CSP-554 PROJECT DRAFT

DATA PIPELINE USING DYNAMODB AND AWS KINESIS

- Rajesh Kumar Bandaru (A20446254)
- Mohammed Jawhar(A20449684)
- Sai Charan Akkena(A20456762)

Section 1: Project Details

1. **Project Topic:** A Data Pipeline using DynamoDB and AWS Kinesis.

2. **Dataset**: Recordings from a Permanent Magnet Synchronous Motor (PMSM)

3. **About the Dataset**:

   The dataset contains readings of various components of the motor. The dataset was taken from Kaggle offered by Paderborn university. The readings are as follows,

   1. Ambient temperature.
   2. Coolant temperature.
   3. Motor speed.
   4. Current components.
   5. Voltage components.
   6. Stator yoke temperature.
   7. Stator tooth temperature.
   8. Stator winding temperature.
   9. Permanent magnet surface temperature.

   These readings are measured by appropriate sensors attached to motor. There are more than 990,000 readings taken over 72 sessions. The data is available as csv file.

4. **Project Goals**:
   - Create a DynamoDB table.
   - Ingest data into a DynamoDB Table.
   - An automated Data pipeline is created using AWS Cloud Formation.
   - Enable Kinesis Data Streaming for DynamoDB.
   - Transform the data as needed using lambda function and kinesis data firehose.
   - Analyzing the data using Amazon Athena if needed.
   - A QuickSight dashboard for analysis.

5. **Technologies**: Amazon DynamoDB, Amazon Kinesis Data Streams, Amazon Kinesis Data Firehose, Amazon Athena, Amazon QuickSight.

Section 2: Literature Review

- **Data Pipeline**

A data pipeline is used to move data from one end to another end transform data in between for analysis at the other end. It consists of a source, a data transformation step and a sink. In this project, the pipeline allows to move data from a DynamoDB table to Athena for querying and QuickSight for visual analysis. It is also called as a series of data processing steps. Streaming should be considered while building a data pipeline. So, we make use of AWS Kinesis services to move data from source to sink. We assume our data is generated from cloud and stored in DynamoDB table (source).

- **DynamoDB**

The data for this project is stored in DynamoDB table. DynamoDB is amazon's NoSQL database service. It delivers single - digit milli second performance at any scale. It's a fully managed, multi-region, multi-active, durable database with built-in security backup and restore and in-memory caching for internet scale applications. DynamoDB is serverless with no servers to provision, patch or manage and no software to install, maintain or operate. It automatically scales tables up and down to adjust for capacity and maintain performance. It supports ACID transactions for building business-critical applications at scale.

- **AWS Kinesis**

For collecting the data from DynamoDB and further processing AWS Kinesis is used. It is used to analyze real time data so that a timely insights can be taken and react fast as per the situation. The application logs of our data set can be ingested real-time for analysis in QuickSight. Amazon kinesis can ingest, buffer and process streaming data in real-time so that insights can be derived in seconds or minutes instead of hours or days. It can handle data from hundreds of thousands of sources with very low latencies.

- **Amazon Athena**

It is a query service offered by amazon to query data in S3 buckets interactively. The querying can be done in Standard SQL. It is a serverless service. The data streamed from DynamoDB is queried for necessary data needed for analysis in QuickSight. Athena uses presto with ANSI SQL support and works with various standard data formats. Athena is really fast as it executes queries in parallel, so more results will come back in less time.

- **QuickSight**

The data extracted from athena is just the data in another format. It should be converted into something visually soothing so that we can get insights from them. QuickSight does the exact same. It is a scalable, serverless service built for cloud. It can easily create interactive dashboards which are machine learning powered.

Section 3: Project Plans & Milestones

WBS	Task Name	Notes	Status
0	Downloading the Dataset		On Schedule
1	Setting up the environment	Creating various accounts necessary and installing required tools	Completed
2	Preparation of Project	Architecture Diagram	Completed
3	Ingesting Data	Loading Data into DynamoDB Table	Completed
4	AWS Cloud Formation	Pipeline Construction	InProgress
4.1	Kinesis Data Streams	Data Streaming from DynamoDB table	InProgress
4.2	Kinesis Data Firehose	A gateway or a pipe to throw the streaming data to S3 bucket	Future Task
4.3	Lamdba function	For the data to be readable by Athena the data from firehose is converted by adding a EOL character	Future Task
5	Athena Queries	Necessary data is extracted by using some queries	Future Task
6	QuickSight Dashboards	Data is transformed into various charts to gain insights	Future Task

Section 4: References

[1] Dataset: https://www.kaggle.com/wkirgsn/electric-motor-temperature

[2] P. O Donovan, K. Leahy, K. Bruton, "An Industrial big data pipeline for data-driven analytics maintenance application in large-scale smart manufacturing facilities" Published in Springer Journal of Big Data, Article Number: 25, Year: 2015.

[3] Eveline van Stijn, David Hesketh, Yao-Hua Tan, Bram Klievink, Sietse Overbeek, Frank Heijmann, Markus Pikart, Tom Butterly, "The Data Pipeline" Conference paper for the United Nations Global Trade Facilitation Conference 2011.

[4] About Data Pipeline: https://hazelcast.com/glossary/data-pipeline/

[5] Stream Processing: https://hazelcast.com/glossary/stream-processing/

[6] How to build a Cloud Formation Stack: https://aws.amazon.com/blogs/devops/automated-cloudformation-testing-pipeline-with-taskcat-and-codepipeline/

[7] Design Strategies for building a Data Pipeline: https://medium.com/@mrashish/design-strategies-for-building-big-data-pipelines-4c11affd47f3