*Dr.G.Y.PATHRIKAR COLLEGE OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY.*

*A PROJECT SYNOPSIS*

*ON*

# *“ Handling Big data using AWS ETL pipeline ”*

***SUBMITTED BY***

**ROHIT SATISH JOSHI**

*B.Sc. (COMPUTER SCIENCE)-6TH SEMESTER*

*YEAR (2023-2024)*

***GUIDED BY***

**Ms.PATIL R.**

***SUBMITTED TO***

*Dr.G.Y.PATHRIKAR COLLEGE OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY.*

***MGM UNIVERSITY, CH.SAMBHAJINAGAR***

***Dr.G.Y.PATHRIKAR COLLEGE OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY.***

॥ अप्पो दीपो भव ॥
महात्मा गांधी मिशन विश्वविद्यालय
MGMUNIVERSITY

# *CERTIFICATE*

This is to certify that the candidates of **B.sc (Computer Science) 3rd year, 6th semester** has satisfactorily completed the project entitled “Handling Big data using AWS ETL pipeline” fulfillment of Bachelor’s Degree of MGM University, Chh.Sambhajinagar for academic year 2023-2024

**Submitted by :-**

**Saili Sunil Malshikhare**

| | | |
|---|---|---|
| **Guide by:** | **Examiner:** | **HOD:** |
| **Ms.Kumbhakarna V.M.** | | **Dr. S.Y. Azade** |

# ACKNOWLEGMENT

At every outset I express my gratitude to almighty lord for showering his grace and blessings upon us to complete this project. Although our name appears on the cover of this project, many people have contributed in some or the form to this project. We could not have done this Project without the assistance or support of each of the following. We thank all of you.

I wish to place o my record my deep sense of gratitude toward my project guide **Ms.Kumbhakarna V.M.** for the constant motivation and valuable help through the project. I also want to express my gratitude towards our Head of department **Dr. S.Y. Azade** , for his invaluable guidance, insightful advice's and continue encouragement.

We also thanks to those who directly as well as indirectly supported us for completion of this Project Report.

Thanking you.

| **Submitted by** | **Seat no’s :-** |
|---|---|
| **Saili Sunil Malshikhare** | ***A0233048*** |

# Table of contents

# 1. INTRODUCTION

## Handling Big data using AWS ETL pipeline:-

Modern businesses tend to generate a lot of data everyday. once the data is generated, it is required to be stored and analyzed so that strategic business decisions can be made based on the insights gained. In today's word, more & more Organization shifting their infrastructure to amazon cloud services (AWS), provides a fully managed cloud data to are warehousing solution which is amazon Redshift.

Amazon Redshift is a fully managed data warehouse on the cloud. It supports massively. parallels processing Architecture (MPP), which allows users to process data parallely. It allows users to load and transform data within Redshift & then make it available for business intelligence tools firstly the raw data is Extracted to amazon S3. Amazon simple storage service (Amazon S3) is an object storage service offering industry. leading scalability, data availability, security & performance . it is an object Storage built to store and retrieve any amount of data from anywhere, further the stored data is Transformed to the amazon EMR which is used to process the data, amazon EMR can cleanse, join / Lookup, Aggregate allocate, compress, partition the data as per the need then further the processed data is load into the amazon S3 bucket which can parquet the data from unauthorized access and once the client data is encrypted then it transfer into the amazon Redshift warehouse for storage.

## *1.2. NEED AND SIGNIFICANCE OF PROPOSED SYSTEM.*

Amazon Redshift is used to manage and analyze the large amount of data. One of the key benefits of using Aws Redshift is that it is a fully managed service. Aws takes care of all underlying infrastructure and maintenance, freeing up organization's it resources to focus on Other task, also Aws red-shift is the cost-effective service, Aws Redshift is the is priced based on date amount of data you store and the amount of data you query and there long-term commitments

Aws Redshift is designed to handle large amount of data and provides high performance & scalability. It can scale up and down in real-time to meet the changing need of your organization Aws Redshift integrates seamlessly with other Aws services, such as AWS Amazon S3, Amazon EMR, and Amazon Athena. This allows you to transfer between these Services & store, process easily, and analyze your data in a single, integrated platform. Aws Redshift support multiple data sources csu, json and Apache parquet. you can easily load data from these sources into your date warehouse & query it using SQL.

Aws Redshift has built - in security features, including network isolation, rest encryption and IAM authentication it also supports real-time data analytic using its columnar Storage & MPP architecture. Instead of this Aws Redshift is easy to used and comes with a user-friendly web-based console and a range of tools and libraries quering date.

## 1.3. Objectives and Motivation

In a modern data architecture, unified analytic enable you to access the date you need wheather its stored in a data lake or a data Warehouse. In particular, we have Observed an increasing numbers of customer who combine and integrate their data into amazon Redshift data their warehouse to analyze huge data and run Complex queries to achieve their business goals.

one of the most common uses for data preparation on amazon Redshift is to ingest and transform data from different. date stores into an amazon Redshift data Warehouse. This is commonly achieved via aws glue, which is server-less, scalable data integration service that make it easier to discover, prepare, move and integrate data from multiple sources. Aws glue provide an extensible architecture that enables uses with different data processing use cases and work well with Amazon Redshift

# 2. System Requirement

**Amazon S3 :-**

Amazon simple storage is a Service (Amazon S3) is a highly scalable object Storage services. Amazon S3 can be used for a wide range of storage solution, including websites, mobile applications, backups and data-lakes.

**Amazon Redshift :-**

Amazon Redshift is fully managed, Petabytes - Scale data warehouse service , with amazon Redshift, you can query petabyte of structured and semi-structured data across your data. warehouse and your data lake using Standard SQL

**Aws glue :-**

Aws glue is fully managed ETL service that makes it easier to prepare and load data for analytics. Aws glue discover your date and stores the associated metadata in the Aws glue Catalog.

**Amazon EMR :-**

Amazon EMR ( Amazon elastic map Reduce ) is managed cluster platform that Simplifies running big data frameworks, on Aws to process and analyze vast amount of data. it also lets you transform and move large amount of date into and out of other Aws data stores and databases.

# 3. Requirement analysis

Thousands of Customers Choose Amazon Redshift to accelerate their time to insights because it is powerful analytics system that integrates well with database and machine learning services, is streamlined to use and can become a Central service  to deliver on all their analytics need. Amazon Redshift Serverless automatically provisions and scales data Warehouse capacity to to deliver high performance for demanding and unpredictable workloads.

Amazon Redshift offers leading price performance for diverse analytics workloads weather it is dash-boarding, application development, data sharing, ETL jobs, with 10,000 of customers running analyties on terabytes to petabytes of data, Amazon Redshift optimize real-world customer workload performance. Amazon Redshift let you get insights from running real-time and predictive analytics on all your data across your operational databases, data lake, data warehouse, streaming data and third party datasets

Amazon Redshift supports industry-leading security  with built-in identity management and federation for single sign-on (**sso**), multi-factor authentication Column level access control, row-level security, role-based access controls, Amazon Virtual private cloud, and faster cluster resize

# 4. Software-Requirement specifications (SRI)

**Aws (Amazon Web Services)**

Aws is an online platform providing cost-effective Scalable cloud computing solution. It offers a range of on-demand operations such as compute power, content delivery, database storage, to help enterprise and organization grow.

# *5. Data Flow Diagram*

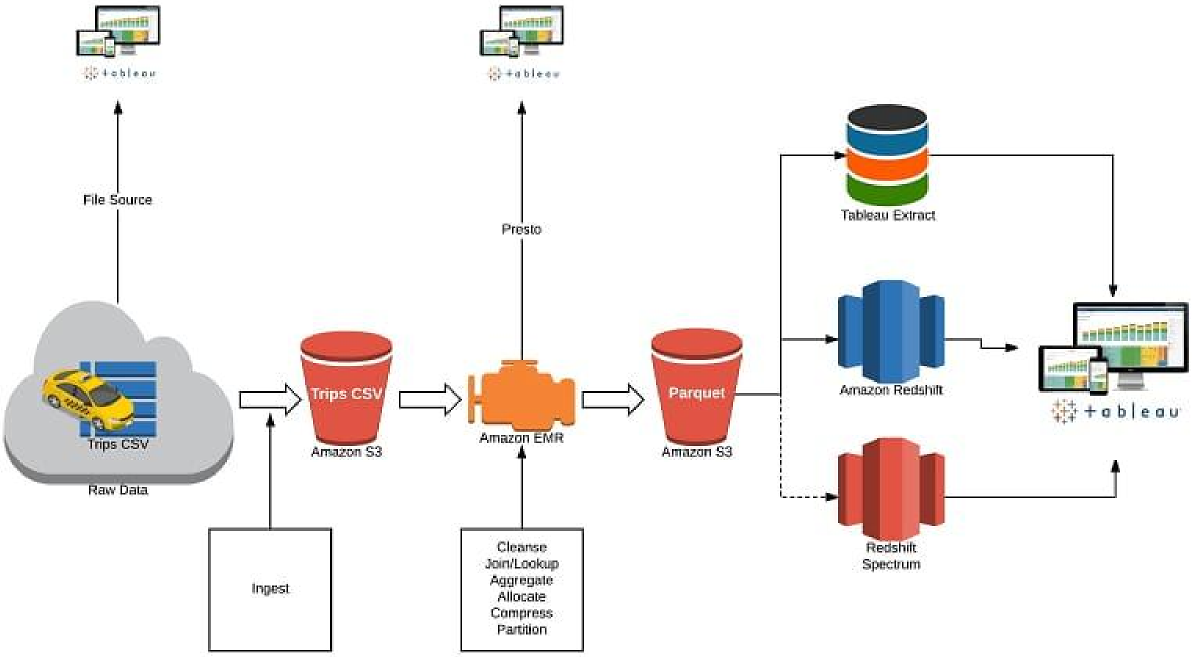

tableau
tableau
File Source
Presto
Tableau Extract
Amazon Redshift
tableau
Trips CSV
Raw Data
Trips CSV
Amazon S3
Amazon EMR
Parquet
Amazon S3
Redshift
Spectrum
Ingest
Cleanse
Join/Lookup
Aggregate
Allocate
Compress
Partition