## Abstract

This project “"Nationwide Distributed Energy Resource Modeling and Analysis.” focuses on the comprehensive extraction, transformation, and loading (ETL) analysis of a dataset containing information on solar energy installations. Through meticulous data cleaning and transformation, the objective is to refine the dataset for meaningful analysis. The exploration includes geospatial and temporal patterns, technical details such as capacity, tracking systems, and efficiency, as well as insights into technology classes and inverter specifications.

The project begins by ingesting and storing data in Amazon S3, a scalable and secure cloud storage solution. AWS Glue, a fully managed extract, transform, and load (ETL) service, is then utilized to process the data. This ETL process is orchestrated using AWS Glue.  
  
Once processed, the data is loaded into Tableau, a fully managed data visualization service, where it is transformed into actionable insights through visualization tools. These insights enable Solar Panel providers to identify trends, patterns, and potential customers, empowering them to implement targeted customer strategies.  
  
To streamline the workflow, the project utilizes AWS glue to schedule and execute the ETL jobs. By establishing connections between AWS services and configuring job definitions, the project ensures a smooth and efficient data pipeline.  
  
Furthermore, the project addresses technical challenges such as configuring security settings, establishing connections between services, and troubleshooting errors. Through careful planning and implementation, these challenges are overcome to ensure the successful execution of the project.  
  
Overall, the Subscriber “Nationwide Distributed Energy Resource Modeling and Analysis” demonstrates the power of cloud computing and data analytics in addressing real-world business challenges.

**Table of Contents**

1. **Abstract……………………………………………………………………….1**
2. **Acknowledgement….…………………………………………………………2**
3. [**Table of Contents………………………………………………………….**](#_bookmark0) **….3**
4. **Introduction……….…………………………………………………………..4**
5. [**Purpose, Scope, Objective, Functionalities**](#_bookmark1) **…………………………………5**
6. [**AWS Architecture**](#_bookmark1) **…………………………………………………………..7**
7. [**Services Used**](#_bookmark2) **………………………………………………………………..8**
8. [**Dataset**](#_bookmark3) **……………………………………………………………………….9**
9. [**Identity Access Management**](#_bookmark4) **………………………………………………11**
10. [**Simple Storage Service S3**](#_bookmark5) **…………………………………………………12**
11. [**AWS Glue - Crawler**](#_bookmark7) **………………………………………………………14**
12. [**AWS Glue**](#_bookmark8) **……………………………………………………………..……16**
13. [**Parquet File format……………………………………………………… ..1**](#_bookmark12)**8**
14. [**Code Snippets ……………………………………………………………...1**](#_bookmark13)**9**
15. [**Tableau**](#_bookmark14) **…………………………………………………………………….25**
16. **Insights On Tableau……………………………………………………….27**
17. **Conclusion………………………………………………………………….29**
18. [**References…………………………………………………………………**](#_bookmark22) **30**

**INTRODUCTION**

The project titled "Nationwide Distributed Energy Resource Modeling and Analysis" is a comprehensive endeavor aimed at extracting, transforming, and loading (ETL) a dataset encompassing data on solar energy installations. With a primary focus on meticulous data cleaning and transformation, the project seeks to refine the dataset to enable meaningful analysis. The exploration delves into various aspects, including geospatial and temporal patterns, technical details such as capacity, tracking systems, and efficiency, as well as insights into technology classes and inverter specifications.

This project commences with the ingestion and storage of data in Amazon S3, a highly scalable and secure cloud storage solution. Leveraging AWS Glue, a fully managed extract, transform, and load (ETL) service, the data undergoes processing. The orchestration of this ETL process is facilitated through AWS Glue, ensuring efficient handling of the dataset.

Upon completion of the processing phase, the refined data is loaded into Tableau, a fully managed data visualization service. Here, it undergoes transformation into actionable insights using visualization tools. These insights empower solar panel providers to discern trends, patterns, and potential customers, enabling the implementation of targeted customer strategies.

The workflow of the project is streamlined through the utilization of AWS Glue to schedule and execute ETL jobs. By establishing seamless connections between AWS services and configuring job definitions, the project ensures the smooth operation of the data pipeline.

Moreover, the project addresses various technical challenges, including configuring security settings, establishing inter-service connections, and troubleshooting errors. Through meticulous planning and implementation, these challenges are effectively overcome to ensure the successful execution of the project.

**Purpose:**

The purpose of the "Nationwide Distributed Energy Resource Modeling and Analysis" project is to extract, transform, and load a dataset containing information on solar energy installations. Through meticulous data cleaning and transformation, the project aims to refine the dataset for meaningful analysis, exploring geospatial and temporal patterns, technical details, and insights into technology classes. Leveraging AWS Glue for ETL processing and Tableau for visualization, the project empowers solar panel providers to identify trends and potential customers, facilitating targeted customer strategies. Streamlining workflows with AWS Glue and addressing technical challenges ensures the successful execution of the project, delivering actionable insights for informed decision-making.

**Scope:**

The project scope is to conduct a robust ETL process on the solar energy installations dataset. It involves thorough data cleaning, transforming raw data into a structured format for analysis. The focus is on addressing missing values, refining data types, and ensuring consistency. The cleaned dataset will serve as a foundation for insightful analysis, emphasizing geospatial patterns, temporal trends, and technical details such as tracking systems and inverter specifications. The ultimate goal is to provide a refined dataset for meaningful analysis, uncover patterns in solar energy installations, and contribute valuable insights to stakeholders in the renewable energy sector.

**Objective of Project:**

1. Conduct a comprehensive ETL process on the solar energy installations dataset.
2. Address missing or irrelevant data and ensure data accuracy.
3. Implement extraction, transformation, and loading phases to refine the dataset.
4. Explore geospatial and temporal patterns in solar energy system distribution.
5. Analyze technical details, including capacity, tracking systems, and efficiency.
6. Investigate technology classes, ground-mounted systems, and inverter specifications.
7. Generate visualizations to represent analysis outcomes effectively.

**Functionalities provided as follows:**

The " Nationwide Distributed Energy Resource Modeling and Analysis " project provides a range of functionalities. Some of the key functionalities offered by the project include:

1. Data Ingestion: Ability to ingest and store customer data from various sources into

Amazon S3.  
2. Data Processing: Utilize AWS Glue for ETL (Extract, Transform, Load) processes to clean, transform, and prepare data for analysis.  
3. Glue Orchestration: Employ AWS Glue to schedule and monitor data workflows, ensuring seamless execution of tasks.  
4. Data Analysis: Utilize Tableau for analyzing large datasets and uncovering insights related to solar panel utilization and patterns.

5. Automation: Implement automated processes for data ingestion, processing, and analysis to minimize manual intervention and improve efficiency.

6. Error Handling: Incorporate error handling mechanisms to identify and resolve issues during data processing and analysis.

7. Scalability: Design the project to be scalable, allowing for the handling of large volumes of data and accommodating future growth.  
8. Security: Implement security measures to ensure the confidentiality, integrity, and availability of data throughout the project lifecycle.

## AWS Architecture

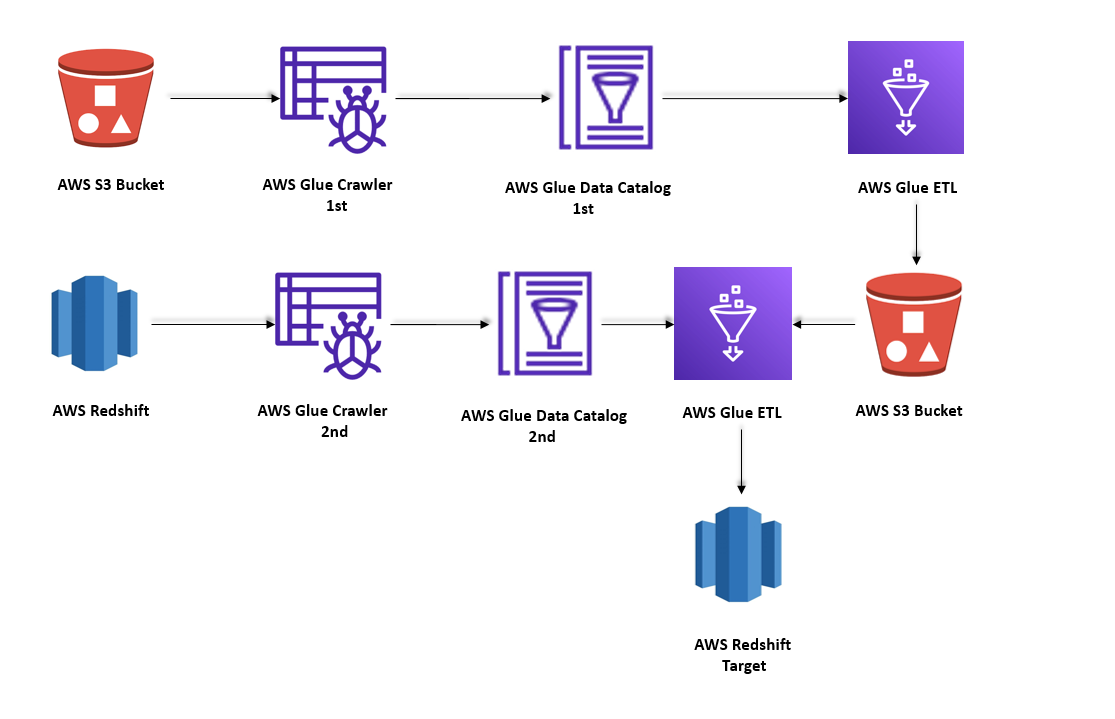


Fig.1 Architecture

## Services Used:

1. Amazon S3: Amazon S3 is an object storage service that provides manufacturing scalability, data availability, security, and performance.
2. AWS IAM: This is nothing but identity and access management which enables us to manage access to AWS services and resources securely.
3. AWS Glue: A serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

4. AWS Athena: Athena is an interactive query service for S3 in which there is no need to load data it stays in S3.

## Dataset

The provided data is a table with various columns containing information related to solar energy installations. Each row represents a specific solar energy system, and the columns provide details such as the system's ID, customer class (e.g., residential or non-residential), location information (zipcode, city, state), the utility service territory, capacity, technology class, and various flags indicating features like ground mount, tracking type, storage, and inverter specifications.

Key columns and their meanings:

ID\_1, ID\_2: System IDs.

customer\_class: Type of customer (RES for residential, NON-RES for non-residential).

zipcode, city, state: Geographic location of the solar energy system.

utility\_service\_territory: The utility service territory the system is a part of.

year, month, day: Installation date.

capacity\_DC\_kW: Capacity of the solar energy system in kilowatts.

ground\_mount\_flag: Indicates whether the system is ground-mounted.

tracking\_type: Type of tracking system used (e.g., single-axis, dual-axis).

azimuth\_1, azimuth\_2, azimuth\_3: Azimuth angles for tracking systems.

tilt\_1, tilt\_2, tilt\_3: Tilt angles for tracking systems.

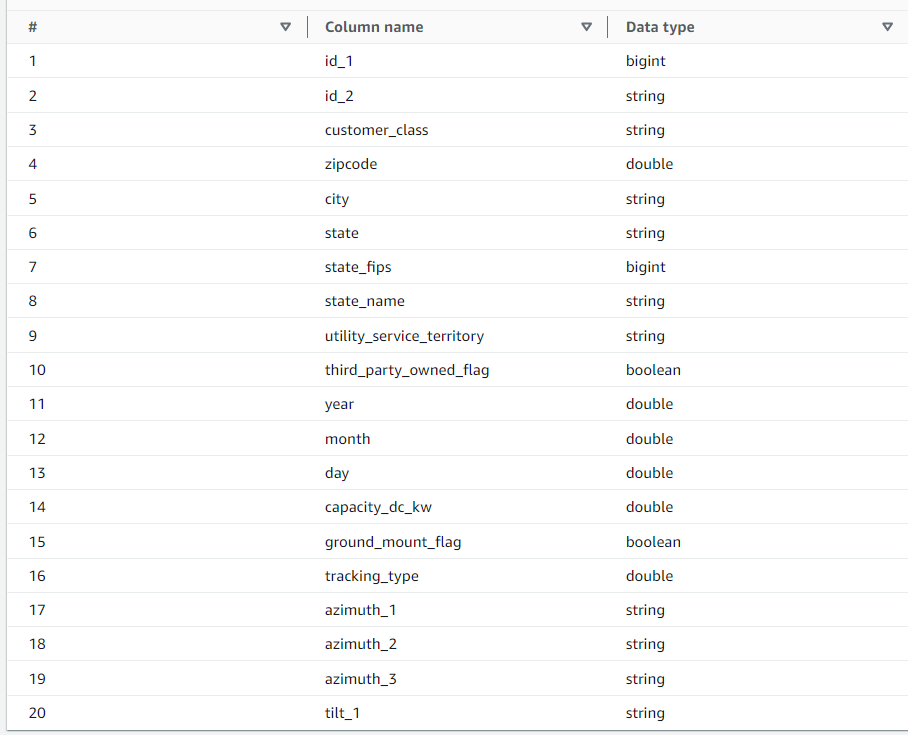
efficiency\_1, efficiency\_2, efficiency\_3: Efficiencies of the system.

tech\_class\_primary, tech\_class\_primary\_2, tech\_class\_primary\_3: Primary technology classes.

storage\_flag, storage\_info: Indicates if the system has storage, and additional storage information.

inverter\_loading\_ratio, inverter\_capacity\_1, inverter\_capacity\_2, inverter\_capacity\_3: Inverter-related information.

The dataset is related to solar energy installations and can be utilized for analyzing and reporting purposes. It offers valuable insights into the characteristics of different solar energy systems, including their locations and various technical details associated with them.



# Identity Access Management

A web service that helps to securely control access to AWS resources. Shared access to your AWS account. Granular permissions. Roles are created by AWS Management Console. Then Policies are attached to manage granular access. An IAM account is created with custom password and admin access in AWS console.

The access key and secret key are generated and the access key credentials are downloaded.

Then the AWS CLI is set up using the access key credentials and configured with the access key, secret key, and region. The configuration is verified by running a command in AWS CLI.

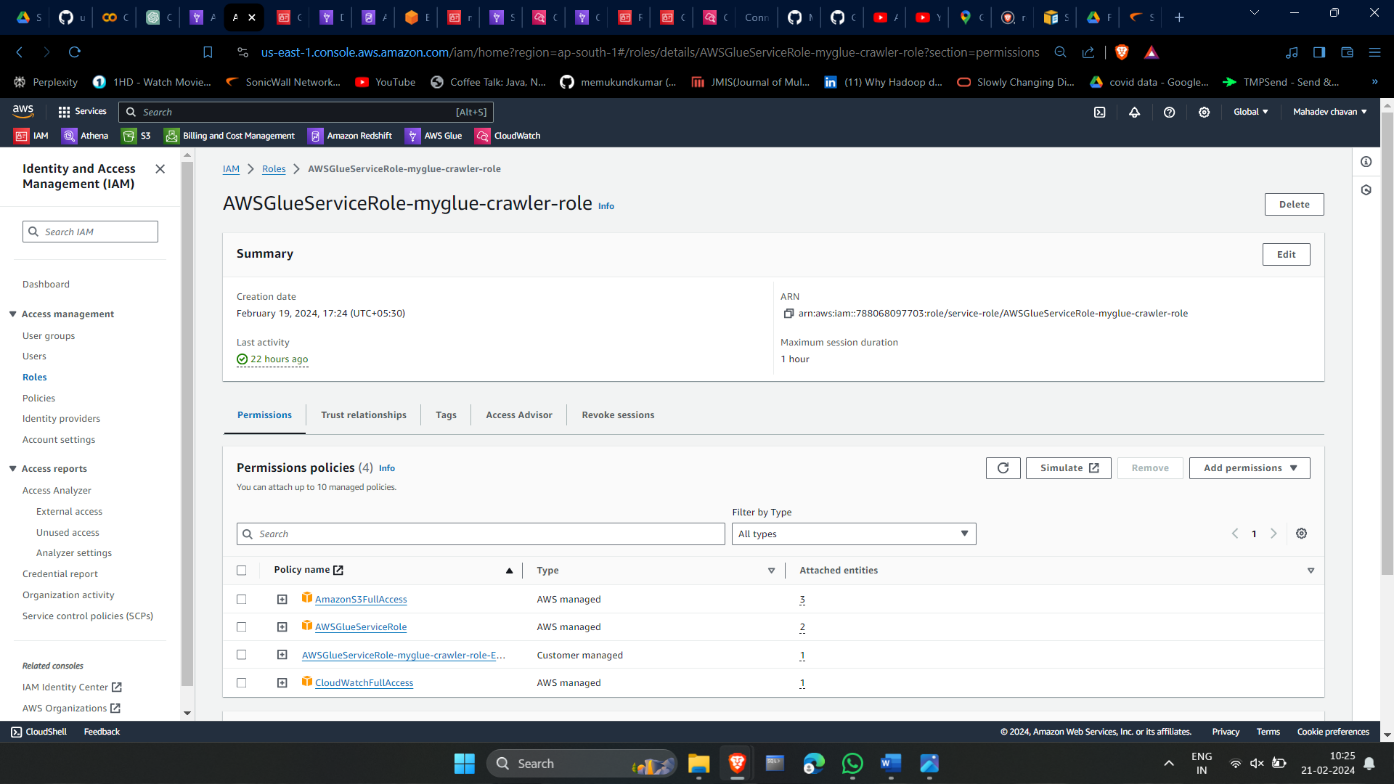


Fig.2- IAM role

# Simple Storage Service S3:

Amazon Simple Storage Service (Amazon S3) is a scalable object storage service provided by Amazon Web Services (AWS). It is designed to store and retrieve any amount of data from anywhere on the web. Here's a brief overview of the key concepts and features of Amazon S3:

Core Concepts:

Buckets:

In Amazon S3, data is stored in containers called "buckets." A bucket is a globally unique namespace within Amazon S3, identified by a DNS-compliant name.

Objects:

Objects are the fundamental entities stored in Amazon S3. An object consists of data, a key (unique within a bucket), and metadata. Objects can range from 0 bytes to 5 terabytes in size.

Keys:

A key is a unique identifier within a bucket for an object. It defines the path of the object within the bucket and can include slashes to mimic a hierarchical structure.

Features and Capabilities:

Durability and Availability:

Amazon S3 is designed for 99.999999999% (11 9's) durability of objects over a given year. It achieves high availability by distributing data across multiple servers and data centers.

Data Lifecycle Management:

You can configure lifecycle policies to automatically transition objects between storage classes or delete them based on criteria such as age or object size.

Storage Classes:

Amazon S3 offers various storage classes, each designed for different use cases. These include Standard, Intelligent-Tiering, Standard-IA (Infrequent Access), One Zone-IA, Glacier, and Glacier Deep Archive.

Versioning:

Versioning allows you to keep multiple versions of an object in the same bucket. This helps protect against accidental deletions or overwrites.

Security and Access Control:

Amazon S3 provides robust access controls to secure your data. You can use Access Control Lists (ACLs) and Bucket Policies to manage access permissions. Additionally, AWS Identity and Access Management (IAM) can be used for fine-grained access control.

Server-Side Encryption:

Amazon S3 supports server-side encryption to encrypt data at rest. Options include SSE-S3, SSE-KMS (Key Management Service), and SSE-C (Customer-Provided Keys).

Event Notifications:

You can configure event notifications that trigger AWS Lambda functions, SQS queues, or SNS topics when specific events, such as object creation or deletion, occur in your bucket.

Multipart Upload:

For large objects, Amazon S3 supports multipart uploads, allowing you to upload parts of an object in parallel, improving efficiency and resiliency

**AWS Glue - Crawler**

A crawler is used to populate the AWS Glue Data Catalog with tables. Crawlers can crawl multiple data stores in a single run. ETL jobs that you define in AWS Glue use these Data Catalog tables as sources and targets. AWS Glue Data Catalog is a managed metadata repository that stores and organizes metadata. It used to define the structure and schema of your data during glue etl job .

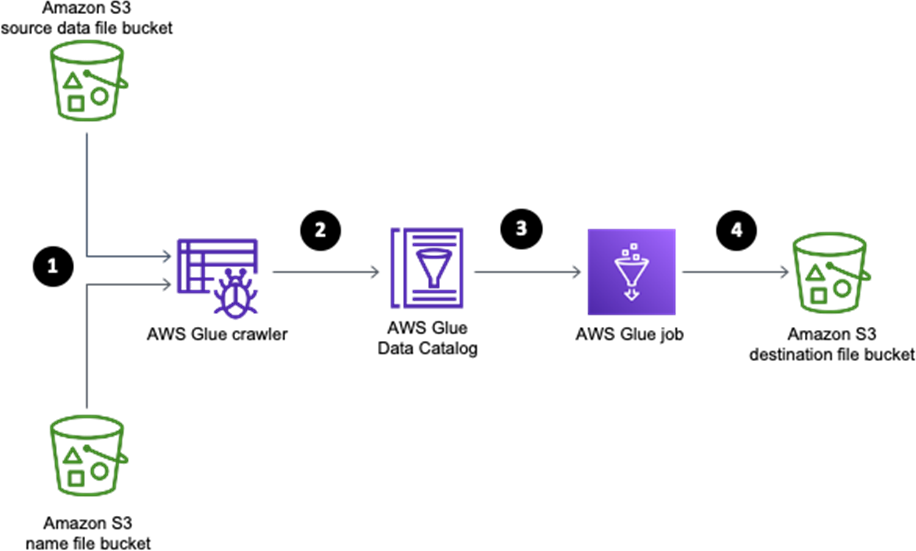


Fig.6- AWS Glue Crawler

Crawler run on S3 object

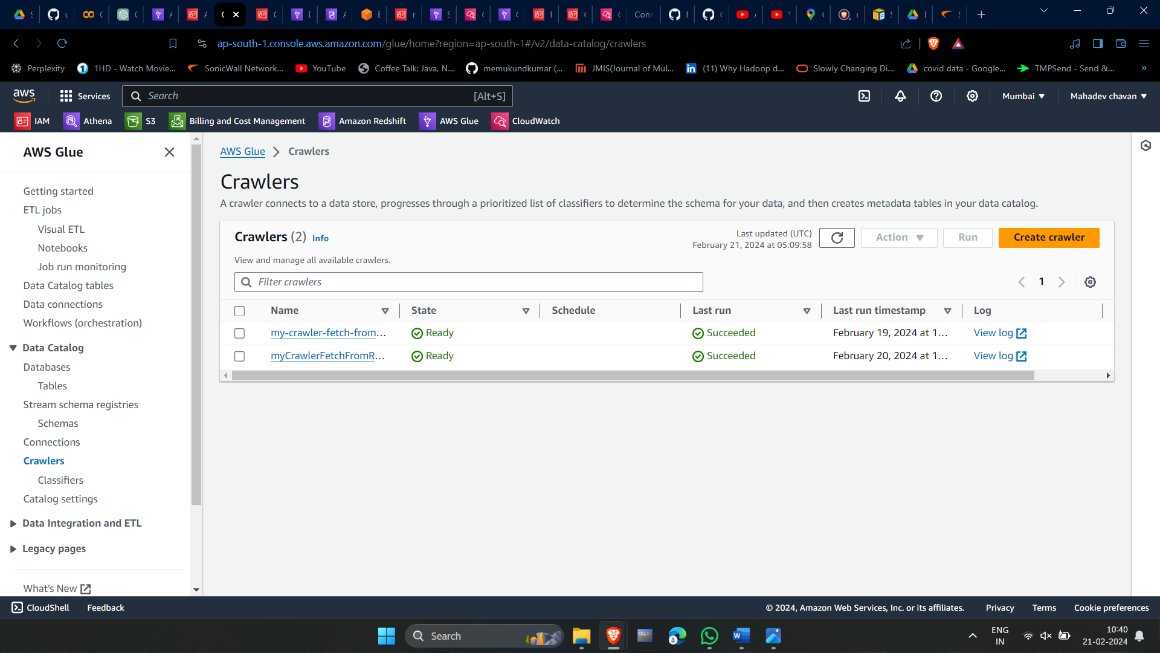


Fig. 7- Glue Crawler with 10 partition regions

Crawled Data

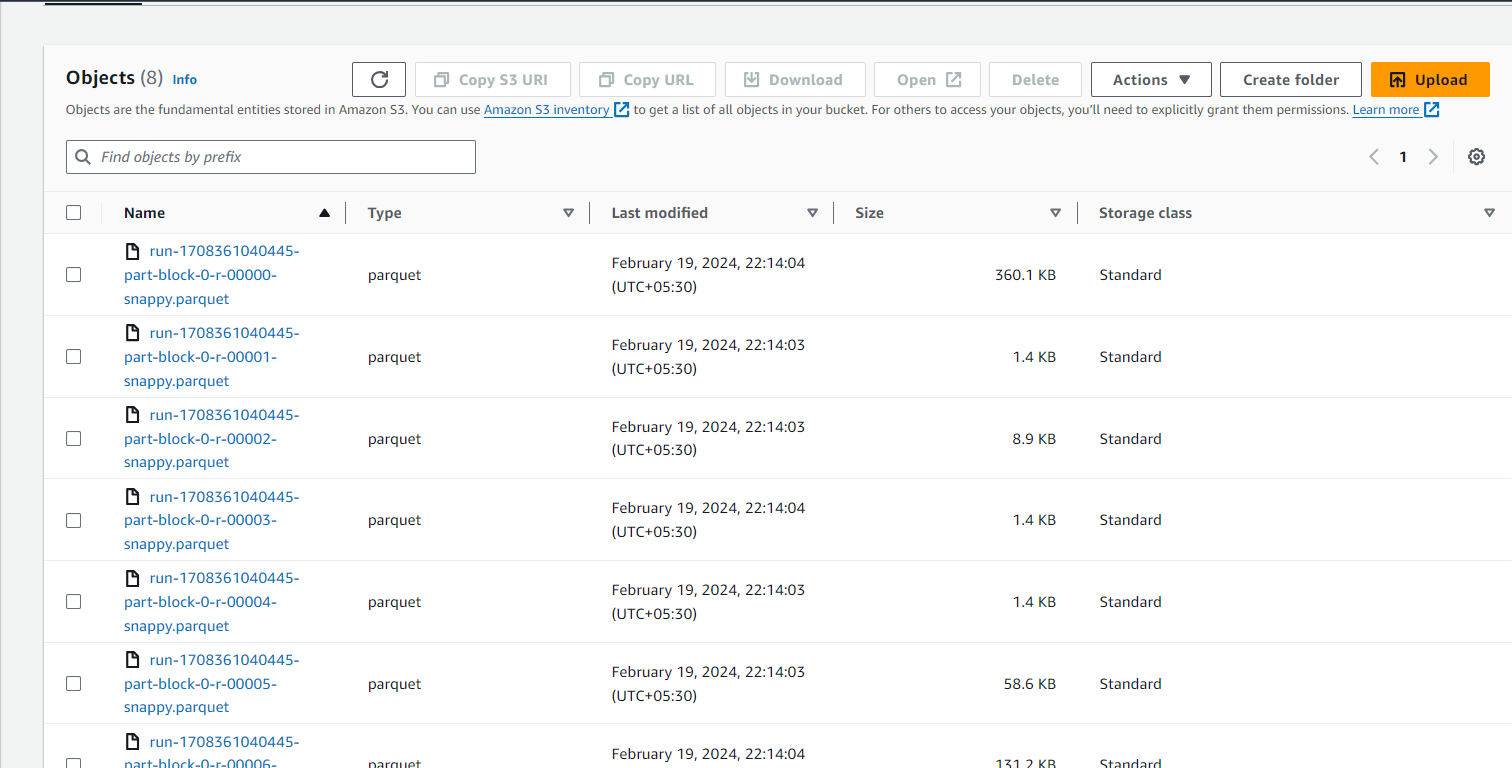


Fig. 8- Crawled Data Tables

# AWS Glue

Glue is a serverless data integration service that makes it easier to discover, prepare, move, and integrate data from multiple sources. It simplifies ETL pipeline development. It supports various processing frameworks and workloads .

An ETL job was performed using AWS glue. The job extracted the raw .csv data from the s3 bucket through tables. Transformed the data by defining schema structure and predicate pushdown to select the region which filters the data before the source is extracted. It improves the ETL process and optimizes resources used.

Then we loaded the transformed data to a new s3 bucket. Serverless Data Integration: AWS Glue provides a fully managed, serverless data integration service that makes it easy to discover, prepare, and load data for analytics. By abstracting away the underlying infrastructure management, AWS Glue allows users to focus more on data processing logic rather than infrastructure provisioning and maintenance.

ETL (Extract, Transform, Load) Operations: AWS Glue is commonly used for ETL operations, which involve extracting data from various sources, transforming it into a desired format, and loading it into a target destination for analysis. This process is crucial for data warehousing, analytics, and business intelligence applications.

Cataloging and Metadata Management: AWS Glue includes a metadata repository that automatically crawls and catalogs data sources to capture metadata such as table definitions, schemas, and partitions. This metadata management capability simplifies data discovery and ensures data consistency across different systems.

Apache Spark Integration: AWS Glue leverages Apache Spark, a powerful distributed data processing engine, to execute ETL workflows at scale. Spark provides a unified analytics engine for large-scale data processing, enabling Glue to handle diverse data processing tasks efficiently.

Job Orchestration and Scheduling: With AWS Glue, users can orchestrate and schedule ETL jobs using built-in features or integrate with AWS Step Functions for more advanced workflow orchestration. This allows for the automation of data processing pipelines, ensuring timely execution of tasks and efficient resource utilization.

Data Transformation and Enrichment: AWS Glue supports various transformation operations, such as filtering, joining, aggregating, and enriching data, using familiar programming constructs like SQL, Python, and Scala. These transformations enable users to prepare data for downstream analytics and reporting requirements.

Integration with AWS Services: AWS Glue seamlessly integrates with other AWS services, such as Amazon S3, Amazon Redshift, Amazon RDS, Amazon Aurora, Amazon DynamoDB, and more. This integration facilitates data movement, storage, and analysis across different AWS environments, providing a comprehensive data ecosystem.

Cost Optimization: As a serverless service, AWS Glue offers cost-effective pricing models based on usage, with no upfront commitments or long-term contracts. Users only pay for the resources consumed during data processing, making it suitable for both small-scale and enterprise-level data integration projects.

## Parquet File format

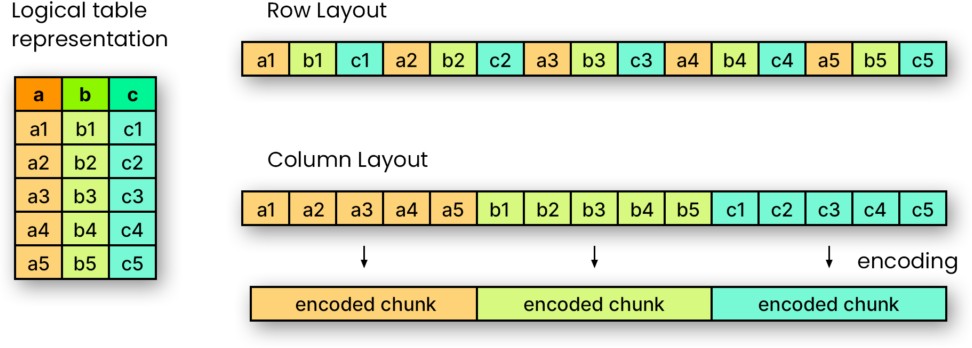


Fig. 12- Parquet file format

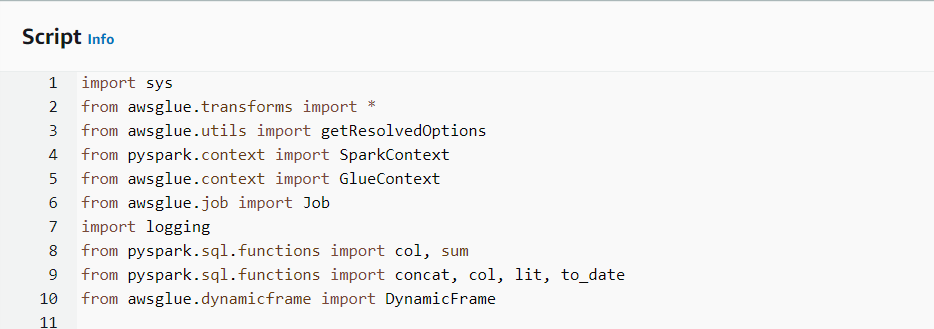
Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval. It provides efficient data compression and encoding schemes with enhanced performance to handle complex data in bulk.

Parquet files are smaller than CSV files, and they can be read and written much faster. Parquet files also support nested data structures, which makes them ideal for storing complex data query services like AWS EMR ( Apache Hive ) or Amazon Athena charge you by the amount of data scanned per query.

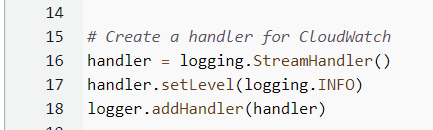
Google and Amazon charge you for the amount of data stored on GS/S3.

## Code Snippets

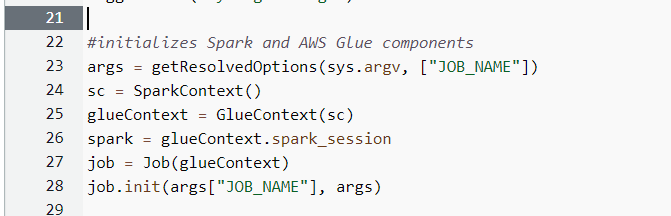
**Importing necessary Packages**



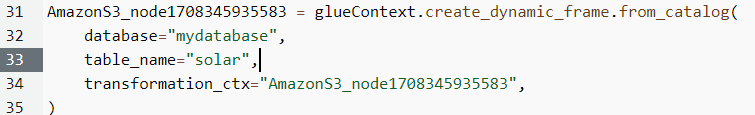
**# Create a handler for CloudWatch**



#**initializes Spark and AWS Glue components**



**# creating dynamic frame from s3 data object**

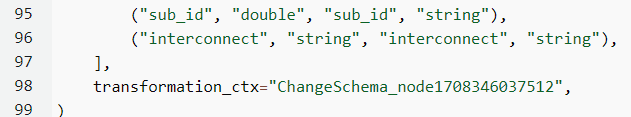


**# Script created for node Change Schema**

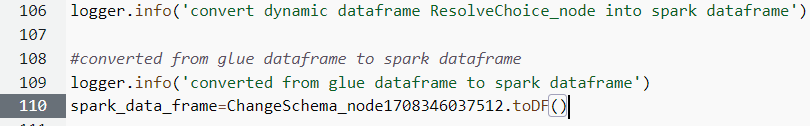




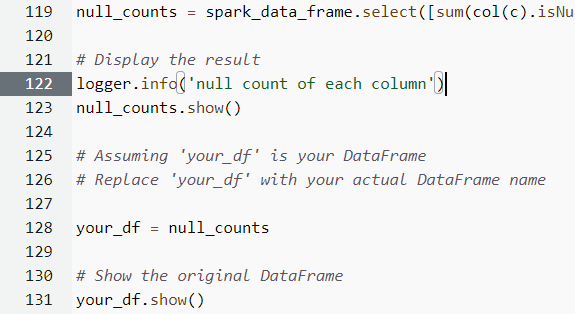
****



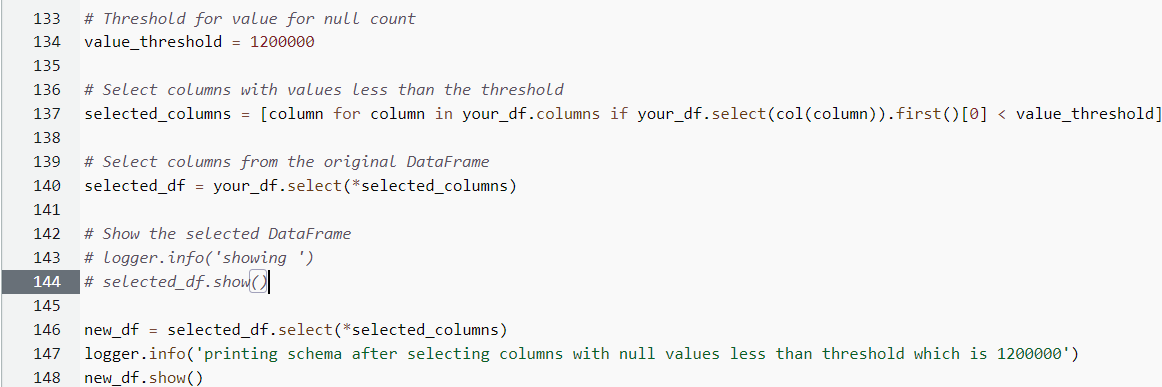
**# Convert dynamic dataframe into spark dataframe**



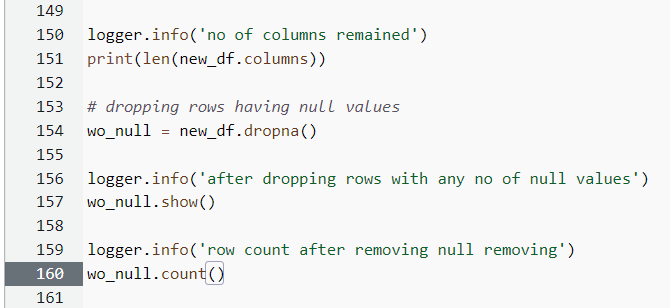
**# null count of each column**



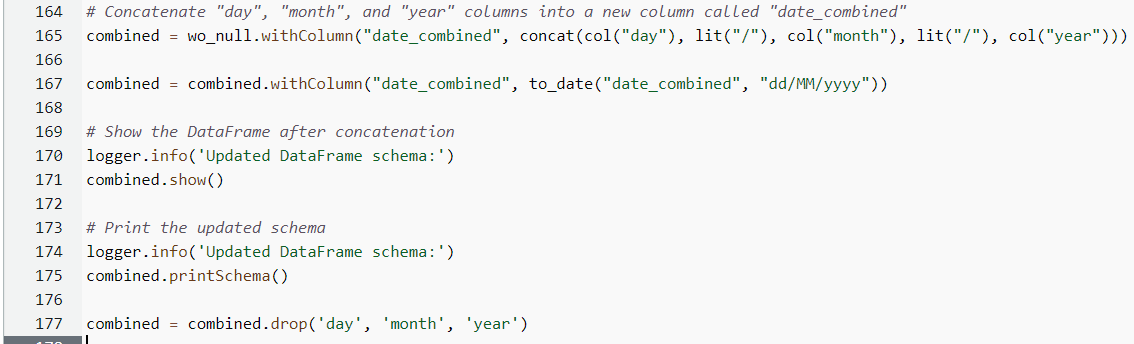
**# Select columns with values less than the threshold**



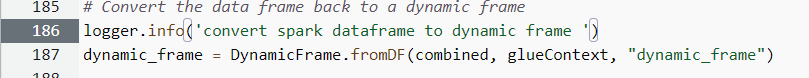
**# dropping rows having null values**



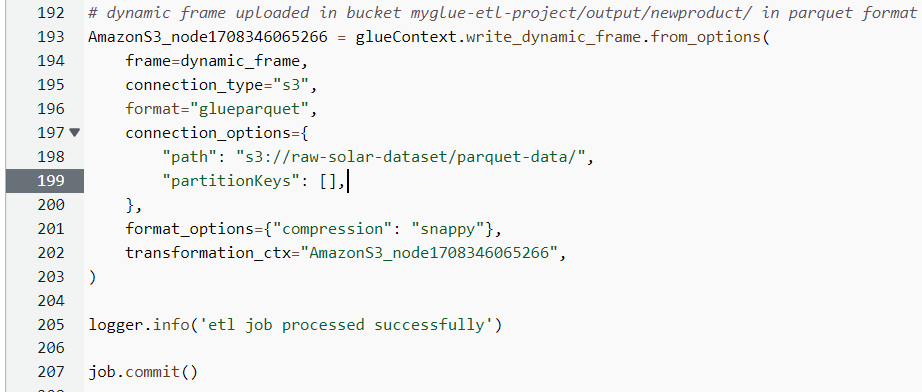
**# Concatenate "day", "month", and "year" columns into a new column called "date\_combined"**



**# Convert the data frame back to a dynamic frame**



**#dynamic frame uploaded in bucket myglue-etl-project/output/newproduct/ in parquet format**



## Tableau

Introduction to Tableau:

Tableau is a robust data visualization and business intelligence tool designed to help users analyze and interact with their data in meaningful ways. With its intuitive interface and powerful features, Tableau has become a popular choice for organizations and individuals seeking to transform raw data into actionable insights.

Connecting to Data Sources:

One of Tableau's key strengths is its ability to connect to a diverse range of data sources. Users can seamlessly integrate databases, spreadsheets, and cloud-based storage, providing flexibility in working with various types of data. Furthermore, Tableau supports both live connections for real-time data querying and extracts for improved performance.

Building Visualizations in Worksheets:

Tableau's main workspace is the worksheet, where users create visualizations by dragging and dropping dimensions and measures onto shelves. The Show Me menu offers a variety of chart types, and users can design interactive and dynamic visualizations to uncover patterns and trends in their data.

Designing Interactive Dashboards:

Dashboards in Tableau bring together multiple worksheets and objects on a single canvas, allowing users to create comprehensive presentations of their insights. These dashboards are interactive, providing a cohesive view of the data, and users can implement filters and parameters for dynamic control over the displayed information.

Data Preparation and Transformation:

Tableau provides tools for managing and preparing data within the platform itself. Users can clean, shape, and transform their data, making it easier to derive meaningful insights. Calculated fields and various functions allow for the creation of custom measures and dimensions to enrich the analysis.

Publishing and Sharing Insights:

Tableau workbooks can be published to platforms such as Tableau Server or Tableau Online, enabling users to share their insights with colleagues or a broader audience. Additionally, Tableau Public allows for the public sharing of visualizations, fostering a community-driven approach to data exploration and analysis.

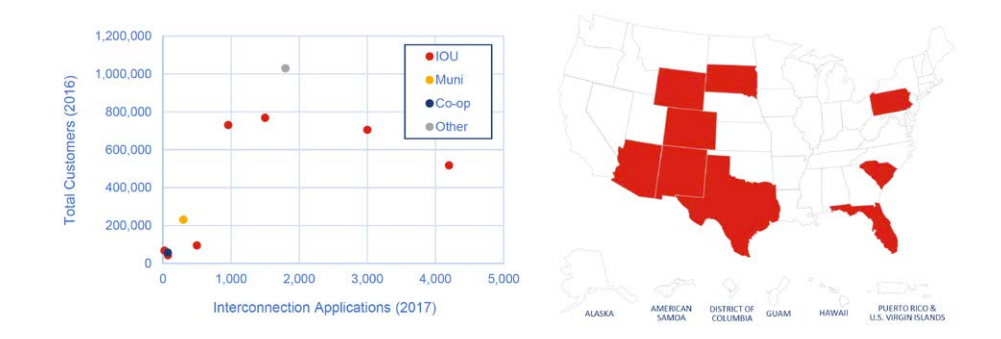
Community and Support:

The Tableau community plays a vital role, offering a platform for users to share knowledge, seek assistance, and access valuable resources. With forums, tutorials, and sample datasets, Tableau encourages collaborative learning and problem-solving among its user base.

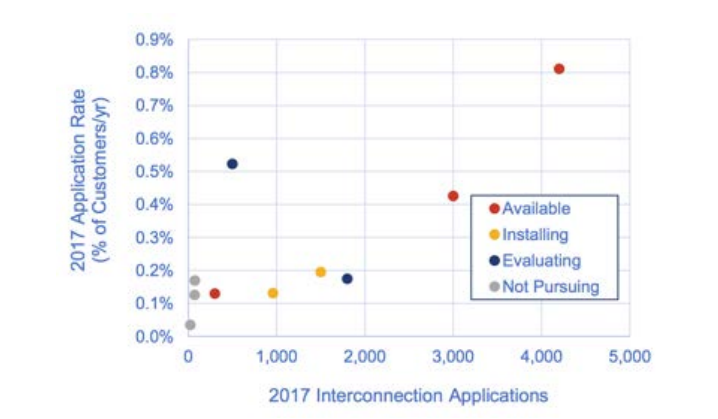
**Insights from Tableau Visualization**

x

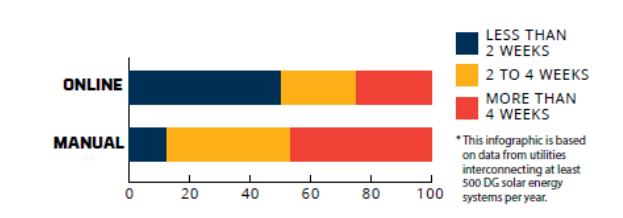
Fig. U.S. annual energy storage deployment history (2012–2017) and forecast (2018–2023), in MW, from GTM Research (2018)

****

**Fig.** Characteristics of utilities interviewed by SEPA about interconnection practices

****

**Fig.** Online interconnection platform status for utilities interviewed by SEPA in July 2018

****

**Fig.** Application processing times: manual processes versus online

Conclusion

The presence of energy storage in a limited number of installations indicated a gradual adoption of storage solutions. Additionally, the prevalence of ground-mounted systems suggested diverse preferences in system deployment, with potential implications for land use planning.

To advance the understanding of solar energy dynamics, future research could delve into the socio-economic factors influencing adoption, the impact of policy changes on the industry, and advancements in emerging technologies. Continued monitoring and analysis will contribute to a dynamic and responsive solar energy landscape.

Based on the findings, recommendations include targeted investments in regions with high solar potential, seasonal adjustments to installation planning, and a focus on technology advancements to enhance overall system efficiency. Insights into customer class preferences can guide tailored marketing strategies, while considerations for storage and ground-mounted systems can inform future project designs.

## References

* 1. <https://en.wikipedia.org/wiki/Web_crawler>
  2. <https://github.com/microsoft/us_der_dataset>
  3. Data Analysis with Python and PySpark by Jonathan Roux
  4. <https://aws.amazon.com/glue/getting-started/>
  5. <https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html>
  6. <https://help.tableau.com/current/pro/desktop/en-us/gettingstarted_overview.html>