Part 1: Distributed Data Preprocessing

Learn how to utilize distributed data processing in your data science projects that supports both interactive serverless development and ephemeral jobs for large scale processing.

In this workshop, we will walk through a series of hands-on labs.

Instructions

In SageMaker Studio, open and run the following notebooks in the sequence shown here:

0. Lab overview
1. Setup workshop dependencies
2. Register parquet data in S3 using AWS Glue and Amazon Athena
3. Visualize data with serverless distributed PySpark on SageMaker notebooks using Glue interactive sessions
4. Analyze data quality with distributed PySpark on SageMaker Processing Jobs

Go to the provided AWS account.
Run the notebooks in SageMaker Studio.

Data Science on AWS