

Visionet-XAnalytics -Develop comprehensive Ops Analytics view to improve Cloud productivity (Cloud+BI tool Systems of engagements - Security/Finance/Ops/Service Mgmt/Engineering)

SYED AFNANUDDIN

*Department of Computer Science and Engineering
Presidency University*

Itgalpur Rajanakunte, Yelahanka, Bengaluru, Karnataka
560064

20191CSE0739@presidencyuniversity.in

Abstract - AWS Glue is a service that allows many data sources to be processed, integrated, and stored into a single storage like S3 or any RDS. This makes it easier to perform analytics or create a dashboard for different types of data. Forecasting these data on a dashboard will therefore enable the industry to receive fully automatic real-time updates on the many statistics it needs, assisting it in avoiding resource waste.

key words - AWS Glue, S3, RDS.

I. INTRODUCTION

In today's world, businesses are generating more data than ever before. This data can come from a variety of sources, such as sensors, social media, and customer transactions. By analyzing this data, businesses can gain valuable insights into their operations and make better decisions.

However, analyzing large amounts of data can be a challenge. This is where cloud computing comes in. Cloud computing provides a scalable and cost-effective way to store, process, and analyze data. This makes it possible for businesses of all sizes to take advantage of big data analytics.

One of the benefits of cloud computing is that it allows businesses to easily integrate data from different sources. This can be done through a variety of methods, such as APIs, data lakes, and data warehouses. Once the data is integrated, it can be analyzed using a variety of tools and techniques.

The results of the analysis can then be displayed on a dashboard. Dashboards provide a visual representation of the data, making it easy to understand and interpret. They can also be used to track progress over time and identify trends.

DHIREN H SHETTY

*Department of Information Science and Technology
Presidency University*

Itgalpur Rajanakunte, Yelahanka, Bengaluru, Karnataka 56006

20191ist0041@presidencyuniversity.in

Overall, cloud computing is a powerful tool that can be used to analyze data from different sources and put it in a dashboard for better understanding. This can help businesses improve their operations, make better decisions, and gain a competitive advantage.

A. What sources do the data on the cloud come from?

1. Internet of Things: These gadgets link to the internet and gather data from a variety of sensors, cameras, and other smart devices before sending it to the cloud.
2. Web apps are programmes that run in a web browser and gather information from clicks and form submissions made by users.
3. Mobile devices: gather information on user behaviour, use patterns, and location
4. Social media: Websites like Facebook, Twitter, and Instagram where users are able to post details about themselves and their activities.
5. Enterprise applications, including HR, ERP, and CRM systems, are those that are used within an organisation to handle business activities.
6. Open data sources, public APIs, and government information that are accessible to the general public
7. Cloud-Native Data: These are data sources made specifically to function with cloud-based services including cloud storage, cloud databases, and cloud analytical tools.



Fig. 1 an image depicting the various sources of data

B. What is monitoring in IT? What data, do we need to capture for a cloud operation?

Checking the functionality and state of various systems, apps, and other components is known as monitoring in the field of information technology.

IT monitoring is crucial because it helps IT professionals to proactively spot issues and fix them before they negatively impact user experience or bring down systems. By spotting possible security risks, it increases system performance, ensures the smooth operation of IT processes, and improves security.

Systems including servers, network equipment, databases, apps, and cloud infrastructure, for instance, may all be monitored. Log analysis, real-time alerts, performance indicators, and predictive analytics are just a few of the monitoring tools and methods available.

The following data must be collected for cloud operations:

1. **Resource consumption:** This offers information on how virtual machines, containers, and other cloud resources are being used in terms of CPU, memory, disc I/O, and network usage.
2. **Performance metrics** for cloud-based applications and services include information on response times, throughput, and other important performance parameters.
3. **Logs:** Logs record actions and events that take place in the cloud environment, such as security-related events, system faults, and application problems.
4. **Data on access control, authentication, and authorization** procedures as well as information on security-related occurrences like breaches, vulnerabilities, and threats are included in the category of security data.

5. Data about consumption and pricing for cloud resources, such as compute instances, storage, and network traffic, is included in this.
6. Data on compliance with legal requirements and audit trails of actions taken inside the cloud environment are included in the compliance and audit categories.

C. What data should be displayed if we wish to offer all data in one dashboard?

The information that should be shown on a dashboard relies on the particular requirements of the organization and the target audience. However, the following are some typical data kinds that may be shown in a dashboard:

1. **Key Performance Indicators (KPIs)** are measurements that show how well an organization is performing. Examples include revenue, sales, customer satisfaction, and website traffic.
2. **Data pertaining to the organization's operations**, such as inventory levels, output rates, or service levels, are referred to as **operational data**.
3. **Financial information**: This comprises information about the company's financial performance, such as sales, costs, profit margins, or cash flow.
4. **Customer information**: This refers to information on how customers behave, such as demographic information, purchasing history, or customer service interactions.
5. **Data about marketing efforts**, such as website traffic, conversion rates, or social media engagement, is included in this category.
6. **Information on the organization's employees**, such as statistics on employee turnover, performance indicators, or training and development, is referred to as **human resources data**.
7. **Data pertaining to the general health of the organization**, such as trends, patterns, and insights gleaned from numerous data sources, is referred to as **business intelligence data**.

D. Which tools may be utilised to gather data from several cloudOps tools that are gathered in one location?

1. *SumoLogic*
2. *DataDog Splunk*
3. *Log Analytics in Azure*
4. *Google Cloud Observation*

E. DASHBOARD OPTED FOR THE ANALYSIS

OPTION 1 (APACHE SUPERSET) :

An open-source platform for data exploration and visualization is Apache Superset. It is constructed using React, Python, and SQLAlchemy. Users of Superset can browse and display data from many different sources, such as databases, cloud storage, and APIs. Additionally, it offers a wide range of dashboard creation and sharing features, such as drag-and-drop chart creation, an integrated SQL editor, and a potent search engine. Superset is an effective tool for exploring and visualizing data. It is simple to use and adaptable to any organization's requirements. In addition to being extremely scalable, Superset can be used to analyze vast amounts of data.

STEPS TO INSTALL IT :

For Debian and Ubuntu

The following command will ensure that the required dependencies are installed:

- `sudo apt-get install build-essential libssl-dev libffi-dev python-dev python-pip libsasl2-dev libldap2-dev default-libmysqlclient-dev`
- `sudo apt-get install build-essential libssl-dev libffi-dev python3-dev python3-pip libsasl2-dev libldap2-dev default-libmysqlclient-dev`

Let's also make sure we have the latest version of pip and setup tools:

- `pip install --upgrade setup tools pip`

Python Virtual Environment

- `pip install virtualenv`

You can create and activate a virtual environment using:

- `# virtualenv is shipped in Python 3.6+ as venv instead of pyvenv.`

`# See https://docs.python.org/3.6/library/venv.html`

`python3 -m venv venv`

`venv/bin/activate`

Installing and Initializing Superset:

First, start by installing `apache-superset`

- `pip install apache-superset`

Package Dependencies are (To overcome errors in superset installation)

- `Typing-Extension=3.10.0`
- `Flask-Limiter=2.0.0`
- `Flask-appbuilder=4.1.3`
- `Alembic=1.6.5`
- `Werkzeug=2.0.0`
- `Flask-WTF=0.14.3`
- `Jinja2=3.0.0`
- `Cryptography=38.0.0`
- `Rich=12.0.0`

`Pip install --upgrade`

Then, you need to initialize the database:

- `superset db upgrade`

Finish installing by running through the following commands:

- # Create an admin user in your metadata database (use `admin` as username to be able to load the examples)

export FLASK_APP=superset

superset fab create-admin

- # Load some data to play with

superset load_examples

- # Create default roles and permissions

superset init

- # Build javascript assets

cd superset-frontend

npm ci

npm run build

cd ..

- # To start a development web server on port 8088, use -p to bind to another port

superset run -p 8088 --with-threads --reload --debugger

F. REASON TO NOT GOING FORWARD WITH APACHE SUPERSET :

There was no Direct connection for neither Jira nor git nor aws cloudwatch so we had to change the dashboard which we had use as it could only connect to databases like Redshift, DynamoDB.

OPTION 2 (GRAFANA) :

A web application for interactive visualisation and analytics is called Grafana. When connected to supported data sources, it offers charts, graphs, and alerts for the web. Through the use of Docker or Docker Compose, installation is simple. A well-liked tool for tracking and visualising metrics, logs, and traces is called Grafana. Numerous businesses, including start-ups, large corporations, and governmental institutions, use it. A strong tool like Grafana can be used for: Monitoring metrics: Grafana can be used to track metrics from Prometheus, InfluxDB, and Elasticsearch, among other sources. Logs can be visualised using Grafana, which can display data from ELK, Splunk, and Graylog, among other sources. Trace: You can use Grafana to track requests coming from different sources, such as Jaeger, Zipkin, and OpenTelemetry.

To install the latest release:

- `sudo apt-get install -y apt-transport-https`
- `sudo apt-get install -y software-properties-common`
- `wget -q -O /usr/share/keyrings/grafana.key https://apt.grafana.com/gpg.key`

Add this repository for stable releases:

```
echo "deb [signed-by=/usr/share/keyrings/grafana.key]
https://apt.grafana.com stable main" | sudo tee -a
/etc/apt/sources.list.d/grafana.list
```

Add this repository if you want beta releases:

```
echo "deb [signed-by=/usr/share/keyrings/grafana.key]
https://apt.grafana.com beta main" | sudo tee -a
/etc/apt/sources.list.d/grafana.list
```

After you add the repository:

```
sudo apt-get update
# Install the latest OSS release:
sudo apt-get install grafana
```

G. FURTHER STEPS TO ANALYZE THE DATA :

Downloading Jira Software Management to get the log data

Jira Software using Docker

Update the packages in the systems

- `sudo apt update`
- `sudo apt-get update`

Creation of volume and assigning the port to it

For the `JIRA_HOME` directory that is used to store application data (amongst other things), we recommend mounting a host directory as a `data volume`, or via a named volume.

Additionally, if running Jira in Data Center mode it is required that a shared file system is mounted. The mount point (inside the container) can be configured with `JIRA_SHARED_HOME`.

To get started you can use a data volume, or named volumes. In this example, we'll use named volumes.

- `docker volume create --name jiraVolume`

- `docker run -v jiraVolume:/var/atlassian/application-data/jira --name="jira" -d -p 8080:8080 atlassian/jira-software`

Success. Jira is now available on <http://localhost:8080>*

Please ensure your container has the necessary resources allocated to it. We recommend 2GiB of memory allocated to accommodate the application server. See [System Requirements](#) for further information.

** Note: If you are using docker-machine on Mac OS X, please use open `http://$(docker-machine ip default):8080` instead.*

DOCKER :

Docker is a tool that aids in the consistent packaging and execution of applications across various environments. Docker containers are small, standalone, executable software packages that contain all the components required to run an application, including the code, runtime, system tools, system libraries, and settings. Docker containers can run concurrently without interfering with one another because they are isolated from the host machine and from one another. Docker is therefore perfect for setting up and operating microservices-based applications. System administrators and developers both use Docker frequently. Numerous businesses, including start-ups, large corporations, and governmental institutions, use it.

Installation of docker

- `snap install docker`
- `apt install docker.io`

With this the installation is completed.

To check the status of the server.

- `systemctl status docker`
- `docker ps`

REASON TO NOT GO AHEAD WITH DOCKER :

We could not go further with this approach as the Docker was not visible after one day of usage. Because of which we had to go with traditional approach of downloading the Jira Service Management on the instance directly

Alternate ways to download Jira Service Management

Update the packages in the systems

- **sudo apt update**
- **sudo apt-get update**

Follow the link below to download the latest version of Jira Service Management

<https://www.atlassian.com/software/jira/service-management/download-archives>

To Download on the EC2 Instance we follow the below command

- **wget**
<https://www.atlassian.com/software/jira/downloads/binary/atlassian-servicedesk-5.7.0-x64.bin>

BIT BUCKET :

A self-hosted version of Bitbucket is called Bitbucket Server. It is a hosting service for Git-based source code repositories that provides a wide range of features for teams of all sizes, from tiny startups to enormous corporations. For groups that have an on-premises requirement for hosting their code, Bitbucket Server is a popular option. Compared to cloud-based solutions, it has a number of benefits, including: Control: Teams have total control over their code thanks to Bitbucket Server. They

have control over its hosting location, configuration, and user access. Security: Bitbucket Server's hardening capabilities can be tailored to each organization's unique security requirements. Compliance: Bitbucket Server can be set up to comply with the regulations of a variety of sectors, including the financial and healthcare industries.

Install Bitbucket Server

1. Download Bitbucket Server

<https://www.atlassian.com/software/stash/downloads/binary/atlassian-bitbucket-8.9.0-x64.bin>

2. Run the installer

1. Make the installer executable.

```
chmod +x atlassian-bitbucket-8.9.0-x64.bin
```

2. Run the installer – we recommend using sudo to run the installer as this will create a dedicated account to run Bitbucket Server and allow you to run Bitbucket Server as a service.

```
sudo ./atlassian-bitbucket-8.9.0-x64.bin
```

3. Follow the prompts to install Bitbucket. You'll be asked for the following info:

a. Type of Bitbucket instance - the type of installation, for these instructions select Standard.

b. Installation directory - where Bitbucket will be installed.

c. Home directory - where Bitbucket application data will be stored.

d. TCP ports - the HTTP connector port and control port Bitbucket will run on.

4. Once the installer completes, head to <http://localhost:7990> in your browser to begin the setup process.

3. Connect to your database

If you've not already done so, it's time to create your database. See the 'Before you begin' section of this page for details.

Select External as your database, then choose a Database Type from the dropdown menu and enter the details of your database.

4. Add your license key

Follow the prompts to log in to my.atlassian.com to retrieve your license, or enter a license key.

You can also set the base URL at this step, (you can elect to do this later).

5. Create your administrator account

Enter details for the administrator account.

Select either Go to Bitbucket to go straight to the Bitbucket interface or Integrate with Jira to create your connection with an existing Jira application.

6. Start using Bitbucket Server

That's it! Your Bitbucket site is accessible from a URL like this:

http://<computer_name_or_IP_address>:<port>

REASON TO NOT GO AHEAD WITH BIT BUCKET :

This approach couldn't be used as the Atlassian bitbucket and Jira service management marketplaces has removed the amazon s3 connector, hence we moved towards another way to send JiraSM audit logs to amazon s3.

H. STEPS THAT LEAD US TO COMPLETE THE ANALYSIS OF DATA FROM VARIOUS SOURCES :

AWS (AMAZON WEB SERVICES) :

AWS SERVICES THAT WERE USED -



Fig . 2 depicts the services that were under use during this project

1. Compute: AWS provides a range of compute services, such as Amazon Lambda, Amazon Elastic Container Service, and Amazon Elastic Compute Cloud (EC2). Virtual machines are available from EC2 and can be used to run a variety of applications. A managed service for running Docker containers is offered by

ECS. Without provisioning or managing servers, code can be run using Lambda's serverless compute service.

2. Amazon Simple Storage Service (S3), Amazon Relational Database Service (RDS), and Amazon Elastic Block Store are just a few of the storage services that AWS provides (EBS). A very scalable object storage service is S3. For a variety of database engines, RDS offers a managed database service. Block storage is offered by EBS and can be connected to EC2 instances.

3. Networking: AWS provides a range of networking services, including Amazon CloudFront, Amazon Route 53, and Amazon Virtual Private Cloud (VPC). A private network is offered by VPC and can be used to isolate AWS resources. A managed DNS service is Route 53. A content delivery network called CloudFront can be used to enhance the functionality of websites and applications.

4. Databases: RDS, Amazon Aurora, and Amazon Redshift are just a few of the database services that AWS provides. For a variety of database engines, RDS offers a managed database service. A MySQL-compatible database engine with a focus on high performance and availability is called Aurora. A data warehouse service called Redshift is made for big data analytics.

5. Analytics: AWS provides a range of analytics services, including Amazon QuickSight, Amazon Athena, and Amazon Simple Analytics Service (S3). A data lake is offered by S3 and can be used to store different types of data. Data analysis can be done using the business intelligence service QuickSight. A serverless query service called Athena can be used to examine S3 data.

ALTERNATE WAY TO EXTRACT THE DATA :

ELT VS ETL

- In contrast to ETL, where processes cannot be carried out concurrently due to interdependencies between them, ELT allows for much greater scalability. These days, businesses prefer to use ELT to increase productivity. In contrast to ETL, which stores data in temporary locations, ELT stores data in a targeted location where users can analyse data at any time by using various filters and business logic.

Jira Service Mngement :

We can create multiple tickets for different issues using the Jira interface, assign them to different users, and download the data as a CSV file. The S3 bucket is then used to store this data.

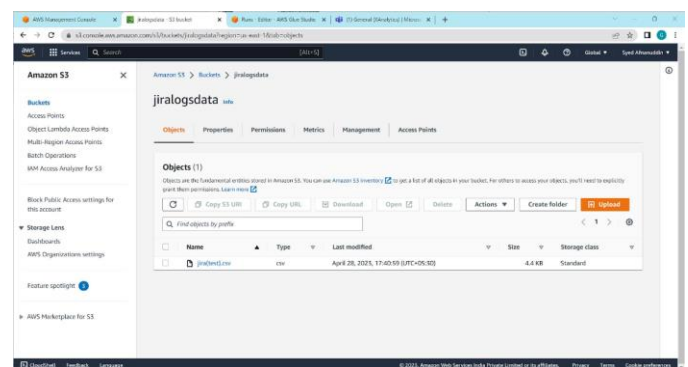
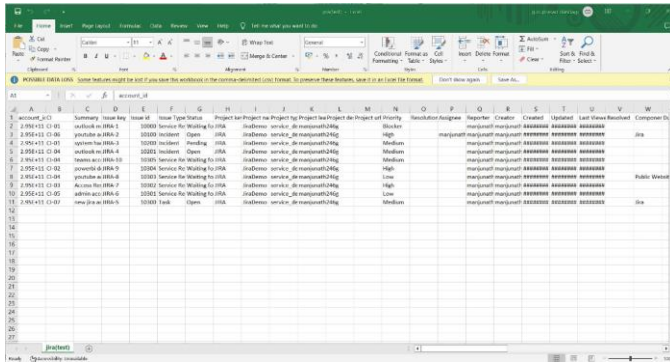
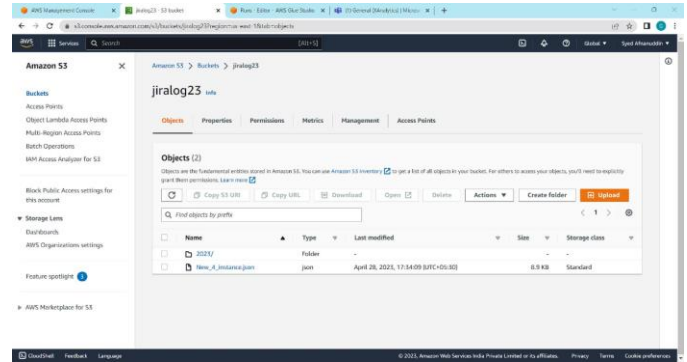


fig . 3 the figure depicts the log data that was collected from jira service management and put into a s3 bucket .



Issue ID	Summary	Issue Key	Issue Type	Status	Project	Assignee
1. JIRA-1001	Summary of Issue 1	JIRA-1001	Task	Open	Project A	Assignee A
2. JIRA-1002	Summary of Issue 2	JIRA-1002	Task	Open	Project A	Assignee B
3. JIRA-1003	Summary of Issue 3	JIRA-1003	Task	Open	Project A	Assignee C
4. JIRA-1004	Summary of Issue 4	JIRA-1004	Task	Open	Project A	Assignee D
5. JIRA-1005	Summary of Issue 5	JIRA-1005	Task	Open	Project A	Assignee E
6. JIRA-1006	Summary of Issue 6	JIRA-1006	Task	Open	Project A	Assignee F
7. JIRA-1007	Summary of Issue 7	JIRA-1007	Task	Open	Project A	Assignee G
8. JIRA-1008	Summary of Issue 8	JIRA-1008	Task	Open	Project A	Assignee H
9. JIRA-1009	Summary of Issue 9	JIRA-1009	Task	Open	Project A	Assignee I
10. JIRA-1010	Summary of Issue 10	JIRA-1010	Task	Open	Project A	Assignee J

fig . 4 this image depicts the data in a table format . the tickets that were raised in jira.



Name	Type	Last modified	Size	Storage class
2022/	Folder			
New_A_Instance.json	File	April 26, 2022, 17:54:09 UTC+05:30	0.9 KB	Standard

fig 5 . this figure depicts the data stored in the jiralog23 s3 bucket.

Amazon Kinesis :

A fully-managed service called Amazon Kinesis makes it simple to gather, process, and analyse streaming data. It is a well-liked option for many use cases, such as: Analytics for streaming data: Kinesis can be used to instantly analyse streaming data. Anomaly detection, real-time recommendations, and fraud detection are just a few uses for this.

On streaming data, machine learning models can be trained and deployed using Kinesis. Sentiment analysis, image recognition, and natural language processing are just a few applications for this. Kinesis can be used to analyse log data from devices and applications. This can be applied to resolve issues, spot trends, and boost efficiency. IoT: Data from IoT devices can be gathered and analysed using Kinesis. This can be applied to track events, keep track of devices, and boost the efficiency of IoT systems.

The Cloudwatch data from multiple ec2 instances is produced and passed through a kinesis stream stored in an S3 bucket



```

{
  "id": "1",
  "name": "John",
  "age": 30,
  "gender": "Male",
  "email": "john.doe@example.com",
  "phone": "1234567890",
  "address": {
    "street": "123 Main St",
    "city": "New York",
    "state": "NY",
    "zip": "10001"
  },
  "education": {
    "school": "ABC High School",
    "degree": "Bachelor's",
    "major": "Computer Science",
    "graduation_year": 2010
  },
  "employment": {
    "company": "XYZ Corp",
    "position": "Software Engineer",
    "start_date": "2015-01-01",
    "end_date": "2020-12-31"
  },
  "hobbies": [
    "Reading",
    "Golfing",
    "Traveling"
  ],
  "pets": [
    {
      "name": "Fido",
      "species": "Dog",
      "breed": "Golden Retriever",
      "age": 5
    },
    {
      "name": "Whiskers",
      "species": "Cat",
      "breed": "Siamese",
      "age": 3
    }
  ],
  "social_media": {
    "facebook": "john.doe.123",
    "twitter": "johndoe123",
    "instagram": "johndoe123"
  },
  "languages": [
    "English",
    "Spanish",
    "French"
  ],
  "skills": [
    "Programming",
    "Data Analysis",
    "Project Management"
  ],
  "last_updated": "2022-04-26"
}

```

fig . 6 this image depicts understanding of data thats been collected using jason parser.

EXTRACT , LOAD AND TRANSFORM USING AWS GLUE :

AWS GLUE :

Data preparation for analytics, machine learning, and application development is simple with AWS Glue, a serverless, fully managed ETL service. You can find, prepare, and combine data from various sources using AWS Glue, which has an easy-to-use interface. A robust ETL engine is also part of AWS Glue, which can automatically extract, transform, and load data into various data stores.

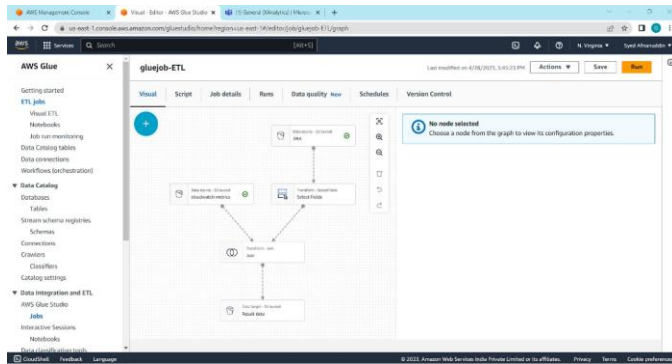


fig 7. depicts how the data is ETL process takes place in AWS Glue

The data from JiraSM and Cloudwatch data from the two S3s, and then the two data sets are joined, the data is filtered out or modified, and the output data is obtained and sent to the S3 (there are other destinations also like RDS). Finally, we can run the glue job and save the process.

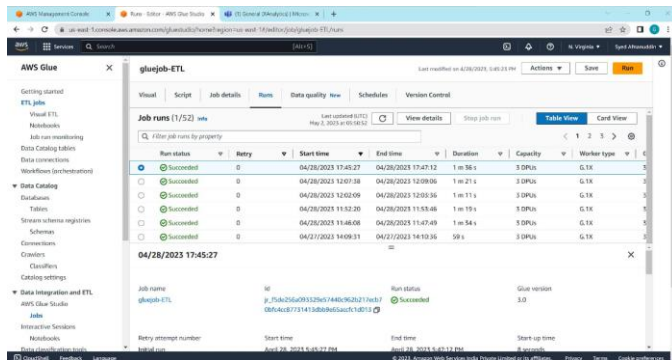


fig 8 . It depicts the gluejob done to transform the data for further analysis.

ARCHITECTURE USED TO ANALYZE DATA FROM MULTIPLE SOURCES :

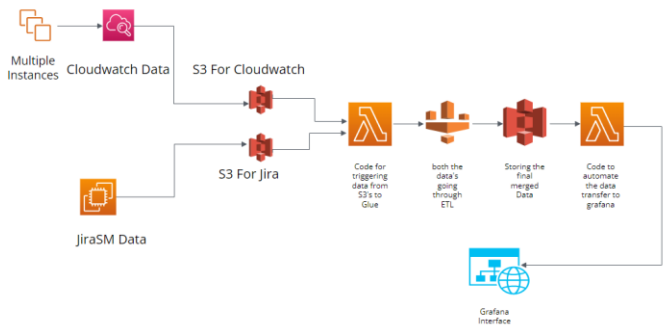


Fig 9 . This image depicts the architecture used to tackle the issue of analyzing the data from multiple sources .

In the architecture shown above, we can see that we have created multiple instances for generating cloudwatch data and, on the other hand, that we have another instance of JiraSM installed. These two sets of data are stored in two separate S3 storage buckets, and after going through the ETL process, the final product is delivered as a single file. The data is stored in an S3 and streamed on Grafana, a third-party dashboard that provides real-time analytics about various services and software. The ETL process only takes place when the lambda is triggered, which can only occur when the data is continuously streamed. This data is being produced to Grafana via a trigger, which creates a pre signed URL of the data in S3 and sends it straight to Grafana.

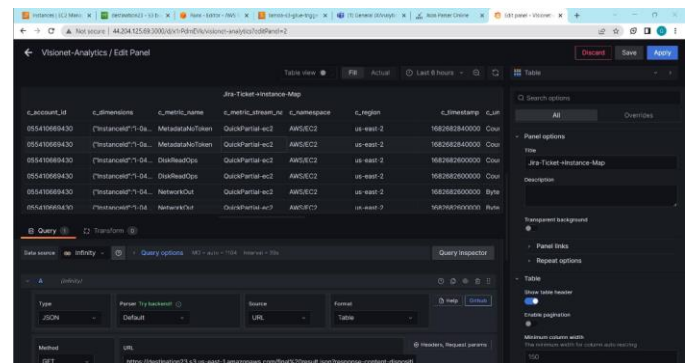


fig 10 . This image depicts the analysis of data from multiple sources on the grafana dashboard .

CONCLUSION :

In conclusion, we can say that this project gives us a clear understanding of how various data sets can be viewed in real-time from a single data source to the dashboard without switching to multiple dashboards, which shortens the time needed to build dashboards or switch tabs. We have learned and used glue, and we used lambda functions to trigger various functions in pushing the data into glue and another trigger to send the data as a link to the third dashboard. We scripted the glue to take two different sets of data and combine them into a single file (Grafana). Through this report, I'd like to draw your attention to the significance of cloud computing and analytics for real-time needs and efficiency.

References :

1. <https://www.cloudzero.com/blog/cloud-financial-management>
2. <https://www.comparitech.com/net-admin/best-cloud-analytics-tools/#:~:text=The%20best%20cloud%20analytics%20tools%201%201.%20AppOptics,5.%20Board%20...%206%206.%20TIBCO%20Spotfire%20>
3. [What is CloudOps and How to implement it? - DevOpsSchool.com](https://www.missioncloud.com/blog/the-top-10-security-tools-for-your-aws-environment)
4. <https://www.missioncloud.com/blog/the-top-10-security-tools-for-your-aws-environment>
5. <https://www.ibm.com/topics/it-operations>
6. <https://www.dnsstuff.com/infrastructure-monitoring-tools>
7. <https://resources.owllabs.com/blog/it-operations/#:~:text=IT%20operations%20are%20the%20activities,internal%20and%20external%20IT%20systems.>
8. [https://www.ibm.com/topics/it-service-management#:~:text=IT%20service%20management%20\(ITSM\)%20is,employees%2C%20customers%20or%20business%20partners.](https://www.ibm.com/topics/it-service-management#:~:text=IT%20service%20management%20(ITSM)%20is,employees%2C%20customers%20or%20business%20partners.)
9. <https://sematext.com/blog/cloud-monitoring-tools/>
10. <https://www.integrate.io/blog/get-data-from-multiple-sources/>
11. <https://www.softwaretestinghelp.com/itsm-tools/>
12. <https://docs.aws.amazon.com/>
13. <https://grafana.com/docs/grafana/latest/>