CAMPERSO Framework - Data Engineering System Design Cheat Sheet

1. **C - Capture (Ingestion)**

   Clarification Questions:

   - What is the incoming event volume per second?

   - Is it real-time (event-time) or batch (ingest-time)?

   - Are events pushed or pulled?

   - Is ordering important per user/session?

2. **A - Audit & Validate**

   Clarification Questions:

   - What data quality rules must be enforced?

   - Are schema changes expected frequently?

   - How do we handle invalid records?

   - Do we need validation at ingestion and after loading?

3. **M - Model (Schema Design)**

   Clarification Questions:

   - Should the model support historical changes (SCD2)?

   - What metrics are needed? Sessions, clicks, journeys?

   - Are we modeling for BI (denormalized) or storage optimized?

   - Is there a need for dimensional modeling (star/snowflake)?

4. **P - Persist (Storage & Format)**

   Clarification Questions:

   - Do we need ACID, schema evolution, or time travel?

   - What?s the retention period for raw data?

   - Do we query from Spark or SQL engines?

   - Do we need incremental loads?

5. **E - Expose (Query Layer)**

   Clarification Questions:

   - Are real-time dashboards required?

   - Who consumes the data ? BI analysts, apps, ML?

   - What are latency requirements (<1 sec or hourly)?

   - Are APIs needed to access processed data?

6. **R - Resilience (Fault Tolerance)**

   Clarification Questions:

   - What failure scenarios are critical to guard against?

   - Should we reprocess failed data or skip?

   - What?s the tolerance for delayed data?

   - Are retries and checkpoints required?

7. **S - Security**

   Clarification Questions:

- Is clickstream data considered PII (e.g., user_id)?

- Do we need to mask or hash identifiers?

- How do we manage access control to raw/curated layers?

- Are we compliant with GDPR or CCPA?

8. **O - Observability**

Clarification Questions:

- How will we detect data pipeline failures?

- Do we track freshness, lineage, and volume drops?

- Do we monitor schema changes over time?

- Should we be alert on processing delays?

YouTube.com/@FullStackMaster | roopkt@gmail.com

bitly.com/rupesh-kit | bitly.com/book-rupesh | bitly.com/bookwithrupesh