1. **Write the definitions for the data terms slides and provide an example use-case for each term.**

**Ans:**

1.Data Catalog: A data catalog is a centralized metadata repository, often implemented using services like AWS Glue, AWS Lake Formation, and Amazon Athena, allowing users to efficiently discover, manage, and analyze data assets across various data sources within their organization.

Test Case: Validate the search functionality of the data catalog by querying metadata attributes such as data source, schema, and tags to locate specific datasets or data entities.

2.Data Anonymization/Data Masking: Data anonymization or data masking involves obscuring sensitive information within datasets to protect the privacy of individuals, typically achieved by replacing identifiable data with fictitious or generalized values while preserving the dataset's utility for analysis or testing purposes.

Test Case: Execute data profiling and statistical analysis to ensure that anonymized data maintains the distribution characteristics and statistical properties of the original dataset.

3.PII Data: Personally Identifiable Information (PII) data refers to any information that can be used to identify an individual, such as names, social security numbers, email addresses, or biometric data, which must be handled with care to comply with privacy regulations.

Test Case: Conduct vulnerability scanning and security assessments to identify potential security gaps or misconfigurations in AWS IAM policies and resource permissions governing access to PII data.

4.Data Democratization: Data democratization refers to the process of providing access to data and analytics capabilities to a broader range of users within an organization, enabling self-service analytics and empowering users to make data-driven decisions without extensive technical expertise.

Test Case: Validate the scalability and performance of data democratization platforms to accommodate increasing user adoption and data query loads during peak usage periods.

5.Data Modeling: Data modeling involves designing the structure and relationships of data entities within a database or data warehouse, typically using techniques such as entity-relationship diagrams or schema design to organize and represent data for efficient storage and retrieval.

Test Case: Execute data validation tests to ensure that data integrity constraints and referential integrity rules defined in the data model are enforced correctly during data ingestion and updates.

6.Dashboards: Dashboards are visual interfaces that display key performance indicators (KPIs), metrics, and other relevant data in a concise and interactive format, allowing users to monitor and analyze data trends, make informed decisions, and track progress towards organizational goals.

Test Case: Conduct usability testing to evaluate the intuitiveness of dashboard layouts, filters, and interactive elements, ensuring a seamless user experience for business stakeholders.

7.Data Ecosystem: A data ecosystem encompasses the interconnected set of tools, services, and processes for managing, analyzing, and deriving value from data across an organization, including data storage, processing, integration, analytics, and governance components.

Test Case: Conduct integration testing to verify interoperability between different components of the data ecosystem, including data lakes, data warehouses, ETL pipelines, and analytics tools.

8.Data Enrichment: Data enrichment involves enhancing existing datasets with additional contextual information or attributes obtained from external sources, such as demographic data, geographic information, or market trends, to improve the quality and relevance of the data for analysis or decision-making.

Test Case: Conduct regression testing to ensure that data enrichment pipelines produce consistent results over time and that changes or updates to enrichment logic do not introduce data quality issues.

9.Data Exchange: Data exchange involves the secure sharing and transfer of data between different systems, applications, or organizations, typically facilitated by APIs, data pipelines, or integration platforms to enable seamless data interoperability and collaboration.

Test Case: Test the compatibility and interoperability of data exchange formats and protocols with external systems and data formats, ensuring seamless integration and data interoperability.

10.Data Extraction: Data extraction refers to the process of retrieving data from various sources, such as databases, files, or APIs, and transferring it to a destination for further processing, analysis, or storage, often performed using ETL (Extract, Transform, Load) tools or services.

Test Case: Conduct performance testing to evaluate the efficiency and throughput of data extraction pipelines, ensuring that data is extracted within specified timeframes and processing SLAs.

11.Data Governance: Data governance encompasses the policies, processes, and controls for ensuring the availability, integrity, security, and compliance of data assets across an organization, including data management, access control, and regulatory compliance measures.

Test Case: Conduct access control testing to verify that role-based access controls (RBAC) are correctly configured to restrict unauthorized access to sensitive data and ensure data privacy and security.

12.Data Ingestion: Data ingestion involves the process of collecting, receiving, and loading data from various sources into a data storage or processing system, such as a data lake, data warehouse, or analytics platform, typically performed using automated pipelines or ingestion tools.

Test Case: Conduct end-to-end testing of data ingestion workflows to verify data integrity and completeness from source systems to target data stores, including error handling and data validation checks.

13.Data Joins: Data joins refer to the operation of combining related datasets based on common keys or attributes to create a unified view of the data, enabling analysis and insights generation across multiple sources, often performed in SQL queries or data processing workflows.

Test Case: Conduct performance testing to evaluate the efficiency and scalability of join operations for large datasets and complex query scenarios, ensuring that queries execute within acceptable timeframes.

14.Data Lineage: Data lineage involves tracking the origins, transformations, and movements of data throughout its lifecycle, providing visibility into how data is created, used, and modified across different systems, processes, and analytical workflows.

Test Case: Conduct metadata validation to ensure consistency and accuracy of data lineage attributes such as data source, transformation steps, and data dependencies stored in the data catalog or metadata repository.

15.Data Mesh: Data mesh is an architectural approach that advocates for decentralizing data ownership and management by treating data as a product, enabling cross-functional teams to manage their own data domains and providing self-serve data infrastructure and tools to enable data democratization and agility.

Test Case: Conduct integration testing to verify the interoperability of data domains and data services within the data mesh ecosystem, ensuring seamless data exchange and collaboration across domains.

16.Data Portability: Data portability refers to the ability to easily move data between different storage systems, services, or regions within the AWS ecosystem, ensuring flexibility and agility in managing data assets.

Test Case: Migrate a database from an on-premises data center to Amazon RDS and validate data integrity post-migration.

17.Data Replication: Data replication involves copying data from one storage location or service to another, typically for purposes such as disaster recovery, high availability, or data distribution across multiple regions or environments.

Test Case: Simulate network failures during data replication between Amazon DynamoDB tables and ensure successful recovery and resumption of replication.

18.Data Privacy: Data privacy involves protecting sensitive information and ensuring compliance with data protection regulations by implementing appropriate security measures, encryption techniques, access controls, and data governance practices to safeguard data against unauthorized access or disclosure.

Test Case: Implement data redaction techniques in Amazon Redshift to mask sensitive information and ensure that masked data remains unreadable.

19.Data Consistency: Data consistency refers to the reliability and accuracy of data across different systems or replicas, ensuring that all copies of the data are synchronized and up to date to maintain data integrity and reliability for applications and users.

Test Case: Conduct stress testing by simulating concurrent read and write operations on a distributed database to ensure data consistency under load.

20.Data Quality: Data quality refers to the level of accuracy, completeness, consistency, and reliability of data stored and processed within the AWS ecosystem, often addressed through data validation, cleansing, enrichment, and monitoring processes to ensure high-quality data for analytics and decision-making.

Test Case: Conduct data profiling and anomaly detection to identify and remediate data quality issues such as missing values or outliers.

21.Data Silo: A data silo refers to a situation where data is stored or managed in isolated or fragmented systems or environments, leading to inefficiencies, duplication, and barriers to data sharing and collaboration across an organization.

Test Case: Evaluate the impact of data silos on data accessibility and collaboration by measuring the time and effort required to access and share data across different.

22.Data Validation: Data validation involves the process of checking and verifying the accuracy, integrity, and compliance of data against predefined rules, standards, or requirements, typically performed using automated validation routines or manual review processes to ensure data quality and reliability.

Test Case: Conduct regression testing to ensure that data validation rules continue to function correctly after system updates, schema changes, or data source modifications.

23.Data Wrangling: Data wrangling refers to the process of preparing and transforming raw or unstructured data into a usable format for analysis or consumption, involving tasks such as cleansing, parsing, aggregating, and structuring data using tools and services like AWS Glue, Data Pipeline, or Amazon EMR.

Test Case: Validate the scalability of data wrangling workflows by processing large volumes of raw data and measuring the performance of data transformation and cleaning operations.

24.Database Schema: A database schema defines the structure, organization, and relationships of data elements within a database, specifying the tables, fields, constraints, and indexes that govern how data is stored, accessed, and manipulated.

Test Case: Conduct performance testing to assess the efficiency of database queries and data retrieval operations against the designed schema.

25.Data Stewardship: Data stewardship involves assigning responsibility and accountability for managing and protecting data assets within an organization, including defining data governance policies, overseeing data usage and access, and ensuring compliance with data privacy regulations and industry standards.

Test Case: Provide data stewardship training and education programs to empower data stewards with the knowledge and skills required to fulfill their roles effectively.

26.EDI Data Standards: Electronic Data Interchange (EDI) data standards define formats, protocols, and syntax for exchanging structured data electronically between different business systems or trading partners, facilitating seamless integration and interoperability of data across supply chains and business processes.

Test Case: Conduct interoperability testing to ensure that EDI messages can be exchanged and processed correctly between different trading partners and EDI systems.

27.Observability: Observability refers to the ability to monitor, measure, and understand the behavior and performance of distributed systems, applications, and services, using metrics, logs, traces, and other telemetry data to identify issues, troubleshoot problems, and optimize resource utilization and user experience.

Test Case: Conduct stress testing to assess the scalability of observability solutions and their ability to handle increased workload and traffic without degradation in performance.

28.Streaming Data: Streaming data refers to continuous and real-time data streams generated by devices, sensors, applications, or online transactions, which are processed, analyzed, and acted upon in near-real-time using streaming data services such as Amazon Kinesis or AWS Lambda to enable use cases such as real-time analytics, monitoring, and alerting.

Test Case: Conduct performance testing to evaluate the throughput, latency, and scalability of streaming data processing systems under varying data volumes and velocity.

29.Data Lake: A data lake is a centralized repository that stores large volumes of structured, semi-structured, and unstructured data in its native format, providing scalable storage and processing capabilities for data analytics, machine learning, and other data-driven applications.

Test Case: Conduct data ingestion testing to verify that data ingestion pipelines can efficiently ingest data from various sources into the data lake while ensuring data integrity and security.

30.Lakehouse Architecture: Lakehouse architecture combines the features and benefits of data lakes and data warehouses, enabling organizations to store and analyze both raw and structured data in a unified platform, leveraging services like AWS Glue, Amazon Redshift, and Apache Spark to support diverse analytics workloads and use cases with improved performance, cost-effectiveness, and ease of management.

Test Case: Conduct concurrency testing to ensure smooth performance under multiple simultaneous queries.

1. **Differentiate between Monolith vs Micro-service Architecture.**

|  |  |  |
| --- | --- | --- |
| **Purpose** | **Monolith** | **Micro-service Architecture** |
| Deployment | Simple and fast deployment of the entire system | Requires distinct resources, making orchestrating the deployment complicated |
| Scalability | It’s hard to maintain and handle new changes, the whole system needs to be redeployed | Each element can be scaled independently without downtime |
| Agility | Not flexible and impossible to adopt new tech, languages, frameworks | Integrate with new technologies to solve business purposes |
| Resiliency | One bug or issue can affect the whole system | A failure in one microservices does nor affect other services |
| Testing | End-to-end testing | Independently components need to be tested individually |
| Security | Communication within a single unit makes data processing secure | Interprocess communication requires API gateways raising issues |
| Development | Impossible to distribute the team’s efforts due to the huge indivisible database | A team of developers can work independently on each component |

1. **Write about the following AWS services.**

* **S3 and S3 Glacier:**
  + **Amazon S3:** S3 or Simple Storage Service is a cloud-based storage service provided by Amazon Web Services (AWS). It's like a giant digital storage warehouse where you can store and retrieve any amount of data from anywhere on the web. S3 is designed to be highly durable, meaning your data is protected against hardware failures, and it's highly scalable, so you can store as much data as you need. It's commonly used to store files, images, videos, backups, and any other type of data that needs to be accessed over the internet.
  + **Amazon S3 Glacier:** Amazon S3 Glacier is a storage service also provided by Amazon Web Services (AWS), but it's designed for storing data that you don't need to access frequently. It's like putting your data in long-term storage. Glacier is optimized for data archiving and backup purposes, offering lower storage costs compared to S3, but retrieval times are slower. Think of it as storing your data in a vault where it's safe and secure, but it may take a bit longer to retrieve when you need it. Glacier is often used for storing data that needs to be kept for compliance reasons or as a backup for disaster recovery scenarios.
  + **Amazon Redshift:** Amazon Redshift is a fully managed data warehouse service provided by Amazon Web Services (AWS). Think of it as a massive, high-performance storage and processing system for your data. Redshift is optimized for running complex queries and analytics on large datasets, making it ideal for businesses that need to analyze big data quickly. It's like having a supercharged database that can handle petabytes of data and deliver fast query results, allowing you to gain valuable insights from your data in real-time.
  + **Amazon RDS (Relational Database Service):** Amazon RDS is a cloud-based service provided by Amazon Web Services (AWS) that makes it easy to set up, operate, and scale relational databases. Essentially, it's like having a team of experts manage your database infrastructure for you. RDS supports several popular database engines like MySQL, PostgreSQL, SQL Server, and Oracle, allowing you to choose the one that best fits your needs. With RDS, you can focus on building your applications without worrying about the underlying database infrastructure, making it a convenient and cost-effective solution for businesses of all sizes.
  + **Amazon DynamoDB:** Amazon DynamoDB is a fully managed NoSQL database service provided by Amazon Web Services (AWS). Unlike traditional relational databases, DynamoDB is designed to scale seamlessly and handle large amounts of data with low latency. It's like having a super-fast and infinitely scalable database that can handle any workload. DynamoDB is ideal for applications that require high performance, such as gaming, advertising, and IoT (Internet of Things) applications. With DynamoDB, you can store and retrieve any amount of data, from any number of users, with predictable performance and single-digit millisecond latency.
  + **Amazon EC2 (Elastic Compute Cloud):** Amazon EC2 is a web service provided by Amazon Web Services (AWS) that allows you to rent virtual servers, known as instances, in the cloud. Think of it as having your own computer in the cloud that you can customize and use however you like. With EC2, you can choose the operating system, configure the hardware specifications (such as CPU, memory, and storage), and install any software you need. EC2 is highly scalable, allowing you to easily increase or decrease the number of instances based on your needs. It's commonly used for hosting websites, running applications, and performing computational tasks in the cloud.
  + **Amazon Lightsail:** Amazon Lightsail is a simplified version of Amazon EC2 that offers an easy way to launch and manage virtual private servers (VPS) in the cloud. It's designed for users who want the simplicity and cost-effectiveness of a pre-configured server without the complexity of managing all the details themselves. Lightsail provides pre-packaged server configurations with fixed monthly pricing, making it easy to predict your costs. It's like renting a small, ready-to-use virtual server that comes with everything you need to get started quickly, such as a specific amount of CPU, memory, storage, and data transfer allowances. Lightsail is ideal for developers, small businesses, and anyone who needs a simple, affordable solution for hosting websites, blogs, or applications in the cloud.
  + **AWS LAMBDA:** AWS Lambda is a serverless computing service provided by Amazon Web Services (AWS). In simple terms, it allows you to run your code without provisioning or managing servers. You just upload your code, and Lambda takes care of everything else, including scaling, monitoring, and maintenance.
* **Amazon SNS:** Amazon Simple Notification Service is a fully managed messaging service that enables message delivery through push notifications, SMS, email, and other endpoints to distributed systems.
* **Amazon Dynamo DB:** DynamoDB is a fully managed NoSQL database service that delivers single-digit millisecond performance at any scale, making it ideal for applications requiring low-latency data access.
* **Amazon Cloudwatch:** CloudWatch provides monitoring and observability for AWS resources and applications, while CloudTrail enables governance, compliance, and audit trail logging by recording AWS API calls and related events.
* **Amazon Sage Maker:** Amazon Sage Maker is a fully managed service that enables developers and data scientists to build, train, and deploy machine learning models at scale.
* **Amazon Step Functions:** AWS Step Functions is a serverless orchestration service that allows you to coordinate multiple AWS services into serverless workflows using visual workflows to automate application workflows and coordinate multiple AWS services.