TresVista

Catalyzing Your Impact

Data Engineering & Cloud Computing

Case Studies

Data Engineering & Cloud Computing

Log-Based CDC to BigQuery

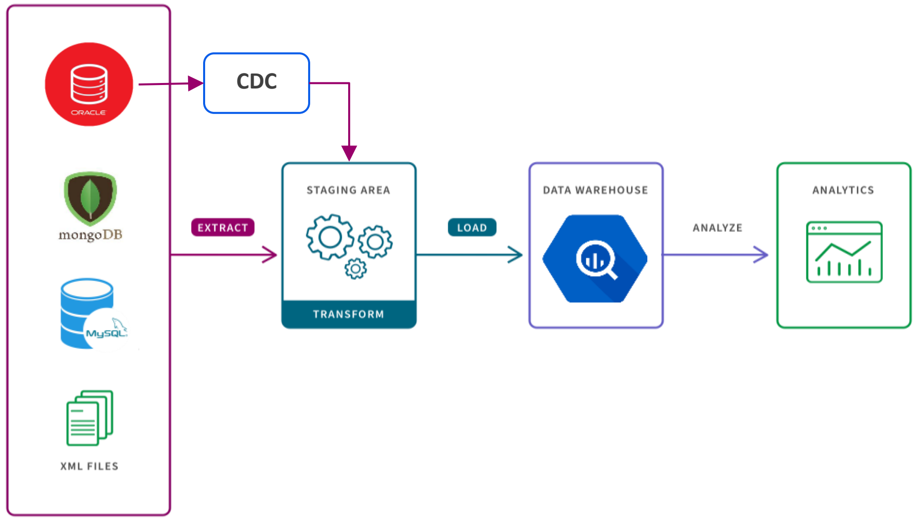
Request and Guidelines Provided

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

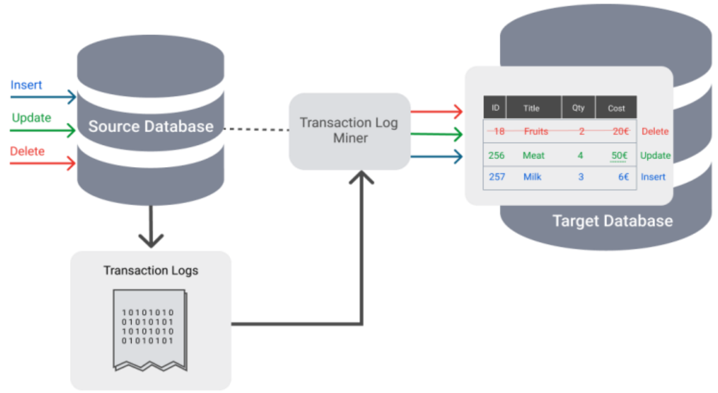
Methodology and Final Deliverable

- Data model across the databases were studied (MySQL had company/employee data, Oracle had tick level accounting data, and MongoDB had ticker specific reviews & newsletters) and BigQuery 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 BigQuery

ETL Process



Log-Based CDC



ID	Title	Qty	Cost	
~~18~~	~~Fruits~~	~~2~~	~~20€~~	Delete
256	Meat	4	50€	Update
257	Milk	3	6€	Insert

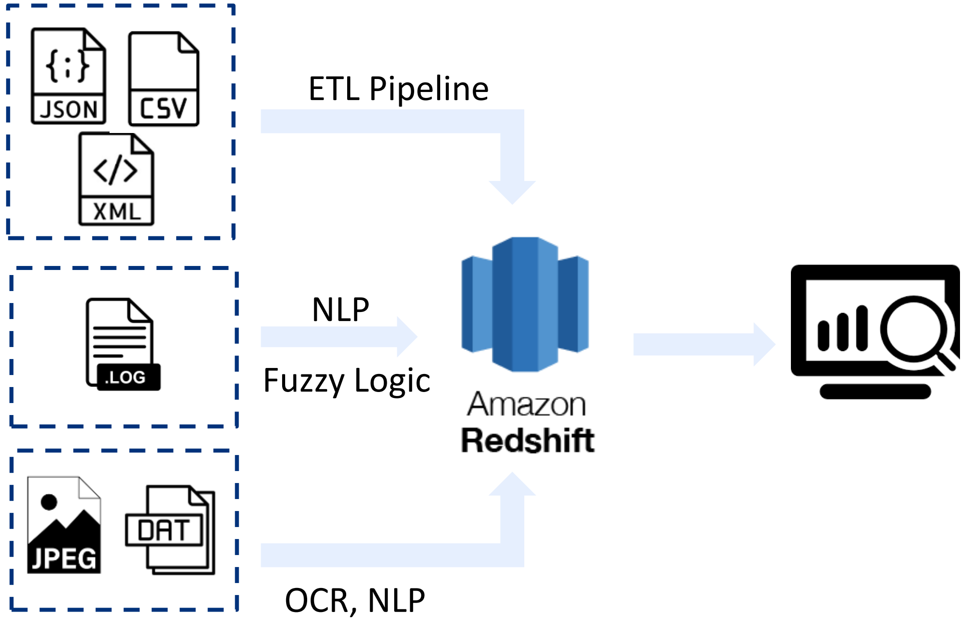
Tools/Technology used: SQL, PySpark, BigQuery

Data Engineering & Cloud Computing

Multiple Data Sources to Redshift

Request and Guidelines Provided

- Client: Private Equity Firm
- Create a centralized datawarehouse by,
  - Extracting data from different sources like flat-files, logs, images, etc.
  - Transforming the data into a standardized format based on business logics
  - Storing it for in-depth analysis



Methodology and Final Deliverable

- Extracted data of different types, such as CSV, .log files, PDFs etc. from multiple data sources like emails, deal documents, meeting logs etc.
- Performed data transformation process using a combination of business logic, NLP, Fuzzy Logic and OCR to extract specific deal details, sentiment of the meetings, text from images etc.
- Implemented a set of rule-based algorithms to structure the data based on specific patterns and created automated pipelines that integrates with the database management system to load the transformed data into Redshift
- Performed data analysis to gather insights to take data-driven decisions using visualization tools

Tools/Technology used: SQL, PySpark, Redshift

Data Engineering & Cloud Computing

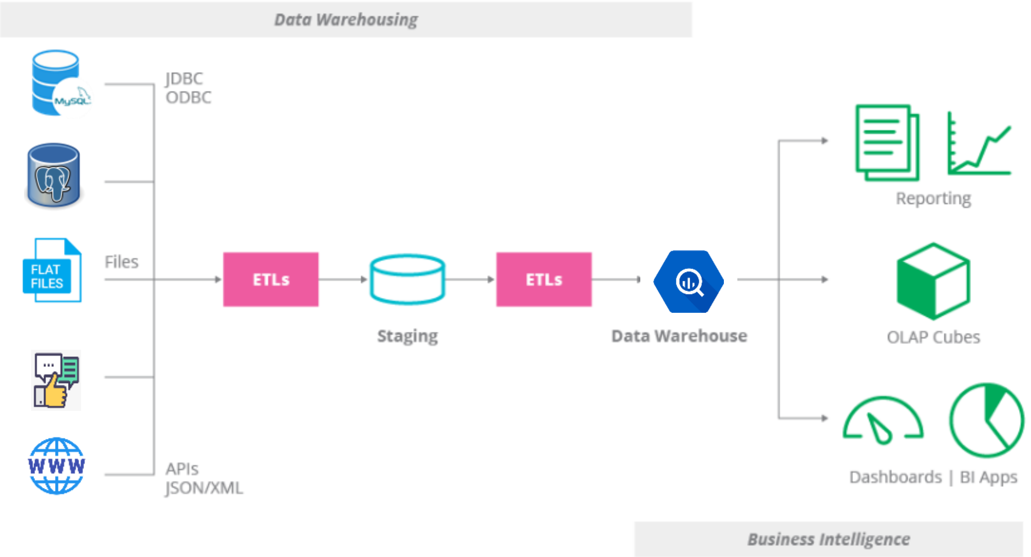
Multiple Data Sources to Redshift

Request and Guidelines Provided

- Client: A mid-sized perfume retail private company
- Currently, the data is spread across multiple databases depending on external stores, central office, e-commerce websites, etc. which is difficult to evaluate and generate insights
- Create a centralized datawarehouse by consolidating data across multiple databases which can be used to perform BI and Analytics
  - The datawarehouse solution should be read-only

Methodology and Final Deliverable

- Post client discussions, data model were built to set-up the BigQuery efficiently
- To maintain read-only datawarehouse, a staging area was set-up to store all the data for temporary period
- Developed simple python based ETL scripts to first move the data into staging area and periodically move it to BigQuery
- All the sales, inventory updates/modifications to the data was done within the threshold period of 1 month post which it was moved to BigQuery
- The datawarehouse was used by the BI and analytics team to generate insights that can be used to improve business



Tools/Technology used: SQL, Python, BigQuery

Data Engineering & Cloud Computing

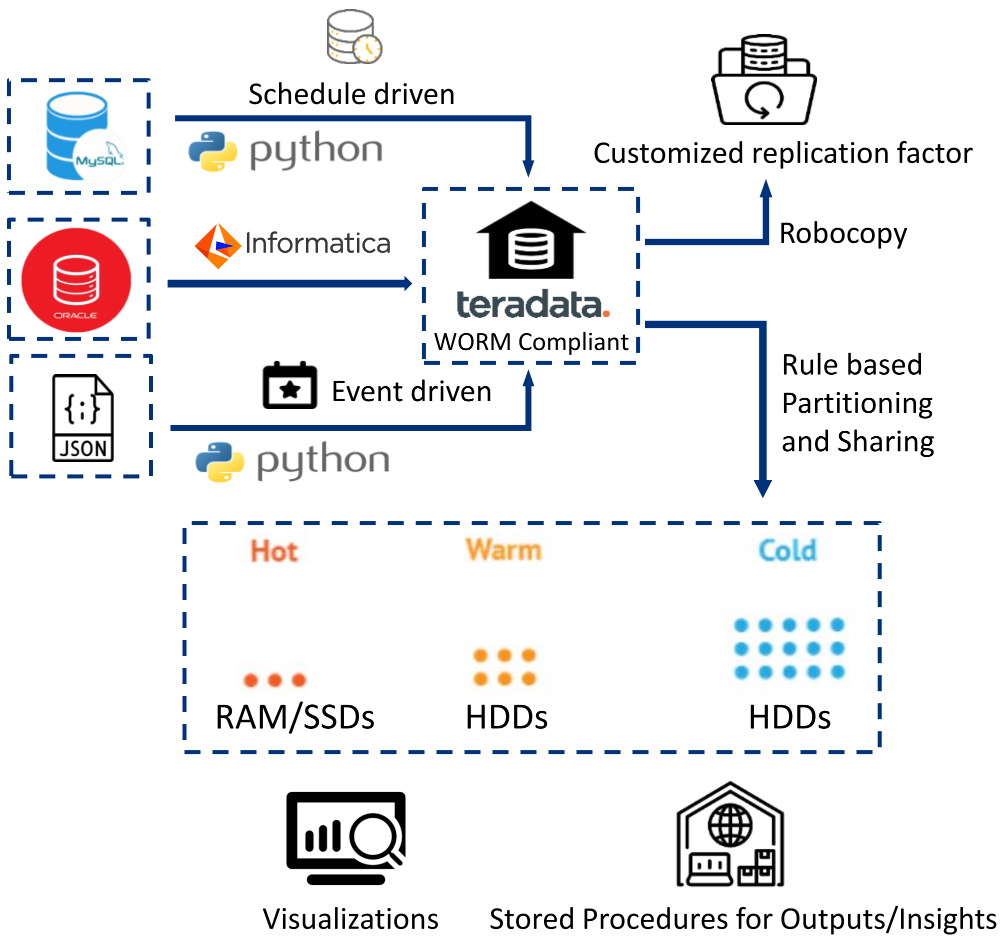
Multiple Data Sources to Teradata

Request and Guidelines Provided

- Client: Private Equity Firm
- Currently, the data is spread across multiple databases depending on their portfolio company and geographic regions, with no single consolidated source to monitor real estate transactions globally
- Consolidate the data across multiple databases enabling in-house BI and Analytics
  - The datawarehouse solution should be WORM compliant
  - Create a pipeline for data backup and storage, optimized as per nature of data
  - Data to be available for analytics services and programming

Methodology and Final Deliverable

- ETL pipelines were setup using python and Informatica to extract data from multiple sources (MySQL, Oracle, JSON files) and load in Teradata (Datawarehouse)
  - Data from MySQL is transferred to Teradata on a monthly basis, the database only stores last 6 months of data
  - Data from Oracle is transferred to Teradata when more than 80% of storage is exceeded, the transfer block size is 5 GB
  - Data storage was WORM compliant; partitions were created to make best use of resources and appropriate replication and backups were done with Robocopy
- The data warehouse and data marts were further utilized by the BI and analytics team to generate insights that can be used for better business decisions



Tools/Technology used: Teradata, SQL, Python, Informatica, Robocopy

Data Engineering & Cloud Computing

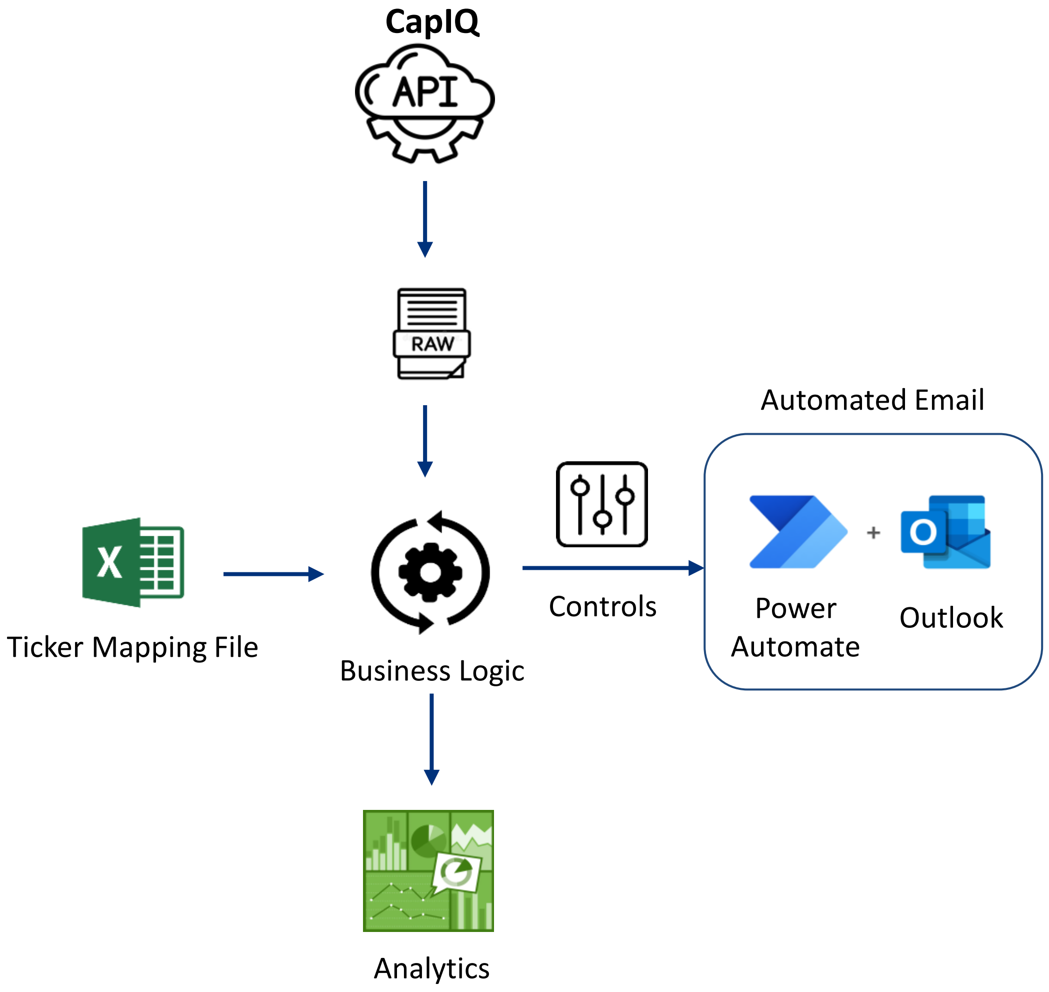
CapIQ API Data Pull with Controls

Request and Guidelines Provided

- Client: Hedge Fund
- Automate the manual task of going period by period to fetch securities based on certain fundamental values using CapIQ API
- Fetch optimal screens/data and develop data cleaning algorithms for different geographies based on the ticker mapping file and business logic
- Based on the business logics and mapping file set controls on the raw data and trigger emails if there is a breach

Methodology and Final Deliverable

- Developed first and most recent trade-date based automated screening methodology and extracted raw data using CapIQ API
- Considering every database has a unique way to name the ticker symbol, a tickerization exercise was conducted on historical data to create Ticker Mapping file
- Leveraged Ticker Mapping file and business logic to clean the data and send automated emails in case of breach based on the controls set using Power Automate and Outlook
- The cleaned data was then moved to a model for further backtesting analysis
- Theis automated process eliminated 75% reduction in time to screen securities across periods with all controls in place



Tools/Technology used: CapIQ, Python, Excel, Power Automate, Outlook

Data Engineering & Cloud Computing

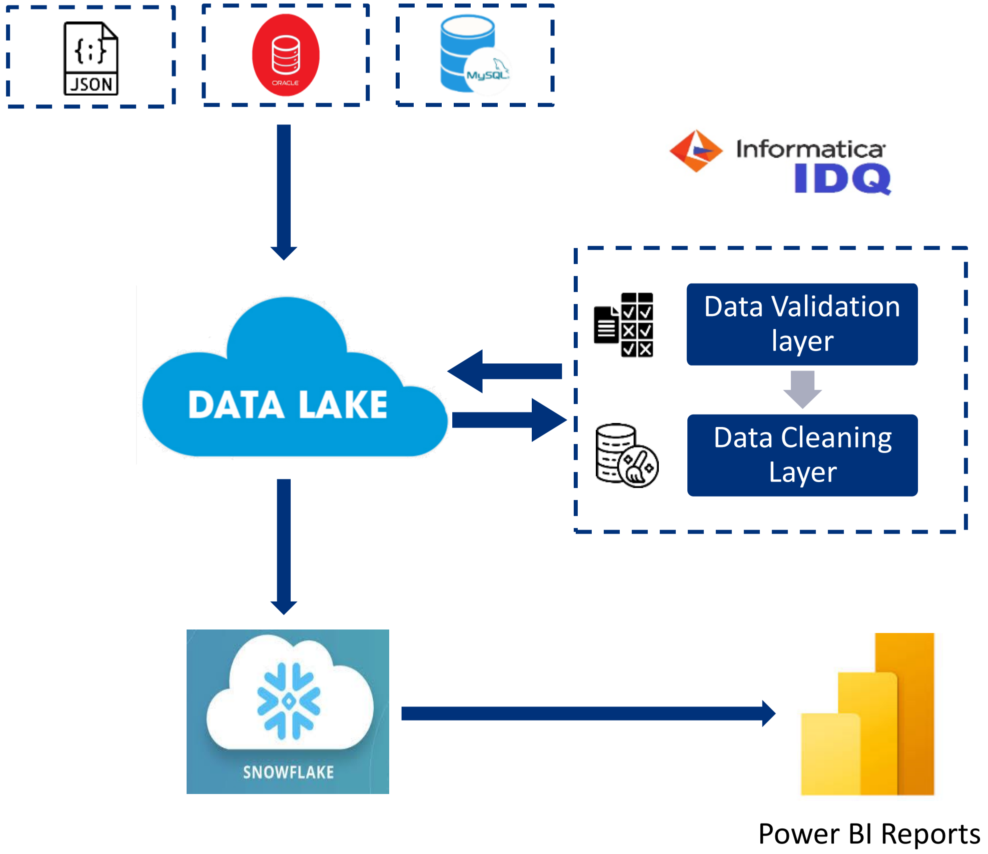
Data Strategy, Architecture & Governance for optimal Data Engineering

Request and Guidelines Provided

- Client: Private Equity Firm
- Currently, the data is spread across multiple databases depending on their plant locations and managing team, reporting based on local standards
- Consolidate the data across multiple databases, run data governance checks and then provide the same for Analytics
  - The data is spread across multiple formats and data sources, many users input data leading to inconsistencies borne out of human errors
  - Data to be available at one place for a central analytics team, to run analytics for key stakeholders

Methodology and Final Deliverable

- ETL pipelines were set up to extract data from multiple sources (MySQL, Oracle, JSON files) and loaded in a Data Lake with multiple layers of data governance framework
- Data quality was analyzed and curated using Informatica IDQ in
  - The "Validation" layer leads to integrity checks, completeness checks, de-duplication, ACID properties validation, and others; it was then passed to the "Cleaning" layer.
  - In the "Cleaning" layer, data was remediated with business logic, operational rules, and a feedback cycle from the operational team
- The processed data was then loaded into the Snowflake warehouse, where it was further used by the BI and analytics team to generate insights



Tools/Technology used: Snowflake, Informatica IDQ, SQL, Python, Power BI

salessupport@tresvista.com | www.tresvista.com