Data Engineering Fellowship Project

Part(1):

I choose AWS due to following reasons:

- Reliability and High Availability: AWS provides multiple availability zones and fault-tolerant infrastructure, ensuring high availability of the app and its data. It also offers built-in backup and disaster recovery mechanisms to protect against data loss.
- Elasticity is one of the AWS advantages. If you use fewer resources and you don't need the rest of them, then AWS itself shrinks the resources to fit your requirement. That is, upsizing and downsizing of resources are easy here. Also, AWS always lets you know how many resources you are using at the moment.
- Extensive Service Offerings: Aws is the most mature service provider having a wide range of services for storage, data processing, analytics, security, and more. This allows for flexibility in choosing the most appropriate services for different components of the solution architecture.
- Scalability and Elasticity: AWS offers auto-scaling capabilities, allowing the app's data storage and processing resources to scale up or down based on demand. This ensures that the app can handle increased user activity and data growth without compromising performance and user experience with it.
- Highly Performant
- High-performance computing (HPC) is the ability to process a massive amount of data at high speed. AWS offers a high-performance computing service so that the companies need not worry about the speed and increases their productivity.
- Innovation:
  AWS has become the most popular cloud platform as it always keeps innovating itself with the latest tools and technologies for better productivity. It also aids in assessing company's current status, planning for the future, and solving any issues that may have arisen.
- No-commitment service:
  Moreover, AWS is a no-commitment service. It does not ask for any time commitment before you start using AWS benefits, so you can start or stop using it at any time without hassles.
- Easy to use - AWS is user-friendly.
- All major software vendors make their programs available on AWS
- AWS is suitable for general-purpose and memory-optimized instances. Its compute capacity is more than any other service provider.
- Security and Compliance:

AWS has a robust security framework and provides various tools and services for data encryption, identity and access management, and compliance with regulations such as GDPR and HIPAA.

- Cost Optimization:
  AWS provides cost optimization tools and services, such as auto-scaling, reserved instances, and pay-as-you-go, which help in optimizing expenses based on actual usage patterns.
- Integration and Interoperability:
  AWS offers integration with a wide range of third-party services and tools, enabling seamless data integration and workflow automation with other systems.

<---------->

Part(3):

Superior Attributes of AWS EC2:

AWS EC2 (Elastic Compute Cloud) is a scalable virtual server in the cloud that allows users to rent and configure virtual machines for running applications and managing workloads.

Safe

It operates in a secure environment called Amazon Virtual Private Cloud which enables users to work in a safe and stable platform.

Cost Effective

It only gets users to pay for the resources which they choose to perceive, and it consists of various purchasing plans like Reserved Instances, Spot Instances and Demand Instances, which can be chosen as needed.

Reliable

It provides the users with a reliable network where the change of instances can easily and quickly be made. The Service Level Agreement obligation is 99.9% availability for every EC2 region.

Integration with Other Amazon Web Services

It operates perfectly well with the Amazon services such as Amazon DynamoDB, Amazon S3, Amazon SQS, and Amazon RDS. It gives the users a total settlement for query processing, computing, and storage over a vast variety of applications.

Reasons for choosing EC2 over GCP cloud compute and Azure VMs :

I choose AWS EC2 storage because of the following reasons:

1. High Availability and Reliability

Due to the presence of availability zones in different regions, it provides high availability and reliability.

2. Hibernation

In order to save money when a VM does not need to be running AWS is the only cloud provider that has the ability to hibernate a VM by saving from RAM to disk, and then resuming from where the VM stopped when it is started up again like the sleep functionality in desktops. While in azure and GCP you have to switch your VM from a running state to a stopped state and then start it again when needed.

3.Auto Scalability

AWS EC2 automatically scaled up or down the resources according to the requirements of the system or user in no time and hassle. While in GCP autoscaling is not available and you have to manually enabled it. While in azure it is also a hectic process and use a wide range of metrics to trigger scaling events, including host-based metrics, application-level metrics (using the App Insights service) and in-guest VM metrics.

4. Predictive Scaling

AWS has the ability of predictive scaling, based on historic trends analyzed by a machine learning algorithm it will predict that which instance or functionality need how much resources and will managed it according to it. While Azure does not offer any predictive scaling.

<---------->

AWS Redshift:

Amazon Redshift is a fully managed solution that automates the administrative tasks related to maintaining backups, security, configuration, and used to analyze large-scale data storage and many more. It allows customers to query petabytes of semi-structured and structured data using standard SQL queries.

Amazon Redshift Top Features:

Here are some key features offered by Amazon Redshift:

1. For Faster performance it uses faster and sophisticated ML algorithms to predict incoming query run times, and assigns them to the optimal queue for the fastest processing. Moreover, result caching increase sub-second response times for repeat queries. It implement massively parallel processing (MPP) data warehouse architecture to parallelize and distribute SQL operations to take advantage of all available resources.
2. Powerful data warehousing.
3. Direct integration with various AWS services.
4. Automatic concurrency scaling for optimal query performance.
5. Serverless option for cost-effective scaling.
6. Streaming data ingestion for real-time data analysis.
7. High reliability and uptime with automated backups and maintenance.
8. Advanced security, access control, and compliance.

Reasons to Choose Amazon Redshift

I use Redshift because:

- Redshift continuously monitors the health of the cluster, and automatically re-replicates data from failed drives and replaces nodes as necessary for fault tolerance.
- Amazon Redshift automatically and continuously backs up your data to Amazon S3. Redshift can asynchronously replicate your snapshots to S3 in another region for disaster recovery. You can use any system or user snapshot to restore your cluster using the Amazon Web Services Management Console or the Redshift APIs. Your cluster is available as soon as the system metadata has been restored, and you can start running queries while user data is spooled down in the background.
- Its main focus is providing fast and efficient querying and data retrieval capabilities.
- It can integrate with other AWS services, such as Amazon SageMaker, Amazon EMR, and Amazon QuickSight, and other industry-leading tools and experts for performing advanced analytics tasks.
- Regardless of skill level, get started in a few clicks. It is Easy to setup, deploy, & manage.
- AWS Redshift usually permits large-scale database migrations.

<---------->

AWS CloudFront :

AWS CloudFront is a fast content delivery network (CDN) service by Amazon Web Services.

- Its superior attributes include high performance, low latency, and global coverage.
- CloudFront securely delivers static and dynamic content, such as web pages, images, videos, and APIs, to end-users with high availability and scalability.
- It offers edge caching, SSL/TLS encryption, real-time log analysis, and integration with other AWS services for efficient content delivery and improved user experience.

<---------->

AWS QuickSight:

- AWS QuickSight offers advanced features such as machine learning-powered anomaly detection, data preparation, and natural language querying. In terms of data integration and connectivity, AWS QuickSight offers integration with a wide range of data sources, including on-premise and cloud-based databases, as well as big data platforms such as Amazon Redshift, Amazon EMR, and Amazon Athena.
- Amazon QuickSight is meant for simplicity and convenience of use. While in Azure Microsoft Power BI Finding proper documentation for user-submitted visualizations may be difficult, and some may be missing documentation entirely. While GCP Looker lack the advanced visualization capabilities.
- It's small and fast, even on huge datasets with a lot of users. It can be readily applied to a variety of data sources, particularly information from other (AWS) products. While GCP Looker and Microsoft Power BI can consume a large amount of computer resources and may run slowly for some users and is less granular than Amazon Quicksight's model.

- Individuals without the need for an Amazon QuickSight license can see charts for a nominal per-use price, making it a viable solution for firms that don't require their staff to have regular access to BI software.

<---------->

AWS Lambda:

- AWS Lambda supports unlimited functions per project and will allow 1000 executions per account for each region. While Google Cloud Functions, lack in this and just provide a cap of 1000 functions per project, with a maximum of 400 executions.
- AWS is the only provider to offer highly customized concurrency management options, while Azure and GCP are a little vague on how concurrent executions are handled.
- AWS Lambda is the more mature and most popular than the other service providers. While Azure functions has the biggest drawback Vendor-lock. It is very difficult to run code deployed in Azure function outside the azure environment.
- AWS Lambda provides high availability and low latency by offering Provisioned Concurrency, which ensures that functions are initialized and ready to respond to events, cutting down the dreaded cold-start time to mere milliseconds. Azure offers a more complex variety of hosting options, while GCP just offers a one-size-fits-all plan.

<---------->

Amazon Kinesis Data Streams :

Amazon Kinesis is for developers looking to build streaming data applications processing big and small streams of real-time data.

Reason Of Choosing Amazon Kinesis Data Streams :

Amazon Kinesis Data Streams easier to use, set up, and administer.

Amazon Kinesis Data Streams meets the needs of the business better than Google Cloud Dataflow.

Quality of product support is more in Amazon Kinesis Data Streams than Google Cloud Dataflow.

AWS Kinesis is more comprehensive, sophisticated, and scalable than Azure IOT hub.

<---------->

Amazon EMR :

Amazon EMR Amazon EMR (previously called Amazon Elastic MapReduce) is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark , on AWS to process and analyze vast amounts of data.

EMR:

Some of the features offered by Amazon EMR are:

- Amazon EMR enables you to quickly and easily provision as much capacity as you need and add or remove capacity at any time. Deploy multiple clusters or resize a running cluster

- Amazon EMR is designed to reduce the cost of processing large amounts of data. Some of the features that make it low cost include low hourly pricing, Amazon EC2 Spot integration, Amazon EC2 Reserved Instance integration, elasticity, and Amazon S3 integration.
- With Amazon EMR, you can leverage multiple data stores, including Amazon S3, the Hadoop Distributed File System (HDFS), and Amazon DynamoDB. While Azure HDInsight 4.0 doesn't support Apache Storm and ML Services cluster type. Google Cloud Dataproc is also not user friendly and the spark job UI is not easily accessible.

<---------->

Amazon Rekognition:

Amazon Rekognition makes it easy to add image and video analysis to your applications using proven, highly scalable, deep learning technology that requires no machine learning expertise to use. Amazon Rekognition, can identify objects, people, text, scenes, and activities in images and videos, as well as detect any inappropriate content.

It has User-friendly interface which makes it suitable for all types of users.

<---------->

Amazon ElastiCache :

AWS ElastiCache is used to "Deploy, operate, and scale an in-memory cache in the cloud".

- ElastiCache improves the performance of web applications by allowing you to retrieve information from fast, managed, in-memory caches, instead of relying entirely on slower disk-based databases.

<---------->

AWS WAF

AWS WAF (Web Application Firewall) is a managed service by Amazon Web Services that provides protection for web applications against common web exploits and attacks through customizable rules and real-time monitoring.

- AWS WAF is easier to use when assessing the two solutions.
- It meets the needs of business better than Azure Firewall.
- AWS WAF has more feature updates and roadmaps, over Azure Firewall.

<---------->

AWS RDS:

AWS RDS (Relational Database Service) is a managed database service offered by Amazon Web Services that makes it easy to set up, operate, and scale relational databases in the cloud, such as MySQL, PostgreSQL, Oracle, and more.
I use RDS because

- It provides maximum storage.
- Amazon RDS is the only provider with direct support to host a PaaS Oracle database.
- RDS generates automated backups of your DB instance.

- You can use the AWS Management Console, the Amazon RDS Command Line Interface, or simple API calls to access the capabilities of a production-ready relational database in minutes.
- Amazon RDS will make sure that the relational database software powering your deployment stays up to date with the latest patches.
- Amazon RDS is a collection of high-performing managed services that allows you to set up, operate, and scale databases in the cloud with complete flexibility and cost efficiency.
- It is ideal for building web applications with high scalability requirements.
- It has the support of heterogeneous migrations, customers can easily migrate their databases to AWS Cloud and later switch to a different DBMS, if required.
- While Azure SQL Database and Google SQL Database lag in terms of product breadth and migration capabilities as compared to AWS. Also, as these are SQL database, it is suitable for structured data only.

<---------->

Amazon DynamoDB:

Amazon DynamoDB NoSQL database service is ideal for building and delivering apps with consistent single-digit millisecond performance backed with automatic multi-region replication.

I Choose it due to following reasons:

- Other features include built-in security, in-memory caching, continuous backups, data import, and export tools.
- I Choose it due to following reasons:
- It develops software applications with high concurrency and connection requirements.
- It creates media metadata stores with high throughput.
- AWS DynamoDB supports OLTP/OLAP and ensures atomicity and consistency. On the other hand GCP Bigtable is not suitable for OLTP/OLAP. Cosmos DB, also does not have fully featured support for OLAP and OLTP.
- DynamoDB's give performance at very low cost. By enabling transactions give strong consistency and flexibility.
- With Cloud Bigtable user must specify a cluster size of at least three nodes. This is far in excess of what any small or modest-sized application needs, making the service unsuitable for low-activity databases hosting small amounts of data.
- With an SLA of 99.999%, it offers high security and reliability.

<---------->

AWS S3:

AWS S3 is an object level storage service by Amazon Web Services. It provides storage for unstructured data rather than file or volume data. It stores its data as objects within buckets. With AWS S3, you can store any amount of data, create a data backup, or retrieve it at any time from any device.

I choose it due to following reasons:

- **Notification Management:**
  When there are events in your AWS S3 environment, you receive notifications through either Amazon SQS or Amazon SNS. Also, the notifications can be delivered directly to Amazon Lambda to involve functions. AWS S3 notifications are set at bucket level either through the REST API, S3 Console, or using an SDK.
  In Google Cloud Storage, Pub/Sub the notification management doesn't provide the level of flexibility as S3. Configuring alerts isn't as straightforward as in S3.
- **Replication:**
  AWS allow Batch Replication to replicate existing objects to different buckets to improve availability and offers protection against data loss. When replicating objects in AWS S3, you can use an API. On the other hand, Google Cloud Storage doesn't have an API. Also, it lacks the flexibility of S3 CRR.

<---------->

AWS SNS:

AWS SNS (Simple Notification Service) is a fully managed messaging service by Amazon Web Services.

- Its functionalities include high reliability, scalability, and support for multiple protocols (HTTP, email, SMS, mobile push).
- SNS enables pub/sub messaging patterns, integrates well with other AWS services, and provides filtering capabilities for targeted message delivery.
- It simplifies the process of sending notifications and alerts to users, applications, and systems.

AWS SES:

AWS SES (Simple Email Service) is a cloud-based email sending service by Amazon Web Services.

- Its functionalities include high deliverability, scalability, and cost-effectiveness.
- SES enables businesses to send transactional emails, marketing messages, and automated emails.
- It provides a reliable infrastructure for email delivery, integrated email analytics, and flexible API options.
- With SES, organizations can easily manage their email communications, track email performance, and ensure successful delivery to recipients.

AWS SQS:

AWS SQS (Simple Queue Service) is a fully managed message queuing service by Amazon Web Services.

- Its functionalities include high scalability, reliability, and decoupling of components in distributed systems.
- SQS enables asynchronous communication between microservices, decouples the sender and receiver, and ensures message durability.

- It supports both standard and FIFO (First-In-First-Out) queues, provides flexible message handling, and integrates well with other AWS services for building robust and scalable architectures.

<---------->

AWS Security Services Used For Project:

IAM:

Amazon IAM (Identity and Access Management) is a service provided by AWS that enables you to manage user access to AWS resources.

Its functionalities include centralized access control, allowing you to define and manage permissions for users and groups. It supports role-based access control (RBAC), granting temporary permissions to trusted entities.

IAM provides granular permission management, ensuring least privilege access. It integrates with other AWS services, enabling seamless access control across the AWS ecosystem.

IAM enforces security and compliance through features like multi-factor authentication, strong password policies, and access logs.

It is highly scalable and available, accommodating the access control needs of applications and users.

Amazon KMS:

Amazon KMS (Key Management Service) is a managed service by AWS that helps you create and control encryption keys for securing your data.

Its functionalities include providing a secure and scalable solution for key management, offering integration with various AWS services, simplifying the encryption process, and supporting compliance with regulatory requirements.
KMS allows you to manage keys centrally, encrypt data at rest and in transit, and control access to keys through fine-grained permissions. It also provides audit logs for key usage and supports key rotation.
KMS is designed to ensure the confidentiality and integrity of your data while minimizing the operational overhead of key management.

Amazon ACM:

Amazon ACM (AWS Certificate Manager) is a service provided by AWS that simplifies the process of provisioning, managing, and deploying SSL/TLS certificates for secure communication over the internet.

Its functionalities include automated certificate management, seamless integration with AWS services, built-in certificate renewal and deployment, and support for both public and private certificates.

ACM eliminates the need for manual certificate management, reduces the risk of misconfiguration, and helps enforce secure communication practices. It offers free SSL/TLS certificates issued by Amazon's trusted Certificate Authority (CA) and supports certificate validation and monitoring. ACM streamlines the process of securing your applications and websites with SSL/TLS encryption, enhancing security and trustworthiness.

Amazon Inspector:

Amazon Inspector is a security assessment service offered by AWS that helps you identify security vulnerabilities and compliance violations in your applications and infrastructure. Its superior attributes include automated security assessments, continuous monitoring, easy integration with AWS resources, and actionable insights.

- Inspector scans your resources, analyzes them against a predefined set of rules, and generates detailed findings reports. It provides a prioritized list of security issues and offers remediation recommendations.
- Inspector helps you proactively identify and address security risks, ensuring the security and compliance of your applications and infrastructure.
- Its integration with other AWS services streamlines the security assessment process, making it easier to maintain a secure environment.

Part(4):

This social media app is designed to cater to users of all types. It incorporates an automatic autoscaling feature, allowing it to dynamically adjust its compute resources based on demand. The app has the capacity to store up to 10 TB of data per month, accommodating the needs of its growing user base.

To ensure data privacy and compliance with regulations, stringent measures are implemented. The app supports secure login for up to 5 million users, safeguarding their personal information. For monitoring the app's health and performance, Amazon CloudWatch is utilized, utilizing a comprehensive set of seven metrics.

The app facilitates seamless data transfer, allowing up to 1 TB of data to be efficiently exchanged between different AWS services as well as between users and the app. Additionally, to prevent the dissemination of inappropriate content, the app processes

approximately 1 million images per month, effectively filtering out any potentially objectionable or explicit material.

Solution:

Quicksight:

$28/ month for author querying

$0.30/ month for reader querying

S3 Standard:

Storage cost =10,000 x 0.023 per GB = $ 230

Data transfer out cost = 5,000 * $0.09 per GB = $ 450 (5TB = 5,000 GB)

Total monthly cost = Storage cost + Data transfer out cost

= 230+450

= $ 680

Route 53: $0.50* per health check / month

Rekognition: 1,000,000 images X $0.0010/image = $1,000

CloudWatch : charges = $21 per month

Amazon API Gateway: $2.80 per millin

Amazon SNS pricing

Data Transfer IN : $0.00 per GB

Data Transfer OUT : $0.09 per GB

So, 1 million data transfers * $0.50 per million publish requests = $0.50

Data size: 1 million messages * 1 KB per message = 1 million KB = 1,000 GB

1,000 GB * $0.09 per GB = $90

Amazon Athena:

Worker DPU-hours = Number of calculations * DPUs used per calculations * execution time of calculation = 6 calculations * 20 DPUs per calculation * (1/60) hours per calculation = 2.0 DPU-hours

Driver DPU-hours = DPUs used per session * session time = 1 DPUs per session * 1 hours per session = 1.0 DPU-hours

Total DPU-hours = Worker DPU-hours + Driver DPU-hours = 2.0 DPU-hours + 1.0 DPU-hours = 3.0 DPU-hours

App charges = $0.35 per DPU-hour * 3.0 DPU-hours = $1.05

Amazon Cognito:

10 million users

Price per MAU $0.0046*10 = $ 0.046

Amazon EC2

Next 100 TB / Month $0.07 per GB Greater than 150 TB / Month$0.05 per GB

DynamoDB :

On-Demand Throughput Type Price

Write Request Units (WRU) $1.25 per million write request units= 1.25*10 = $ 12.5

Read Request Units (RRU) $0.25 per million read request units= 0.25*20 = $ 5

AWS GLUE

ETL job: Consider an AWS Glue Apache Spark job that runs for 15 minutes and uses 6 DPU. The price of 1 DPU-Hour is $0.44. Since your job ran for 1/4th of an hour and used 6 DPUs, AWS will bill you 6 DPU * 1/4 hour * $0.44

$0.44 * 10 *30 =$ 132

AWS EMR:

m7g.xlarge $0.1632 per hour * 10 *30 = $ 34.2

Amazon ElastiCache

Cache Node Type Memory Price

cache.t4g.micro 2 2 GiB $0.034 per hour x 730 hours = $24.82 per month

AWS Lambda :

$0.20 per 1 million requests.

WAF:

Request $0.60 per 1 million requests (for inspection up to 1500 WCUs and default body size*)

The total estimated cost for developing this project is $1799.The estimated time for the development of this project is about 1 week.

<---------->

Part(2):

Architecture Diagram: