# Data Engineering Interview Flashcards (12-Problems Playbook)

## 1. CDC Upserts with Delta (MERGE + Audit)

- Delta MERGE on id + updated_at.
- Deletes: whenMatchedDelete(op='D').
- Audit log (id, op, updated_at, processed_at).
- MERGE = idempotent re-run safe.
- Extras: late data, schema drift (mergeSchema).

## 2. Skewed Joins / Shuffle Optimization

- Issue: skewed key dominates.
- Fix: broadcast small table, salting, separate hot keys.
- AQE auto handles skew (Spark 3+).
- Salt size ~ heavy_rows / target_rows_per_task.

## 3. File Ingestion & Validation (Bronze Layer)

- Checks: presence, size, schema, corruption.
- Good -> Bronze, Bad -> Quarantine.
- Schema drift: mergeSchema (nullable ok) vs strict.
- Metadata: ingestion_time, file_name.
- Idempotency: file hash + MERGE.
- Scale: Auto Loader for millions of files.

## 4. Incremental Processing (Silver Layer)

- Dedup: window -> latest updated_at.
- Merge: idempotent updates.
- Late data: watermark/SLA cutoff.
- Partition by event_date.

## 5. Orchestration & Monitoring (Airflow)

- Pipeline: Bronze -> Silver -> Gold.
- Best practices: max_active_runs=1, retries=2-3, SLA alerts.
- Backfill with catchup=True + run_date param.
- Observability: row counts, lineage, logs.
- CI/CD: Git + deploy DAGs.

## 6. Data Modeling & Warehousing

- Star schema (facts + dims).
- SCD: Type 1 overwrite, Type 2 new row with validity.
- Surrogate keys for dim joins.
- Perf: partition facts by date, Z-ORDER, pre-aggregates.
- Star vs Snowflake tradeoff.

## 7. Data Quality & Testing

- Checks: null PK, ranges, uniqueness, FK integrity.
- Tools: Great Expectations, Deequ, DLT expectations.
- Tests: unit -> integration -> end-to-end.

- Streaming: foreachBatch + watermark.
- Remediation: quarantine -> reprocess.

### 8. SQL Query Optimization

- Diagnosis: EXPLAIN, Spark UI.
- Optimize: partition (low-card), ZORDER (high-card).
- Other: column pruning, broadcast joins, compaction, summary tables.
- Skew: salting, AQE.
- Platforms: BigQuery partition+cluster; Redshift sort/dist keys.

### 9. File Validation & Guardrails

- Checks: presence, schema, size, corruption.
- Actions: quarantine, alert, audit logs.
- Schema drift: allow nullable, reject breaking.
- Late files: manifest table + SLA cutoff.
- Idempotency: staging overwrite + MERGE.
- Monitoring: ingestion dashboard.

### 10. DAG Reliability & Retry Logic

- Retries: 2-3, exponential backoff.
- Idempotency: MERGE, audit keys, unique IDs.
- Sensors: reschedule mode.
- Timeouts: execution_timeout, dagrun_timeout.
- Concurrency: max_active_runs=1, pools.
- Retries vs reschedules: retry reruns, reschedule waits.

### 11. Security & Governance

- Access: IAM, Unity Catalog, least privilege.
- Encryption: at rest (SSE/CMEK), in transit (TLS).
- PII: mask, hash, tokenize.
- Row-level vs column-level security.
- GDPR compliance: right-to-be-forgotten (delete/anonymize).
- Audit: logs + lineage.
- Non-prod: synthetic/masked data only.

### 12. Debugging & Production Issues

- Check Spark UI: stragglers, skew, OOM.
- Skew: salting, broadcast, AQE.
- OOM: repartition, filter early, more memory.
- Small files: OPTIMIZE/compaction.
- Delta issues: schema drift, concurrency, vacuum.
- Streaming: checkpoint corruption -> reset + replay.
- Proactive: monitoring, SLA alerts, quarantine bad data.