Cloud ETL

Data Boot Camp

Lesson 22.3

Class Objectives

By the end of today's class, you will be able to:

01

Use Amazon Web Services (AWS) to host data in S3 buckets.

02

Create and use databases in the cloud.

03

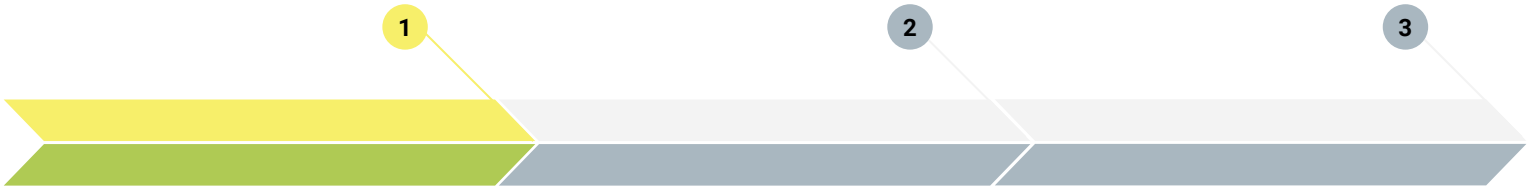
Define and create ETL pipelines in the cloud.

ETL in the Cloud

Cloud Extract

Files are stored in a cloud location such as an AWS S3 bucket.

These files are extracted from S3 and read into PySpark DataFrames using Apache ZEPL.



Files are stored in a cloud location such as an AWS S3 bucket.

S3

Cloud Transformation

Once the files are extracted into ZEPL, transformations can take place.

PySpark is used to transform the data.

Files are stored in a cloud location such as an AWS S3 bucket.

S3

ZEPL uses PySpark to perform transformations.

ZEPL Notebooks

Cloud Load

After the data is transformed, ZEPL loads the data.

ZEPL creates a connection to an RDS instance and loads the DataFrame.

Files are stored in a cloud location such as an AWS S3 bucket.

S3

ZEPL uses PySpark to perform transformations.

ZEPL Notebooks

ZEPL loads the transformed dataset to an RDS database.

PostgreSQL RDS

AWS S3

S3: Simple Storage Service

- S3 is Amazon’s cloud file storage service.
- S3 is a key-value store of objects: **key** is the object name, and **value** is the content stored.
- Files are stored on multiple servers, providing redundancy.
- High (> 99.99%) rate of availability guaranteed by Amazon.
- Files are organized by **buckets** (more on this later).

S3 Buckets

- S3 buckets are like computer folders or directories.
- An S3 bucket can contain multiple files.
- Unlike directories, S3 buckets must have unique names.
- The bucket name is a part of the file URL.

`https://s3.us-east-2.amazonaws.com/data-bootcamp-001/important_data.csv`

S3 Settings

S3 provides fine-grained control over files, including read and write permission for buckets and files.

Read and/or write permission can be granted to individuals and/or groups.

Questions?