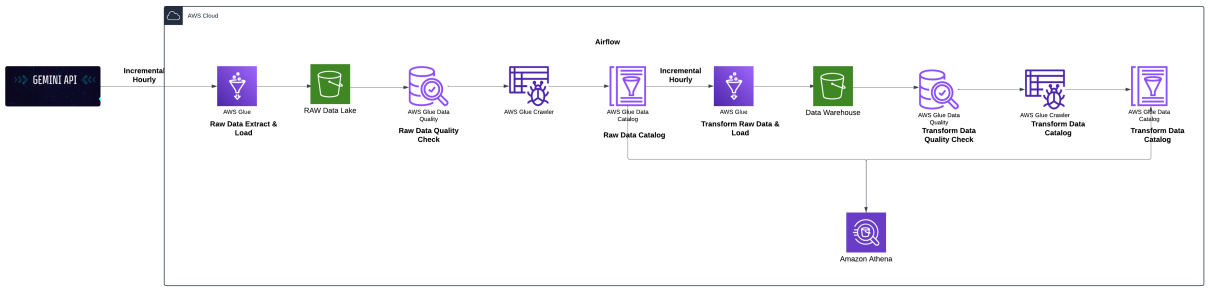
ETL Design

Batch Data Pipeline Design

I have designed the pipeline in AWS cloud using all the AWS services. I have used below components.

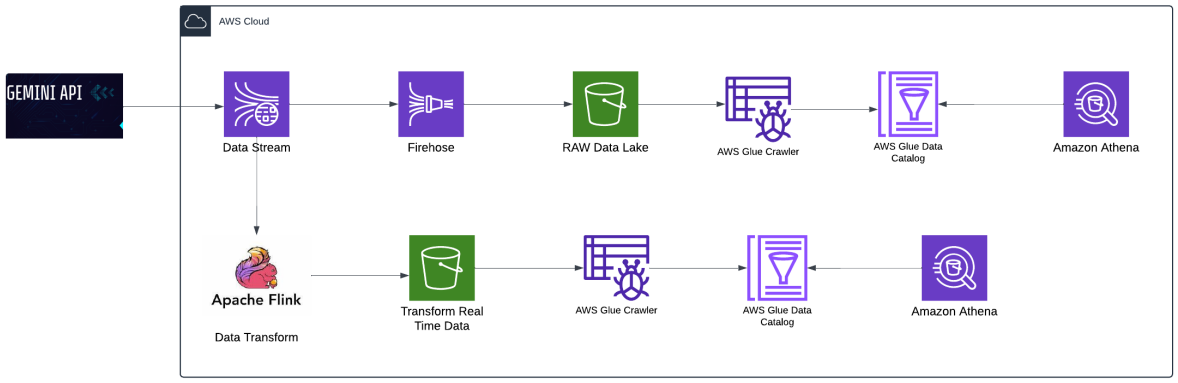
Frequency: Hourly
Pull Type: Incremental
Orchestration: AWS managed Airflow



1. AWS ETL Glue serve-less service for processing the data from Gemini exchange and write it to S3 bucket for raw data. It will be incremental pull and will be running every hour considering the volume and the mid size<less than a million records> volume can be pulled using Glue as it can easily scale without managing infrastructure. This stage, we will not manipulate the data.
2. Execute the Glue Data quality service for verifying the data quality.
3. Execute the Glue crawler for creating the Glue catalog. It can be access using AWS Athena for query using SQL.
4. We will run the another AWS ETL Glue job for transforming the data every hour and store in another S3 bucket. The RAW and transformed data bucket will be different.
5. Execute the Glue Data quality service for verifying the data quality.
6. Execute the Glue crawler for creating the Glue catalog. It can be access using AWS Athena for query using SQL.

Streaming Data Pipeline Design

I have designed the pipeline in AWS cloud using all the AWS services. I have used below components.



1. The AWS Kinesis stream service is used for streaming the real time data which can hold it for defined time before it loss on the producer.
2. The streamed data will be consumed by AWS Kinesis firehose for storing on AWS S3 in tabular readable format.
3. The same stream will be processed in Apache Flink for advance data manipulation and implementing the business logic in real time data which further store on Amazon s3.
4. Both the data will be cataloged using Glue catalog which can be accessed using Amazon Athena through SQL in near real time.

All the services scale up and down which is efficient to handle the large volume.