gathr.ai

Fortune 100 telecommunications company seamlessly migrates from Teradata to Amazon Redshift

Enables massive scale analytics and improves query performance by 20% using Gathr.ai

Enterprises today have massive volumes of data coming in from multiple sources. To leverage this data with agility and speed, businesses are increasingly looking to migrate their on-premises data warehouses to the cloud. This helps lower operational costs, improve scalability, and realize faster time-to-insight.

About the Customer

A US-based Fortune 100 broadband connectivity company and cable operator serving more than 30 million customers.

In their existing setup, the telecom company received data from multiple sources that was fed into an SFTP server. Once ETL was performed, the data was read by an Informatica workload and persisted to their Teradata data warehouse. Business analysts then accessed this data and ran queries to gather insights. This workflow was posing several technical and business challenges, and the client wanted to make strategic shi to the cloud with the following objectives:

Enhance scalability: Ability to handle rapidly growing volumes of data, manage peaks in traic, and run new business use cases with greater ease

- **Reduce costs:** Lower Teradata licensing costs and maintenance overheads

- **Improve query performance:** Ability to query raw data and generate business insights faster

- **Realize a unified view:** Ability to track workloads across the end-to-end data transformation journey using a single platform

- **Simplify management:** Overcome the management complexities of Teradata and Informatica by leveraging cloud-based services like Amazon Redshi, S3, and Athena

- **Seamless integrations:** Integrate with other cloud-native services to load data and visualize insights

- **Automate workflows for CI/CD:** Ability to swily move changes from the development environment to staging and production environments
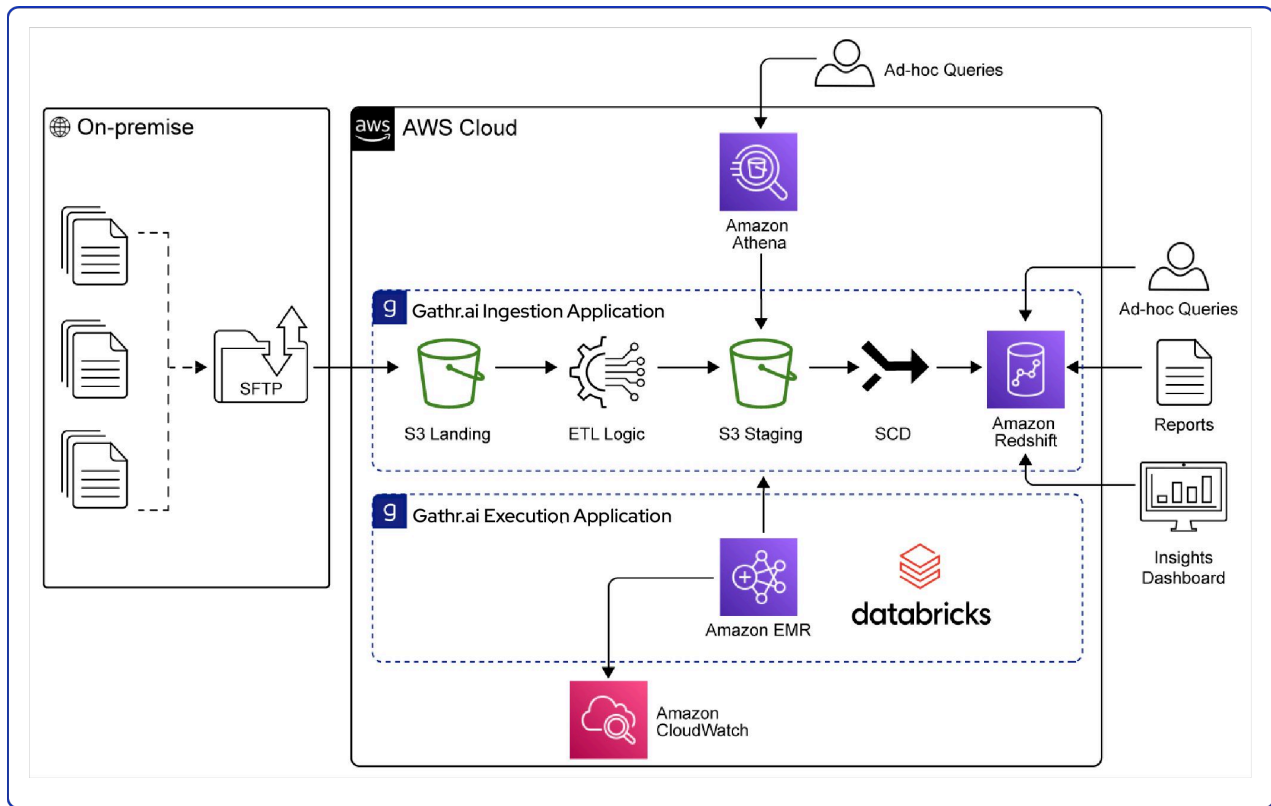
Solution

To meet the customer's business requirements, we built an end-to-end data flow solution leveraging the following key components:

1. Amazon Redshi: Scalable, cloud-native data warehouse to collect and store data for all workloads, support query requirements, and accommodate varying business use cases
2. Amazon S3: Cloud-native storage for the gathered data feeds
3. Amazon Athena/EMR/Redshi: Ad-hoc query engine to query data feeds directly and generate insights
4. Gathr.ai, the all-in-one data pipeline platform, was used to:

   - Configure ETL flows, ingest data, perform full load, incremental load, and CDC (SCD type 1 and 2) from Teradata to Redshi, leveraging a visual interface

   - Transform and persist data feeds to Amazon S3 with an auto-scalable execution engine

   - Enable one-time migration by directly loading Teradata tables into Amazon Redshi

   - Validate data post migration

   - Provide a unified view of the complete workflow

   - Set up CI/CD for upgrading ETL flows and moving them seamlessly from one environment to another

   - Schedule and trigger the data flow process at a pre-configured frequency

Here's how Gathr helped simplify the ingestion pipeline workflow:

- Moved the raw data feeds arriving on the SFTP server to the Amazon S3 bucket

- Converted the raw data feeds to Parquet format and persisted them in the Amazon S3 landing bucket, enabling users to directly query the data in Amazon S3 using Athena

- Gathr's ingestion application managed the incremental load of data feeds, wherein users only needed to specify the endpoint details of each stage

- Used the self-service platform's visual, drag-and-drop interface to configure ETL flows and clean, enrich, transform, and persist data in the S3 staging area

- Configured a data validation job, and gathered statistics from data in the landing and staging areas, which were used to validate the ingestion setup for each job

- Upon successful validation, Gathr connected Amazon S3 staging locations to the Amazon Redshi tables, enabling users to configure CDC behavior for the table on Amazon EMR or Databricks with auto-scaling

High-level overview of the solution

Business benefits

As a next-gen data pipeline platform, Gathr helped the client modernize their legacy data warehouse and move analytics from an on-premises environment to the AWS Cloud. The solution enabled them to accelerate and de-risk the entire migration process, seamlessly integrate with Amazon's cloud-native services, and switch to a cost-effective pay-as-you-go business model. Some of the strategic benefits they realized are given below:

- Ability to handle 30 billion rows and easily scale to manage fluctuating production loads

- 20% better query performance for analytical queries

- Support for 40% more analytics users across the enterprise

- 15% increase in the number of queries executed, enabling users to unlock new business opportunities

- Ability to perform ETL with minimal hand-coding using pre-built operators

- 360-degree visibility with a unified, configurable view of all workloads

- Lower licensing and infrastureture cost

Scan and
start free 14-day trial

Scan to
schedule a demo