Data-Intensive Distributed Computing
CS 431/631 451/651 (Winter 2021)

Part 6: Analyzing Relational Data (1/3)

Ali Abedi

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

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States
See http://creativecommons.org/licenses/by-nc-sa/3.0/us/ for details
Structure of the Course

“Core” framework features and algorithm design

Analyzing Text
Analyzing Graphs
Analyzing Relational Data
Data Mining
Evolution of Enterprise Architectures

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

Monolithic Application
users

- Frontend
- Backend
Edgar F. Codd

• Inventor of the relational model for DBs
• SQL was created based on his work
• Turing award winner in 1981
Why is this a good idea?
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

BI tools

database

Backend

Frontend

users

analysts
Why is my application so slow?

Why does my analysis take so long?
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!
OLTP database for user-facing transactions

OLTP database

ETL (Extract, Transform, and Load)

OLAP database for data warehousing

Data Warehouse

BI tools

analysts

Frontend

Backend

users
A Simple OLTP Schema
A Simple OLAP Schema

- Dim_Customer
- Dim_Date
- Dim_Product
- Fact_Sales
- Dim_Store
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

Meh.

My data is a day old…
What do you actually do?

Report generation
Dashboards
*Ad hoc* analyses
OLAP Cubes

Common operations
slice and dice
roll up/drill down
pivot

https://youtu.be/LRdsZqrwOrc
Fast forward...
“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
database

ETL
(Extract, Transform, and Load)

Data
Warehouse

BI tools

analysts

Facebook context?
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
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
Virtuous Product Cycle

a useful service

$\ (\text{hopefully})$

transform insights into action

analyze user behavior to extract insights

What do you actually do?

Report generation
Dashboards
Ad hoc analyses
“Descriptive”
“Predictive”
Data products
“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
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
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!
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

Frontend
Backend

users
Frontend
Backend

users
Frontend
Backend

OLTP database

OLTP database

OLTP database

ETL
(Extract, Transform, and Load)

Data Warehouse

BI tools

analysts
What’s Next?
external APIs

Frontend

Backend

OLTP database

users

Frontend

 Backend

OLTP database

users

Frontend

 Backend

OLTP database

ETL
(Extract, Transform, and Load)

Data Warehouse

BI tools

analysts

OLTP database

OLTP database
My data is a day old...
I refuse to accept that!
What if you didn’t have to do this?
HTAP

Hybrid Transactional/Analytical Processing (HTAP)

Coming back full circle?
IaaS / Load balance aaaS

DBaaS (e.g., RDS)

DBaaS (e.g., RedShift)

Everything In the cloud!

ETL (Extract, Transform, and Load)

"Cloudified" tools

"Data Lake"

S3

OLTP database

OLTP database

OLTP database

Other tools

SQL on Hadoop

"Traditional" BI tools

"Data Warehouse"