slideshare a Scribd company

Search

Upload Login Signup

Snowflake SnowPro Certification - Dustin Liu

By liucong.haonan@gmail.com

Key Concepts & Architecture

Data Platform as a Cloud Service

- OLAP SaaS
- New SQL query based on public cloud with no-premises option
- Managed Service
- decoupled storage and computation

Architecture

A hybrid of traditional **shared-disk** and **shared-nothing** database architectures

Storage

- Stored managed by Snowflake
- Data objects are not directly visible (but through SQL)

Query processing

- Using 'Virtual warehouses' to process queries
- Each VW is a **independent** MPP(Massive Parallel Processing) **compute** cluster

Cloud Services

- Services that coordinate activities across Snowflake
  - Authentication, Infrastructure management, Metadata management etc

  from the cloud provider(AWS/Azure/GCP)
- Query planning and optimisation

Connecting to Snowflake

- Web/ CLI / ODB & JDBC / Native Connectors (Kafka/Python/Spark/Node.js/Go/.NET/PHP) / 3rd Party Connectors (Fivetran, Tableau etc)

Supported Cloud Platforms/Regions

Pricing

- Seperate pricing for computation and storage
- Pricing factors: region, editions, On-Demand or Capacity account
- On-Demand Storage:$40/TB Capacity Storage: $23/TB

Computation	Storage
- Charged for active warehouses **per hour** - Depending on the size of the warehouse - Billed by second (minimum of 1 min) - Charged in Snowflake credits	- **Monthly** storage fees - Based on average storage used per month - Cost calculated after compression - Cloud Providers

Regions

North America Europe Asia Pacific

- Supported platforms: AWS/Azure/GCP
- Regions(Data geographically/Compute resources geographically provisioned):
  - AWS (North America -4, EU-4, Asia Pacific-5)
- Supported platforms: AWS/Azure/GCP
- Regions(Data geographically/Compute resources geographically provisioned):
  - AWS (North America -4, EU-4, Asia Pacific-5)
  - Azure (North America -4, EU-3, Asia Pacific-3)
  - GCP(North America -1, EU-2, Asia Pacific-0)

* Excluding Government Regions
* Across region account is not supported, single account for each region.

Organisations

An organization is a first-class Snowflake object that links the accounts owned by your business entity

- By default, the maximum number of accounts in an organisation cannot exceed 25.
- Each account in your organisation can have its own set of users, roles, databases, and warehouses - billed for usage in all of your accounts on a single bill.

Virtual Warehouses

Overview of Warehouses

- Resources provision for operation SELECT and DML(DELETE, INSERT, UPDATE, COPY INTO)
- Started and stopped at anytime, resize at anytime, running queries not affected but only new queries
  - Auto-suspension and Auto-resumption are enabled by default and configurable

10 t-shirt sizes with X-Large as default:

- Impact on credit usage and billing

Warehouse Size	Credits / Hour	Credits / Second	Notes
X-Small	1	0.0003	Default size for warehouses created using CREATE WAREHOUSE.
Small	2	0.0006	
Medium	4	0.0011	
Large	8	0.0022	
X-Large	16	0.0044	Default for warehouses created in the web interface.
2X-Large	32	0.0089	
3X-Large	64	0.0178	
4X-Large	128	0.0356	
5X-Large	256	0.0711	Preview feature; currently only available on Amazon Web Services (AWS).
6X-Large	512	0.1422	Preview feature; currently only available on AWS.

- Impact on Data Loading(larger size **doesn't always** improve data loading performance
- Impact on Query Processing

The approach to choose size: starting with a smaller size -> monitor cost/performance -> resize if needed

The number of queries that a warehouse can concurrently process is determined by the size and complexity of each query

Multi-cluster Warehouses - Scale Out

Only available starting with Enterprise Edition

- Up to 10 warehouses
- Support all the same properties and actions as single warehouses

Mode:

- Maximised - same value for both max and min # of warehouses (value must be larger than 1)
  - Effective for statically controlling the available compute resources
- Auto-scale: different max and min for # of warehouses
  - Based on scaling policy

Scaling Policy	Strategy	Cluster Starts	Cluster Shuts Down
Standard(default)	- Favoring starting immediately - Costy but fast scaling	- immediately when a query is queued or one more query than the currently running	- After 2 to 3 consecutive successful checks (performed at 1 minute intervals)
Economy	- Favoring keeping running clusters fully loaded - Cheaper but longer waiting time	- Only if the system estimates there's enough query load to keep the warehouse busy for at least 6 minutes	- After 5 to 6 consecutive successful checks (performed at 1 minute intervals)

Scale-up vs Scale-out

Scale-up: Resizing to a larger warehouse is to accommodate more complex queries/workload

Scale-out: To accommodate more concurrent queries/consumption. Size and complexity of query determines the concurrency (also # queries being exectued)

Caching

Snowflake Cache Layers

Snowflake has three types of caching to optimize performance. Select the three types of caches from the list: Metadata, Warehouse and Results

Snowflake Caching

Result Cache

Local Disk Cache

Remote Disk

Service-layer:

Result Cache: Which holds the results of every query executed in the past 24 hours and available across virtual warehouses. (result caches can last for 31 days maximum)

Metadata Cache: Improve compile time for queries against commonly used tables

Virtual Warehouse Local Disk Cache

Local Disk Cache: Which is used to cache data used by SQL queries, data retrieved from the Remote Disk storage, and cached in SSD and memory.

Storage-layer:

Remote Disk: Which holds the long term storage. This level is responsible for data resilience

Release

About Release

- Release frequency: Weekly
- During a release, new customer requests/queries/connections transparently move over to the newer version
- Types: full, patch and behavior change release
- Staged Release Process: Three stages - early access for designated Enterprise account -> regular access for

Edition/Service Plan

Comparison

The Snowflake Edition that your organisation chooses determines the unit costs for the credits and the data storage you use.

Standard	Enterprise	Business Critical	Virtual Private
$2.7/Credit	$4/Credit	$5.4/Credit	Contact Snowflake

Note: The cost here is an estimation only

Monitoring

Resource Monitoring

- Only visible to ACCOUNTADMIN role by default (Website UI -> Account->Resource Monitors), can be set on either account or warehouse level.
- Monitor credit usage by user-managed VW and VW used by cloud services.
- Create by ACCOUNTADMIN only but can enable other roles to view/modify
- Limit can be set for a specified interval or date range:
  - Sending alter notification or
  - Suspending or both
- Not intended for hourly basis but day, week, month etc
- Not intended for setting precise limits

Notes: The credits used by SERVERLESS features are **not**

Loading Data

Types of Loading

Bulk Loading	Continuous Loading
- Most frequent method - Loading from stages - COPY command - Transformations possible	- Designed to load small volumes of data - Automatically once they are added to stages - Latest results for analysis - Snowpipe (Serverless feature)

Files:

External Stage	Internal Stage
- External cloud provider - S3 - GCP - Azure - Database object created in Schema	- Local storage maintained by Snowflake

File types:

- Structured: CSV(The most performant for loading), TSV ETC
- Semi-structured: JSON, Avro, ORC, Parquet, XML

Compression:

- uncompressed files -> automatically compressed to gzip(configurable)
- SF can automatically detect the compressed files: gzip, bzip2, deflate, raw_deflate
- SF can't automatically detect the following compression format: Brotli, Zstandard

Encryption:

- Unencrypted File: Auto encryptions using 128-bit keys or 256 bit keys (additional configuration)
- Encrypted File: key needs to be provided to SF

Planning a Data Load

Loading can affect query performance.
Dedicating separate warehouses for loading and querying operations to optimise performance.

- Normally, a smaller warehouse (Small, Medium, Large) would be sufficient.
- To optimise the parallel loading, produce data file roughly 100 - 250MB, compressed is recommended
- Loading very large files (e.g. 100 GB or larger) is not recommended.

Staging Data

Both internal and external references can include a path.
SF recommendations:

- Partitioning the data into logical path (such as geographical location, source identifiers, date)
- Organising data files by path will take advantage of parallel operations.

Loading Data

The COPY command supports:

- By path (internal stages) / prefix (Amazon S3 bucket)
- Specifying a list of specific files to load (fastest approach)
  - max 1,000 files
- The max file size default: 16MB
- Using pattern matching to identify specific files by pattern (slowest approach)
  - Applied differently for bulk loads vs Snowpipe loads
- copyOptions parameter:
  - Validation Mode: RETURN_N_ROWS, RETURN_ERRORS,RETURN_ALL_ERRORS - validate errors without actually loading into table
  - ON_ERROR - actions to follow
- Removing data from stage once loading is completed (REMOVE command)
- Loading semi-structured files
  - as type **VARIANT**
  - compounded values can be flatten into rows using FLATTEN function
- Transformation along loading is supported but **NOT** for all functions

Bulk Loading

Bulk Loading from a Local File System

- Step 1: Upload one or more data files to a Snowflake stage (various types) using the PUT command from local machine (not executable in Web UI)
- Step 2: Use the COPY INTO command to load the contents from stage to table

Types of Internal Stages

By default, each user and table is automatically allocated an internal stage. In addition, you can create named internal stages.

- User Stages (@~): convenient option for files accessed by single user but copying into multiple tables
  - limits: can't be altered or dropped; format setting not supported
- Table Stages (@%): convenient option for files accessed by multiple users but copying into single table
  - limits: can't be altered or dropped; format setting not supported
- Named Stages(@): Data objects, greatest degree of flexibility for data loading. Recommended when planning regular data loads that could involve multiple users or

Bulk Loading from AWS S3

- Creating S3 bucket
- Uploading files in S3
- Creating policies: multiple options
- Create integration object
- Loading from S3

Bulk Loading from GCP

Bulk Loading from Azure

Snowpipe

Loading Continuously Using Snowpipe

Snowpipe enables loading data from files as soon as they're available in a stage (Incremental loading), rather than manually executing COPY statements on a schedule to load larger batches (bulk loading)

- Using PIPE which is a named, first-class Snowflake object that contains a COPY statement used by Snowpipe
- Different mechanisms for detecting the staged files are available:

  object that contains a COPY statement used by Snowpipe
- Different mechanisms for detecting the staged files are available:
  - Automating Snowpipe using cloud messaging
  - Calling Snowpipe REST endpoints
- Snowpipe generally loads older files first but there is no guarantee that files are loaded in the same order they are staged
- Snowpipe uses **file loading metadata** associated with each pipe object to prevent reloading the same files
- Snowpipe charges are calculated by per-second/per-core granularity
- Check usages and charges:
  - Account -> Usage/Bill
  - SQL -> information schema -> PIPE_USAGE_HISTORY

Bulk Loading VS Snowpipe

Areas	Bulk Data Loading	Snowpipe
Authentication	Relies on the security options supported by the client for authenticating and initiating a user session	When calling the REST endpoints: Requires key pair authentication with JSON Web Token (JWT). JWTs are signed using a public/private key pair with RSA encryption
Load History	Stored in the metadata of the target table for 64 days. Available upon completion of the COPY statement as the statement output.	Stored in the metadata of the pipe for 14 days. Must be requested from Snowflake via a REST endpoint, SQL table function, or ACCOUNT_USAGE view
Transactions	Loads are always performed in a single transaction. Data is inserted into table alongside any other SQL statements submitted manually by users	Loads are combined or split into a single or multiple transactions based on the number and size of the rows in each data file. Rows of partially loaded files (based on the ON_ERROR copy option setting) can also be combined or split into one or more transactions
Compute Resources	Requires a user-specified warehouse to execute COPY statements	Uses Snowflake-supplied compute resources
Cost	Billed for the amount of time each virtual warehouse is active	Billed according to the compute resources used in the Snowpipe warehouse while loading the files.

Unloading Data

Overview

Similar to data loading, Snowflake supports bulk export (i.e. unload) of data from a database table into flat, delimited text files.

- Step1: COPY INTO command
  - Supporting queries with SELECT statement
  - Unloading into Single or Multiple Files (MAX_FILE_SIZE: default 16MB and max size 5G for AWS/GCP and 256MB for Azure)
  - Partitioned Data Unloading
  - Parameter SINGLE = TRUE to force unloading into a single file
- Step2: Download the file from stage

Unloading Features

Output Data File Details

Locations: Local files/S3/GCP Storage/Azure
Formats: CSV/TSV/JSON/Parquet
Encoding: UTF-8
Compression: gzip, bzip2, Brotli, Zstandard
Encryption:

- Internal: 128-bit by default or 256-bit keys enabled with additional configuration
- External: User-supplied key

Unloading into a Snowflake Stage

Step 1: COPY INTO command to copy the data from into one or more files in a Snowflake stage. This step will require virtual warehousing running

Step 2: GET command to download the data

Unloading into Amazon S3

Step 1: Use the COPY INTO command to copy data into S3 bucket

Step 2: Use the interfaces/tools provided by Amazon to download the files from the S3 bucket.

Databases, Table & View Types

Micro-partition

All SF tables are divided into micro-partition

Traditional DW	Snowflake
- static partitioning	- micro-partition(immutable)
Issues: - maintenance - data skew	Benefits: - no need to explicitly define up-front - small in size: 50 - 500MB before compression(16MB with compression) - range of value overlap

Order of data ingestion is used to partition and metadata of the micro-partitions are stored in the Cloud Service layer:

- User can query without active warehouse - column column, which enables Snowflake to perform the following actions for queries on the table:
- First, prune micro-partitions that are not needed for the query.
- Then, prune by column within the remaining micro-partitions.

Clustering Keys

Considering clustering keys when:

- table contains multiple terabytes (TB) of data
- The queries can take advantage of clustering
- A high percentage of the queries can benefit from the same clustering key(s)
- Table is large enough to reside on many micro-partitions

Indicators:

- Queries are running slower
- Clustering ration for the table is very low and clustering depth is very large

Note: Clustering keys are not intended for all tables and changing clustering key on a table incur cost.

Type of Tables

	Permanent	Transient	Temporary
Command	CREAT TABLE	CREATE TRANSIENT TABLE	CREATE TEMPORARY TABLE
	- Fail safe supported(7 days)	supported - Persists until dropped - contribute to the overall storage charges	supported - data is purged after session ends - Not visible to other users and not recoverable - Contribute to storage cost can have the same name as another non-temp tables
User cases	Permanent data	Only for data that doesn't need to be protected	Non-permanent data

External table:
Allows access to data stored in external stage through COPY INTO command

Type of Views

- Non-materialized views (usually simply referred to as "views")
  - Result are not stored
- Materialized views.
  - Result is stored
  - Faster performance than normal views
  - Contributes towards storage cost
  - Can query only a single table. Joins, self-joins are not supported
- Secure view
  - Why: No wish to expose the underlying tables or internal structural details for a view
  - When: Specifically designated for data privacy
  - Both non-materialized and materialized view can be defined as Secure view
  - Performance is impacted as internal optimisation is bypassed

Access Control

Approach

- Discretionary Access Control (DAC): Each object has an owner, who can in turn grant access to that object.
- Role-based Access Control (RBAC): Access privileges are assigned to roles, which are in turn assigned to users.

Access to securable objects is allowed via privileges assigned to roles, which are in turn assigned to other roles or users.

System Defined Roles

Role	Desc
ORGADMIN	Organization Administrator, organization level operation
ACCOUNTADMIN	Encapsulates the SYSADMIN and SECURITYADMIN
SECURITYADMIN	Role that can manage any object grant globally, as well as create, monitor, and manage users and roles
USERADMIN	Dedicated to user and role management only
SYSADMIN	Create warehouses and databases
PUBLIC	When explicit access control is not needed(default role for a new user)

Privileges

ACCOUNTADMIN SECURITYADMIN SYSADMIN USERADMIN Custom Role PUBLIC

For best practice, Custom Roles are normally created by SECURITYADMIN and then assigned to SYSADMIN.

Data Sharing

Overview

The following Snowflake database objects can be shared:

- Tables
- External tables
- Secure views
- Secure materialised views
- Secure UDFs

All database objects shared are **read-only**, no actual data is copied or transferred, all sharing is accomplished through unique **services layer** and **metadata store**.

- No storage cost
- Compute resource cost only
- Only AccountAdmin can create sharing
- No limit of number of shares but 1 database only

Share

- Between SF accounts
- The privileges that grant access to the database(s) and the schema containing the objects to share.
- The privileges that grant access to the specific objects in the database.
- The consumer accounts with which the database and its objects are shared.

Reader Accounts

- For users who does not already have a Snowflake account.
- DML tasks not allowed.

Security

Features

The following features are specific to either Business Critical or Enterprise Edition, otherwise supported by All editions

Enterprise Edition	Business Critical Edition
- Configure session policies - Periodic rekeying of encrypted data	- Private communication between the VPC and the Snowflake - Private communication to internal stage - Support for encrypting data using customer-managed keys - HIPAA, PCI DSS, HITRUST CSF, FredRAMP Moderate, IRAP Protected compliance

Authentication

Federated authentication enables your users to connect to Snowflake using secure SSO (single sign-on)

Authentication

Federated authentication enables your users to connect to Snowflake using secure SSO (single sign-on)

Native supported	Other SAML 2.0-compliant vendors
- Okta - Microsoft ADFS	- Google G Suite - Microsoft Azure Active Directory - OneLogin - Ping Identity PingOne

MFA

At a minimum, **ACCOUNTADMIN** role should be assigned to use MFA.

By using **DUO Mobile** app should be sufficient and MFA token caching can be combined with connection caching in federated single sign-on.

MFA is supported for following connection method:Web/Python/SnowSQL/ODBC/JDBC

Admin& Authorization

Sessions

A session is independent of an identity provider(IdP) session.

open a worksheet -> max 4 hours(default) session -> Snowflake logs the user of out the web interface after the 4 hour limit expires for **any open** worksheet

Session Policies

A session policy defines the idle session timeout period in minutes and provides the option to override the default idle timeout value of 4 hours.
It is **schema-level** objects and current **NOT** applicable to Snowsight

Data encryption

Snowflake supports both internal (Snowflake-provided, image B) and external (customer-provided, image A) stages for data files, Data is end-to-end encrypted.

(A) Customer-provided Staging Area

(B) Snowflake-provided Staging Area

1. A user uploads one or more data files to a stage.
   - Image A - user may optionally encrypt the data files
   - Image B - automatically encrypted by the Snowflake client on the user's local machine.
2. The user loads the data from the stage into a table.
3. The user can unload query results into an external or internal stage.
4. The user downloads data files from the stage and decrypts the data on the client side.

Snowflake managed keys:
Hierarchical Key Model, all Snowflake-managed keys are automatically rotated by Snowflake when they are more than **30** days.

Root Key
Account Master Keys
Table Master Keys
File Keys

Zero-copy Cloning

Zero-Copy Cloning

- A way to quickly take a "snapshot" of any table, schema, or database
- No additional costs (until changes are made to the cloned object)
- Each change to the clone results in new micro-partitions (new cost)
- Cloned object does not inherit the source's privileges
  - Only child objects retain privileges of the source (if database is cloned -> schema and tables retain; if schema is cloned -> tables retains)
- Cloning a database or schema:
  - External named stages are cloned
- Objects for clone:
  - Database Objects
  - External named stages
  - Snowpipes
  - Sequences
  - Streams
  - Tasks
- Tables are cloned
- Internal named stages are **NOT** cloned

Time Travel & Fail-Safe

Overview

Snowflake Time Travel enables accessing historical data at any point within a defined period.

Continuous Data Protection Lifecycle

Time travel:

- Standard Edition: 0 -1 day
- Enterprise Edition: 0 - 90 days

Fail-Safe

Fail-safe: 0 - 7 days, post the end of time travel period.

Continuous Data Protection Lifecycle

- Not provided as a means for accessing historical data after the Time Travel retention period has ended. failures.

Data Governance

Overview

high-level summary of the features

Features	Editions
Masking columns - masking policies	Enterprise or higher
Row access control - row access policies	Enterprise or higher
Access history - Access History view	Enterprise or higher
Tracking sensitive data - tags	Enterprise or higher

Column-level Security

- Dynamic Data Masking
  - schema level through masking policies
- External Tokenization
  - Enable accounts to tokenize data before loading it into SF, this is through external function and it is on schema-level

Row-level Security

A row access policy is a schema-level object that determines whether a given row in a table or view can be viewed from the following types of statements:

- SELECT statements
- Rows selected by UPDATE, DELETE, and MERGE statements.

It is defined through row access policy

Access History

Access History:

Through `SNOWFLAKE.ACCOUNT_USAGE.ACCESS_HISTORY` and this view is last for 1 year.

Others

Query History

- UI history will be saved for 14 days
- select * from table(information_schema.query_history()) can query history back in 7 days
- snowflake.account_usage.query_history can query up to 1 year

Access History

- select * from snowflake.account_usage.access_history

Virtual Warehouses

- Credit consumption based on warehouse size and # of clusters for the warehouse

Data Sharing

Consumer accounts must reside in the same Cloud Provider Region as the Provider account

Show vs Information Schema

The INFORMATION_SCHEMA views provides a SQL interface to the same information provided by the SHOW commands but with some key differences.

Show:

- Not required to execute WH but Information_schema does.
- Case insensitive when filter but information_schema is case-sensitive
- Result limited to the current schema by default and with information_schema, views display all objects in the current database

UDF

Supported programming language: SQL, Java, JavaScript

Warehouse resizing

The size of the cache is determined by the compute resources in the warehouse

- Cache is dropped when the warehouse is suspended -> trade-off between saving credits vs maintaining the cache
- Cache is dropped while the size of warehouse is decreased -> trade-off between saving credits vs maintaining the cache

Search Optimisation

The search optimisation service aims to significantly improve the performance of selective point lookup queries on tables.
User case:

- Business users who need fast response times for critical dashboards with highly selective filters.
- Data scientists who are exploring large data volumes and looking for specific subsets of data.

Snowflake SnowPro Core Cert CheatSheet.pdf

Dec 30, 2022 • 8 likes • 5,533 views

Download Now

Download to read offline

Data & Analytics

Thanks for your reading/downloading, it would be highly appreciated if you could clap on the original post of this cheatsheet at https://medium.com/@grdustin/snowflake-snowpro-core-cert-cheatsheet-ead4a01a5428. Cheers!

D Dustin Liu Follow

Recommended

Snowflake SnowPro Certification Exam Cheat Sheet
Jeno Yamma
78.9K views • 7 slides

An overview of snowflake
Sivakumar Ramar
1.1K views • 23 slides

Snowflake essentials
qureshihamid
1.4K views • 14 slides

Snowflake Overview
Snowflake Computing
20.6K views • 21 slides

Snowflake Data Loading.pptx
Parag860410
290 views • 19 slides

Snowflake Best Practices for Elastic Data Warehousing
Amazon Web Services
5.5K views • 35 slides

More Related Content

What's hot (20)

Let's get to know Snowflake
Knoldus Inc. • 325 views

Zero to Snowflake Presen…
Brett VanderPla… • 2.4K views

Load & Unload Data TO an…
Faysal Shaarani … • 8.5K views

Snowflake Data Governance
ssuser538b022 • 261 views

Azure purview
Shafqat Turza • 2.4K views

Slideshows for you (20)

How People Are Leveragin…
Roy Ahuja • 1.4K views

Data Design: Where Math …
Trina Chiasson • 90.1K views

Data cleaning and visualiz…
Tapan Gautam • 326 views

Crowdsourcing presentati…
kleepg23 • 4.8K views

Machine Learning: A Fast …
Ahmad Ali Abin • 20.1K views

SCRIBD + slideshare

Access the ultimate library with Scribd.

Try it free for 60 days

Recently uploaded (20)

SAC 24 HALO 2022 Nation…
Elaine Werffeli • 6 views

Journey of Generative AI
thomasjvarghese49 • 5 views

The Business Tycoons (Ja…
Global India Busin… • 11 views

CIV208 __ FLUID MECHANI…
Mrunmayee Manjari • 6 views

From Lab to Life - Lessons…
StephenLeo7 • 15 views

About Support Terms Privacy Copyright Cookie Preferences Do not sell or share my personal information English

© 2023 SlideShare from Scribd