UWE Bristol | University of the West of England

Presented by

Dr. Trupti Padiya

School of Computing and Creative Technologies

Advanced Databases UFCFU3-15-3

Big Data & Data Lakes

Slides with modification adapted from Dr. Elias Pimenidis

Big Data?

- Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis.
- It’s not the amount of data that’s important. It’s what organizations do with the data that matters. Data can be analysed for insights that lead to better decisions and strategic business moves.

Big Data – the Three Vs

- **Volume**. Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data.
- In the past, storing it would've been a problem – but new technologies have eased the burden. Still important though as it affects processing

Big Data – the Three Vs (2)

- **Velocity.** Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.
- **Variety.** Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

Big Data – two more Vs

- **Variability.** In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads **can be challenging to manage**. Even more so with unstructured data.

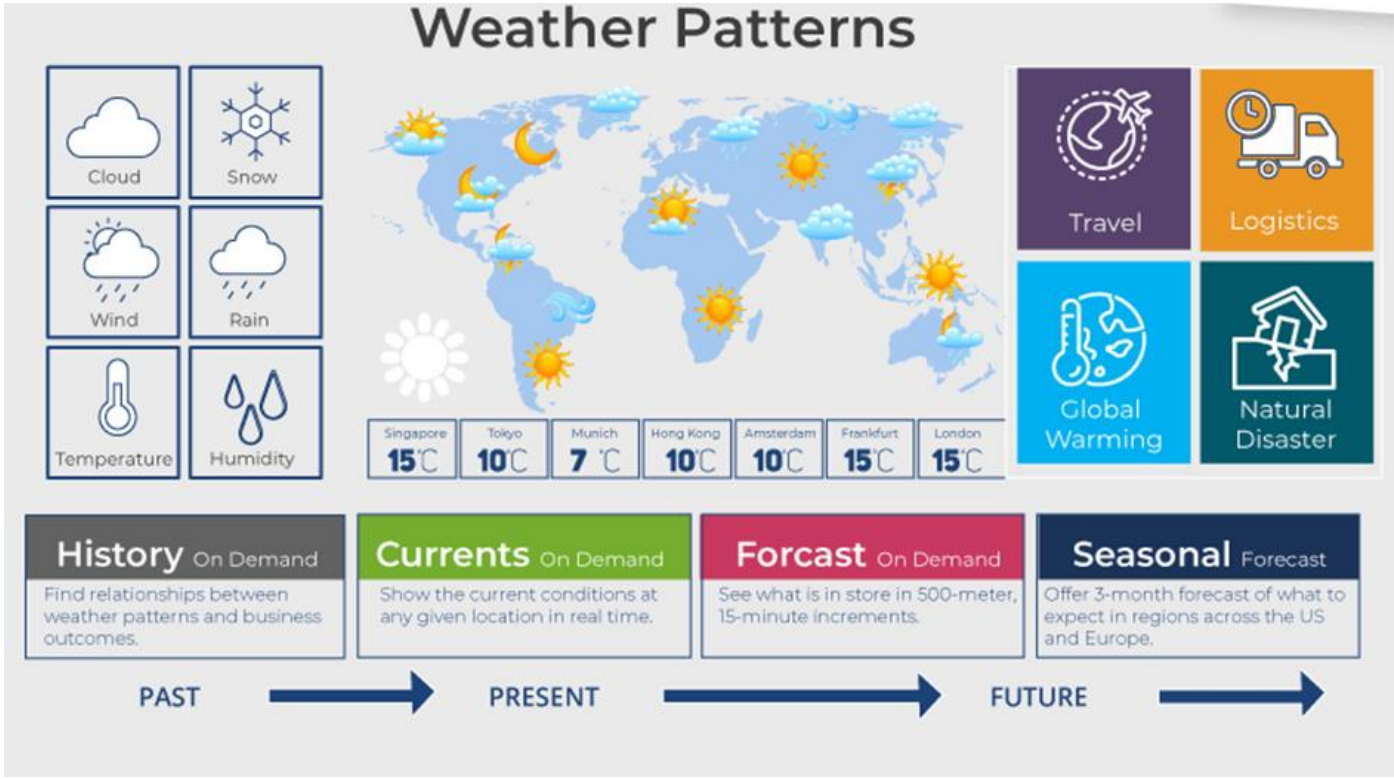
Big Data – two more Vs (2)

- **Complexity**. Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's **necessary to connect and correlate relationships, hierarchies and multiple data linkages** or your data can quickly spiral out of control.

The importance of Big Data

- Doesn’t revolve around how much data you have, but what you do with it. You can take data from any source and analyse it to find answers that enable
- 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smart decision making.

Big Data in Weather Patterns



Big Data in Weather Patterns

- There are weather sensors and satellites deployed all around the globe. A huge amount of data is collected from them, and then this data is used to monitor the weather and environmental conditions.

- All of the data collected from these sensors and satellites contribute to big data and can be used in different ways such as:

- In weather forecasting
- To study global warming
- In understanding the patterns of natural disasters
- To make necessary preparations in the case of crises
- **To predict the availability of usable water around the world**

Big Data in education

Big Data in Education

- Customized and Dynamic Learning Programs using the data collected on the bases of each student's learning history. This improves the overall student results.
- Reframing Course Material according to the data that is collected on the basis of what a student learns and to what extent by real-time monitoring of the components of a course is beneficial for the students.
- Grading Systems as a result of a proper analysis of student data.
- Career Prediction Study of every student's records will help understand each student's progress, strengths, weaknesses, interests, and more. It would also help in determining which career would be the most suitable for the student in future.

Big Data in Government Sector

Big Data in Government Sector

- Governments come face to face with a very huge amount of data on almost daily basis. They have to keep track of various records and databases regarding their citizens, their growth, energy resources, geographical surveys, and many more. All this data contributes to big data. Some areas where this can contribute are:
- Welfare Schemes
  - In making faster and informed decisions regarding various political programs
  - To identify areas that are in immediate need of attention
  - To stay up to date in the field of agriculture by keeping track of all existing land and livestock.
  - To overcome national challenges such as unemployment, terrorism, energy resources exploration, and much more.
- Cyber Security
  - Big Data is hugely used for deceit recognition.
  - It is also used in catching tax evaders.

Big Data in Media and Entertainment Industry

Big Data in Media and Entertainment Industry

- With people having access to various digital gadgets, generation of large amount of data is inevitable and this is the main cause of the rise in big data in media and entertainment industry.
- Social media platforms are another way in which huge amount of data is being generated.
- Some of the benefits extracted from big data in the media and entertainment industry are given below:
  - Predicting the interests of audiences
  - Optimized or on-demand scheduling of media streams in digital media distribution platforms
  - Getting insights from customer reviews
  - Effective targeting of the advertisements

Big Data Contributions to Healthcare

Healthcare contributions

- Healthcare is yet another industry which is bound to generate a huge amount of data. Here's how Big data has contributed to healthcare:
- Reduces costs of treatment since there is less chances of having to perform unnecessary diagnosis.
- Predicting outbreaks of epidemics and also in deciding what preventive measures could be taken to minimize the effects of the same.
- Avoid preventable diseases by detecting them in early stages. It prevents them from getting any worse which in turn makes their treatment easy and effective.
- Patients can be provided with evidence-based medicine which is identified and prescribed after doing research on past medical results.

IoT, Cloud & Big Data

- IoT is generating massive volumes of structured and unstructured data, and an increasing share of this data is being deployed on cloud services. The data is often heterogeneous and lives across multiple relational and non-relational systems.
- Innovations in storage and managed services have sped up the capture process.
- Accessing and understanding the data itself still pose a significant last-mile challenge.

IoT, Cloud & Big Data

- As a result, demand is Growing for analytical tools that seamlessly connect to and combine a wide variety of cloud-hosted data sources.
- Such tools enable businesses to explore and visualize any type of data stored anywhere, helping them discover hidden opportunity in their IoT investment.

What is a data lake?

- A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data.
- A data puddle is basically a single-purpose or single-project data mart built using big data technology. It is typically the first step in the adoption of big data technology. The data in a data puddle is loaded for the purpose of a single project or team.

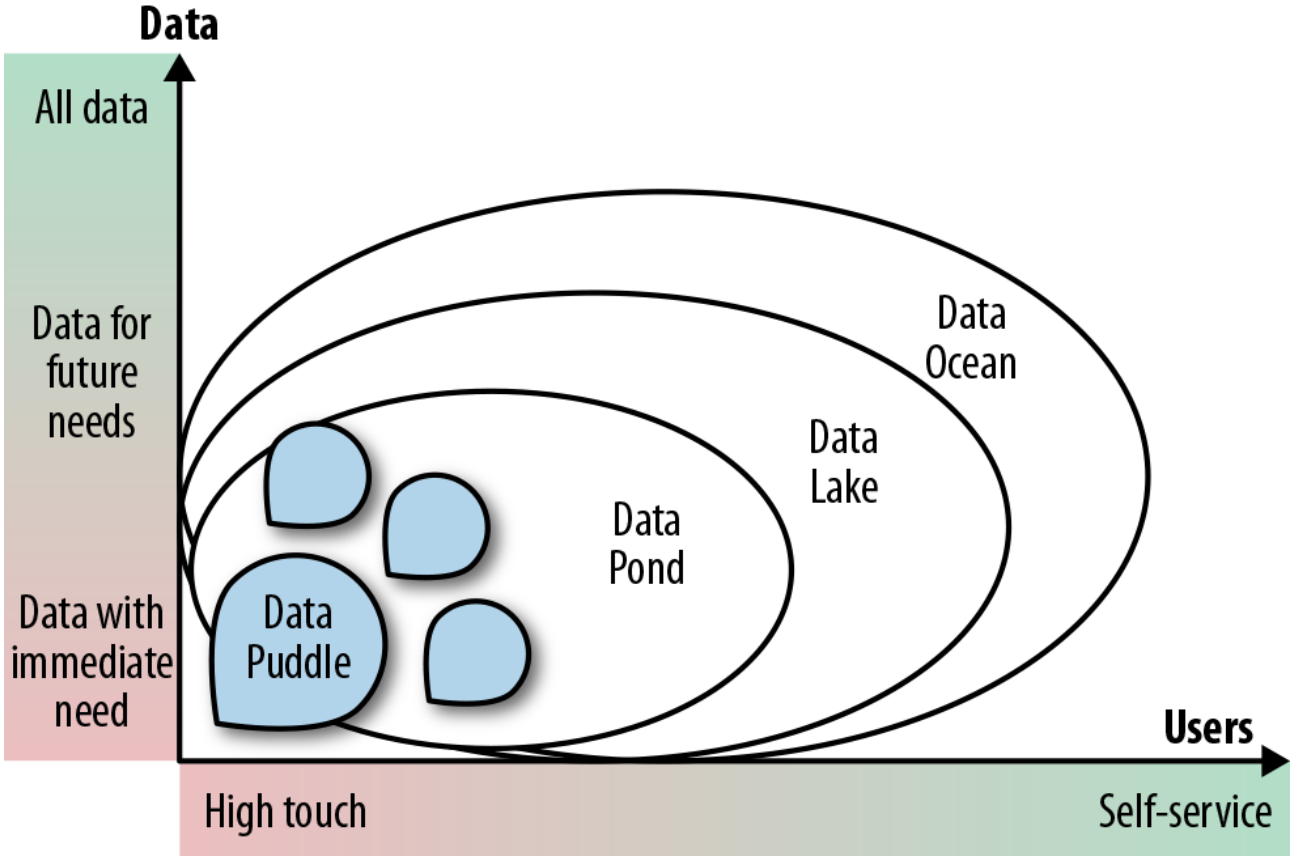
A Data Pond?

- A data pond is a collection of data puddles. It may be like a poorly designed data warehouse, which is effectively a collection of collocated data marts, or it may be an offload of an existing data warehouse.
- Data ponds limit data to only that needed by the project, and use that data only for the project that requires it.
- Given the high IT costs and limited data availability, data ponds do not really help us with the goals of democratizing data usage or driving self-service and data-driven decision making for business users.

Data Lakes and Data Oceans?

- A data lake is different from a data pond in two important ways. First, it supports self-service, where business users are able to find and use data sets that they want to use without having to rely on help from the IT department. Second, it aims to contain data that business users might possibly want even if there is no project requiring it at the time.
- A data ocean expands self-service data and data-driven decision making to all enterprise data, wherever it may be, regardless of whether it was loaded into the data lake or not.

Maturity growth from a puddle to a pond to a lake to an ocean

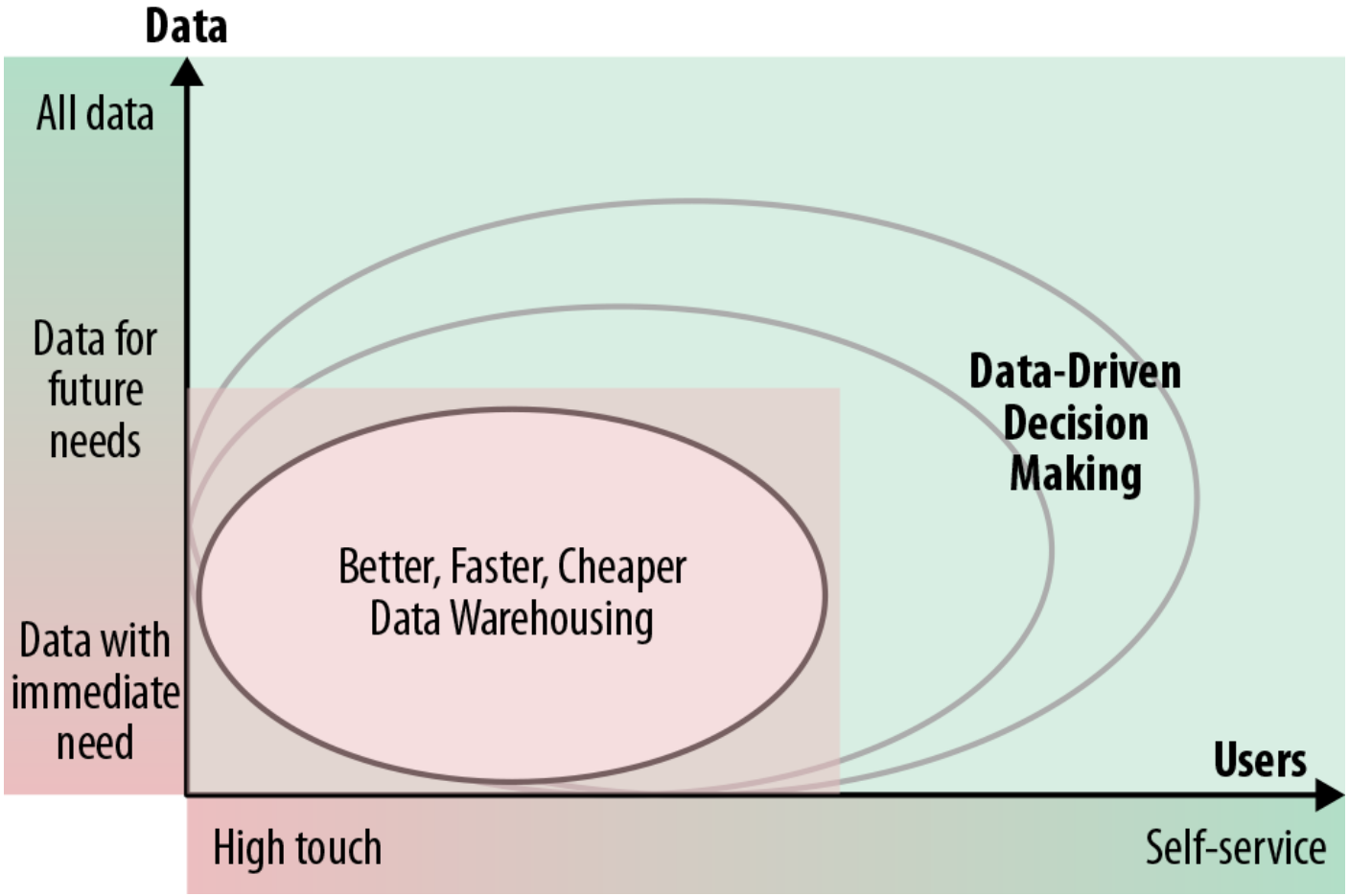


The key difference between the data pond and the data lake is the focus.

Data ponds provide a less expensive and more scalable technology alternative to existing relational data warehouses and data marts.

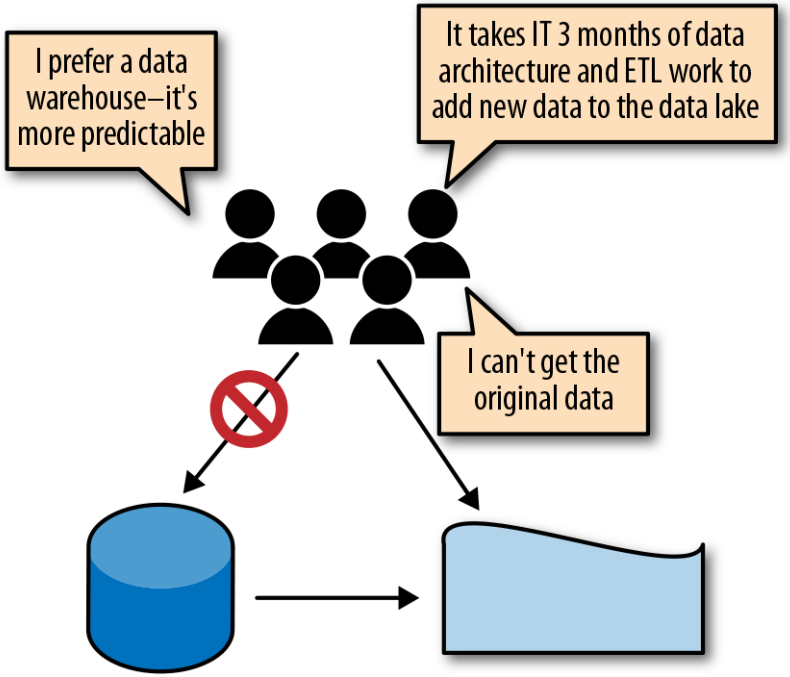
Whereas the latter are focused on running routine, production-ready queries, data lakes enable business users to leverage data to make their own decisions by doing ad hoc analysis and experimentation with a variety of new types o.f data and tools

Value proposition of the data lake



limitations of data ponds

ETL - Extract, transform, load. A three-phase process where data is extracted, transformed and loaded into an output data container.



Lack of:
Predictability,
Agility,
and Access to the Original Untreated Data

The Data Swamp

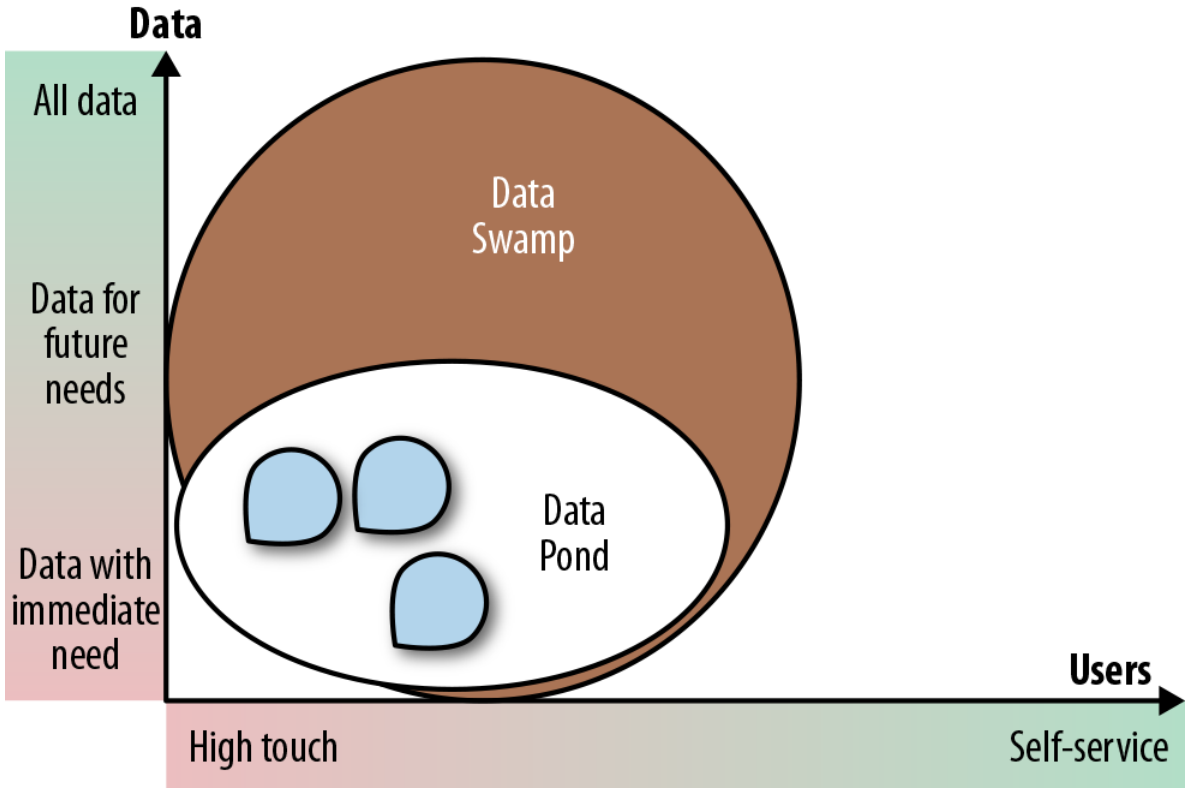
While data lakes always start out with good intentions, sometimes they take a wrong turn and end up as data swamps.

A data swamp is a data pond that has grown to the size of a data lake but failed to attract a wide analyst community, usually due to a lack of self-service and governance facilities.

At best, the data swamp is used like a data pond, and at worst it is not used at all.

Often, while various teams use small areas of the lake for their projects (the white data pond area in the image on the following slide), the majority of the data is dark, undocumented, and unusable.

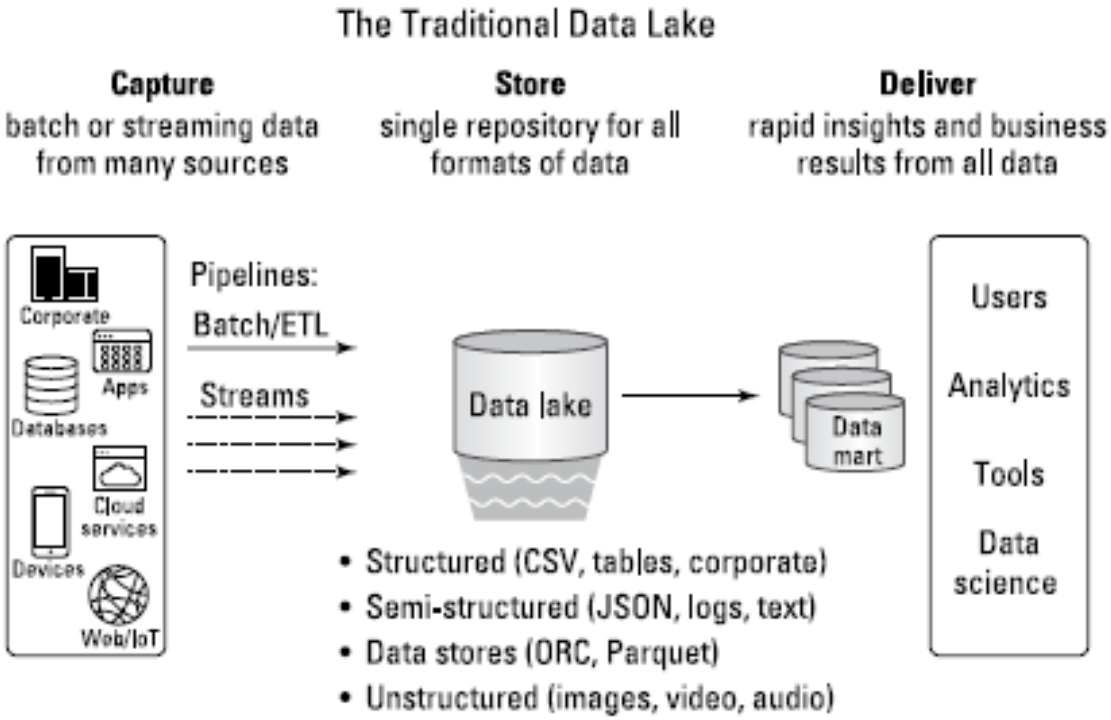
The Data Swamp



Creating a Successful Data Lake

- There are three key prerequisites
  - The right platform - like Hadoop, Amazon Web Services (AWS), Microsoft Azure, and Google Cloud.
    - Volume. These were designed to scale out—i.e. to scale indefinitely without any significant degradation in performance.
    - Cost. Usually at one-tenth to one-hundredth the cost of a commercial relational database
    - Variety. need to know its schema, but that's only when you use the data. This approach is called schema on read and it's one of the important advantages of big data platforms, enabling what's called "frictionless ingestion."
    - Future Proofing. Because big data technology is evolving rapidly, this gives people confidence that any future projects will still be able to access the data in the data lake.
  - The right data - save as much data as possible in its native format.
  - The right interfaces - providing data at the right level of expertise for the users, and ensuring the users are able to find the right data.

A Traditional Data Lake Structure



UWE Bristol University of the West of England

Some extra resources/links for further reading

https://cloud.google.com/learn/what-is-big-data

https://www.oracle.com/uk/big-data/what-is-big-data/

https://www.ibm.com/think/topics/big-data-analytics

https://cloud.google.com/learn/what-is-a-data-lake

https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake

https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction