The **Glue Crawler** described here is configured to catalog raw YouTube data from an S3 bucket into the Glue Data Catalog. Here's a detailed explanation of the properties and its purpose:

Glue Crawler Properties

1. Name:

- **`de-on-youtube-raw-glue-catalog-1`**
  - This crawler specifically catalogs **raw statistics reference data** from the S3 bucket.

2. IAM Role:

- **`de-on-youtube-glue-s3-role`**
  - This IAM role grants the crawler permissions to:
    - Access the S3 bucket (`s3://de-on-youtube-raw-us-east-1-dev-vignesh/youtube/raw_statistics_reference_data/`).
    - Update the Glue Data Catalog with the inferred schema and metadata.

3. Database:

- **`de_youtube_raw`**
  - The crawler saves the metadata (table definitions) in this Glue database.
  - **Purpose**: This database logically groups all tables related to raw YouTube data.

4. State:

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

5. S3 Target:

- **`s3://de-on-youtube-raw-us-east-1-dev-vignesh/youtube/raw_statistics_reference_data/`**
  - This is the **source location** in S3 where the raw data files are stored.
  - The crawler will scan this location for files, infer their schema, and catalog them.

6. Recrawl Behavior:

- **`Recrawl all`**
  - The crawler will **recrawl the entire dataset** each time it is run. This means:
    - It will check for changes in the data, such as new files, updated files, or schema modifications.
    - If new columns or files are detected, it updates the existing Glue table or creates a new table.

What This Crawler Does

1. **Scans the S3 Source**:
   - It scans the raw data files in `s3://de-on-youtube-raw-us-east-1-dev-vignesh/youtube/raw_statistics_reference_data/`.
2. **Infers Schema**:
   - Based on the raw files (likely CSV or JSON), the crawler determines:
     - Column names
     - Data types (e.g., `string`, `int`, `float`)
     - Partition columns (if applicable)
3. **Creates/Updates Metadata**:
   - It catalogs the inferred schema in the Glue Data Catalog under the database **`de_youtube_raw`**.
   - If the table already exists, it updates the schema if changes are detected.
4. **Makes Raw Data Queryable**:
   - Tools like Athena or Redshift Spectrum can query the raw data directly using the cataloged table.

Workflow Example

Step 1: Raw Data Ingestion

- Raw data files (e.g., CSV or JSON) are uploaded to the S3 bucket: `s3://de-on-youtube-raw-us-east-1-dev-vignesh/youtube/raw_statistics_reference_data/`

Step 2: Crawler Execution

- The crawler runs and:
  - Scans the bucket for raw data.
  - Infers the schema.
  - Creates or updates a Glue table (e.g., `raw_statistics_reference_data`) in the **`de_youtube_raw`** database.

Step 3: Catalog Update

- Metadata about the raw data (e.g., columns, data types, S3 location) is stored in the Glue Data Catalog.

Step 4: Query Raw Data

- You can use Athena to query the raw data directly:

```
SELECT *
FROM "de_youtube_raw"."raw_statistics_reference_data"
LIMIT 10;
```

Step 5: Downstream ETL

- A downstream process (like a Lambda function or ETL job) can transform this raw data into a cleaned format (e.g., Parquet), ready for analytics.

Advantages of This Crawler Configuration

1. **Automates Metadata Management**:
    - Eliminates manual schema management by dynamically updating the Glue Data Catalog.
2. **Keeps Data Updated**:
    - The **recrawl all** setting ensures that any new files or schema changes are detected and cataloged automatically.
3. **Enables Querying of Raw Data**:
    - Immediate querying of raw data via Athena or other AWS services.
