TresVista
Catalyzing Your Impact

# Data Warehouse

Log-Based CDC to Snowflake

# Data Warehouse

## Log-Based CDC to Snowflake

### Request and Guidelines Provided

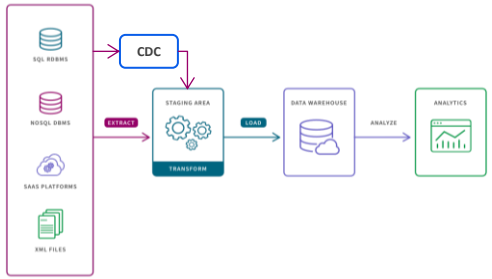
- Client: A leading North American Quant Hedge Fund
- Design a Data Warehouse to capture data across multiple sources like MySQL, Oracle, and MongoDB into Snowflake
  - Oracle has been the main database which is used by the entire firm, the ETL process should not impact the database performance
- The data from Snowflake would be further used by the reporting and analytics team to generate insights and share with the stakeholders

### Methodology and Final Deliverable

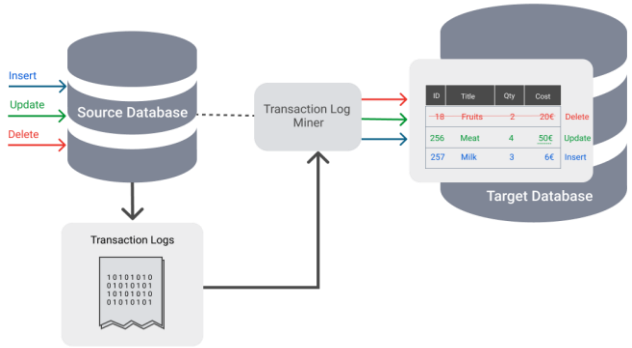
- Data models across the databases were studied (MySQL had company/employee data, Oracle had accounting data, and MongoDB had ticker-specific reviews & newsletters) and Snowflake schema was finalized
- PySpark-based ETL scripts were built for MySQL and MongoDB and log-based CDC scripts were built and scheduled for periodic updates
  - Log-based CDC was a highly efficient approach that limited the impact on Oracle DB with minimal/zero-downtime
- Multiple data and business-specific controls were added in the ETL scripts to maintain the quality and integrity of the data moving into Snowflake

### Output Snapshot

**ETL Process**




**Log-Based CDC**




**Tools/Technology used: MySQL, Oracle, MongoDB, Goldengate, PySpark, Snowflake**

salessupport@tresvista.com | www.tresvista.com