1. Have you worked on snowflake sampling? Why do we use sampling?

Yes, we have used block and row sampling for various use cases, mostly, for smaller tables, we’ve used cloning but for larger tables, we’ve used sampling.

We use sampling when we want to test some new feature on production data but don’t want to incur the cost on querying the whole dataset. Once the feature is completely developed, we can finally run the final query on the actual prod data.

1. How is sampling better than cloning?  
   Sampling creates smaller data out of bigger chunk unlike cloning which will run queries on production-like data and cost compute. Sampling costs you storage but reduce compute costs.
2. How to manage resource monitors?
3. What all can be configured for a user profile in Snowsight?  
   Deafult role and warehouse, default experience, language, notifications and multi factor authentication
4. What is the prerequisite for **Search Optimization service?**

**To be on** Enterprise Edition (or higher)

1. The Search Optimization service can be used to improve the performance of which type of queries?
   * Selective point lookup queries on tables.
   * Substring and regular expression searches
   * Queries on fields in VARIANT, OBJECT, and ARRAY (semi-structured) columns
   * Queries that use selected geospatial functions with GEOGRAPHY values
2. You are unloading data from a multi-gigabyte table to an external stage; which of the following statements regarding the exported file(s) are correct? Select all that apply.

* Each exported file is of 16 MB
* Data is exported to multiple files
* Files are not compressed unless the option is specified
* Largest file size is 5GB
* Data can be exported in a single file with parameter single=true

1. What happens when you resize a running warehouse?

Resizing a running warehouse adds or removes compute resources in each cluster in the warehouse. All usage and credit rules with starting and suspending a warehouse apply to resizing a started warehouse, such as:

**Compute resources added to a warehouse start using credits when they are provisioned**, however, the additional compute resources don’t start executing statements until they are all provisioned, unless some of the resources fail to provision.

Compute resources are removed from a warehouse **only when they are no longer being used to execute any current statements.**

1. How are credit and usage rules applied to resizing a suspended warehouse?  
   Resizing a suspended warehouse does not provision the new compute resources for the warehouse, it simply instructs the snowflake to provision the additional compute resources when the warehouse is next resumed.
2. Which are caching mechanisms in snowflake?

Query result caching: For already executed queries

Warehouse caching: Within virtual warehouse and based on queries that have already been executed

Metadata caching: For queries that can be fulfilled directly from metadata.

1. What are the snowflake cache layers?  
   Result Cache: Holds the results of every query executed in last 24 hours upto a period of 31 days after which if the query is submitted a new result is generated and persisted. These are available across virtual warehouses so query results returned to one user is available to any other user on the system who executes the same query provided the underlying data has not changed.

Local disk cache: Cache data used by SQL queries. Whenever data is needed for a given query it is retrieved from the remote disk storage and cached in SSD and memory

Remote disk: Holds long term storage and is responsible for data resilience which in the case of AWS means 99.999999999% durability, even in the event of an entire data centre failure.

1. What are Transaction processing council benchmark tables?
2. What determines the cost for credits and data storage you use?

The snowflake edition that your organization chooses

Region where the snowflake account is located and whether it is an On Demand or Capacity account.

1. What drivers does snowflake currently provide?

.net, php pdo, python, go lang, odbc, jdbc, spark, kafka

1. What is supported by snowflake for the purpose of auto provisioning users and group membership?

Snowflake supports SCIM 2.0 which is compatible with Okta and Azure active directory. It is an open standard that provides automatic user provisioning and role synchronization based on identity provider information.

When a new user is created in the identity provider the SCIM automatically provisions the user in snowflake.

Additionally, SCIM can sync groups defined in an identity provider with snowflake roles.

1. What is the fail safe feature of snowflake?

Fail safe is a data recovery service that is provisioned on a best effort basis and is intended for use when all other recovery options have been attempted.

It is **NOT** provided as a means for accessing historical data after time travel retention period has ended. It is for use only by snowflake to recover the data that may have been lost or damaged due to extreme operational failures.

Data recovery through fail safe may take several hours to several days to complete.

1. Snowflake architecture

3 layers :

Cloud services : brain of the system , manages infrastructure access control, security, optimizer and metadata

Query processing : muscle of the system performs massive parallel processing

and

Storage : Hybrid columnar storage , data stored in blobs

1. Sizes of virtual warehouses:
   * XS : 1
   * S : 2
   * M : 4
   * L : 8
   * XL : 16
   * 4XL : 128

Billed by second with 1 minute as minimum

1. What is the Snowpark optimized virtual warehouse

Each node on SOVW has 16x more memory and 10x the local cache as compared to a single node of a standard warehouse

1. What is the default time for auto-suspend for a warehouse?

10 minutes

1. How can you create a warehouse from SQL?

CREATE WAREHOUSE COMPUTE\_WAREHOUSE

WITH

WAREHOUSE\_SIZE=XSMALL

WAREHOUSE\_TYPE = STANDARD

MAX\_CLUSTER\_COUNT=3

SCALING\_POLICY=STANDARD

AUTO\_SUSPEND=300 (OR NULL if don’t require to suspend)

AUTO\_RESUME = TRUE

INITIALLY\_SUSPENDED=TRUE

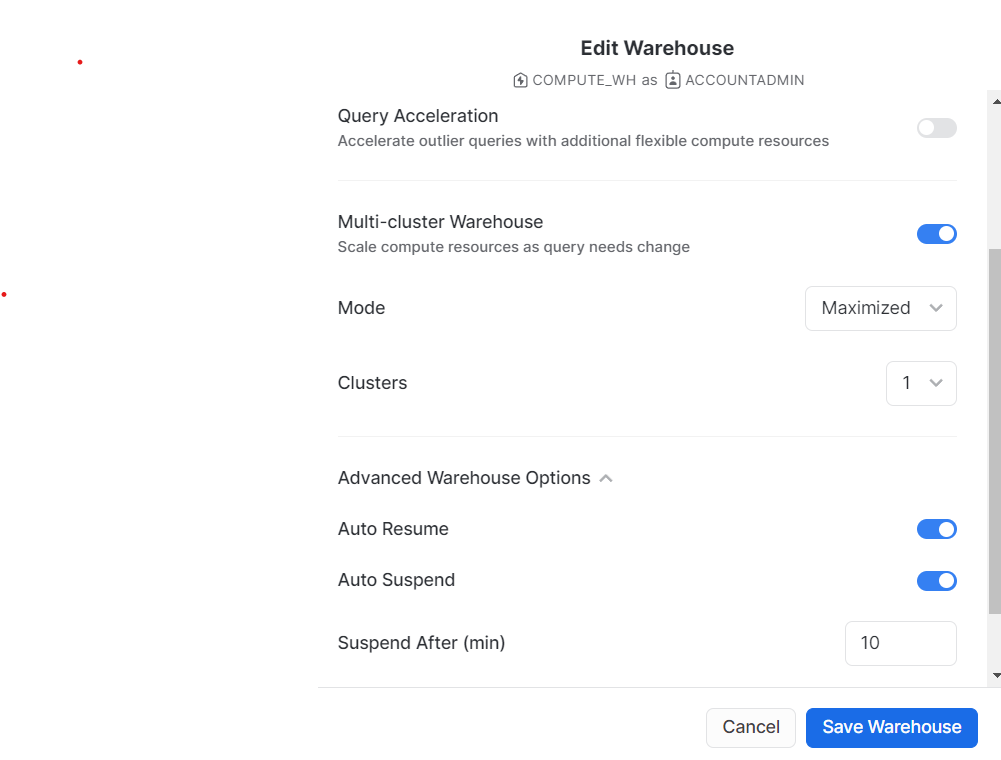
COMMENT=’My warehouse’

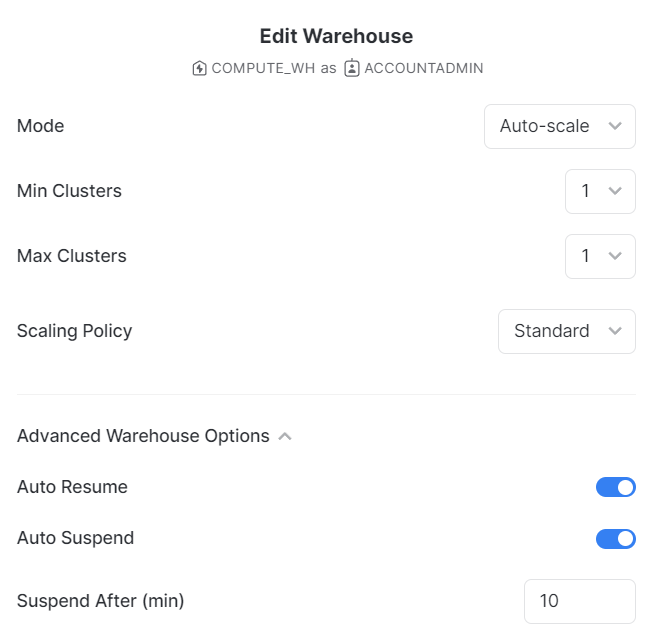
RESOURCE\_MONITOR=<monitor\_name>

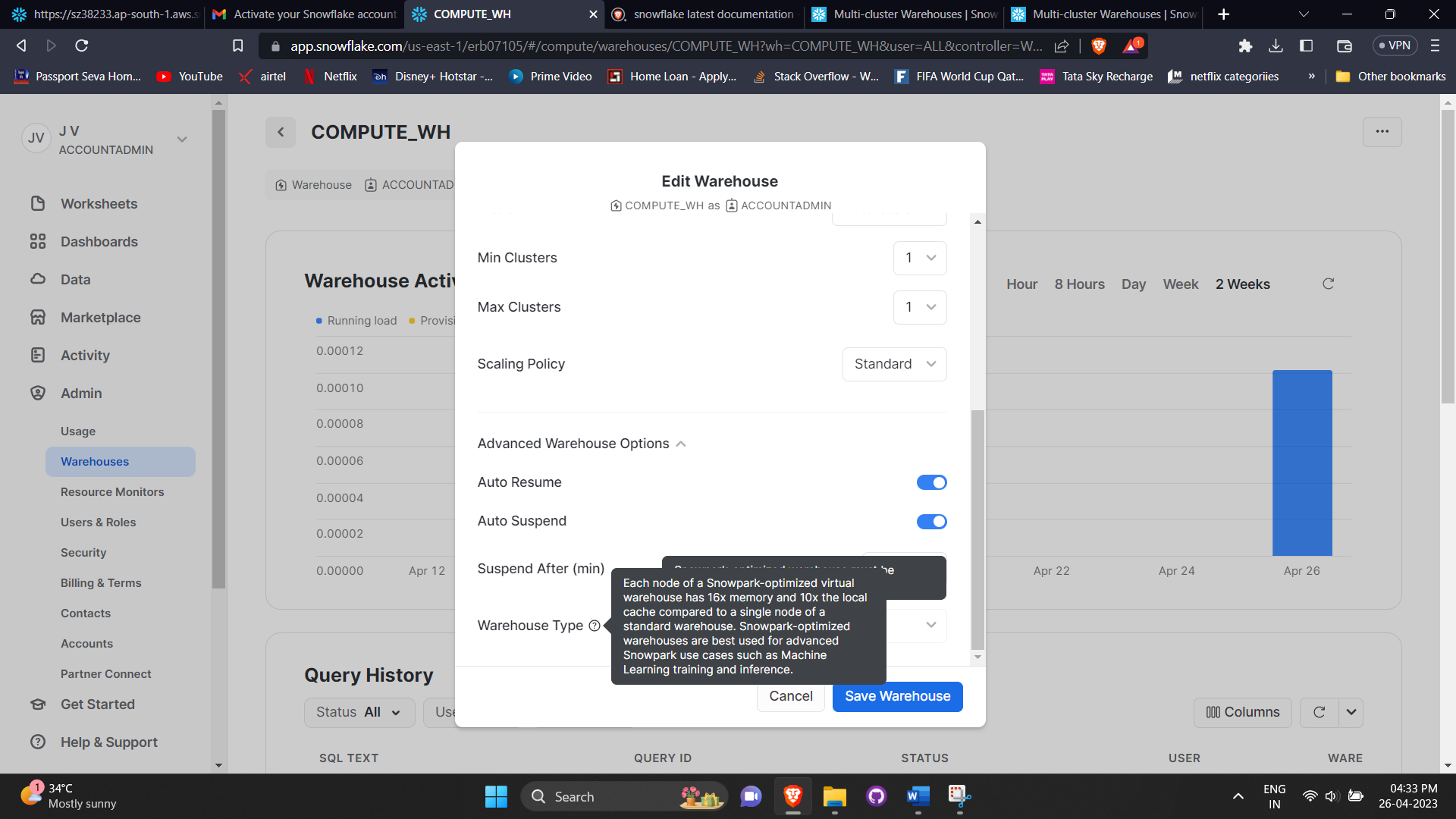
ENABLE\_QUERY\_ACCELERATION=TRUE

QUERY\_ACCELERATION\_MAX\_SCALE\_FACTOR=4

1. What is multi clustering?







1. How is the amount of compute resources in each cluster determined?  
   The total number of clusters for a multi-cluster warehouse is the product of warehouse size(XS :1, S:2 …) and the maximum number of clusters(defined by you in max clusters variable, usually started with 3)
2. What happens when a multi cluster warehouse is resized?  
   all the clusters of the warehouse including the ones running and those started after the multi-cluster warehouse is resized
3. When do you use multi cluster warehouse and when do you opt for resizing?  
   Multi-cluster warehouses are best utilized for scaling resources to improve concurrency for users/queries. They are not as beneficial for improving the performance of slow-running queries or data loading. For these types of operations, resizing the warehouse provides more benefits.
4. What kind of warehouse did you use in your project?  
   We used a multi cluster warehouse with auto-scaling, with 1 min and 3 max clusters and a standard scaling policy with auto resume and auto suspension time of 1 minute.
5. What do you mean by SaaS?  
   Snowflake provides with   
    Application – databases, tables etc  
    Software, data, OS 🡪 snowflake manages   
    Physical servers, virtual machines, physical storage 🡪 Cloud provider manages
6. What are available snowflake editions?  
   Standard : Introductory level  
   Enterprise : additional features for large scale enterprises  
   Business critical : even higher levels of data protection for organizations with extremely sensitive data  
   Virtual Private : highest level of security
7. Explain features of all these editions  
   1. Standard: $2/credit
   * Complete DWH
   * Automatic data encryption
   * Time travel up to 1 day
   * Disaster recovery for 7 days time travel
   * Secure data share
   * Premier support 24/7

2. Enterprise: $3/Credit

1. All standard features
2. Multi cluster warehouse
3. Time travel up to 90 days
4. Materialized views
5. Search optimization
6. Column level security

3. Business critical : $4/Credit

1. All enterprise features
2. Additional security features such as data encryption everywhere
3. Extended support
4. Database failover and disaster recovery

4. Virtual private : Check with snowflake

1. All business-critical features
2. Dedicated virtual servers and completely separate snowflake env

The prices are for AWS platform in US east northern virginia region

1. What are snowflake credits?  
    We exchange dollars and euros with snowflake to buy snowflake credits first and then consume those credits.
2. Explain snowflake pricing

Pay only what you need

Scalable amount of storage at affordable cloud price

Pricing depending on the region

Compute and storage costs are decoupled

Compute  
 Charged for active warehouses per hour

Depending on the size of the warehouse

Billed by second (min of 1 min)

Charged in snowflake credits

Storage

Monthly storage fees

Based on avg storage used per month

Cost calculated after compression

Cloud providers

1. On Demand Storage: Pay only for what you use, $40/TB
2. Capacity Storage: Pay only for defined capacity upfront, $23/TB (prices as per AWS platform in US east northern virginia region)

Example: We think we need 1 TB of storage

Scen 1: 100 GB of storage used

ODS : 0.1 TB \* 40 = $4

CS : 1\*23 = $23

Scen 2: 800 GB of storage used

ODS : 0.8 TB\*40 = $32

CS : 1\*23 = $23

Start with on demand and once you are sure about your usage use capacity storage

1. How are credits utilized/charged in snowflake?  
   It charges 1 credit per hour , Billed by second with 1 minute as minimum
   * Diff between standard and economy scaling policy  
     Cluster in economy starts only if the system estimates there is enough query load to keep the cluster busy for atleast 6 minutes but in standard The first cluster starts immediately when either a query is queued or the system detects that there’s one more query than the currently-running clusters can execute.  
       
     Each successive cluster waits to start 20 seconds after the prior one has started. For example, if your warehouse is configured with 10 max clusters, it can take a full 200+ seconds to start all 10 clusters.
   * Standard policy prevents/minimizes queuing by favouring starting additional clusters over conserving credits unlike economy which conserves credits by favoring keeping running clusters fully-loaded rather than starting additional clusters, which may result in queries being queued and taking longer to complete.
   * Cluster shuts down in standard policy when after 2 to 3 consecutive successful checks (performed at 1 minute intervals), which determine whether the load on the least-loaded cluster could be redistributed to the other clusters without spinning up the cluster again, where as for economy it waits for 5-6 successful checks.
2. When are you charged for data transfer in snowflake?  
   When we get the data out of snowflake, not when data is loaded from aws into our account

Not even when we share the data with another snowflake account with same cloud provider in the same region

1. What are different snowflake roles?  
   AccountAdmin: Top level role

Security Admin: Useradmin role is granted to security admin, can manage users and roles and can manage any object grant globally

User Admin: dedicated to user and role management only and can create users and roles

Public: Assigned to every user automatically and can create own objects like every other role

Sysadmin: Create warehouses and dbs and more objects and recommended that all custom roles are assigned

1. What are snowflake stages?  
   Location of data files where data can be loaded from   
   External: External cloud provider like S3, DCP and Microsoft azure, database object created in schema, created with create stage (url , access settings), additional costs If region/platform differs

Internal: local storage maintained by snowflake

1. What is a Snowflake cloud data warehouse?

Snowflake is an analytic data warehouse implemented as a SaaS service. It is built on a new SQL database engine with a unique architecture built for the cloud. This cloud-based data warehouse solution was first available on AWS as software to load and analyze massive volumes of data. The most remarkable feature of Snowflake is its ability to spin up any number of virtual warehouses, which means the user can operate an unlimited number of independent workloads against the same data without any risk of contention.

1. Is Snowflake an ETL tool?

Yes, Snowflake is an ETL tool. It’s a three-step process, which includes:

Extracts data from the source and creates data files. Data files support multiple data formats like JSON, CSV, XML, and more.

Loads data to an internal or external stage. Data can be staged in an internal, Microsoft Azure blob, Amazon S3 bucket, or Snowflake managed location.

Data is copied into a Snowflake database table using the COPY INTO command.

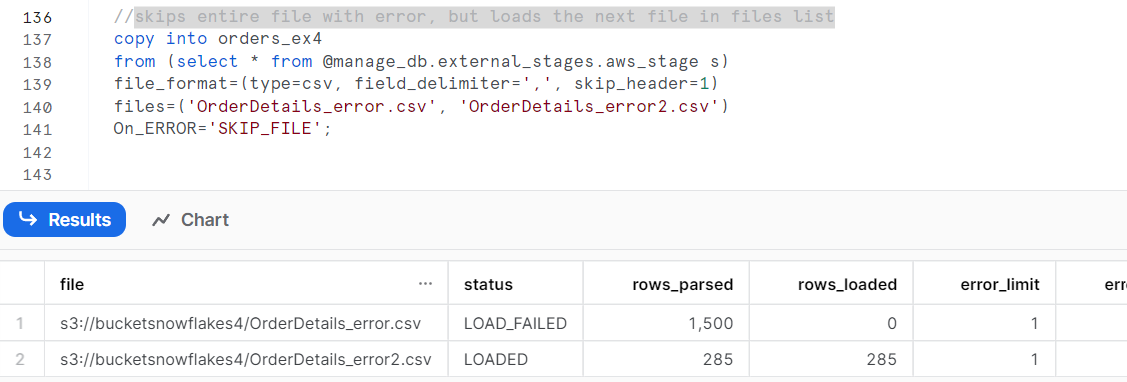
1. How is Snowflake distinct from AWS?

Snowflake offers storage and computation independently, and storage cost is similar to data storage. AWS handles this aspect by inserting Redshift Spectrum, which enables data querying instantly on S3, yet not as continuous as Snowflake.

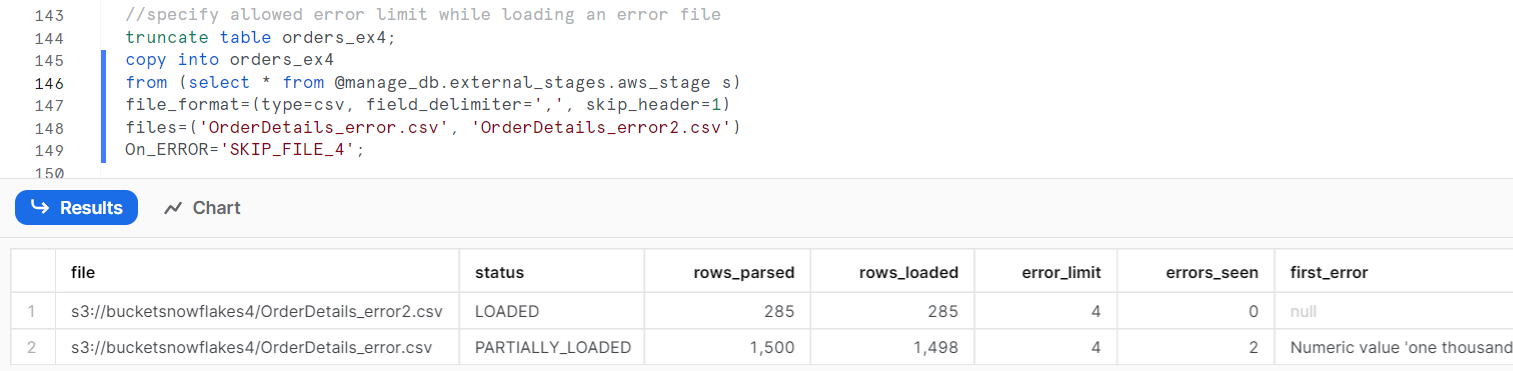
1. Can AWS glue connect to Snowflake?

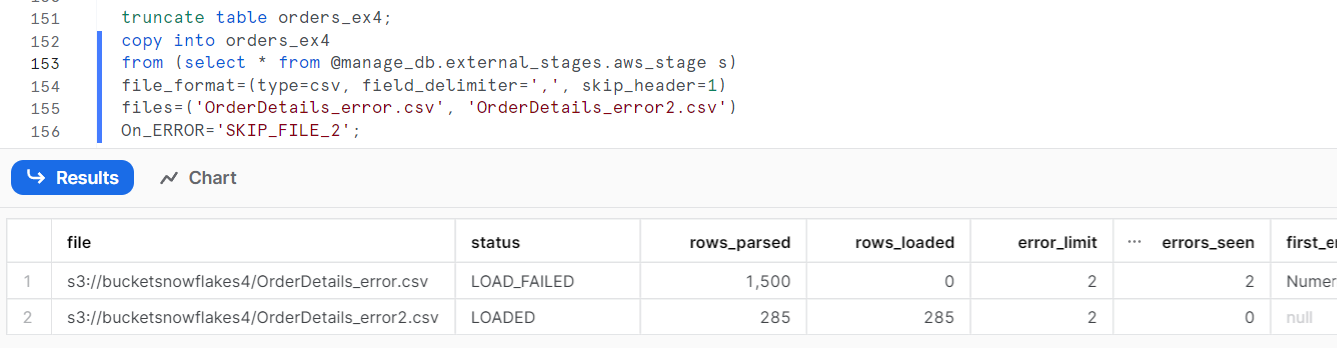
Definitely. AWS glue presents a comprehensive managed environment that easily connects with Snowflake as a data warehouse service. These two solutions collectively enable you to handle data ingestion and transformation with more ease and flexibility.

1. What happens when you try to load the data from the same file again using copy command?  
   It will not load the data if the data is already copied, the message would be -> load skipped, file was loaded before.
2. How can you handle errors while loading data from multiple S3 files into snowflake?  
   We can define 3 behaviours  
    1. ABORT\_STATEMENT: Default, if error occurs on the first file, second file will not be loaded
3. CONTINUE: Continues to load other records from the file, skipping the records with error
4. SKIP\_FILE: skips entire file with error, but loads the next file in files list

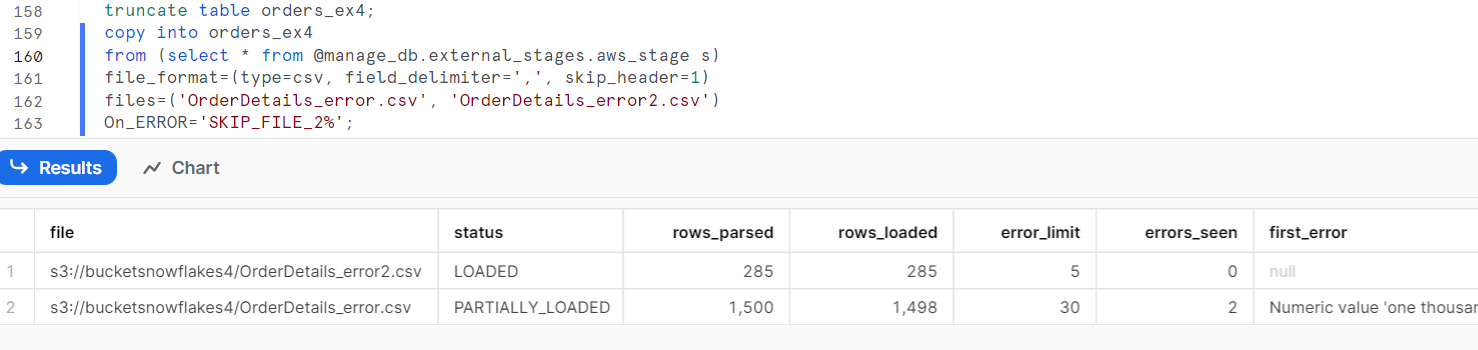


1. SKIP\_FILE\_<number>: Allow <number> of errors while loading the data, skip file if number limit is crossed.





1. SKIP\_FILE\_<number>%: Allow <number>% of errors while loading the data, skip file if the percentage limit is crossed.



1. Is it possible to change the file format type for a file format you have already created?  
   No, you need to recreate the file format with create or replace, but you can add available file format option in the copy command file\_format option like below

copy into orders\_ex\_ff

from @manage\_db.external\_stages.aws\_stage

file\_format=(format\_name=manage\_db.file\_formats.my\_file\_format, skip\_header=1)

files=('OrderDetails\_error.csv')

ON\_ERROR='SKIP\_FILE\_3';

1. What are the properties in the stage object?  
   STAGE\_FILE\_FORMAT, STAGE\_COPY\_OPTIONS and STAGE\_LOCATION
2. What is the validation\_mode used for?  
   For data validation without loading them.  
   RETURN\_ERRORS: Returns 0 records when there are no errors.  
   RETURN\_5\_ROWS: Returns first 5 error free records  
   If there are files with no errors in the list of files being loaded, you might see some result here because the order in which files are read by copy command is not the same everytime.
3. What is the purpose of size\_limit option in copy?  
   To limit the size of the files loaded in snowflake, it is the sum of all files loaded, so if limit is reached by the time 2 files are loaded, 3rd file will not be loaded.
4. What is the RETURN\_FAILED\_ONLY option in copy?  
   If set to true, need to use on\_error=continue to see what were the failed records along with the passed records. Useless if specified as true without this on\_error option.
5. TRUNCATECOLUMNS?  
   Specifies whether to truncate text strings that exceed the target column length  
   if true, strings are automatically truncated to the target column length  
   if false, copy produces an error if a loaded string exceeds the target column length(DEFAULT)
6. FORCE option  
   Specifies to load all files regardless of whether they’ve been loaded previously and have not changed since they were loaded.  
   Potentially duplicates the data.  
   False by default
7. Load History  
   Enabled you to retrieve the history of the data loaded using copy into command  
   LOAD\_HISTORY is present as a view in the information\_schema and in snowflake.account\_stage.  
     
   The global LOAD\_HISTORY view will have global data, for example, if a table/view is just recreated, this view will have a table\_id which would be different for the same table/view, meaning the table was recreated, which is not visible in the information\_schema view.