FEBUARY,2021

SNOWFLAKE USE CASES

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

SNOWFLAKE USE CASES

TOGETHER WE KNOW HOW!!

SNOWFLAKE – OBSERVED USAGE SCENARIOS

Snowflake offers a cost-effective alternative to standard Data warehouse solution like Teradata, Oracle or IBM Netezza. There are observed usage scenarios which are driving the adoption of Snowflake Data warehouse across enterprises.

Data warehouse modernization
Snowflake architecture allows storage and compute capacity to grow independently and provides a fully managed service solution for a cloud Data warehouse

Building Time Series Databases
A time series database is optimized for time-stamped (time series) data and for measuring change over time.

Handling variety of workload
Snowflake virtual data warehouse allows running variety of workloads against the same Snowflake database

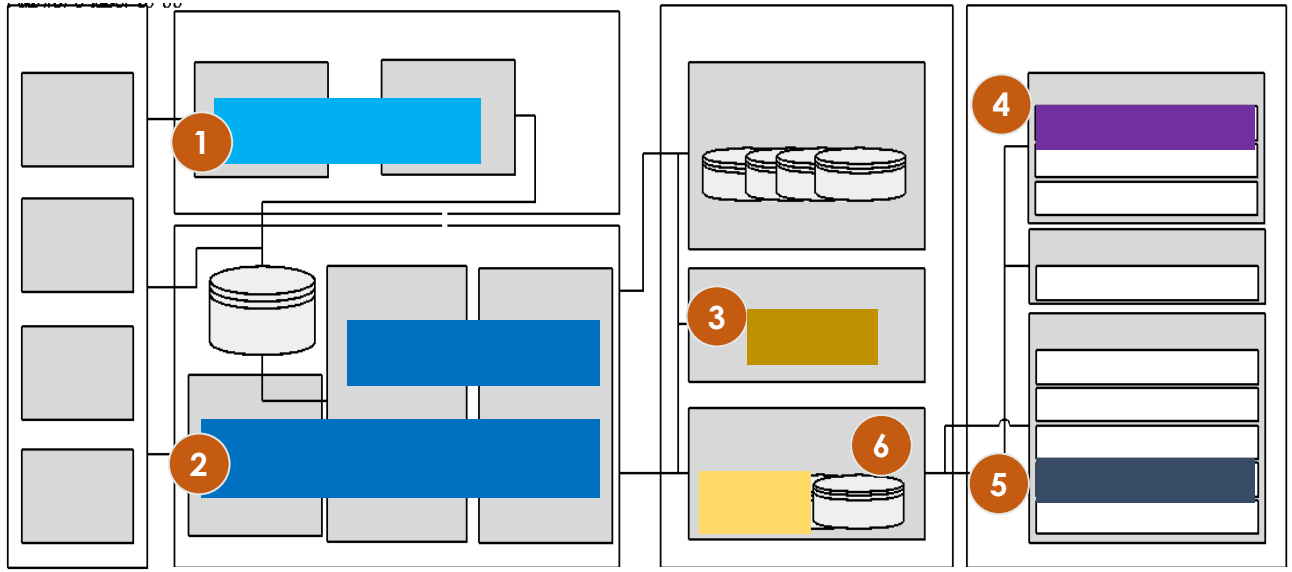
Standard SQL Access
Snowflake provides standard ANSI SQL access based on organized data tier for easy data consumption without much complexity in data architecture

brillio

DATA WAREHOUSE MODERNIZATION ON SNOWFLAKE

TOGETHER WE KNOW HOW!!

THERE ARE KEY CHALLENGES OPERATING A LEGACY WAREHOUSE IN THE NEW AGE OF DATA



- Non-Relational Data Store
- Relational Data Store
- Stored procedures /OLAP cubes
- Analytics Sandbox
- BI and Reporting Layer
- Data Archive Store

1. Difficult to maintain and expand custom extract and transformations processing non-relational data.
2. Slow running ETLs that address data ingestion, data quality, data preparation and aggregation become a very expensive scaling proposition – requiring more compute resources
3. Pushdown SQLs for relevant DW technologies take up expensive resources and starving critical workloads
4. Inflexible data for analytics as requirements evolve. Unmanageable and slow ad hoc analytical queries.
5. Slow BI reporting jobs dropping out of SLAs all the time. Unacceptable response time for queries.
6. Unused and dormant data living around taking up expensive storage space

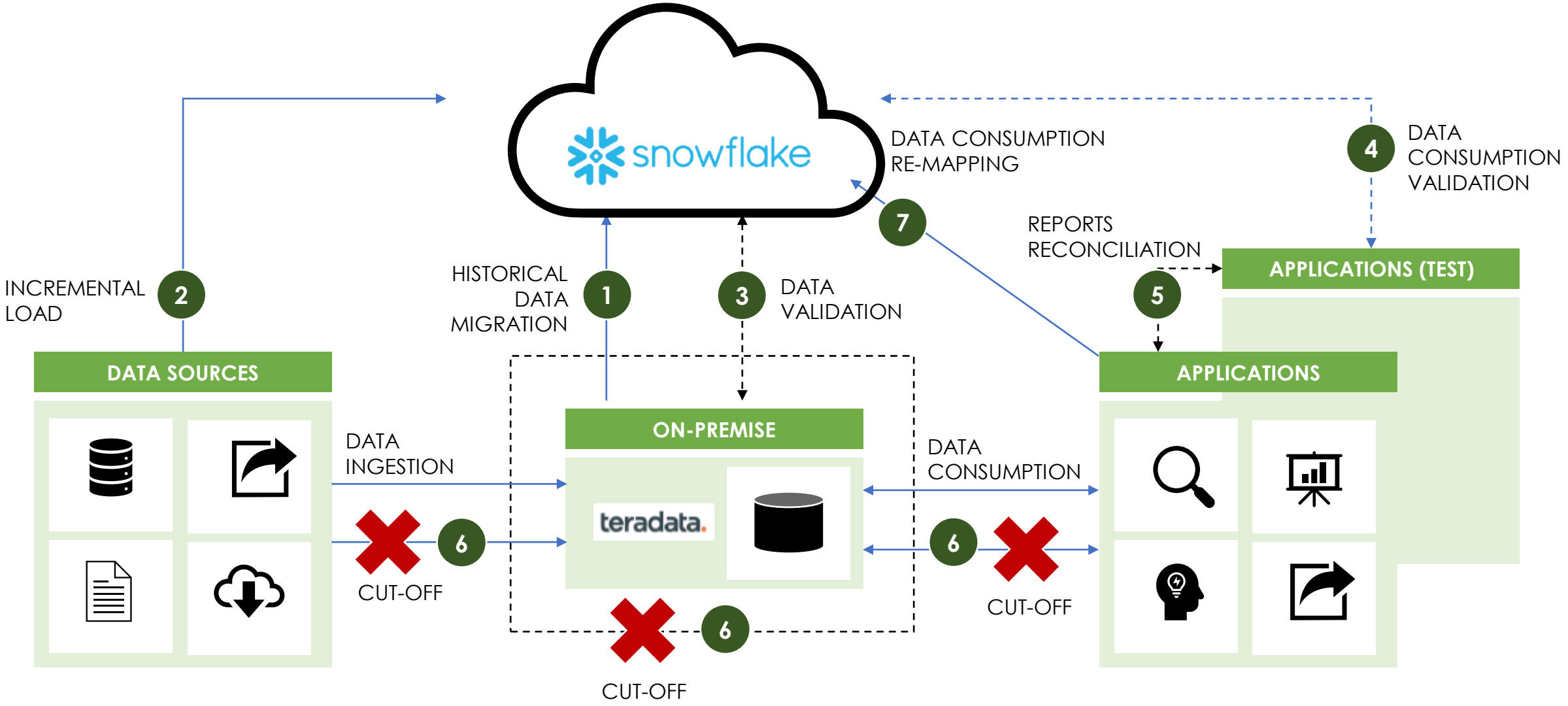
ENTERPRISES ARE MOVING TO SNOWFLAKE FOR MODERNIZING THEIR LEGACY DATA WAREHOUSE ON CLOUD

Snowflake is preferrable over cloud native data warehouses like Redshift.

#	Feature	Redshift	Snowflake
1	Setup	Sizing appropriate cluster as storage & compute are not separated--Complex	No sizing required as storage & compute are separated--Easy
2	Maintenance	Required Vacuuming/Analyzing tables periodically this can be a huge headache with Redshift	Low Maintanance
3	Management	Difficult to manage without skilled AWS Architect	Zero Management from end user
4	Scaling	Does not easily scale up and down	Auto Scaling available
5	Data Cloning	Feature not available	Available
6	Pricing	High Compare to Snowflake	Very Low
7	Multi-Warehouse Concurrency against same data	No	Yes
8	Query Latency	Low at speed Compare to Snowflake	High
9	Time Travel	Not available	Available

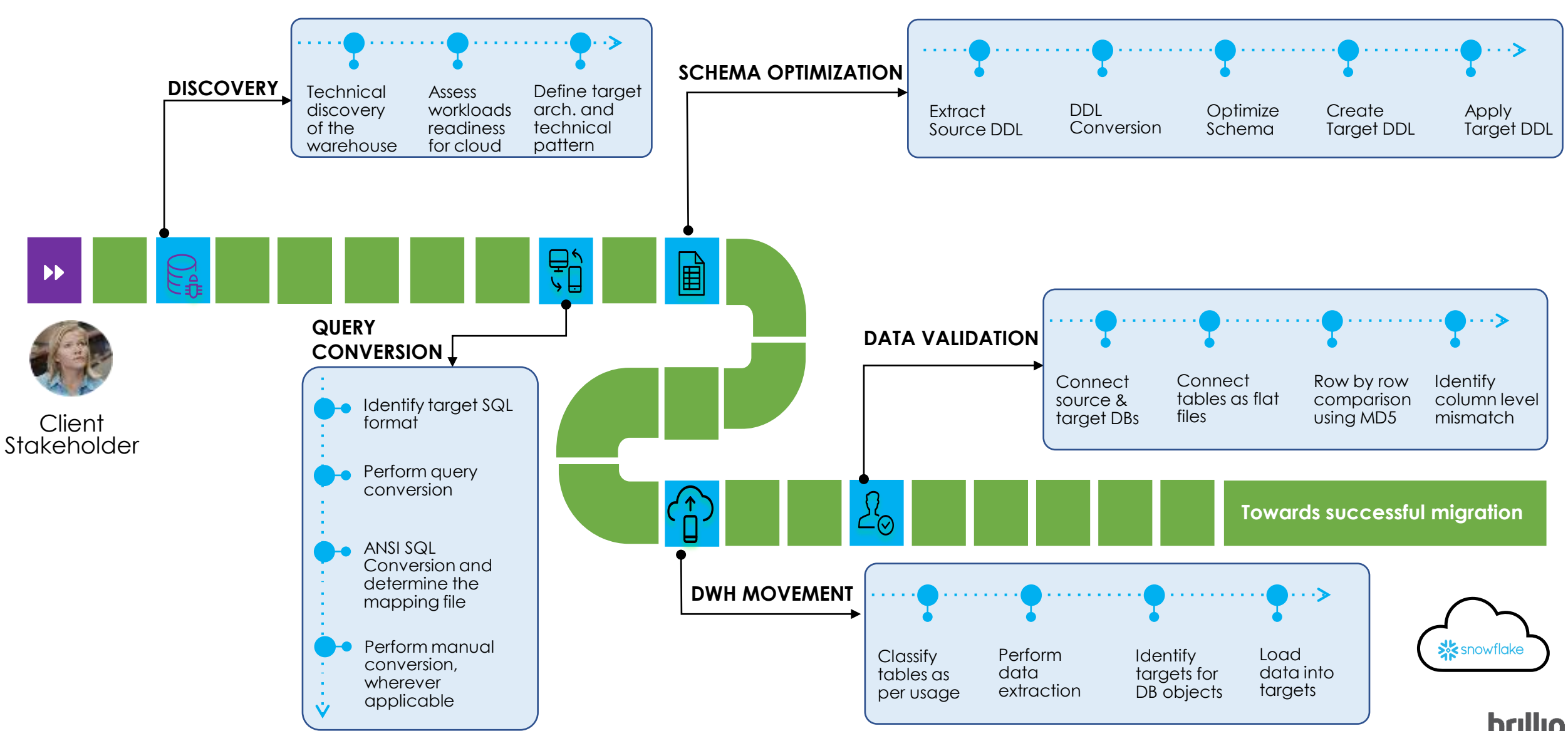
- ANSI SQL support and ACID transactions
- Peta-byte scale
- A fully managed solution
- Seamless scaling capability, ideally ability to scale independently compute and storage
- Cost effective

DATA WAREHOUSE MODERNIZATION ON SNOWFLAKE – PROCESS FLOW



brillio

MODERNIZATION ON SNOWFLAKE - KEY MIGRATION STEPS

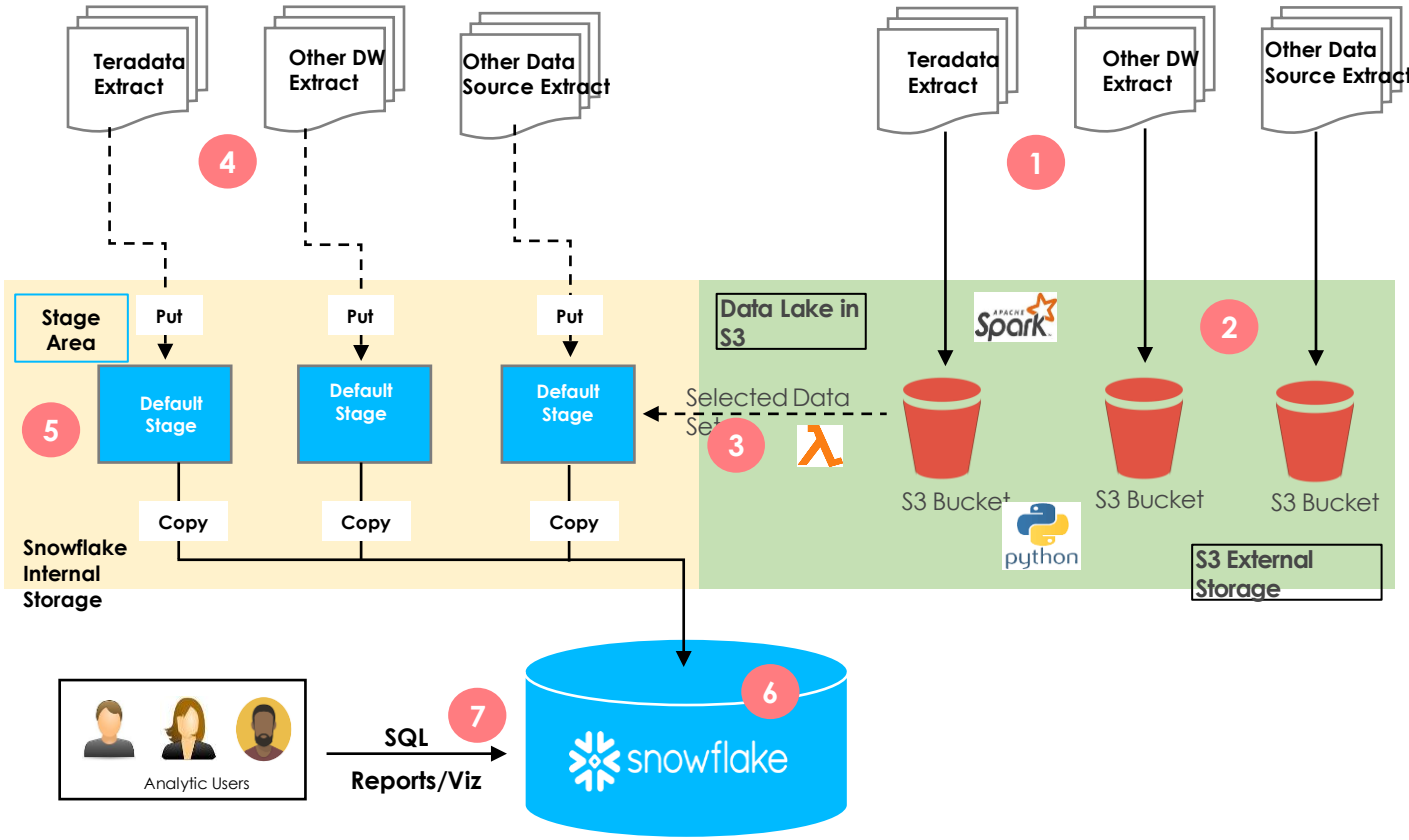


MODERNIZATION ON SNOWFLAKE – MARKET TOOLS AVAILABLE

	CLOUD NATIVE	THIRD PARTY
DATA DISCOVERY	AWS SCT, DB Migration Assistant, Data Transfer Service, DB Conversion Workbench	DATOMETRY, datametica, Cloudera Navigator
QUERY CONVERSION	AWS SCT, DB Migration Assistant, Data Transfer Service, DB Conversion Workbench	RazorSQL, DBConvert, datametica,
SCHEMA OPTIMIZATION	AWS SCT, DB Migration Assistant, Data Transfer Service, DB Conversion Workbench	ATTUNITY REPLICATE, DBConvert, datametica, nifi
DATA MOVEMENT	AWS DataSync, AWS DMS, AWS Data Pipeline, Azure Data Factory, Google Dataflow	ATTUNITY REPLICATE, ORACLE GoldenGate, StreamSets, talend, nifi, Informatica
DATA VALIDATON	AWS DMS, Azure Data Factory, Google Dataflow	QuerySurge, Informatica ETL testing, Datagaps, datametica

MODERNIZATION ON SNOWFLAKE – TRANSITION APPROACH

TERADATA TO SNOWFLAKE TRANSITION



1. Create data extract from various source systems including Teradata and other data warehouses. Send data extract to various buckets in S3 periodically.

2. S3 serves the purpose of Data Lake. Various data processing may take place in S3 using Spark or Python for data transformation based on business requirement. That may produce new data sets.

3. Selected data sets from S3 send to staging area in Snowflake using AWS Lambda, Python or other scripts.

4. It's possible to bring data warehouse extract directly to Snowflake staging area based on business requirement.

5. Staging area in Snowflake serves the purpose of internal landing area. Data is coped periodically from staging area to Snowflake database

6. Snowflake provides organized ways to store data in tables and columns.

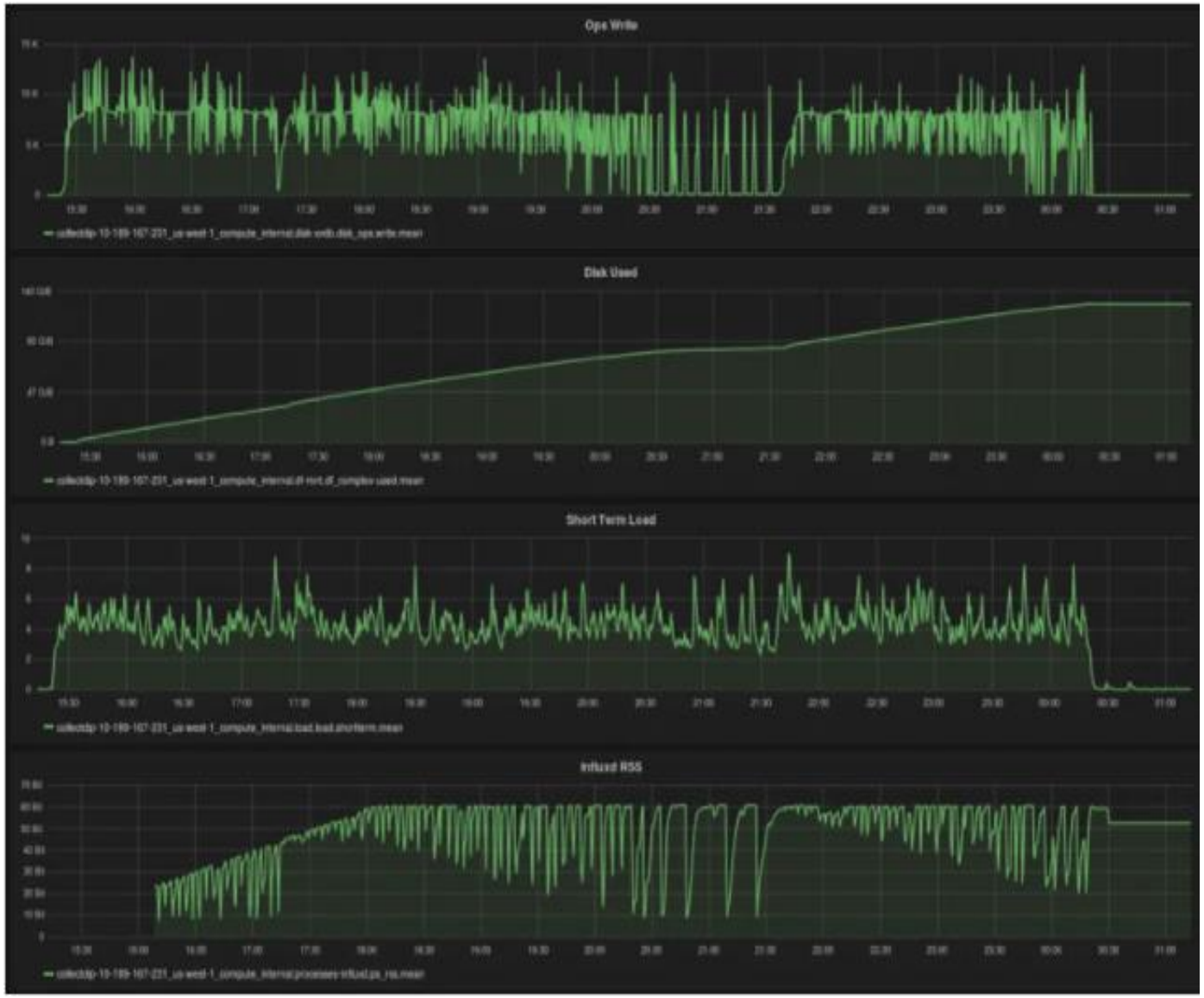
7. Users can run standard SQL and other reporting tools including data visualization tools for data analysis.

brillio

BUILDING TIME SERIES DATABASES ON SNOWFLAKE

TOGETHER WE KNOW HOW!!

WHAT IS A TIME SERIES DATABASE

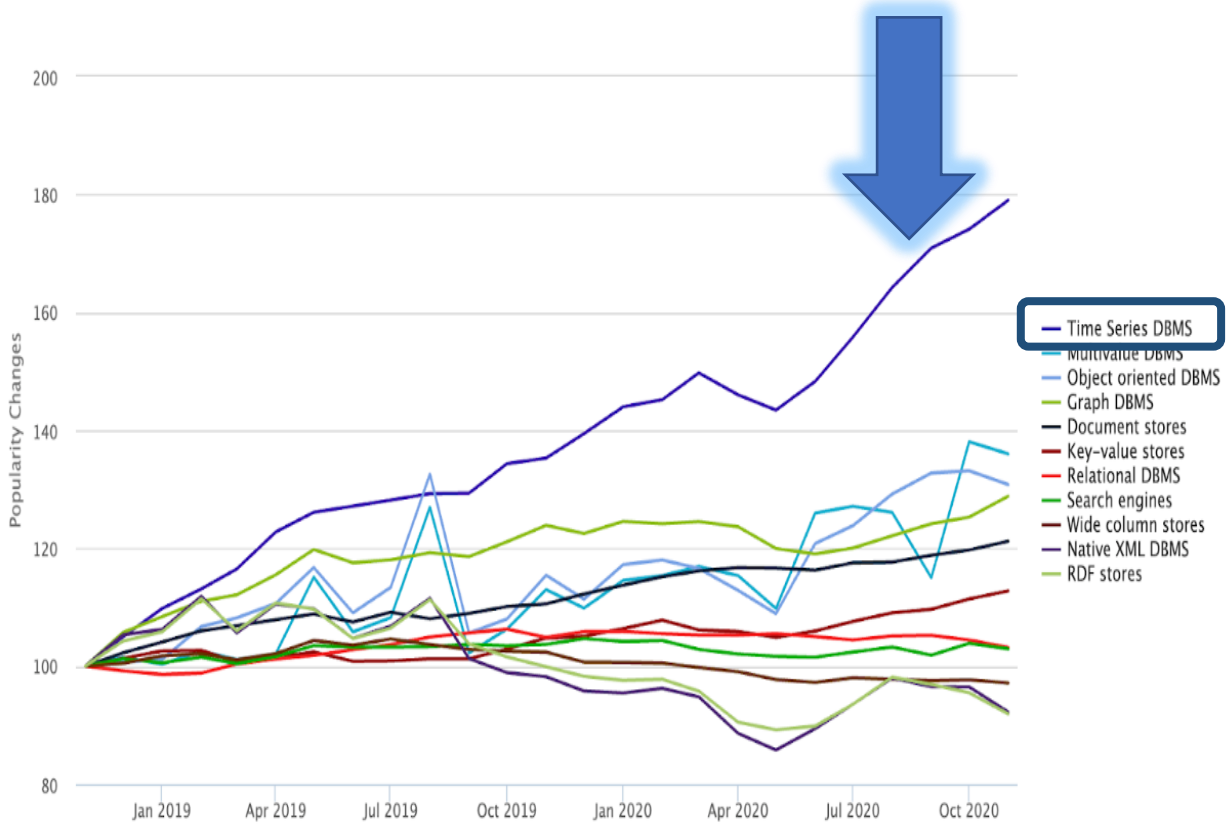


A time series database (TSDB) *is a database optimized for time-stamped or time series data.*

Time series data are simply measurements or events that are tracked, monitored, down-sampled, and aggregated over time.

TIME SERIES DATABASES ARE BECOMING IMPORTANT NOW

In the past 2 years, the rise of the time series database (TSDB) has been meteoric – growing faster than any other database model*



*Reference : DBEngines.com - https://db-engines.com/en/ranking_categories

KEY DRIVERS FOR INCREASED ADOPTION

Data Explosion

- Data universe at 44ZB by 2020 & 180ZB by 2025[1]
- New data sources: Sensors, Connected Devices, Social Media

Technology Advancements

- Cloud: Enabling cost reduction
- Digital: Enabling faster accessibility
- NEW paradigms – AI, ML, AR/VR

Need for Increased Data Analysis

- Data driven business models (Google, Apple, Facebook, Amazon)

brillio

KEY CAPABILITIES OF A TIME SERIES DATABASE

CONCURRENT WRITE PERFORMANCE

Time-series data is more frequently written than read, with 95%-99% of operations being writes.

Therefore, we should focus on a TSDB's ability to write. In most cases, a TSDB must be able to support highly-concurrent and high-throughput data writes.

INTERACTIVE AGGREGATE QUERY

The latency of interactive queries must be very low even when the queried data is enormous in size (measured in TB)

LARGE SCALE DATA STORAGE

The ability to store massive amounts of data: The data size is determined by the characteristics of the scenarios. In most cases, time series data is measured in TB, and even PB.

HIGH AVAILABILITY

Time Series Databases have high availability requirements – to support query throughput and low latency

DISTRIBUTED ARCHITECTURE

Distributed architecture: Considering the requirements of data writes and storage, the underlying layer must be the distributed architecture.

THERE ARE DIFFERENT MARKET SOLUTIONS FOR TIME SERIES DATABASES

InfluxDB is very efficient at collecting, storing, querying, visualizing, and taking action on streams of time series data, events, and metrics in real-time.

TimescaleDB is an open-source relational database that makes SQL scalable for time-series data. Can scale to petabytes horizontally and writes millions of data points per second

QuestDB is a relational column-oriented database that can perform real-time analytics on time series data. It works with SQL and some extensions to create a relational model for time series data.

AWS Timestream is a serverless time series database service that is fast and scalable. It is used majorly for IoT applications to store trillions of events in a day and 1000 times faster with 1/10th cost of relational databases.

OpenTSDB is a scalable time-series database that has been written on top of HBase. It can store trillions of data points at millions of writes per second.

HOWEVER, THEY HAVE A SIMILAR SET OF CHALLENGES

KEY CHALLENGES

NON-STANDARD INTERFACE

Fine for connecting from custom applications or for data science, but hard to connect to standard systems like BI tools and standard ETL/ELT tools.

LACK OF MATURITY

While there are 'cool technologies' in the time series database world, the overall space is still relatively niche and low volume and therefore most of these systems don't have the deep maturity from having a very large customer base

RELATIVELY IMMUTABLE

Updates are either not supported or very slow – this is because one of the ways that a TSDB can be very fast is to store data in a way optimized for write and read but making updates extremely difficult

NARROW FOCUS

The problem with time series databases is that they ONLY support the time series data. When there is other information to store, another database will be needed.

Using two different databases doesn't really help since there is still no ability (without the complexity and performance hit of putting something like Presto over the top of both) to run queries and analytics using time series and non time series data at once.

SNOWFLAKE DELIVERS ON THE PROMISE OF TIME SERIES DATABASES (1/2)

HOW SNOWFLAKE ADDRESSES KEY TSDB GAPS

STANDARD INTERFACE

Snowflake provides a standard SQL interface for connecting BI applications, ETL/ELT tools along with custom apps and data science models

PRODUCT MATURITY

Snowflake provides a powerful, scalable, robust data storage and retrieval solution – and ensures sure that data is validated and fully accessible to stakeholders when needed.

RELATIVELY IMMUTABLE

Snowflake's unique architecture makes it possible to run multiple workloads concurrently, without performance impact, by using more than one concurrently running warehouse.

Query performance and table optimization is taken care of with micro-partitions and data clustering

NARROW FOCUS

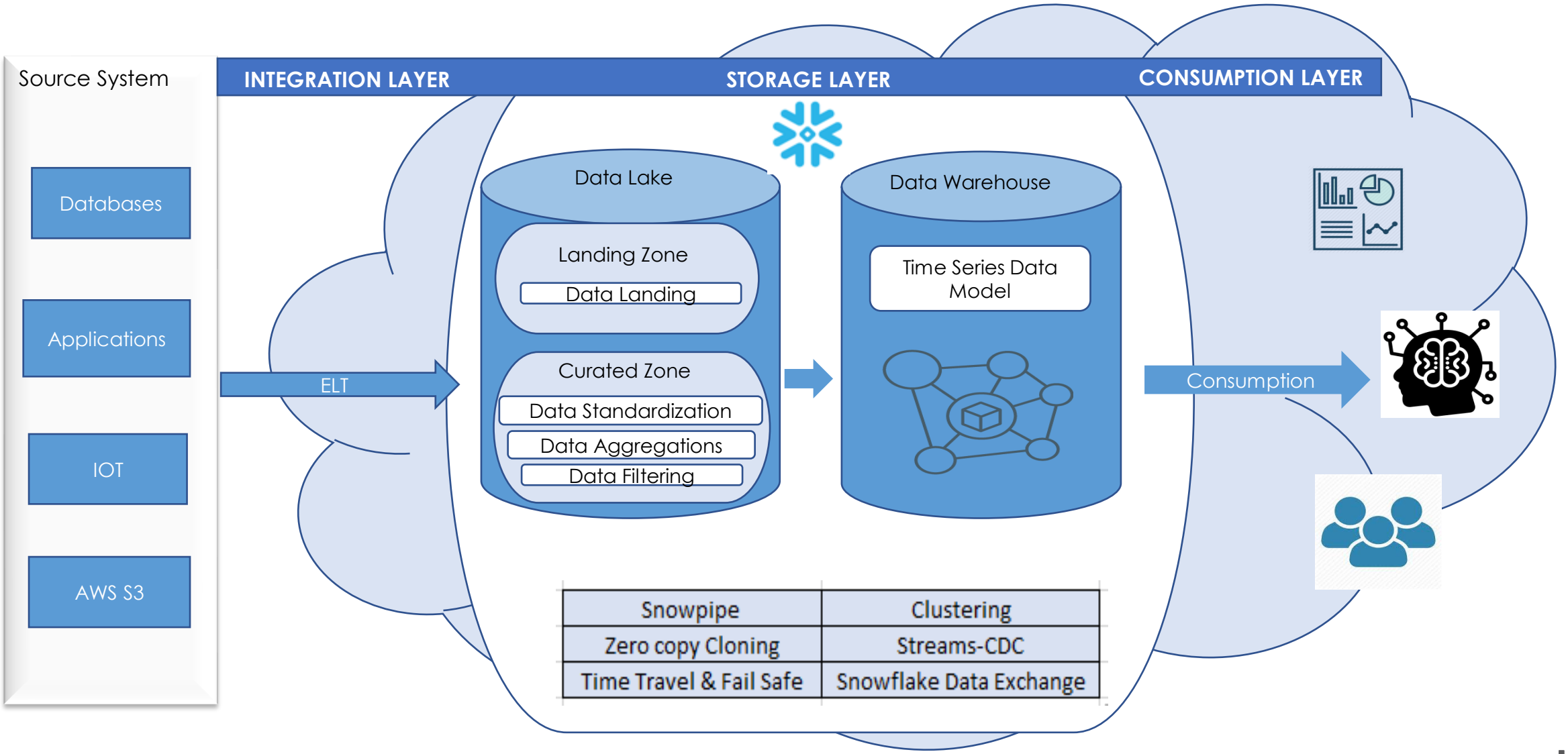
Snowflake Cloud Data Platform supports storage of different data sets and sharing through Snowflake Data Exchange – this allows users to interact with both time series and non-time series datasets

SNOWFLAKE DELIVERS ON THE PROMISE OF TIME SERIES DATABASES (2/2)

Dimension	Key Capabilities
1 DATA MANAGEMENT	From a data management perspective, Snowflake can serve as the central time-series data repository or data can be loaded in inexpensive cloud storage from Amazon S3, Azure Data Lake Storage or Google Cloud storage.
2 SCALIBILITY	From a scalability perspective, Snowflake's multi-cluster shared data architecture provides superior performance that is only paid for when it is used. Compute resources can be spun as XS or 4XL compute notes based on analytic workloads
3 DATA WRANGLING	From a data wrangling perspective, Snowflake allows time-series data to be prepared and combined with other content in a common data platform. Repetitive quality checking, restructuring and integrations can be automated as they mature so that data scientists spend more time analyzing rather than manipulating data
4 ANALYTICS	From an analytics and data science perspective, Snowflake has connectors to a wide variety of tools to support exploratory activities. Tools such as Power BI and Spotfire are supported out of the box
5 DATA SHARING	From a training perspective, Snowflake is extremely efficient and fast and has powerful technology for data sharing". "This allows us to store information once and share with many different clients, cutting down on the total storage we require and avoiding duplicating information."

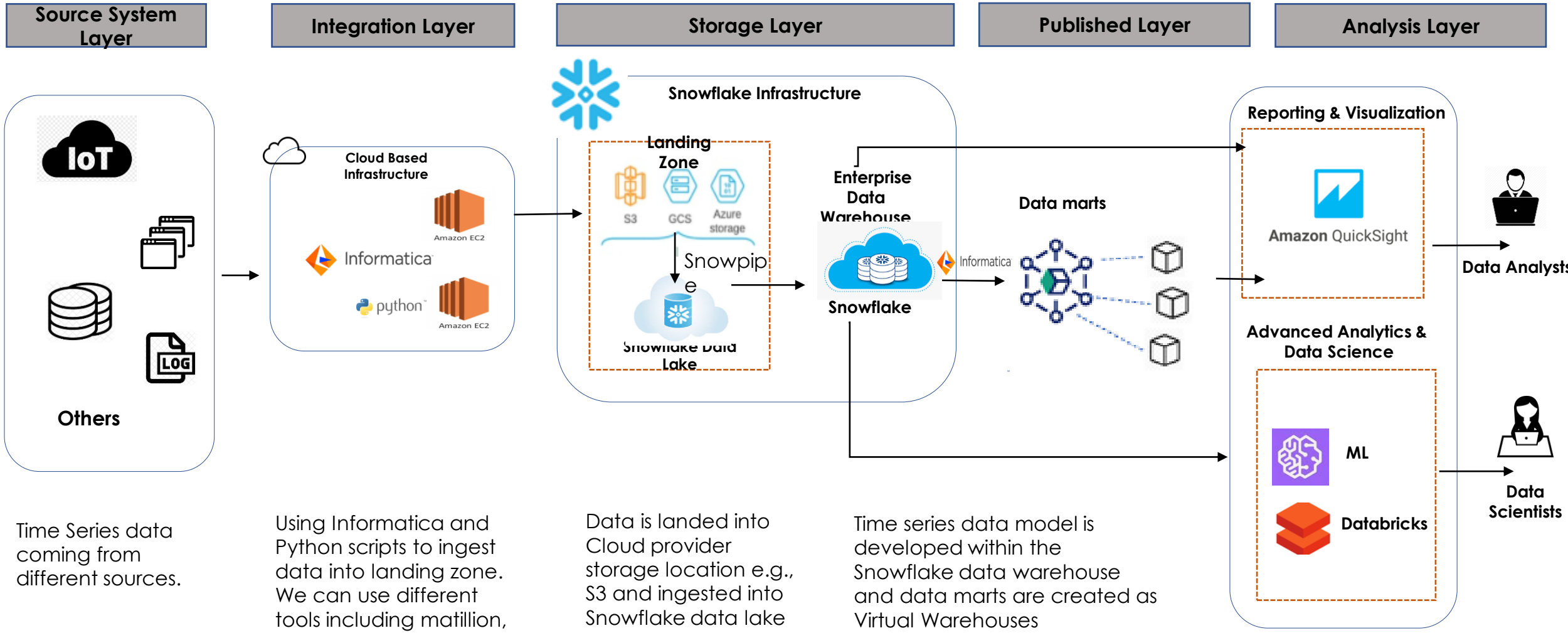
BUILDING A TIME SERIES DATABASE ON SNOWFLAKE

CONCEPTUAL VIEW



BUILDING A TIME SERIES DATABASE ON SNOWFLAKE

TECHNICAL ARCHITECTURE



Time Series data coming from different sources.

Using Informatica and Python scripts to ingest data into landing zone. We can use different tools including matillion, StreamSets and native Snowflake utilities like Snowpipe.

Data is landed into Cloud provider storage location e.g., S3 and ingested into Snowflake data lake using Snowpipe.

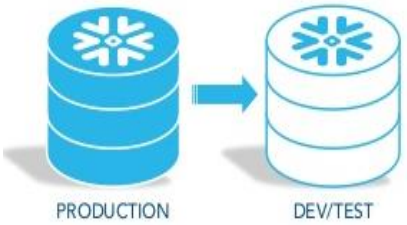
Time series data model is developed within the Snowflake data warehouse and data marts are created as Virtual Warehouses

brillio

BUILDING A TIME SERIES DATABASE ON SNOWFLAKE

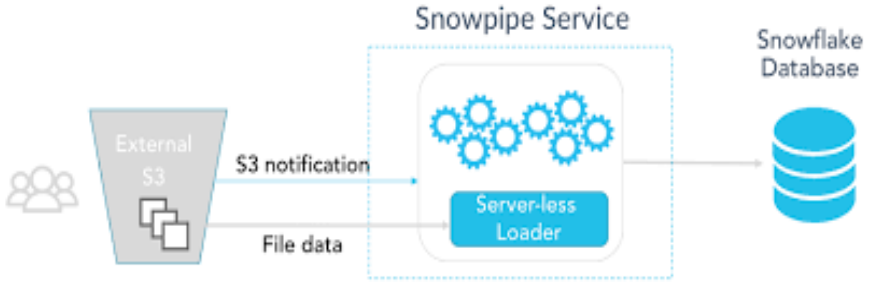
SNOWFLAKE UTILITIES (1/2)

Snowpipe



Snowpipe enables loading time series data from files as soon as they're available in a stage. This means you can load data from files in micro-batches, making it available to users within minutes, rather than manually executing COPY statements on a schedule to load larger batches.

Zero Copy Cloning



Zero copy cloning is a snowflake implementation where a simple keyword CLONE lets you create copy of your tables, schemas, databases without copying the actual data. So, you can have almost real time TS data from production cloned into your dev and stage environments to be able to perform various actions

Time Travel & Fail Safe

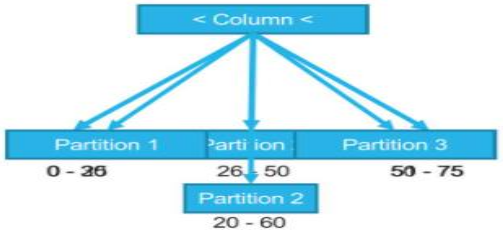
Snowflake provides powerful CDP features for ensuring the maintenance and availability of your historical time series data (i.e. data that has been changed or deleted):

- Querying, cloning, and restoring historical data in tables, schemas, and databases for up to 90 days through Snowflake Time Travel.
- Disaster recovery of historical data (by Snowflake) through Snowflake Fail-safe.

BUILDING A TIME SERIES DATABASE ON SNOWFLAKE

SNOWFLAKE UTILITIES (2/2)

Clustering



Snowflake has a way of partitioning data to optimize read-time performance by allowing the query engine to prune unneeded data quickly. In Snowflake, the partitioning of the data is called clustering, which is defined by cluster keys you set on a table.

Data Marketplace

Snowflake Data Marketplace utilizes Snowflake Secure Data Sharing to connect providers of data with consumers. Providers can publish and monetize their Time series data and Consumers can use that data for analysis.

Streams-CDC

A stream is a new Snowflake object type that provides change data capture (CDC) capabilities to track the delta of changes in a table, including inserts and data manipulation language (DML) changes, so action can be taken using the changed data. A table stream allows you to query a table and consume a set of changes to a table, at the row level, between two transactional points in time.

THANK YOU