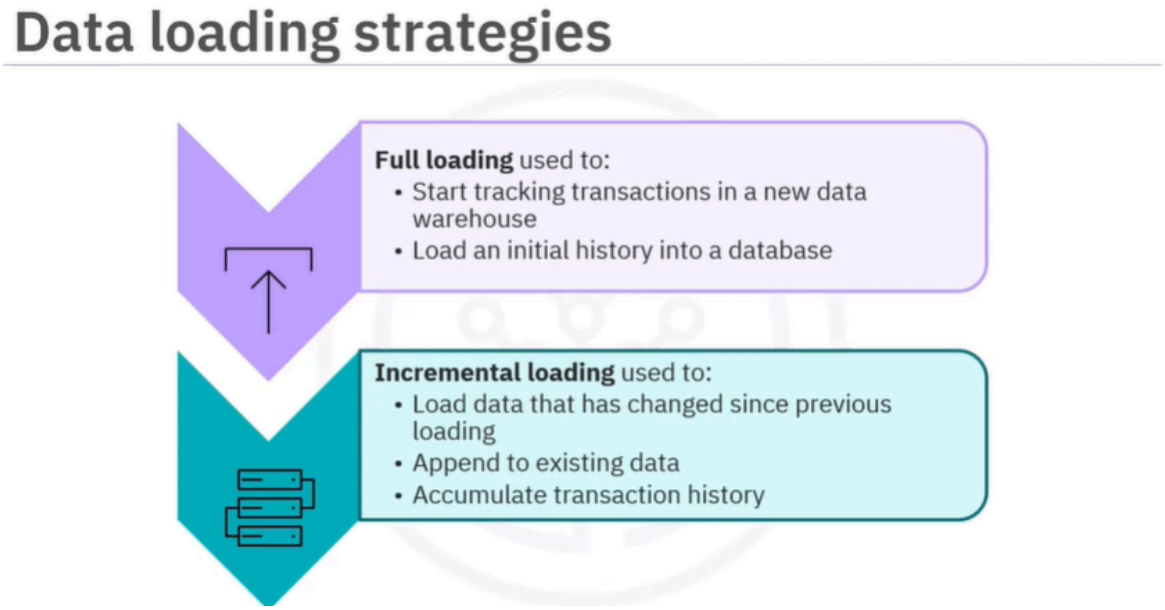
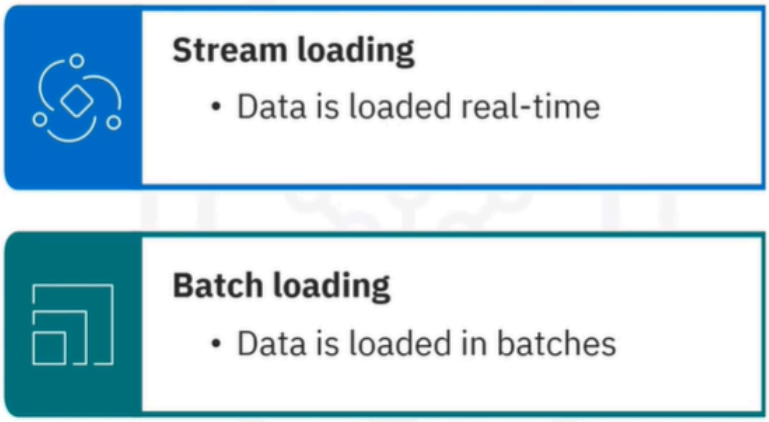
# Data Transformation Techniques



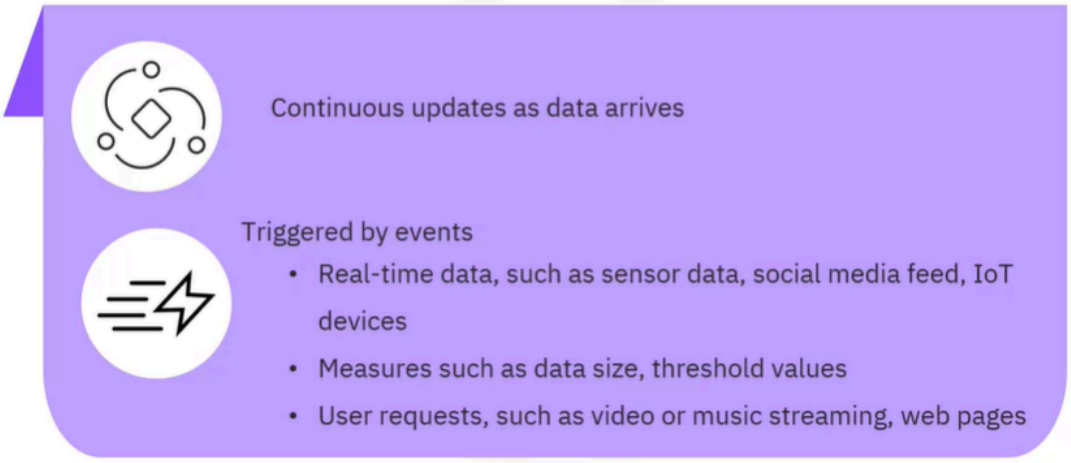

- There are two main data loading strategies, full loading and incremental loading. Full loading is used when you want to start tracking transactions in a new data warehouse or when you want to load an initial history into a database. To reiterate, there is no existing content when you use full loading. After full loading is complete, you can use incremental loading to insert data that has changed since the previous loading. With incremental loading strategy, data is appended in the database and not overwritten. It is useful for accumulating transaction history.

## Types of incremental loading

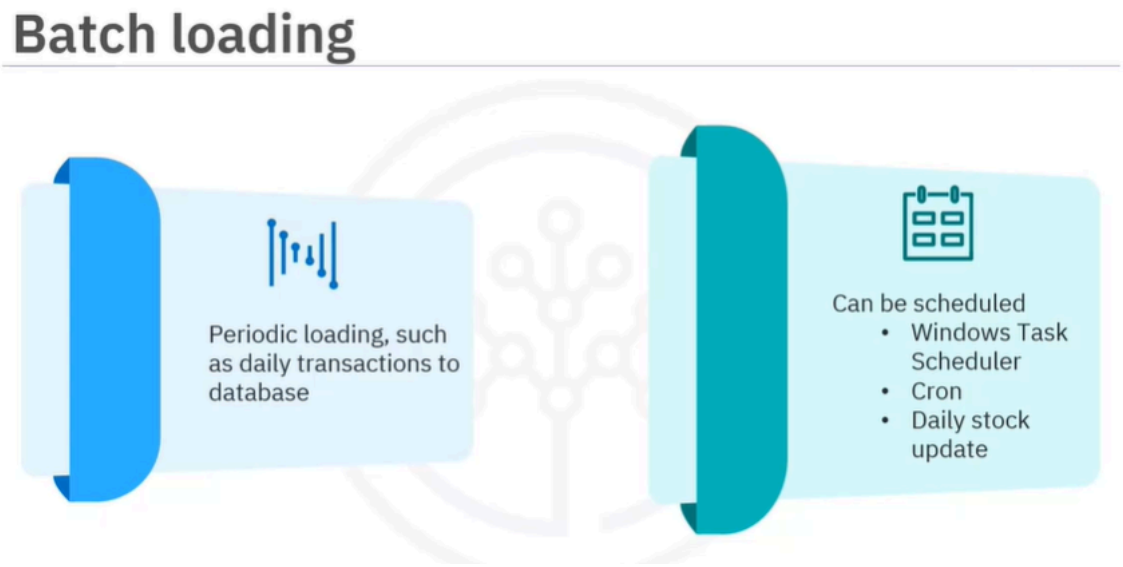



- You can categorize incremental loading into stream loading and batch loading, depending on the volume and velocity of data. Stream loading is used when the data is to be loaded real time. Batch loading is used when it's efficient and effective to load data in batches.

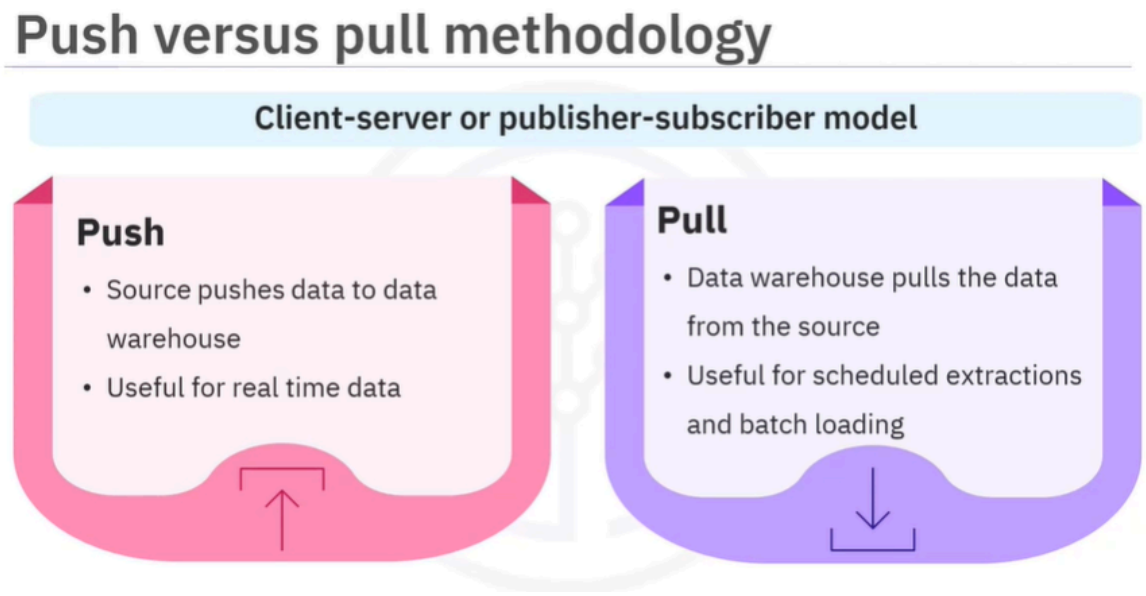
## Stream loading



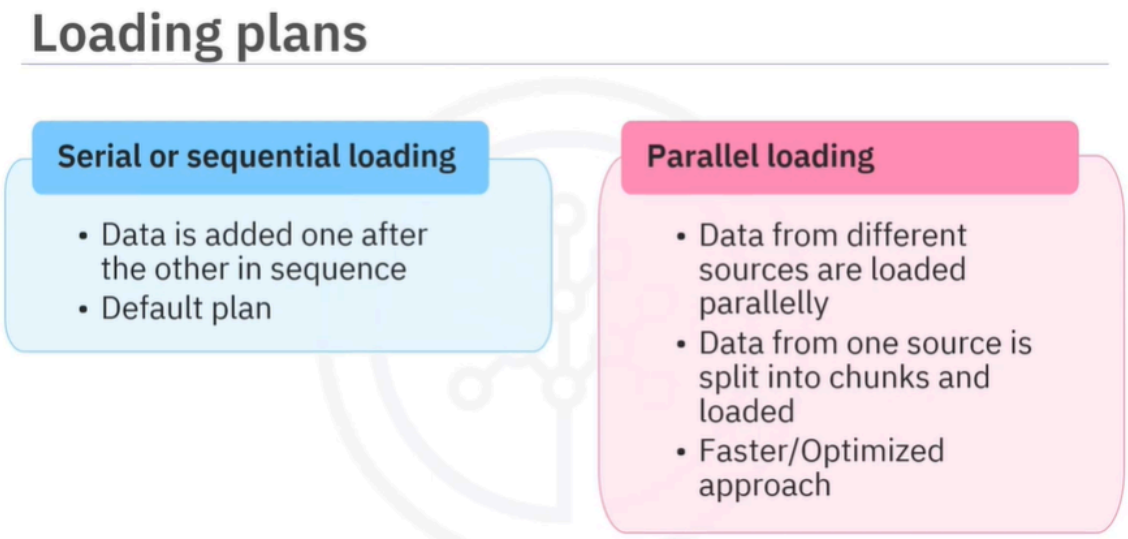
- Let's look at each of these loading strategies in detail. Stream loading refers to continuous data updates performed in the data warehouse or other storage systems as new data arrives. It is usually triggered by events, such as real-time data from sensors, like thermostat or motion sensors, social media feed, and IoT devices, and measures, such as data size when a certain amount of data is collected, or threshold values, or when a user requests data, such as online videos, music, or web pages.



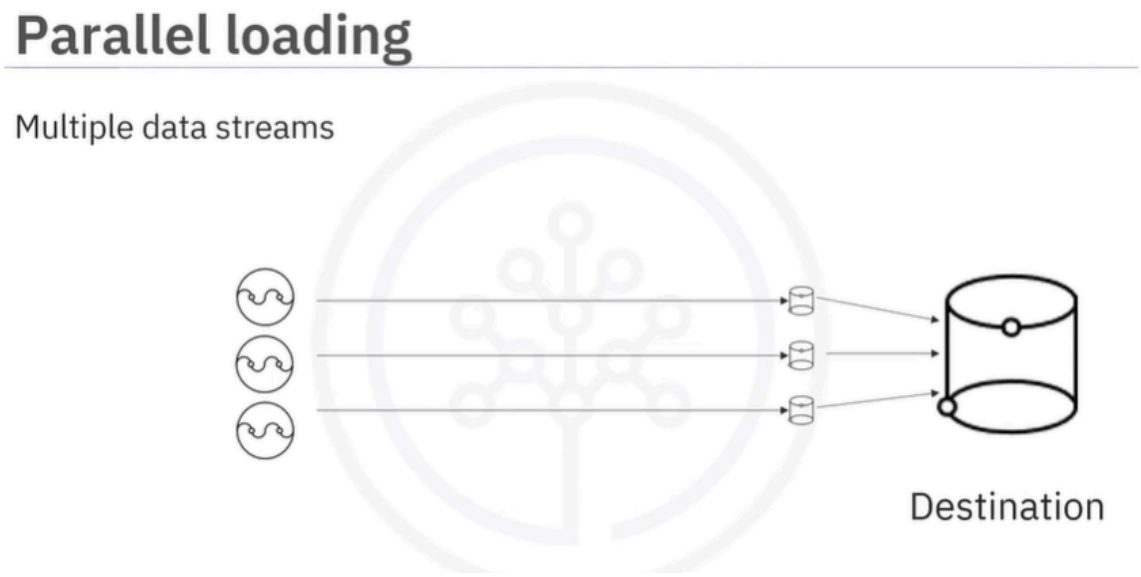

- Batch loading refers to periodic updates made/pushed to the data in the data warehouse or other storage systems, such as daily updates, hourly updates, or weekly updates. Batch data can be scheduled. Some examples include Windows Task Scheduler, Cron jobs in Linux, and daily stock update.




- Next, let's review push and pull data loading methodologies. Push and pull data loading methodologies are based on a client-server or publisher-subscriber model. A push method is used when the source pushes data into the data warehouse or other storage. While push method can be used for batch loading, it is most suited for stream loading involving real-time data. A pull method is used when the data warehouse pulls the data from the source by subscribing to receive the data. It is useful for scheduled transactions and batch loading.




- Loading can be serial or parallel. Serial loading is when the data is copied sequentially, one after the other. This is how data loads in the data warehouse by default. You can use parallel loading when you need to load data from different sources parallelly or to split data from one source into chunks and load them parallelly. When compared with serial loading, parallel loading is a faster and optimized approach.



## Parallel loading

File partitioning


- Parallel loading can be employed on multiple data streams to boost loading efficiency, particularly when the data is big or has to travel long distances. Similarly, by splitting a single file into smaller chunks, the chunks can be loaded simultaneously.

## Recap

In this video, you learned that:

- Full and incremental are data loading strategies
- Data can be loaded in batches or streamed continuously
- Both pull and push methodologies can be used for data loading
- Parallel loading can boost loading efficiency