Why?

Why data systems?

What about data structures from algorithms class?

Spreadsheets? Text files?

Why so many design choices?

SQL? Big data? (MapReduce, Hadoop, Spark)

Learning systems? (TensorFlow, deep learning)

What's exciting for future?

Shiva@stanford.edu -- “Shiva” Shivakumar

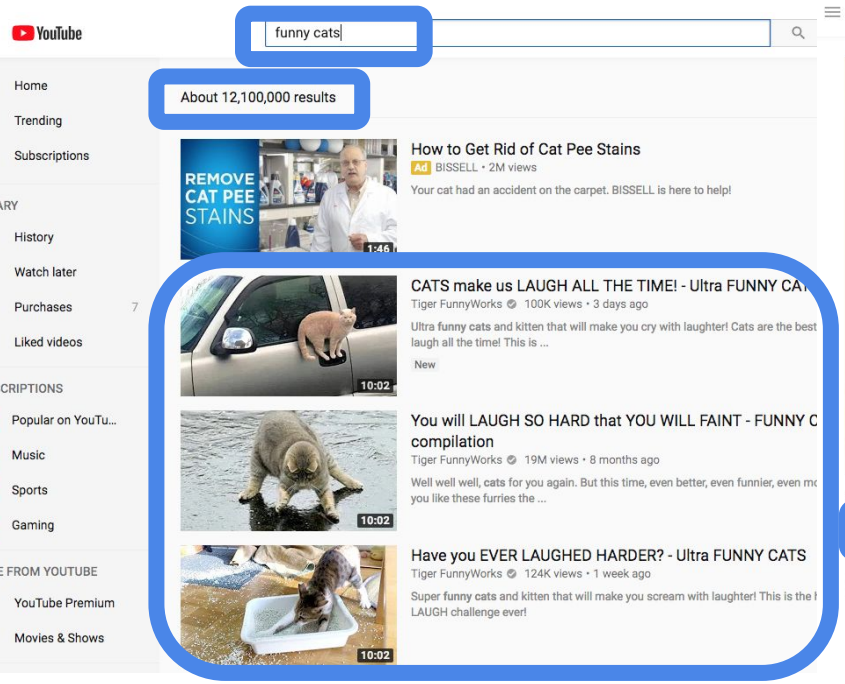
Interests

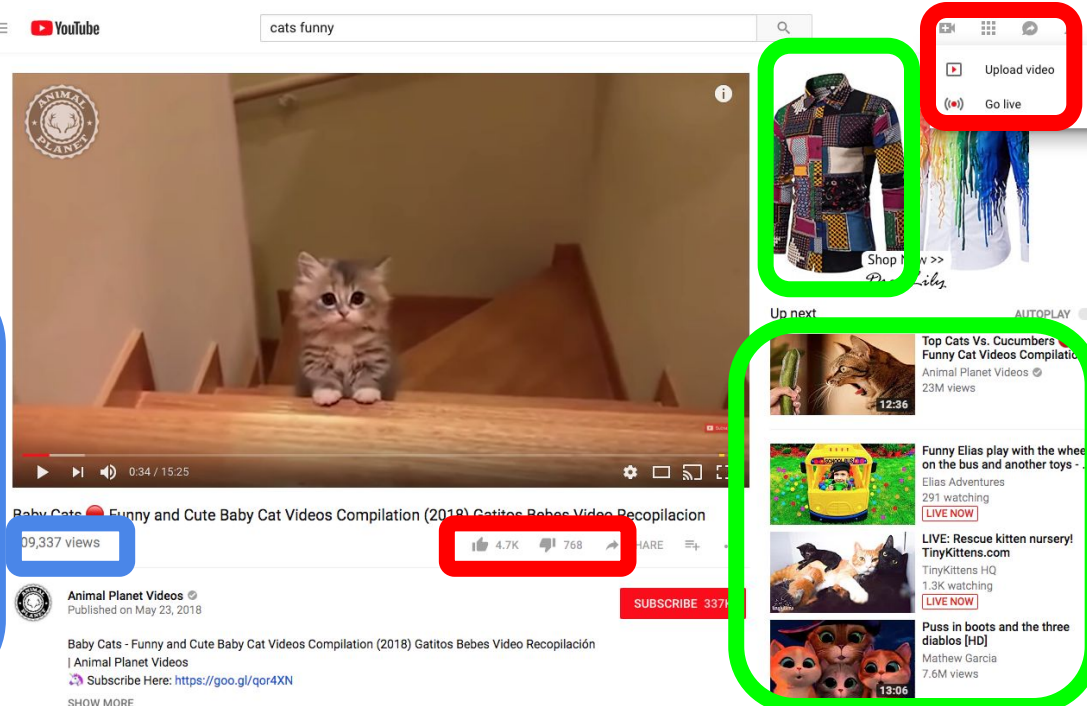
Peta scale data systems (from cs145 -> Infolab -> now)

Building new data systems, products (and teams)

- Scaled to billions of consumers, billions of ad $s, millions of web publishers, trillions of data rows, million QPS systems
- E.g., AdSense, Search, Dremel/BigQuery, Gmail/Google Apps, Sitemaps, Warp, Google Maps, etc.
- VP @Google, Founder of 2 startups (Urban Engines, Gigabeat), angel investor/advisor in more data/AI startups

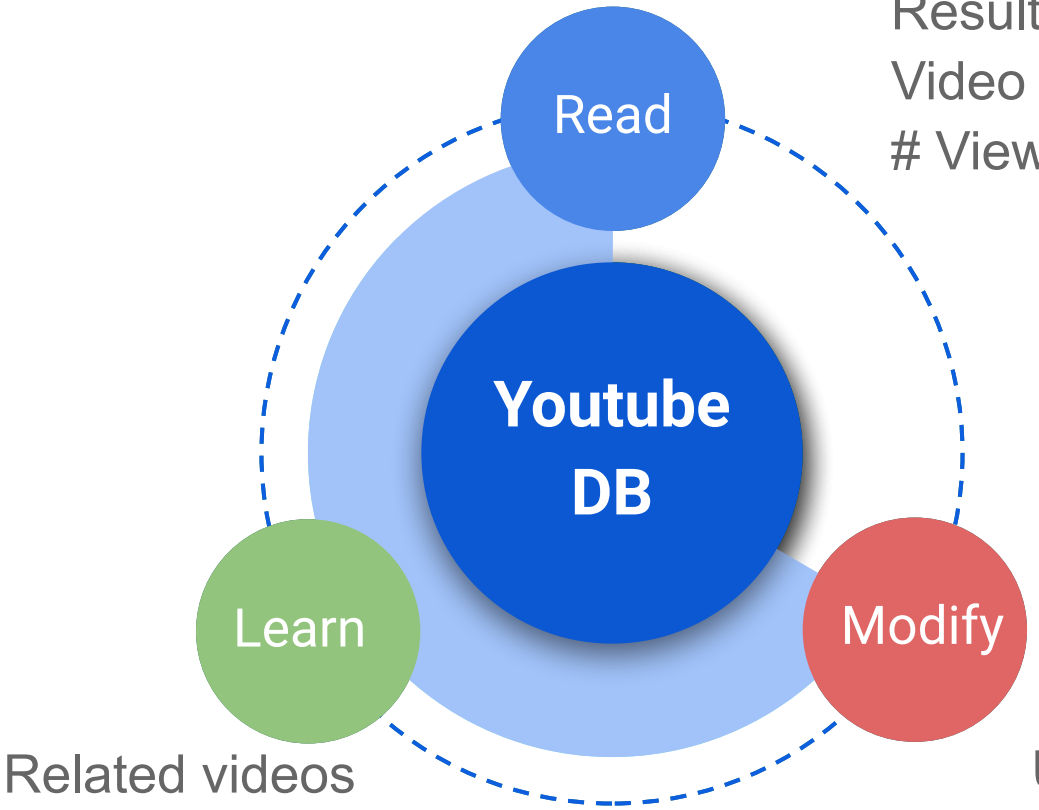
Example: Youtube DB





Example

Unpack
Youtube DB

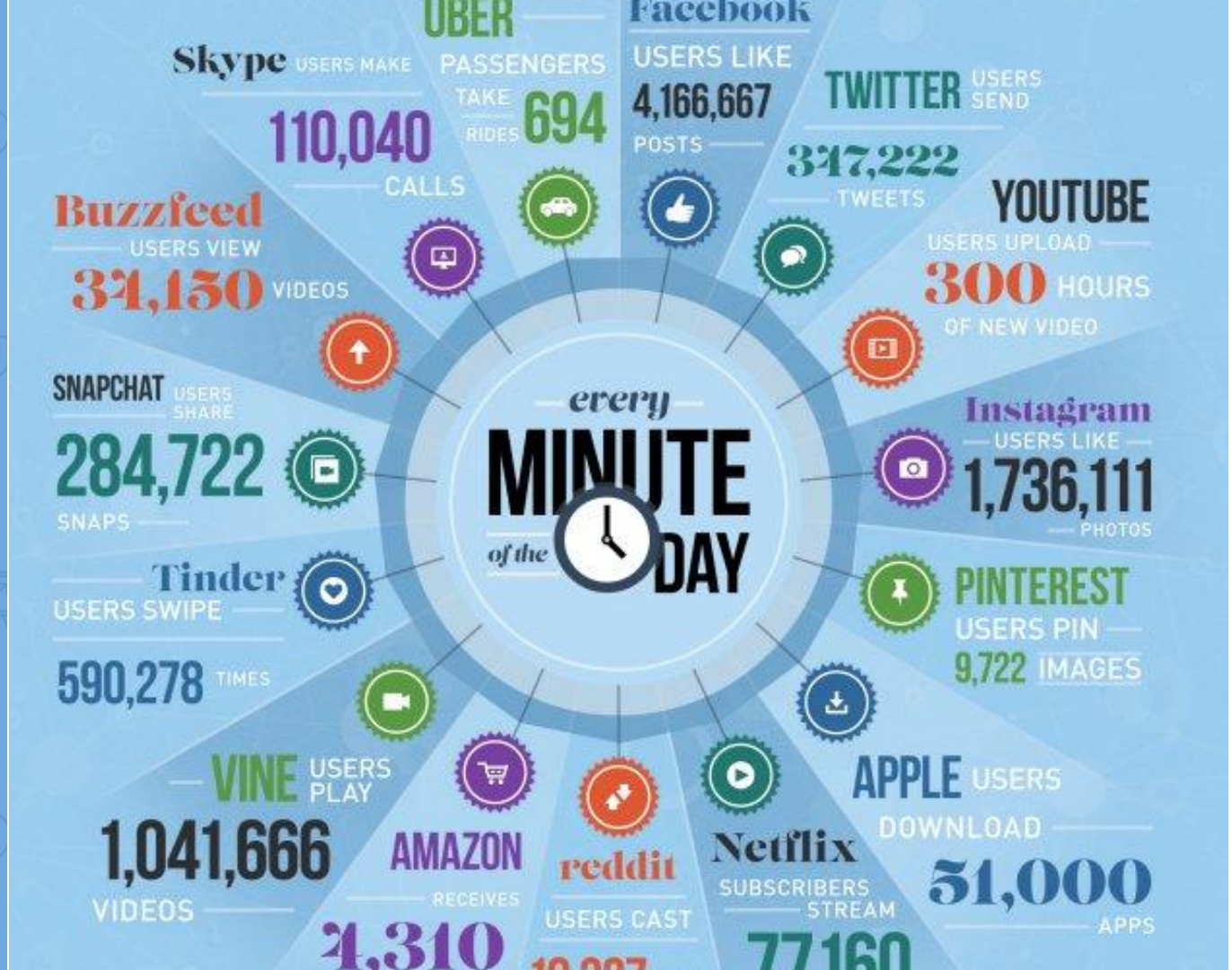


Result list
Video & description,
Views, Likes

Related videos
Relevant ads

Upload
Like, Review

Every minute on the Internet



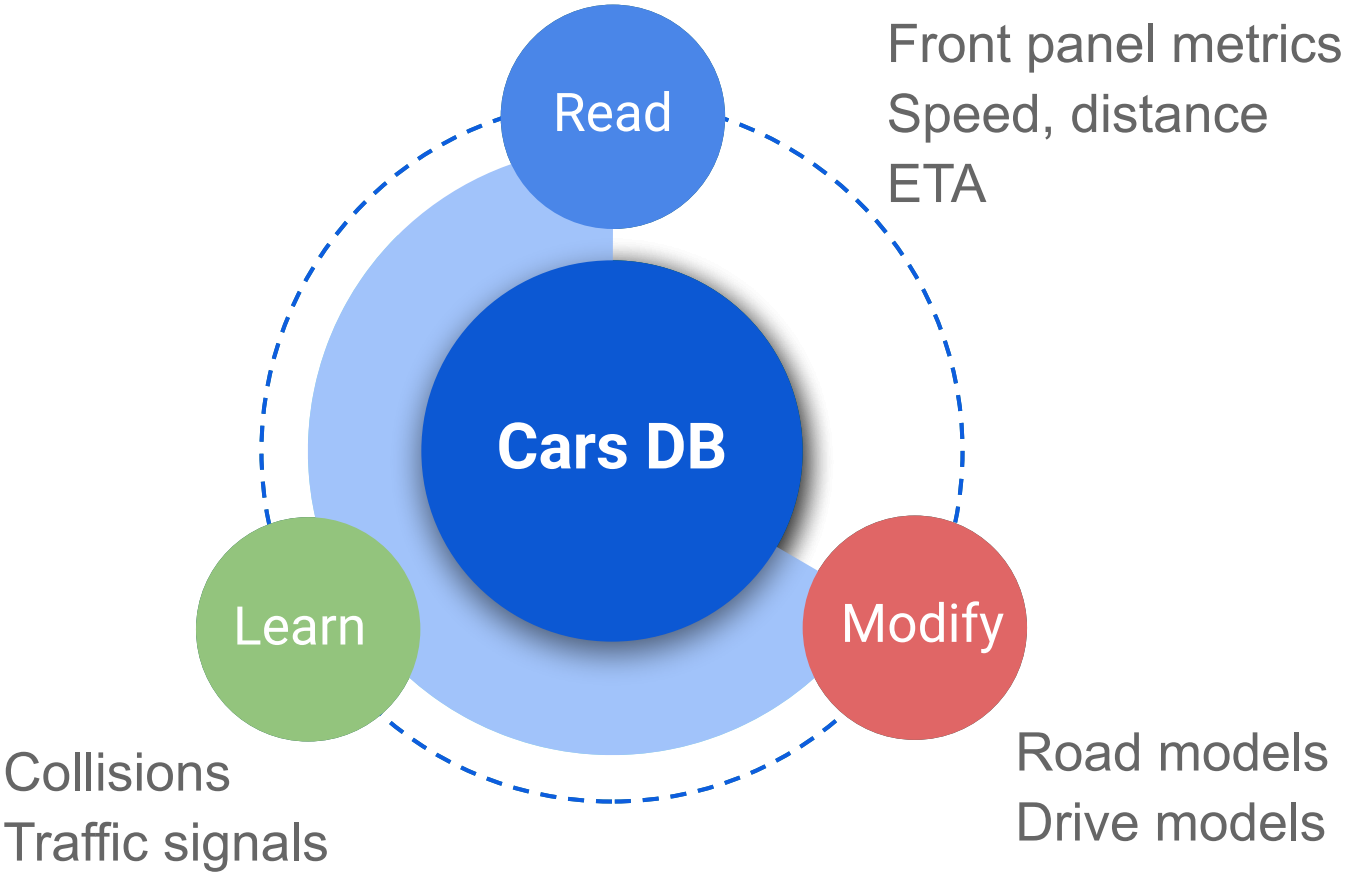
Example

Self Driving Cars

THE COMING FLOOD OF DATA IN AUTONOMOUS VEHICLES

Example

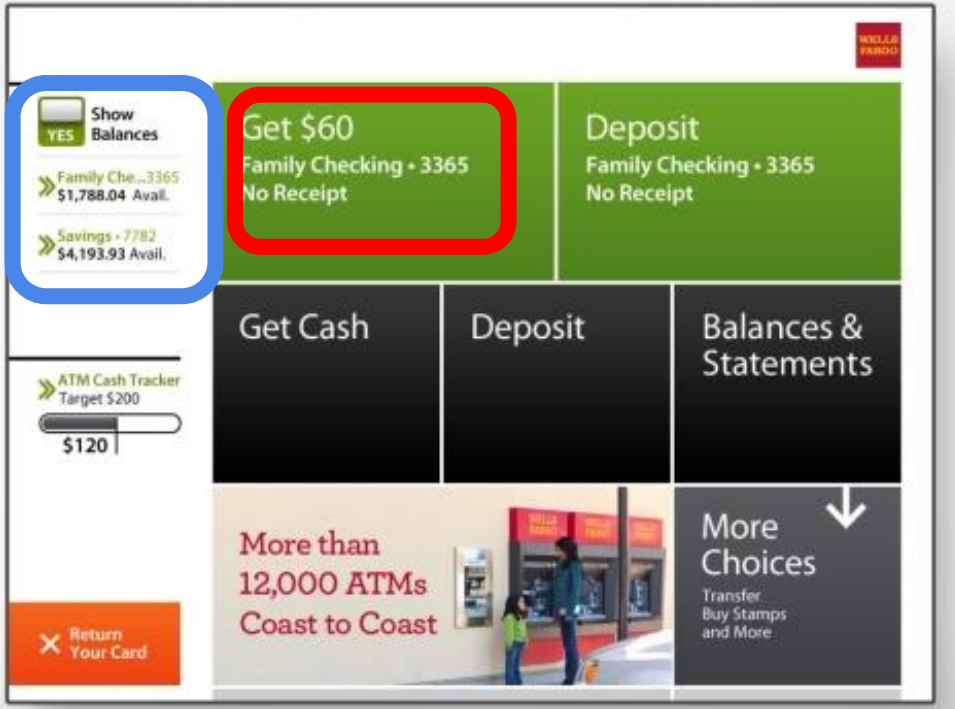
Unpack
Cars DB



Example

Unpack ATM DB:

Transaction



WELLS FARGO

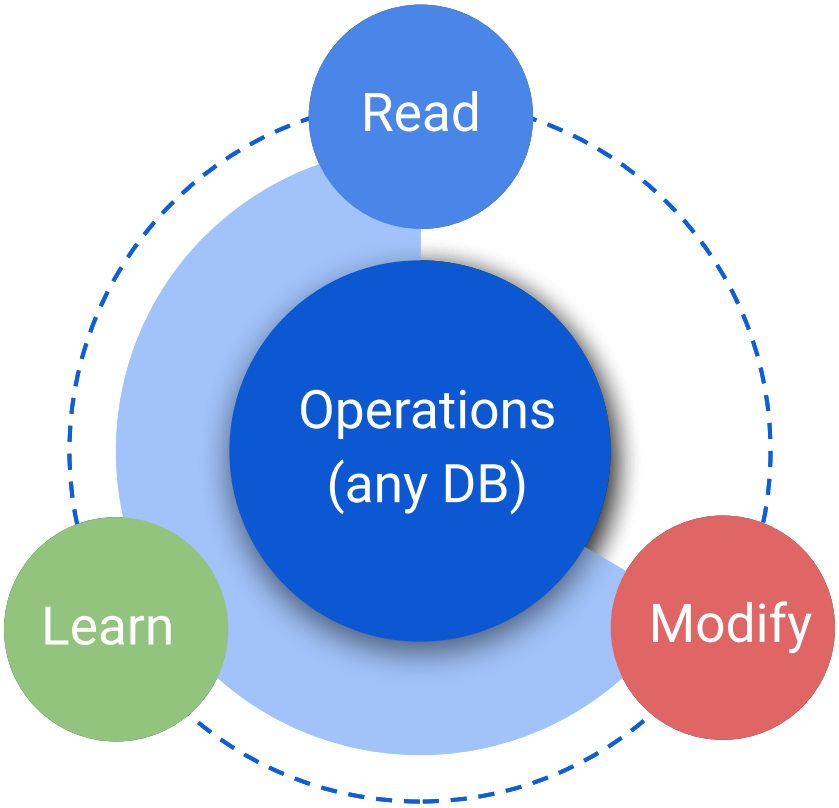
Read Balance
Give money
Update Balance

vs

Read Balance
Update Balance
Give money

Goals of Databases

Platform to store, manage data

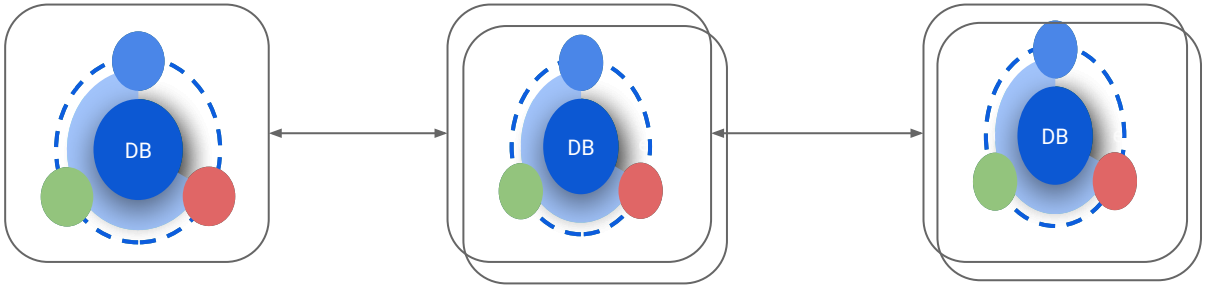


Supporting

- Scale
- Speed
- Stability
- Evolution
- Reliability
- Cost efficiency

Goals of Data systems

Connect one/many DBs for custom system



Store current data (e.g., lot of reads)

Optimize historical data (e.g., logs)

Run batch Workloads (e.g. training)

When to build a custom data system?

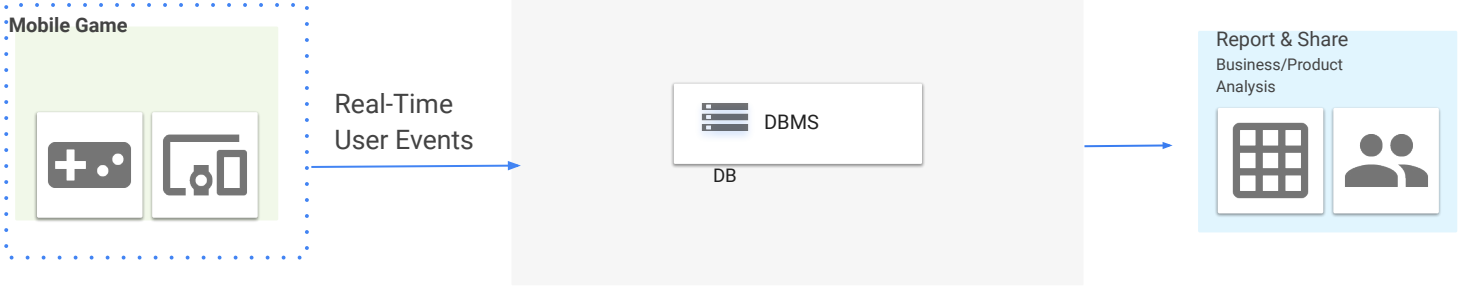
How?

For Whom?

How?

Example Game App

DB v0



Q1: 1000 users/sec?
Q2: Offline?
Q3: Support v1, v1' versions?

App designer

Q7: How to model/evolve game data?
Q8: How to scale to millions of users?
Q9: When machines die, restore game state gracefully?

Systems designer

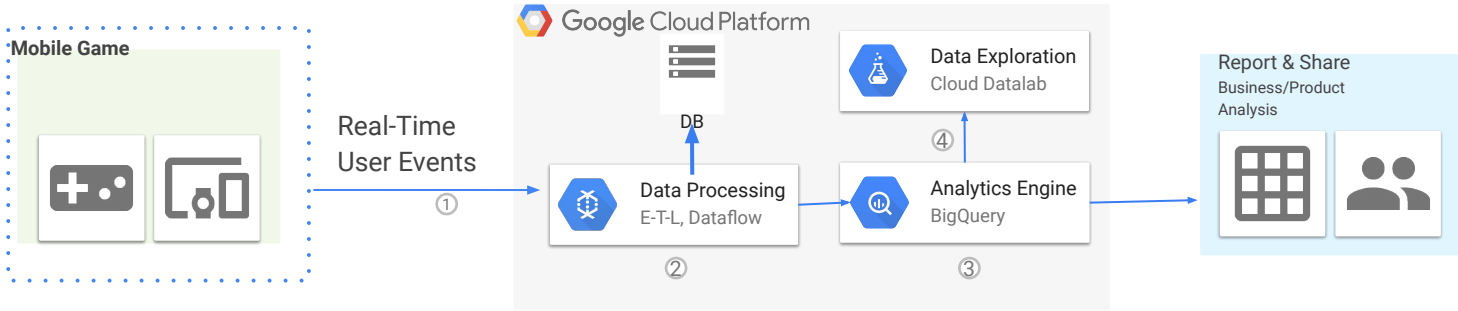
Q4: Which user cohorts?
Q5: Next features to build? Experiments to run?
Q6: Predict ads demand?

Product/Biz designer

How?

Example Game App

Data system “v1” on Cloud



1 Log user actions

2 Store in DB, after Extract-Transform-Load

3 Run queries in a peta scale analytics system

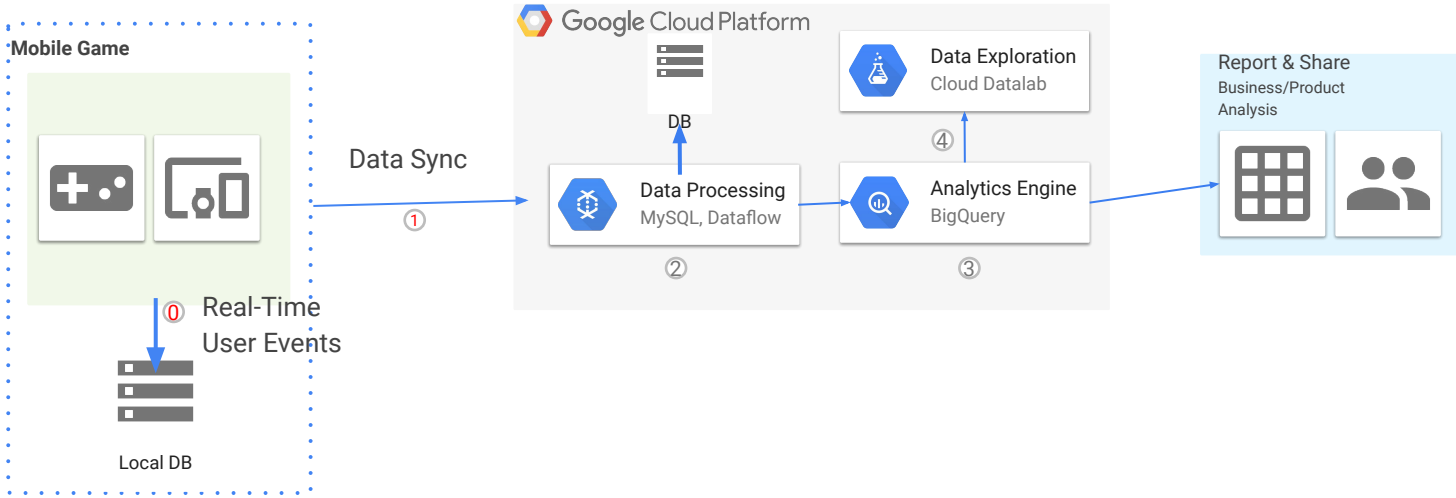
4 Visualize query results

How?

Example Game App

Data system

“v2” Cloud + Local



0 Log user actions In local DB

1 Data sync to cloud

2 Store in DB, after ETL

3 Run queries in a petabyte scale analytics system

4 Visualize query results

Course Summary

We'll learn…

- **How to query a database? With concurrent users and crashes / aborts?**

1. Intro

2-3. SQL

4. ER Diagrams

5-6. DB Design

7-8. TXNs

11-12. IO Cost

14-15. Joins

16. Rel. Algebra

Course Summary

We'll learn…

- How to query a database? With concurrent users and crashes / aborts?
- **How to design a database?**

1. Intro

2-3. SQL

4. ER Diagrams

5-6. DB Design

7-8. TXNs

11-12. IO Cost

14-15. Joins

16. Rel. Algebra

Course Summary

We'll learn…

- How to query a database? With concurrent users and crashes / aborts?
- How to design a database?
- **How to optimize performance?**

1. Intro

2-3. SQL

4. ER Diagrams

5-6. DB Design

7-8. TXNs

11-12. IO Cost

14-15. Joins

16. Rel. Algebra

Logistics

cs145.stanford.edu

Assessment Projects: 35% (5 + 10 + 20), Midterm Exam: 25%, Final Exam: 40%. We will be offering extra credit for (a) insightful in-class and OH participation (b) high quality answers to fellow student questions in piazza.

Exam dates

- Midderm: 1 hr in class midterm on Nov 1st(Thurs)
- Final: Friday, Dec 14th from 3:30-6:30 p.m. No alternate times, plan ahead.
- **Accommodations** Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE) and notify us at least 7 days (ONE week) prior to the Midterm and/or Final Exam.

Homeworks/Suggested reading

- Posted on cs145 webpage. We'll use Gradiance. It's an auto-grader to help reinforce concepts AS-You-GO. Strongly encouraged for prepping materials for exams.