

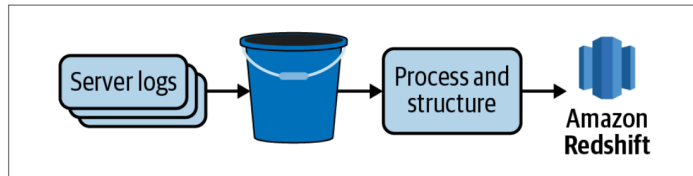
CHAPTER 1 Introduction to Data Pipelines

Friday, 19 September 2025 07:45

What are data pipelines ?

Data pipelines are set of processes that move data from source into destination where new value can be derived. They are foundation of analytics, reporting, and machine learning capabilities.

The complexity of data pipeline depends on the size, state and structure of source data as well as needs of the analytics project.



A simple data pipeline extracting data from server logs and loading in amazon S3 bucket and does some simple processing and structuring , there by storing the result in amazon redshift database.

Basic System Administration

A data engineer is said to be proficient in linux command line and be able to perform tasks such as

- Analyze applications logs
- Schedule cron jobs
- Troubleshoot firewall and
- Other security settings