Data Lake Architecture Diagram

A diagram of data processing

Description automatically generated with low confidence

Data Ingestion:

1. IOT Sensors’ real-time data can be ingested using a combination of Amazon API Gateway and Amazon Kinesis Data Firehose. API Gateway acts as a front-end service and IOT sensors can send data to the API Gateway using HTTP. Amazon Kinesis Firehose can receive the data from API Gateway and deliver it to Amazon S3 for storage.
2. To ingest data from a database with historical records, we can use AWS Glue. AWS Glue can also perform necessary data processing/transformation and it can move the data to Amazon S3 for storage.
3. Third-party Supplemental data can be ingested using AWS Data Exchange. AWS Data Exchange can automatically deliver the data to Amazon S3 for storage.

Data Lake Storage:

Amazon S3 is the storage layer for the data lake. The IOT sensor’s real-time data, historical data from the database, and third-party supplemental data can all be stored in S3 as separate buckets.

Initial Data Cleaning and Processing:

As data is stored in Amazon S3, we can trigger a lambda function to perform initial data-cleaning tasks. AWS Lambda can be configured to run small functions that perform various data-cleaning tasks. The cleaned data is stored in S3 as separate buckets.

ETL / Data Processing:

One of the requirements is to use Apache Hadoop-based software for data processing. Hence, we use Amazon EMR for ETL (Extract, Transform, Load) and data processing. The processed/curated data can be stored in Amazon S3 and is ready for consumption.

Consumption:

Once the data is processed, it can be visualized using Amazon QuickSight. With Amazon QuickSight, we can create dashboards with curated data. Dashboards make it easy to analyze and gain insights from large data stored in the data lake.