AWS DATA ENGINEERING SERVICES (PIPELINE)

MAY 17, 2023

BYTEWISE LIMITED

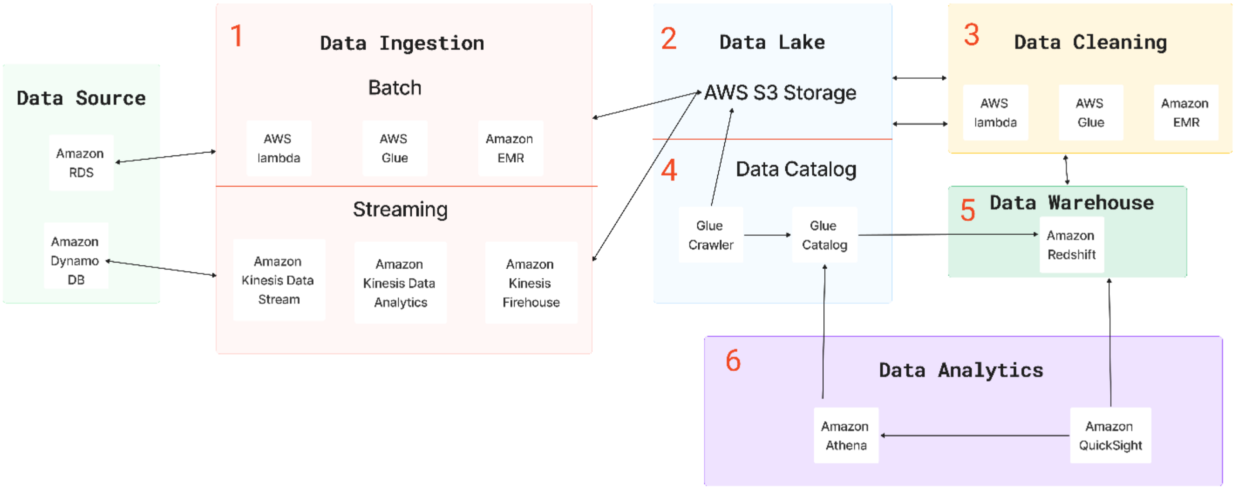

Fig 1.1 AWS Pipeline

Step 1: Ingest the source data from these application databases.

2 types of ingestion,

- **Batch ingestion**:
  - When you are bringing lot of data at once. Kind of Bulk ingestion to get all the historical data. So if tables are small then we sue database lambda to ingest this data.
  - Run code without thinking about servers.

  - ### AWS Lambda:
    - Is a computing service that runs code in response to events & automatically manages the computing resources required by that code
  - ### AWS GLUE:
    - Simple, scalable, and server less data integration.
    - Create, run, and monitor extract, transform, and load (ETL) pipelines to load data into your data lakes.
    - Discover and organize data.
    - Transform, prepare, and clean data for analysis.
    - Build and monitor data pipelines.
  - ### Amazon EMR:
    - You can choose your own open source big data framework.

- **Streaming ingestion**:

  Amazon kinesis:

  Analyze real-time video and data streams.

  - ### Amazon Kinesis Data Firehouse
    - Load data streams in AWS data stores.
  - ### Amazon kinesis Data Analytics
    - Process & Analyze streaming data using SQL.
  - ### Amazon kinesis Data Streams
    - Build custom applications that analyze data streams using popular stream-processing frameworks.

Step 2: Storage

- ### AWS S3:
  - Store all the files that we have ingested from our data sources.

- Data is in data lake raw state, not ready for consumption but analytics can be performed on it.

Step 3: Data cleaning

- Optimize storage format in s3 for querying.
- Process our data further using **AWS Lambda, AWS GLUE, AWS EMR.**
- Data is written back to S3 in a process zone.

Step 4: Data Catalog

- **AWS Glue Catalog**
    - Exists within AWS glue service & is central catalog of various data sets.
- **Glue crawlers**
    - It can scan data in all kinds of repositories, classify it, extract schema information from it, and store the metadata automatically in the AWS Glue Data Catalog

Step 5: Data Warehouse

- **Amazon Redshift**
    - Fast, simple, cost-effective data warehouse service
    - OLAP database designed to process large datasets

Step 6: Data Analytics

- **AWS Athena:**
    - Perform ad-hoc queries on data
    - Query data in Amazon S3 using SQL.
    - Integrates with glue catalog so you can perform SQL queries on table.
- **Amazon QuickSight:**
    - Create various graphs & charts.
    - Integrates with AWS Glue Catalog and AWS Athena to create dashboards.