Data Engineering Latam

snowflake

Query Processing (Virtual Warehouse) Compute Layer

Yazmin Abat A.

October 5th, 2022

Data Engineering Latam

snowflake

- **Agenda**
  - Compute Warehouses
  - WareHouse Sizes
  - Multi-clustering – Table Structures
  - Strategies
  - Monitoring
- **Additional Resources**
  - Literature & Quickstarts
  - Github and Bibliography

Data Engineering Latam

snowflake

Agenda

1	**Introduction**
2	Compute Warehouses
3	Warehouse Sizes
4	MultiClustering –Table Structures
5	Strategies
4	Monitoring Snowflake Compute Layer

Snowflake
= SaaS **Datawarehouse**

Data Warehouse

- Structured data
- Stored in tables
- Schema on read
- Huge volume

Data Lake

- Unstructured data (files)
- Stored in object storage
- Schema on read
- Massive volume

SaaS= Software as a Service

New Pizza as a Service

Traditional On-Premises Deployment	Infrastructure as a Service (IaaS)	Platform as a Service (PaaS)	Software as a Service (SaaS)
Kitchen	Kitchen	Kitchen	Kitchen
Gas	Gas	Gas	Gas
Oven	Oven	Oven	Oven
Pizza Dough	Pizza Dough	Pizza Dough	Pizza Dough
Toppings	Toppings	Toppings	Toppings
Cook the Pizza	Cook the Pizza	Cook the Pizza	Cook the Pizza
Made In-House	Kitchen-as-a-Service	Walk-In-and-Bake	Pizza-as-a-Service

You Manage Vendor Manages

SaaS= Software as a Service

On-Premises (Private Cloud)	Infrastructure (as a Service)	Platform (as a Service)	Software (as a Service)
Data & Access	Data & Access	Data & Access	Data & Access
Applications	Applications	Applications	Applications
Runtime	Runtime	Runtime	Runtime
Operating System	Operating System	Operating System	Operating System
Virtual Machine	Virtual Machine	Virtual Machine	Virtual Machine
Compute	Compute	Compute	Compute
Networking	Networking	Networking	Networking
Storage	Storage	Storage	Storage

You Manage

Cloud Provider Manages

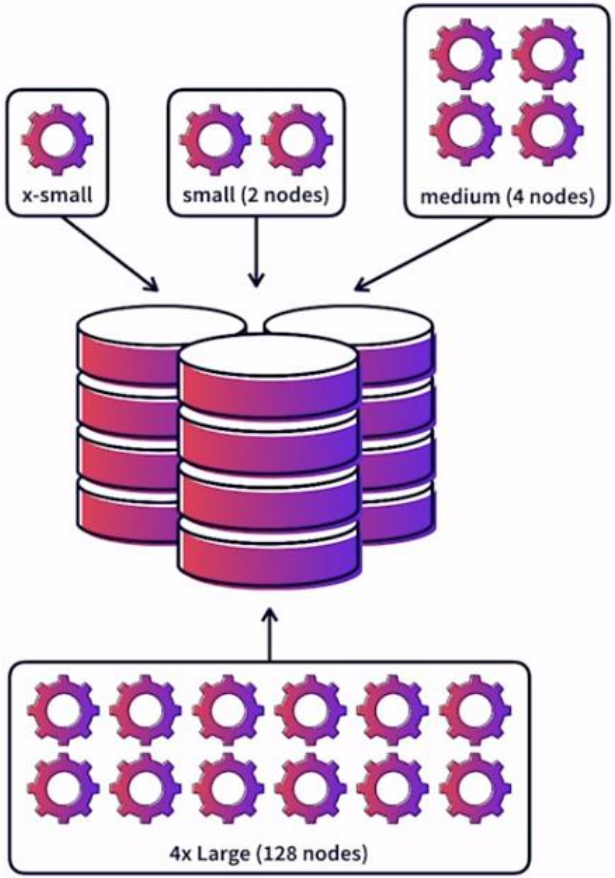
Source https://learn.microsoft.com/en-us/training/modules/fundamental-azure-concepts/categories-of-cloud-services

From Ingestion to Reporting

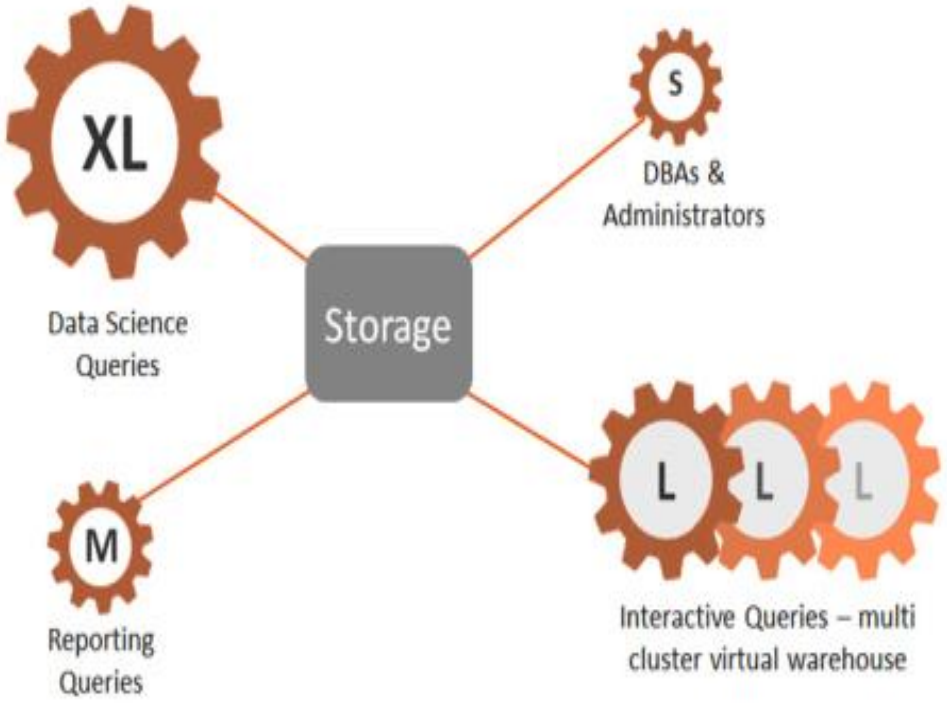
Key Features of SnowflakeDB

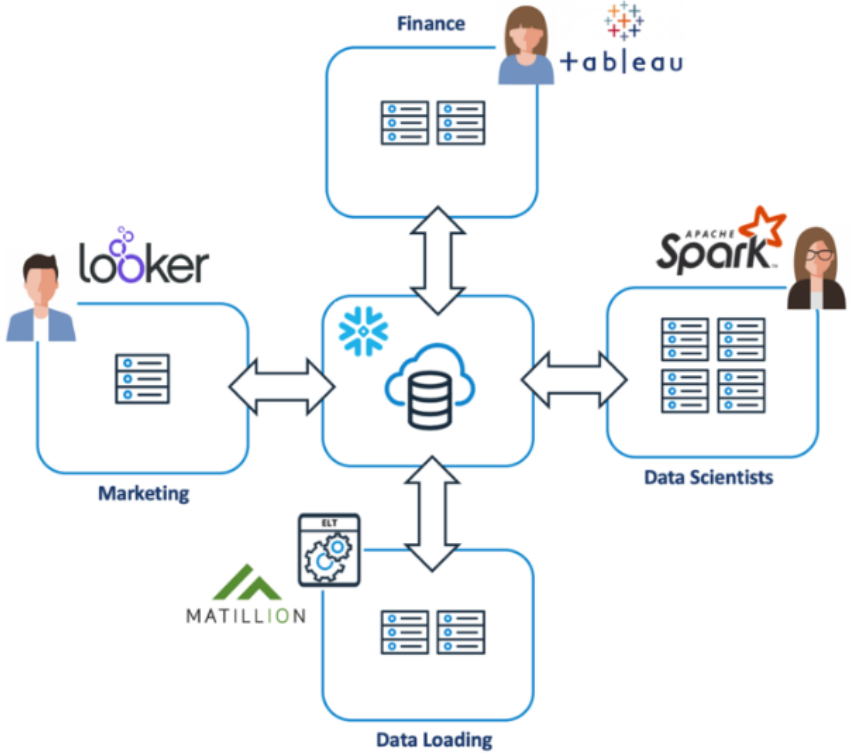
1. Virtual and multicluster warehouses
2. Time travel and zero-copy cloning
3. Data sharing and marketplace
4. High availability via cross-cloud replication
5. Granular security and automatic encryption

*ALSO: Strong support for JSON data



Separation of Snowflake workloads





Source: Snowflake /Mahmood & Sharif (2021)

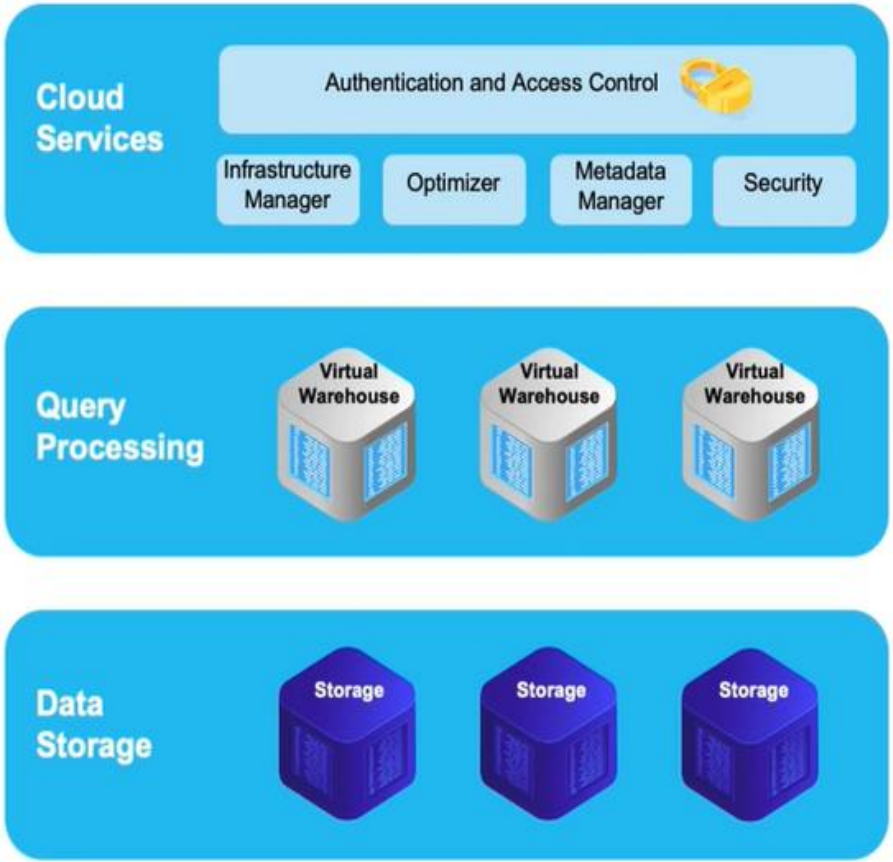
Data Engineering Latam

snowflake

Agenda

1	Introduction
2	**Compute Warehouses**
3	Warehouse Sizes
4	MultiClustering –Table Structures
5	Strategies
4	Monitoring Snowflake Compute Layer

Query Processing = Virtual Warehouses



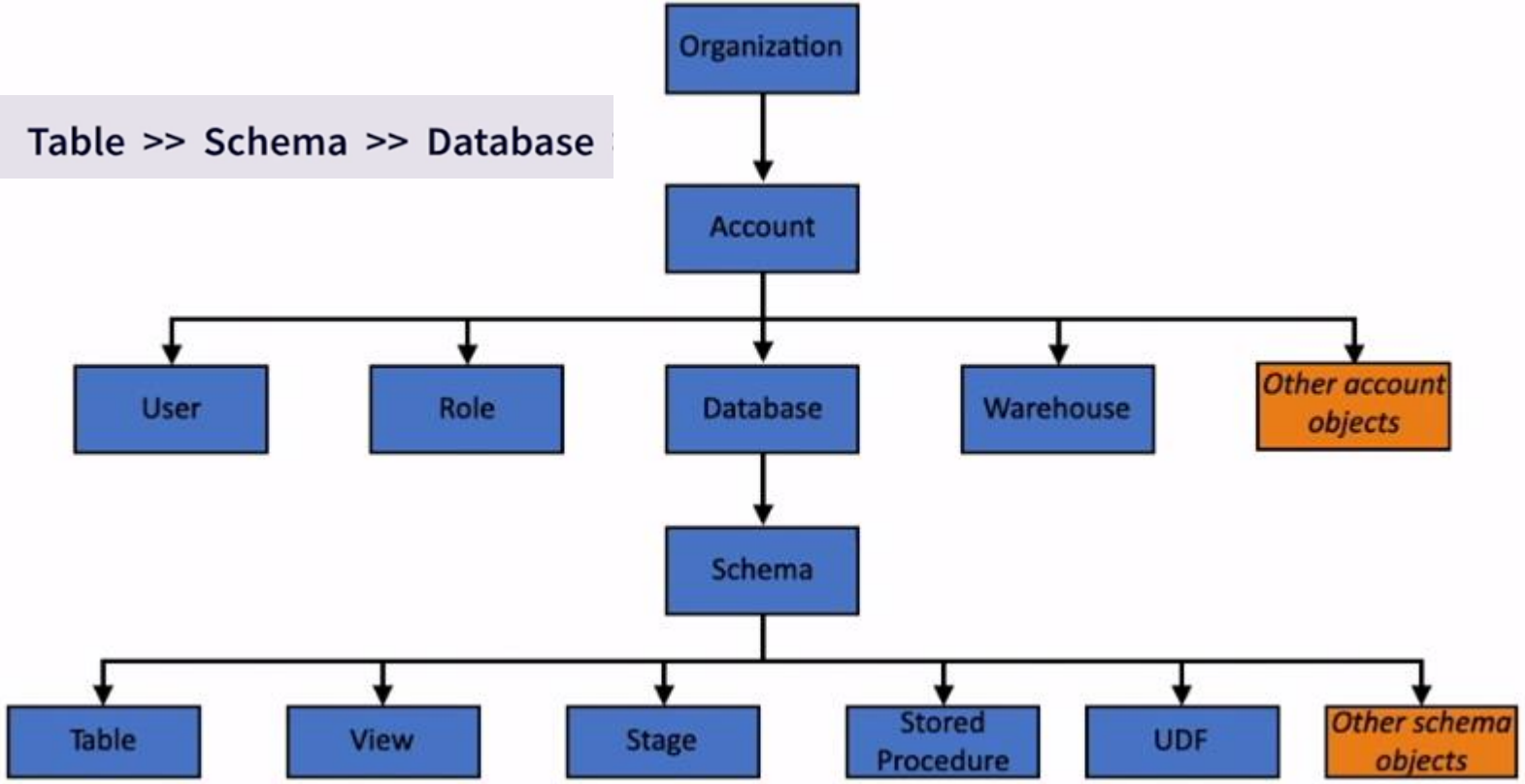
- Compute/Queries
- Billed by seconds of usage
- Paused when not in use
- Flexible and scalable (account-> usage)

- Cluster of servers
  - CPU, memory, temp storage
- Query caching
- Elastic sizing
- On-demand
  - Powered on/off
  - Millisecond availability
- Pay-as-you-go
  - Per second (min 60 seconds)
- Automatic suspend and resume

```
USE ROLE SYSADMIN;
CREATE WAREHOUSE WH_DEL
WITH WAREHOUSE_SIZE =
MEDIUM Auto_suspend = 300
Auto_resume = true
Initially_suspended = true;
```

Object Hierarchy

Table >> Schema >> Database



- Access data/objects from the **past** which may have been:
  - Deleted
  - Modified
  - Dropped
- Restore dropped tables, schemas, or databases
- Clone tables, schemas, or databases **at** or **before** a specific time

CDP(Continuous Data Protection) & Time Travel

Snowflake Tables	Permanent	Transient	Temporary	External
Persistence	Until explicitly dropped	Until explicitly dropped	Remainder of session	Until explicitly dropped
Time Travel Retention (Days)	0 – 90 days*	0 or 1	0 or 1	0
Fail-Safe Period (Days)	7	0	0	0
Cloning Possible	Yes	Yes	Yes	No
Create Views Possible	Yes	Yes	Yes	Yes
*Enterprise Edition and above 0-90 days. Standard Edition 0 or 1 day.				

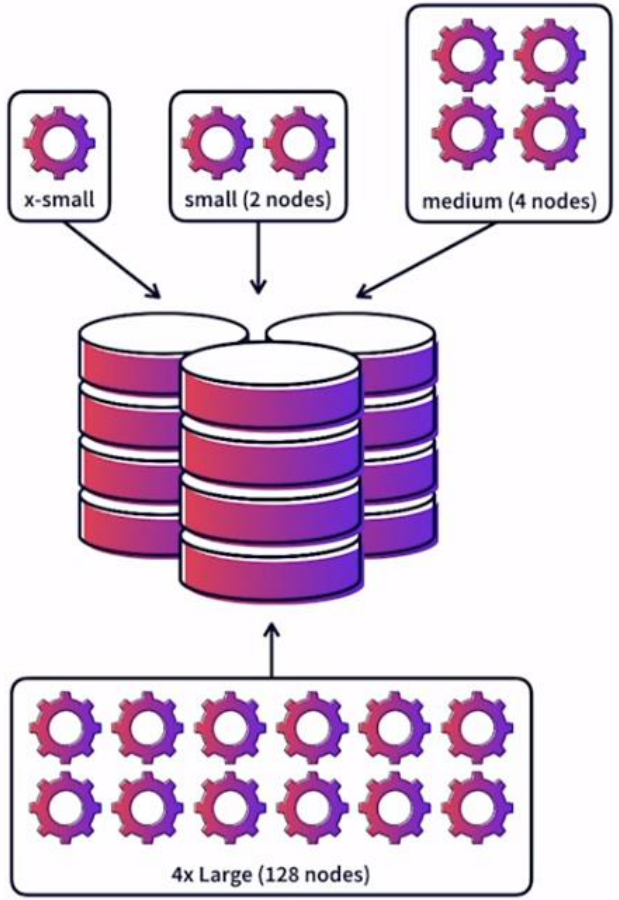
Data Engineering Latam

snowflake

Agenda

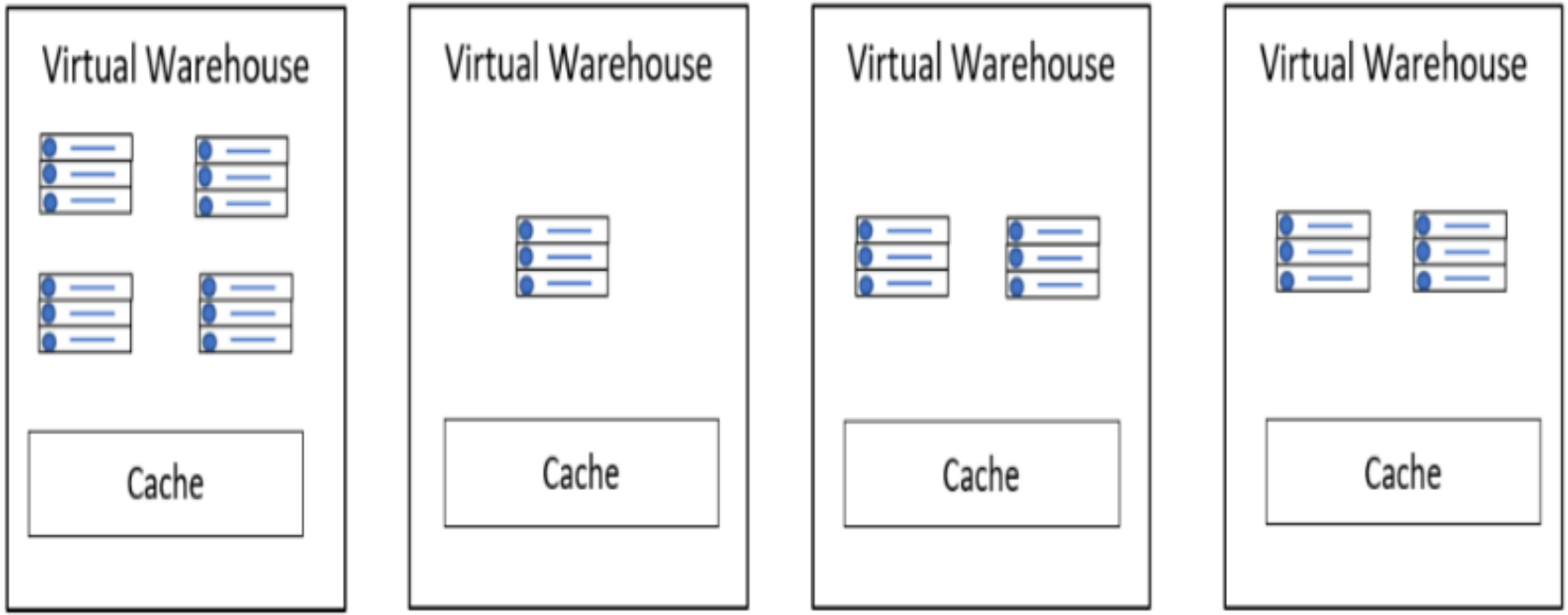
1	Introduction
2	Compute Warehouses
3	**Warehouse Sizes**
4	MultiClustering –Table Structures
5	Strategies
4	Monitoring Snowflake Compute Layer

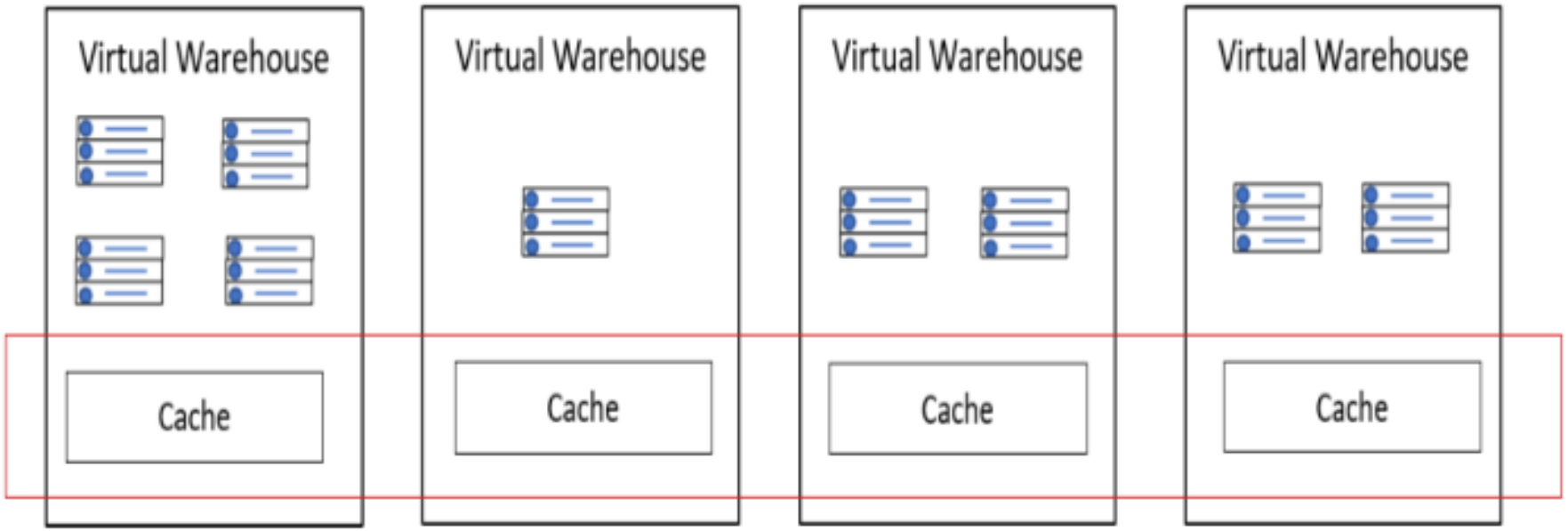
Credits



Virtual Warehouse	Credits
X-Small	1
Small	2
Medium	4
Large	8
X-Large	16
2X-Large	32
3X-Large	64
4X-Large	128
5X-Large	256
6X-Large	512

Cache


Virtual Warehouse
Cache
Virtual Warehouse
Cache
Virtual Warehouse
Cache
Virtual Warehouse
Cache



Cache (SSD or Solid-State Drive): Which holds the results of every query executed in the past 24 hours.

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

Scaling up a Warehouse Increases Cluster Size

Original Warehouse

Virtual Warehouse Small

Scaled Up version of Original Warehouse

Virtual Warehouse Medium

Further Scaled Up version of Original Warehouse

Virtual Warehouse Large

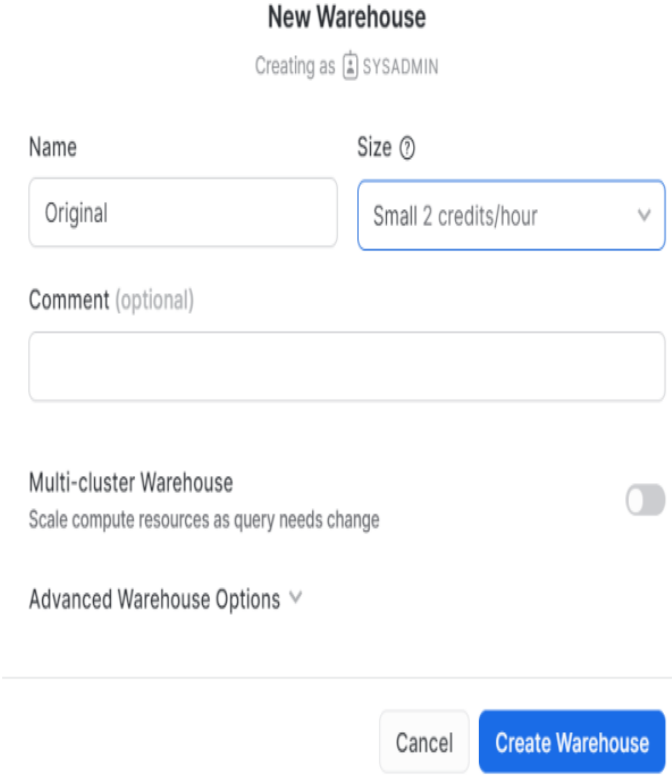
Scaling out a Warehouse Increases Number of Clusters

Original Warehouse	Scaled Out version of Original Warehouse	Further Scaled Out version of Original Warehouse
Virtual Warehouse Small	Virtual Warehouse Small	Virtual Warehouse Small
	Virtual Warehouse Small	Virtual Warehouse Small
		Virtual Warehouse Small

Creating a new warehouse using SQL or WEB UI

```
CREATE WAREHOUSE ACCOUNTING WITH Warehouse_Size = MEDIUM MIN_CLUSTER_COUNT = 1 MAX_CLUSTER_COUNT = 6
SCALING_POLICY = 'STANDARD';
```

SNOWFLAKE METADATA:
Includes table definitions and references to the micro-partition files for that table. The range of values in terms of MIN and MAX, the NULL count, and the number of distinct values are captured from micro-partitions and stored in Snowflake. As a result, any queries which return the MIN or MAX
value, for example, will not need a running warehouse.



Data Engineering Latam

snowflake

Agenda

1	Introduction
2	Compute Warehouses
3	Warehouse Sizes
4	**MultiClustering –Table Structures**
5	Strategies
4	Monitoring Snowflake Compute Layer

Table Structures in Snowflake

Snowflake stores the total number of **micro-partitions** and
the depth of overlapping micro-partitions to provide information about **clustering.**

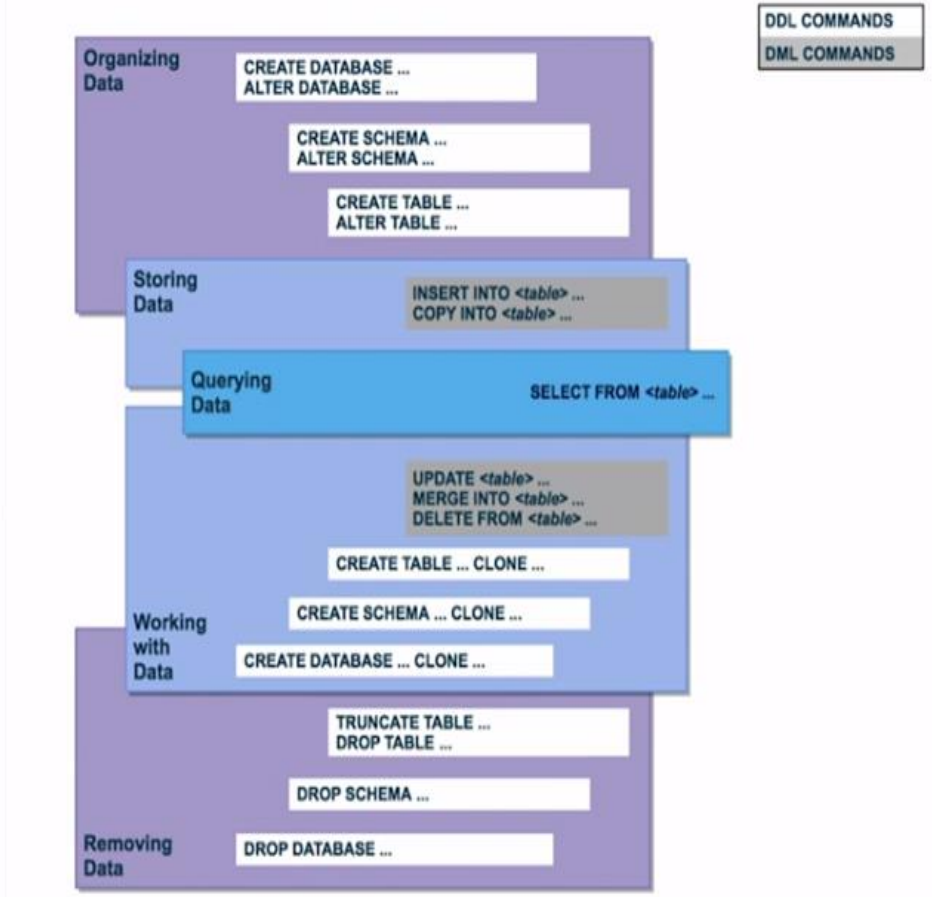
Micropartitions

All data in Snowflake automatically divided into micro-partitions - no explicit partitioning done by user

Each micro-partition contains 50MB to 500MB uncompressed data

Rows in tables mapped to micro-partitions and organized in a columnar fashion

Large tables can contain thousands, even millions of micro-partitions



Clustering Benefits

Improve queries by skipping data that does not match the filter predicate

Better column compression

Only configuration required is to correctly define the clustering key

Improves performance of equality searches as well as range searches

Clustering Benefits

Logical Structure

type	name	country	date
2	A	UK	11/2
4	C	SP	11/2
3	C	DE	11/2
2	B	DE	11/2
3	A	FR	11/2
2	C	SP	11/2
3	Z	DE	11/2
2	B	UK	11/2
4	C	NL	11/2
5	X	FR	11/3
1	A	NL	11/3
5	A	FR	11/3
2	X	FR	11/2
4	Z	NL	11/2
2	Y	SP	11/2
1	B	SP	11/3
5	X	DE	11/3
3	A	UK	11/4
1	C	FR	11/3
4	Z	NL	11/4
5	Y	SP	11/4
5	B	SP	11/5
3	X	DE	11/5
2	Z	UK	11/5

Physical Structure

	Micro-partition 1 (rows 1-6)			Micro-partition 2 (rows 7-12)			Micro-partition 3 (rows 13-18)			Micro-partition 4 (rows 19-24)		
type	2	4	3	3	2	4	2	4	2	1	4	5
	2	3	2	5	1	5	1	5	3	5	3	2
name	A	C	C	Z	B	C	X	Z	Y	C	Z	Y
	B	A	C	X	A	A	B	X	A	B	X	Z
country	UK	SP	DE	DE	UK	NL	FR	NL	SP	FR	NL	SP
	DE	FR	SP	FR	NL	FR	SP	DE	UK	SP	DE	UK
date	11/2	11/2	11/2	11/2	11/2	11/2	11/2	11/2	11/2	11/3	11/4	11/4
	11/2	11/2	11/2	11/3	11/3	11/3	11/3	11/3	11/4	11/5	11/5	11/5

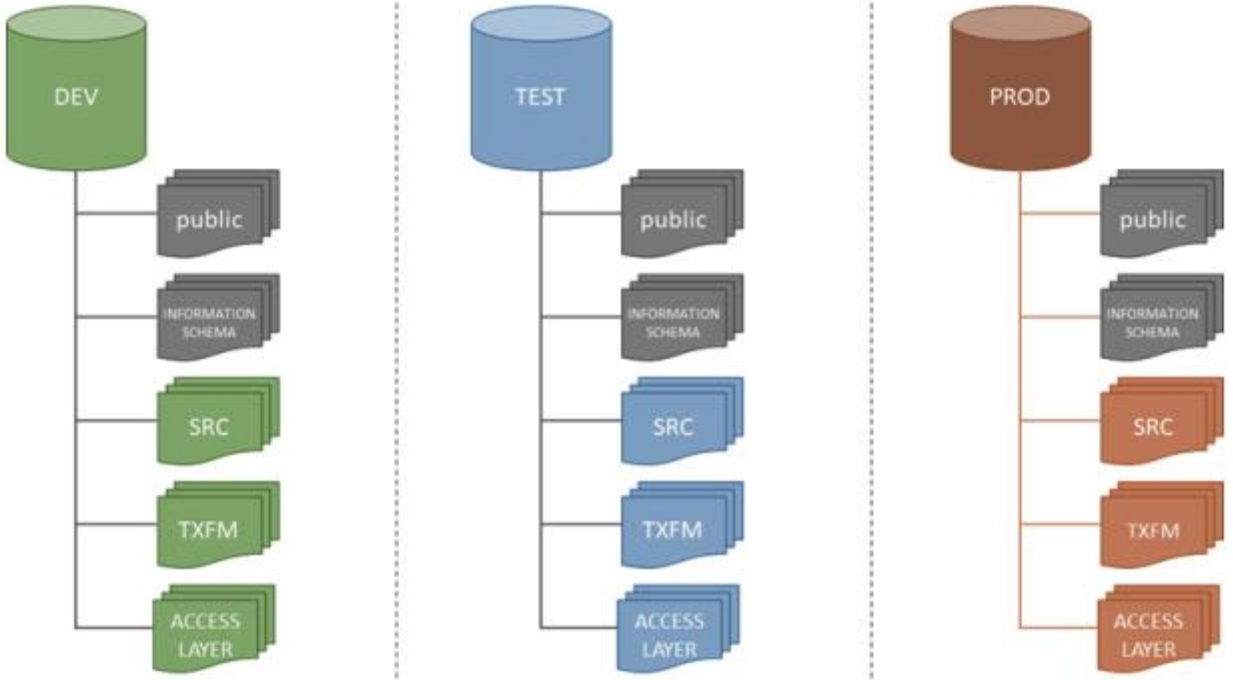
Source: https://docs.snowflake.com/en/user-guide/tables-clustering-micropartitions.html

Data Engineering Latam

snowflake

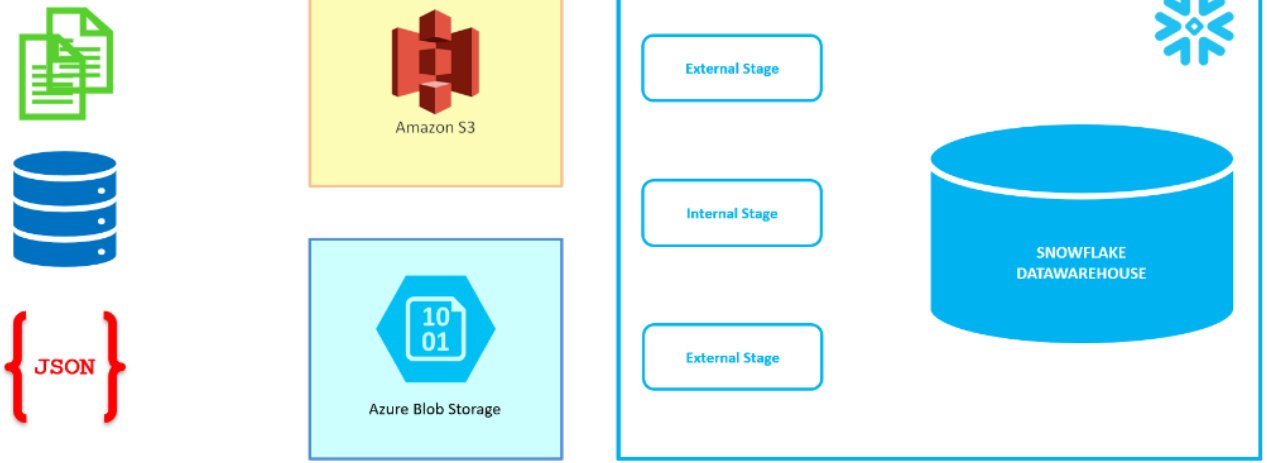
Agenda

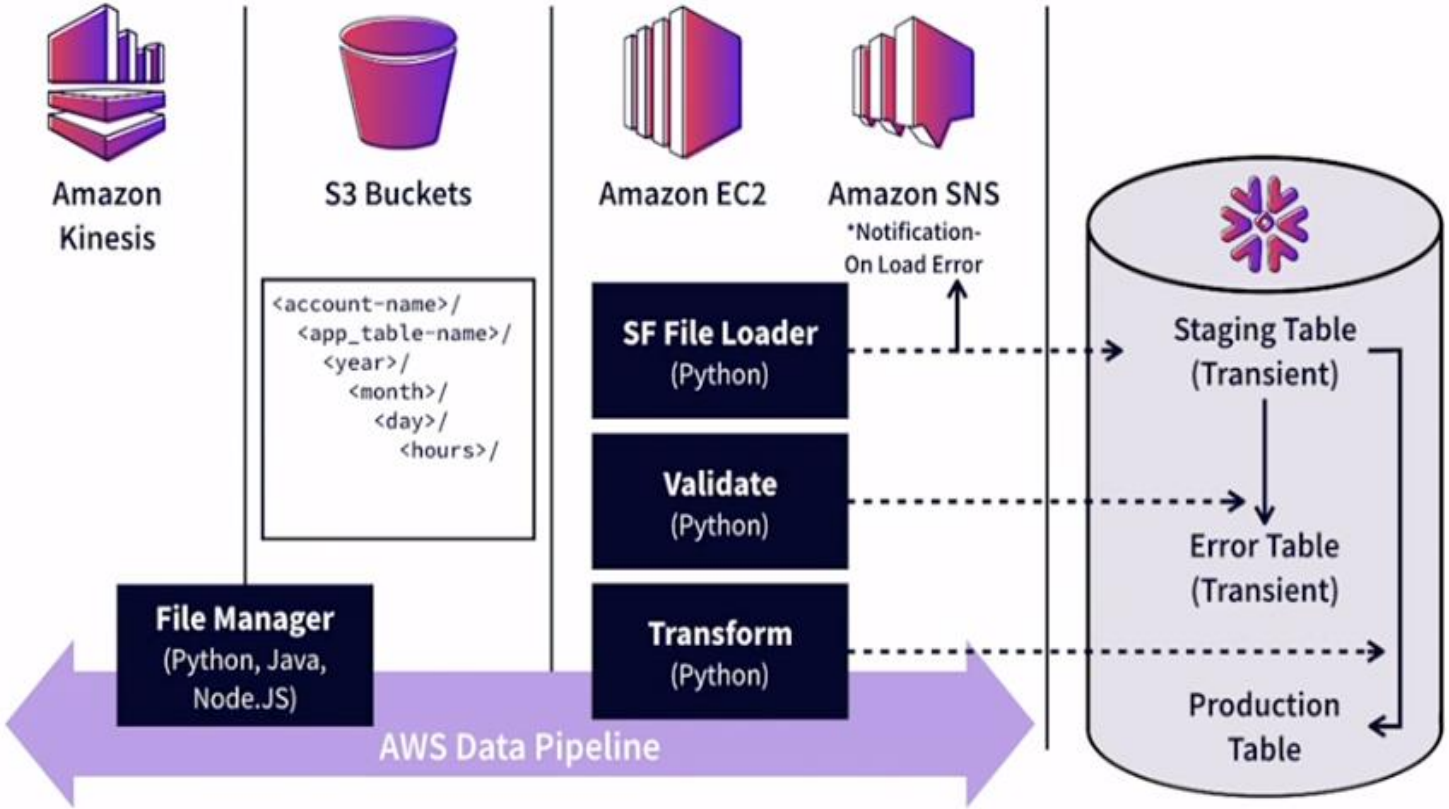
1. Introduction
2. Compute Warehouses
3. Warehouse Sizes
4. Multiclustering -Table Structures
5. **Strategies**
4. Monitoring Snowflake Compute Layer

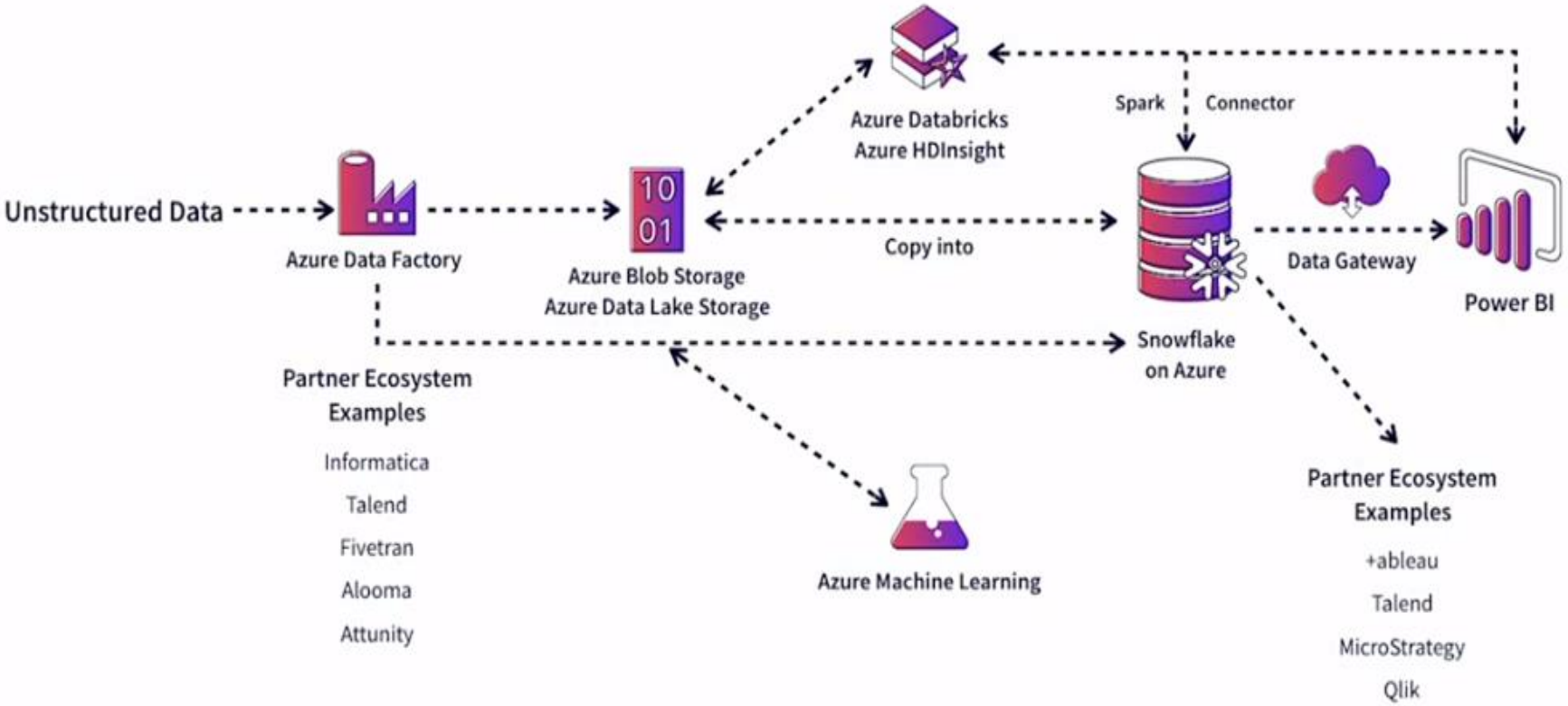

DEV
public
INFORMATION SCHEMA
SRC
TXFM
ACCESS LAYER
TEST
public
INFORMATION SCHEMA
SRC
TXFM
ACCESS LAYER
PROD
public
INFORMATION SCHEMA
SRC
TXFM
ACCESS LAYER

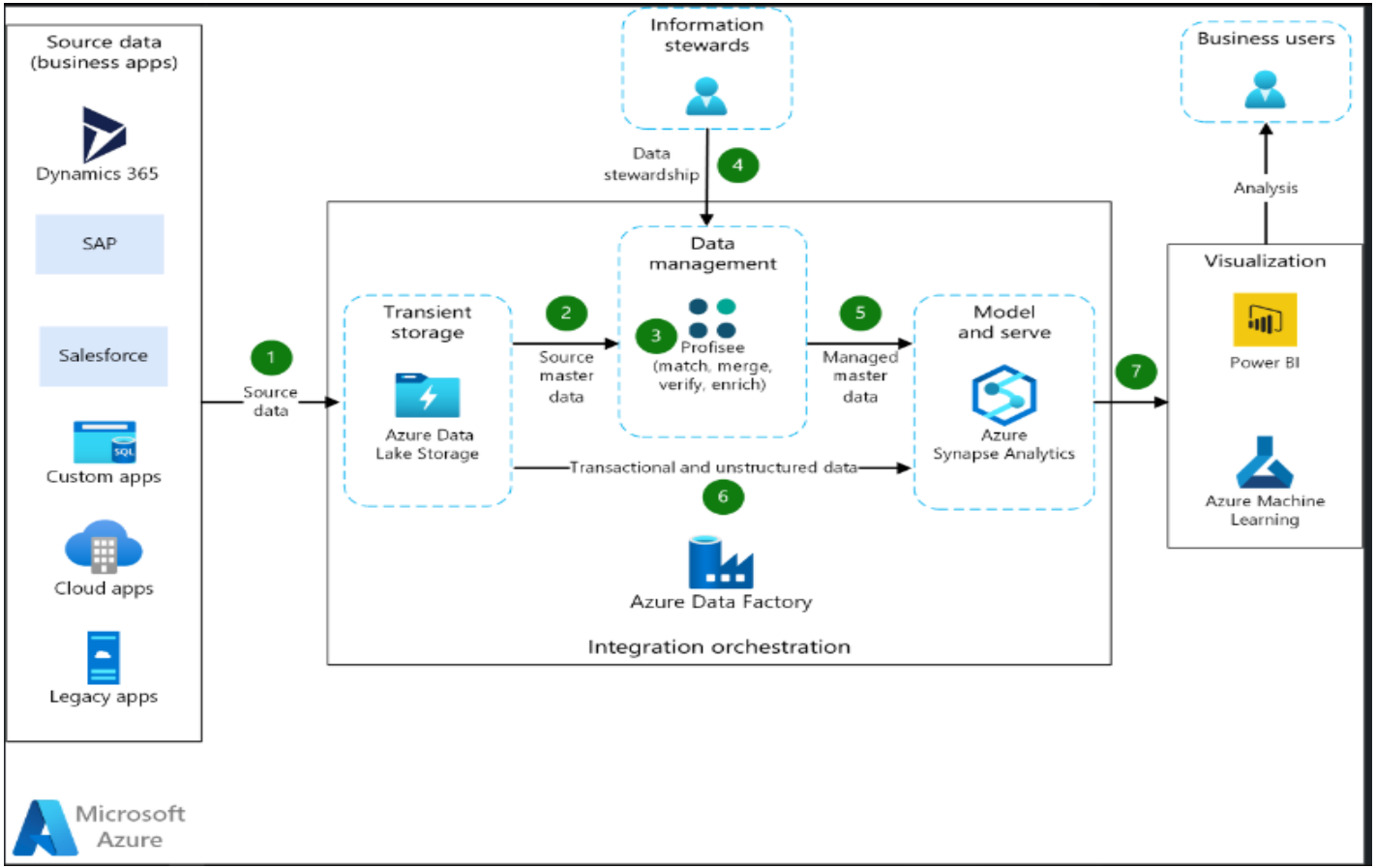
Load Process

1. Source from bucket.
2. Create or use file format.
3. Load with Web UI (max 100 MB).




Amazon Kinesis
S3 Buckets
Amazon EC2
Amazon SNS
*Notification-On Load Error
<account-name>/
<app_table-name>/
<year>/
<month>/
<day>/
<hours>/
SF File Loader (Python)
Validate (Python)
Transform (Python)
File Manager (Python, Java, Node.JS)
AWS Data Pipeline
Staging Table (Transient)
Error Table (Transient)
Production Table


Unstructured Data
Azure Data Factory
10
01
Azure Blob Storage
Azure Data Lake Storage
Azure Databricks
Azure HDInsight
Spark
Connector
Copy into
Snowflake
on Azure
Data Gateway
Power BI
Partner Ecosystem
Examples
Informatica
Talend
Fivetran
Alooma
Attunity
Azure Machine Learning
Partner Ecosystem
Examples
+ableau
Talend
MicroStrategy
Qlik


Source data (business apps)
Dynamics 365
SAP
Salesforce
Custom apps
Cloud apps
Legacy apps
1
Source data
Information stewards
Data stewardship
4
Data management
Transient storage
Azure Data Lake Storage
2
Source master data
3
Profisee (match, merge, verify, enrich)
5
Managed master data
Model and serve
Azure Synapse Analytics
Transactional and unstructured data
6
Azure Data Factory
Integration orchestration
7
Business users
Analysis
Visualization
Power BI
Azure Machine Learning
Microsoft Azure


CSV
JSON
AVRO
ORC
Parquet
XML
File Formats

CSV, JSON, XML

```
COPY INTO [TABLENAME] ([COLUMN]) FROM @[STAGENAME]
FILE_FORMAT=(TYPE = 'CSV' FIELD_DELIMITER = '|' SKIP_HEADER = 1)
PATTERN= '.*.csv';

COPY INTO [TABLENAME] ([COLUMN]) FROM @[STAGENAME]
FILE_FORMAT=(TYPE = 'JSON' STRIP_OUTER_ARRAY = TRUE)
PATTERN= '.*.json';

COPY INTO [TABLENAME] ([COLUMN]) FROM @[STAGENAME]
FILE_FORMAT=(TYPE = 'XML')
PATTERN= '.*.xml';
```

AVRO, ORC, Parquet

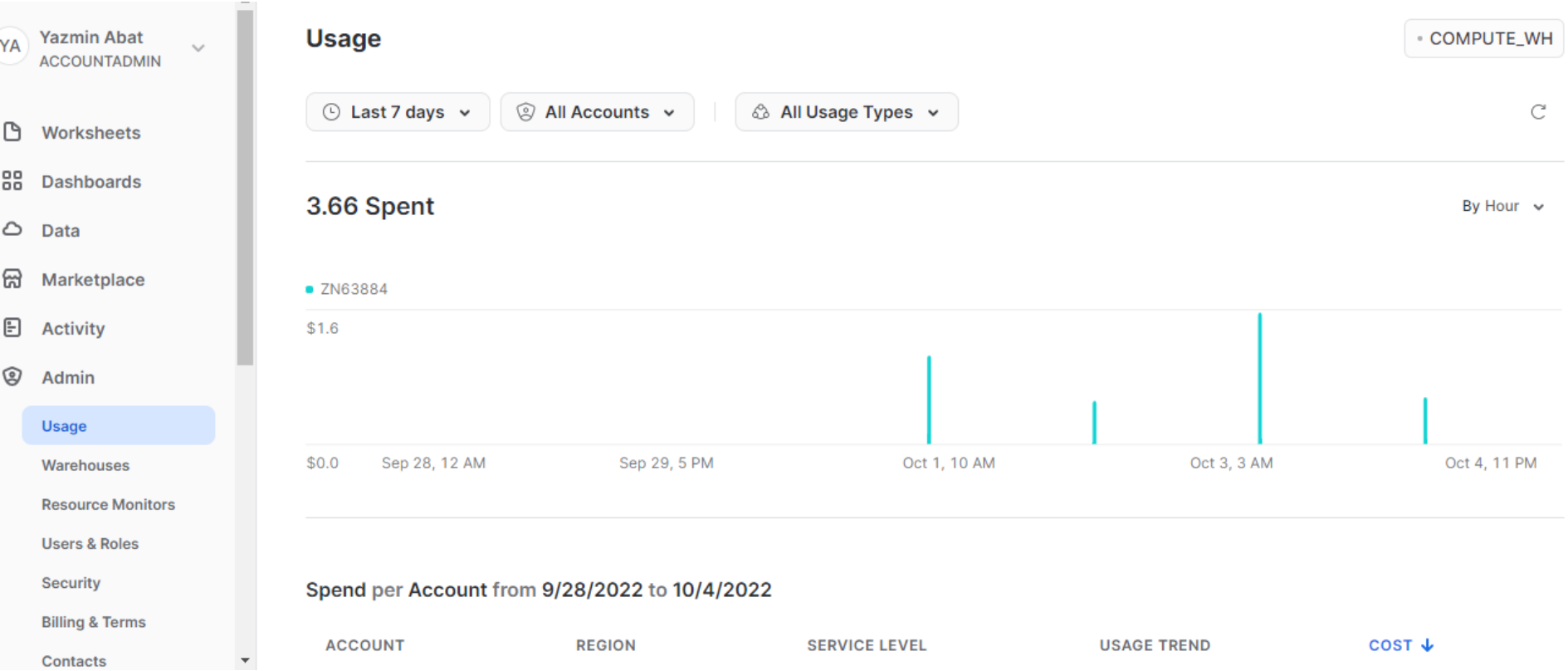
```
COPY INTO [TABLENAME] ([COLUMN]) FROM @[STAGENAME]
FILE_FORMAT=(TYPE = 'AVRO')
PATTERN= '.*.avro';

COPY INTO [TABLENAME] ([COLUMN]) FROM @[STAGENAME]
FILE_FORMAT=(TYPE = 'ORC')
PATTERN= '.*.orc';

COPY INTO [TABLENAME] ([COLUMN]) FROM @[STAGENAME]
FILE_FORMAT=(TYPE = 'PARQUET')
PATTERN= '.*.parquet';
```

Data Engineering Latam

snowflake

Agenda

1. Introduction
2. Compute Warehouses
3. Warehouse Sizes
4. Multiclustering -Table Structures
5. Strategies
4. **Monitoring Snowflake Compute WH**


YA
Yazmin Abat
ACCOUNTADMIN
Worksheets
Dashboards
Data
Marketplace
Activity
Admin
Usage
Warehouses
Resource Monitors
Users & Roles
Security
Billing & Terms
Contacts
Usage
COMPUTE_WH
Last 7 days
All Accounts
All Usage Types
3.66 Spent
By Hour
ZN63884
$1.6
$0.0
Sep 28, 12 AM
Sep 29, 5 PM
Oct 1, 10 AM
Oct 3, 3 AM
Oct 4, 11 PM
Spend per Account from 9/28/2022 to 10/4/2022
ACCOUNT
REGION
SERVICE LEVEL
USAGE TREND
COST

Cost of Virtual Warehouses

Snowflake deployed in
Azure in the **Australia East** region

STANDARD

Complete SQL data warehouse
Secure Data Sharing across regions / clouds
Premier Support 24 x 365
1 day of time travel
Always-on enterprise grade encryption in transit and at rest
Customer-dedicated virtual warehouses
Federated authentication
Database replication
Snowsight
Create your own Data Exchange
Data Marketplace access

$2.75
cost per credit

GET STARTED

ENTERPRISE

Standard +

Multi-cluster warehouse
Up to 90 days of time travel
Annual rekeying of all encrypted data
Materialized views
Search Optimization Service
Dynamic Data Masking

$4.05
cost per credit

GET STARTED

BUSINESS CRITICAL

Enterprise +

HIPAA support
PCI compliance
Tri-Secret Secure using customer-managed keys
Database failover and failback for business continu
Azure PrivateLink support

$5.50
cost per credit

GET STARTED

ON-DEMAND STORAGE
Pay for usage month to month.
$46 per TB / per month
LEARN MORE

CAPACITY STORAGE
Pay for usage up front.
$25 per TB / per month
LEARN MORE

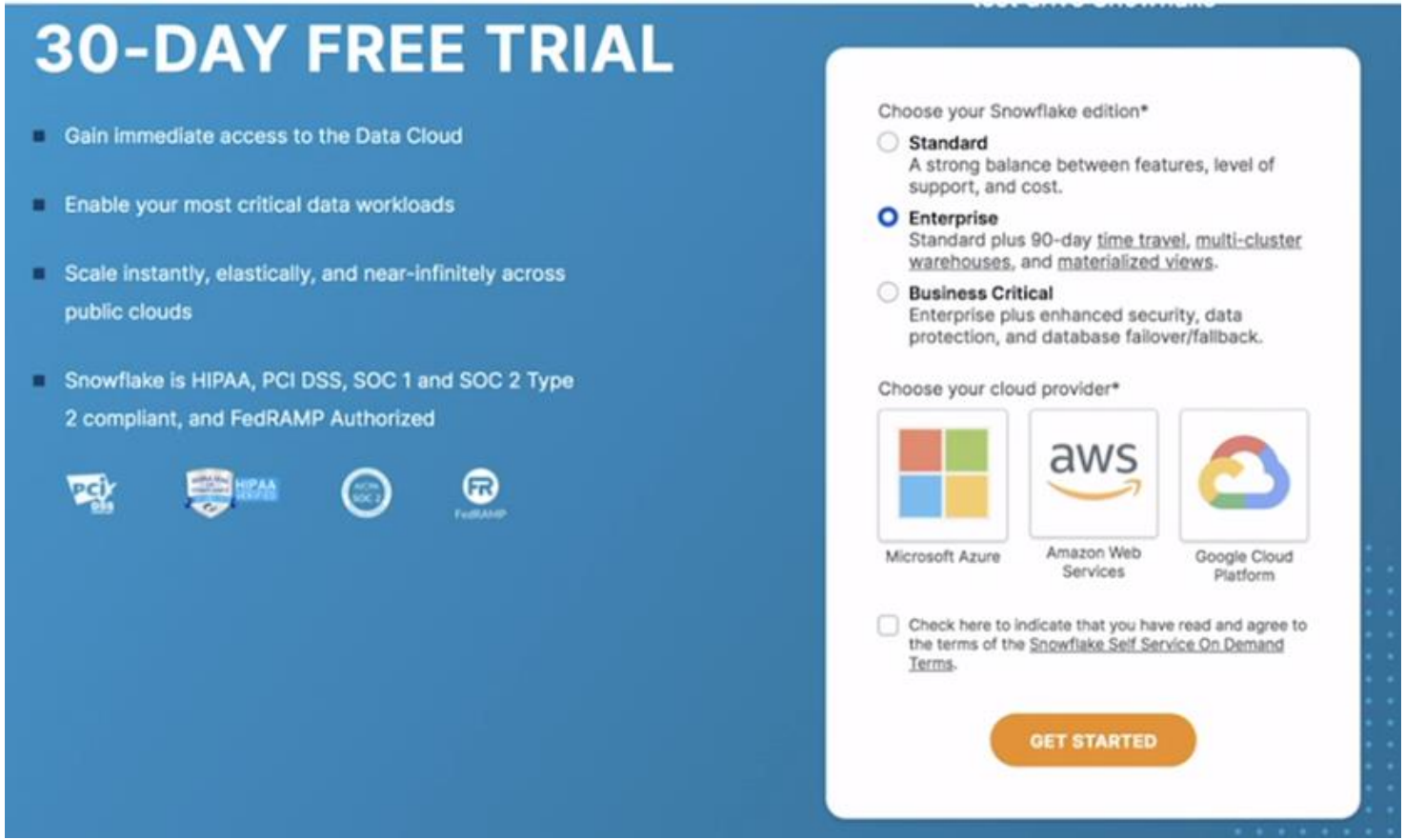
https://www.snowflake.com/pricing

Cost of Virtual Warehouses

Snowflake deployed in **AWS** in the **US East** region

STANDARD	ENTERPRISE	BUSINESS CRITICAL	VIRTUAL PRIVATE SNOWFLAKE (VPS)
Complete SQL data warehouse	Standard +	Enterprise +	Business Critical +
Secure Data Sharing across regions / clouds	Multi-cluster warehouse	HIPAA support	Customer-dedicated virtual servers wherever the encryption key is in memory
Premier Support 24 x 365	Up to 90 days of time travel	PCI compliance	Customer-dedicated metadata store
1 day of time travel	Annual rekeying of all encrypted data	Tri-Secret Secure using customer-managed keys	
Always-on enterprise grade encryption in transit and at rest	Materialized views	AWS PrivateLink support	
Customer-dedicated virtual warehouses	Search Optimization Service	Database failover and failback for business continuity	
Federated authentication	Dynamic Data Masking	External Functions - AWS API Gateway Private Endpoints support	
Database replication	External Data Tokenization		
External Functions			
Snowsight			
Create your own Data Exchange			
Data Marketplace access			
$2.00 cost per credit	**$3.00** cost per credit	**$4.00** cost per credit	
GET STARTED	GET STARTED	GET STARTED	CONTACT US

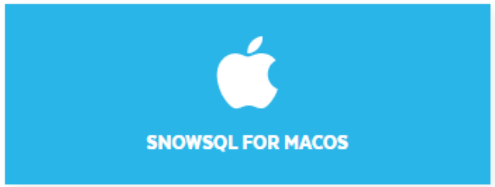
ON-DEMAND STORAGE Pay for usage month to month.	$40 per TB / per month	LEARN MORE	CAPACITY STORAGE Pay for usage up front.	$23 per TB / per month	LEARN MORE


30-DAY FREE TRIAL
Gain immediate access to the Data Cloud
Enable your most critical data workloads
Scale instantly, elastically, and near-infinitely across public clouds
Snowflake is HIPAA, PCI DSS, SOC 1 and SOC 2 Type 2 compliant, and FedRAMP Authorized
Choose your Snowflake edition*
Standard
A strong balance between features, level of support, and cost.
Enterprise
Standard plus 90-day time travel, multi-cluster warehouses, and materialized views.
Business Critical
Enterprise plus enhanced security, data protection, and database failover/fallback.
Choose your cloud provider*
Microsoft Azure
Amazon Web Services
Google Cloud Platform
Check here to indicate that you have read and agree to the terms of the Snowflake Self Service On Demand Terms.
GET STARTED

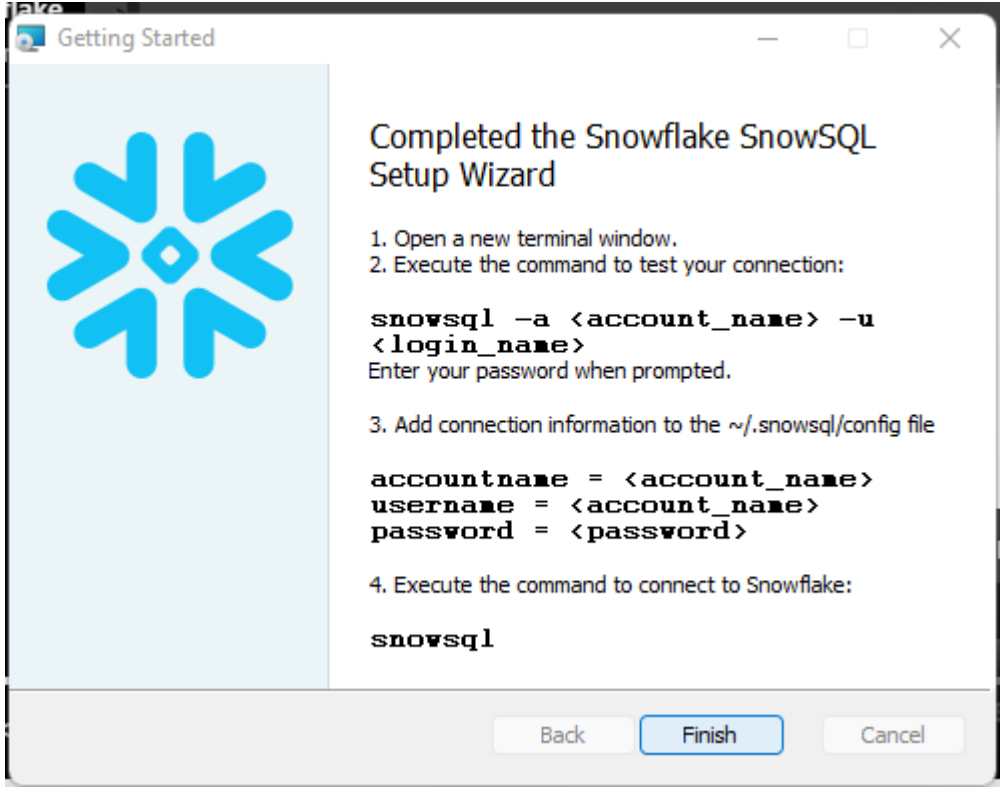
SNOWSQL CLI USAGE

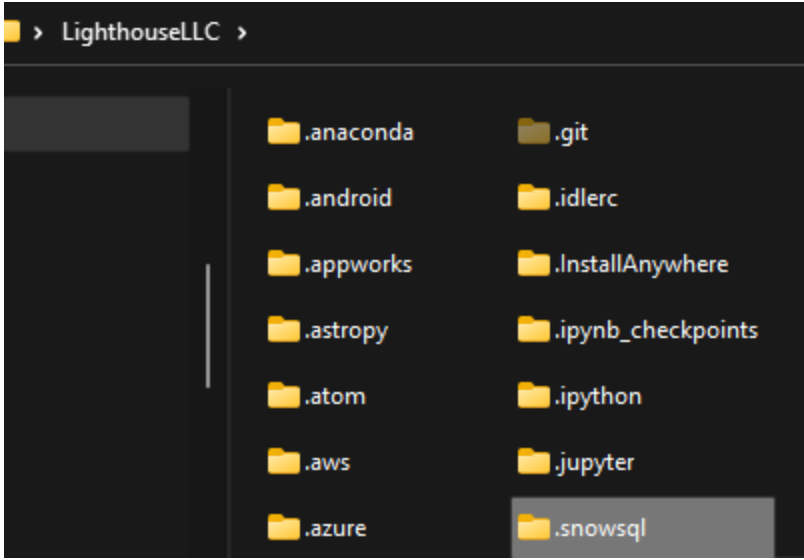
Connecting and Authenticating

```
snowsql --accountname aza12345 --username username

snowsql -a accountname -u username -d databasename -s schemaname
```

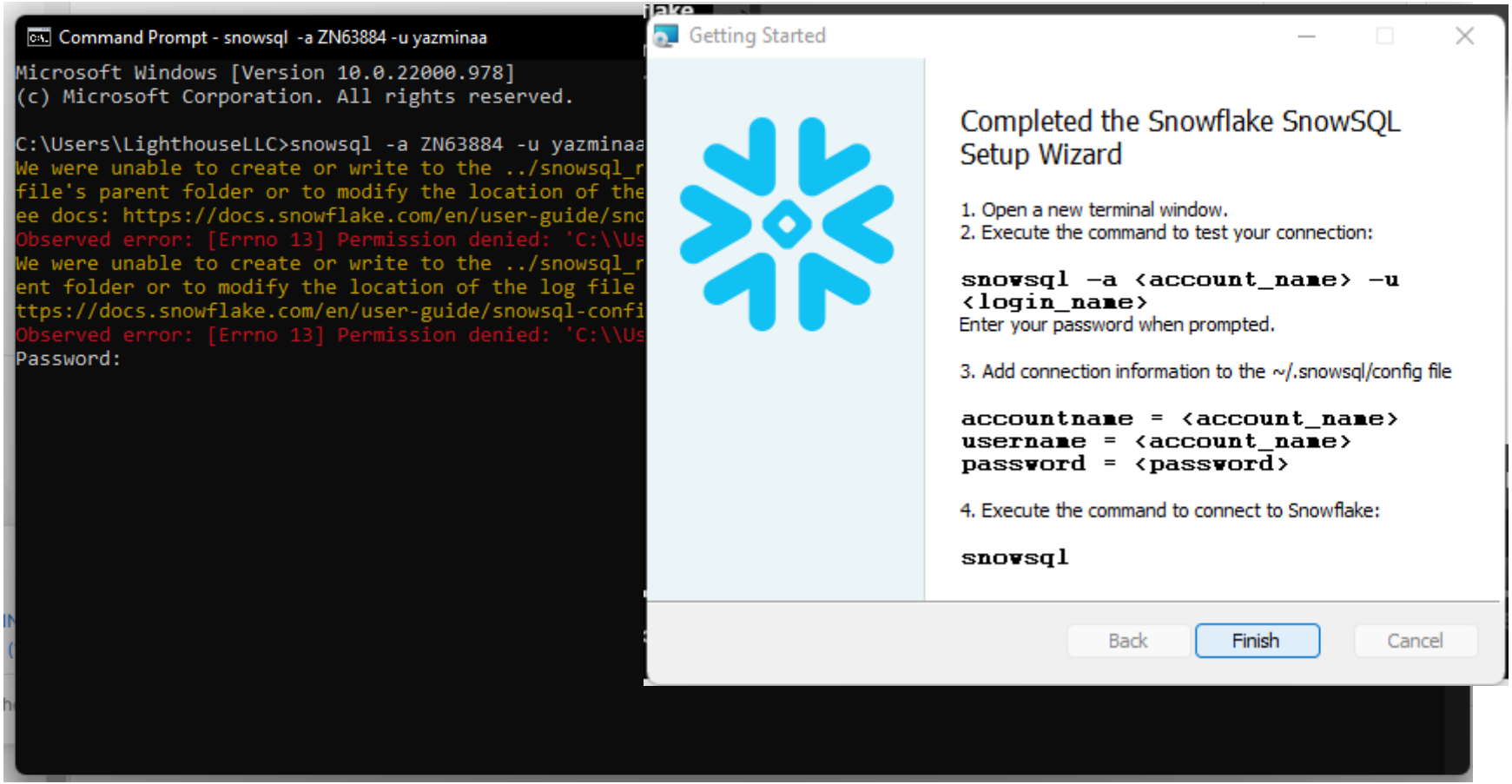


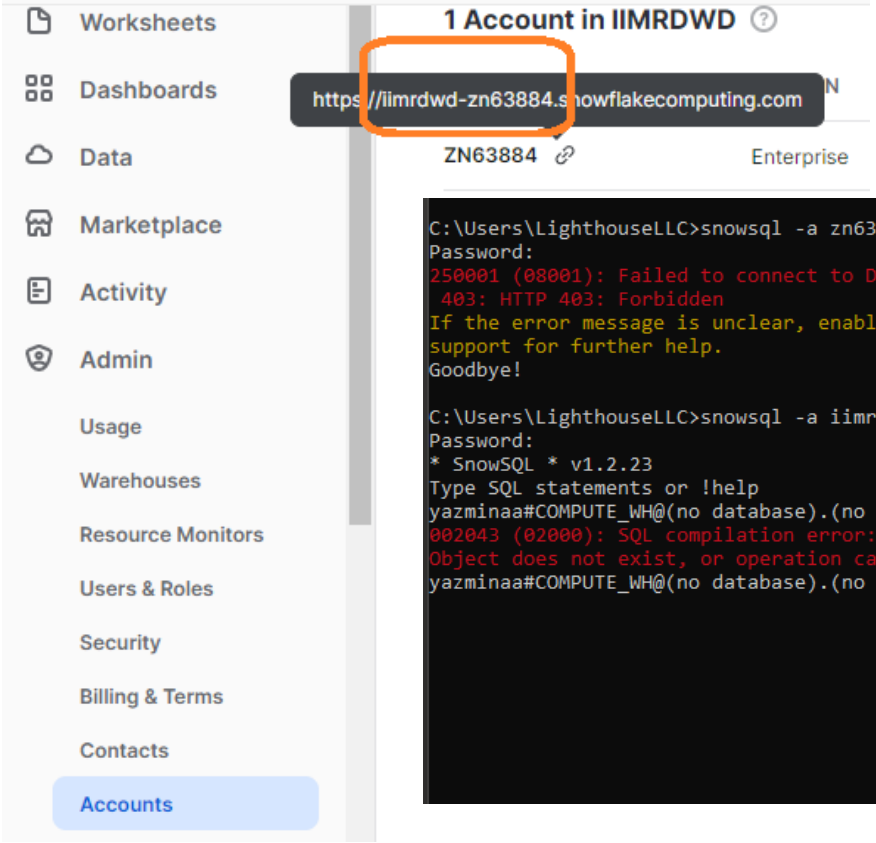
Download All Versions


Getting Started
Completed the Snowflake SnowSQL Setup Wizard
1. Open a new terminal window.
2. Execute the command to test your connection:
snowsql -a <account_name> -u <login_name>
Enter your password when prompted.
3. Add connection information to the ~/.snowsql/config file
accountname = <account_name>
username = <account_name>
password = <password>
4. Execute the command to connect to Snowflake:
snowsql
Back
Finish
Cancel


LighthouseLLC
.anaconda
.git
.android
.idlerc
.appworks
.InstallAnywhere
.astropy
.ipynb_checkpoints
.atom
.ipython
.aws
.jupyter
.azure
.snowsql

Error with just code of account_name






```
C:\Users\LighthouseLLC>snowsql -a zn63884 -u yazminaa
Password:
250001 (08001): Failed to connect to DB. Verify the account name is correct: zn63884.snowflakecomputing.com:443. 000403:
 403: HTTP 403: Forbidden
If the error message is unclear, enable logging using -o log_level=DEBUG and see the log to find out the cause. Contact
support for further help.
Goodbye!

C:\Users\LighthouseLLC>snowsql -a iimrdwd-zn63884 -u yazminaa
Password:
* SnowSQL * v1.2.23
Type SQL statements or !help
yazminaa#COMPUTE_WH@(no database).(no schema)>USE DATABASE SALES_DB;
002043 (02000): SQL compilation error:
Object does not exist, or operation cannot be performed.
yazminaa#COMPUTE_WH@(no database).(no schema)>USE DATABASE
                                               CHAPTER3_TDB1
                                               COVID19_DATA_ATLAS
                                               MARKETPLACE_EMISSIONS
                                               PC_MATILLIONLOADER_DB
                                               PC_MATILLION_DB
                                               SNOWFLAKE
                                               SNOWFLAKE_SAMPLE_DATA
                                               US_STOCKS_DAILY
```

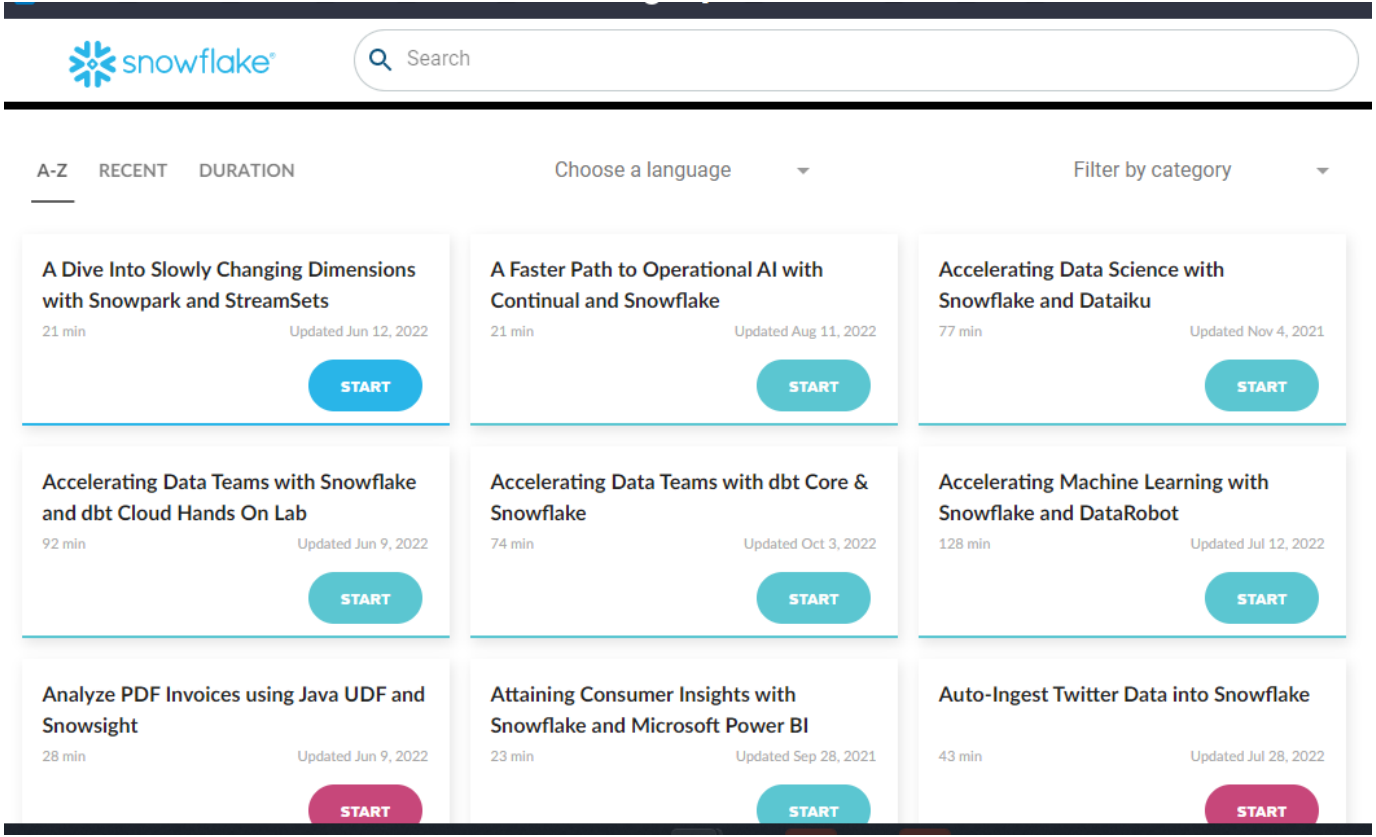
snowflake

Additional Resources

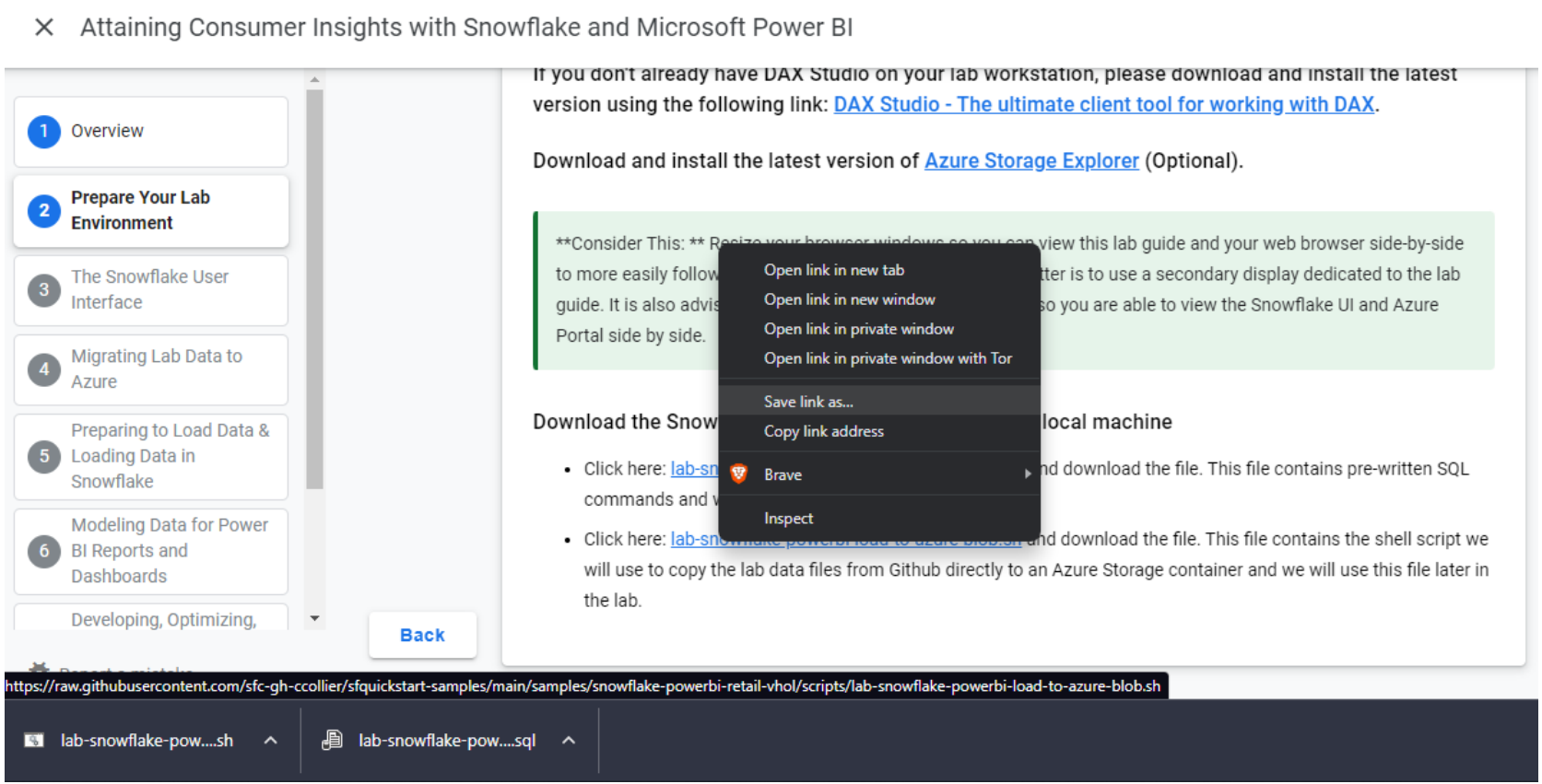
Conclusions

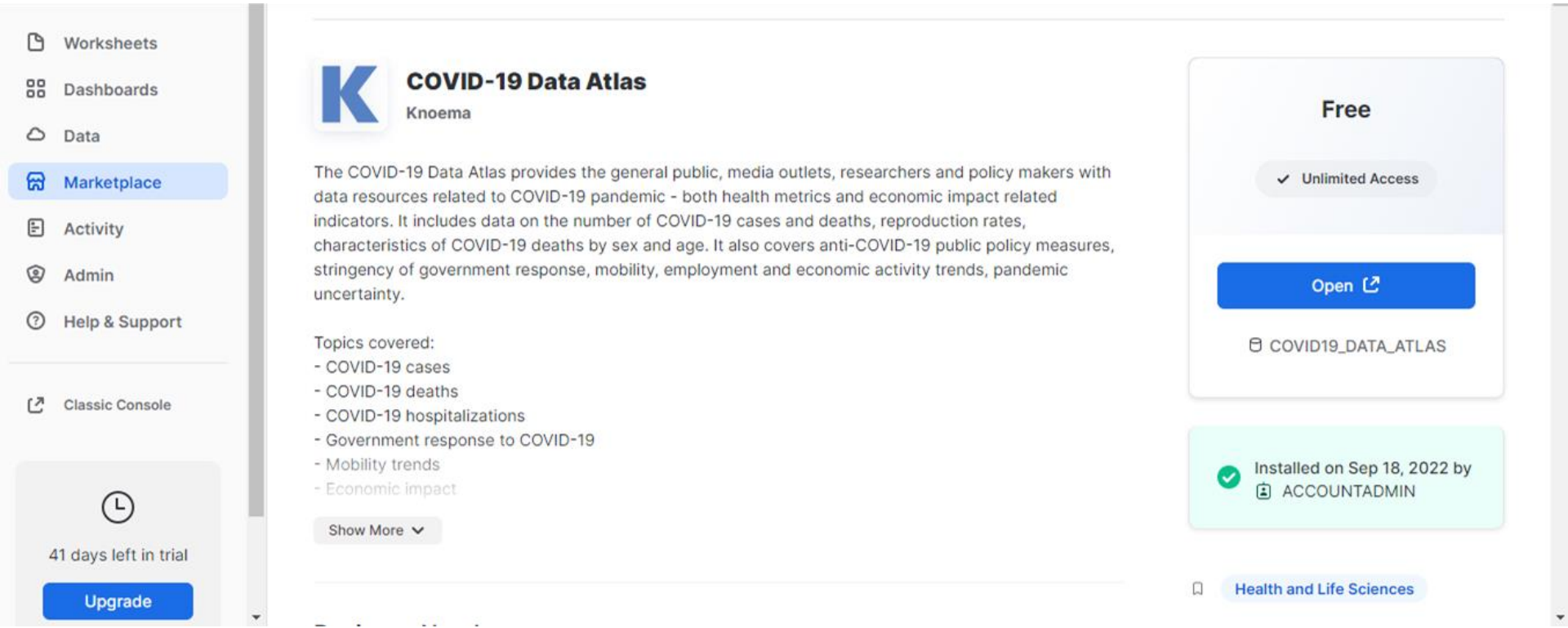
- Snowflake is a managed big data platform
- Tables in Snowflake can be permanent, temporary, or transient
- Snowflake allows you to create non-materialized and materialized views
- Queries are optimized using caching, clustering, and search optimization
- Snowflake allows you to manage and query semi-structured data
- Snowflake enables access control using roles

https://quickstarts.snowflake.com/

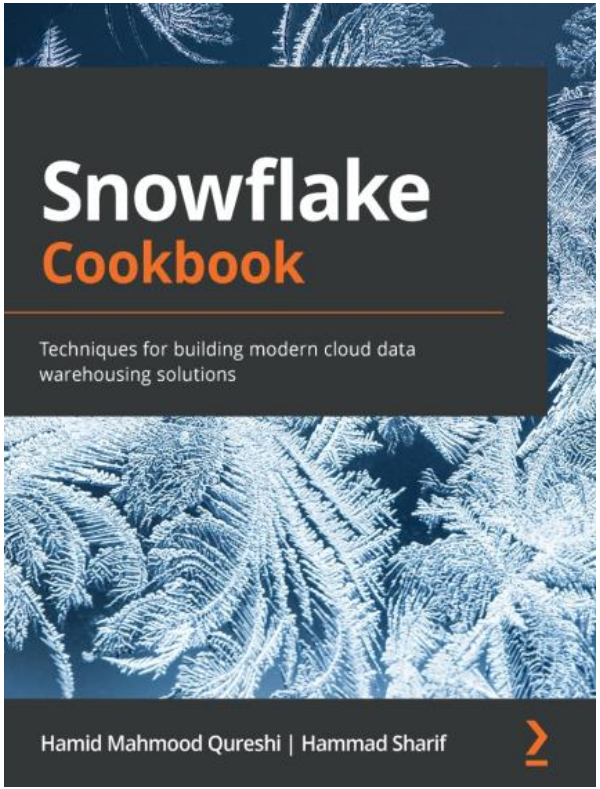


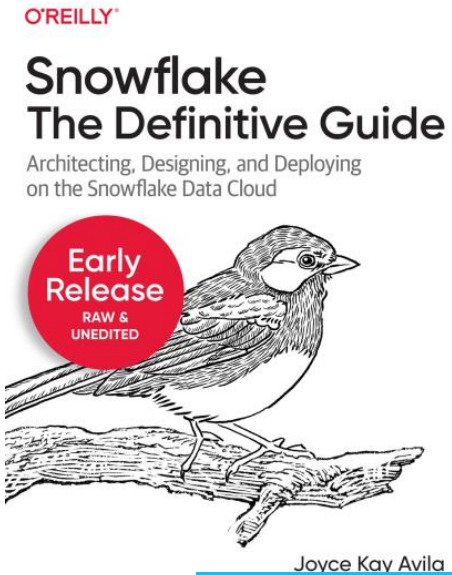
https://quickstarts.snowflake.com/




Worksheets
Dashboards
Data
Marketplace
Activity
Admin
Help & Support
Classic Console
41 days left in trial
Upgrade
COVID-19 Data Atlas
Knoema
The COVID-19 Data Atlas provides the general public, media outlets, researchers and policy makers with data resources related to COVID-19 pandemic - both health metrics and economic impact related indicators. It includes data on the number of COVID-19 cases and deaths, reproduction rates, characteristics of COVID-19 deaths by sex and age. It also covers anti-COVID-19 public policy measures, stringency of government response, mobility, employment and economic activity trends, pandemic uncertainty.
Topics covered:
- COVID-19 cases
- COVID-19 deaths
- COVID-19 hospitalizations
- Government response to COVID-19
- Mobility trends
- Economic impact
Show More
Free
Unlimited Access
Open
COVID19_DATA_ATLAS
Installed on Sep 18, 2022 by
ACCOUNTADMIN
Health and Life Sciences

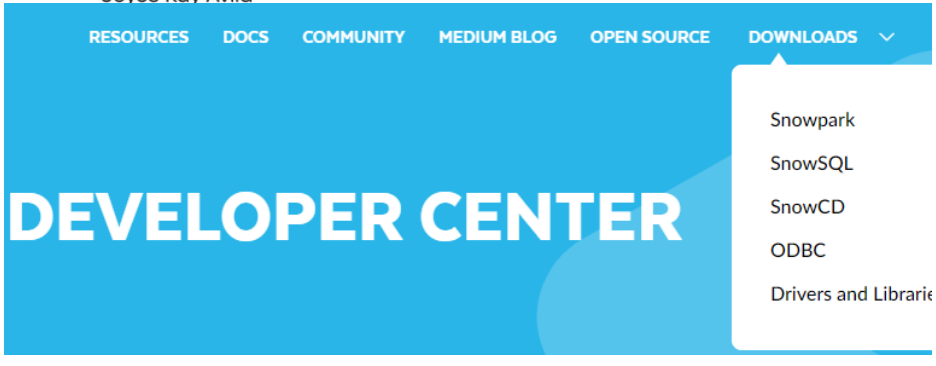
Literature





https://resources.snowflake.com/

https://developers.snowflake.com/



Tutorial: Bulk Loading from a Local File System Using COPY

This tutorial describes how to load data from files in an internal Snowflake stage into a table.

You will learn how to:

- Create named file formats that describe your data files.
- Create named stage objects.
- Stage your data files to internal Snowflake stages.
- Load your data into Snowflake tables.
- Resolve errors in your data files.

The tutorial covers loading of both CSV and JSON data.

Related Topics

- Loading Data into Snowflake
- Data Loading Considerations

Interface: SnowSQL (CLI Client)

Cost

Bulk data load: Billed for the amount of time each virtual warehouse is active.

Snowpipe: Billed according to the compute resources used in the Snowpipe warehouse while loading the files.

Recommended Load File Size

For the most efficient and cost-effective load experience with Snowpipe, we recommend following the file sizing recommendations in File Sizing Best Practices and Limitations and staging files once per minute. This approach typically leads to a good balance between cost (i.e. resources spent on Snowpipe queue management and the actual load) and performance (i.e. load latency). For more information, see Continuous Data Loads (i.e. Snowpipe) and File Sizing.

Load Order of Data Files

For each pipe object, Snowflake establishes a single queue to sequence data files awaiting loading. As new data files are discovered in a stage, Snowpipe appends them to the queue. However, multiple processes pull files from the

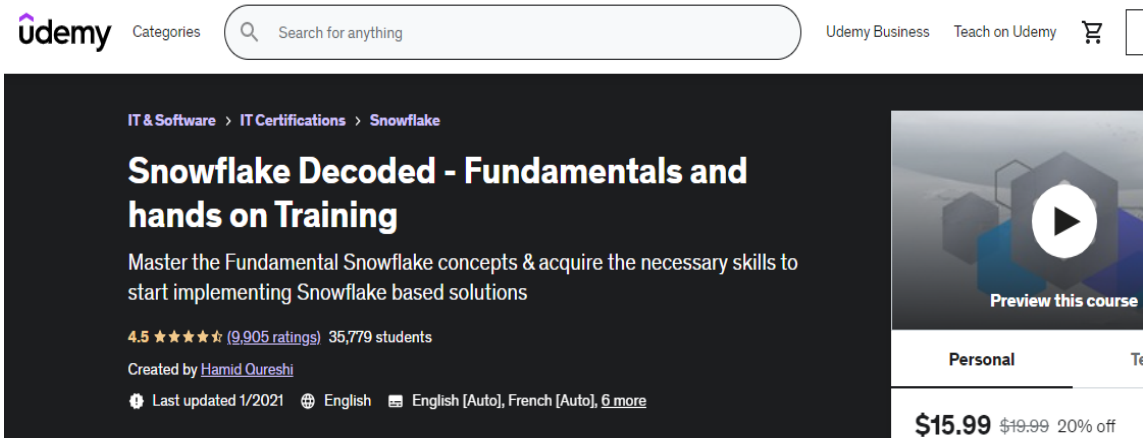
Video Courses (English)

Advanced Snowflake
With Janani Ravi · Liked by 23 users

https://cloudacademy.com/course/introduction-snowflake/course-intro/

https://www.linkedin.com/learning/advanced-snowflake?trk=learning-serp_learning-search-card_search-card&upsellOrderOrigin=default_guest_learning

https://www.youtube.com/watch?v=AR88dZG-hwo&list=PLba2xJ7yxHB7SWc4Sm-Sp3uGN74ull4pS



https://www.udemy.com/course/snowflake-essentials/

snowflake

Thank You!