Conversation Summary

Topics Discussed

1. **Data Versioning Principles**
- Clarity of purpose (reproducibility, auditability, rollback, collaboration).
- Granularity (dataset, partition, row level).
- Immutability, metadata-driven tracking, logical vs physical separation.
- Storage efficiency, retention, compaction.
- Time travel and reproducibility.
- Atomic commits, governance, compliance, discoverability.
2. **Reference Architecture for Data Versioning (Iceberg on AWS)**
- Ingestion (batch, streaming, CDC).
- Storage on S3 with lifecycle management.
- Apache Iceberg as table format for snapshots, schema evolution, time travel.
- Glue Catalog as metadata store.
- Query engines (Athena, Spark, Trino).
- Governance: IAM, Lake Formation, audit logs.
- Monitoring: CloudWatch/Prometheus, compaction jobs, retention policies.
- Runbooks for restore, snapshot expiry, compaction.
- Retention policies, schema evolution, pitfalls and next steps.
3. **Architecture Diagram**
- Generated a clean diagram showing Iceberg tables, S3 storage, Glue Catalog, IAM, and versioned data.
4. **User Request for Downloadable Format**
- Created downloadable PNG diagram for the reference architecture.
5. **PDF Export**
- User requested summarization of the entire conversation into a downloadable PDF.

Key Outcomes

- User received principles and best practices for designing a data versioning strategy.
- A detailed reference architecture document for Iceberg on AWS was created.
- A visual architecture diagram was generated and shared.
- Conversation summarized into a structured document (this PDF).