DATA ENGINEERING BOOTCAMP

BYTEWISE

SQL(structured query language)

- What is SQL?
- SQL Functions
- Single-row Functions
- Case-Conversion Functions
- Number Functions
- Group Functions
- Having Clause
- Cast()

SQL JOINS

- TYPES OF JOINS
- INNER JOINS
- LEFT JOIN
- RIGHT JOIN
- FULL OUTER JOIN
- CROSS JOIN
- SQL SELF JOIN
- SQL UNION
- UNION ALL
-

SQL GROUP BY AND SUM OVER()

- GROUP FUCNTIONS
- GROUP BY
- HAVING CLAUSE
- OVER CLAUSE
- SUM OVER CLAUSE

HACKERRANK

DATA ENGINEERING

- Data Engineering basics?
- Data Engineering end-to-end architecture?
- Role of a Data Engineer?
- What is Data Warehouse?
- Data Lakes, Layers

WHAT IS DATA ENGINEERING?

A data engineer is an individual responsible for managing, optimizing, overseeing, and monitoring data retrieval, storage, and distribution.

ROLE OF DATA ENGINEER

They can broadly be categorized into three main categories: generalist, pipeline-centric, and database-centric.

1. Generalist

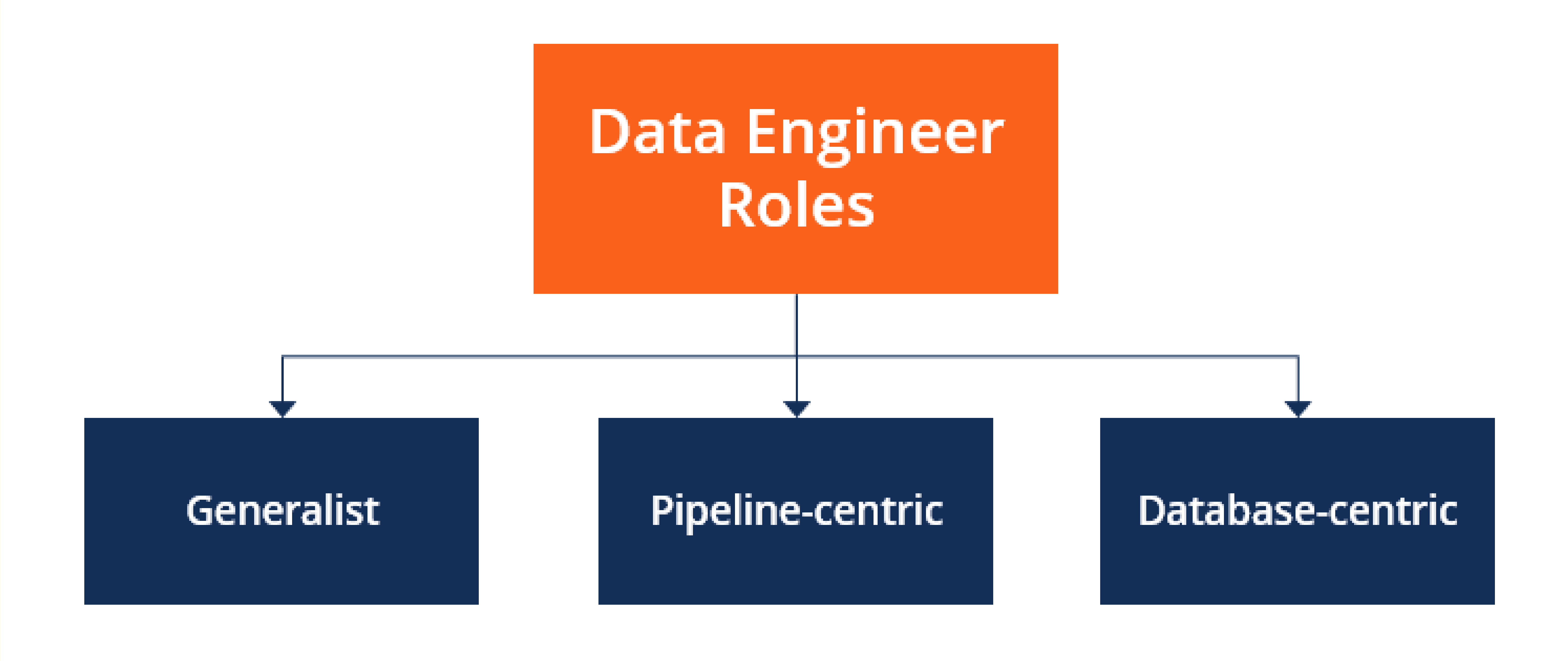
They will likely need to do more end-to-end work, such as following through with the entire process of ingesting the data, processing it, and getting involved in data analysis.

2. Pipeline-centric

Pipeline-centric data engineers are often found in larger, midsize companies. They are responsible for working with other data scientists to interpret and use the data collected.

3. Database-centric

Database-centric data engineers are found in some of the largest companies and conglomerates, and their job is to focus on setting up and populating analytics. There are usually large databases involved, and the data engineers work with data warehouses across multiple databases.



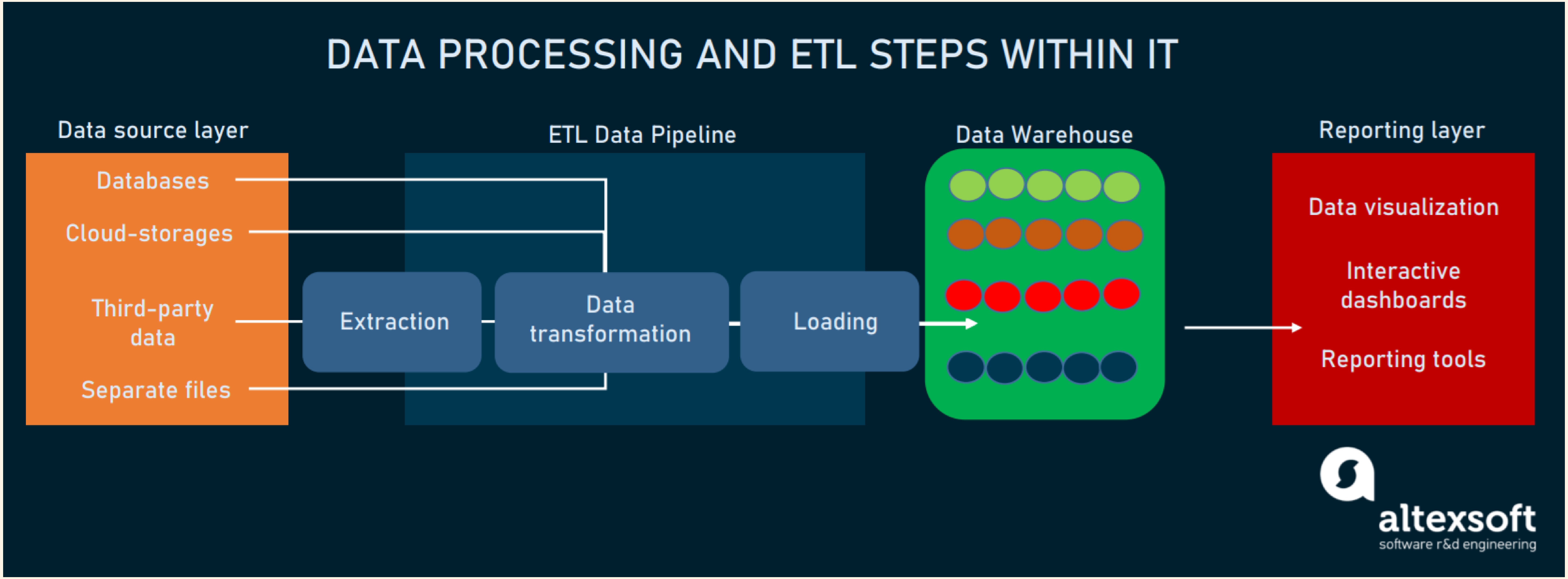
Data Engineer Roles

Generalist

Pipeline-centric

Database-centric

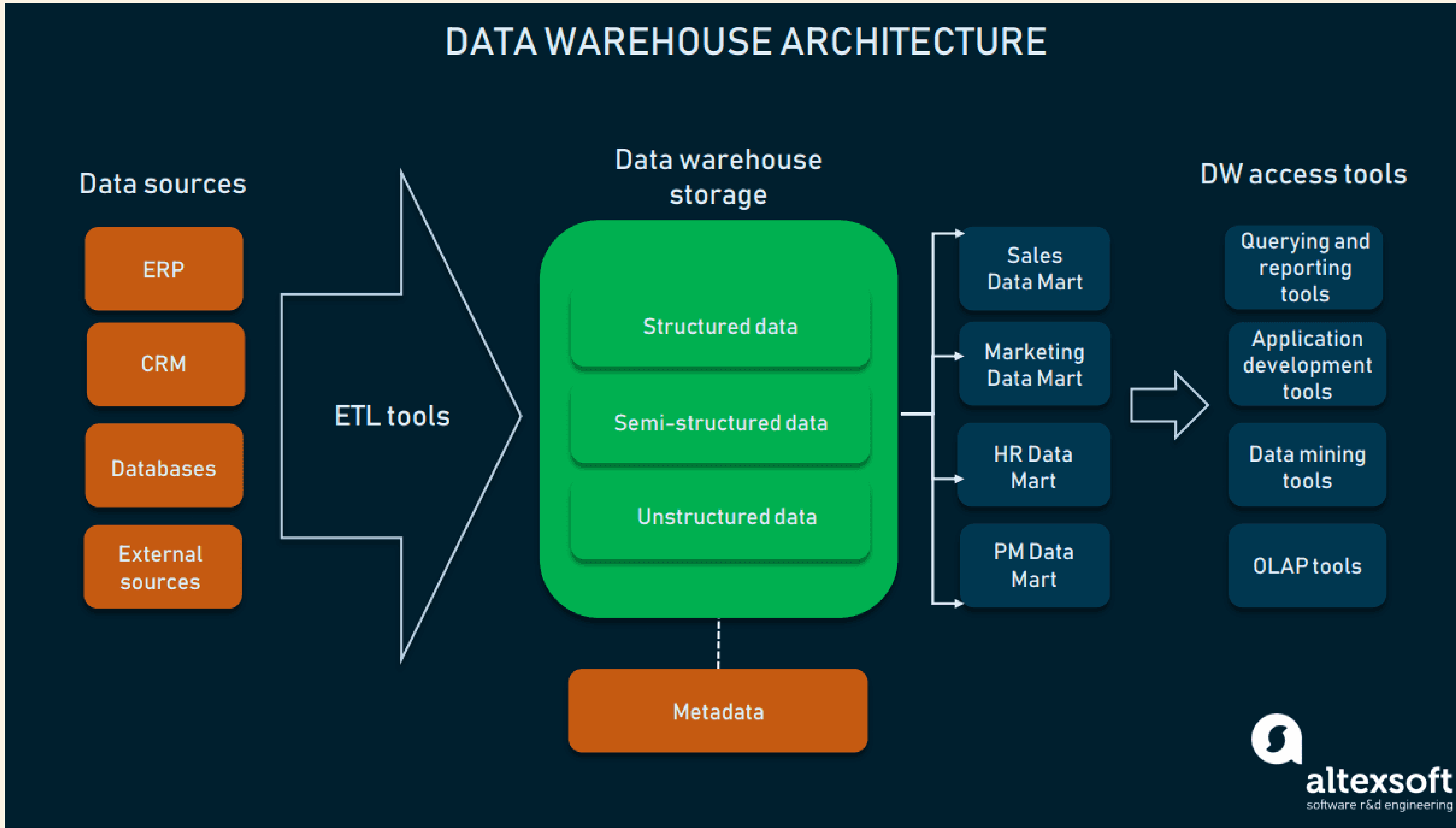
END TO END ARCHITECTURE



DATAWAREHOUSE ARCHITECTURE

A data warehouse is a central repository of information that can be analyzed to make more informed decisions.

Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence.



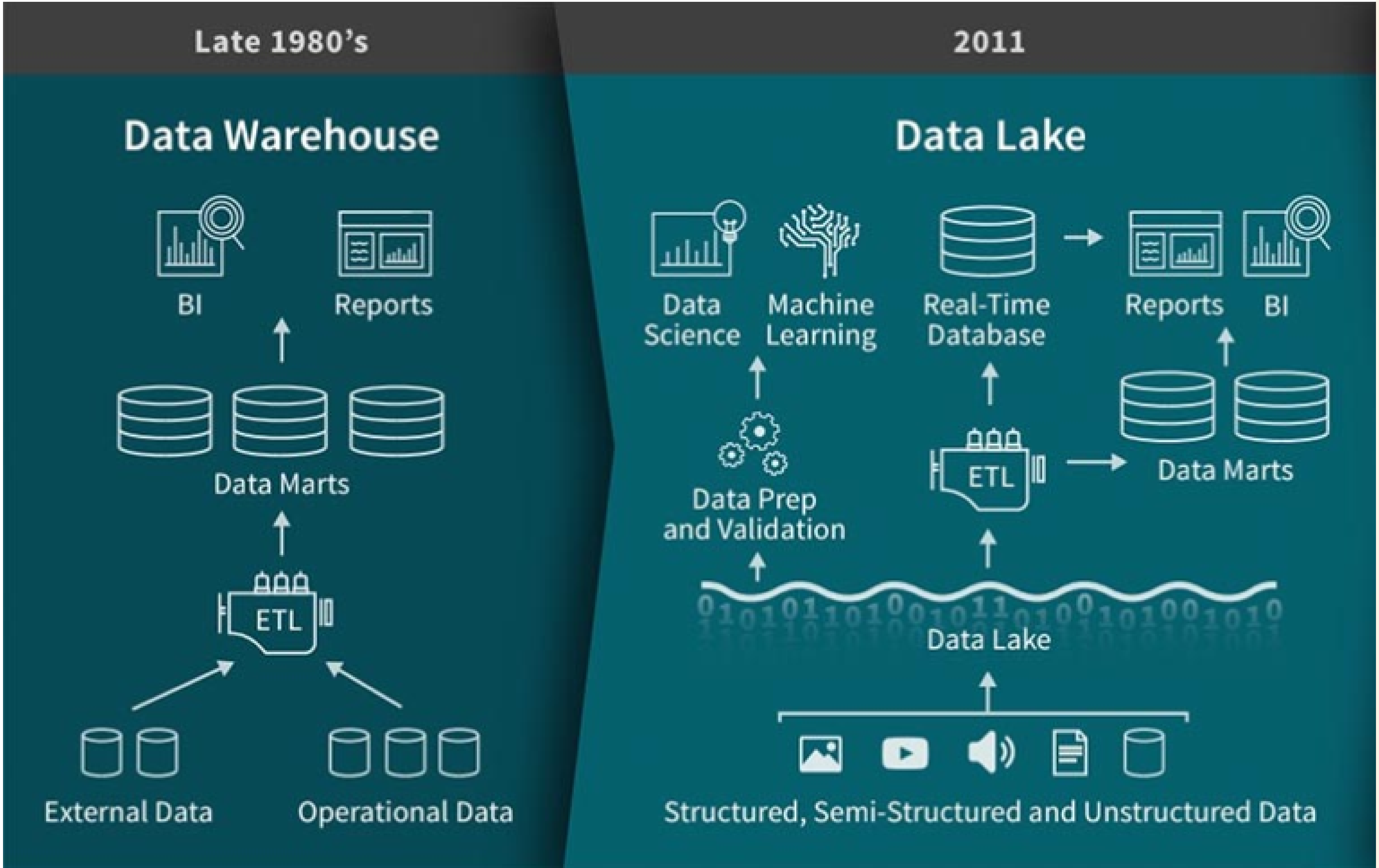
DATA LAKES

- A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions

- A data lake provides a scalable and secure platform that allows enterprises to:
- ingest any data from any system at any speed—even if the data comes from on-premises, cloud, or edge-computing systems;
- store any type or volume of data in full fidelity;
- process data in real time or batch mode;
- analyze data using SQL, Python, R, or any other language, third-party data, or analytics application.
-

DATA WAREHOUSE

A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making



Late 1980's

Data Warehouse

BI

Reports

Data Marts

ETL

External Data

Operational Data

2011

Data Lake

Data Science

Machine Learning

Real-Time Database

Reports

BI

Data Marts

ETL

Data Prep and Validation

Data Lake

Structured, Semi-Structured and Unstructured Data

DATA LAYERS BRONZE,SILVER,GOLD

- Bronze Layer:
  A one-on-one copy of the data from the source into the data lake. ‘Bronze data’ is raw untransformed unmodified data and all your sources land into this layer.

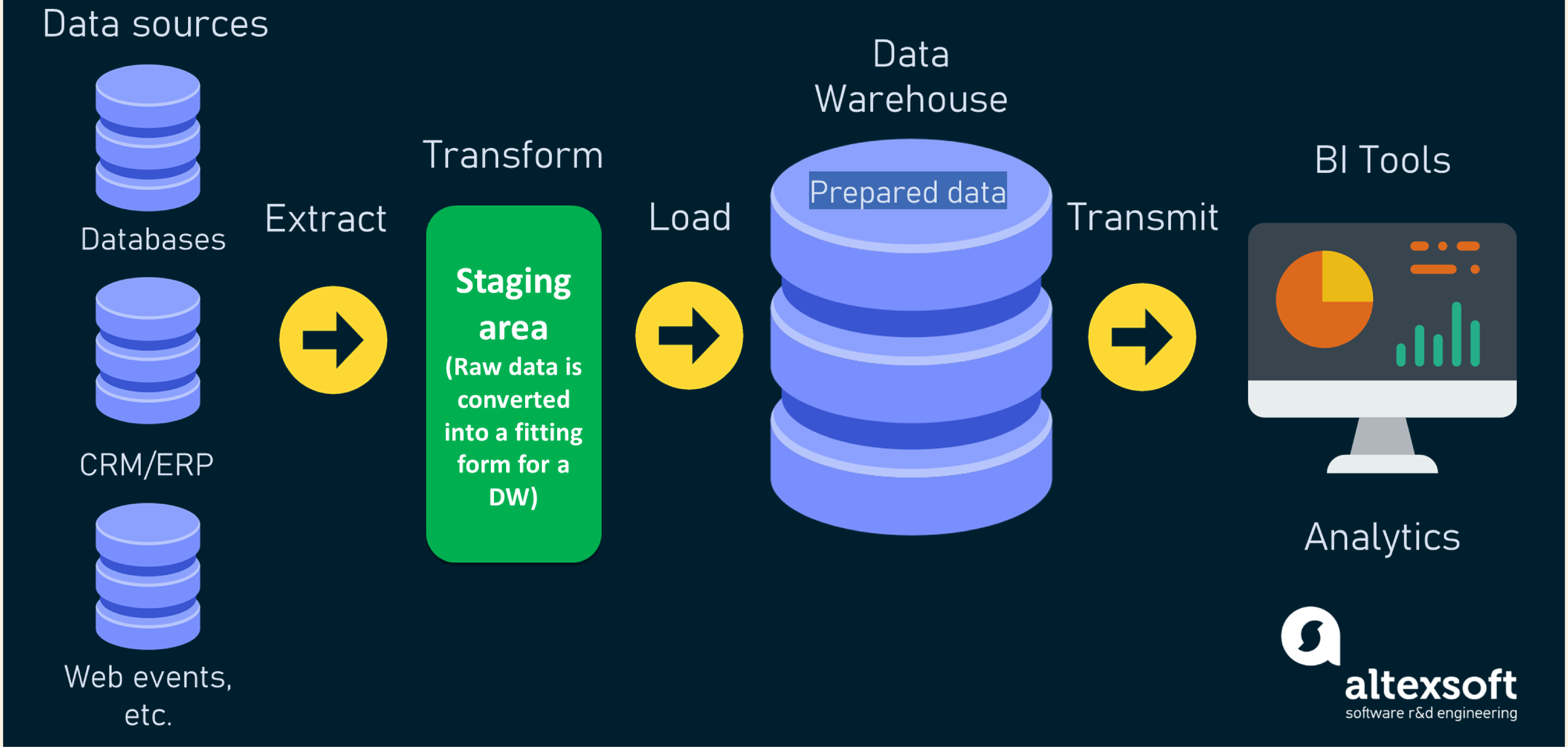
- Silver Layer:
  Once a business case has been identified and requires analysis, the ‘raw Bronze data’ is transformed into sets of data that add additional values. This can imply replacements of codes to meaningful values, adding sanity constraints, filtering-out unneeded information. Hence, resulting in concise useful datasets.

- Gold Layer:
  The gold layer then provides a well-constructed dataset ready for analysis by data scientists and business analysts. The data is presented in such a way that appeals to them the most, which may include aggregations, joins and merges, encoding, etc.

ETL

- ETL stands for Extract, Transform, Load
- ETL is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other unified data repository.
- ETL provides the foundation for data analytics and machine learning work streams.
- Through a series of business rules, ETL cleanses and organizes data in a way which addresses specific business intelligence needs, like monthly reporting, but it can also tackle more advanced analytics, which can improve back-end processes or end user experiences. ETL is often used by an organization to:
- Extract data from legacy systems
- Transform or Cleanse the data to improve data quality and establish consistency
- Load data into a target database

ETL: EXTRACT, TRANSFORM, LOAD



Data sources

Databases

CRM/ERP

Web events, etc.

Extract

Transform

Staging area

(Raw data is converted into a fitting form for a DW)

Load

Data Warehouse

Prepared data

Transmit

BI Tools

Analytics

altexsoft
software r&d engineering

Importance of ETL in Data Engineering

- Data engineers use ETL processes to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.
- First, data engineers construct a data warehouse. The tried and true process that data engineers use is called ETL — Extract, Transform, Load.
- ETL is required for data engineers. Traditional ETL experience may be valuable as data engineers may need to gather data from various sources, transform structured and unstructured data into useful information, clean up data, etc.

AZURE DATABRICKS

- azure subscription
- Azure Databricks
- Apache Saprk
- Saprk Architecture
- Databricks Spark
- Create Azure Databricks Services

Databricks

Azure Databricks is a data analytics platform optimized for the Microsoft Azure cloud services platform. Azure Databricks offers three environments for developing data intensive applications: Databricks SQL, Databricks Data Science & Engineering, and Databricks Machine Learning. It is a collection of virtual machines.

CLUSTER

A Databricks cluster is a set of computation resources and configurations on which you run data engineering, data science, and data analytics workloads, such as production ETL pipelines, streaming analytics, ad-hoc analytics, and machine learning

Pools

Databricks pools reduce cluster start and auto-scaling times by maintaining a set of idle, ready-to-use instances. When a cluster is attached to a pool, cluster nodes are created using the pool’s idle instances. If the pool has no idle instances, the pool expands by allocating a new instance from the instance provider in order to accommodate the cluster’s request. When a cluster releases an instance, it returns to the pool and is free for another cluster to use. Only clusters attached to a pool can use that pool’s idle instances.

Notebooks

A notebook is a web-based interface to a document that contains runnable code, visualizations, and explanatory text

Databricks Workflows, Filter & Join Transformation, Aggregation

Data Fundamentals Certifications

Hackerrank

Azure Fundamental Training

AZ-900 exam