Chapter 10: Data Warehouses

Chapter 10 of the SnowFlake SnowPro Core Certification Course. Data Warehouses.

Data warehouses in SnowFlake are one of the most important parts of this technology, and as we discussed in previous chapters, they will allow us to compute. In this chapter, we are going to see them in more detail:

WAREHOUSES IN SNOWFLAKE

A Data Warehouse is a cluster of computing resources in Snowflake. It provides the required resources, such as CPU, memory, and temporary storage, to perform queries and DML operations (like loading data into tables). A warehouse must be running and in use for the session to perform these operations. While a warehouse is running, it consumes Snowflake credits.

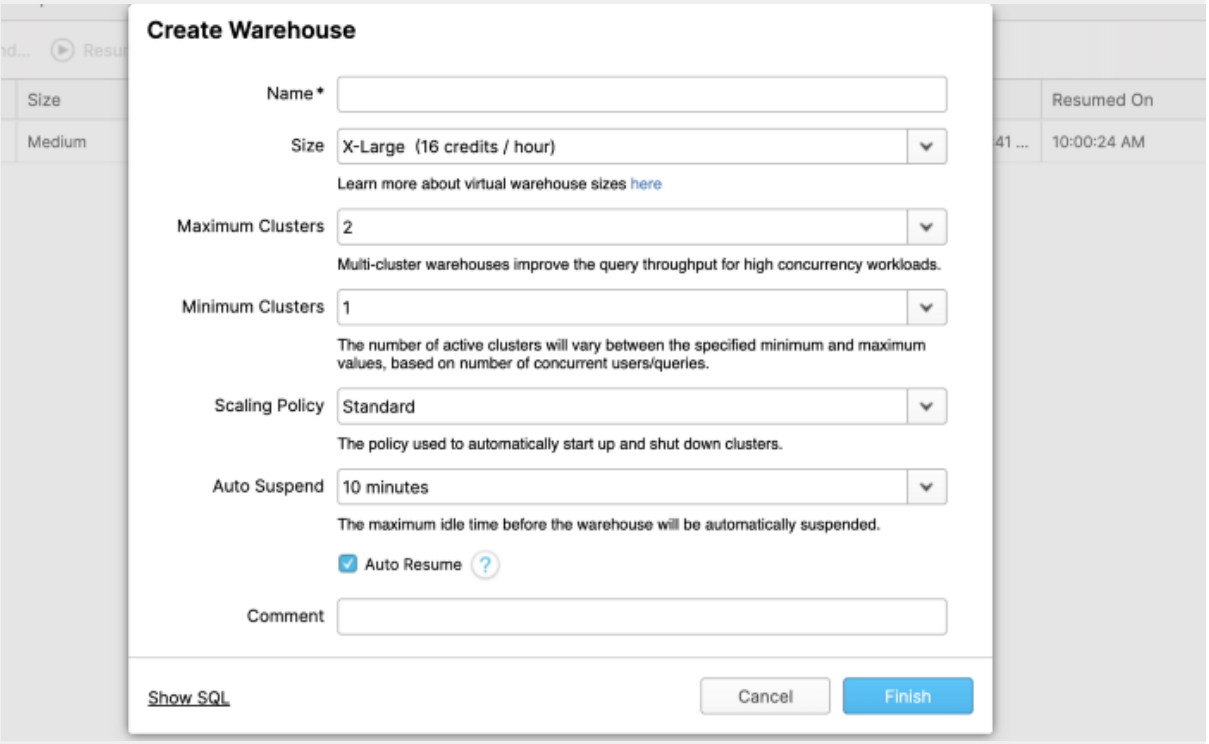
Virtual Warehouses in SnowFlake (via snowflake.com)

Snowflake utilizes per-second billing (with a 60-second minimum each time the warehouse starts), so warehouses are billed only for the credits they consume:

- If we run the warehouse for 10 seconds, we will be billed for 60 seconds.
- If we run the warehouse for 59 seconds, we will be billed for 60 seconds.
- If we run the warehouse for 138 seconds, we will be billed for 138 seconds.

PROPERTIES FROM THE WAREHOUSES

A Warehouse is defined by its size and other properties that we can set to help control and automate warehouse activity. Let's take a look:



SnowFlake Warehouse properties.

LinkedIn: https://www.linkedin.com/in/avinash-sharma-553378151/

Size

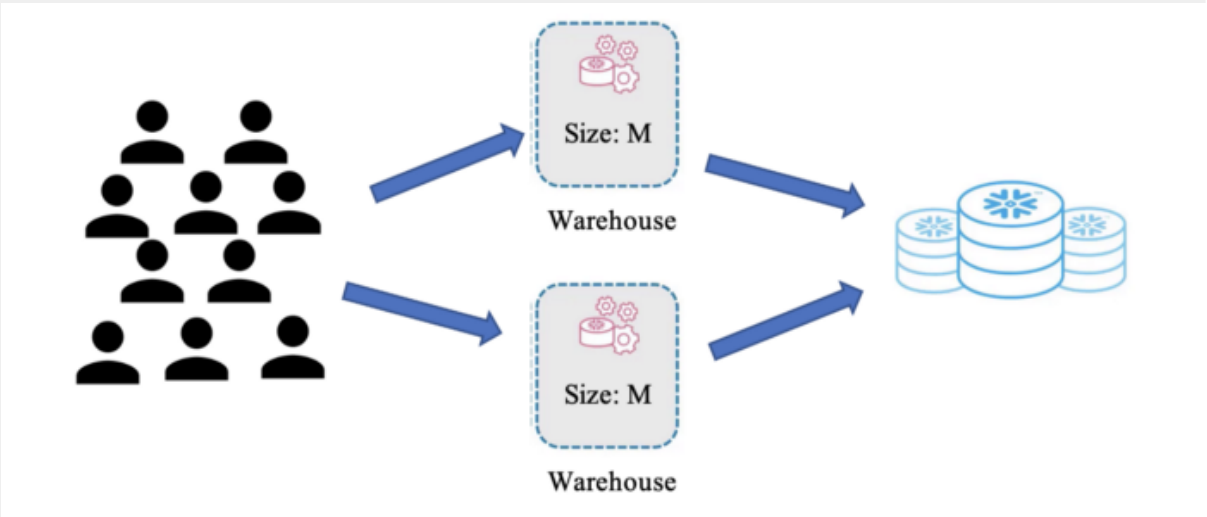
As we saw in the chapter about pricing, they might consume additional credits per hour depending on the warehouse size. The size of a warehouse can impact the amount of time required to execute queries submitted to the warehouse. So, a query will be executed faster in a Large warehouse than in a Small warehouse, but you will pay more credits per second. We can see some of the most common sizes in the following image:

Warehouse Size	Credits / Hour	Credits / Second
X-Small	1	0.0003
Small	2	0.0006
Medium	4	0.0011
Large	8	0.0022
X-Large	16	0.0044

Credits/Hour of a Warehouse depending on the size.

Multi-Cluster Warehouses

With multi-cluster warehouses, you can scale compute resources to manage query concurrency during, for example, peak hours. You can add additional warehouses to make a larger pool of computing resources available. To activate this option, you need (at least) the SnowFlake Enterprise Edition (this is a typical question in the exam). You can see how they look like in the following picture:



Multi-cluster warehouse diagram.

For a multi-cluster warehouse, the number of credits billed is calculated based on the size and the number of warehouses that run within the time period. If we run two "S" warehouses, we will be billed for 2 credits/hour * 2 warehouses of "S" size, which is 4 credits/hour.

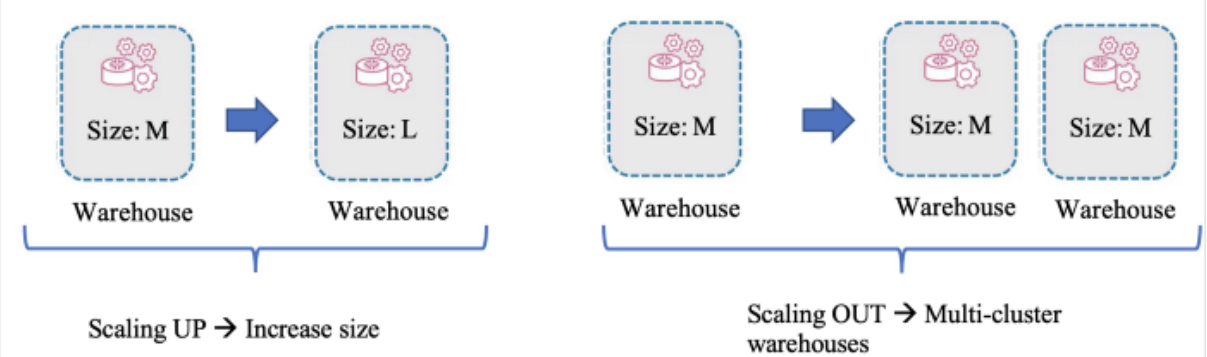
To create multi-cluster warehouses, you need to specify the following properties:

- Maximum clusters (1–10)
- Minimum clusters (≤ than the maximum)

Scale Up vs Scale-Out

What is the difference between increasing the Warehouse size and multi-cluster warehouses? Multi-cluster warehouses are best utilized for scaling resources to improve concurrency for users/queries, also known as scale OUT/IN. If we wanted to improve the performance of the queries, we should resize the warehouse, also known as scale UP/DOWN the Data Warehouse.

LinkedIn: https://www.linkedin.com/in/avinash-sharma-553378151/



Scale Up vs Scale Out a SnowFlake Warehouse.

Multi-warehouse modes

Depending on the maximum and the minimum number of clusters, we have two different modes in which the warehouse can run:

- Maximized → We will enable this mode specifying the SAME value for the maximum and the minimum number of clusters. SnowFlake will start all the warehouses so that the maximum resources are available while they run.
- Auto-Scale → We will enable this mode specifying DIFFERENT values for the maximum and the minimum number of clusters. SnowFlake will start and stop warehouses as needed. To control how the warehouse scales, we can define Scaling Policies.

Scaling policy

When you create a multi-cluster warehouse, you need to specify a scaling policy, which will help you control the credits consumed by the multi-cluster warehouse.

- Standard policy → It prioritizes starting additional warehouses over conserving credits.
- Economy policy → A more restrictive policy that prioritizes conserving credits over conserving starting additional warehouses.

Auto Suspend & Auto Resume

SnowFlake can suspend the warehouse if it's inactive for a specific period of time. By default, the auto suspend option is enabled. So imagine that we don't use the warehouse for 10 minutes, it will automatically be suspended.

Auto Resume allows SnowFlake to automatically resume the warehouse when any statement requires the use of the warehouse, like any query or DML command. By default, it's also enabled.

SNOWFLAKE SNOWPRO EXAM QUESTIONS

1. How is query processing done in SnowFlake?

1. AWS EMR with Spark
2. AWS EC2 with Spark
3. Virtual Warehouses

Solution: 3

2. You have two virtual warehouses in your SnowFlake account. If one of them updates the data in the storage layer, when will the other one see the data?

1. Immediately
2. After an average time of 5 seconds
3. After the sync process

Solution: 1. All the warehouses of your account share the storage layer, so if the data is updated, all the warehouses will be able to see it.

3. Can you resize the warehouse once you have selected the size?

1. True
2. False

Solution: 1. You can always change the size of the warehouse depending on your needs.

4. If you want a dedicated virtual warehouse, which is the lowest SnowFlake edition you should opt for?

1. Standard
2. Enterprise

3. Business Critical
4. Virtual Private Snowflake

Solution: 1. In SnowFlake, all the Virtual Warehouses are dedicated to the users. If you create a virtual warehouse, you will only be the one using it.

5. If you want a multi-cluster warehouse, which is the lowest SnowFlake edition that you should opt for?

1. Standard
2. Enterprise
3. Business Critical
4. Virtual Private Snowflake

Solution: 2.

6. Queries in SnowFlake are getting queued on the warehouses and delaying the ETL processes of the company. What are the possible solution options that you can think of considering we have the SnowFlake Enterprise edition?

1. Resize the warehouse
2. Use multi-cluster warehouse
3. Set auto-resize parameter to TRUE
4. Contact the SnowFlake support to increase the size of the warehouse

LinkedIn: https://www.linkedin.com/in/avinash-sharma-553378151/

Solution: 1, 2. Resizing the warehouse, your company will scale up, reducing the time to execute big queries. Using multi-cluster warehouses, you will have more queries running simultaneously and a high concurrency when they execute.

7. A warehouse ran for 62 seconds, and it was suspended. After some time, it ran for another 20 seconds. For how many seconds will you be billed?

1. 20 seconds
2. 62 seconds
3. 92 seconds
4. 122 seconds

Solution: 4. 62 seconds + 60 seconds because warehouses are billed for a minimum of one minute. The price would be different if the warehouse wasn't suspended before executing the second query.

8. Can two different virtual warehouses from the same account access the same data simultaneously without any contention issue?

1. True
2. False

Solution: 1

9. Can virtual warehouses be resized while they are running?

1. True
2. False

Solution: 1

10. A medium (M) warehouse has auto-suspend configured after 15 minutes. You have noticed that all of the queries that run on this warehouse finish within a minute. What will you do to optimize compute costs?

1. Delete the warehouse after a minute
2. Reduce the auto-suspend time to 1 minute
3. Use another data-warehouse

Solution: 2

11. What happens to the incoming queries when a warehouse does not have enough resources to process them?

1. Queries are aborted
2. Queries are queued and executed when the warehouse has resources
3. SnowFlake resizes the warehouse

Solution: 2

12. Which function returns the name of the warehouse of the current session?

LinkedIn: https://www.linkedin.com/in/avinash-sharma-553378151/

1. ACTIVE_WAREHOUSE()
2. RUNNING_WAREHOUSE()
3. CURRENT_WAREHOUSE()
4. WAREHOUSE()

Solution: 3. I'm not a big fan of learning commands by heart, and they are unlikely to appear on the exam, but this one may be useful.

Thanks for Reading!