MARCH,2021

CLOUD DATA WAREHOUSE COMPARISON

BRILLIO DATA COE - POINT OF VIEW

ACCELERATE WHAT MATTERS. NOW.

brillio

OUR DIGITAL TRANSFORMATION ENABLERS

DRIVEN BY PRODUCT MINDSET

DESIGN THINKING

CONTENT COLLABORATION

DESIGN STUDIO

PRODUCT ENGINEERING

- OMNI CHANNEL APPS
- MICROSERVICES/MESH ARCHITECTURE
- MODERN APPS & CONTAINERIZATION
- DEVOPS
- LOW/NO CODE SOLUTION
- COGNITIVE TESTING

CUSTOMER EXP PLATFORMS

- CRM IMPLEMENTATION
- MARKETING/SERVICE CLOUD
- SERVICE BOT
- HYBRID INTEGRATION
- INTELLIGENT SALES & E-COMMERCE

DATA & ANALYTICS

- MASTER DATA MANAGEMENT
- DATA MIGRATION
- DATA LAKE ON CLOUD
- AI/ML
- ANALYTICS AS A SERVICE

DIGITAL INFRASTRUCTURE

- CLOUD STRATEGY & MIGRATION
- DIGITAL OPERATIONS
- ROBOTIC PROCESS AUTOMATION
- MANAGED SERVICES
- ZERO OPS
- SECURITY & COMPLIANCE

ADVANCED TECHNOLOGY GROUP

TECH STRATEGY & CONSULTING | TECH LABS | ENTERPRISE ARCHITECTURE | BLOCKCHAIN | EDGE | SERVERLESS COMPUTING

Microsoft apigee salesforce SAP paloalto NETWORKS Digital Asset snowflake

Accelerators: brillio BOLT™ brillio SMARTEST™ brillio DEVOPS TOOLKIT brillio ONE™ brillio CLIP

brillio

CLOUD DATA WAREHOUSE SOLUTION PROVIDERS

TOGETHER WE KNOW HOW!!

CLOUD DATA WAREHOUSE SOLUTION PROVIDERS

A cloud data warehouse is a database delivered in a public cloud as a managed service that is optimized for analytics, scale and ease of use.

Top Cloud Data Warehouse Solution Providers

Amazon Redshift
Amazon Redshift is a data warehouse product which forms part of the larger cloud-computing platform AWS. It is a simple and cost-effective data warehouse solution that analyses all the user data across their on-premise data warehouses and data lakes.

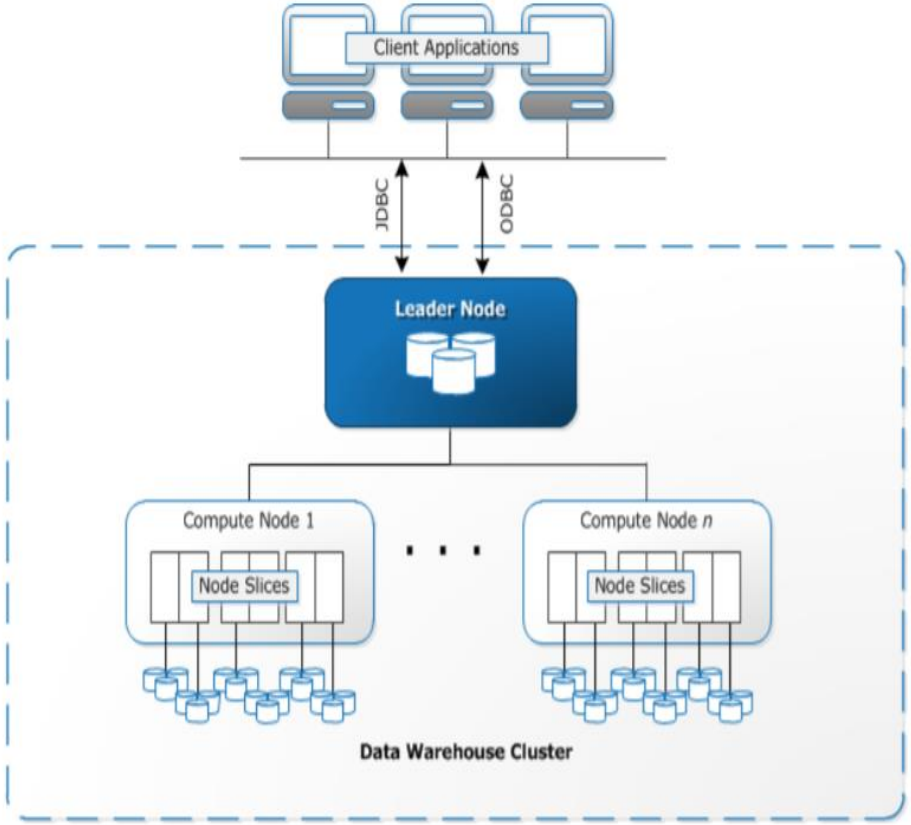
Google BigQuery
Google's BigQuery is an enterprise-grade cloud-native data warehouse. BigQuery has evolved into a more economical and fully-managed data warehouse which can run blazing fast interactive and ad-hoc queries on datasets of petabyte-scale.

Snowflake
Snowflake offers a cloud-based data storage and analytics service, generally termed "data warehouse as a service". It allows corporate users to store and analyze data using cloud-based hardware and software.

Azure Synapse
Synapse is a state-of-the-art analytics solution that combines enterprise data warehousing with the latest in big data analytics. With Microsoft's technology, you can easily query data according to your individual needs.

AMAZON REDSHIFT

Amazon Redshift is a fully managed, petabyte-scale data warehouse service on the cloud. Regardless of the size of the data set, Amazon Redshift offers fast query performance using the same SQL-based tools and business intelligence applications.



Client Application
Amazon Redshift provides integration with various ETL tools including custom applications through industry standard PostgreSQL, ODBC and JDBC drivers.

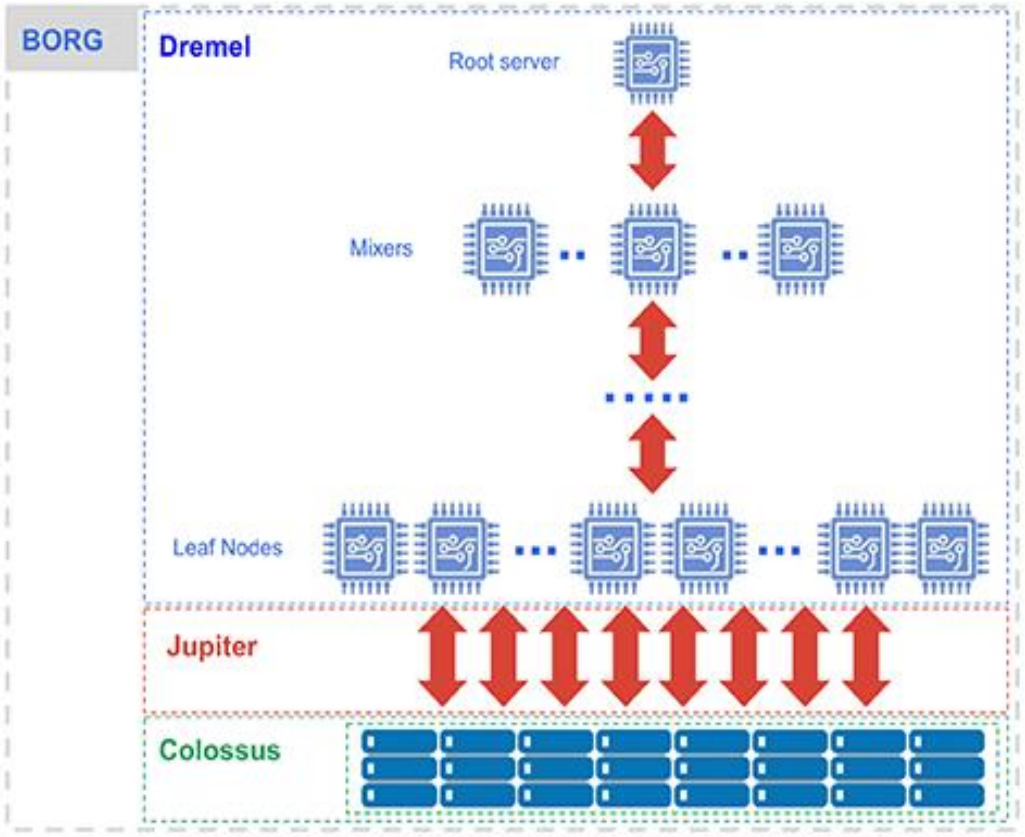
Leader Node
Leader node manages communication between client application and compute nodes. Leader node is responsible for parsing queries and determining SQL execution plan. Leader node understands how tables are distributed in Compute nodes. It sends query to appropriate compute node for execution based on reference of the table.

Compute Node
Compute nodes are responsible for execution of queries, building result sets and sending results back to Leader node. Each Compute node is an independent machine that has it's own storage, memory and computational capacity. Compute nodes are designed to be horizontally scalable. Compute nodes can be added, as and when storage and CPU demand capacity grow.

Databases
A Redshift cluster may contain one or more databases. Databases are stored in Compute nodes and distributed across multiple nodes. Redshift databases are relational databases based on PostgreSQL database kernel.

GOOGLE BIGQUERY

BigQuery is a Cloud-Powered Massively Parallel Query Service with a true server-less database. It leverages columnar storage and tree architecture with Root, mixer and leaf nodes.



Dremel - The Execution Engine
BigQuery is the public implementation of Dremel, used widely in Google - from search to ads, from YouTube to Gmail. Dremel turns SQL queries into an execution tree. Leaves, or 'slots', read data from Colossus & perform computation and branches, or 'mixers', perform aggregation. In between is 'shuffle', uses Jupiter to move data. The mixers and slots all run by Borg, that assigns hardware resources. Dremel dynamically apportions slots to queries, amongst multiple users.

Colossus - Distributed Storage
Google's distributed file system (columnar storage). Each Google datacenter has its own Colossus cluster, which in turn has disks for every user. It also handles replication, recovery & distribution management.

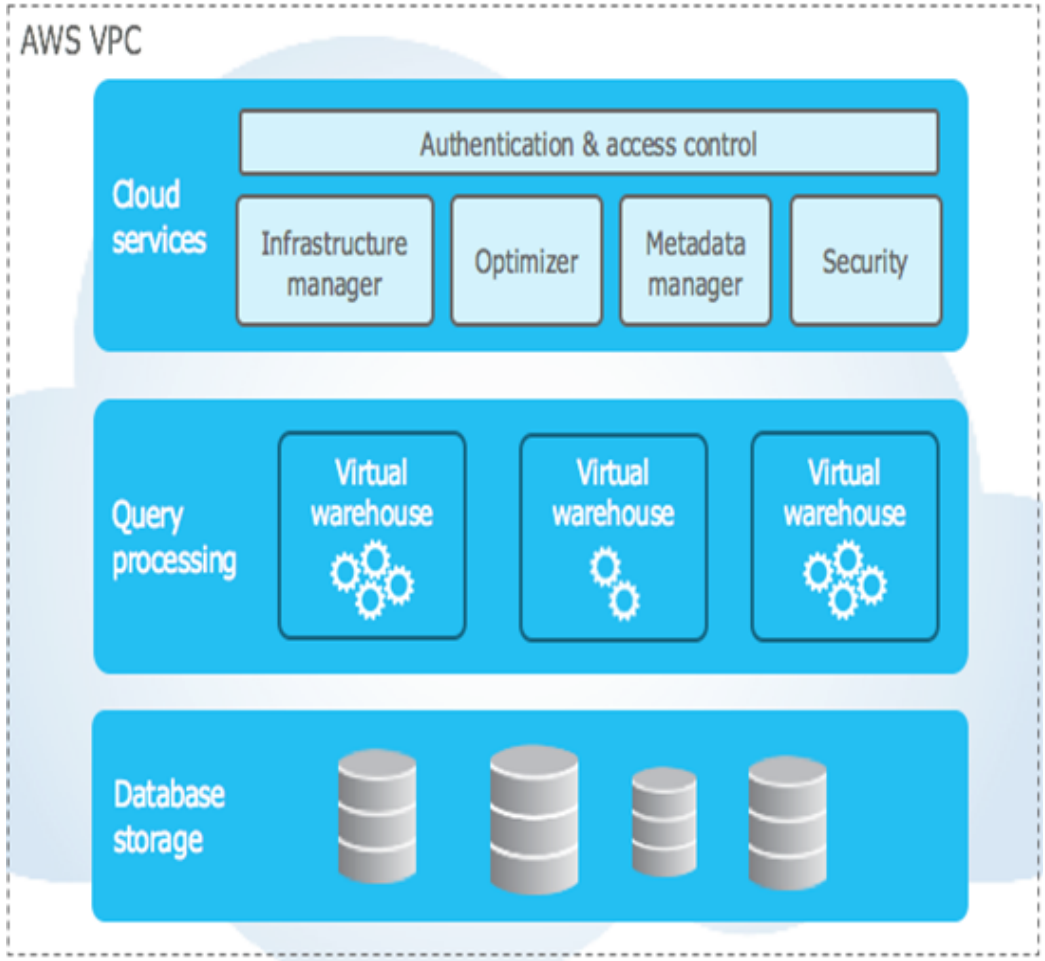
Borg – Compute
Google's large-scale cluster management system. Borg assigns server resources to jobs. It also routes around for network & machine failures.

Jupiter – The Network
Google's Jupiter can efficiently and quickly distribute large workloads.

BigQuery – The Service
Low level infrastructure components are combined with several dozen high-level technologies, APIs, and services — like Bigtable, Spanner, and Stubby — to make one transparent and powerful analytics database — BigQuery.

SNOWFLAKE

Snowflake is an elastic data warehouse running on three public clouds – AWS, Azure and Google Cloud. It is a fully managed solution with scale out capabilities to allow processing over multiple compute nodes.



Snowflake architecture consists of three tiers:

Database Storage

Snowflake reorganizes data into its internal optimized, compressed, columnar format. Snowflake stores this optimized data using Amazon Web Service's S3 cloud storage.

Snowflake manages all aspects like the organization, file size, structure, compression, metadata, statistics, and other aspects of data storage.

Query Processing

Query execution is performed in the processing layer. Snowflake processes queries using "virtual warehouses". Each virtual warehouse is an MPP compute cluster composed of multiple compute nodes allocated by Snowflake from Amazon EC2.

Cloud Services

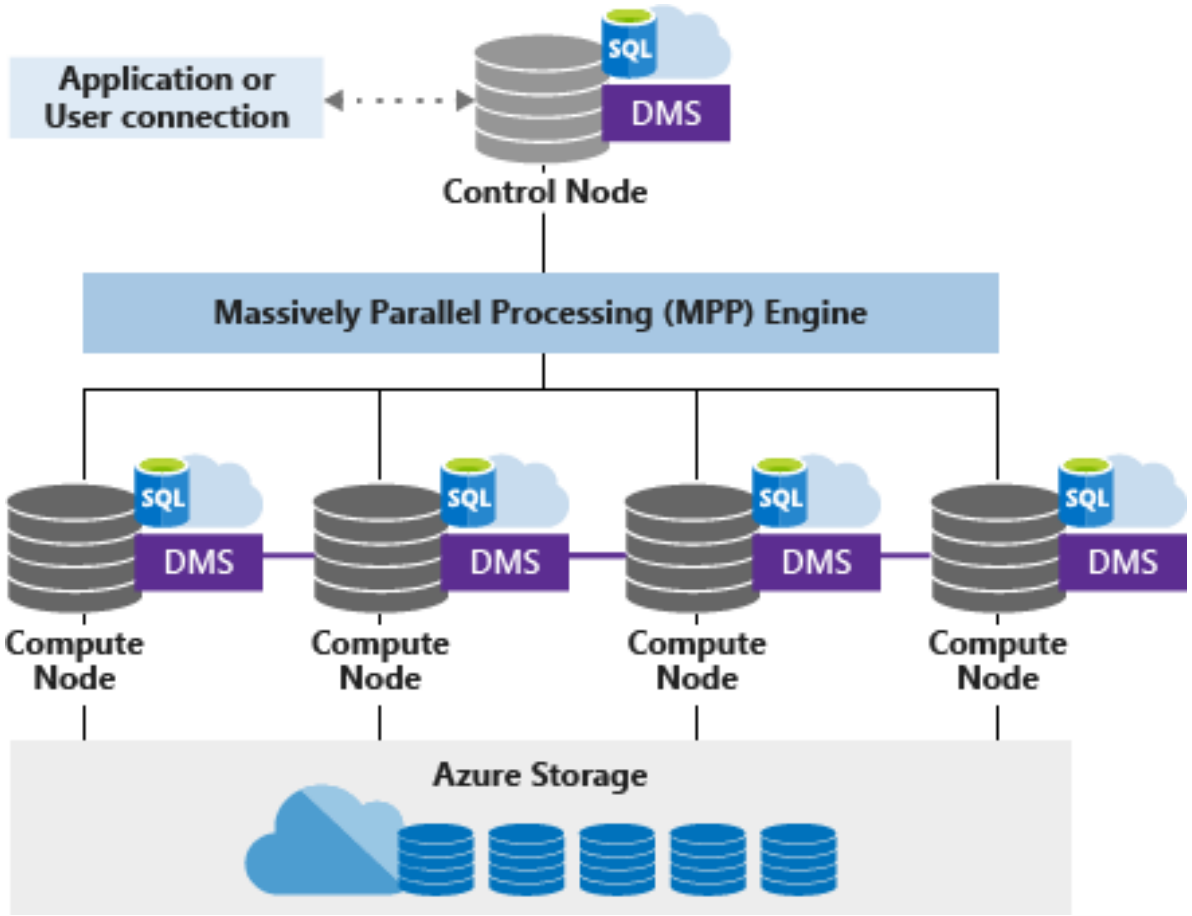
The cloud services layer is a collection of services that coordinate activities across Snowflake.
The Services in this layer are:

- Authentication
- Infrastructure management
- Metadata management
- Query parsing and optimization
- Access control

.

AZURE SYNAPSE

Azure Synapse combines massively parallel processing (MPP) with Azure storage. It leverages a scale out architecture to distribute computational processing of data across multiple nodes.



- Applications connect and issue **T-SQL** commands to a Control node, which is the single point of entry for the data warehouse.
- The **Control node** runs the MPP engine which optimizes queries for parallel processing, and then passes operations to Compute nodes to do their work in parallel.
- The **Compute nodes** store all user data in **Azure Storage** and run the parallel queries.
- The **Data Movement Service** (DMS) is a system-level internal service that moves data across the nodes as necessary to run queries in parallel and return accurate results.
- With decoupled **storage and compute**, Synapse Analytics can:
    - Independently size compute power irrespective of your storage needs.
    - Grow or shrink compute power, within a SQL pool, without moving data.
    - Pause compute capacity while leaving data intact, so you only pay for storage.
    - Resume compute capacity during operational hours.

brillio

CLOUD DATA WAREHOUSE COMPARISON

TOGETHER WE KNOW HOW!!

CLOUD DATA WAREHOUSE COMPARISON (1/7)

	Snowflake	Azure Synapse	Google Big Query	AWS Redshift
DB Architecture	True segregation of compute from storage and separate metadata cluster	MPP with multiple compute nodes .Although storage is decoupled and nodes can be paused , but addition or deletion of nodes results in data movement	Serverless highly scalable and cost effective multi cloud data warehouse designed for business agility. Storage and compute are separated to achieve high performance	Redshift is enterprise level petabyte scale fully managed DWH solution with MPP support. Key features - columnar data store, optimum data compression for storage and processing
Compute & Storage	Completely decoupled and allows scaling	Limited degree of decoupling	Completely decoupled and allows scaling	Provides various options for cluster design focusing 1)dense storage 2)dense compute 3)decoupled(storage and compute)-pay for storage & compute separately
Semi structure File Format	JSON, AVRO, ORC, Parquet, XML	Delimited, Hive RC, Hive ORC, Parquet	JSON(New Line Delimited Only), AVRO, ORC, Parquet, CSV	AVRO, CSV, JSON, Parquet, ORC & Delimited
Bulk Loading Support	Supported through COPY command in AWS, AZURE and GCP	Supported through BCP,SQL Bulk Copy, API, COPY Command	Load API and BQ Transfer service support	COPY command (API Feature introduced newly)

CLOUD DATA WAREHOUSE COMPARISON (2/7)

	Snowflake	Azure Synapse	Google BigQuery	AWS Redshift
Continuous Data Load	Snow Pipeline support	ADF Support	Load API and BQ Transfer service support .BQ Streaming data	Lambda, Glue, Kinesis, CLI Command and with AWS SDK for external tools
Storage Requirements	Time Travel config needs extra storage, but SF provides excellent compression	Store in ADLS Gen2 and Optimiza in Synapse	Internally uses GCS storage	Various options 1.SSD 2.HDD 3.S3 using redshift spectrum and external tables
Scaling Up/Out	Instant scalability for multi cluster DW. No downtime for scale up but no impact on running queries	Unavailability window during scaling No auto scaling need manual intervention	Serverless scales (auto) based on the query requirement/pattern No setting/config .Taken care by BQ	can scale using classic scale up method. Initially it provision new node and migrate data New feature of elastic scale in Redshift also concurrency scale which add storage/compute as needed
Concurrency	Automatic concurrency scaling with multi cluster DW	Depending on DWU config.4 to 128 concurrent user queries (gen2)	100 concurrent users .Need to approach GCP support to increase limit	Automatic concurrency scale with multi node cluster
Query Flashback	Fully supported	Query datastore feature captures history of queries	Supports Query history by default	Automatic snapshot in redshift take incremental snapshots which can be used for restore

CLOUD DATA WAREHOUSE COMPARISON (3/7)

	Snowflake	Azure Synapse	Google Big Query	AWS Redshift
Data Distribution	At table level clustering can be specified .SF has provisioning for auto clustering -Less Developer dependence	Effort needed for partitioning and determining how to distribute the data	Partition and clustering feature is supported .Users has to specify/design the tables as per need	Data distribution happens through distribution key define while designing tables
Performance Tuning effort	Optimize only SQL no table level tuning required Auto clustering takes care of micro partitions -less intervention	Table and query level tuning required High manual intervention to tune	Table and query level tuning required	Table and query level tuning required RS provided advisory service as feature to optimize cluster config, distribution/sort key etc.
Surrogate Generation	Can be handled inside SF	Needs to be externalized or it needs Azure SQL to generate surrogate keys	Two types of surrogate keys possible. Alphanumeric and numeric	Create as identity column to generate value
Scale Down	Can specify auto shutdown policy that can scale down cluster or shutdown during in activity -automatic with no downtime	Has to be done through explicit commands and removing of nodes have implications on running queries and new queries to be submitted	Serverless taken care of by BQ	With elastic resize feature scale down can be done in minutes. Alternatively use classic resize feature which creates new instances in parallel and switch primary instance to new instance during this period it will not be available

CLOUD DATA WAREHOUSE COMPARISON (4/7)

	Snowflake	Azure Synapse	Google BigQuery	AWS Redshift
Stored Proc support	Available only through Python or JavaScript UDF	TSQL compatibility -Rich support	Rich support with SQL compatibility	Yes -Newly introduced feature
Data Caching	Hot Data SSD Cache+ Exact Query	Hot Data SSD Cache+ Exact Query	Only Exact Data + In Development	Hot Data SSD Cache+ Exact Query
View Scalability	Materialized views available from Enterprise + editions	Materialized views enhances performance of complex join queries	Materialized views enhances performance of complex join queries	Materialized views enhances performance of complex join queries
Workload Isolation	Excellent and provided isolation at VW level	Needs to be managed which is a manual process-Priority and allocation-admin effort	Can be handled either segregating by projects or going for reserved capacity option	Need to be managed manually or to be automated using workload manager
Compute Pricing	Based on credits consumed ,lowest unit is minutes for first one minute, followed by seconds level credits	Minimum unit is hour	Charged only for Bytes processed on each query and not for the compute	Depending on cluster type pricing will vary , compute will be charged based on hourly usage .New feature introduced for per second billing as well for concurrent scaling

CLOUD DATA WAREHOUSE COMPARISON (5/7)

	Snowflake	Azure Synapse	Google BigQuery	AWS Redshift
Cost Controls	Provide resource monitors for setting thresholds of usage - multi tenant charge back scenario	None	Realtime API and data feeds are possible to visualize the cost at granular levels. Also has the option to trigger notifications based on usage limits	1.Reserved instance 2.Monitor utilization 3.start with lower capacity and increase as needs (elastic resize)
Ease Of cost allocation by usage	fine-grained at VWH level	Not easy to apportion-only can be approximated	Dashboards can be created from the internal data BQ TO SHOWCASE THE COST AT granular levels	User defined label (tagging) can be assigned to each resource for cost allocation
Data Security - Row and column level	Implemented through secure views	Row and column level security support using RLS AND CONTROLLING ACCESS table columns at user level	Implemented through authorized views	Row and column level security supported
Data Security - Masking	Out of the box dynamic data masking capabilities for enterprise +	Out of the box dynamic data masking capabilities	Has to implement thru Data catalogue and user role combinations	Out of the box dynamic data masking capabilities
Data Security- Encryption	Column level encryption handled through tokenization available through enterprise +	Can enable double encryption using a customer-managed key" option on the "Security" tab when creating your new workspace	Column level encryption using custom key or google based key	Column level encryption using UDF

CLOUD DATA WAREHOUSE COMPARISON (6/7)

	Snowflake	Azure Synapse	Google BigQuery	AWS Redshift
Secure Data Sharing	Allows for data sharing to other snowflake instance without need to replace data	NA	possible using share dataset option	Allows for data sharing to other accounts with replicate/snapshot restore
Across Geo zone replication	Provides replication capabilities for moving the data across geo zones -US East to US west. Data is replicated but their roles and metadata has to be created	Supports HA through DB snapshots and geo backup for DR to a paired DC	Only data replication only within the region For the data to be copied to diff geo this requires manual intervention	Supports HA through DB snapshots on S3 which can be stored in any region
Time Travel Flashback	Configure property in a table for up to how many days' Time Travel can happen .Can query previous versions	NA	7 days of history the user can query and get the data	NA
Data Clones	Ability to create data subsets without copying the data through cloning	NA	Copy data set is available but it would incur separate storage cost	Can be implemented using custom solution for data replication

CLOUD DATA WAREHOUSE COMPARISON (7/7)

	Snowflake	Azure Synapse	Google BigQuery	AWS Redshift
ML Support	Supports ML and AI workloads with its unlimited scalability and integration capabilities with ML/AI products like Dataiku	Azure Synapse Analytics also evolves rapidly to address Advanced Analytics use cases leveraging the power of Apache Spark	Supports ML algorithms using SQL language	Supports through Sagemaker/Glue
API Support	Supports CLI, JDBC, ODBC	Supports . Net , JDBC, ODBC	Supports all the BQ SQL operations using REST API	Supports all Data API
Maintenance	Low Maintenance Automatic and rapid provisioning of grater compute resources	Low Maintenance Limitations Smoothly integrates data from diverse sources, but can be prone to bugs	Low maintenance Limitations No Indexes No Column constrains No performance tuning capabilities	Requires Vacuuming/Analyzing tables periodically
Federated Table support	You can use the SQL Gateway to configure a MySQL remoting service and set up federated tables for Snowflake data	You can use the SQL Gateway to configure a MySQL remoting service and set up federated tables for Azure Table data.	Supports federated table for cloud SQL Big Table and google sheets	Supports federate query to source data from object store and RDBMS

THANK YOU