DATA MANAGEMENT FOR DATA SCIENCE

DATA WAREHOUSING

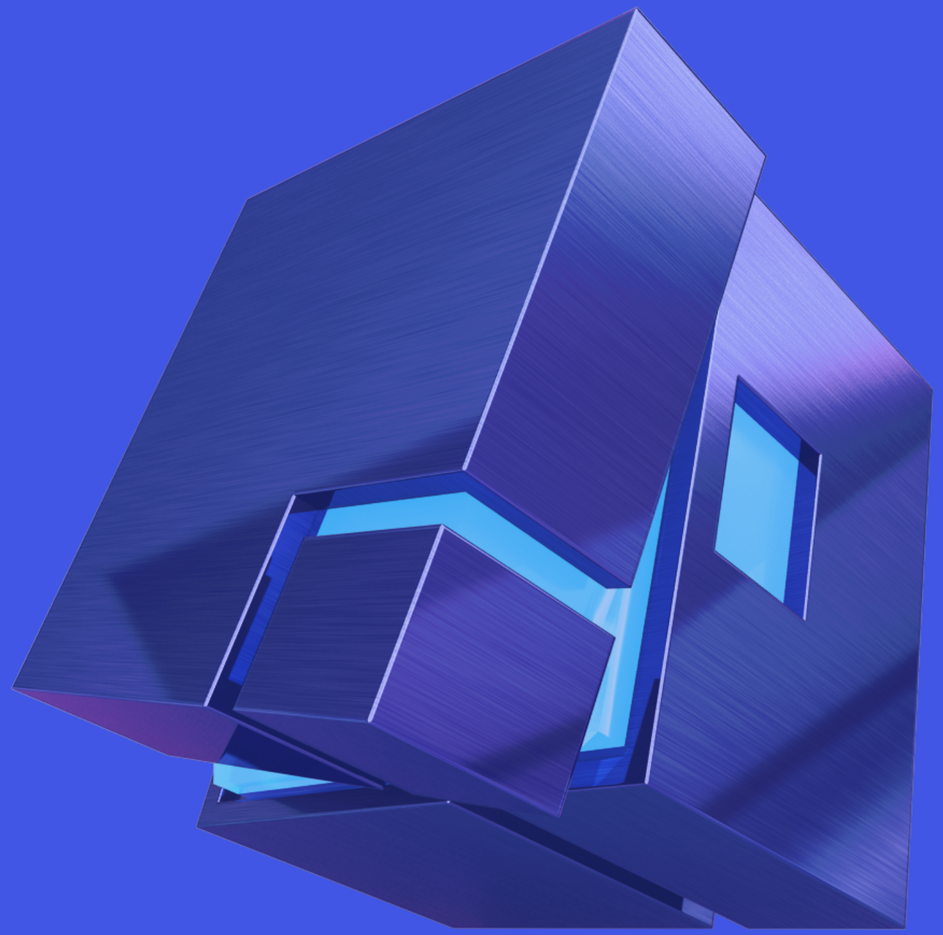
Presented by Giuliano Tocilj & Alice Schirinà

-DATA

The NYC taxis dataset

We worked on the same dataset used in the last 3 homeworks: All NYC taxi trips in january 2018.

- 1 Gb Size
- Over 8.000.000 records



-APPROACH

Multidimensional analysis

We worked with multidimensional models, in which every trip is treated as a FACT, that has some measures and dimensions

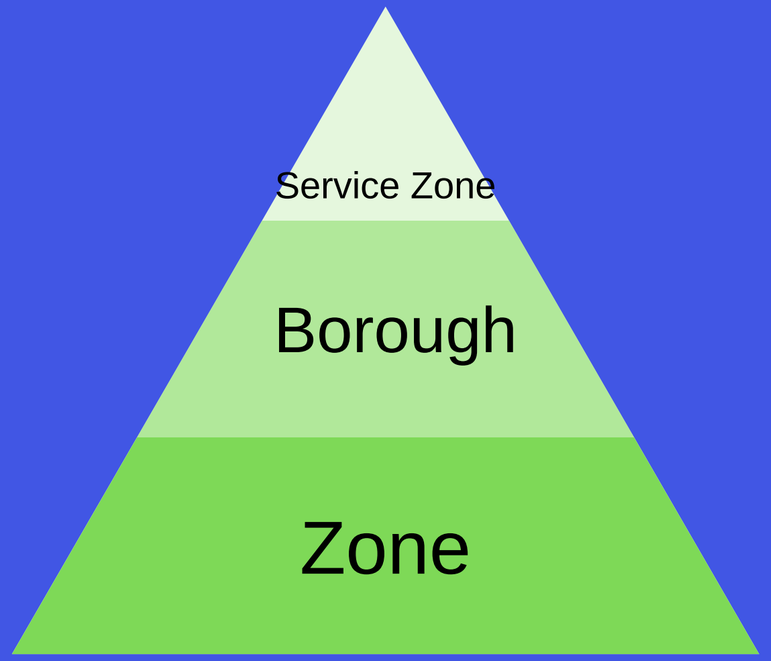
STAR DIAGRAM

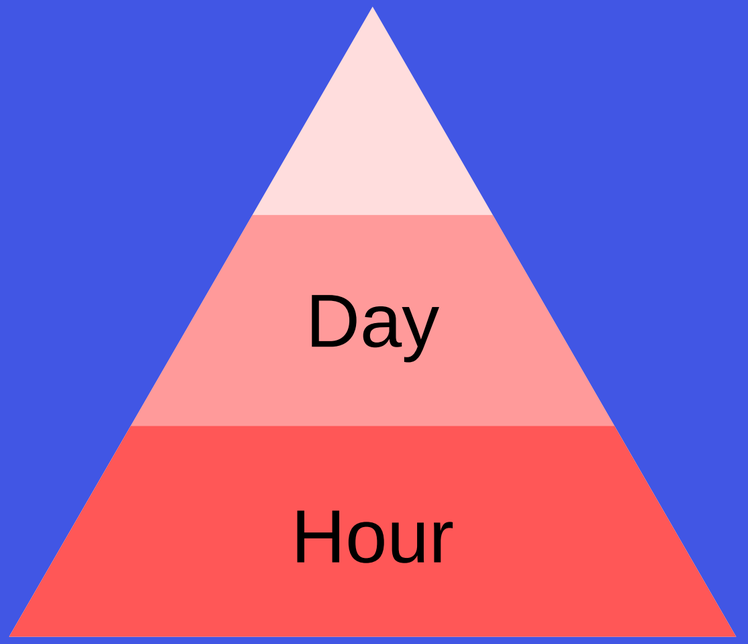
Star schema

We used a star schema with 4 dimensions in which:

Fact = Trip (with dimensions "Distance", "Passenger number", "Total amount")

Dimension 1 = Provider
Dimension2 = Time Info
Dimension3 = Location
Dimension4 = Payment type





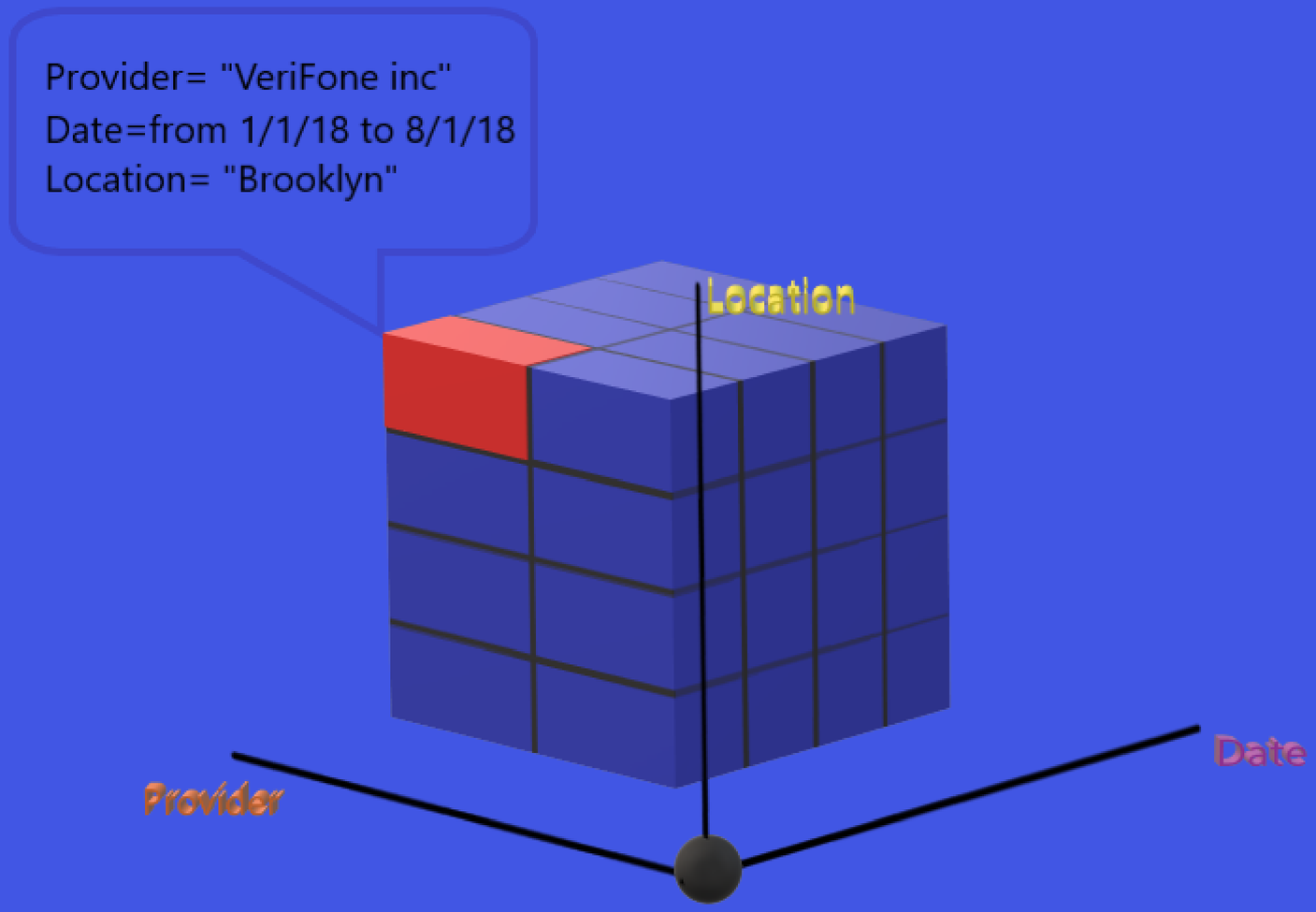
HIERARCHY

Dimension hierarchy

Each dimension has a hierarchy and we can aggregate our data by greater or smaller attribute (by *roll-up* & *dril down*)

SLICING

2° Query: Slicing



After had diced our data by 3 dimension (Provider, Location, Date) we *slice* the 3D cube selecting only the attributes that we are interested in:

```
SELECT borough, count(fact_trip.id) as num_trips, provider
FROM fact_trip, time_info, lookup, provider
WHERE fact_trip.id = time_info.id and fact_trip.pu_loc =
lookup.loc_id and fact_trip.vendor_id = provider.vendor_id
and provider = 'VeriFone Inc' and date_pick between '2018-01-01' and
'2018-01-08' and borough = 'Brooklyn';
```

	borough	num_trips	provider
▶	Brooklyn	14016	VeriFone Inc