Session 3- Day3- Snowpro Core certification Snowflake Deep Dive

Santosh Ubale, GSI Sales Engineer 25 May 2021

AGENDA

Day 3

- Snowflake Citibike demo
- Snowflake AWS/Azure/GCP Enablement & Reference Architectures
- 6 workload overview section
- Snowflake migration approach , different phases and Business use case discussion
- Teradata To Snowflake Migration Approach Using Automated Tools

Snowflake AWS/Azure/GCP Enablement

Some Key Terminology

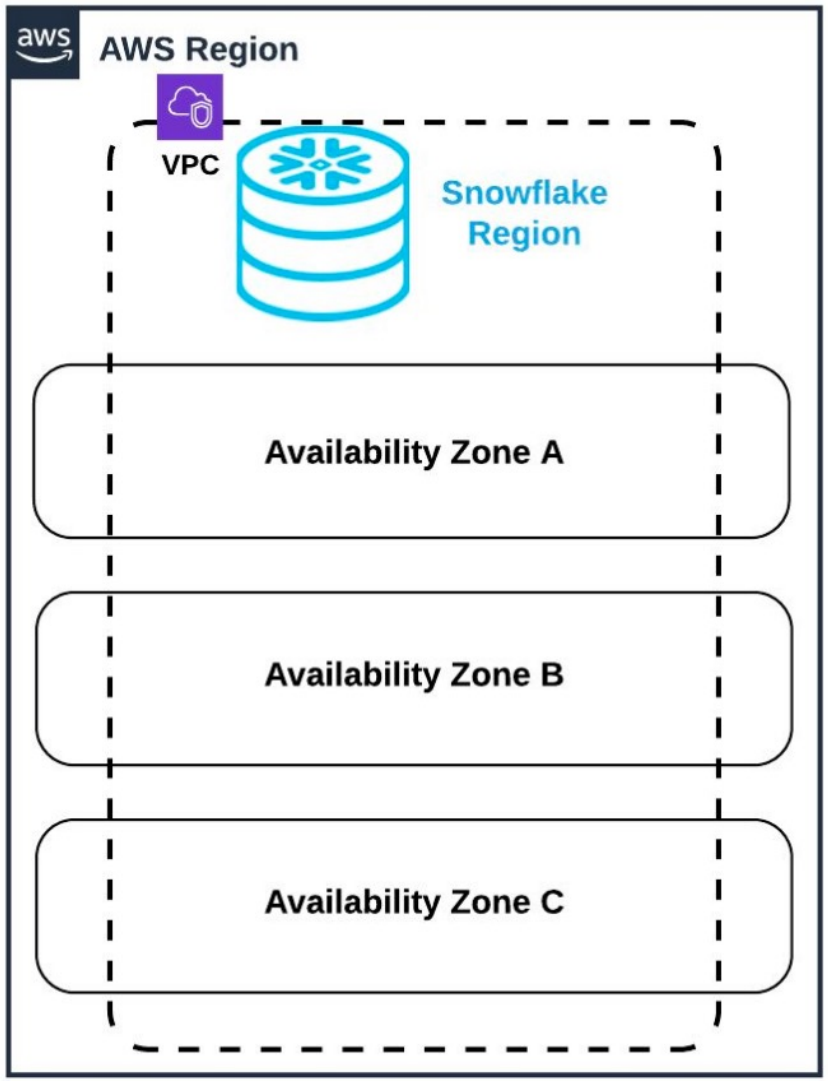
Service	AWS	Azure	GCP
Organizing Entity	Account	Subscription	Project
Object Storage	S3	Azure Storage	Google Cloud Storage (GCS)
Compute	EC2	Azure Compute	Google Compute Engine (GCE)
Storage Container	Bucket	Container	Bucket
Virtual Private Cloud	VPC	VNet	VPC
Dedicated Connection	Direct Connect	ExpressRoute	Cloud Interconnect
Serverless Compute	AWS Lambda	Azure Functions	Cloud Functions
Identity Management	IAM	Azure AD	Cloud IAM / Cloud Identity

Snowflake on AWS

- Engineered to provide equivalent features, and price / performance parity, with other cloud platforms (GCP, Azure)
- Built on standard AWS services such as Elastic Compute Cloud (EC2), Simple Storage Service (S3), Amazon Virtual Private Cloud (VPC)
- In 2020 Snowflake is deployed in 11 AWS regions - 5 NA, 2 EMEA & 4 APAC
- Integrates with a number of AWS services through native integration or leveraging the Snowflake connectors and drivers (with some configuration) - (e.g. AppFlow, MSK, EMR, SageMaker, Glue, etc.)
- Replication across regions / clouds for data migration or business continuity

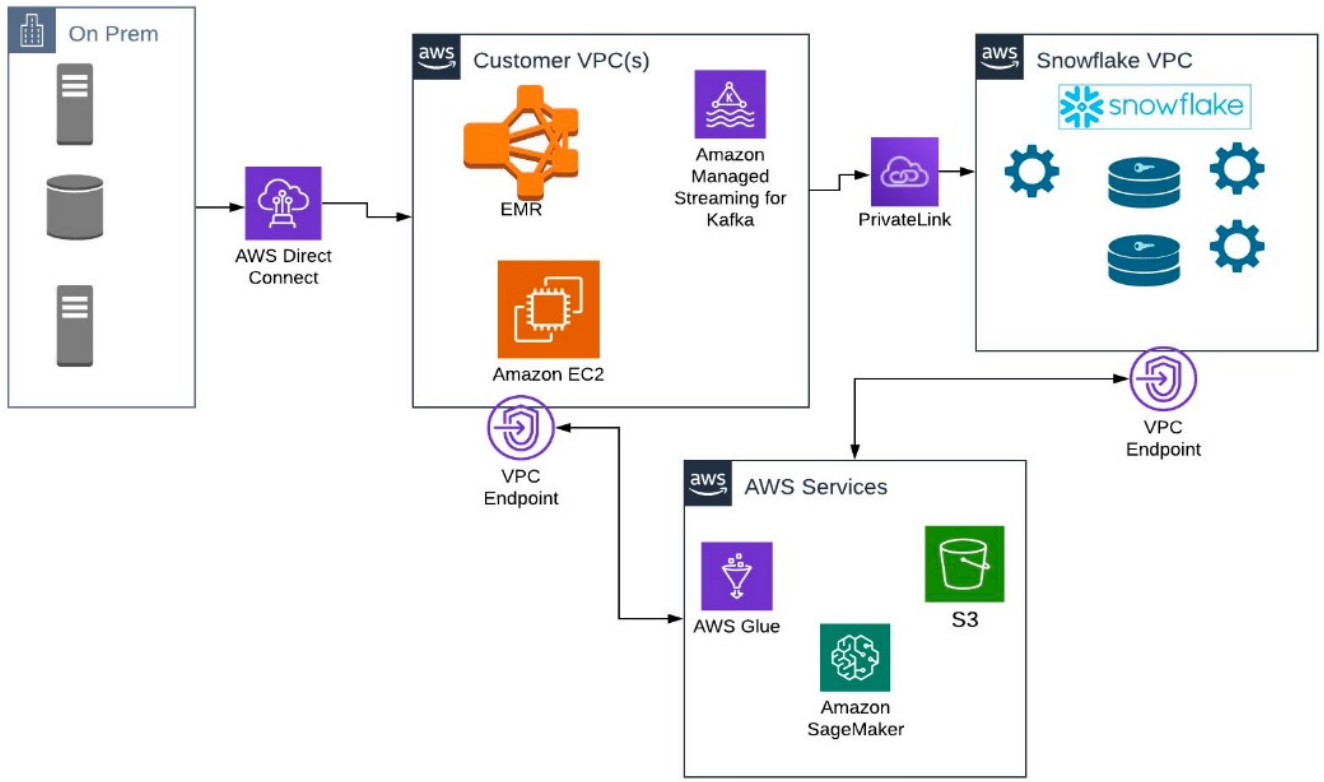
Snowflake deployment in an AWS region

- AWS regions have 3-4 availability zones (AZ)
- An AZ is one or more discrete data centers that are separated by a meaningful distance
- A VPC spans all AZs in a region
- A Snowflake region is deployed in a VPC owned and managed by Snowflake, across all AZs for availability and scale

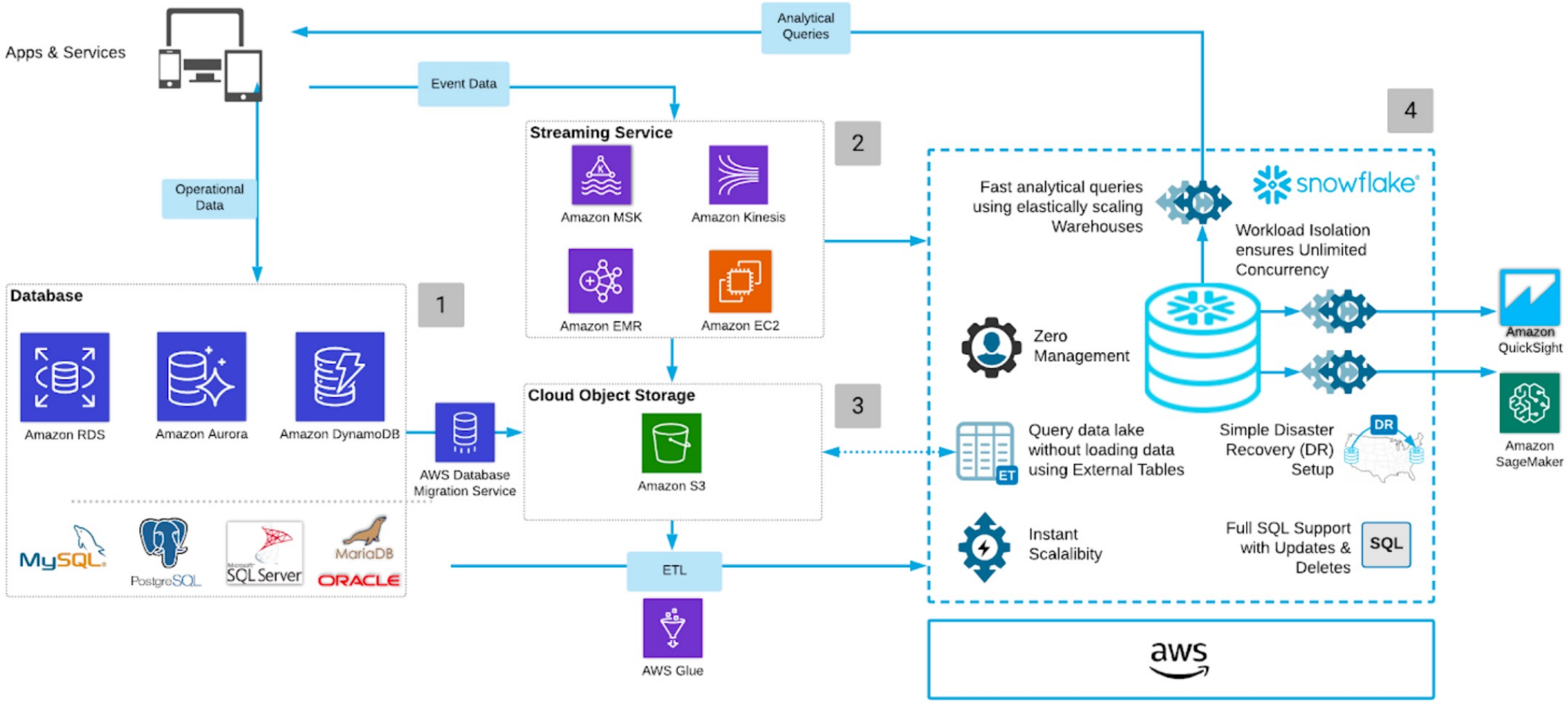


Customer and AWS Services connecting to Snowflake

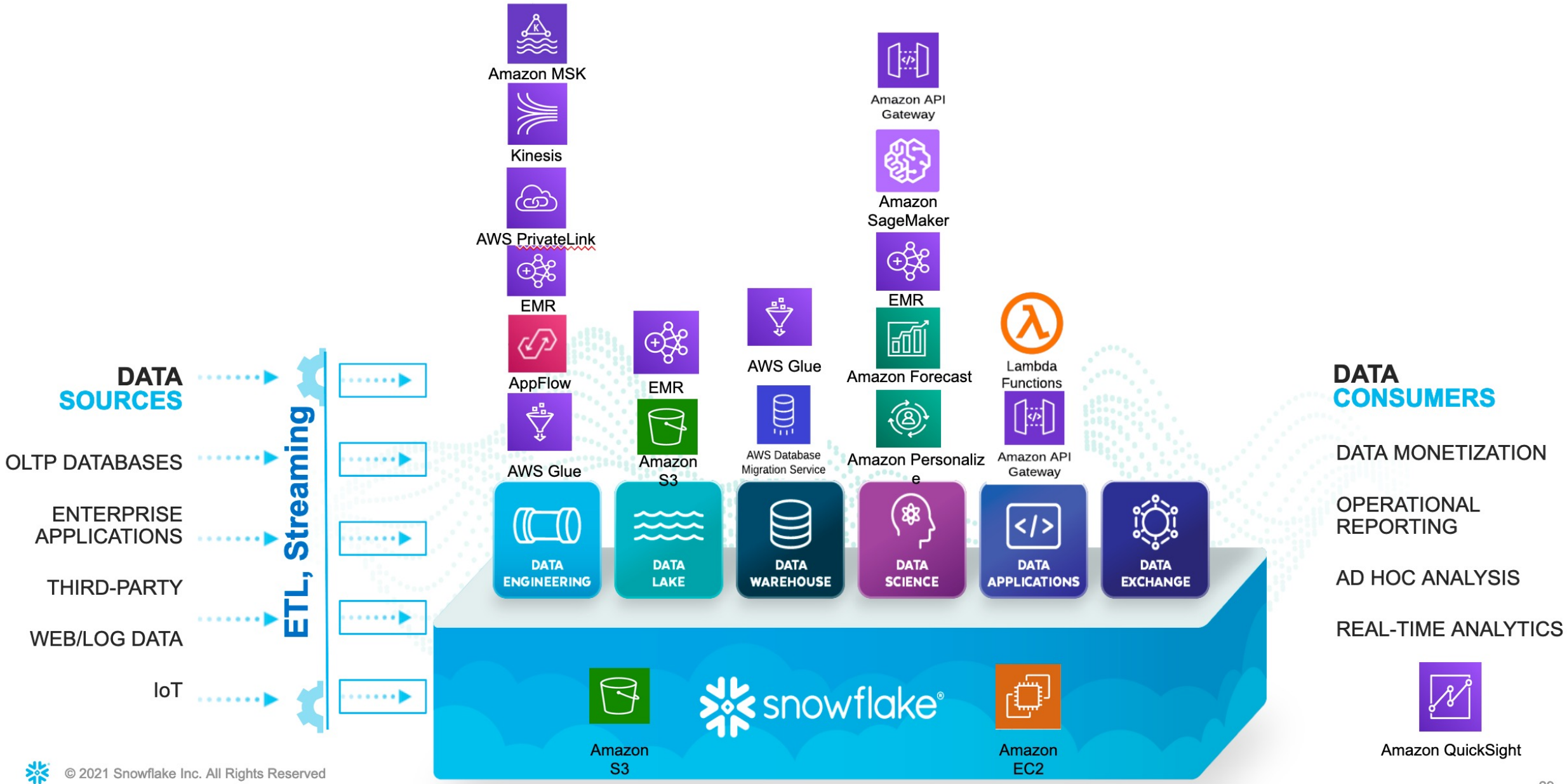
- Customers need to connect their on prem and their own VPCs to Snowflake
- They also need to connect AWS Services to Snowflake
- Various AWS Services are available for secure networking connectivity



Cloud Data Warehouse Stack With Snowflake



AWS Services for different Snowflake workloads



Snowflake on Azure

- Engineered to provide equivalent features, and price / performance parity, with other cloud platforms (AWS, GCP)

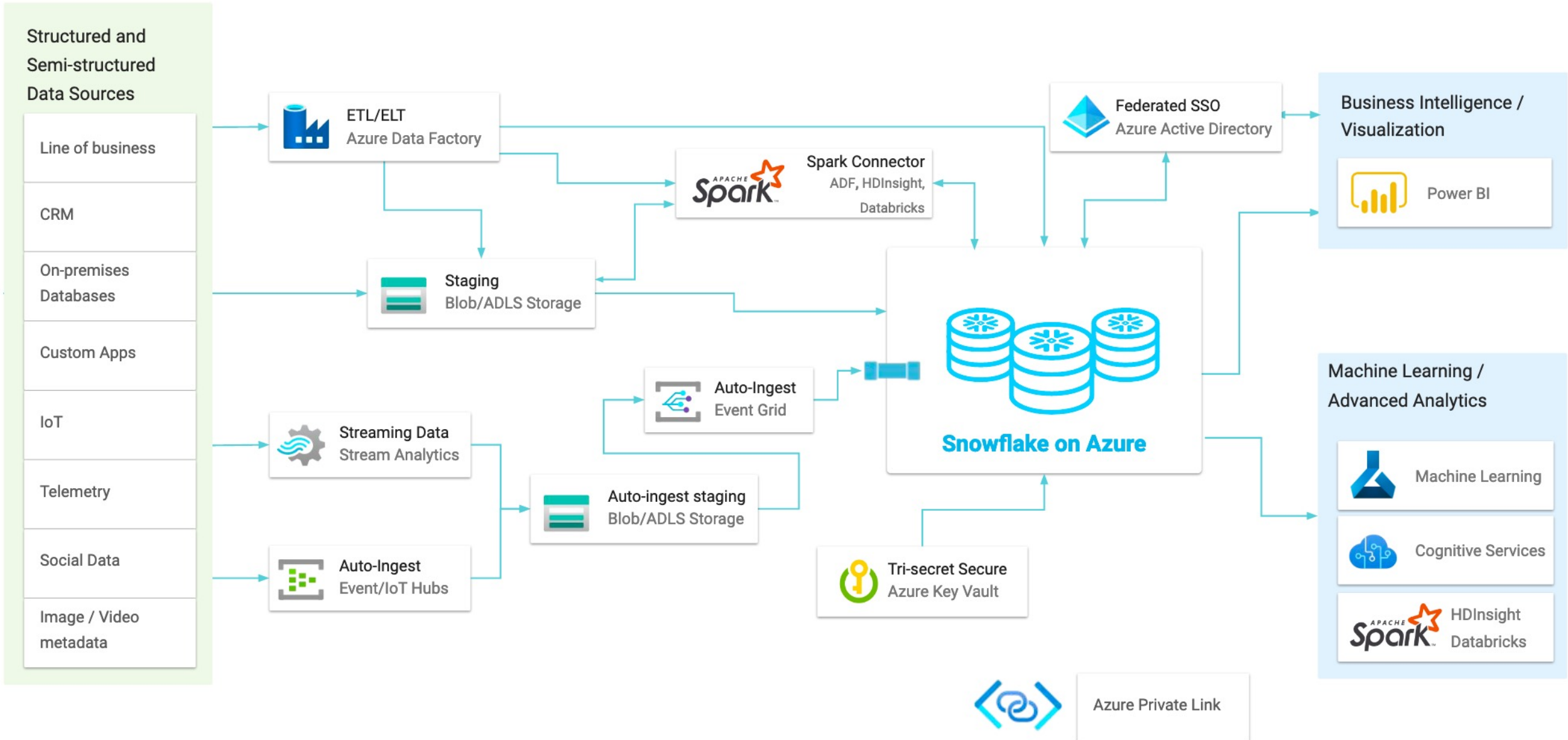
- Built on standard Azure services such as Azure VMs, Azure Blob Storage and Azure Virtual Networks(VNet)

- Today, Azure Snowflake is Generally Available in East US 2 (Virginia), West US 2 (Washington), Canada Central (Toronto), US Gov (Virginia), West Europe (Netherlands), Switzerland North (Zurich), Southeast Asia (Singapore), Australia East (New South Wales)

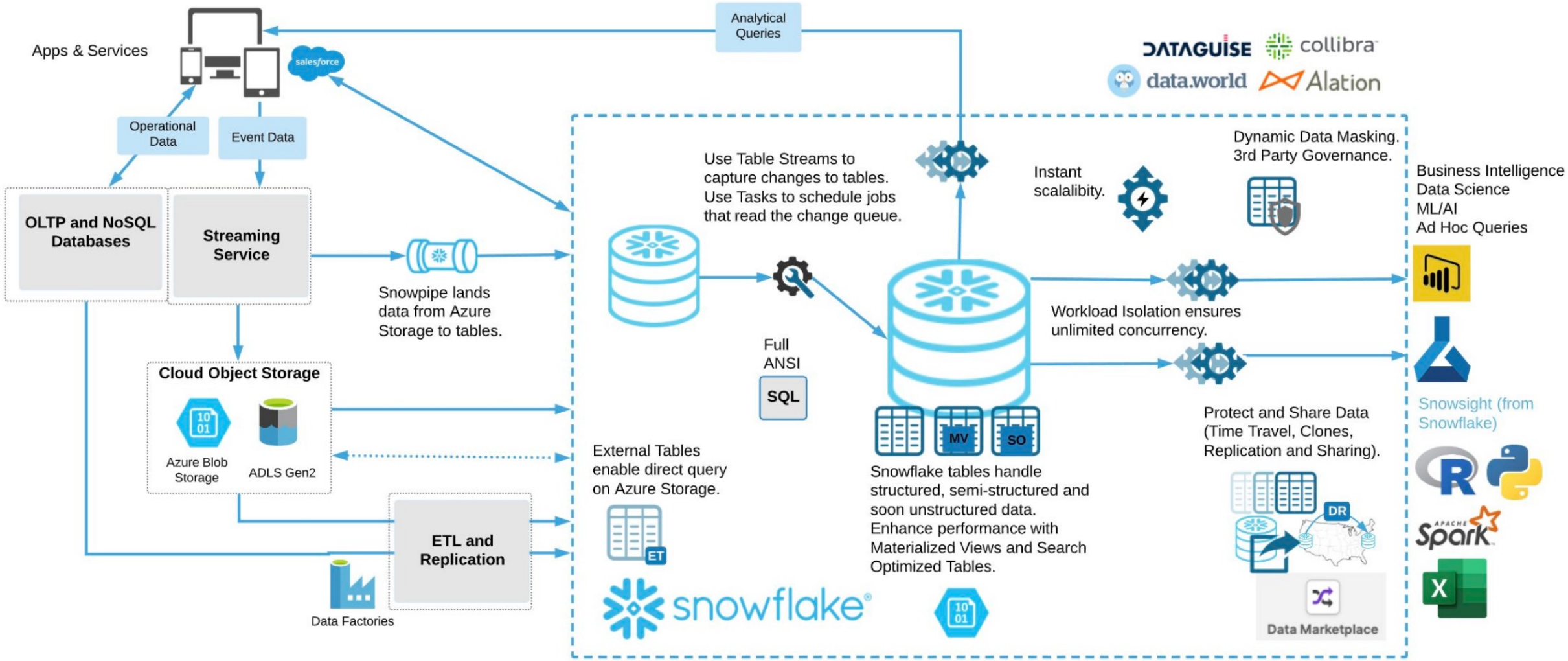
- Integrates with Azure services: Azure Data Factory (as a Native Sink) Power BI (with support for SSO) Storage (ADLS Gen 2 Blob and Azure Blob) Security (Azure PrivateLink)

- Replication across regions / clouds for data migration or business continuity

Snowflake on Azure Architecture



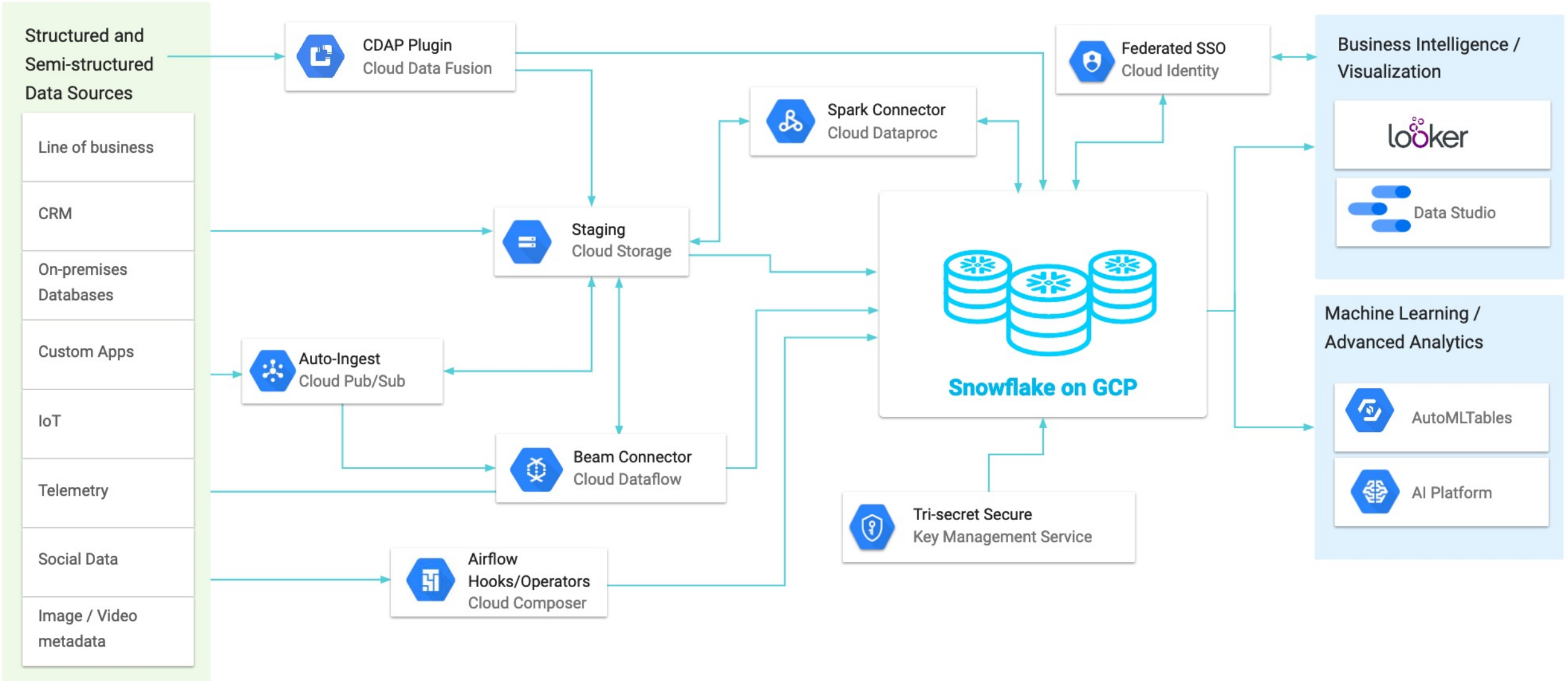
Snowflake on Azure Integration Architecture



Snowflake on Google Cloud Platform

- Engineered to provide equivalent features, and price / performance parity, with other cloud platforms (AWS, Azure)
- Built on standard GCP services such as Google Compute Engine (GCE), Google Cloud Storage (GCS), Google Virtual Private Cloud (VPC)
- As of April, 2020, Generally Available in us-central1 (Iowa), europe-west2 (London) and europe-west4 (Netherlands)
- Integrates with GCP analytic services (e.g. Cloud Dataflow, Dataproc, etc.)
- Replication across regions / clouds for data migration or business continuity

Snowflake on GCP Integration Architecture



➢6 workload over view section

SNOWFLAKE PLATFORM

Under the hood

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

AD HOC ANALYSIS

REAL-TIME ANALYTICS

SNOWFLAKE'S DATA ENGENNING PLATFORM

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

AD HOC ANALYSIS

REAL-TIME ANALYTICS

DATA ENGINEERING WITH SNOWFLAKE

	1	2	3
WORKFLOW	**Data Ingestion**	**Data Transformation**	**Delivery**
DATA ENGINEERING FEATURES	Auto-load stage data using **Snowpipe** Highly performant, cloud-based **storage ingest** Ingest streaming data using **Kafka Connector** **Schema Detection**	Transformation with **ANSI SQL** or **Snowpark** Continuous transformation with **Streams & Tasks** Enhanced extensibility with **functions & stored procedures** Live backup access through **Time Travel**	Easily collaborate with **Live Data Sharing** Built-in support for common **drivers** (ODBC, JDBC, Python, Spark,etc.) Support an open ecosystem with **Data Lake Export**
PLATFORM FEATURES	Unmatched support for **semi-structured data** with **schema-on-read** **The Data Cloud**	Optimized pipeline performance with **instant warehouse resizing** Eliminate data duplication with **Zero Copy Cloning**	Concurrency scaling with **multi-clustered warehouse** **The Data Cloud**

SNOWFLAKE CLOUD DATA PLATFORM

- **One platform, one copy of data, many workloads**
- **Secure and governed access to all data**
- **Near-zero maintenance, as a service**
- **Unlimited performance and scale**

SNOWFLAKE'S DATA LAKE PLATFORM

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

AD HOC ANALYSIS

REAL-TIME ANALYTICS

DATA LAKE WITH SNOWFLAKE

CLICK ON NAV BAR TO GO TO DEEP DIVE

WORKFLOW	1 Data Ingestion	2 Data Transformation	3 Consumption
DATA LAKE FEATURES	Auto load stage data using **Snowpipe** Highly performant, cloud-based storage ingest Incremental ingestion from your data lake using **Materialized Views** over External Tables	Fast & simplified transformation using **ANSI SQL** Enhanced extensibility with **functions and stored procedures** Continuous transformation with **Streams & Tasks** Support an open ecosystem with **Data Lake Export** and Spark connector	Directly query data in your data lake with **External Tables** Better query performance using **Materialized Views** over External Tables Synchronize External Tables with **Apache Hive** metastore Accelerate data exploration with **Snowsight**
PLATFORM FEATURES	Unmatched support for **semi-structured data** with **schema-on-read**	Optimized pipeline performance with **Instant warehouse resizing** Eliminate data duplication with **Zero Copy Cloning**	**Elastic Scaling** with multi-clustered warehouse Secure **data sharing** & **data exchange** Encryption, RBAC, PCI, HIPAA, FedRAMP Dynamic **Data Masking** & External **Tokenization**

SNOWFLAKE CLOUD DATA PLATFORM

- **One platform, one copy of data, many workloads**
- **Secure and governed access to all data**
- **Near-zero maintenance, as a service**
- **Unlimited performance and scale**

SNOWFLAKE'S DATA WAREHOUSE PLATFORM

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

AD HOC ANALYSIS

REAL-TIME ANALYTICS

DATA WAREHOUSE WITH SNOWFLAKE

	1	2	3	4
WORKFLOW	Source Data	Data Processing	Modeling and Deploying	Consumption and Sharing
DATA WAREHOUSE FEATURES	**Snowpipe** and **Kafka** streaming **MVs** on **External Tables** **Schema-less** load	External Functions SQL transformation	Search Optimization Service Dynamic Data Masking & External Tokenization Time travel Zero copy cloning	**Drivers** for (ODBC, JDBC, Python, etc.) **Snowsight UI** for Dashboards & visualizations Extensive **BI Partner ecosystem**
PLATFORM FEATURES	Automatic data compression and partitioning Variant data type **semi-structured data** (i.e., JSON) **Data Sharing & Data Marketplace** for access to external datasets	Extensive **partner ecosystem** for ETL, ELT, data cataloging	Elastic scaling with multi-cluster warehouses Encryption, RBAC, PCI, HIPAA, FedRAMP Replication cross-cloud & region Near-zero maintenance (no indexing, partitioning, vacuuming) Automatic updates with no downtime	Data Lake Export Full support for **ANSI SQL** **External** functions Secure data sharing through **Data Exchange**

SNOWFLAKE CLOUD DATA PLATFORM

- **One platform, one copy of data, many workloads**
- **Secure and governed access to all data**
- **Near-zero maintenance, as a service**
- **Unlimited performance and scale**

SNOWFLAKE'S DATA SCIENCE PLATFORM

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

AD HOC ANALYSIS

REAL-TIME ANALYTICS

CHALLENGES WITH DATA SCIENCE

Data silos make data collection time-consuming

Relevant data is scattered across multiple systems or unavailable within organization

Compute bottlenecks slow down data preparation

Lengthy data preparation process made longer by slow or complex systems

Isolated teams and tools fail to generate value

Lack of collaboration and cohesive toolchain prevents ML models from going beyond experiments

DATA SCIENCE WITH SNOWFLAKE
BEST PRACTICES

Enrich datasets using **Data Marketplace** for improved model accuracy

Use **Streams & Tasks** to build end-to-end ML pipelines

Create datasets without loading data into Snowflake via **External Tables**

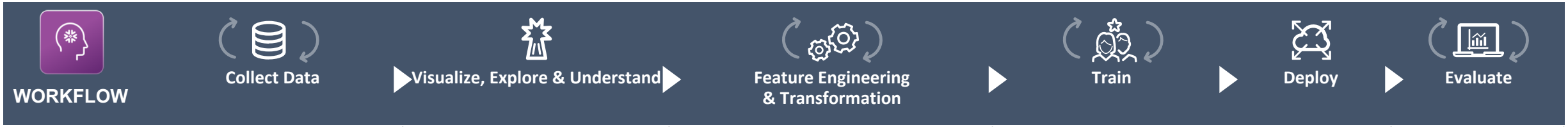
Leverage **External Functions** to trigger training or get predictions

Use **Zero-Copy Clones** for training snapshots

Use regular or Materialized **Views** to create repository of ML features used for training and prediction

Optimize training instance memory usage by using **Snowflake SQL** for aggregation & sampling

DATA SCIENCE WITH SNOWFLAKE



SNOWFLAKE FEATURES

Data Sharing & Data Marketplace for access to external datasets

Schema-on-Read for semi-structured data (eg. JSON)

Process streaming data using **Kafka Connector** and **Snowpipe**

Query cloud storage without loading data using **External Tables**

Dashboards & visualizations using **Snowsight** & BI partners

Quick & easy feature engineering using **ANSI SQL Views**

Functional data engineering in Java/Scala via **Snowpark***

Data enrichment & pipeline orchestration via **External Functions**

Transformation using **Streams and Tasks**

Private sandboxes without duplicating data with **Zero-Copy Cloning**

Extensive Partner Ecosystem

Amazon SageMaker · DataRobot · H2O.ai · data iku · zepl · alteryx

Bulk ML inference using **Java UDFs***

Built-in support for common drivers (ODBC, JDBC, Python, and more)

Easily Connect Your ML Toolchain

TensorFlow · PyTorch · Keras · scikit-learn · Spark · RStudio · XGBoost

Model Management

Amazon SageMaker · DataRobot · H2O.ai · data iku · zepl · alteryx

Persist predictions and ground-truth in Snowflake for easy evaluation

SNOWFLAKE PLATFORM

One platform, one copy of data, many workloads

Secure and governed access to all data

Near-zero maintenance, as a service

Unlimited performance and scale

SNOWFLAKE'S DATA APP PLATFORM

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

AD HOC ANALYSIS

REAL-TIME ANALYTICS

WHAT APP BUILDERS NEED

Meet Scalability and Concurrency Demands

Easily scale apps horizontally and vertically, cost-effectively—and deliver fast response times regardless of load.

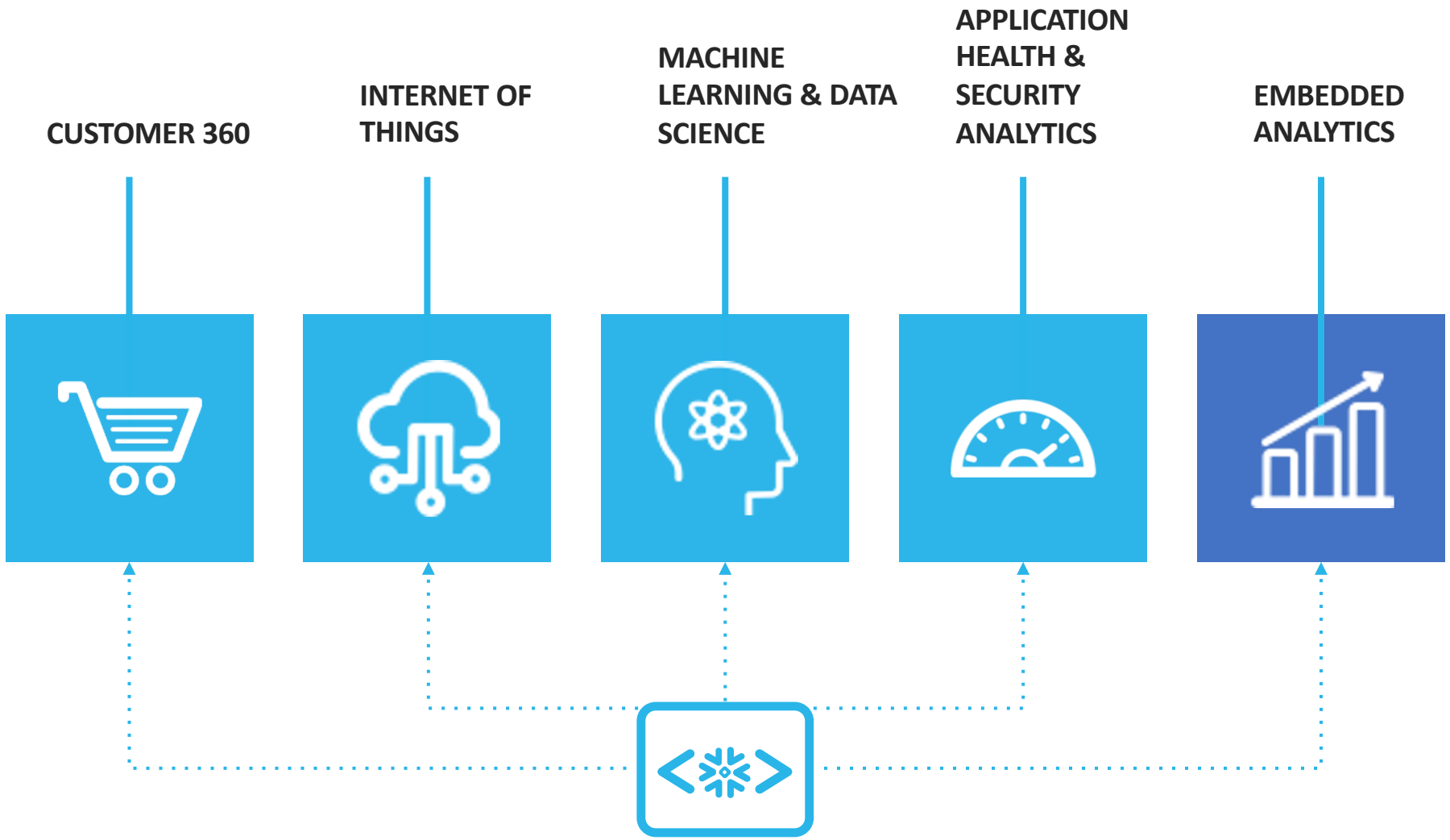
Simple, Fast and Flexible Data Ingestion

Ingest structured and semi-structured data, including JSON, without having to hire engineers to maintain data pipelines.

Reduced Operational Burden

Get higher developer productivity–there is no need to provision and maintain infrastructure.

USE CASES



SNOWFLAKE'S DATA SHARING PLATFORM

DATA SOURCES

OLTP DATABASES

ENTERPRISE APPLICATIONS

THIRD-PARTY

WEB/LOG DATA

IoT

DATA CONSUMERS

DATA MONETIZATION

OPERATIONAL REPORTING

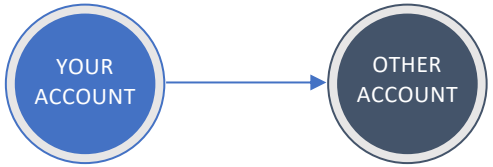
AD HOC ANALYSIS

REAL-TIME ANALYTICS

SECURE DATA SHARING

A single, live copy of the data: no copying, no moving, no delays.

DIRECT SHARE



For sharing data 1-to-1

Square

Square sharing marketing data with a partner

DATA EXCHANGE

YOUR DATA EXCHANGE IN YOUR ACCOUNT

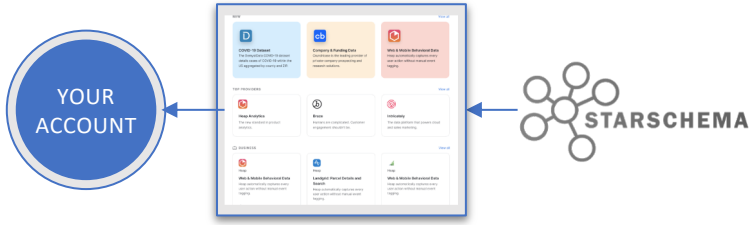
For data sharing AT SCALE:

many parties, many data sets

CISCO

Cisco's data exchange for its suppliers

SNOWFLAKE DATA MARKETPLACE



SNOWFLAKE DATA MARKETPLACE

Consuming data from 3rd parties:

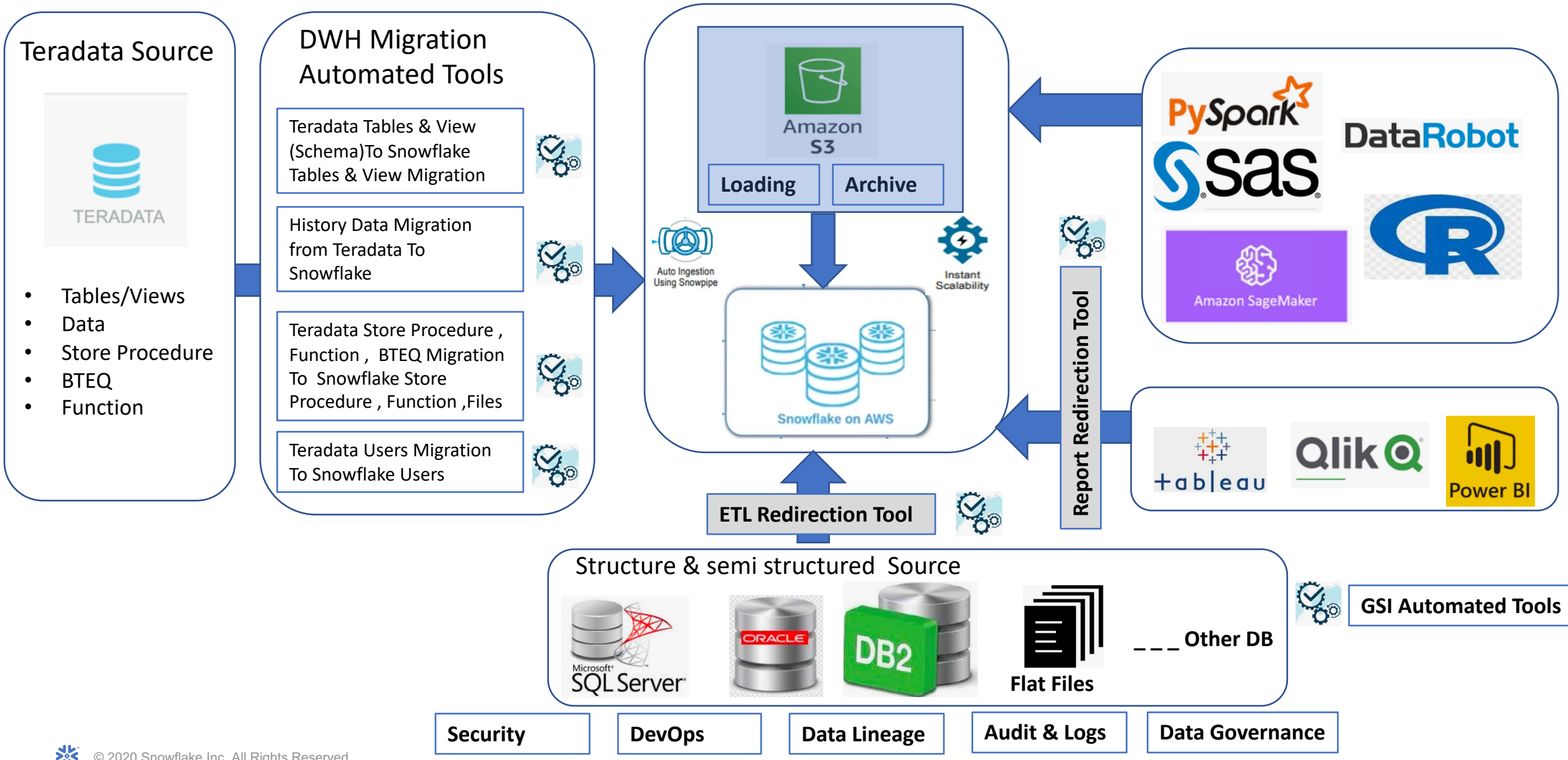
Access data from over 150 providers

STARSCHEMA

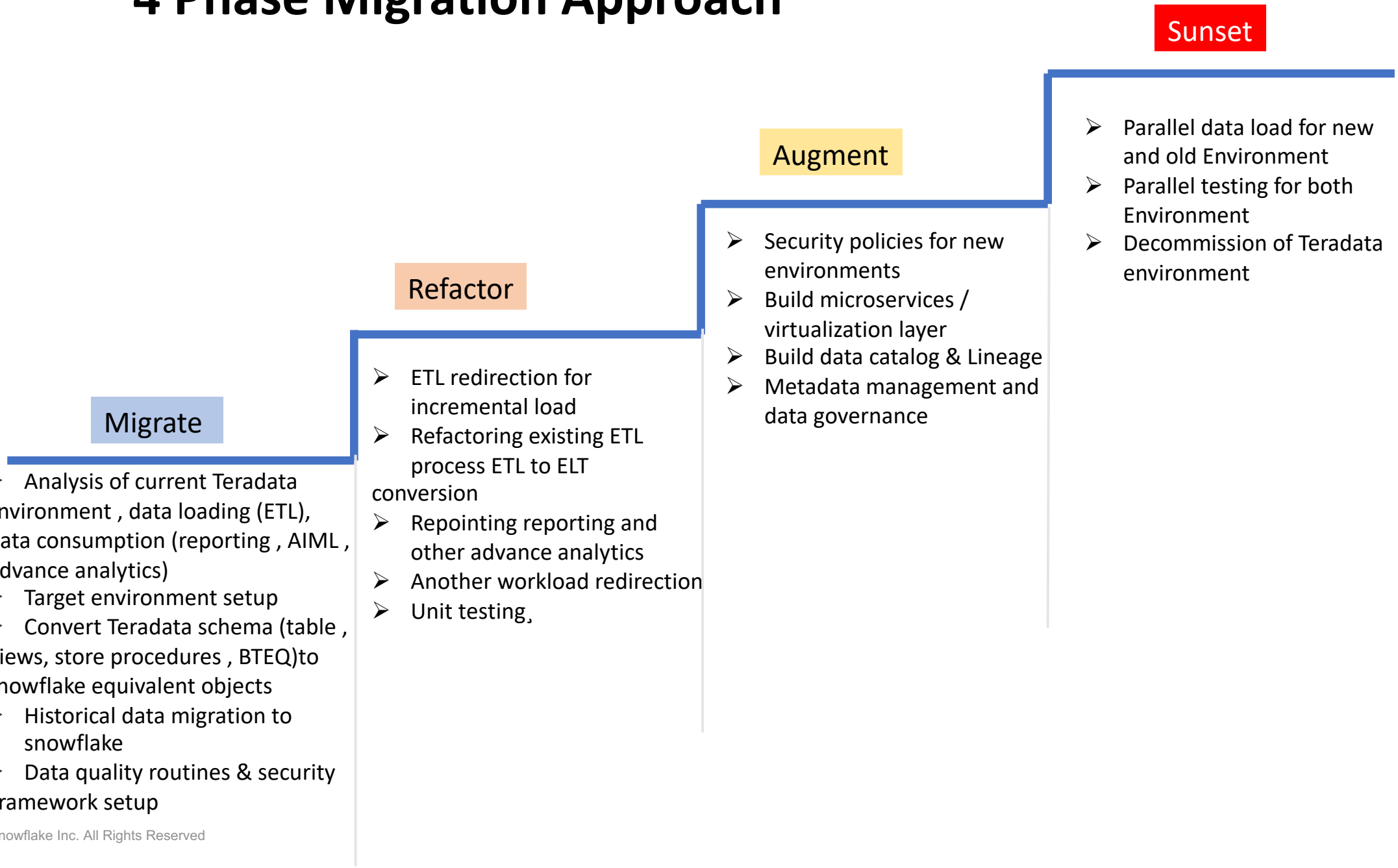
Starschema COVID data

Teradata To Snowflake Migration Approach Using Automated Tools

Teradata To Snowflake Migration Approach Using Automated Tools



4 Phase Migration Approach



- Analysis of current Teradata Environment , data loading (ETL), Data consumption (reporting , AIML , Advance analytics)
- Target environment setup
- Convert Teradata schema (table , Views, store procedures , BTEQ)to Snowflake equivalent objects
- Historical data migration to snowflake
- Data quality routines & security framework setup

- ETL redirection for incremental load
- Refactoring existing ETL process ETL to ELT conversion
- Repointing reporting and other advance analytics
- Another workload redirection
- Unit testing.

- Security policies for new environments
- Build microservices / virtualization layer
- Build data catalog & Lineage
- Metadata management and data governance

- Parallel data load for new and old Environment
- Parallel testing for both Environment
- Decommission of Teradata environment

E2E Indicative Migration Delivery Plan

Wave 1: Environment Setup & Core Component Migration (Migrate + Refactor + Augment)

Wave 2: Report Migration (Sunset)

Wave 3: Incremental Data Mart Migration

Lift & Shift – LOB Wise – Incremental Approach

7 To 9 Months Indicative Delivery Timelines for 10-12 resources

1-4 Months

4-7 Months

2-9 Months

Migration Plan based On Below Inventory

Data Volume = 300 TB
Tables = 3000
Reports = 9000
LOB = 3
Store Procedures = 300
BTEQ= 100
ETL Jobs = 100

Planning & Analysis

Environment Setup

Scheme Migration

Store Procedure , Function , BTEQ Script Migration

Security & Governance Setup

History Data Migration

Delta Incremental Data Migration

ETL & Report Redirection

Business/Mart wise Incremental Data Rollout

Data Copy

Business/Schema 1

Business/Schema 1

Business/Schema 1

Cutover: Business1/Schema1 1-5 months

Cutover: Business2/Schema2 2-6 months

Cutover: Business3/Schema3 3-7 months

Report Migration

Due Diligence

Repointing To Snowflake

Report Metadata Extraction & Change

Report Testing

Sprint Wise Report Delivery

Started to Delivery for business benefits by redirecting history data reporting

Indicative Target Operating Model

Governance

Migration Strategy & Architecture alignment

Program Management

SLA , KPI managements | Knowledge Management | Service Transition

Demond

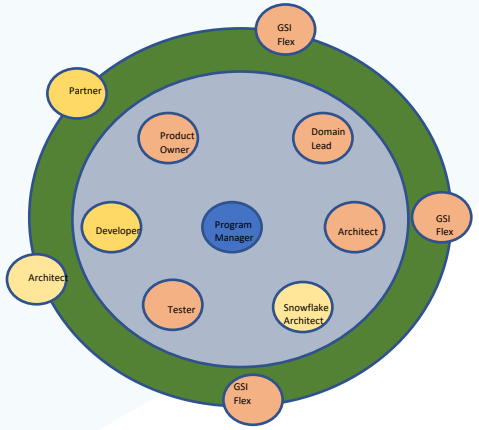
Migration Prioritization | Migration Identification | Migration Planning

Design

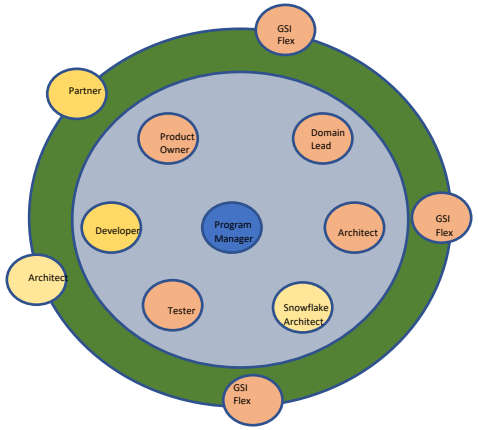
Deliver

Assurance

Migration Pods



ETL & Report Repointing Pod

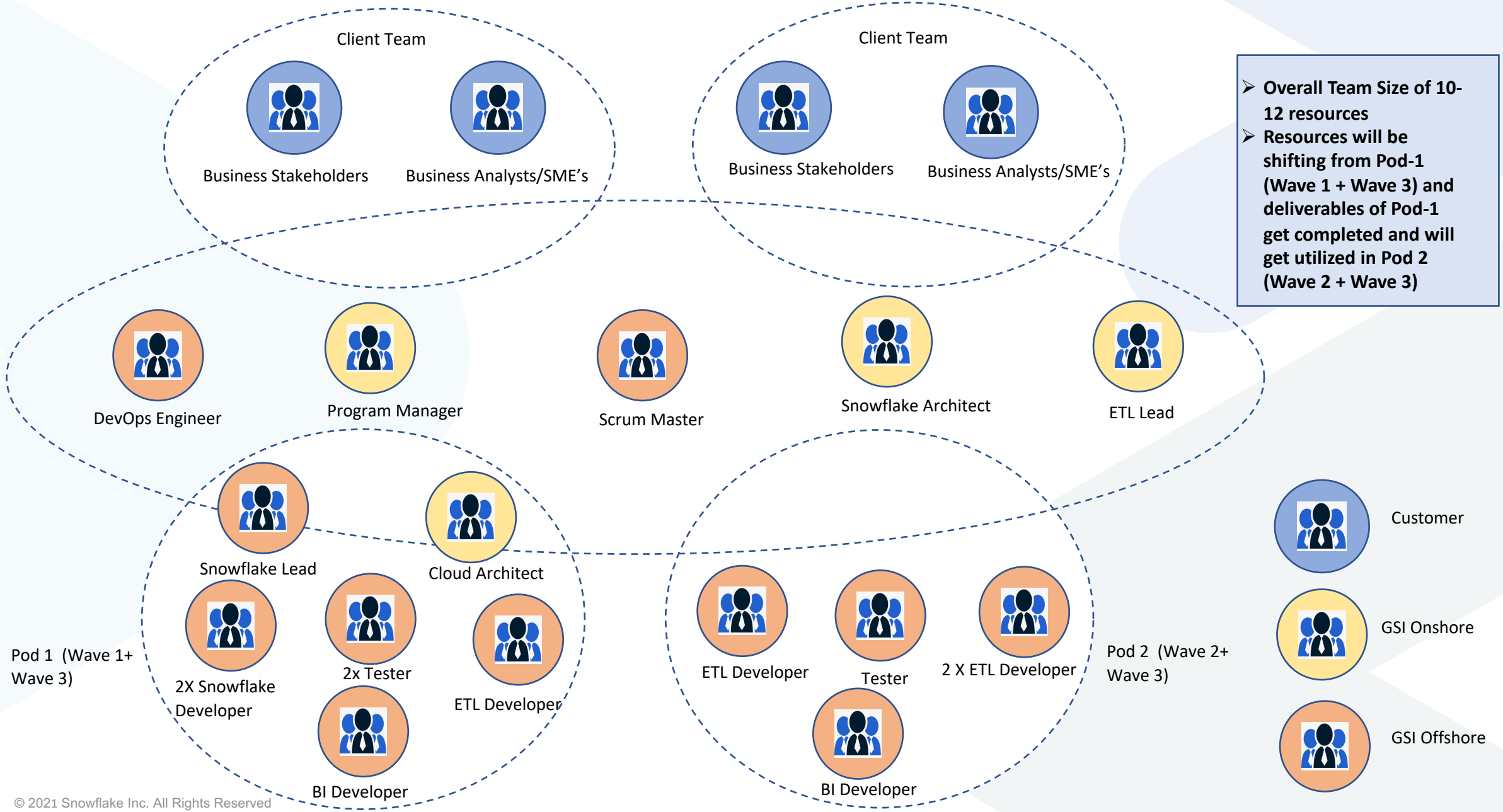


Innovation Lighthouse

Snowflake Partnership | GSI Automation tools & Accelerators | GSI COE

- Fully Integrated governance and transformation structure for better transparency
- Pod based delivery team as per skilled requirement
- Proven automated tools will be used for migration
- Integrated program management to help address seamless execution
- Core and flex team operating models the spikes in demand from various business units / products in case of scope increases

Indicative Pod Based Team Structure



- **Overall Team Size of 10-12 resources**
- **Resources will be shifting from Pod-1 (Wave 1 + Wave 3) and deliverables of Pod-1 get completed and will get utilized in Pod 2 (Wave 2 + Wave 3)**

Thank You