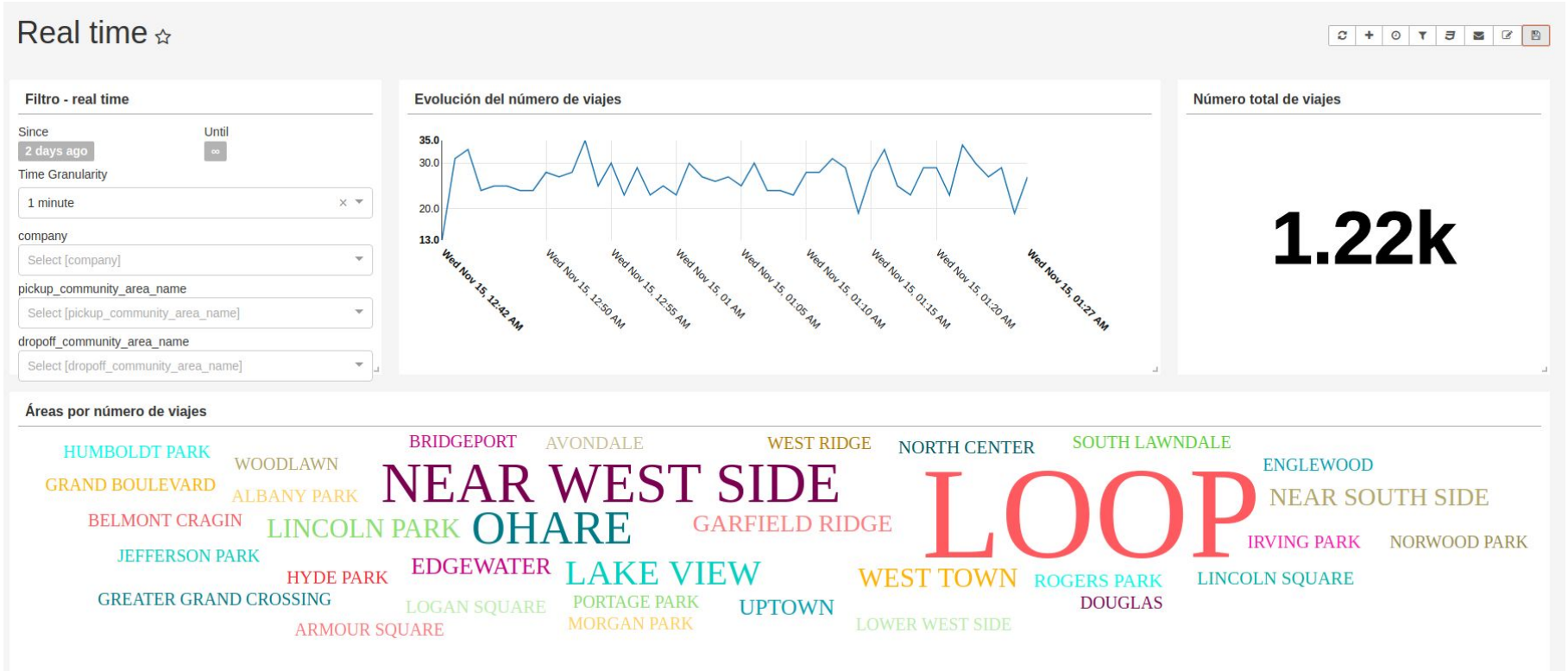
Superset


Historico empresas
Ingresos totales
456M
Filtro compañias
Since
Until
Time Grain
month
company
Select [company]
Distribución de los ingresos por empresas
Northwest Management LLC
Choice Taxi Association
Blue Ribbon Taxi Association Inc.
Dispatch Taxi Affiliation
Taxi Affiliation Services
Empresas por ingresos
company
sum__trip_totals
Taxi Affiliation Services 106M
Dispatch Taxi Affiliation 36.1M
Blue Ribbon Taxi Association Inc. 30.3M
Choice Taxi Association 29.2M
Chicago Elite Cab Corp. (Chicago Carriag 20.1M
Northwest Management LLC 16.0M
KOAM Taxi Association 7.66M
Top Cab Affiliation 7.20M
Chicago Medallion Leasing INC 5.44M
Suburban Dispatch LLC 5.01M
Chicago Medallion Management 3.00M
T.A.S. - Payment Only 1.15M
Blue Cab Co 210k
Dispatch Taxi Affiliation (credit hold) 163k
0118 - 42111 Godfrey S.Awir 141k
5129 - 87128 134k
6743 - 78771 Luhak Corp 134k
Star North Management LLC 119k
Media mensual de ingresos por empresa
10.2k
8.00k
6.00k
4.00k
2.00k
0.00
2016
April 2016
July 2016
October 2016
2017
April 2017
July 2017
Evolución ingresos mensuales por empresa
8.79M
8.00M
6.00M
4.00M
2.00M
3.75
2016
April 2016
July 2016
October 2016
2017
April 2017
July 2017

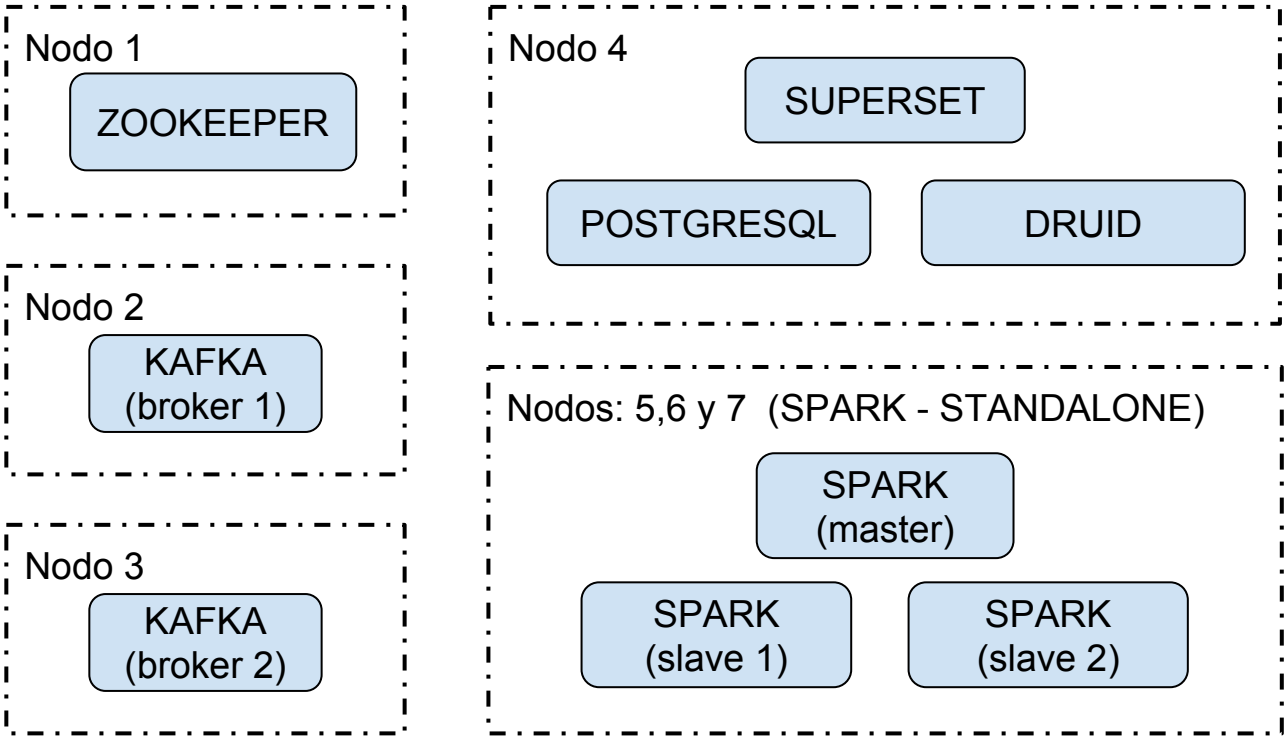
Superset


Histórico areas
Distribución de los viajes por área de inicio
LOOP
LINCOLN PARK
LAKE VIEW
OHARE
NEAR WEST SIDE
Filtro - áreas
Since
Until
Time Grain
week
dropoff_community_area_name
Select [dropoff_community_area_name]
Número total de viajes
27.6M
Distribución de los viajes por área de fin
LOOP
OHARE
WEST TOWN
LINCOLN PARK
NEAR WEST SIDE
LAKE VIEW
Evolución de número de viajes por área de incio
102k
80.0k
60.0k
40.0k
20.0k
1.00
Mon Dec 28
April 2016
July 2016
October 2016
2017
April 2017
July 2017
Mon Nov 13
Evolución número de viajes por área de fin
86.8k
80.0k
60.0k
40.0k
20.0k
1.00
Mon Dec 28
April 2016
July 2016
October 2016
2017
April 2017
July 2017
Mon Nov 13

Superset


Real time
Filtro - real time
Since
2 days ago
Until
∞
Time Granularity
1 minute
company
Select [company]
pickup_community_area_name
Select [pickup_community_area_name]
dropoff_community_area_name
Select [dropoff_community_area_name]
Evolución del número de viajes
35.0
30.0
20.0
13.0
Wed Nov 15, 12:42 AM
Wed Nov 15, 12:50 AM
Wed Nov 15, 12:55 AM
Wed Nov 15, 01 AM
Wed Nov 15, 01:05 AM
Wed Nov 15, 01:10 AM
Wed Nov 15, 01:15 AM
Wed Nov 15, 01:20 AM
Wed Nov 15, 01:27 AM
Número total de viajes
1.22k
Áreas por número de viajes
HUMBOLDT PARK
WOODLAWN
BRIDGEPORT
AVONDALE
WEST RIDGE
NORTH CENTER
SOUTH LAWNDALE
ENGLEWOOD
GRAND BOULEVARD
ALBANY PARK
NEAR WEST SIDE
LOOP
NEAR SOUTH SIDE
BELMONT CRAGIN
LINCOLN PARK
OHARE
GARFIELD RIDGE
IRVING PARK
NORWOOD PARK
JEFFERSON PARK
HYDE PARK
EDGEWATER
LAKE VIEW
WEST TOWN
ROGERS PARK
LINCOLN SQUARE
GREATER GRAND CROSSING
LOGAN SQUARE
PORTAGE PARK
UPTOWN
DOUGLAS
ARMOUR SQUARE
MORGAN PARK
LOWER WEST SIDE

AWS


Nodo 1
ZOOKEEPER
Nodo 2
KAFKA
(broker 1)
Nodo 3
KAFKA
(broker 2)
Nodo 4
SUPERSET
POSTGRESQL
DRUID
Nodos: 5,6 y 7 (SPARK - STANDALONE)
SPARK
(master)
SPARK
(slave 1)
SPARK
(slave 2)

Problemas / Soluciones

- Consultas desde Superset: lentitud de las consultas a PostgreSQL. / Ubicación del servidor de PostgreSQL en la misma máquina que Superset.

- S3: lentitud en la escritura cuando existen muchas carpetas, ralentizaba en gran medida la finalización del job de Spark. / Inicialmente se particionaba la información por Año/Mes/Día, se resolvió el problema, eliminado el día de la partición quedado por Año/Mes.

- AWS: Configuración de los distintos servicios en el cluster.

DEMO