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 analyses data assets across various data sources within their organization.

**Test case**: Evaluate the integration of the data catalog with data governance tools and policies to enforce access controls, data lineage tracking, and compliance requirements.

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**: Conduct penetration testing to assess the effectiveness of data masking controls in preventing unauthorized access to sensitive information stored in databases or data repositories.

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:** Validate compliance with regulatory requirements such as GDPR, HIPAA, or CCPA by performing data privacy impact assessments and verifying adherence to data handling and disclosure regulations.

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**: Test the integration of self-service analytics tools with AWS data sources such as Amazon Redshift, Amazon Athena, or Amazon RDS to ensure seamless data connectivity and interoperability.

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**: Perform schema evolution tests to verify that database schema changes, such as table alterations or column additions, are applied consistently across development, staging, and production environments.

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**: Test the functionality of dashboard drilldowns and cross-filtering features to ensure users can explore underlying data and uncover insights effectively.

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**: Test the resilience and fault tolerance of data processing pipelines and data stores within the ecosystem by introducing failure scenarios and verifying the effectiveness of recovery mechanisms.

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**: Test the scalability of data enrichment pipelines to handle increasing data volumes and processing loads, ensuring that performance remains within acceptable limits.

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**: Verify the accuracy and completeness of data transferred between systems during data exchange operations by comparing checksums or hash values of source and destination datasets.

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:** Verify the consistency and reliability of data extraction across different data sources, formats, and environments, ensuring that data quality is maintained regardless of source complexity.

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**: Test data retention policies and archival procedures to ensure that data is retained for the required duration and archived securely to comply with legal and regulatory retention requirements.

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**: Validate data transformation and normalization logic applied during data ingestion to ensure consistency and compatibility of data formats and structures across different data sources and target systems.

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**: Validate the compatibility of join operations with different data formats and structures, ensuring that joins can be performed seamlessly across heterogeneous data sources and schemas.

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**: Test the resilience of data lineage tracking mechanisms to handle data pipeline failures, schema changes, or data format updates, ensuring that lineage information remains accurate and up to date.

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**: Validate the effectiveness of data governance and access control mechanisms in ensuring data security, privacy, and compliance within the data mesh ecosystem.

**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**: Ensure compatibility of data formats between AWS Glue and Azure Data Factory by exchanging sample datasets.

**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**: Verify the scalability of data replication solutions by increasing the volume of data and measuring performance under load.

**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**: Conduct data leakage prevention tests to prevent accidental exposure of sensitive data in AWS services such as Amazon S3 buckets.

**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**: Monitor data consistency across distributed caches in Amazon Elastic ache to ensure synchronized data access and updates.

**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**: Evaluate data enrichment techniques such as geocoding or sentiment analysis to enhance the quality and relevance of datasets.

**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**: Validate data governance policies and access controls to ensure that data sharing and collaboration efforts comply with regulatory requirements and organizational policies.

**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**: Validate error handling and reporting mechanisms for data validation failures, ensuring that errors are captured, logged, and communicated to stakeholders for resolution.

**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**: Verify the interoperability of data wrangling tools with various data sources and formats, ensuring that data can be imported, transformed, and exported seamlessly.

**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**: Test schema migration and evolution processes to ensure that changes to the database schema can be applied seamlessly without disrupting ongoing operations.**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:** Validate the effectiveness of data stewardship tools and platforms in supporting data stewardship activities, including data profiling, data quality monitoring, and data remediation.

**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**: Verify the accuracy of EDI data translations and conversions by comparing translated data with source data and validating data mappings and business rules.

**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**: Evaluate the accuracy and completeness of telemetry data collected by observability tools by comparing it against ground truth data and expected system behavior.

**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:** Validate data enrichment and transformation processes applied to streaming data to enrich and contextualize raw data streams with additional metadata or derived attributes.

**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:** Assess resource utilization during real-time data querying.

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

| **Aspect** | **Monolithic Architecture** | **Microservice Architecture** |
| --- | --- | --- |
| **Architecture** | Single-tier architecture | Multi-tier architecture |
| **Size** | Large, all components tightly coupled | Small, loosely coupled components |
| **Deployment** | Deployed as a single unit | Individual services can be deployed independently |
| **Scalability** | Horizontal scaling can be challenging | Easier to scale horizontally |
| **Development** | Development is simpler initially | Complex due to managing multiple services |
| **Technology** | Limited technology choices | Freedom to choose the best technology for each service |
| **Fault Tolerance** | Entire application may fail if a part fails | Individual services can fail without affecting others |
| **Maintenance** | Easier to maintain due to its simplicity | Requires more effort to manage multiple services |
| **Flexibility** | Less flexible as all components are tightly coupled | More flexible as components can be developed, deployed, and scaled independently |
| **Communication** | Communication between components is faster | Communication may be slower due to network calls |

**3) AWS services:**

**Amazon S3 (Simple Storage Service):**

Amazon 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.

**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.

Think of Lambda as a magic box where you can put your code, and whenever something triggers it (like an event or an API call), Lambda automatically runs your code in response. It's like having a team of invisible elves ready to execute your commands whenever you need them, without you having to worry about the underlying infrastructure.

Overall, Lambda simplifies the process of building and deploying applications, allowing developers to focus on writing code without getting bogged down by server management tasks.

**Amazon SNS (Simple Notification Service):**

Amazon SNS is a messaging service provided by Amazon Web Services (AWS) that enables you to send notifications to your applications or users. Think of it as a messaging system that helps you send messages or alerts to various endpoints, such as mobile devices, email addresses, or other applications.

In simpler terms, Amazon SNS is like a broadcast system that allows you to send messages to multiple recipients at once. You can create topics (channels) for different types of messages and subscribe endpoints (devices or applications) to these topics. When you publish a message to a topic, Amazon SNS delivers the message to all subscribed endpoints.

For example, if you have a mobile app and you want to notify users about new updates or events, you can use Amazon SNS to send push notifications to their devices. Similarly, if you have an email newsletter and you want to notify subscribers about new content, you can use Amazon SNS to send email notifications to their email addresses.

Overall, Amazon SNS simplifies the process of sending notifications, allowing you to reach your audience quickly and efficiently across different channels.

**DynamoDB:**

DynamoDB is a fully managed NoSQL database service provided by Amazon Web Services (AWS). In simple terms, it's like a super-fast, highly reliable digital filing cabinet for storing and retrieving any amount of data, from anywhere in the world.

Imagine you have a giant digital notebook where you can write down information about anything you want - like customer profiles, product details, or game scores. DynamoDB is like that notebook, but it's incredibly fast and can handle a massive amount of data.

Overall, DynamoDB makes it easy to build fast, reliable, and scalable applications without worrying about the underlying infrastructure. It's like having a digital assistant that takes care of all your data storage needs, so you can focus on building awesome applications.

**Amazon CloudWatch:**

Amazon CloudWatch is a monitoring and observability service provided by Amazon Web Services (AWS). In simple terms, it's like a digital dashboard that helps you keep an eye on your AWS resources and applications.

Imagine you have a control room with screens showing the status of all your systems - that's CloudWatch. It collects and displays metrics (data points) about your AWS resources, such as CPU usage, disk space, and network traffic. You can set up alarms to notify you if any metrics exceed certain thresholds, helping you quickly identify and respond to issues.

Overall, CloudWatch helps you monitor, troubleshoot, and optimize your AWS infrastructure, ensuring your applications run smoothly and efficiently.

**AWS CloudTrail**:

AWS CloudTrail is a logging service provided by Amazon Web Services (AWS). In simple terms, it's like a digital trail that records every action taken on your AWS account.

Imagine you have a detective following your every move and writing down everything you do - that's CloudTrail. It keeps track of who did what, when they did it, and from where. This includes actions like launching instances, creating buckets, and modifying security groups.

Overall, CloudTrail helps you maintain security, track changes, and ensure accountability in your AWS environment.

**Amazon Sage maker:**

**Sage maker:** Amazon Sage maker is a machine learning service provided by Amazon Web Services (AWS). In simple terms, it's like having a virtual workshop where you can build, train, and deploy machine learning models without needing to be an expert in machine learning or managing infrastructure.

Imagine you want to build a model that can predict whether a customer will buy a product based on their past behavior. With Sage Maker, you can upload your data, choose a machine learning algorithm, and let Sage Maker handle the rest. It automatically scales your computer resources, optimizes your model's performance, and provides tools to monitor and analyze its behavior.

Overall, Sage Maker simplifies the process of building, training, and deploying machine learning models, making it accessible to developers and businesses of all sizes. It's like having a team of machine learning experts at your fingertips, ready to help you bring your ideas to life.

**AWS Step Functions:**

AWS Step Functions is a serverless orchestration service provided by Amazon Web Services (AWS). In simple terms, it's like a conductor that helps you coordinate and automate the execution of multiple tasks or processes in your applications.

Imagine you have a series of steps that need to be performed in a specific order to complete a task - that's where Step Functions comes in. You define these steps as a state machine, which is like a flowchart that describes the sequence of tasks and the conditions for transitioning between them.

With Step Functions, you can build complex workflows for a variety of use cases, such as data processing pipelines, order fulfillment systems, or workflow automation. It helps you streamline your application logic, improve reliability, and scale your workflows as needed.

Overall, AWS Step Functions simplifies the process of orchestrating and automating complex tasks in your applications, allowing you to focus on building great experiences for your users.