Preliminaries

AnHai Doan

What We Will Discuss

- **Overview of data management at organizations**
  - where DS and data exploration/cleaning/integration fit in
- **Relational data**
- **ML**
- **Others (cover later as the topic arises)**
  - Python, big data scaling, crowdsourcing, keyword search

Overview of Data Management at Organizations

- **Organizations**
  - companies, non-profit organizations (including UW), government agencies, militaries
  - others: big research groups, etc.
- **An organization has**
  - employees, products, customers, suppliers
  - even UW fits this template, it offers courses to students
- **Business needs of organizations**
  - bookkeeping, logistics (eg keep track of employees, customers, products, suppliers, etc.)
  - managing business processes/transactions/workflows
    - eg how a customer purchases a product, travel requests/approvals, payroll
  - data analysis (historical analytics, predictive analytics)
  - gathering more information
    - eg find all information about a customer, for marketing/customer service
    - eg find weather/crime related information for building security purposes

How Data Management Evolves as an Organization Grows

- **Let's take a look at the evolution of an organization X, from small to big**
  - and how the data management needs/solutions grow along the way
- **When X is very small, e.g., startup**
  - need some bookkeeping/logistics
  - eg GreenBay, a data integration company (using AI/ML)
  - payroll using Gusto on cloud

How Data Management Evolves as an Organization Grows

- **When X becomes bigger (medium-size organizations)**
  - hundreds of employees, tens of products, hundreds to thousands of customers
  - need RDBMSs to keep track of employees/products/customers/suppliers/etc.
  - a lot more need for business processes/workflows/transactions (= business functions)
    - CRM: customer relationship management
    - ERP: enterprise resource planning
    - communication: Zoom, Teams/Slack, Outlook/Gmail
    - customer facing: Web page, social media, email, chat, phone call logs, etc.
    - and a lot more (see “Business software” on Wikipedia)
- **All these software generate a lot of data**
  - capture and store them somewhere
- **Raises the need for analytics, three main goals**
  - infer insights to do better business: eg sell more cheese in Arizona
  - infer insights to improve internal business processes
  - infer stuff for security, intrusion detection, hacking prevention, etc.

How Data Management Evolves as an Organization Grows

- **Two different kinds of analytics**
- **Historical analytics**
  - find patterns in existing (aka historical) data
  - eg people in Arizona consume a lot more cheese than people in Wisconsin
  - eg cluster customers who have obtained a loan from a bank to detect interesting patterns
    - maybe the biggest cluster is college students from Ohio
- **Predictive analytics**
  - develop a MODEL to predict what can happen next
  - eg predict the amount of cheese people in Arizona will consume next winter
  - eg predict if a given potential customer will repay a loan or not
- **Since X has a lot of analytic needs, may create certain kinds of data warehouses to help with these**
  - data warehouse, data lake, lakehouse, etc.
- **X may also want more information about certain topics**
  - e.g., customers, products, so build Customer 360s, Product 360s

How Data Management Evolves as an Organization Grows

- **When X becomes really big (the largest organizations today)**
  - eg GreenBay now become national, spanning multiple states, acquiring other startups
  - eg Informatica, insurance companies, hospital chain, Google, Microsoft, etc.
- **Now there is a lot more to worry about**
  - lot of need to move data efficiently: replicate, synchronize, transform (eg split, merge, etc.)
  - may need to process some data very fast, eg process log to detect intrusion
  - data quality, profiling
  - data security, privacy, logging
  - data catalog
  - data governance, compliance
  - meta-data
    - business terms, reference tables, knowledge graphs, etc.
  - acquiring external data

Summary of Needs

Needs	Solutions
Bookkeeping, logistics + RDBMSs Business functions + aka processes/transactions/workflows + CRM, ERP, Zoom, etc. Analytics & report + historical, predictive Find more information about … + customers, products, etc. Move data efficiently + replicate, synchronize, transform (split, merge, …) Process fast data Data quality, profiling Security, privacy, logging Data catalog Data governance, compliance Meta data (business terms, ref tables, KGs, …) Acquiring external data	

Solutions

Needs	Solutions
Bookkeeping, logistics + RDBMSs Business functions + aka processes/transactions/workflows + CRM, ERP, Zoom, etc. Analytics & report + historical, predictive Find more information about … + customers, products, etc. Move data efficiently + replicate, synchronize, transform (split, merge, …) Process fast data Data quality, profiling Security, privacy, logging Data catalog Data governance, compliance Meta data (business terms, ref tables, KGs, …) Acquiring external data	Hardware, machines, cloud infrastructure Software, data artifacts + RDBMSs + business software: CRM, ERP, etc. + logging + data warehouse, lake, lakehouse, etc. + Customer 360s, Product 360s, etc. + software to manage and create - data catalog - data governance - meta data (business terms, ref tables, KGs) Hire people to manage all of these + software engineers + data engineers, data scientists + data stewards, subject matter experts (SME) + business users/analysts + and more

Solutions

- **Organizations increasingly maintain multiple hardware infrastructures**
  - on-prem, cloud
- **Software split into on-prem and cloud software**
  - but things are increasingly being moved to cloud: both data and software
- **More on people**
  - many different roles, and a person may play multiple roles
  - roles are often also called “personas”
- **Main categories for role**
  - general software roles: backend, frontend, full-stack
  - data related roles: data scientists, data engineers, data quality experts, etc.
    - data stewards, data owners
  - business related roles: subject matter experts (SMEs), business analysts/users
  - ML/AI related roles: ML engineers, MLOps engineers, etc.
  - data acquisition roles: to buy data from outside
  - crowdsourcing roles: to manage crowdsourcing tasks, eg label data for ML
- **Another split: business users vs technical users**

Recent Trends

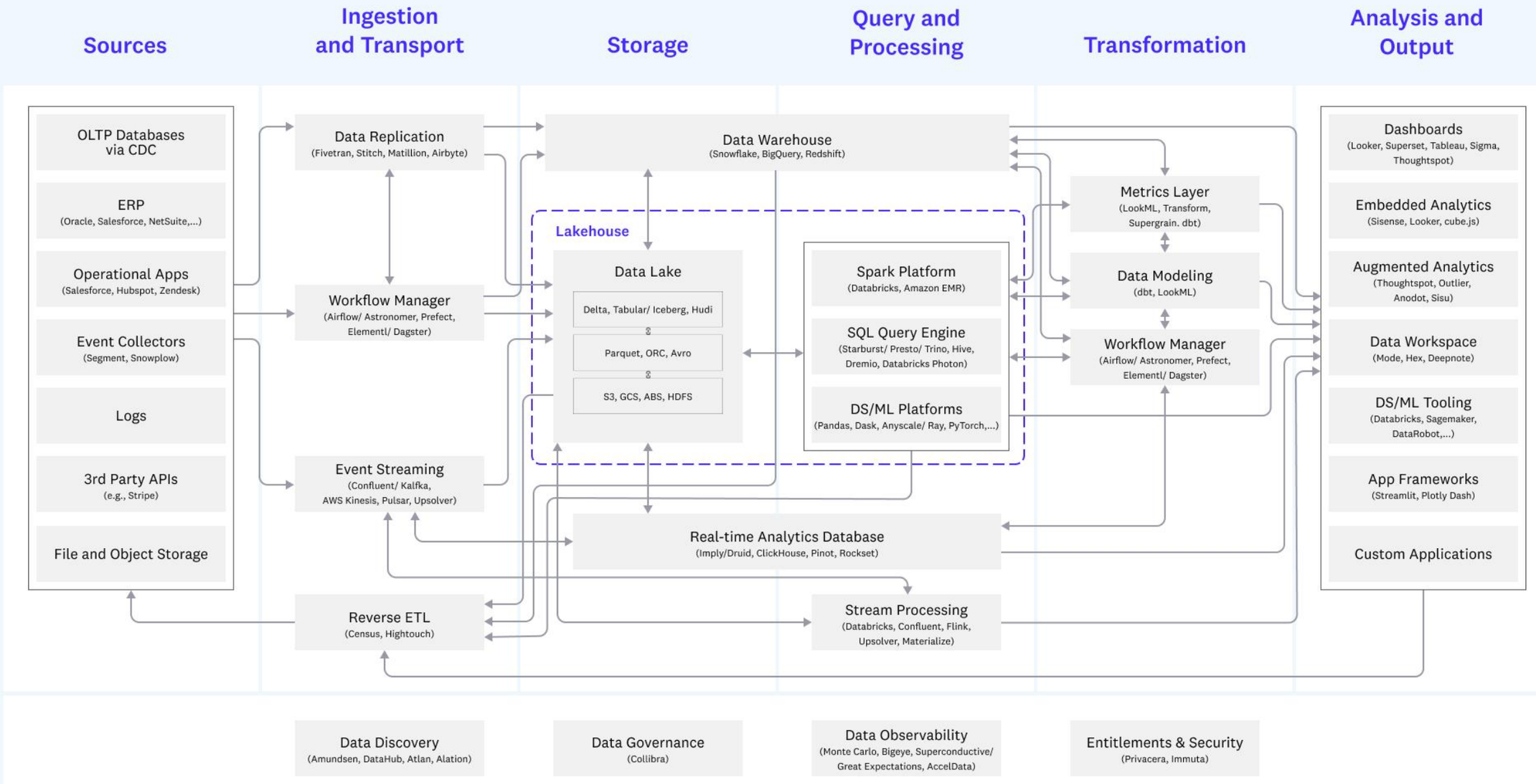
- On prem vs cloud, mobile
- Self service
- AI/ML
- Data streams

Levels of Data Maturity

8	Drive Innovation	Using data analytics to spark innovations that differentiate the organization in a competitive market.
7	Drive Organizational Learning	Using data science, artificial intelligence, and machine learning to adapt to changing business conditions, recommend and/or automate decisions and actions, and increase efficiencies and competencies.
6	Look into the Future	Using data mining and predictive analytics to understand probable future conditions and events, and to inform and guide the strategies and tactics that shape the organization's future.
5	Understand Cause & Effect	Using data for causal analysis, to understand why things happen, and to identify leverage points to effect change—knowing how to create more of desired outcomes and to reduce or eliminate the undesirable.
4	Understand Patterns & Trends	Using data for analysis and visualization to understand correlations and connections among business variables, and to see behavior over time of various business metrics.
3	Know What Has Happened	Using data for descriptive and retrospective analysis of business outcomes (know what has happened), to quantify the outcomes (know how much), and to see the outcome historically (know when).
2	Reporting	Using data to produce reports about business entities, events, and results—from internal reference and management reporting to external corporate, compliance, and enterprise reporting.
1	Record Keeping	The most basic level of data management uses data simply as a digital record of business transactions and events.

Google for "seven core responsibilities of a chief data officer eckerson"

Unified Data Infrastructure (2.0)



From the Web, google for “unified data infrastructure a16z”

Where Does DS and DI Fit?

Needs	Solutions
Bookkeeping, logistics + RDBMSs Business functions + aka processes/transactions/workflows + CRM, ERP, Zoom, etc. Analytics + historical, predictive Find more information about … + customers, products, etc. Move data efficiently + replicate, synchronize, transform (split, merge, …) Data quality, profiling Security, privacy, logging Data catalog Data governance, compliance Meta data (business terms, ref tables, KGs, …) Acquiring external data	Hardware, machines, cloud infrastructure Software, data artifacts + RDBMSs + business software: CRM, ERP, etc. + logging + data warehouse, lake, lakehouse, etc. + Customer 360s, Product 360s, etc. + software to manage and create - data catalog - data governance - meta data (business terms, ref tables, KGs) Hire people to manage all of these + software engineers + data engineers, data scientists + data stewards, subject matter experts (SME) + business users/analysts + and more

Where Does DS and DI Fit?

- **A big part of DS is to do analysis, or build data artifacts to help do analysis**
  - a major function as organizations get bigger
- **To do DS, need to work with a lot of data**
  - internally, externally
- **Data comes in all kinds of format, uses different vocabularies, often is hard to understand and dirty**
  - need data exploration, cleaning, integration

- **Big Data**
  - volume
  - velocity
  - variety
  - veracity
  - we focus on the last two

What We Will Discuss

- **Overview of data management at organizations**
  - where DS and data exploration/cleaning/integration fit in
- **Relational data**
- **ML**
- **Others (cover later as the topic arises)**
  - Python, big data scaling, crowdsourcing, keyword search

Relational Data

- **Ideally, you should have taken an ugrad DB class, eg 564**
- **If not, here's the minimum to know for this class**
- **Most data at most organizations are tables**
- **Each table has schema and data**
  - schema: table name, column names, constraints on the tables
  - data: the tuples/rows
- **A database is a set of tables, it also has schema and data**
  - schema of the database is the set of table schemas: table names, col names, constraints
  - data of the database is the tuples of the tables
- **If you know SQL, great, otherwise can learn just the basic select-project-join queries over tables, or else we will cover it in the class**

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

data cleaning: GE revenue: 351 ⟶ 35.1

schema matching: name = cname
loc = address

schema merging: X(name, loc)
Y(cname, address, rev)
} Z(name, loc, rev)

data matching:

M

xid	yid
x_1	y_2
x_2	y_1

data merging: for name, return the longer string from X.name and Y.cname
for loc, return X.loc

schema mapping: Z = select merge_name(X.name, Y.cname), X.loc, Y.rev
from X, Y, M
where X.id = M.xid and Y.id = M.yid

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

Z

name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25

Machine Learning

- **Not a course on ML, we only use ML to solve DI tasks**
- **Main ML topics: classification, clustering, reinforcement learning**
- **We just mostly use classification, aka supervised learning**
- **Key concepts to know**
  - example, label
  - training, testing, ML model, trained model
  - search for the model that best fits the training data
  - popular models: decision tree, random forest
  - latest sexiest model: deep learning (DL) models
    - here notion of output label is generalized
  - we will also touch upon foundation models, which are very large DL models that have been trained on a very large amount of data

Foundation Models

A dog looking curiously in the mirror, seeing a cat.

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A giant cobra snake on a farm. The snake is made out of corn.

A marble statue of a Koala DJ in front of a marble statue of a turntable. The Koala has wearing large marble headphones.

Others

- **Python**
  - pandas, Jupyter notebook
  - why Python? Lot of data science stuff
- **Big data scaling, eg Spark**
  - you don’t have to know this, we will introduce the minimum if necessary
- **Other topics will be introduced as they arise**