One■Page Data Engineering System Design – Interview Crib Sheet

ARCHITECTURE (Say First)

- Kappa■first + Medallion (Bronze immutable, Silver correct, Gold fast)
- Batch = bounded streaming
- Design for replay, not perfection

INGESTION

- At■least■once ingestion + dedupe = effectively exactly■once
- Capture event_time, ingest_time, metadata
- DLQ/quarantine for bad data

BACKFILL & REPLAY

- Immutable raw data
- Deterministic pipelines
- Partition/event■time backfills, not full recompute

LATE DATA & WATERMARKS

- Always event■time processing
- Dataset■level late tolerance (e.g., 48h)
- Beyond watermark → backfill workflow

CDC & SCD

- Bronze stores I/U/D with sequence
- Silver MERGE by key + latest seq
- SCD1 overwrite, SCD2 history

SCALABILITY LEVERS

- Parallelism: partitions, micro■batches
- Storage: partition + clustering
- Compute: broadcast joins, pre■agg
- Fix skew, avoid small files

RELIABILITY

- Idempotent writes
- Checkpoints + retries
- Fail = delayed data, not wrong data

METRICS TO SAY

- Freshness (data age)
- Completeness (expected vs actual)
- Correctness (quality pass rate)
- Lag, throughput, cost

MONITORING

- Alert on SLA breach
- Separate late vs wrong vs down
- Trend■based anomaly alerts

60■SECOND SCRIPT

“I ingest data immutably into Bronze with contracts and checkpoints. I process in Silver using event■time, watermarking, deduplication, and idempotent merges so the system is replayable and fault■tolerant. I publish Gold tables optimized per consumer and monitor freshness, completeness, correctness, and cost against SLAs.”