Kafka Interview - Day 5 Revision Notes (May 2)

1. Difference between current offset and committed offset

Current offset is the next message the consumer will read. Committed offset is the last acknowledged (processed) offset. Lag = latest offset - committed offset.

2. Auto vs Manual Commit

Auto-commit is easier but risky; if crash happens after process and before commit, you get duplication. Manual commit gives control using commitSync/commitAsync, suitable for critical pipelines.

3. Lag Recovery Strategy

Use manual commit. Tune session.timeout.ms, heartbeat.interval.ms, max.poll.records. Retry failures with exponential backoff. Persistent state store helps recovery.

4. Fault Tolerance Best Practices

Use replication.factor >= 3, min.insync.replicas = 2, acks=all, and disable unclean.leader.election. Monitor ISR and lag. Enable idempotent producers.

5. Real-Time Pipeline Design

Kafka -> Spark Structured Streaming (1-min batch) -> Sink (Snowflake/S3). Consider checkpointing, schema evolution, end-to-end latency targets, topic partitioning, and throughput.