UNIVERSITY OF WATERLOO

Data-Intensive Distributed Computing

CS 431/631 451/651 (Winter 2019)

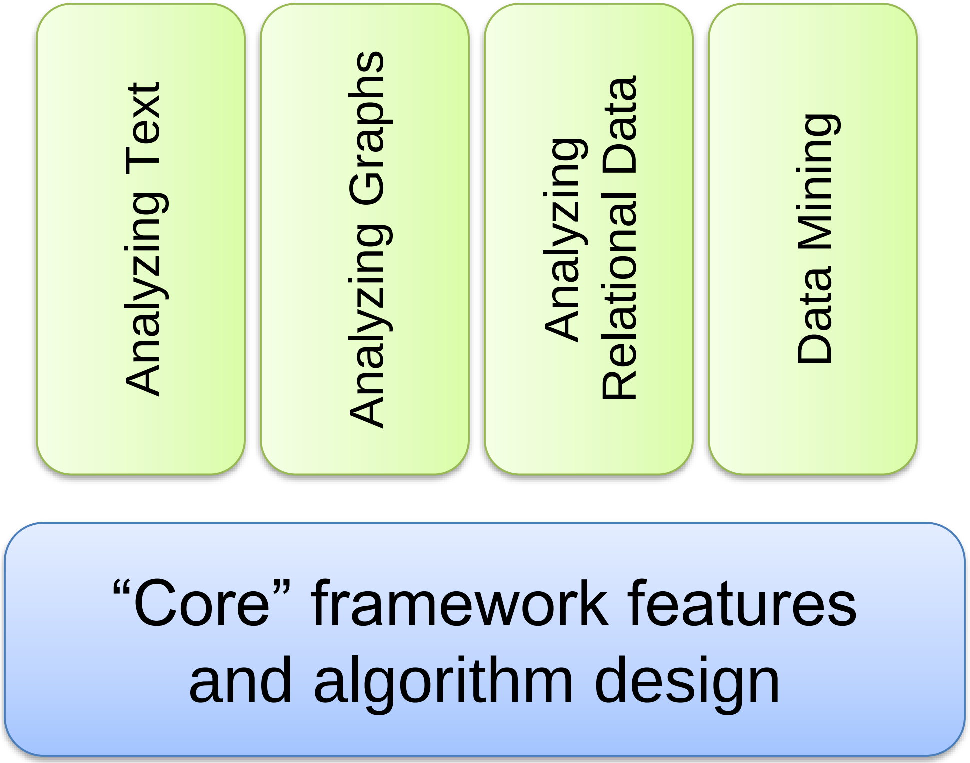
Part 5: Analyzing Relational Data (1/3)

February 12, 2019

Adam Roegiest
Kira Systems

These slides are available at http://roegiest.com/bigdata-2019w/

Structure of the Course



Analyzing Text

Analyzing Graphs

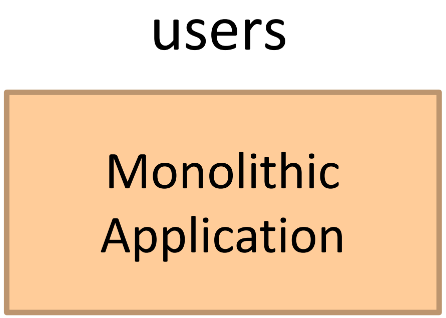
Analyzing Relational Data

Data Mining

“Core” framework features and algorithm design

Evolution of Enterprise Architectures

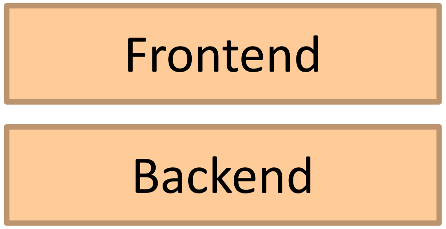
Next two sessions: techniques, algorithms, and optimizations for relational processing



users

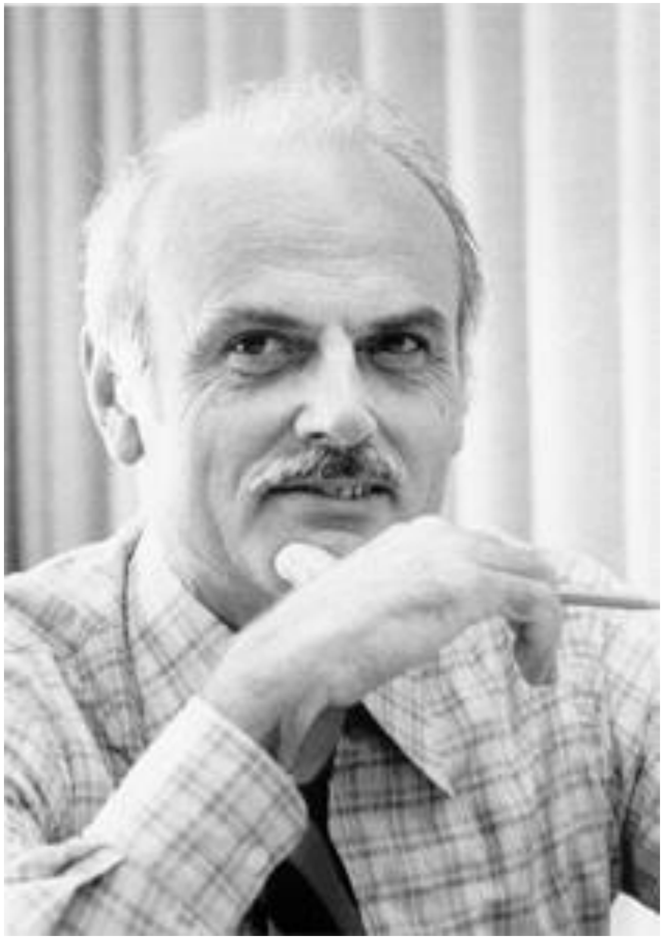
Monolithic Application

users

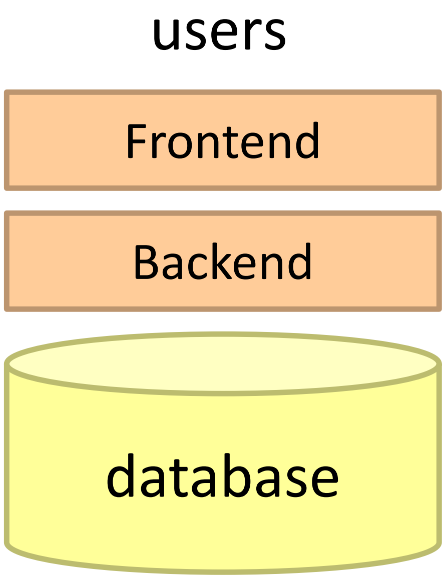


Frontend

Backend



Source: Wikipedia



users

Frontend

Backend

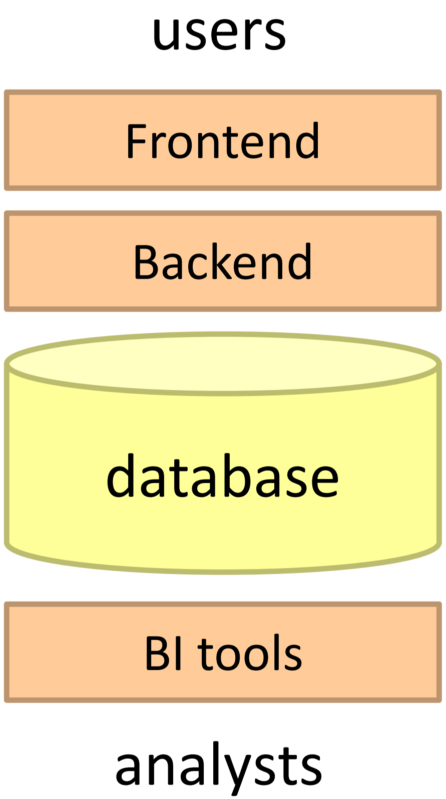
database

Why is this a good idea?

Business Intelligence

An organization should retain data that result from carrying out its mission and exploit those data to generate insights that benefit the organization, for example, market analysis, strategic planning, decision making, etc.

Duh!?



users

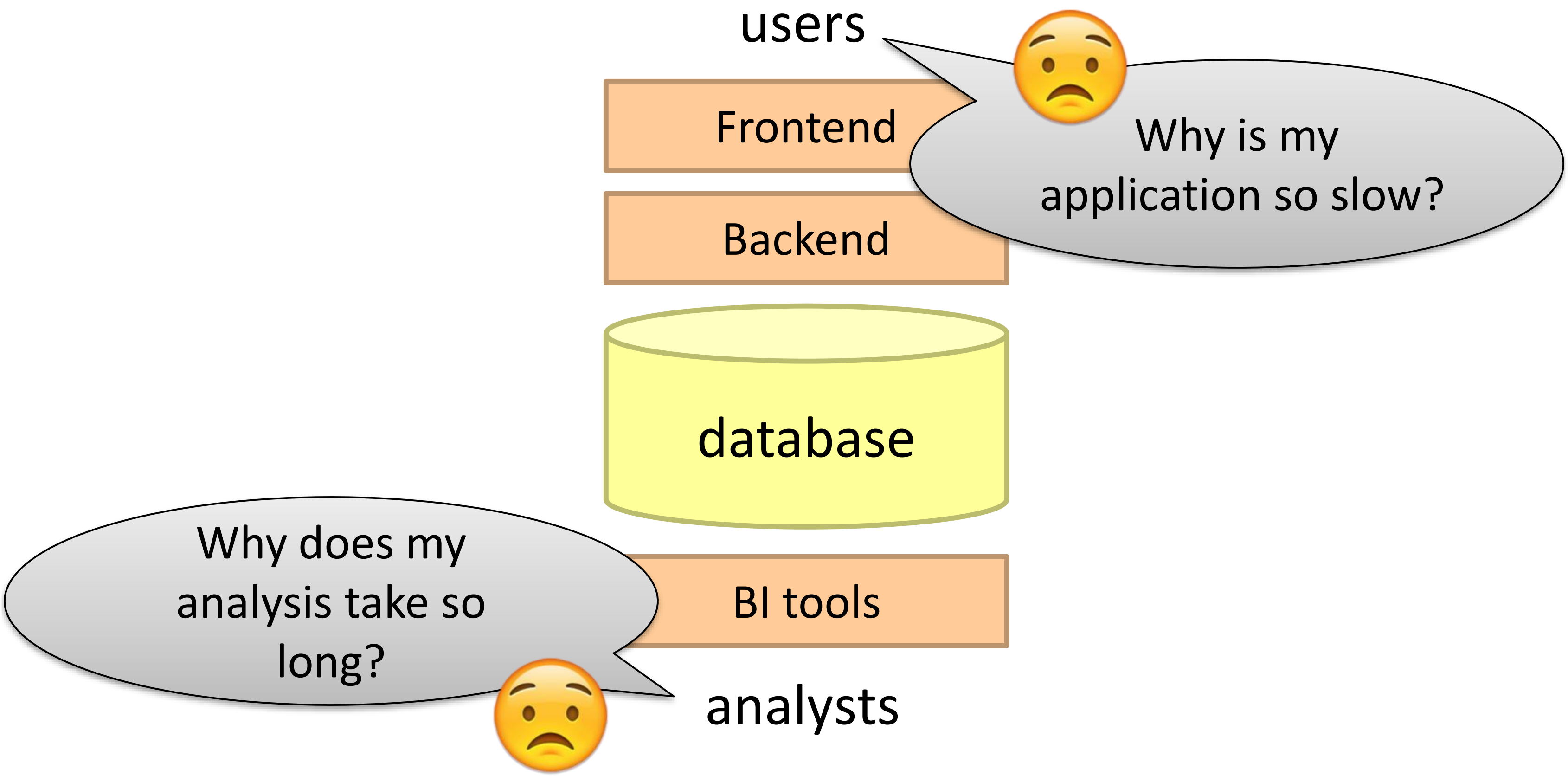
Frontend

Backend

database

BI tools

analysts



users

Frontend

Backend

database

BI tools

analysts

Why is my application so slow?

Why does my analysis take so long?

Database Workloads

OLTP (online transaction processing)

Typical applications: e-commerce, banking, airline reservations

User facing: real-time, low latency, highly-concurrent

Tasks: relatively small set of “standard” transactional queries

Data access pattern: random reads, updates, writes (small amounts of data)

OLAP (online analytical processing)

Typical applications: business intelligence, data mining

Back-end processing: batch workloads, less concurrency

Tasks: complex analytical queries, often ad hoc

Data access pattern: table scans, large amounts of data per query

OLTP and OLAP Together?

Downsides of co-existing OLTP and OLAP workloads

Poor memory management

Conflicting data access patterns

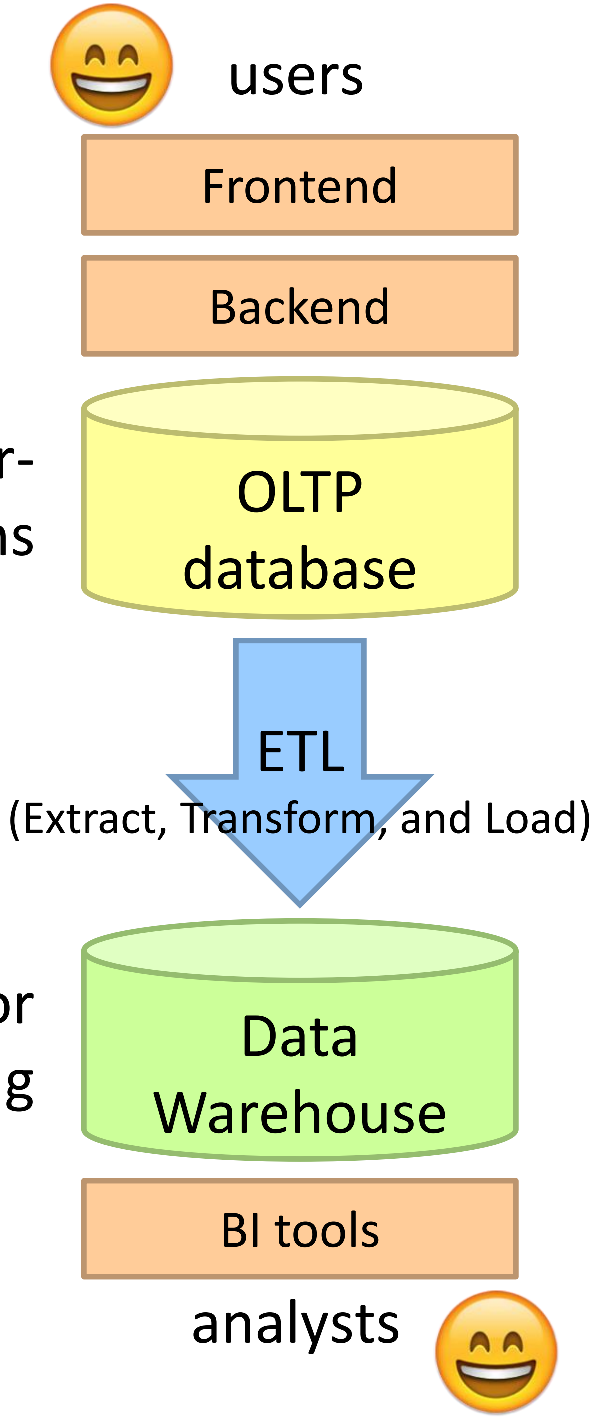
Variable latency

😟 users and analysts

Solution?

Build a data warehouse!

Source: Wikipedia (Warehouse)



users

Frontend

Backend

OLTP database

ETL

(Extract, Transform, and Load)

Data Warehouse

BI tools

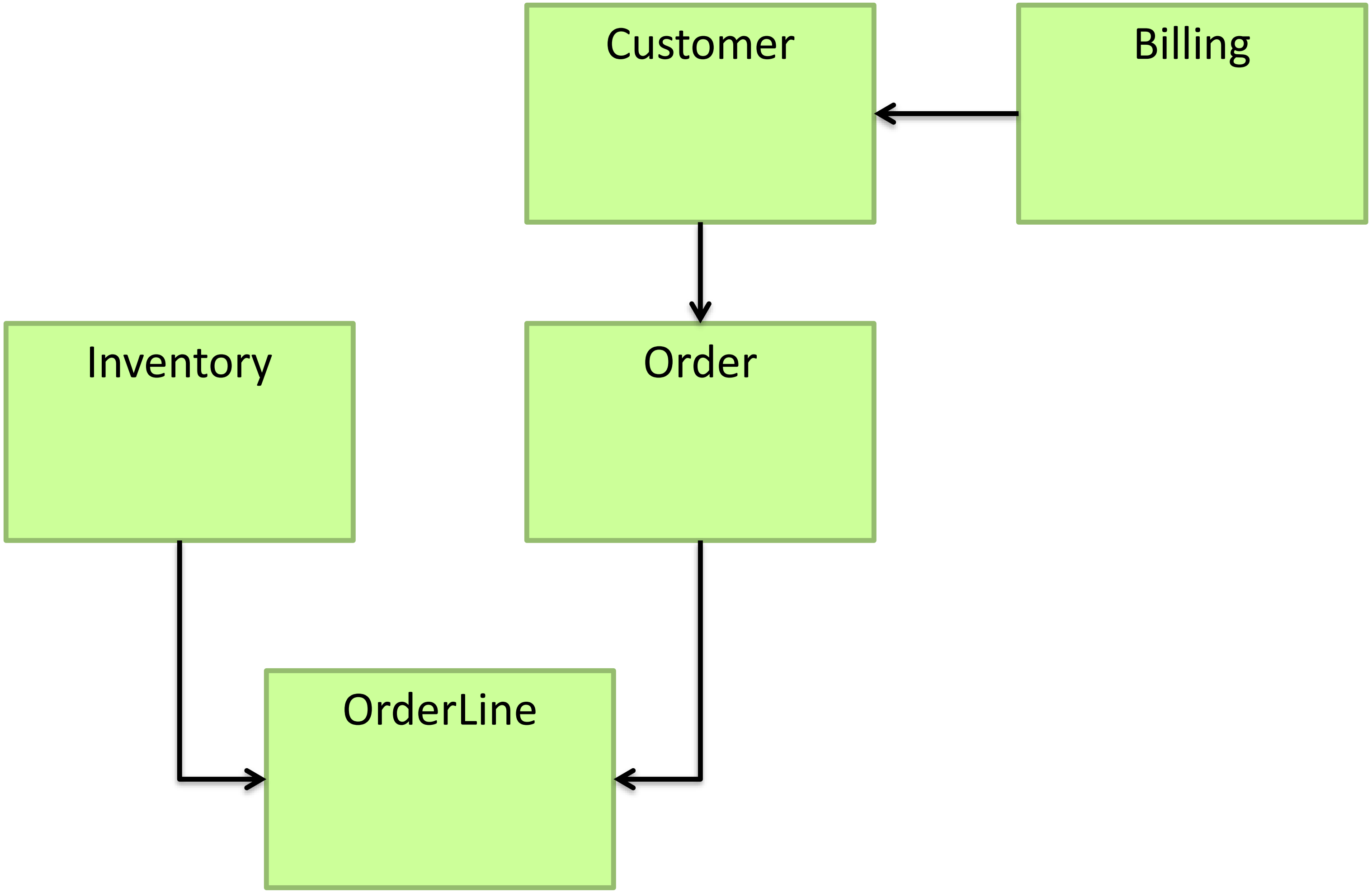
analysts

OLTP database for user-facing transactions

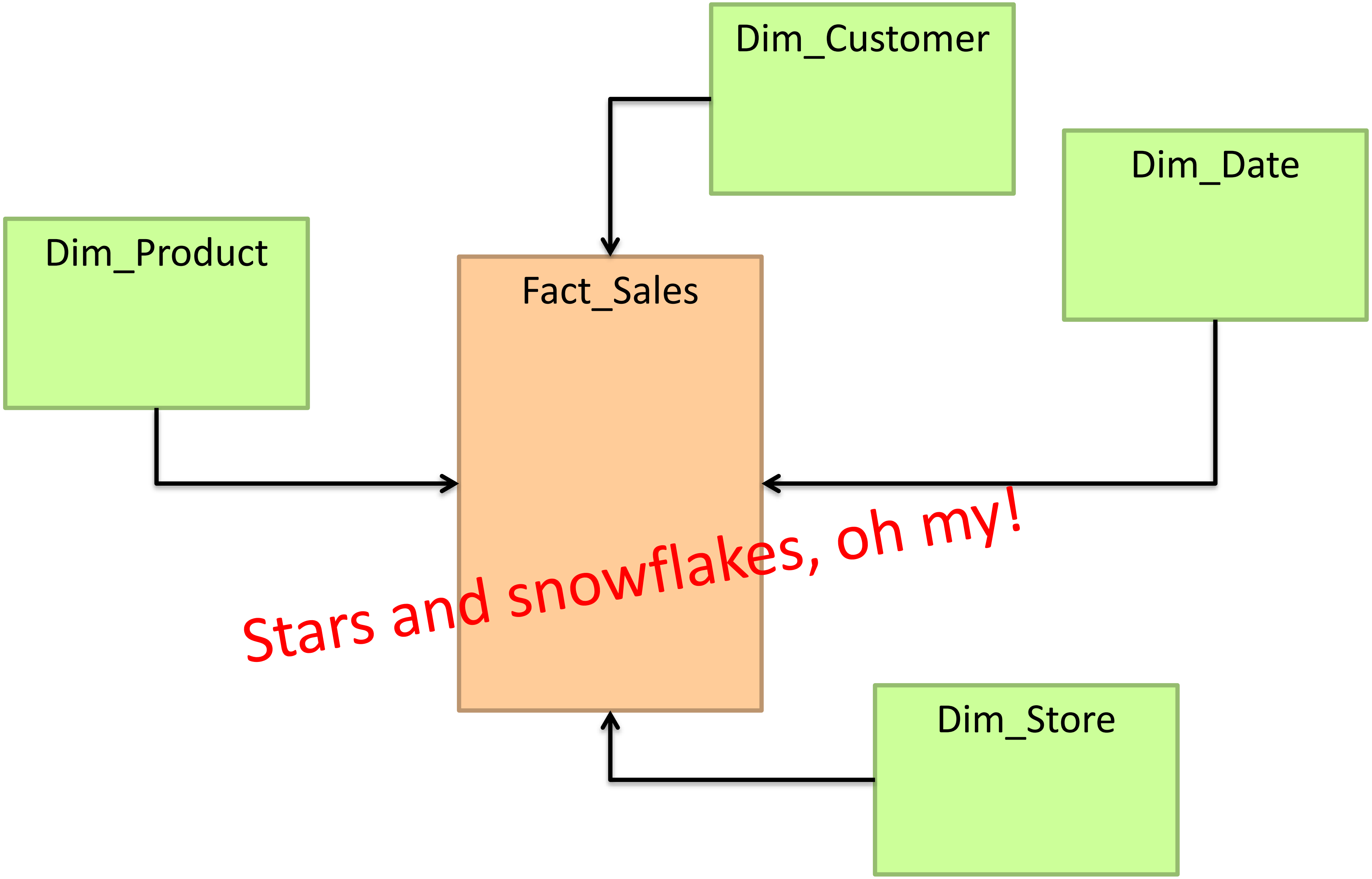
OLAP database for data warehousing

What’s special about OLTP vs. OLAP?

A Simple OLTP Schema



A Simple OLAP Schema



Stars and snowflakes, oh my!

ETL

Extract

Transform

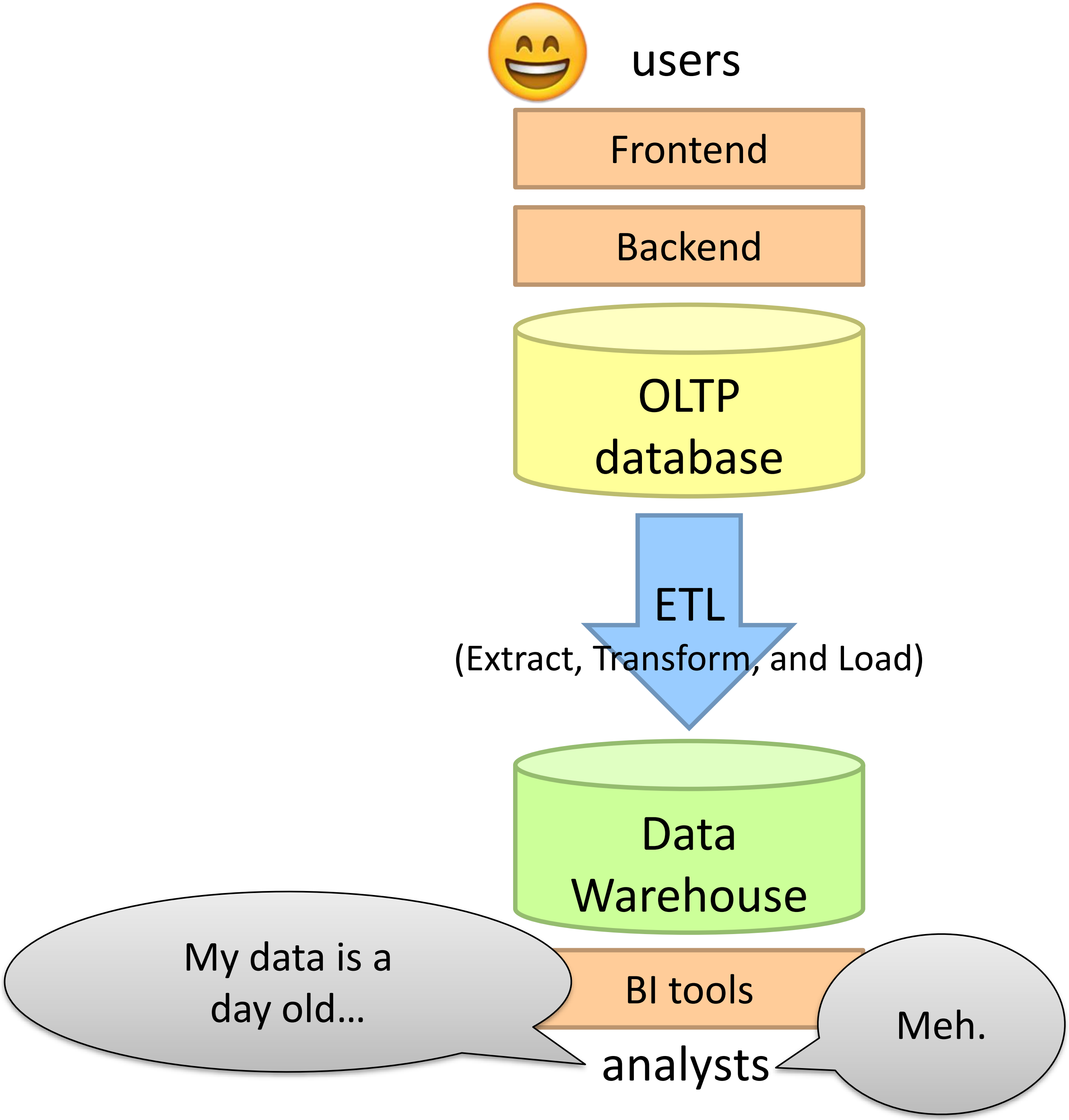
Data cleaning and integrity checking

Schema conversion

Field transformations

Load

When does ETL happen?



users

Frontend

Backend

OLTP database

ETL

(Extract, Transform, and Load)

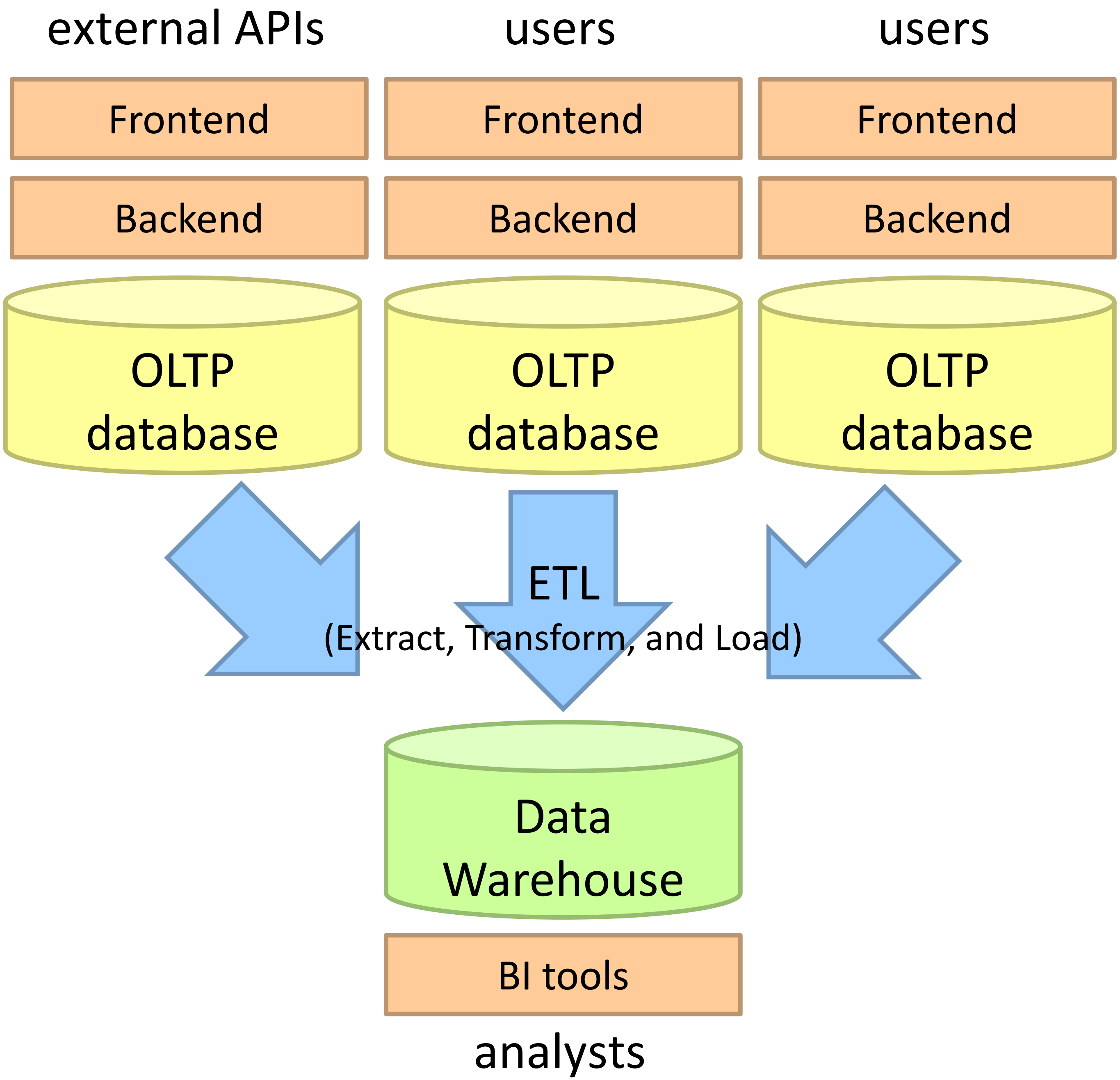
Data Warehouse

BI tools

analysts

My data is a day old…

Meh.


external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
Data Warehouse
BI tools
analysts

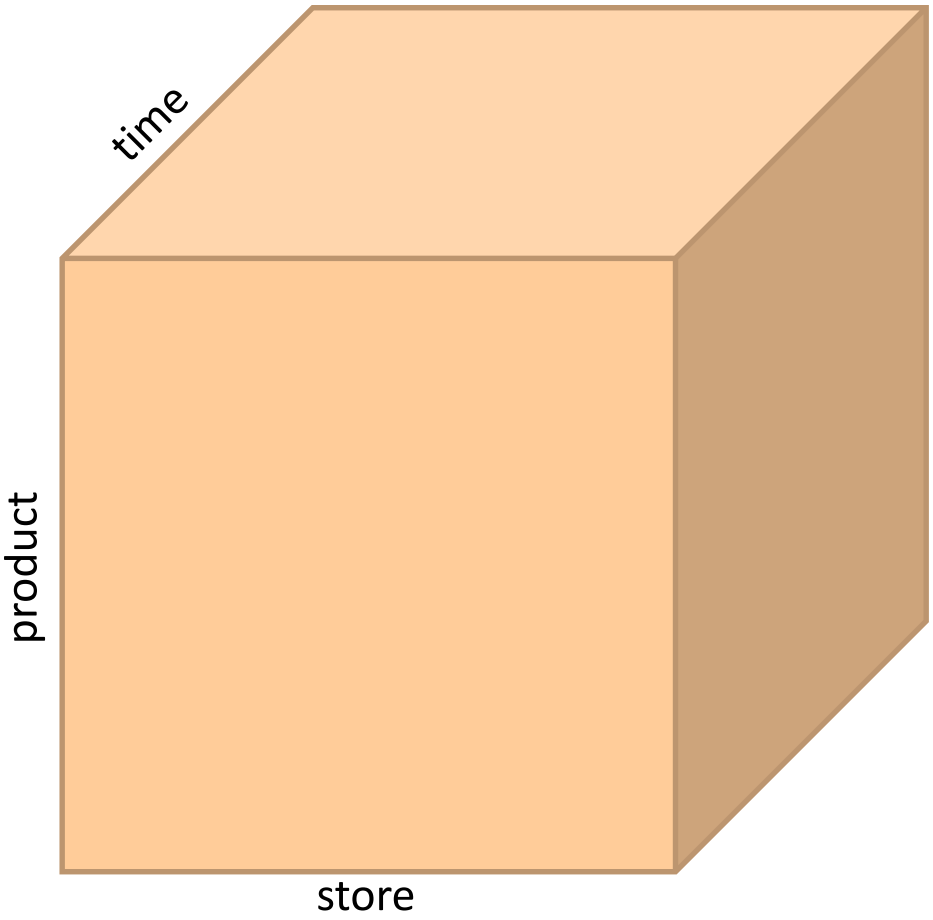
What do you actually do?

Report generation

Dashboards

Ad hoc analyses

OLAP Cubes



Common operations

- slice and dice
- roll up/drill down
- pivot

OLAP Cubes: Challenges

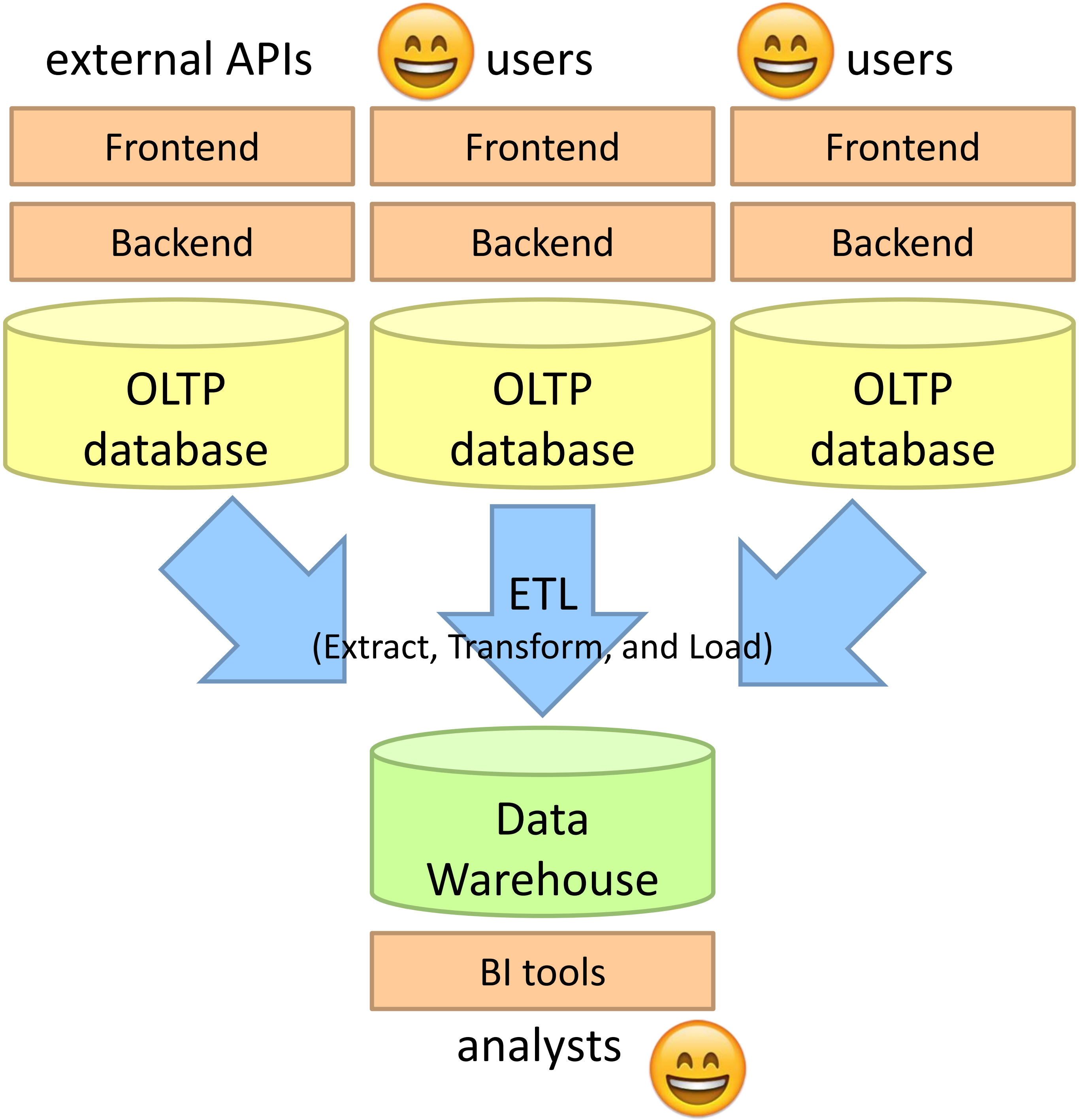
Fundamentally, lots of joins, group-bys and aggregations

How to take advantage of schema structure to avoid repeated work?

Cube materialization

Realistic to materialize the entire cube?

If not, how/when/what to materialize?



external APIs

users

users

Frontend

Frontend

Frontend

Backend

Backend

Backend

OLTP database

OLTP database

OLTP database

ETL

(Extract, Transform, and Load)

Data Warehouse

BI tools

analysts

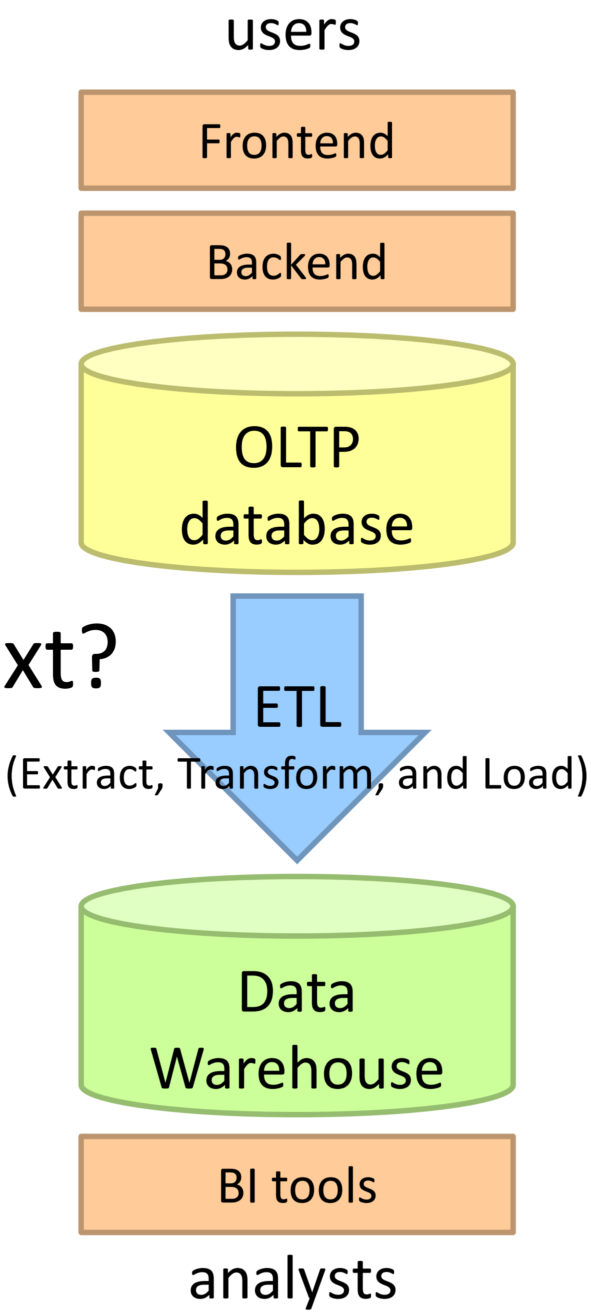
Fast forward…

facebook®

Jeff Hammerbacher, Information Platforms and the Rise of the Data Scientist.
In, *Beautiful Data*, O'Reilly, 2009.

> "On the first day of logging the Facebook clickstream, more than 400 gigabytes of data was collected. The load, index, and aggregation processes for this data set really taxed the Oracle data warehouse. Even after significant tuning, we were unable to aggregate a day of clickstream data in less than 24 hours."

Facebook context?



users

Frontend

Backend

OLTP database

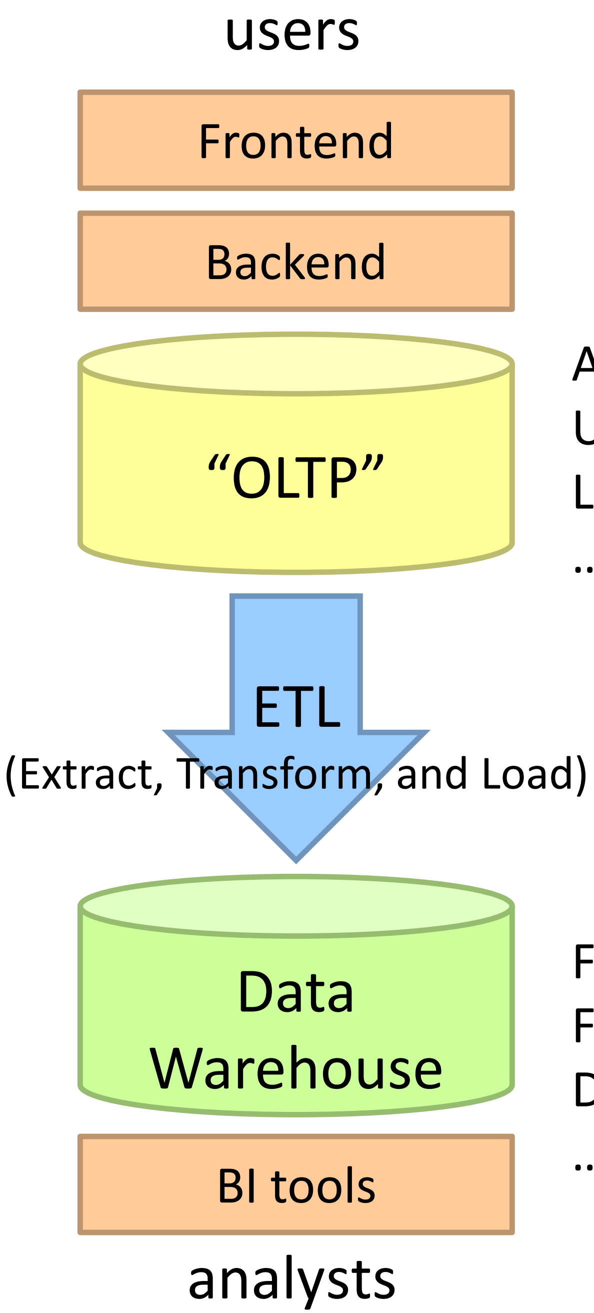
ETL

(Extract, Transform, and Load)

Data Warehouse

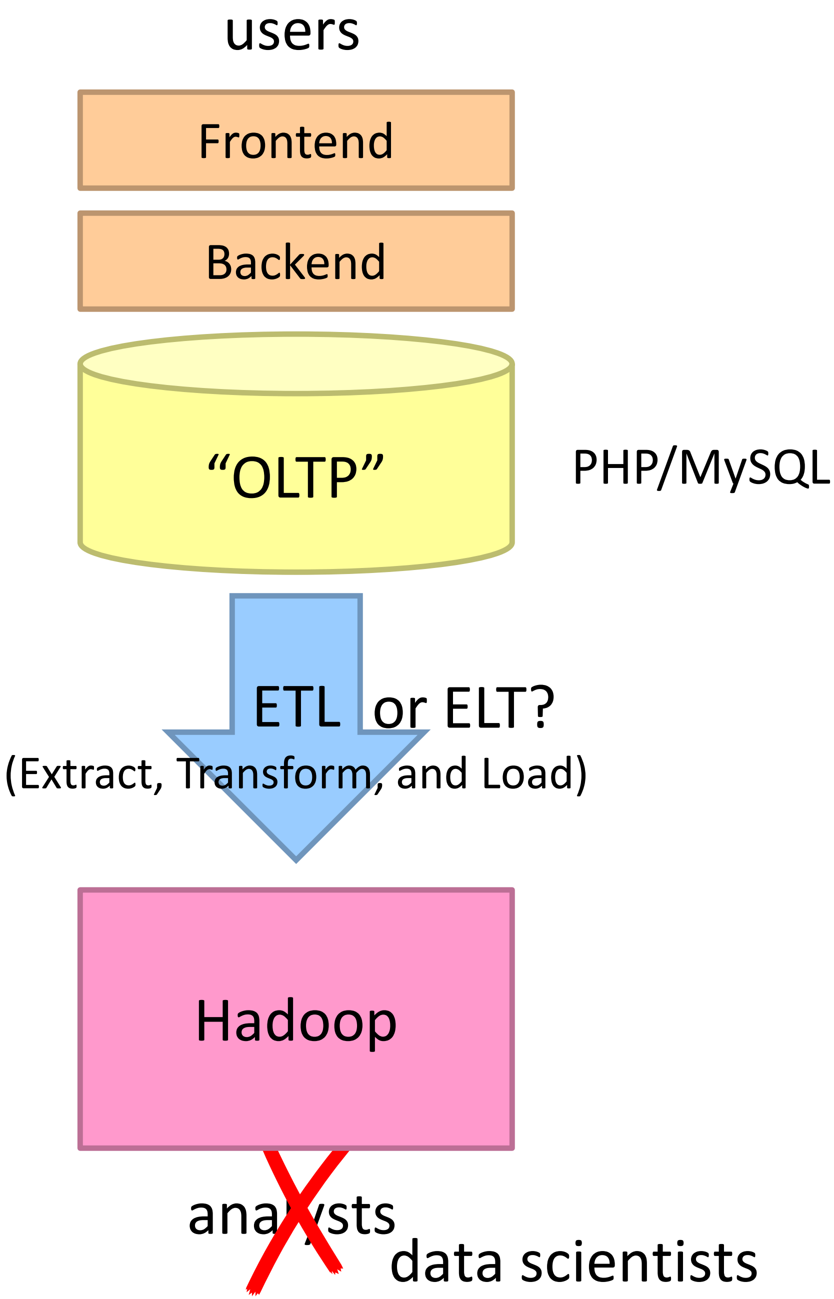
BI tools

analysts



Adding friends
Updating profiles
Likes, comments
…

Feed ranking
Friend recommendation
Demographic analysis
…



users

Frontend

Backend

“OLTP”

PHP/MySQL

ETL or ELT?

(Extract, Transform, and Load)

Hadoop

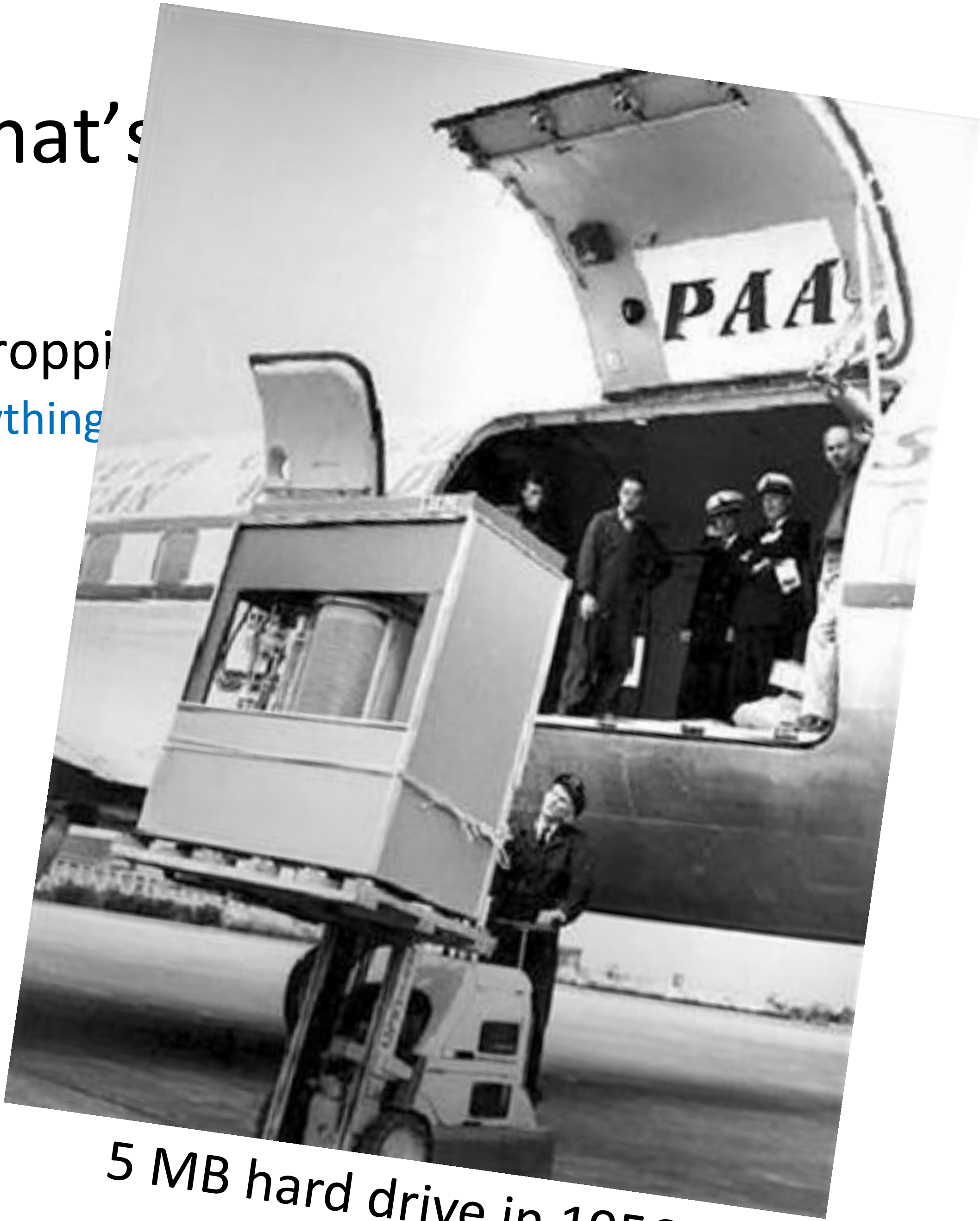
~~analysts~~

data scientists

What’s

Droppi

Cheaper to store everything



5 MB hard drive in 1956

What’s changed?

Dropping cost of disks

Cheaper to store everything than to figure out what to throw away

Types of data collected

From data that’s *obviously* valuable to data whose value is less apparent

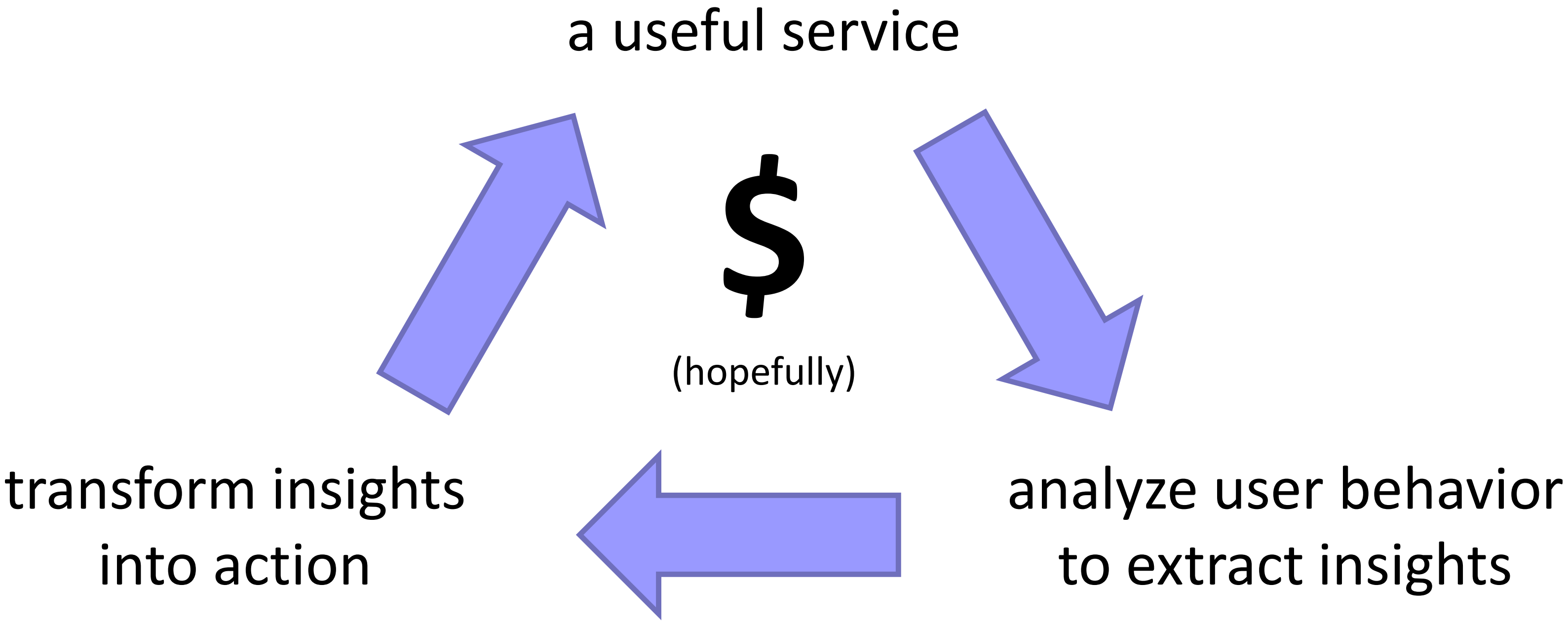
Rise of social media and user-generated content

Large increase in data volume

Growing maturity of data mining techniques

Demonstrates value of data analytics

Virtuous Product Cycle



a useful service

$

(hopefully)

analyze user behavior
to extract insights

transform insights
into action

Google. Facebook. Twitter. Amazon. Uber.

What do you actually do?

Report generation

Dashboards

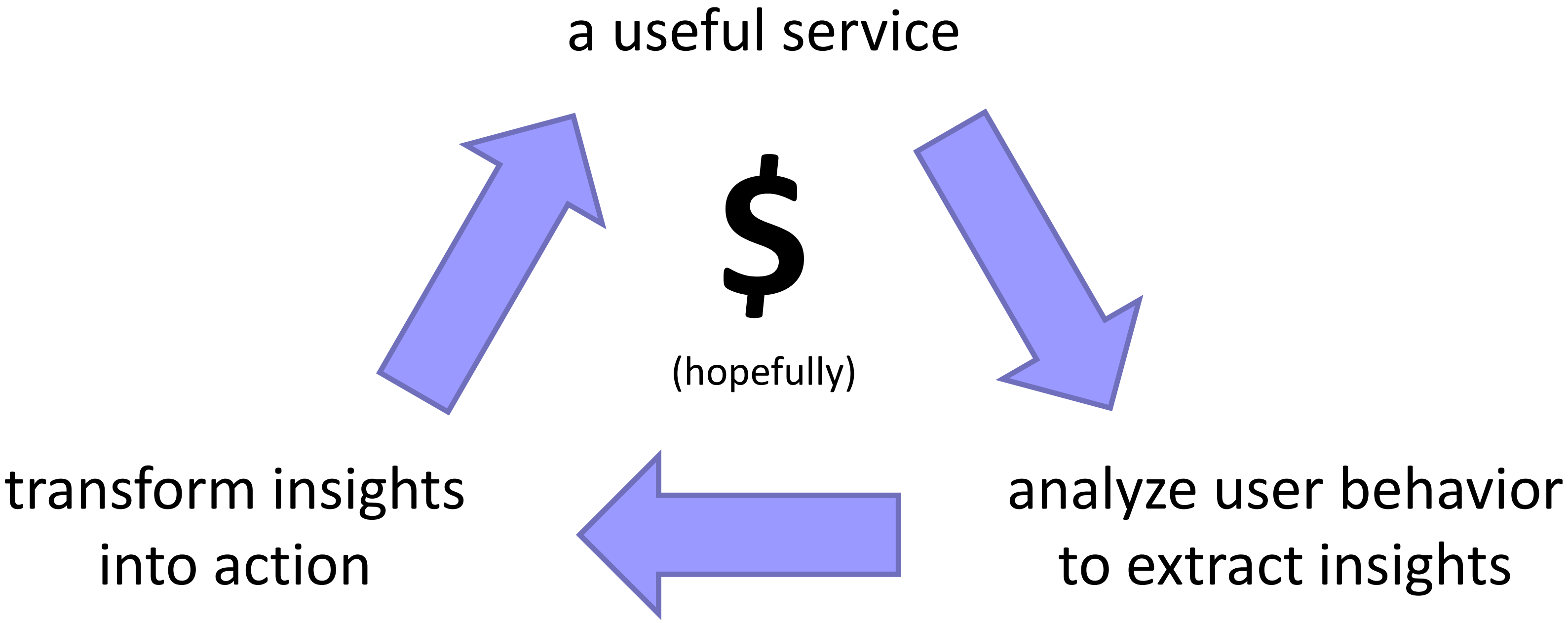
Ad hoc analyses

“Descriptive”

“Predictive”

Data products

Virtuous Product Cycle



a useful service

$

(hopefully)

analyze user behavior
to extract insights

transform insights
into action

Google. Facebook. Twitter. Amazon. Uber.

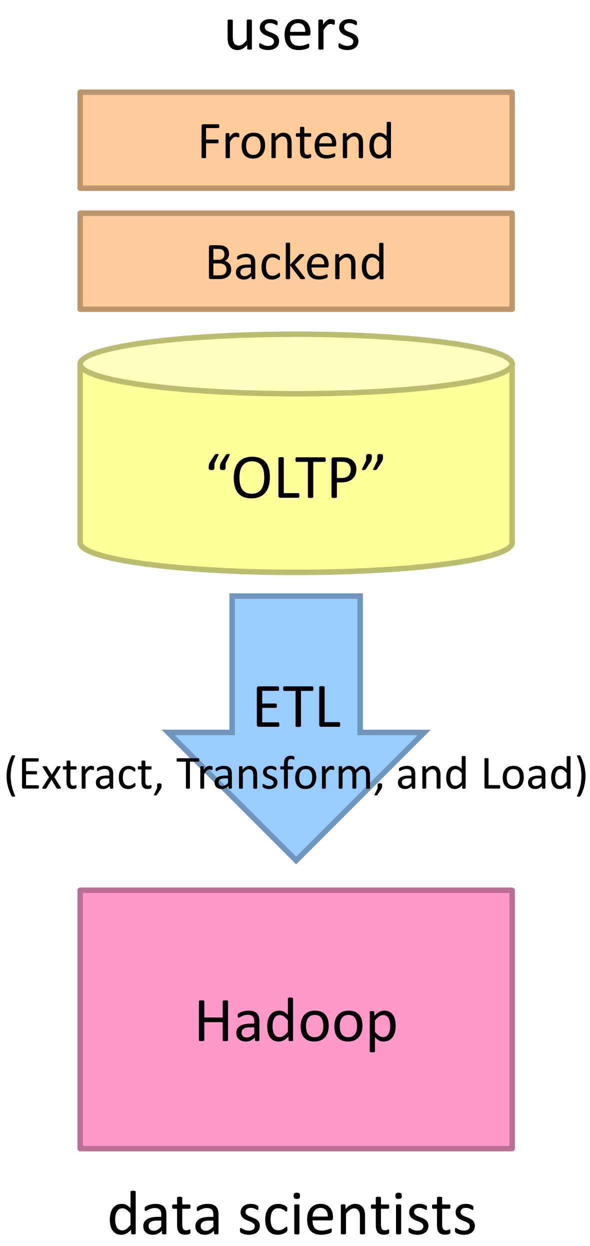
data products

data science

facebook®

Jeff Hammerbacher, Information Platforms and the Rise of the Data Scientist.
In, *Beautiful Data*, O'Reilly, 2009.

> "On the first day of logging the Facebook clickstream, more than 400 gigabytes of data was collected. The load, index, and aggregation processes for this data set really taxed the Oracle data warehouse. Even after significant tuning, we were unable to aggregate a day of clickstream data in less than 24 hours."



users

Frontend

Backend

“OLTP”

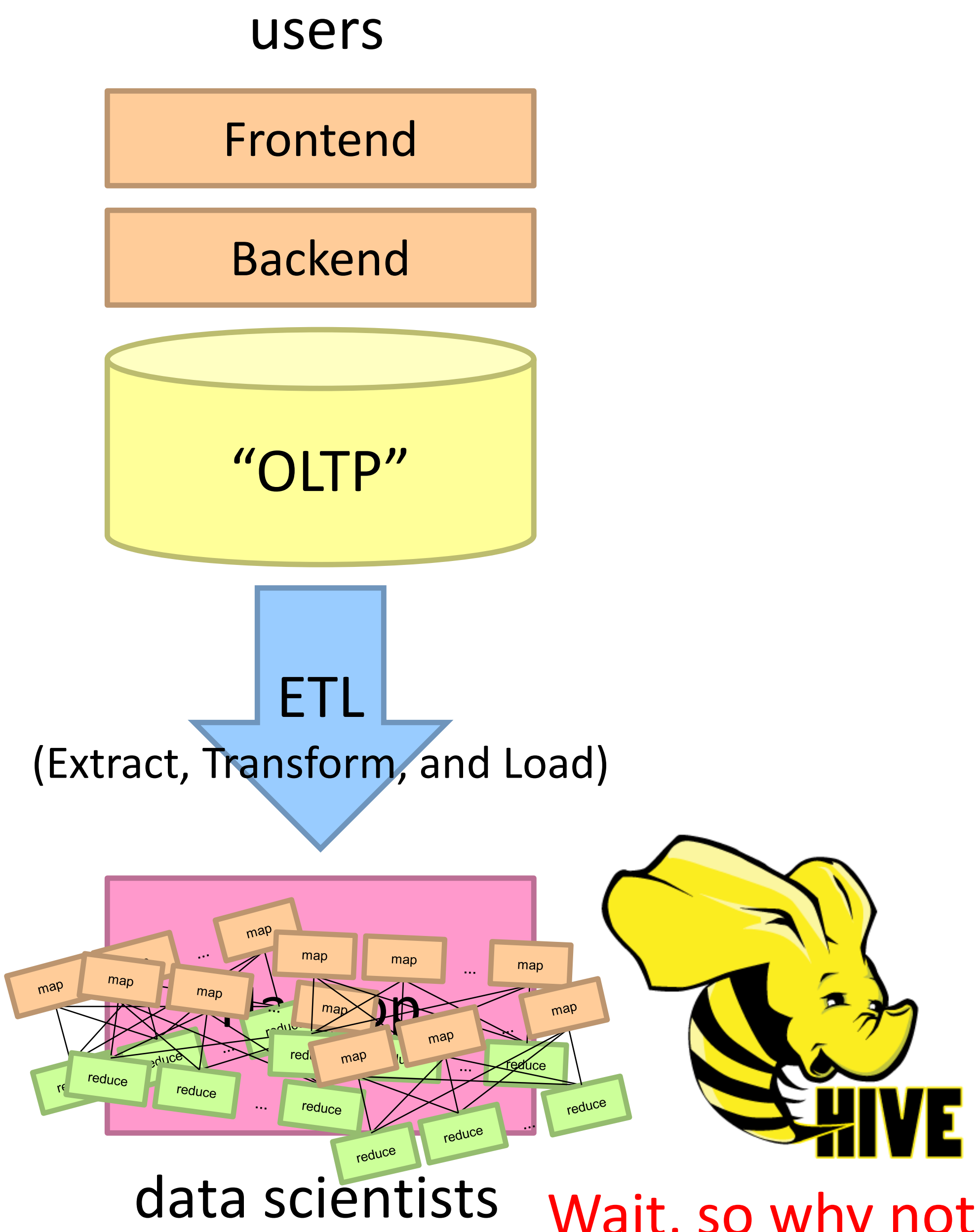
ETL

(Extract, Transform, and Load)

Hadoop

data scientists

The Irony…



Wait, so why not use a database to begin with?

Why not just use a database?

SQL is awesome

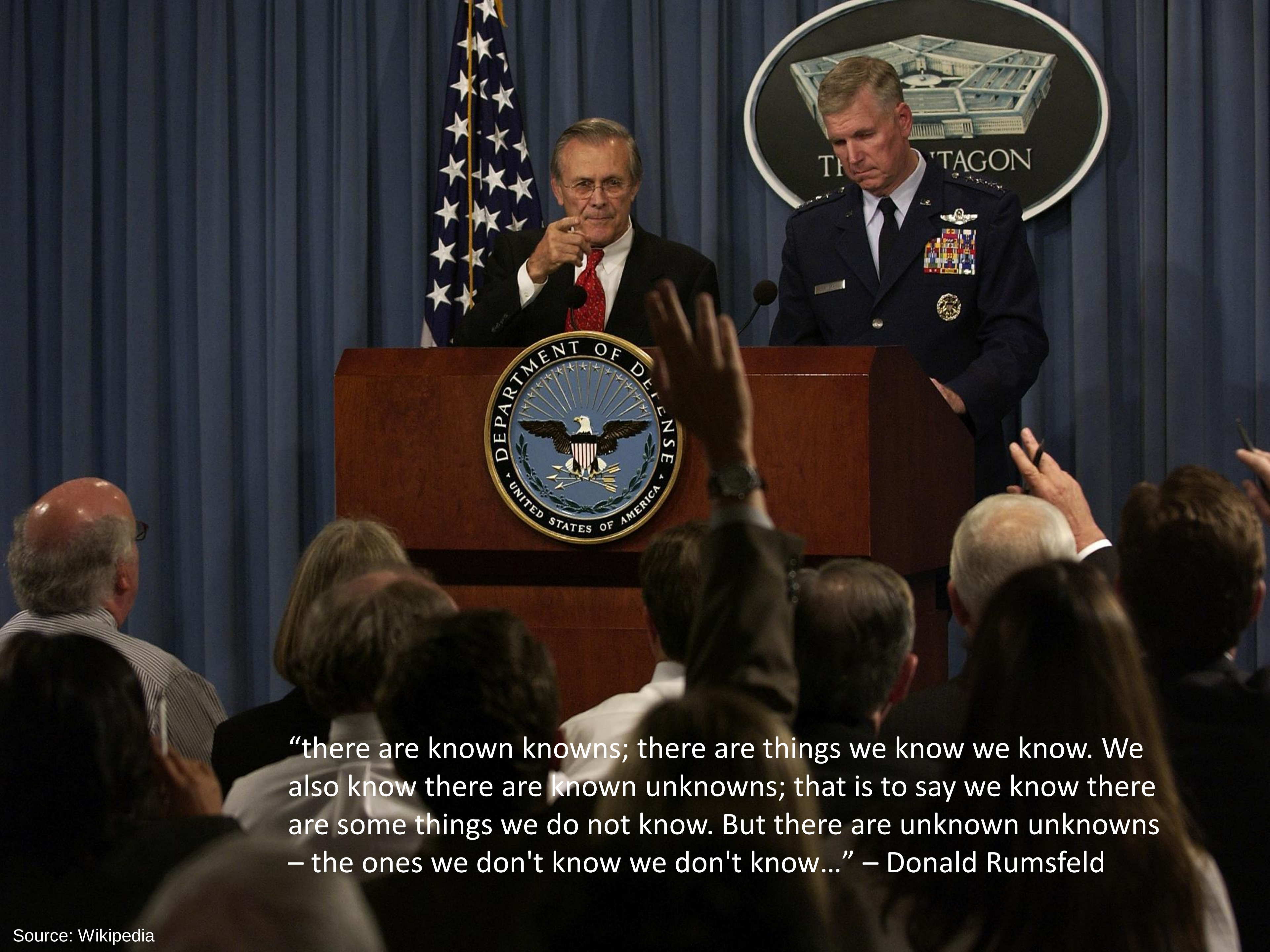
Scalability. Cost.

Databases are great…

If your data has structure (and you know what the structure is)
If your data is reasonably clean
If you know what queries you're going to run ahead of time

Databases are not so great…

If your data has little structure (or you don't know the structure)
If your data is messy and noisy
If you don't know what you're looking for



“there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are unknown unknowns – the ones we don't know we don't know…” – Donald Rumsfeld

Source: Wikipedia

Databases are great…

If your data has structure (and you know what the structure is)

If your data is reasonably clean

If you know what queries you're going to run ahead of time

Known unknowns!

Databases are not so great…

If your data has little structure (or you don't know the structure)

If your data is messy and noisy

If you don't know what you're looking for

Unknown unknowns!

Advantages of Hadoop dataflow languages

Don’t need to know the schema ahead of time

Raw scans are the most common operations

Many analyses are better formulated imperatively

Much faster data ingest rate

What do you actually do?

Report generation

Dashboards

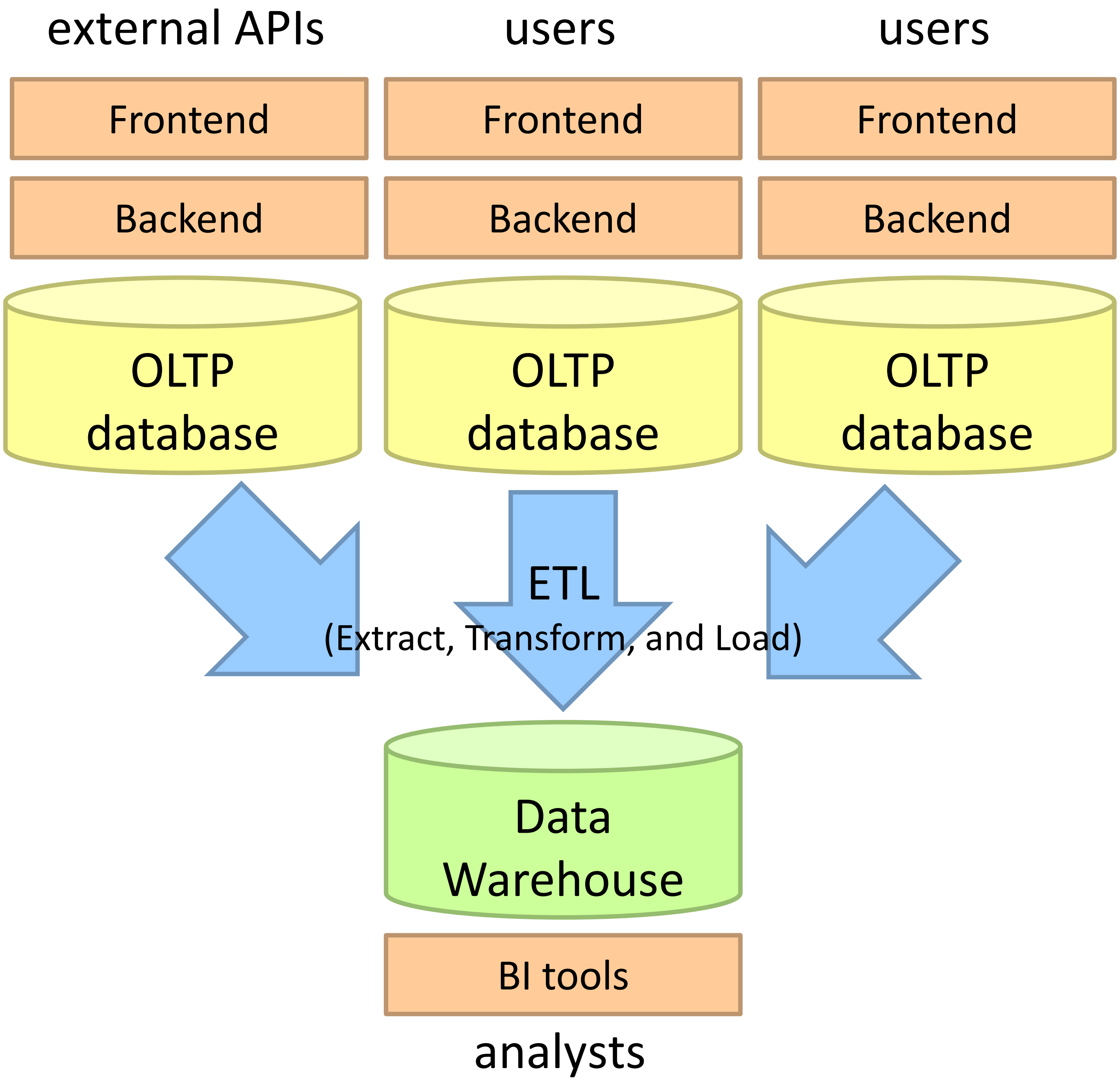
Ad hoc analyses

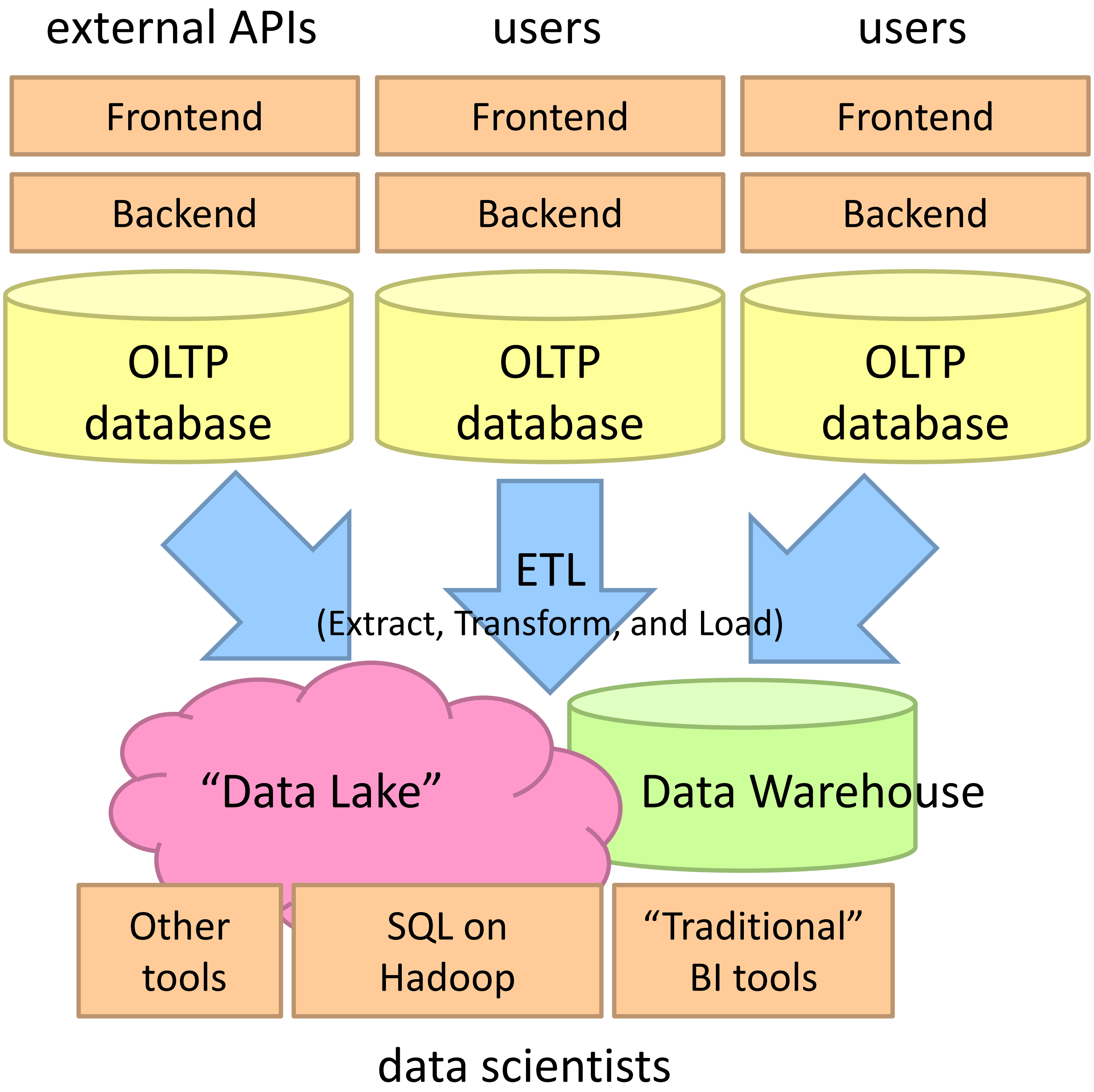
“Descriptive”

“Predictive”

Data products

Which are known unknowns and unknown unknowns?


external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
Data Warehouse
BI tools
analysts



external APIs

users

users

Frontend

Frontend

Frontend

Backend

Backend

Backend

OLTP database

OLTP database

OLTP database

ETL
(Extract, Transform, and Load)

"Data Lake"

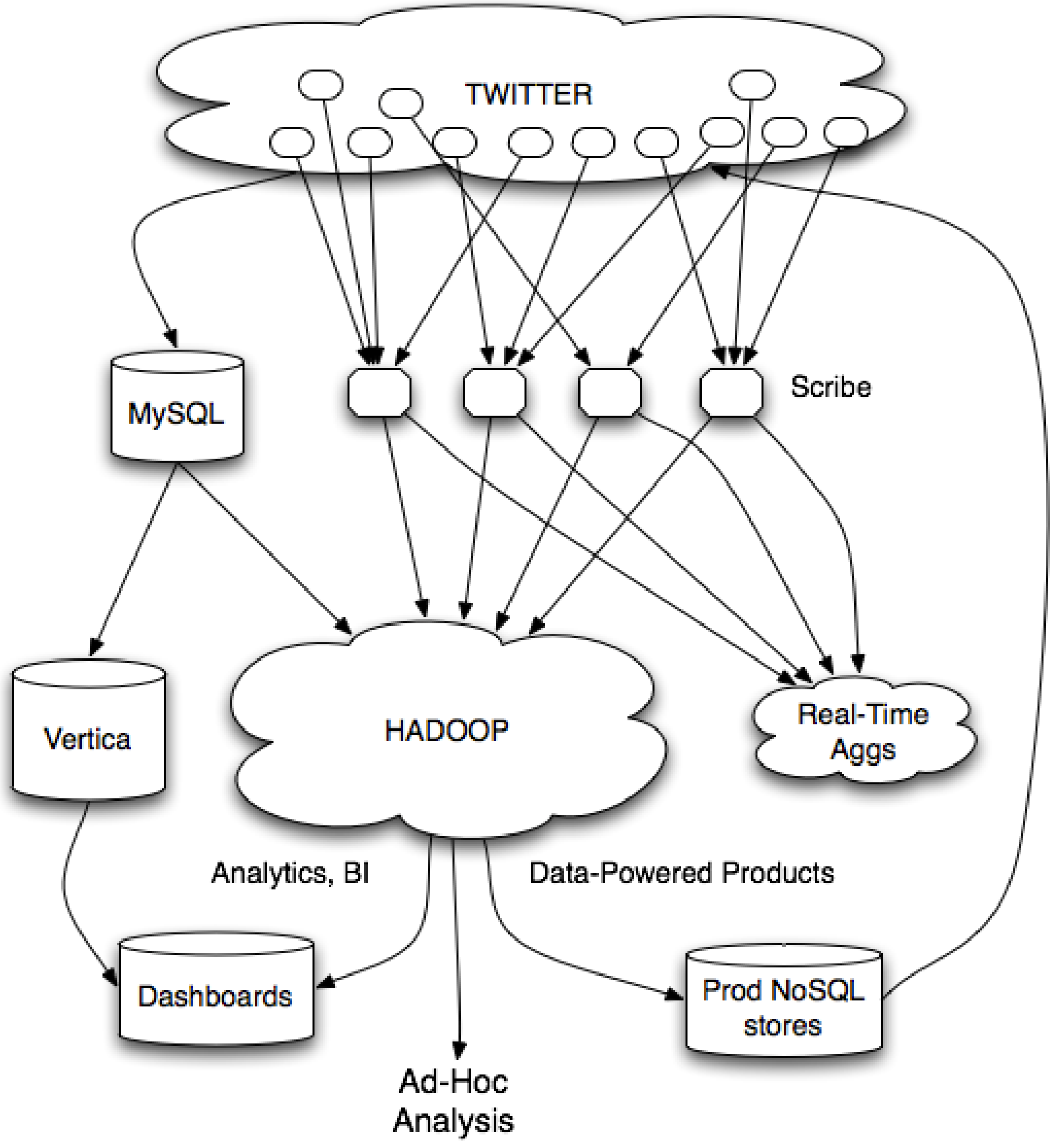
Data Warehouse

Other tools

SQL on Hadoop

"Traditional" BI tools

data scientists



Twitter’s data warehousing architecture (circa 2012)

circa ~2010

~150 people total

~60 Hadoop nodes

~6 people use analytics stack daily

circa ~2012

~1400 people total

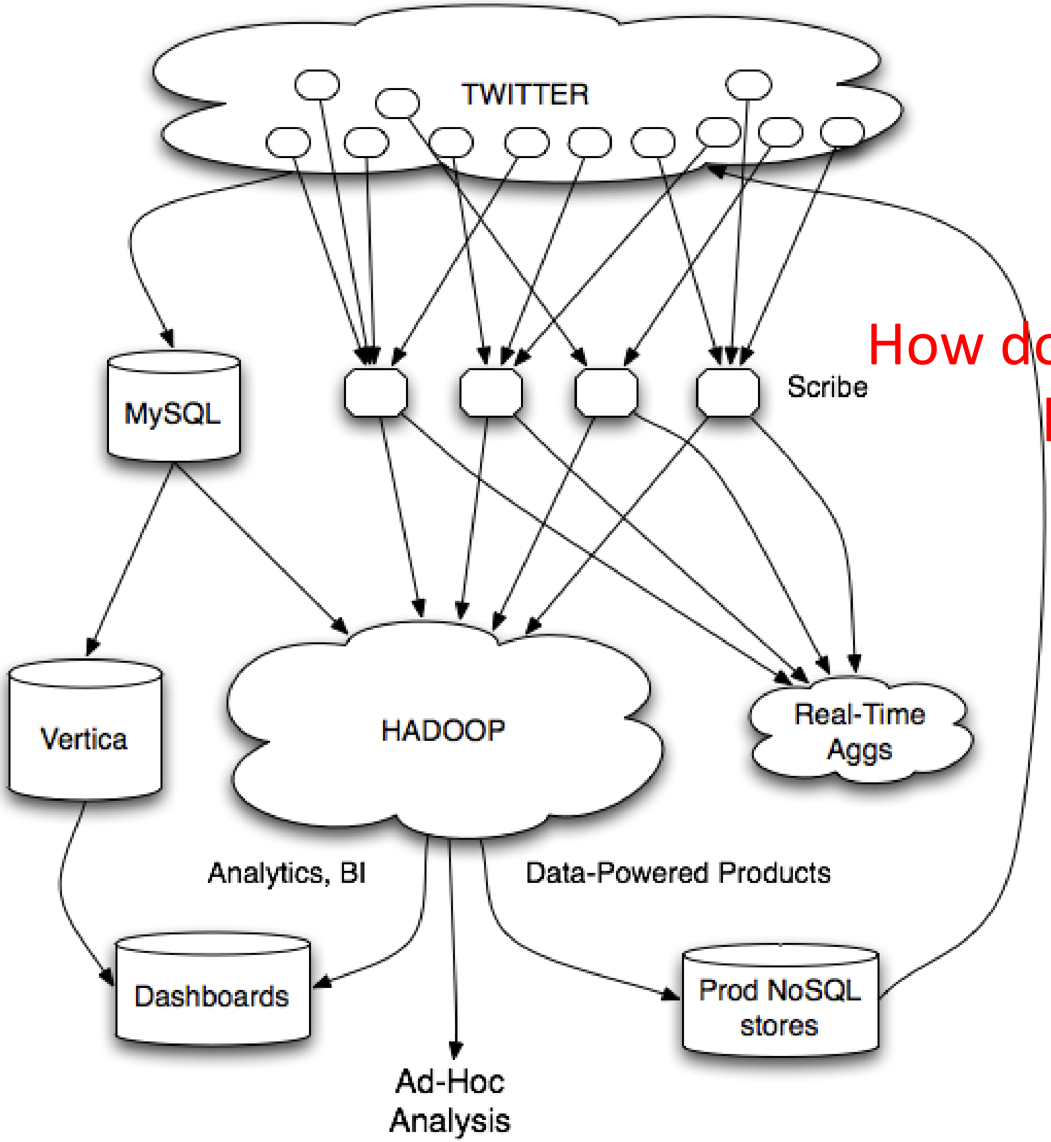
10s of Ks of Hadoop nodes, multiple DCs

10s of PBs total Hadoop DW capacity

~100 TB ingest daily

dozens of teams use Hadoop daily

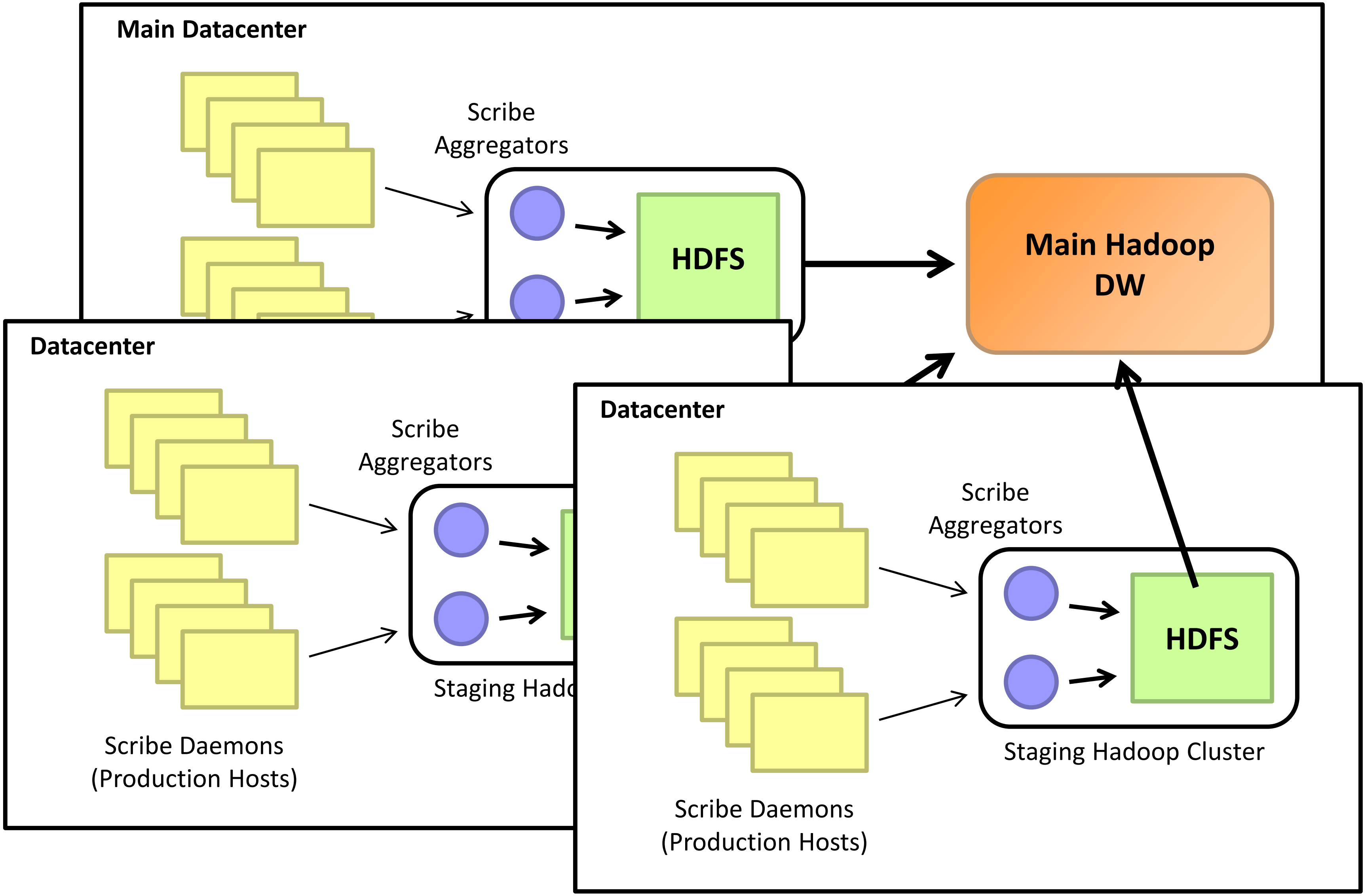
10s of Ks of Hadoop jobs daily

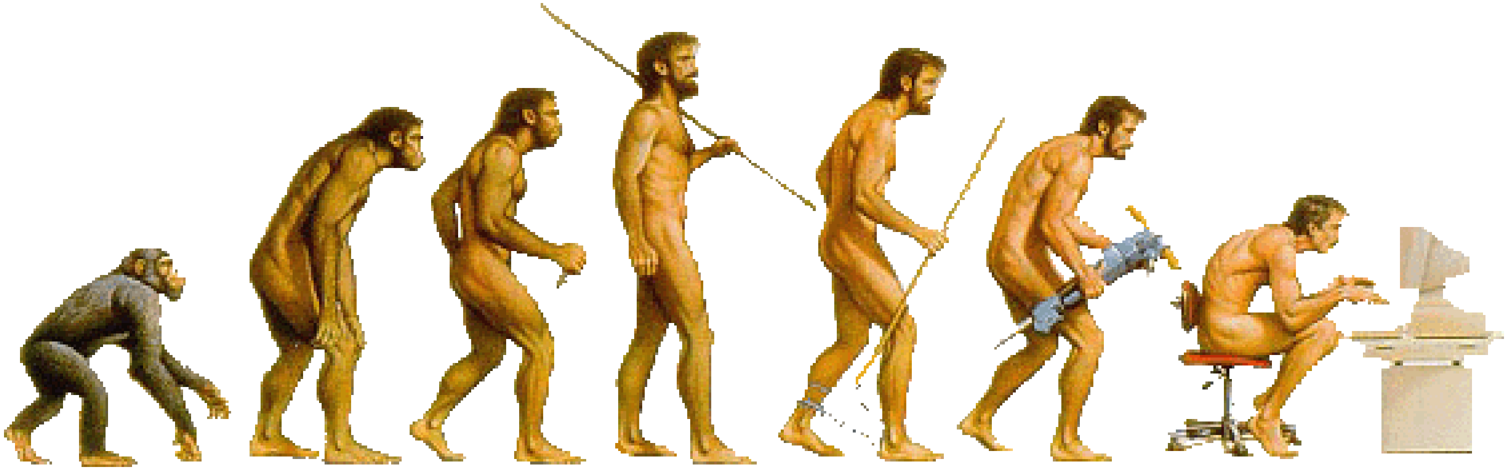


How does ETL actually happen?

Twitter’s data warehousing architecture (circa 2012)

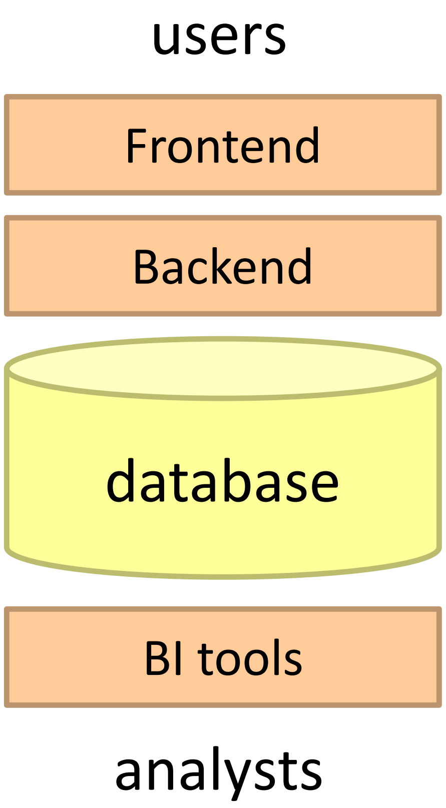
Importing Log Data





What’s Next?

Two developing trends…



users

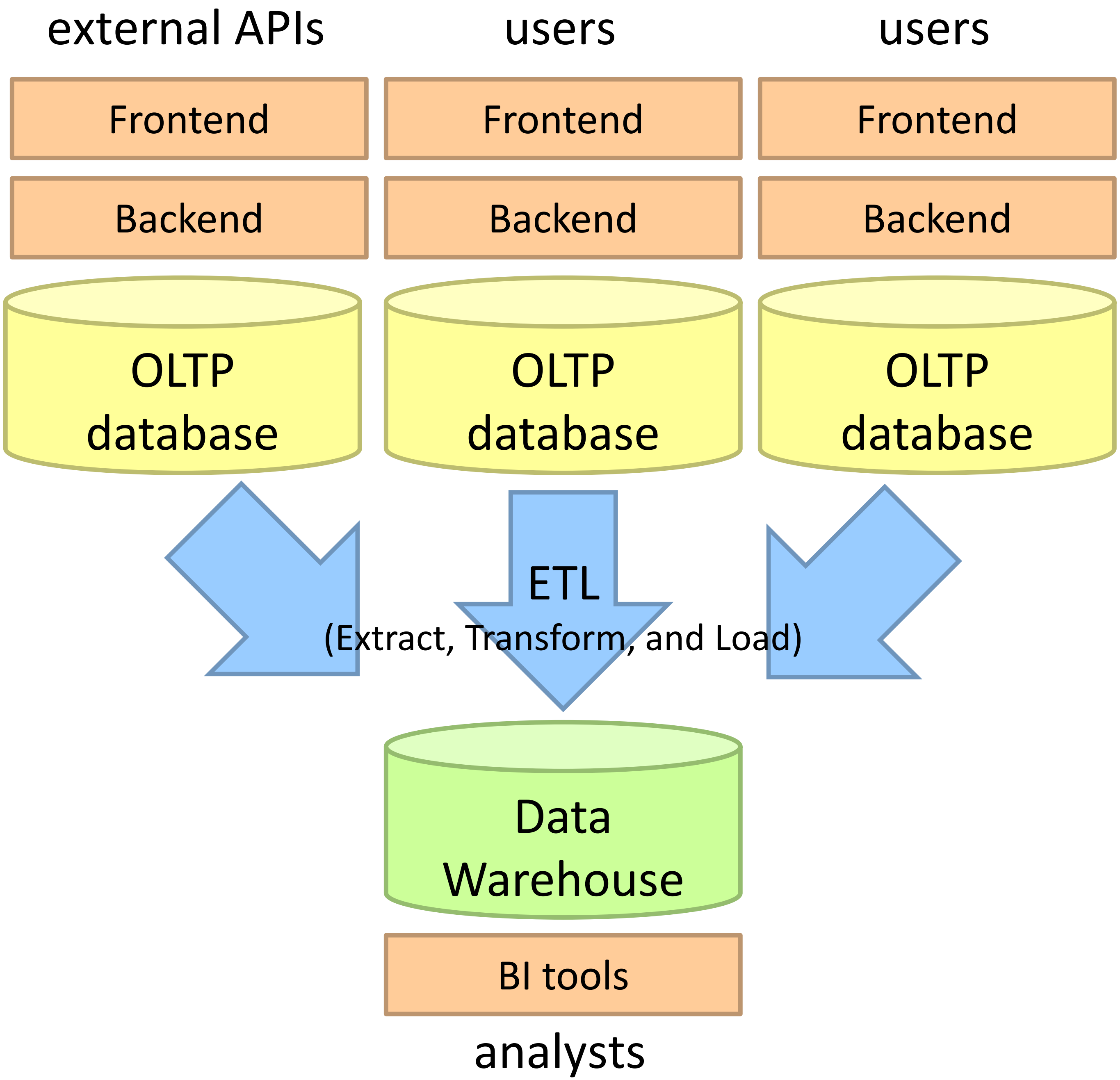
Frontend

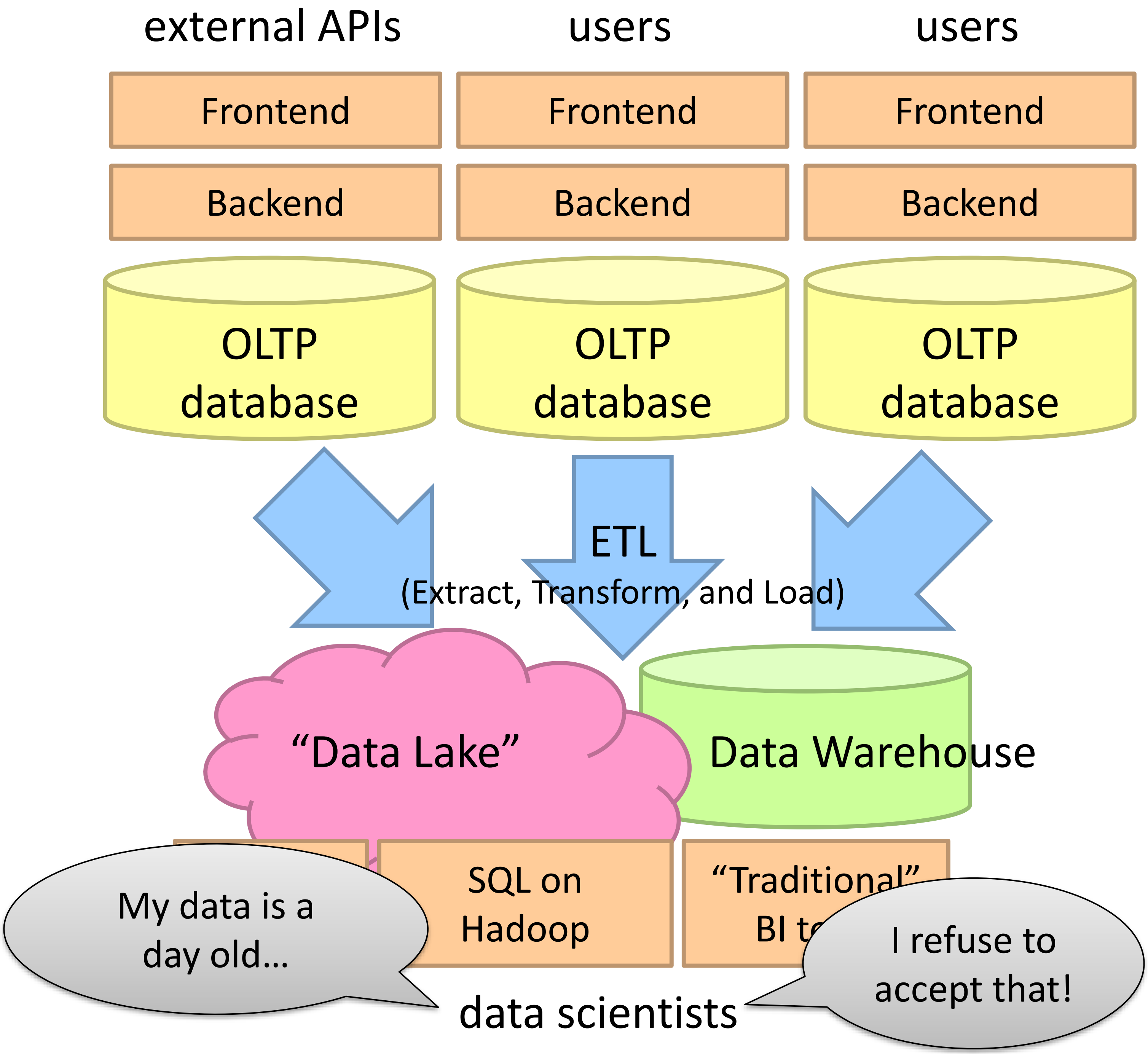
Backend

database

BI tools

analysts


external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
Data Warehouse
BI tools
analysts



external APIs

users

users

Frontend

Frontend

Frontend

Backend

Backend

Backend

OLTP database

OLTP database

OLTP database

ETL

(Extract, Transform, and Load)

“Data Lake”

Data Warehouse

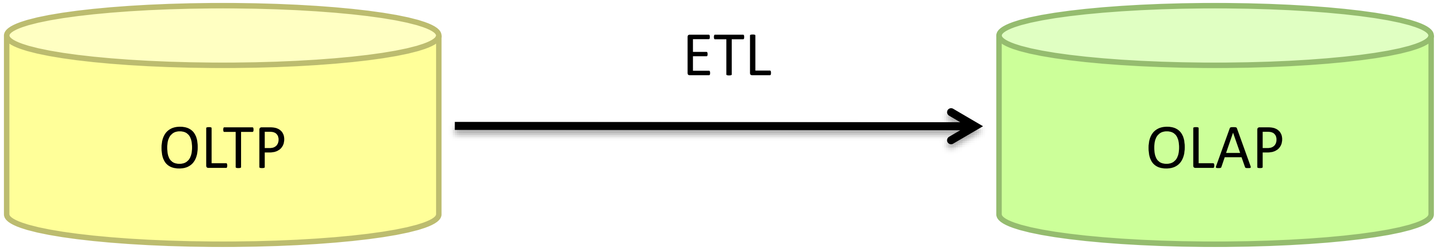
SQL on Hadoop

“Traditional” BI t

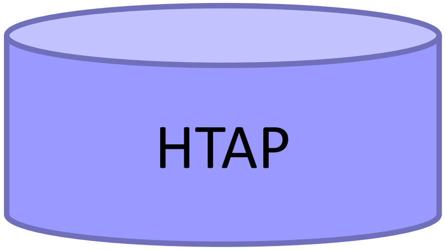
My data is a day old…

I refuse to accept that!

data scientists



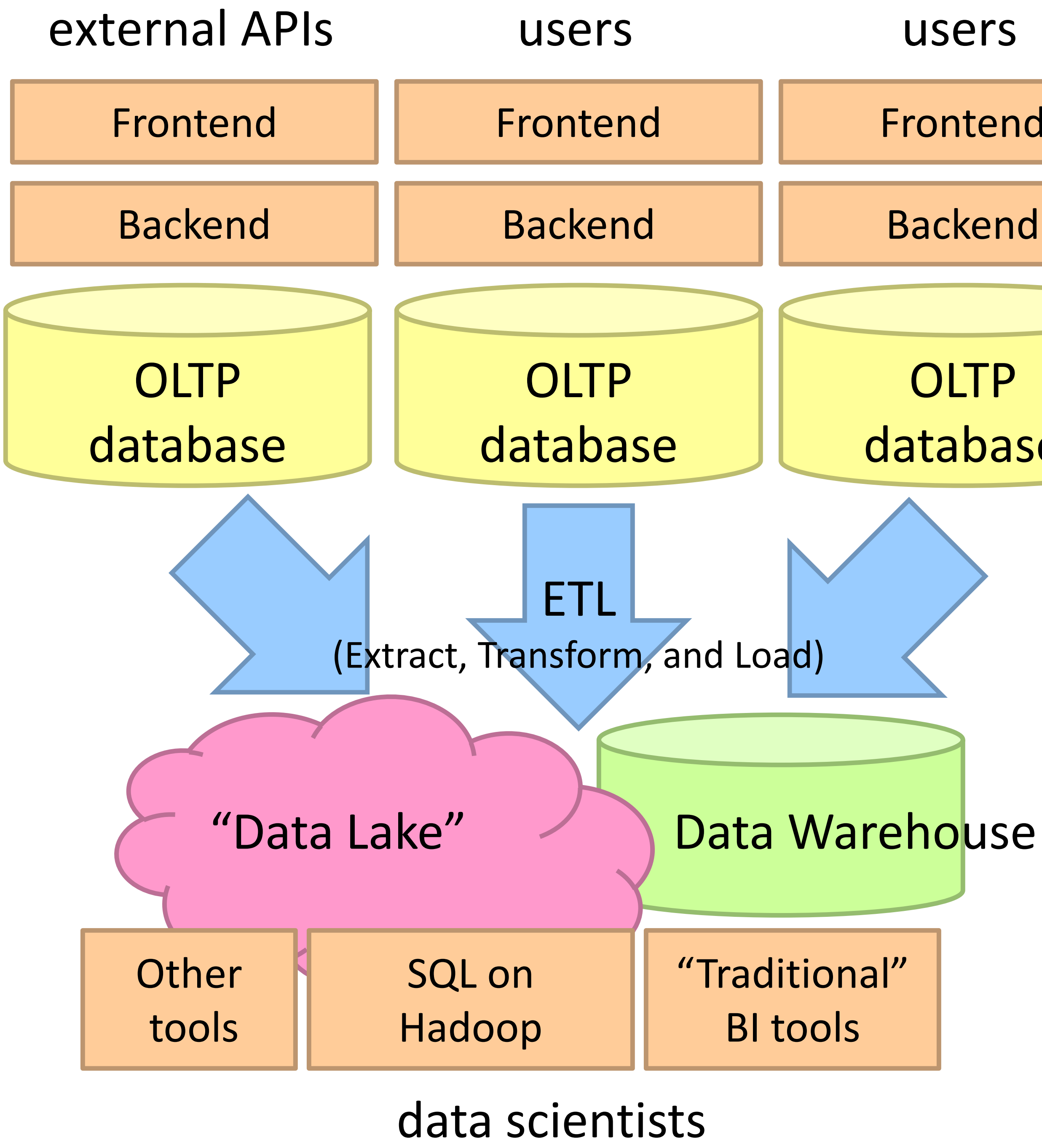
What if you didn’t have to do this?

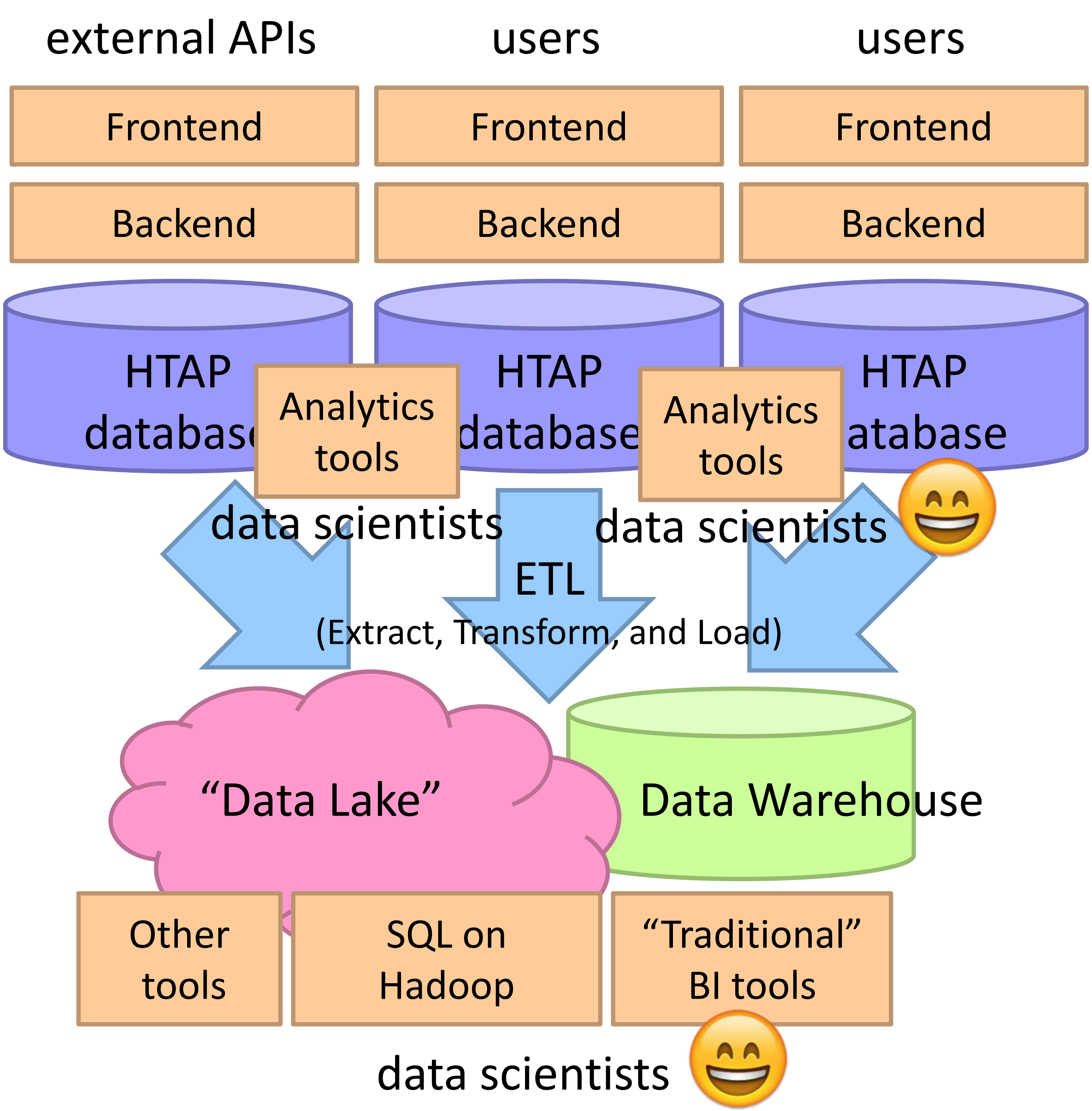


HTAP

Hybrid Transactional/Analytical Processing (HTAP)

Coming back full circle?


external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
“Data Lake”
Data Warehouse
Other tools
SQL on Hadoop
“Traditional” BI tools
data scientists



external APIs

users

users

Frontend

Frontend

Frontend

Backend

Backend

Backend

HTAP database

HTAP database

HTAP database

Analytics tools

Analytics tools

data scientists

data scientists 😄

ETL

(Extract, Transform, and Load)

“Data Lake”

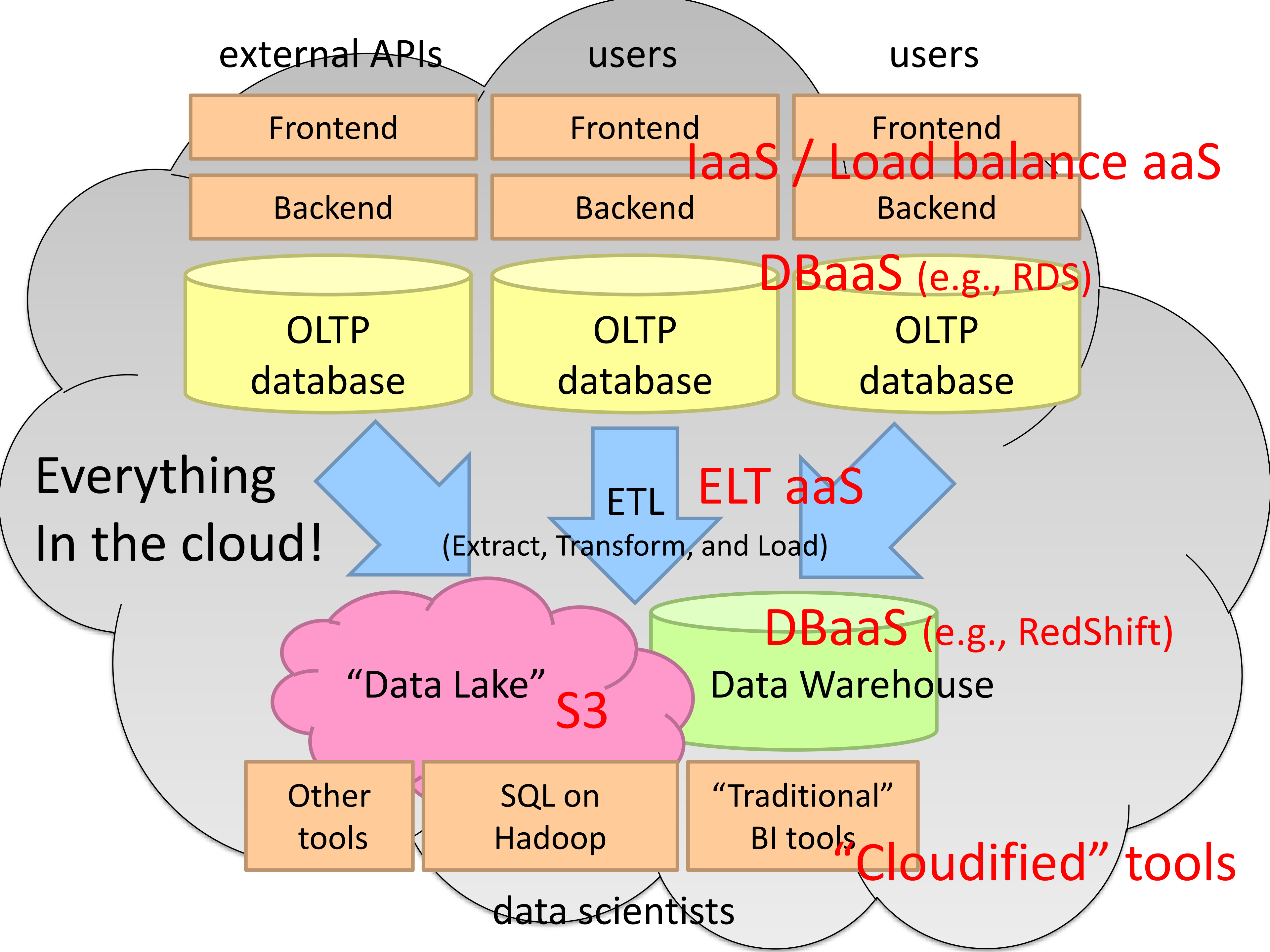
Data Warehouse

Other tools

SQL on Hadoop

“Traditional” BI tools

data scientists 😄


external APIs
users
users
Frontend
Frontend
Frontend
IaaS / Load balance aaS
Backend
Backend
Backend
DBaaS (e.g., RDS)
OLTP database
OLTP database
OLTP database
Everything In the cloud!
ETL
ELT aaS
(Extract, Transform, and Load)
"Data Lake"
S3
DBaaS (e.g., RedShift)
Data Warehouse
Other tools
SQL on Hadoop
"Traditional" BI tools
"Cloudified" tools
data scientists

Source: Wikipedia (Japanese rock garden)