# Broadcasting Corporation

**THE DATA PLATFORM ENGAGEMENT DISCUSSED IN THIS DOCUMENT IS A COLLABORATIVE EXERCISE INVOLVING STAKEHOLDERS FROM AWS, A NORTH AMERICAN BROADCASTING CORPORATION AND PRESIDIO. THE DESIGN, LOW LEVEL SOLUTION, DEPLOYMENT AND IMPLEMENTATION ACTIVITIES INVOLVED STAKEHOLDERS FROM ALL THE TEAMS MENTIONED ABOVE.**

### The Customer

A broadcasting corporation that provides satellite radio and online radio services in the United States and Canada.

### The Challenge

The broadcasting corporation was challenged with understanding their IoT data and its dynamic complex structure. They were processing their data in Glue for a specific timeframe in a bookmark enabled job. Another challenge the company faced was performing ACID transactions on S3 data lake.

### Why AWS

The broadcast company selected AWS primarily to build out a centralized ETL pipeline. Their primary goal is to manage large streams of data coming from their connected Vehicles, Airship and other systems, availing the data to operational and application intelligence solutions. This aggressive move to the cloud allows the broadcasting corporation to take advantage of the whole AWS ecosystem that allows them to scale to their data needs.

### The Solution

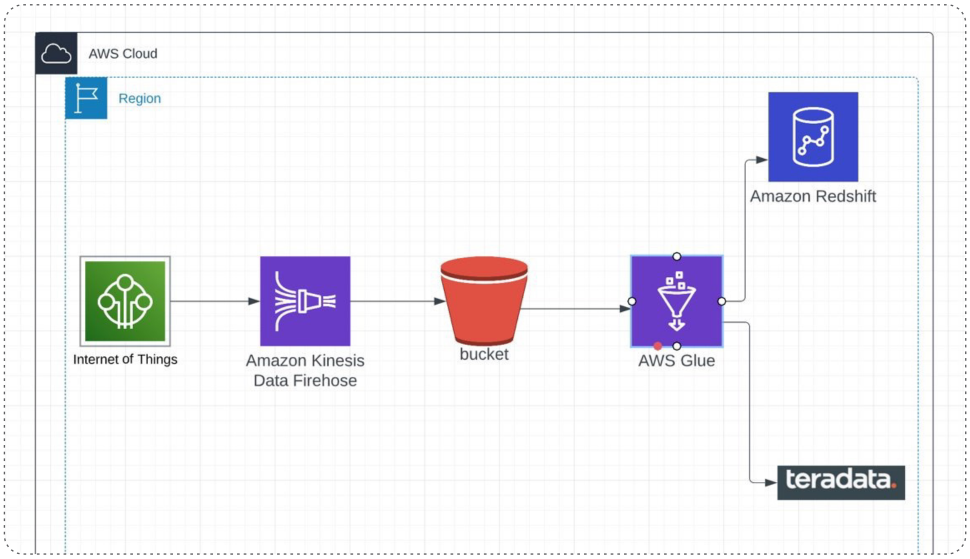
Presidio, in partnership with the broadcasting corporation, built an end-to-end data pipeline to ingest, transform and consume IoT data. Pipeline is orchestrated utilizing AWS services like Kinesis Firehose, Glue, Redshift,Teradata, S3.

### Data Flow

The IoT data is ingested into S3 Raw zone. A pre-processing Glue Job is invoked using Raw zone S3 event to perform initial data validation and cleanse the data.  Once the data have been cleansed, the main glue job will be called, this will be responsible for transformation and ingesting the transformed data into Teradata/Redshift.

The current data capture (CDC) and batch job have been implemented on the same Glue job using Glue Job Bookmarks.  This have reduced their billing consistently on their workloads.

### ETL Data Pipeline



### The Impact

Utilizing AWS Analytics tools such as AWS Redshift, Amazon Glue, and AWS EMR, the broadcast company & Presidio's engineering team were able to set up a secure and error-free data pipeline. This allowed the broadcasting company to gain confidence in the data exposed to the end consumer.

### Technologies Used

AWS Services: Amazon S3 ,AWS Lambda, Amazon EMR, AWS Glue, Amazon Redshift, Teradata, Amazon SNS, Amazon Kinesis Firehose, Amazon VPC, AWS VPN, AWS CloudWatch

Other Services/Frameworks/ Technologies: Spark, Python, Scala, Deltalake

## NON-FUNCTIONAL REQUIREMENTS

### Recovery Point Objective (RPO)

RPO is the maximum tolerable amount of data loss in case of a failure. To implement RPO in Amazon Redshift:

- **Regular Backups** – Regular backups of Amazon Redshift clusters were taken using automated snapshots. A backup schedule was setup to align with the agreed RPO objectives with the customer. 8 hours in this case. Amazon Redshift allows us to configure the retention period for automated snapshots.
- **Manual Backups** – We took manual backups before making major changes or updates to our Redshift cluster. This ensured that we had a known-good state to restore to, minimizing potential data loss.
- **Snapshot Frequency** – Adjusted the frequency of automated snapshots to align with your RPO requirements. More frequent snapshots helped us reduce the potential cause of data loss.

### Recovery Time Objective (RTO)

RTO is the maximum tolerable downtime allowed to restore a system after a failure. To implement RTO in Amazon Redshift:

- **Snapshot Restore** – Restore from automated or manual snapshots to recover the Redshift cluster to a known-good state. Automated the restore process to minimize manual intervention and reduce RTO. AWS lambda was used for the automated restoration process. A cloudwatch alarm was setup for a few key error messages/events from redshift. Application code will also push a certain metrics to cloudwatch and when each of these metrics go beyond a certain threshold, then they trigger the corresponding alarm and inturn trigger the automatic restore of the redshift cluster.
- **Incremental Load Processes** – Implement incremental data loading processes to restore only the changed or missing data, reducing the time required for recovery. Incremental data loading process has been implemented using AWS Glue job bookmarks.
- **Optimize Data Loading** – Optimized ETL processes to load data efficiently and quickly, reducing the overall time required for data restoration.

### Partners

aws

### About Presidio

Presidio is a global digital services and solutions provider accelerating business transformation through secured technology modernization. Highly skilled teams of engineers and solutions architects with deep expertise across cloud, security, networking, and modern data center infrastructure help customers acquire, deploy, and operate technology that delivers impactful business outcomes. Presidio is a trusted strategic advisor with a flexible full life cycle model of professional, managed, and support and staffing services to help execute, secure, operationalize, and maintain technology solutions. For more information, visit www.presidio.com.

**For more information on how we connect IT of today to IT of tomorrow, visit presidio.com**

| April 2024