Data Modernization

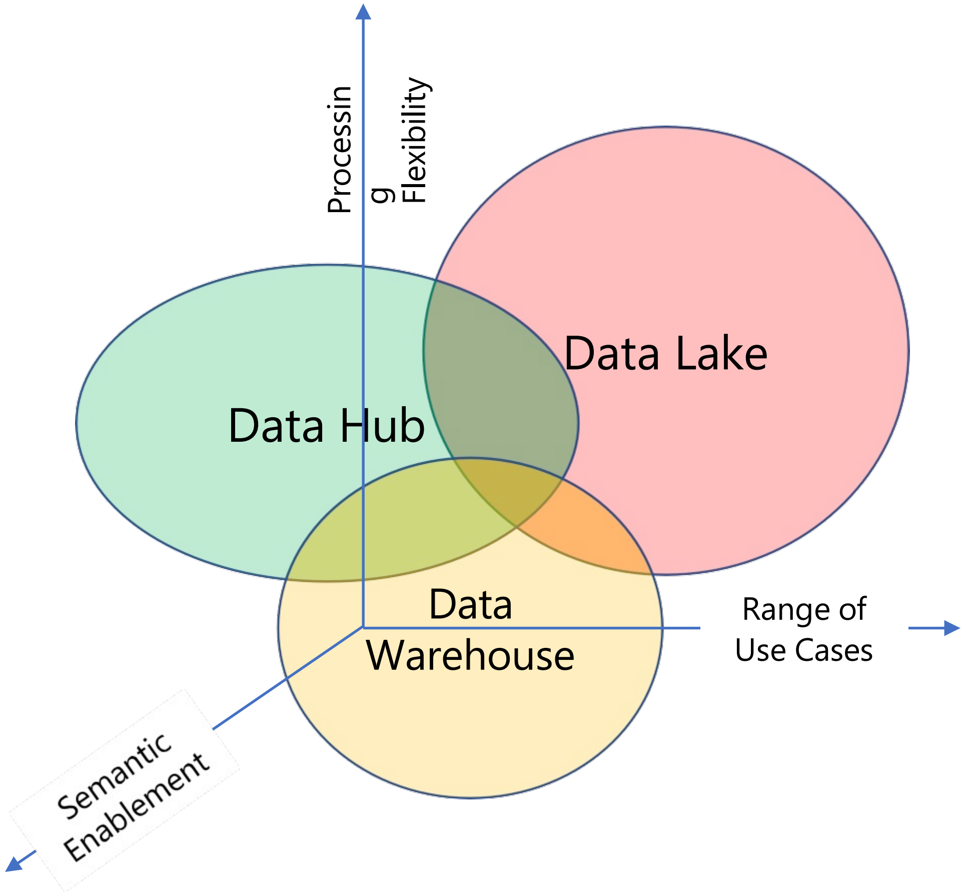
Sales Playbook

Data Modernization

Data Strategy

Digital Data Platform in Data Strategy

Data Warehouse, Data Lake & Data Hub is coexistent in Digital Data Platform, which support diversified requirement of Range of Use Cases, Processing Flexibility and Semantic Enablement.



Requirement to Consider

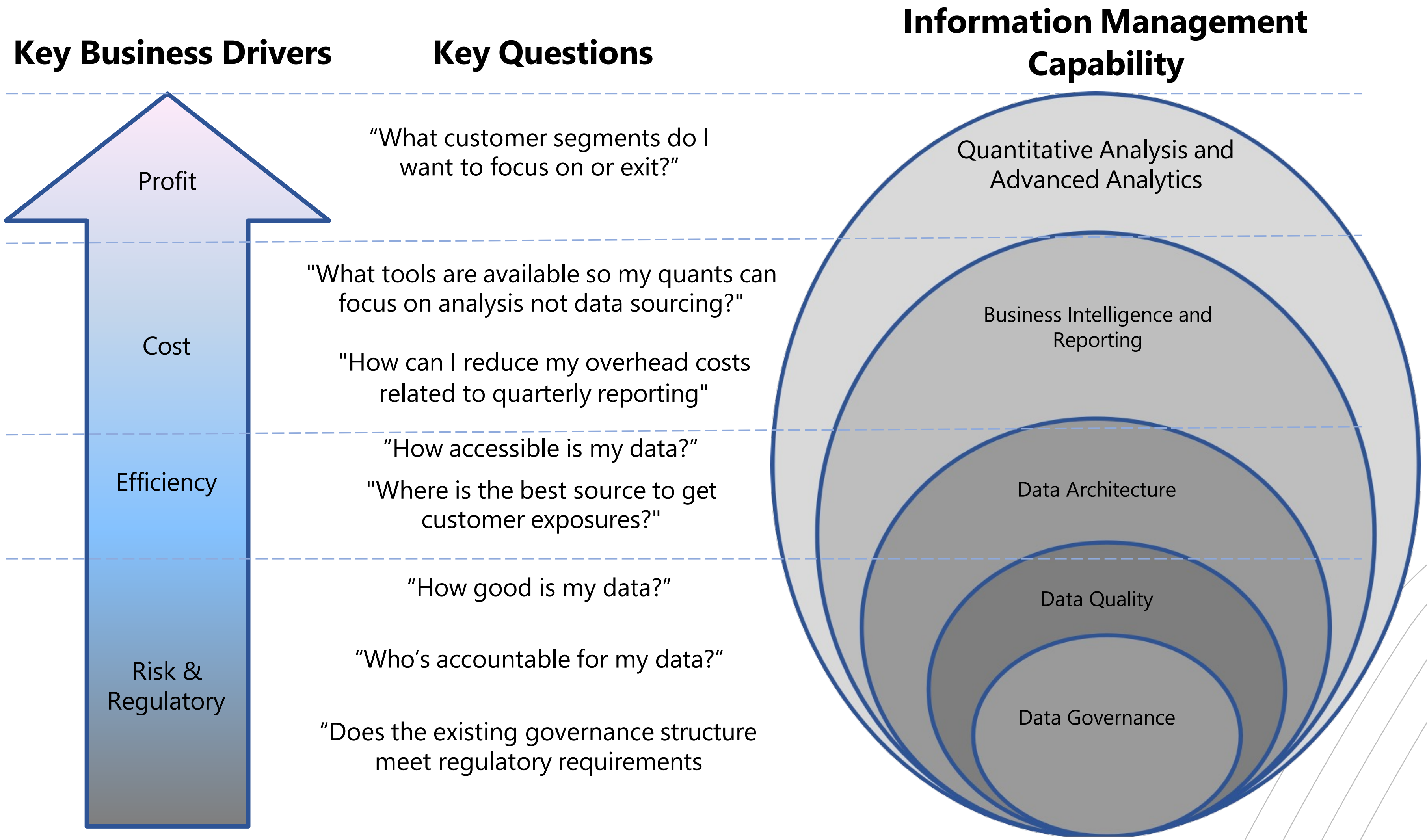
- ***Processing Flexibility***: Rigid (fixed and high optimized) vs. Flexible (diversified options but less optimized)
- ***Semantic Enablement***: Consistent (stable and reliable) vs. Variant (context-specific)
- ***Range of Use Cases***: Specific (targeted use cases and high optimized) vs. Generic (user-driven context but less optimized)

Patterns/Components to Select

- ***Data Warehouse***: supporting mostly known data and known questions to deliver consensus for running business.
- ***Data Lake***: supporting unknown data and unknown questions to enable exploration and innovation.
- ***Data Hub***: enabling manageable and governed sharing of data between producing and consuming systems and process.

Understand Drivers in Data-Driven Business

Key take-away: Representative business questions often help illustrate how investment in information capabilities support key business drivers



Key Business Drivers	Key Questions	Information Management Capability
Profit	“What customer segments do I want to focus on or exit?”	Quantitative Analysis and Advanced Analytics
Cost	"What tools are available so my quants can focus on analysis not data sourcing?" "How can I reduce my overhead costs related to quarterly reporting"	Business Intelligence and Reporting
Efficiency	“How accessible is my data?” "Where is the best source to get customer exposures?"	Data Architecture
Risk & Regulatory	“How good is my data?” “Who’s accountable for my data?” “Does the existing governance structure meet regulatory requirements	Data Quality Data Governance

Key Success Factors of Digital Data Platform

By understanding the requirement and having strong experience to build up Digital Data Platform with many customers, FPT is aware that the most challenges of developing Digital Platform are

- Having **strong and extendable architecture** that support to **flexibly develop large number of use cases**.
- But still be **easy and flexible to start with small number of components** and **focus on immediate success of the first use cases**.

The below key factors to ensure to delivery successfully Digital Data Platform solution & project.

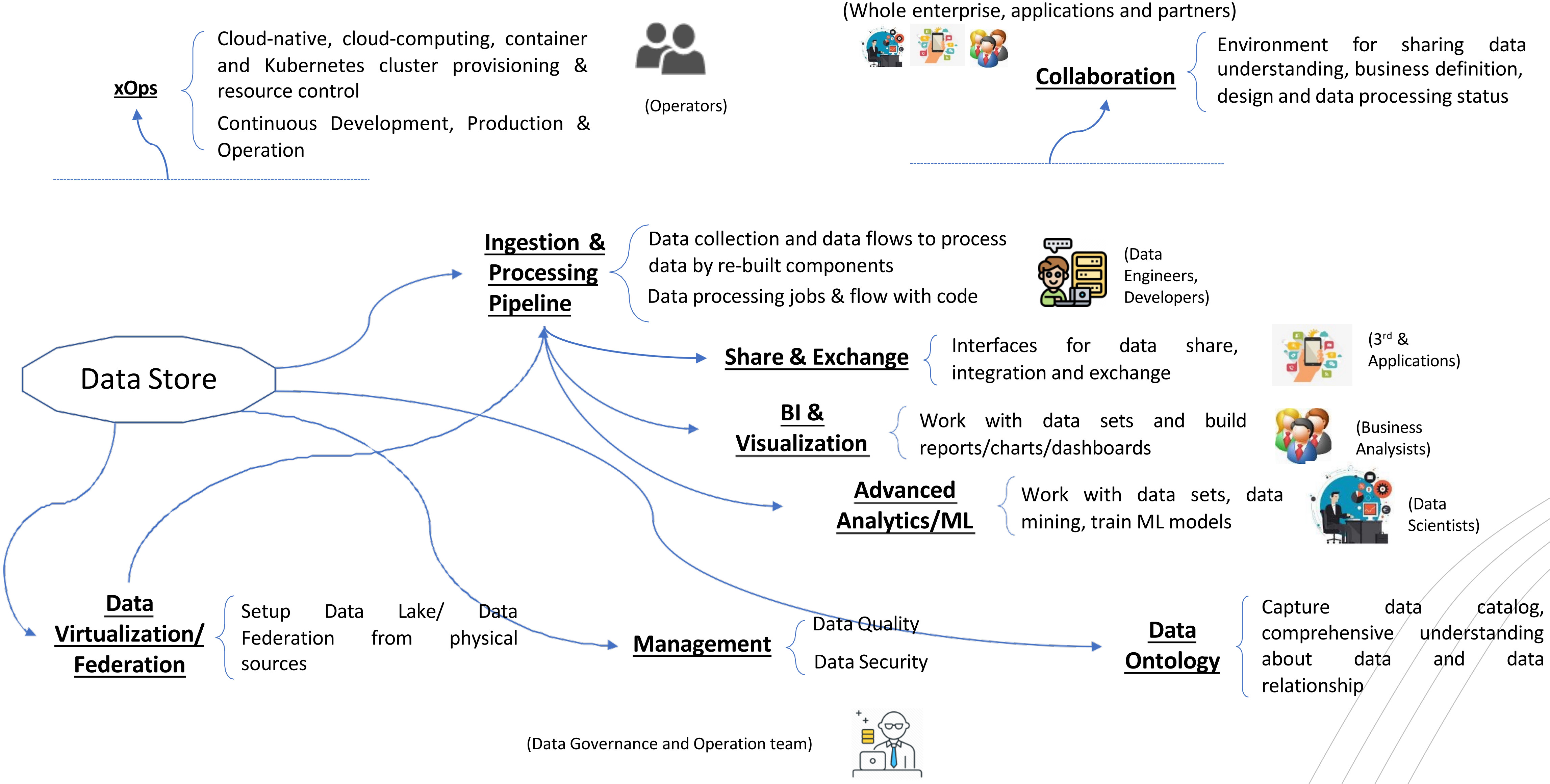
Solution & Technology Aspect

- The **reference architecture** which **combines benefit of Data Warehouse, Data Lake and Data Hub** patterns.
- **Loose-coupling** design which help to easily deploy and extend.
- High **security and compliance**.
- Leverage of cloud services for data analytics and big data: Azure Synapse, Azure Databricks, Azure Analysis Services, Azure Streaming Analytics, Data Factory, Azure ML, Azure Cognitive Services

Project Execution Aspect

- **Well-architecture for long-time** but **focus on small and doable scope** to make it done in short period.
- Lightweight and agile process to implement use case on top of Digital Data Platform with **use case discovery and implementation model**

Digital Data Platform – Implementation & Usage View



xOps

- Cloud-native, cloud-computing, container and Kubernetes cluster provisioning & resource control
- Continuous Development, Production & Operation

(Operators)

(Whole enterprise, applications and partners)

Collaboration

- Environment for sharing data understanding, business definition, design and data processing status

Data Store

Ingestion & Processing Pipeline

- Data collection and data flows to process data by re-built components
- Data processing jobs & flow with code

(Data Engineers, Developers)

Share & Exchange

- Interfaces for data share, integration and exchange

(3rd & Applications)

BI & Visualization

- Work with data sets and build reports/charts/dashboards

(Business Analysists)

Advanced Analytics/ML

- Work with data sets, data mining, train ML models

(Data Scientists)

Data Virtualization/ Federation

- Setup Data Lake/ Data Federation from physical sources

Management

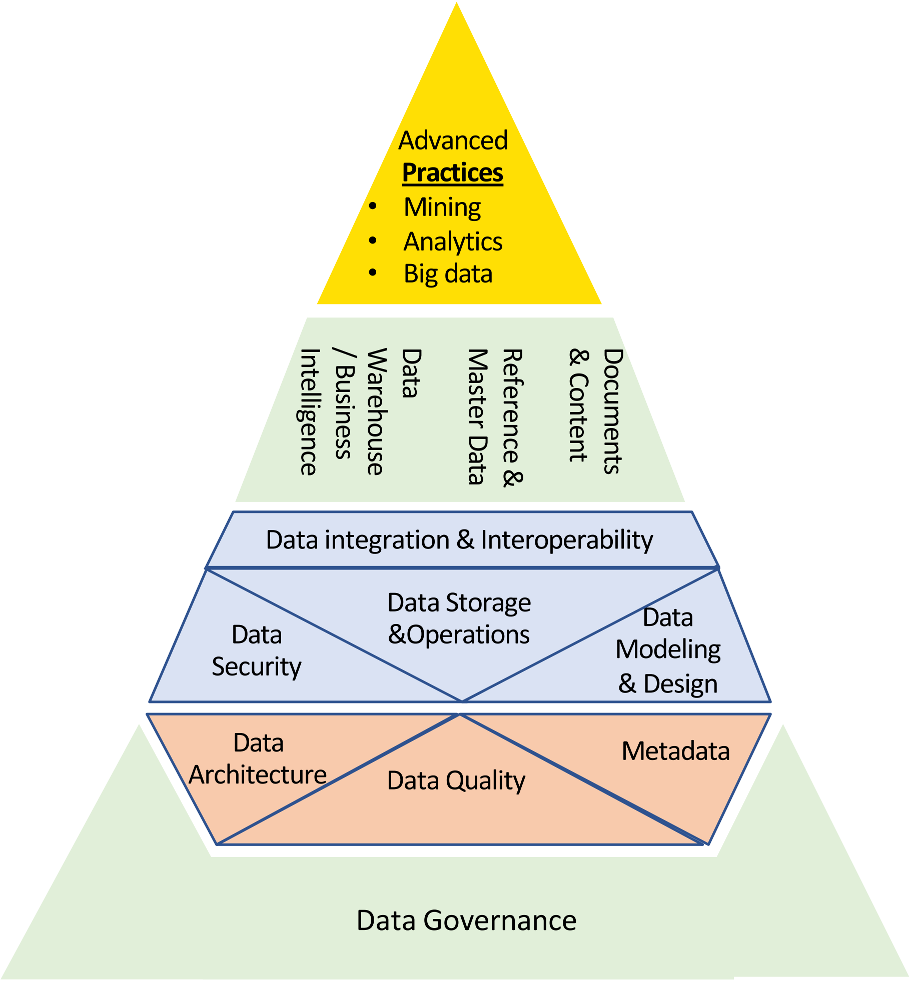
- Data Quality
- Data Security

Data Ontology

- Capture data catalog, comprehensive understanding about data and data relationship

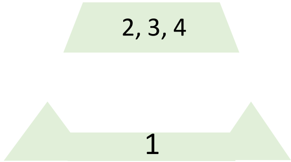
(Data Governance and Operation team)

The Golden Pyramid – Maturity Phases



1. **Data Governance:**
provides direction and oversight

2. **Reference & Master Data:**
include ongoing reconciliation

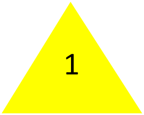


3. **Document and Content Management**
planning, implementation and control activities to manage lifecycle of data found in a range of unstructured media.

4. **Data Warehouse & Business Intelligence**
planning, implementation and control activities to manage decision support data

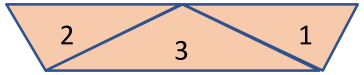
Phase 3

1. **Advance Practices:**
Mining, Analytics, Big data,



Phase 4

1. **Metadata**
HQ integrated metadata

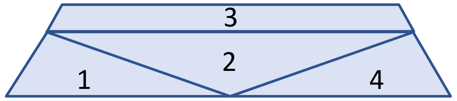


2. **Data Architecture**
Define blueprint for managing data assets

3. **Data Quality**
measure and improve fitness of data

Phase 2

1. **Data Security**
ensures that data privacy and confidentiality are maintained, that data is not breached , and that data is accessed appropriately.



2. **Data storage & Operations**
includes the design, implementation, and support of stored data to maximize its value. Operations provide support throughout the data lifecycle from planning to disposal.

3. **Data integration & Interoperability**
includes processes related to the movement and consolidation of data within and between data stores, applications, and organization.

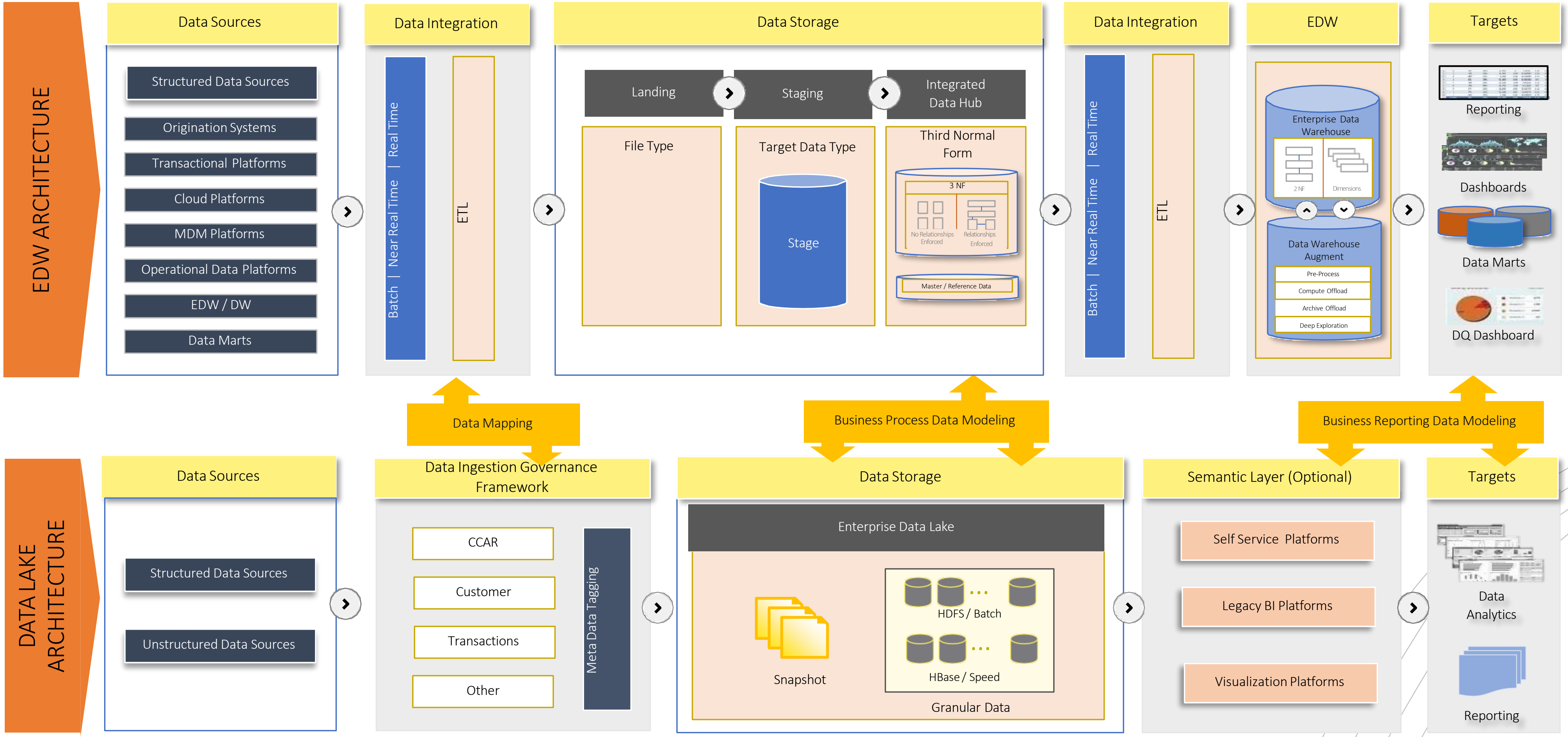
4. **Data Modeling & Design**
is the process of discovering, analyzing, representing, and communicating data requirements in a precise form called the data model.

Phase 1

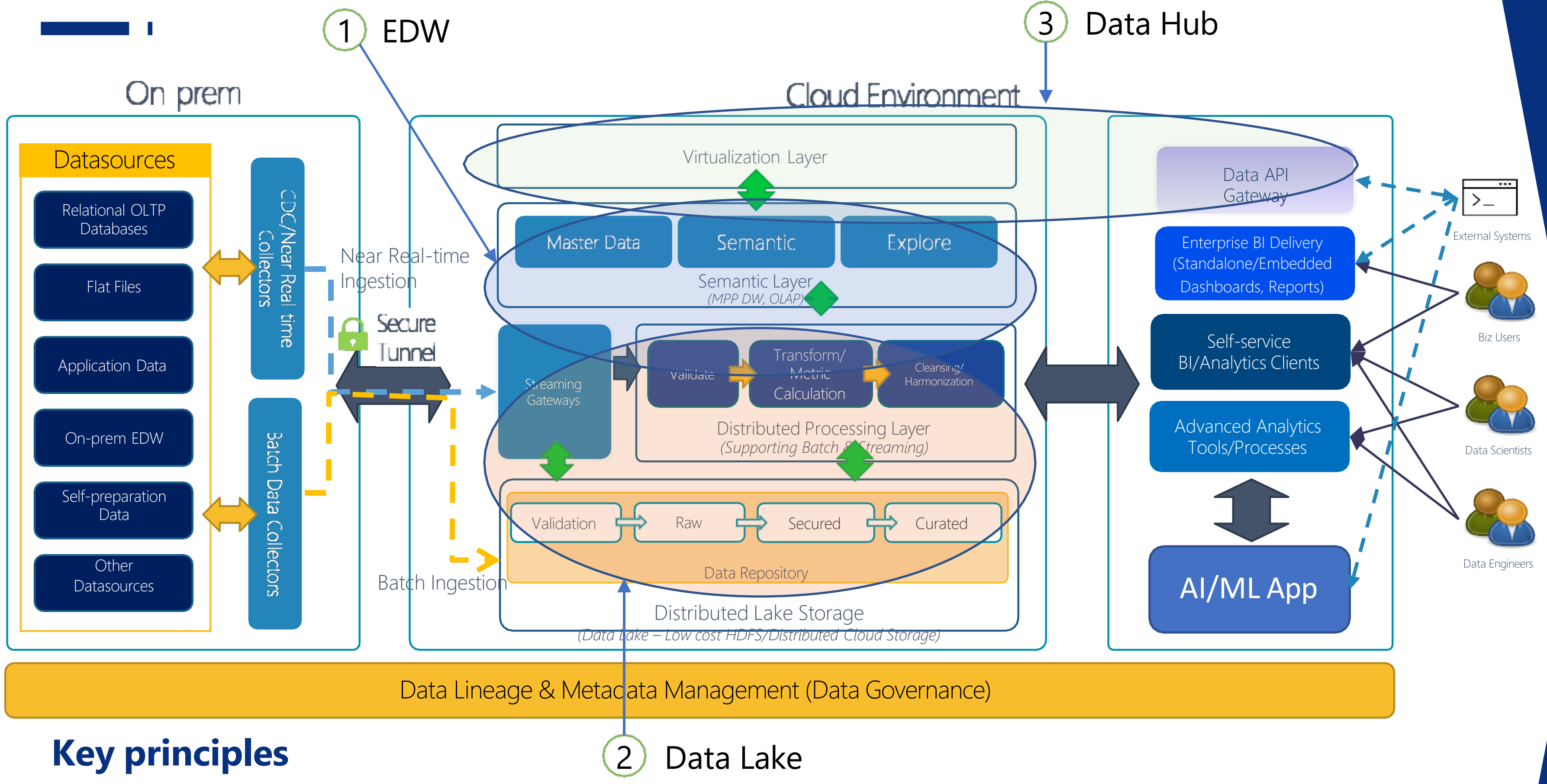
Data Modernization

Architecture

Data Platform – General Architectures



Data Modernization - Design Principles



Key principles

1 Data Warehouse + data Lake + Data Hub: don’t use one to replace other

2 Loose-coupling: each component can work independently

3 Independent data: data becomes independent from underlying systems

4 Open-end: standardize for data storage but open for data analyzing and consuming

Key Requirements

1 Support structured & unstructured data sources

2 Support batch and streaming (near-real time) data integration

3 Secured hybrid Cloud infrastructure

4 Scalable, cost-effective and well-managed data repository

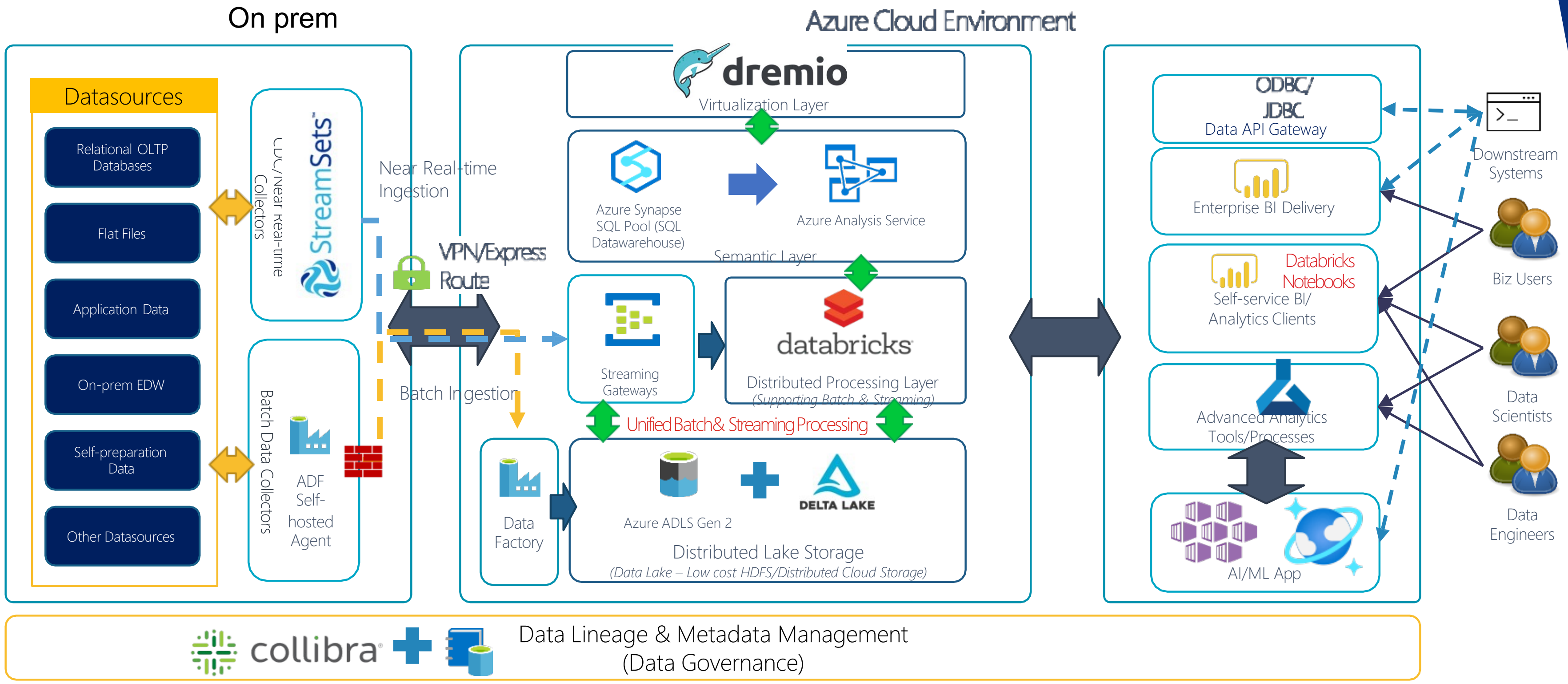
5 Support big data processing & AI

6 MPP & In memory Data Warehouse

7 Self-services BI & downstream integration

8 Unified operations

Data Modernization – Sample Azure Desig

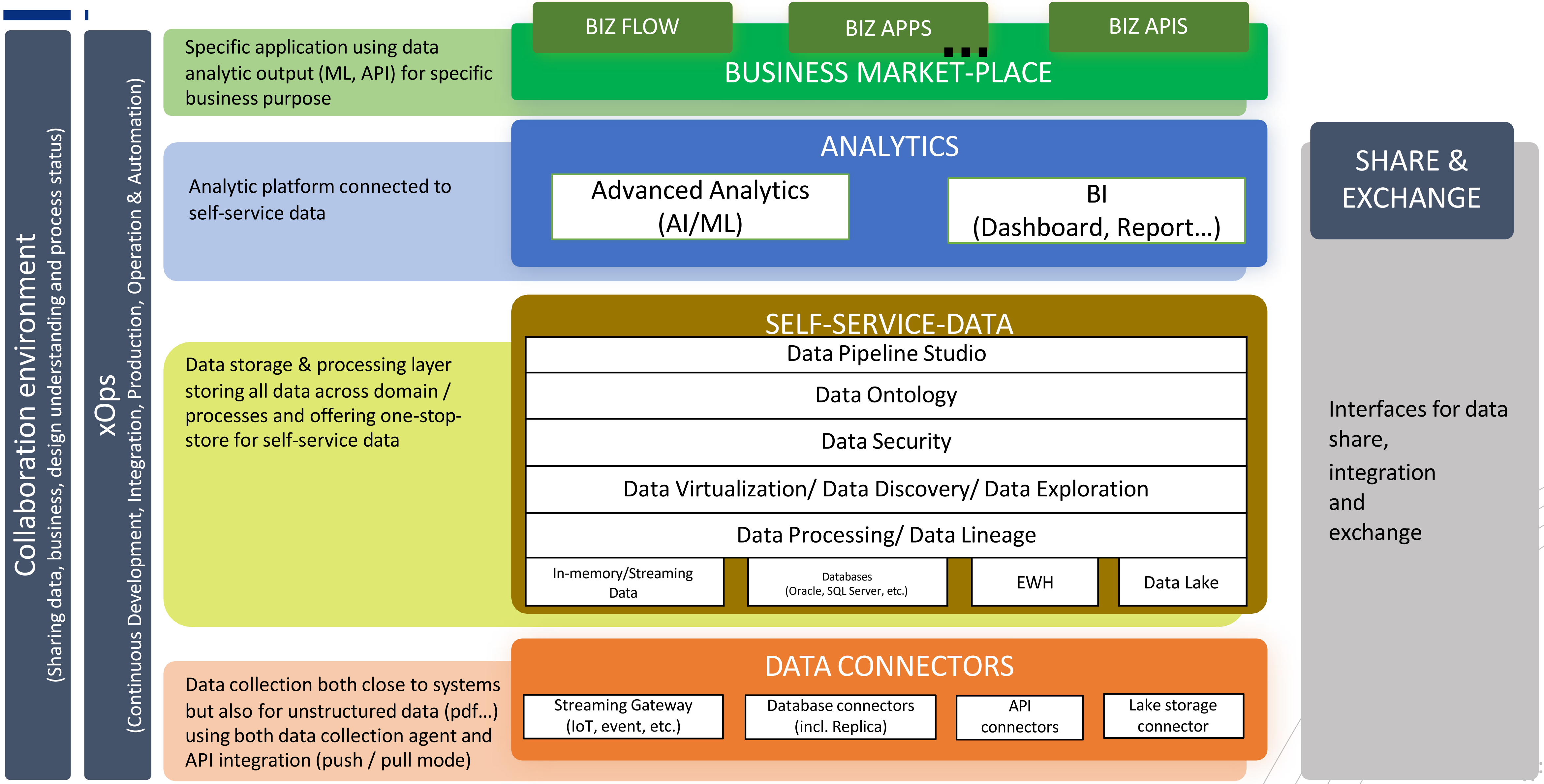


- All components of Digital Data Platform will be deployed in Azure by leveraging Azure native services
- StreamSets for data streaming and data capture change requirement
- Dremio for data sematic layer & data virtualization layer
- PowerBI for data visualization and Self service BI analytics (PowerBI can be replaced by Tableau)
- Delta Lake for incremental processing to optimize big data analytics
- Graph DB & Kubernetes related services for AI/ML related use cases

FPT already applied this architecture to many customers *. We also used it to build our own Digital Data Platform to server FPT Corp. and all subsidiaries.

** Please see Appendix for more detailed reference projects*

Data Platform Functional Stack – a Digital Platform



Collaboration environment

(Sharing data, business, design understanding and process status)

xOps

(Continuous Development, Integration, Production, Operation & Automation)

Specific application using data analytic output (ML, API) for specific business purpose

BIZ FLOW

BIZ APPS

BIZ APIS

…

BUSINESS MARKET-PLACE

Analytic platform connected to self-service data

ANALYTICS

Advanced Analytics (AI/ML)

BI (Dashboard, Report…)

Data storage & processing layer storing all data across domain / processes and offering one-stop-store for self-service data

SELF-SERVICE-DATA

Data Pipeline Studio

Data Ontology

Data Security

Data Virtualization/ Data Discovery/ Data Exploration

Data Processing/ Data Lineage

In-memory/Streaming Data

Databases (Oracle, SQL Server, etc.)

EWH

Data Lake

Data collection both close to systems but also for unstructured data (pdf…) using both data collection agent and API integration (push / pull mode)

DATA CONNECTORS

Streaming Gateway (IoT, event, etc.)

Database connectors (incl. Replica)

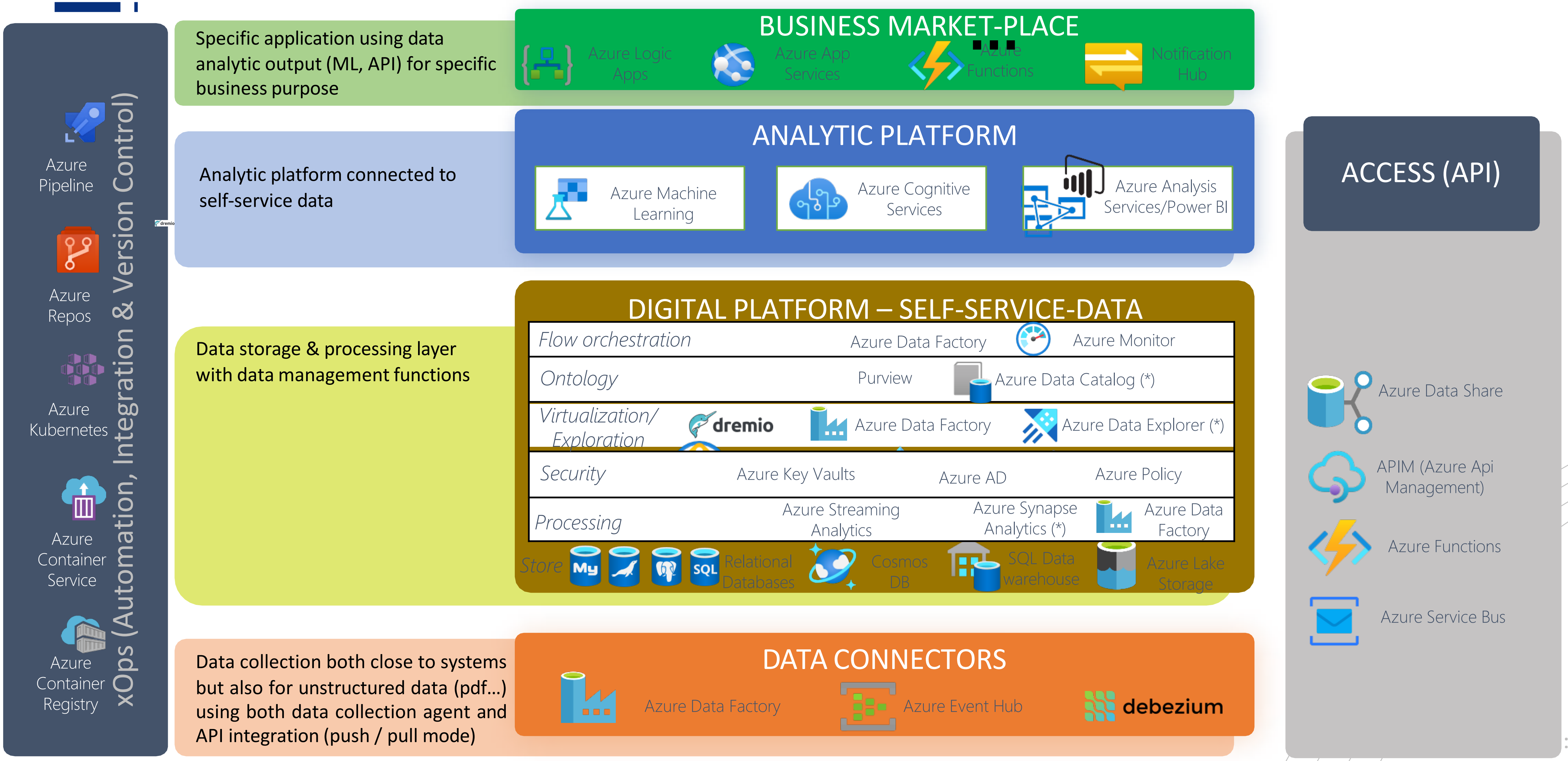
API connectors

Lake storage connector

SHARE & EXCHANGE

Interfaces for data share, integration and exchange

Data Modernization – Azure Implementation



xOps (Automation, Integration & Version Control)

Azure Pipeline

Azure Repos

Azure Kubernetes

Azure Container Service

Azure Container Registry

Specific application using data analytic output (ML, API) for specific business purpose

BUSINESS MARKET-PLACE

Azure Logic Apps

Azure App Services

Azure Functions

Notification Hub

Analytic platform connected to self-service data

ANALYTIC PLATFORM

Azure Machine Learning

Azure Cognitive Services

Azure Analysis Services/Power BI

Data storage & processing layer with data management functions

DIGITAL PLATFORM – SELF-SERVICE-DATA

Flow orchestration	Azure Data Factory	Azure Monitor	
Ontology	Purview	Azure Data Catalog (*)	
Virtualization/ Exploration	dremio	Azure Data Factory	Azure Data Explorer (*)
Security	Azure Key Vaults	Azure AD	Azure Policy
Processing	Azure Streaming Analytics	Azure Synapse Analytics (*)	Azure Data Factory
Store	Relational Databases	Cosmos DB	SQL Data warehouse Azure Lake Storage

Data collection both close to systems but also for unstructured data (pdf…) using both data collection agent and API integration (push / pull mode)

DATA CONNECTORS

Azure Data Factory

Azure Event Hub

debezium

ACCESS (API)

Azure Data Share

APIM (Azure Api Management)

Azure Functions

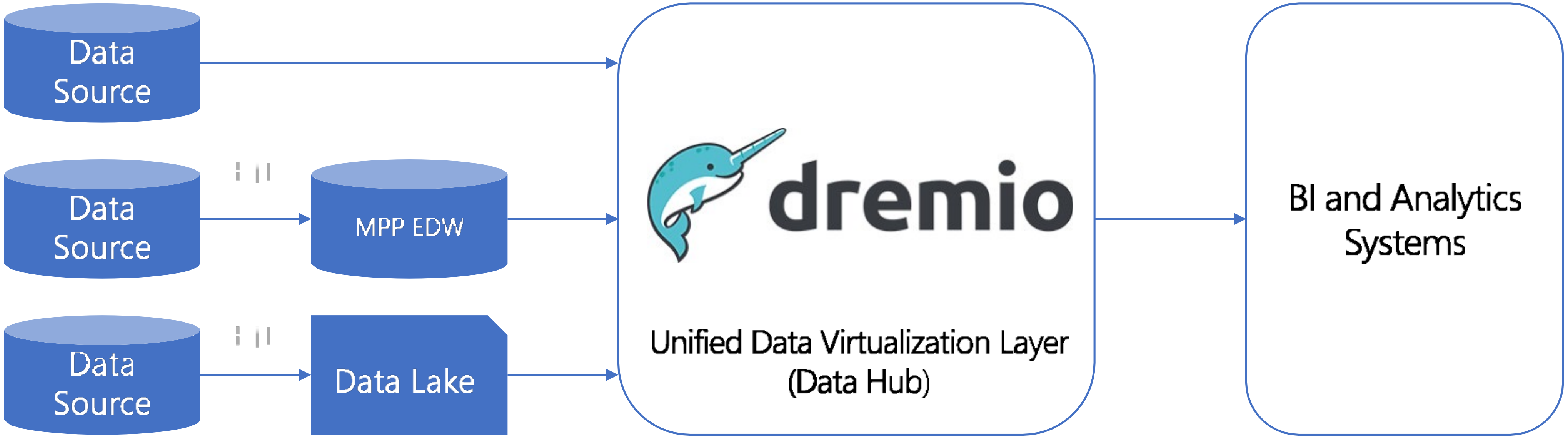
Azure Service Bus

Data Modernization

Technical Topics

Data Virtualization

A sample of leveraging Dremio to unify data virtualization for various data sources, even for sources not in Data Lake or EDW.



Well integration with

- MPP EDW including Azure Synapse SQL Pool, AWS Redshift
- Data Lake including Azure Data Lake Storage, AWS S3
- Other on premises and cloud data sources

Lightning-fast engine for

- Accelerating data query, BI queries: up to 1000x faster than SQL engine
- Eliminating the need for data cubes

Well security support: Masking, Row Level Security, AD integration

Accelerate time to implement analytics use case

- Directly connect to data source (even existing data warehouse) to utilize existing ETLs, cubes, pre calculated data to build analytics use cases before we migrate them to a central Digital Data Platform.
- Logically combine data from many sources to build analytics use cases without developing complex ETLs.

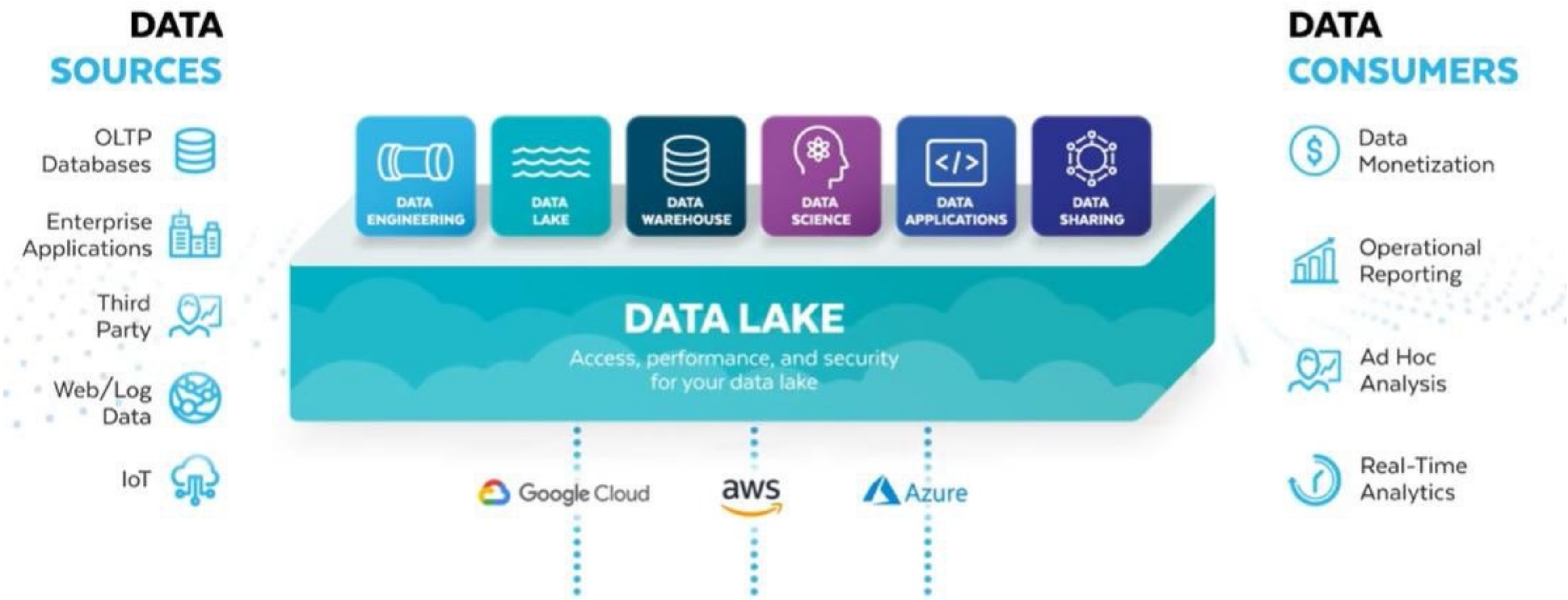
Building Data Lake – Snowflake example

Building data lake with a data storage and a semantic layer :

For example, with Snowflake using external table referring to data store on Cloud or Hadoop to build data lake:

- Apache Hive Metastores
- Data files located in a cloud storage (Amazon S3, Google Cloud Storage, or Microsoft Azure)

Design and implement a single platform with unified technology landscape for many types of data workloads

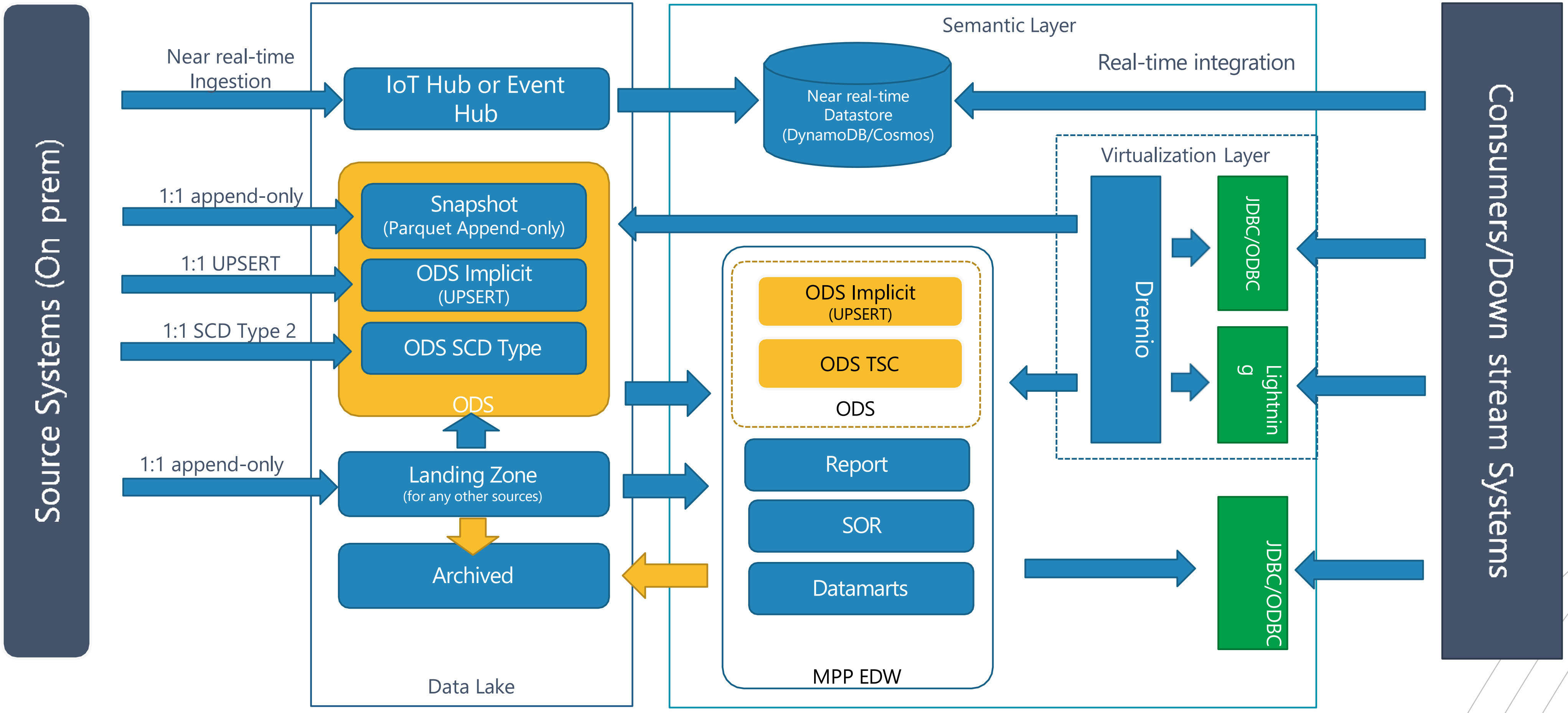


Others choices of combined facilities, e.g. using suitable data type (e.g. Snowflake's single VARIANT column) for semi-structured data, supporting JSON, XML, ORC, Parquet and Avro:

- ORC (Optimized Row Columnar) binary file format for efficient compression and improved performance storing Hive data
- Parquet is a compressed, efficient columnar data representation designed for projects in the Hadoop ecosystem
- Avro is a data serialization, Avro schema consists of a JSON string, object, or array

Data Storage on Cloud

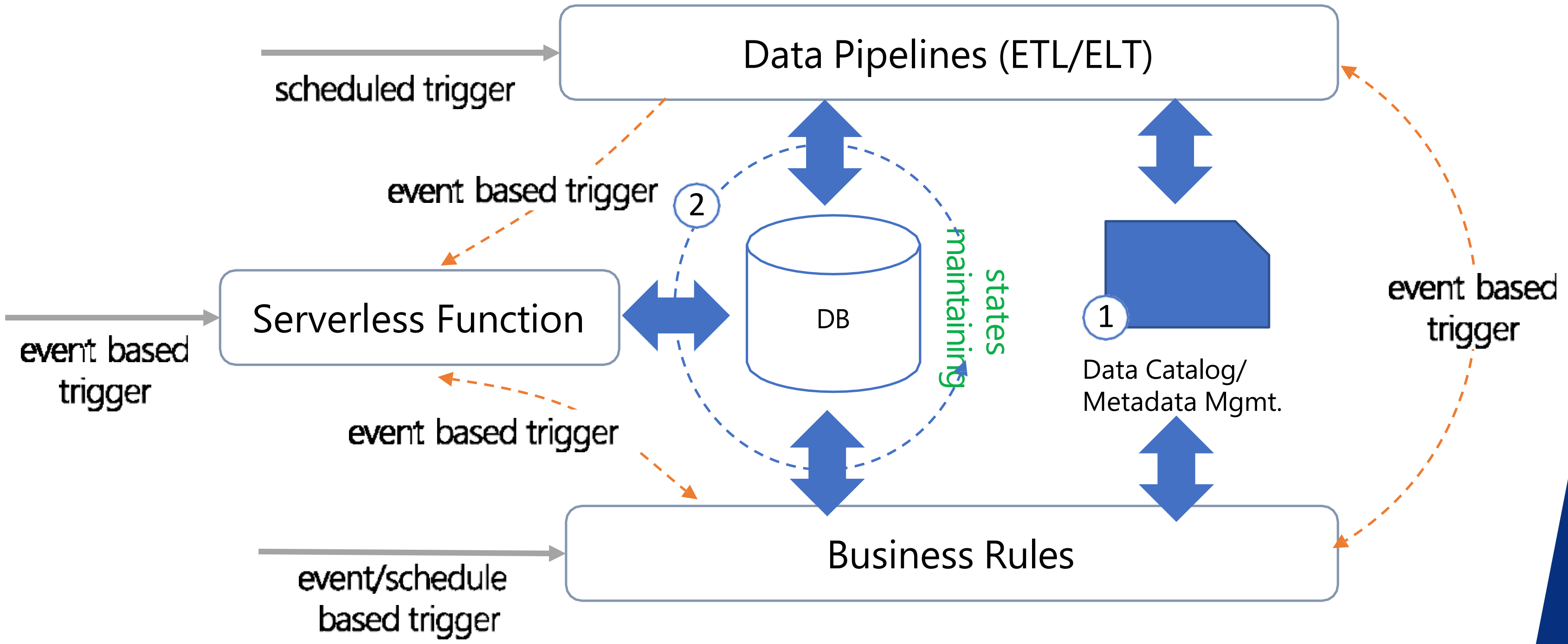
Combine capacity of Data Lake and MPP DW help to create flexible and cost-effective Data Storage on Cloud.



* ODS: Operational Data Store designed to support high volumes of simple queries on subset of data to feed operational processes, dashboard or APIs. In traditional architecture ODS might be a part of EDW, but in Digital Data Platform to develop ODS we leverage feature from Data Lake and EDW.

Building Data Pipelines

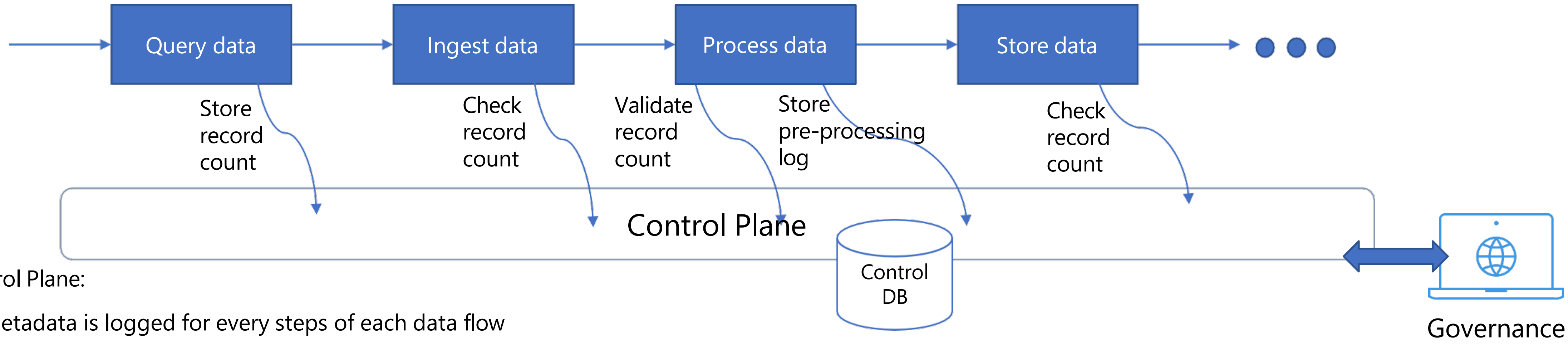
The Digital Platform is built as a complex mixing of services and the network topology is spanning between on-premise and cloud environments. So that, we need to consider flow orchestration with triggers/coordinators.



1. Manage data dictionary, metadata with Data Catalog/ Metadata Mgmt.
2. Custom control framework (built with Serverless Functions, Business Rules and SQL Database) to control data flow and data quality. Activities log for every step is key principle to keep track and analyze problem when it happens.

Data Management along Data Pipelines

Considering about data management over data store and data pipeline with a control plane



Control Plane:

- Metadata is logged for every steps of each data flow
- Operation and governance activities can be done, e.g.:
  - Monitoring: e.g. notify to related stakeholder for data not being able to process or violated access
  - Exceptional data is stored for analysis then recovered/corrected for preventing data loss
  - Controlling: e.g. Stop, continue or re-run the data jobs
  - Validating, de-duplicating… for data quality
  - Adding tags, indexes for searching
  - Classifying data to build data catalogue
  - Adding additional data for lineage management
  - Tokenizing, encrypting for data privacy and security

-> Data quality, lineage, security and compliance are managed for each step

Sample additional generated data

Pipeline_id	Source_id	Step_id	Status	No_of_record	No_of_invalid_record	Evaludation_Log_File	…
000…0001	PI System	0001	OK	12345	0		
000…0001	PI System	0002	OK	12345	0		
000…0001	PI System	0003	NOK	12340	200	ADLS_Log_00…0001_Log_...	
…	…						

Data Integration to Cloud – Snowflake example

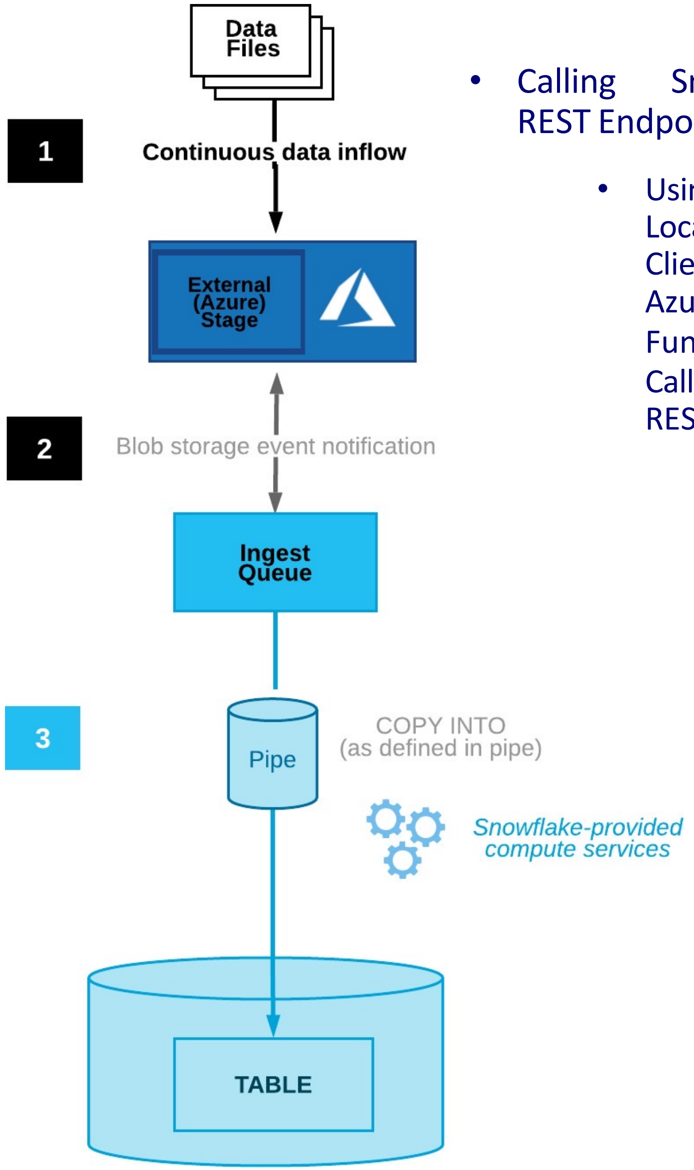
Leveraging 3rd party ETL, DWH tool and Cloud-based ETL: For example, Snowpipe with cloud service (e.g. Azure) for automation and flow orchestration as below:

- Automating Snowpipe Using Cloud Messaging:
  - Configuring Secure Access to Cloud Storage
  - Configuring Automated Snowpipe Using Azure Event Grid

❖ Notification Integration

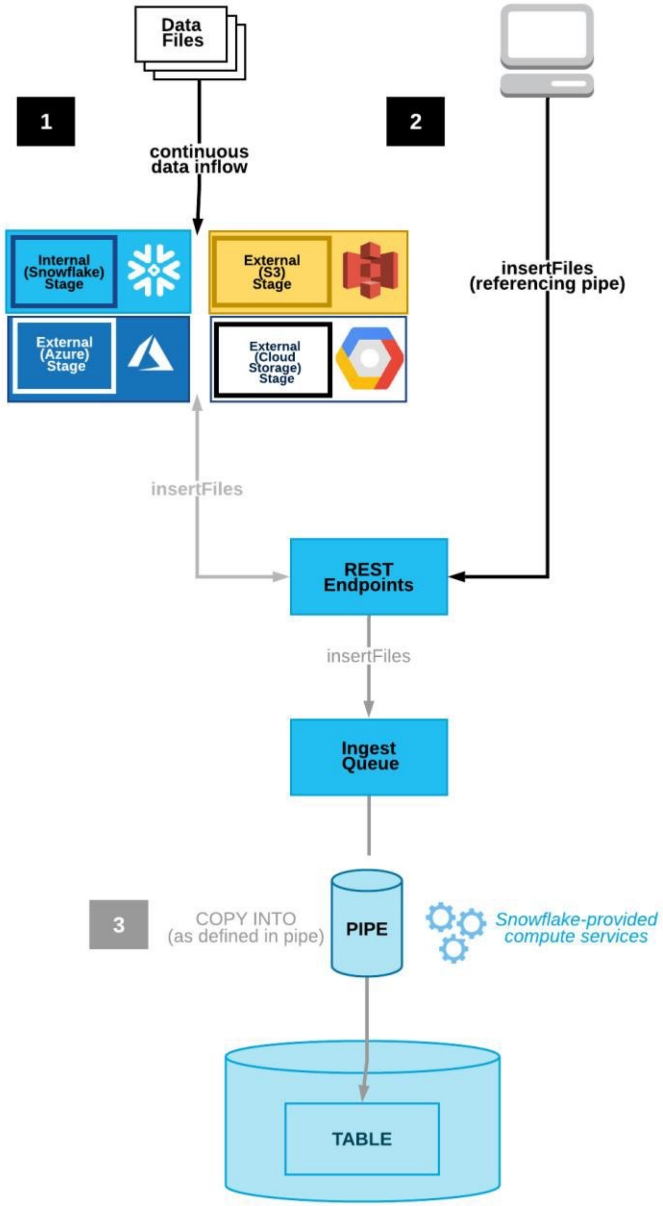
❖ Authentication & Workflow Ingest Enabled

❖ Load Historical Files



Data Files
1
Continuous data inflow
External (Azure) Stage
2
Blob storage event notification
Ingest Queue
3
Pipe
COPY INTO (as defined in pipe)
Snowflake-provided compute services
TABLE

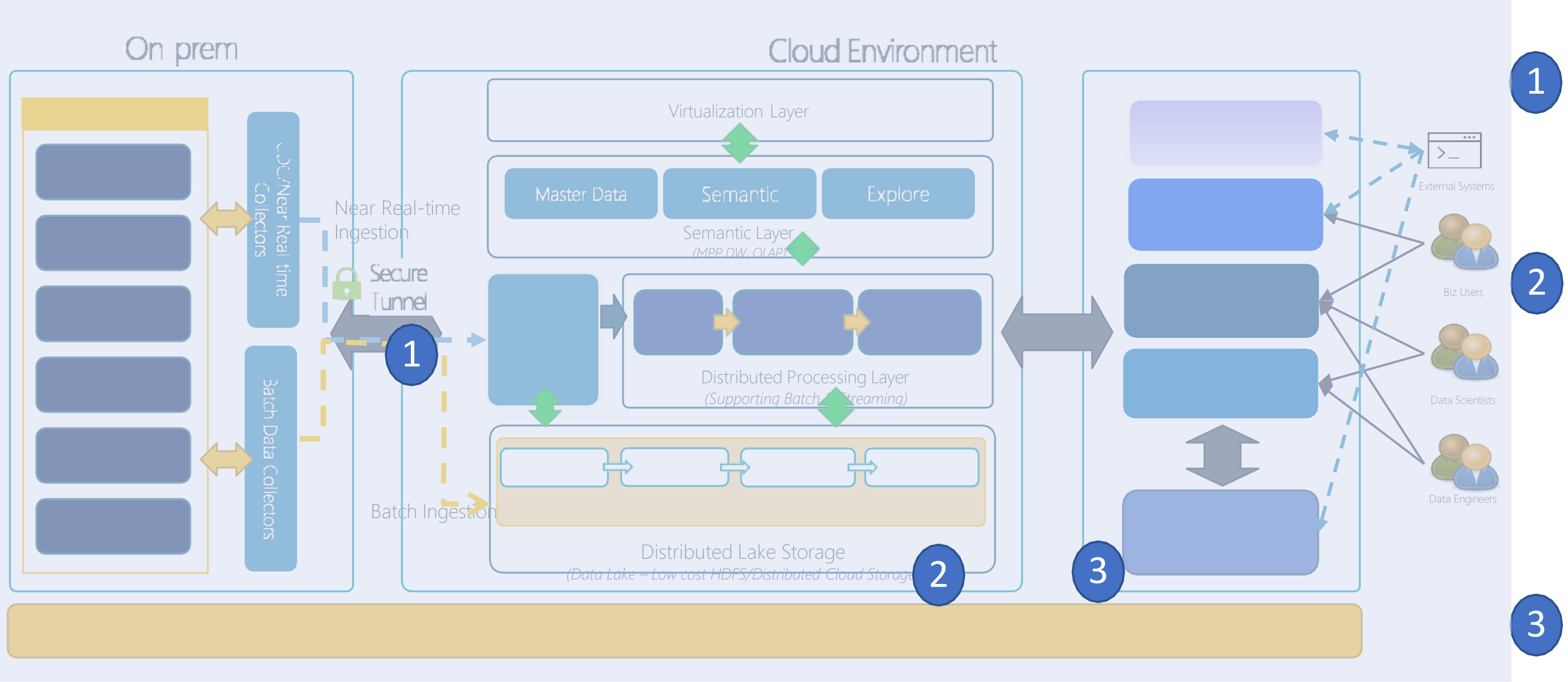
- Calling Snowpipe REST Endpoints:
  - Using a Local Client/ Azure Function to Call the REST API



Data Files
1
continuous data inflow
2
Internal (Snowflake) Stage
External (S3) Stage
External (Azure) Stage
External (Cloud Storage) Stage
insertFiles (referencing pipe)
insertFiles
REST Endpoints
insertFiles
Ingest Queue
3
COPY INTO (as defined in pipe)
PIPE
Snowflake-provided compute services
TABLE

Network Integration Considerations

Digital data platform requires moderate network connection because of huge data transferring via internet. Dedicated network (ExpressRoute/ DirectConnect) might need to consider adapting with growth of data size.



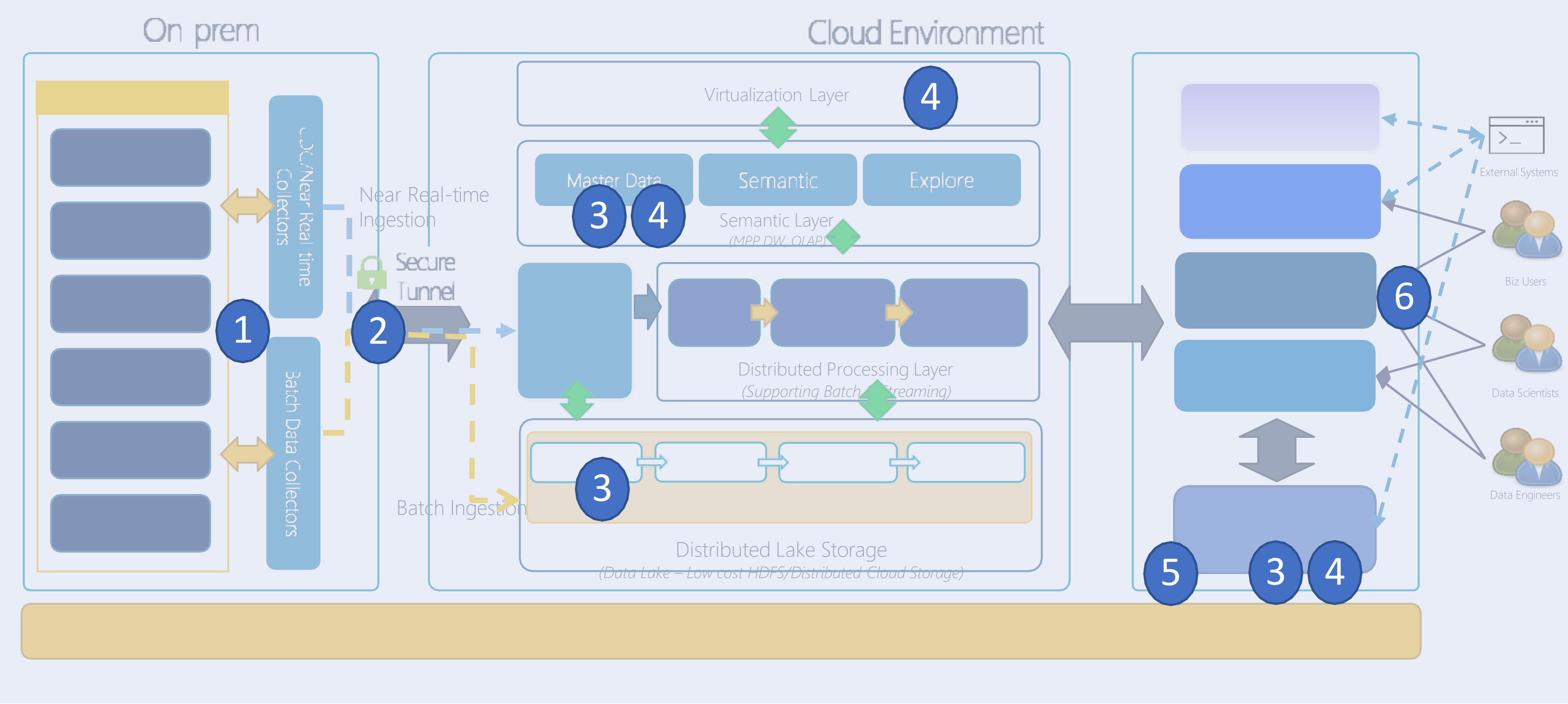
1. IPSec VPN connection between on-premise and Cloud

2. Separated Virtual Network for Digital Data Platform, all services will be deployed in this Virtual Network (no direct internet connection here). This environment uses for data analytics, data processing, data scientist, data engineering only

3. Separated Virtual Network for services and applications. Peer to Peer VPN will be used to connect to Digital Data Platform

Data & Security Compliance

Compliance (ISO27001, GDPR, etc.) are top requirement need to be handled in data analytics and recommendation system. All encryption algorithms must be compliant with FIPS 140-2 (Symmetric Key: AES, Triple-DES; Asymmetric Key: DSA, RSA, ECDSA; Hash Standards: SHA-1, SHA-256, SHA-512)



1 Any PII data must be tokenized

2 Data must be encrypted in transit (SSL/ TLS), encrypted at rest (no plain text) before sending out data center (to internet)

3 Data must be encrypted at rest in any kind of storage

4 Any PII data must be tokenize/ masking, data storage is encrypted, might support row-level, column-level encryption

5 Service must support to encrypt sensitive data in application layer

6 Services/Apps must be secured when integrate with other system

Compliance Standards

Ensure data compliance is the top priority in development and operation of data platform. Implementing automation engine to scan, notify and fix compliance issues is the most relevant and efficient approach to simplify how to deal with security compliance. Compliance check report should be visible and transparent to relevant stakeholders.

Regarding to data compliance

One of the newest and most-wide-ranging standards, it's been hard to ignore the European Union's General Data Protection Regulation (GDPR) over the last year. Coming into force on May 25th 2018.

HIPAA, or more formally the Health Insurance Portability and Accountability Act of 1996, sets out how US organizations that deal with individuals' healthcare and medical data need to ensure the safety and confidentiality of these records.

For businesses dealing with customers' financial information, the Payment Card Industry Data Security Standard (PCI DSS) is a vital part of any compliance process, as it sets out rules regarding how companies handle and protect cardholder data such as credit card numbers.

The Personal Information Protection and Electronic Documents Act (PIPEDA) is a Canadian law relating to data privacy.

The California Consumer Privacy Act (CCPA) is a state statute intended to enhance privacy rights and consumer protection for residents of California, United States

Regarding to security compliance

ISO/IEC 27001 is an international standard on how to manage information security, riginally published jointly by the International Organization for Standardization (ISO) and the International Electro-technical Commission (IEC) in 2005 and then revised in 2013.

The NIST Cybersecurity Framework provides a policy framework of computer security guidance for how private sector organizations in the United States can assess and improve their ability to prevent, detect, and respond to cyber attacks.

The Center for Internet Security (CIS) is a nonprofit organization, formed in October, 2000. Its mission is to "identify, develop, validate, promote, and sustain best practice solutions for cyber defense and build and lead communities to enable an environment of trust in cyberspace".

The HITRUST CSF (Common Security Framework) is a prescriptive set of controls that meet the requirements of multiple regulations and standards. The framework provides a way to comply with standards such as ISO/IEC 27000-series and HIPAA.

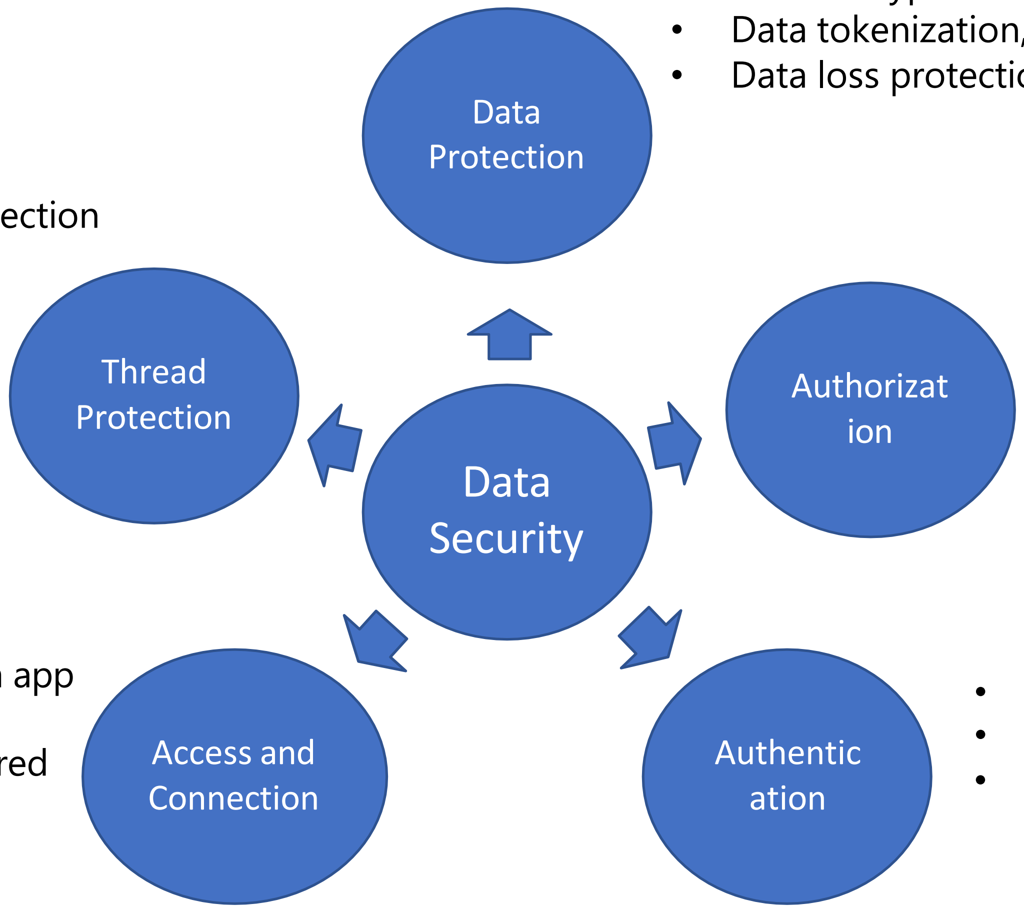
SOC 2 (Systems and Organizations Controls 2) is both an audit procedure and criteria. It’s geared for technology-based companies and third-party service providers which store customers’ data in the cloud.

Cloud Security Alliance (CSA) is a not-for-profit organization with the mission to “promote the use of best practices for providing security assurance within cloud computing, and to provide education on the uses of cloud computing to help secure all other forms of computing.

Example: FPT’s AkaClaud security compliance check

Compliance	Policies	Coverage
NIST	151	100% for Azure & AWS
CSA	136	
SOC 2	141	
CCPA	119	
GDPR	111	
PCI DSS	95	
ISO27001	141	
PIPEDA	294	
CIS	141	
HIPAA	82	
HITRUST	130	

Data Security Considerations



- Data encryption in transit
- Data encryption at rest
- Data tokenization, masking
- Data loss protection

- SQL thread protection
- Data auditing
- Vulnerability

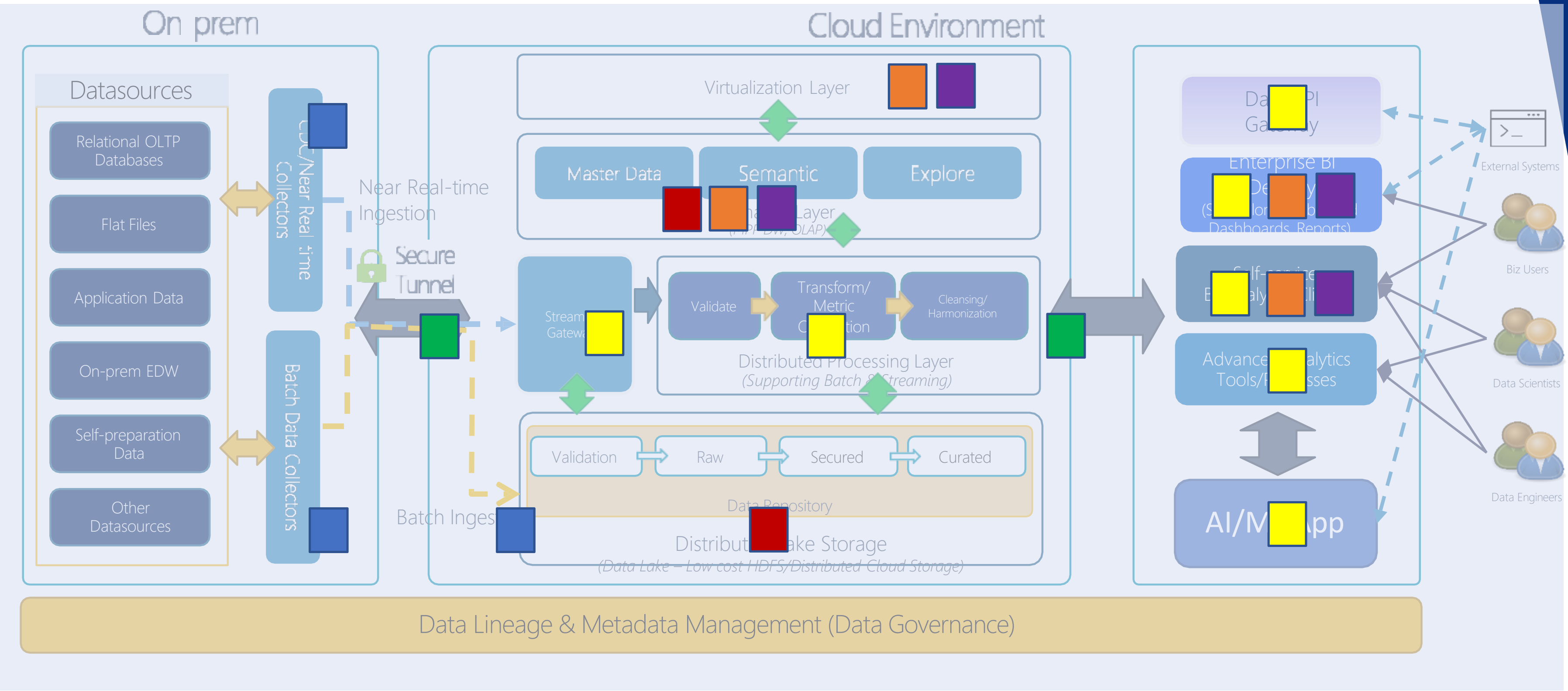
- Row level security
- Column level security
- Schema, table & view security
- RBAC access policy

- Database & data app firewall
- Private end secured network

- SQL authentication
- AD/ADFS integration
- Multi-factor authentication

Comprehensive considerations to ensure data security

Data Security Design



🟩 SSL/TSL: network security

🟥 TDE: transparent data encryption

🟪 Row level Encryption

🟧 Column level Encryption

🟨 RBAC: role based access control

🟦 Tokenization: encrypt sensitive data to align with security compliance

Beside applying security best-practices, we also need to enable common security components

1 **Key Vault**: centralized solution to store and manage key

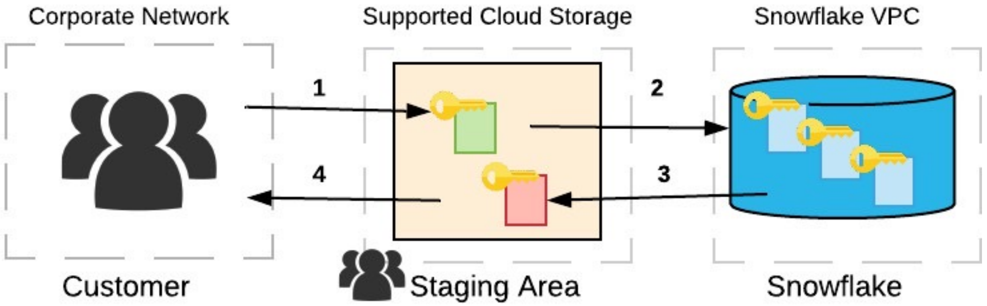
2 **Active Directory**: single sign-on and multi-factor authentication to protect users

3 **Security Center**: unified security management systems

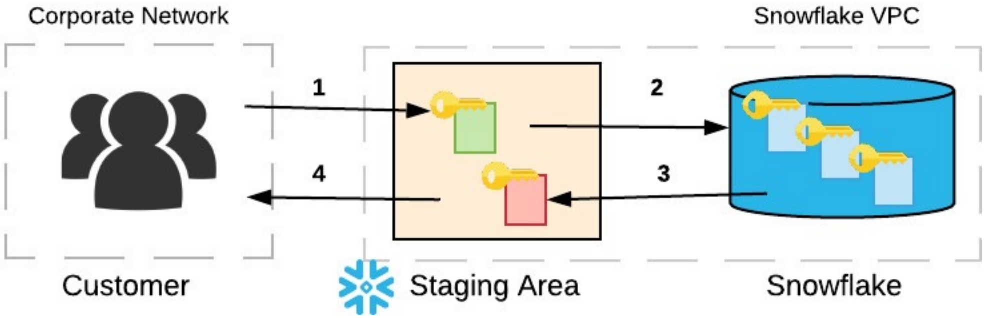
Data Encryption & Protection – Snowflake example

Consider an End-to-end encryption (E2EE)

(A) Customer-provided Staging Area

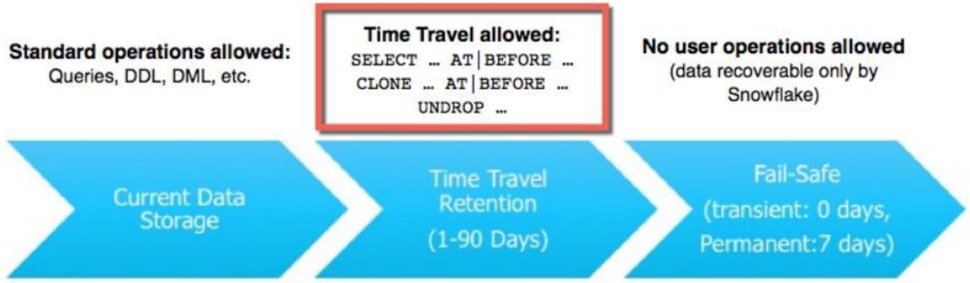


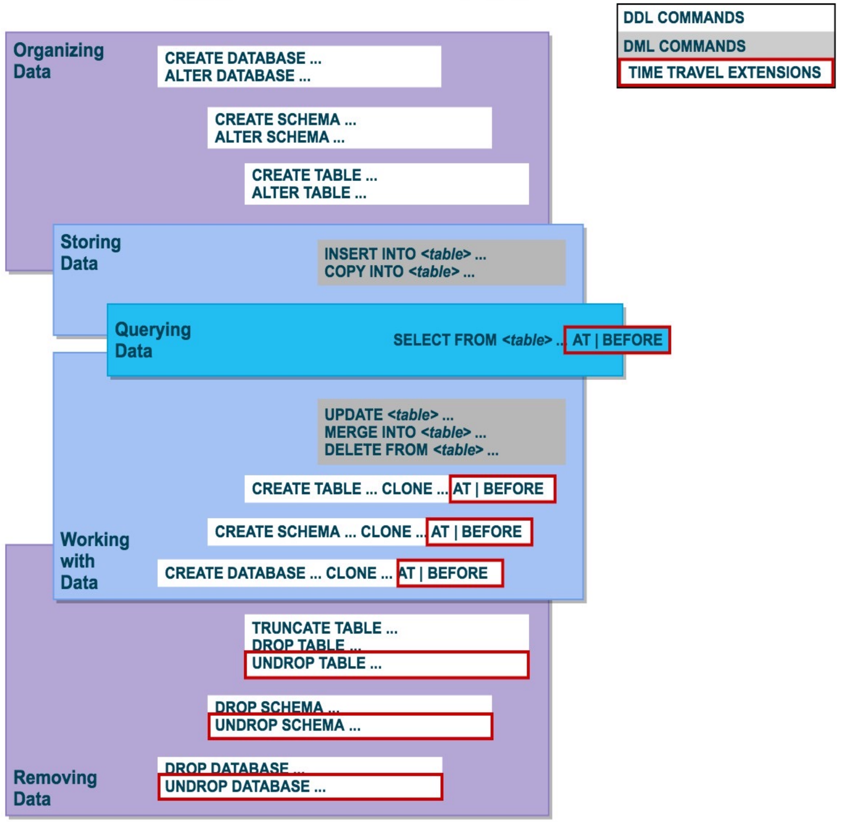
(B) Snowflake-provided Staging Area



- Using a client-side master key when reading or writing data between a cloud storage service stage and Snowflake.
- Using hierarchical key model consists of levels of keys (root key, account master keys, table master keys, file keys) and key rotation + periodic rekeying

Protecting data over time travel to make data to be accessible and recoverable in the event of accidental or intentional modification, removal, or corruption

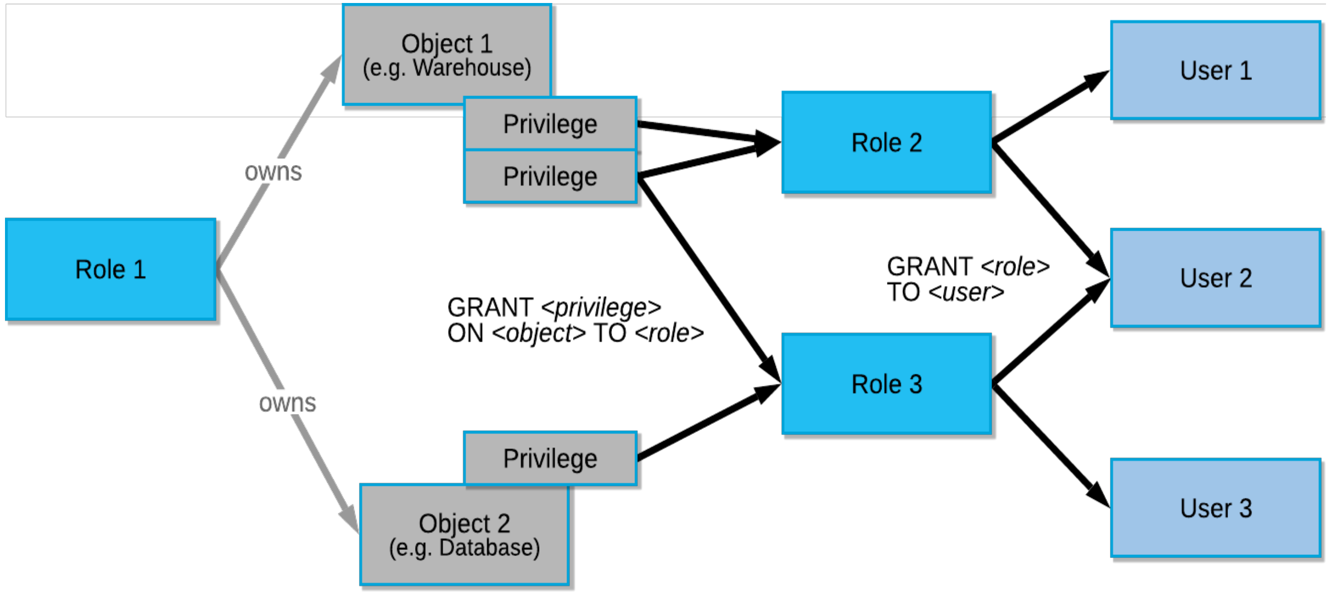




Access Control – Snowflake example

Leverage both access control mechanisms:

- 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.



Secure object down to row or column:

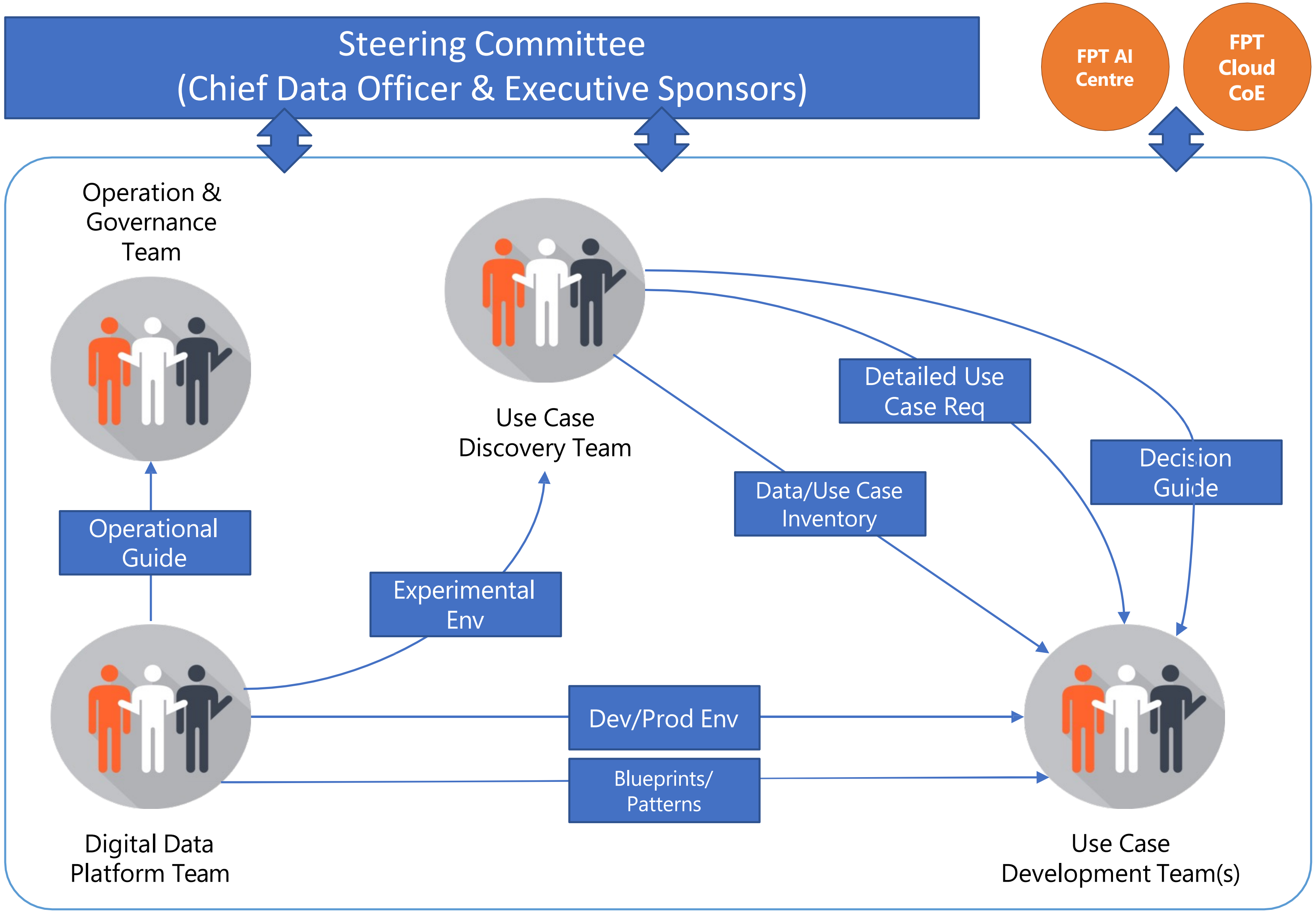
- Row-level Security: row in a table or view can be viewed from SELECT, UPDATE, DELETE, and MERGE statements.
- Column-level Security
  - Dynamic Data Masking, Hashing, Cryptographic, and Encryption Functions in Masking Policies
  - External Tokenization

Access governance: The user access history in a Access History View (a single record per query describing the columns the query accessed directly i.e. the base table, and indirectly i.e. derived objects, such as views, but Snowflake cannot record write operations)

Digital Data Platform

Project Execution

Organization



1 **Flexible Digital Data Platform** which helps data engineers, data scientists and business users can develop and experience scenarios

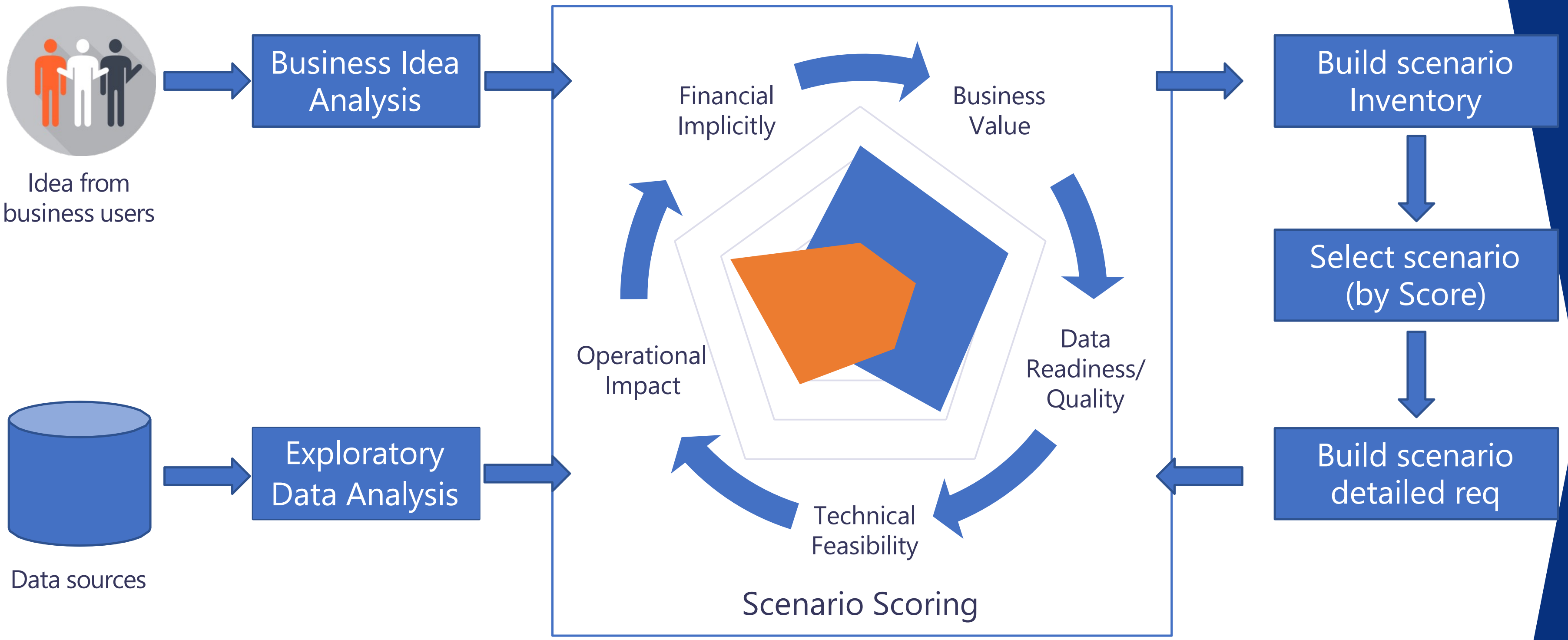
2 **Effective scenario discovery** is the approach to identify and select the relevant scenario to develop and bring to production.

3 **Optimal scenario development process** to control and manage when and why to bring a given scenario to next step.

4 **FPT AI Centre** is centralized pool of Data Scientists across various AI and data science domain.

5 **FPT Cloud CoE**: this team own practices and researches regarding Cloud technology and can support for cloud topics in case needed.

Scenario/ Use Case Discovery



Score will be 1 to 5 for each perspective, scenario with higher score is more relevant to select. Scenario scoring activity will be handled by brainstorming in focus group.

1 **Business Value**: what is the benefit will be brought to end-users or company if we can develop a scenario?

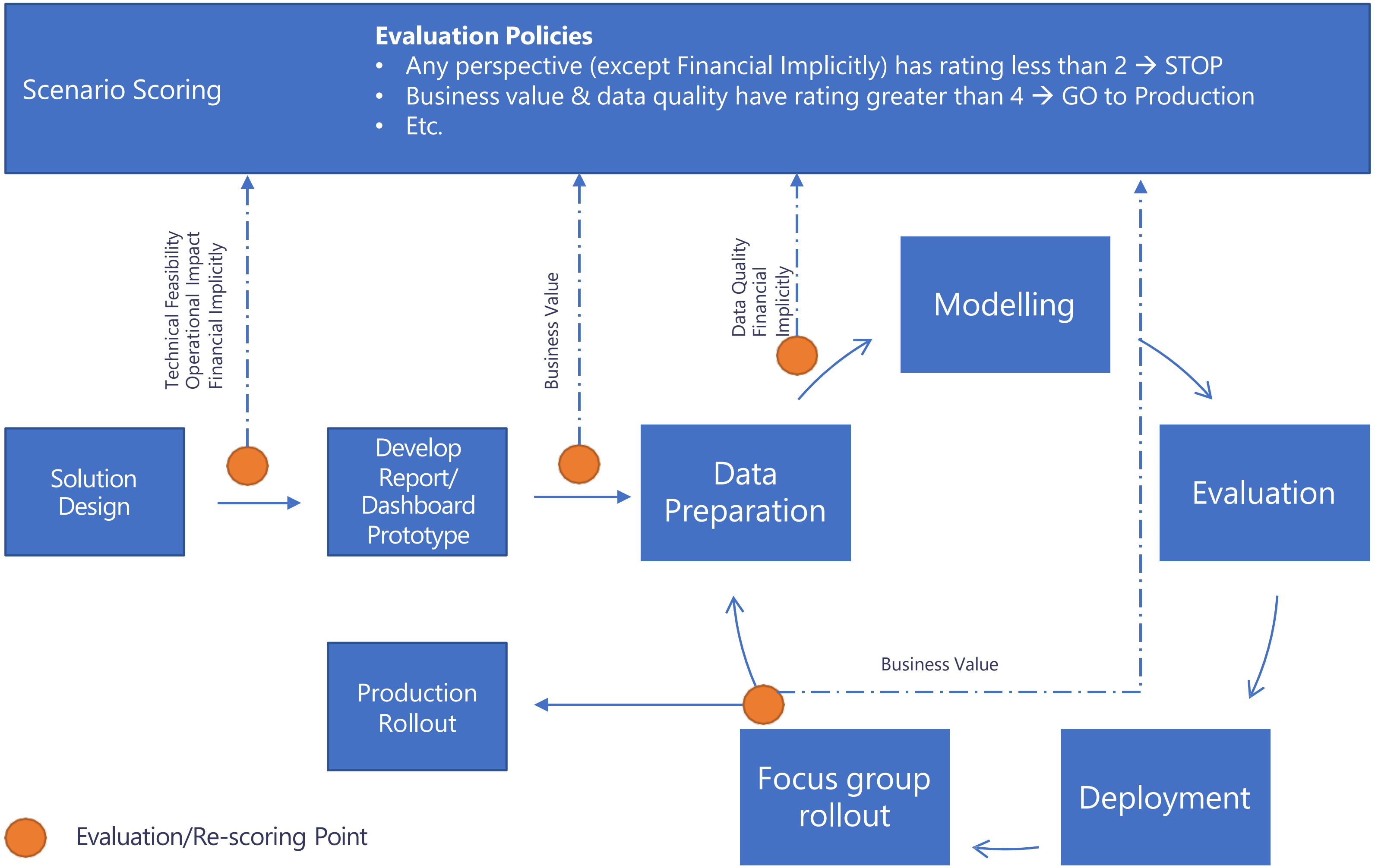
2 **Data Readiness/ Quality**: how good of data which is used to develop the scenario?

3 **Technical Feasibility**: how difficult regarding to technical aspect to develop the scenario?

4 **Operational Impact**: what is the impact to daily operation of not only this system but also other systems or data source?

5 **Financial implicitly**: what is the cost to develop the scenario (including development cost, hardware cost, license cost, data acquisition cost, etc.

Scenario/ Use Case Development



Scenario Scoring

Evaluation Policies

- Any perspective (except Financial Implicitly) has rating less than 2 → STOP
- Business value & data quality have rating greater than 4 → GO to Production
- Etc.

● Evaluation/Re-scoring Point

Agile will be methodology to develop and rollout a scenario. A scrum team will take responsibility to develop only one scenario at any point of time. If a scenario will be stopped by any reason, new scenario will be assigned to team. Sprint duration will be 2 weeks as usual.

Services & Offers

Service Offerings

Engagement	Scope & Delivery Targets
Assessment & Consulting	• Assess and propose Digital Platform (the modernized data platform) architecture and roadmap • Provide detailed solution, design and implementation plan for Digital Platform or existing issues • Provide a POC for complex projects • Data governance consulting
Platform Implementation	• Migrate the traditional EDW to Digital Platform on prem or cloud, including migrating data and rebuild data processing and analytics/BI • Renovate/optimize existing Data Platform • Building new data layers and data pipelines from ingestion, processing to analytics/BI on existing Data Platform
Use Case Implementation	• Implement E2E data pipeline or a part, handling scheduled/event-triggering batch and streaming data with flow orchestration • Implement BI reports & dashboards • Implement big data processing • Implement AI/ML use cases
Operation & Support	• Provide managed Digital Data Platform operation services, data administration. • Supporting services for managing, critical problem investigation and troubleshooting. • Data Quality Management • Data Governance

Capability

Our Professional Resources

Microsoft Gold Partner

aws partner network

Premier Consulting Partner

Migration Competency
MSP Partner
Solution Provider
Marketplace Seller
Channel Partner

500+

- ✓ Cloud Consultant
- ✓ Big Data Consultant
- ✓ Data Architect
- ✓ Data Engineer
- ✓ Data Analyst
- ✓ Data Scientist

1500+ Cloud Professionals

Microsoft CERTIFIED EXPERT

aws certified Solutions Architect Professional

Google Cloud Certified PROFESSIONAL CLOUD ARCHITECT

Optimized for Cloud

aws

Technology Stack

Power BI
tableau
MicroStrategy
dremio
SQL
Azure Synapse Analytics
snowflake
amazon REDSHIFT
druid
APACHE HBASE
HDInsight
amazon EMR
Informatica
hadoop
APACHE Spark
CLOUDERA
databricks
HIVE
talend
Azure Data lake
amazon S3
hadoop HDFS
collibra

Security Compliance

CIS Center for Internet Security
PCi Security Standards Council
HIPAA COMPLIANT
ISO 27001

Thank you.

FPT Digital Kaizen ™