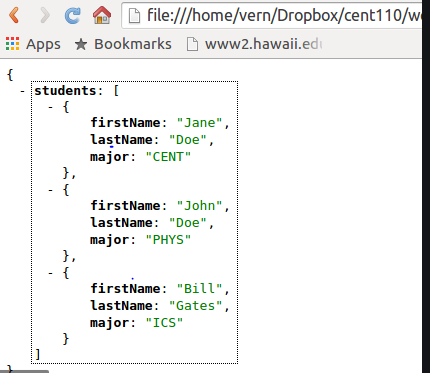
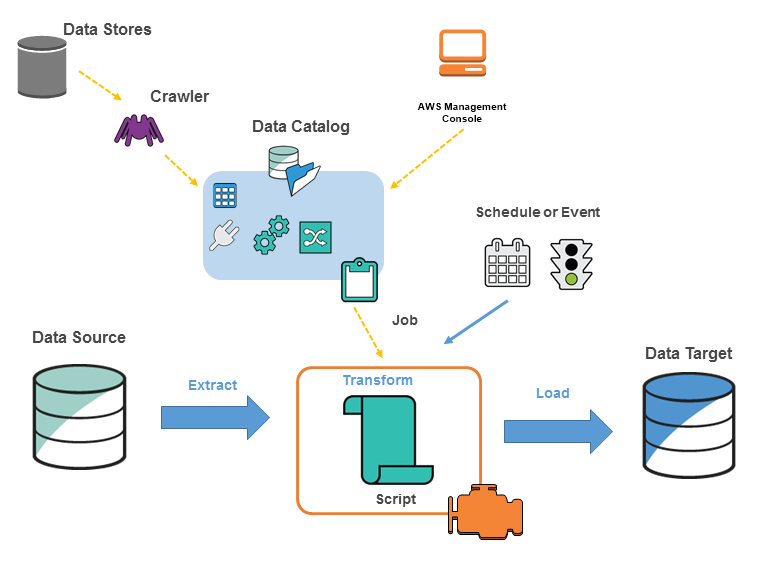
**JSON  file :- JSON (JavaScript Object Notation) is a lightweight data-interchange format. It take various forms like-An object is an unordered set of name/value pairs. An object begins with**{**left brace and ends with**}**right brace. Each name is followed by**:**colon and the name/value pairs are separated by, comma.**



Convert CSV / JSON files to Apache Parquet using AWS Glue



AWS Glue is fully managed and serverless ETL service from AWS. Now many projects are based on Parquet file format to reduce the file size and the amount of data to be scanned. Json /CSV file are easy to understand, But the users want to reduce the cost at the end of the day. Parquet is the perfect solution for this. For converting these files, AWS EMR cluster and GCP DataProc cluster are used. But these clusters are chargeable till the conversion done. So We wanted to use a solution with Zero Administrative skills. And now we are using Glue for this.

**Why Parquet?**

Parquet is a columnar file format and provides efficient storage. Better compression for columnar and encoding algorithms are also available. So we can have a better control in Performance and the Cost.

Let’s start to convert the files to Parquet.

* CSV file location: s3://searce-bigdata/etl-job/csv\_files
* Parquet file location: s3://searce-bigdata/etl-job/parquet\_files
* Sample CSV files :

<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-01.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-02.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-03.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-04.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-05.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-06.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-07.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-08.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-09.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-10.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-11.csv>  
<https://s3.amazonaws.com/nyc-tlc/trip+data/fhv_tripdata_2015-12.csv>

**Create the crawlers:**

We need to create and run the Crawlers to identify the schema of the CSV/JSON files.

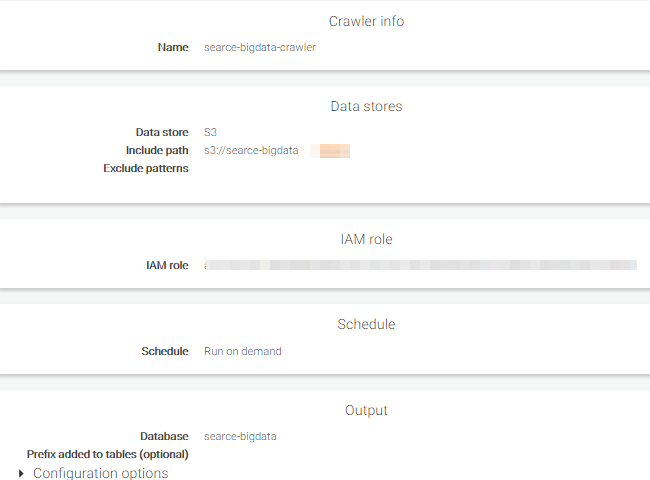
* Go to AWS Glue home page.
* From the Crawlers → add crawler.
* Give a name for you crawler.
* Data source S3 and the Include path should be your CSV files folder.
* The next step will ask to add more data source, Just click NO.
* Next one for selecting the IAM. The crawlers needs read access of the S3, but the Parquet files, needs the Write access too. So create a role along with the following policies.

.

AWSGlueServiceRole  
S3 Read/Write access for your bucket.

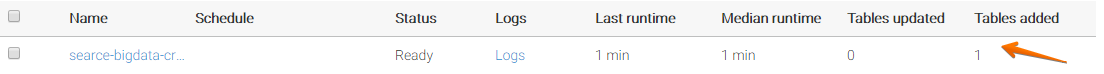
NOTE :- AWS Identity and Access Management (IAM) enables you to manage access to AWS services and resources securely. Using IAM, you can create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources

* In the next step just let the crawler as Run as On Demand.
* Then it’ll ask a database name to create a table schema for the CSV/JSON file.
* Give the path name as: s3://searce-bigdata/etl-job/csv\_files



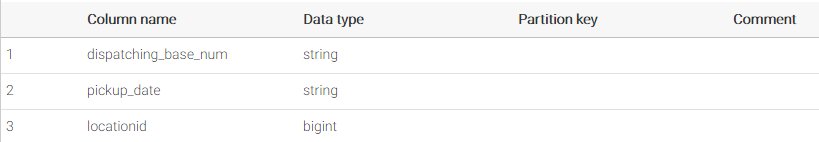
**Run the Crawler**

Once its created, it’ll ask to run. Click run and wait for few mins, then you can see that it’s created a new table with the same schema of your CSV files in the Data catalogue.



Tabled added to the data cataloge

Then it’ll create the table name as the CSV file location. [csv\_file]

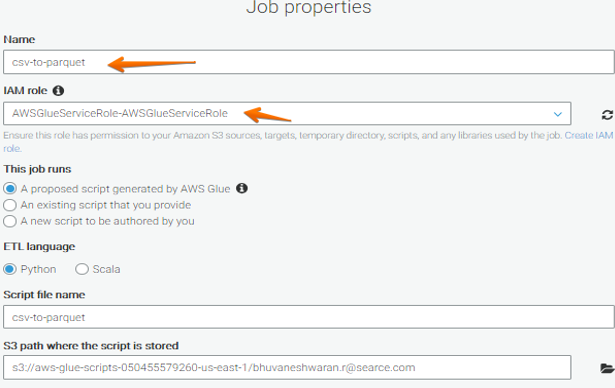


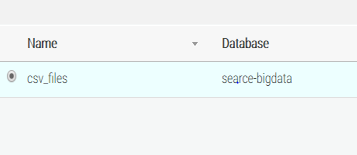
csv\_files table created in the database(CSV files and table schema is same)

Note: AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics. You can create and run an ETL job with a few clicks in the AWS Management Console

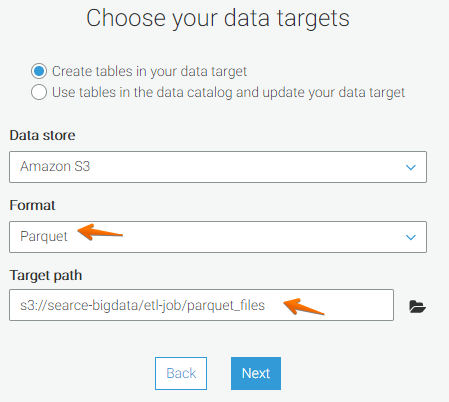
**Create Parquet conversion Job:**

* In the ETL Section, go to Jobs → add Job.
* Give a name for your job and select the IAM role(select the one which we have created in the previous step).

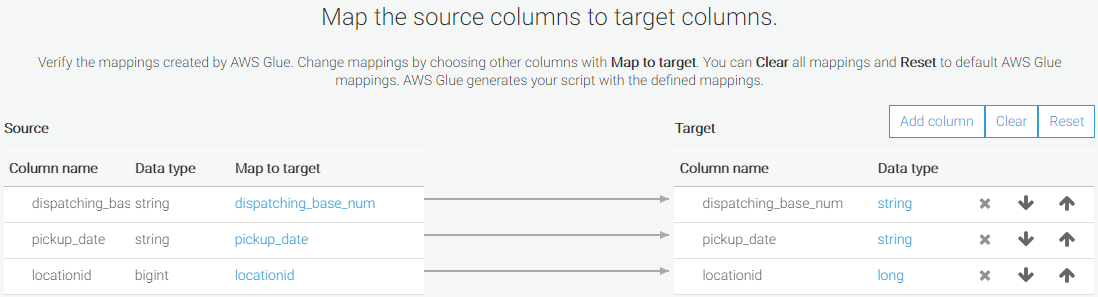


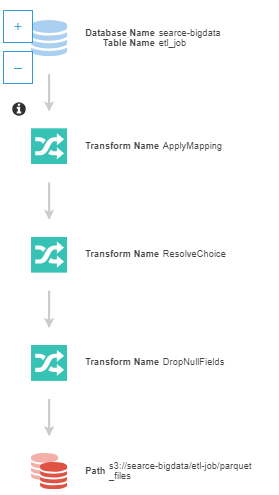
* Choose Data Source: Select the datasource which is created by the crawler.
* 
* In Choose your data target,

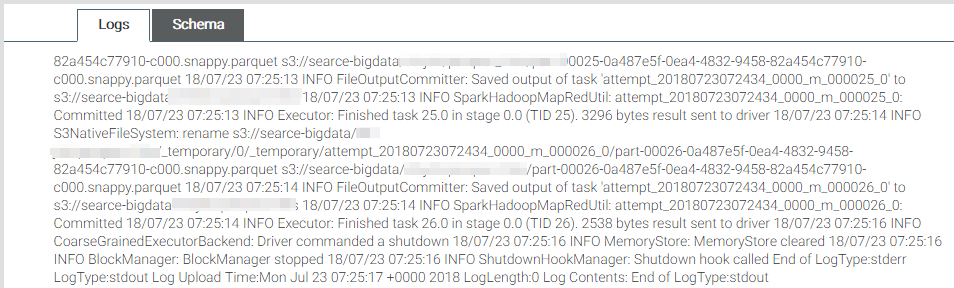
***DataStore****: S3****Format****: Parquet****TargetPath****: s3://searce-bigdata/etl-job/parquet\_files*



* The next windows is for column mapping. If you need to remap any column or remove any columns from CSV, you can achieve it from here.

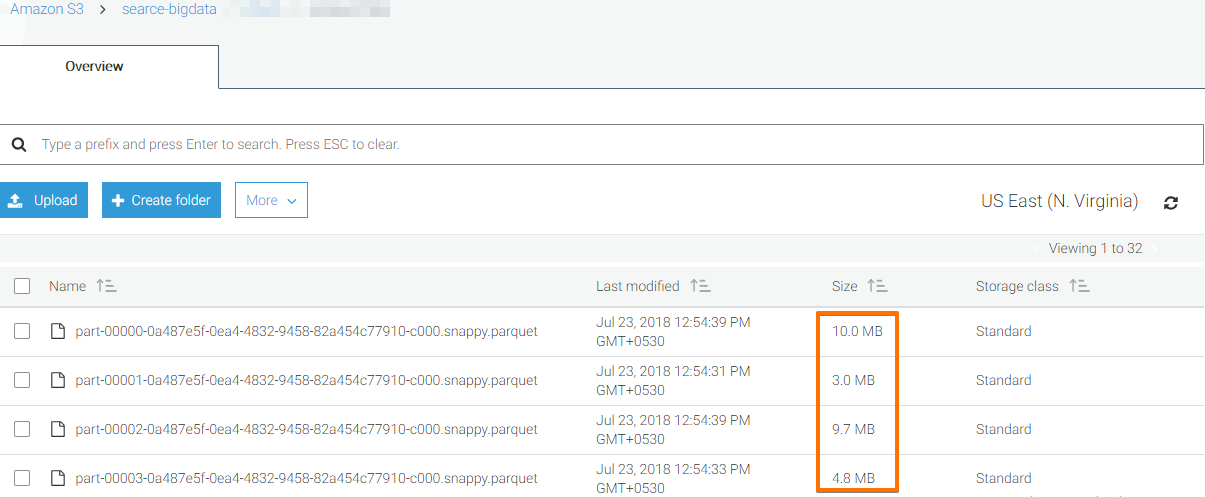


* Now the next one will show you the Diagram and source code for the job. Just click the run job button.
* 
* Wait for few mins (its based on your total amount of data) to complete the job. You can see the logs from the bottom.



**You can now check the files in S3.**

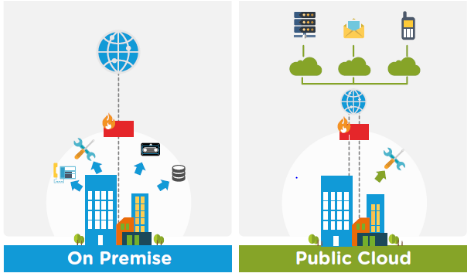
Go to *s3://searce-bigdata/etl-job/parquet\_files* and see the converted files and their size.



Thanks ….

**On-premise to public cloud**

**Cloud**-based software is hosted on the vendor's servers and accessed through a web browser. On-**premise** software is installed locally, on a company's own computers and servers. Some vendors also offer “hybrid” deployments, in which **cloud** software is hosted on an organization's private servers.



**Public clouds** are the most common way of deploying cloud computing. The cloud resources (like servers and storage) are owned and operated by a third-party cloud service provider and delivered over the Internet. **Microsoft Azure** is an example of a public cloud.

**When to use the public cloud**  
The Public cloud is most suitable for situations with these needs: Predictable computing needs, such as communication services for a specific number of users. Apps and services necessary to perform IT and business operations. Additional resource requirements to address varying peak demands. Backups are there and independent of location problems.

### On-Premise Storage

* Storage resources are procured, owned and managed by the enterprise.
* The enterprise is responsible for securing the storage resources and data.
* Storage resources remain dedicated to the company.
* The investment is considered CapEx, which is a typically a high cost.

### Cloud Storage

* Storage resources are owned and managed by a third party.
* Storage resources may be purchased on a pre-paid or pay-as-you-go basis.
* Storage resources may be shared in a multi-tenant environment.
* Software is kept up-to-date as part of an active subscription.
* IT does not have to install software updates and patches.
* The investment is considered OpEx, which is a lower monthly cost.

**Is cloud cheaper than on premise?**

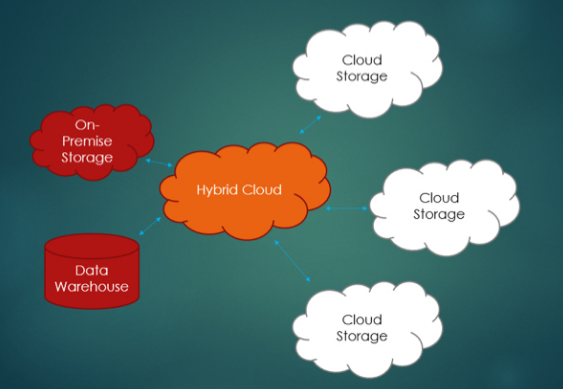
Cloud storage can be considerably cheaper than on premise at lower data levels. But as the total amount of storage increases, so does the total cost.

**Why public cloud?**

Scalability – Cloud technologies provide greater flexibility as you only pay for what you use and can easily scale to meet demand, for example adding and scaling back licenses. Lower energy costs - When you move to the cloud, you no longer have to pay to power on premise servers or to maintain their environment.

## Hybrid Cloud Storage Strategies are Popular

Most enterprises have some type of a [hybrid storage strategy](https://www.enterprisestorageforum.com/cloud-storage/hybrid-cloud-storage-guide-for-businesses.html) that involves some combination of on premise and cloud storage. Storing all data on premise may not be feasible or cost-effective given the sheer amount of data enterprises have to manage.



Thanks……………..