UNIVERSITY OF WATERLOO

Data-Intensive Distributed Computing

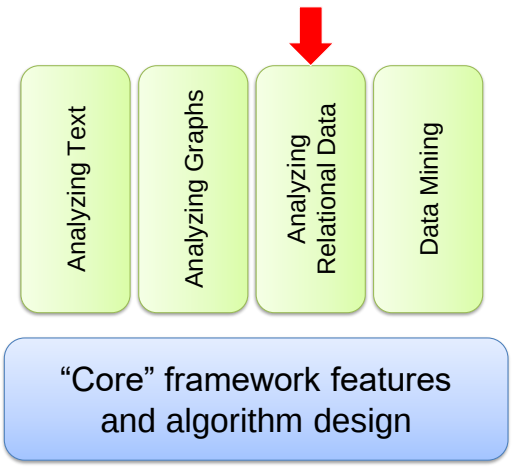
CS 431/631 451/651 (Fall 2021)

Part 7: Analyzing Relational Data (1/3)

Ali Abedi

These slides are available at https://www.student.cs.uwaterloo.ca/~cs451

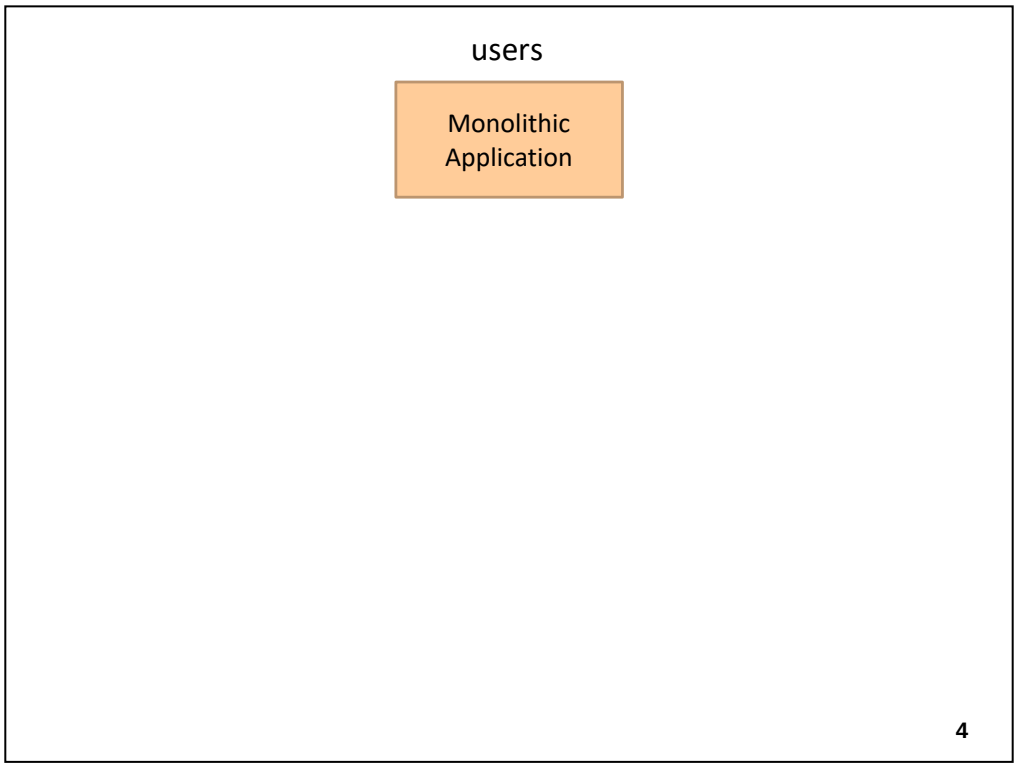
Structure of the Course

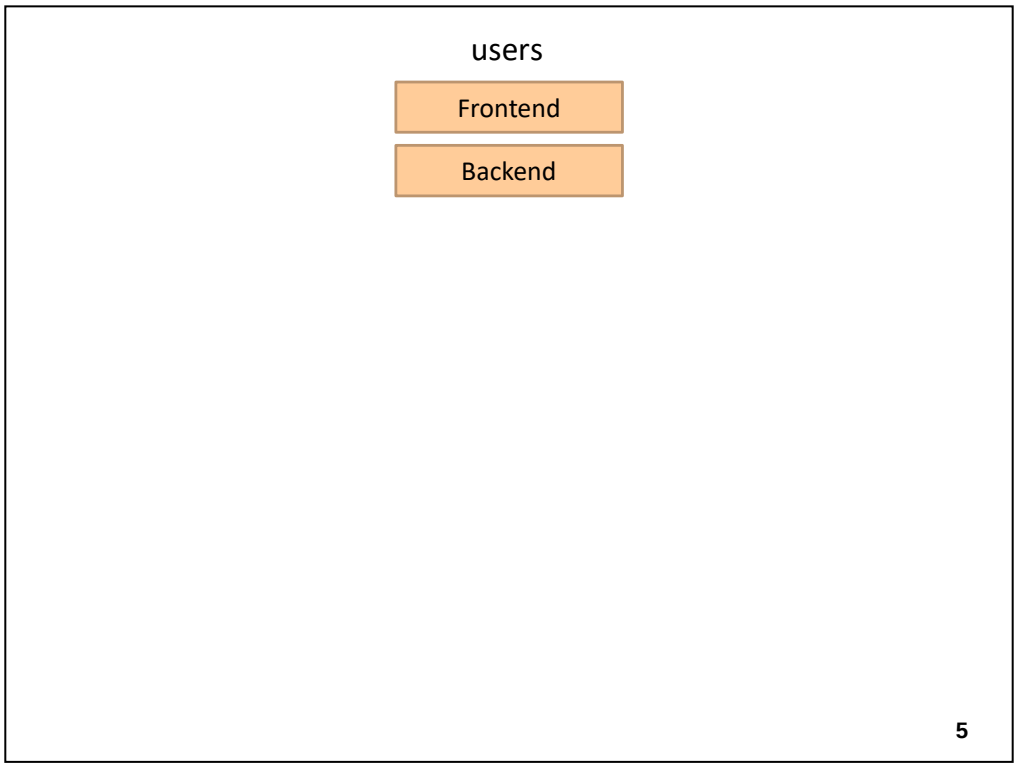


Evolution of Enterprise Architectures

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

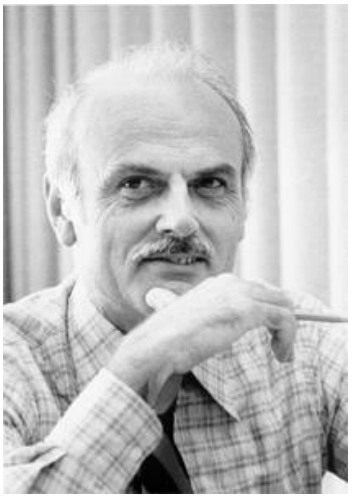
3


users
Monolithic
Application
4

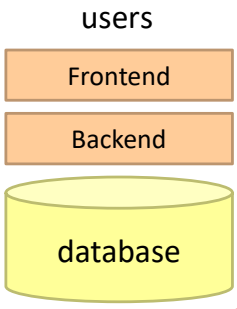

users
Frontend
Backend
5

Edgar F. Codd

- Inventor of the relational model for DBs
- SQL was created based on his work
- Turing award winner in 1981



6

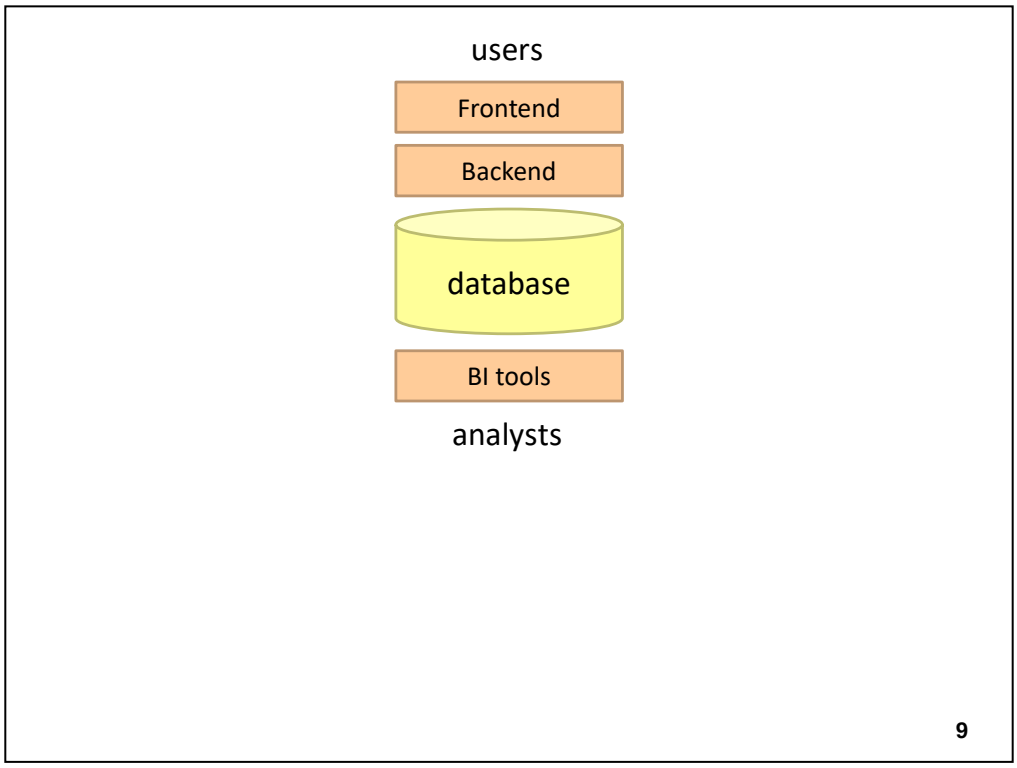


Why is this a good idea?

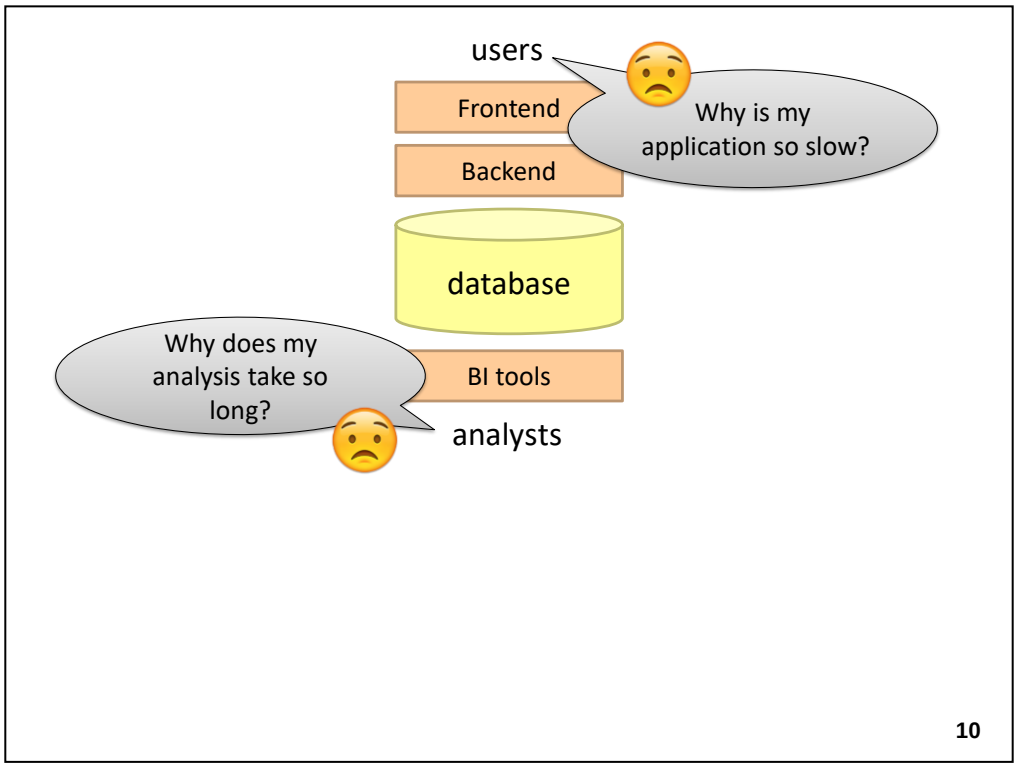
Business Intelligence

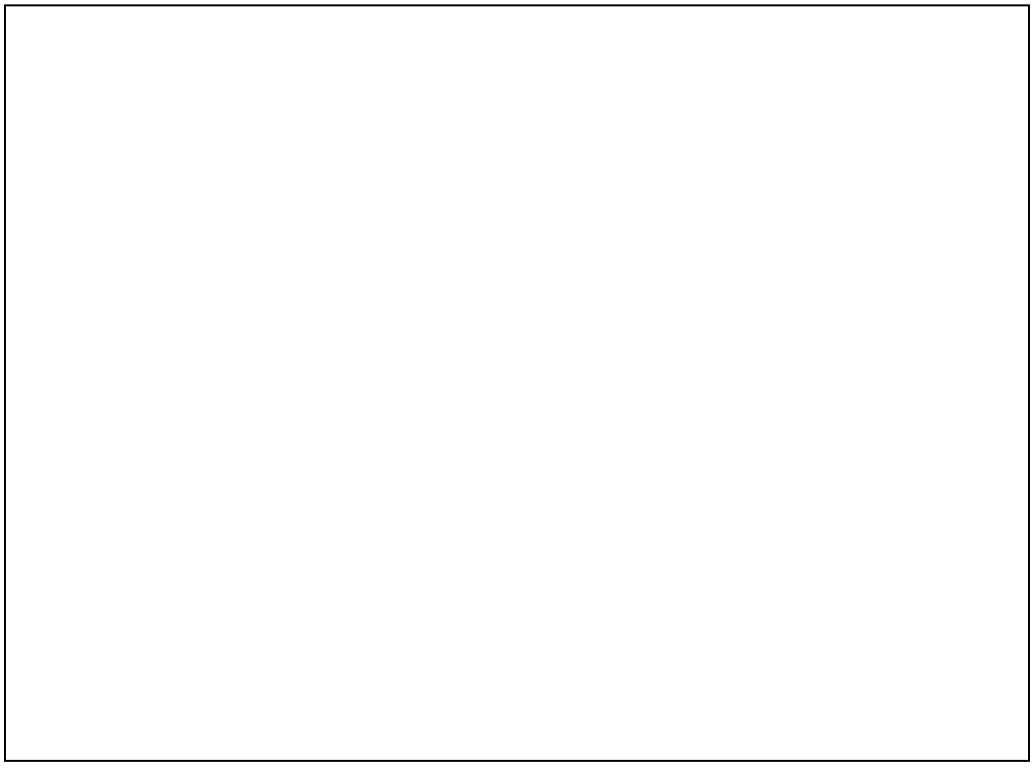
An organization should retain data that result from carrying out its mission and exploit those data to generate insights that benefit the organization, for example, market analysis, strategic planning, decision making, etc.

Duh!?



BI: Business intelligence


users
Frontend
Backend
database
BI tools
analysts
Why is my application so slow?
Why does my analysis take so long?
10



OLTP and OLAP Together?

Downsides of co-existing OLTP and OLAP workloads

Poor memory management

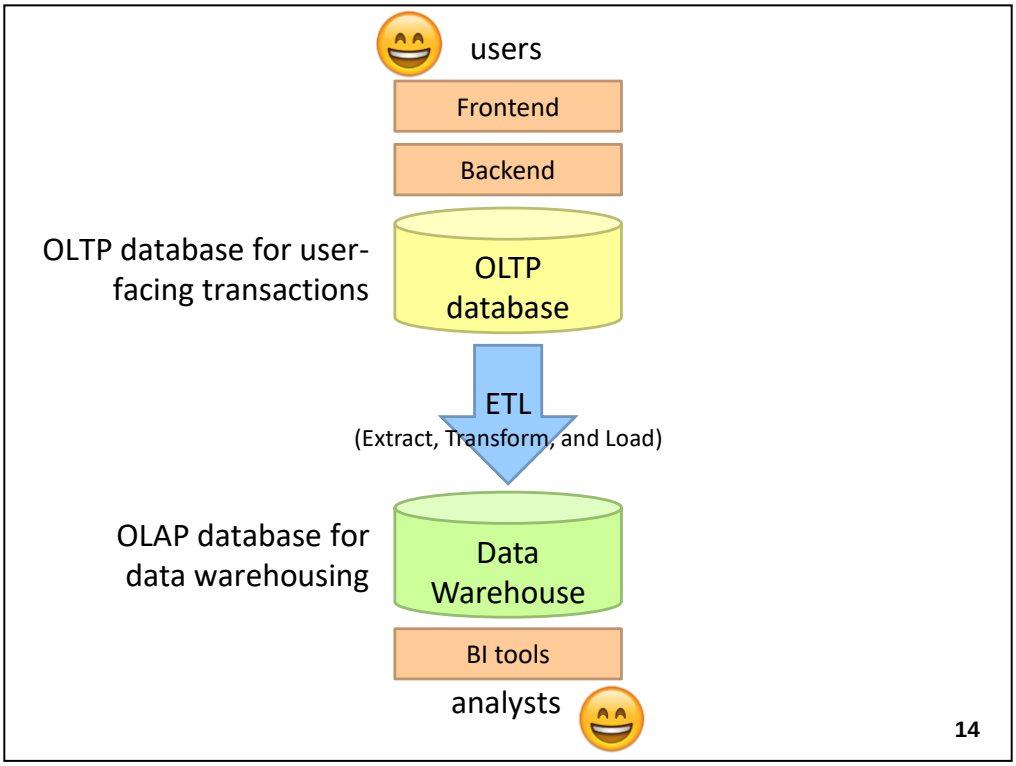
Conflicting data access patterns

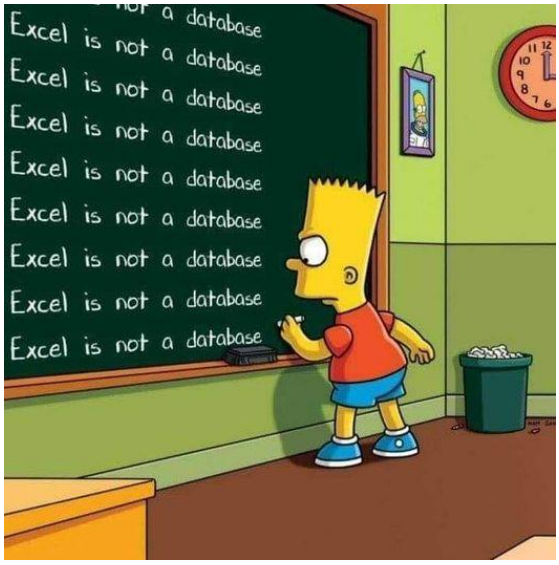
Variable latency

 users and analysts

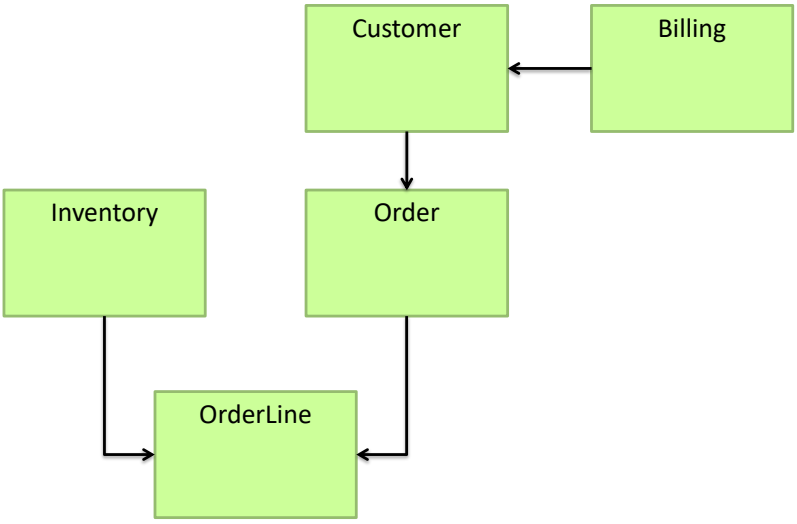
Solution?


M
L
K
E
Build a data warehouse!
13
Source: Wikipedia (Warehouse)


users
Frontend
Backend
OLTP database for user-facing transactions
OLTP database
ETL
(Extract, Transform, and Load)
OLAP database for data warehousing
Data Warehouse
BI tools
analysts
14

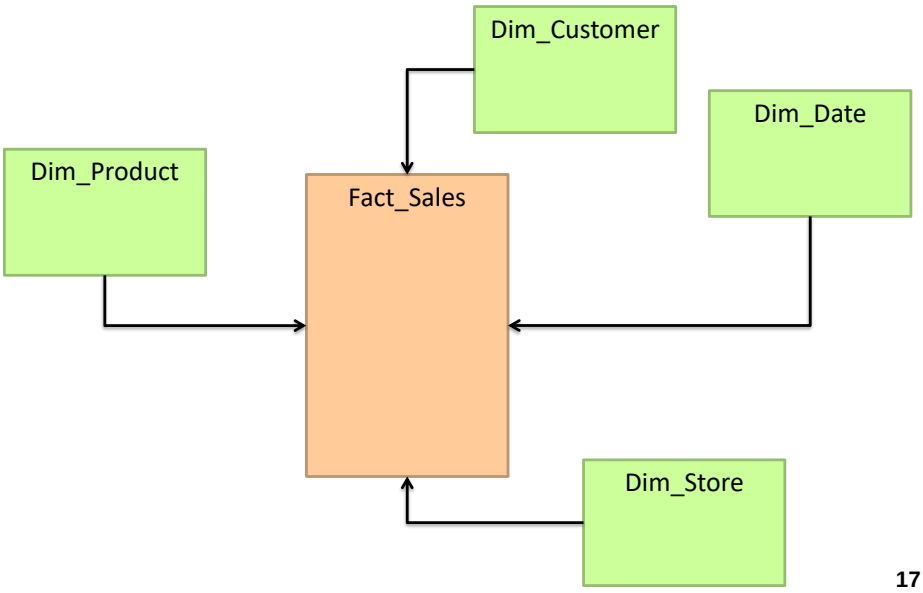

Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database
Excel is not a database

A Simple OLTP Schema



16

A Simple OLAP Schema



17

ETL

Extract

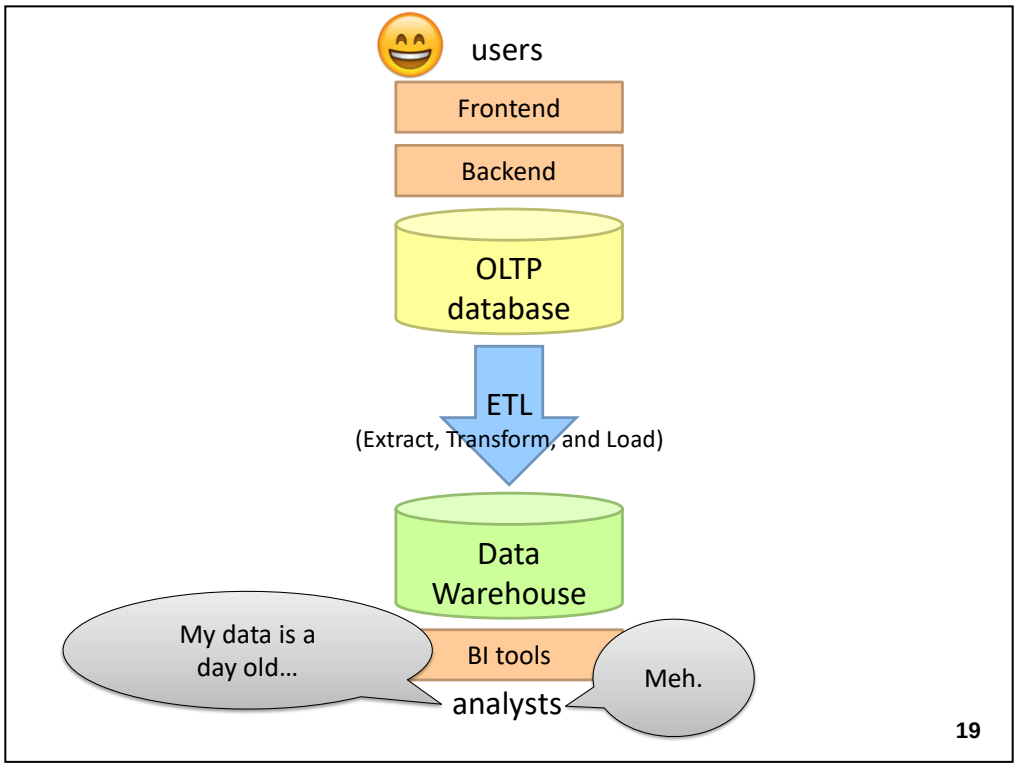
Transform

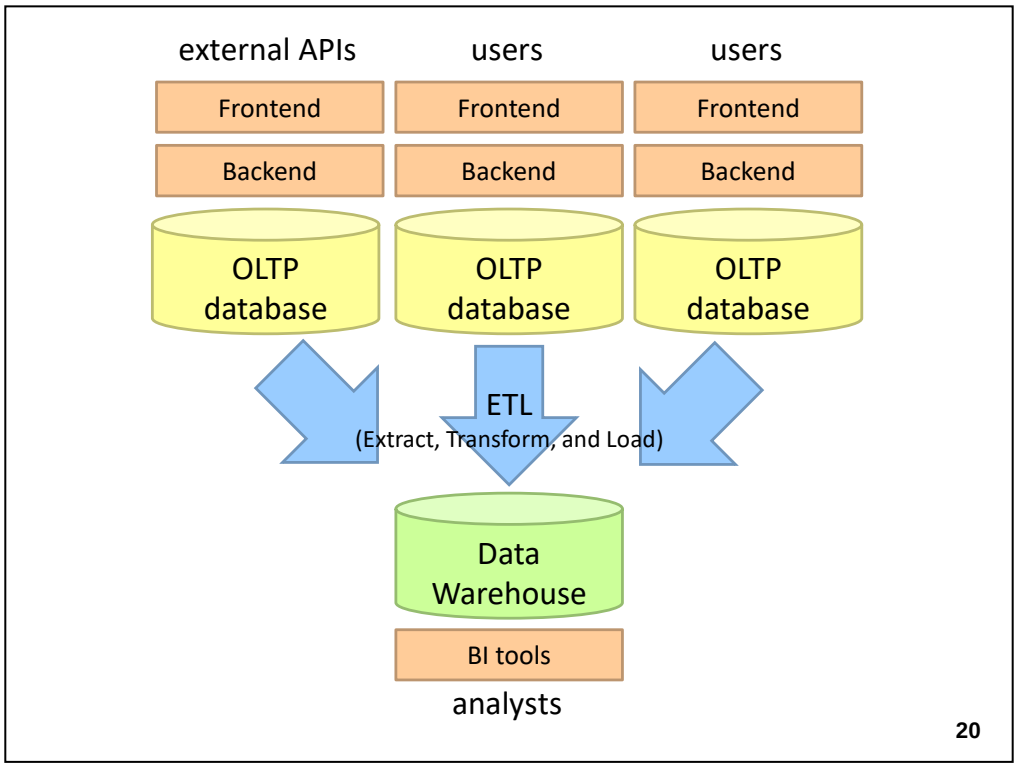
Data cleaning and integrity checking
Schema conversion
Field transformations

Load

When does ETL happen?

18


users
Frontend
Backend
OLTP database
ETL
(Extract, Transform, and Load)
Data Warehouse
My data is a day old…
BI tools
Meh.
analysts
19


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
20

What do you actually do?

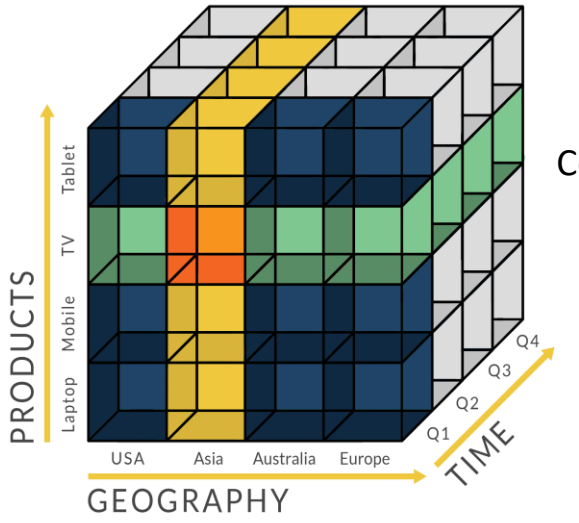
Report generation

Dashboards

Ad hoc analyses

21

OLAP Cubes

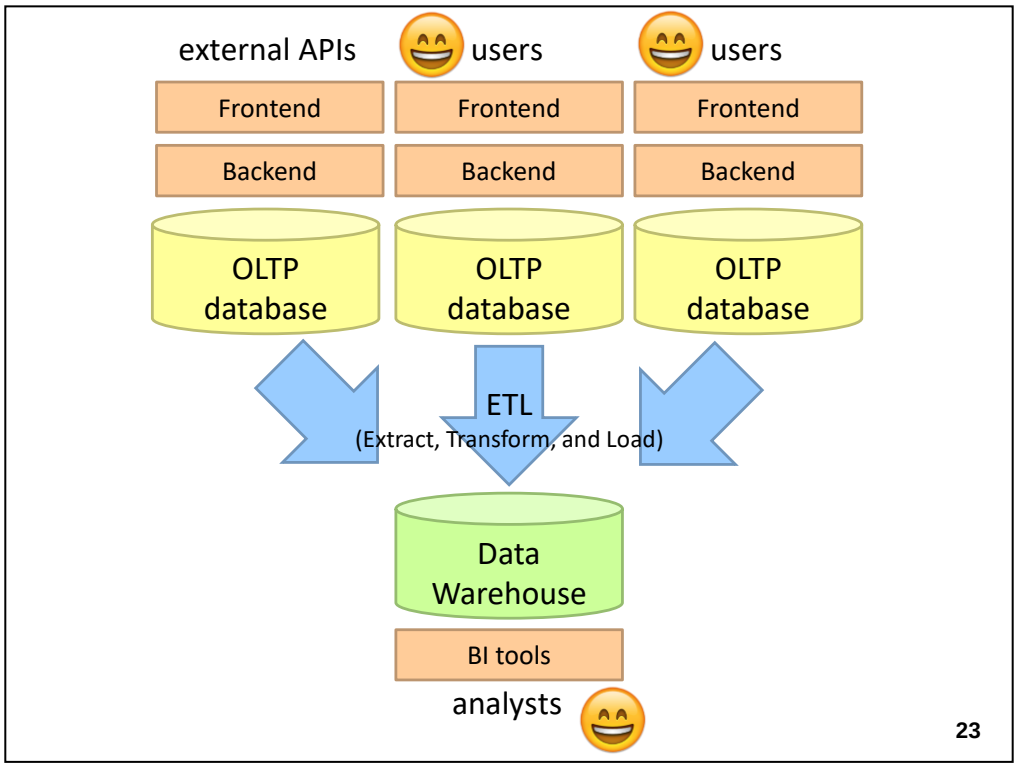


Common operations

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

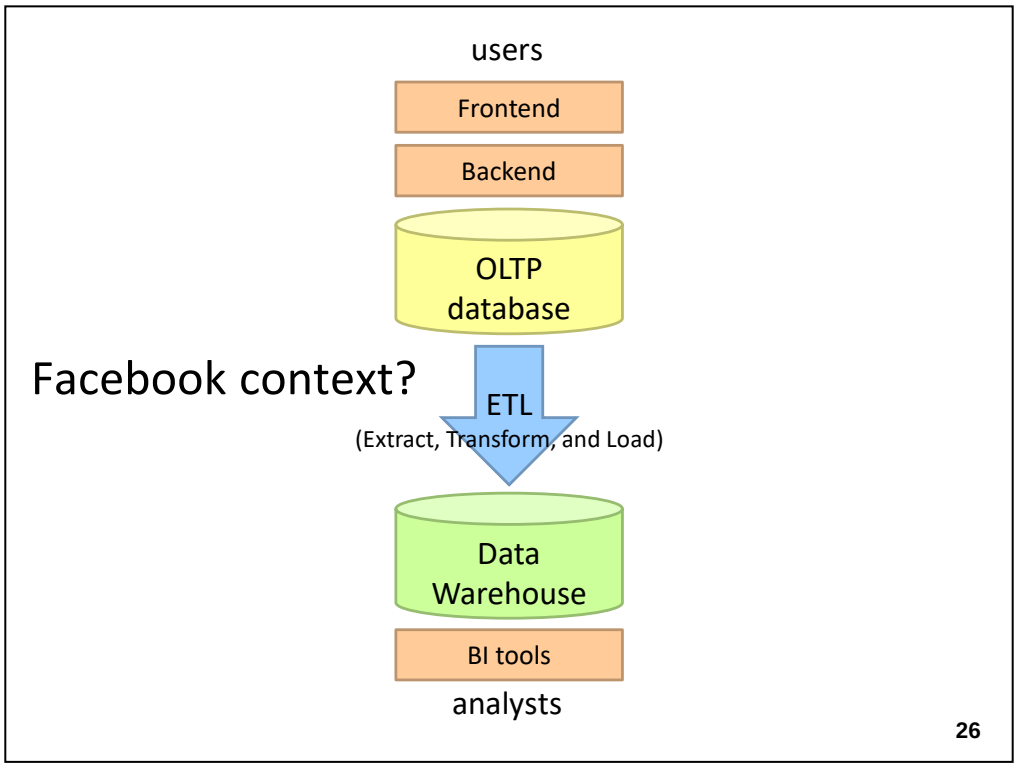
22

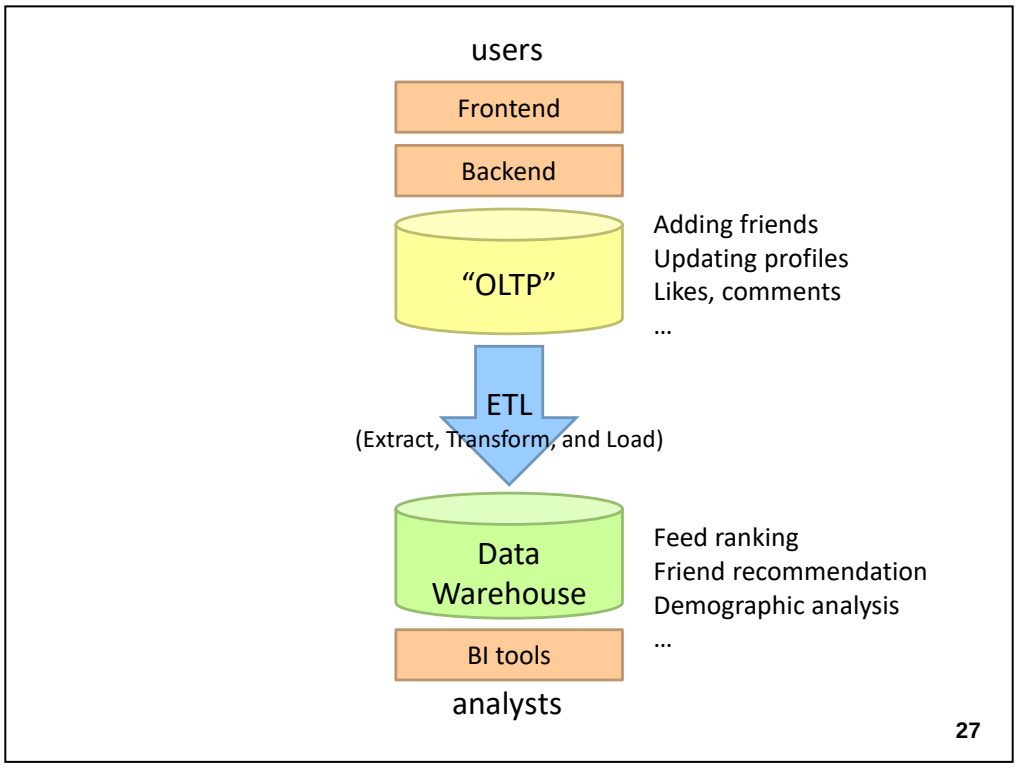
https://youtu.be/LRdsZqrwOrc


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
23

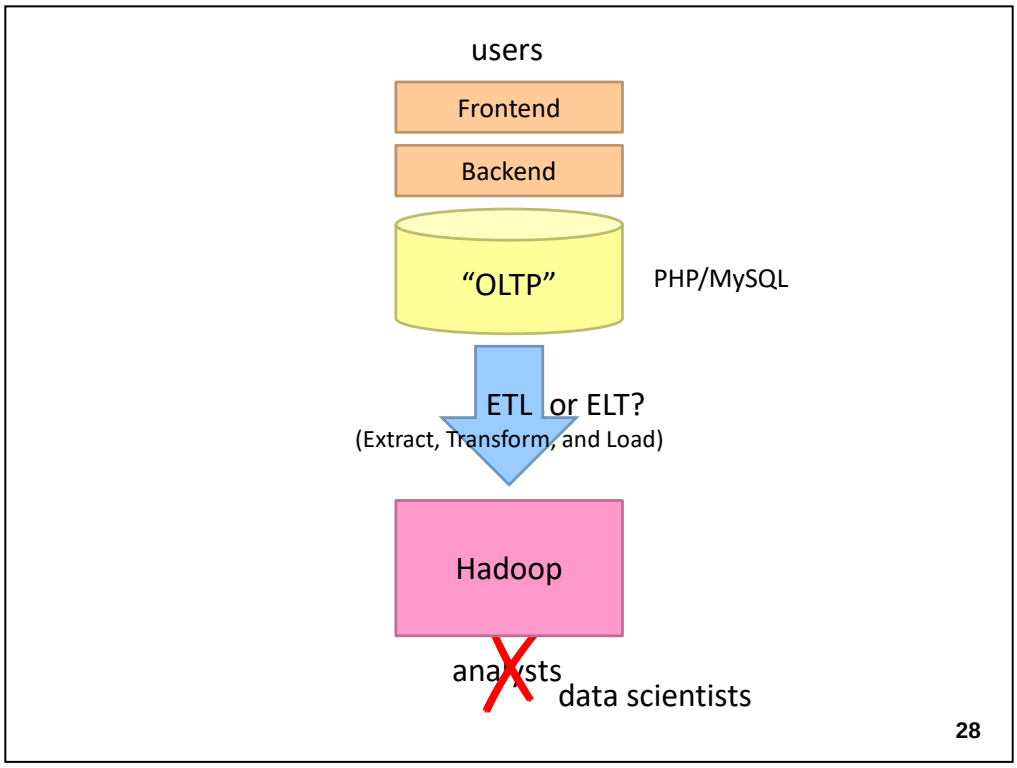
Fast forward…


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."
25


users
Frontend
Backend
OLTP database
Facebook context?
ETL
(Extract, Transform, and Load)
Data Warehouse
BI tools
analysts
26



But we have tools to deal with this, right?


users
Frontend
Backend
"OLTP"
PHP/MySQL
ETL or ELT?
(Extract, Transform, and Load)
Hadoop
analysts
data scientists
28

What's changed?

Dropping cost of disks

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

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

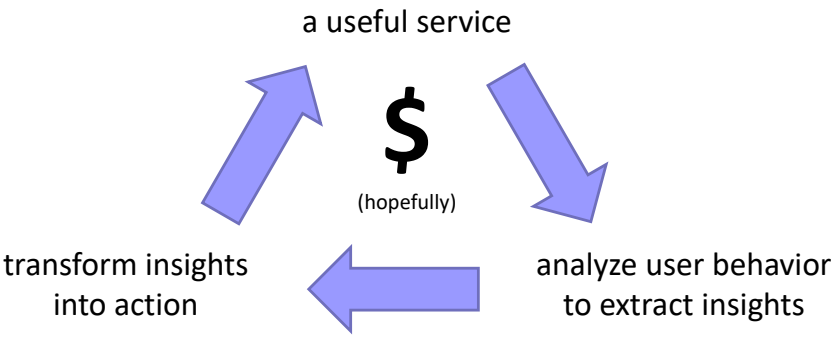
Large increase in data volume

Growing maturity of data mining techniques

Demonstrates value of data analytics

30

Virtuous Product Cycle



Google. Facebook. Twitter. Amazon. Uber.

31

What do you actually do?

Report generation

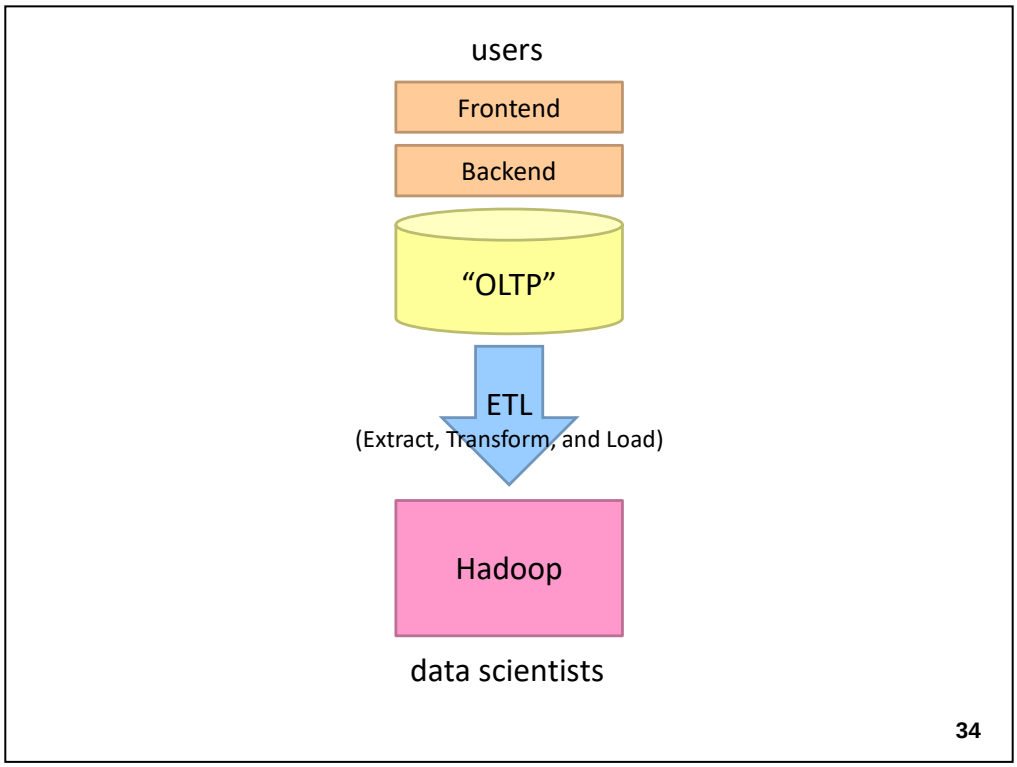
Dashboards

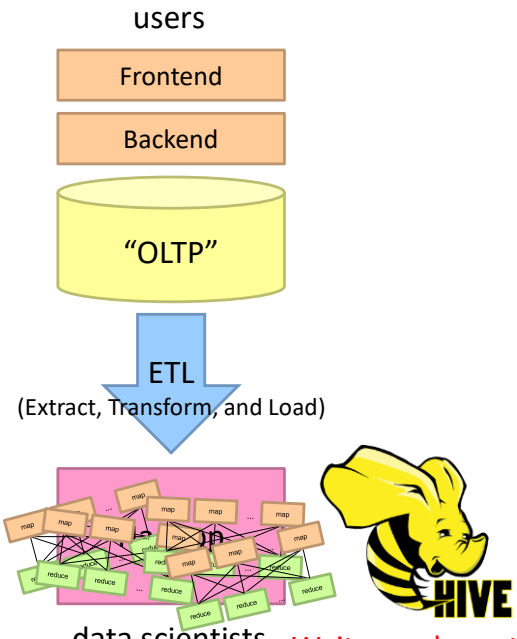
Ad hoc analyses
"Descriptive"
"Predictive"

Data products

32


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."
33


users
Frontend
Backend
“OLTP”
ETL
(Extract, Transform, and Load)
Hadoop
data scientists
34



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

35

Why not just use a database?

SQL is awesome

Scalability. Cost.

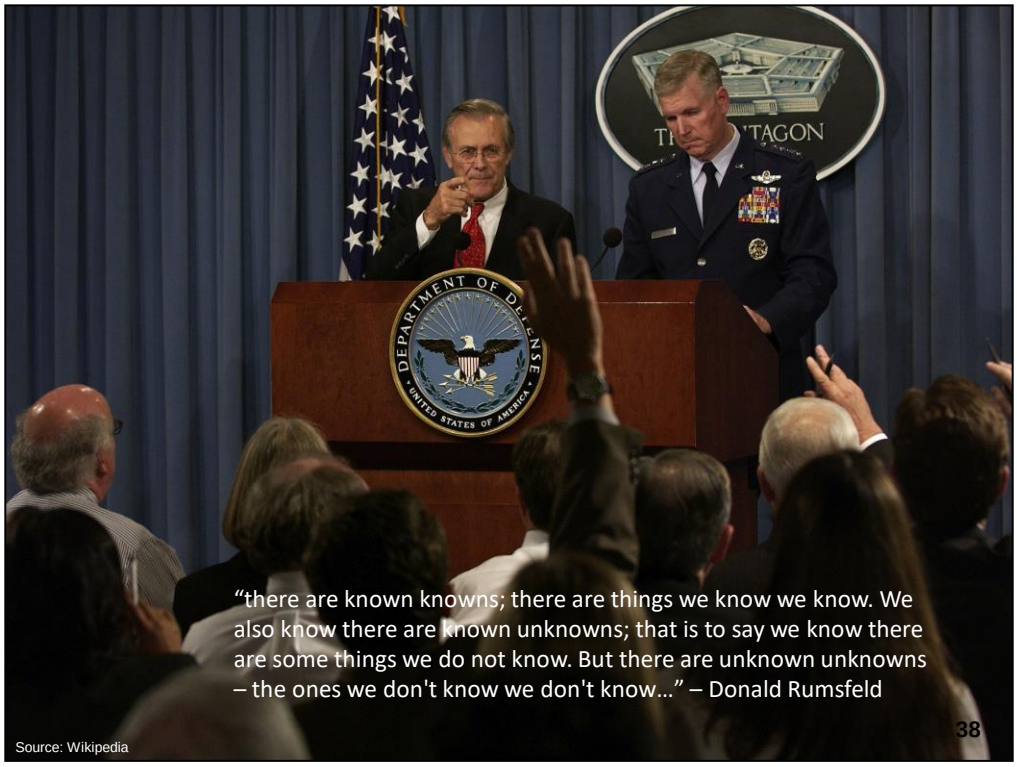
Databases are great…

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

37



"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

38

Source: Wikipedia

One who knows and knows that he knows
His horse of wisdom will reach the skies

One who doesn't know, but knows that he doesn't know
His limping mule will eventually get him home

One who doesn't know and doesn't know that he doesn't know
He will be eternally lost in his hopeless ignorance!

Ibn Yamin (1286-1368)

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!

40

What do you actually do?

Report generation

Dashboards

Ad hoc analyses
"Descriptive"
"Predictive"

Data products

Which are known unknowns and unknown unknowns?

41

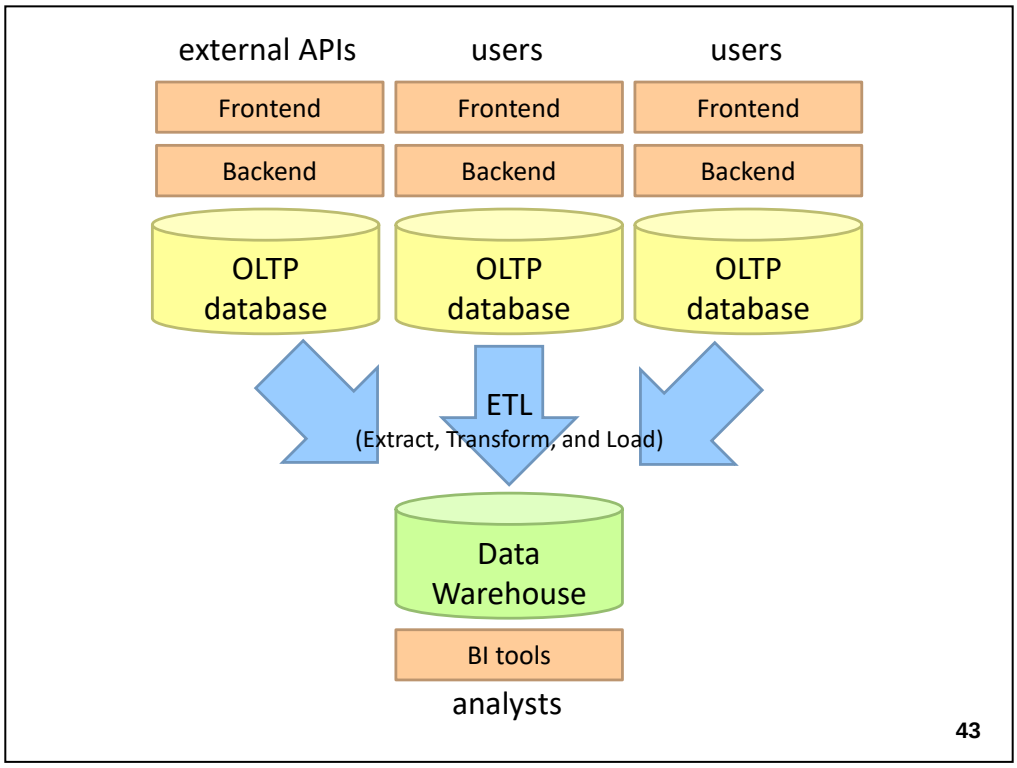
Advantages of Hadoop dataflow languages

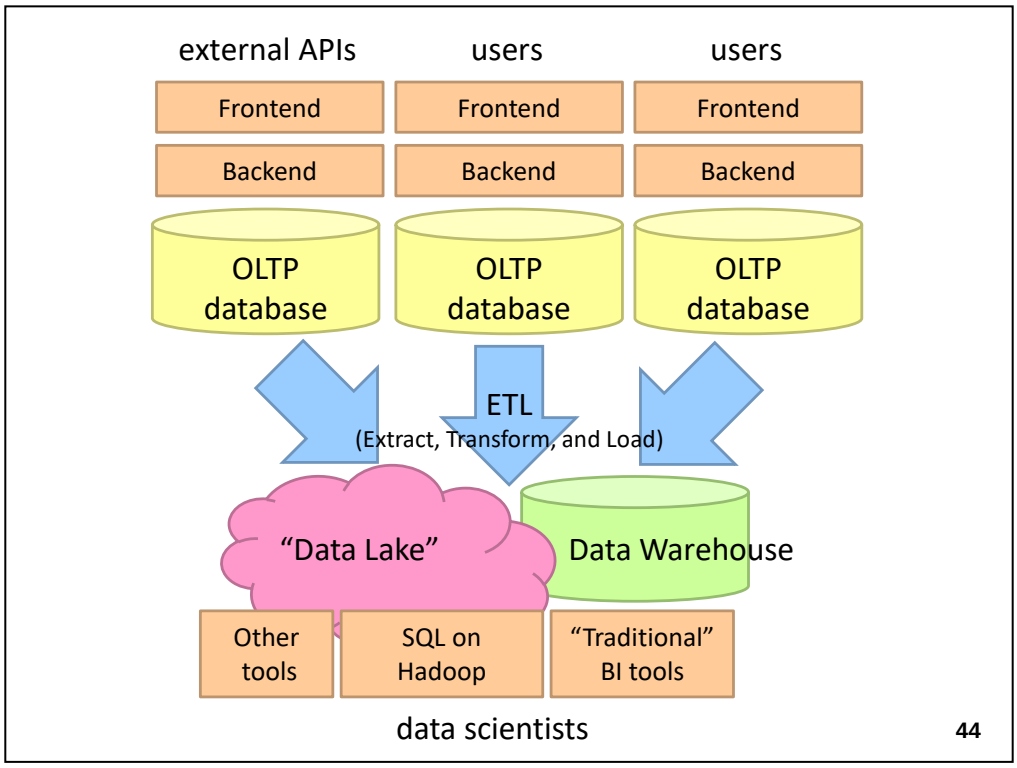
Don’t need to know the schema ahead of time

Raw scans are the most common operations

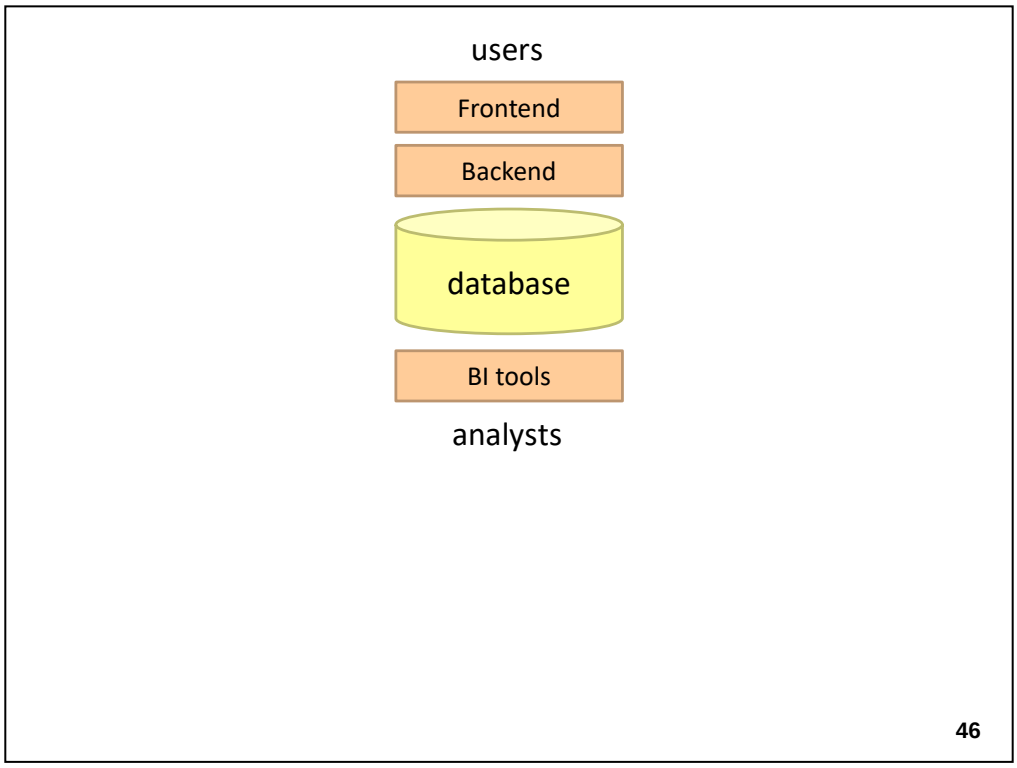
Many analyses are better formulated imperatively

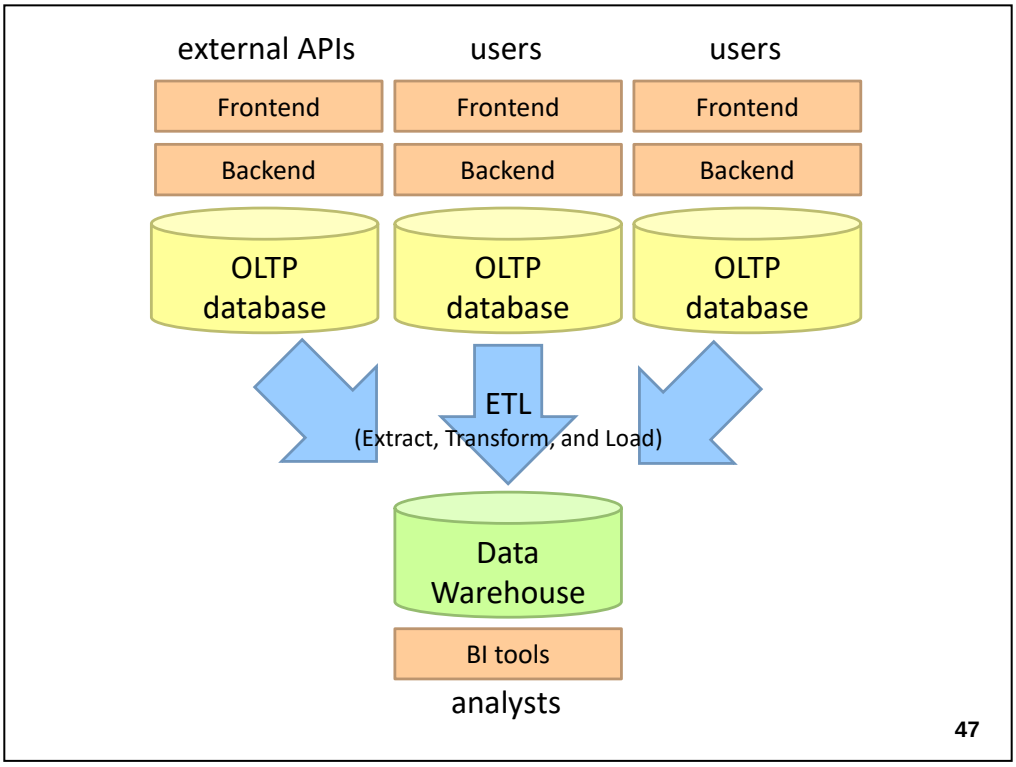
Much faster data ingest rate

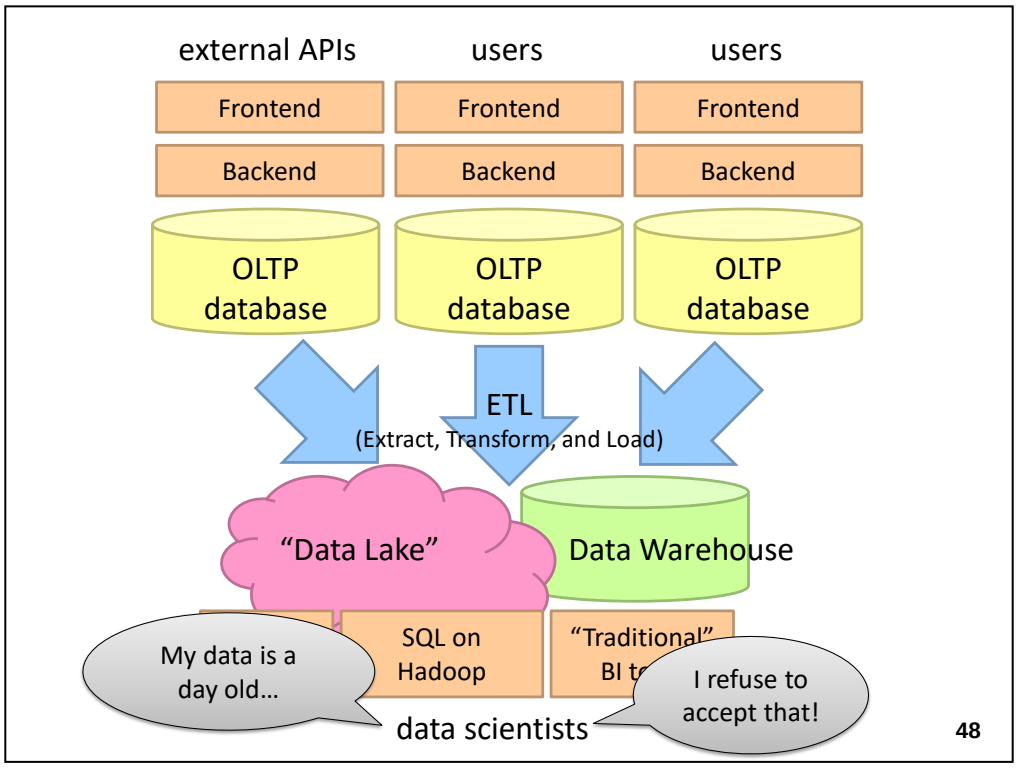

external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
Data Warehouse
BI tools
analysts
43

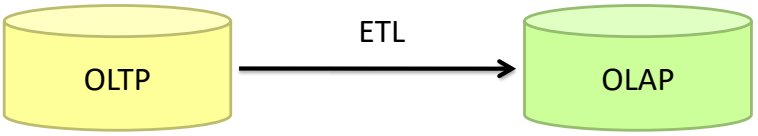

external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
"Data Lake"
Data Warehouse
Other tools
SQL on Hadoop
"Traditional" BI tools
data scientists
44


What's Next?
45


users
Frontend
Backend
database
BI tools
analysts
46

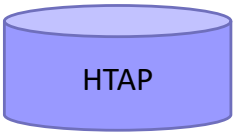

external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
OLTP database
OLTP database
OLTP database
ETL
(Extract, Transform, and Load)
Data Warehouse
BI tools
analysts
47


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 tr
My data is a day old…
I refuse to accept that!
data scientists
48



What if you didn't have to do this?

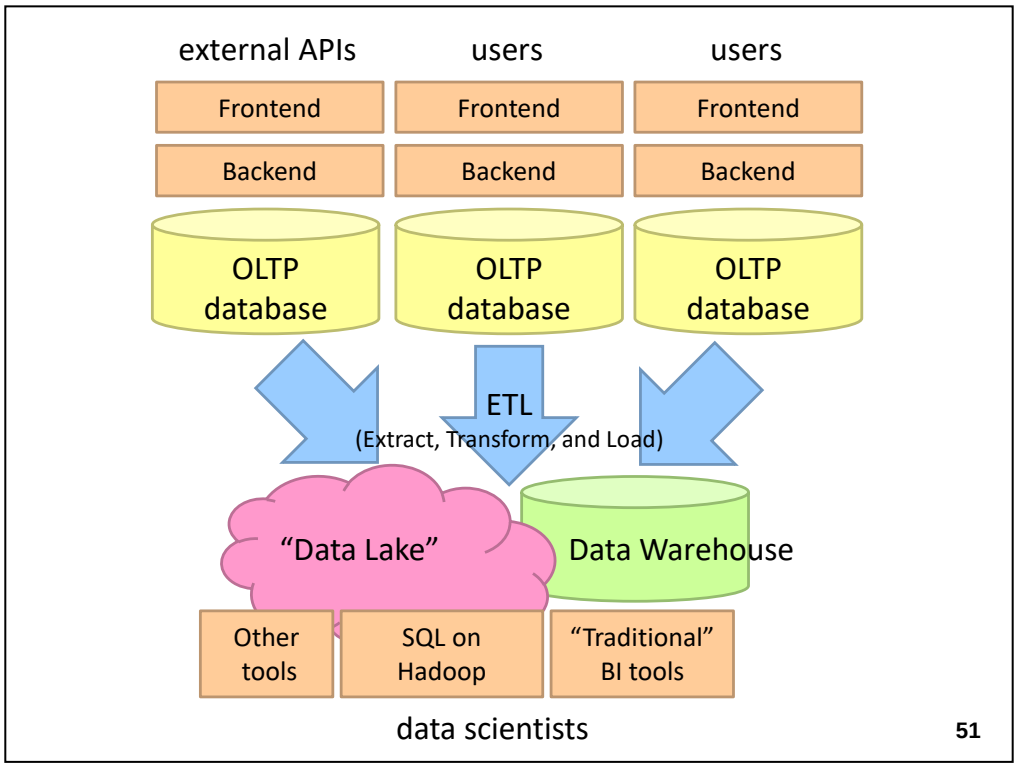
49

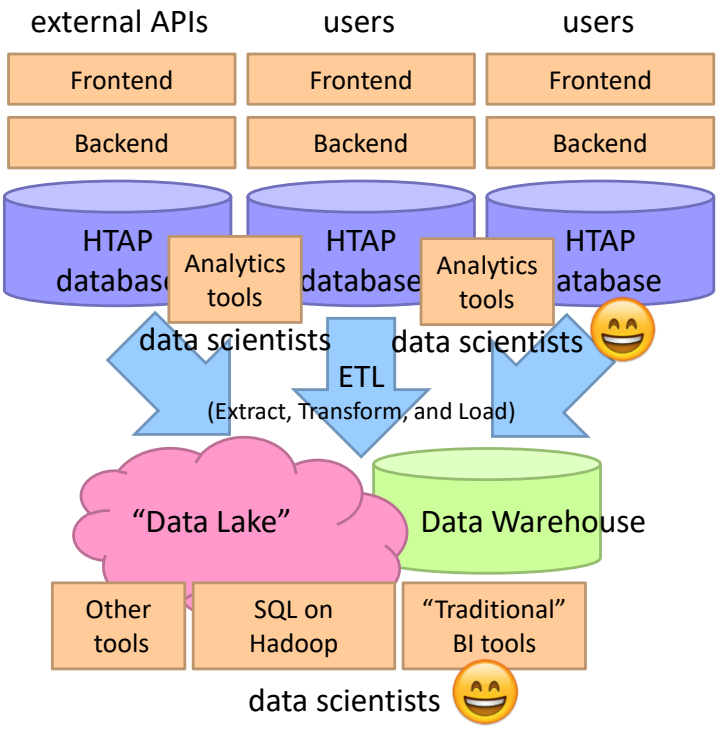


Hybrid Transactional/Analytical Processing (HTAP)

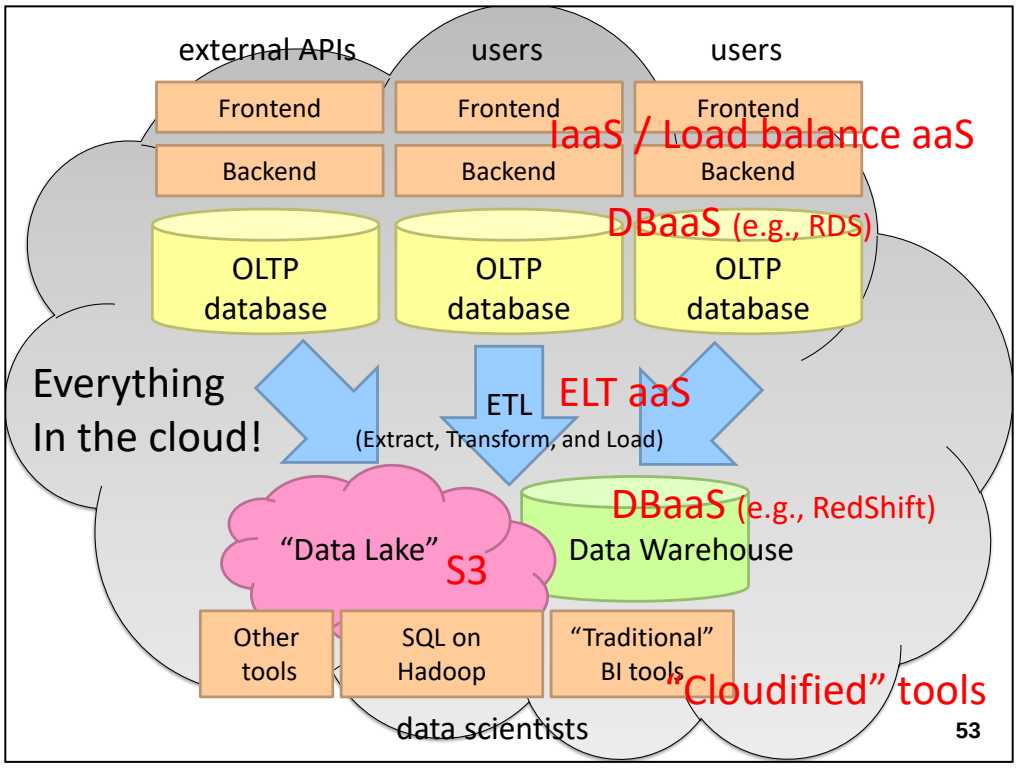
Coming back full circle?

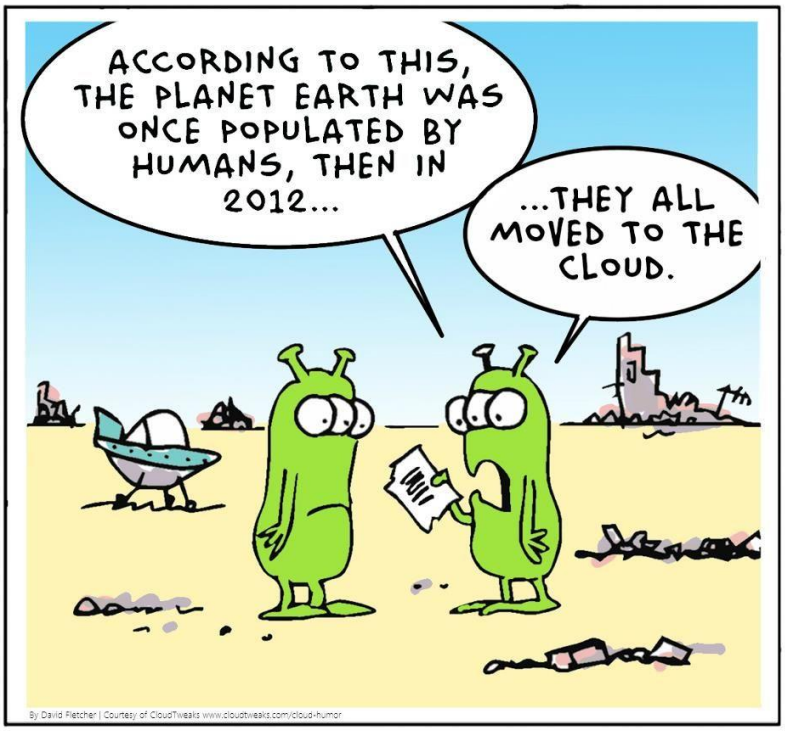
50


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
51


external APIs
users
users
Frontend
Frontend
Frontend
Backend
Backend
Backend
HTAP
HTAP
HTAP
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

52


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
(Extract, Transform, and Load)
ELT aaS
DBaaS (e.g., RedShift)
"Data Lake"
S3
Data Warehouse
Other tools
SQL on Hadoop
"Traditional" BI tools
"Cloudified" tools
data scientists
53


ACCORDING TO THIS, THE PLANET EARTH WAS ONCE POPULATED BY HUMANS, THEN IN 2012...
...THEY ALL MOVED TO THE CLOUD.
By David Fletcher | Courtesy of CloudTweaks www.cloudtweaks.com/cloud-humor