**API6 to SDP Data Flow Architecture Document**

**Executive Summary**

This document outlines the architecture and data flow for capturing LLM model request/response logs through API6, streaming them via Kafka to the Strategic Data Platform (SDP), and distributing them to SIP (AIF) tenant-specific Hive databases .

**1. System Overview**

1. **Capture Phase**: API6 intercepts all LLM model traffic and generates structured logs
2. **Streaming Phase**: Logs are published to a dedicated Kafka topic in real-time
3. **Ingestion Phase**: Streaming process continuously syncs data from Kafka to Hadoop cluster (SDP) in SIP (AIF) temporary location.
4. **Distribution Phase**: Batch processor reads data from temporary location and writes to appropriate tenant Hive databases in SIP (AIF)

**2. Architecture Comparison**

**2.1 Approach A: Direct Streaming to Tenant Hive Databases**

**Flow**: Kafka Topic → Direct Stream → Tenant Hive Databases (SIP AIF)

**Advantages**

* **Lower Latency**: Data reaches tenant databases faster (near real-time)
* **Simplified Architecture**: Fewer components and transformation stages
* **Reduced Storage**: No intermediate staging storage required

**Disadvantages**

* **Limited Data Validation**: Difficult to validate or cleanse data before final write
* **Error Handling Complexity**: Failed writes require complex retry mechanisms
* **No Rollback Capability**: Difficult to correct data once written
* **Tenant Coupling**: Direct dependency on all tenant Hive databases being available
* **Performance Risk**: Slow tenant database can impact entire streaming pipeline
* **Security Concerns**: Direct access to production databases from streaming process
* **Missing Data Recovery**: Harder to replay or reprocess historical data

**2.2 Approach B: Staging Location with Batch Processing (Recommended)**

**Flow**: Kafka Topic → Stream → Staging Location (SIP AIF) → Batch Job → Tenant Hive Databases

**Advantages**

* **Data Quality Control**: Ability to validate, cleanse, and transform data before final load
* **Error Recovery**: Failed batches can be retried without data loss
* **Audit Trail**: Staging provides complete data lineage and history
* **Decoupling**: Streaming and database writes are independent operations
* **Flexible Scheduling**: Batch jobs can run during optimal windows
* **Data Replay**: Easy to reprocess historical data if needed
* **Better Monitoring**: Clear separation of streaming vs. loading metrics
* **Graceful Degradation**: System continues to collect data even if tenant databases are unavailable

**Disadvantages**

* **Higher Latency**: Additional time between data capture and availability (batch frequency dependent)
* **Additional Storage**: Requires staging area storage
* **Increased Complexity**: More components to develop, deploy, and monitor
* **Resource Overhead**: Both streaming and batch processes consume resources

**3. Key Assumptions and Considerations**

**Assumption 1:**

**SIP (AIF) Availability Timeline**

**Assumption**: SIP (AIF) will be available by November. New AIF provisioning requires significant lead time.

**Scenario Analysis**

**If SIP (AIF) is NOT available by November:**

**Temporary AIF with Migration Plan (Recommended)**

* Suggest one AIF as interim solution
* Data will be store in one database and later it will be migrated to tenant specific DB at SIP.

**Action Items**:

* Escalate SIP (AIF) provisioning request immediately
* Identify dependencies and bottlenecks in provisioning process
* Prepare contingency plan for temporary AIF deployment
* Document migration strategy from temporary to permanent solution

**Assumption 2:**

**Hive Database Mapping Integrity**

**Scenario**: API6 logs contain a column specifying the target Hive database, but the value may be:

* Incorrect/invalid database reference
* Missing/null database reference
* Database that doesn't exist yet

**Assumption 3:**

**Data stored in tenant specific AIF does not require any encryption.**

**Assumption 4:**

**For debugging any issues by using SDP stored logs , team need to check the logs store in different tenants Hive Database**