Production-Ready Data Pipeline Checklist

1. Orchestration & Reliability

1.1 Pipeline Scheduling & Orchestration

- Pipeline managed by reliable orchestrator.
- Clear DAG structure.
- Automated schedules.

1.2 Logging

- Structured logging with timestamps and IDs.
- Logs for start/end, row counts, inputs/outputs, errors.
- Centralized log storage.

1.3 Monitoring

- Track success rate, latency, data volume, cost.
- Dashboards available.

1.4 Alerting

- Alerts for failures, long jobs, SLO breaches, anomalies.

1.5 Retry Mechanisms

- Automatic retries, exponential backoff, retry limits.

1.6 Idempotency

- Safe re-runs, MERGE/UPSERT, global run IDs.

1.7 Dead Letter Queue

- Malformed records with metadata.
- Regular review.

1.8 SLO Compliance

- Defined and monitored SLOs.

2. Scalability & Performance

2.1 Resource Allocation

- CPU/memory limits configured.

2.2 Horizontal Scaling

- Supports scaling workers/executors/slots.

2.3 Incremental Processing

- CDC or timestamp logic.
- Partition pruning.

2.4 Performance Testing

- Stress tests and benchmarks.

3. Data Quality & Change Management

3.1 Schema Validation

- Automated type/column/nullability checks.

3.2 Data Integrity Checks

- PK uniqueness, referential integrity, valid ranges.

3.3 Data Drift Monitoring

- PSI, KS-test, baseline comparisons.

3.4 Schema Change Management

- Graceful handling of new/deprecated fields.

3.5 SCD Strategy

- Type 1/2/3 defined.

3.6 SCD Implementation

- MERGE logic, validity intervals, tests.

3.7 Data Versioning

- Versioned datasets and training data.

4. Security & Compliance

4.1 Secrets Management

- Stored in secret managers.

4.2 Access Control

- RBAC enforced.

4.3 Encryption in Transit

- TLS/SSL everywhere.

4.4 Encryption at Rest

- KMS-managed keys.

4.5 Compliance

- Masking, tokenization, retention policies.

5. Development, Testing & CI/CD

5.1 Unit & Integration Tests

- Transform tests and E2E tests.

5.2 CI/CD

- Automated lint, tests, deploy.

5.3 Runbook Documentation

- Common failures, debugging, backfill steps.

5.4 Data Lineage

- Source→Raw→Staging→Curated.

6. Cost Governance

6.1 Cost Monitoring

- Dashboards and alerts.

6.2 Storage Lifecycle

- Archive old data.

6.3 Compute Optimization

- Auto-shutdown, optimized queries.

7. Backfill & Reprocessing

7.1 Backfill Strategy

- Documented method, isolated runs.

7.2 Reproducibility

- Versioned data/code/config.

7.3 Checkpointing & Recovery

- Resume from last checkpoint.