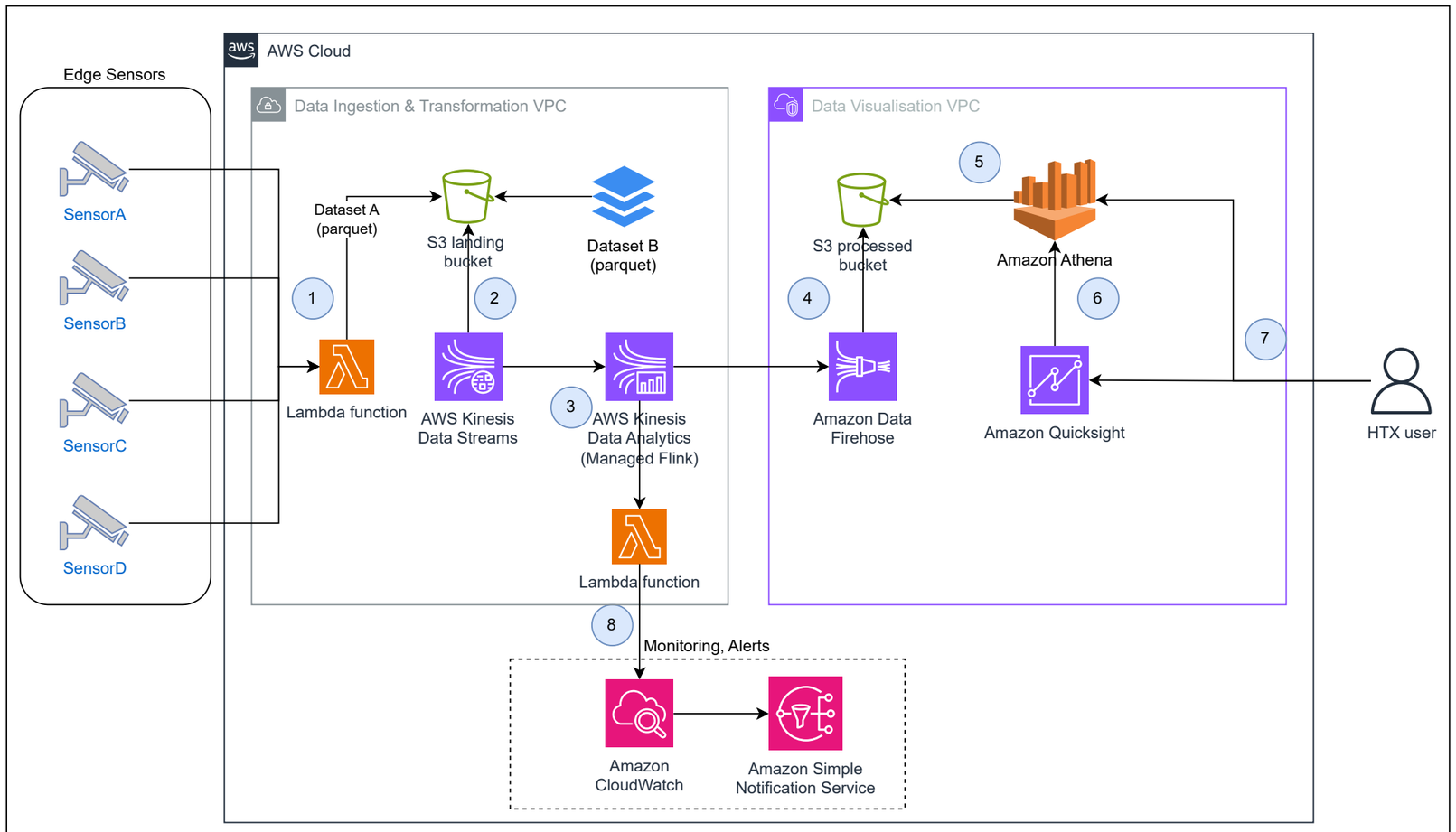
Amazon Web Service (AWS): Data Design Architecture



Data flow

1. Use AWS Lambda function to pull data from sensor API and create a parquet file (Dataset A)
2. AWS Kinesis Data Streams will ingest the data once the parquet is placed in the landing S3 bucket, which can be an updated Dataset B
3. A managed Flink service from AWS Kinesis Data Analytics will perform de-duplication for Dataset A and then join the Dataset B.
4. Processed data will then be send to a processed S3 bucket via the Amazon Data Firehose.
5. Amazon Athena is used as a query engine to directly access the processed S3 bucket by using SQL.
6. Amazon QuickSight is a visualisation tool that can create a data source with Amazon Athena to query the processed data S3 bucket and creates dashboards for users.
7. HTX users can access both the dashboards in Amazon QuickSight and also Amazon Athena if they want to run SQL query on the processed data in the S3 bucket.
8. A lambda function is created to send metrics, logs to Amazon CloudWatch for monitoring and alerting purposes with Amazon SNS

References:

1. *https://d1.awsstatic.com/architecture-diagrams/ArchitectureDiagrams/monitoring-streaming-data-with-mlv3-ra.pdf*
2. *https://aws.amazon.com/blogs/big-data/streaming-etl-with-apache-flink-and-amazon-kinesis-data-analytics/*

Assumptions & Questions

1. The architecture diagram is designed in Amazon Cloud (AWS).
2. The security of the architecture is not taken into consideration as this is a very high-level design.
3. The sensors' data can be pulled via API calls using a Lambda function. If the sensors has Kafka streaming services, then the lambda may not be used and can be ingested directly by AWS Kinesis.
4. The visualisation tool will be Amazon QuickSight and not other tools e.g. Tableau.
5. Will the records be delayed due to sensors has weak connectivity when sending data via API calls?

Tech stacks explanation

1. Lambda is used mainly to pull data from APIs of the sensors to be placed in S3 buckets.
2. Amazon Kinesis suite of services is used for real-time data ingestion and also transformation on-the-fly (using managed Apache Flink).
3. Amazon QuickSight is a native AWS visualisation tool that can ben used for simple dashboarding. It also uses Amazon IAM Identity Center to manage users' access.
4. Parquet file format is used because it is a columnar storage format and has the following benefits:
    1. Reduced I/O: because it is stored in columns, it will only read those columns that you require and do not need to read all the rows then select the columns that you need.
    2. It also has better compression that can be used, e.g. Snappy, ZSTD
    3. It is also a common big data storage format that is widely used by analytics tool e.g. AWS Athena, Google BigQuery, Microsoft Fabric