DATA ENGINEERING PROJECT

JUNE 18, 2023

BYTEWISE LIMITED

Contents

Project Statement:

As a client with a social media app, I require a comprehensive solution architecture for cloud-based data management. The goal is to optimize the **app's data storage**, **processing**, and **retrieval capabilities**, **ensuring scalability, reliability**, and **security**. The solution should leverage cloud technologies and services to enable efficient **data handling**, **analysis**, and **integration with other systems**. The architecture should address **data governance**, **data privacy**, and **compliance requirements**, while also considering **performance optimization** and **cost-effectiveness**. Ultimately, the aim is to enhance the overall **user experience**, **streamline data workflows**, and enable future growth and innovation within the app.

What you need to do:

i. Select a cloud platform for the project, providing a reason for the specific choice made.

ii. Using the chosen cloud platform, devise a comprehensive end-to-end solution for the project, including recommendations for storage services, ETL (Extract, Transform, Load) processes, security measures, visualization tools, and more.

iii. Justify the selection of these services over alternatives, highlighting their superior attributes and benefits.

iv. Conduct a thorough cost analysis for the entire project, including a breakdown of expenses and the estimated duration of the project.

Solution:

1. Feature based comparison of AWS, Azure & GCP:

The following table enlists the features of 3 big cloud providers and how each of the following cloud providers provide these features listed below in table 1.1,

Feature	**AWS**	**Azure**	**Google cloud platform**
Define	Amazon web services is a cloud computing platform that manages and maintains hardware and infrastructure reducing the expense and quality.	Microsoft azure is a cloud computing service for building testing and managing applications in cloud	GCP offers a variety of cloud computing services for building, deploying scaling, monitoring and operating a cloud.
Compute services	EC2 (Elastic Compute Cloud).	Azure virtual machine	Google Compute Engine.
Security	AWS security Hub IAM	Azure security center	Cloud security command center
Storage	Amazon S3 Amazon RDS	Azure Blob Storage SQL , MySQL, PostgreSQL	SQL based Cloud storage
Service Integration (is a set of tools and technologies that connects applications, systems, repositories and data in process interchange in real time.)	AWS makes it simpler for users to combine services such as Amazon EC2, S3 and Beanstalk.	Allows customers to effortlessly combine Azure VM, App service and databases.	Users can utilize GCP to combine services such as compute engine, cloud storage and SQL.
Data warehouse	Redshift	SQL warehouse	Big Query

Table 1.1 Comparison of Different Cloud Providers

2. Which one to Choose AWS, Azure or GCP?

I would recommend using ***Amazon Web Services (AWS)*** as the cloud platform for this project as it offers a wide range of services, including storage, computing, networking, analytics.AWS is also highly scalable and reliable, making it a good fit for a social media app that needs to handle large volumes of data.

The following are the reasons as depicted in Table 2.1 that why AWS is suitable,

	AWS	**Azure**	**Google cloud platform**
Why to Choose	1) Dominant market position 2) Extensive, mature offerings 3) Support for large organizations 4) Global reach 5) Flexibility and a wider range of services 6) Considered the best for reliability and security. 7) More computational capacity than Azure and GCP. 8) Provides most services, from networking to robots. 9) It offers detailed documentation for every tool and service.	1) Second largest provider 2) Integration with Microsoft tools and software 3) Broad feature set 4) Hybrid cloud 5) Support for open source 6) Ideal for startups and developers 7) Includes multiple useful Microsoft tools.	1) Designed for cloud-native businesses 2) Commitment to open source and portability 3) Its object storage allows you to store any amount of data and retrieve it as it is. 4) You can automatically manage, scale and deploy containers.
Why not to Choose	It doesn't have free technical support difficult to use for beginners.	Fewer service choices compared to AWS. Less efficient management tools	Limited services compared to AWS and Azure.

Table 2.1 which cloud providers to choose

3. Comprehensive end-to-end solution architecture

Source:

1. **Amazon S3 (Simple Storage Service):**
   - S3 is use for scalable, durable, and cost-effective object storage to store user-generated content such as images, videos, and files.

ETL (Extract, Transform, Load) Processes:

2. **AWS Glue:**
   - Employ Glue for automated ETL processes. It enables data extraction from various sources, transformation using built-in or custom scripts, and loading into target data stores.
   - AWS Glue is a scalable, serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.

Source	Transformation	Target
Data Lake	ETL	DataWarehouse
S3 bucket	AWS glue	Amazon Redshift

Table 2.1 Pipeline

3. **AWS Lambda:**
   - With AWS Lambda, you can run code without provisioning or managing servers.
   - Just upload your code and Lambda takes care of everything required to run and scale your code with high availability.

Data Storage

4. **Amazon Redshift:**
   - Use Redshift as a data warehousing solution for storing and analyzing large datasets. It offers high-performance querying and integrates well with visualization tools.
   - Fast, simple, cost-effective data warehouse service.

Security Measures:

5. AWS Identity and Access Management (IAM):

- AWS Identity and Access Management (IAM) is a web service for securely controlling access to AWS services. With IAM, you can centrally manage users, security credentials such as access keys, and permissions that control which AWS resources users and applications can access.

Visualization and Reporting:

6. Amazon QuickSight:

- Utilize Quick Sight for data visualization, dashboards, and interactive analytics. It integrates well with various data sources, including S3 and Redshift.

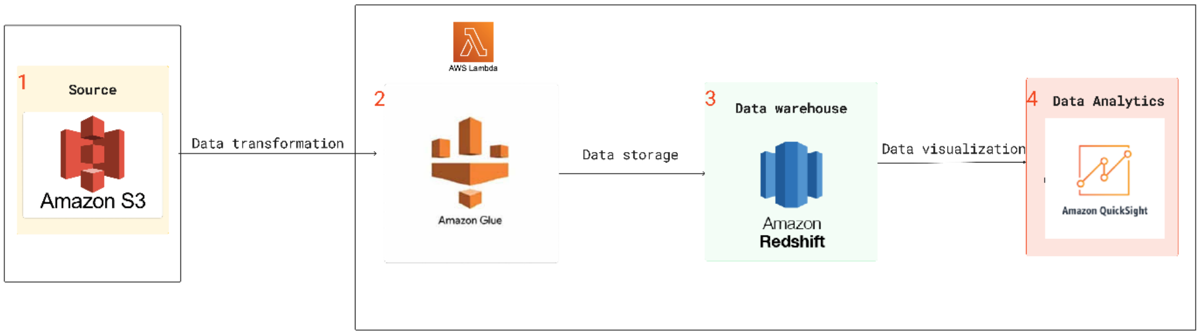
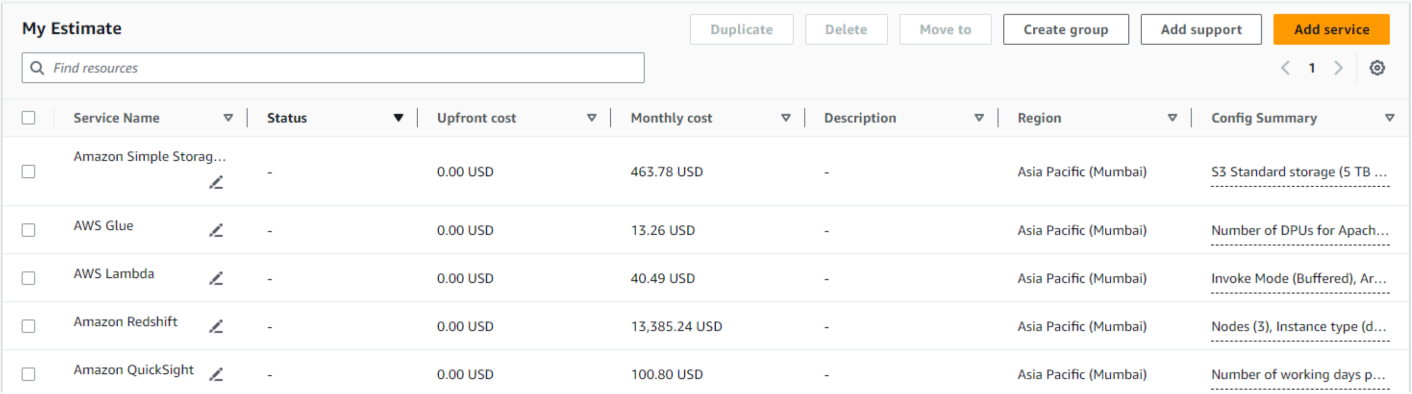


Fig 3.1 AWS Architecture

4. Cost Analysis

1. Breakdown of expenses

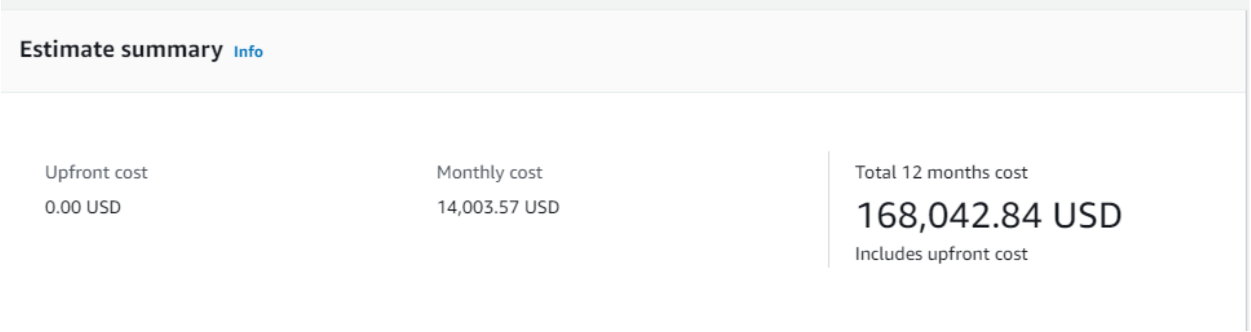
The following table shows which services are used and how cost of each service being used is calculated,



My Estimate

□	Service Name	Status	Upfront cost	Monthly cost	Description	Region	Config Summary
□	Amazon Simple Storag...	-	0.00 USD	463.78 USD	-	Asia Pacific (Mumbai)	S3 Standard storage (5 TB ...
□	AWS Glue	-	0.00 USD	13.26 USD	-	Asia Pacific (Mumbai)	Number of DPUs for Apach...
□	AWS Lambda	-	0.00 USD	40.49 USD	-	Asia Pacific (Mumbai)	Invoke Mode (Buffered), Ar...
□	Amazon Redshift	-	0.00 USD	13,385.24 USD	-	Asia Pacific (Mumbai)	Nodes (3), Instance type (d...
□	Amazon QuickSight	-	0.00 USD	100.80 USD	-	Asia Pacific (Mumbai)	Number of working days p...

Fig 4.1 Cost Analysis through AWS calculator



Estimate summary

Upfront cost
0.00 USD

Monthly cost
14,003.57 USD

Total 12 months cost
168,042.84 USD
Includes upfront cost

Fig 4.2 Monthly & Yearly Cost Analysis through AWS calculator

2. Duration of project

However, it's worth noting that building a comprehensive cloud-based data management solution for a social media app can be a significant undertaking that may span 2-3 months. Factors such as

1. Selecting a cloud platform for the project.
This phase may require 1-2 weeks as we have to decide which cloud platform will best suits the project ensuring cost and budget is within it.

2. Designing the architecture
When cloud platform is selected then we have to gather all the requirements according to the chosen cloud platform and start designing the architecture such as build a schema and a rough sketch of how AWS will do all processing. This can be completed within 1-2 weeks

3. Developing and integrating various components
If the projects require any API then after thorough cost analysis there would be successful integration of components within 2 weeks.

4. Testing

This will take time as if changes would be required so testing will be done again.

5. Deployment

Successful completion of project satisfying all customer requirements will lead to deployment of project.