**Integrated IoT Data Processing and Analysis Pipeline using AWS and Azure**

**Team members**

Sandeep Kandagatla #700761272

Pulligila Sai Sudheer Chary #700760830

Khareedu Gopi #700756699

Katta Karthik #700757117

Medasani Likhitha #700756468

Presentaion URL: <https://drive.google.com/file/d/1geu4_TNGTG4Sh5mJr9TR96v7gcydt4-M/view?usp=sharing>

**Table of Contents:**

1. Overview
2. Project Objectives - Rationale
3. Cloud service providers and services
4. Architecture and components
5. Detailed System Architecture
6. Conclusion
7. **Overview**

There has been a rise in the usage of Internet of Things (IoT) devices in recent years which has brought about innovative technologies across all industries and the transport sector is not left behind. Notably, the development of IoT technology has significantly impacted transit vehicle management which stands to benefit greatly from real-time data gathered by sensors and other gadgets on board for better operations, safety, and overall efficiency purposes. Nonetheless, taking full advantage of this wealth of transit vehicle IoT data calls for the establishment of a strong, adaptable data processing and analysis pipeline.

The project purpose is to address this gap with an integrated solution that processes seamlessly as well as analyzing transit vehicle IoT data using the cloud services by Amazon Web Services (AWS) collaborating with Microsoft Azure. The idea here is to build a total solution based on the best cloud providers today so as to match the requirements for modern mobility systems.

The main objective of this project is to build a cloud application using resilient design patterns that can ingest process or analyze transit vehicle related data. This approach will ensure organizations are able to manage and use various aspects of public transport information such as logistics optimization and transportation management among others.

In today's fast-paced, data-centric landscape, the capacity to harness and derive insights from data is increasingly vital for organizations striving to maintain a competitive edge and respond swiftly to dynamic market conditions. This project addresses this imperative by proposing a comprehensive cloud-based solution for managing vehicle transit data. The following sections delve into the project's scope, objectives, and significance, providing a detailed overview of its aims and potential impact.

**﻿Scope of the Project**

The scope of the assignment encompasses the whole lifecycle of actual-time vehicle transit statistics, from ingestion to evaluation. It includes the design, implementation, and deployment of a scalable and records pipeline that could take care of high volumes of data with minimal latency. Key elements of the undertaking scope consist of:

**Data Ingestion:** Developing mechanisms for securely consuming automobile transit records from various resources, inclusive of IoT devices and sensors, into the cloud environment.

**Data Processing:** Implementing data records processing with good judgment to cleanse, validate, and transform incoming data streams, ensuring statistics are pleasant and integrity.

**Data Storage:** Implementing real-time data processing logic to cleanse, validate, and transform incoming data streams, ensuring data quality and integrity.

**Data Analysis:** Making the data structured in SQL database for further integration with analytics tools for getting actionable insights.

1. **Project Objectives (Rationale):**

﻿One area profoundly impacted with the aid of development is transit vehicle control, in which the information generated by way of onboard sensors and devices presents a wealth of possibilities for optimization and enhancement. However, in spite of the full-size capability of this information, many groups struggle to effectively harness and put it to use due to the dearth of a robust and scalable data processing and analysis pipeline.

The primary motivation behind this project lies in addressing this essential gap by way of providing an incorporated answer that seamlessly processes and analyzes transit vehicle IoT data. By leveraging the blended talents of cloud computing, we intend to broaden a comprehensive platform able to ingest, processing, and analyze transit data successfully and effectively.

The primary objectives of the project are as follows:

* Aims to optimize operations, enhance safety, and improve efficiency through real-time insights and compliance adherence.
* Develop an end-to-end data processing pipeline capable of handling transit vehicle IoT data efficiently.
* Enable seamless integration between AWS and Azure services to facilitate data transfer and processing.
* Implement robust data validation mechanisms to ensure the integrity and quality of the processed data.
* Store validated data in Azure SQL Database for further analysis and reporting.
* Ensure high availability, scalability, and security of the entire system to meet the demands of a production environment.
* **Improved Data Management:** Centralizing transit vehicle data in Azure allows for better data governance, easier access, and simplified analytics.
* **Cost Optimization:** Leveraging Azure services can potentially reduce storage and processing costs compared to on-premise solutions.
* **Enhanced Scalability:** The cloud-based architecture can seamlessly scale to accommodate growing data volumes from the transit vehicles.
* **Streamlined Data Validation:** The Azure Function automates data validation, ensuring data quality for downstream applications.

1. **Cloud service providers and services**

**AWS Services:**

* AWS API Gateway
* AWS S3
* IAM Role
* IAM Policy
* CloudWatch

**Azure Services:**

* Azure Data Factory
* Azure Data Lake Storage Gen2
* Azure Blob Trigger Function
* Azure SQL
* Azure Linked Service
* Azure Key Vault
* Application Insights

1. **Components and Features**

**AWS Services:**

* **AWS API Gateway:** Facilitates the upload of transit data to AWS S3 securely.
* **AWS S3 (Simple Storage Service):** Stores transit data in object storage format, ensuring durability, availability, and scalability. Stores the raw transit vehicle IoT data in a structured folder hierarchy.
* **IAM Role (Identity and Access Management Role):** Manages access control and permissions for AWS services, ensuring secure data handling.
* **IAM Policy:** Defines permissions and access levels for IAM roles, ensuring granular control over data access.
* **CloudWatch:** Monitors and logs the performance of AWS resources, providing insights into system health and performance.

**Azure Services:**

* **Azure Data Factory:** Orchestrates data movement and transformation between cloud services, enabling efficient data processing workflows. Orchestrates the data movement between AWS S3 and Azure Data Lake Storage Gen2.
* **Azure Data Lake Storage Gen2:** Provides scalable and secure storage for big data analytics, allowing for efficient data ingestion and storage. Serves as the landing zone for the ingested data in Azure.
* **Azure Blob Trigger Function:** Triggers data processing logic upon the arrival of new data in Azure Data Lake Storage, ensuring real-time data processing.
* **Azure SQL (SQL Database):** Stores processed transit data in a relational database format, enabling structured querying and analysis.
* **Azure Linked Service:** Connects Azure Data Factory to external data sources, facilitating seamless data integration and movement.
* **Azure Key Vault:** Securely stores and manages cryptographic keys and secrets, ensuring data encryption and protection.
* **Application Insights:** Monitors and troubleshoots the performance of Azure applications, providing insights into application behavior and usage patterns.
* **Azure Function:** A serverless compute unit triggered by new data arrival in ADLS Gen2. The function validates the data format and schema.
* **Azure SQL Database:** Stores the validated and processed transit vehicle IoT data for querying and analysis.

**Major Features/Functions:**

1. **Data Ingestion and Storage:**

* Utilizes AWS API Gateway to securely upload transit data to AWS S3.
* Ensures data durability, availability, and scalability through AWS S3 storage.
* IAM Role and Policy manage access control and permissions for secure data handling.

1. **Data Processing and Transformation:**

* Azure Data Factory, cleansing, and enrichment processes to ensure data quality and integrity.

1. **Data Loading and Analysis:**

* Azure Data Factory loads processed transit data from Azure Data Lake Storage Gen2 into Azure SQL for structured querying and analysis.
* Azure Linked Service facilitates seamless integration with external data sources for comprehensive analysis.
* Application Insights monitors application performance, providing insights into system behavior and usage patterns for optimization.

1. **Security and Compliance:**

* Azure Key Vault securely stores cryptographic keys and secrets, ensuring data encryption and protection.
* IAM Role and Policy in AWS enforce access control and permissions, adhering to security best practices and regulatory requirements.

1. **Monitoring and Logging:**

* CloudWatch in AWS and Application Insights in Azure monitor and log system performance, providing real-time insights into system health and behavior.
* Enables proactive troubleshooting and optimization to ensure smooth operation of the cloud-based transit data solution.

1. **Detailed System Architecture**

A screenshot of a computer

Description automatically generated

The integrated IoT data processing and analysis pipeline designed for transit vehicle management leverages the capabilities of both Amazon Web Services (AWS) and Microsoft Azure to create a comprehensive solution. The system architecture comprises several components distributed across AWS and Azure cloud platforms. It begins with data ingestion from IoT devices and sensors onboard transit vehicles, followed by data processing, transformation, storage, and analysis. The solution ensures scalability, reliability, and security throughout the data lifecycle.

The process starts with securely ingesting real-time transit vehicle data from various sources, such as IoT devices, into the cloud environment in a JSON file format. AWS API Gateway facilitates the upload of data to AWS S3, ensuring durability, availability, and scalability. IAM Role and Policy manage access control and permissions for secure data handling during ingestion.

Once the data is ingested into AWS S3, Azure Data Factory orchestrates data movement and transformation between cloud services. Data processing logic is implemented using Azure Functions, ensuring real-time processing triggered by new data arrivals in Azure Data Lake Storage Gen2. Azure Blob Trigger Function is responsible for triggering data processing logic, ensuring seamless and efficient processing.

The processed data is then stored in Azure SQL Database, providing a structured format for querying and analysis. Azure Linked Service facilitates seamless integration with external data sources for comprehensive analysis. Application Insights monitors application performance, providing insights into system behavior and usage patterns for optimization.

Security measures are paramount throughout the system. Azure Key Vault securely stores cryptographic keys and secrets, ensuring data encryption and protection. IAM Role and Policy in AWS enforce access control and permissions, adhering to security best practices and regulatory requirements.

Both AWS and Azure offer monitoring and logging capabilities. CloudWatch in AWS and Application Insights in Azure monitor and log system performance, providing real-time insights into system health and behavior. This enables proactive troubleshooting and optimization to ensure smooth operation of the cloud-based transit data solution.

Overall, this integrated IoT data processing and analysis pipeline ensures efficient handling of transit vehicle data, enabling organizations to unlock valuable insights and drive informed decision-making for optimized operations, enhanced safety, and improved efficiency in transit vehicle management.

1. **Conclusion:**

The proposed project aims to deliver a scalable and efficient cloud-based solution for managing vehicle transit data. By leveraging the capabilities of AWS and Azure, the solution provides organizations with the tools to process large data in an efficient pipeline which can be used to unlock valuable insights from transit data and drive informed decision-making.