Principal Data Engineer – Golden Design Checklist

1. Outcomes and Constraints

- Business goal and acceptance criteria
- Stakeholders and target consumers
- Non-functional requirements (latency, freshness, availability)
- SLA/SLO definitions
- Compliance and data residency constraints

2. Sources and Data Contracts

- Source types, ownership, and access method
- Volume, velocity, and retention
- Data contracts: schema, semantics, keys, nullability
- Contract versioning and compatibility rules
- Idempotency and ordering guarantees

3. Ingestion Design

- Batch vs streaming vs micro-batch decision
- Retries, deduplication, replay, backfills
- CDC strategy if applicable
- DLQ / quarantine patterns
- Delivery guarantees and compensations

4. Storage Architecture

- Bronze / Silver / Gold layering or data products
- File formats, compression, partitioning
- Table formats for schema evolution and time travel
- Warehouse vs lake responsibilities
- Retention and lifecycle policies

5. Processing and Transformation

- Compute framework selection and trade-offs
- Modular, versioned transformations
- Incremental processing and late data handling
- Dependency graph and partial reruns
- Backfill strategy

6. Data Modeling and Serving

- Dimensional vs wide vs event models
- Semantic layer and KPI definitions
- Serving patterns (BI, APIs, ML)
- Downstream consumer contracts

7. Data Quality and Correctness

- Schema, completeness, and integrity checks
- Reconciliation with source systems
- Drift and anomaly detection
- Bad data handling strategy

8. Observability and Operations

- Pipeline metrics and alerting
- Logs, traces, and lineage correlation
- Runbooks and on-call ownership
- Resilience and backpressure handling

9. Security, Privacy, and Governance

- IAM and least privilege access
- Encryption and key management
- PII masking and row/column security
- Audit trails and access workflows
- GDPR deletion strategies

10. CI/CD and Infrastructure as Code

- Environment separation and parity
- Automated testing and deployments
- Roll-forward vs rollback strategy
- Secrets management
- Terraform modules and drift detection

11. Cost Engineering

- Cost drivers by pipeline layer
- Query and compute optimization
- Autoscaling and concurrency limits
- Retention vs recompute decisions
- Budget alerts and cost visibility

12. Documentation and Change Management

- System diagrams and ownership mapping
- Data dictionary and lineage
- Operational documentation
- Deprecation and migration strategy
- Stakeholder communication plan