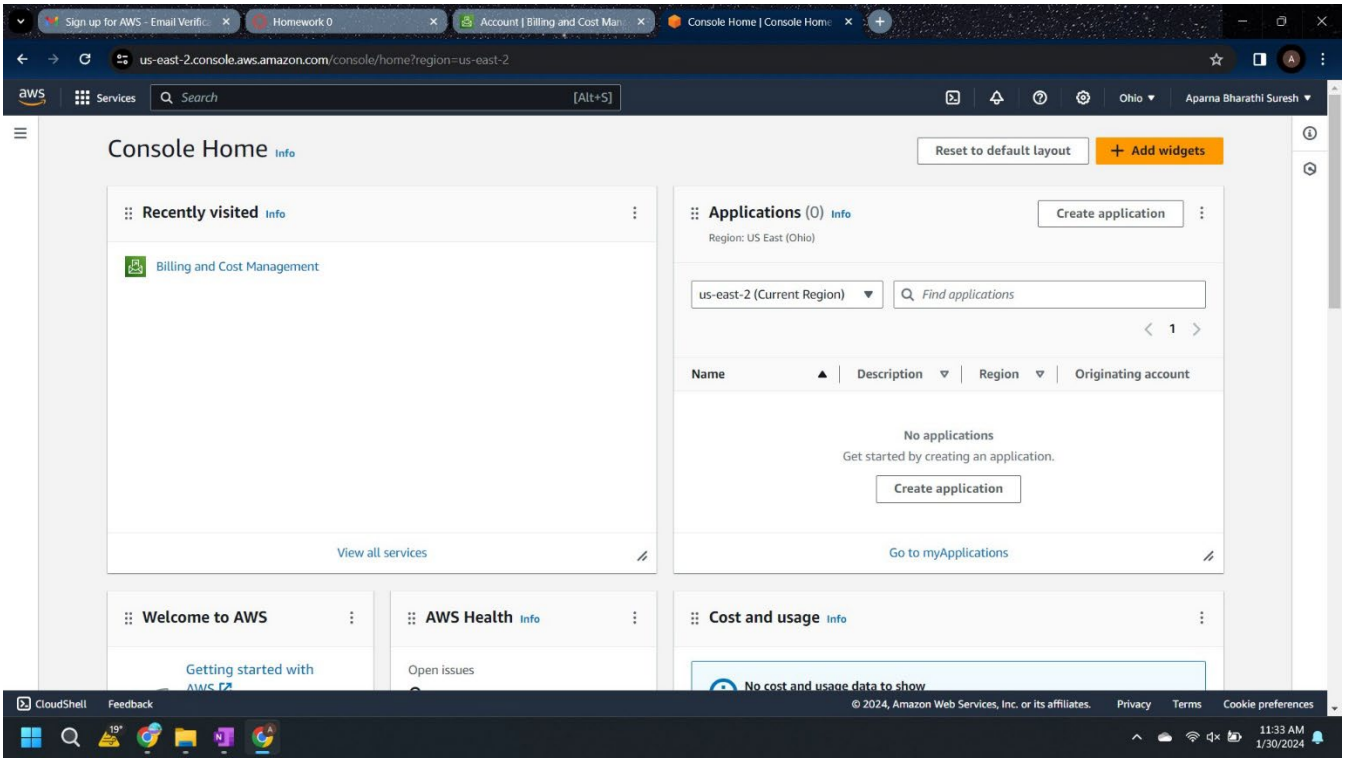
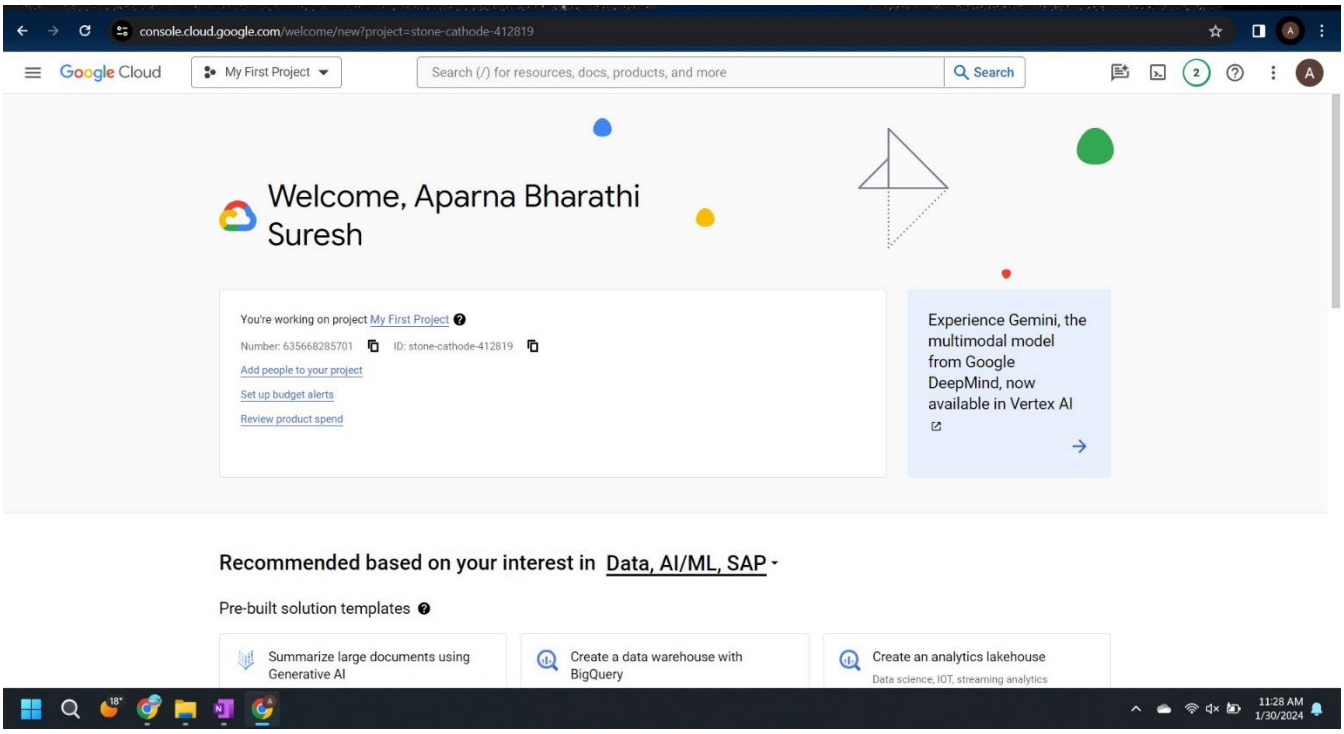
Homework0_AparnaBharathiSuresh_DATA-225

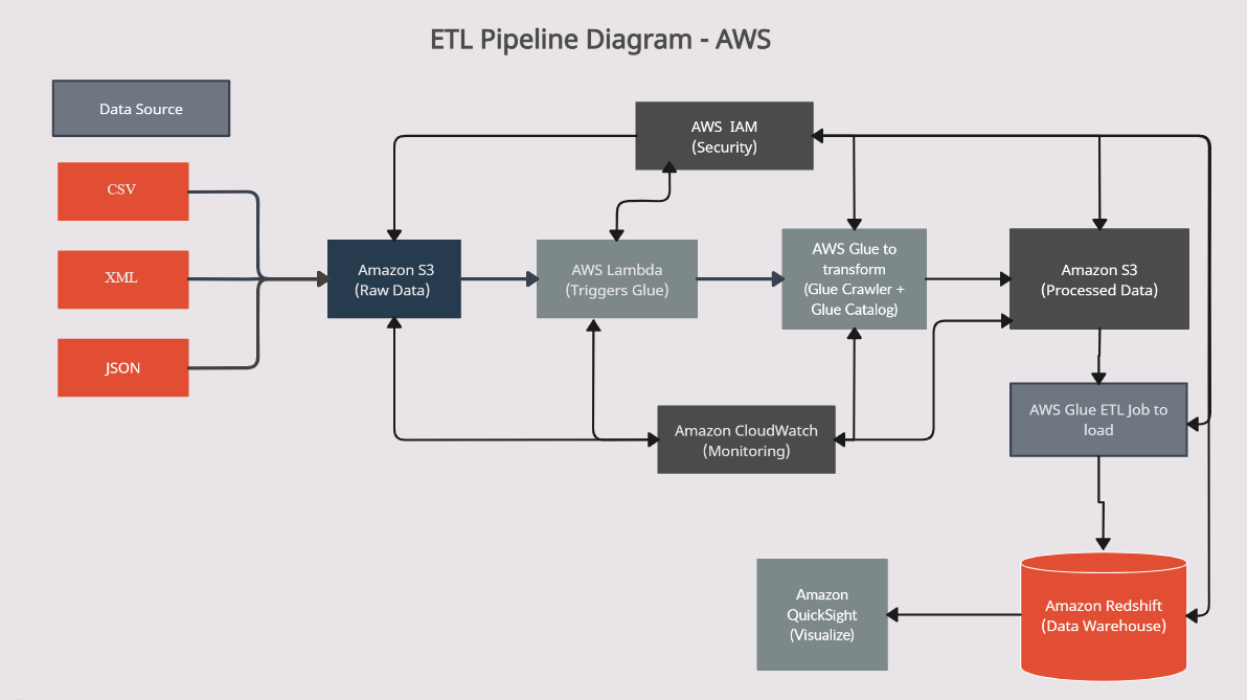
1.AWS Screenshot:



2.GCP Screenshot:



3. ETL Pipeline of AWS:



AWS - ETL Pipeline Description:

Below is the explanation of each service and the reason for choosing it.

1. Amazon Simple Storage Service (Amazon S3):
   - S3 can store and retrieve unlimited amounts of data from anywhere on the web.
   - S3 supports a wide range of data types and formats unlike Dynamo DB and RDS.
   - S3 is compatible with many AWS services, and it integrates with the other AWS services.
   - S3 allows versioning, which allows users to retrieve and restore data.

2. AWS Lambda:
   - Lambda allows users to create event driven architecture. When a file is uploaded in the S3 bucket user can trigger the AWS Glue jobs.
   - Lambda follows a serverless execution model.
   - Lambda scales automatically depending on the number of incoming events without manual intervention.
   - Lambda seamlessly integrates with AWS Glue and other AWS services.

3. AWS Glue:
   - Glue has very powerful and valuable tools like Glue crawler and Glue catalog.

- Glue crawler automates data discovery and metadata generation.
- Glue Catalog is the centralized repository of metadata.
- Glue is easy to use because of its visual interface and automatic code generation compared to AWS Data Pipeline.
- Glue supports different languages, and it performs basic data validation which maintains data quality.

4. Amazon S3(Processed Data):
   - Easy to set up, requires minimal configuration, unlike Amazon Elasticsearch Service which is more complex.
   - Offers low storage cost, unlike Amazon DynamoDB.
   - Handles large datasets efficiently.

5. AWS Glue ETL Job to Load:
   - It is serverless, infrastructure management is not required, unlike AWS Batch.
   - Integrates with AWS services like S3, Redshift, Athena.
   - Provides user friendly interface.

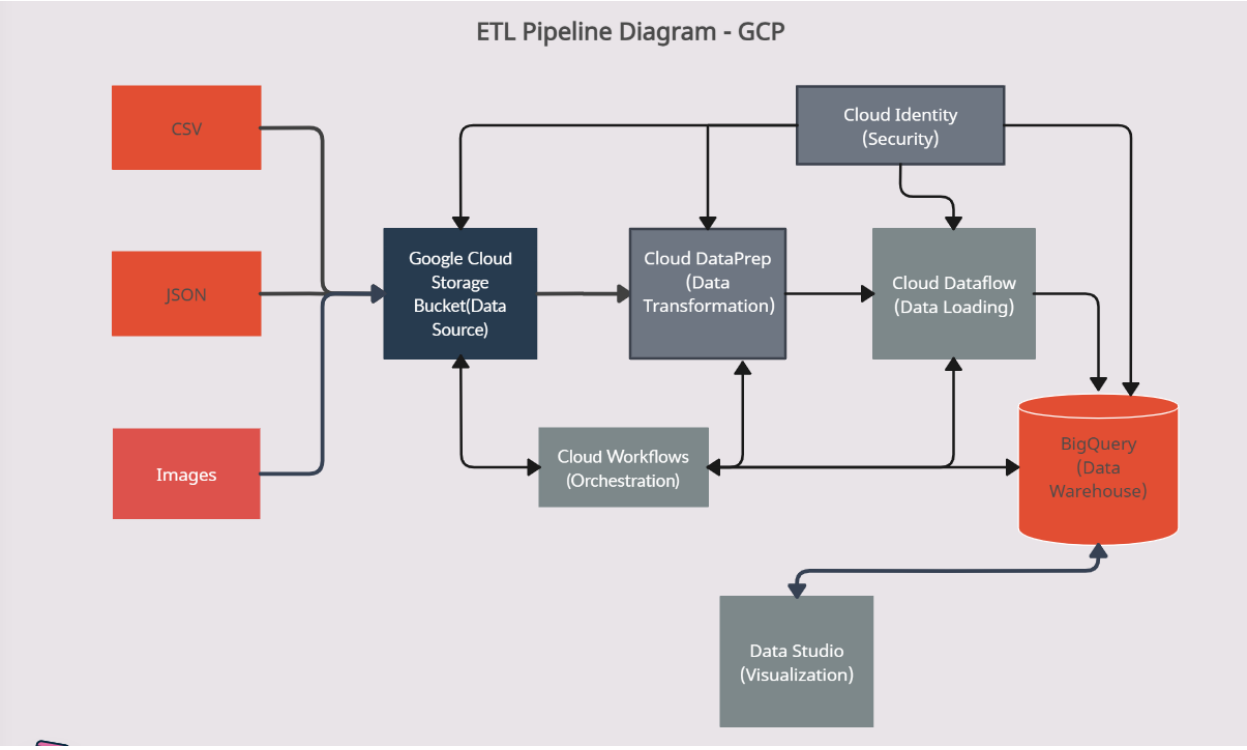
6. AWS Redshift:
   - Redshift has high query performance, that is it fetches data faster compared to Azure Synapse.
   - Redshift is easy to use and more secure compared to Snowflake.
   - Integrates with various data visualization tools.

7. AWS CloudWatch:
   - Monitors metrics and logs for a wide range of AWS services.
   - It's flexible and customizable with filtering options which avoid information overload.
   - Can analyze the resource utilization and ensure efficient resource allocation to improve the performance.

8. AWS IAM:
   - IAM is flexible as it allows role-based access control.
   - It simplifies administration by providing centralized location to manage user identities, groups and their permissions across all AWS services.
   - It provides detailed logs, which can be used for security analysis to identify potential threats.

9. Amazon QuickSight:
   - It's faster and easy to use because of its drag and drop interface.
   - Integrates with other AWS services unlike Amazon Redshift Data Studio which focused only on Redshift data.
   - It offers a wide range of visualization (charts, tables, maps and more). It adheres to AWS security and compliance requirements.

4. ETL Pipeline of GCP:



GCP - ETL Pipeline Description:

Below is the explanation of each service and the reason for choosing it.

1. Google Cloud Storage:
   - Handles very large datasets like petabytes of data, unlike Cloud Pub/Sub which is not suitable for storing large datasets.
   - Supports different types of data unlike Cloud SQL.
   - Follows strict security.

2. Cloud Dataprep:
   - It's easy to use unlike Cloud Dataflow, and it has built-in functions to clean, filter and collect the data.
   - It also validates the data and improves the quality of it.
   - Supports different data formats and easily integrates with other google cloud services.

3. Cloud Dataflow:
   - Supports different programming languages and frameworks. It handles large datasets.
   - Have good documentation and readily available support, reliable.

- Easy integration with other google services, flexible data loading options unlike Cloud Storage Transfer Service.
- It handles the errors without manual intervention, and it automatically retries to load the failed attempts.

4. BigQuery:
   - BigQuery is a powerful data warehouse. It handles petabytes of data effortlessly.
   - It automatically compresses the data which reduces the storage cost.
   - Highly secure and reliable.
   - It has SQL interface which is familiar for data analysts and makes it easy to query data unlike Cloud Storage which is not optimized for querying.

5. Cloud Workflows:
   - It is serverless, provides centralized view for monitoring the flow and it's cost-effective.
   - Google takes care of the configuration and maintenance unlike Cloud Monitoring.
   - Secure and supports various language making it more flexible.

6. Cloud Identity:
   - It follows Multifactor authentication which improves security unlike Cloud IAM which has limited user management and authentication features.
   - It is simple because it manages all user identities and groups from a single platform.
   - It has improved user experience like self-service password management and Single sign-on.

7. Data Studio:
   - Easy to use with drag and drop interface, unlike Cloud BigQuery ML which requires familiarity with BigQuery and machine learning concepts.
   - It provides a wide range of visualization types (charts, tables, maps, and more), and it creates shareable reports.
   - It is cloud based so we can edit and access from any device.