Data preparation pipeline

Cloud-based data pipeline hosted on Amazon Web Services (AWS)

Host data pipelines in a highly-scalable cloud environment by utilising cost-efficient serverless computing and storage resources on-demand.

Employ a structured approach for building maintainable and reusable data pipelines. Build modular data pipelines by using a simple and accessible template.

Create complex and concurrent workflows employing computing power and orchestration capabilities of AWS services.

Add layers of quality control to pipelines and cloud infrastructure. Catch unexpected changes in data quality and infrastructure security vulnerabilities.

Project Objectives

The client organisation is Shine Solutions Group, a technology consultancy firm specialising in Cloud Solutions. The project's main deliverable is a cloud-based data pipeline that prepares employment profile data for analysis. Cloud-based data pipelines can provide organisations with a more efficient, flexible, and cost-effective way to process and manage their data.

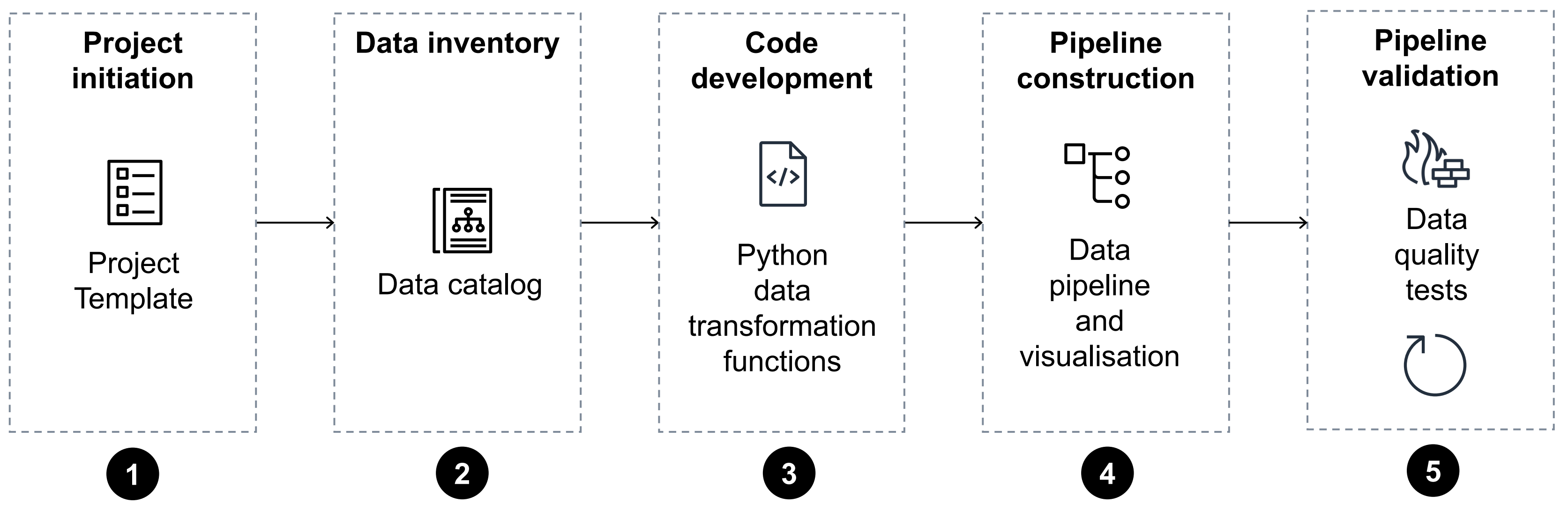
shine
solutions group

How it works

This project uses AWS resources and Python's Kedro framework to build and host the data preparation pipeline. Cloud infrastructure is defined with Terraform, an infrastructure as code (IaC) tool that allows specifying cloud resources and their configurations. AWS resources are deployed through a CI/CD pipeline including validation stage that scans the infrastructure configurations to find misconfigurations and security vulnerabilities.

Kedro is an open sourced Python framework for creating reproducible, maintainable and modular data pipeline code. In combination with AWS resources and other tools, it gives a robust, scalable, and secure approach to developing data workflows. AWS computing solutions used in this project enable high-level parallel computing and scalability by orchestrating large-scale, complex workflows to execute concurrently in a highly scalable manner.

Data transformation code development



1. The project is initiated with Kedro's standardised project template.

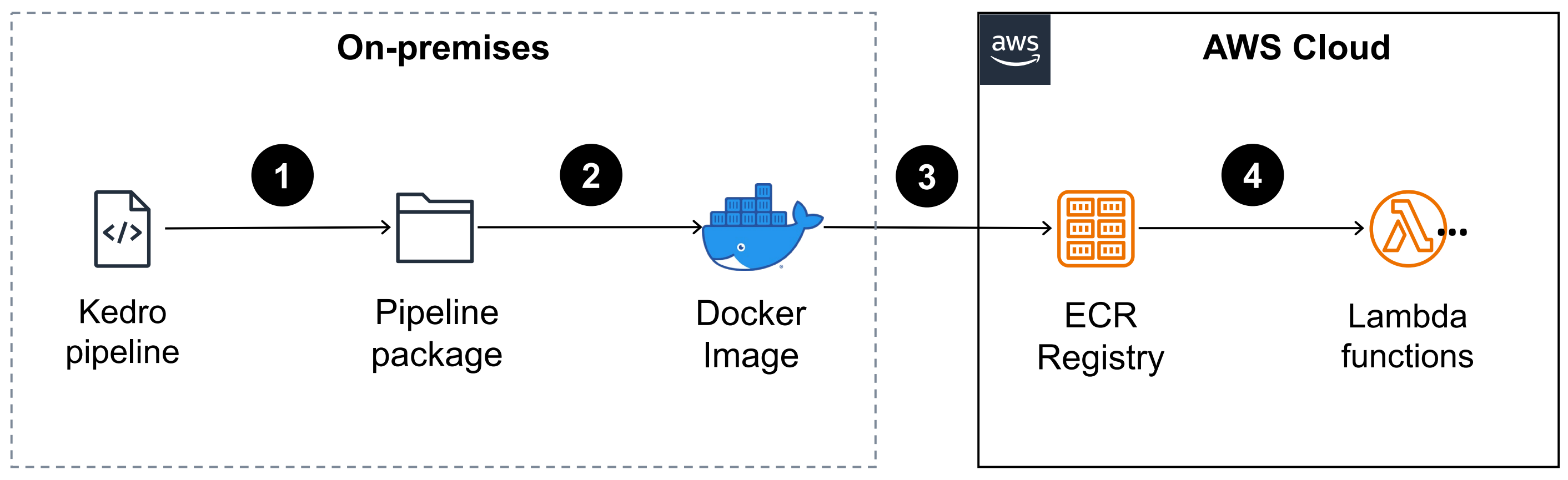
2. A data catalog is used to version and interact with data.

3. Define nodes, the building blocks of pipelines. A node is a wrapper for a Python function that names the inputs and outputs of the function.

4. A pipeline organises the execution order of a collection of nodes. Kedro's pipeline visualisation plugin can show a blueprint of your data workflows.

5. Modular pipelines run incremental computations, allowing for quality tests and easy maintenance.

Data transformation code deployment



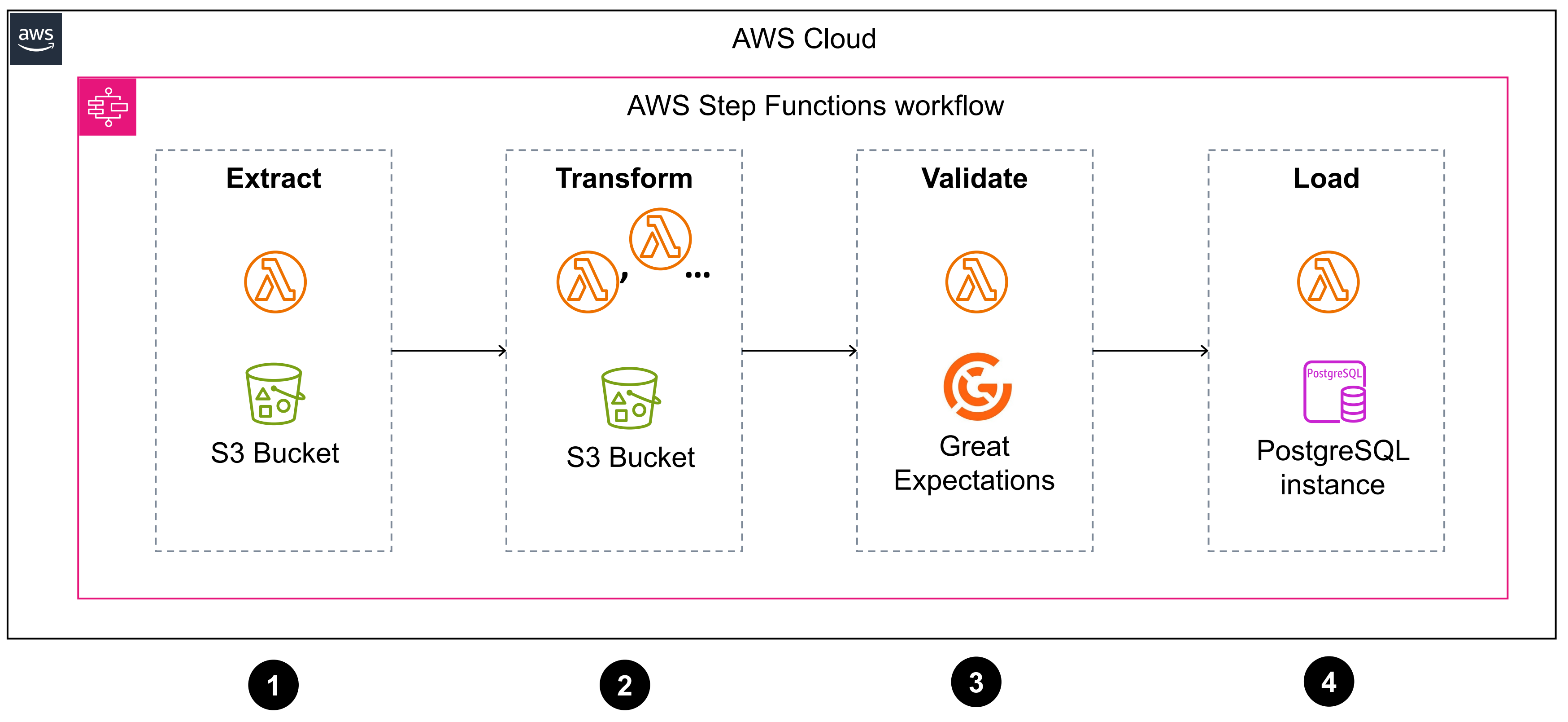
1. Locally developed Kedro pipeline is packaged using a default command.

2. The package is converted into an AWS Lambda-compliant Docker image.

3. The Docker image is pushed into AWS Elastic Container Registry (ECR). AWS ECR is a container registry offering high-performance hosting in the cloud.

4. Each Kedro node is deployed as an AWS Lambda function orchestrated by AWS Step Functions. AWS Lambda is a compute service that runs code without provisioning or managing servers. AWS Step Functions is a workflow service that automates and orchestrates processes. Each function will use the same pipeline Docker image from the AWS ECR and run a single Kedro node associated with it.

Data pipeline on AWS



1. Raw data, uploaded to a designated AWS Simple Storage Service (S3) bucket, is extracted by a Lambda function. AWS S3 is an object storage service that can store any amount and type of data. This function also triggers the pipeline as the data is uploaded.

2. Lambda functions compute data transformation functions in parallel. Intermediate data that has been transformed is stored in a separate S3 bucket.

3. Data quality is evaluated and monitored by Great Expectations checkpoints. Great Expectations is a tool for validating and profiling data to maintain its quality.

4. The validated data is uploaded into a PostgreSQL database hosted on AWS RDS. AWS RDS is a managed relational database service in the cloud.

Author: Gundalai "Dale" Batkhuu