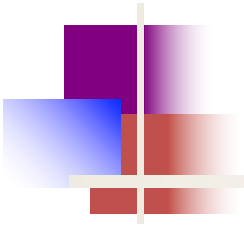


Data-Driven Decisions

Chapter 4

Using DBMS

Two main types:

- Operational (or Transactional) Databases
  - E.g. “what courses does student x have?”
  - E.g. “is the account balance large enough to permit this withdrawal at an ATM?”
  - Lots of queries from lots of users.
  - Tight time (usually real-time) constraints.
  - But most queries are quick.

Using DBMS (2)

- Analytic Databases or Data Warehouses
  - Not used for business operations
  - Used for data analysis and decision-making
  - Most queries are long-running, and require aggregating over many data elements
  - Make all queries read only
  - Have a separate (periodic) update process
  - Very different performance parameters compared to operational DBMSs

Schema for Data Warehouses

- One central FACT table with billions of entries.
  - Columns classified as DIMENSION or MEASURE. Primary key is all dimensions.
- Multiple DIMENSION tables, each joining with exactly one dimension column.
- Dimension tables are typically small.
- Not all dimension columns in the FACT table have a dimension table.

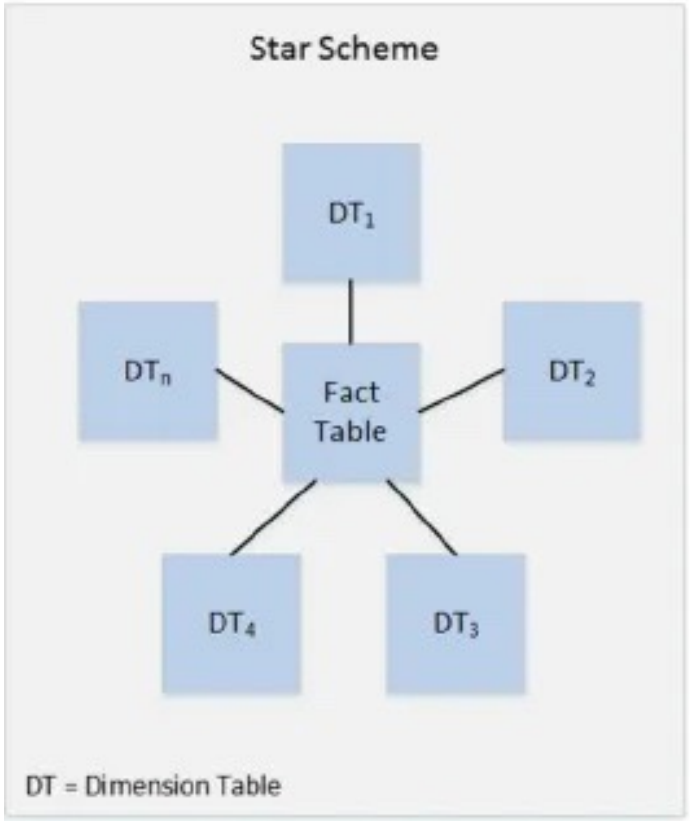
Star Schema

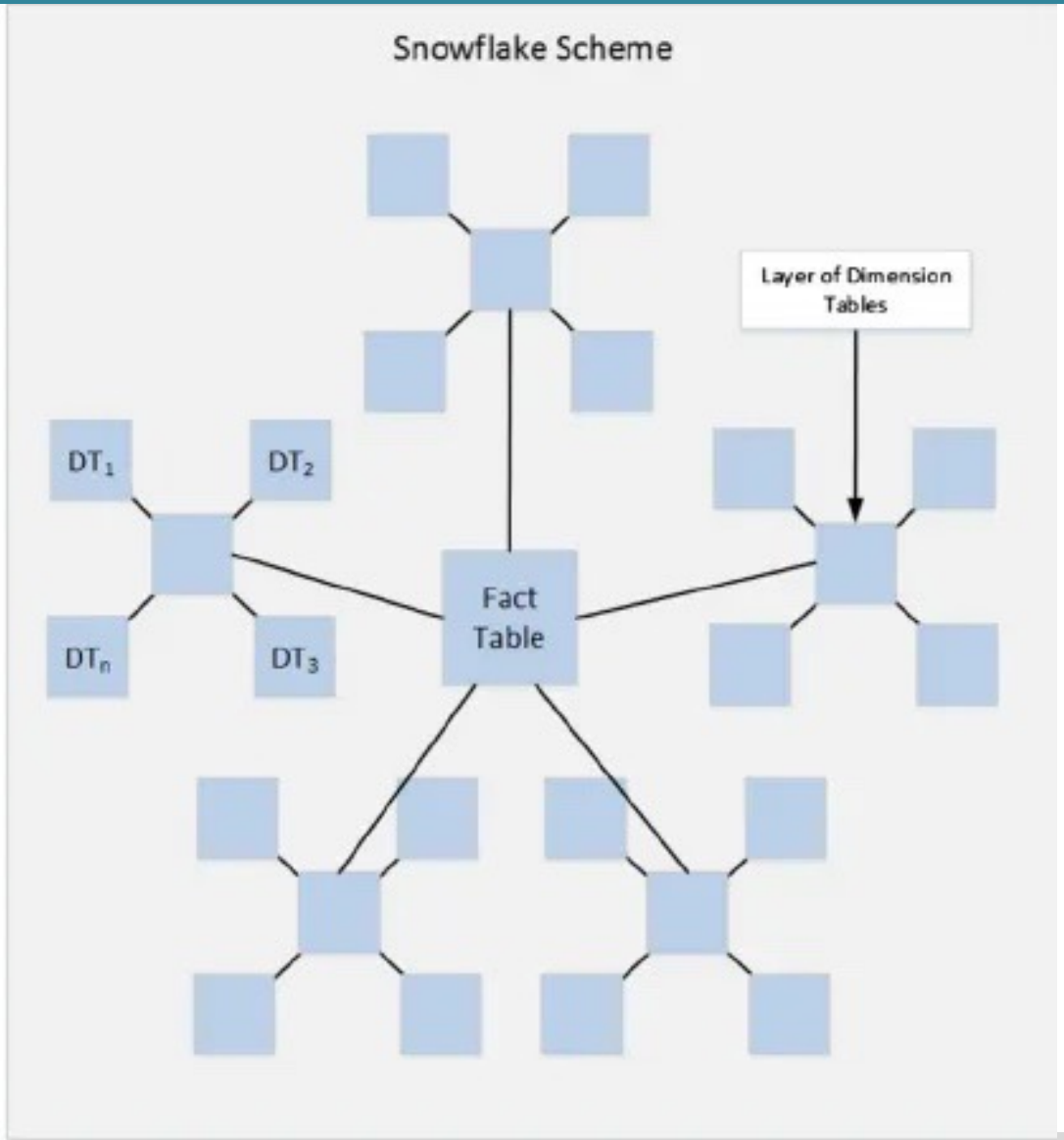
- Purchase(Itemid, Custid, Registerid, Time, Quantity, Price) ← Measure Attributes
- Item(Itemid, Description, Supplier, Size)
- Customer(Custid, Name, Address)
- CashRegister(Registerid, Storeid)

Snowflake Schema

- Purchase(Itemid, Custid, Registerid, Time, Quantity, Price)
- Item(Itemid, Description, Supplier, Size)
- Customer(Custid, Name, Address)
- CashRegister(Registerid, Storeid)
- Store(Storeid, Address, City, Stateid)
- State(Stateid, Name, Regionid)
- Region(Regionid, Name)

Star vs Snowflake





The Value of Data

- Data sets are bought and sold every day.
- Value is in the organization of data.
- Increase value by doing work:
  - Data cleaning
  - Data integration
  - More convenient access tools
  - …
- Very poor theory on how to price.

Surveillance Capitalism

- Do we really get stuff for free from companies?
- How do TV networks make money?
- How do Google, Facebook, etc. make money?
- How do these differ?

- Shoshana Zuboff

Privacy

- Ability to control sharing of information about self.
- Basic human need.
  - Even for people who have “nothing to hide”

Privacy and the Law

- The fifth amendment is a basis for privacy in the US, even though it does not explicitly say anything about privacy.
- What is in your head is your private information, and you cannot be compelled to give it out, when that may incriminate you.
- What is in your home is also somewhat private. You can be compelled to show it, but only with a warrant obtained through due process.
- But, no protection for non-private information.

Scale Matters

- I don’t mind being observed by passers-by when I am on a public street.
  - My expectation is that any one observer will only get a single fleeting observation of me.
- What if observers could pool data?
- They could track all my movements!
  - I certainly object to that!!

Loss of Privacy

- Due to loss of control over personal data.
- I am OK with you having certain data about me that I have chosen to share with you or that is public, but I really do not want you to share my data in ways that I do not approve.
- Data I voluntarily give to a service provider may only be used for the purpose specified.
  - GDPR in the EU
  - HIPAA in the US

Facebook/Cambridge Analytica

- Your preferences can be predicted by the app, better than by your roommate, based on 70 ”like”s on Facebook. (Better than your spouse with 300 “like”s).
- Once someone has such a powerful app, they really know you, and can “push your buttons”.
- We need to limit such use if we are to feel free to share in the datafied world.

Choice May not be Yours to Make

"The Golden State killer," Joseph DeAngelo, was identified on account of partial matches with DNA his cousins had entered at a genealogy website.

No Option to Exit

- In the past, one could get a fresh start by:
  - Moving to a new place
  - Waiting till the past fades
    - Reputations can be rebuilt over time.
- Big Data is universal and never forgets anything!!
  - Way back machine for the web
- Can we develop techniques to forget?
  - In Europe, right to forget.
  - Implemented via search engines.

Anonymity

NETFLIX PRIZE

Closeted Lesbian Sues Netflix For Potential Outing

By Laura Northrup on December 19, 2009 3:00 PM

Here's the problem with anonymized data: if it were truly anonymized, it wouldn't be useful to anyone for anything. With enough data about a person—say, their age, gender, and zip code—it's not hard to narrow down who someone is. That's the idea behind a class-action lawsuit against Netflix regarding the customer data they released to the public as part of the Netflix Prize project, a contest to help create better movie recommendations. A closeted lesbian alleges that the data available about her could reveal her identity.

Consumerist.com

Anonymity is Impossible

- Anonymity is virtually impossible, with enough other data.
  - Diversity of entity sets can be eliminated through joining external data
  - Random perturbation works only if we can guarantee a one-time perturbation
  - Aggregation works only if there is no known structure among entities aggregated
- Faces can be recognized in image data.
  - Progressively, even under challenging conditions, such as partial occlusion

Anonymity Techniques

- K-Anonymity
  - Require at least k entries in a group about which information is revealed.
  - Hope that is enough to hide details about any one individual.
  - But not provably safe.
- Differential Privacy
  - Only respond to aggregate queries about the data.
  - Add carefully calibrated noise to the aggregate value being reported.
  - Can guarantee (with high probability) not revealing detail data about presence of any individual in the data set.

Differential Privacy

- Elegant concept
- Provides a sound basis for someone collecting detailed data to release aggregates safely
- Widely used today
  - E.g. US Census Bureau
- Really important to be able to find aggregate patterns without delving into individual details
- Aggregate value has enough noise added that an adversary cannot tell whether any particular data item has been included in the aggregate.

Example of Diff. Priv.

Sailors

sid	sname	rating	age
22	Dustin	7	45.0
29	Brutus	1	33.0
31	Lubber	8	55.5
32	Andy	8	25.5
58	Rusty	10	35.0
64	Horatio	7	35.0
71	Zorba	10	16.0
74	Horatio	9	35.0
85	Art	3	25.5
95	Bob	3	63.5

Suppose we want to know the average age of sailors with rating 3.
That average is 44.5.

Can we disclose this without revealing any individual's rating?

Since most sailors are younger than that, we can guess that Bob, who is an older sailor, must have rating 3

So we output 44.5+n instead.
(n is a noise term: + or –ve)

Differential Privacy Shortcomings

- Needs strong assumptions
  - Privacy “budget” for repeated release
  - No one else can collect and release correlated data
- Complaints about added noise from some researchers
- Works only for an all-or-nothing model of privacy.
  - Does not support sharing with control

Algorithmic Fairness

- Do the data “speak for themselves”?
- Can algorithms be biased?
- Can we make algorithms unbiased?
  - Is training data set representative of the population?
  - Is past population representative of future population?
  - Are observed correlations due to confounding processes?

Validity

- Bad data leads to bad decisions.
- But most data are dirty.
- If decision-making is opaque, results can be bad in the aggregate, and catastrophic for an individual.
- What if someone has a loan denied because of an error in the data analyzed?

Third Party Data

- Material decisions can often be made on the basis of public data or data provided by third parties.
- There often are errors in these data.
- Does the affected subject have a mechanism to correct errors?
  - Credit rating data on steroids.
- Does the affected subject even know what data were used?
- “Right of recourse”

Biased Data

- Data collection mechanisms often result in biases.
  - Whether these matter requires thought.
- Social media posts are not representative of the general population
  - Skew younger, better educated, more tech-savvy.
  - Over-represent people with strong opinions
- Medical tests often at one (or a few) local site(s)
  - But results are claimed to apply throughout the world.
  - Most humans are indeed alike.
  - But what about racial/genetic differences?
  - Environmental differences between rich and poor nations.

Equity

Treat people differently based on their circumstances to achieve comparable outcomes.

Equity ≠ Fairness

Fairness

Equity

MPCA Photos, pexels.com, CC-BY-NC 2.0

Example of Equity

- It is fair to give each student in the class the same amount of time to take an exam.
- Equity requires allowing extra time for some students.

Example of Model Equity

- It is fair to measure every applicant's knowledge/potential through a standardized test, such as GRE.
- Equity requires taking into account studies showing the strong correlation between test performance and socio-economic status (and gender and race and …).

Example of Data Equity

- It is fair to create a training data set that is an unbiased sample of the population: each minority group is represented in proportion to its size in the population.
- Equity may require over-sampling of small minorities. If a small minority group “behaves” differently than others, the model may minimize aggregate error by ignoring the minority group.

Conclusion

Data-driven automation can do a lot more, and do it a lot faster. But the “it” needs to be defined carefully.