



**Data Lake Solution:
Twitter Data Analysis using AWS Services
Project Report**

DATA608: Developing Big Data Applications

Group 1

Members:

Raj Bhanvadia, Prashant Mittal, Soma Dipti, Somnath Bhattacharjee,
Allhad Abhyankar

Table of Contents

Chapter 1: Introduction:	5
1.1 Twitter and Its Relevance in Society:	5
1.2 AWS (Amazon Web Services):	6
1.3 Tableau:	7
1.4 Terminologies:	7
Data Lake:	7
Why Data Lake?	8
Database vs Data Warehouse vs Data Lake?	8
OLAP vs OLTP:	9
1.5 Project Scope:	10
1.6 Out of Scope:	11
1.7 Objectives:	11
1.8 Learning Objectives:	12
Chapter 2: Dataset	13
2.1 Dataset Information:	13
2.2 Strength and Weakness of Dataset:	13
2.3 Methodology:	13
2.4 Data Lake Conceptual Architecture:	15
2.5 ETL	16
Importance of ETL:	16
Data Ingestion:	17
2.6 Data Processing & Transformation:	19
Create Classifier:	19
Create Crawlers:	20
Create Amazon Glue Jobs:	21
Data Load to Redshift:	23
Create Redshift Cluster and Database:	24
Create ETL job for Redshift data load:	25

2.7 Redshift Database:.....	26
<i>Chapter 3: Data Insights and Visualization in Tableau</i>	28
3.1 Future Scope:	29
3.2 Summary:	30
3.3 Limitations:	30
<i>References:</i>	31

List of Figures:

Figure 1 Data Lake Information	8
Figure 2 Difference between DB, DW, and DL	9
Figure 3 OLAP vs OLTP	10
Figure 4 Terminologies	14
Figure 5 Conceptual Architecture of AWS Data Lake	15
Figure 6 AWS Services	16
Figure 7 Amazon Glue Service	17
Figure 8 Data Ingestion	17
Figure 9 S3 Bucket	19
Figure 10 Data Process and Transform	19
Figure 11 Classifier	20
Figure 12 Crawlers	20
Figure 13 Schema analysis	20
Figure 14 Amazon Glue Jobs	21
Figure 15 First job configuration	22
Figure 16 Successful execution	22
Figure 17 Clean Zone	23
Figure 18 Glue crawler data	24
Figure 19 Redshift cluster	25
Figure 20 Redshift database connection	26
Figure 21 Redshift table	27
Figure 22 Tableau Integration	28
Figure 23 Dashboard for Sentiment Analysis on Twitter Covid Tweets	29
Figure 24 Future Scope	30

Chapter 1: Introduction:

1.1 Twitter and Its Relevance in Society:

Twitter is a social media platform and a social networking service that allows users to post texts, images, or videos. These posts are called tweets. Users can create new posts, forward posts (also called re-tweets), or direct messages to other users. For our project **we shall be considering text tweets, either original or retweets**).

Twitter was created by Jack Dorsey of New York University with the idea of using SMS (Short Messaging Service) to communicate with small groups. From SMS service to a social media platform and from mere 20000 tweets per day in 2007 to 200 million tweets per day in 2022(Sayce, 2010), Twitter started to grow. Today, users post tweets about their own life, such as activities they do or events that they attend as well as things that make an impact on our society in general.

Twitter users range from individuals to political leaders, musicians, movie stars, sports celebrities, organizations, and much more. On the controversial side, Twitter has been used to promote hate speech, extreme ideologies, fake news, and conspiracy theories. However, on the positive side, tweets are posts about well-being (status about our friends and family), Emergency use like helping in epidemics, crisis, or natural disasters, education, social help like food banks, environmental concerns, equality, and other social responses. It can be used by companies for the promotion of their products.

A tweet can be linked to another tweet about the same topic. This is done using the keyword ‘#’ in front of any unbroken word. This is known as a hashtag and the action is called hash-tagging. This makes all the tweets about any topic searchable. Moreover, users (individuals, companies, or organizations) can mention other users in their posts (tweets) using the ‘@’ (at the rate) sign. This links the mentioned user to the tweet. (Dorney, n.d.)

With tweets posted **at a rate of generally 200 million per day**, a lot of data is generated. This data tells us about more than what is posted. For example, data about the covid topic hash-tagged ‘#COVID’ can tell us about the pandemic, the government can inform people about the situation, individuals can post about how they feel about their situation and social organizations can promote their activities to connect people in a similar situation.

And with this posted data and information contained in it, an analysis of the trends can be done. Again, for the example above ('#COVID') and tweets made by any government, we can analyze if the government was aware of the situation, perform an analysis of the timeline of actions being taken, and search for any shortcomings. Similarly, tweets from individuals can be analyzed using Natural Language Processing (NLP) to understand their emotional condition, their awareness of the situation, and if they require any help.

The analysis is complex as there are two major issues. Firstly, the data is large (we already know that there are about 200 million tweets per day) and secondly, processes required for analysis require large computations. Meaning that there can be many processes like counting tweets to understand trends and the processes may be complex like actual understanding of tweets (if they are positive or negative language) using NLP. **Thus arises a need of using Big Data Applications for this process.**

1.2 AWS (Amazon Web Services):

Amazon Web Services is a provider of on-demand cloud computing platforms and APIs for individuals, businesses, and governments. The term **on-demand cloud computing system** means that computer systems resources like computing power and storage (cloud storage) are provided to the users by the provider as a service. The on-demand term here means that the provider (amazon) uses a **pay-as-you-go model** for its consumers.

Why is AWS used by us for Twitter data?

As we have discussed earlier regarding the amount of data collected by Twitter, a web service with the capability to store and process this data. **Creating a technical solution** to store and process structured and unstructured data is the primary objective of our project. Amazon Web Services is one of the web services that has infrastructure big enough to handle such large quantities of data. Other similar services are Microsoft Azure, Google Cloud Platform, Alibaba Cloud, Oracle Cloud Infrastructure, IBM Cloud, and others (Zelleke, 2021). Amazon provides cloud computing services in more than 20 countries and has about 100+ locations (Zhang, 2022). The storage and computing capability and its worldwide outreach make **AWS the service of choice for our project.**

1. Some other advantages of using AWS were: The simplicity of use.

2. Speed.
3. Variety of services.

Although AWS provides a wide range of storage and computational solutions, we used the following,

1. **Amazon S3:** Simple Storage Services (S3) is a scalable, performance-oriented, and secure data storage service by Amazon (Amazon Web Services, n.d.). All the data collected from the Twitter website needs to be stored in cloud storage and cleaned. As the tweets grow the storage capacity increases, this is where the scalability feature of S3 plays a major role.
2. **Amazon Glue:** Extract, Transform, and Load service in AWS. The raw data from Twitter is pulled, cleaned, and transformed into clean data. This is done with Glue. Currently, glue only supports Python and Scala. Our project uses Python for this purpose.
3. **Amazon Redshift:** Data that has been structured, cleaned, and transformed can now be stored in a data warehouse. The Data Warehouse service in AWS is called Redshift. This is where our atomic data will be stored. Redshift can be integrated with Tableau for visualization.

1.3 Tableau:

Tableau is a data visualization tool used by many industry experts. Users can create charts, maps, dashboards, and stories for data analysis which can be a further help in making business decisions (Biswal, 2023). Tableau is one of the most popular data visualization tools in the data science industry (Andy Patrizio, 2021).

Its ability to integrate with all the major advanced databases, like Teradata, SAP, MySQL, Amazon AWS, and Hadoop makes it ideal for our project.

1.4 Terminologies:

Data Lake:

A data lake is a centralized repository of structured, semi-structured, and unstructured data. It requires data pipelines to carry copies of data from a source to a central repository where data can be processed, and analysts or data scientists derive insights from the processed data. (AWS, 2019)

Why Data Lake?

Twitter data may contain various structured information like username, time of post, tweet, and other such data, as well as unstructured data like a video, picture, or voice recording. Depending on the needs of customers these data may be pulled for analysis. Thus, a data lake is required.

Amazon S3 in our case provides the data lake solution to us.

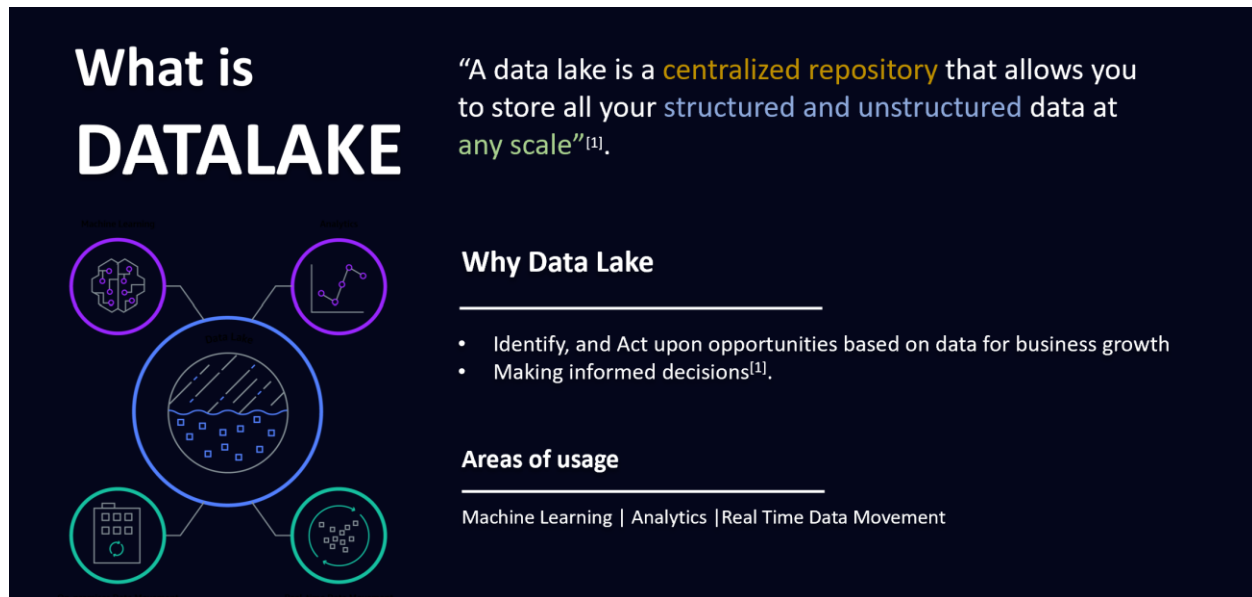


Figure 1 Data Lake Information

Amazon S3 service has been used as the data repository which stores any kind of object.

Database vs Data Warehouse vs Data Lake?

A database is a house for structured data. It stores relational, transactional data for the day-to-day digital store of information. It can only handle structured data and can be used by anyone. A typical example of a database is debit card statements, where simple reporting of the debit transaction is done and can be accessed by any person authorized by the bank.

Data Warehouse is the digital storage of a huge amount of information (structured) intended for query and analysis by Business Analysts. It involves a process of transforming data into information. It provides meaningful business insights. This too can handle structured data and is typically used by IT/Business Users. In banking analogy, this is information that any bank

employee (business user) can see like debit/credit statements, user information, bank locker information, etc.

Data Lake is a data repository. It can hold any type of data of any size and any format. Typically used by data scientists, these stores both structured and unstructured data. A core banking application can be a data lake. This will store all data that is in the data warehouse and anything that is concerned with banking. This may include phone calls to banking call centers, video recording of their security cameras, and any other information.



Figure 2 Difference between DB, DW, and DL

Data Lake Icon Reference: <https://aws.amazon.com/solutionspace/data-lake-foundation-with-aws-services>

OLAP vs OLTP:

Online Analytical Processing (OLAP) and Online Transaction Processing (OLTP) both are transaction processing systems used for processing huge amounts of data.

OLTP collects, stores, and processes data from real-time transactions.

OLAP is for analytics. It works on historical data to find data insights.

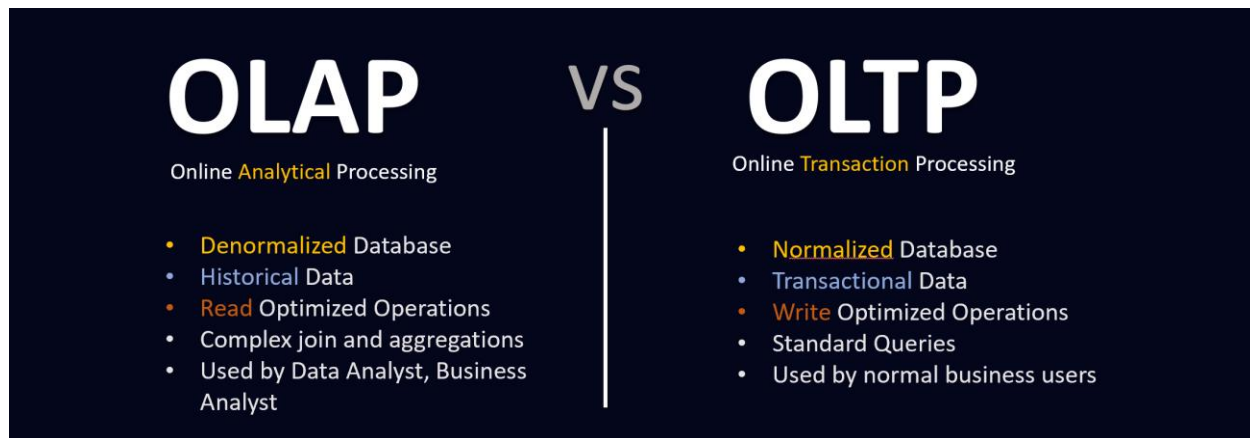


Figure 3 OLAP vs OLTP

1.5 Project Scope:

The scope of the project is to create a Data Lake technical solution using AWS cloud services.

1. Create a Data Lake Technical Solution using AWS Cloud Services.

The objective of this project is to create a process on AWS Cloud to collect data from Twitter using an API. This data will be stored, cleaned, processed, and finally displayed on Tableau.

2. Use offline Twitter message data collected from Kaggle

Currently, we will be using offline data from Kaggle. The data is scraped by Gabriel Preda (Preda, 2020) starting in February 2020. The data was collected using the hashtag #covid19. The collection script used is stored in “GitHub - gabrielpreda/covid-19-tweets: Covid-19 tweets”. And the data was collected using Python.

3. Setup Extract, Transform, and Load process using AWS services

The ETL or Extract, Transform, and Load process is done in Amazon Glue. In Amazon S3, data is stored in two folders (Raw Zone and Clean Zone) and finally in reporting stage data is stored in Redshift (Data Warehouse). The objective of this step is to pull, transform and push data from Raw Zone to Clean Zone and once more, to pull data from Clean Zone, process and transform and push data to Redshift so that it can be reported. This process is done using Amazon Glue.

4. Use Tableau for visualization

Finally, in the reporting stage, data will be pulled from Redshift to Tableau and will be used to create the visualization.

1.6 Out of Scope:

1. Deep analysis insights and make a decision from used data:

Although data is pulled from the Kaggle website and is related to COVID-19, the overall objective is to create a generalized solution where users can create their own analysis. Hence analysis or decision-making is not done using this data.

2. Deep dive into data wrangling:

Some cleaning of data is done in the data refinery (Raw Zone to Clean Zone) and from Clean Zone to Atomic Zone (Redshift). This is done to transform data as supplied by Twitter API (or Kaggle Data by Preda) to a ready-to-use stage. No further cleaning, restructuring, enriching (replacing Nan, None, Null values) or any other wrangling process is done.

1.7 Objectives:

The aim is to create a Technical Solution in AWS Cloud which will store and process a large amount of structured or unstructured data. The solution brings many benefits in data analytics and machine learning. It helps process a large amount of data and apply statistical methodologies to process it.

The 'Extract, Transform, and Load' process prepares the data through a technical data refinery and stores it in a data warehouse in a ready-to-use format. The solution creation using AWS is the major focus and secondary purpose to get the insights using tableau visualization.

The proposed solution has 4 steps:

- **Ingest Data:** Upload Twitter data snapshot extracted from Kaggle.
- **Process or Transform data:** Perform some simple data wrangling steps on the raw data format.
- **Store Data:** Store clean/processed data in a data warehouse and NO-SQL database.
- **Data Analytics in Tableau:** Create visualization in Tableau from the Data Warehouse.

1.8 Learning Objectives:

- Create a process to ingest raw data in Amazon S3. Manually upload offline Twitter messages
- Create a data processing ETL job using Amazon Glue. Which will process data and move it to the Clean Zone of the S3 bucket
- Create an ETL job using Amazon Glue to move data from the clean zone (S3 Bucket) to Amazon Redshift.
- Configure database mapping with clean data schema.
- Provision a Redshift cluster.
- Create a table in Redshift to store data.
- Create data mapping between table schema and processed data.
- Integrate Tableau as a visualization tool with the Amazon Redshift cluster.
- Pull data from the Data Warehouse table into Tableau.
- To test our project we will,
 - Creating a visualization of Covid-tweets data in Tableau.
 - Finding insights from sentiment analysis.
- Connection troubleshooting for Cloud security group.

Chapter 2: Dataset

2.1 Dataset Information:

The COVID-19 Tweets dataset is a collection of tweets related to the COVID-19 pandemic that was collected using the Twitter API. The dataset was **created** by **Gabriel Preda** and is **hosted** on **Kaggle**.

URL: <https://www.kaggle.com/datasets/gpreda/covid19-tweets>

The dataset contains approximately 1.7 million tweets that were posted between January 2020 and July 2020. Each tweet in the dataset includes various information such as the user ID, the date and time of the tweet, the text of the tweet, the number of retweets, and the number of likes. The dataset is licensed under the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International (CC BY-NC-SA 4.0) license.

2.2 Strength and Weakness of Dataset:

The dataset's strengths include its large size and coverage of a wide range of dates, which enables tracking changes in public opinion or the spread of misinformation over time. The inclusion of metadata, such as the number of retweets and likes, also allows for the measurement of the popularity of specific topics or sentiments.

user_name	user_location	user_description	user_created	user_follow	user_friend	user_favou	user_verified	date	text	hashtags	source	is_retweet
Time4fisticuffs	Pewee Valley, KY	#Christian #Catholic	2/28/2009 18:57	9275	9525	7254	FALSE	7/25/2020 12:27	@diane344	['COVID19']	Twitter for	FALSE
ethel mertz	Stuck in the Mid	#Browns #Indians #	3/7/2019 1:45	197	987	1488	FALSE	7/25/2020 12:27	@brookbar	['COVID19']	Twitter for	FALSE
DIPR-J&K	Jammu and Kash	ðŸŒ™-ðŸŒ™ Official Twitte	2/12/2017 6:45	101009	168	101	FALSE	7/25/2020 12:27	25 July :	['CoronaVir	Twitter for	FALSE
Franz Schubert		ðŸŒ™ #ðŸŒ™ðŸŒ™ðŸŒ™	3/19/2018 16:29	1180	1071	1287	FALSE	7/25/2020 12:27	#coronavir	['coronavir	Twitter We	FALSE
hr bartender	Gainesville, FL	Workplace tips and	8/12/2008 18:19	79956	54810	3801	FALSE	7/25/2020 12:27	How #COVI	['COVID19',	Buffer	FALSE
Member of Christ	location at link	I just as the body is c	8/17/2014 4:53	55201	34239	29802	FALSE	7/25/2020 12:26	POPE AS	['Hurricane	Twitter for	FALSE
SEXXLYPPS	Hotel living - var	My ink "My	3/25/2010 21:16	0	8	32	FALSE	7/25/2020 12:26	ðŸŒ™ðŸŒ™@P	['COVID19']	Twitter We	FALSE
Africa Youth Advi	Africa	Official account of t	5/13/2019 6:27	830	254	3692	FALSE	7/25/2020 12:26	Let's all	['COVID19']	Twitter We	FALSE
DailyaddaaNews	New Delhi	Breaking news alert	10/22/2016 9:18	546	29	88	FALSE	7/25/2020 12:26	Rajasthan Government	Twitter We	FALSE	FALSE
Dimapur 24/7.	Nagaland, India	strive to promote	11/11/2019 12:02	274	32	378	FALSE	7/25/2020 12:26	Nagaland	['Covid19',	Twitter for	FALSE

However, the dataset's weaknesses include its limited time frame, which may not accurately reflect the current state of the pandemic or public opinion, and its exclusion of deleted or private tweets, which could potentially impact the accuracy of the analysis. Additionally, the lack of demographic information about users who posted the tweets limits the ability to analyze differences in opinions or behaviors across different groups.

2.3 Methodology:

- **Amazon Glue:** Extract, Transform, and Load service in AWS.

- **Amazon Athena:** Data query service in AWS for executing queries from Amazon S3 bucket.
- **Amazon Dynamo DB:** NoSQL database service in AWS.
- **Amazon Redshift:** Data Warehouse service in AWS.
- **Tableau:** A visualization tool to give insight into Twitter data for COVID.

Analyzing COVID-related data, we can use Amazon Glue as an Extract, Transform, and Load (ETL) service to prepare and process data from various sources, including social media platforms such as Twitter. The processed data can then be stored in Amazon Redshift, a data warehouse service, depending on the specific use case. Finally, Tableau can be used as a visualization tool to gain insights into COVID-related data and present the findings in a user-friendly and interactive way.

Simplified definitions of AWS services used in this project are as follows.

Amazon S3:

AWS provides a service used for object storage.

Amazon Glue:

AWS provided service used for ETL (Extract, Transform, Load).

Amazon Redshift:

AWS provided Data Warehouse service.

Tableau:

Visualization tool



Figure 4 Terminologies

2.4 Data Lake Conceptual Architecture:

There are mainly three phases to make the data journey through different data zones.

Raw Zone: This is the landing zone of raw data.

Clean Zone: This is clean and processed data.

Atomic Zone: This is the stage of data that is ready for use in any integrated system or directly for analysis.

Data Ingestion: Source data from various sources and gather in a raw zone. In our project, it is offline Twitter data uploaded manually to Amazon S3.

Data Processing & Transformation: Clean data and normalize data structure to the required format.

Reporting: Fetch data from the atomic zone and use visualization tools (Tableau in our case) to create data insights.

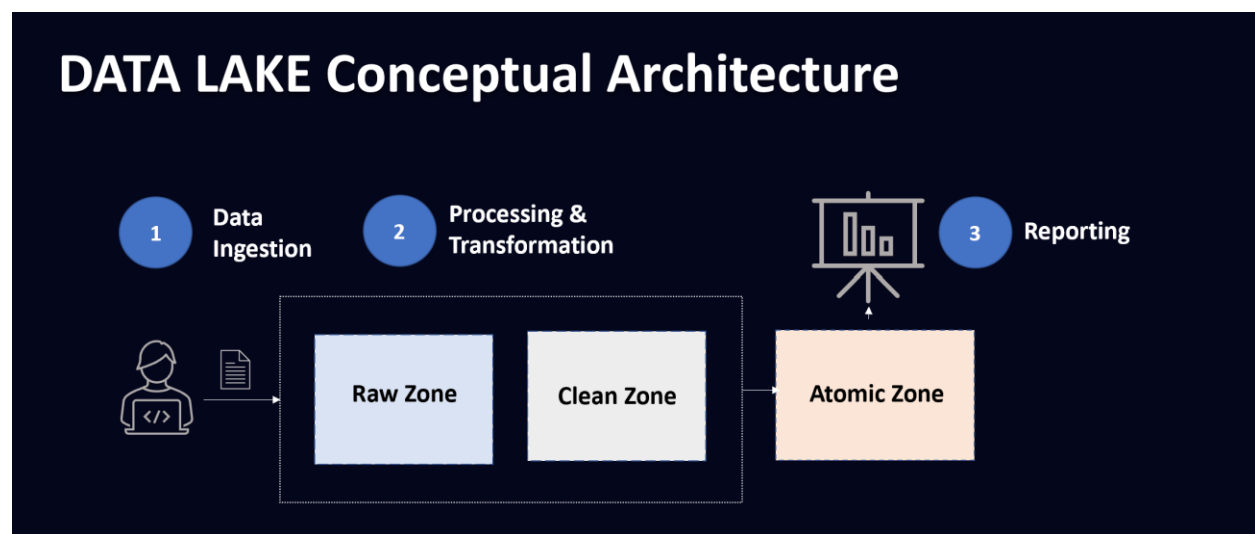


Figure 5 Conceptual Architecture of AWS Data Lake

Below is the flow through the actual AWS services used to create the data lake solution. As the below diagram shows. The user manually uploads a CSV file to S3 in the 'raw zone'. The Glue job does the data transformation and takes the data from the raw zone and loads it to the clean zone in S3. The next Glue job loads the data from the clean zone to Data Warehouse. Tableau has been integrated with Redshift and fetches data from a data warehouse.

DATA LAKE with AWS SERVICES

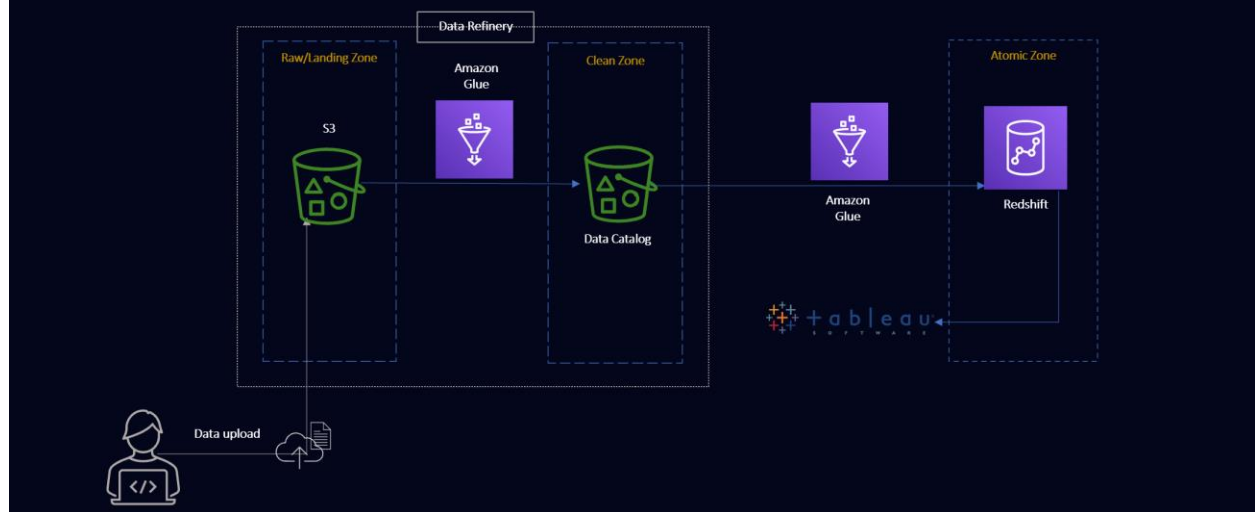


Figure 6 AWS Services

2.5 ETL

‘ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system’. (IBM, n.d.)

Importance of ETL:

- Extract data from legacy systems.
- Cleanse the data to improve data quality and establish consistency.
- Load data into a target database. (IBM, n.d.)

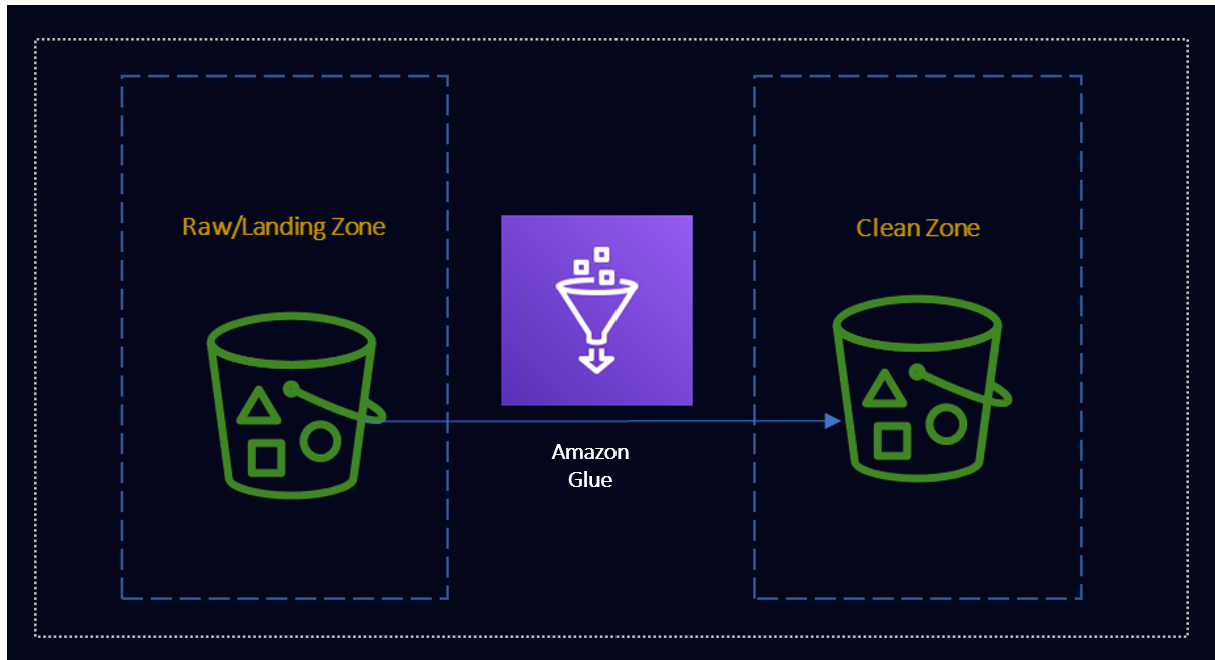


Figure 7 Amazon Glue Service

Amazon Glue is the service used for creating ETL processes.

Data Ingestion:

Data Ingestion is the first phase of a data lake. In our case, we have manual data upload by a user to the AWS S3 bucket in 'raw-zone' as shown in highlighted (dotted box) below figure.

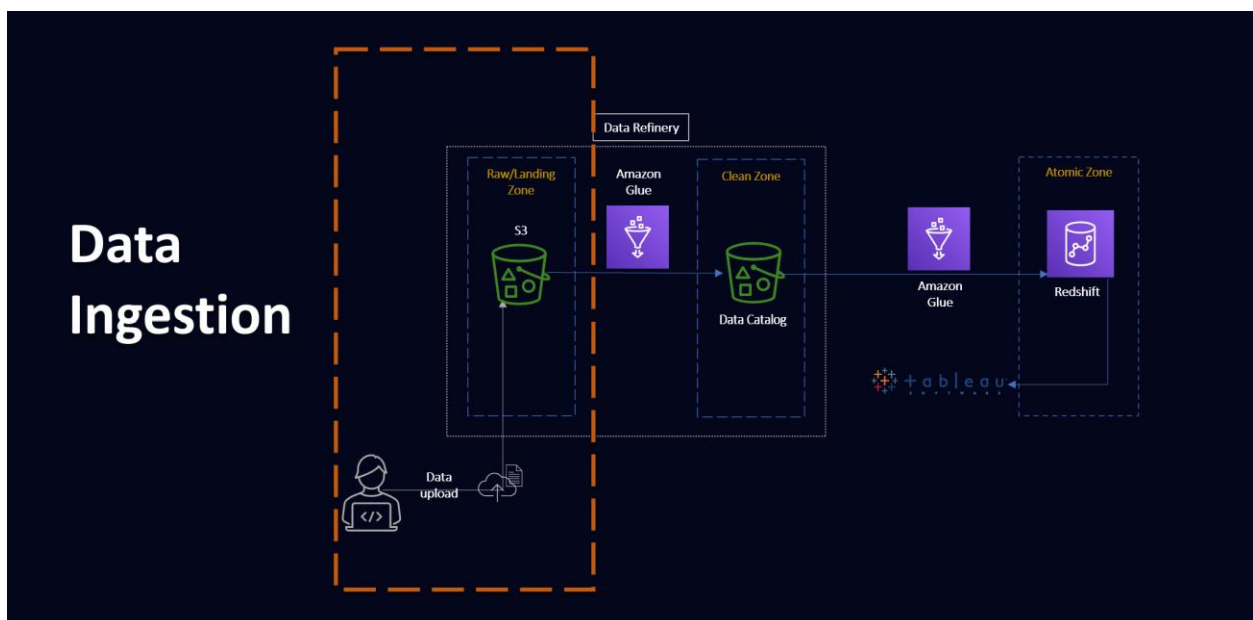


Figure 8 Data Ingestion

Ingest data to S3 raw zone.

Manually upload offline Twitter data (twitter_tweets_data.csv, 55.9 MB) to the S3 bucket 'ucal-datalake' in the 'raw-zone/streaming-data' folder.

The image shows two screenshots from the Amazon S3 console. The top screenshot displays the 'Buckets' page for the 'ucal-datalake' bucket. It includes an 'Account snapshot' section with metrics: Total storage (1.1 GB), Object count (657), and Average object size (1.7 MB). Below this, a table lists the bucket details: Name (ucal-datalake), AWS Region (US East (N. Virginia) us-east-1), Access (Bucket and objects not public), and Creation date (March 1, 2023, 19:35:21 (UTC-05:00)). The bottom screenshot shows the 'ucal-datalake' bucket's 'Objects' tab. It lists two folders: 'clean-zone/' and 'raw-zone/'. The 'raw-zone/' folder is highlighted, indicating it is the selected view.

Name	AWS Region	Access	Creation date
ucal-datalake	US East (N. Virginia) us-east-1	Bucket and objects not public	March 1, 2023, 19:35:21 (UTC-05:00)

Name	Type	Last modified	Size	Storage class
clean-zone/	Folder	-	-	-
raw-zone/	Folder	-	-	-

S3 bucket 'ucal-datalake' contains 'raw-zone' and 'clean-zone' as shown in the above figure.

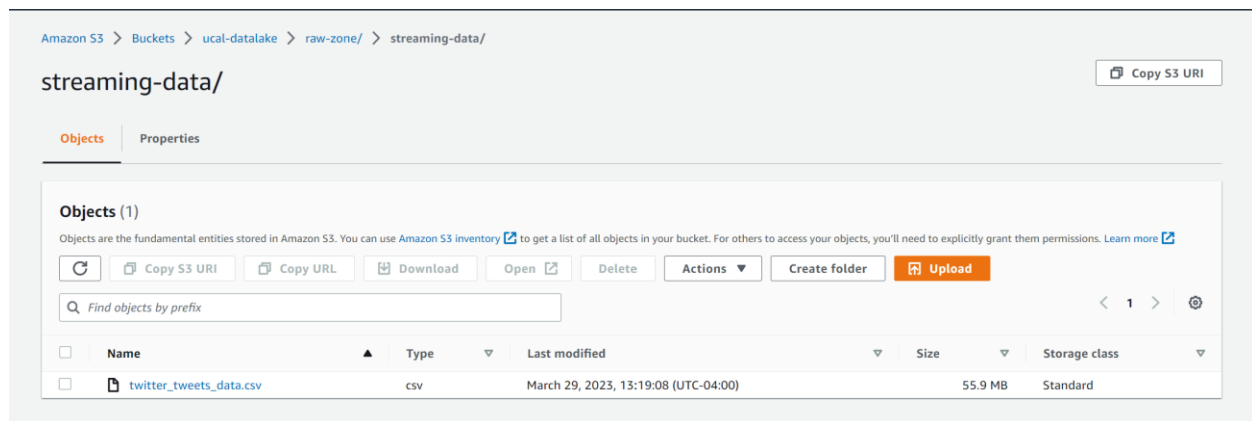


Figure 9 S3 Bucket

2.6 Data Processing & Transformation:

After data ingestion, the next phase is the data processing and transformation phase. A glue job is created to take data from the 'raw-zone' and perform data cleaning, sentiment analysis, and load into the clean-zone as highlighted in the below diagram.



Figure 10 Data Process and Transform

Create Classifier:

Classifier is the identifier that specifies which kind of input file is expected. In our case, the file type is a 'CSV' file. That's why we created a crawler as below figure.

Classifiers

Classifiers are triggered during a crawl task. A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Classifiers (3) [Info](#)

View and manage all available classifiers.

Last updated (UTC) April 5, 2023 at 23:54:19 [Refresh](#) [Edit](#) [Delete](#) [Add classifier](#)

Q X 1 match [<](#) 1 [>](#) [Settings](#)

<input type="checkbox"/>	Name	Type	Classification	Last updated (UTC)
<input type="checkbox"/>	twitter-message-classifier-csv	CSV	-	March 14, 2023 at 23:30:52

Figure 11 Classifier

Create Crawlers:

The purpose of Crawler is to automatically analyze the schema of the source data and create a glue table in the glue database.

Crawlers

A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (9) [Info](#)

View and manage all available crawlers.

Last updated (UTC) March 31, 2023 at 18:38:05 [Refresh](#) [Action](#) [Run](#) [Create crawler](#)

Q X 2 matches [<](#) 1 [>](#) [Settings](#)

[ucal2](#) X [Clear filters](#)

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from last run
<input type="checkbox"/>	ucal2-crawler-redshift	Ready		Succeeded	March 26, 2023 at 21:02:00	View log	1 created
<input type="checkbox"/>	ucal2-crawler-s3	Ready		Succeeded	March 26, 2023 at 20:55:38	View log	1 created

Figure 12 Crawlers

The glue database as shown in the figure below is created to hold the tables which are created by the crawler after its automated schema analysis.

ucal2-crawler-s3

Last updated (UTC) March 31, 2023 at 18:38:25 [Refresh](#) [Run crawler](#) [Edit](#) [Delete](#)

Crawler properties

Name ucal2-crawler-s3	IAM role ucal-glue-crawler-role	Database ucal2-glue-db-s3	State READY
Description -	Security configuration -	Lake Formation configuration -	Table prefix -
Maximum table threshold -			

[Advanced settings](#)

[Crawler runs](#) [Schedule](#) [Data sources](#) [Classifiers](#) [Tags](#)

Data sources (1) [Info](#)

The list of data sources to be scanned by the crawler.

[Refresh](#) [Edit](#) [Remove](#) [Add a data source](#)

Type	Data source	Parameters
<input type="radio"/> S3	s3://ucal-datalake/raw-zone/streaming-data/	Recrawl all

Figure 13 Schema analysis

Create Amazon Glue Jobs:

After creating Classifier, Crawlers, and Glue database we have created two separate jobs for the ETL process. ‘**ucal2-message-raw-to-clean-job**’ is for data transferring from raw to clean-zone. Secondly, ‘**ucal2-message-rs-loading-job**’ is for loading clean data to the Redshift data warehouse.

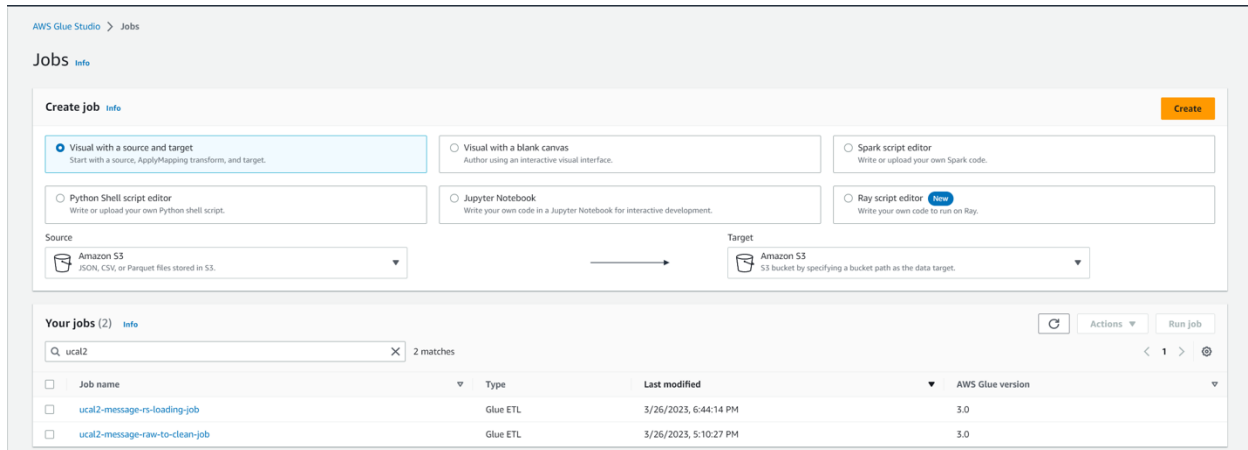


Figure 14 Amazon Glue Jobs

The below figure shows the configuration of the first job configuration. Which takes the data from the raw zone, processes and loads it to the clean zone. Here our source was Raw Zone and the target was the Clean Zone path. In the visual diagram below, we see the stages and on the right side the data catalog table which was created by ‘crawler’.

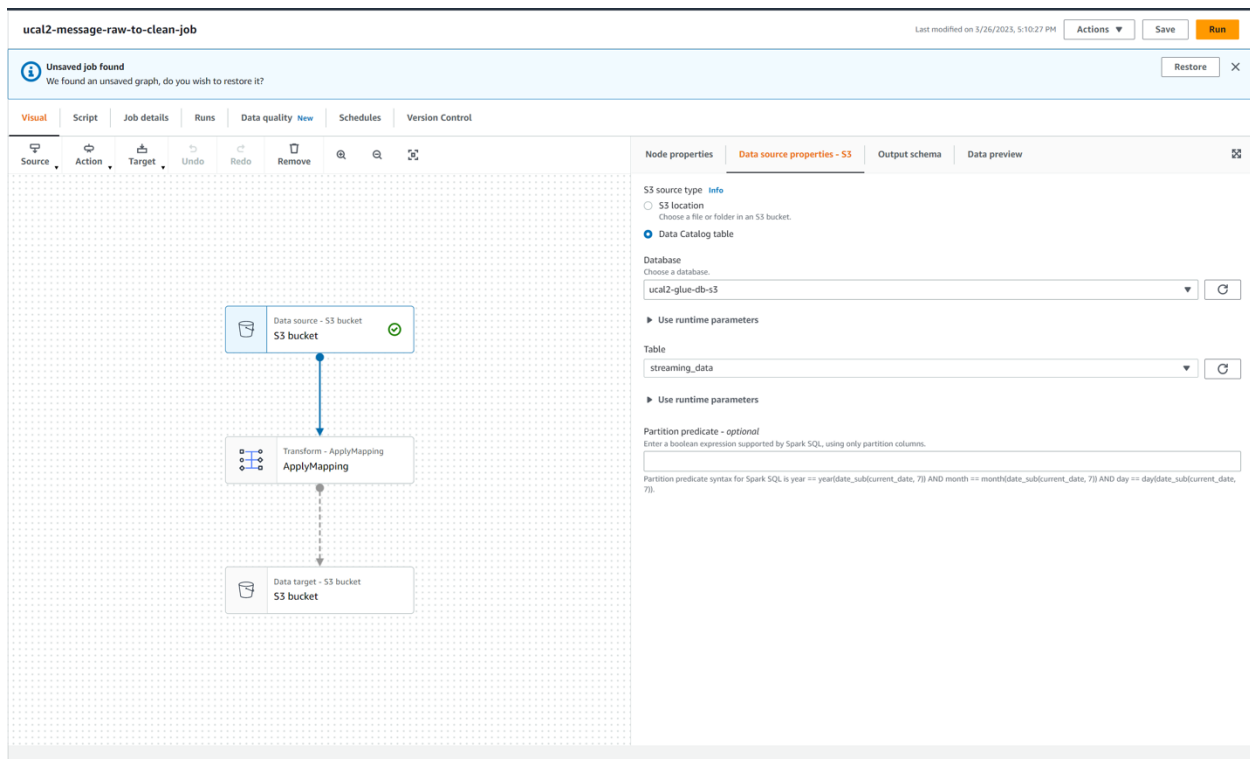


Figure 15 First job configuration

The below figure shows the successful execution of the first ETL job.

The screenshot shows the 'Runs' tab for the job 'ucal2-message-raw-to-clean-job'. It displays a table of recent job runs. The first run is highlighted, showing details such as 'Job name', 'Id', 'Run status' (Succeeded), 'Start time', 'End time', 'Execution time', 'Timeout', 'Execution class', 'Log group name', 'Glue version', 'Start-up time', 'Trigger name', 'Security configuration', 'Worker type', and 'Performance and debugging recommendations'.

Job name	Id	Run status	Glue version
ucal2-message-raw-to-clean-job	j_r_ddc0cce4326ab5236870f4812f5d9964455669ed4833c673ac6e695c6549c97	Succeeded	3.0

Figure 16 Successful execution

After the successful execution of the above ETL job, the output in Clean Zone looks like the below figure. We had one CSV for 56MB but after transformation, it created 20 files of 2.8MB.

Amazon S3 > Buckets > ucal-datalake > clean-zone/ > streaming-data/

streaming-data/ Copy S3 URI

Objects Properties

Objects (20)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Refresh
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	run-1680147582812-part-r-00000	-	March 29, 2023, 23:39:57 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00001	-	March 29, 2023, 23:39:57 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00002	-	March 29, 2023, 23:39:57 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00003	-	March 29, 2023, 23:39:57 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00004	-	March 29, 2023, 23:39:58 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00005	-	March 29, 2023, 23:39:58 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00006	-	March 29, 2023, 23:39:58 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00007	-	March 29, 2023, 23:39:58 (UTC-04:00)	2.9 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00008	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00009	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00010	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00011	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00012	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00013	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00014	-	March 29, 2023, 23:39:59 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00015	-	March 29, 2023, 23:40:00 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00016	-	March 29, 2023, 23:40:00 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00017	-	March 29, 2023, 23:40:00 (UTC-04:00)	2.8 MB	Standard
<input type="checkbox"/>	run-1680147582812-part-r-00018	-	March 29, 2023, 23:40:00 (UTC-04:00)	2.8 MB	Standard

Figure 17 Clean Zone

Data Load to Redshift:

In this part, we need an ETL job that will take data from the clean zone and load it into Redshift. The schema mapping is required here.

We created a crawler for analyzing the data schema in Clean Zone as it is shown in the below figure.

AWS Glue > Crawlers > ucal2-crawler-redshift

ucal2-crawler-redshift Last updated (UTC) March 31, 2023 at 18:39:10 Refresh Run crawler Edit Delete

Crawler properties

Name ucal2-crawler-redshift	IAM role ucal-glue-crawler-role	Database ucal2-gluedb-redshift	State READY
Description -	Security configuration -	Table prefix -	

► Advanced settings

Crawler runs | Schedule | **Data sources** | Classifiers | Tags

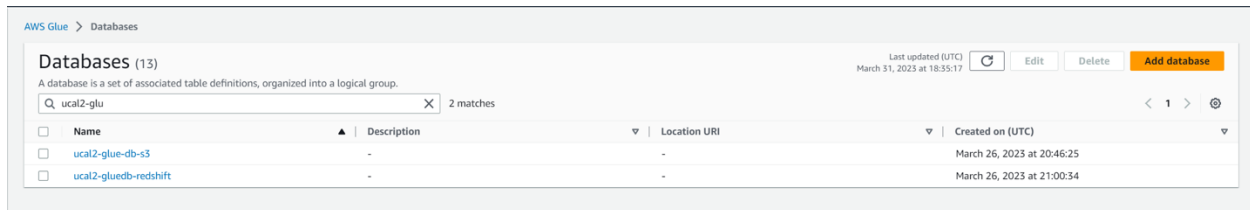
Data sources (1) [Info](#)

The list of data sources to be scanned by the crawler.

Type	Data source	Parameters
<input type="radio"/> JDBC	e2esa-rs-db/public/ucal2-twitter-messages	-

Refresh Edit Remove Add a data source

A Glue database has been created to hold the Glue crawler data. ‘**ucal2-gluedb-redshift**’ has been created for the same.



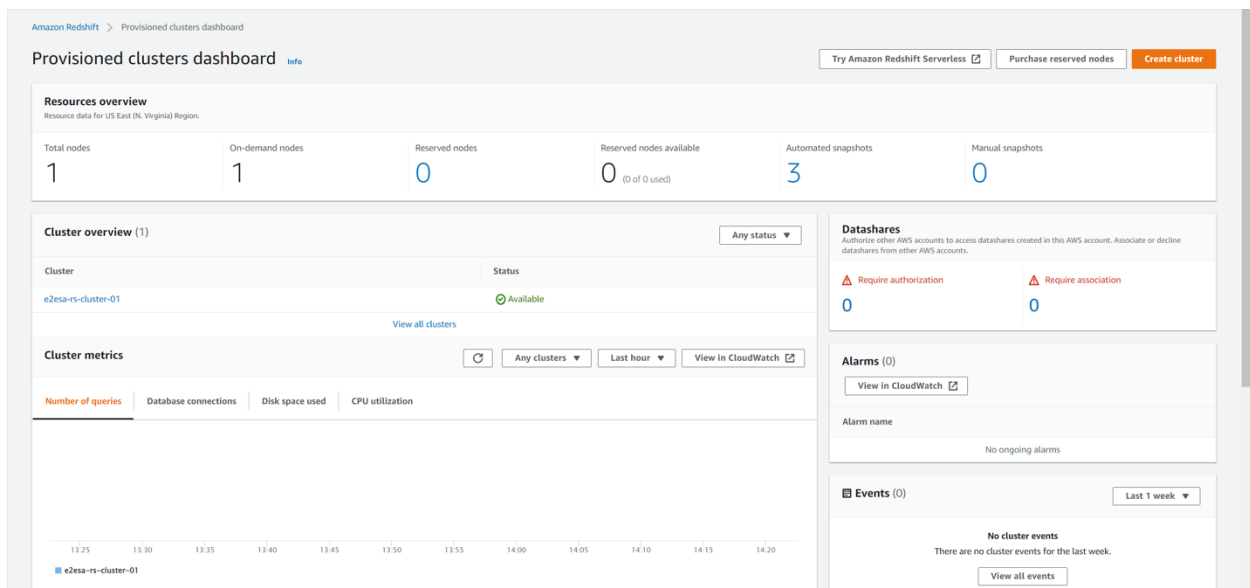
The screenshot shows the AWS Glue 'Databases' page. It features a search bar with 'ucal2-glu' and a table listing two databases: 'ucal2-glue-db-s3' and 'ucal2-gluedb-redshift'. The table columns are Name, Description, Location URI, and Created on (UTC). The 'ucal2-gluedb-redshift' database was created on March 26, 2023, at 21:00:34.

Name	Description	Location URI	Created on (UTC)
ucal2-glue-db-s3	-	-	March 26, 2023 at 20:46:25
ucal2-gluedb-redshift	-	-	March 26, 2023 at 21:00:34

Figure 18 Glue crawler data

Create Redshift Cluster and Database:

To load data to Amazon Redshift we needed a Redshift database to be provisioned. We created a Redshift cluster and a database in that cluster as shown in the figure below.



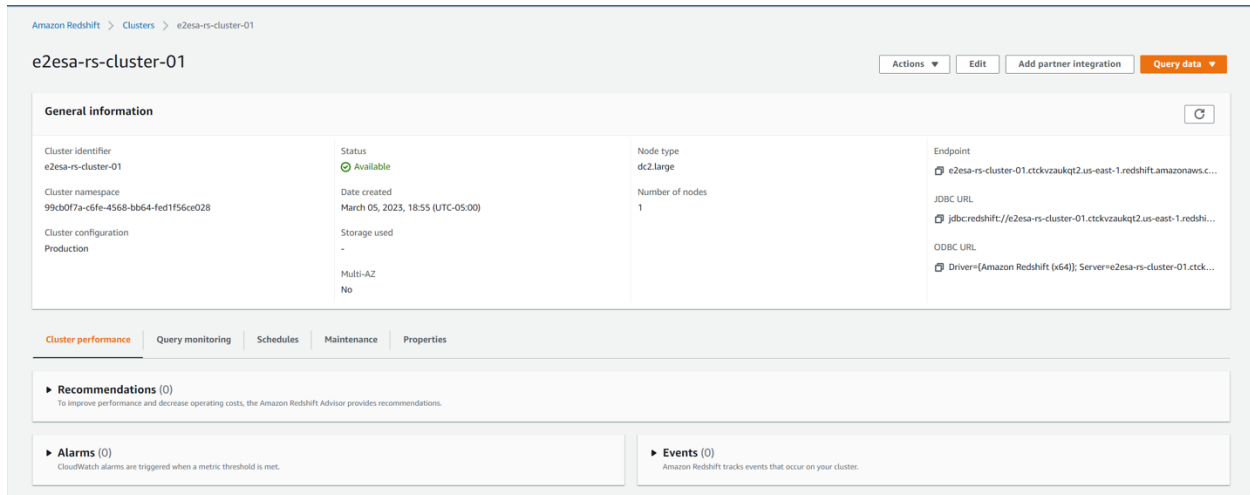


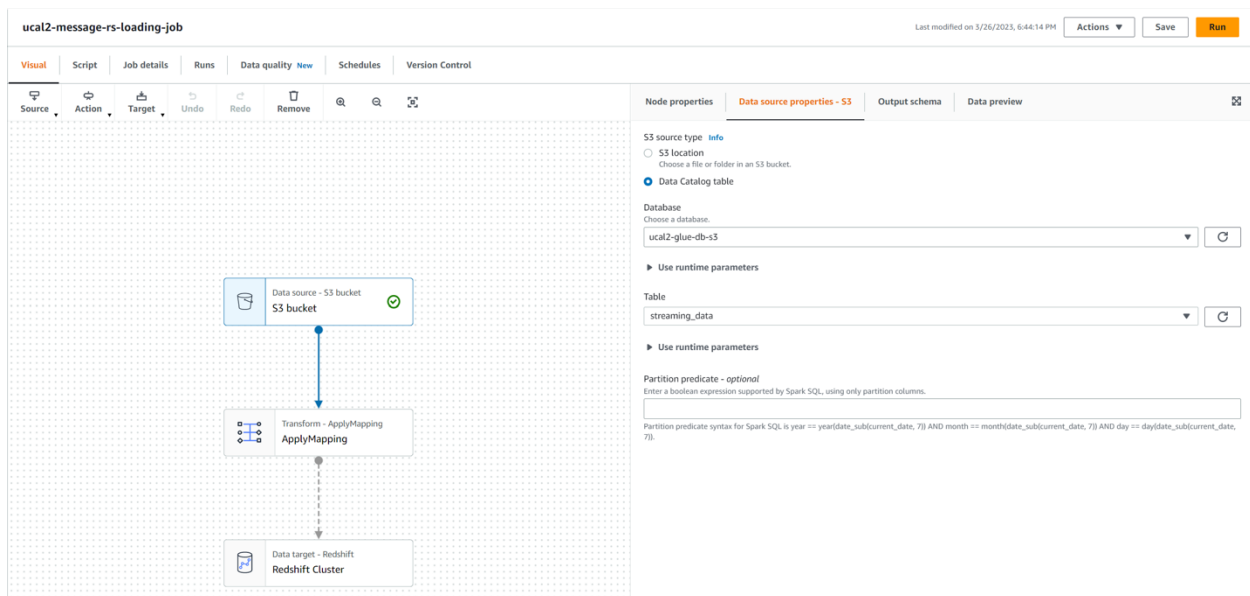
Figure 19 Redshift cluster

Create ETL job for Redshift data load:

Here is the part where we need to create the next ETL job which will take data from the clean zone and load it to Redshift.

There are three parts here.

1. **Define the source:** In our case, it is the Clean Zone. As in the below figure, it should be the data catalog table. It was created by the second Crawler which analyzed clean zone data schema.
2. Create a mapping of the variable between the clean file schema and the Redshift table schema in 'AppMapping'.
3. The target for us is the Redshift table. We did the mapping with the Redshift database connection.



The below figure shows the successful execution of the Redshift data load ETL job.

The screenshot displays the AWS Glue console interface for the 'ucal2-message-rs-loading-job'. The 'Recent job runs' section shows a single run with a status of 'Succeeded'. The table below provides details about the job run.

Recent job runs (13)			
Filter job runs by property			
March 29, 2023 11:42:08 PM			
Job name	Id	Run status	Glue version
ucal2-message-rs-loading-job	jx_0c6453e93a052d8046ca67ca9dc6f1024d5a9aebf116f3000e757e05be3f7	Succeeded	3.0
Retry attempt number	Start time	End time	Start-up time
Initial run	March 29, 2023 11:42:08 PM	March 29, 2023 11:46:21 PM	1 minute 59 seconds
Execution time	Last modified on	Trigger name	Security configuration
2 minutes 13 seconds	March 29, 2023 11:46:21 PM	-	-
Timeout	Max capacity	Number of workers	Worker type
2880 minutes	2 DPU's	2	G.1X
Execution class	Log group name	Cloudwatch logs	Performance and debugging recommendations
STANDARD	/aws-glue/jobs	<ul style="list-style-type: none"> All logs Output logs Error logs 	<ul style="list-style-type: none"> View in CloudWatch

Figure 20 Redshift database connection

2.7 Redshift Database:

After the successful execution of the data load ETL process above we ran a select query from the redshift table as shown in the below figure. The data is uploaded as it is shown below.

The screenshot displays the AWS Redshift Query Editor v2 interface. On the left, a sidebar shows the database structure with a tree view containing 'e2esa-rs-cluster-01', 'e2esa-rs-db', and 'ucl2-twitter-messages'. The main area shows a SQL query: `SELECT * FROM "e2esa-rs-db"."public"."ucl2-twitter-messages";`. Below the query, the results are displayed as a table with 100 rows. The table has columns: id, user_name, user_location, user_description, user_created, user_followers, user_friends, user_favourites, date, and text. The data includes various user profiles and their associated information.

id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	date	text
30	Ma Paz	United States		9/15/2019 18:10	127	974	30217		
100	VideoChinaTV	Beijing	A window brings a real Ch...	7/15/2016 1:57	366	202	256		
184	Michael Schindeler	Sarasota, FL	Air Force Captain Veteran...	11/22/2009 4:35	10127	10231	992		
254	GlobalPandemic.NET	WORLDWIDE	Breaking News & Critical I...	7/13/2010 21:58	26106	26003	2		
322	G V Vinod Kumar	Hyderabad	Senior Journalist	5/16/2010 17:50	543	330	173		
400	TechGenyz	Global	A tech community #New...	2/3/2015 4:49	19567	14662	2109		
469	The Lancet Infectious Dis...	London	The Lancet Infectious Dis...	3/6/2014 16:55	41714	242	760		
542	Concha Chulita		Conchass+Tacos+Truth+Fal...	1/15/2019 0:18	391	4996	34035		
612	Come Closer	Carol City, FL	That natural Beauty Beau...	6/11/2014 18:28	334	576	10356		
680	Costanza Hermanin	Europe	Insegnante. Europeista. F...	8/12/2009 20:20	2831	1118	7221		
745	Outbreak Science	United States	A nonprofit to advance th...	2/12/2016 17:40	1204	41	14		
804	Doha News	Doha, Qatar	Covering breaking news, ...	3/6/2009 21:48	512111	2682	5476		
862	Source of the Spring	Silver Spring, MD	Silver Spring & Takoma P...	7/8/2016 20:12	2768	832	5578		
965	Don West	Washington, DC	Journalist, Author, Futurist...	3/29/2016 1:19	131	176	69		
1042	J-F Claude, M.S.M.	Ottawa (Algonquin Territory)	Leading Canadian Differ...	10/22/2012 12:50	10304	1543	105258		
1127	TamilanCinema	Chennai, India	புதுபி @tamilcinema	11/14/2014 13:16	49177	0	379		
1212	Ian Winter	2020theatre Hull City	https://t.co/dRVmQQuc3...	9/3/2009 18:20	1897	2689	5983		
1300	The Forest of Dean Const...	Regent Hall, Bath Place, ...	Welcome to the official Tw...	5/1/2017 13:27	281	290	182		
1385	James Twiss #GTT0 #O...	Stafford	Old bald guy. Hate injuncti...	7/30/2012 19:32	2206	3162	24816		

Figure 21 Redshift table

Chapter 3: Data Insights and Visualization in Tableau

The final part of our project is to fetch redshift data into Tableau and create a visualization from that.

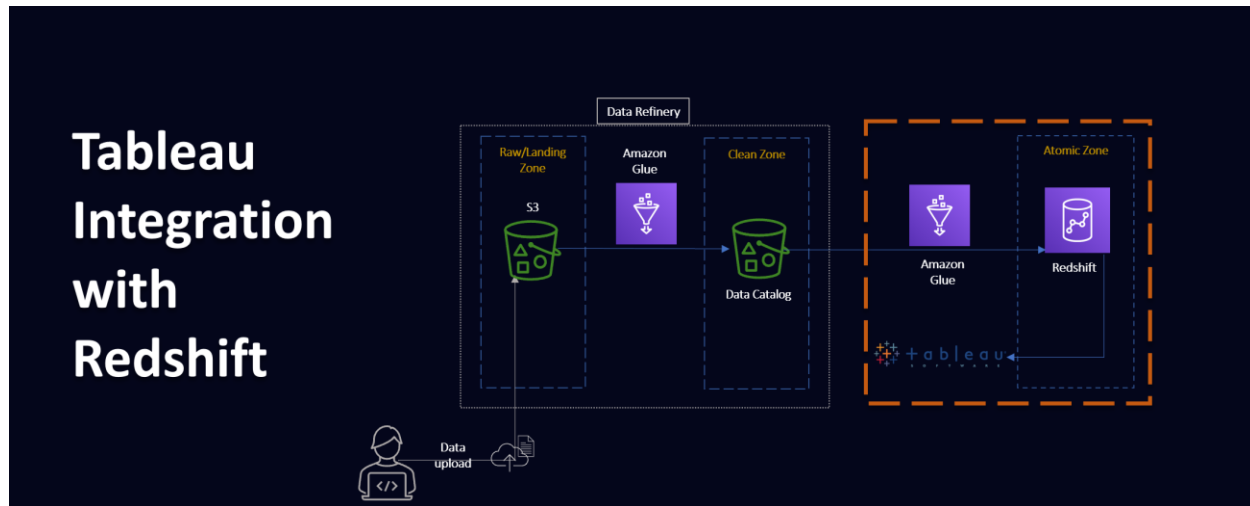


Figure 22 Tableau Integration

Here sentiment analysis is target-oriented, aiming to identify opinions or attitudes from the tweets related to covid. Sentiment analysis focuses here on the polarity of a tweet.

We have used polarity categories *Negative*, *Neutral*, *Positive*, *Slightly Positive*, and *Slightly Negative*.

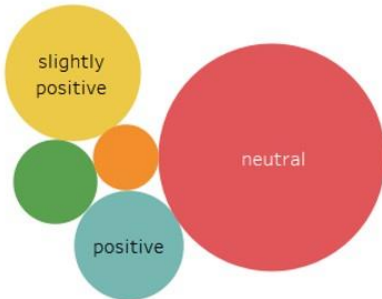
The bubble chart shows the frequency of the polarity categories. Where it clearly shows maximum tweets were normal. very few negative.

Now below the visualization, we have plotted the Polarity vs Subjectivity:

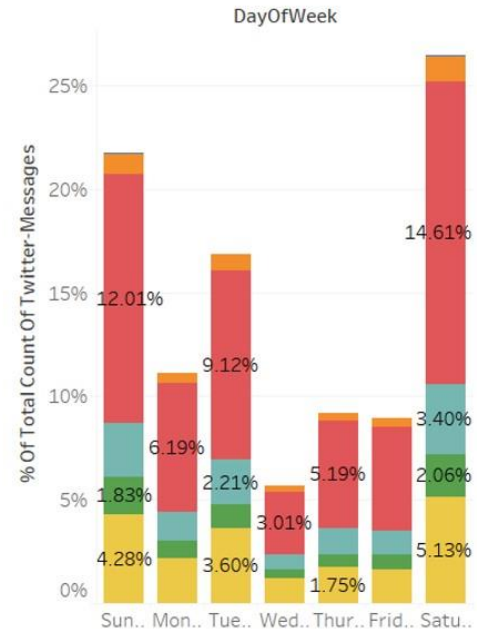
- Polarity is the output that lies between $[-1,1]$, where -1 refers to negative sentiment and +1 refers to positive sentiment.
- Subjectivity is the output that lies within $[0,1]$ and refers to personal opinions and judgments from the tweets.

2020 Covid Tweets Sentiment Analysis Visualisation

Bubbles of Sentiment



Tweets Distribution



Polarity vs Subjectivity

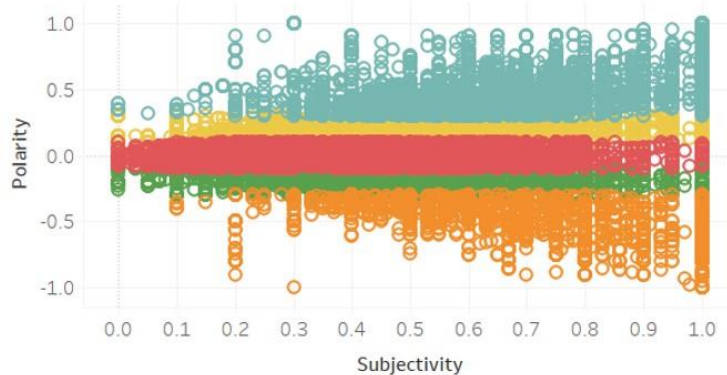


Figure 23 Dashboard for Sentiment Analysis on Twitter Covid Tweets

- The team was able to create a Data Lake solution using AWS cloud resources within budget and timeline
- The solution is fully running as planned
- Our data lake solution can process any structured or unstructured data

3.1 Future Scope:

The future scope mentioned below shows the capability of the data lake and how far it can be extended in the future. The extension depends on the requirement and the skill set. There are a few features like automated job triggering, and real-time data ingestion for real data sources are demonstrated in the below figure.

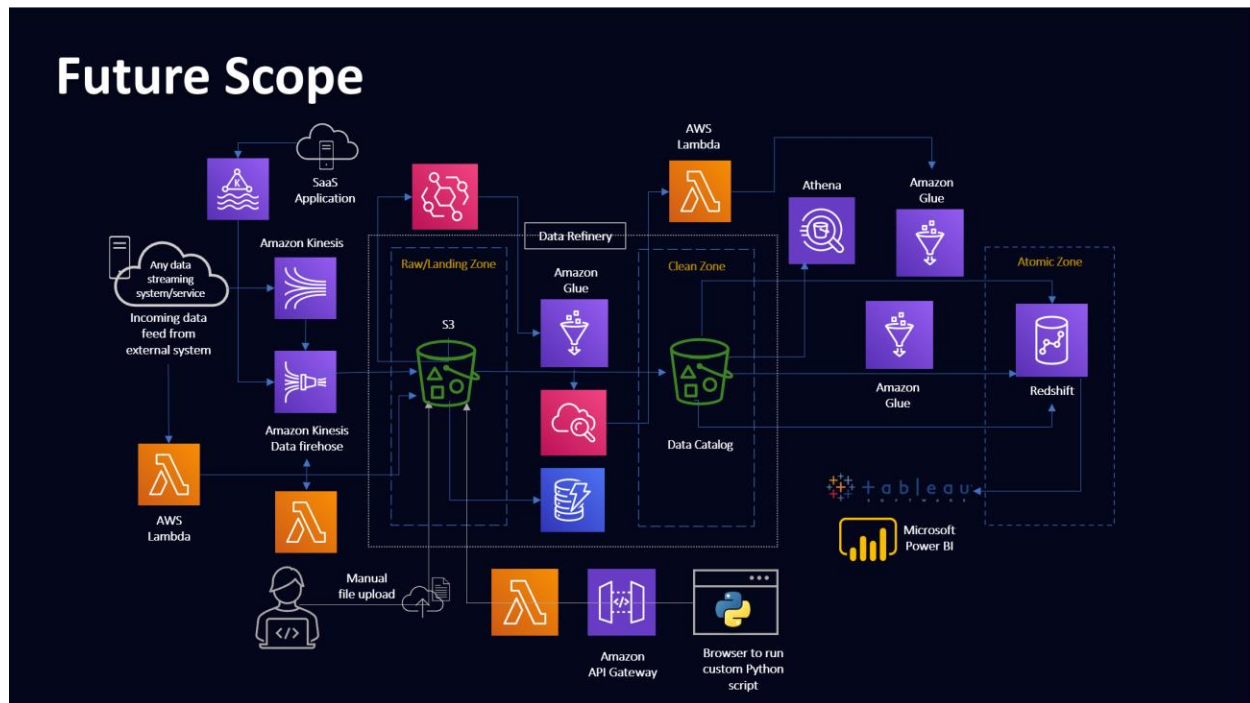


Figure 24 Future Scope

Of these processes, the Amazon API Gateway to collect data using API and Amazon Athena to do preliminary analysis is now complete.

3.2 Summary:

- The team was able to create a Data Lake solution using AWS cloud resources within budget and timeline.
- The solution is fully running as planned.
- Our data lake solution is capable of processing any structured or unstructured data.

3.3 Limitations:

- Currently, ETL Jobs are on-demand; it could be event-driven.
- Currently, purging processed files is manual which could be automated.

References:

- Sayce, D. (2010, March 3). Number of tweets per day? | David Sayce. David Sayce; Paper Gecko Limited. <https://www.dsayce.com/social-media/tweets-day/>
- Dorney, H. (n.d.). How to create and use hashtags. [Www.twitter.com](https://www.twitter.com); Twitter. Retrieved March 25, 2023, from <https://business.twitter.com/en/blog/how-to-create-and-use-hashtags.html#:~:text=On%20Twitter%2C%20adding%20a%20%E2%80%9C%23,that%20they're%20interested%20in.>
- Zelleke, L. (2021, January 21). The 10 Best AWS Alternatives: Amazon Web Services competitors ranked. [Www.itprc.com](https://www.itprc.com); IT Professionals Resource Center. <https://www.itprc.com/best-aws-alternatives/>
- Zhang, M. (2022, June 15). Amazon Web Services (AWS) Data Center Locations: Regions and Availability Zones. Dgtl Infra. <https://dgtlinfra.com/amazon-web-services-aws-data-center-locations/>
- Amazon Web Services. (n.d.). *What is Amazon S3? - Amazon Simple Storage Service*. Docs.aws.amazon.com. Retrieved April 6, 2023, from <https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html>
- Biswal, A. (2023, January 28). *What is Tableau: The Ultimate Guide To Know All About Tableau in 2021*. Simplilearn.com. <https://www.simplilearn.com/tutorials/tableau-tutorial/what-is-tableau>
- Preda, G. (2020, February 28). *COVID19 Tweets*. [Www.kaggle.com](https://www.kaggle.com); Kaggle (Alphabet Inc.). <https://www.kaggle.com/datasets/gpreda/covid19-tweets?resource=download>
- *What is ETL (Extract, Transform, Load)? | IBM*. (n.d.). [Www.ibm.com](https://www.ibm.com); IBM. <https://www.ibm.com/topics/etl>
- Andy Patrizio. (2021, October 8). *Top Data Visualization Tools*. EWEEK. <https://www.eweek.com/big-data-and-analytics/data-visualization-tools/>
- AWS. (2019). *What Is a Data lake?* Amazon Web Services, Inc.; Amazon Inc. <https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/>
- Koivisto, T. (2019). *Efficient Data Analysis Pipelines* (pp. 1–4). University of Helsinki. http://www.edahelsinki.fi/dsns2019/a/dsns2019_koivisto.pdf

- Rahman, M. M., & Hasibul Hasan, M. (2019, October 1). *Serverless Architecture for Big Data Analytics*. IEEE Xplore. <https://doi.org/10.1109/GCAT47503.2019.8978443>