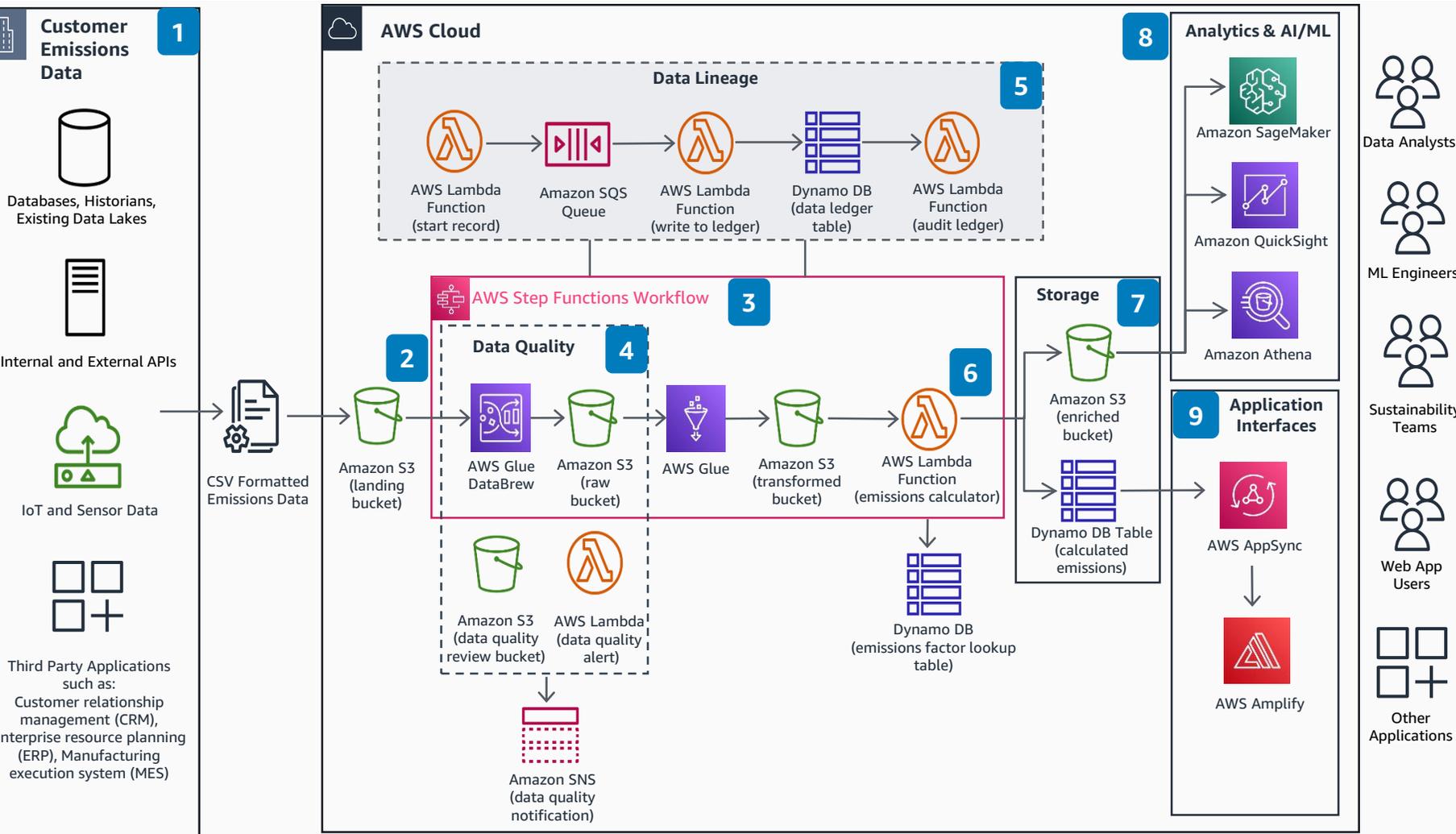


# Guidance for Carbon Data Lake on AWS

The Carbon Data Lake reduces the difficulties of ingesting, standardizing, transforming, calculating, tracing, and analyzing your carbon and greenhouse gas emissions data. A Carbon Data Lake provides a platform to build low-carbon products, solutions, and tools that leverage customer domain expertise and adds unique business value. Carbon Data Lake deploys a data pipeline, data quality check, data transformation ledger, greenhouse gas calculator, prebuilt analytics and machine learning tools, and a serverless web application.



- Customer emissions data from various sources is mapped to a standard CSV upload template. The CSV is uploaded, either directly to the **Amazon Simple Storage Service (Amazon S3)** landing bucket, or through the user interface.
- Amazon S3** landing bucket provides a single landing zone for all ingested emissions data. Data ingress to the landing zone bucket triggers the data pipeline.
- AWS Step Functions** workflow orchestrates the data pipeline including data quality check, data compaction, transformation, standardization, and enrichment with an emissions calculator **AWS Lambda** function.
- AWS Glue DataBrew** provides data quality auditing and an alerting workflow, and **Lambda** functions provide integration with **Amazon Simple Notification Service (Amazon SNS)** and **AWS Amplify** web application.
- Lambda** functions provide data lineage processing, queued by **Amazon Simple Queue Service (Amazon SQS)**. **Amazon DynamoDB** provides NoSQL pointer storage for the data ledger, and a **Lambda** function provides data lineage audit functionality, tracing all data transformations for a given record.
- A **Lambda** function outputs calculated CO2 equivalent emissions by referencing a **DynamoDB** table with Customer provided emissions factors.
- Amazon S3** enriched bucket provides data object storage for analytics workloads and the **DynamoDB** calculated emissions table provides storage for GraphQL API (a query language for users API).
- Customers can deploy a prebuilt **Amazon SageMaker** notebook and a prebuilt **Amazon QuickSight** dashboard with artificial intelligence and machine learning stacks, and business intelligence stacks. Deployments come with prebuilt **Amazon Athena** queries to query data stored in **Amazon S3**. Each service includes **Amazon S3** enriched object storage.
- Customers can deploy a Web Application stack that uses **AWS AppSync** for a GraphQL API backend to integrate with web applications and other data consumer applications. **Amplify** provides a serverless, pre-configured management application that includes basic data browsing, data visualization, data uploader, and application configuration.