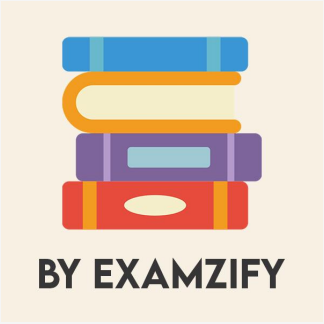
AWS Data Analytics Practice Test (Sample)

Study Guide



Everything you need from our exam experts!

Questions

1. **What is the optimal data source for Amazon QuickSight when needing to create heat maps for visualizing sales data in S3?**

   A. Amazon S3

   B. Amazon Athena

   C. Amazon Redshift

   D. Amazon Glue

2. **Which analytical tool is integrated with Amazon Elasticsearch Service for querying?**

   A. Tableau

   B. Kibana

   C. Data Studio

   D. Amazon Redshift

3. **What method should a data analytics team implement for sharing dashboard analysis while restricting access to external product owners?**

   A. Separate the data and use IAM policies for secure authorization.

   B. Create a manifest file to enforce row-level security.

   C. Utilize dataset rules with row-level security in Amazon QuickSight.

   D. Use S3 bucket policies to control user access per product.

4. **Which solution will improve the data loading performance for a sales data dashboard using Amazon Redshift?**

   A. Compress .csv files and use an INSERT statement to ingest data into Amazon Redshift.

   B. Split large .csv files, then use a COPY command to load data into Amazon Redshift.

   C. Use Amazon Kinesis Data Firehose to ingest data into Amazon Redshift.

   D. Load the .csv files in an unsorted key order and vacuum the table in Amazon Redshift.

5. In a scenario where query performance is prioritized over cost, which AWS service combination is recommended for analyzing scanned documents?

- **A. Amazon RDS and Amazon S3**
- **B. Amazon Elasticsearch Service and Amazon S3**
- **C. Amazon Athena and Amazon DynamoDB**
- **D. Amazon Redshift and Amazon S3**

6. Which feature of Amazon QuickSight would enhance data visualization for a global company's sales?

- **A. Interactive dashboards with real-time data**
- **B. Static reports generated on a weekly basis**
- **C. Manual data entry for accuracy**
- **D. Pre-set templates without customization**

7. How can a data engineer ensure real-time access to the most current data stored in S3 for analytics?

- **A. Set up a cron job to refresh the data catalog hourly.**
- **B. Run the AWS Glue crawler from a Lambda function triggered by S3:ObjectCreated events.**
- **C. Utilize Amazon EventBridge to trigger the crawler.**
- **D. Manually invoke the crawler after every data upload.**

8. What solution will allow Amazon QuickSight in ap-northeast-1 to access Amazon Redshift in us-east-1?

- **A. Configure cross-Region snapshots for Amazon Redshift**
- **B. Set up a VPN connection between Regions**
- **C. Use Amazon CloudFront for data distribution**
- **D. Create a security group with appropriate rules for access**

9. Which approach should a company take when looking for an out-of-the-box machine learning solution for forecasting metrics?

- **A. Use AWS Glue machine learning transforms for forecasting.**
- **B. Use Amazon QuickSight for visualizing key business metrics.**
- **C. Deploy a custom ML model using AWS Lambda.**
- **D. Utilize a pre-built machine learning AMI for analysis.**

10. What information can internal data analysts retrieve from the scanned documents using the proposed solution?

A. Document images only

B. Applicant name and document type only

C. Application date and application type only

D. Applicant name, application date, and application text

Answers

1. B
2. B
3. C
4. B
5. B
6. A
7. B
8. D
9. B
10. D

Explanations

1. What is the optimal data source for Amazon QuickSight when needing to create heat maps for visualizing sales data in S3?

A. Amazon S3

B. Amazon Athena

C. Amazon Redshift

D. Amazon Glue

When creating heat maps for visualizing sales data that is stored in Amazon S3, using Amazon Athena is the optimal choice because it allows for efficient querying of data stored directly in S3. Athena is a serverless interactive query service that enables users to analyze data in S3 using standard SQL without the need to set up any infrastructure or data warehousing solution like Amazon Redshift. By utilizing Athena, you can easily create ad-hoc queries against the sales data in S3 and prepare the data for visualization in Amazon QuickSight. This shortens the time from data storage to visualization, enhances flexibility, and allows for seamless integration with QuickSight, which can directly connect to results produced by Athena for creating various types of visualizations, including heat maps. While other options like Amazon Redshift provide more extensive data warehousing capabilities, they generally involve additional setup and data loading steps that may not be necessary for quick and periodic analysis of data stored in S3. Amazon Glue, on the other hand, serves primarily as a data catalog and ETL service, which means it is useful for data preparation but does not directly participate in visual analysis. Thus, Amazon S3, while serving as the data storage, typically interfaces with Athena to enable visualization, but

2. Which analytical tool is integrated with Amazon Elasticsearch Service for querying?

A. Tableau

B. Kibana

C. Data Studio

D. Amazon Redshift

The integration of Kibana with Amazon Elasticsearch Service serves as a powerful tool for querying, visualizing, and exploring the data stored in Elasticsearch. Kibana provides an intuitive user interface that is specifically built to leverage the capabilities of Elasticsearch, allowing users to create dashboards, charts, and graphs based on the data housed within their Elasticsearch clusters. By connecting Kibana to Amazon Elasticsearch Service, users can easily navigate through vast datasets, conduct real-time searches, and gain insights through visual representations. This synergy makes it an ideal choice for data analytics tasks that demand interactive exploration and visualization. The support for various types of visualizations and analytics within Kibana complements the core functionalities provided by Elasticsearch, thereby creating a robust analytics environment. In contrast, while Tableau and Google Data Studio are popular data visualization tools, they do not have native integration with Amazon Elasticsearch Service. Amazon Redshift, on the other hand, is a data warehouse solution, which serves a different purpose within the data ecosystem and does not directly integrate with Amazon Elasticsearch Service for querying purposes. This distinction underscores why Kibana is the most suitable analytical tool for users of Amazon Elasticsearch Service.

3. What method should a data analytics team implement for sharing dashboard analysis while restricting access to external product owners?

A. Separate the data and use IAM policies for secure authorization.

B. Create a manifest file to enforce row-level security.

C. Utilize dataset rules with row-level security in Amazon QuickSight.

D. Use S3 bucket policies to control user access per product.

Utilizing dataset rules with row-level security in Amazon QuickSight is the most effective method for sharing dashboard analysis while managing access restrictions. This approach allows the analytics team to create specific access controls at the row level based on user attributes. By implementing dataset rules, the team can define which data each user or group can access, ensuring that external product owners only see the information relevant to them. This capability is crucial in scenarios where sensitive data must remain confidential and access needs to be tailored according to the roles or needs of the audience. In contrast, separating the data and using IAM policies for secure authorization can be effective for overall access control but may not provide the granularity needed to restrict data visibility effectively at the dashboard level. A manifest file can help enforce security, but it's not as integrated into the analytical workflow as the row-level security capabilities of QuickSight. Finally, using S3 bucket policies to control access is generally more suitable for managing file access at a storage level rather than directly related to data analysis visualizations. This doesn't offer the same level of detailed control as row-level security for dashboard sharing.

4. Which solution will improve the data loading performance for a sales data dashboard using Amazon Redshift?

A. Compress .csv files and use an INSERT statement to ingest data into Amazon Redshift.

B. Split large .csv files, then use a COPY command to load data into Amazon Redshift.

C. Use Amazon Kinesis Data Firehose to ingest data into Amazon Redshift.

D. Load the .csv files in an unsorted key order and vacuum the table in Amazon Redshift.

The selected solution, which involves splitting large .csv files and using a COPY command to load data into Amazon Redshift, significantly enhances data loading performance due to several key factors inherent to Redshift's design and optimization strategies. Firstly, the COPY command in Amazon Redshift is specifically designed for bulk loading of data. It processes data in parallel, which means it can handle multiple files simultaneously, leading to much faster ingestion compared to using an INSERT statement. The parallel processing capabilities of Redshift leverage its architecture, allowing it to effectively distribute the workload across multiple nodes. Secondly, splitting large .csv files into smaller sizes optimizes the loading process further. Smaller files are easier to process and can also improve the data distribution across slices in a cluster. This distribution is crucial for performance because when data is spread evenly, it reduces the chances of bottlenecks and maximizes the use of available resources. As a result, loading split files can decrease the time to ingest large volumes of data significantly. Moreover, using the COPY command supports various optimizations such as automatic compression and support for file formats that can optimize storage and query performance. This means that data can not only be loaded quickly but also be stored efficiently, which contributes to better overall database performance for analytics workloads,

5. In a scenario where query performance is prioritized over cost, which AWS service combination is recommended for analyzing scanned documents?

A. Amazon RDS and Amazon S3

B. Amazon Elasticsearch Service and Amazon S3

C. Amazon Athena and Amazon DynamoDB

D. Amazon Redshift and Amazon S3

The emphasis on query performance over cost in analyzing scanned documents makes the combination of Amazon Elasticsearch Service and Amazon S3 a suitable choice. Amazon Elasticsearch Service is specifically designed for full-text search, enabling fast and scalable searches of large volumes of unstructured data, like scanned documents. Its capability to index and retrieve data quickly makes it ideal for scenarios where the speed of query responses is critical. When paired with Amazon S3, which serves as a versatile and cost-effective storage solution for a vast array of data types, this combination allows users to store their scanned documents efficiently while leveraging the powerful search functionalities of Elasticsearch. This integration facilitates rapid data access and facilitates insightful querying, essential for handling structured and unstructured data alike. Therefore, for scenarios prioritizing performance in analyzing documents, this pairing provides the necessary tools to address the complexities of querying while ensuring speed and efficiency in data retrieval.

6. Which feature of Amazon QuickSight would enhance data visualization for a global company's sales?

A. Interactive dashboards with real-time data

B. Static reports generated on a weekly basis

C. Manual data entry for accuracy

D. Pre-set templates without customization

The feature that would enhance data visualization for a global company's sales is interactive dashboards with real-time data. This capability allows decision-makers and stakeholders to engage actively with the data, exploring various dimensions and metrics dynamically. Interactive dashboards can highlight trends, patterns, and anomalies as they emerge, enabling timely responses to changing business conditions. Real-time data integration is particularly valuable for a global company, as sales information can vary significantly across different regions and markets. By providing the ability to drill down into the data, filter views, and respond to questions on-the-fly, interactive dashboards empower users to gain deeper insights into sales performance and make informed decisions promptly. In contrast, static reports generated on a weekly basis only provide a snapshot of data at a given time and lack the dynamism needed for fast-paced decision-making. Manual data entry introduces the possibility of errors and inefficiencies, which can compromise data integrity. Lastly, pre-set templates without customization limit flexibility and might not cater to the specific analytical needs or unique aspects of a company's operations.

7. How can a data engineer ensure real-time access to the most current data stored in S3 for analytics?

A. Set up a cron job to refresh the data catalog hourly.

B. Run the AWS Glue crawler from a Lambda function triggered by S3:ObjectCreated events.

C. Utilize Amazon EventBridge to trigger the crawler.

D. Manually invoke the crawler after every data upload.

The choice involving running the AWS Glue crawler from a Lambda function triggered by S3:ObjectCreated events is correct because it enables a real-time response to data uploads in Amazon S3. When new objects are created in an S3 bucket, the S3:ObjectCreated event can trigger a Lambda function. This function can then start an AWS Glue crawler, which will update the data catalog with the newly available data without any manual intervention. This approach ensures that the data catalog is updated immediately after new data is added, allowing analytics tools and users to access the most current data for their analysis. The integration of Lambda and S3 events allows for an efficient and automated workflow that minimizes latency between the data being uploaded and its availability for analytics. On the other hand, the alternative methods don't provide the same level of efficiency or immediacy: - Setting up a cron job to refresh the data catalog hourly introduces a delay in data availability, as it only updates at specified time intervals rather than in real-time. - Utilizing Amazon EventBridge for triggering the crawler, while it is a serverless event bus, does not directly correlate to immediate data access since the event would still require proper configuration and might not trigger on the exact data changes needed. - Manually invoking the crawler

8. What solution will allow Amazon QuickSight in ap-northeast-1 to access Amazon Redshift in us-east-1?

A. Configure cross-Region snapshots for Amazon Redshift

B. Set up a VPN connection between Regions

C. Use Amazon CloudFront for data distribution

D. Create a security group with appropriate rules for access

To enable Amazon QuickSight in one AWS region to access Amazon Redshift in another region, creating a security group with the appropriate rules for access is essential. When QuickSight needs to communicate with Redshift, it must have the necessary permissions set up to ensure secure connectivity. A security group acts as a virtual firewall that controls incoming and outgoing traffic to your resources. By defining the security group with rules that allow traffic from the QuickSight IP ranges, you can facilitate the connection between QuickSight in the ap-northeast-1 region and Redshift in the us-east-1 region. Other options don't directly address the need for proper connectivity. Cross-region snapshots do not facilitate real-time access; they are primarily for data backup and recovery. A VPN connection may provide a secure channel between regions, but it is not the most efficient solution for QuickSight to access Redshift, as it requires more complex setup and is not inherently necessary. Using Amazon CloudFront is inappropriate in this context, as it is a content delivery network designed for distributing web content, not for directly enabling database access. Thus, effective communication between services across regions relies significantly on correctly configured security measures like security groups.

9. Which approach should a company take when looking for an out-of-the-box machine learning solution for forecasting metrics?

A. Use AWS Glue machine learning transforms for forecasting.

B. Use Amazon QuickSight for visualizing key business metrics.

C. Deploy a custom ML model using AWS Lambda.

D. Utilize a pre-built machine learning AMI for analysis.

The selected answer focuses on utilizing Amazon QuickSight for visualizing key business metrics, highlighting an important aspect of forecasting in business. QuickSight is a business intelligence tool that allows organizations to create interactive dashboards and perform data analysis, making it useful for visualizing metrics that are essential for forecasting. Effective forecasting relies heavily on visual representation of data to identify trends, patterns, and anomalies. By using QuickSight, businesses can manipulate data derived from various sources and present it in a format that aids in decision-making. While it doesn't perform machine learning forecasting on its own, it allows users to visualize data outputs from machine learning models or other analytical processes effectively. The other options may seem applicable but focus on varying methodologies that might not provide a straightforward, out-of-the-box solution tailored specifically for forecasting metrics. AWS Glue, for example, is great for data preparation and ETL processes but is not exclusively aimed at forecasting. Deploying a custom ML model using AWS Lambda involves more complex setup and maintenance, which may not suit a need for a simple out-of-the-box solution. Utilizing a pre-built machine learning AMI could provide some analytical capabilities; however, it often requires a more technical background to implement effectively and may not be the best for immediate business visualization needs.

10. What information can internal data analysts retrieve from the scanned documents using the proposed solution?

A. Document images only

B. Applicant name and document type only

C. Application date and application type only

D. Applicant name, application date, and application text

The correct answer is that internal data analysts can retrieve the applicant name, application date, and application text from the scanned documents using the proposed solution. This is critical in scenarios where data extraction from documents is required for further analysis or processing. The ability to extract applicant names facilitates identifying individuals associated with applications, which is fundamental in many processing workflows such as loans, admissions, or service requests. Likewise, retrieving application dates helps in tracking the timelines of submissions and understanding how long an application has been in the system. Most importantly, accessing the application text is essential for content analysis, enabling data analysts to assess the quality, circumstances, and specifics of the application itself. Focusing solely on document images or only partial data such as applicant name and document type limits the functionality to mere identification. Similarly, restricting the extraction to application date and type omits vital qualitative information contained within the application text, which is often the most significant portion of data when making informed decisions. Hence, retrieving all three parameters provides a comprehensive dataset for thorough analysis.