Big Data

(Uber's Success Story)

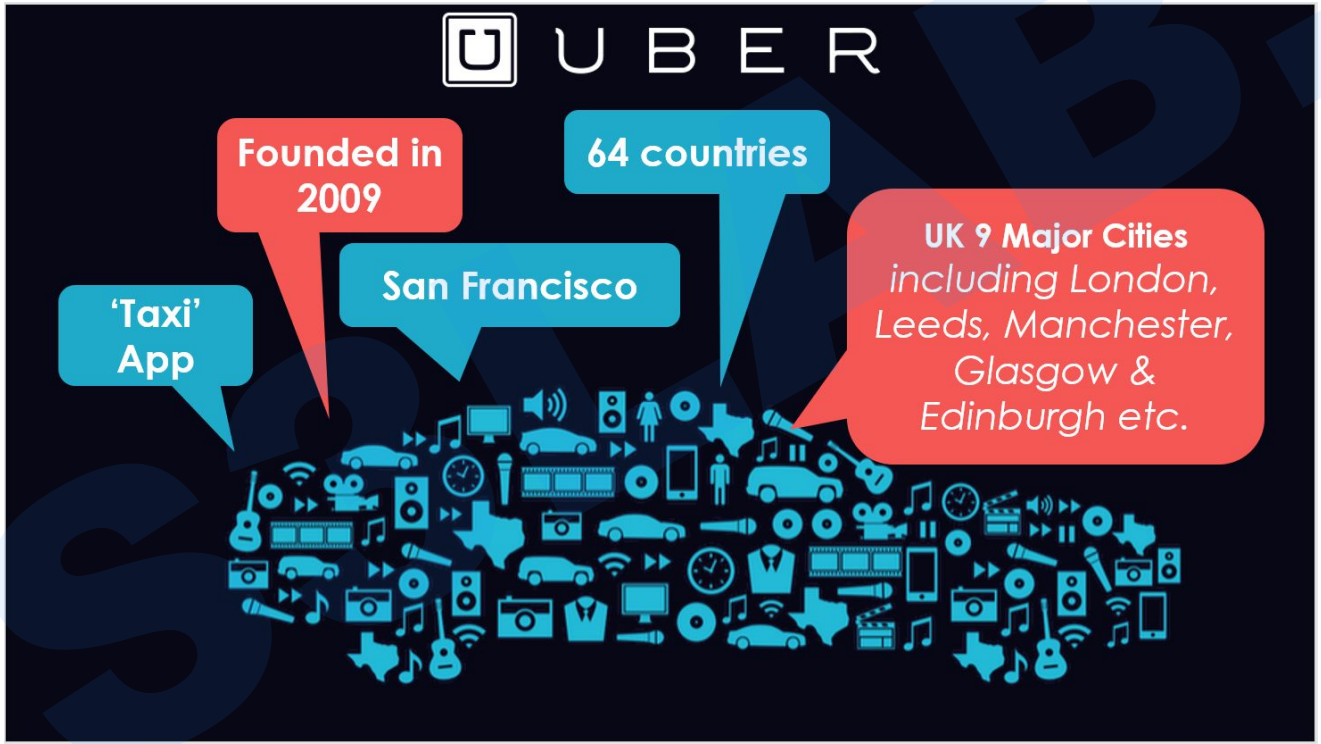
Instructor: Thanh Binh Nguyen

September 1st, 2019

S³Lab
Smart Software System Laboratory

> “Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming.”
>
> – Chris Lynch, Vertica Systems

Introduction



Introduction

Driver Partner

Our partner in the ride sharing business

Riders

Folks like you and me who request a ride on any of Uber's transportation products. e.g. UberX, uberPool

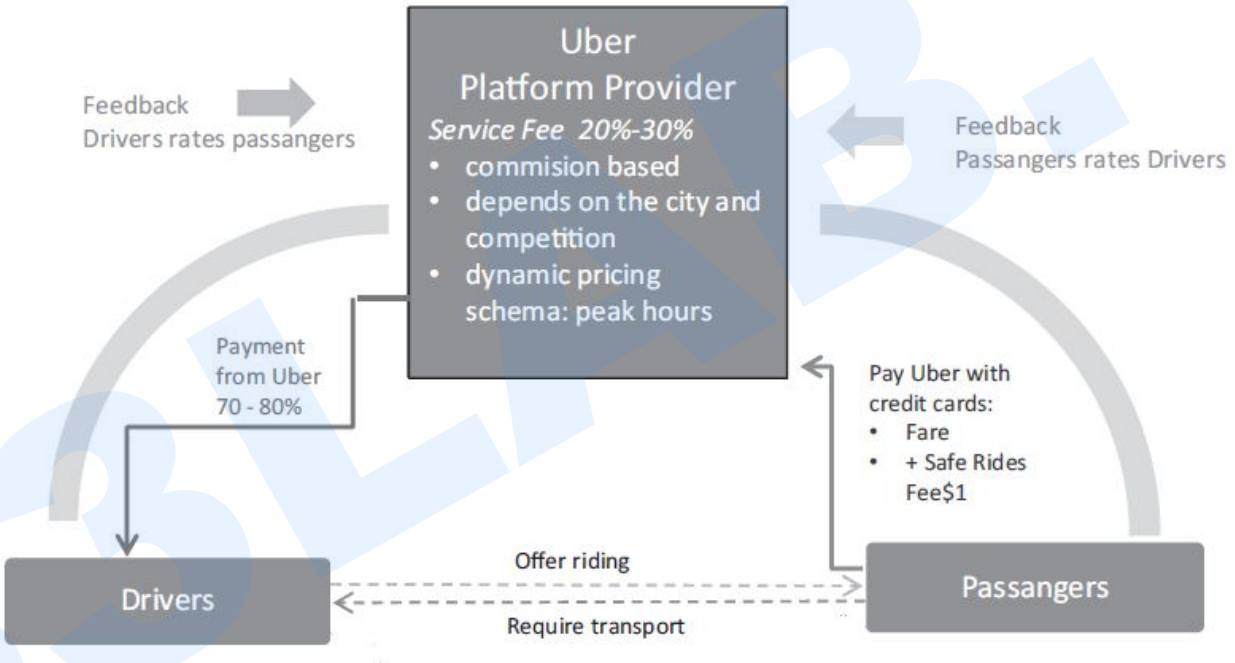
Merchants

Restaurants or shops that have signed on to the Uber platform.

Marketplace

Uber's logistic platform

Introduction



- what determines demand (required transport) and supply (offer riding) for uber cars?
- how should Uber determine the optimal pricing policy?

Introduction

"Transportation as reliable as running water, everywhere, for everyone"

Uber
Mission

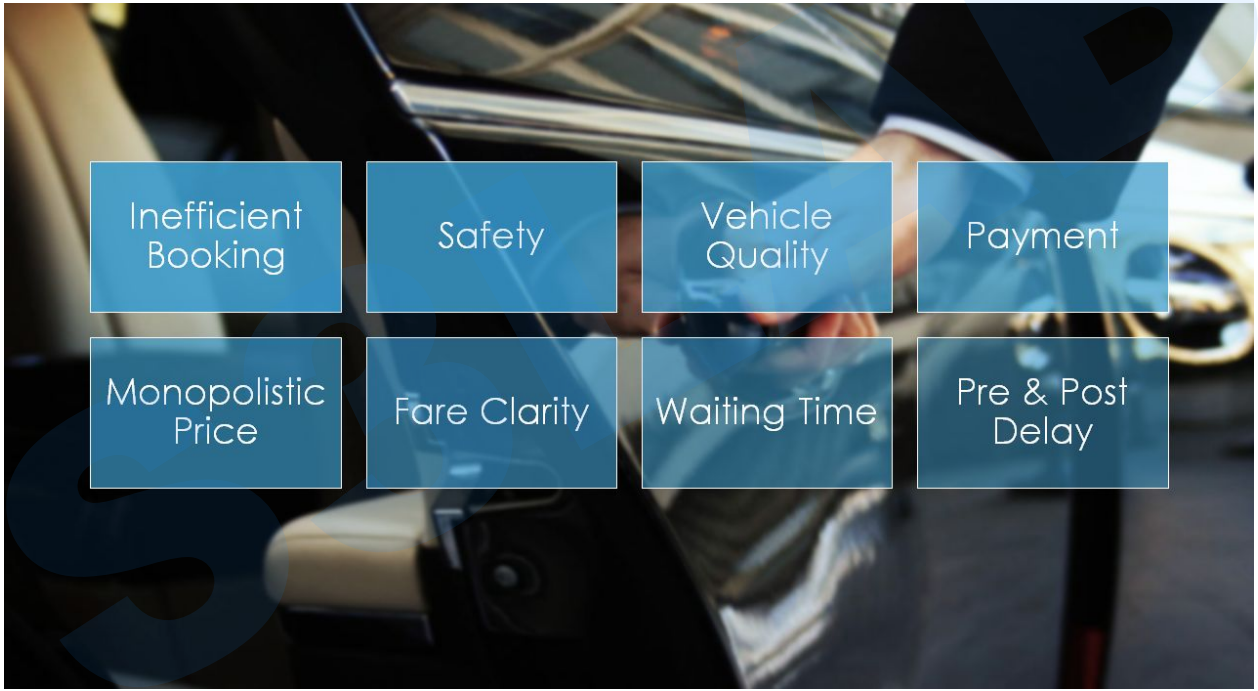
Introduction

Market

4 factors

Market

Consumers

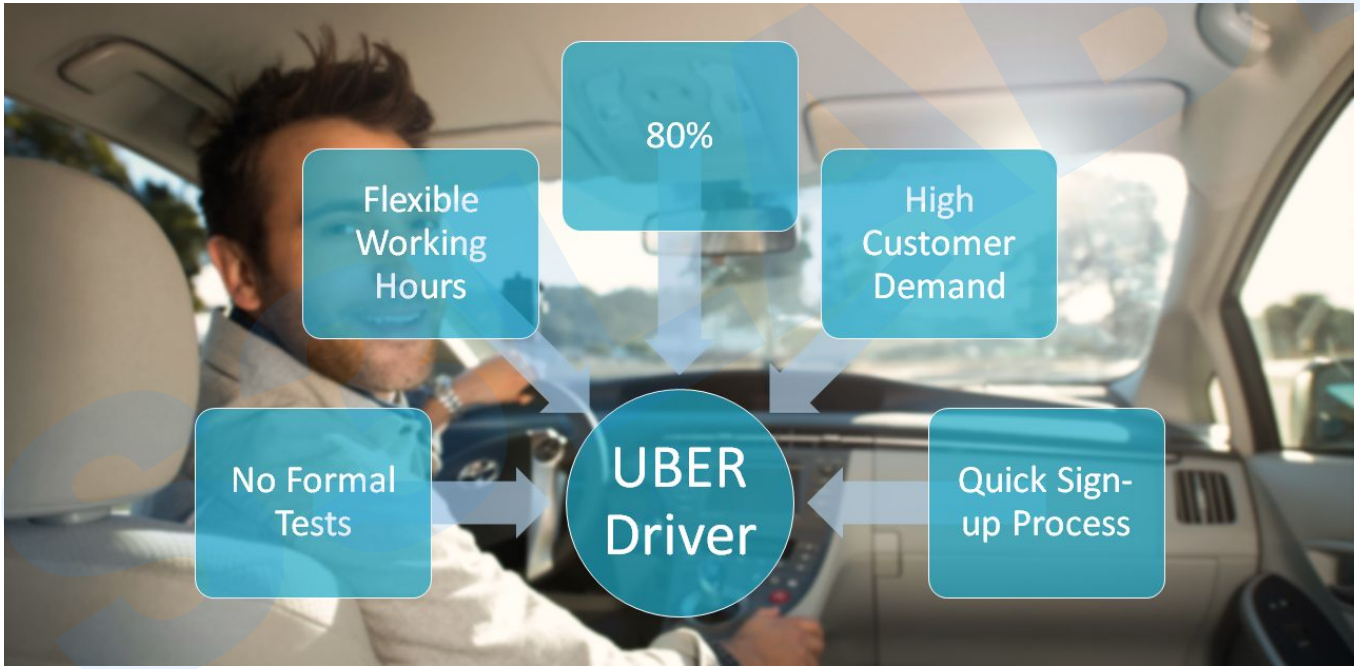


Market

Consumers

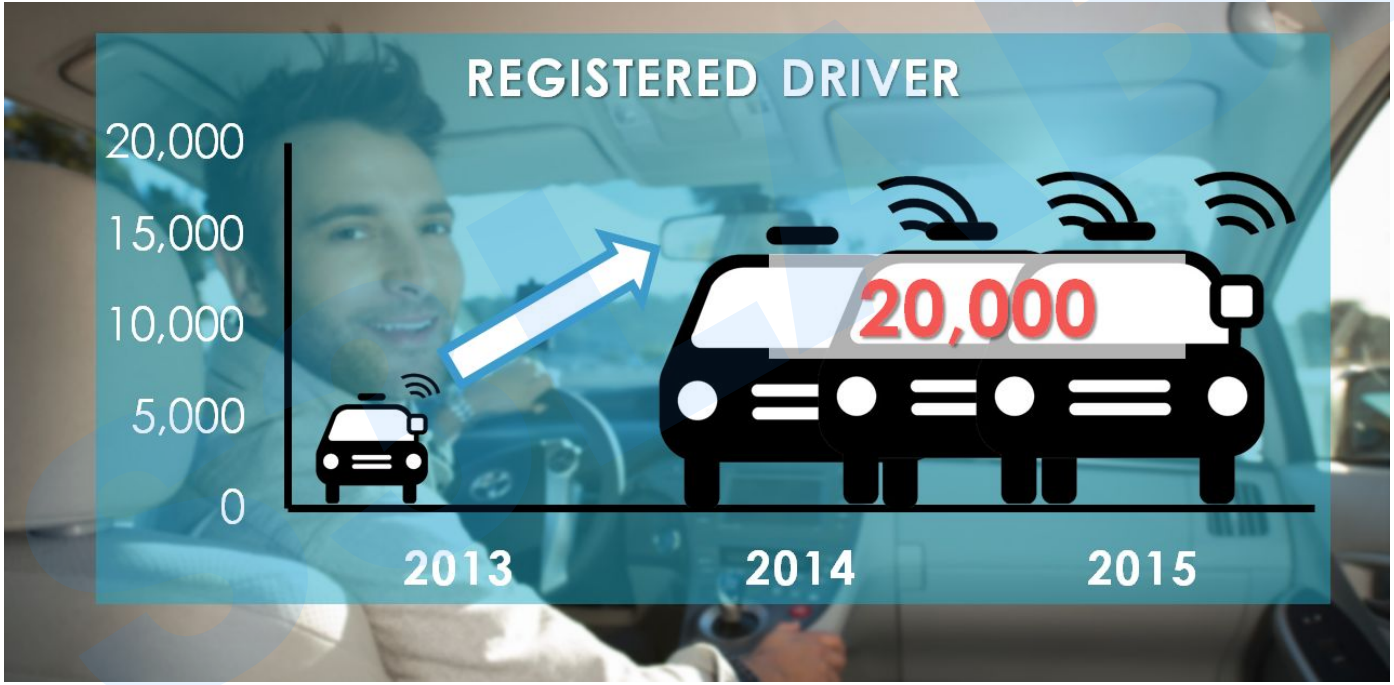
Market

Drivers



Market

Drivers

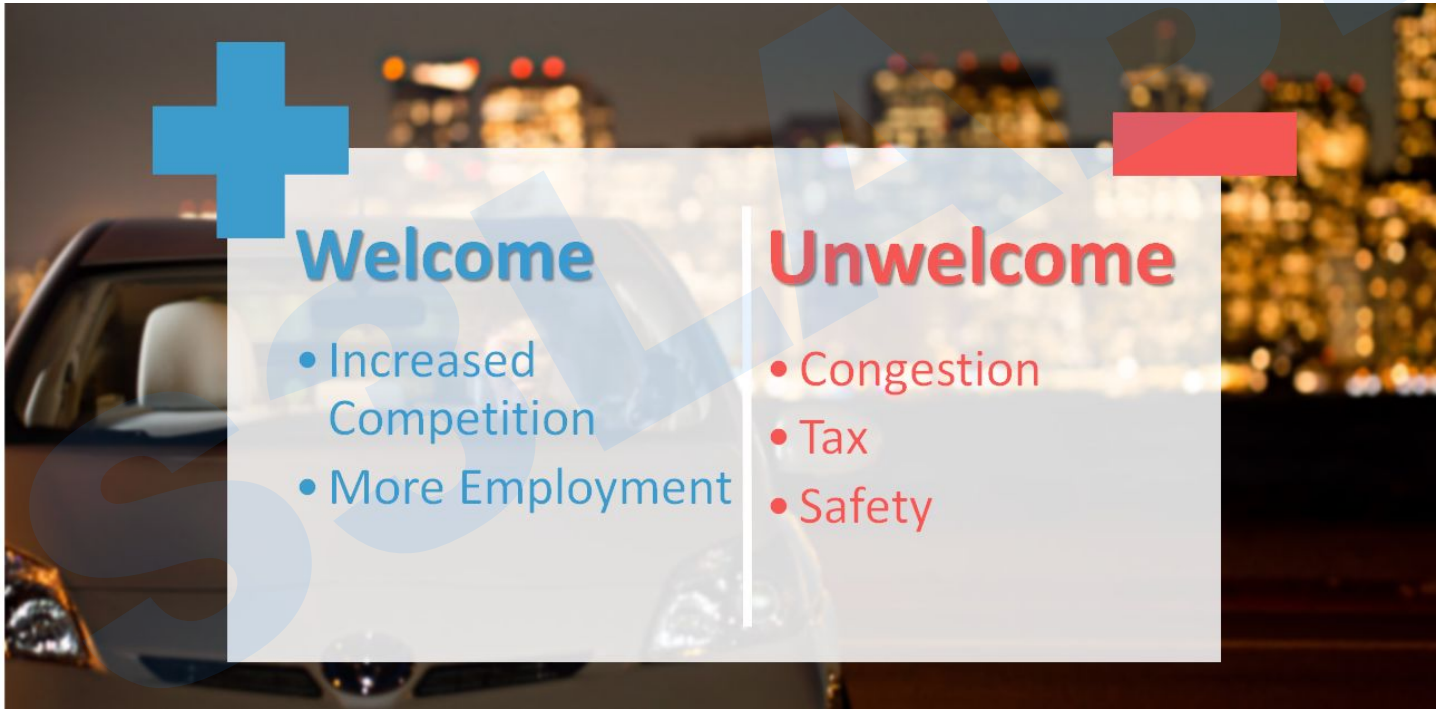


Market

Drivers

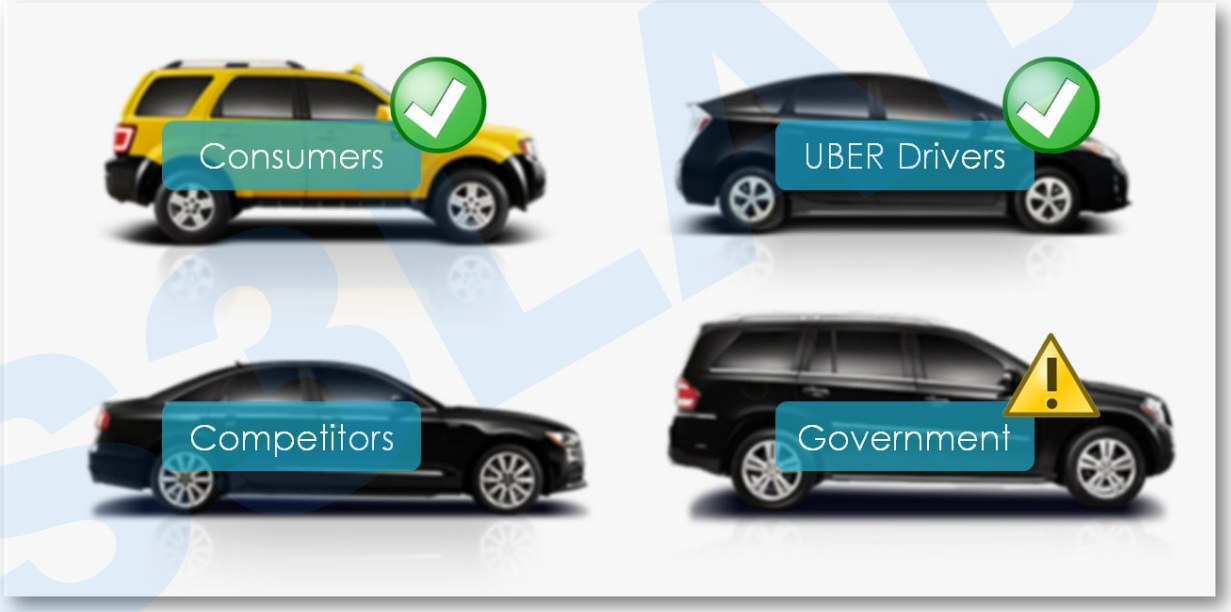
Market

Government



Market

Government



Market

Competitors

Market

Competitors

Market

Competitors

Market

Competitors

Market

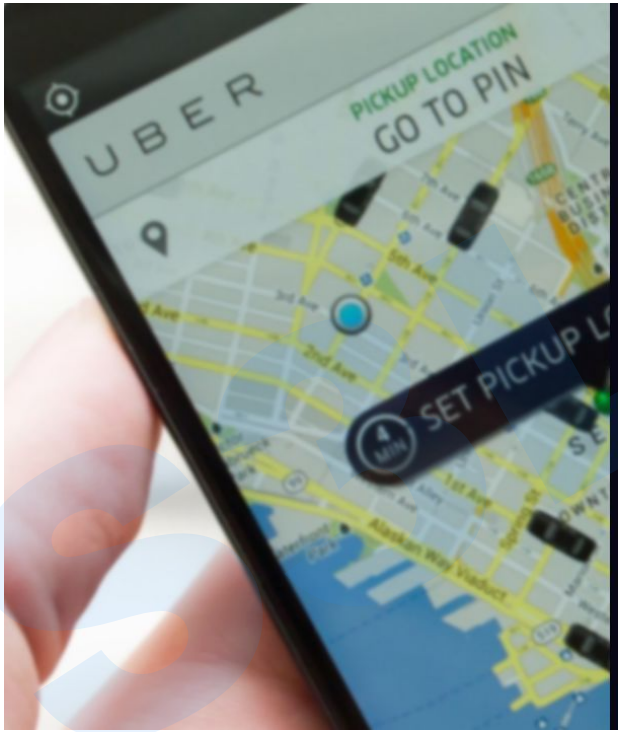
Competitors

Technology

TECHNOLOGY

- Only requires smartphone;
- All hiring and payments are handled through Uber;
- Distances are calculated using GPS in accordance with Uber's prescribed rates.
- But has it grown too fast?

The Present and Future

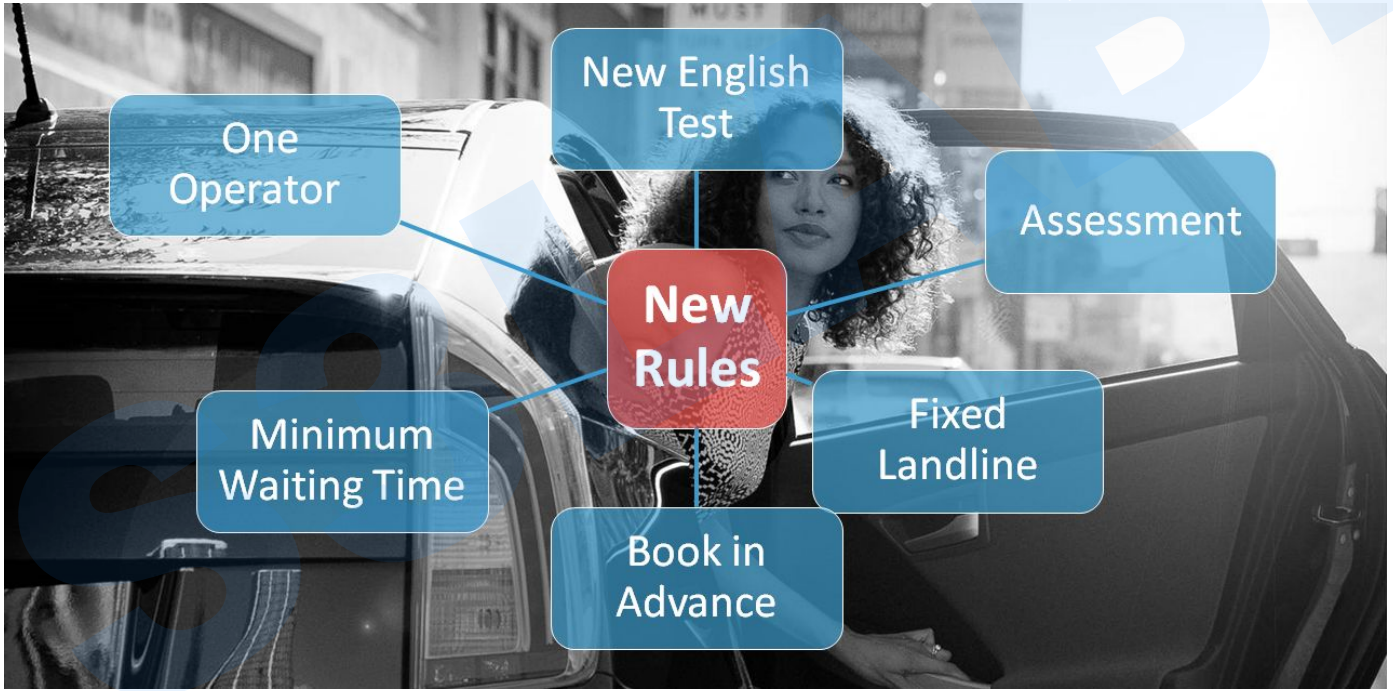


THE PRESENT

- Recently won a high court ruling on their use of GPS technology and whether it constitutes as a meter
- Their own drivers recently protested over pay in London, with Uber increasing their service fee from 20 to 25%

The Present and Future

Future



ML Problems

Why do we need machine learning?

- Subscribers Mapping (Routes, ETAs, ...)
- Fraud and Security
- uberEATS Recommendations
- Marketplace Optimizations
  - Forecasting
  - Driver Positioning
  - Health, Trends, Issues, ...
- And more ...

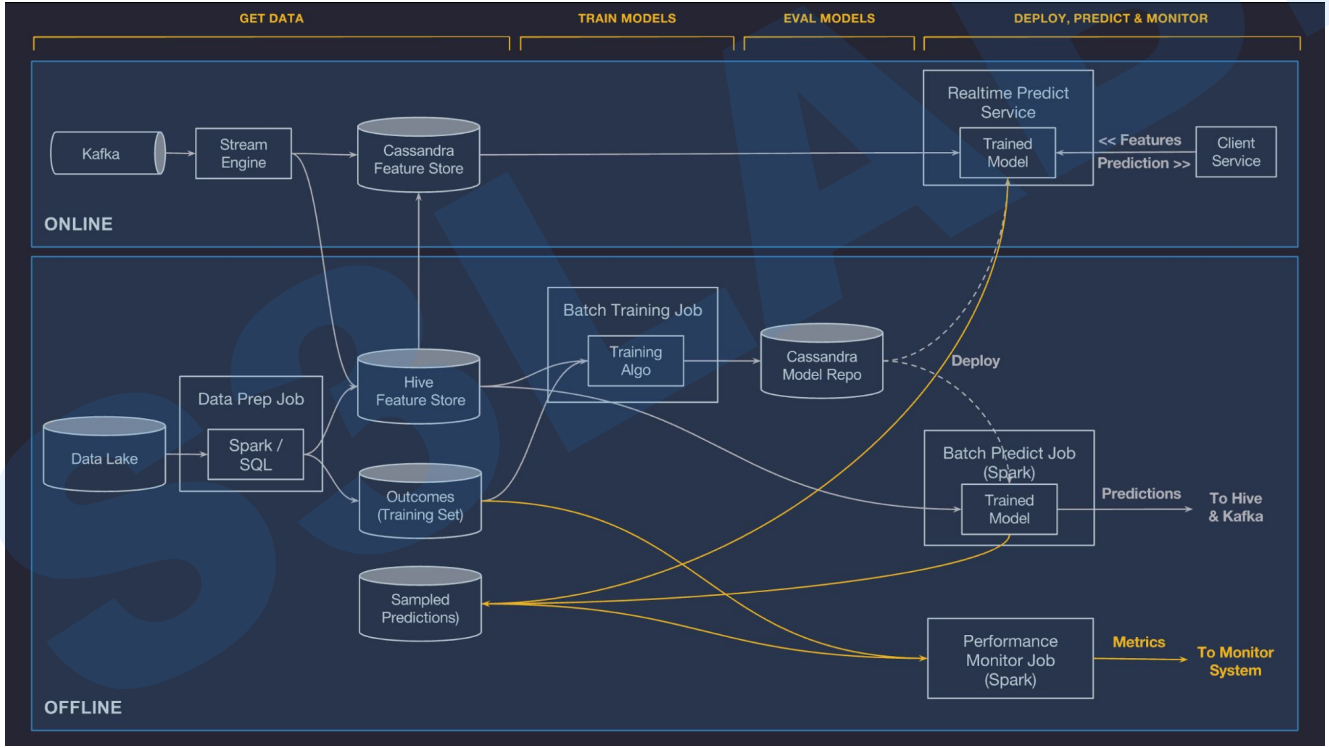
ETA, Route Optimization, Pickup Points, Pool rider matches

Examining holiday traffic trends in Manila

Manila, Phillipines

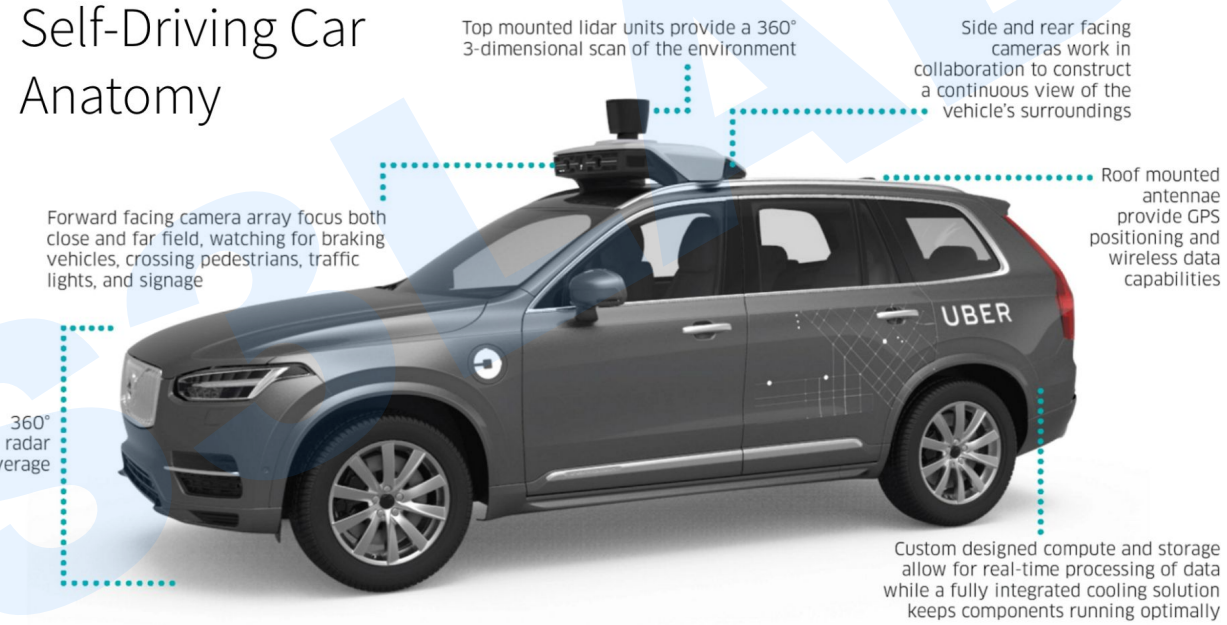
ML Problems

Michelangelo: an internal ML-as-a-service



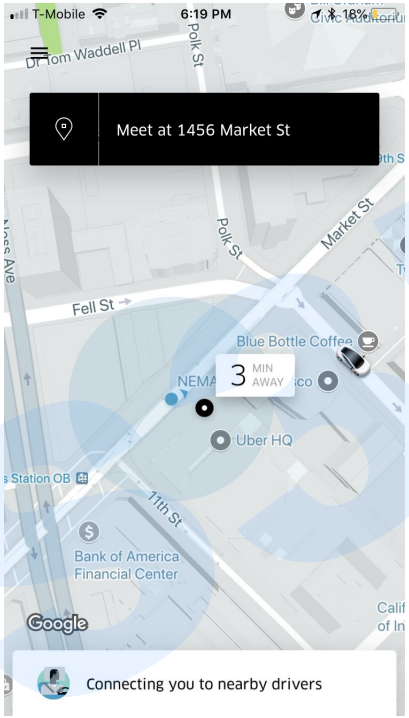
ML Problems

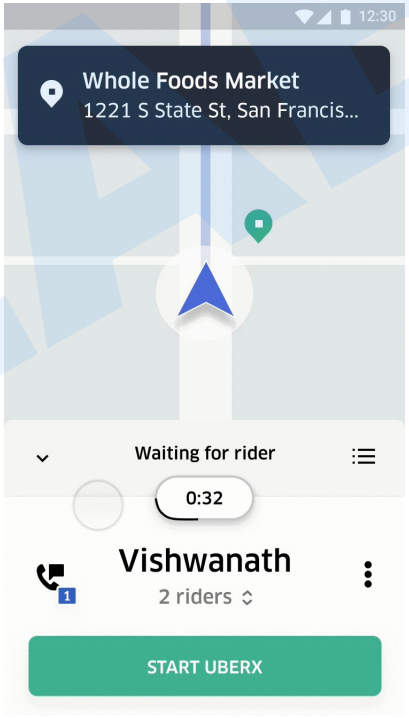
Self-Driving Car Anatomy



ML Problems

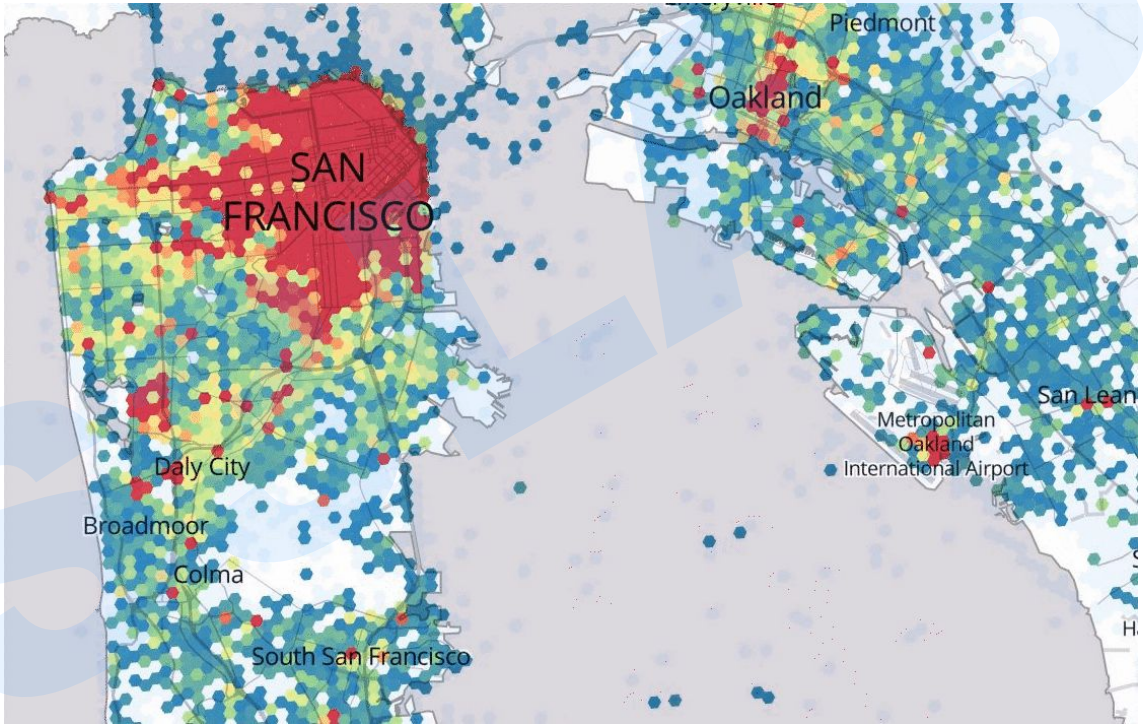
ETAs and One-Click Chat





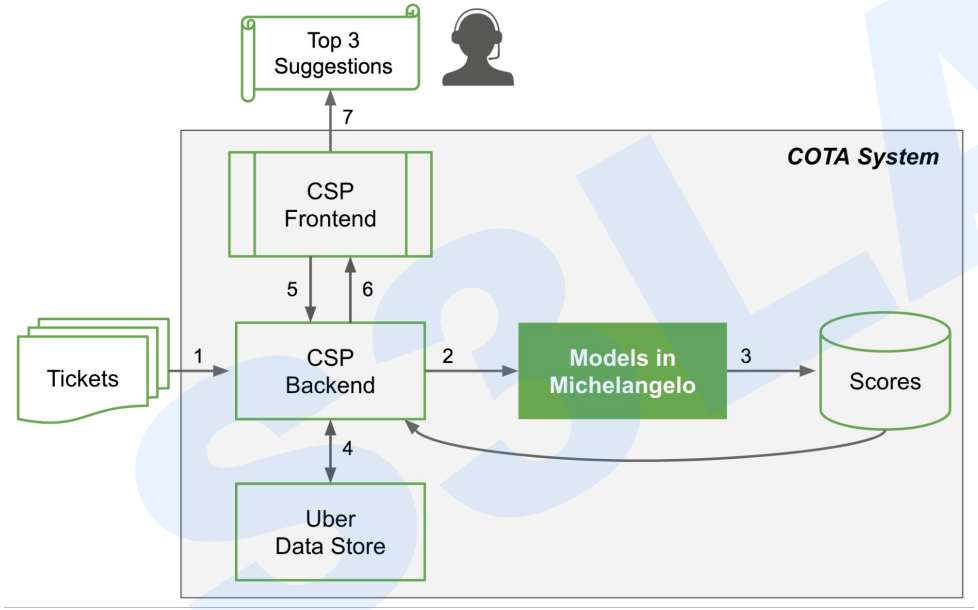
ML Problems

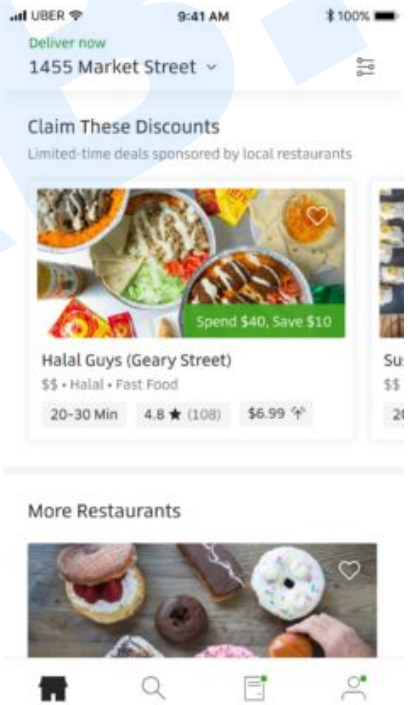
Marketplace Forecasting



ML Problems

Uber Eats & Customer Support

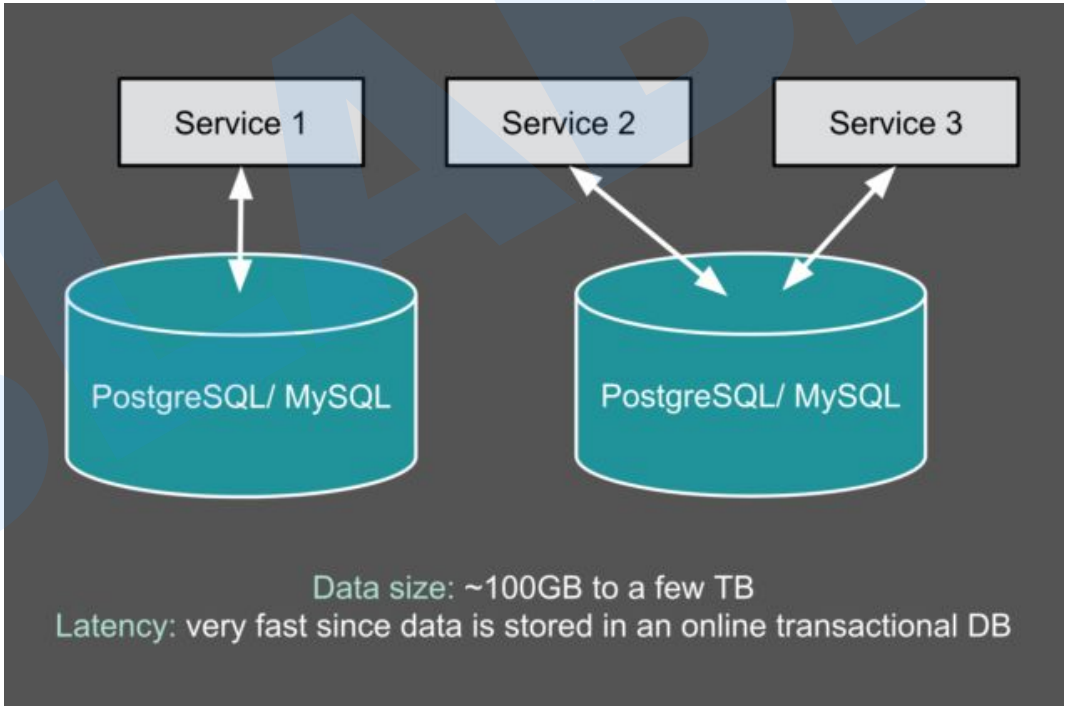




Big Data Platform: 100+ Pe with Minute Latency

Generation 1: Traditional Online Transaction Processing (OLTP) Before 2014

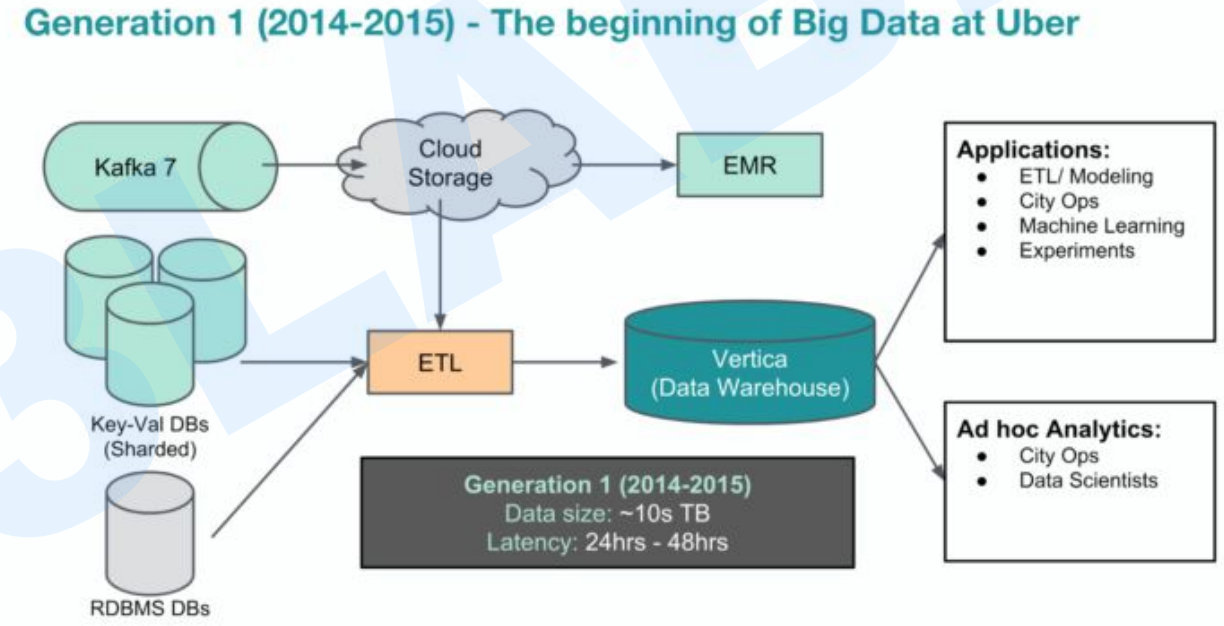
- Monolithic
- Growing exponentially
  - Cities / countries
  - Riders / drivers
- Uber as data-driven
- https://eng.uber.com/uber-big-data-platform/



Big Data Platform: 100+ Pe with Minute Latency

Generation 1: 2014-2015 The beginning of Big Data at Uber

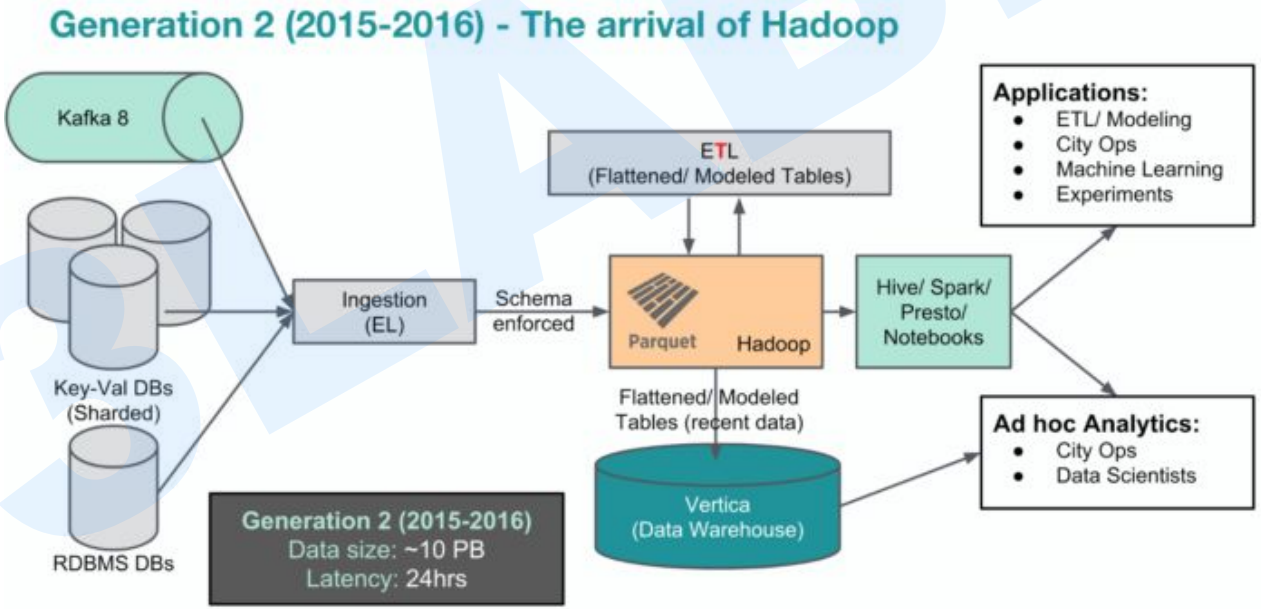
- Aggregate all of Uber's data in one place & provide standard SQL interface



Big Data Platform: 100+ Pe with Minute Latency

Generation 2: 2015-2016

- Service-Oriented Architecture with about 100s of services



Big Data Platform: 100+ Pe with Minute Latency

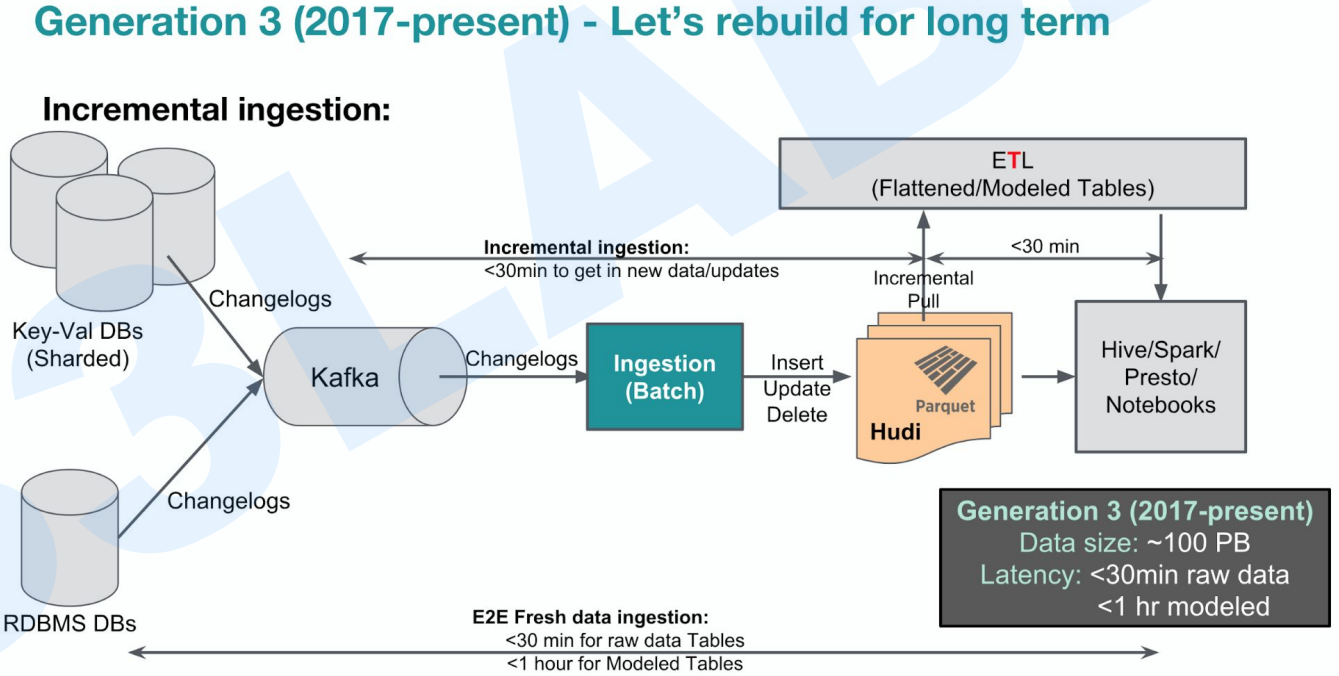
Generation 3: 2017 - present

- **HUDI**:

  Hadoop

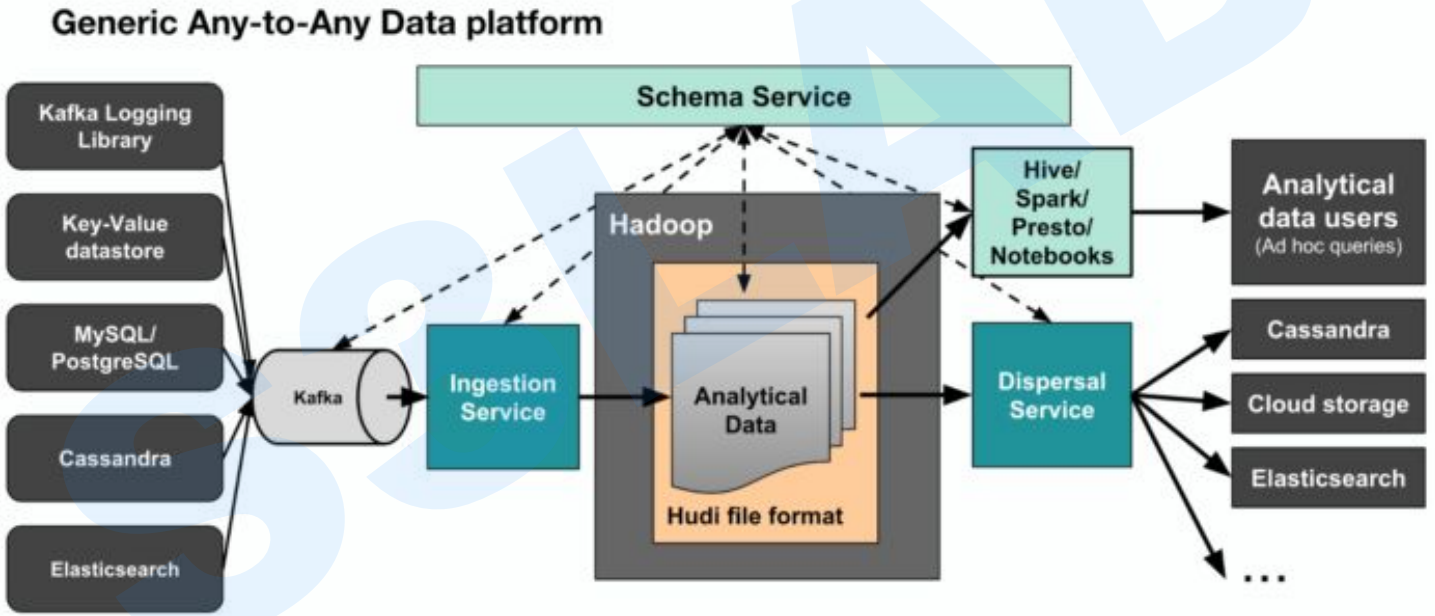
  Upserts and

  Incremental



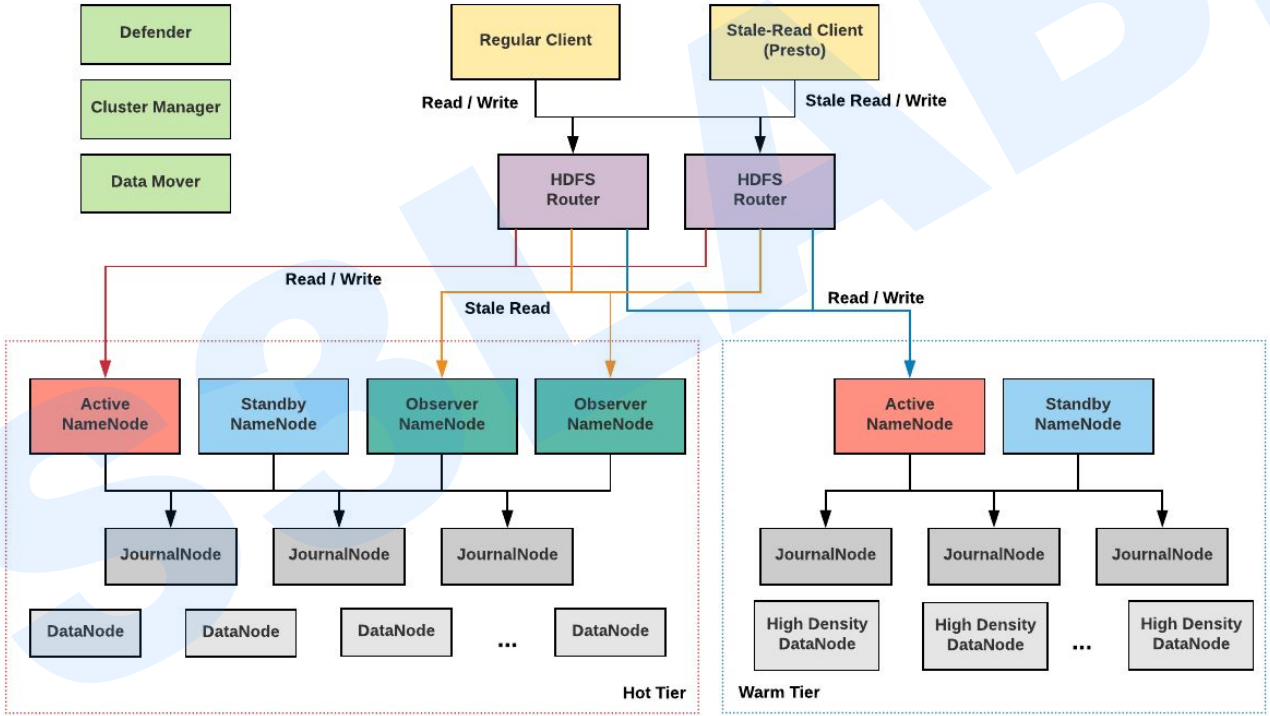
Big Data Platform: 100+ Pe with Minute Latency

Generation 3: 2017 - present



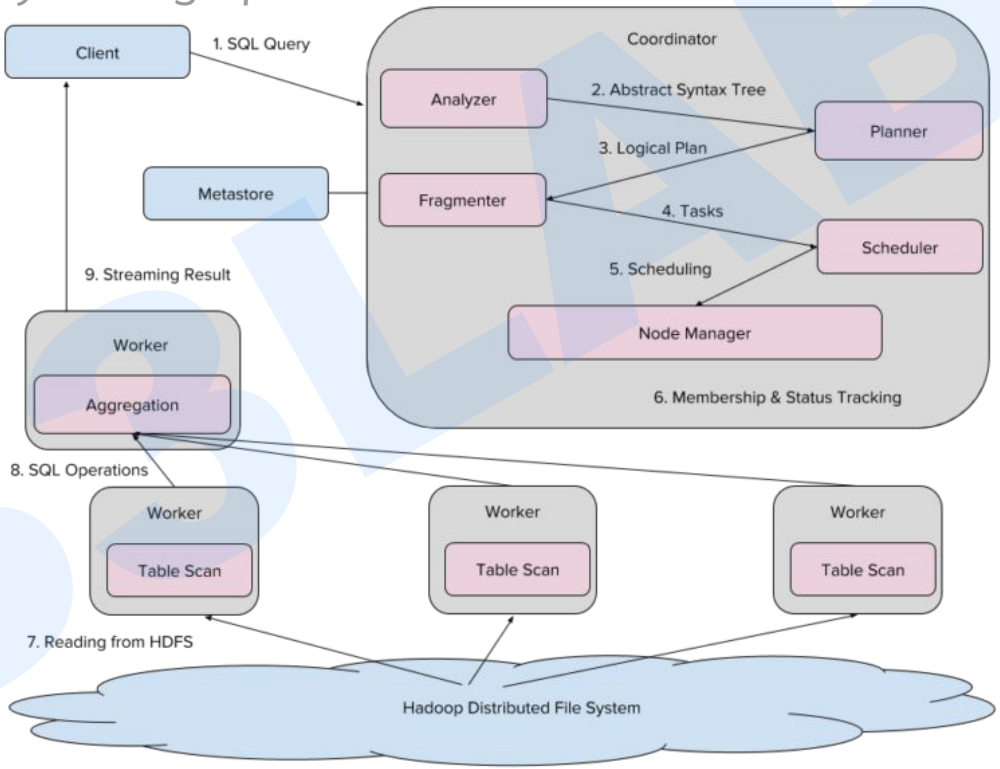
Big Data Platform: 100+ Pe with Minute Latency

Generation 3: Storage



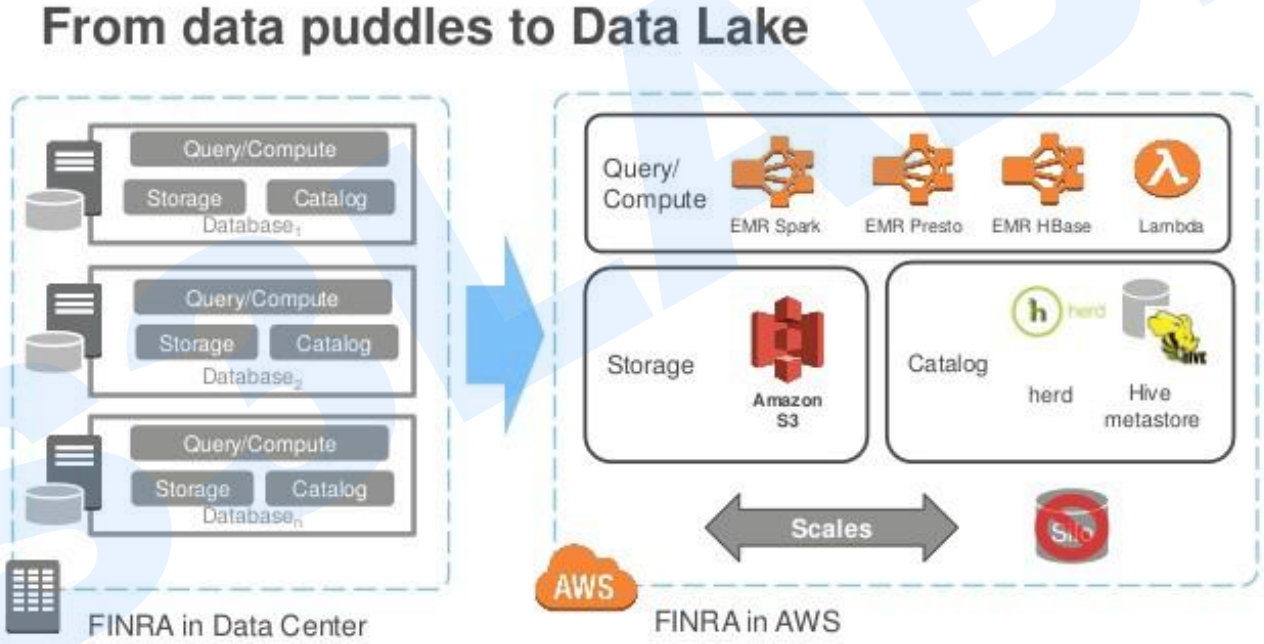
Big Data Platform: 100+ Pe with Minute Latency

Generation 3: Query through presto



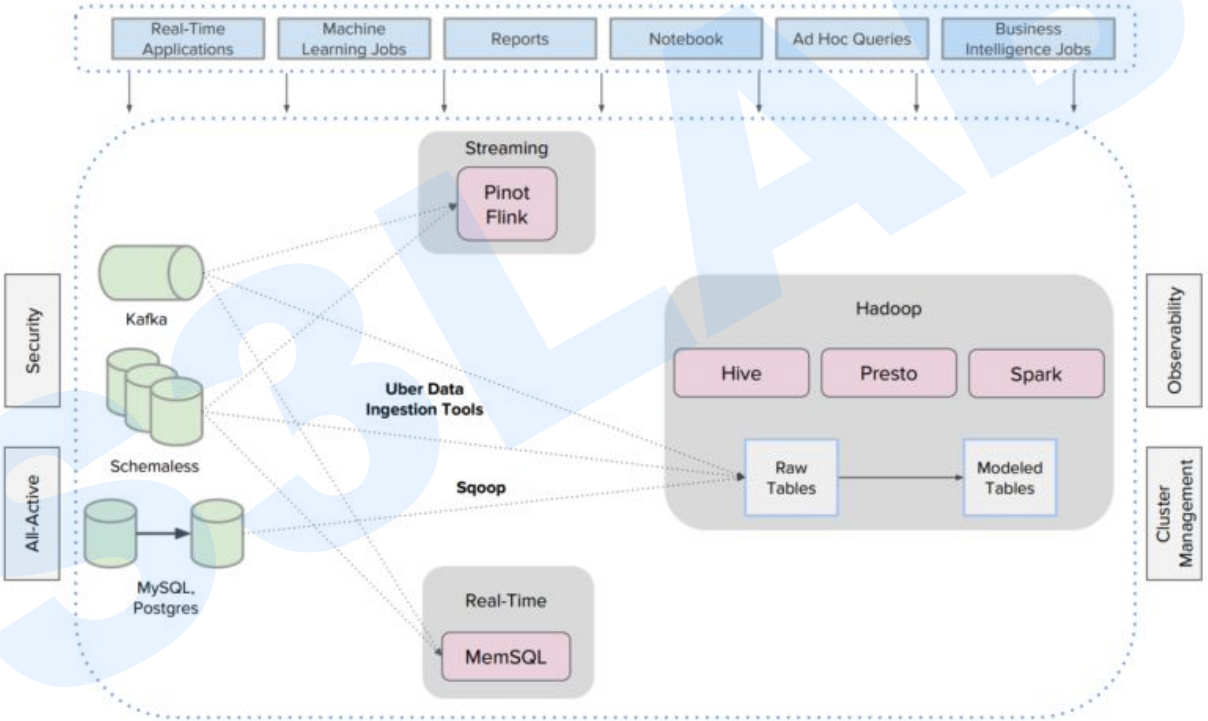
Big Data Platform: 100+ Pe with Minute Latency

Generation 3: data lake using AWS



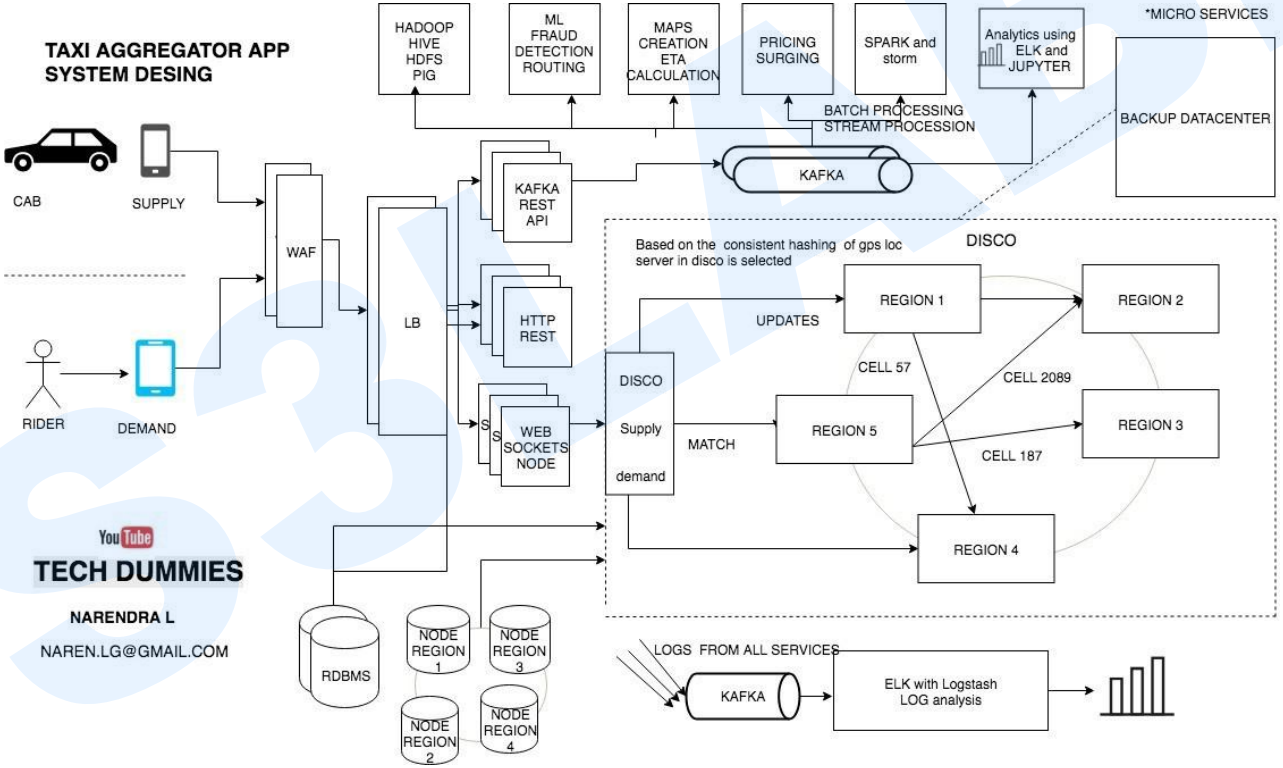
Big Data Platform: 100+ Pe with Minute Latency

Analytics infrastructure



Big Data Platform: 100+ Pe with Minute Latency

Generation 3: System Design



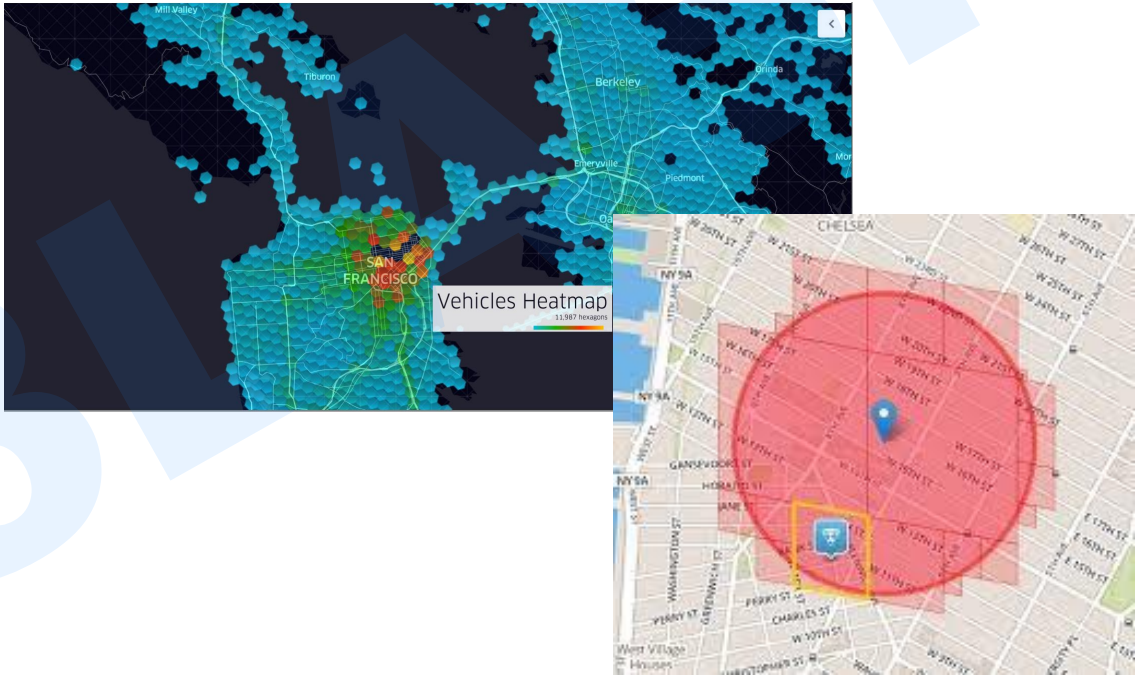
Big Data Platform: 100+ Pe with Minute Latency

Generation 3: DISCO - Dispatch Optimization

- Supply Service
  - Track cars using geolocation(lat and long)
  - Cab keep on sending lat-long to the server every 5 sec once
  - State machines of all of the supply also kept in memory.
  - Vehicle model: seats, type, child seat, wheelchair fit, ...
  - Tracked allocation: vehicle have 3 seats but 2 of those are occupied.
- Demand Service
  - Track GPS of user when requested.
  - Order: small / big car, pool, ...
  - Vehicle model: seats, type, child seat, wheelchair fit, ...
  - Demand requirements must be matched against supply inventory.

Big Data Platform: 100+ Pe with Minute Latency

- Supported by Google S2
- Indexing, Lookup, Rendering
- Symmetric Neighbors
- Convex & Compact Regions
- Equal Areas
- Equal Shape



Big Data Platform: 100+ Pe with Minute Latency

Generation 3: DISCO - Dispatch Optimization

- Core requirements
  - Reduce extra driving
  - Reduce waiting time
  - Lowest overall Estimated Time of Arrival (ETA)

Big Data Platform: 100+ Pe with Minute Latency

Generation 3: Price and Surge

- The price is increased when there are more demand and less supply with the help of prediction algorithms.
- According to UBER surge helps to meet supply and demand. by increasing the price more cabs will be on the road when the demand is more.

Big Data Platform: 100+ Pe with Minute Latency

Generation 3: Log collection and analysis

- Service logging services are configured to push logs to a distributed Kafka cluster and then using logstash we can apply filters on the messages and redirect them to different sources, for example, Elasticsearch to do some log analysis using Kibana / Graphana: ***Track HTTP APIs***, ***To manage profile***, ***To collect feedback and ratings***, ***Promotion and coupons etc***, ***FRAUD DETECTION***, ***Payment fraud***, ***Incentive abuse***, ***Compromised accounts***.

Q & A

Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.