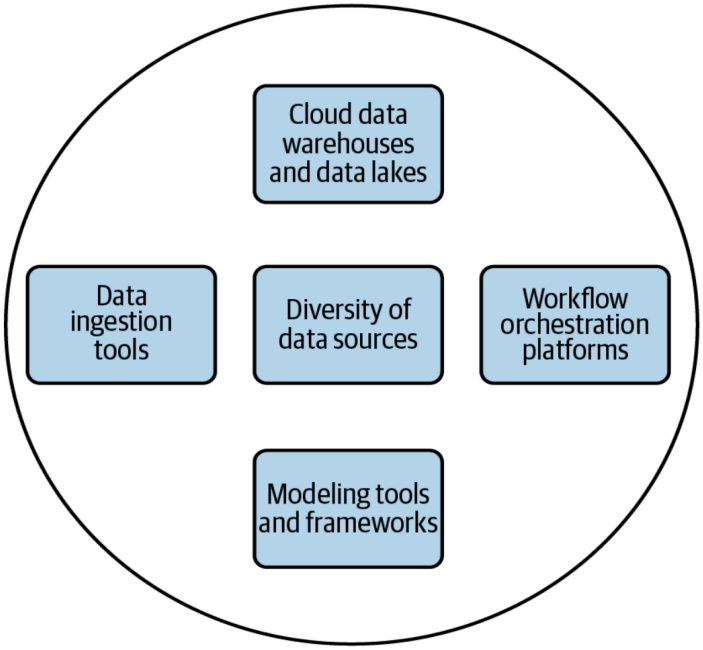
CHAPTER 2 Modern Data Infrastructure

Friday, 19 September 2025 12:59

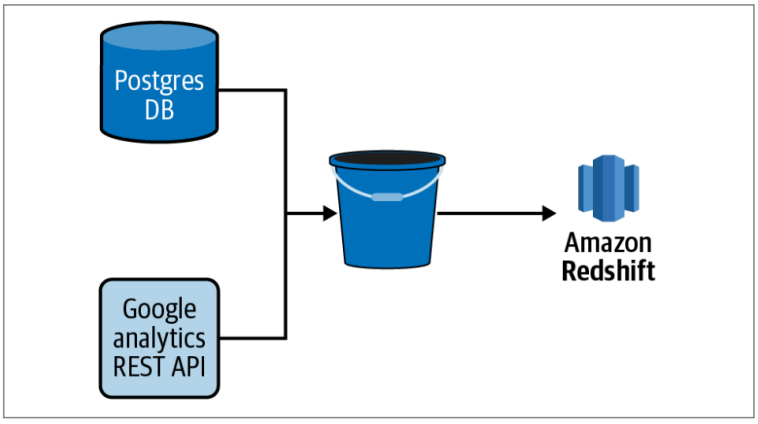
Key Components of Modern Data Architecture



These are affecting factors of design and implementation of data pipelines practices.

Source System Ownership

- Example : An ecommerce company stores their shopping information in backend such as in PostgreSQL and also can use third party web analytics tool like google analytics to track usage on their website.
- The combination of two data sources is required to get full understanding of customer behavior leading up to the purchase.
- So the data pipelines starts with the ingestion of both the sources



- Ownership of the source system is more important, since if in case of third party clients, its likely limited to what data we access and how we access it. Some company might provide REST API's or some with SQL database access.
- It all depends on what data we can access and at what granularity.

Ingestion Interface and Data Structure

Regardless of who owns data, the first thing a data engineer makes concern about it how you get it and in what form you get it.

Interface to data common platforms:

- A database behind application (MySQL or PostgreSQL)
- A datalake or data warehouse
- Shared network file system
- Rest API
- Data in HDFS

Common Structure of Data

- CSV files
- JSON from REST API's
- Semistructured
- Stream output from kafka

Note : Data engineers value small data same as they deal with big data, it is safe to think in a wide spectrum rather than confined terms like low or high in case of data volume.

Data Cleanliness & Validity

There are great diversity in data sources, the quality of data sources varies greatly.

Common characteristics of messy data includes :

- Missing values
- Incorrect data
- Duplicate values
- Inconsistent formatting

Modern pipelines tend to follow extract load transform rather than ETL, it is sometimes optimal to load data into the data lakes and worry about structuring and data cleaning later in the pipeline.

Even if you don't clean your data, validate it as much as often as we never know which will go wrong when.

Data Warehouses

Data Warehouses have structured and query optimized data stored from different systems used for reporting and analytics.

Data Lakes

Where as data lakes have varieties of data in the form of JSON, texts, csv and so on.

Data Ingestion Tools

Some of the common data ingestion tools are

1. Singer
2. Fivetran
3. Stitch

At last what is important is that its not about transformation or anything else, its about extracting data from source and loading them into the destination.

Data Transformation

Transforming data broadly is signified by the term T in ETL and ELT, transformation can be as simple as changing the timestamp of one timezone into another time zone or it may be deriving an entirely different column by transforming an existing column based on aggregated business logics.

Data Modelling

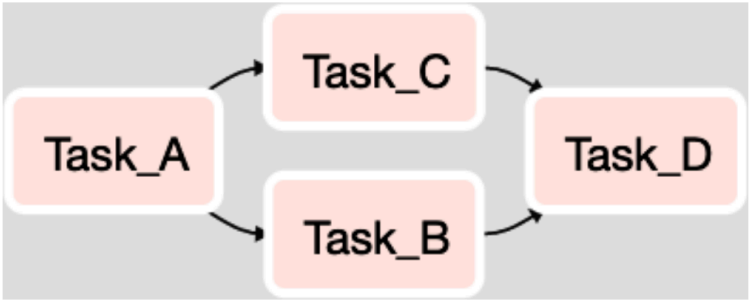
It is much different and specific type of data transformation, generally data models are two or more tables in a data warehouse.

Example : For the sake of protecting personally identifiable information (PII) say email id of a client, we can transform the email data into hashing code thereby such transformation is usually performed during the ingestion process.

Workflow Orchestration Platforms

- As the complexity and number of data pipelines grow, it is important to use a workflow orchestration platform to our data infrastructure
- These platforms manages scheduling and flow of tasks in pipelines, particularly imagine to execute a tasks in sequence, where you have to deal with data ingestion written in python and data transformation written in SQL.
- Some platforms such apache airflow, luigi and AWS Glue are designed for more of general purpose use cases.
- **Directed Acyclic Graphs (DAG's)**

  The orchestration pipelines are also referred to as graphs, and they often start with specific task and end with destination tasks. Thats why they are named directed and also to ensure the graphs don't cycle back to the completed tasks they are called acyclic as well.



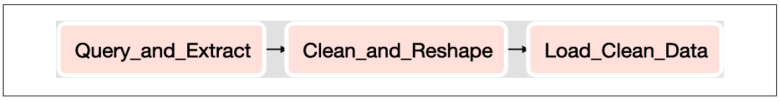


Figure 2-4. A DAG with three tasks that run in sequence to extract data from a SQL database, clean and reshape the data using a Python script, and then load the resulting data into a data warehouse.