Surge pricing & traffic congestion estimation using NYC-Taxi data

By Amith Das

Contents

- ▶ Introduction of use case
- ▶ NYC Dataset - Dimensions
- ▶ Approach for Surge-pricing estimation
- ▶ Approach for traffic congestion estimation
- ▶ Technical Process flow diagram
- ▶ Surge-pricing estimation results
- ▶ Traffic congestion estimation results
- ▶ Weather effect on surge pricing and traffic congestion
- ▶ Code walk through – Live Demo

Introduction of use case

▶ Surge pricing estimation based on supply-demand ratio at each location

    ▶ Supply – Number of drivers arriving at each location

    ▶ Demand – Number of booking requests at each location

▶ Traffic Congestion based on average speed at each location

▶ Effect of weather on surge pricing and traffic congestion

▶ Batch and Streaming versions for both surge pricing and traffic congestion
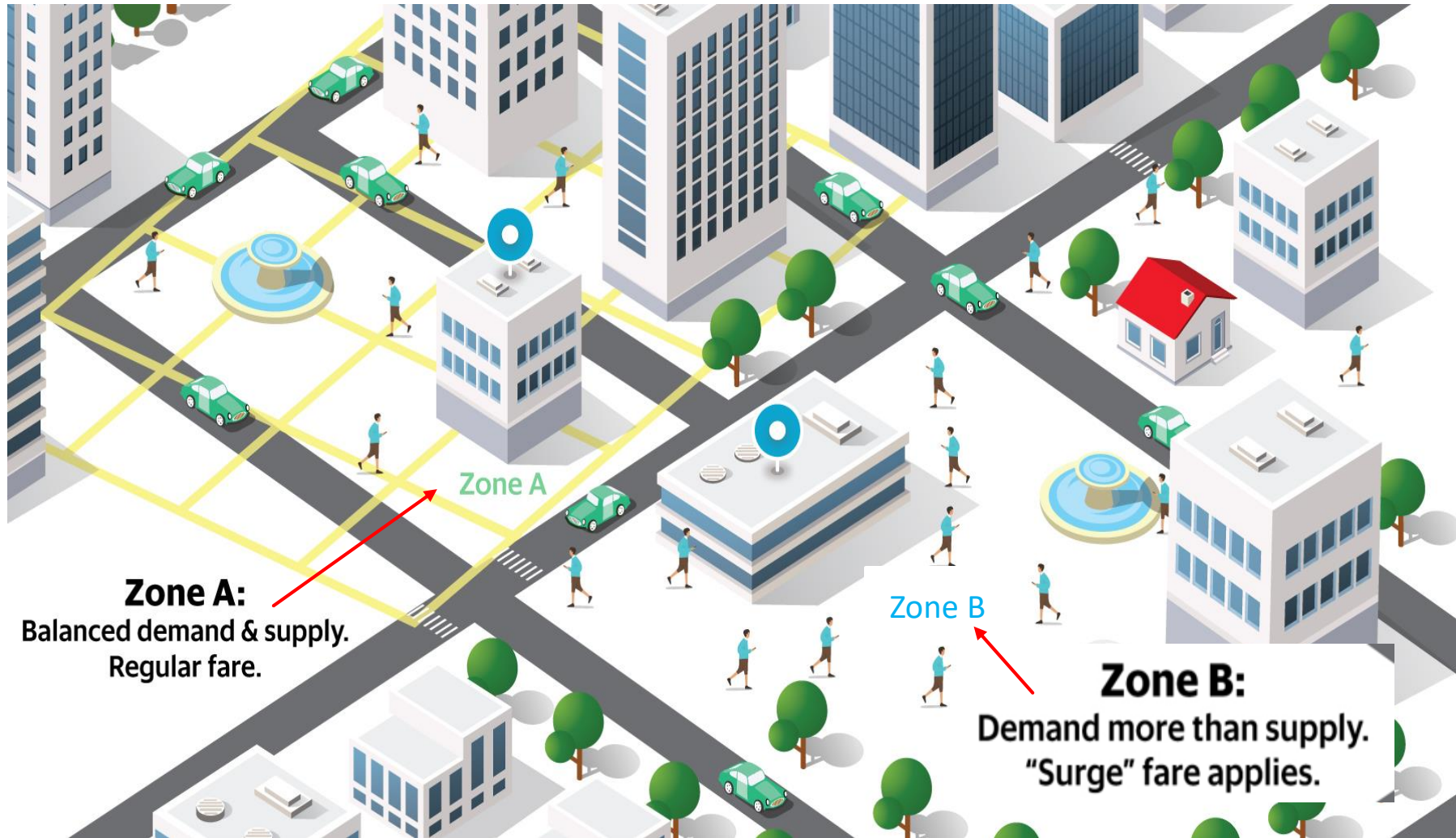
NYC Dataset – Dimensions

▶ **Dimensions:**

    ▶ pickup_datetime,dropoff_datetime,pickup_longitude,pickup_latitude,dropoff_longitude ,dropoff_latitude,trip_duration

pickup_datetime	dropoff_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	trip_duration
01-01-2016 00:00	01-01-2016 00:14	-73.98174286	40.71915817	-73.93882751	40.82918167	849
01-01-2016 00:00	01-01-2016 00:22	-73.98508453	40.74716568	-73.95803833	40.71749115	1294
01-01-2016 00:01	01-01-2016 00:03	-73.97333527	40.76407242	-73.97485352	40.76173401	114
01-01-2016 00:01	01-01-2016 00:21	-73.99310303	40.75263214	-73.9539032	40.81653976	1204

▶ **Derived Dimensions:**

    ▶ geohash,trip_distance,avg_speed_per_day,avg_speed_per_hour,weather_condition

Surge Pricing Example

Zone A:
Balanced demand & supply.
Regular fare.

Zone B:
Demand more than supply.
"Surge" fare applies.

Approach for surge pricing estimation

▶ Geohash Conversion (Precision Level 6)

▶ Cleaning outliers (zero distance trip, Distance > 200 km etc.)

▶ Surge pricing at geohash = (supply / demand)

  ▶ Supply – no of drop offs at each geohash

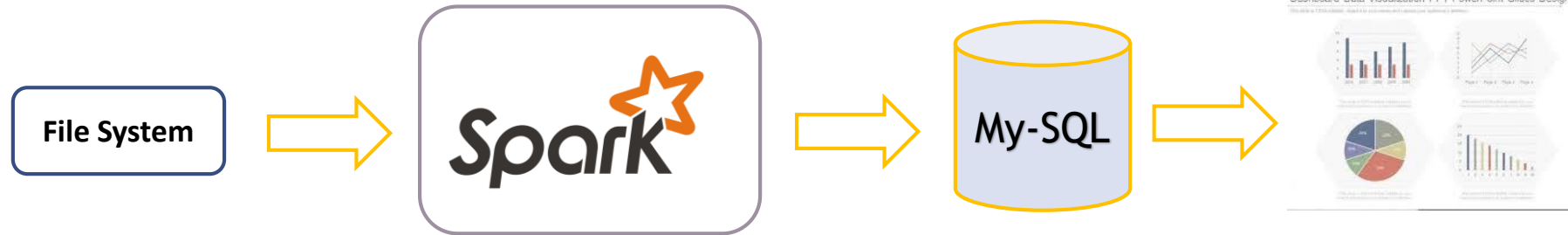 ▶ Demand – no of pickup requests at each geohash

▶ Sample output :

geo-hash	trip_date	trip_hour	avilable cabs	booking requests	surge price
hfugpk	07-01-2016	17	11	12	1.09
hfugp8	28-01-2016	18	7	1	1
hfufyx	15-01-2016	13	1	8	5

Approach for traffic congestion estimation

- Traffic congestion – By average speed difference

  - Speed diff = (avg speed per day – avg speed per hour)

  - If Speed diff > 10   – Heavy traffic on that hour

  - If Speed diff between 5 and 10 - Moderate traffic

  - If speed diff < 5   - Less traffic

- Traffic congestion – By cab density

  - Density diff= (avg cab density– avg density per hour)

  - If Density diff > 10   – Heavy traffic on that hour

  - If Density diff between 5 and 10 - Moderate traffic

  - If Density diff < 5   - Less traffic

Technical Process flow diagram

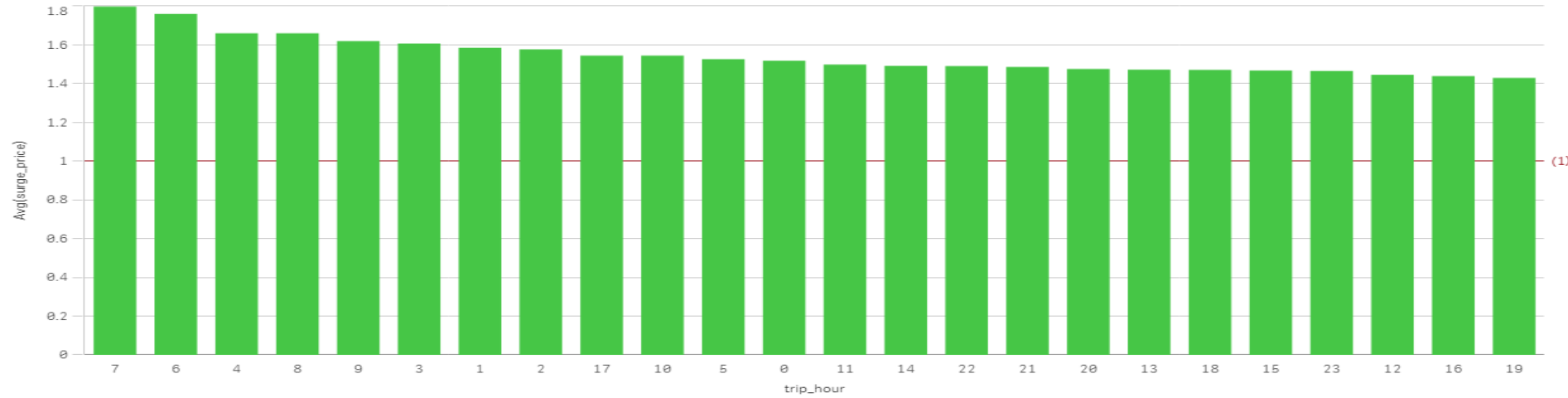▶ Batch version:

| File System | → | **Spark** | → | My-SQL | → | Qlik Sense |

▶ Streaming Version

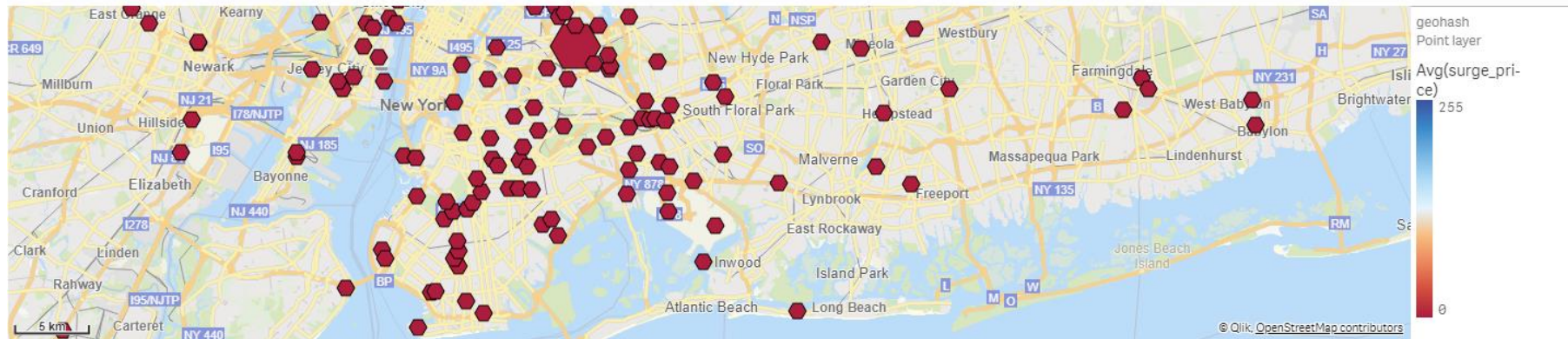| Kafka | → | **Spark** *Streaming* | → | My-SQL | → | Qlik Sense |

Surge pricing estimation results

Hourly Surge pricing by GeoHash

Geo Hash Plotting Based on Surge price

Traffic Congestion Results
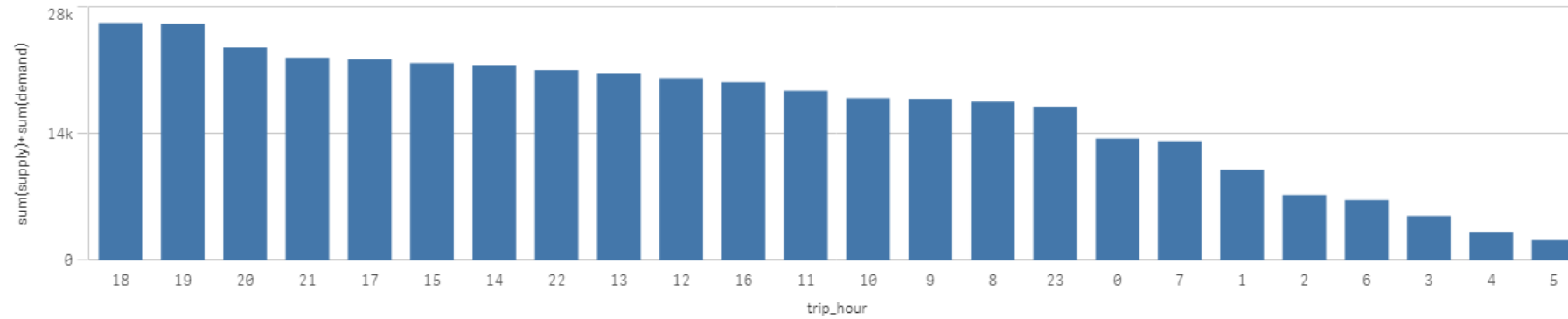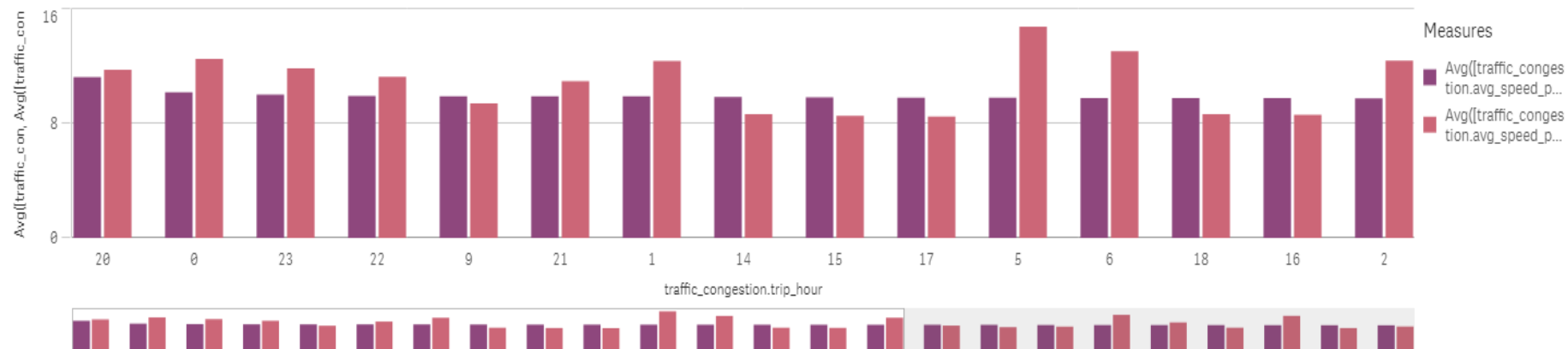
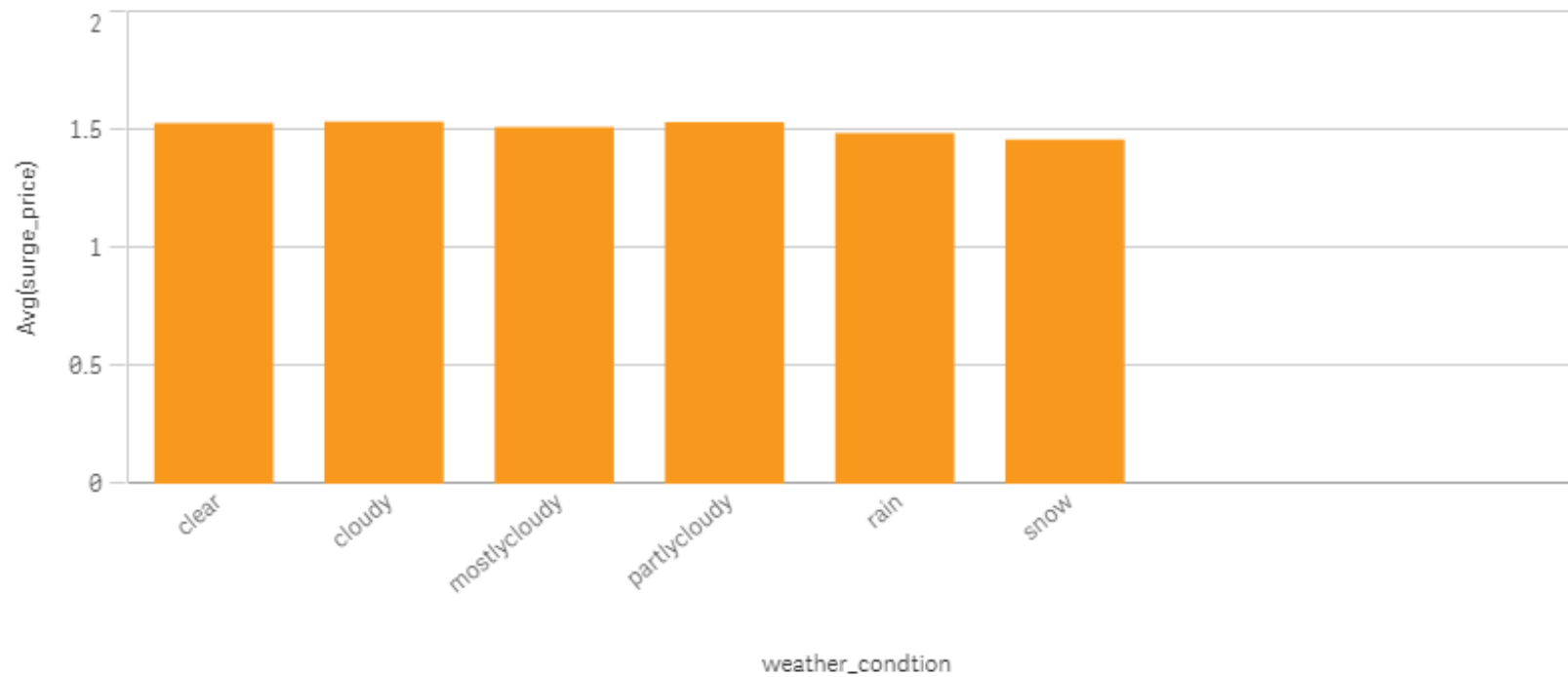# Weather effect on surge pricing

Weather effect on surge price