Data Engineering Exercise: Real-Time Software Engineering Analytics

Background

You're part of a platform engineering team responsible for monitoring the health and performance of multiple microservices in a large-scale enterprise application (e.g., similar to Netflix, GitHub, or Adobe). Product teams rely on your metrics to detect service degradation, slow response times, or increasing error rates.

Objective

Build a data pipeline that captures service logs and metrics in real time, processes them, and provides dashboards with key reliability indicators (SLI/SLOs).

Simulated Dataset

Generate JSON-formatted service telemetry logs with fields such as:
{ "service": "auth-service", "timestamp": "2025-04-30T10:01:00Z", "status_code": 200, "latency_ms": 350, "error": false }

And a config file mapping service name to owning team, expected latency thresholds, and error budgets.

Tasks

1. Ingestion Layer:
- Simulate streaming logs using Python or Kafka
- Consume logs into a central processing function

2. Transformation Layer:
- Flag slow responses (e.g., latency > threshold)
- Mark error logs (status_code >= 500)
- Aggregate per-service metrics: average latency, error rate, 95th percentile

3. Storage Layer:
- Store in a columnar format (Parquet) partitioned by service and date
- Alternatively, use Prometheus for metrics storage

Data Engineering Exercise: Real-Time Software Engineering Analytics

4. Serving Layer:

- Visualize key metrics using Grafana or Streamlit
- Create alerts for SLO violations

Stretch Goals

- Create a status dashboard that shows green/yellow/red for each microservice
- Add auto-scaling simulation: Increase load and observe metrics
- Integrate with an orchestration tool like Airflow or Dagster

Deliverables

- Code simulating service logs and ingestion
- A short architecture diagram and README
- Dashboard screenshots or visualizations

Interview Value

- Prepares you for platform/infra roles in observability and SRE
- Helps demonstrate practical logging, monitoring, and data pipeline skills
- Aligns with real-world systems at high-scale companies like Google, Meta, and Netflix