REAL-TIME DATA STREAM PIPELINE

PRESENTED BY TEAM THUNDERCLOUD TECHIES

PROJECT OVERVIEW

Introduction

AIM:

- The project aims to build a serverless data stream pipeline using Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose.

OBJECTIVE:

- To process, and load data into Snowflake, a popular analytic data warehouse, to achieve low-latency and near real-time data processing for actionable insights.

BENIFITS:

- Real-time insights, scalability, cost-effectiveness, and simplified data ingestion.

Key & Technologies

01 Amazon Kinesis Data Streams: Collect, process, and analyze real-time streaming data. Partitioned for scalability and parallel processing.

02 Amazon Kinesis Data Firehose: Ingest, transform, and load data into destinations such as Amazon S3, Redshift, Elasticsearch, or Splunk.

03 Snowflake: Automatically loads data from files into a designated stage, enabling continuous data ingestion.

04 Amazon S3: Used as intermediate storage for data before loading it into Snowflake.

Problem Statement

Organizations today generate large streams of real-time data from various sources, such as web applications, IoT devices, and social media feeds. Analyzing this data in real-time is crucial to gain timely insights, identifying abnormalities, and respond promptly to changing conditions. However, building a robust and scalable real-time data stream pipeline that can handle high-velocity data while ensuring low-latency processing presents significant challenges.

Solution Overview

The solution to the above problem involves building a real-time data stream pipeline using Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose, integrated with Snowflake. This pipeline will capture, process, and load streaming data into Snowflake for real-time analysis and reporting.

Solution Overview

Here is a high-level overview of the solution:

- **Data Collection:** The pipeline will continuously collect data records from various sources using Amazon Kinesis Data Streams. Data producers, such as web applications and IoT devices, will emit data records to the stream. Producers can utilize AWS SDK, Amazon Kinesis Agent, or third-party tools to send data to the stream.
- **Data Processing:** Amazon Kinesis Data Streams applications deployed on Amazon EC2 instances will read and process the data records in real time. The Kinesis Client Library will ensure fault-tolerant and scalable consumption of data.

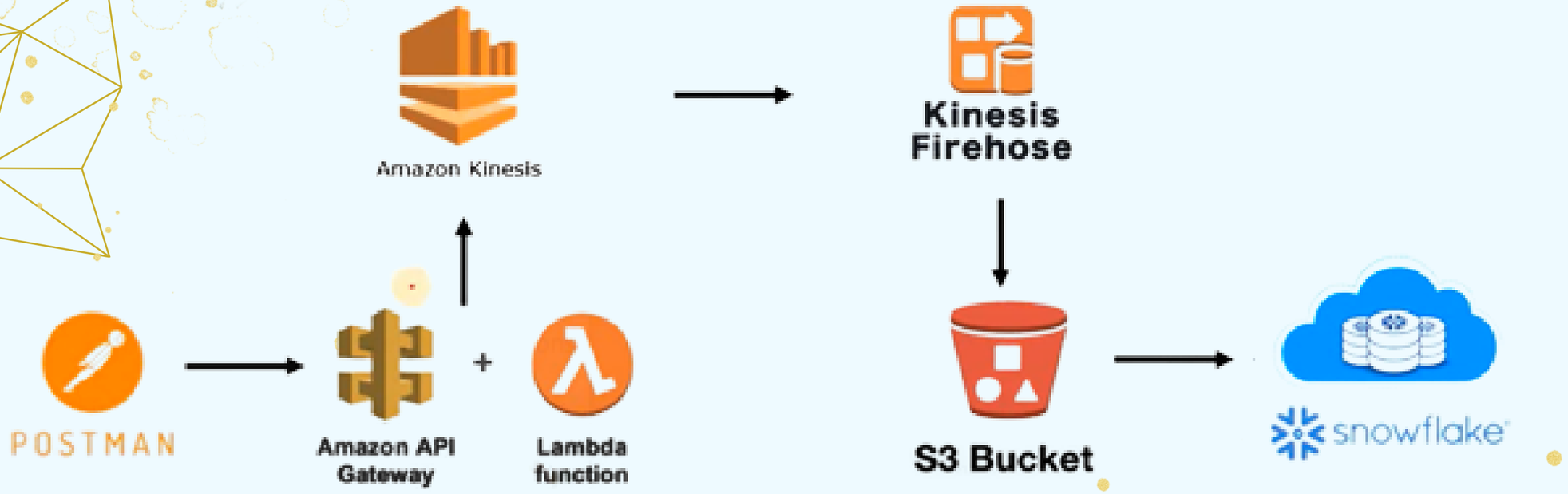
Solution Overview

- **Data Delivery:** Processed data records will be delivered to Amazon S3 using Amazon Kinesis Data Firehose. Kinesis Data Firehose will handle data transformation, if required, before loading it into S3.
- **pipeline Integration: it** will be configured to monitor the designated S3 bucket for new data files. As soon as new data files become available, Snowpipe will automatically load the data into the designated Snowflake stage.
- **Snowflake Loading:** Data loaded into the Snowflake stage by Snowpipe will be automatically ingested into the Snowflake database, enabling real-time data availability for analysis and reporting.

Workflow:

1. **Data Collection:** Data producers emit data records to Amazon Kinesis Data Streams using AWS SDK or other third-party tools.
2. **Data Processing**: Amazon Kinesis Data Streams applications read and process data records in real-time using Kinesis Client Library.
3. **Data Delivery:** Processed data records delivered to Amazon S3 using Amazon Kinesis Data Firehose, which also handles data transformation.
4. **pipeline Integration:** pipeline monitors S3 for new data files and automatically loads data into the designated Snowflake stage.
5. **Snowflake Loading:** Data is ingested into the Snowflake database from the Snowflake stage, enabling real-time data availability for analysis.

Project Architecture



Practical Presentation

Our Project

Understanding the Problem and Getting known to all services

Dividing the tasks Implementation

Integrating and completing the project

Testing the Project

THANKS AND REGARDS
- THUNDERCLOUD TECHIES