Data Flow, Data Warehouses, and Data Lakes

Justin Post

Data Flow, Data Warehouses, and Data Lakes

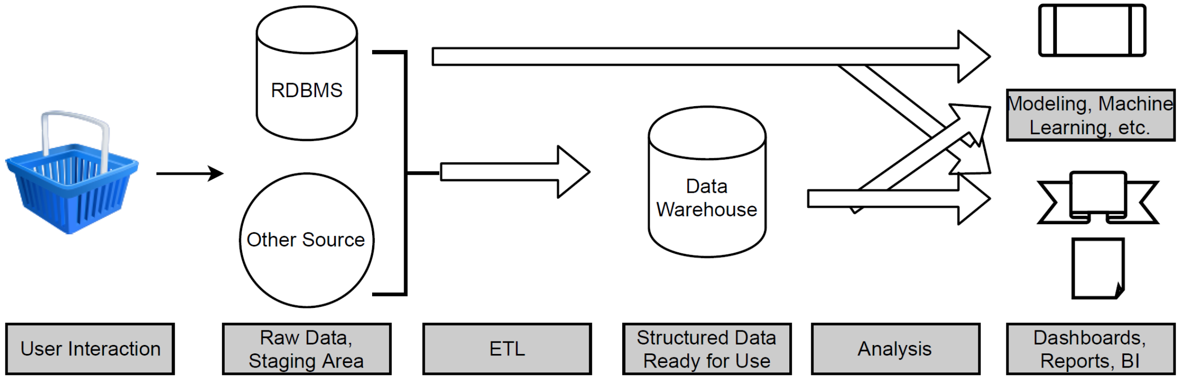
Justin Post

Databases

- We've discussed the use of a relational database (data, management system, and applications associated)
- The term database is really a bit more general
  - Object oriented databases
  - NoSQL databases
  - Cloud databases
  - Self-driving databases
- Data Warehouse

Data Flow (Non-big data)

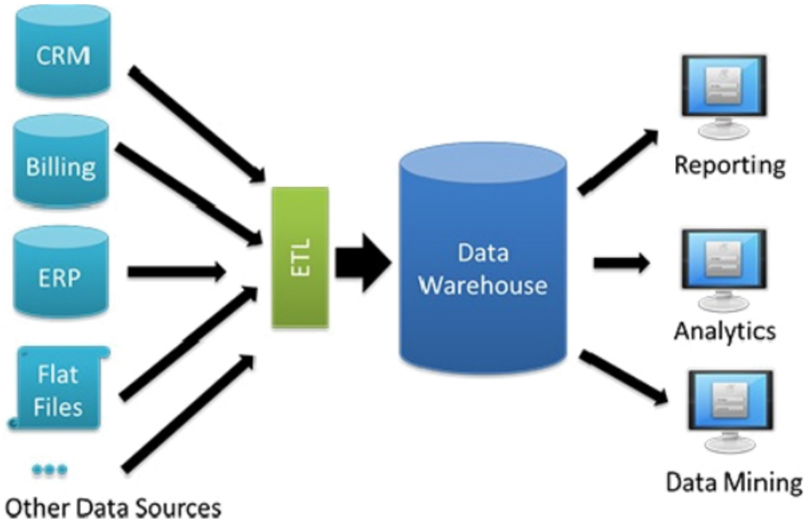
- As data comes in, it may be placed directly into a database (say RDBMS)
    - Highly structured schema, normalized data
- Or ETL (Extract, Transform, and Load) may be done and the data stored in a **data warehouse**
    - Structured schema, denormalized data ready for dashboards/analysis/etc.



Data Warehouse

Data warehouses are databases which are designed to:

- Store large amounts of data in a central database – and in a standard format.
- Integrate data from many different sources and standardize it, so it’s ready for analytics or reporting.
- Maintain historical records, since it can store months or even years of data.
- Keep data secure by storing it in a single location. Access can be granted only to those who need specific data.
- Provide quick, easy access to data to enable faster business decisions.



Databases vs Data Warehouses

Processing Types: OLAP vs OLTP

- Databases (like SQLite) use OnLine Transactional Processing (OLTP) to insert, replace, update, or delete records quickly
    - Optimized to add, modify, or delete records a lot

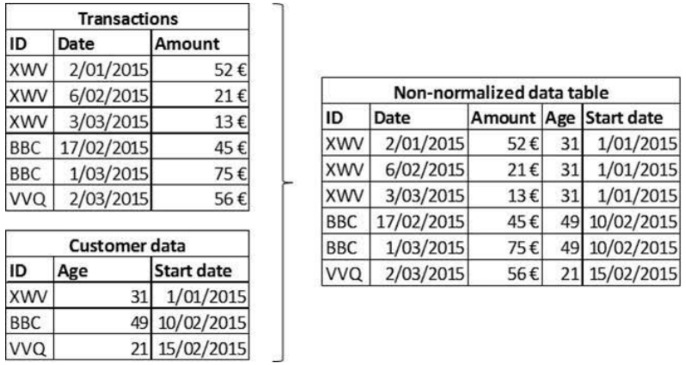
Databases vs Data Warehouses

Processing Types: OLAP vs OLTP

- Databases (like SQLite) use OnLine Transactional Processing (OLTP) to insert, replace, update, or delete records quickly
  - Optimized to add, modify, or delete records a lot
- Data Warehouses use OnLine Analytical Processing (OLAP) processing to analyze large amounts of data quickly
  - Optimized to exectue a smaller number of complex quieries

Databases vs Data Warehouses

- Databases often have data in a **normalized** format
  - Reduces redundancy and increases consistency as data isn't stored in multiple places
- Data Warehouses usually have **denormalized** that is ready to be analyzed
  - More query efficient, but data may exist in multiple places (and become inconsistent)



Transactions

ID	Date	Amount
XWV	2/01/2015	52 €
XWV	6/02/2015	21 €
XWV	3/03/2015	13 €
BBC	17/02/2015	45 €
BBC	1/03/2015	75 €
VVQ	2/03/2015	56 €

Customer data

ID	Age	Start date
XWV	31	1/01/2015
BBC	49	10/02/2015
VVQ	21	15/02/2015

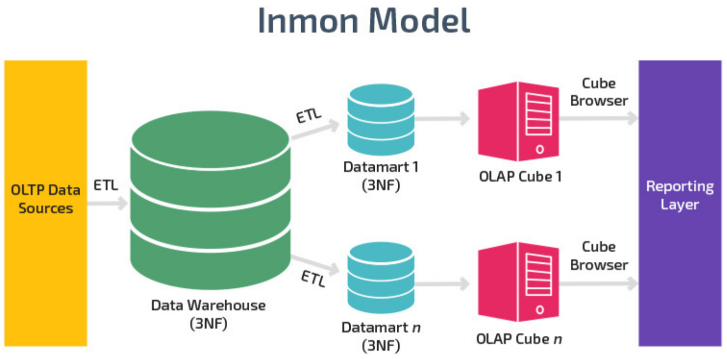
Non-normalized data table

ID	Date	Amount	Age	Start date
XWV	2/01/2015	52 €	31	1/01/2015
XWV	6/02/2015	21 €	31	1/01/2015
XWV	3/03/2015	13 €	31	1/01/2015
BBC	17/02/2015	45 €	49	10/02/2015
BBC	1/03/2015	75 €	49	10/02/2015
VVQ	2/03/2015	56 €	21	15/02/2015

https://bit.ly/3LGTnP0

Data Marts & MDM

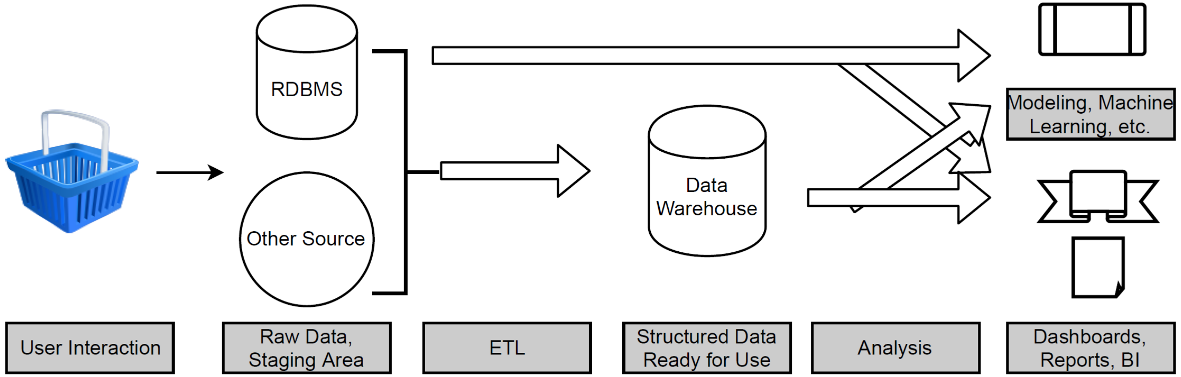
- Data Marts are focused versions of a data warehouse for special teams or departments
- You may also hear the term MDM (Master Data Management)
  - Another data source created that incorporates information about all **master** data sources
  - Provides a single consistent view of all business entities' information (a gold standard for their information)



https://bit.ly/3HUQwju

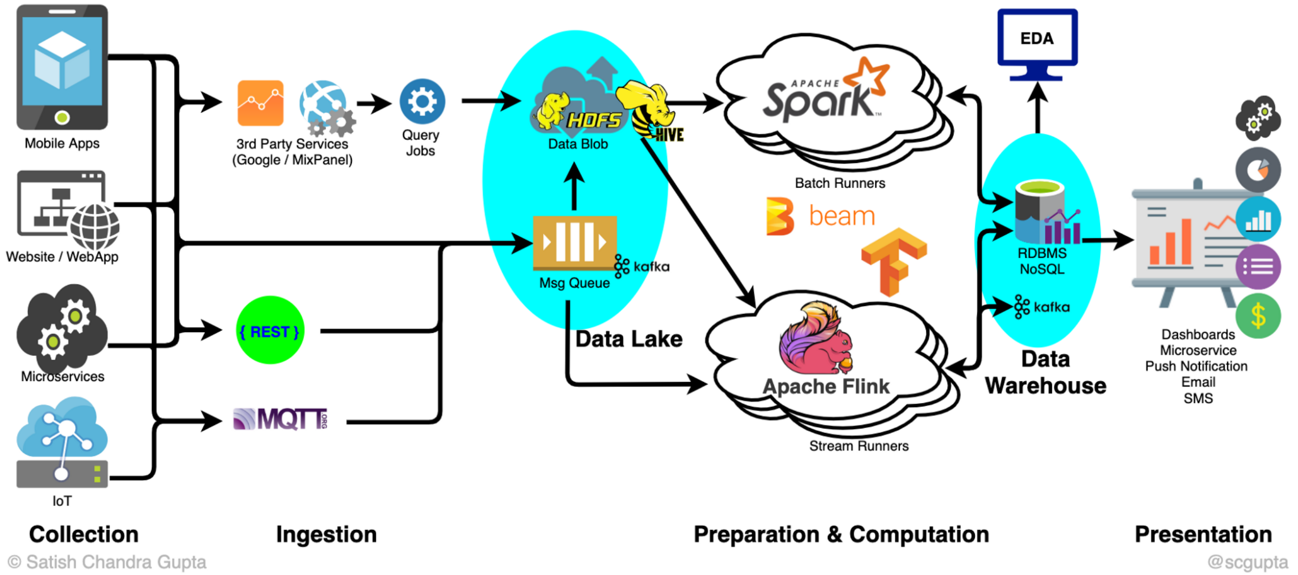
Data Flow (Non-big data)

- As data comes in, it may be placed directly into a database (say RDBMS)
  - Highly structured schema, normalized data
- Or ETL (Extract, Transform, and Load) may be done and the data stored in a **data warehouse**
  - Structured schema, denormalized data ready for dashboards/analysis/etc.



Data Flow (Big Data)

- (Often) All data stored in a stored in a **data lake**
  - Place for raw data to go until it is needed (schema is defined on read)
- ETL (Extract, Transform, and Load) is then done on the data to prepare it for use
  - Data may be placed into a database or a data warehouse within the data lake

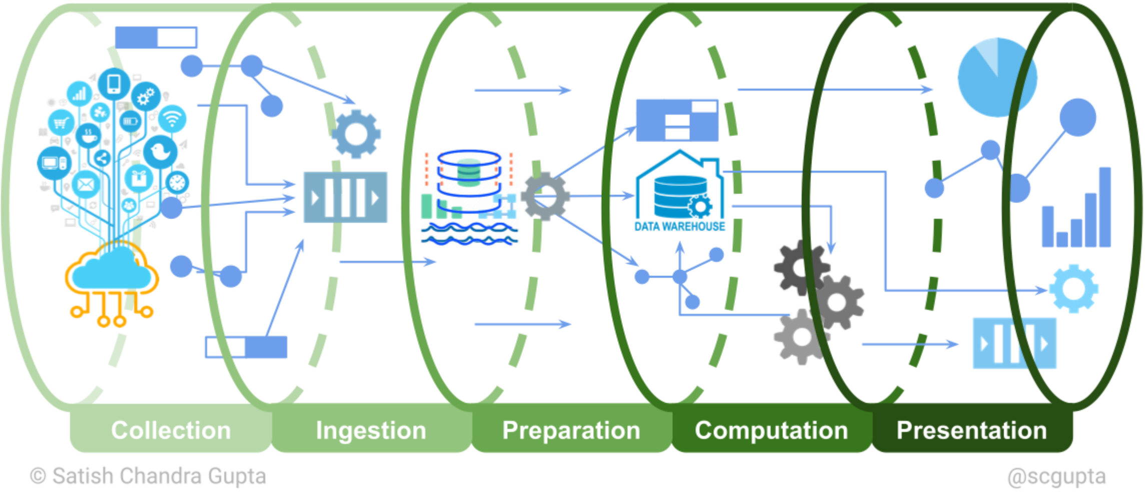


https://bit.ly/357wbcg

Data Lakes

A central repository to store all data in

- Can handle unstructured, semi-structured, or structured data
- Usually includes raw data and data after ETL
    - Raw data kept for long term archival and for data scientists to use



https://bit.ly/357wbcg

Data Sources

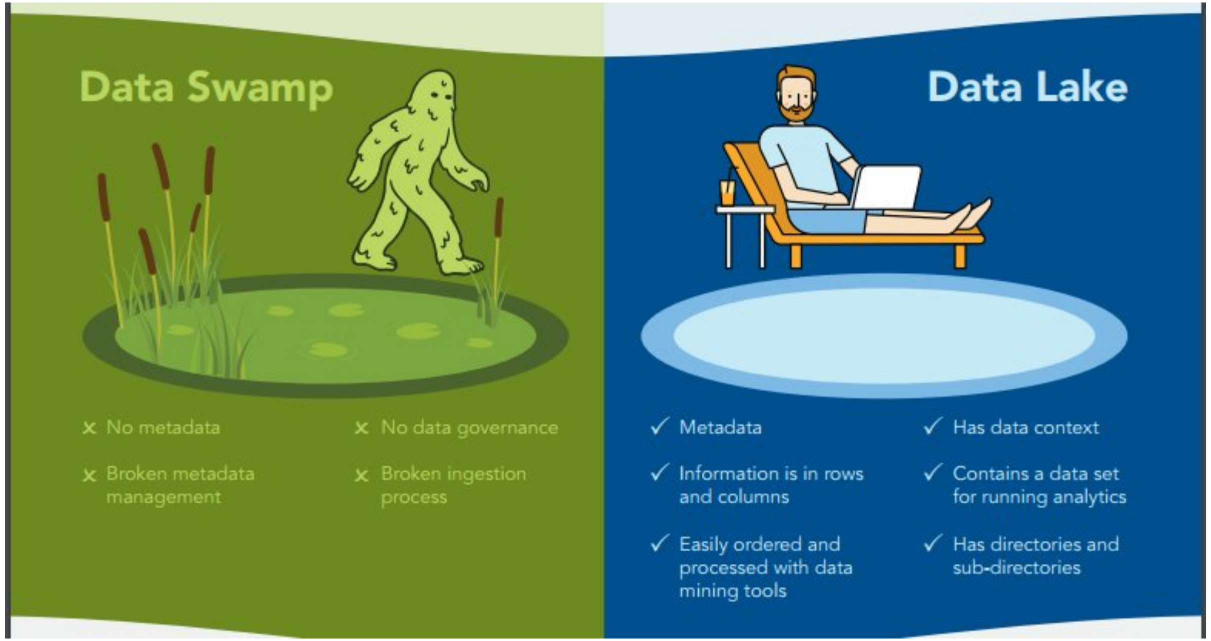
Data ingestion can happen via a batch or streaming process

- **batch**: data is updated in bulk
    - Say once each day at 3am
- **streaming**: data is updated in 'real-time'
    - Jobs run 24/7, waiting for new events to be published

Data Swamps

Flexibility of Data Lakes can also cause problems

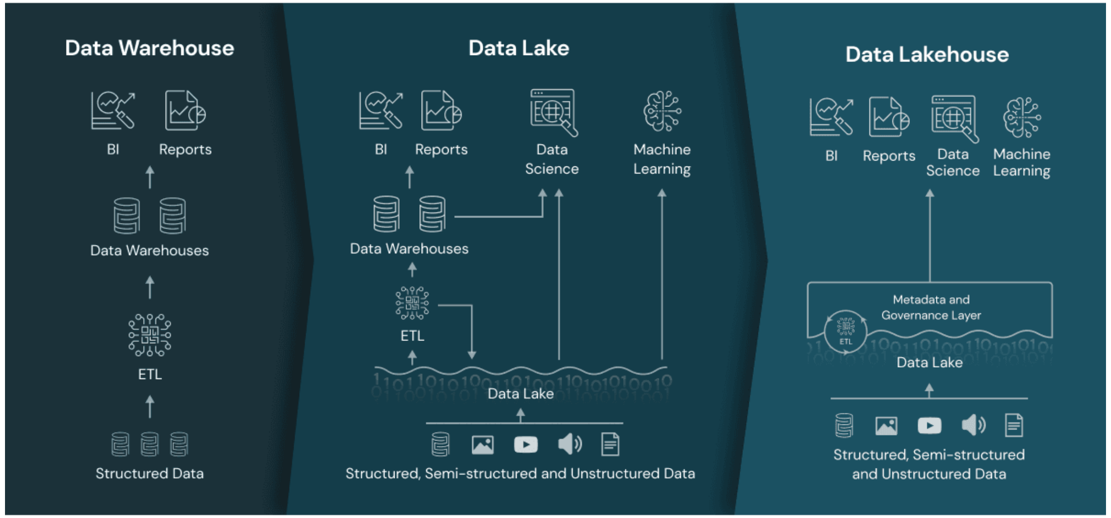
- **Data Swamp**: a data lake with poor data management
  - Data not well tagged or lacks structure


Data Swamp
✗ No metadata
✗ No data governance
✗ Broken metadata management
✗ Broken ingestion process
Data Lake
✓ Metadata
✓ Has data context
✓ Information is in rows and columns
✓ Contains a data set for running analytics
✓ Easily ordered and processed with data mining tools
✓ Has directories and sub-directories

https://bit.ly/3BuHuHl

Lake House

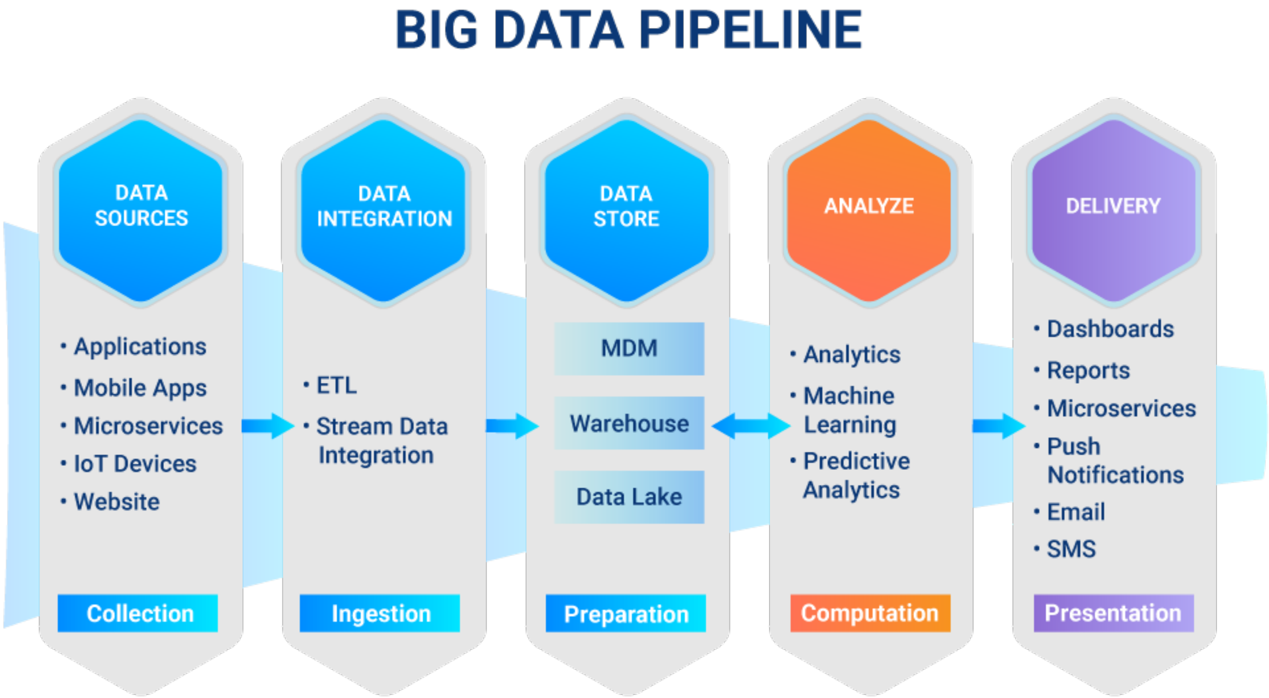
- **Lake House**: an intermediary between the unstructured data lake and the very structured database/data warehouse
- Delta Lake storage technology can power the lake house
  - Guarantees ACID transactions



https://bit.ly/3sKOFYa

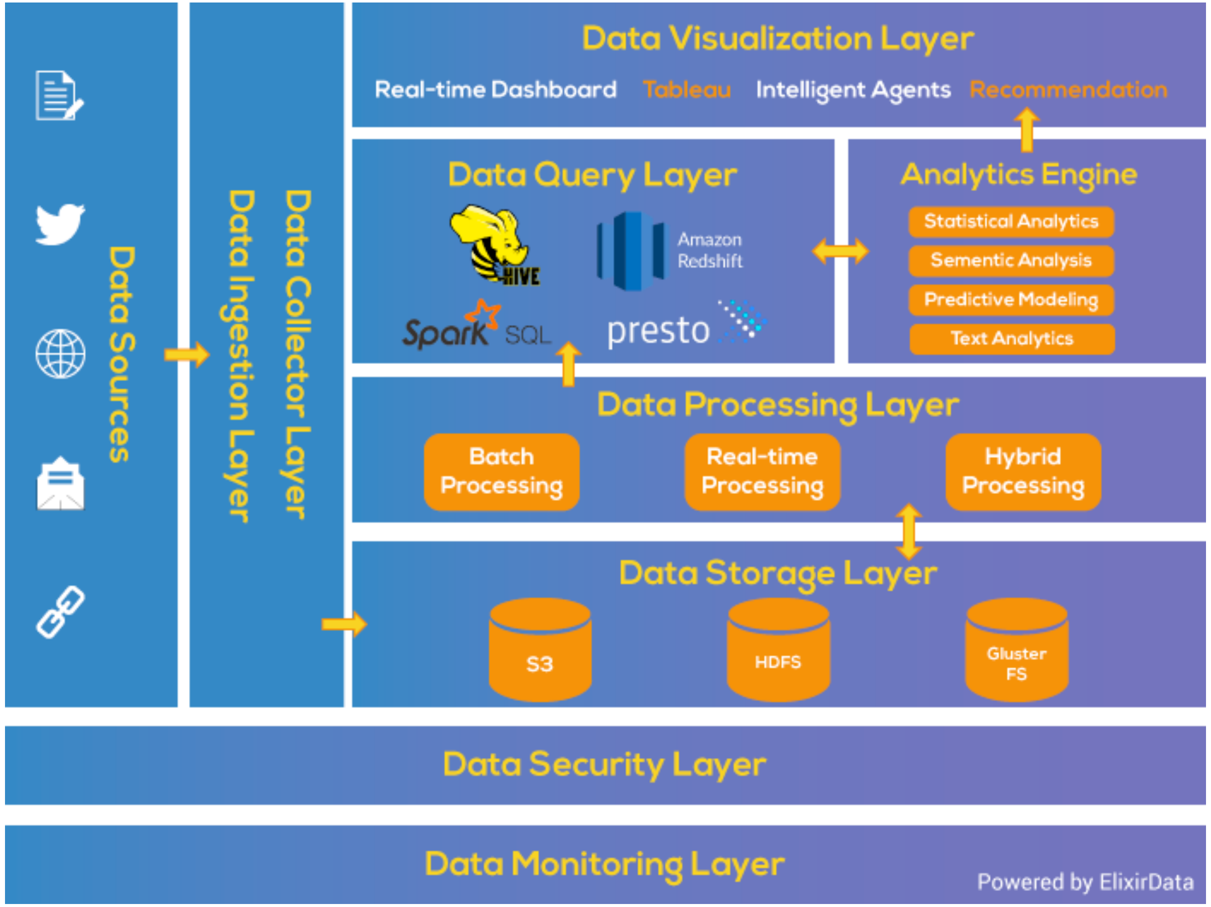
Big Data Flow

- A lot to manage in the data pipeline!



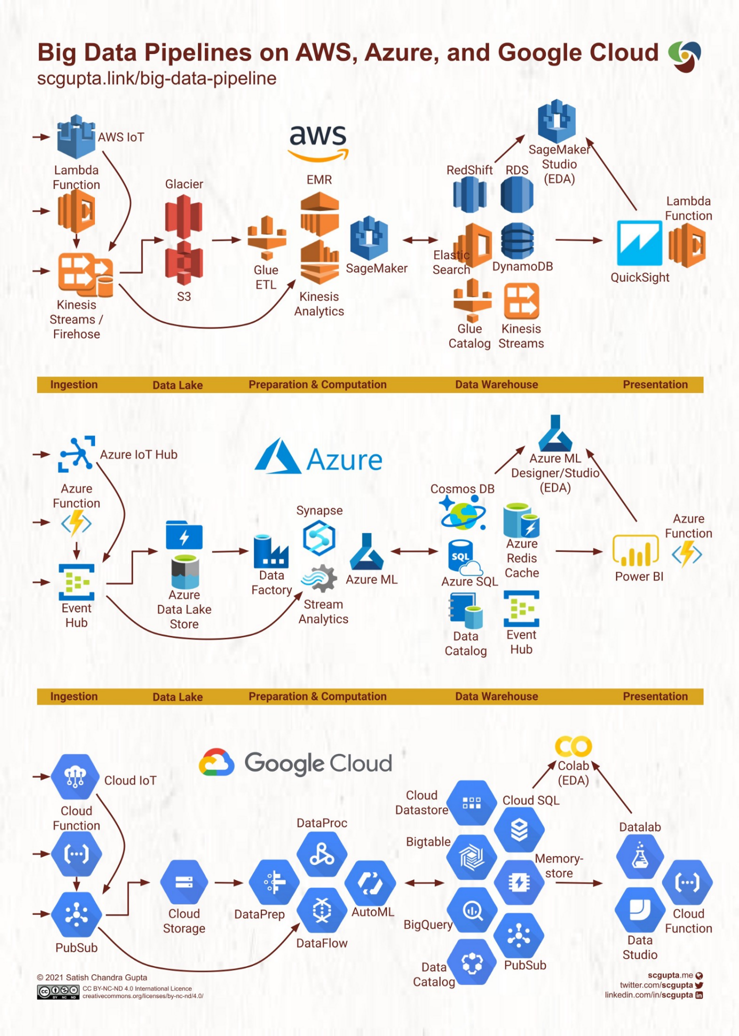
https://bit.ly/357wbcg

Another Flow Chart



https://bit.ly/36kiG9E

Companies that Manage the Process



Big Data Pipelines on AWS, Azure, and Google Cloud

scgupta.link/big-data-pipeline

scgupta.me
twitter.com/scgupta
linkedin.com/in/scgupta

Open Source Tools

- Also a host of open source tools that require management (and you can mix and match)
- Maybe feeling a little less overwhelmed?

Recap

- Important to understand the basic data pipeline (big data and non-big data)
- Data lakes, data warehouses, data marts, MDM, and lake houses
- Lots of competing options
    - Open source and company managed
    - Cloud managed