Chisom Nnamani

Travel Agency Data Platform

Building a robust Data Platform
for predictive analytics

Presented By:

Chisom Nnamani

Chisom Nnamani

Table of Contents

Chisom Nnamani

Overview

A travel Agency reached out to Core Data Engineers, their business model involves recommending tourist locations to their customers based on different data points.

They wanted our Data team to build a Data Platform that will process the data from the Country REST API here into their cloud-based Database/Data Warehouse for predictive analytics by their Data Science team.

Chisom Nnamani

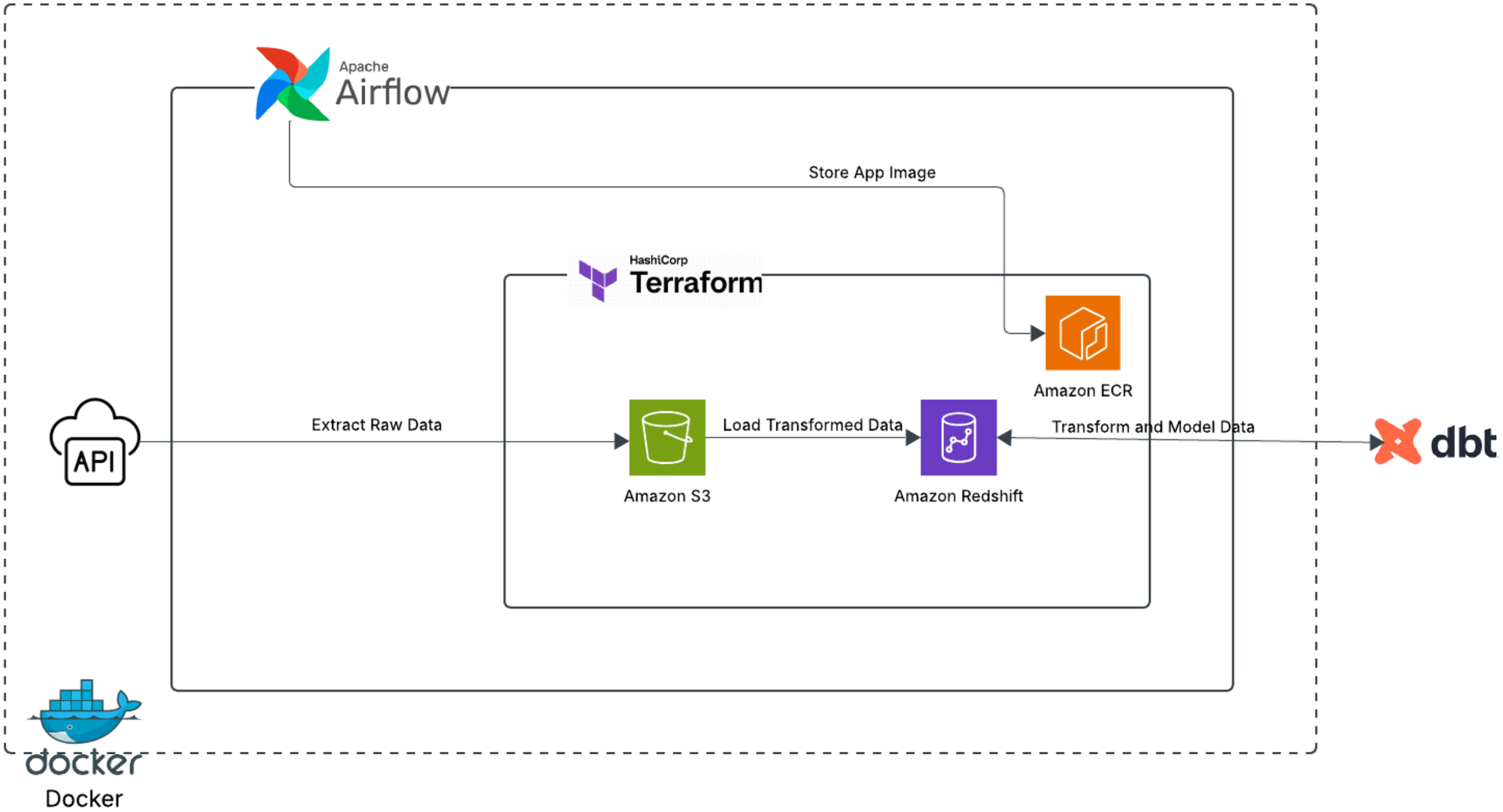
Methodology

Having carefully assessed the requirements, **Docker** was used to host **Airflow**, which served as the orchestration tool for this project. The dataset was extracted from the Country REST API and stored in Parquet format in **Amazon S3** to ensure future extensibility.

Relevant columns were then selected from the raw data and loaded into a **Redshift** table, which functioned as the Data Warehouse. **dbt** was utilized to model the transformed data into Fact and Dimension tables, enabling efficient querying. Additionally, **Terraform** was employed as an Infrastructure as Code (IaC) tool to provision all necessary AWS resources.

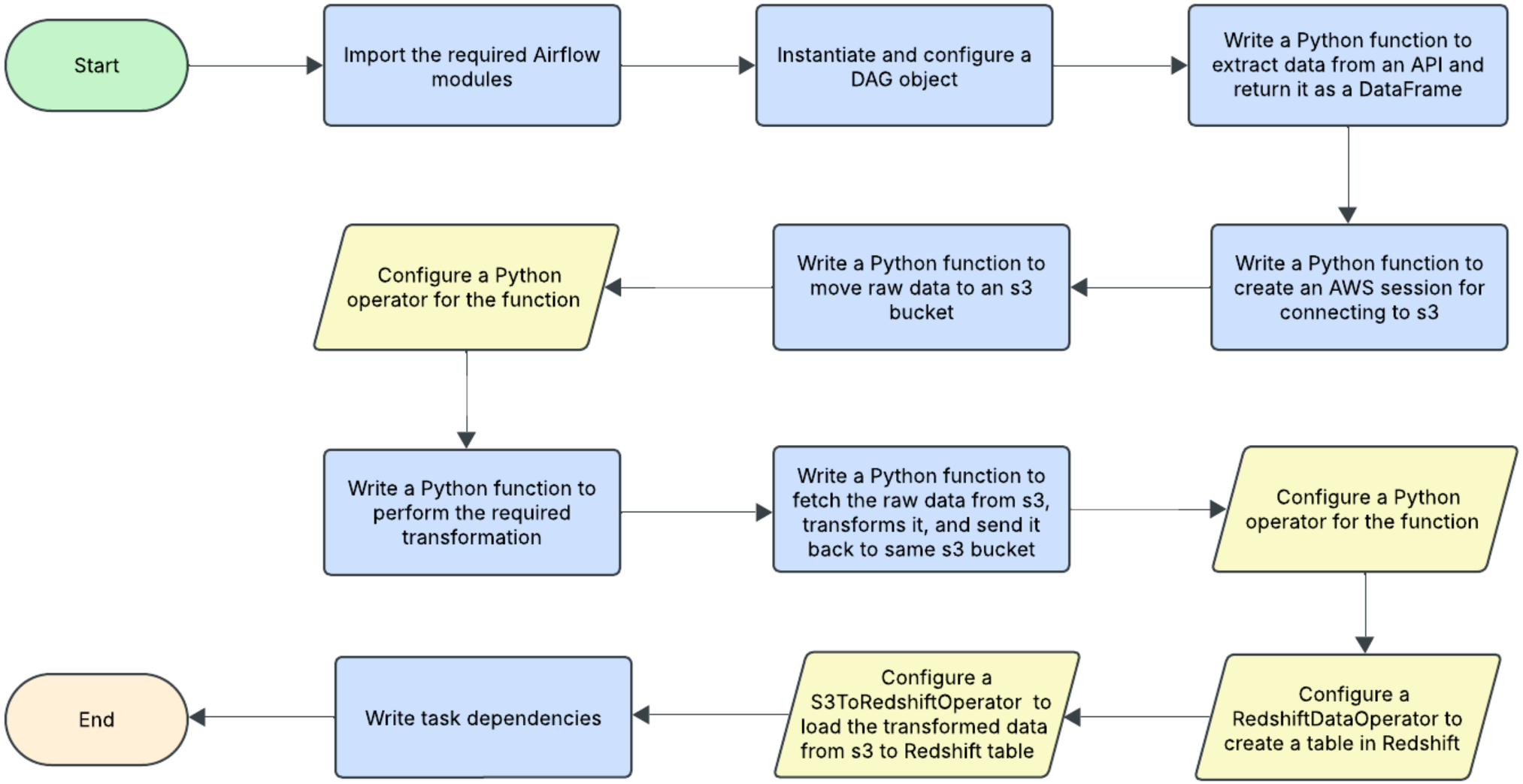
Travel Agency Architectural Diagram

Chisom Nnamani



Travel Agency Orchestration Flow Chart

Chisom Nnamani



Chisom Nnamani

Choice of Tools

Infrastructure as a Code (IAC)

HashiCorp Terraform

Terraform

Purpose: Used for Infrastructure as Code (IaC) to provision and manage cloud resources like AWS S3, Redshift, IAM roles, and VPC.

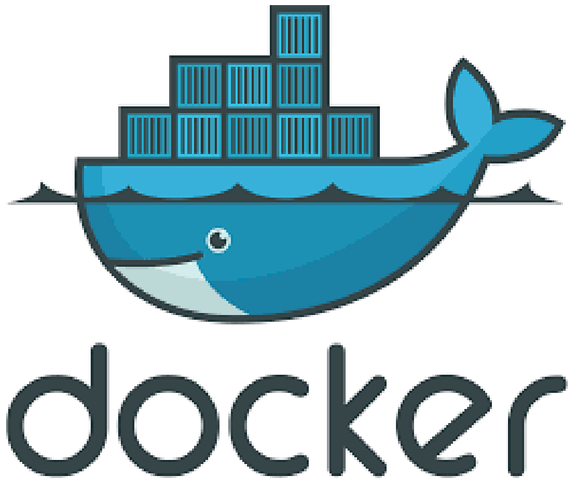
Why Terraform?

- It ensures scalability, consistency, and reproducibility in infrastructure deployment.
- It helps avoid the manual creation of resources which leads to a waste of time

Chisom Nnamani

Choice of Tools

Containerization Platform



Docker

Purpose: Used to containerize Airflow by building from an Apache Airflow Image found here. It was also used to build the app image (that contains the code for extracting and loading the raw data to Amazon s3).

Why Docker?

- It provides lightweight, portable, and consistent environments, making application deployment seamless across different systems.
- It is an open-source technology, making it cost-effective, widely supported, and adaptable for various use cases.

Choice of Tools

Chisom Nnamani

Orchestration Tool

Apache Airflow

Airflow

Purpose: Used as an orchestration tool to automate and manage the ETL pipeline.

Why Airflow?

- Scalable – It is an open-source technology and handles large data workflows efficiently.
- Flexible – Allows defining workflows as code, and provides a visual interface to monitor and manage workflows.
- Automated – Provides logging, alerts, and retries.

Choice of Tools

Chisom Nnamani

Data Lake

Amazon S3

Purpose: Used as the cloud-based Object Storage for the Data Lake to store both raw and transformed data.

Why Amazon S3?

- Highly durable and scalable storage system.
- Supports Parquet format for efficient data storage and query performance.
- Integrates seamlessly with other cloud and data-processing tools.

Choice of Tools

Chisom Nnamani

Container Registry

Amazon ECR

Purpose: Used to store Docker images for the packaged API extraction and data writing process.

Why ECR?

- Fully managed Docker container registry.
- Seamless integration with other AWS services and CI/CD pipelines.
- Supports secure and scalable image storage.

Chisom Nnamani

Choice of Tools

CI/CD

GitHub and GitHub Actions

Purpose:

- GitHub for source code management.
- GitHub Actions for CI/CD to automate code checks, builds, and deployments.

Why GitHub?

- Industry-standard version control system with extensive community support.
- GitHub Actions simplifies CI/CD pipeline setup, ensuring high-quality code and streamlined deployment.

Chisom Nnamani

Conclusion

This project lays the foundation for a robust, scalable, and automated data ecosystem that enables the Travel Agency to drive data-driven recommendations efficiently. Future enhancements will further optimize performance and expand analytical capabilities.