**Fundamental Concepts:**

Data Warehouse & Lakes:

Data Warehouse:

* Structured and/or processed.
* Ready to use
* Rigid – structures are hard to change

Data Lake:

* All raw and/or unstructured.
* Ready to analyze
* More flexible – no structure enforcement.

Data Warehouse & Lakes are just abstract definition to think about when designing data systems.

OLTP vs OLAP:

* On-line Transactional Processing vs On-Line Analytical Processing.

OLTP:

* + Processes a high volume of short transactions. – e.g SELECT, INSERT statements in basic SQL terms.
  + Queries should be fast.
  + System Should Maintain a High Level of Data Integrity.
  + Example: A DB that records purchase tx is an OLTP

OLAP:

* Low volume of long-running queries.
* Queries may also be run over aggregated historical data.
* E.g System could be used to produce purchases or purchasing trends over time.

\*OLTP Systems Modify DB ; OLAP systems query DBs.

\*ETL – contain data from OLTP system & move it into OLAP system.

SQL vs NoSQL:

* NoSQL – catch all term for everything that is not SQL.
* NoSQL DBs include: Key Value Stores, JSON Document Stores e.tc
* NoSQL examples – MongoDB, Apache Cassandra.

Batch & Streaming: - Ways to ingest data into a System:

Batch:

* Data Gathered within a defined window of time.
* Large volumes of data.
* Mainly used when importing data from legacy systems.

Streaming:

* Continuous collection of data.
* Collected as it happens.
* It’s possible to stagger the input of streaming data with time windows & micro-batches.
* Can be used for near real-time analytics.

**Data Processing Pipelines:**

* Built by Data Engineer to enable data to get from A-B
* 4 main stages to a data pipeline – Ingestion, Storage, Processing, Visualization.
* Ingestion – move data into our processing system.
  + Examples of Data Sources: Archive Data, Log Collection, Device Data
  + Technical Challenges: Big Data is Big: - Ingestion Method (HTTP/FTP), Storage Options; Data should have value, security implications of Data Processing Designs,
* Processing: ETL
  + ETL- Data is taken from a source and manipulated to fit the parameters of the destination system.
  + ELT – data is loaded into a data lake and transformation take place later
  + Common Transformations:
    - Formatting – to match destination systems
    - Labeling
    - Filtering -
    - Validating – it meets certain requirements.
* Visualization – using various dashboards tool.