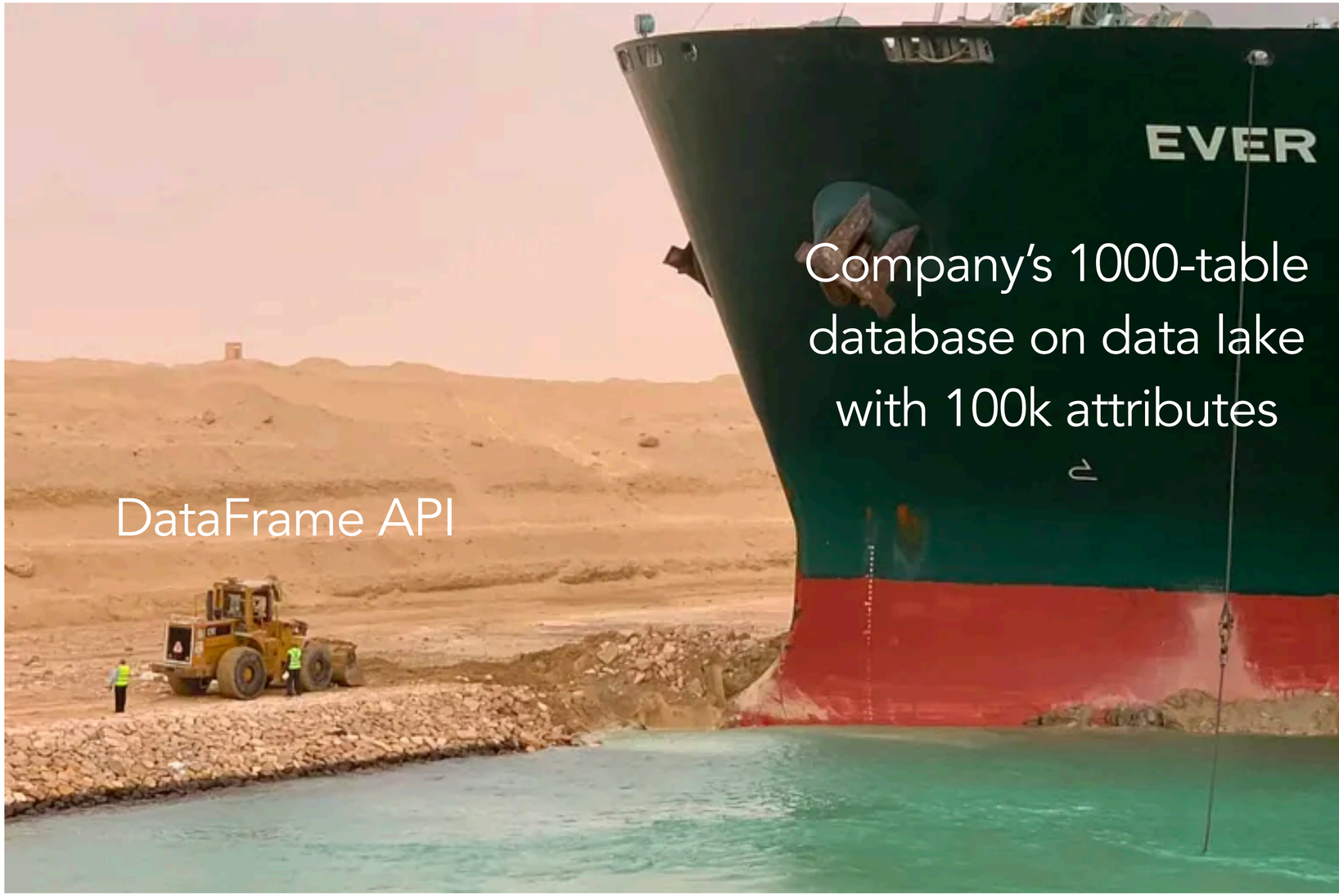
DSC 204a Scalable Data Systems

- Haojian Jin



Meme idea credit: https://datasystemsfun.tumblr.com/

Where are we in the class?

Foundations of Data Systems (2 weeks)

- Digital representation of Data → Computer Organization → Memory hierarchy → Process → Storage

Scaling Distributed Systems (3 weeks)

- Cloud → Network → **Distributed storage** → Partition and replication (HDFS) → Distributed computation

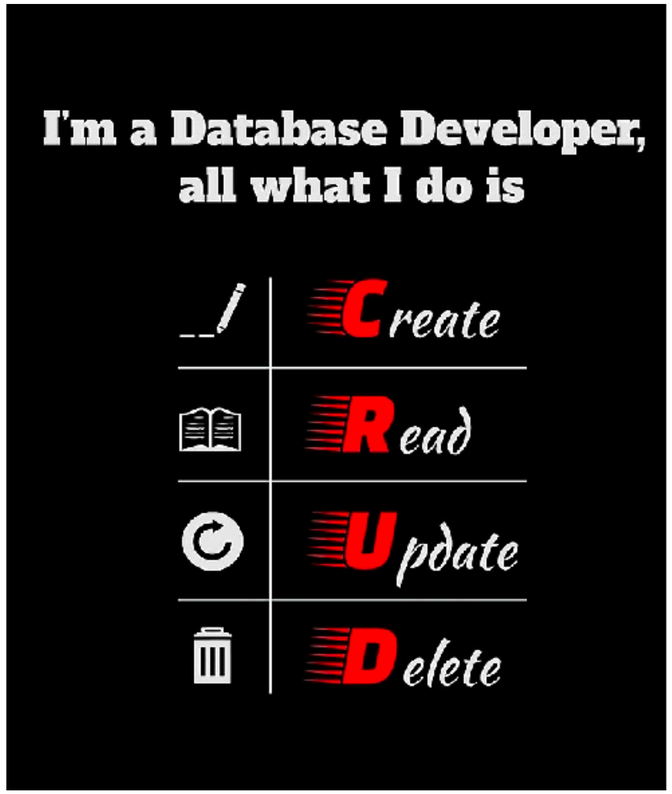
Data Processing and Programming model (5 weeks)

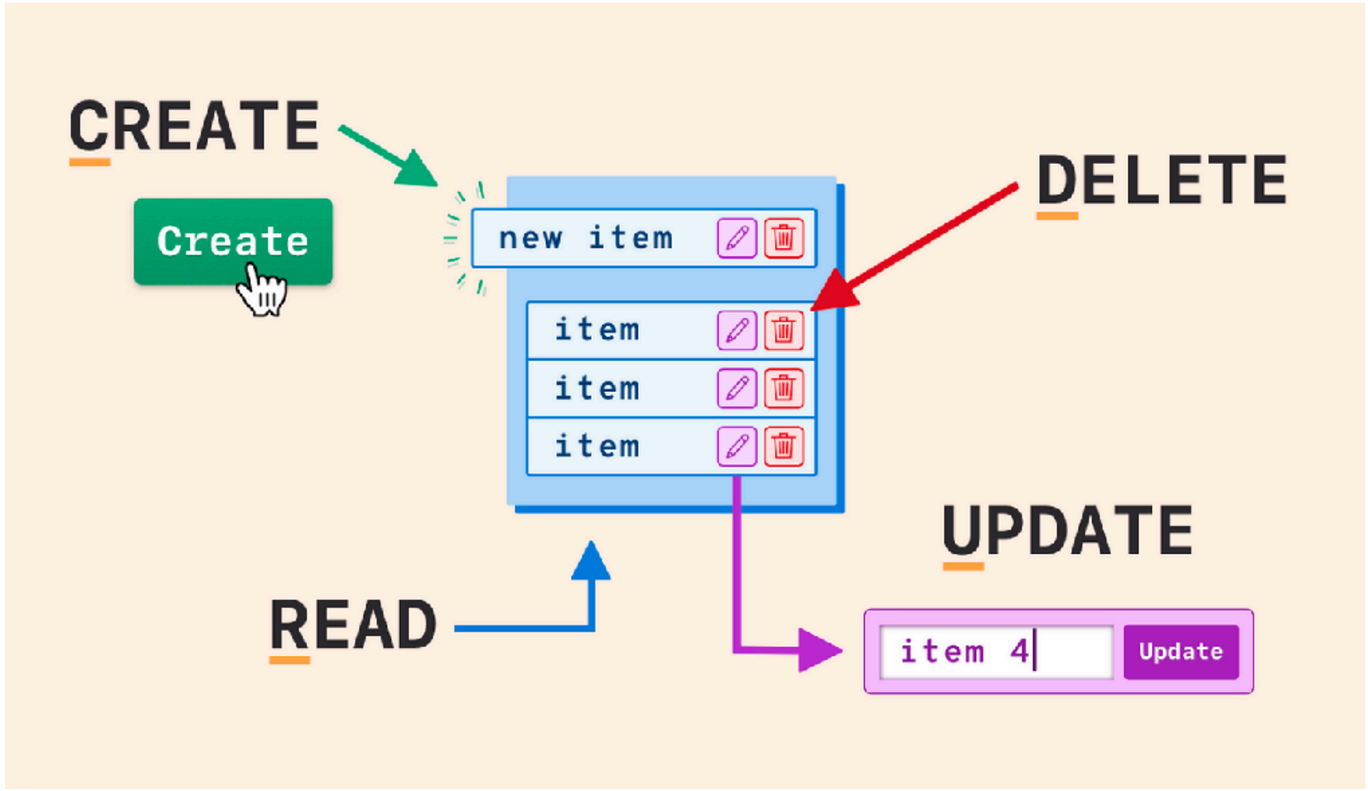
- Data Models evolution → Data encoding evolution → → IO & Unix Pipes → Batch processing (MapReduce) → Stream processing (Spark)

Today's topic: Column-oriented storage

- OLTP v.s. OLAP
- Data warehousing
- Schemas for Analytics
- Column-oriented storage
- Data cubes and materialized views

CRUD


I'm a Database Developer,
all what I do is
Create
Read
Update
Delete


CREATE
Create
new item
item
item
item
DELETE
UPDATE
READ
item 4
Update

Database transactions

- Make sale
- Place an order
- Pay an employee’s salary
- Comment a blog post
- Act in games
- Add/remove contract to an address book

Online transaction processing (OLTP)

Walmart Beer and Diaper (1988)

Forbes 1988

- Unexpected correlation:
  - Sales of diapers and beer

https://www.forbes.com/forbes/1998/0406/6107128a.html?sh=2574a9316260

Data analytics

- What was the total revenue of each of our stores in Jan?
- How many more bananas that usual did we sell during our latest data?
- Which brand of baby food is most often purchased together with brand X diapers?

Online analytic processing (OLAP)

OLTP v.s. OLAP

Property	Transaction processing systems (OLTP)	Analytic systems (OLAP)
Main read pattern	Small number of records per query, fetched by key	Aggregate over large number of records
Main write pattern	Random-access, low-latency writes from user input	Bulk import (ETL) or event stream
Primarily used by	End user/customer, via web application	Internal analyst, for decision support
What data represents	Latest state of data (current point in time)	History of events that happened over time
Dataset size	Gigabytes to terabytes	Terabytes to petabytes