Data Architecture Design

Architecture Overview

1. **Data Ingestion**:
    - Use Amazon Kinesis Data Firehose to capture streaming data from video camera sensors.
    - Ensure Kinesis Data Firehose is configured for high-throughput and low-latency.
2. **Data Processing**:
    - Use Amazon Kinesis Data Analytics (with Apache Flink) for real-time data processing.
    - Apply deduplication logic to remove duplicate events.
    - Join Dataset A (events) with Dataset B (static reference table) using a lookup table.
3. **Data Storage**:
    - Store joined results in Amazon Redshift for analytics and dashboarding.
    - Use Redshift's streaming API for real-time data ingestion.
4. **Data Visualisation:**
    - Use Amazon QuickSight for dashboarding and visualization.

Tech Stack Justification

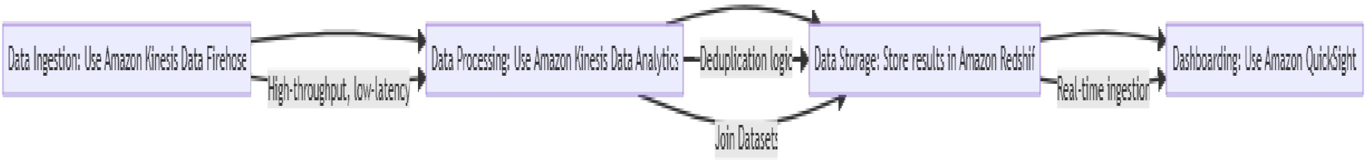
1. **Amazon Kinesis Data Firehose**: Scalable, low-latency, and high-throughput data ingestion service.
2. **Amazon Kinesis Data Analytics**: Fully managed service for real-time data processing with Apache Flink.
3. **Amazon Redshift**: Scalable, secure, and cost-effective data warehousing solution.
4. **Amazon QuickSight**: Fast, cloud-powered business intelligence service with Redshift integration.

Considerations and Assumptions

1. **Data Volume and Velocity**: Assume 10,000 events per second, with potential spikes.
2. **Data Quality**: Assume some duplicate events due to retries or other issues.
3. **Joining Logic**: Assume a simple join between Dataset A and B based on a common key.
4. **Dashboard Requirements**: Assume the PM wants a real-time dashboard with filtering and aggregation capabilities.
5. **Security and Compliance**: Assume necessary security measures are in place for data ingestion, processing, and storage.

Additional Considerations

1. **Monitoring and Logging**: Set up monitoring and logging for the pipeline using Amazon CloudWatch.
2. **Scalability**: Ensure the pipeline can scale with increasing data volumes using Kinesis Data Firehose and Redshift's scaling features.
3. **Data Retention**: Determine the necessary data retention period for compliance and analytics purposes.
4. **Cost Optimization**: Optimize costs by using Kinesis Data Firehose's and Redshift's pricing models.


Data Ingestion: Use Amazon Kinesis Data Firehose
High-throughput, low-latency
Data Processing: Use Amazon Kinesis Data Analytics
Deduplication logic
Join Datasets
Data Storage: Store results in Amazon Redshif
Real-time ingestior
Dashboarding: Use Amazon QuickSight

Questions for the End User

1. What is the expected data volume and velocity growth rate?
2. Are there any specific join logic requirements or complexities?
3. What is the desired latency for data availability in the dashboard?
4. Are there any specific dashboarding requirements or features needed?
5. Are there any data quality or data cleansing requirements?

Drafted by: Kader