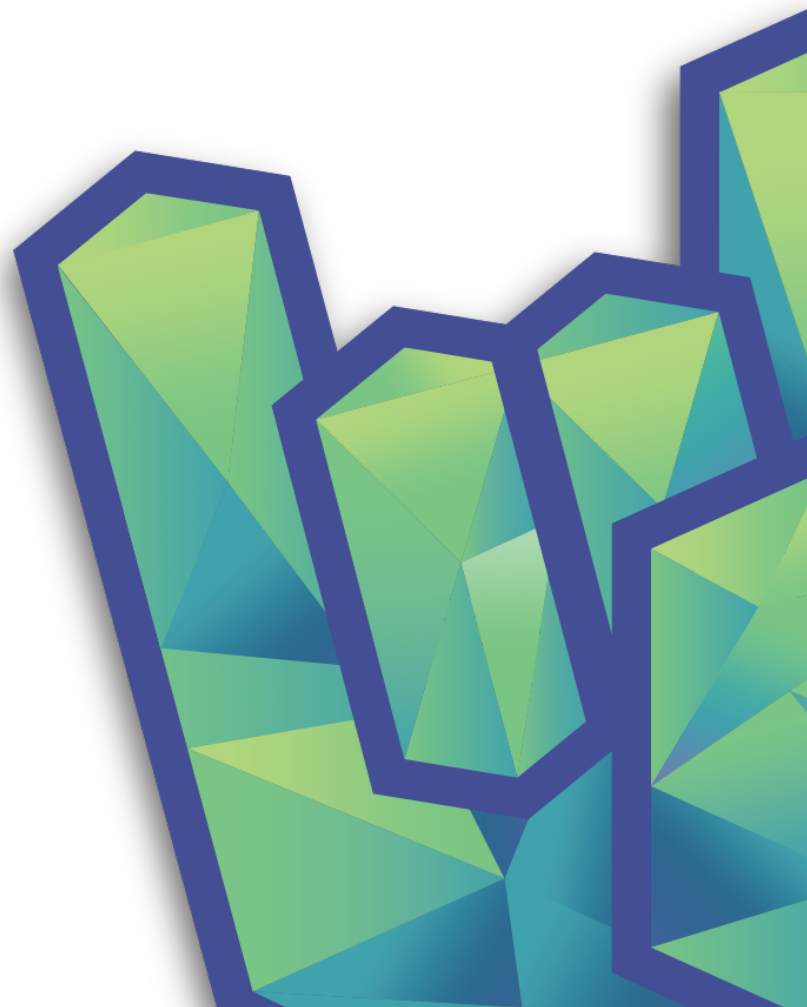
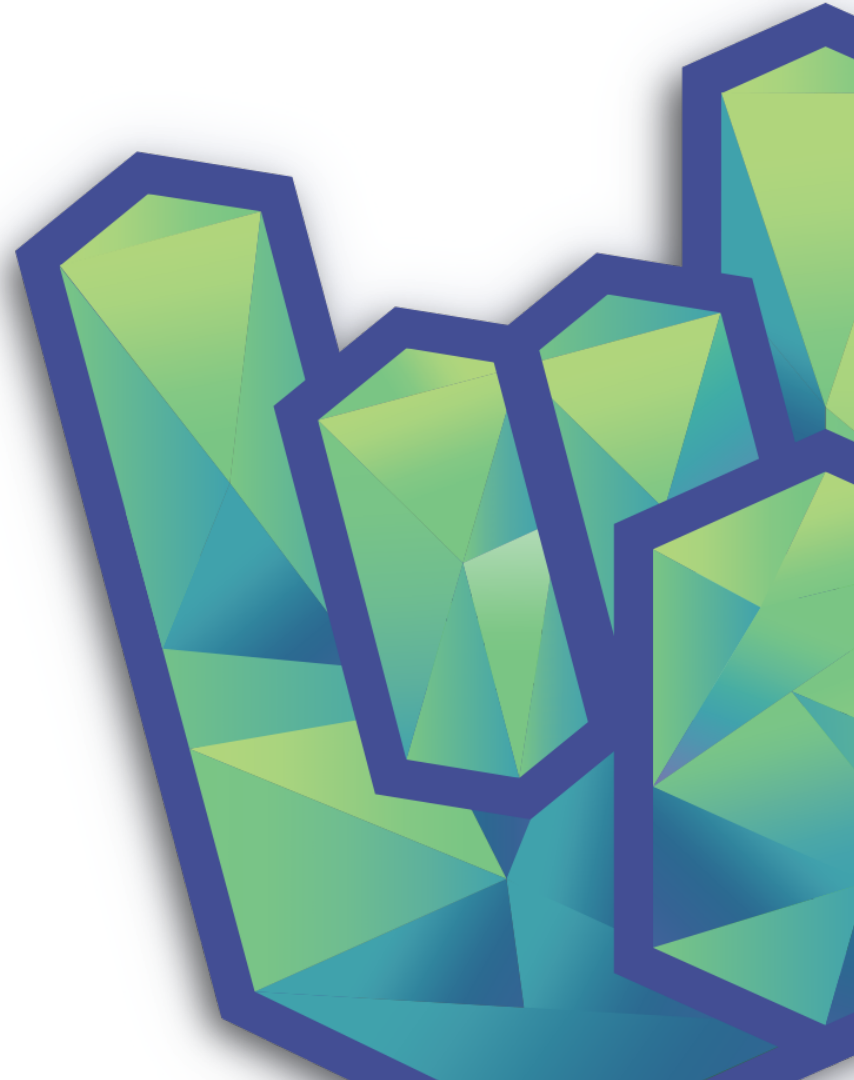
Taxi!



Objective

Run our first truly big data project

- data gathering
- investigative analysis
- package application
- deploy and run on Amazon EMR



The Scenario

The NYC taxi want to improve their business. They hired you to advise them.

Data: all the taxi rides between 2009 and 2016

Requirements

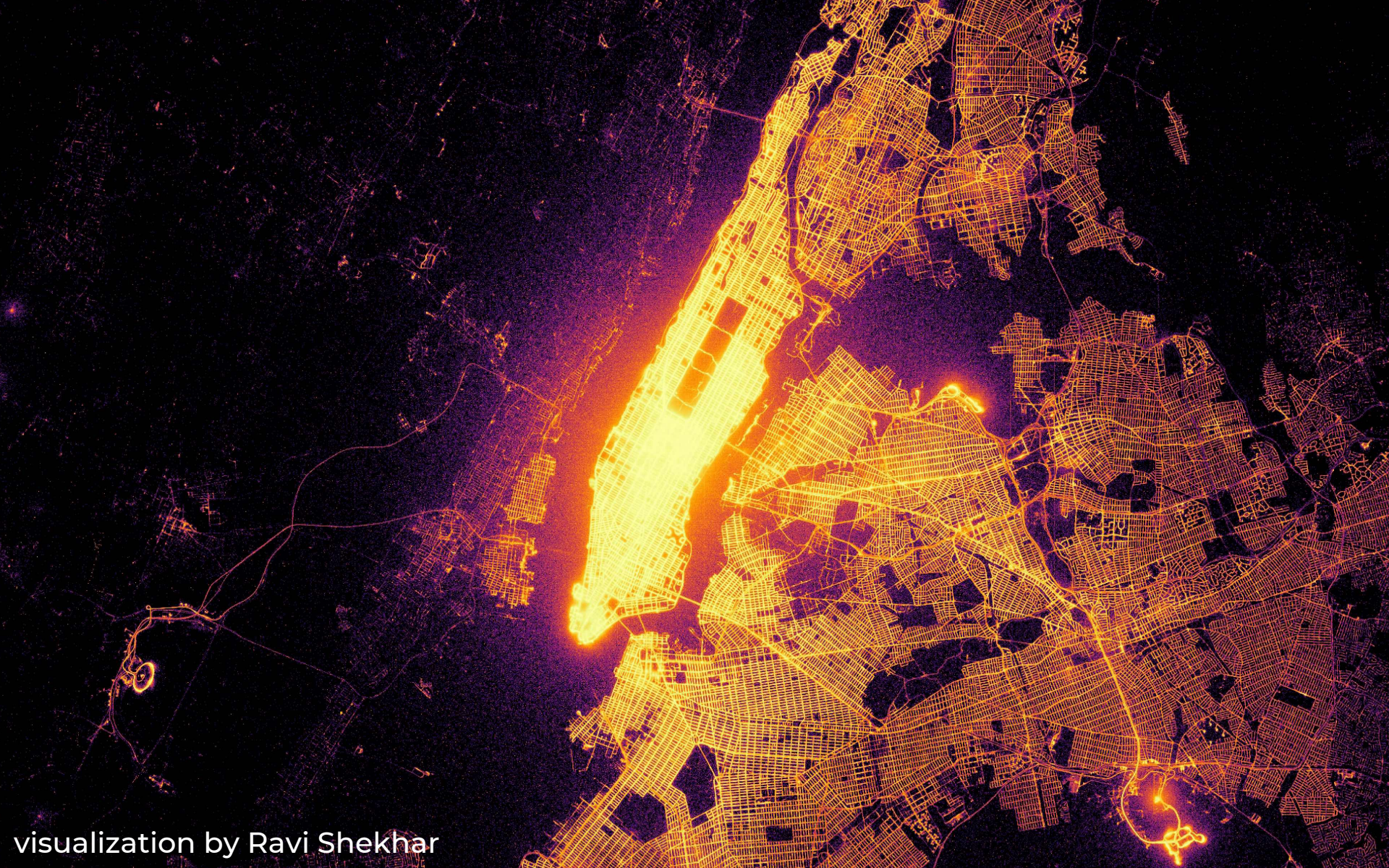
- gather data insights
- make proposals for business improvements
- suggest one potential approach
- evaluate the impact over the existing data

The Data

Data: 1.4 billion taxi rides between 2009 and 2016

35GB snappy Parquet, ~400GB uncompressed CSV

Breaking the big data barrier


visualization by Ravi Shekhar

The Data

A big DataFrame containing all taxi rides

- tpep_pickup_datetime = pickup timestamp
- tpep_dropoff_datetime = dropoff timestamp
- passenger_count
- trip_distance = length of the trip in miles
- RatecodeID = 1 (standard), 2 (JFK), 3 (Newark), 4 (Nassau/Westchester) or 5 (negotiated)
- PULocationID = pickup location zone ID
- DOLocationID = dropoff location zone ID
- payment_type = 1 (credit card), 2 (cash), 3 (no charge), 4 (dispute), 5 (unknown), 6 (voided)
- total_amount
- ... and 8 other columns

Plus a smaller DataFrame with the taxi zone descriptions

The Questions

1. Which zones have the most pickups/dropoffs overall?
2. What are the peak hours for taxi?
3. How are the trips distributed? Why are people taking the cab?
4. What are the peak hours for long/short trips?
5. What are the top 3 pickup/dropoff zones for long/short trips?
6. How are people paying for the ride, on long/short trips?
7. How is the payment type *evolving with time*?
8. Can we explore a ride-sharing opportunity by grouping close short trips?

Proposal 1

The data is extremely skewed towards Manhattan

Proposal: differentiate prices according to the pickup/dropoff area, and by demand

Proposal 2

There are clear peak hours with increased demand

Proposal: differentiate prices according to demand

Proposal 3

There is a clear separation of long/short trips

Short trips in between wealthy zones (bars, restaurants)

Long trips mostly used for airport transfers

To the NYC town hall: airport rapid transit

To the taxi company: separate market segments and tailor services to each

Strike a partnerships with bars/restaurants for pickup service

Proposal 4

Cash is dying!

Make sure the card payment processor works 24/7

Proposal 5

Lots of close taxi rides

Incetivize people to take a grouped ride, at a discount

- lower costs
- more competitive with lower prices
- fewer emissions – can ask for a subsidy on the project

Proposal 5

A simple model for estimating potential economic impact over the dataset

Parameters

- 5% of taxi trips detected to be groupable at any time
- 30% of people actually accept to be grouped
- $5 discount if you take a grouped ride
- $2 extra to take an individual ride (privacy/time)
- if two rides grouped, reducing cost by 60% of one average ride

Big Taxi Project

5 proposals worth pursuing

Best proposal worth ~100000000 dollars in economic impact

Consultation project worth 10000 – 100000 dollars

Spark rocks