Data Lake for Financial Services

Solution Request

Data engineer in a bank is tasked to create a centralized analytics platform using the data coming from various banking systems into a centralized data lake and analyze it in near real time. In order to query the data from the data lake he needs to catalog the data. Which AWS services are needed and how could the data engineer design the architecture?

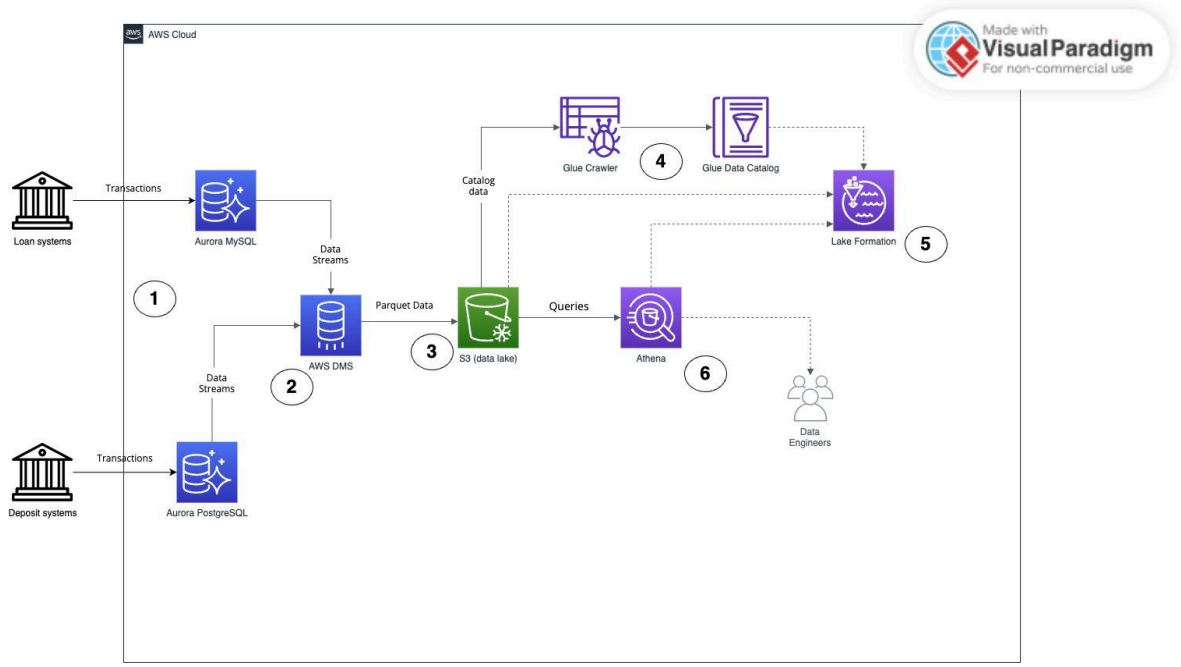
Main AWS Services to be used

The *core of the solution* would be to use AWS **Database Migration Service**, or AWS DMS, to replicate the data from core banking systems to an **Amazon S3** data lake in near real-time. AWS DMS can continuously monitor source databases and replicate any changes to the target data lake in Amazon S3. This way, the data in the data lake will be as fresh as possible, so it would be easy to extract insights that reflect the current state of banking operations.

For *cataloging and querying* the data lake we'll use **AWS Glue**. First, we'll set up an AWS Glue crawler to automatically scan the data in your S3 data lake and populate the AWS Glue Data Catalog with metadata about the data, such as table definitions and data types.

After the data is cataloged, you'll be able to use **Amazon Athena** to efficiently *query the data* in the data lake. Athena is a serverless, interactive query service that helps you use standard SQL to analyze the data without the need to set up any infrastructure.

Architecture



Explanation of the Architecture

This solution provides a centralized analytics environment that extracts insights from the bank's loan and deposit systems (and database) and combines the data into a data lake.

Data engineers can use the combined data in this data lake, located in Amazon S3, to run on-demand queries and obtain near real-time insights.

1. The solution starts with bank transaction data being housed in Amazon Aurora MySQL-Compatible Edition and Amazon Aurora PostgreSQL-Compatible Edition databases.
2. AWS Database Migration Service (AWS DMS) captures a copy of this data, from these multiple sources, in near real time.
3. The replicated data is stored, in Parquet format, to form the Amazon S3 data lake.
4. AWS Glue then catalogs the raw data, using an AWS Glue crawler, to create an AWS Glue Data Catalog.
5. AWS Lake Formation provides central access controls for the data in the data lake so that data engineers and applications have access to only what they need.
6. Amazon Athena can issue SQL-based queries against the data lake, using the AWS Glue Data Catalog. In this way, data engineers can use the data, combined from multiple sources into a data lake, to generate on-demand reports and near real-time insights.