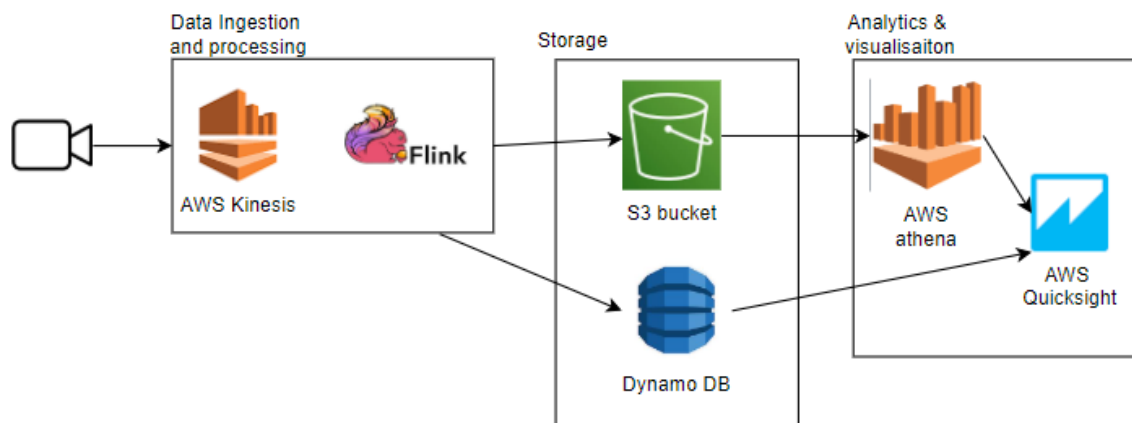


Questions to ask stakeholders:

1. What is the expected growth rate of the data ingestion?
2. How long should the data be stored?
3. Will there be a need to store unstructured data like images or videos from the sensors?
4. Is there a need to archive older data for historical analysis?



Dataset A is formed by storing data from all the video cameras. This is represented by the data ingestion handled by Kinesis accompanied with Apache Flink. As there are around 10000 events per second, Kinesis would be able to provide real time event streaming with low latency while supporting event replay in case of failures. Apache Flink will be responsible for removing any duplication from incoming events by filtering out retries and joining dataset A and B(referenced from S3). Any further transformation to the data can also be made here before storing it in DynamoDB.

Amazon DynamoDB is designed for high-throughput transactional processing, ensuring it can ingest and store data at a speed that matches Kinesis streams. As a real-time, low-latency storage solution, it meets the requirement that the joined results of Dataset A and B must be available as soon as events in Dataset A are published. With millisecond-level access times, DynamoDB enables instant retrieval of processed data, allowing dashboards to display the latest updates without delay.

Since DynamoDB is optimized for real-time data retrieval but not for historical data storage, older data will be offloaded to Amazon S3 for long-term storage. Dataset B, the static reference table, is also stored in S3 to support batch processing and historical queries. AWS Athena will efficiently query structured historical data from S3, enabling cost-effective and scalable analytics. The query results from Athena are then fed into

AWS QuickSight, ensuring accurate, interactive, and up-to-date visualizations that combine both real-time (DynamoDB) and historical (S3) insights.