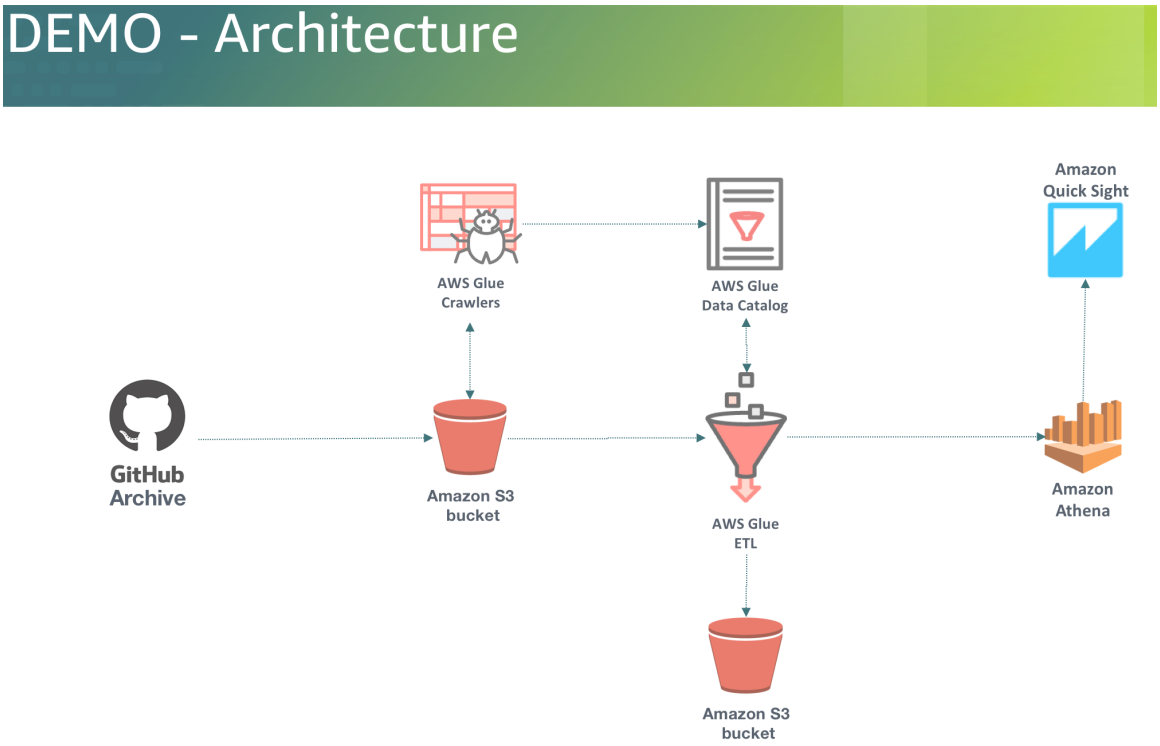
Building Data Lake on AWS with S3, Glue and Athena

Author: Rudi Suryadi

Architecture - Diagram



Prerequisites:

- Have access to an AWS account in which your user has **AdminstratorAccess**
- This lab should be executed in **ap-southeast-1** region. Best is to **follow links from this guide**
- Access to a modern browser 😜

Part 1 : Ingest and Storage

Download Sample Data from GitHubArchive

- Download this file : http://data.gharchive.org/2018-03-01-15.json.gz

Create S3 Bucket

In this step we will navigate to S3 Console and create the S3 bucket used throughout this demo.

Login to AWS Console : https://console.aws.amazon.com/console/home?region=ap-southeast-1

Navigate to S3 Console & Create a new bucket in ap-southeast-1 region :

- Goto : https://console.aws.amazon.com/s3/home?region=ap-southeast-1
- Click - **Create Bucket**
- Bucket Name : **yourname-etl-demo-bucket**
- Region : **Asia Pacific (Singapore)**
- Click **Create** (bottom left)

Upload Sample Data to S3 Bucket

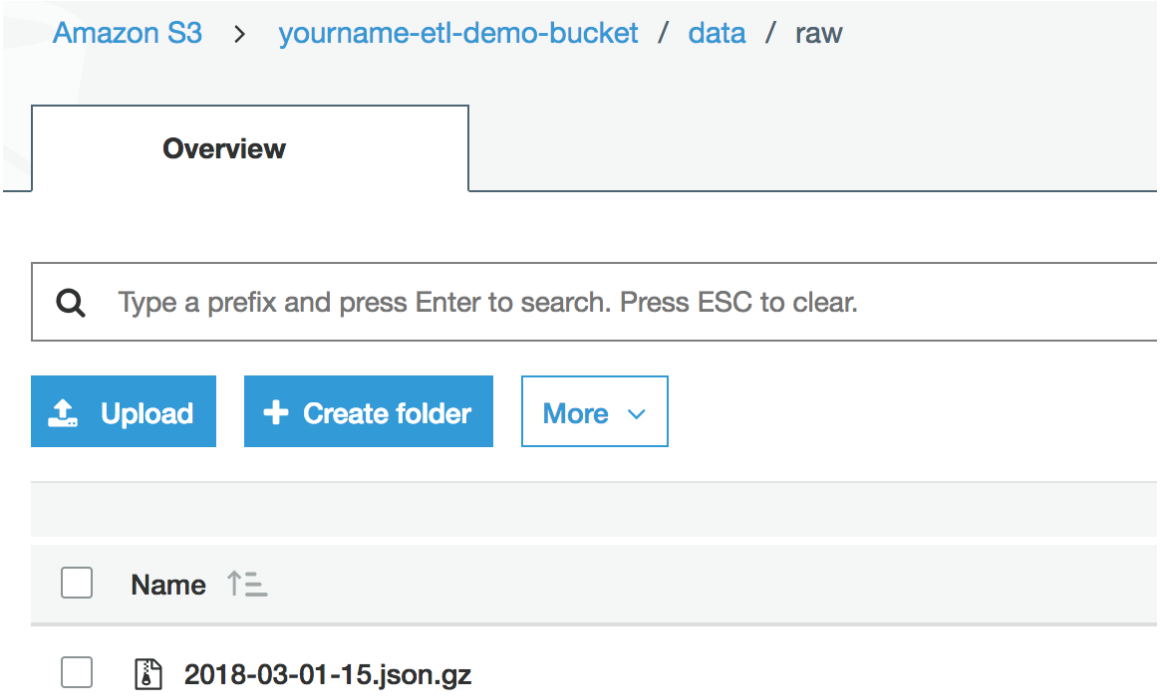
In this step we will navigate to S3 Console and upload the sample data used in this lab.

- GoTo : https://console.aws.amazon.com/s3/home?region=ap-southeast-1
- Open Bucket : **yourname-etl-demo-bucket**
- Click : **Create folder**
- Folder Name : **data**
- Click : **Save**
- Open Folder : **data**
- Click : **Create folder**
- Folder Name : **raw**
- Click : **Save**

You should have a folder structure similar to this

- Click : **Upload**
- Click : **Add Files** > Navigate & upload the downloaded **2018-03-01-15.json.gz** file
- Click : **Upload**

By now your S3 bucket should look like this



Part 2 : Catalog and Transform

Create IAM Role

In this step we will navigate to IAM Console & create a new Glue service role, this allows AWS Glue to access data sitting in S3 and create necessary entities in Glue catalog.

- Goto: https://console.aws.amazon.com/iam/home?region=ap-southeast-1#/roles
- Click - **Create role**
- Choose the service that will use this role: **Glue**
- Click - **Next: Permissions**
- Search for - **AmazonS3FullAccess**
- Select **Checkbox**
- Search for - **AWSGlueServiceRole**
- Select **Checkbox**
- Click - **Next: Review**
- Role name: **AWSGlueServiceRoleDefault**
- make sure that are two policies attached to this role (**AmazonS3FullAccess**, **AWSGlueServiceRole**)
- Click - **Create role**

Create AWS Glue Crawlers

In this step, we will navigate to AWS Glue Console & create glue crawlers to discovery the newly ingested data in S3.

- Goto: https://ap-southeast-1.console.aws.amazon.com/glue/home?region=ap-southeast-1
- On the left panel, click on **Crawlers** > Click on **Add Crawler**

- Crawler info
  - Crawler name: **innovate-crawler**
  - Click - **Next**
  - Data store
  - Data store: **S3**
  - Crawl data in: **Specified path in my account**
  - Include path: **s3://yourname-etl-demo-bucket/data/**
  - Click - **Next**
- Add another data store : **No**
- Click - **Next**
- IAM Role
  - Choose: **Choose an existing IAM role**
  - Role Name: **AWSGlueServiceRoleDefault**
  - Click - **Next**
- Schedule
  - Frequency: **Run on demand**
  - Click - **Next**
- Output
  - Click - Add database
  - Database name: **innovate-db**
  - Click - **Create**
  - Click - **Next**
- Review all steps
  - Review the configuration & make sure its as mentioned above
  - Click - **Finish**

You should see a message : Crawler **innovate-crawler** was created to run on demand.

- Click - **Run it Now?** this will run the crawler

Wait for few minutes

Verify newly created tables in catalog

Navigate to Glue Catalog & explore the crawled data:

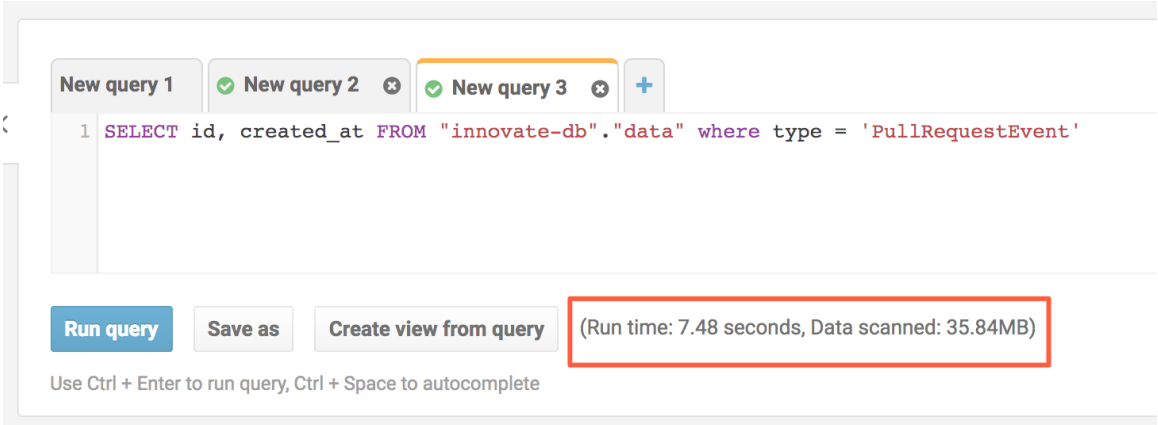
- Goto : https://ap-southeast-1.console.aws.amazon.com/glue/home?region=ap-southeast-1#catalog:tab=databases
- Click - **innovate-db**
- Click - **Tables in innovate-db**
- Click - **data**
- Look around and explore the schema for your dataset
- Look for the average recordSize, recordCount, compressionType

Query newly ingested data using Amazon Athena

- Goto : https://ap-southeast-1.console.aws.amazon.com/athena/home?region=ap-southeast-1#query
- On the left panel (**Database**) drop down , select **innovate-db** > select table **data**
- Cick on **3 dots** (3 vertical dots) > Select **Preview Table**
- In query editor, paste the following query > Click on **Run Query**

```
SELECT id, created_at FROM "innovate-db"."data" where type = 'PullRequestEvent'
```

- One the query execution finishes, note down the**Run time** & **Data scanned**** statistics



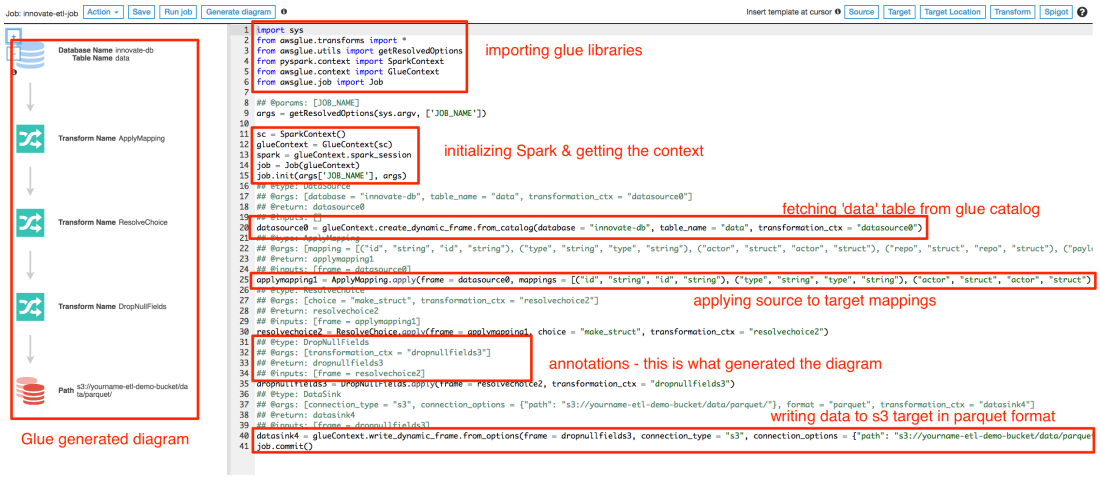
Transform data - write your ETL job

In this step you will convert the JSON files to parquet

Navigate to Glue Console and Transform your data:

- Goto : https://ap-southeast-1.console.aws.amazon.com/glue/home?region=ap-southeast-1#etl:tab=jobs
- Click - **Add job**
- Job properties:
  - Name: **innovate-etl-job**
  - IAM Role: **AWSGlueServiceRoleDefault**
  - This job runs: **A proposed script generated by AWS Glue**
  - ETL language: **Python**
  - Leave everything else to default
  - Expand **Security configuration, script libraries, and job parameters (optional)**
  - Concurrent DPUs per job run :**2** (this is the capacity of underlying spark cluster that Glue uses)
  - Click - **Next**
- Choose your data sources:
  - Select : [Name] = **data** | [Database] = **innovate-db**
  - Click : **Next**
- Choose your data targets:
  - Select: **Create tables in your data target**
  - Data store : **Amazon S3**
  - Format : **Parquet**
  - Target Path : **s3://yourname-etl-demo-bucket/data/parquet/**
  - Click : **Next**
- Map the source columns to target columns:
  - Leave default options
  - Click : **Next**
- Review:
  - Review the job configuration and properties & ensure its same as mentioned above
  - Click : **Save job and edit script**

  This is the **AWS Glue Script Editor**. Here is where you will author your ETL logic.



- Review the code in the editor & explore the UI (do not make any changes to the code at this stage)
- Click - **Save**
- Click - **Generate Diagram**
- Click - **Run** > **Run Job**

First time execution of the Job takes up to 10-20 minutes.

Validate that processed data has arrived in S3

Once the ETL script has ran successfully.

- Goto : S3 console : https://s3.console.aws.amazon.com/s3/home?region=ap-southeast-1#
- Navigate :
  - Click - **yourname-etl-demo-bucket > data**

There should be a folder called **parquet** created here

Catalog transformed data

Now that we have transformed the raw data and put it in parquet folder in our S3 bucket, we should re-run the crawler to update the catalog information.

- Goto : https://ap-southeast-1.console.aws.amazon.com/glue/home?region=ap-southeast-1#catalog:tab=crawlers
- Click : **innovate-crawler** > **Run Crawler**

Wait for few minutes

Once the crawler has stopped make, 2 new table has been added to the catalog. (Table 1 - raw , Table 2 - parquet -> as the crawler is crawling the parent **data** directory)

Part 3 : Analyze

Explore our data set using Athena

In this step we will analyze the transformed data using Athena

- Goto : https://ap-southeast-1.console.aws.amazon.com/athena/home?region=ap-southeast-1#query
- On the left panel (**Database**) drop down , select **innovate-db** > select table **parquet**
- Cick on **3 dots** (3 vertical dots) > Select **Preview Table**
- In query editor, paste the following query > Click on **Run Query**

```
SELECT id, created_at FROM "innovate-db"."parquet" where type = 'PullRequestEvent'
```

- One the query execution finishes, note down the **Run time** & **Data scanned**** statistics

Homework : Find out why this happened ?
How can you further optimize how Athena reads from S3.

- Explore the Athena UI and try running some queries

Part 4: Visualize

Setting up data access

Login to Amazon Quick Sight Console & complete the registration & sign-up https://ap-southeast-1.quicksight.aws.amazon.com/sn/start

- Goto : https://us-east-1.quicksight.aws.amazon.com/sn/console/resources
- Select All options
- Click : **S3 Bucket** > select **yourname-etl-demo-bucket** . Click **Select Bucket**
- Click : **Apply**

Using Amazon Quick Sight to visualize our processed data

In this step we will visualize it using QuickSight

- Change region : **Singapore**
- Goto: https://ap-southeast-1.quicksight.aws.amazon.com/sn/start
- Adding a new dataset:
  - On top right, Click - **Manage Data**
  - Click - **New Data Set**
  - Click - **Athena**
  - New Athena data source
  - Data source name: **innovate-db**
  - Click - **Create data source**
- Choose your table:
  - Database: contain sets of tables: select - **innovate-db**

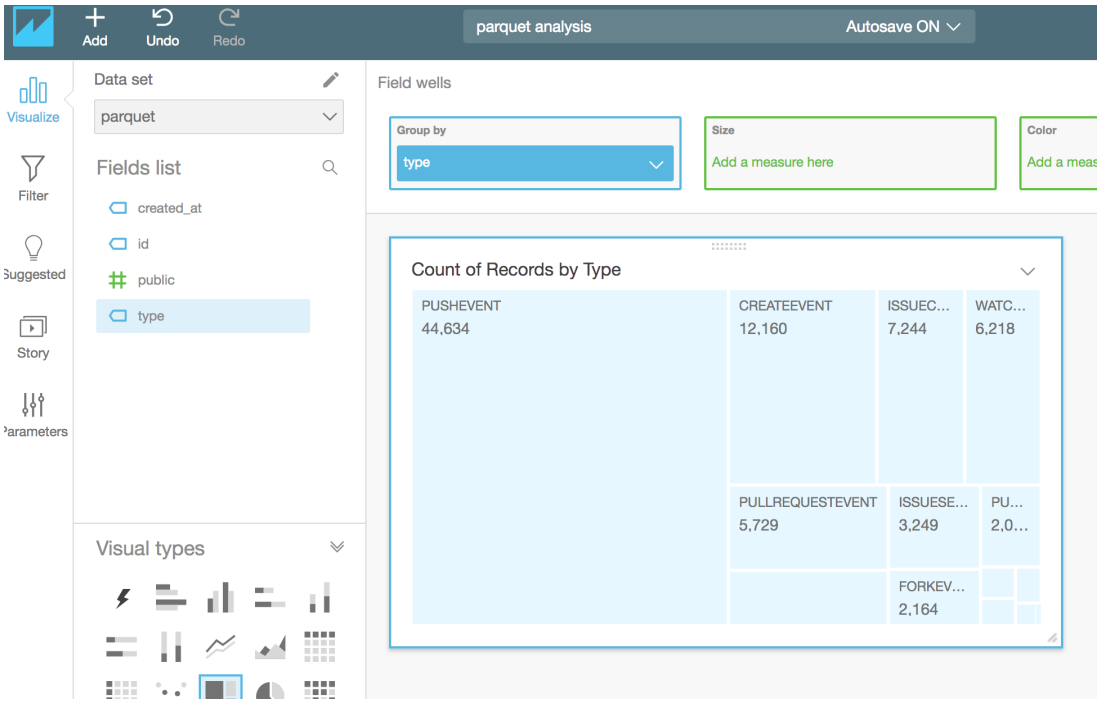
- Tables: contain the data you can visualize : select -**parquet**
- Click - **Select**

- Finish data set creation:
  - Select - **Directly query your data**
  - Click **Visualize**

Visualization : Tree map of most played Artist Names

In this step we will create a visualization that shows who are the host played artists

- On the bottom-left panel -**Visual types**
- Hover on icon there to see names of the visualizations
- Click on - **Tree Map**
- On top-left panel - **Fields list**
- Click - **type**



Play around and explore Amazon QuickSight Console. Try out filters, other visualization types, etc.

Clean Up

- Delete Glue Database - **innovate-db**
- Delete Glue Job - **innovate-etl-job**
- Delete S3 bucket - **yourname-etl-demo-bucket**
- Delete QuickSight account https://docs.aws.amazon.com/quicksight/latest/user/closing-account.html