The Glue crawler, based on the details you provided, is used to **automatically detect and catalog data in its raw form** from a source location (usually S3) into the AWS Glue Data Catalog. Here's an explanation of its properties and purpose:

Glue Crawler Properties

1. Name:

- `de-on-youtube-cleaned-csv-to-parquet-etl`
  - This name indicates the crawler is part of a pipeline where raw CSV data is transformed into Parquet format. It scans the source location (S3) and creates or updates metadata in the Glue Data Catalog for further processing or querying.

2. IAM Role:

- `de-on-youtube-glue-s3-role`
  - The IAM role allows the Glue crawler to access:
    - The **S3 bucket** where the raw data resides.
    - The **Glue Data Catalog** to create or update metadata.
    - Any other associated AWS resources required for its operation.

3. Database:

- `db_youtube_cleaned`
  - This is the Glue database where the crawler stores metadata about the tables it discovers. A Glue database is a logical grouping for organizing tables within the Glue Data Catalog.

4. State:

- `READY`
  - Indicates the crawler is configured and ready to run.

5. Description:

- `-`
  - No description provided, but it would typically outline the crawler's purpose (e.g., scanning raw YouTube data in CSV format).

6. Security Configuration:

- `-`
  - No additional security configuration applied. If needed, encryption or Lake Formation permissions could be set here for sensitive data.

7. Lake Formation Configuration:

- -
    - Not explicitly configured, meaning the crawler relies on Glue-native permissions for data access rather than Lake Formation.

8. Table Prefix:

- **`(not mentioned)`**
    - If a prefix is specified, the crawler adds it to table names it creates. For example, if the prefix is `raw_`, tables might be named `raw_table_name`.

Purpose of This Crawler

1. **Discover and Catalog Raw Data**:
    - The crawler scans raw data (likely CSV files) in the specified S3 bucket or folder.
    - It infers the schema (column names, data types, etc.) and creates or updates metadata in the Glue Data Catalog.
2. **Enable Querying of Raw Data**:
    - Once the metadata is cataloged, tools like Amazon Athena can query the raw data directly.
3. **Integration with ETL Pipelines**:
    - This crawler is likely the **first step** in your ETL pipeline. It processes raw data, and a subsequent process (like your Lambda function) transforms it into a cleaned Parquet format for analytics.

How It Works

1. **Scan the Source**:
    - The crawler scans the S3 location specified during its configuration.
    - For example, if it's configured to scan `s3://de-on-youtube-raw-us-east-1/`, it processes files in that path.
2. **Infer the Schema**:
    - Based on the raw data (e.g., CSV headers), it determines:
        - Column names
        - Data types (e.g., `string`, `int`, `float`)
        - Table structure
3. **Create or Update Tables**:
    - It creates new tables in the **`db_youtube_cleaned`** database or updates existing ones with any new data or schema changes.

Example Workflow

1. **Raw Data Ingestion**:

    - Raw CSV files are uploaded to S3 (e.g., `s3://de-on-youtube-raw-us-east-1/`).
2. **Run the Crawler**:
    - The crawler scans the bucket and creates a table in the Glue database (`db_youtube_cleaned`) with the raw data schema.
3. **Transformation**:
    - The Lambda function reads this raw data and transforms it into Parquet format, writing the results to a different location in S3.
4. **Cleaned Data Cataloging**:
    - The Lambda function updates the Glue Data Catalog with metadata about the cleaned Parquet data.
5. **Query Data**:
    - Athena or other tools can query both raw and cleaned data as needed.
