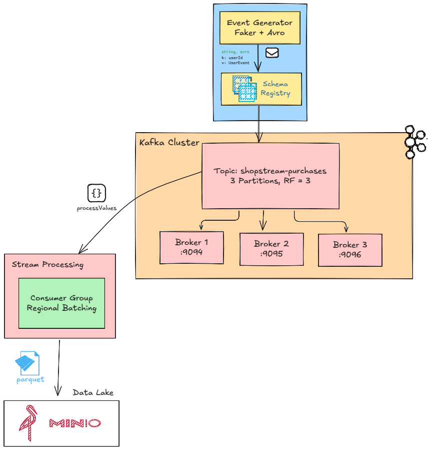
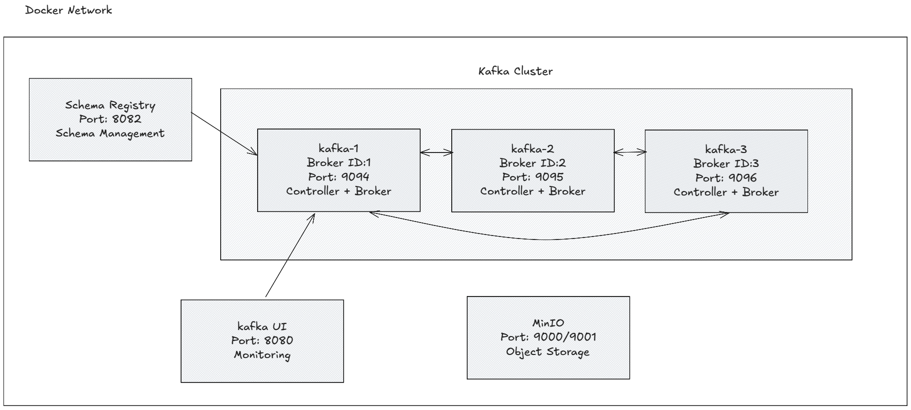
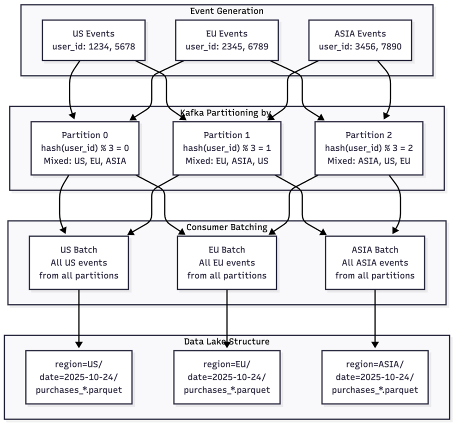


- Real-world e-commerce streaming pipeline simulation
- Demonstrates complete data journey from generation to storage
- Multi-technology integration (Kafka, Schema Registry, Object Storage)

- Event generation with realistic data
- Schema validation at entry point
- Distributed storage across 3 brokers
- Analytics-optimized storage





Event Generation

Three regions generate purchase events: US, EU, ASIA
Message key = user_id (not region) for ordering guarantees
Avro serialization with Schema Registry validation
Realistic data using Faker library

Kafka Partitioning

hash(user_id) % 3 determines partition assignment
All regions mixed across all 3 partitions
Same user always goes to same partition
Replication factor 3 = every message on every broker

Consumer Processing

Single consumer group reads from all partitions
Events reorganized by region field in message
Separate in-memory batches: US, EU, ASIA
Manual offset commits after successful storage

Batch Triggers

Size trigger: 50 events per batch
Timeout trigger: 30 seconds maximum wait
Balances efficiency with data freshness
Prevents memory overflow and stale data

Stream-to-Batch Conversion

Pandas DataFrame → PyArrow Table → Parquet
Columnar format optimized for analytics
Snappy compression for storage efficiency
In-memory processing without disk I/O

Data Lake Organization

Hive partitioning: region=US/date=2025-10-24/
Business logic grouping vs technical distribution
Enables partition pruning for efficient queries
Bronze layer for raw ingested data