sogeti labs
Part of Capgemini

Data engineering best practices with Snowflake for Utilities project

Author – Swati Bhole

Introduction:

In a utility project which generally handles gas and electricity, has a huge data for processing and analysis. The data is generally generated by the end user so majorly has anomalies and discrepancies in the formatting, outliers, information carried, etc. So, to perform the analysis on this data is very difficult without processing the same. Lot of technical operations must be done on this data.

This data is generally on different on-premises databases, which is extracted and loaded to a data warehouse for the further processing.

While doing the data engineering in data warehouse the data has to undergo different stages like transformation, profiling, cleansing, standardization, loading the data vaults, fetching the data in the data visualizations tool for further analysis during AI and ML.

The process of data engineering should be done using the best practices with the proper tool to suffice the goal of the same.

Business Case:

- Increased efficiency and productivity: Following best practices for data engineering with Snowflake can help businesses optimize their data workflows and pipelines. This can increase efficiency and productivity by reducing the time and effort required to process and analyze data. For example, using Snowflake's automatic query optimization and multi-cluster processing features can help businesses process large volumes of data quickly and efficiently.
- Improved data quality: Best practices for data engineering with Snowflake can help businesses improve the quality of their data. This can include implementing data validation checks, ensuring data consistency and accuracy, and addressing data quality issues early in the data pipeline. By improving data quality, businesses can reduce the risk of errors and make more informed decisions based on accurate data.
- Enhanced collaboration and governance: Best practices for data engineering with Snowflake can help businesses improve collaboration and governance. This can include setting up a centralized data repository with clear data governance policies, implementing role-based access control, and using version control for data engineering code. By improving collaboration and governance, businesses can ensure that everyone has access to the same data and that data is managed and used in a consistent and secure manner.
- Greater agility and flexibility: Best practices for data engineering with Snowflake can help businesses become more agile and flexible in their data processing and analysis. This can include using automation and monitoring tools to streamline data workflows, using cloud-based data warehousing to scale up or down as needed, and using APIs to integrate Snowflake with other data tools and platforms. By becoming more agile and flexible, businesses can respond quickly to changing data requirements and stay ahead of the competition.
- Lower costs: Best practices for data engineering with Snowflake can help businesses reduce costs by optimizing data workflows and pipelines, reducing data quality issues, and minimizing manual intervention. By reducing costs, businesses can free up resources to invest in other areas of their operations.

Problem Statement:

Data extraction for utility projects is a complex process due to a variety of reasons. Here are some of the difficulties that may arise during data extraction for a utility project:

- Data quality issues: Utility companies often have large and complex data sets, with data coming from a variety of sources. This can lead to data quality issues, such as missing data, inconsistent data formats, and duplicates. Addressing these data quality issues can be time-consuming and require specialized tools and expertise.
- Data integration challenges: Utility companies often have multiple systems and databases that store data. Extracting data from these disparate sources and integrating it into a single data repository can be challenging. This may require the use of specialized ETL tools, as well as mapping and transformation of data to ensure consistency and accuracy.
- Regulatory compliance: Utility companies are often subject to regulations and data privacy laws, which can impact the data extraction process. Compliance with regulations may require additional steps, such as obtaining consent from customers before extracting their data, ensuring that sensitive data is properly secured, and tracking and auditing data access.
- Data volume and complexity: Utility projects often involve large volumes of data, with complex relationships between data entities. This can make the extraction process time-consuming and resource-intensive, requiring significant computing power and storage capacity.
- Legacy systems and data formats: Utility companies often have legacy systems and data formats that can be difficult to work with. Extracting data from these systems may require specialized expertise and tools, as well as additional steps to ensure compatibility with modern data systems and formats.
- Data sources: Utility companies often have a large number of disparate data sources, including databases, spreadsheets, and legacy systems. Extracting data from these sources and integrating it into a centralized system can be a challenging task.
- Data variety: Utility projects may involve a wide variety of data types, including structured and unstructured data, as well as data in different formats such as text, images, and video. Extracting and integrating this data can require specialized tools and expertise.
- Data security: Utility data may be sensitive and confidential, requiring strict security measures to be put in place to ensure that the data is not compromised during the extraction process.

Proposed Solution:

- Data Integration: The first step is to integrate the data from various sources into the Snowflake data warehouse. This can be done by creating an ETL pipeline through Matillion/ SSIS that extracts the data from different sources, transforms it as per the data quality (Informatica/ Cloud Data Quality) and MDM standards, and loads it into Snowflake. To ensure the quality of the data, the MDM system can be used to validate the data, standardize it, and match and merge it to ensure consistency across the organization.
- Data Modelling: The next step is to model the data in Snowflake, which involves defining tables, views, and other database objects based on the business requirements. This can be done using the standard data modelling techniques, such as ER diagrams or dimensional modelling.

- Data Quality: To ensure the data quality, Snowflake offers built-in data quality checks that can be customized to meet the specific requirements of the utility project. This includes data validation rules, data profiling, and data lineage.
- Data Governance: To ensure that the data is managed effectively, a data governance framework can be implemented, which includes policies and procedures for managing data, defining roles and responsibilities, and establishing standards for data quality and security.
- Data Analytics: With the data integrated, modeled, and validated, it can be used for data analytics and reporting. This can be done using the standard BI tools that integrate with Snowflake or by using Snowflake's built-in analytics capabilities, such as Snowflake's worksheet, which allows for interactive SQL queries, and Snowflake's machine learning capabilities, which allows for predictive analytics.
- Continuous Improvement: Finally, it's essential to monitor and continuously improve the data engineering process with Snowflake. This can be done by regularly reviewing the data quality reports, monitoring performance metrics, and conducting data governance audits.

Introduction of Solution:

By using the latest cloud technology it is possible to handle a very high volume with respect to the usage of the resources of the tool and be cost effective. Leverage the inhouse skills to do the development and testing.

Application of Solution:

- Integration of Utility Data: A utility project requires integration of data from various sources such as smart meters, IoT devices, and customer data. Snowflake can be used to integrate this data into a single, scalable, and secure data warehouse. This data can then be transformed and cleansed to ensure that it meets the MDM standards and can be used for further analysis.
- MDM Validation and Standardization: To ensure data quality, it is essential to validate the data against the MDM standards. Snowflake can be used to perform data quality checks, data profiling, and data standardization. This can help ensure that the data is consistent, accurate, and up-to-date.
- Analytics and Reporting: With the data integrated and validated, it can be used for analytics and reporting. Snowflake's built-in analytics capabilities can be used to generate reports and dashboards that help utility companies make informed decisions about their operations. For example, they can use the data to optimize energy usage, identify maintenance issues, and improve customer service.
- Machine Learning and AI: Snowflake's machine learning capabilities can be used to develop predictive models that help utility companies anticipate future demand, identify trends, and optimize their operations. For example, machine learning algorithms can be used to predict energy usage based on weather patterns, time of day, and customer behavior.
- Customer Analytics: Snowflake can be used to integrate customer data from various sources, such as CRM systems, billing systems, and customer surveys. By analyzing this data, utility companies can gain insights into customer behavior, preferences, and satisfaction levels. This can help them improve customer service, reduce customer churn, and identify cross-selling opportunities.

- Real-time Data Processing: In a utility project, real-time data processing is critical for monitoring operations and making quick decisions. Snowflake's real-time data processing capabilities can be used to process streaming data from IoT devices and smart meters in real-time. This can help utility companies identify issues quickly and take corrective actions in a timely manner.
- Regulatory Compliance: Utility companies must comply with various regulations related to energy usage, emissions, and safety. Snowflake can be used to integrate data from different sources, such as regulatory reports, environmental data, and safety records, to ensure that the utility company is meeting its compliance requirements.

Future/Long Term Focus:

- Snowflake optimization: Snowflake is a cloud-based data warehousing platform that provides a range of features and functionality to optimize data storage, query performance, and overall scalability. To ensure that the utility project using MDM with Snowflake remains performant and cost-effective over the long-term, data engineering teams will need to focus on continually optimizing Snowflake usage. This may involve implementing best practices for data storage, partitioning, and indexing, as well as leveraging features like auto-scaling and auto-tuning to optimize query performance.
- Advanced analytics: With Snowflake's native support for semi-structured data, data engineering teams can focus on building systems that support advanced analytics, such as machine learning, natural language processing, and graph analysis. This may involve building data pipelines that preprocess and transform data for use in advanced analytics workflows, or building infrastructure to support training and deploying advanced analytics models.
- Data governance: As with any MDM project, ensuring data governance and data quality is critical. In the context of Snowflake, data engineering teams will need to focus on building systems that support data lineage, data quality checks, and data profiling to ensure that data is accurate, complete, and consistent.
- Real-time processing: As organizations move towards real-time decision-making, data engineering teams may need to focus on building systems that can process data in real-time or near real-time using Snowflake. This may involve building data pipelines that ingest and process data in real-time using Snowflake's real-time data ingestion capabilities, or integrating Snowflake with streaming frameworks like Apache Kafka to support real-time processing.
- Cloud migration: As more organizations move to the cloud, data engineering teams will need to focus on building systems that are cloud-native and can take advantage of Snowflake's cloud-based architecture. This may involve migrating existing on-premise systems to Snowflake, or building new systems from scratch using Snowflake's cloud-based service

Conclusion:

When it comes to data engineering on utility projects with Snowflake and MDM, there are several best practices that you can follow for effective data profiling, cleansing, standardization, and data vaulting. Some of these best practices include:

- Start with a Data Profiling Exercise: Before beginning any data engineering work, it is essential to profile the data to understand its quality, completeness, and accuracy. This exercise helps in identifying data inconsistencies and data gaps that need to be addressed.
- Implement Data Cleansing Techniques: Data cleansing is a crucial step in any data engineering project. This process involves the removal of duplicate records, standardization of data, and the correction of inconsistencies in data fields. It is essential to automate data cleansing techniques as much as possible to minimize manual errors.
- Standardize Data Formats: Standardizing data formats across different data sources and systems is essential to ensure data consistency and quality. This can be achieved through the use of data transformation tools or manual mapping and translation of data fields.
- Use MDM for Master Data Management: Master data management (MDM) is a critical component of any utility project. It involves the creation and maintenance of a centralized repository for master data such as customer, product, and vendor information. MDM helps to ensure data accuracy, consistency, and completeness.
- Implement Data Vault Architecture: Data vault architecture is a highly scalable and flexible approach to data warehousing that enables you to store historical data in a structured manner. Data vaults can be used to store raw data, perform data integration, and support analytical reporting.
- Implement Data Governance Processes: Data governance is essential to ensure the quality, consistency, and accuracy of data. It involves defining data standards, policies, and procedures for managing data across the organization. A well-defined data governance process ensures that data is properly managed, secured, and compliant with regulatory requirements.
- Use Snowflake for Cloud Data Warehousing: Snowflake is a cloud-based data warehousing solution that offers scalability, performance, and cost-effectiveness. Snowflake provides native support for MDM, data cleansing, and data transformation, making it an ideal choice for utility projects.

Appendix A – Scenarios:

Test cases covered for **Exception** rules

- Validate of the exception rules are defined with respect to entity type (Asset, Location, Work Order)
- Validate that the attributes are populated with respect to the rule developed
- Validate the total view count is equal to Valid and InValid count of views
- Validate the Invalid count is equal to the Exception table count
- Validate the data in the exceptional tables is same as Invalid record from the views.
- Data validation Scenario

Test cases covered for Rule Based rules

- The valid data counts with respect to the rules created
- Verify the invalid data counts with respect to the rules created
- Verify the total number of records present in source tables (Satellite tables)

Test cases covered for Initial Views

- Verify the number of rows in Landing table with the View table.
- Verify hash values in the View with the Hash values from the view table script
- Verify the joined table
- Verify the Hash columns in the respective views
- Verify the newly added column 'Source Change Identifier' fetches correct hash value after the concatenation of all columns except the Derived columns
- Verify that hash values creation is correct in 'Source Change Identifier'
- Verify the data navigation from landing table to View table
- Data validation with except statement with Landing tables and respective view.

Test cases covered for Column Profiling

- Verify that the profiling report name is as per naming convention standard
- Verify that these tabs are available in the report
- Verify total number of records in database is same in the report
- Verify the number of columns in report are same with respective landing table
- Verify the max length for column matches with the max length for the respective table
- Verify that tab has the metadata information of the profiling report like Name of the report, Location, Row Count, Username, time stamp etc.

Test cases covered for Landing Tables

- Validation of SQL query used for extracting data
- Verify the number of columns with respect to source database
- Verify the number of records with respect to source database
- Verify the columns description in Landing tables
- Verify the datatype of the columns with respect to source database
- Validation of the data migrated from source to target

About Sogeti

Part of the Capgemini Group, Sogeti operates in more than 100 locations globally. Working closely with clients and partners to take full advantage of the opportunities of technology, Sogeti combines agility and speed of implementation to tailor innovative future-focused solutions in Digital Assurance and Testing, Cloud and Cybersecurity, all fueled by AI and automation. With its hands-on 'value in the making' approach and passion for technology, Sogeti helps organizations implement their digital journeys at speed.

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of over 325,000 team members more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2021 global revenues of €18 billion. Get The Future You Want | www.capgemini.com

Visit us at www.sogeti.com