ASSIGNMENT
Understanding Data Engineering and Key Concepts

Question 1

Research and compare two popular cloud service providers and their offerings for data engineering. Analyze the features, pricing models, and scalability options provided by each platform.

Two popular Cloud service providers are;

1) Amazon Web Service

Amazon Web Service is an Amazon company that was launched in 2002. AWS is the most popular cloud service provider in the world. It's the world's most comprehensive and broadly adopted cloud platform, offering over 165 fully- featured services from data centers globally. Millions of customers use this service.

2) Microsoft Azure

Microsoft Azure is one of the fastest-growing clouds among them all. Azure was launched years after the release of AWS and Google Cloud but is still knocking on the door to become the top cloud services provider. Microsoft Azure recently won a $10 billion US government contract.
Microsoft's Azure revenue is expected to grow between $33 billion to $35 billion. This makes Azure one of the most profitable cloud services in the world.

Comparison	**AWS**	**Microsoft Azure**
Features	● Owned by Amazon Company ● Launched in 2002 ● Covers 25 Geographical Regions	● Owned by Microsoft ● Launched in 2010 ● Covers 53 Geographical Regions ● Available in 140

	• Available in 78 zones • Other features like compute, storage, database, analytics,security, • Provide virtual services in the cloud • Scalable storage in the cloud • High percentage managed relational database • Managed Nosql database • Managed relational database service for MySQL, postgreSQL,Oracle, , SQL server	countries • Other features like compute, storage, data management,, security • Managed intelligent SQL in the cloud, • Globally distributed multi model database in the cloud • Power BI embedded • Azure Cosmos DB • •
Pricing Models	Offers pay as you go approach There are multiple ways to pay for Amazon EC2 instances: On-Demand, Savings Plans, Reserved Instances, and Spot Instances. You can also pay for Dedicated Hosts, which provide EC2 instance capacity on physical servers dedicated for your use.	MA offers pay as you go , reserved instances, and spot instances
Scalability	• AWS Auto Scaling is an Amazon service that lets you configure automatic scaling of AWS resources. It increases computing power or storage resources	The Azure storage platform is designed to be massively scalable to meet the data storage and performance needs of modern applications. Data services in the Azure storage platform are: Azure Blob - A massively

	available for applications when loads increase, and reduces it when no longer needed.	scalable object store for text and binary data. Includes support for big data analytics through Data Lake Storage Gen2. Azure Files - Managed file shares for cloud or on-premises deployments. Azure Queue - A messaging store for reliable messaging between application components. Azure Tables - A NoSQL store for schemaless storage of structured data. Azure Disks - Block-level storage volumes for Azure VMs.
DSS (Data Storage Service	• enables users to store and retrieve any amount of data at any time or place, giving developers access to highly scalable, reliable, fast, and inexpensive data storage	Azure storage is divided into three tiers: Basic, Standard, and Premium. The Basic tier offers 1TB of storage for $5/month, the Standard tier offers 10TB of storage for $50/month, and the Premium tier offers 50TB of storage for $2,000/month. Each tier offers different features and benefits.

Question 2

Select a real-world use case that exemplifies each of the 5Vs (Volume, Velocity, Variety, Veracity, Value) of big data. Explain how data engineering techniques can be applied to address the challenges and opportunities associated with each V.

5 vs of Big data in Banking System

Volume

Volume in big data refers to the quantity/size or amount of data to be stored.

For Example

Banks generate terabytes of data everyday to store new customers details and transactions. Financial services during the years handle high volume of data and always have been with the biggest datasets

Banks use Big Data and BI technologies such as Hadoop and RDBMS in all of their processes, changing the face of banking for the better.

Challenges associated with Volume is Storage, while banking structured data are continually growing, the unstructured data is growing faster and is becoming a more important source for customers insights. This increases the need for having unstructured terabyte databases.

Another example is Walmart, which deals with big data. They handle more than 1 million customer transactions every hour,importing more than 2.5 petabytes of data into their database. This is about 167 times the amount of information contained in all the books in the US Library of Congress
To handle all of this data, DE uses data lakes and a warehouse or data management system. DE store it on cloud or use service providers such as Google cloud,AWS
Apache's hadoop splits big data into chunks, saving it across clusters of servers.

Velocity

Velocity refers to the speed at which data is entered into a system and must be processed. It is a concept which deals with the speed of the incoming data from different sources, it is the data is frequently updated and can be quickly analyzed with possibility to real time analyzes

Example Amazon captures every click of the mouse while shoppers are browsing on its website. This happens rapidly.

In the banking sector Hitting a threshold of 100 transactions per minute is easy for a respectable bank.

Variety

It refers to the complexity of data formats. The data can range from structured to unstructured. Big Data technology allows users to analyze not only the structured data we find in the financial sector, but also the more complex unstructured data that is becoming more relevant in order to discover new conclusions and findings

For Example when a telecommunication company like MTN records data on calls to its call center, this data includes both Structured data which conforms to a predefined data model (e.g your customer Id, the timestamps of your call, your service type), and Unstructured data e.g the recording of the call, notes that the call center operator makes during the call, the problem history related to your call

Another example is CCTV audio and video files that are generated at various locations in a city.

Banks do have to deal with huge numbers of various types of data. From transaction details, bank statements which are structured (name, date and amount) and history to credit scores and risk assessment reports.

Veracity

It refers to the quality of the data that is being analyzed or the trustworthiness of data especially if it is obtained from third-party public sources. Veracity burden can rise exponentially with data volumes.

Example is communication with customers that fails to convert to sales due to incorrect customer information. Poor data quality or incorrect customer data can result in the targeting of wrong customers and communications, which ultimately cause a loss in revenue.

Another Example is Social Media like Instagram, facebook etc
It is particularly challenging to verify the truthfulness of posts on social media platform, as we do not always know the posters backgrounds and their intentions,
In fact, detecting fake reviews, fake news, and fake friends is currently an active research area.

Data can only help organizations if it's clean. That is if it's accurate, error-free, reliable , consistent,bias -free, and complete.

Scalable Apache Spark is good for quick queries across data sizes.

Walmart is making sure its data helps with privacy issues, ensuring sensitive details are encrypted while customer contact information is segregated.

Value

Value amounts to how worthy the data is of positively impacting a company's business. Value is by predicting new trends based on the analysis of data, banks and financial services can create value for customers by offering them new services
Banks can apply the results of big data analysis real time and make business decisions accordingly. This can be applied in the following activities

- Discovering the spending pattern of the customers
- Identifying the main channels of transactions (ATM withdrawal, credit/debit card payment
- Splitting the customers into segments according to their profiles
- Product cross-selling based on the customers segmentation
- Fraud management and prevention
- Risk assessment, compliance and reporting
- Customer feedback analysis and application

In Banking system, the size of data is increasing at a rapid rate. Uses of big data in the banking sector are mitigation of risk factors, clarity in business, misuse of debit/credit cards, and money laundering.

Another example
Walmart uses its big data to make its pharmacies more efficient, help it improve store checkout, personalize its shopping experience, manage its supply chain, and optimize product assortment among other ends.

Splunk Enterprise helps businesses analyze data from different points of view and has advanced monitoring features that come at a price.

Question 3

Analyze a real-world business case and propose a hybrid solution that combines OLTP and OLAP systems to meet both transactional and analytical data processing requirements. Justify your solution and discuss potential benefits and challenges.

Example of Hybrid Solution that combines OLTP and OLAP is ATM Machine

OLTP - Online Transaction Processing has day to day transaction data which keeps changing.
In an ATM, transactions happen everyday and data will be stored in the system called OLTP, from the OLTP, the data is sent to the OLAP system. From data stored in the OLAP, we will create the reports.

OLTP system collects the data from day to day transactions in ATM machines. Hence, it contains current or present transactional data. Like withdrawing or depositing of cash, and transferring of cash or making use of the Quickteller for airtime etc. OLTP will be sending this information to the OLAP system. The OLAP system does not contain present data but it contains historical or old data of the customers like Name, date of birth, address etc. OLAP data is being analyzed in order to generate reports.

The transactional engine keeps track of data including the inventory list and products and the registered customers and manages all the purchases.

An analyst can use this online data to understand customer behavior and come up with better strategies for product placement, optimized pricing, discounts, and personalized recommendations as well as to identify products which are in high demand and do proactive inventory refill.

Another example of hybrid solutions that uses the two techniques is Paying of Utility bills for example Electricity.

Customers use different means to pay for electricity via POS machine, Mobile app, etc. Transactions are stored in the OLTP system which collects data from day to day transactions. From OLTP DE request for data stored in OLTP and moved to OLAP system for analysis. The OLAP system already contains customers' history information like Name, address, account no, phone number etc.
From the OLAP system the DE can fetch out customers that do not make payment for electricity for a certain period.

Question 4

Explain the key differences between a traditional relational database, a data warehouse, and a data lake. When would you recommend using each of these storage solutions in a data engineering project?

Database is a storage location that houses structured data. Popular databases are:Oracle, PostgreSQL, MongoDb, Redis.

Data warehouses are large storage locations for data that you accumulate from a wide range of sources. For decades, the foundation for business intelligence and data discovery/storage rested on data warehouses. Their specific, static structures dictate what data analysis you could perform.

Data warehouses are popular with mid- and large-size businesses as a way of sharing data and content across the team- or department-siloed databases. Data warehouses help organizations become more efficient. Organizations that use data warehouses often do so to guide management decisions—all those “data-driven” decisions you always hear about. Popular companies that offer data warehouses include: Snowflake, Yellowbrick, Teradata

A data lake is a large storage repository that holds a huge amount of raw data in its original format until you need it. Data lakes exploit the biggest limitation of data warehouses: their ability to be more flexible. For a coPopular data lake companies are: Hadoop,Azure, Amazon S3

Comparison Table Between Database, Data warehouse, Data Lake

	Traditional Database	Data warehouse	Data lake

Data	Structured	Structured	Raw and unstructured and structured
Processing	Schema-on-write	Schema-on-write	Schema-on-read
Users	Anyone	IT/Business users	Data Scientist
Use cases	Reporting, analysis, automation	Machine learning	Data Science and research

Question 5

Discuss the advantages and challenges of ETL in the context of data engineering.

ETL process improves data quality as data is cleansed before being loaded onto the final repository for further analytics. An automated data processing pipeline is provided to collect and format data without having to pass on data transformation tasks to other tools.

However, ETL is a time-consuming batch operation, which is recommended for building smaller data repositories that do not need to be updated frequently. Other data integration tools like ELT, CDC, and data virtualization can be appropriately used to integrate larger volumes of data that require real-time updating.