# AWS Data Pipeline Setup: S3 ➝ Glue ➝ Redshift

## Overview

We used the following services:  
- Amazon S3: Stores the CSV files.  
- AWS Glue (Python Shell Job): Executes ETL script.  
- Amazon Redshift: Destination for loaded data.  
- IAM Role: Grants Redshift access to S3.

## 1. Amazon S3

- Bucket: s3://covid-19-data-de/  
- Folder: output/  
- CSV Files:  
 - factCovid.csv  
 - dimHospital.csv  
 - dimRegion.csv  
 - dimDate.csv  
- File Format: CSV, uncompressed, with headers.

## 2. IAM Role

- Name: redshift-s3-access  
- Attached to: Redshift cluster

## 3. Amazon Redshift

- Cluster Name: redshift-cluster-1  
- Database: dev  
- User: awsuser  
- Port: 5439  
- Public Access: Yes  
- IAM Role Attached: redshift-s3-access

## 4. AWS Glue Job

- Job Name: s3\_glue\_covid\_data  
- Type: Python Shell  
- Python Version: 3.9  
- Python Wheel File: redshift\_connector-2.0.889-py3-none-any.whl (uploaded to S3)  
- Libraries Used: redshift\_connector  
- No Connection Object Used: Redshift cluster was public

## 5. Network Setup

- VPC: Redshift in a VPC with:  
 - Public subnets  
 - Active Internet Gateway  
- Security Group:  
 - Port 5439 open to 0.0.0.0/0 or Glue job security group  
- Glue Job Access: Internet (no VPC connection)

All tables were created successfully and populated with data from S3 using AWS Glue.