Data Pipelines

In this lesson, you will learn about data pipelines.

We'll cover the following

- What are data pipelines?
- Features of data pipelines
- What is ETL?

What are data pipelines?#

> *Data pipelines* are the core component of a data processing infrastructure. They facilitate the efficient flow of data from one point to another and also enable developers to apply filters on the data streaming-in in real-time.

Today’s enterprise is data driven. This makes data pipelines key in implementing scalable analytics systems.

Features of data pipelines#

Speaking of some more features of the data pipelines -

- They ensure a smooth flow of data.
- They enable the business to apply filters and business logic on

streaming data.

- They avert any bottlenecks and redundancy in the data flow.
- They facilitate the parallel processing of data.
- They avoid data being corrupted.

These pipelines work on a set of rules predefined by the engineering teams, and the data is routed accordingly without any manual intervention. The entire flow of data extraction, transformation, combination, and validation and the convergence of data from multiple streams into one are completely automated.

Data pipelines also facilitate parallel processing of data via managing multiple streams. I'll talk more about distributed data processing in the upcoming lesson.

Traditionally we used *ETL* systems to manage all of the data's movement, but one major limitation is they don't really support handling real-time streaming data, which is possible with new era-evolved data processing infrastructure powered by the data pipelines.

What is ETL?#

If you haven't heard of ETL before, it means Extract Transform Load.

Extract means fetching data from single or multiple data sources.

Transform means transforming the *extracted* heterogeneous data into a standardized format based on the rules set by the business.

Load means moving the *transformed* data to a data warehouse or another data storage location for further processing of data.

The *ETL* flow is the same as the *data ingestion* flow. The difference is just that the entire movement of data is done in batches as opposed to

streaming it through the data pipelines in real-time.

Also, at the same time, it doesn't mean the *batch processing* approach is obsolete. Both real-time and batch data processing techniques are leveraged based on the project requirements.

You'll gain more insight into it when we go through the *Lambda* and *Kappa* architectures of distributed data processing in the upcoming lessons.

In the previous lesson, I brought up a few of the popular data processing tools, such as *Apache Flink*, *Storm*, *Spark*, *Kafka*, etc. All these tools have one thing in common they facilitate processing data in a cluster in a distributed environment via data pipelines.

This Netflix case study is a good read on how they migrated batch ETL to Stream processing using Kafka and Flink (https://www.infoq.com/articles/netflix-migrating-stream-processing/)

What is distributed data processing? How does it work? We are going to look into this in the next lesson. Stay tuned.

← Back

Data Ingestion Use Cases

Next →

Distributed Data Processing

Mark as Completed

Report an Issue