Data Pipeline Design - Interview Cheat Sheet

Data Source Considerations

- Types: APIs, DBs, files, logs, IoT
- Formats: JSON, CSV, Parquet, Avro
- Volume: Real-time vs. batch loads
- Schema changes & compatibility
- Source SLAs and data freshness

Ingestion Strategy

- Batch vs. streaming (Kafka, Spark)
- Trigger mechanisms: CDC, pub/sub
- Partitioning strategies (time, region)
- Retry logic, backpressure, DLQs

Transformation & Enrichment

- ETL vs. ELT strategy
- Data normalization & cleansing
- Joins, enrichments, business rules
- Handling schema evolution

Data Quality & Validation

- Schema & type validation (e.g., GE, dbt)
- Nulls, duplicates, outliers
- Freshness & completeness checks
- Automated data tests

Security & Privacy

- Encryption at rest & in transit
- Access control (IAM, row-level)
- Data masking & anonymization
- Audit logs and classification

Orchestration & Scheduling

- Scheduling cadence: hourly/daily/event
- Task dependencies and DAGs
- Retries & idempotency handling
- Alerting and failure notification

Monitoring & Observability

- Track metrics: latency, volume, errors
- Lineage tracking tools (e.g., DataHub)
- Logging & dashboards (Grafana, ELK)
- Real-time health checks

Scalability & Performance

- Parallelism and partitioning
- Caching and materialized views
- Efficient data formats (Parquet)
- Compute/storage separation

Integration & Interoperability

- Tooling compatibility (dbt, Airflow)

- Schema versioning and APIs
- Avoiding vendor lock-in (Iceberg, Arrow)
- Cross-platform data sharing

Compliance & Governance

- Data retention & archival
- Metadata cataloging & tagging
- Regulations: GDPR, HIPAA, CCPA
- Approval workflows and audits