SOP of KBZ MS Data Lake Creation (Thamawaddy Data)

Flow Chart

Background Theory

In this SOP, Data Lake will be created on AWS with static files primarily stored in S3 storage and directly run SQL query against files stored in S3 with Athena service. Athena supports most of the popular file format such as CSV, JSON, Parquet, ORC etc. CSV format will be used mainly for this SOP. Redshift offers the similar functions as Athena, but Athena is more suitable for Ad-hoc data discovery and SQL querying.

Data Source

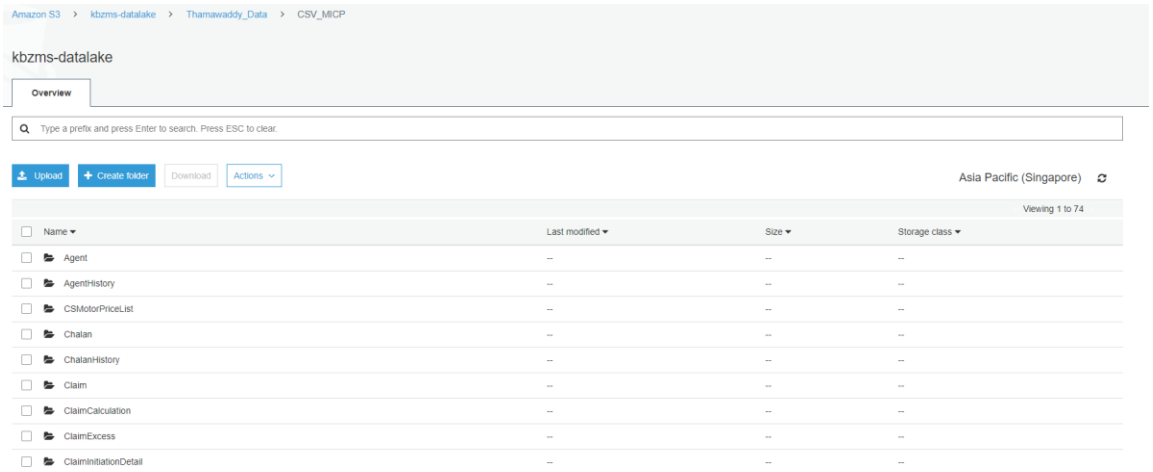
1. Source files are in sql format directly from Thamawaddy databases. They are stored in the following location in S3:
   s3://kbzms-datalake/Thamawaddy_Data/sql/
   Data dictionaries and corresponding table creation sql files are also included in the above location.
2. Since sql is not supported by Athena, sql file must be converted to CSV format.
3. Dummy Databases have been created on local machine to convert the sql file to CSV.
4. Relating sql files for dummy databases creation can also be found in the directory mentioned in (1).

Note:

- CSV must be in UTF-8 format in order to display Myanmar fonts correctly.
- Since converted CSV files use (,) as separating identifier, comma in the file content has been replaced with (|). This process can be reverse if the exported csv file uses (|) as separating identifier.

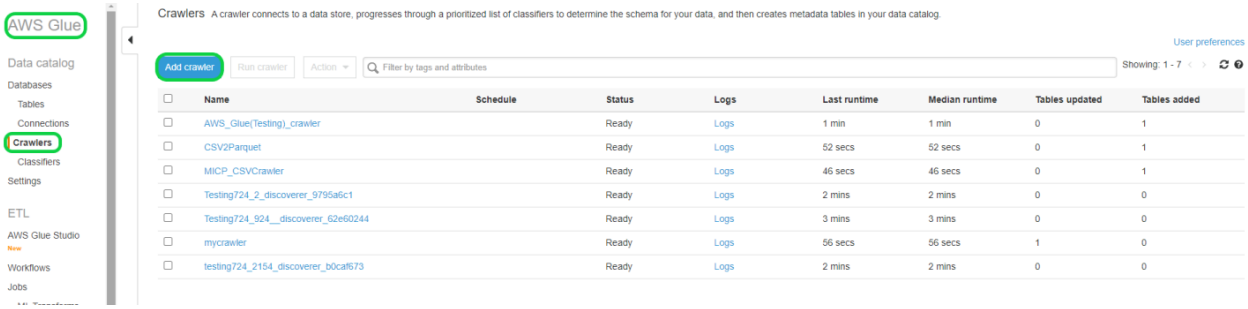
Data Lake location

1. Converted CSV files are available in the following location in S3:
   s3://kbzms-datalake/Thamawaddy_Data/CSV_MICP
   s3://kbzms-datalake/Thamawaddy_Data/CSV_Others
2. Each CSV file will represent a table to be queried by Athena. But Glue service cannot find the correct metadata when there is more than one csv file in the target folder, hence, separated folders are created to accommodate each table.



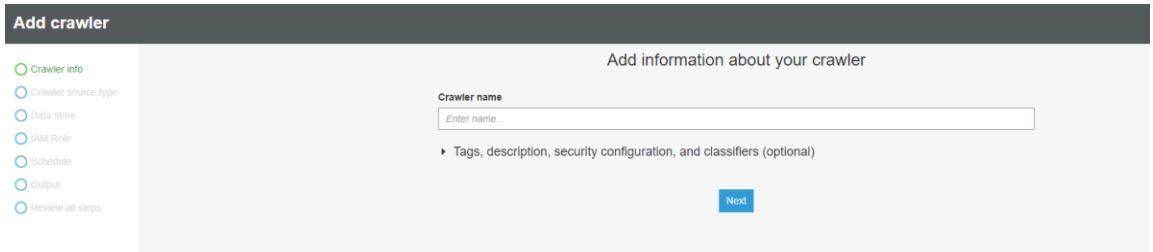
Glue Service

- Glue is the most critical part of this SOP.
- Glue service can be found in AWS under “Analytics”.

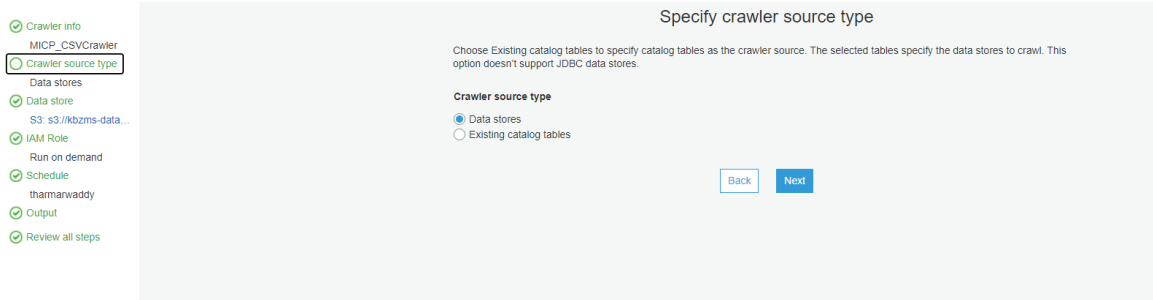


- From Glue main page, a new crawler will be created.
  AWS Glue > Crawlers > Add crawler

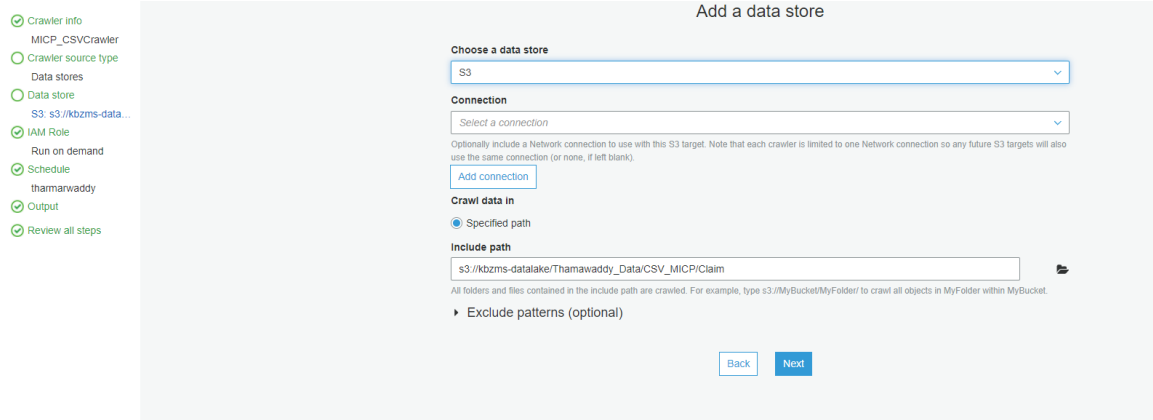
- Crawler name can be anything user desire. Then click ‘Next’.



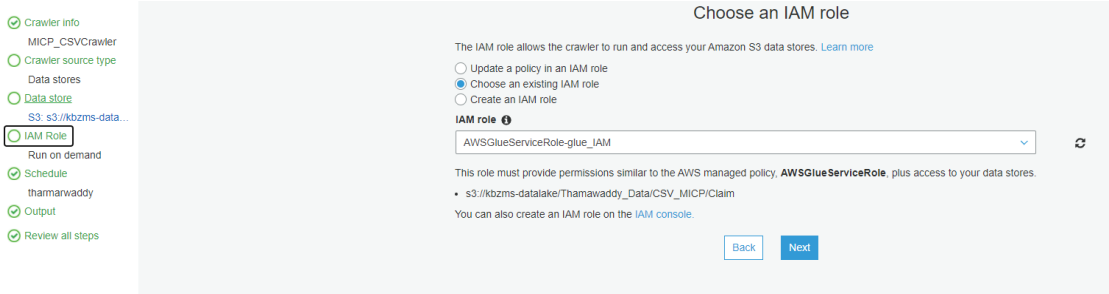
- Data Stores will be chosen for crawler source type in the next page as there is no existing catalog tables.



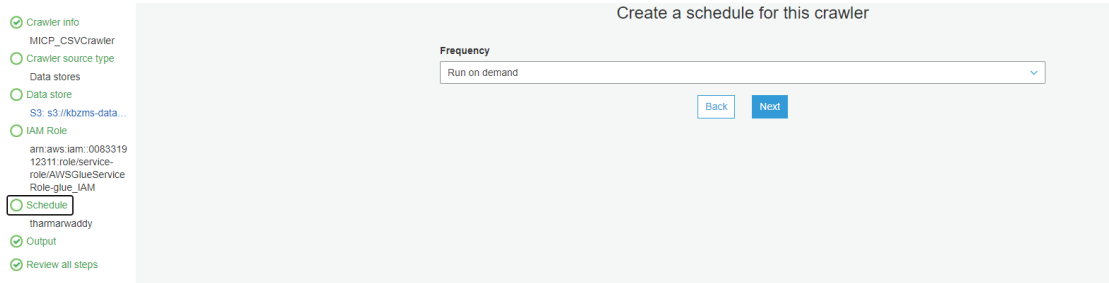
- The next page, ‘Data Store’ is important as Glue’s crawler will be informed what the data store is according to this page.
- Here, since the data source is simple S3 bucket, and connection is not required (connection must be established for database type which uses Java DB connectivity (JDBC) or DB type is Dynamo or Mongo).
- Specified path will be declared for crawler to know the location of the source CSV file in S3.



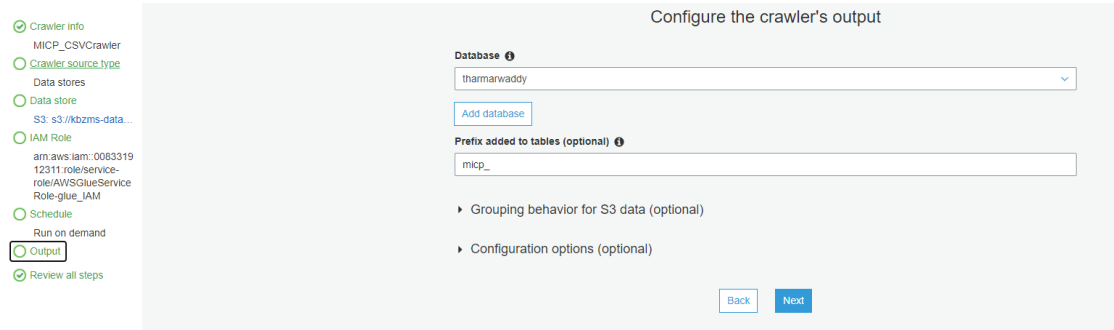
- IAM role for crawler has to be chosen in order to let crawler read and write for both S3 and Athena.
- AWS default IAM role can be used or add/edit IAM role under AWS IAM service. IAM role name must not include whitespace.
- Here, ‘AWSGlueServiceRole-glue_IAM’ will be used which has permission to KBZ MS Data Lake in S3.



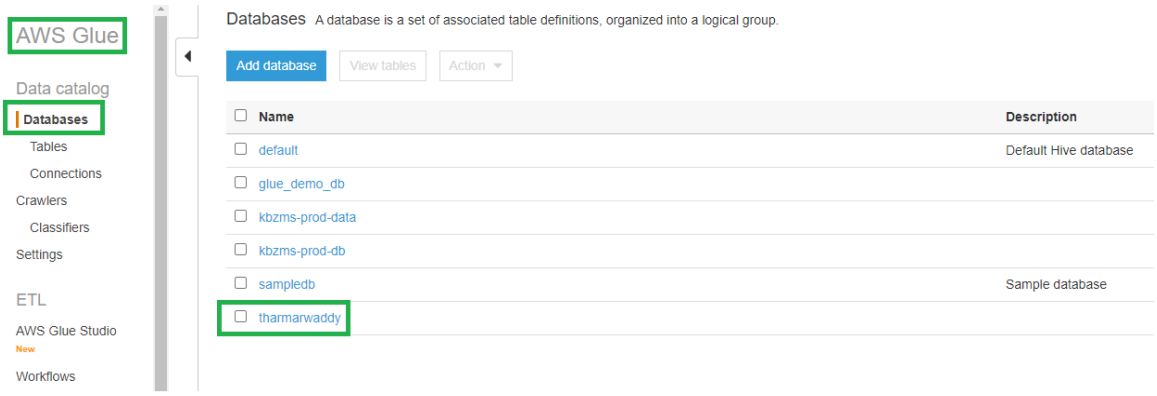
- Crawler can be run by demand or by schedule. Since this operation will use static files from Thamawaddy databases, crawler will be run by 'On Demand'.



- Output page defines the database name where tables will be added. As an option, table prefix can be added.
- For this operation, 'micp_' will be added to the tables related to motor insurance and the rest of the tables will have no prefix.

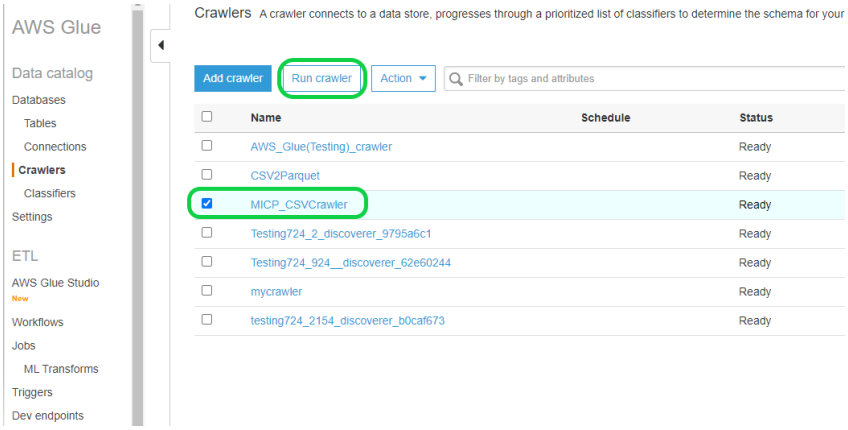


- Added database can be visible and editable from AWS Glue main page.



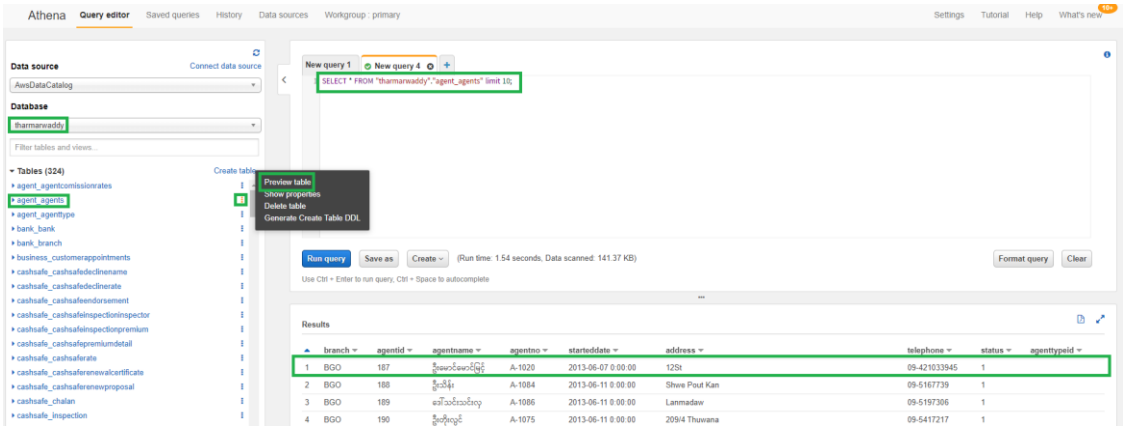
- After reviewing all the steps, crawler will be ready to run.

- To run the crawler, select a crawler on the crawlers main page and click on ‘Run Crawler’.



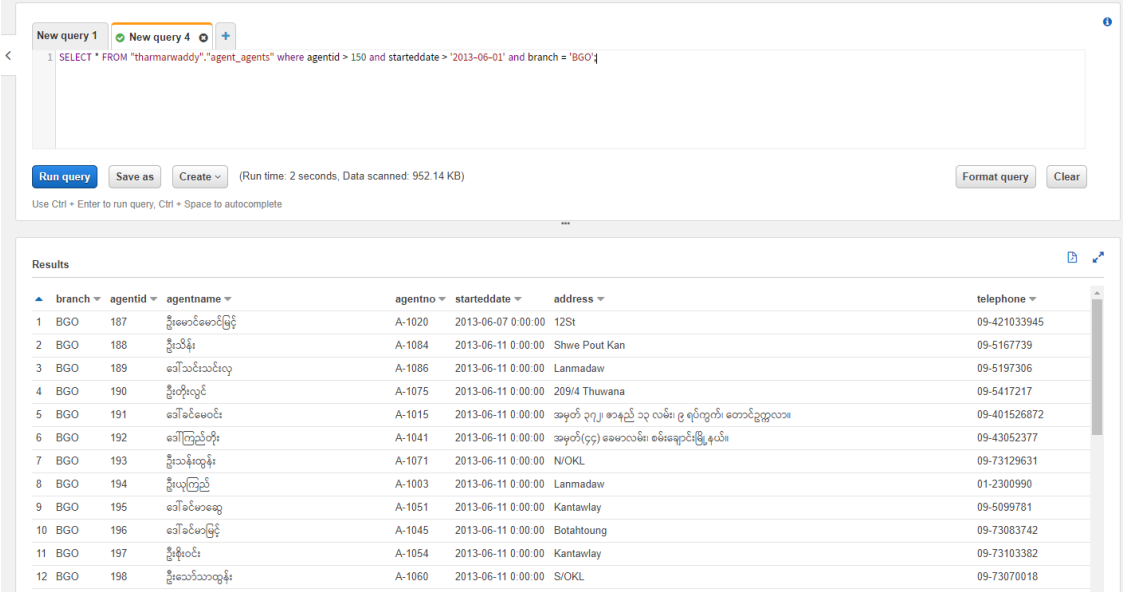
Athena

- SQL can be used for Athena in order to run query directly against object stored in S3 bucket.
- It is a serverless service offered by AWS.
- Go to Athena service under ‘Analytics’ from AWS Web Console.
- Choose the database where query will be run against a table. Click on little 3 dots on the right side of desired table and click on ‘Preview table’.
- A default query will be run for the first 10 rows.



- Check the data are correctly read by crawler or not in the Results.

- Standard query language can be used against CSV stored in S3.



New query 1 | New query 4

```
SELECT * FROM "tharmarwaddy"."agent_agents" where agentid > 150 and starteddate > '2013-06-01' and branch = 'BGO';
```

Run query | Save as | Create | (Run time: 2 seconds, Data scanned: 952.14 KB) | Format query | Clear

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	branch	agentid	agentname	agentno	starteddate	address	telephone
1	BGO	187	ဦးမောင်မောင်မြင့်	A-1020	2013-06-07 0:00:00	12St	09-421033945
2	BGO	188	ဦးသိန်း	A-1084	2013-06-11 0:00:00	Shwe Pout Kan	09-5167739
3	BGO	189	ဒေါ်သင်းသင်းလှ	A-1086	2013-06-11 0:00:00	Lanmadaw	09-5197306
4	BGO	190	ဦးတိုးလွင်	A-1075	2013-06-11 0:00:00	209/4 Thuwana	09-5417217
5	BGO	191	ဒေါ်ခင်မေဝင်း	A-1015	2013-06-11 0:00:00	အမှတ် ၃၇၂၂ ဈာနည် ၁၃ လမ်း၊ ၉ ရပ်ကွက်၊ တောင်ဥက္ကလာ။	09-401526872
6	BGO	192	ဒေါ်ကြည်တိုး	A-1041	2013-06-11 0:00:00	အမှတ်(၄၄) မောလမ်း၊ စမ်းချောင်းမြို့နယ်။	09-43052377
7	BGO	193	ဦးသန်းထွန်း	A-1071	2013-06-11 0:00:00	N/OKL	09-73129631
8	BGO	194	ဦးယုကြည်	A-1003	2013-06-11 0:00:00	Lanmadaw	01-2300990
9	BGO	195	ဒေါ်ခင်မာဆွေ	A-1051	2013-06-11 0:00:00	Kantawlay	09-5099781
10	BGO	196	ဒေါ်ခင်မာမြင့်	A-1045	2013-06-11 0:00:00	Botahtoung	09-73083742
11	BGO	197	ဦးစိုးဝင်း	A-1054	2013-06-11 0:00:00	Kantawlay	09-73103382
12	BGO	198	ဦးသော်သာထွန်း	A-1060	2013-06-11 0:00:00	S/OKL	09-73070018