DATA ENGINEERING CONSULTANCY FOR XYZ CORPORATION

Aryan Zodge, Isabella Recine, Joseph Cocozza
Group 5
July 12, 2024

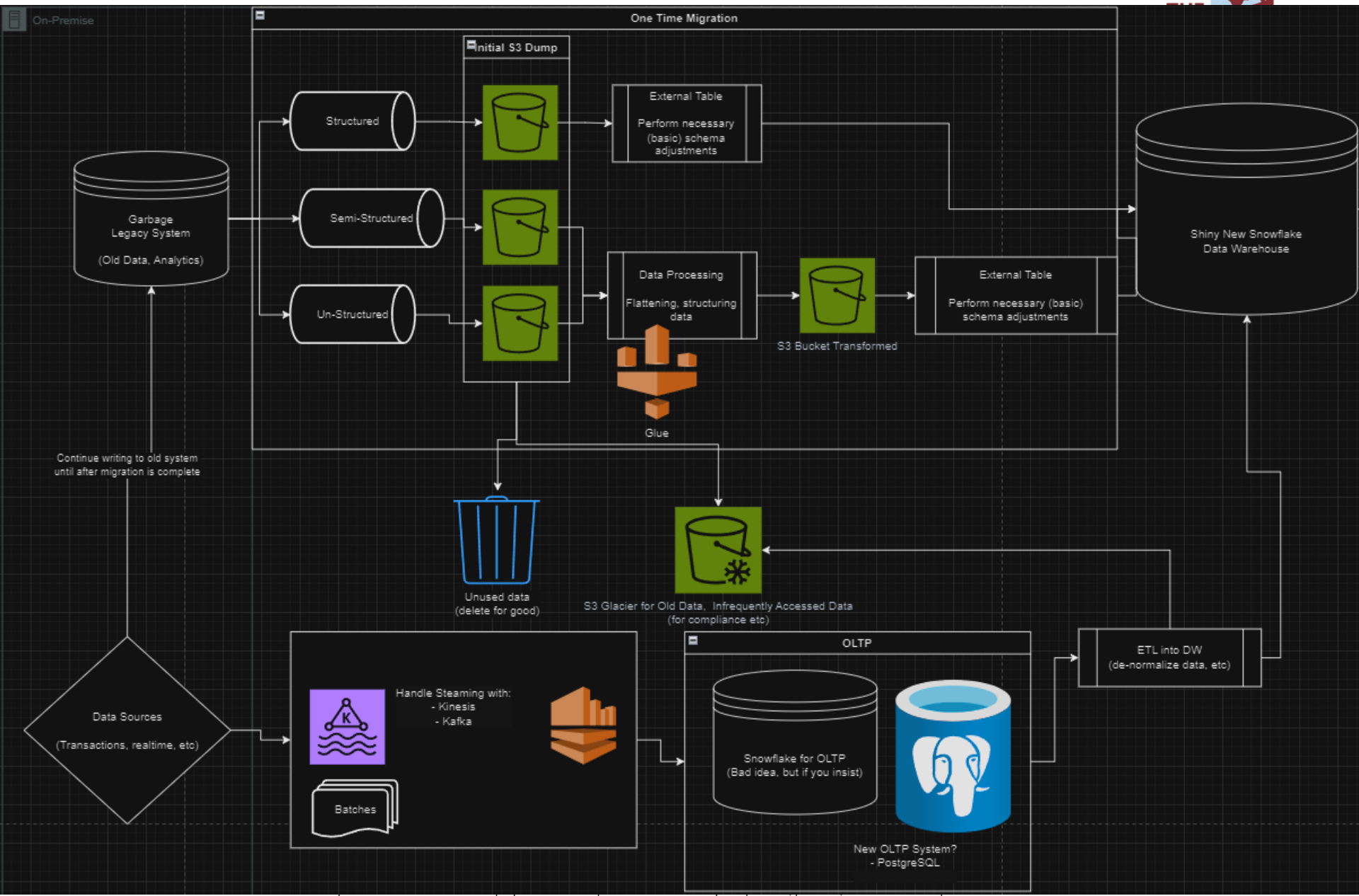
Executive Summary - Current Challenges

- Current Challenge:
  - Issues with their current data storage and management issues
  - Their existing infrastructure is built on a legacy on-premise data warehouse, which has led to inefficiencies, scalability issues, and high maintenance costs
  - Looking to modernize their data architecture and wants to learn more about snowflake

Executive Summary - Benefits of Moving to Snowflake

- Pricing
  - Minimum of $25 and up to $40 per terabyte of data stored in its system per month
    - Can be around $300 to $480 a year
  - Could be anywhere from $2 to $5.40 per credit for users on a standard plan
- Push all data to one location
  - Can mix corporate, sensor, social, and ecosystem data
  - With no service windows, you can run data engineering, data lake, data warehousing, and cyber-security workflows
- Loading data
  - You can load as much or as little data as you need, and do not have to worry about managing it
    - Largest customers have between 25-40 PB of data stored while others have around 1 TB of data
- Architecture
  - Build in the cloud so it uses cloud-native services that guarantee instant elasticity and scalability
    - Not on-premise technology adapted for the cloud

Architecture Diagram



Technical Proposal – Storage Solutions

- Internal Tables
  - Permanent Tables
    - Store critical frequently accessed data that needs to be highly available
    - Transactional data, customer records, financial information
  - Transient Tables
    - For temporary data that doesn't require the same level of durability as permanent tables
    - Staging and intermediate processing
  - Temporary Tables
    - For data that only needs to persist for the duration of a user session.
    - Analytical needs, results are only needed
- External Tables
  - To ingest or querying data stored outside Snowflake without loading to Snowflake
  - Historical records, logs, infrequently accessed data

Technical Proposal – Data Ingestion

- Bulk Loading
  - Use COPY INTO to load large volumes of data from external stages into Snowflake tables
- AWS Kinesis
  - To capture and process real-time data from OLTP databases
- S3
  - For landing and preprocessing

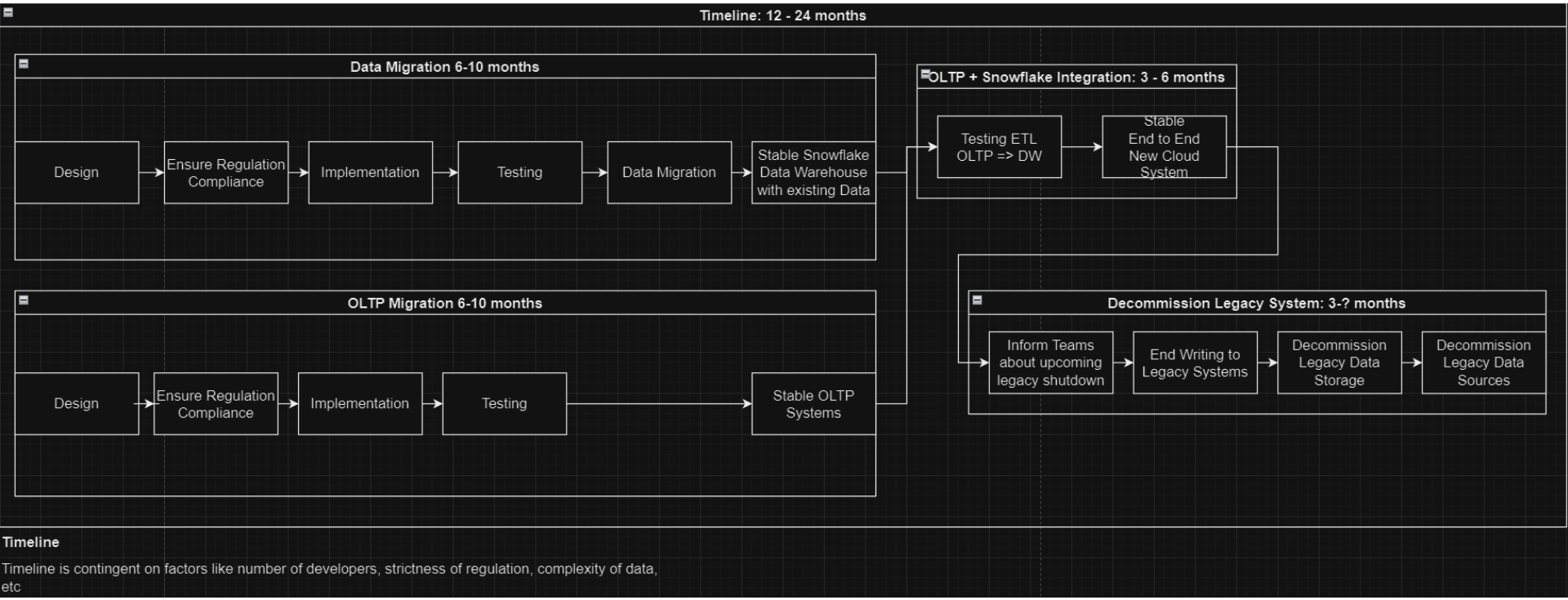
Technical Proposal – Optimization Techniques

- Define Clustering Keys
  - Used to improve query performance by reducing the amount of data scanned on large, frequently queried tables
  - Helps organize data based on specified columns to enhance query efficiency
- Materialized Views
  - Used to precompute and store complex queries
  - Provides fast query performance for repeated queries on large datasets, as the results are stored and updated incrementally
- Result Caching
  - Used to speed up query performance
  - Stores the result of previous queries allowing future similar queries to retrieve results without having to execute the query again

Technical Proposal – Data Security and Governance

- Built in data security features
  - Data encryption
    - Uses an algorithm to convert text characters into an unreadable format, scrabbling the data
    - Authorized users use a decryption key to be able to read the data
  - Data masking
    - Eliminates the need to store and manage multiple versions of the same data
    - Maks the sensitive information without changing the underlying data
      - Can make it so different roles have access to different data
  - Data loss prevention
    - Encompasses other security tools that spot abnormal usage
    - Recover data that has been stolen, corrupted, or lost via natural disaster
- Governance tools
  - Information schema
  - Object tagging
  - Access History
    - See what data exists, where it is, who has access to it and who has accessed it

Implementation Plan



Timeline

Timeline is contingent on factors like number of developers, strictness of regulation, complexity of data, etc

Implementation Plan – Potential Risks and Mitigation Strategies

- Data Quality Issues
- Performance Bottlenecks
- Data Security and Compliance Breaches