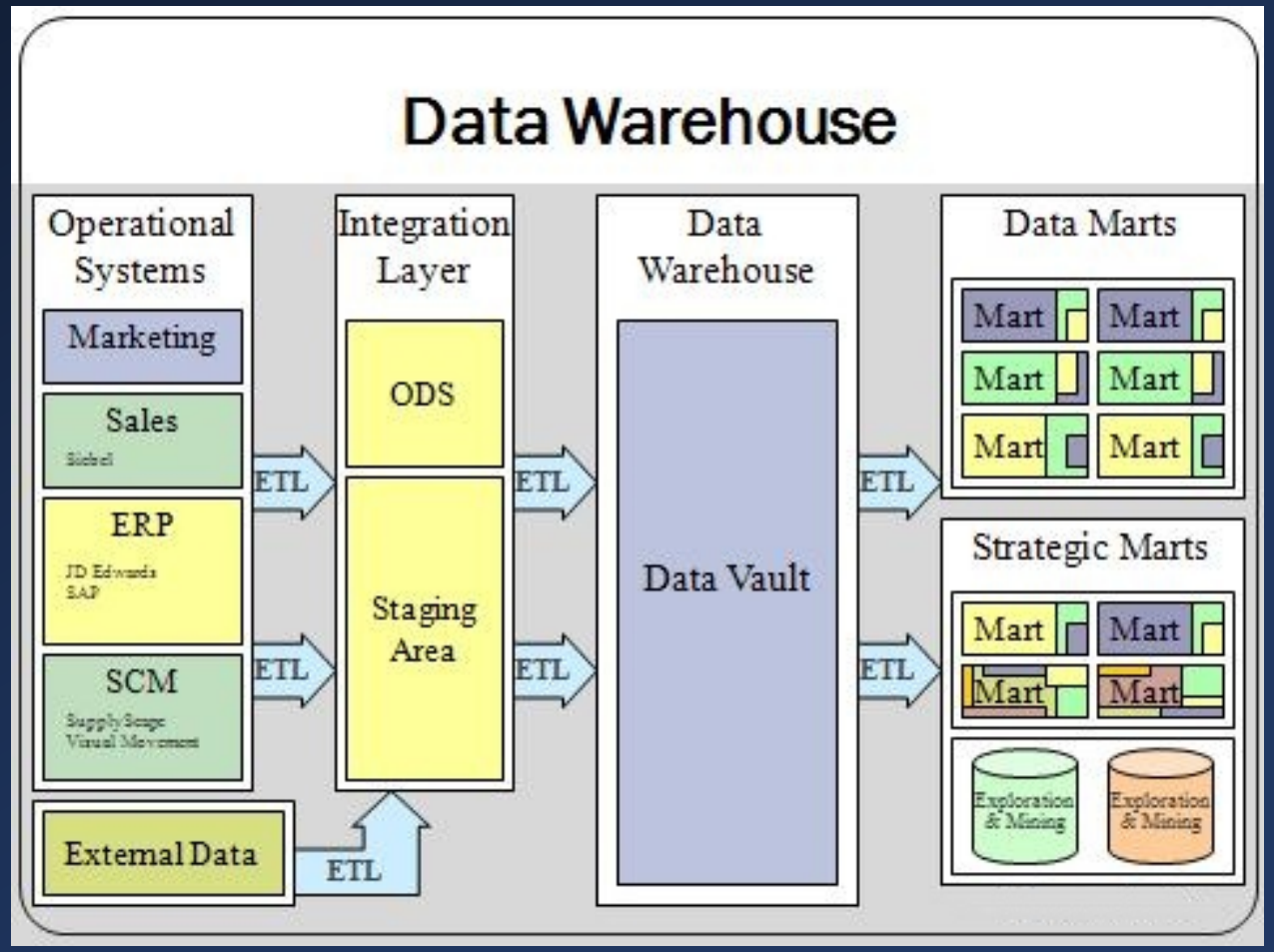
Big Data (MHI222956/MHI225101)

5.1 Big Data Integration & Processing Pipeline

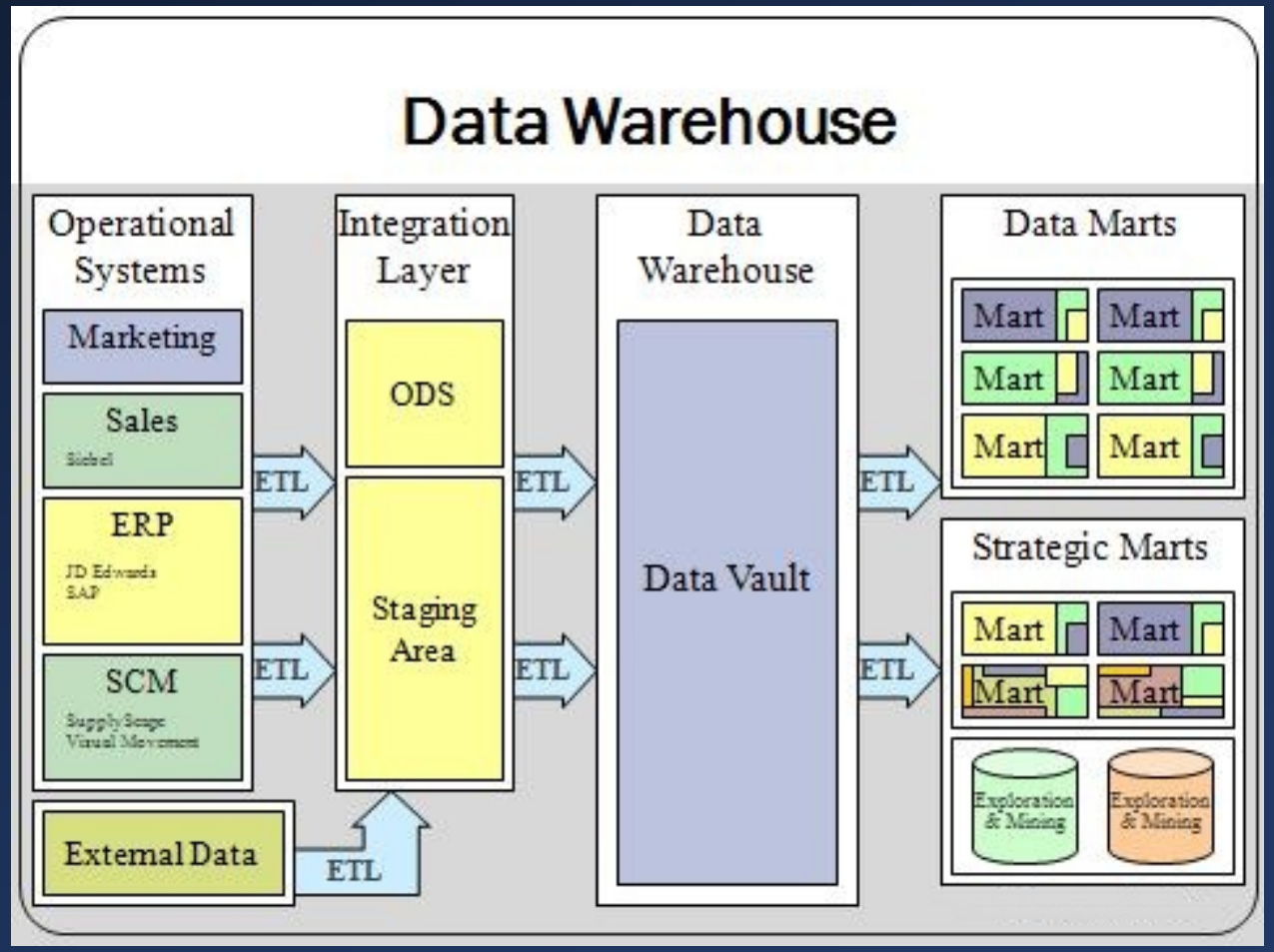
Data Warehouses

- A system used for reporting and data analysis
- Core component of business intelligence
- Central repository of integrated data from one or more disparate sources



ETL (Extract, Transform, Load)

- The general procedure of copying data from one or more sources into a destination system
- Data is represented differently from the source(s) or in a different context than the source(s)
- A popular concept in the 1970s

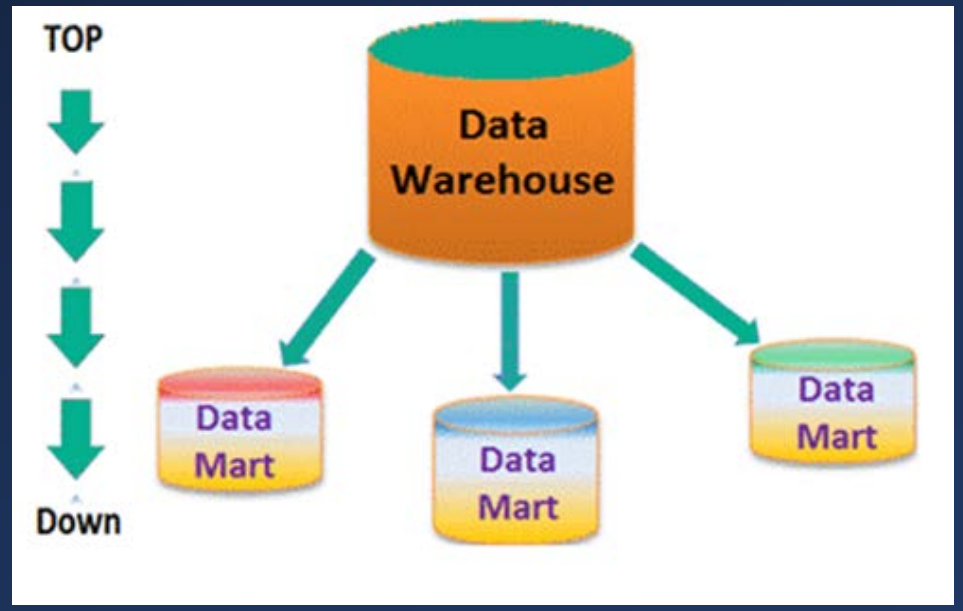


- Data mart

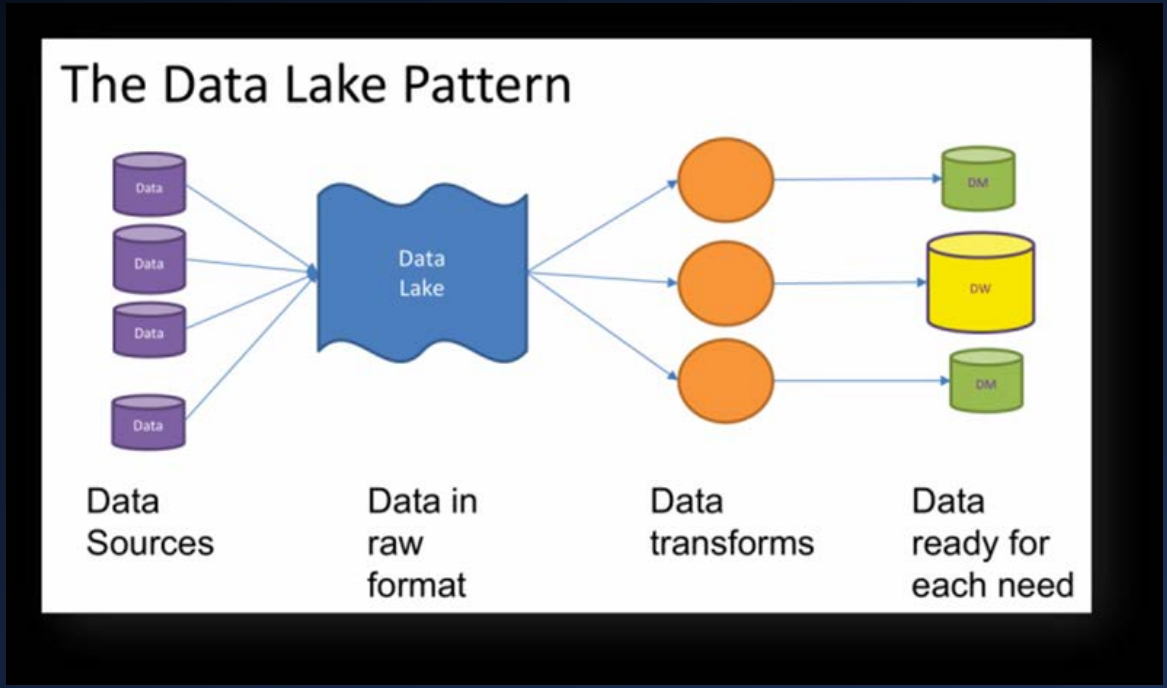
  A simple form of a data warehouse that is Focused on a single subject (or functional area)

- Difference between Data warehouse and data Mart:

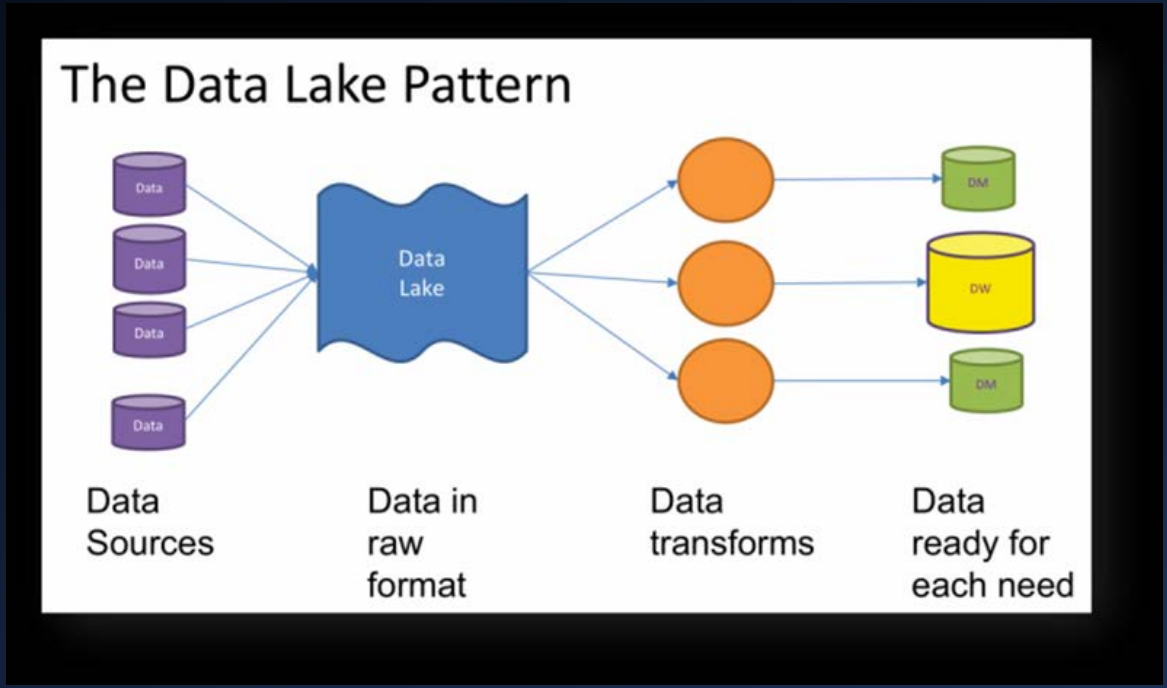
  https://www.guru99.com/data-warehouse-vs-data-mart.html



Data Lakes



- A storage repository of data holds a vast mount of raw data in its native format until it is needed: there is no hierarchy or organization among the individual pieces of data.
- Accepts and retains all data from all data sources: can include structured data from relational databased, semi-structured data (e.g., CSV, XML, JSON) and unstructured data.



ELT (Extract, Load, Transform)

- An alternative to ETL used with data warehouse implementations
- In ELT models the data is not processed on entry to the data lake which enables faster loading times.

- *“If you think of a data mart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”*
- *---- By: James Dixon*
- Data swamp
  - Highly disorganised data repository
  - A data lake equips companies to retrieve and use their data effectively. But, data swamps can make both those tasks exceptionally difficult and perhaps impossible.

Data Lakes vs Data Warehouses

Data Lake	Data Warehouse
Reason for storing data is undefined	Reason fro storing data is pre-defined
Data is left raw until it is needed	Data is processed and ready to be queried
Schema-on-Read	Schema-on-Write
Used by data scientist	Used by business professionals
Emerging technology	Strong maturity model
Adapt easily to changes	Difficult to change the structure

“Data science starts with data. So, your big data initiative needs all the right pieces: the *data lake*, the *analytic applications*, and likely an multi-cloud *architecture*. All that remains is *integrating* with lightning-fast, scalable, and elastic connections. ”

Data Integration

- Data integration is a set of processes used to *retrieve* and *combine* data from disparate sources into meaningful and valuable information.

- Traditional data integration techniques was mainly based on ETL (extract, transform and Load) process to ingest and clean data then load it into a data warehouse.

- Traditionally, ETL has been used with batch processing (data on the rest) in data warehouse environments.

Big Data Integration

Big Data Integration

Big data integration can be done in Real-time or with batch processing. Which make the ETL phases reordered to become ELT in some cases, so the data is extracted, loaded into distributed file systems, and then transformed before being used.

Big Data Integration

Three basic techniques:

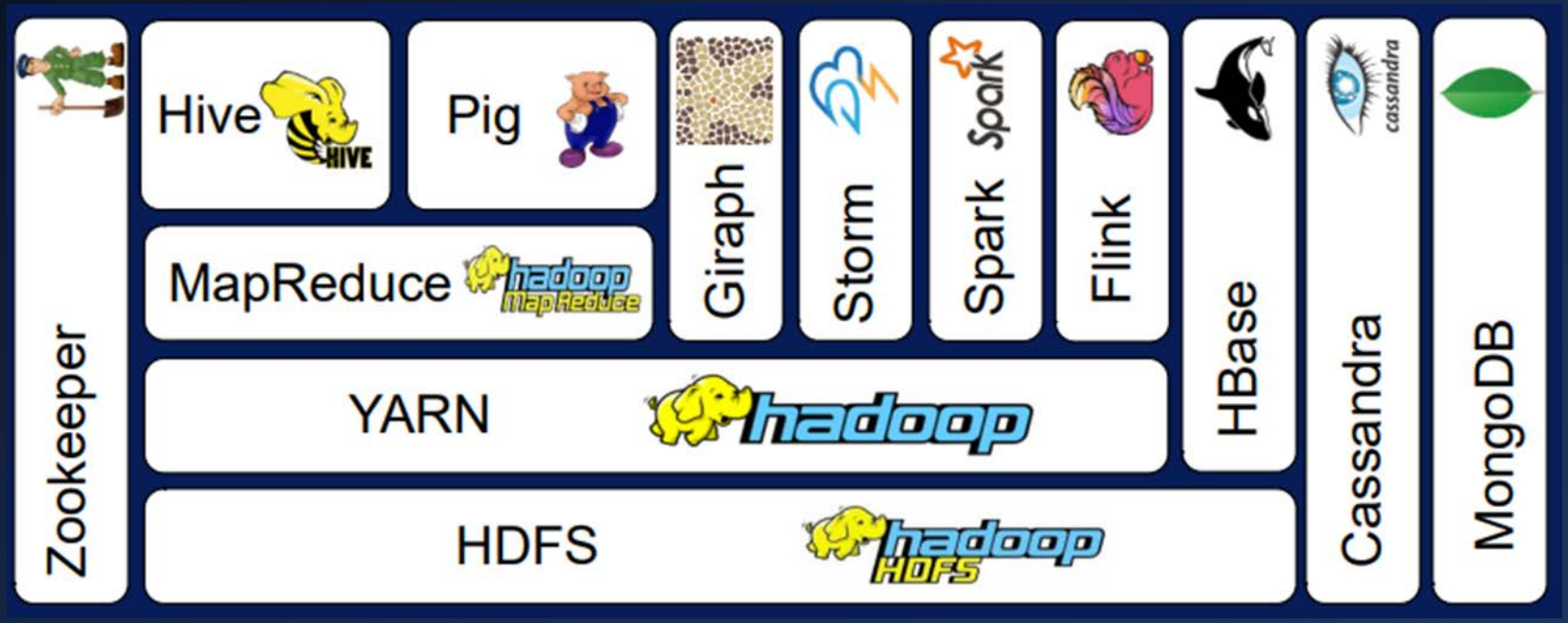
- Schema Mapping
  - First, creating a mediated (global) schema that are most relevant to your business
  - Then, identifying the mappings between the mediated schema and the local schema of the data sources to determine which (sets of) attributes contain the same information.
- Record Linkage
  - Identify records that refer to the same logical entity across different data sources
  - Techniques used: Pairwise matching, Clustering, Blocking, etc.
- Data Fusion
  - A combination of techniques that aims to resolve conflicts from a collection of sources and to find the truth that reflects the real world.
  - It is a new field that has emerged recently and motivated by the veracity of data
  - Techniques used: Copy detection, Voting and Source quality.

Evolution of Hadoop

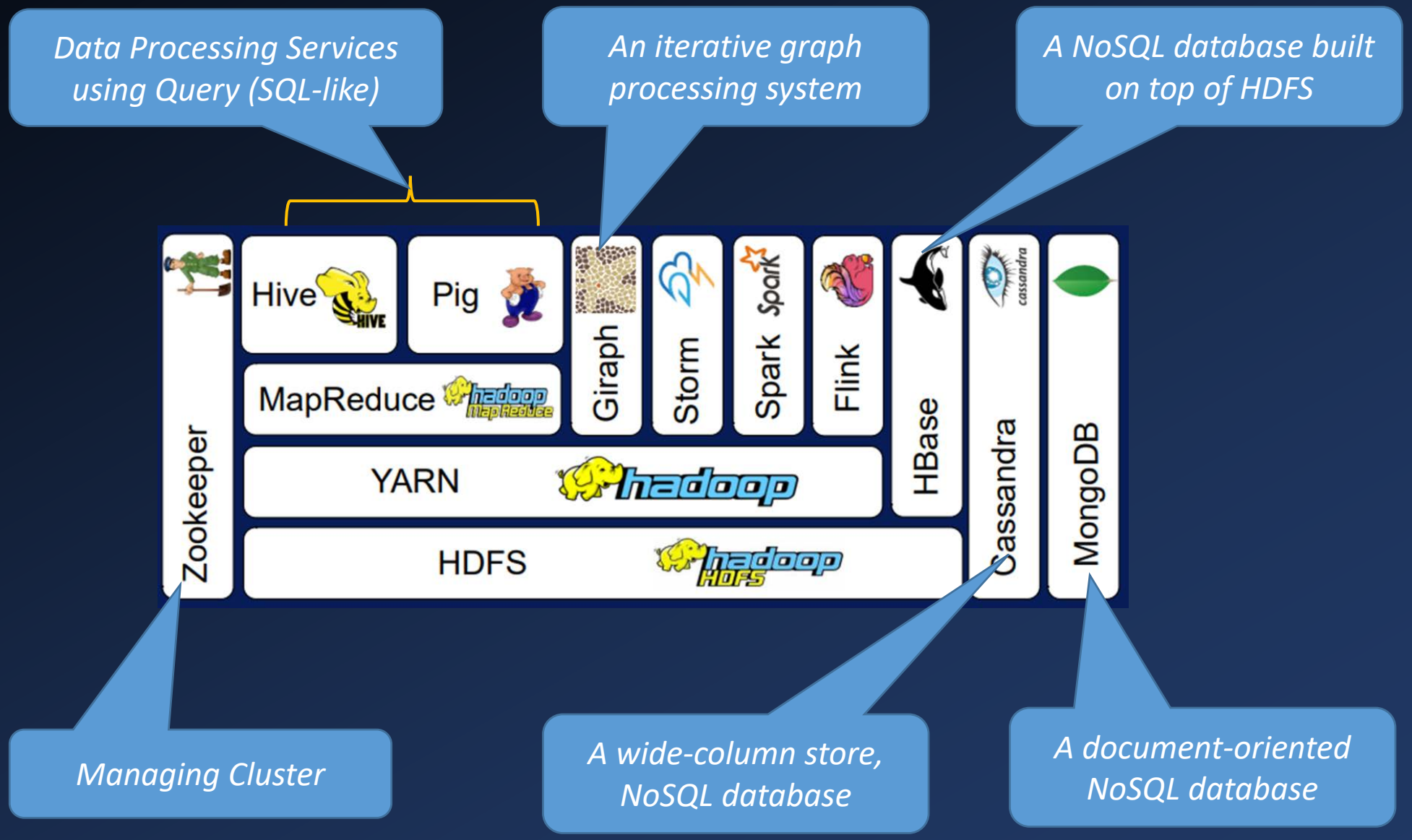
- In Oct 2003 – Google releases papers with GFS (Google's distributed File System).
- In Dec 2004, Google releases papers with MapReduce.
- In 2005, Nutch used GFS and MapReduce to perform operations.
- In 2006, Yahoo created Hadoop based on GFS and MapReduce with Doug Cutting and team.
- In Jan 2008, Yahoo released Hadoop as an open source project to Apache Software Foundation.

- Later More big data frameworks released, now there is over a 100!

The Hadoop Ecosystem



Hadoop Ecosystem is a platform or framework which solves big data problems. You can consider it as a suite which encompasses a number of services (ingesting, storing, analyzing and maintaining) inside it.


Data Processing Services using Query (SQL-like)
An iterative graph processing system
A NoSQL database built on top of HDFS
Hive
Pig
MapReduce
Giraph
Storm
Spark
Flink
HBase
Cassandra
MongoDB
Zookeeper
YARN
HDFS
Managing Cluster
A wide-column store, NoSQL database
A document-oriented NoSQL database

MapReduce is a programming model that simplifies parallel computing

MapReduce

YARN (Yet Another Resource Negotiator) provides flexible scheduling and resource management over the HDFS storage.

YARN

HDFS provides scalable and reliable storage. It is the foundation for many big data frameworks.

HDFS

MapReduce

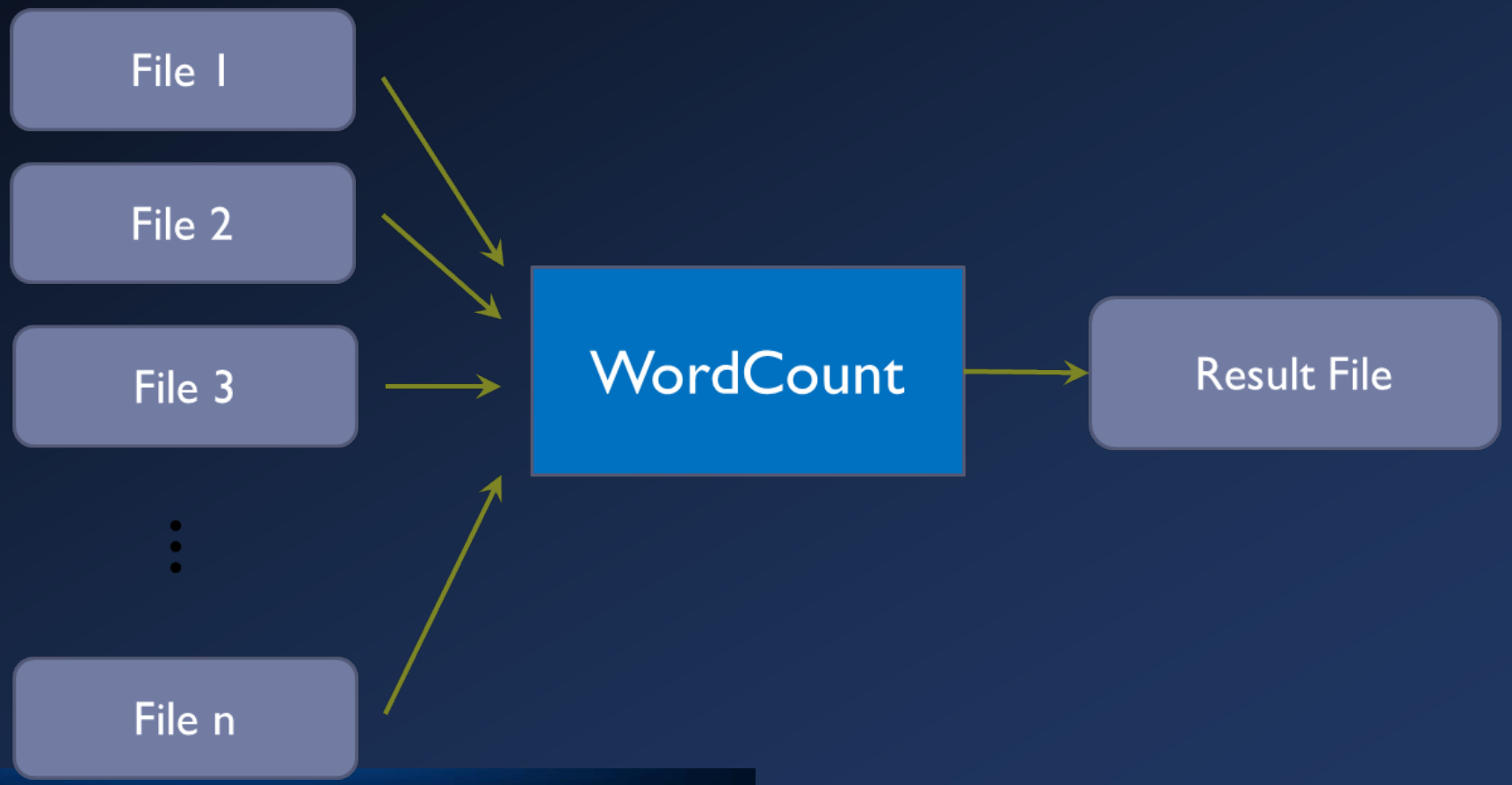
- Traditional parallel programming requires expertise on a number of computing and systems concepts.
  - e.g., synchronization mechanisms are essential
  - High learning curve
- The MapReduce programming model greatly simplifies running code in parallel
  - only need to create and map and reduce tasks
  - don't have to worry about multiple threads, synchronization, or concurrency issues.

MapReduce = Only Map and Reduce!

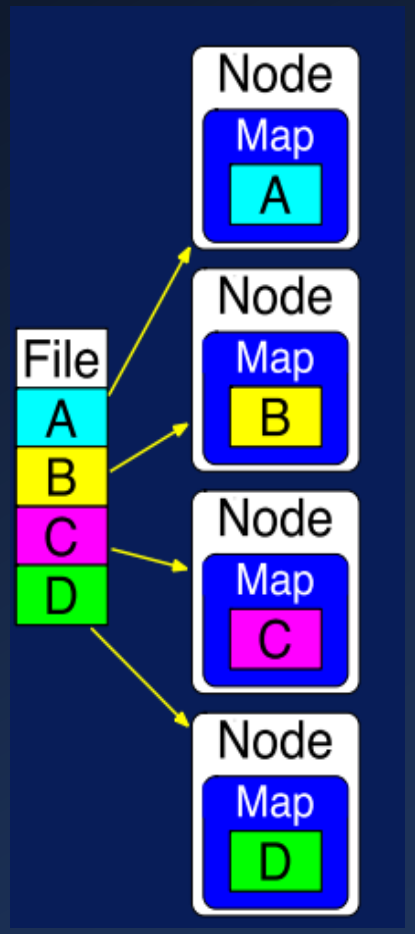
Map = apply operation to all elements

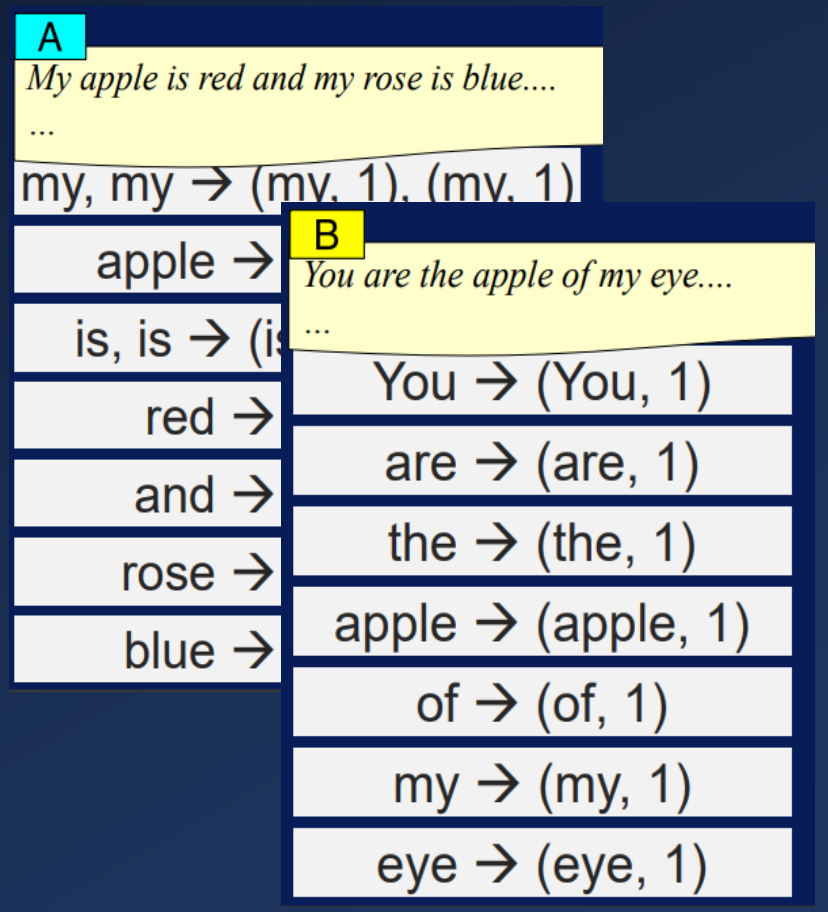
Reduce = summarize operation on elements

The “Hello World” of MapReduce



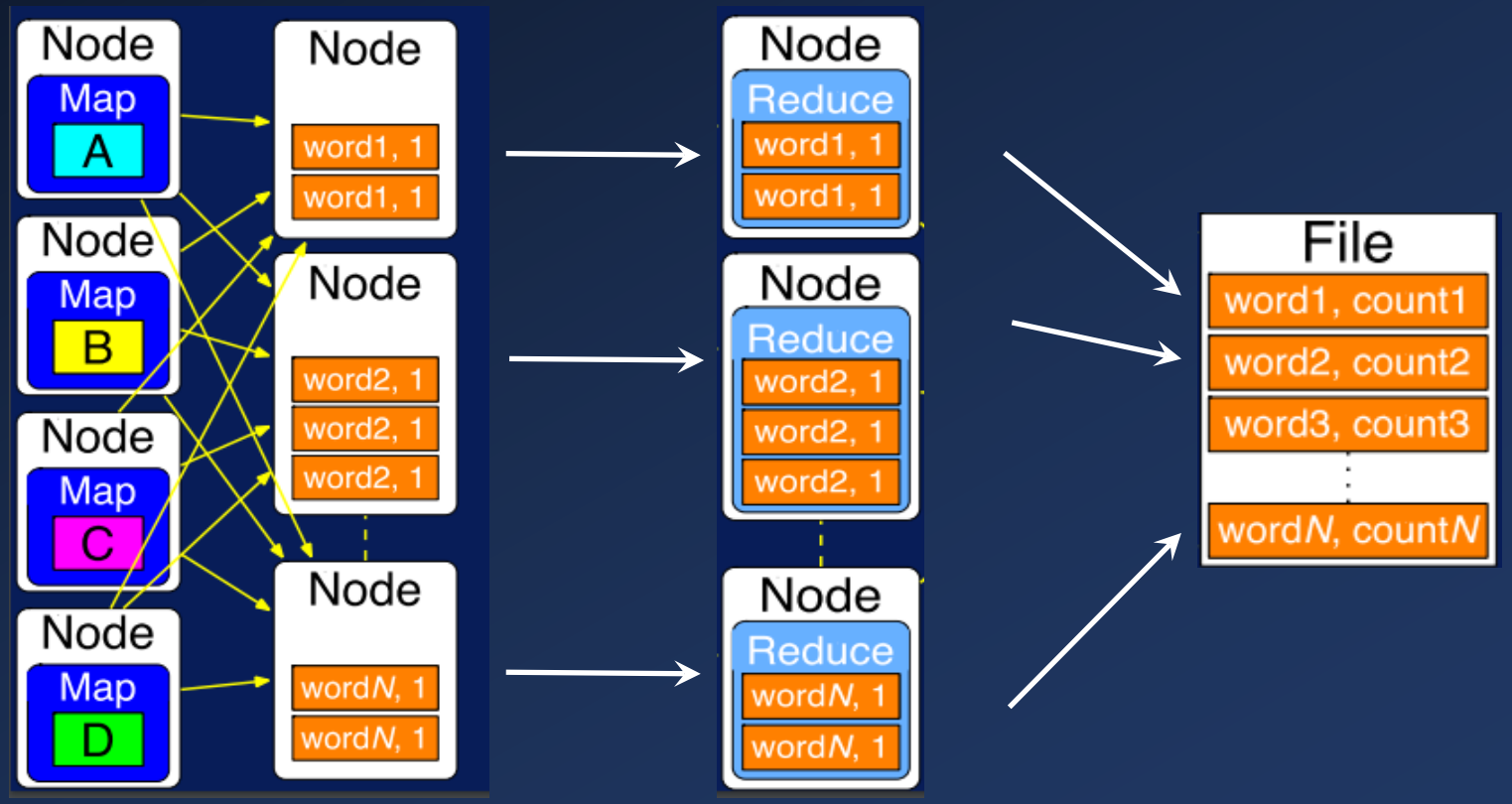
- Step 0: File is stored in HDFS
- Step 1: Map on each node – generate key-value pairs



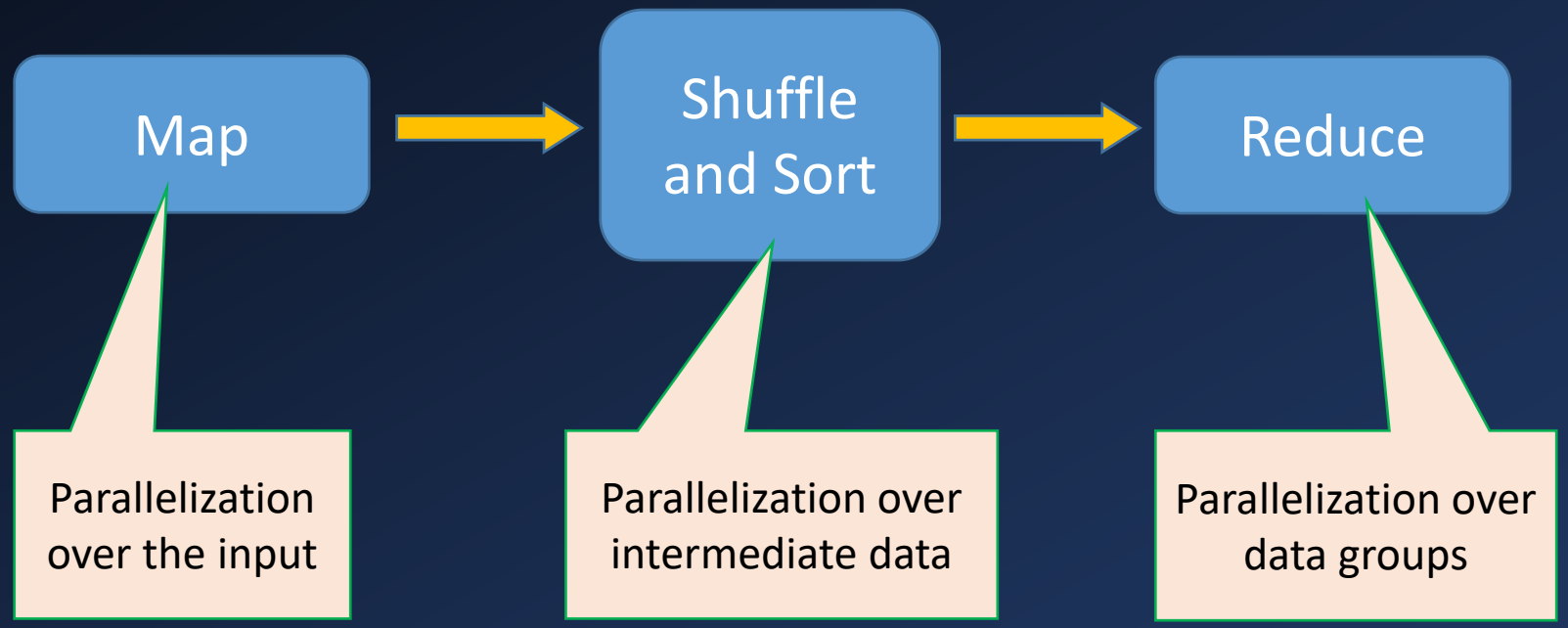


...

- Step 2: Sort and Shuffle – pairs with sane key moved to same node
- Step 3: Reduce – Add value for same keys



The word count application



MapReduce - Simplified parallel programming

Big Data Processing Pipeline

- Apply a specific function
- Work the data from one format to another
- Filter data values out of a data set
- Analytical operations -- analyze the data to discover meaningful trends and patterns, in order to gain insights into the problem being studied
  - Machine learning
  - Graph analytics

Summary

- Data Lake & Data Warehouse
- Data Integration
- Hadoop MapReduce
- Big Data Processing Pipeline