

DATA PIPELINES

1. Data Pipelines

00:00 - 00:05

Let's closely examine what data pipelines are and how they are used.

2. Data pipelines

00:05 - 00:37

DevOps Architecture uses data pipelines to ingest, transform, and move data between systems and microservices. Data Pipelines are used to ingest and merge data from different sources. Best not to confuse them with the CI/CD pipelines. The CI/CD pipelines are used within the DevOps Change Management Model to take developers' code, build it, test it, and deploy it automatically. Data Pipelines, on the other hand, are used for data processing purposes.

3. ETL

00:37 - 01:01

The basic functionality of Data Pipelines is formalized as ETL, short for Extract, Transform, and Load. Data Pipelines could be used to extract the data from the private service database of a microservice, transform the data to prepare for the destination, and load the data into the independent databases for analytics purposes.

1. ¹ <https://www.istockphoto.com/nl/vector/data-extract-transform-load-gm1361894912-434102842?phrase=ETL>

4. Batch processing

01:01 - 01:44

One of the essential types of Data Pipelines is used to move stored data in batches. Batch data refers to a large amount of data. Microservices can generate a substantial amount of data. For example, a microservice responsible for user authentication may store the names of the signed in users. It is a good idea to replicate these records into a centralized database for analytics purposes. Since this use case does not need real-time data, it makes sense to handle the data in batches, after it gets accumulated, for example, moving the data at midnight daily.

5. User connections

01:44 - 02:06

Not all data can be handled in batches. Stream Processing helps us handle real-time data. User interactions have to be in real-time. When a user clicks a button, a lot of computing happens within milliseconds. Many users connect to the backend at the same time.

6. Ingestion API

02:06 - 02:15

The backend has an ingestion API that allows the user to connect to the backend. Through the ingestion API, the user requests enter into the backend.

7. Stream processing

02:15 - 02:50

Once the user requests are ingested, they still need to be categorized and distributed to the relevant microservices in real time. Streaming data pipelines categorize the incoming user requests and send them to the relevant microservices. Once microservices do their magic, the result is sent to the users. In this example, data pipelines are ingested from the ingestion pipeline, transformed within the data pipeline as they are categorized, and loaded into different microservices.

8. Recap

02:50 - 03:23

Data Engineering is an integral part of Infrastructure Engineering. Data pipelines are one of the most robust tools for modern software infrastructure. Batch processing pipelines work on regular schedule like a certain part of the day to handle data whereas streaming processing works continuously. It never stops like a stream. Batch processing handles large amounts of accumulated data. Streaming processing handles data in real time.

9. Let's practice!

03:23 - 03:28

Let's hop onto the exercises and put into use what we learned.