2. Modern Data Architecture

Data Warehouse

A data warehouse is a central repository designed for storing and analyzing large volumes of structured data from various sources. It's optimized for query performance and analytics workloads rather than transaction processing.

Key characteristics of data warehouses:

- **Subject-oriented**: Organized around major subjects like customers, products, or sales
- **Integrated**: Consolidates data from multiple sources
- **Time-variant**: Maintains historical data
- **Non-volatile**: Data doesn't change once it's loaded
- **Structured data**: Works with defined schemas and structured data

Popular data warehouse technologies include:

- **Snowflake**: Cloud-native data warehouse
- **Google BigQuery**: Google's serverless data warehouse
- **Amazon Redshift**: AWS's data warehouse solution
- **Microsoft Azure Synapse**: Microsoft's analytics service
- **Teradata**: Enterprise data warehouse platform

Data Lake

A data lake is a storage repository that holds a vast amount of raw data in its native format until it's needed. Unlike data warehouses, data lakes can store structured, semi-structured, and unstructured data.

Key characteristics of data lakes:

- **Schema-on-read**: No predefined schema; structure applied only when data is read
- **All data types**: Can store structured, semi-structured, and unstructured data
- **Massive scale**: Can scale to petabytes of data
- **Low-cost storage**: Typically uses object storage for cost efficiency
- **Raw data**: Stores data in its original format

Popular data lake technologies include:

- **Amazon S3**: Object storage service used as the foundation for many data lakes
- **Azure Data Lake Storage**: Microsoft's scalable data lake solution
- **Google Cloud Storage**: Google's object storage service
- **Hadoop HDFS**: Distributed file system for storing large datasets

Data Lakehouse

A data lakehouse is a relatively new architecture that combines the best elements of data warehouses and data lakes. It provides structure and data management features to data lakes while maintaining the flexibility, scalability, and lower cost of data lakes.

Key characteristics of data lakehouses:

- **ACID transactions**: Support for atomicity, consistency, isolation, and durability
- **Schema enforcement**: Ability to enforce schemas on data
- **Data governance**: Tools for managing data quality and metadata
- **BI support**: Direct connectivity to business intelligence tools
- **Storage decoupled from compute**: Separates storage and compute resources
- **Open formats**: Uses open file formats like Parquet or ORC

Popular data lakehouse technologies include:

- **Databricks Delta Lake**: Open-source storage layer that brings ACID transactions to data lakes
- **Apache Iceberg**: Table format for huge analytic datasets
- **Apache Hudi**: Data lake storage abstraction with record-level updates and deletes
- **Amazon Athena** with **AWS Lake Formation**: Query service that works with structured data in S3

ETL vs ELT

ETL (Extract, Transform, Load)

ETL is the traditional data integration process:

1. **Extract**: Data is extracted from source systems

2. **Transform**: Data is cleaned, enriched, and transformed in a separate processing server
3. **Load**: Transformed data is loaded into the target system (usually a data warehouse)

Characteristics of ETL:

- **Transformation before loading**: Data is transformed before it enters the data warehouse
- **Separate transformation server**: Requires additional infrastructure for transformation
- **Well-suited for complex transformations**: When transformations are CPU-intensive
- **Good for limited target system resources**: When the target system has limited processing power
- **Traditional approach**: Common with on-premises data warehouses

ELT (Extract, Load, Transform)

ELT is a modern approach that leverages the power of modern data warehouses:

1. **Extract**: Data is extracted from source systems
2. **Load**: Raw data is loaded directly into the target system
3. **Transform**: Data is transformed within the target system

Characteristics of ELT:

- **Loading before transformation**: Raw data is loaded first, then transformed
- **In-database transformation**: Uses the computing power of the data warehouse
- **Flexibility**: Easier to transform data in different ways for different needs
- **Fast initial loading**: Minimal processing before data is available
- **Modern approach**: Common with cloud data warehouses that can scale compute resources

DBT is designed for the ELT approach, focusing exclusively on the "T" (Transform) part of the process. It assumes data is already loaded into your data warehouse and provides tools to transform that data efficiently.

Modern Data Stack

The Modern Data Stack refers to a collection of tools and technologies that together form a complete data pipeline from source systems to analytics. It's characterized by cloud-native, API-first, modular components that can be combined to create a flexible data platform.

Key components of the Modern Data Stack:

1. **Data Ingestion/ETL Tools**:
   - **Fivetran**: Managed ELT service
   - **Airbyte**: Open-source ELT platform
   - **Stitch**: Data pipeline service
   - **Meltano**: Open-source ELT platform
2. **Data Warehouses/Lakes**:
   - **Snowflake**: Cloud data warehouse
   - **BigQuery**: Google's serverless data warehouse
   - **Redshift**: Amazon's data warehouse
   - **Databricks**: Unified analytics platform
3. **Transformation Layer**:
   - **DBT**: Transforms data in the warehouse
   - **Dataform**: SQL-based data transformation
4. **Orchestration**:
   - **Airflow**: Workflow management platform
   - **Prefect**: Modern workflow management system
   - **Dagster**: Data orchestrator for ML, analytics, and ETL
5. **Data Quality & Observability**:
   - **Great Expectations**: Data validation
   - **Monte Carlo**: Data observability platform
   - **Datafold**: Data diff and monitoring
6. **Analytics & BI Tools**:
   - **Looker**: Business intelligence platform
   - **Tableau**: Data visualization tool
   - **Power BI**: Microsoft's business analytics service
   - **Mode**: Analytics platform for data teams

DBT's role in the Modern Data Stack is central to the transformation layer. It sits between the data loading tools and the analytics tools, ensuring that raw data is properly modeled and transformed for analysis.