Single■Source Data Engineering System Design Interview Guide

This document is a complete, interview■ready reference for Senior, Staff, and Principal Data Engineer roles (Atlassian, SEEK, FAANG). It covers scalability, reliability, maintainability, efficiency, optimization, and explicit trade■offs at every stage of a modern data platform, with spoken answers and Databricks + Spark + Delta Lake mappings.

1. Ingestion

Scalability

Incremental ingestion and CDC ensure ingestion cost grows with change rate, not data size.

Reliability

Checkpointing and atomic Delta writes guarantee safe retries.

Efficiency & Optimization

Bound windows, avoid full scans, compact small files.

Trade■offs

CDC adds operational complexity; incremental loads risk late data.

Databricks Example

Spark Structured Streaming + Delta Change Data Feed.

Spoken Answer

> I scale ingestion by reducing scanned data and using CDC or incremental pipelines.

2. Error Handling

Scalability

Dead■letter tables isolate bad records and prevent cascading failures.

Reliability

Windowed deduplication avoids duplicate amplification.

Maintainability

Centralized error handling reduces ad■hoc fixes.

Trade■offs

Extra storage and monitoring overhead.

Databricks Example

Quarantine Delta tables for bad records.

Spoken Answer

> At scale, failures are normal—I isolate them instead of amplifying them.