Data warehouse (IS 422)

Lecture 1

Introduction to course

Dr. Wael Abbas

2023 - 2024

Course Info.

- **Lectures:** Wednesday (8-10)

- **Instructor:** Dr. Wael Abass
- **Office hours**: Monday(10 - 2)
  Contact: WaelMohamed@fci.helwan.edu.eg

Assessment Scheme

- **Mid-Term Exam:** 20%
- **Section :** 20%
- **Final Exam:** 60%

Ground Rules

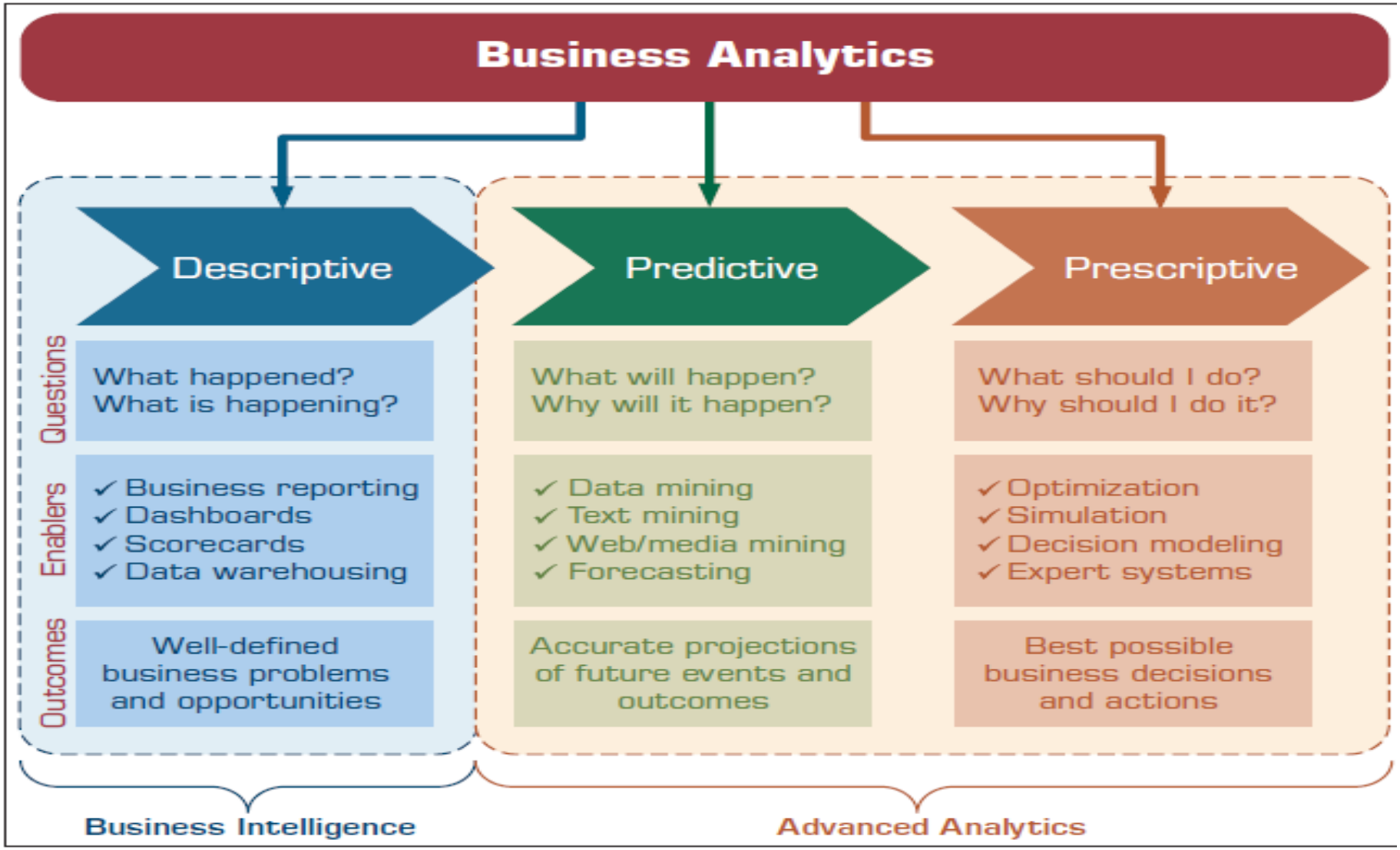
Course Syllabus

Week	Lecture	Lab
1	Introduction to course	
2	Introduction to data warehouse	ETL (SSIS)
3	DW Architectures (part 1)	
4	DW Architectures (part 2)	
5	Dimensional modeling 1	
6	Dimensional modeling 2	Power BI
7	Midterm	
8	Dimensional modeling 3	
9	Extract transform load (ETL)	
10	OLAP	
11	Introduction to Business intelligence	

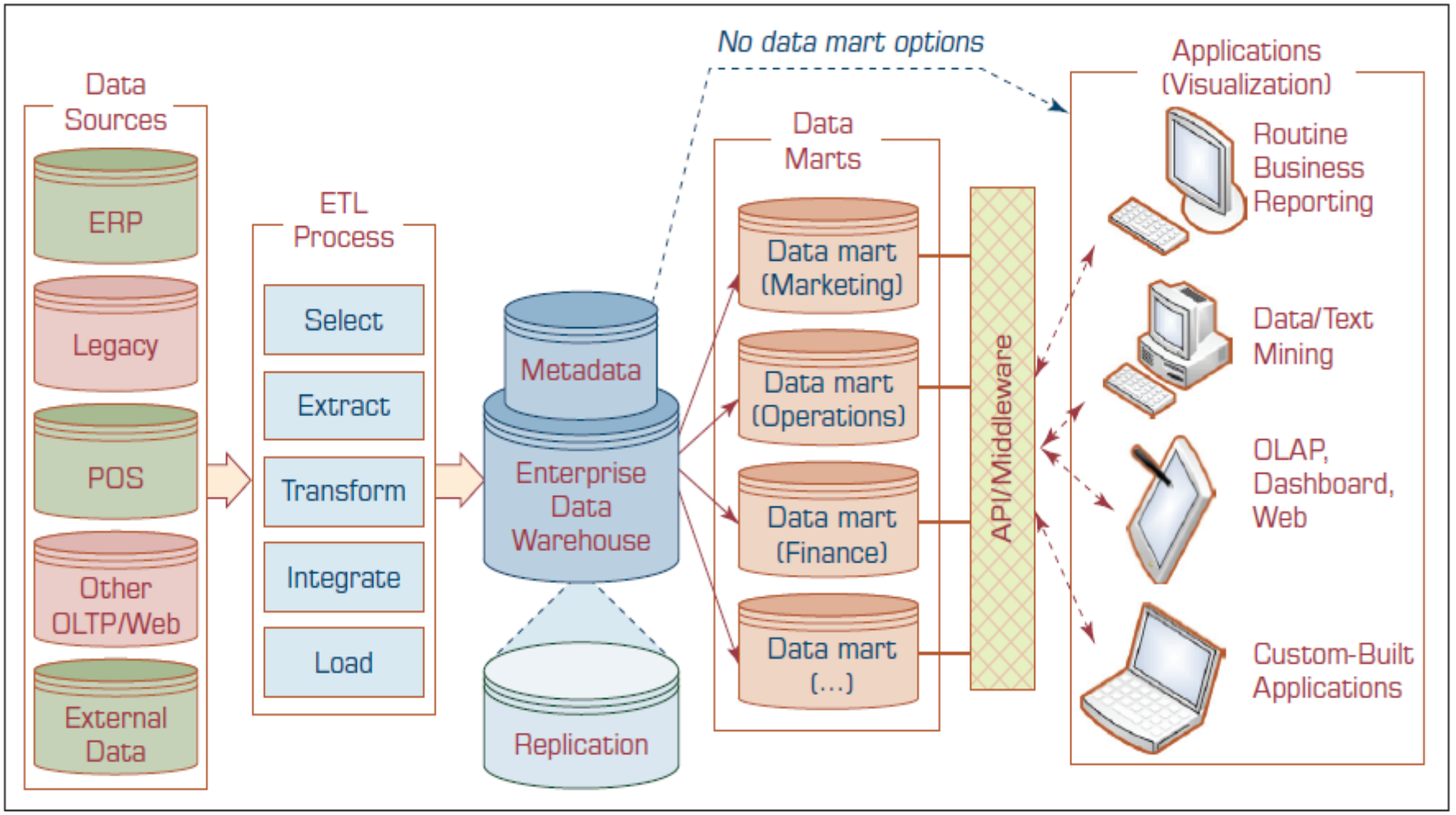
Text books

- W. H. Inmon, **Building the Data Warehouse (Fourth Edition),** John Wiley & Sons Inc., NY.

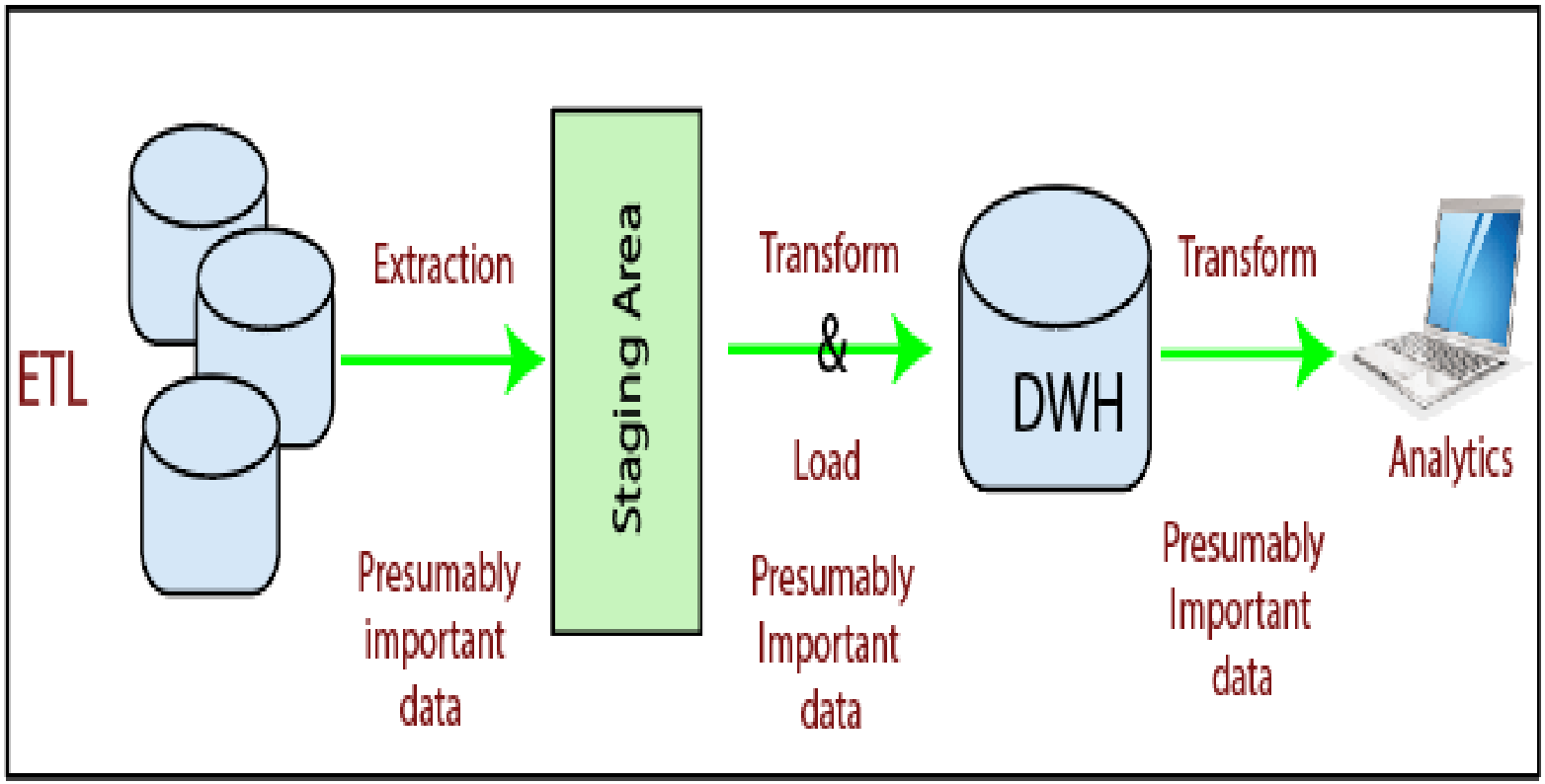
Relationship between BA, BI , DW



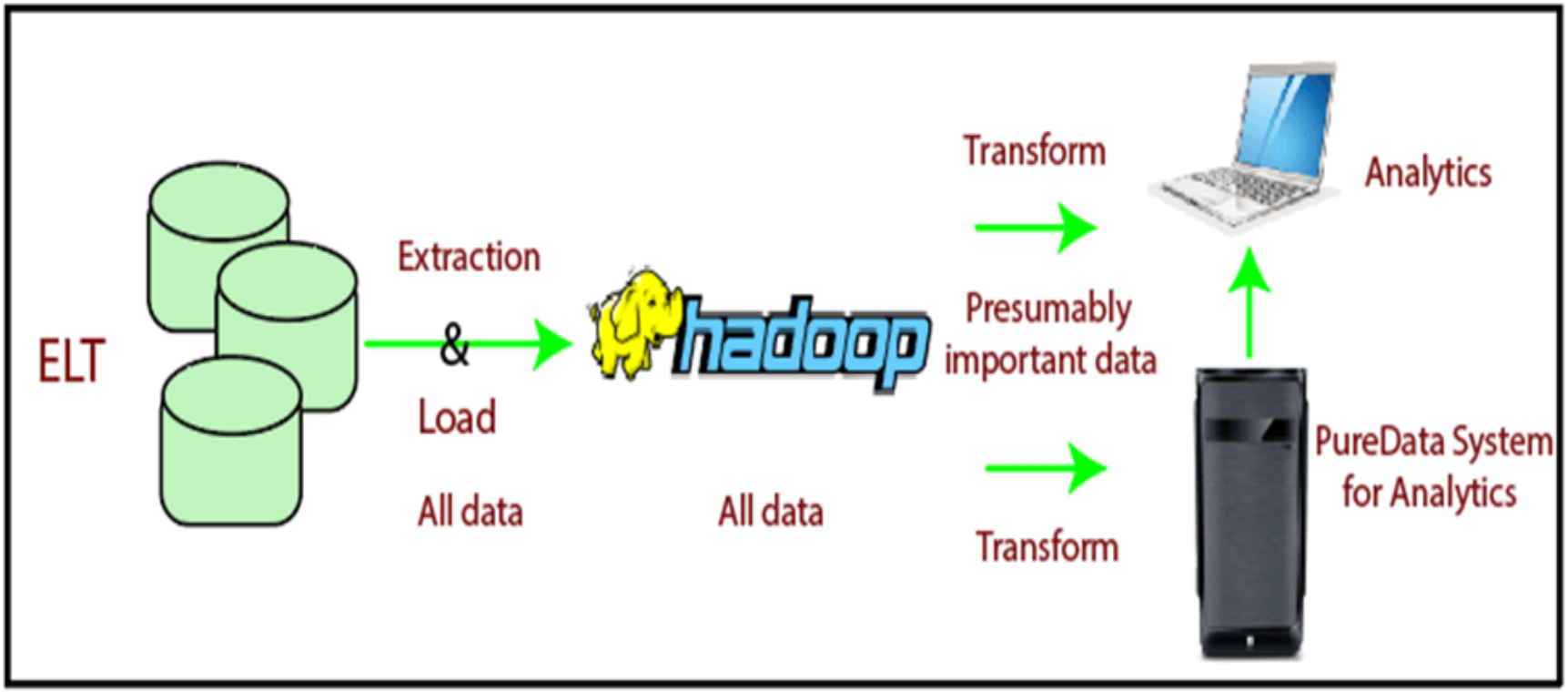
Data warehouse framework



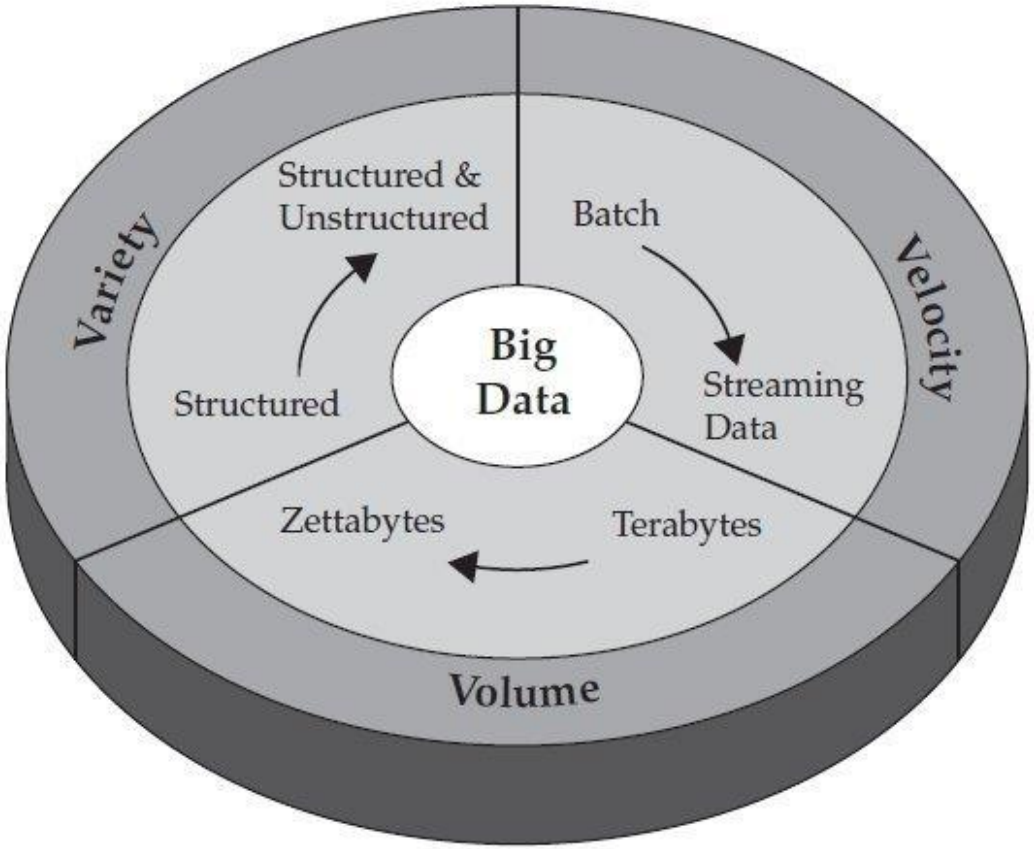
Extract Transform load (ETL)



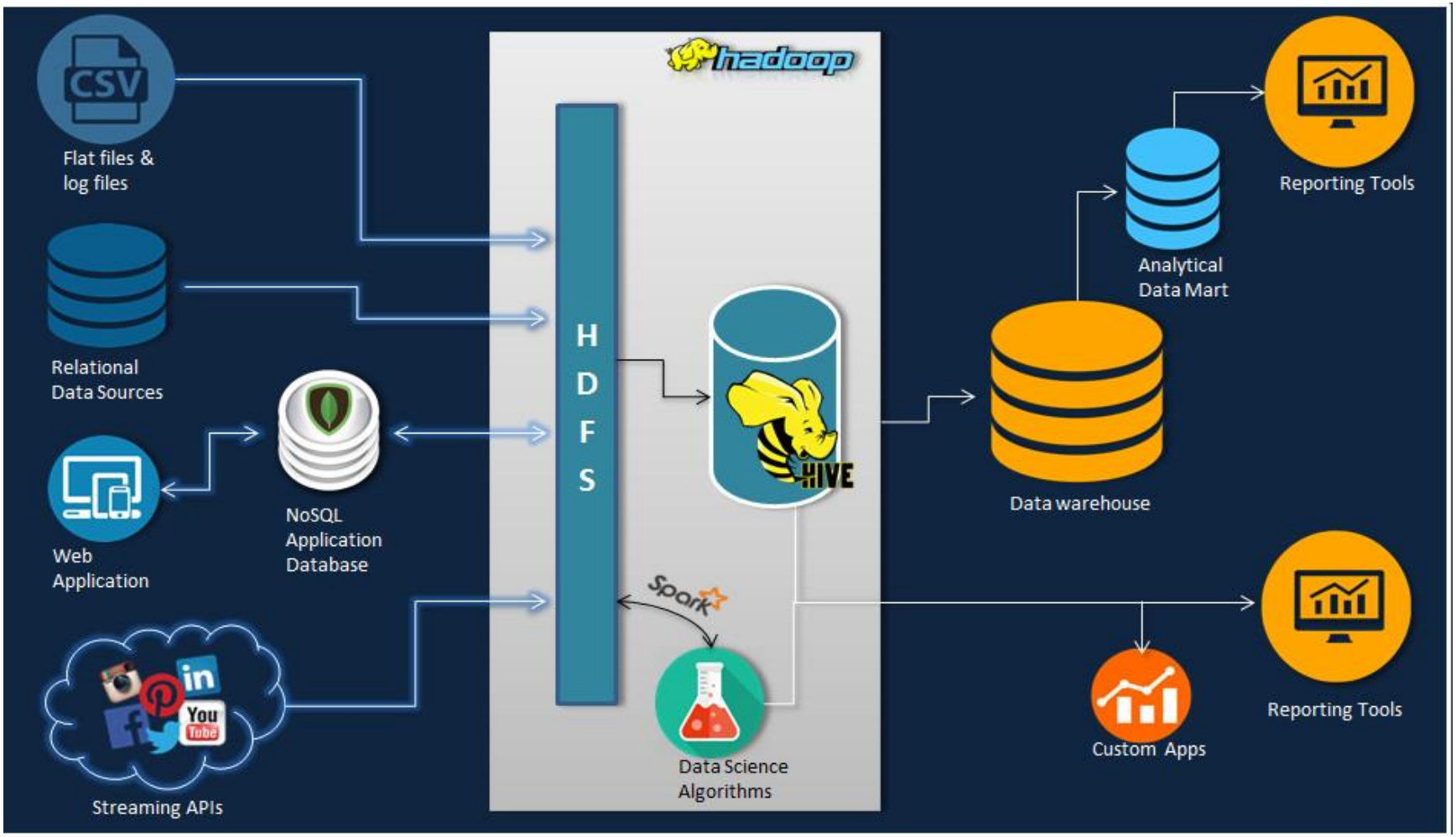
Extract load Transform (ELT)



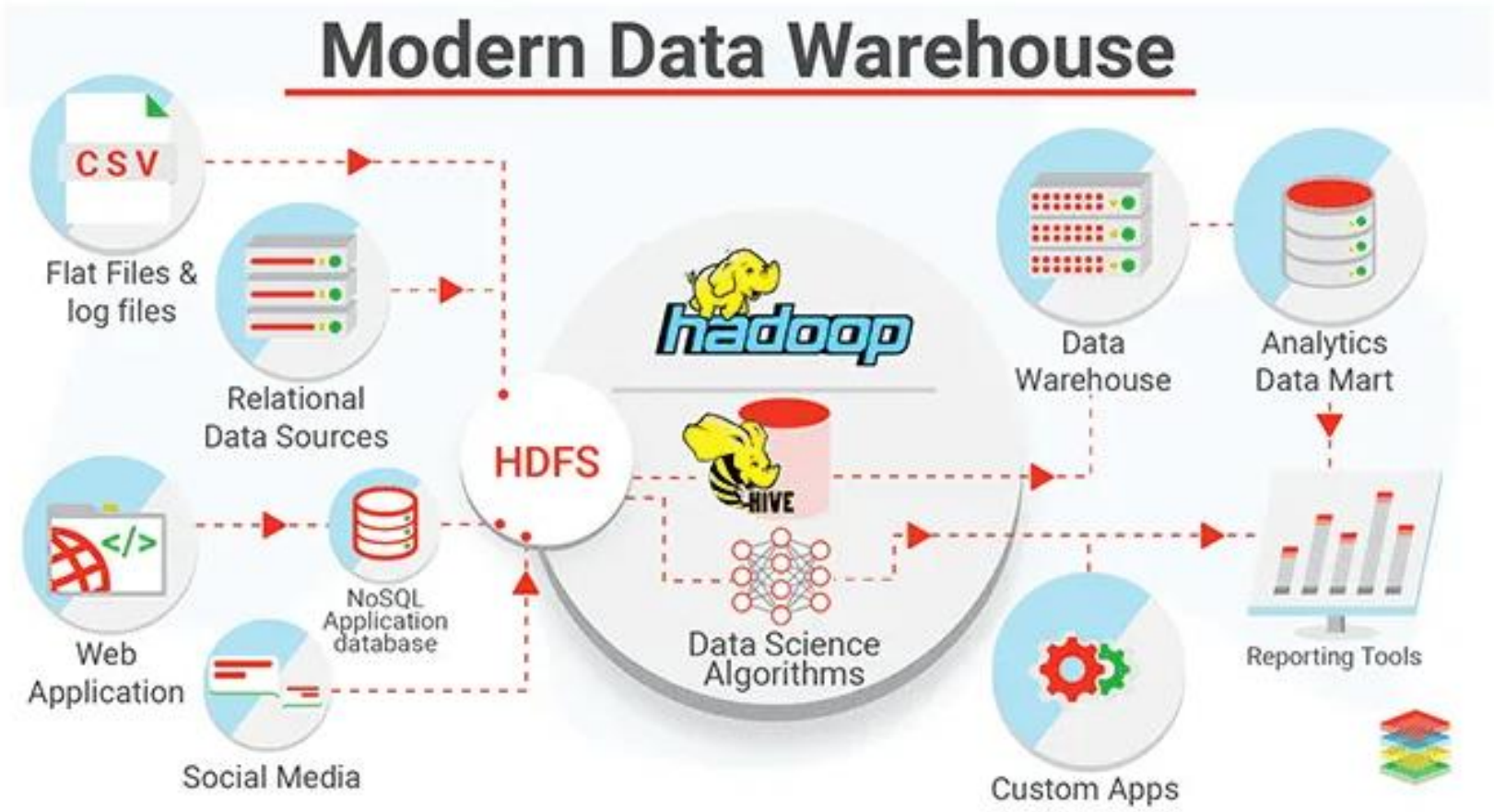
Bigdata


Variety
Structured &
Unstructured
Batch
Velocity
Big
Data
Structured
Streaming
Data
Zettabytes
Terabytes
Volume

Data warehouse and Big Data



Data warehouse and Big Data



Data lakes

- **Data lakes :** With the emergence of **Big Data**, there came a new data platform: data lake, which is a large storage location that can hold vast quantities of data (mostly unstructured) in its native/raw format for future/potential analytics consumption.
- Traditionally speaking, whereas a data warehouse stores structured data, a data lake stores all kinds of data. While they are both data storage mechanisms, a data warehouse is all about structured/tabular data and a data lake is about all types of data.
- Although much has been said and written about the relationship between the two (some of which suggests that data lake is the future name of data warehouses), as it stands, a data lake is not a replacement for a data warehouse; rather, they are complementary to one another.

Data lakes

Data lakes vs Data warehouse

- **Data.** A data warehouse only stores data that has been modeled/aggregated/structured, whereas a data lake stores all kinds of data—structured, semi structured, and unstructured—in its native/raw format.
- **Processing.** Before loading data into a data warehouse, we first need to give it some shape and structure— that is, we need to model it into a star or snowflake schema, which is called schema-on-write. With a data lake, we just load in the raw data, as-is, and then when we are ready to use the data, we give it a shape or structure, which is called schema-on-read. These are two very different processing approaches.
- **Retrieval speed.** For more than two decades, many algorithms have been developed to improve the speed at which the data is retrieved from large and feature-rich data warehouses. Such techniques included triggers, columnar data representation, in-database processing. As of now, the retrieval of data (which can be in any form or fashion—including unstructured text) is a time-demanding activity.

Data lakes

Data lakes vs Data warehouse

- **Storage.** One of the primary features of Big Data technologies like Hadoop is that the cost of storing data is relatively low as compared to the data warehouse. There are two key reasons for this: First, Hadoop is open source software, so the licensing and community support is free. And second, Hadoop is designed to be installed on low-cost commodity hardware.
- **Agility.** A data warehouse is a highly structured repository, by definition. It's not technically hard to change the structure, but it can be very time consuming given all the business processes that are tied to it. A data lake, on the other hand, lacks the structure of a data warehouse—which gives developers and data scientists the ability to easily configure and reconfigure their models, queries, and apps on-the-fly.

Data lakes

Data lakes vs Data warehouse

- **Novelty/newness.** The technologies underlying data warehousing have been around for a long time. Most of the innovations have been accomplished in the last 20–30 years. Therefore, there is very little, if any, newness coming out of data warehousing (with the exclusion of the technologies to leverage and use Big Data within a data warehouse). On the other hand, data lakes are new and are going through a novelty/ innovation phase to become a mainstream data storage technology.
- **Security.** Because data warehouse technologies have been around for decades the ability to secure data in a data warehouse is much more mature than securing data in a data lake. It should be noted, however, that there is a significant effort being placed on security right now in the Big Data industry. It is not a question of if, but when, the security of the data lakes will meet the needs and wants of the analytics professionals and other end users.

Data lakes

Data lakes vs Data warehouse

Dimension	Data Warehouse	Data Lake
The nature of data	Structured, processed	Any data in raw/native format
Processing	Schema-on-write (SQL)	Schema-on-read (NoSQL)
Retrieval speed	Very fast	Slow
Cost	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, flexible configuration
Novelty/newness	Not new/matured	Very new/maturing
Security	Well-secured	Not yet well-secured
Users	Business professionals	Data scientists