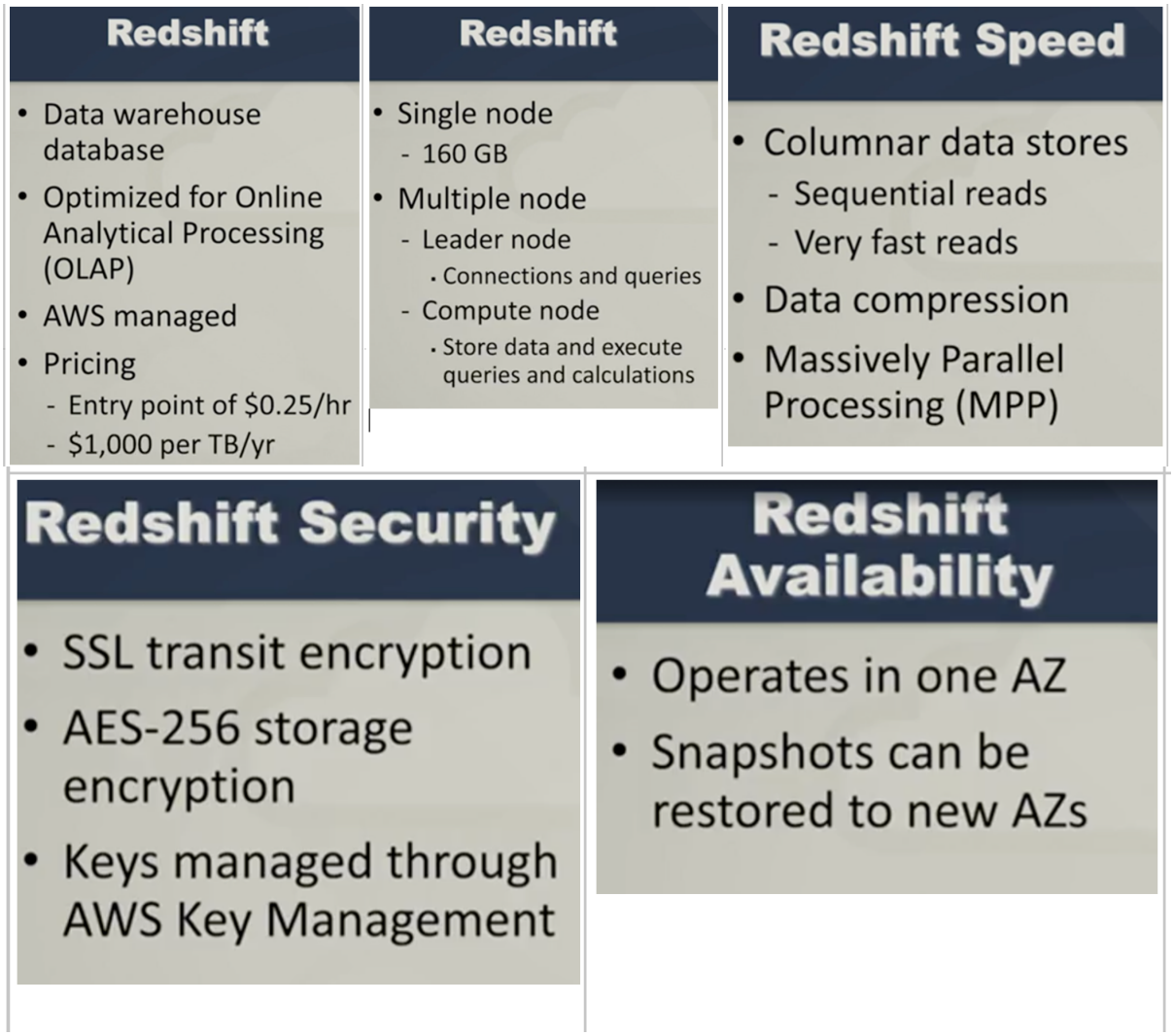
AWS - Redshift

- Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze all your data across your data warehouse and data lake.

- Redshift delivers ten times faster performance than other data warehouses by using machine learning, massively parallel query execution, and columnar storage on high-performance disk.

- **OLAP** (Online Analytical Processing)- database
- Gives fast reads
- Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard structured query language (SQL) and your existing business intelligence tools.



https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html

https://docs.aws.amazon.com/redshift/latest/gsg/getting-started.htm

https://tutorialsdojo.com/aws-cheat-sheet-amazon-redshift/

Here is a case study on finding the most suitable analytical tool - Kinesis vs EMR vs Athena vs Redshift:

https://youtu.be/wEOm6aiN4ww

Definitions

MODULER ARCHITECTURE of REDSHIFT
https://aws-quickstart.s3.amazonaws.com/quickstart-amazon-redshift/doc/modular-architecture-for-amazon-redshift.pdf

https://github.com/aws-quickstart/quickstart-amazon-redshift

- **Redshift Enhanced VPC Routing**
- When you use Amazon Redshift Enhanced VPC Routing, Amazon Redshift forces all COPY and UNLOAD traffic between your cluster and your data repositories through your Amazon VPC.
- By using Enhanced VPC Routing, you can use standard VPC features, such as VPC security groups, network access control lists (ACLs), VPC endpoints, VPC endpoint policies, internet gateways, and Domain Name System (DNS) servers.
- Hence, **enabling Enhanced VPC routing on your Amazon Redshift cluster** is the correct answer.
- You use these features to tightly manage the flow of data between your Amazon Redshift cluster and other resources.
- When you use Enhanced VPC Routing to route traffic through your VPC, you can also use VPC flow logs to monitor COPY and UNLOAD traffic.
- If Enhanced VPC Routing is not enabled, Amazon Redshift routes traffic through the Internet, including traffic to other services within the AWS network.

Redshift - Workload Management

- **Amazon Redshift workload management (WLM)** enables users to flexibly manage priorities within workloads so that short, fast-running queries won't get stuck in queues behind long-running queries.
- Amazon Redshift WLM creates query queues at runtime according to service classes, which define the configuration parameters for various types of queues, including internal system queues and user-accessible queues.
- From a user perspective, a user-accessible service class and a queue are functionally equivalent.

Reference:

https://docs.aws.amazon.com/redshift/latest/dg/c_workload_mngmt_classification.html

Check out this Amazon Redshift Cheat Sheet:

https://tutorialsdojo.com/amazon-redshift/

AWS - Redshift SPECTRUM

- **Redshift Spectrum** is primarily used to directly query open data formats stored in Amazon S3 without the need for unnecessary data movement, which enables you to analyze data across your data warehouse and data lake, together, with a single service.
- It does not provide the ability to process your data in real-time, unlike Kinesis.

Reference:

https://aws.amazon.com/s3/features/#Query_in_Place

Check out these AWS Cheat Sheets:

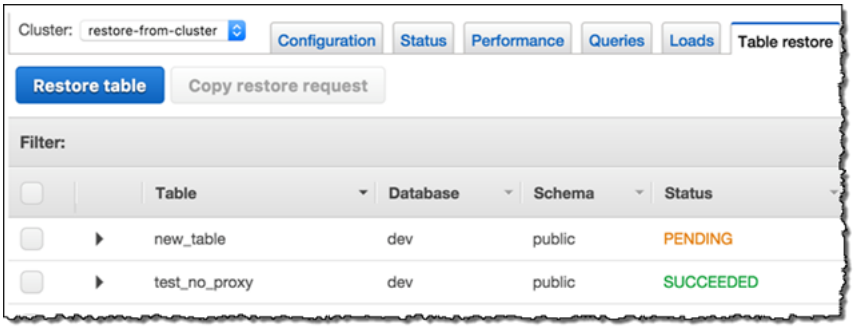
https://tutorialsdojo.com/aws-cheat-sheet-amazon-s3/

https://tutorialsdojo.com/aws-cheat-sheet-amazon-athena/

https://tutorialsdojo.com/aws-cheat-sheet-amazon-redshift/

AWS - Redshift CROSS REGION replication

- Redshift Cross Region Replication
  - You can configure Amazon Redshift to copy snapshots for a cluster to another region.
  - To configure cross-region snapshot copy, you need to enable this copy feature for each cluster and configure where to copy snapshots and how long to keep copied automated snapshots in the destination region.
  - When cross-region copy is enabled for a cluster, all new manual and automatic snapshots are copied to the specified region.



AUTOMATED-SNAPSHOTS

When automated snapshots are enabled for a cluster, Amazon Redshift periodically takes snapshots of that cluster, usually every eight hours or following every 5 GB per node of data changes, or whichever comes first.

Automated snapshots are enabled by default when you create a cluster.

These snapshots are deleted at the end of a retention period. The default retention period is one day, but you can modify it by using the Amazon Redshift console or programmatically by using the Amazon Redshift API.

When you enable Amazon Redshift to automatically copy snapshots to another region, you specify the destination region where you want snapshots to be copied.

In the case of automated snapshots, you can also specify the retention period that they should be kept in the destination region.

After an automated snapshot is copied to the destination region and it reaches the retention time period there, it is deleted from the destination region, keeping your snapshot usage low.

You can change this retention period if you need to keep the automated snapshots for a shorter or longer period of time in the destination region.

Reference:

https://docs.aws.amazon.com/redshift/latest/mgmt/managing-snapshots-console.html

Check out this Amazon Redshift Cheat Sheet:

https://tutorialsdojo.com/aws-cheat-sheet-amazon-redshift/

AWS - Redshift Query Improvements

In Redshift, if your query operation hangs or stops responding, below are the possible causes as well as its corresponding solution:

Connection to the Database Is Dropped

Reduce the size of maximum transmission unit (MTU). The MTU size determines the maximum size, in bytes, of a packet that can be transferred in one Ethernet frame over your network connection.

Connection to the Database Times Out

Your client connection to the database appears to hang or timeout when running long queries, such as a COPY command. In this case, you might observe that the Amazon Redshift console displays that the query has completed, but the client tool itself still appears to be running the query. The results of the query might be missing or incomplete depending on when the connection stopped. This effect happens when idle connections are terminated by an intermediate network component.

Client-Side Out-of-Memory Error Occurs with ODBC

If your client application uses an ODBC connection and your query creates a result set that is too large to fit in memory, you can stream the result set to your client application by using a cursor. For more information, see DECLARE and Performance Considerations When Using Cursors.

Client-Side Out-of-Memory Error Occurs with JDB

When you attempt to retrieve large result sets over a JDBC connection, you might encounter client-side out-of-memory errors.

There Is a Potential Deadlock

If there is a potential deadlock, try the following:

- View the STV_LOCKS and STL_TR_CONFLICT system tables to find conflicts involving updates to more than one table.

- Use the PG_CANCEL_BACKEND function to cancel one or more conflicting queries.

- Use the PG_TERMINATE_BACKEND function to terminate a session, which forces any currently running transactions in the terminated session to release all locks and roll back the transaction.

- Schedule concurrent write operations carefully.

Reference:

https://docs.aws.amazon.com/redshift/latest/dg/queries-troubleshooting.html#queries-troubleshooting-query-hangs

As shown above, the correct answers in this scenario are the following options:

1. Reduce the size of maximum transmission unit (MTU).

2. View the STV_LOCKS and STL_TR_CONFLICT system tables to find conflicts involving updates to more than one table.

3. Use the PG_CANCEL_BACKEND function to cancel one or more conflicting queries.

Running the VACUUM command and ***increasing the available memory by increasing the*** ***wlm_query_slot_count*** are incorrect as you should only do this for situations where your query takes too long to process.

Using a single COPY command is incorrect because you only have to do this if the load operation takes too long