AWS Big DATA Services

14 March 2024 14:19

The Three Vs of Big Data:

- Volume: Refers to the sheer amount of data, often in terabytes, petabytes, or exabytes.
- Variety: Encompasses structured, semi-structured, and unstructured data from various sources and formats.
- Velocity: Describes the speed at which data is generated, processed, and analyzed, often in real-time or near-real-time.

Amazon Redshift:

- Fully managed data warehouse service in the cloud capable of handling massive amounts of data, up to 16 petabytes.
- Based on the PostgreSQL database engine, but tailored for analytical and data warehousing tasks.
- Offers high performance, up to 10 times faster than other cloud-based data warehouse options, due to column-based data storage.
- Supports SQL queries and integrates with familiar BI tools.

High Availability, Snapshots, and Disaster Recovery:

- Redshift now supports multi-availability zone deployments for enhanced availability.
- Snapshots provide point-in-time recovery and can be automated or manually initiated, stored in S3.
- Currently, there's no option to convert between Single-AZ and Multi-AZ deployments, so plan architecture accordingly.

Redshift Spectrum and Enhanced VPC Routing:

- Redshift Spectrum enables querying and retrieving data from S3 without loading it into Redshift tables, leveraging massive parallelism.
- Enhanced VPC Routing directs copy and unload operations through your VPC, enhancing security and enabling VPC features like Endpoints and Flow Logs.

Exam Tips:

- Use Redshift for data warehousing, not as a direct replacement for RDS.
- Consider Multi-AZ deployments for enhanced availability.
- Utilize Redshift Spectrum for efficient querying of data in S3.
- Remember to create snapshots for point-in-time recovery and restoration to other regions.

Overall, understanding Redshift's capabilities and best practices is crucial for AWS Solutions Architect exam preparation.

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

ETL (Extract, Transform, Load):

- ETL processes involve systematically extracting data from various sources, transforming it to meet specific business requirements, and loading it into a target database or data warehouse.
- ETL processes are critical for data management and analysis, enabling valuable insights for business decisions.

Amazon EMR (Elastic MapReduce):

- EMR is a managed service by AWS designed to simplify the management of infrastructure and execution of ETL processes in the cloud.
- It's a big data platform that supports popular open-source tools like Spark, Hive, and HBase for data processing.
- EMR integrates with various AWS services and storage options for efficient data processing.

EMR Storage:

- Three types of storage within EMR: Hadoop Distributed File System (HDFS), EMR File System (EMRFS), and local file system.
- HDFS is scalable and distributed, EMRFS enables direct access to S3 data, and local file system is instance store volumes.

Clusters and Nodes:

- Clusters are groups of EC2 instances within EMR, with nodes serving different purposes: primary, core, and task nodes.
- Core nodes store data and run tasks, while task nodes only run tasks without storing data and are often used with spot instances.

Purchasing Options and Cluster Types:

- EMR instances can be on-demand, reserved, or spot instances, with spot instances being the cheapest option.
- Clusters can be long-running or transient (temporary), with transient clusters often referred to as transient clusters in AWS.

Architecture and Connectivity:

- EMR clusters need access to the EMR service and S3, which can be provided through internet access or VPC endpoint access.
- Leveraging VPC endpoints for S3 access is recommended to avoid data transfer costs.

Exam Tips:

- EMR is suitable for processing vast amounts of data and is commonly used for ETL processes, web indexing, machine learning, and genomics projects.
- Look out for scenarios involving Hive, Spark, HBase, Presto, Apache Hadoop, and Apache Spark, as they often indicate the use of EMR.
- Understand the different storage options within EMR, especially HDFS and EMRFS for accessing S3.

Overall, understanding EMR's capabilities, storage options, cluster types, and connectivity is essential for AWS Solutions Architect exam preparation

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

What is Kinesis?

- Kinesis is a service provided by AWS that facilitates real-time or near-real-time data streaming and processing.
- It acts as a highway for moving large volumes of data from one point to another in real-time.

Kinesis Data Streams:

- Kinesis Data Streams is a real-time data streaming service that offers high-throughput, low-latency data delivery.
- It requires setting up producers, Kinesis Data Streams, and consumers, which involves managing shards and scaling resources.

Kinesis Data Firehose:

- Kinesis Data Firehose is a simplified version of Kinesis Data Streams that handles scaling and data delivery to predefined endpoints like S3, Redshift, Elasticsearch, etc.
- Data Firehose is more managed and requires less configuration compared to Data Streams.

Kinesis Data Analytics:

- Kinesis Data Analytics allows processing and analyzing streaming data using standard SQL queries.
- It can be seamlessly integrated with both Data Streams and Data Firehose for real-time data analysis.

Comparison with SQS:

- SQS (Simple Queue Service) is simpler and easier to set up compared to Kinesis.
- Kinesis is preferred for real-time message delivery and big data applications, while SQS is suitable for asynchronous messaging with no real-time requirements.

Exam Tips:

- If real-time message delivery is required, choose Kinesis over SQS.
- For near real-time scenarios, consider Kinesis Data Firehose; for real-time, choose Kinesis Data Streams.
- Use Kinesis Data Analytics for processing streaming data with SQL queries.
- Data Firehose automatically scales, whereas Data Streams requires manual scaling with shards.
- Always look for Kinesis options in scenarios involving real-time data streaming and processing.
  Overall, understanding the capabilities and differences between Kinesis services is essential for building real-time data processing solutions on AWS.

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

Amazon Athena:

- Athena is an interactive query service that allows analyzing data stored in Amazon S3 using standard SQL queries.
- It enables querying data directly from S3 without the need to load it into a traditional database.
- Athena is serverless, fully managed by AWS, and requires minimal configuration.
- It is ideal for querying structured, semi-structured, and unstructured data stored in S3 buckets.
- Use cases include BI analytics, log analysis, and querying large datasets without the need for complex ETL processes.
  AWS Glue:
- Glue is a serverless data integration service that automates the extract, transform, and load (ETL) process.
- It allows creating and managing data catalogs, discovering data, and generating ETL code without the need for managing infrastructure.
- Glue simplifies the process of transforming raw data into a structured format suitable for analysis.
- It integrates seamlessly with other AWS services like Athena, Redshift, and S3 to enable end-to-end data processing pipelines.
- Glue is particularly useful for handling ETL tasks for big data and data lake architectures.
  Using Athena and Glue Together:
- Glue can be used to structure and catalog data stored in S3, making it queryable by Athena.
- Athena can then execute SQL queries directly on the structured data cataloged by Glue, enabling real-time data analysis without the need for a traditional database.
- Together, Athena and Glue provide a powerful solution for querying and analyzing data stored in S3 buckets.
  Exam Tips:
- If you encounter scenarios requiring serverless SQL querying of data stored in S3, think of using Athena.
- Both Athena and Glue are serverless and fully managed by AWS, requiring minimal configuration.
- Athena is used for serverless SQL querying, while Glue is used for serverless ETL processes.
- Understanding the integration between Athena and Glue is important for building end-to-end data processing pipelines on AWS.
- Always keep in mind the high-level functionalities of Athena and Glue when tackling exam questions involving data analytics and ETL processes.
  Overall, Athena and Glue offer powerful capabilities for querying and transforming data stored in Amazon S3, making them essential tools for building data analytics solutions on AWS

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

Amazon QuickSight:

- QuickSight is a fully managed, serverless business intelligence and data visualization

service provided by AWS.
- It simplifies the process of creating interactive dashboards, sharing insights, and collaborating on data analysis.
- QuickSight integrates seamlessly with various AWS services such as Amazon RDS, Aurora, Athena, and S3, enabling easy access to data stored in these platforms.
- The service offers SPICE (Super fast Parallel In-memory Calculation Engine), an in-memory engine that accelerates data analysis and supports advanced calculations.
- QuickSight pricing is based on a per-session and per-user model, providing flexibility for different usage scenarios.
- For enhanced security, the enterprise edition of QuickSight offers features like Column-Level Security (CLS) to safeguard sensitive data.

Dashboards, Users, and Groups:
- QuickSight allows creating users and groups within its ecosystem for streamlined user management.
- Users can create tailored dashboards with stored configurations and filters to meet their specific requirements.
- Dashboards and analysis results can be shared with specific users and groups, facilitating collaboration and information sharing.

Architecture Example:
- A high-level architecture example involves storing data in Amazon S3 and using Glue crawlers to catalog the data and create data catalogs.
- Redshift Spectrum and Amazon Athena can be integrated with the data catalog to enable querying and analysis of data stored in S3.
- QuickSight seamlessly integrates with Athena, allowing users to build dashboards based on Athena queries for data visualization and analysis.

Exam Tips:
- Amazon QuickSight is the go-to solution for business intelligence and data visualization needs.
- Remember to understand the integration of QuickSight with various AWS services like RDS, Aurora, Athena, and S3.
- Familiarize yourself with the architecture example involving S3, Glue, Redshift Spectrum, Athena, and QuickSight.
- Understand the pricing model of QuickSight and the advanced features available in the enterprise edition.
- SPICE (Super fast Parallel In-memory Calculation Engine) is a key component powering QuickSight's analytics capabilities.

Overall, Amazon QuickSight offers powerful capabilities for visualizing data and gaining insights from various data sources stored in AWS.

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

AWS Data Pipeline:
- AWS Data Pipeline is a managed extract, transform, and load (ETL) service used for automating the movement and transformation of data.
- It allows users to define data-driven workflows, automate tasks, and create dependencies between tasks and activities.
- Data Pipeline enforces specified business logic and manages compute resources needed for data processing.
- The service integrates seamlessly with various AWS storage services such as Amazon DynamoDB, RDS, Redshift, and S3, as well as compute services like EC2 and EMR.
- Users can configure notifications via Amazon SNS for both successful and failed tasks.
- Data Pipeline supports scheduling of tasks, allowing users to define when specific activities should be performed.

Components of AWS Data Pipeline:

- Pipeline definition: Specifies the business logic of data management workflows.
- Managed compute: Automatically creates and manages EC2 instances for performing activities.
- Task runners: EC2 instances that pull tasks from the pipeline and execute them.
- Data nodes: Define the locations and types of data input and output for the pipeline.
- Activities: Pipeline components that define the work to be performed.
  Popular Use Cases:
- Processing data in Amazon EMR using Hadoop streaming.
- Importing or exporting data from DynamoDB tables.
- Copying CSV files or data between S3 buckets.
- Exporting RDS data to Amazon S3.
- Copying data to Amazon Redshift for analysis.
  Example Diagram:
- Exporting MySQL data to Amazon S3 to generate reports.
- Utilizes task runners to authenticate into the database, export data to S3, and generate reports in Amazon EMR.
- Tasks can be scheduled to run daily, weekly, or as needed.
  Exam Tips:
- Understand that AWS Data Pipeline is a managed ETL service used for automating data movement and transformation.
- Know its key features, including data-driven workflows, automatic retries, integration with AWS storage and compute services, and support for notifications.
- Look out for keywords related to managed ETL services, automatic retries, and data-driven workflows in exam questions.
  Overall, AWS Data Pipeline provides a powerful solution for automating data processing tasks and creating complex data workflows in the AWS ecosystem

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

Amazon MSK Overview:

- Amazon MSK is a fully managed service for Apache Kafka, used to build and run applications that process streaming data.
- It provides control-plane operations for cluster management, allowing users to focus on data-plane operations for producing and consuming data.
- Amazon MSK is compatible with existing Apache Kafka applications, tools, and plugins.
  Important Components and Concepts:
- Broker nodes: Specify the number of broker nodes per Availability Zone during cluster creation.
- ZooKeeper nodes: Automatically created for you.
- Producers, consumers, and topics: Used for Kafka data-plane operations.
- Flexible cluster operations: Allow manual control or automation of cluster operations via console, CLI, or SDKs.
  Resiliency in Amazon MSK:
- Automatic recovery: Detects and recovers from common failure scenarios with minimal impact.
- Mitigation of broker failures: Automatically replaces unhealthy nodes and reuses storage from older brokers to reduce data needing replication.
- Low impact time: Recovery is automated, resulting in minimal downtime.
  Good Things to Know:
- MSK Serverless: Offers serverless cluster management compatible with Apache Kafka.
- MSK Connect: Allows easy streaming of data to and from Apache Kafka clusters.
  Security and Logging:
- Integration with Amazon KMS: Provides server-side encryption for data at rest by

default.

- TLS 1.2 encryption: Used for all in-transit communications between brokers in the cluster.
- Logging: Broker logs can be delivered to services like Amazon CloudWatch, S3, or Kinesis Data Firehose. API calls are logged to AWS CloudTrail.

Exam Tips:

- Amazon MSK is a fully managed service for Apache Kafka, handling control-plane operations.
- It supports automatic recovery from common failure scenarios and integrates with AWS security services for encryption and logging.
- Keep Amazon MSK in mind when dealing with Apache Kafka-related scenarios on the exam.

Overall, Amazon MSK provides a robust solution for managing Apache Kafka clusters, ensuring high availability, scalability, and security for streaming data applications.

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

1. What is OpenSearch?: OpenSearch is a managed service for running search and analytics engines. It's the successor to Amazon Elasticsearch Service.
2. Service Features:
   - Quick analysis: Ingest, search, and analyze data within clusters.
   - Scalability: Easily scale cluster infrastructure while running OpenSearch services.
   - Integration with IAM, VPC, security groups, encryption, etc.
   - Stability: Multi-AZ capable with automated snapshots.
   - Flexibility: Supports SQL for business intelligence applications.
   - Integrations: Easily integrates with services like CloudWatch, CloudTrail, S3, and Kinesis.
3. Example Diagram:
   - Data sources on the left, feeding into OpenSearch input.
   - Analytics and processing performed within OpenSearch.
   - Output for various purposes like application monitoring or real-time insights.
4. Exam Tips:
   - OpenSearch is commonly used for logging solutions, log file analytics, and business intelligence reports.
   - Understand OpenSearch as a managed analytics and visualization service.
   - Note that AWS may still reference its predecessor Elasticsearch, but the concepts remain the same.

Overall, OpenSearch is a versatile service for analyzing and visualizing data, particularly useful for log analysis and monitoring solutions.

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>

1. Four Questions to Ask Yourself:
   - What type of database is suitable for the scenario?
   - How much data needs to be manipulated and analyzed?
   - Is serverless a requirement?
   - How can costs be optimized?
2. Redshift and EMR:
   - Redshift is for large-scale data warehousing and analytics, not a replacement for RDS.
   - Supports both single and multi-AZ deployments.
   - EMR is built on EC2 instances, allowing for cost-saving measures like savings plans and reserved instances.
3. Kinesis, Athena, and Glue:
   - Kinesis is for real-time streaming data.
   - Athena for serverless SQL queries on data stored in S3.
   - Glue is a serverless ETL service, integrates well with Athena for querying.
4. QuickSight, OpenSearch, and Elasticsearch:
   - QuickSight for visualizing data and creating dashboards.
   - OpenSearch (successor to Elasticsearch) for analyzing files and log data.

- AWS phasing out Elasticsearch, but important to understand the purpose.

5. Data Pipeline:
   - Managed ETL service for automating data movement and transformation.
   - Integrates with various AWS storage and compute services.
6. Amazon MSK:
   - Managed service for Apache Kafka streaming applications.
   - Handles control-plane operations; users manage data-plane operations.
   - Supports pushing broker logs to CloudWatch, S3, or Kinesis Data Firehose.
   - API calls logged to CloudTrail.

   Remembering these tips can help you identify the most appropriate services and solutions for different big data scenarios on the exam.

From <https://chat.openai.com/c/d6fe90a1-a00b-4374-9bfc-54c5891880f3>