Data Lake Overview

James Serra

Data & AI Architect

Microsoft, NYC MTC

JamesSerra3@gmail.com

Blog: JamesSerra.com

About Me

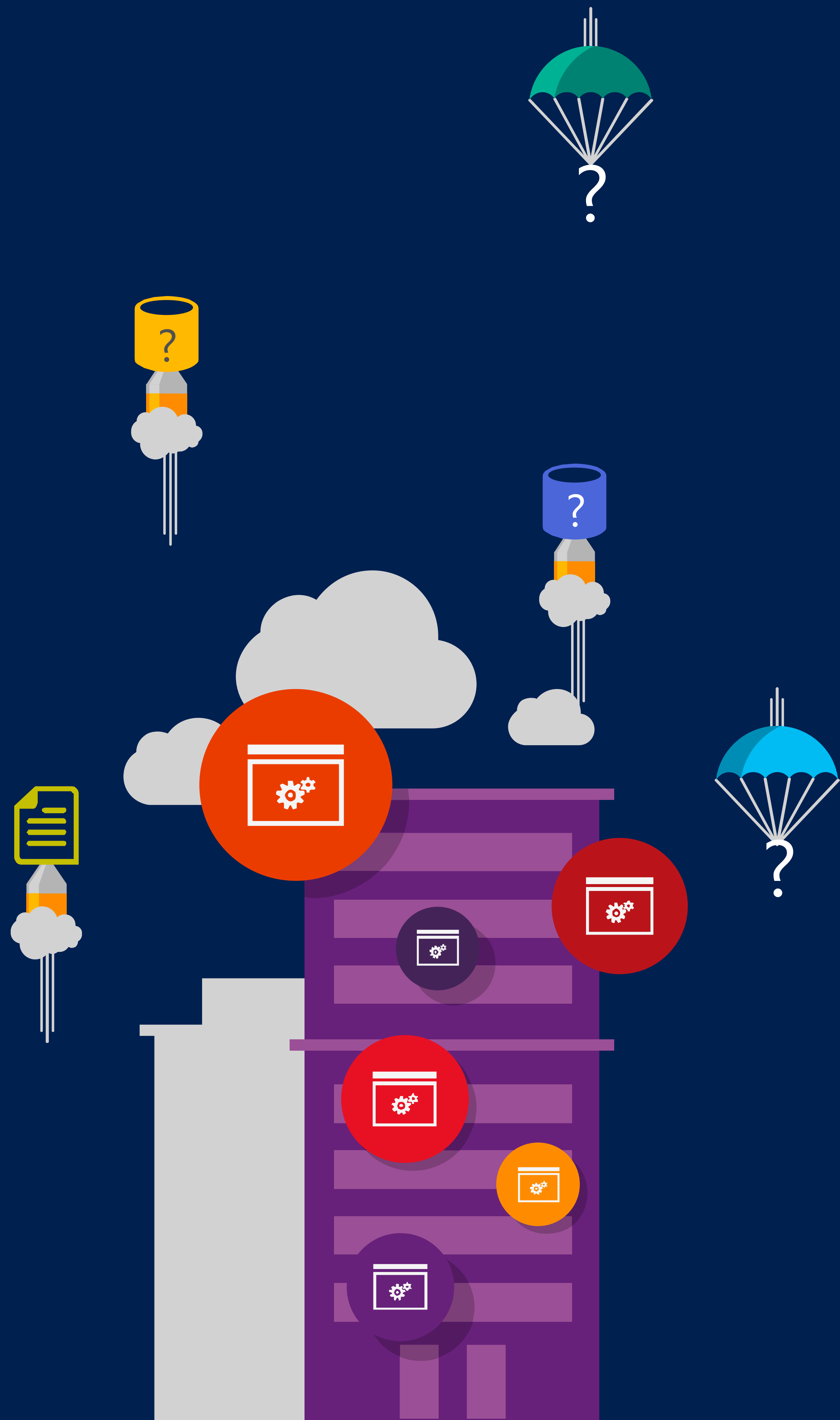
- Microsoft, Big Data Evangelist
- In IT for 30 years, worked on many BI and DW projects
- Worked as desktop/web/database developer, DBA, BI and DW architect and developer, MDM architect, PDW/APS developer
- Been perm employee, contractor, consultant, business owner
- Presenter at PASS Business Analytics Conference, PASS Summit, Enterprise Data World conference
- Certifications: MCSE: Data Platform, Business Intelligence; MS: Architecting Microsoft Azure Solutions, Design and Implement Big Data Analytics Solutions, Design and Implement Cloud Data Platform Solutions
- Blog at JamesSerra.com
- Former SQL Server MVP
- Author of book "Reporting with Microsoft SQL Server 2012"

MVP Microsoft® Most Valuable Professional

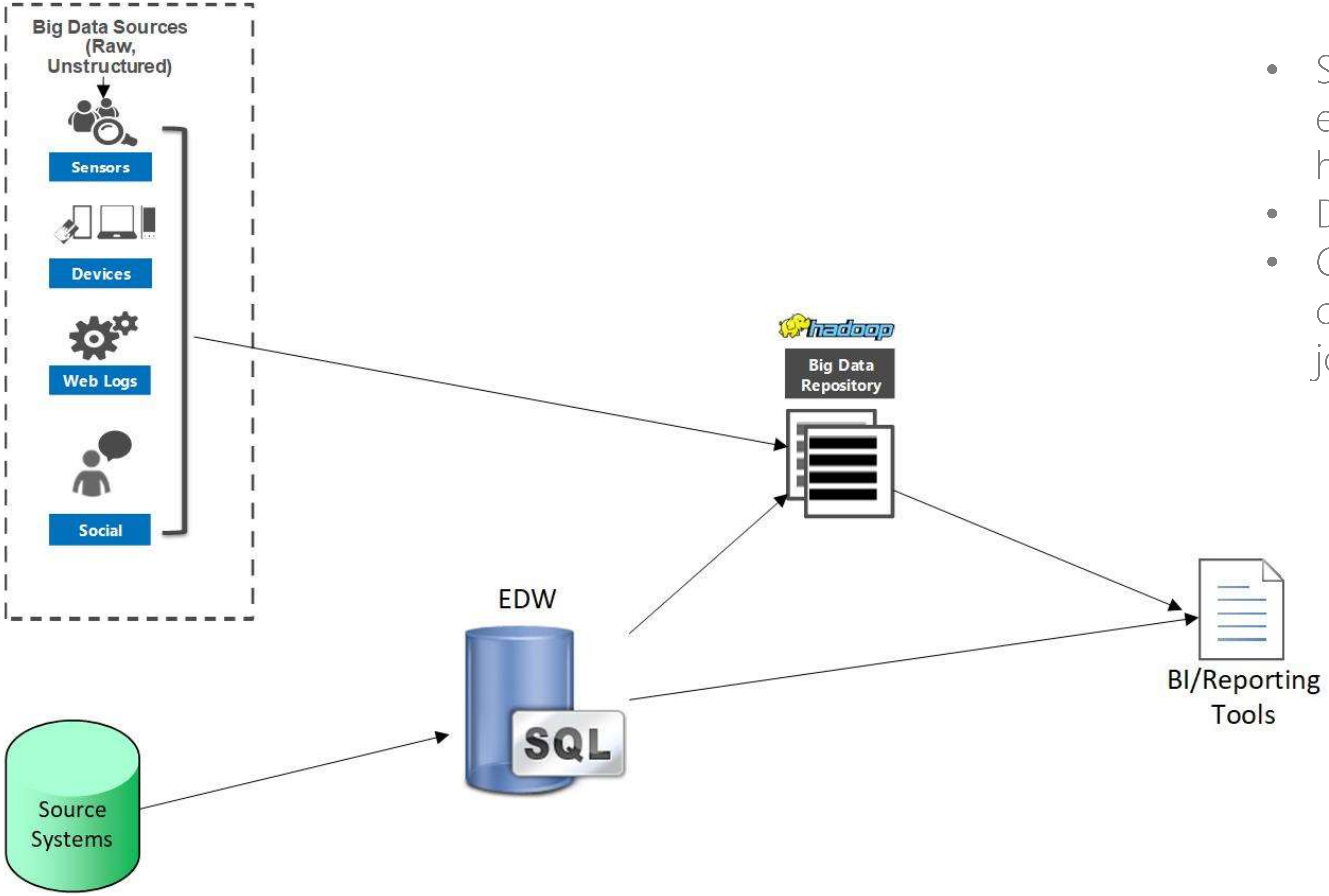
Agenda

- Big Data Architectures
- Why data lakes?
- Top-down vs Bottom-up
- Data lake defined
- Creating ADLS Gen2
- Data Lake Use Cases

Big Data Architectures

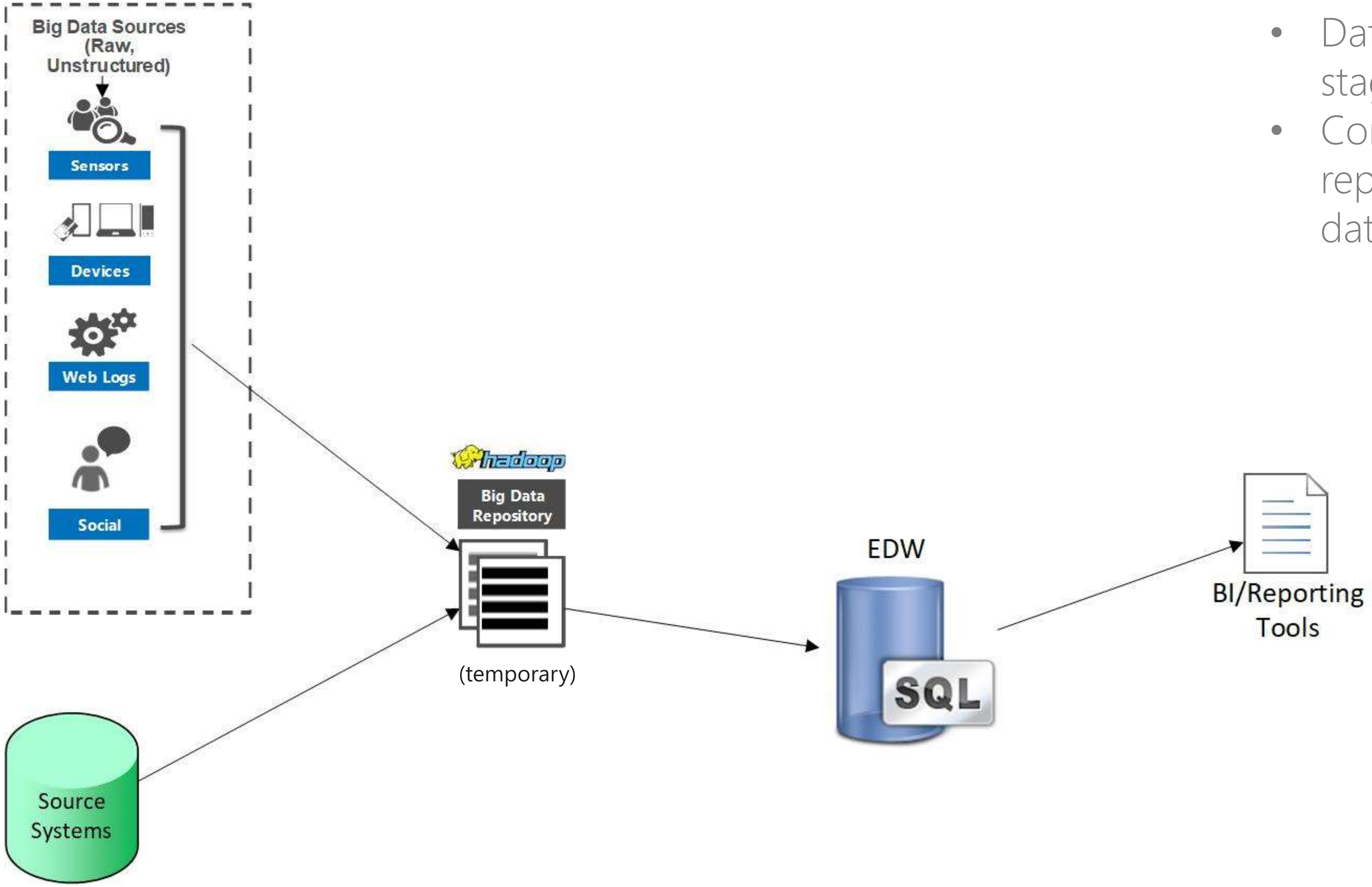


Enterprise data warehouse augmentation



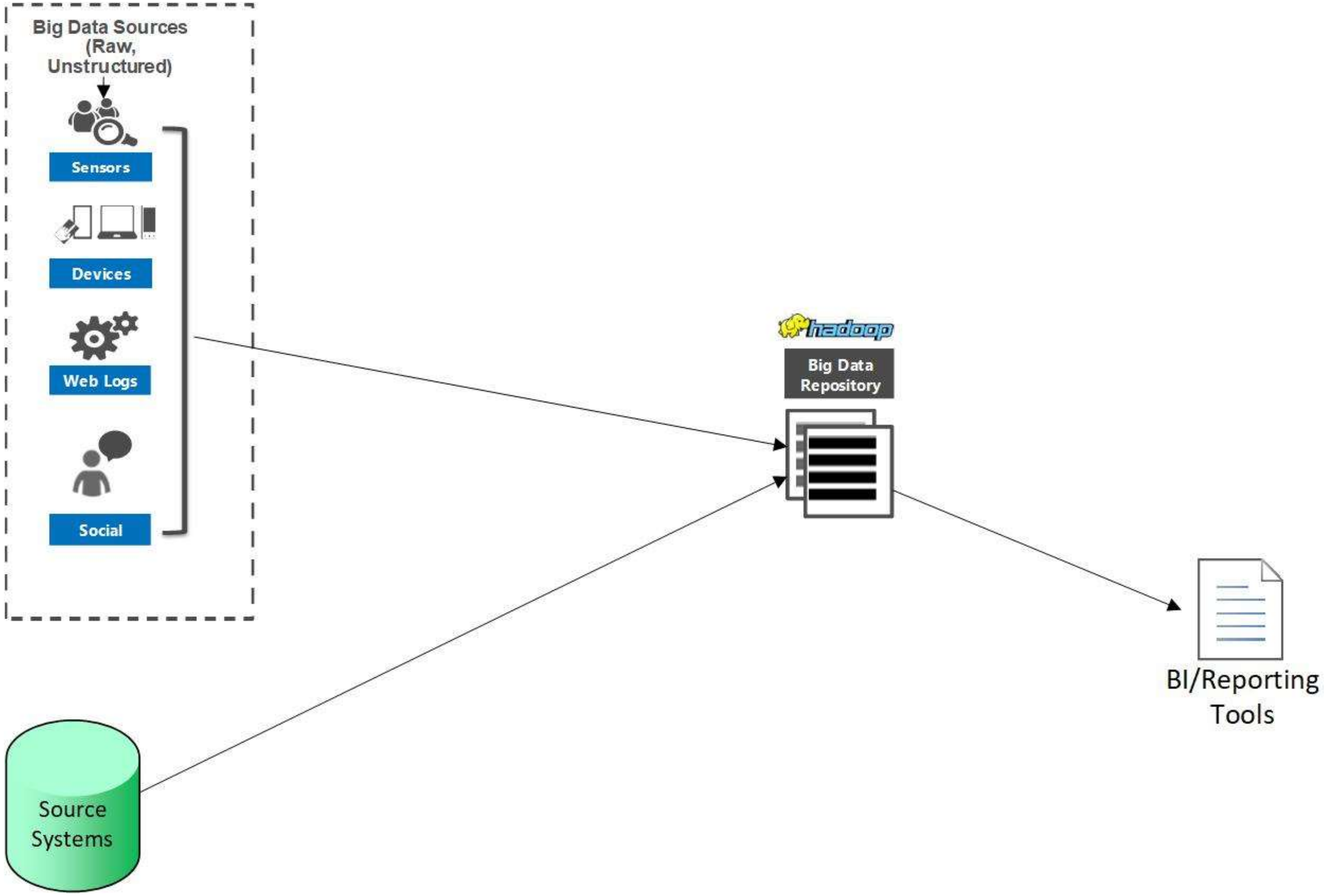
- Seen when EDW has been in existence a while and EDW can't handle new data
- Data hub, not data lake
- Cons: not offloading EDW work, can't use existing tools, difficulty joining data in data hub with EDW

Data hub plus EDW



- Data hub is used as temporary staging and refining, no reporting
- Cons: data hub is temporary, no reporting/analyzing done with the data hub

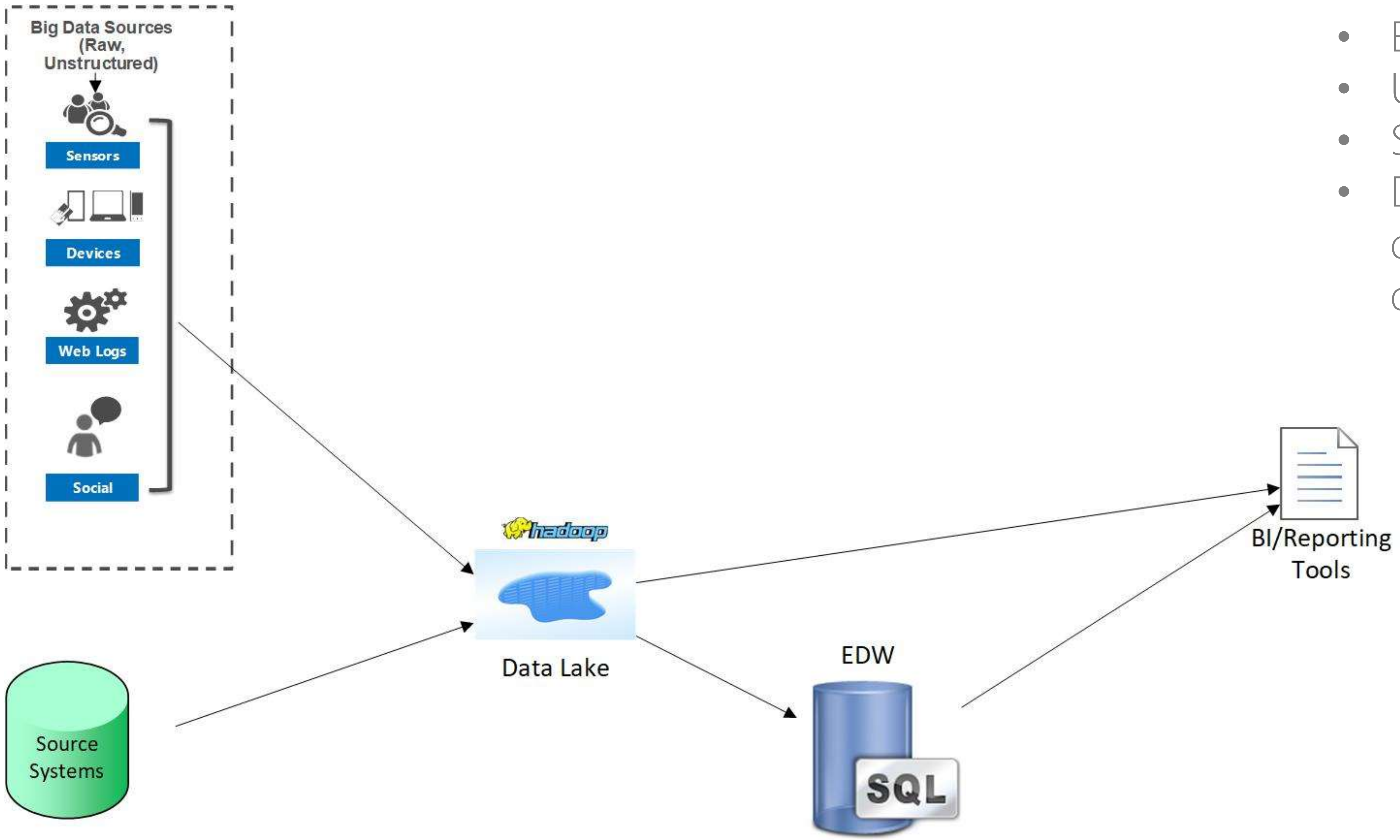
All-in-one



- Data hub is total solution, no EDW
- Cons: queries are slower, new training for reporting tools, difficulty understanding data, security limitations

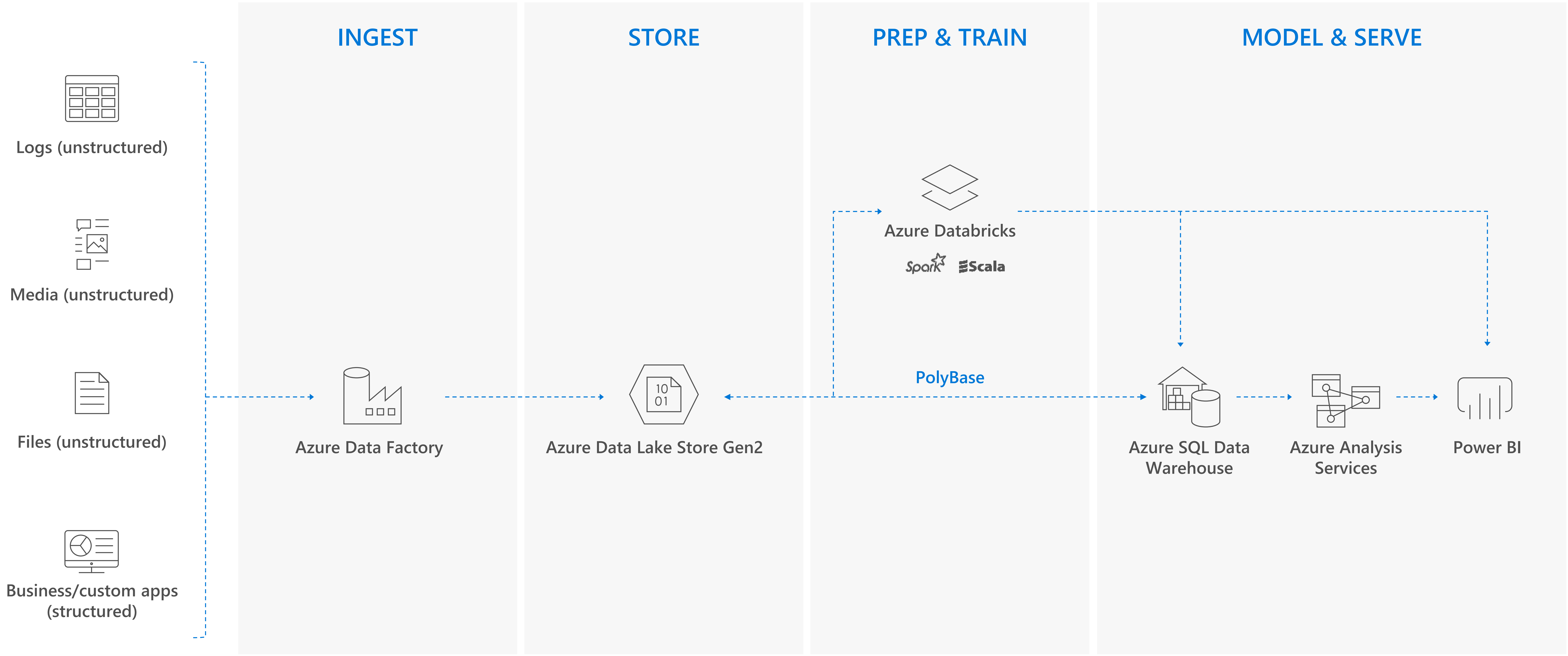
Is the traditional data warehouse dead? https://www.jamesserra.com/archive/2017/12/is-the-traditional-data-warehouse-dead/

Modern Data Warehouse



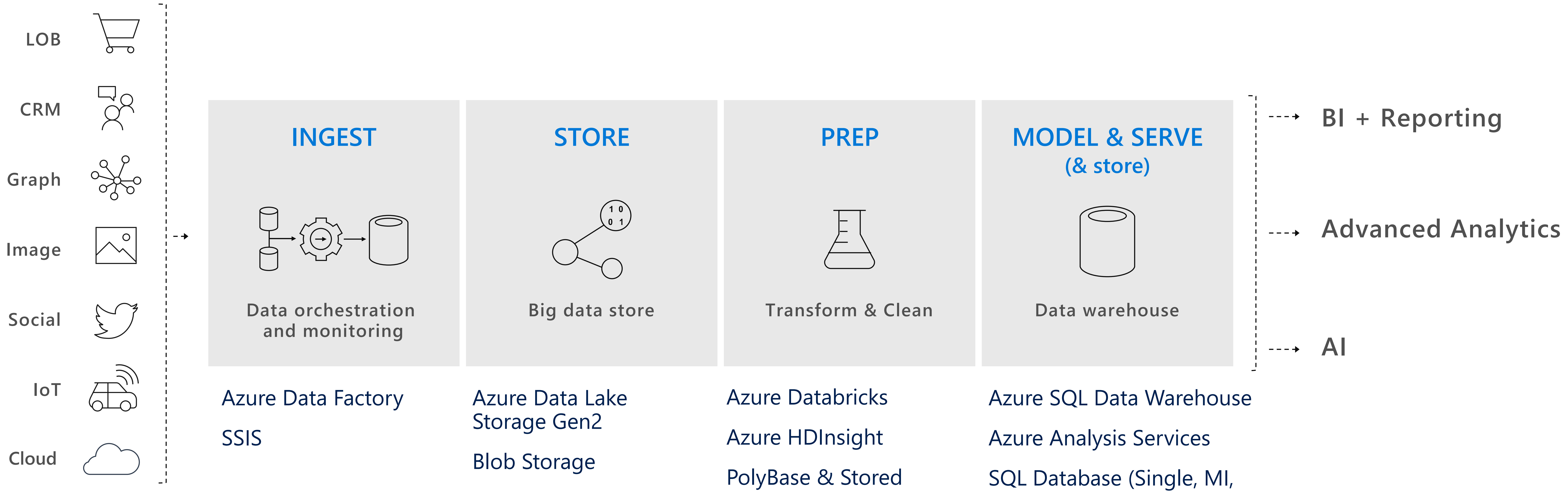
- Evolution of three previous scenarios
- Ultimate goal
- Supports future data needs
- Data harmonized and analyzed in the data lake or moved to EDW for more quality and performance

MODERN DATA WAREHOUSE



Microsoft Azure also supports other Big Data services like Azure HDInsight to allow customers to tailor the above architecture to meet their unique needs.

Modern Data Warehouse (possible products by four areas)



Azure Data Factory

SSIS

Azure Data Lake Storage Gen2

Blob Storage

SQL Server 2019 Big Data Cluster

Azure Databricks

Azure HDInsight

PolyBase & Stored Procedures

Power BI Dataflows

Azure SQL Data Warehouse

Azure Analysis Services

SQL Database (Single, MI, HyperScale, Serverless)

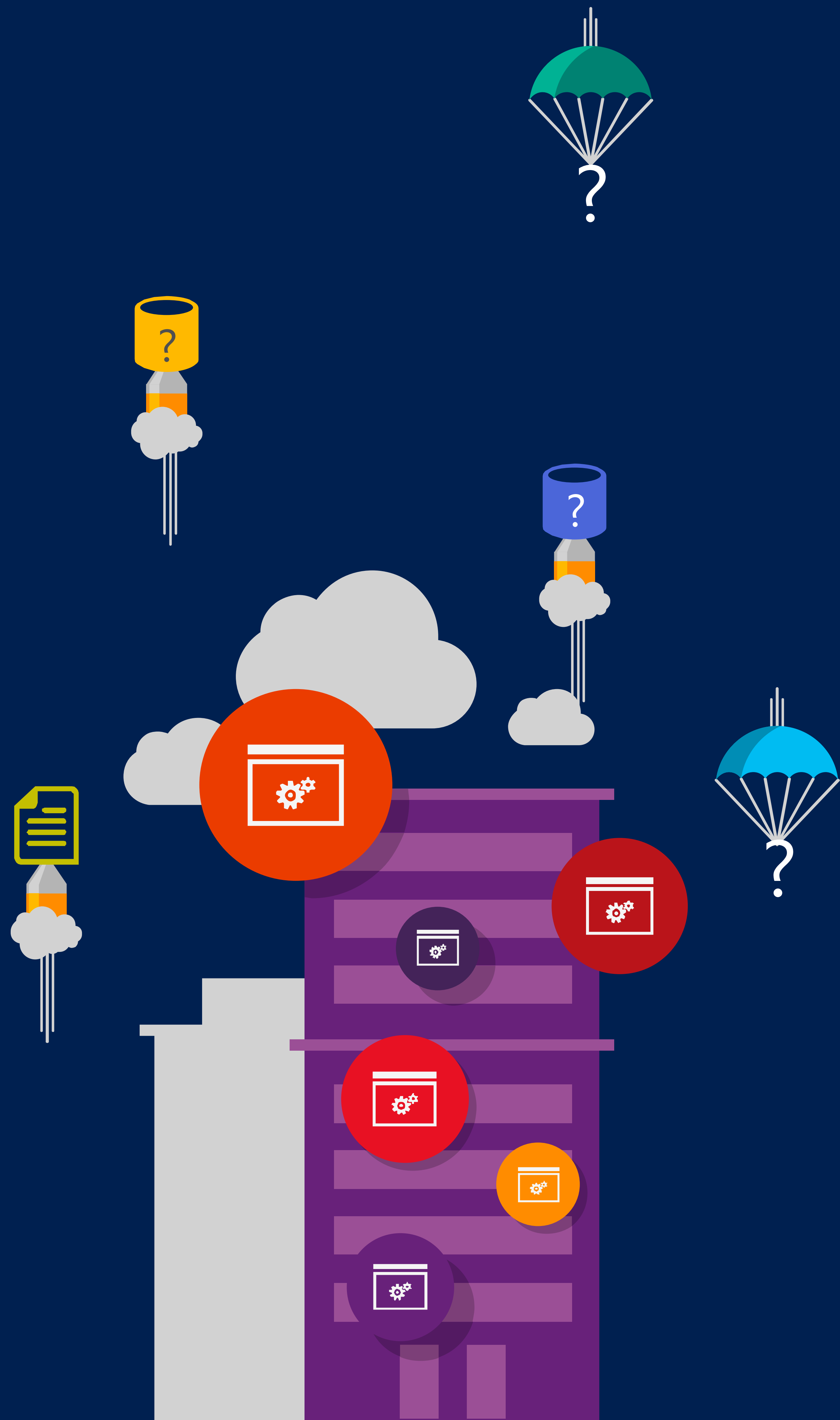
SQL Server in a VM

Cosmos DB

Power BI Aggregations

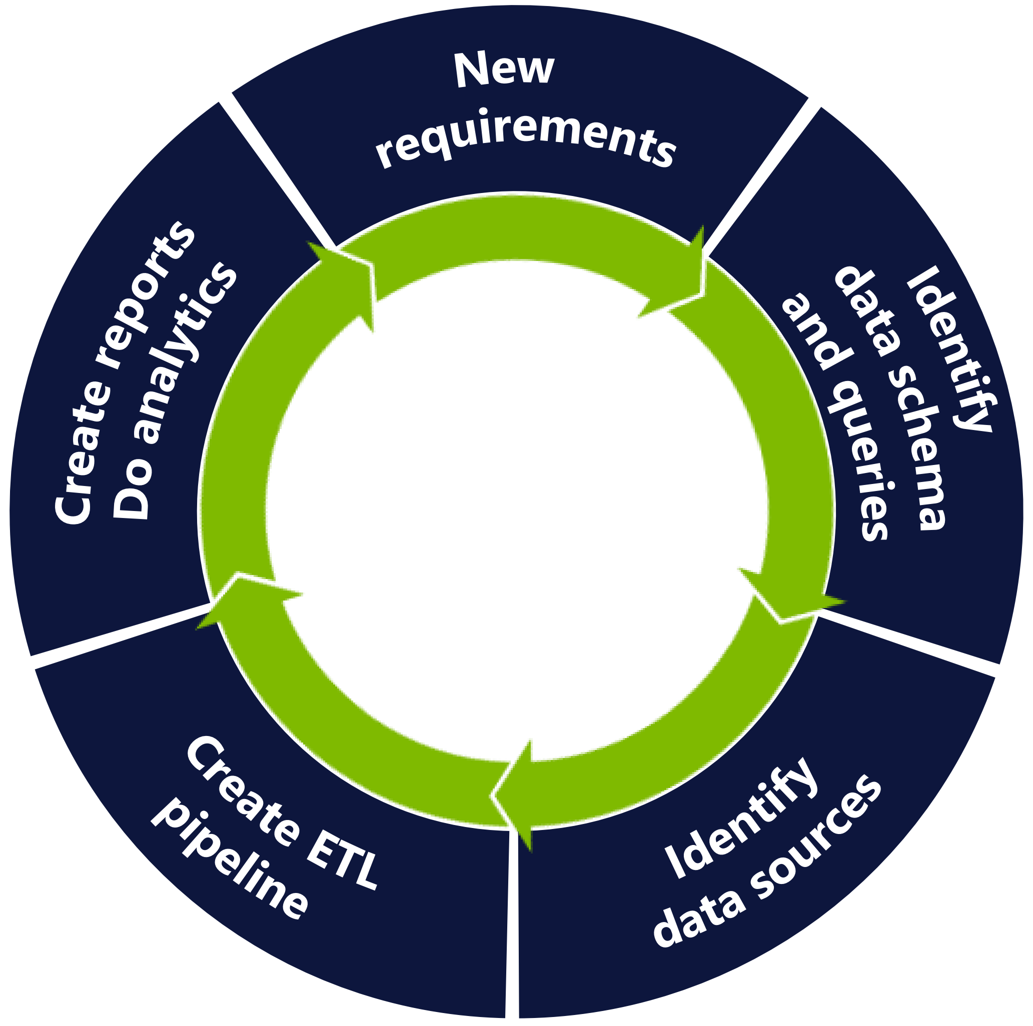
Note: Those products that span more than one area are listed in there primary area

Why data lakes?



Traditional business analytics process

1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding database schema and queries
3. Identify the required data sources
4. Create a Extract-Transform-Load (ETL) pipeline to extract required data (curation) and transform it to target schema (‘*schema-on-write*’)
5. Create reports. Analyze data



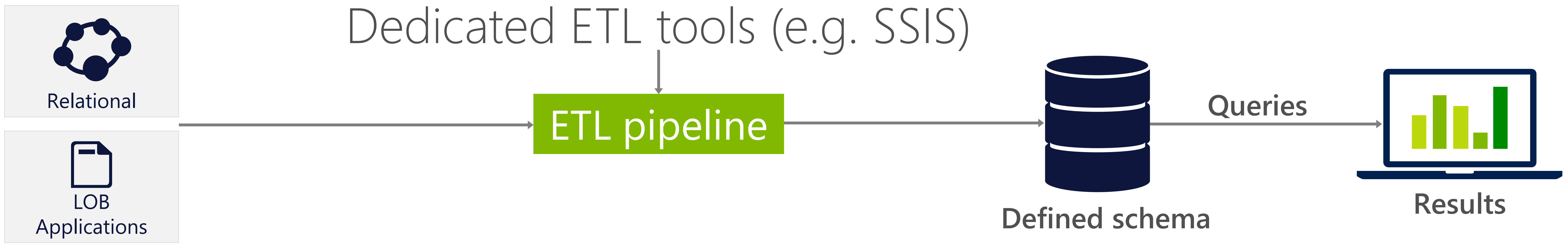
New requirements

Identify data schema and queries

Identify data sources

Create ETL pipeline

Create reports Do analytics



Dedicated ETL tools (e.g. SSIS)

Relational

LOB Applications

ETL pipeline

Defined schema

Queries

Results

All data not immediately required is discarded or archived

Need to collect any data

Harness the growing and changing nature of data

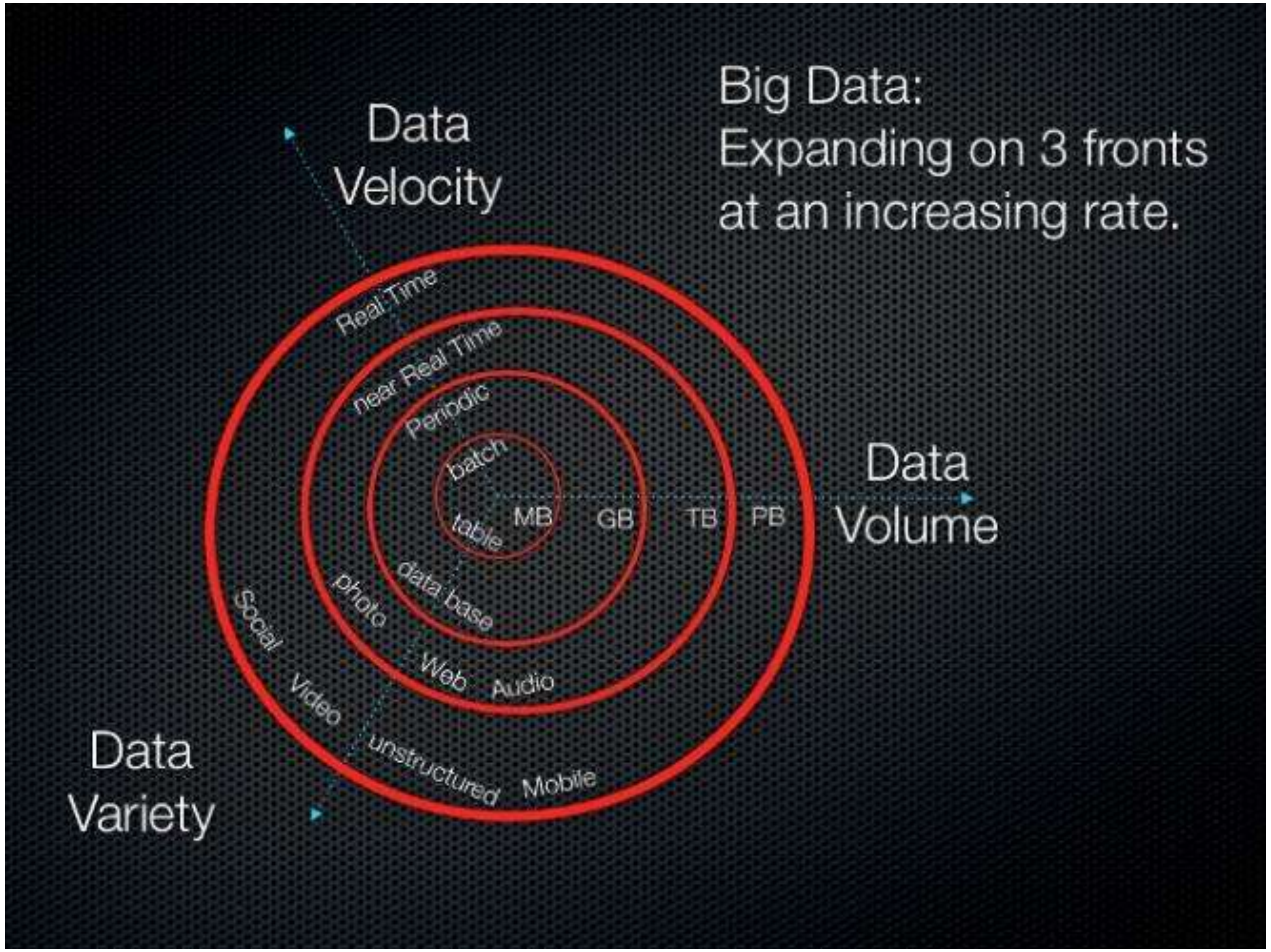
Structured

Unstructured

Streaming

- Challenge is combining transactional data stored in relational databases with less structured data
- *Big Data = All Data*
- Get the right information to the right people at the right time in the right format

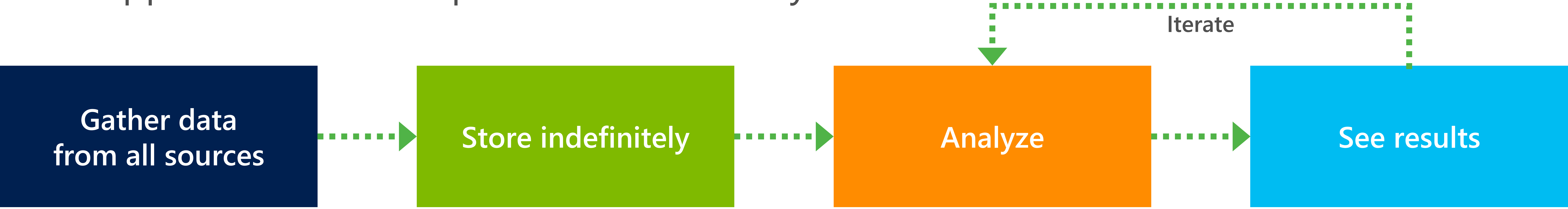
The three V's


Big Data:
Expanding on 3 fronts
at an increasing rate.
Data Velocity
Real Time
near Real Time
Periodic
batch
Data Volume
MB
GB
TB
PB
Data Variety
table
data base
photo
Web
Audio
Social
Video
unstructured
Mobile

New big data thinking: All data has value

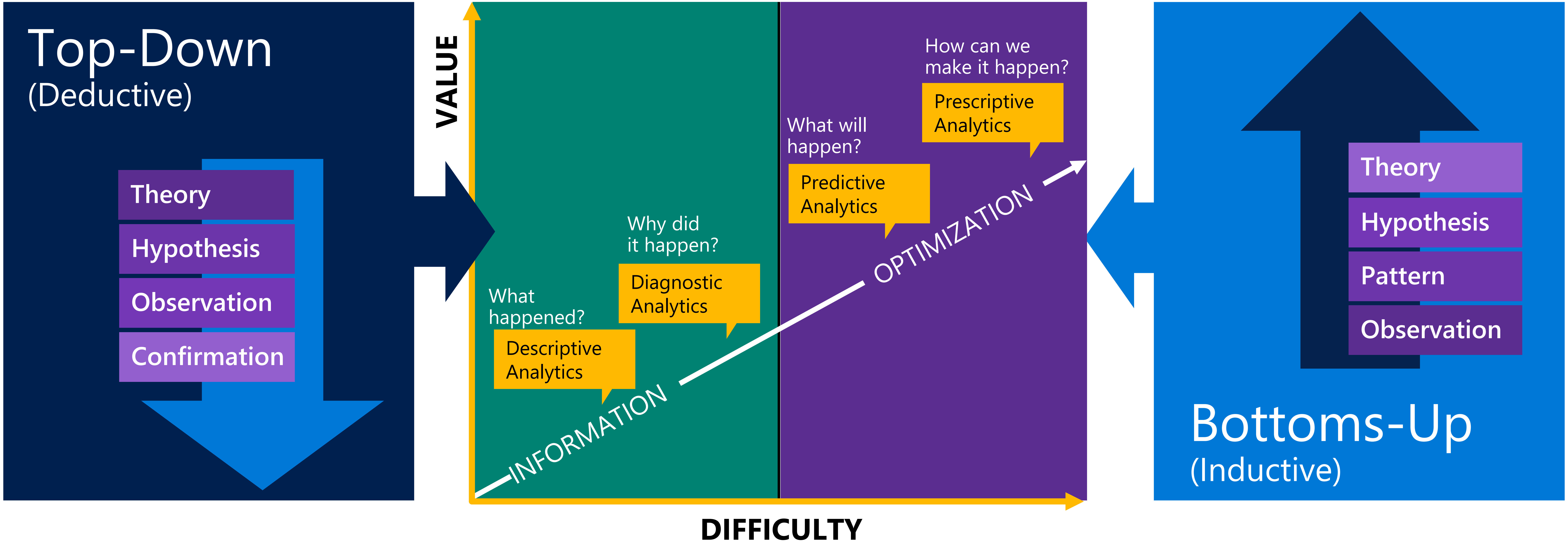
Use a data lake:

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read)*.
- Apps and users interpret the data as they see fit



Top-down vs Bottom-up

Two Approaches to getting value out of data: Top-Down + Bottoms-Up



Top-Down
(Deductive)

Theory

Hypothesis

Observation

Confirmation

What happened?
Descriptive Analytics

Why did it happen?
Diagnostic Analytics

What will happen?
Predictive Analytics

How can we make it happen?
Prescriptive Analytics

Theory

Hypothesis

Pattern

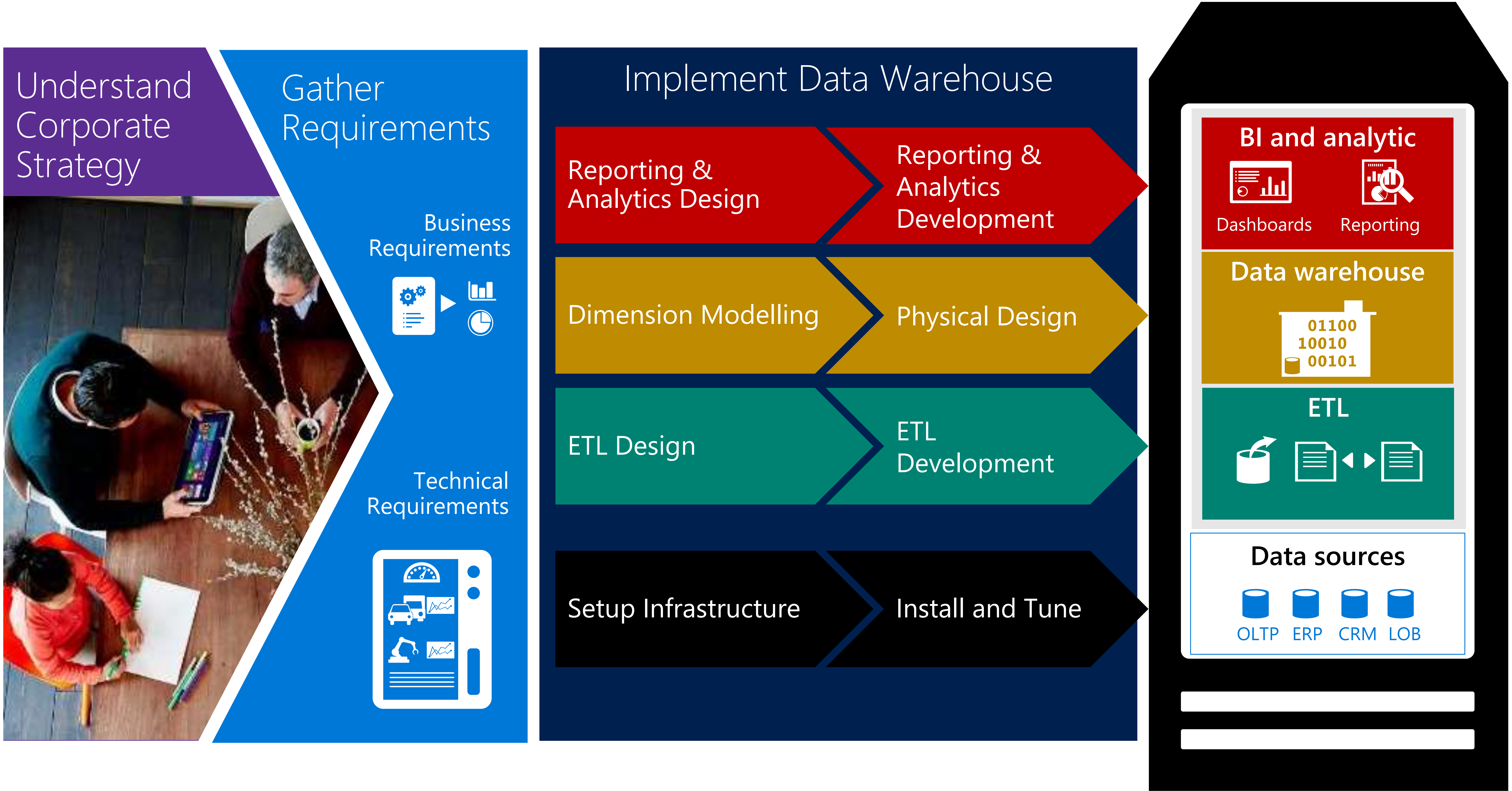
Observation

Bottoms-Up
(Inductive)

- Know the questions to ask
- Lot's of upfront work to get the data to where you can use it
- Model first

- Don't know the questions to ask
- Little upfront work needs to be done to start using data
- Model later

Data Warehousing Uses A Top-Down Approach



Understand Corporate Strategy

Gather Requirements

Business Requirements

Technical Requirements

Implement Data Warehouse

Reporting & Analytics Design

Reporting & Analytics Development

Dimension Modelling

Physical Design

ETL Design

ETL Development

Setup Infrastructure

Install and Tune

BI and analytic

Dashboards

Reporting

Data warehouse

01100
10010
00101

ETL

Data sources

OLTP

ERP

CRM

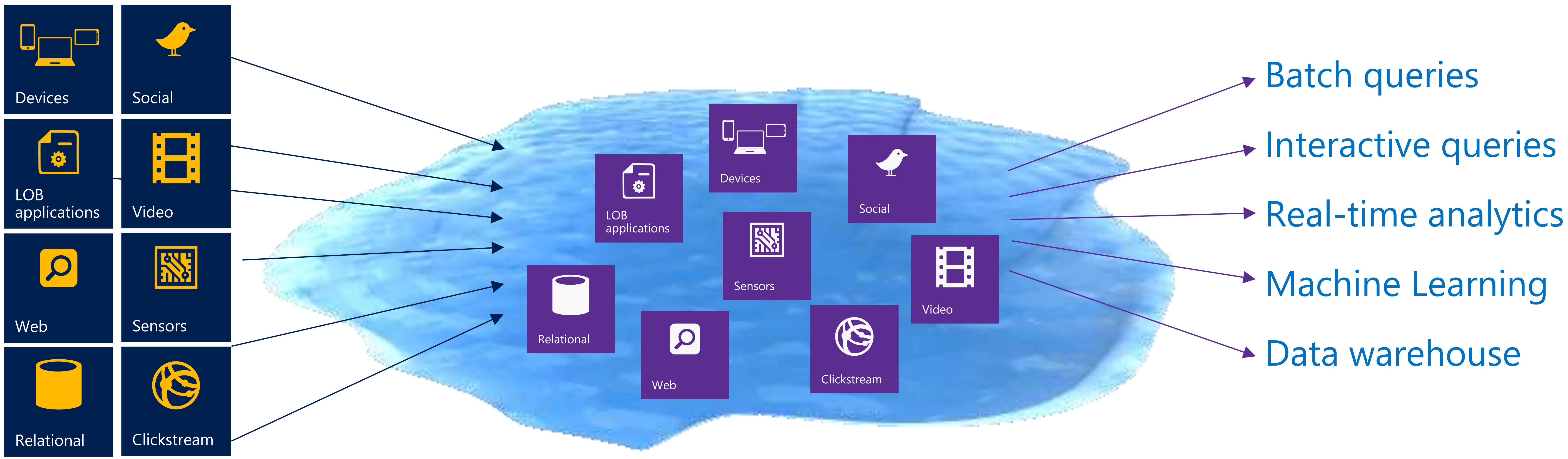
LOB

The "data lake" Uses A Bottoms-Up Approach

Ingest all data
regardless of requirements

Store all data
in native format without schema definition

Do analysis
Using analytic engines like Hadoop



- Batch queries
- Interactive queries
- Real-time analytics
- Machine Learning
- Data warehouse

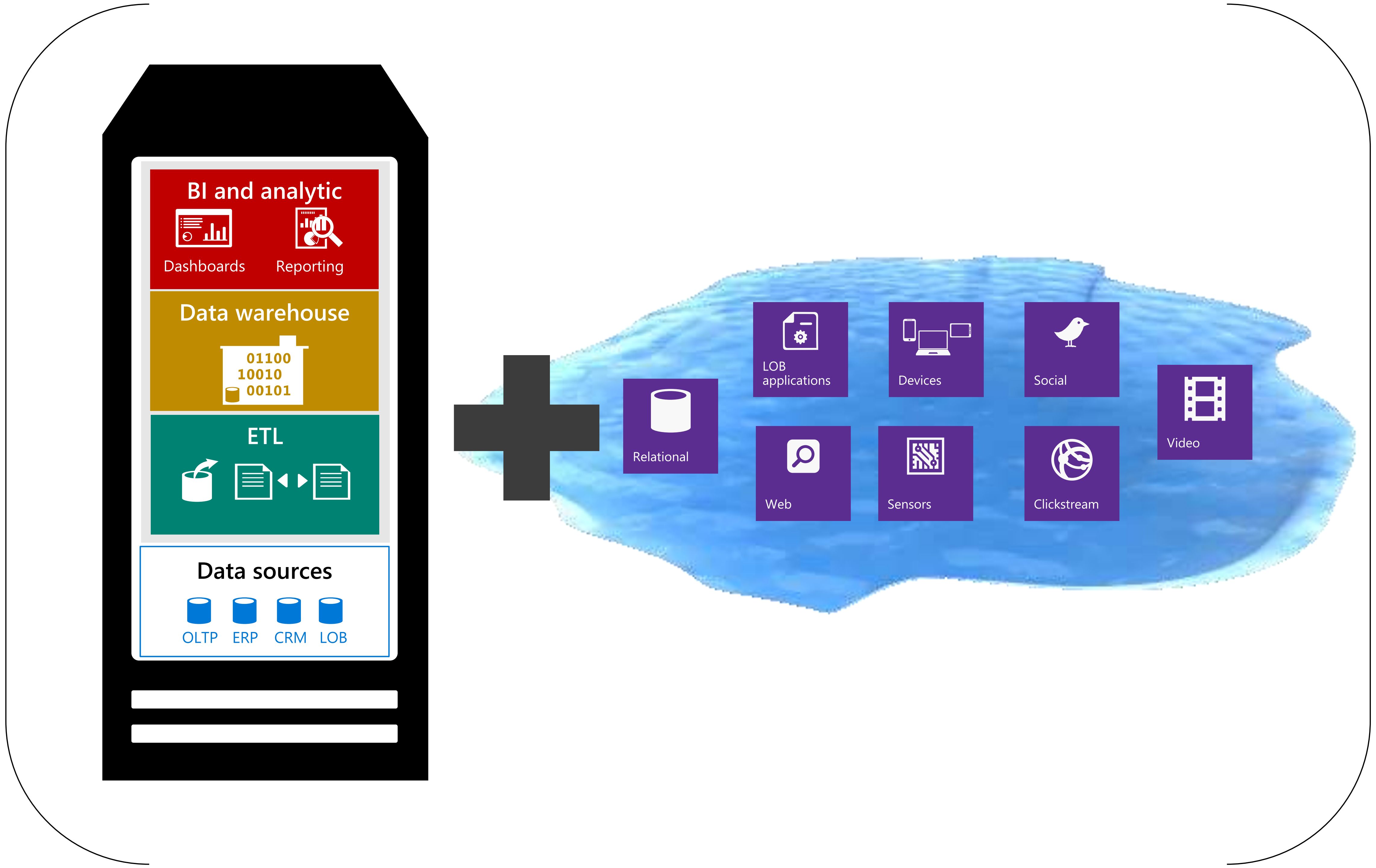
Data Lake + Data Warehouse Better Together

What happened?

Descriptive Analytics

Why did it happen?

Diagnostic Analytics



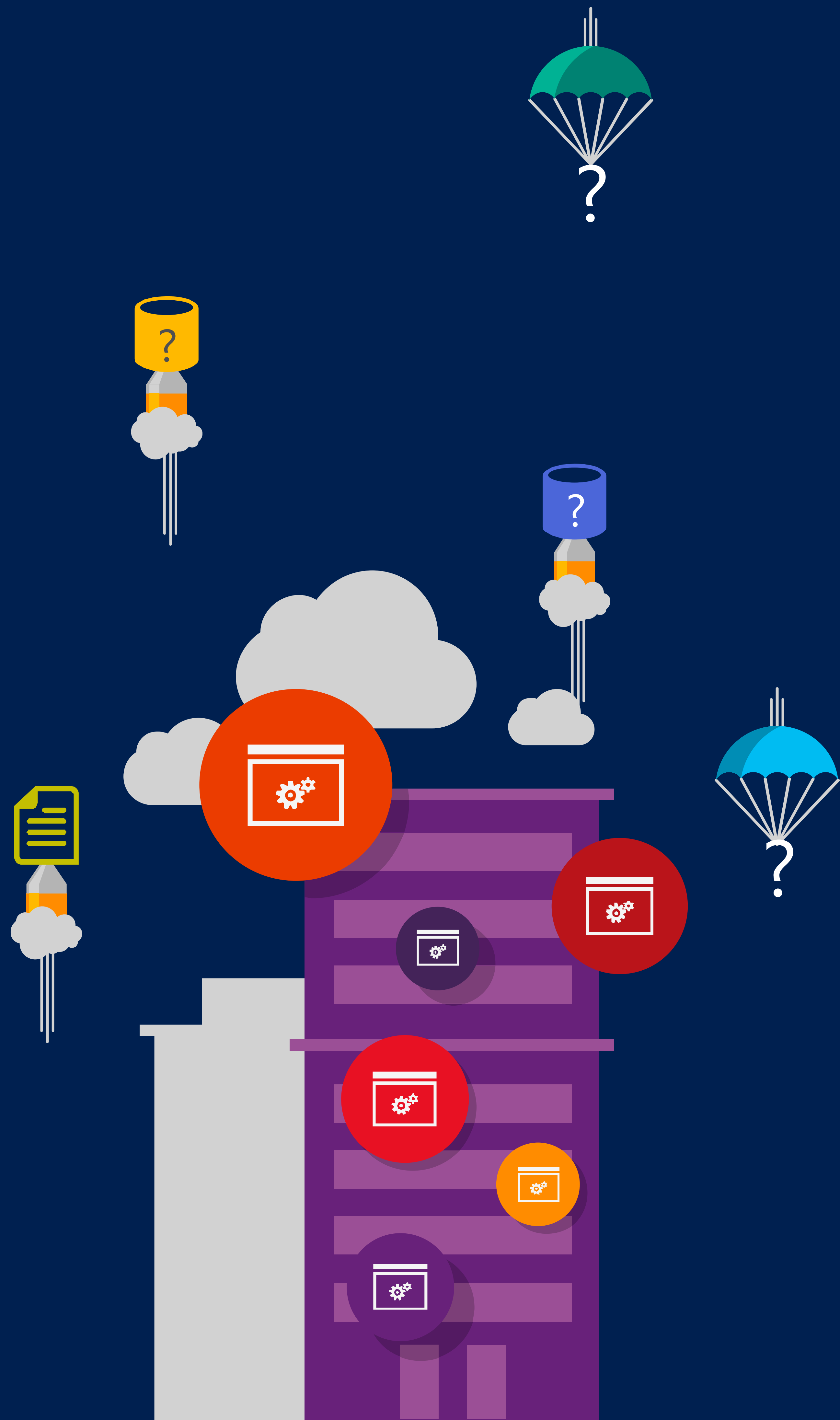
What will happen?

Predictive Analytics

How can we make it happen

Prescriptive Analytics

Data lake defined



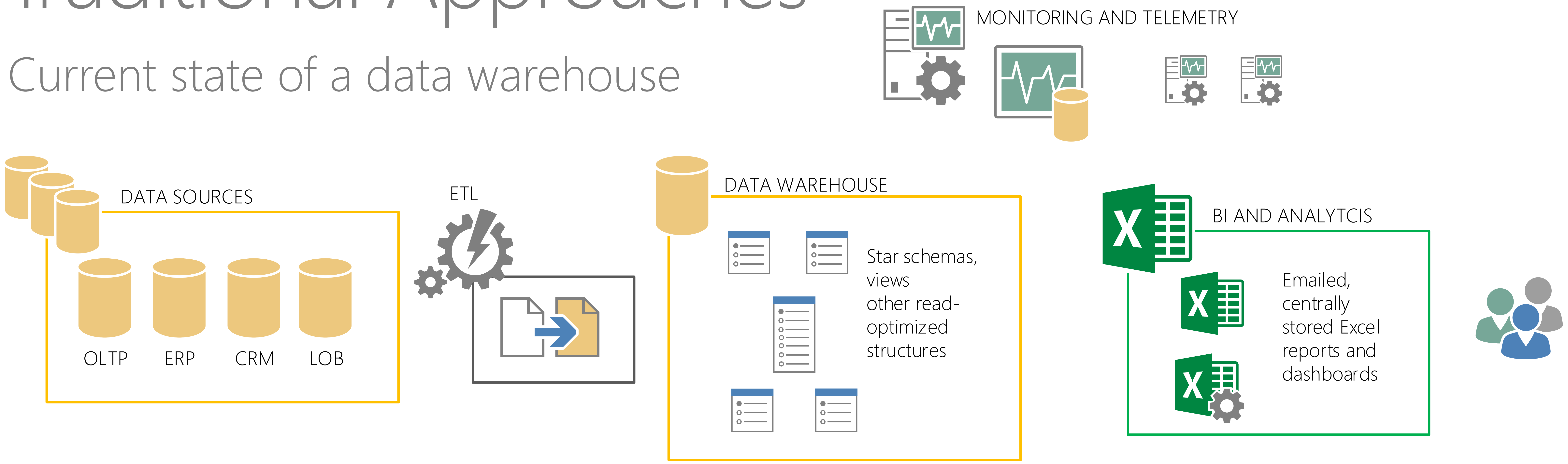
Exactly what is a data lake?

A storage repository, usually Hadoop, that holds a vast amount of raw data in its native format until it is needed.

- Inexpensively store unlimited data
- Centralized place for multiple subjects (single version of the truth)
- Collect all data “just in case” (data hoarding)
- Easy integration of differently-structured data
- Store data with no modeling – “Schema on read”
- Complements enterprise data warehouse (EDW)
- Frees up expensive EDW resources for queries instead of using EDW resources for transformations (avoiding user contention)
- Hadoop cluster offers faster ETL processing over SMP solutions
- Quick user access to data for power users/data scientists (allowing for faster ROI)
- Data exploration to see if data valuable before writing ETL and schema for relational database, or use for one-time report
- Allows use of Hadoop tools such as ETL and extreme analytics
- Place to land IoT streaming data
- On-line archive or backup for data warehouse data
- With Hadoop/ADLS, high availability and disaster recovery built in
- Keep raw data so don’t have to go back to source if need to re-run
- Allows for data to be used many times for different analytic needs and use cases
- Cost savings and faster transformations: storage tiers with lifecycle management; separation of storage and compute resources allowing multiple instances of different sizes working with the same data simultaneously vs scaling data warehouse; low-cost storage for raw data saving space on the EDW
- Extreme performance for transformations by having multiple compute options each accessing different folders containing data
- The ability for an end-user or product to easily access the data from any location

Traditional Approaches

Current state of a data warehouse



MONITORING AND TELEMETRY

DATA SOURCES

OLTP ERP CRM LOB

ETL

DATA WAREHOUSE

Star schemas, views other read-optimized structures

BI AND ANALYTCIS

Emailed, centrally stored Excel reports and dashboards

Well manicured, often relational sources

Known and expected data volume and formats

Little to no change

Complex, rigid transformations

Required extensive monitoring

Transformed historical into read structures

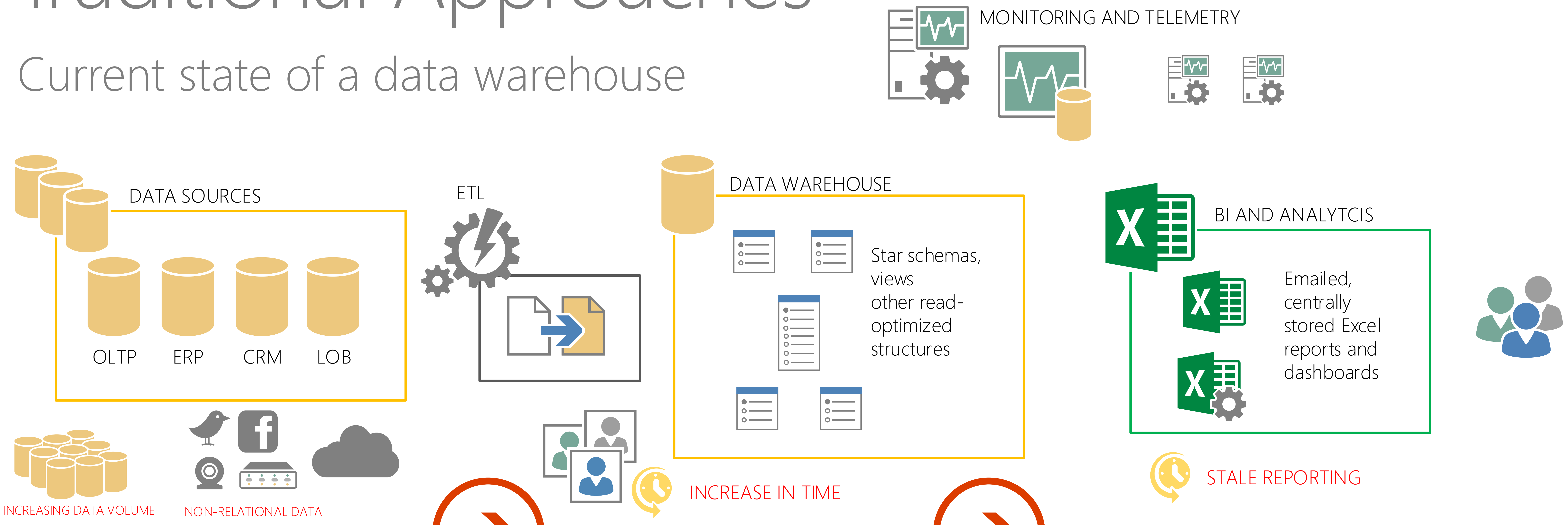
Flat, canned or multi-dimensional access to historical data

Many reports, multiple versions of the truth

24 to 48h delay

Traditional Approaches

Current state of a data warehouse



MONITORING AND TELEMETRY

DATA SOURCES

OLTP ERP CRM LOB

ETL

DATA WAREHOUSE

Star schemas, views
other read-optimized structures

BI AND ANALYTCIS

Emailed, centrally stored Excel reports and dashboards

INCREASING DATA VOLUME

NON-RELATIONAL DATA

INCREASE IN TIME

STALE REPORTING

Increase in variety of data sources

Increase in data volume

Increase in types of data

Pressure on the ingestion engine

Complex, rigid transformations can't longer keep pace

Monitoring is abandoned

Delay in data, inability to transform volumes, or react to new sources

Repair, adjust and redesign ETL

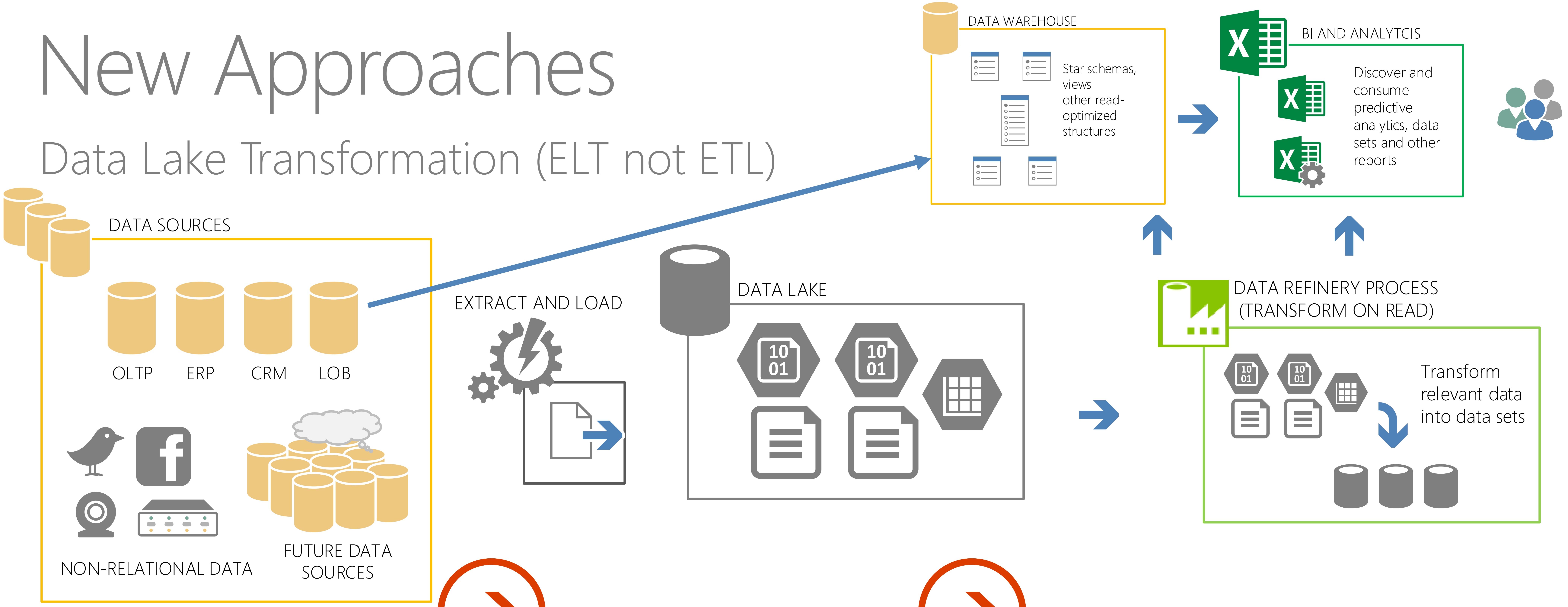
Reports become invalid or unusable

Delay in preserved reports increases

Users begin to "innovate" to relieve starvation

New Approaches

Data Lake Transformation (ELT not ETL)


DATA SOURCES
OLTP
ERP
CRM
LOB
NON-RELATIONAL DATA
FUTURE DATA SOURCES
EXTRACT AND LOAD
DATA LAKE
10 01
DATA WAREHOUSE
Star schemas, views other read-optimized structures
BI AND ANALYTCIS
Discover and consume predictive analytics, data sets and other reports
DATA REFINERY PROCESS (TRANSFORM ON READ)
Transform relevant data into data sets

All data sources are considered

Leverages the power of on-prem technologies and the cloud for storage and capture

Native formats, streaming data, big data

Extract and load, no/minimal transform

Storage of data in near-native format

Orchestration becomes possible

Streaming data accommodation becomes possible

Refineries transform data on read

Produce curated data sets to integrate with traditional warehouses

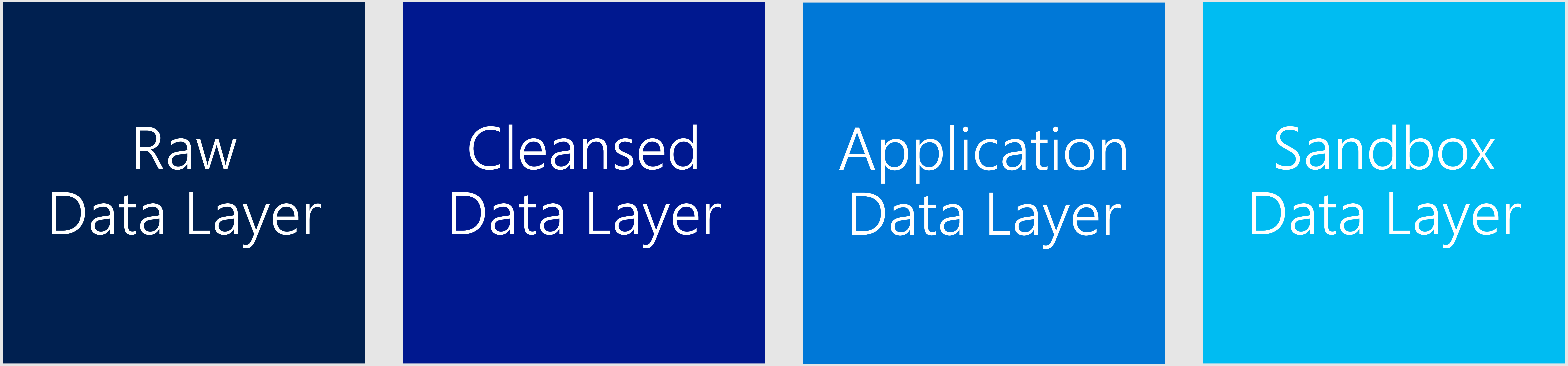
Users discover published data sets/services using familiar tools

Data Analysis Paradigm Shift

OLD WAY: Structure -> Ingest -> Analyze

NEW WAY: Ingest -> Analyze -> Structure

Data Lake Layers



Raw Data Layer

Cleansed Data Layer

Application Data Layer

Sandbox Data Layer

Needs data governance so your data lake does not turn into a data swamp!

Organizing a Data Lake – Folder structure

Objectives

✓ Plan the structure based on optimal data retrieval

✓ Avoid a chaotic, unorganized data swamp

Special thanks to:
Melissa Coates
CoatesDataStrategies.com

Common ways to organize the data:

Time Partitioning

Year/Month/Day/Hour/Minute

Subject Area

Security Boundaries

Department
Business unit
etc…

Downstream App/Purpose

Data Retention Policy

Temporary data
Permanent data
Applicable period (ex: project lifetime)
etc…

Business Impact / Criticality

High (HBI)
Medium (MBI)
Low (LBI)
etc…

Owner / Steward / SME

Probability of Data Access

Recent/current data
Historical data
etc…

Confidential Classification

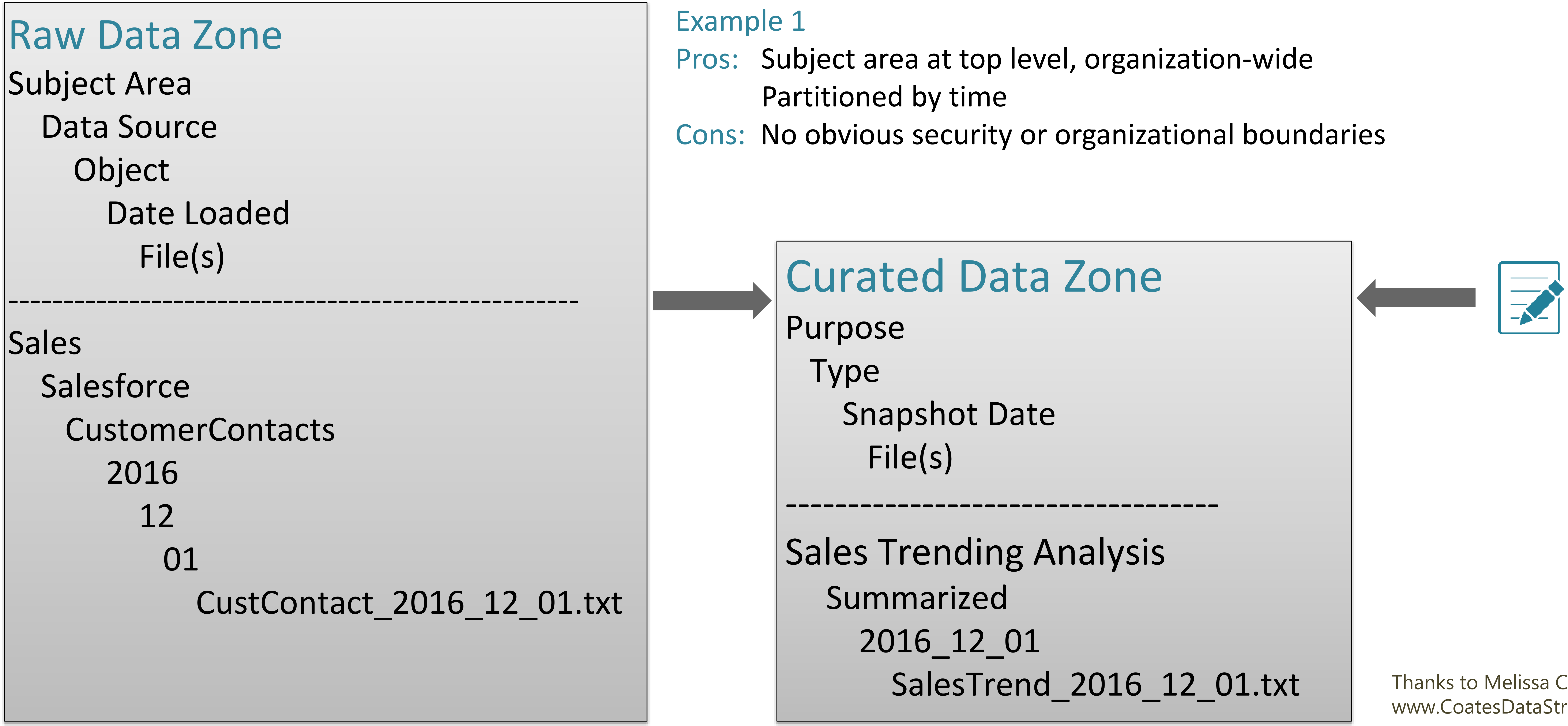
Public information
Internal use only
Supplier/partner confidential
Personally identifiable information (PII)
Sensitive – financial
Sensitive – intellectual property
etc…

Organizing a Data Lake

Example 1

Pros: Subject area at top level, organization-wide
Partitioned by time

Cons: No obvious security or organizational boundaries



Raw Data Zone

Subject Area
Data Source
Object
Date Loaded
File(s)

Sales
Salesforce
CustomerContacts
2016
12
01
CustContact_2016_12_01.txt

Curated Data Zone

Purpose
Type
Snapshot Date
File(s)

Sales Trending Analysis
Summarized
2016_12_01
SalesTrend_2016_12_01.txt

Thanks to Melissa Coates,
www.CoatesDataStrategies.com

Data Lake with DW use cases

Data Lake

Staging & preparation

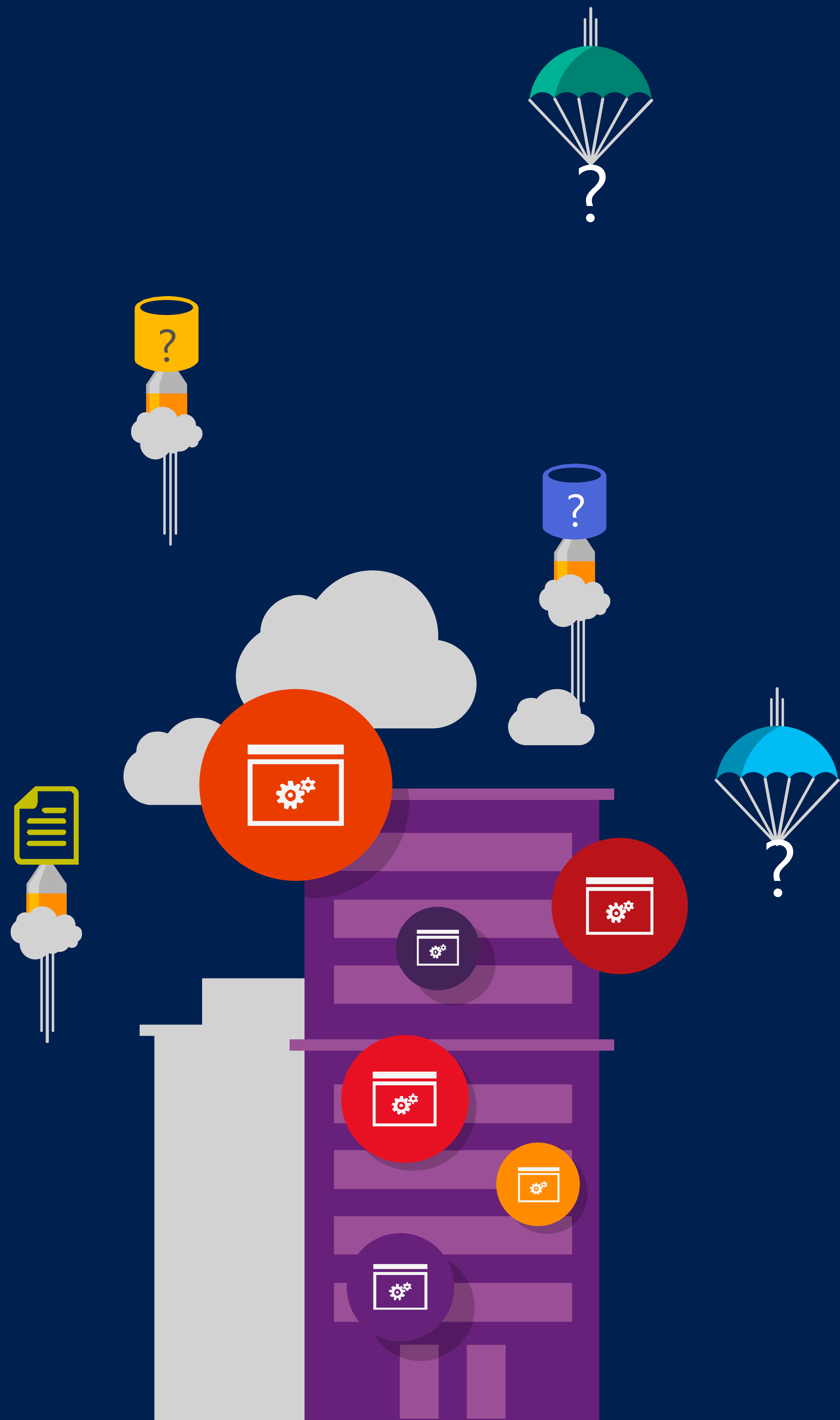
- Data scientists/Power users
- Batch processing
- Data refinement/cleaning
- ETL workloads
- Store older/backup data
- Sandbox for data exploration
- One-time reports
- Quick access to data
- Don't know questions

Data Warehouse

Serving, Security & Compliance

- Business people
- Low latency
- Complex joins
- Interactive ad-hoc query
- High number of users
- Additional security
- Large support for tools
- Dashboards
- Easily create reports (Self-service BI)
- Know questions

Creating ADLS Gen2



Azure Data Lake Storage Gen2

A "**no-compromises**" Data Lake: secure, performant, massively-scalable Data Lake storage that brings the cost and scale profile of object storage together with the performance and analytics feature set of data lake storage

SECURE

✓ Support for fine-grained ACLs, protecting data at the file and folder level

✓ Multi-layered protection via at-rest Storage Service encryption and Azure Active Directory integration

MANAGEABLE

✓ Automated Lifecycle Policy Management

✓ Object Level tiering

FAST

✓ Atomic file operations means jobs complete faster

SCALABLE

✓ No limits on data store size

✓ Global footprint (50 regions)

COST EFFECTIVE

✓ Object store pricing levels

✓ File system operations minimize transactions required for job completion

INTEGRATION READY

✓ Optimized for Spark and Hadoop Analytic Engines

✓ Tightly integrated with Azure end to end analytics solutions

Convergence of two Storage Services

Blob Storage

General Purpose Object Storage

- Large partner ecosystem
- Global scale – All 50 regions
- Durability options
- Tiered - Hot/Cool/Archive
- Cost Efficient

Data Lake Store

Optimized for Big Data analytics

- Built for Hadoop
- Hierarchical namespace
- ACLs, AAD and RBAC
- Performance tuned for big data
- Very high scale capacity and throughput

Azure Data Lake Storage Gen2

The best of Blobs and ADLS

- Large partner ecosystem
- Global scale – All 50 regions
- Durability options
- Tiered - Hot/Cool/Archive
- Cost Efficient

- Built for Hadoop
- Hierarchical namespace
- ACLs, AAD and RBAC
- Performance tuned for big data
- Very high scale capacity and throughput

Remaining known limitations with ADLS Gen2

Missing blob storage features:

- Archive and Premium tier
- Soft Delete
- Snapshots
- Some features in preview

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-known-issues

Missing ADLS Gen1 features:

- Microsoft product support: ADC, Excel, AAS
- 3rd-party products: Informatica, Attunity, Alteryx
- Some features in preview

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-upgrade

Azure Data Lake Store – Distributed File System

Files of any size can be stored because ADLS is a distributed system which file contents are divided up across backend storage nodes.

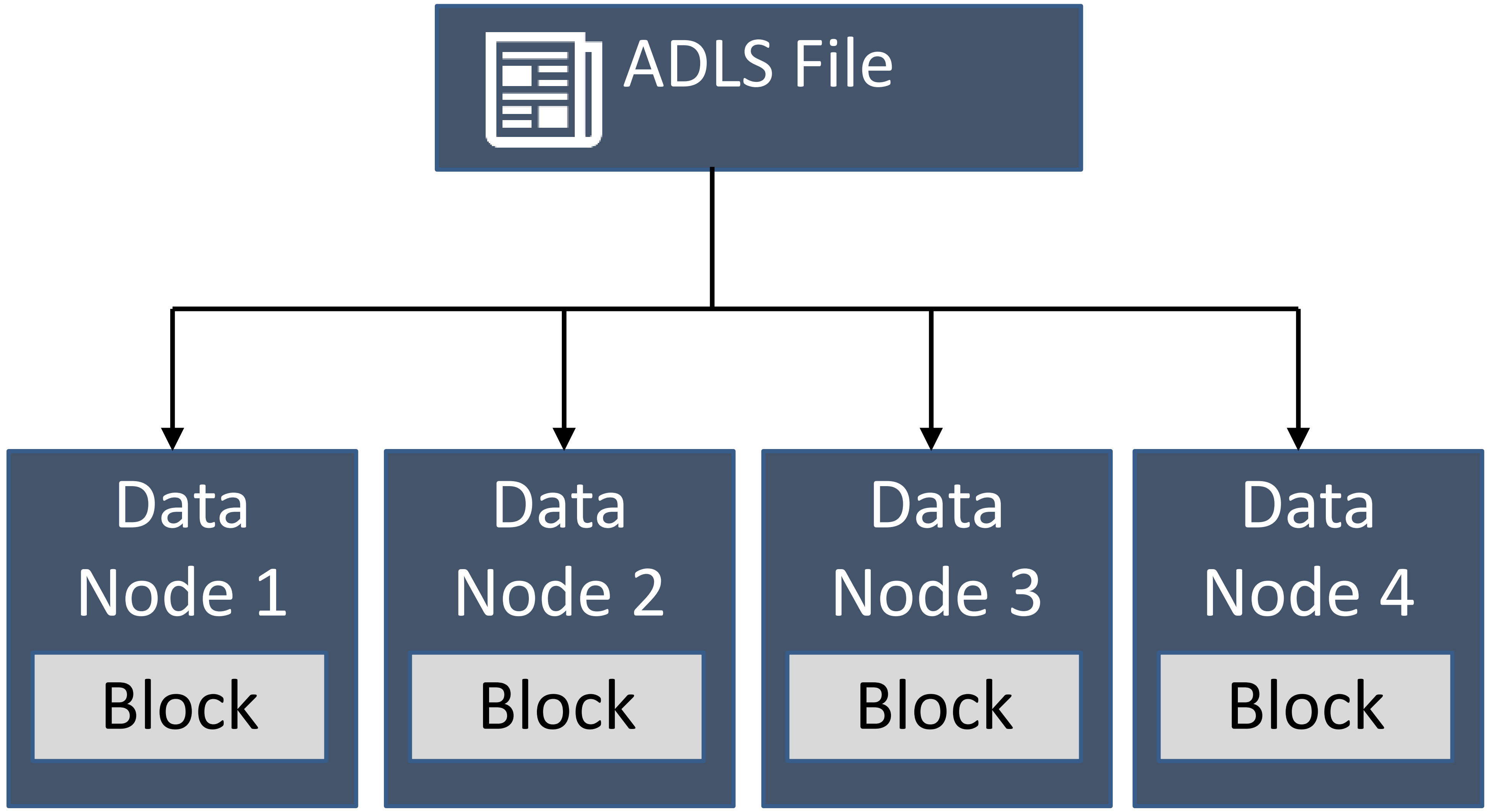
A read operation on the file is also parallelized across the nodes.

Blocks are also replicated for fault tolerance.

The ideal file size in ADLS is 256MB – 2GB in size.

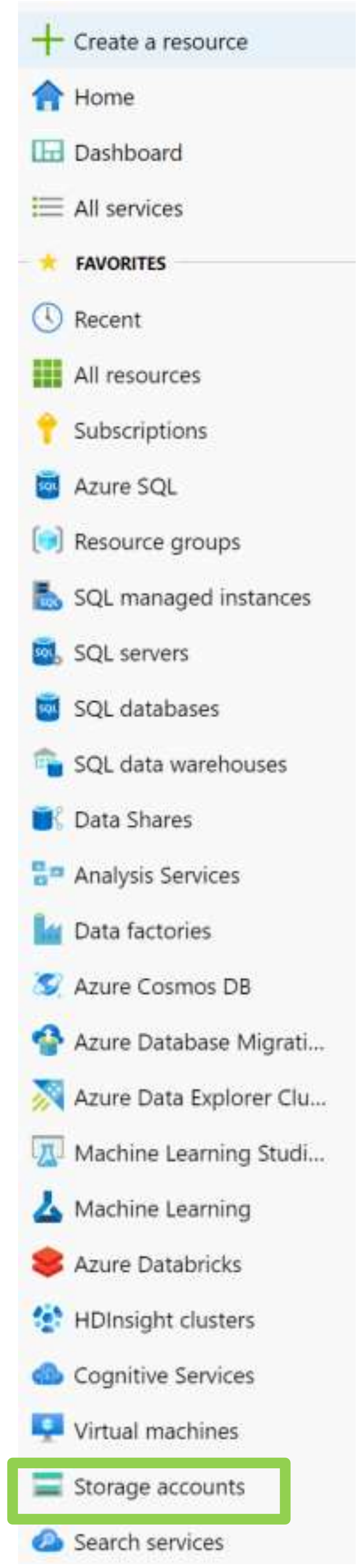
Many very tiny files introduces significant overhead which reduces performance. This is a well-known issue with storing data in HDFS. Techniques:

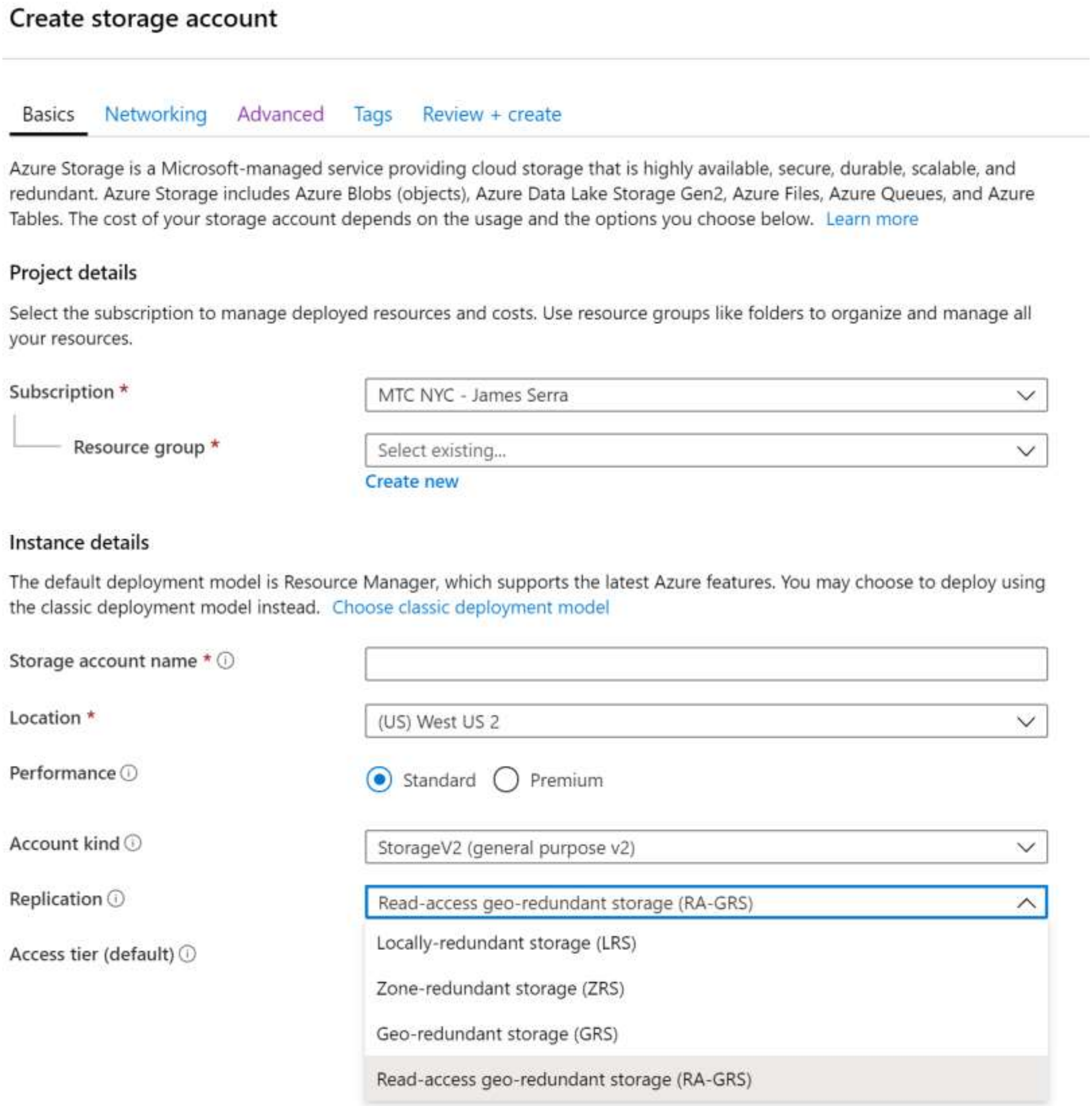
- Append-only data streams
- Consolidation of data into larger files



Thanks to Melissa Coates, www.CoatesDataStrategies.com

Create ADLS Gen2





Create storage account

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more

Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription * MTC NYC - James Serra

Resource group * Select existing...

Create new

Instance details

The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. Choose classic deployment model

Storage account name *

Location * (US) West US 2

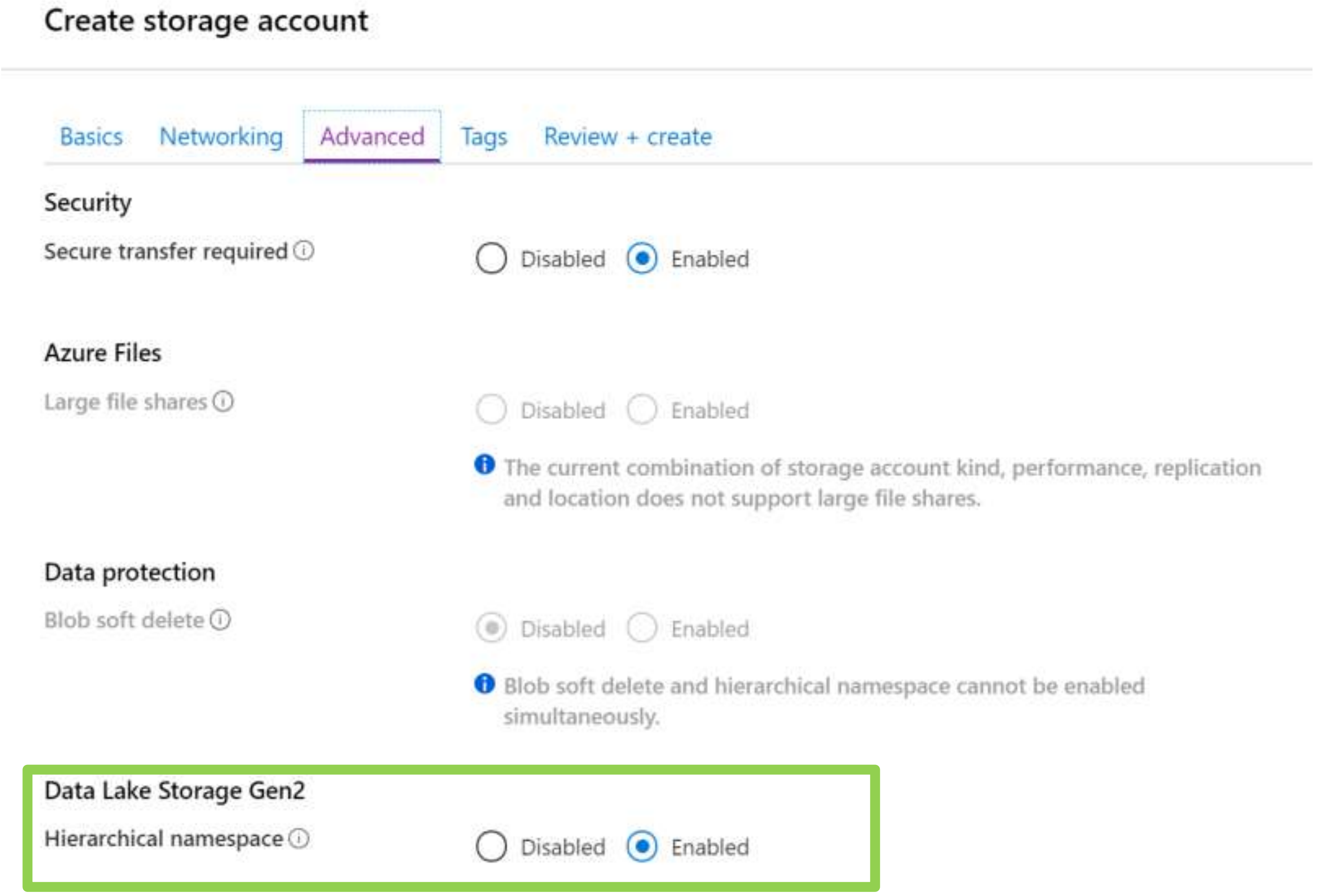
Performance ● Standard ○ Premium

Account kind StorageV2 (general purpose v2)

Replication Read-access geo-redundant storage (RA-GRS)

Locally-redundant storage (LRS)
Zone-redundant storage (ZRS)
Geo-redundant storage (GRS)
Read-access geo-redundant storage (RA-GRS)

Access tier (default)



Create storage account

Security

Secure transfer required ○ Disabled ● Enabled

Azure Files

Large file shares ○ Disabled ○ Enabled

The current combination of storage account kind, performance, replication and location does not support large file shares.

Data protection

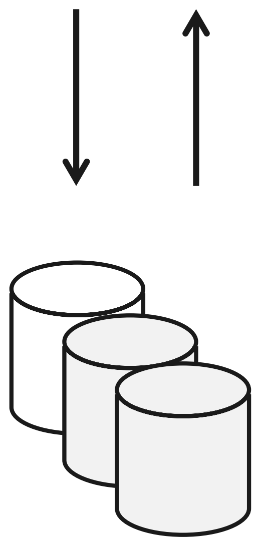
Blob soft delete ● Disabled ○ Enabled

Blob soft delete and hierarchical namespace cannot be enabled simultaneously.

Data Lake Storage Gen2

Hierarchical namespace ○ Disabled ● Enabled

ADLS Gen2 Replication Options



LRS

Multiple replicas across a datacenter

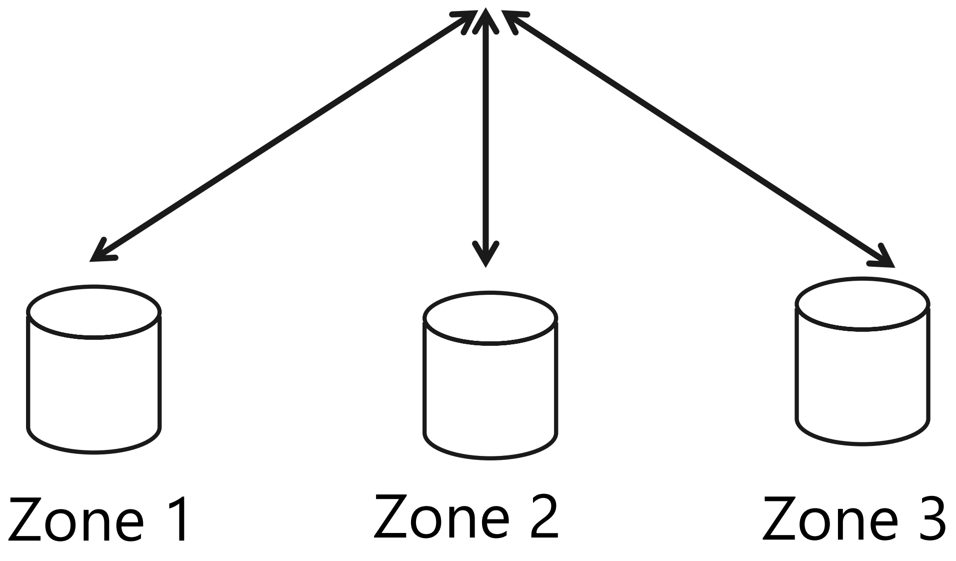
Protect against disk, node, rack failures

Write is ack'd when all replicas are committed

Superior to dual-parity RAID

11 9s of durability

SLA: 99.9%



ZRS

Replicas across 3 Zones

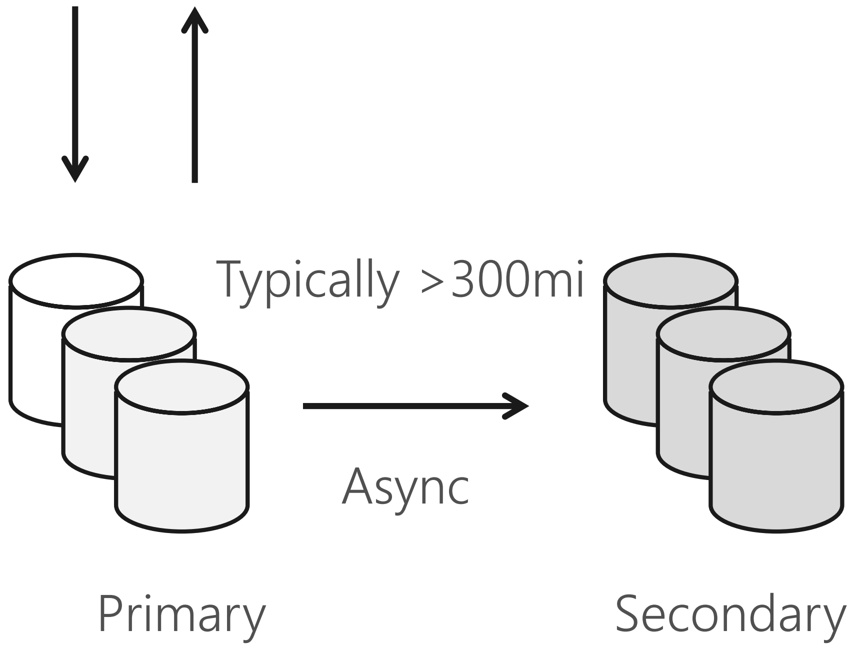
Protect against disk, node, rack and zone failures

Synchronous writes to all 3 zones

12 9s of durability

Available in 8 regions

SLA: 99.9%



GRS

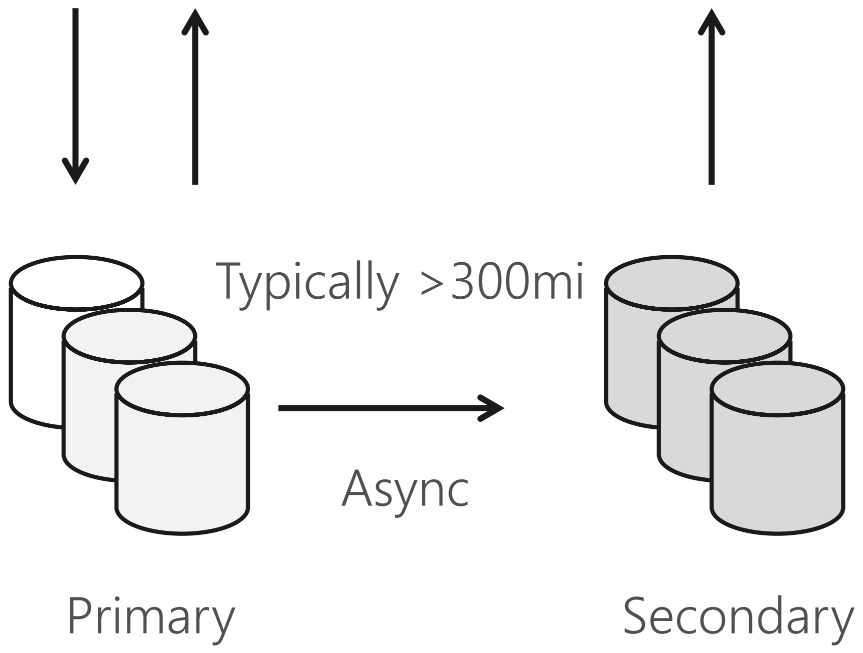
Multiple replicas across each of 2 regions

Protects against major regional disasters

Asynchronous to secondary

16 9s of durability

SLA: 99.9%



RA-GRS

GRS + Read access to secondary

Separate secondary endpoint

RPO delay to secondary can be queried

SLA: 99.99% (read), 99.9% (write)

Preview: Customer controlled failover to GRS location
Geo-zone-redundant storage (GZRS, RA-GZRS)

Data Transport Methods

File Sync

- Windows Srv <-> Azure
- Local caching
- With offline (Databox) can 'sync' remainder

Fuse

- Mount blobs as local FS
- Commit on write
- Linux

Site Replication

- On premise & cloud
- Windows, Linux
- Physical, virtual
- Hyper-V, VMWare

Network Acceleration

- Aspera
- Signiant

AZCopy

- Throughput +30%
- S3 to Azure Blobs
- Sync to cloud
- Hi Latency 10-100%

NetApp

- CloudSync
- SnapMirror
- SnapVault

Data Factory

- On premise & cloud sources
- Structured & unstructured
- Over 60 connectors
- UI design data flow

Partners

- Peer Global File Service
- Talon FAST
- Zerto
- …

Offline

- Data Box
- Data Box Heavy
- Data Box Disk
- Disk Import / Export

Fast Data Transfer
microsoft.com/en-us/garage/profiles/fast-data-transfer/

Azure Data Box Family

Offline Data Transfer

Data Box

- Capacity: 100 TB
- Weight: ~50 lbs
- Secure, ruggedized appliance

- Data Box enables bulk migration to Azure when network isn't an option.

Data Box Disk

- Capacity: 8TB ea.; 40TB/order
- Secure, ruggedized USB drives orderable in packs of 5 (up to 40TB).

- Perfect for projects that require a smaller form factor, e.g., autonomous vehicles.

Data Box Heavy PREVIEW

- Capacity: 1 PB
- Weight 500+ lbs
- Secure, ruggedized appliance

- Same service as Data Box, but targeted to petabyte-sized datasets.

Online Data Transfer

Data Box Gateway

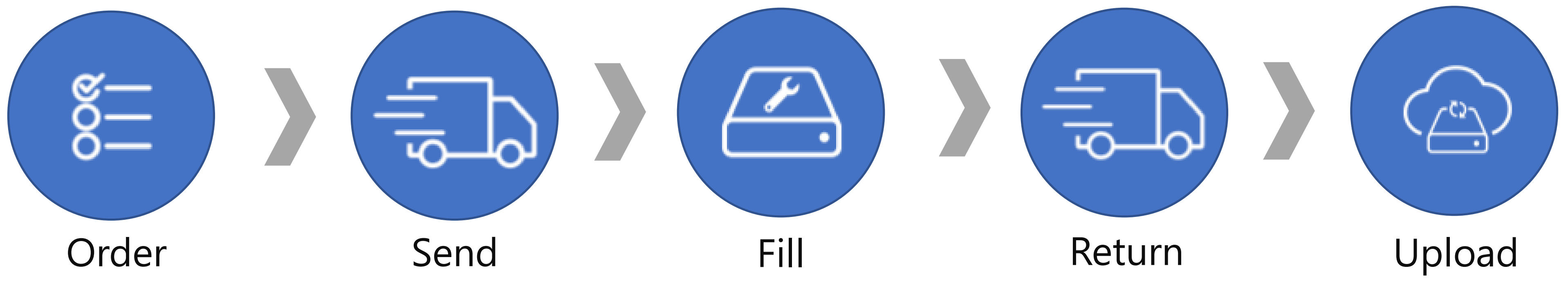
- Virtual device provisioned in your hypervisor
- Supports storage gateway, SMB, NFS, Azure blob, files

- Virtual network transfer appliance (VM), runs on your choice of hardware.

Data Box Edge

- Local Cache Capacity: ~12 TB
- Includes **Data Box Gateway** and **Azure IoT Edge**.

- Data Box Edge manages uploads to Azure and can pre-process data prior to upload.



Network Data Transfer

Cloud to Edge

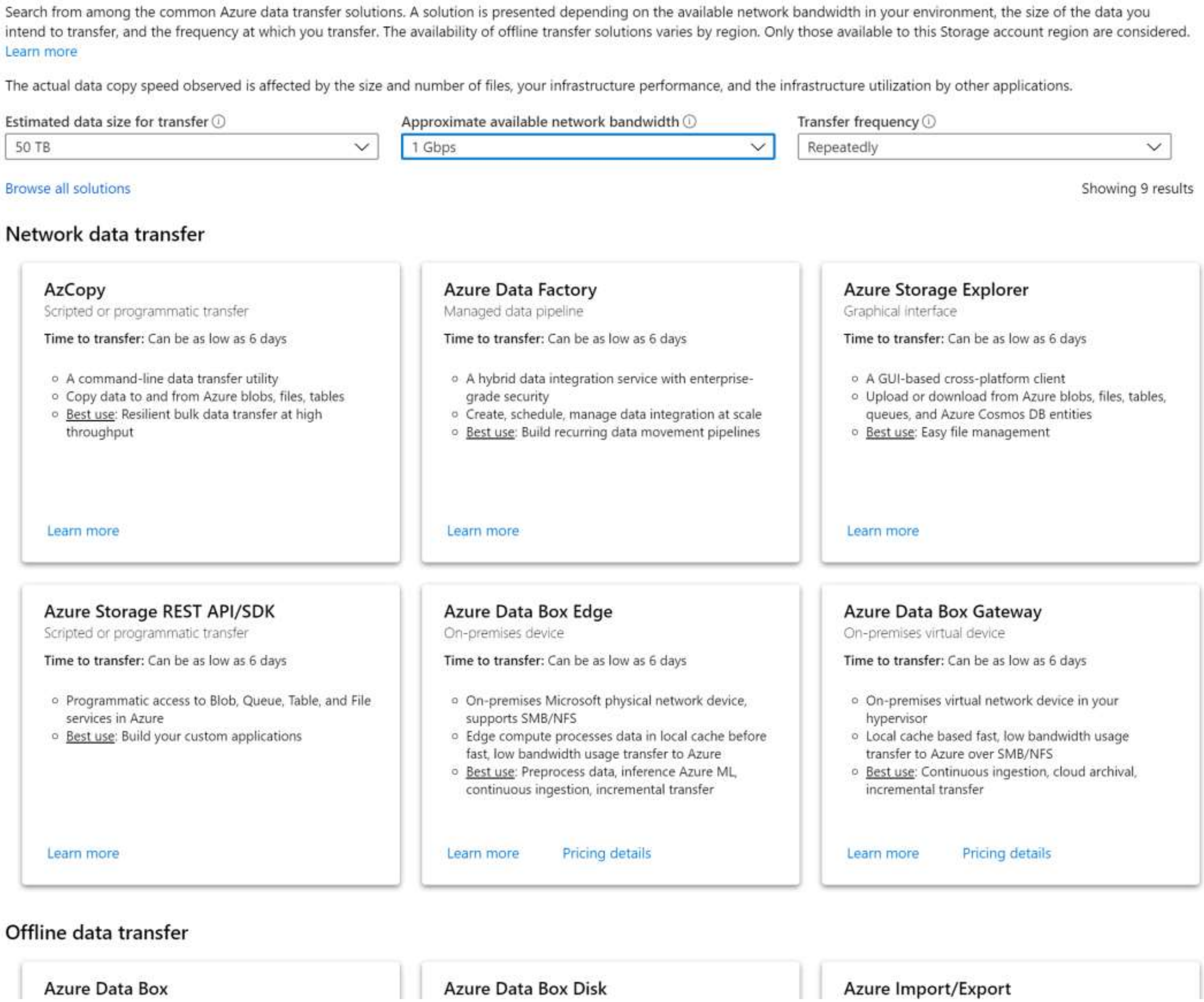
Edge to Cloud

Edge Compute

Pre-processing

ML Inferencing

Data transfer



Search from among the common Azure data transfer solutions. A solution is presented depending on the available network bandwidth in your environment, the size of the data you intend to transfer, and the frequency at which you transfer. The availability of offline transfer solutions varies by region. Only those available to this Storage account region are considered. Learn more

The actual data copy speed observed is affected by the size and number of files, your infrastructure performance, and the infrastructure utilization by other applications.

Estimated data size for transfer ⓘ: 50 TB

Approximate available network bandwidth ⓘ: 1 Gbps

Transfer frequency ⓘ: Repeatedly

Browse all solutions

Showing 9 results

Network data transfer

AzCopy
Scripted or programmatic transfer

Time to transfer: Can be as low as 6 days

- A command-line data transfer utility
- Copy data to and from Azure blobs, files, tables
- Best use: Resilient bulk data transfer at high throughput

Learn more

Azure Data Factory
Managed data pipeline

Time to transfer: Can be as low as 6 days

- A hybrid data integration service with enterprise-grade security
- Create, schedule, manage data integration at scale
- Best use: Build recurring data movement pipelines

Learn more

Azure Storage Explorer
Graphical interface

Time to transfer: Can be as low as 6 days

- A GUI-based cross-platform client
- Upload or download from Azure blobs, files, tables, queues, and Azure Cosmos DB entities
- Best use: Easy file management

Learn more

Azure Storage REST API/SDK
Scripted or programmatic transfer

Time to transfer: Can be as low as 6 days

- Programmatic access to Blob, Queue, Table, and File services in Azure
- Best use: Build your custom applications

Learn more

Azure Data Box Edge
On-premises device

Time to transfer: Can be as low as 6 days

- On-premises Microsoft physical network device, supports SMB/NFS
- Edge compute processes data in local cache before fast, low bandwidth usage transfer to Azure
- Best use: Preprocess data, inference Azure ML, continuous ingestion, incremental transfer

Learn more Pricing details

Azure Data Box Gateway
On-premises virtual device

Time to transfer: Can be as low as 6 days

- On-premises virtual network device in your hypervisor
- Local cache based fast, low bandwidth usage transfer to Azure over SMB/NFS
- Best use: Continuous ingestion, cloud archival, incremental transfer

Learn more Pricing details

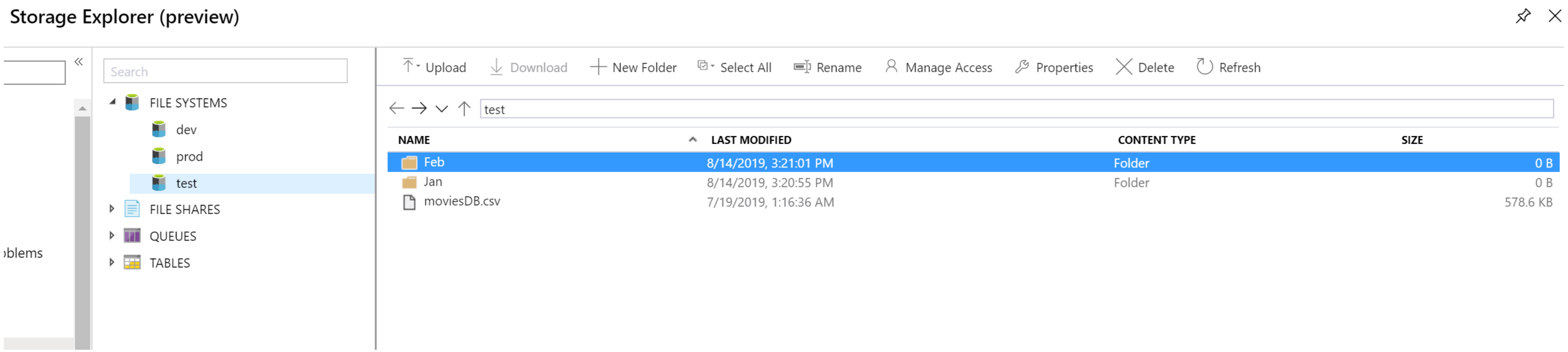
Offline data transfer

Azure Data Box

Azure Data Box Disk

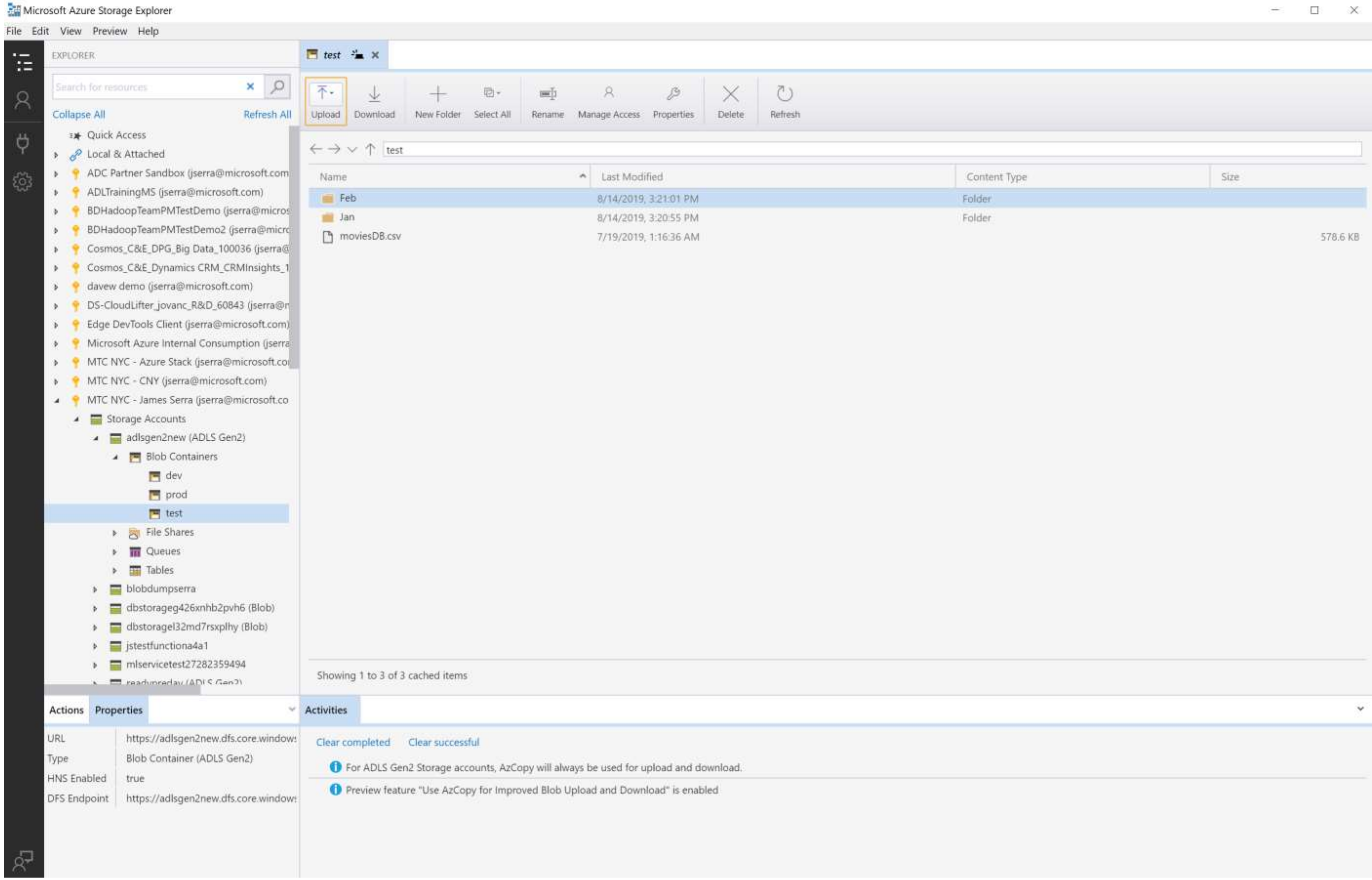
Azure Import/Export

Storage Explorer (preview)



NAME	LAST MODIFIED	CONTENT TYPE	SIZE
Feb	8/14/2019, 3:21:01 PM	Folder	0 B
Jan	8/14/2019, 3:20:55 PM	Folder	0 B
moviesDB.csv	7/19/2019, 1:16:36 AM		578.6 KB

Microsoft Azure Storage Explorer



- Quick Access
- Local & Attached
- ADC Partner Sandbox (jserra@microsoft.com
- ADLTrainingMS (jserra@microsoft.com)
- BDHadoopTeamPMTestDemo (jserra@micros
- BDHadoopTeamPMTestDemo2 (jserra@micro
- Cosmos_C&E_DPG_Big Data_100036 (jserra@
- Cosmos_C&E_Dynamics CRM_CRMInsights_1
- davew demo (jserra@microsoft.com)
- DS-CloudLifter_jovanc_R&D_60843 (jserra@r
- Edge DevTools Client (jserra@microsoft.com)
- Microsoft Azure Internal Consumption (jserra
- MTC NYC - Azure Stack (jserra@microsoft.co
- MTC NYC - CNY (jserra@microsoft.com)
- MTC NYC - James Serra (jserra@microsoft.co
  - Storage Accounts
    - adlsgen2new (ADLS Gen2)
      - Blob Containers
        - dev
        - prod
        - test
      - File Shares
      - Queues
      - Tables
    - blobdumpserra
    - dbstorageg426xnhb2pvh6 (Blob)
    - dbstoragel32md7rsxplhy (Blob)
    - jstestfunctiona4a1
    - mlservicetest27282359494
    - readypreday (ADLS Gen2)

Name	Last Modified	Content Type	Size
Feb	8/14/2019, 3:21:01 PM	Folder	
Jan	8/14/2019, 3:20:55 PM	Folder	
moviesDB.csv	7/19/2019, 1:16:36 AM		578.6 KB

Showing 1 to 3 of 3 cached items

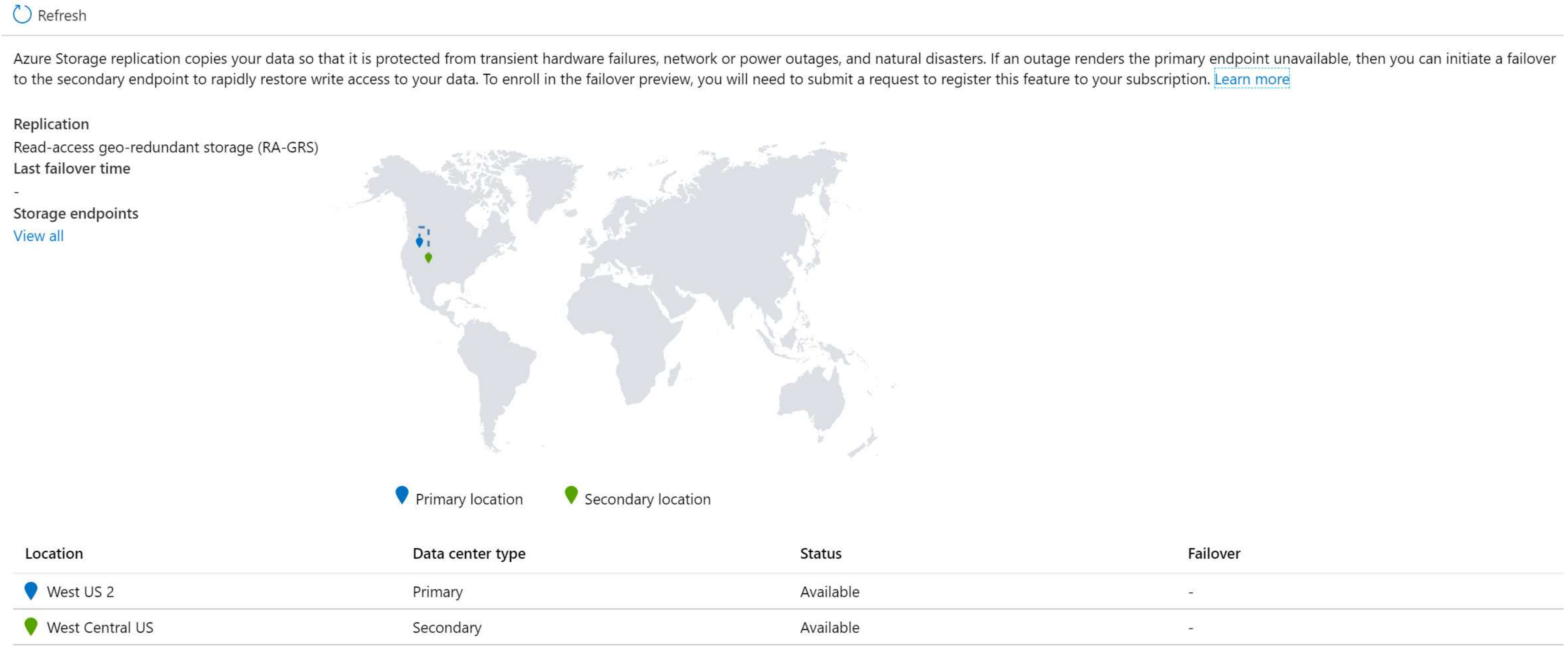
URL	https://adlsgen2new.dfs.core.window:
Type	Blob Container (ADLS Gen2)
HNS Enabled	true
DFS Endpoint	https://adlsgen2new.dfs.core.window:

For ADLS Gen2 Storage accounts, AzCopy will always be used for upload and download.

Preview feature "Use AzCopy for Improved Blob Upload and Download" is enabled

https://azure.microsoft.com/en-us/features/storage-explorer/

Geo-replication



Azure Storage replication copies your data so that it is protected from transient hardware failures, network or power outages, and natural disasters. If an outage renders the primary endpoint unavailable, then you can initiate a failover to the secondary endpoint to rapidly restore write access to your data. To enroll in the failover preview, you will need to submit a request to register this feature to your subscription. Learn more

Replication

Read-access geo-redundant storage (RA-GRS)

Last failover time

-

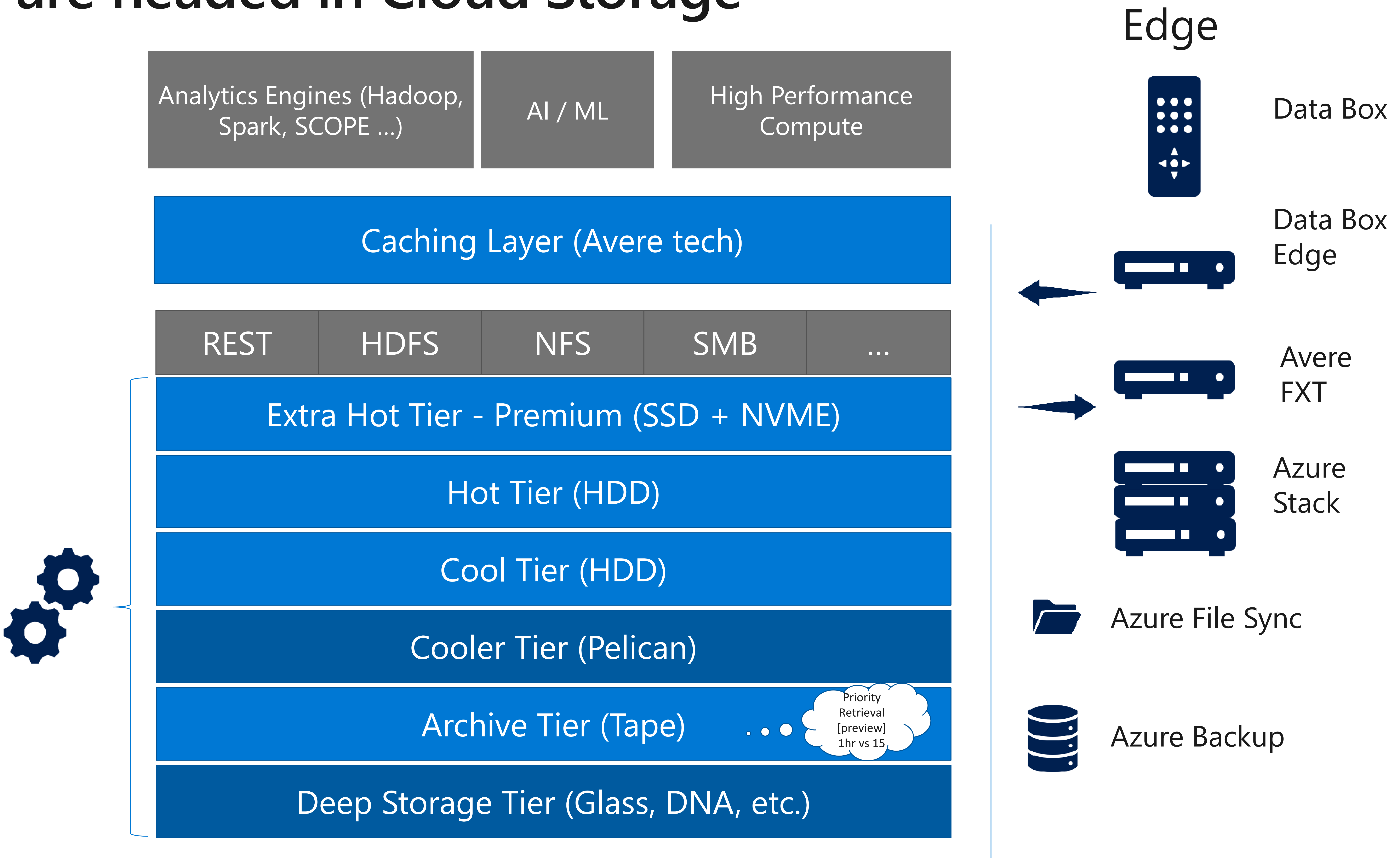
Storage endpoints

View all

● Primary location ● Secondary location

Location	Data center type	Status	Failover
● West US 2	Primary	Available	-
● West Central US	Secondary	Available	-

Where we are headed in Cloud Storage



Edge

Analytics Engines (Hadoop, Spark, SCOPE ...)

AI / ML

High Performance Compute

Caching Layer (Avere tech)

REST | HDFS | NFS | SMB | ...

Extra Hot Tier - Premium (SSD + NVME)

Hot Tier (HDD)

Cool Tier (HDD)

Cooler Tier (Pelican)

Archive Tier (Tape)

Priority Retrieval [preview] 1hr vs 15

Deep Storage Tier (Glass, DNA, etc.)

Data Box

Data Box Edge

Avere FXT

Azure Stack

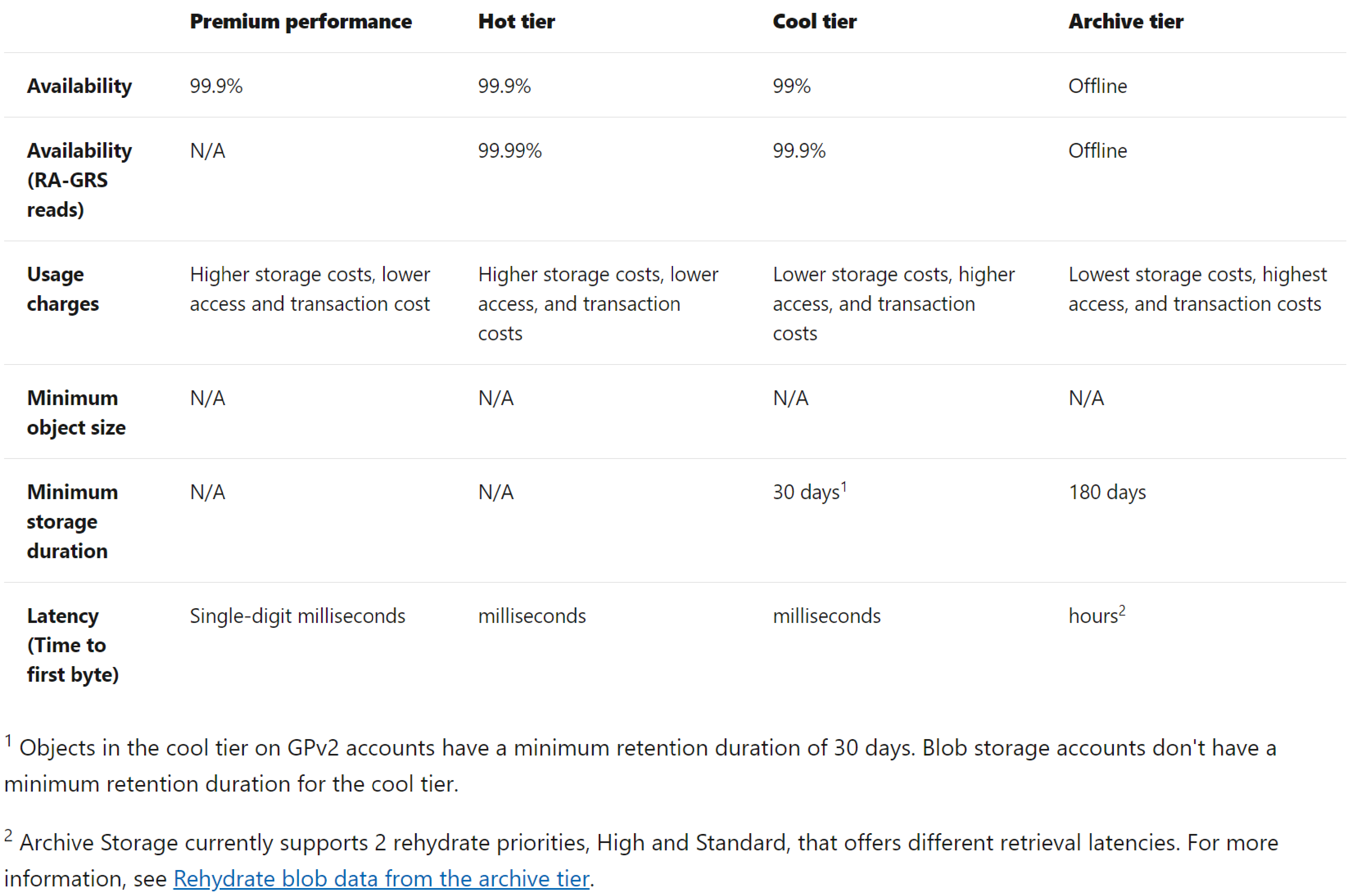
Azure File Sync

Azure Backup

Automatic Lifecycle Management

Current ● Future ●

Comparing storage options

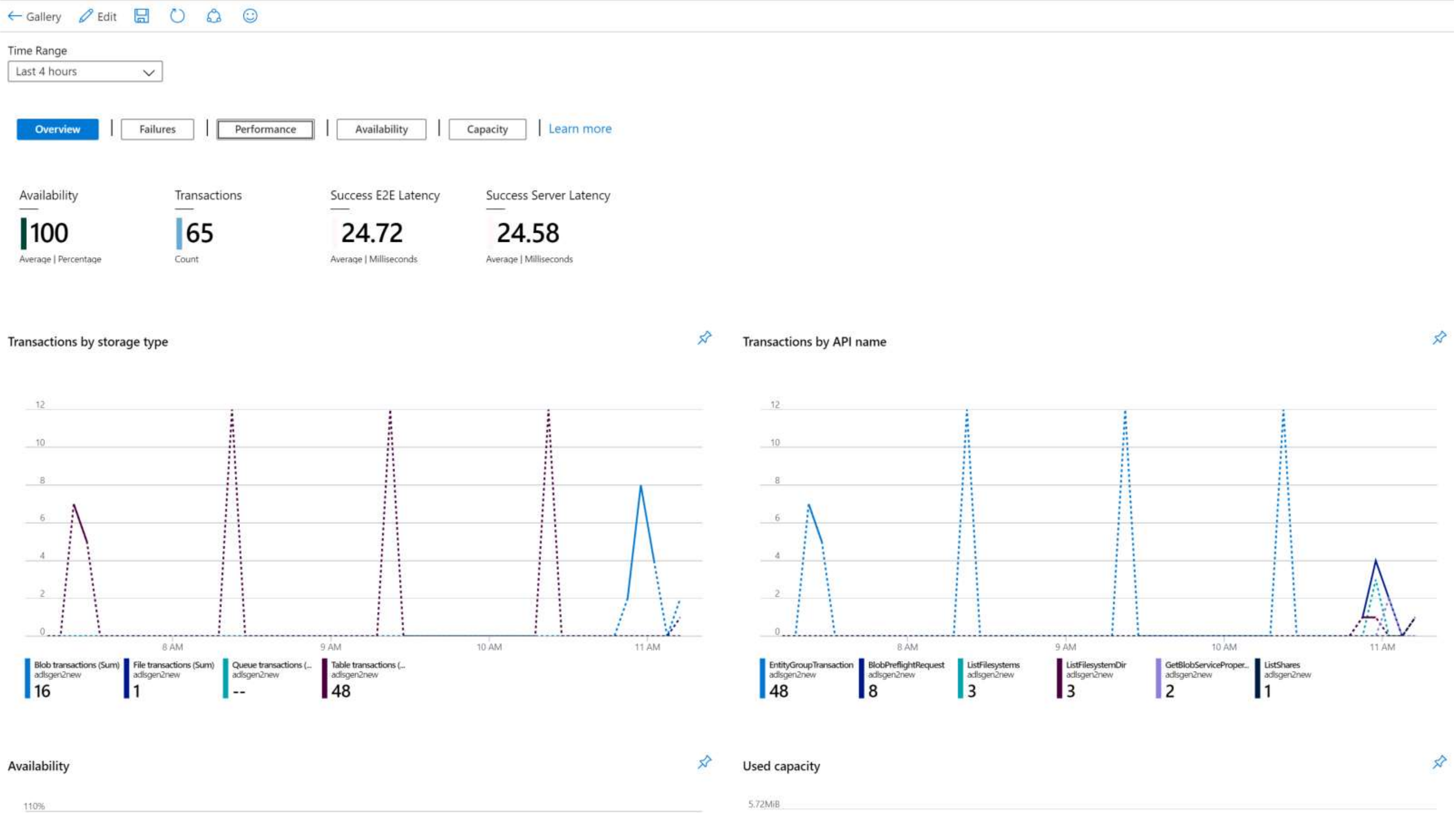


	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum object size	N/A	N/A	N/A	N/A
Minimum storage duration	N/A	N/A	30 days[1]	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours[2]

[1] Objects in the cool tier on GPv2 accounts have a minimum retention duration of 30 days. Blob storage accounts don't have a minimum retention duration for the cool tier.

[2] Archive Storage currently supports 2 rehydrate priorities, High and Standard, that offers different retrieval latencies. For more information, see Rehydrate blob data from the archive tier.

Insights (preview)



← Gallery ✎ Edit

Time Range

Last 4 hours

Overview | Failures | Performance | Availability | Capacity | Learn more

Availability	Transactions	Success E2E Latency	Success Server Latency
100	65	24.72	24.58
Average | Percentage	Count	Average | Milliseconds	Average | Milliseconds

Transactions by storage type

Blob transactions (Sum) adlsgen2new	File transactions (Sum) adlsgen2new	Queue transactions (... adlsgen2new	Table transactions (... adlsgen2new
16	1	--	48

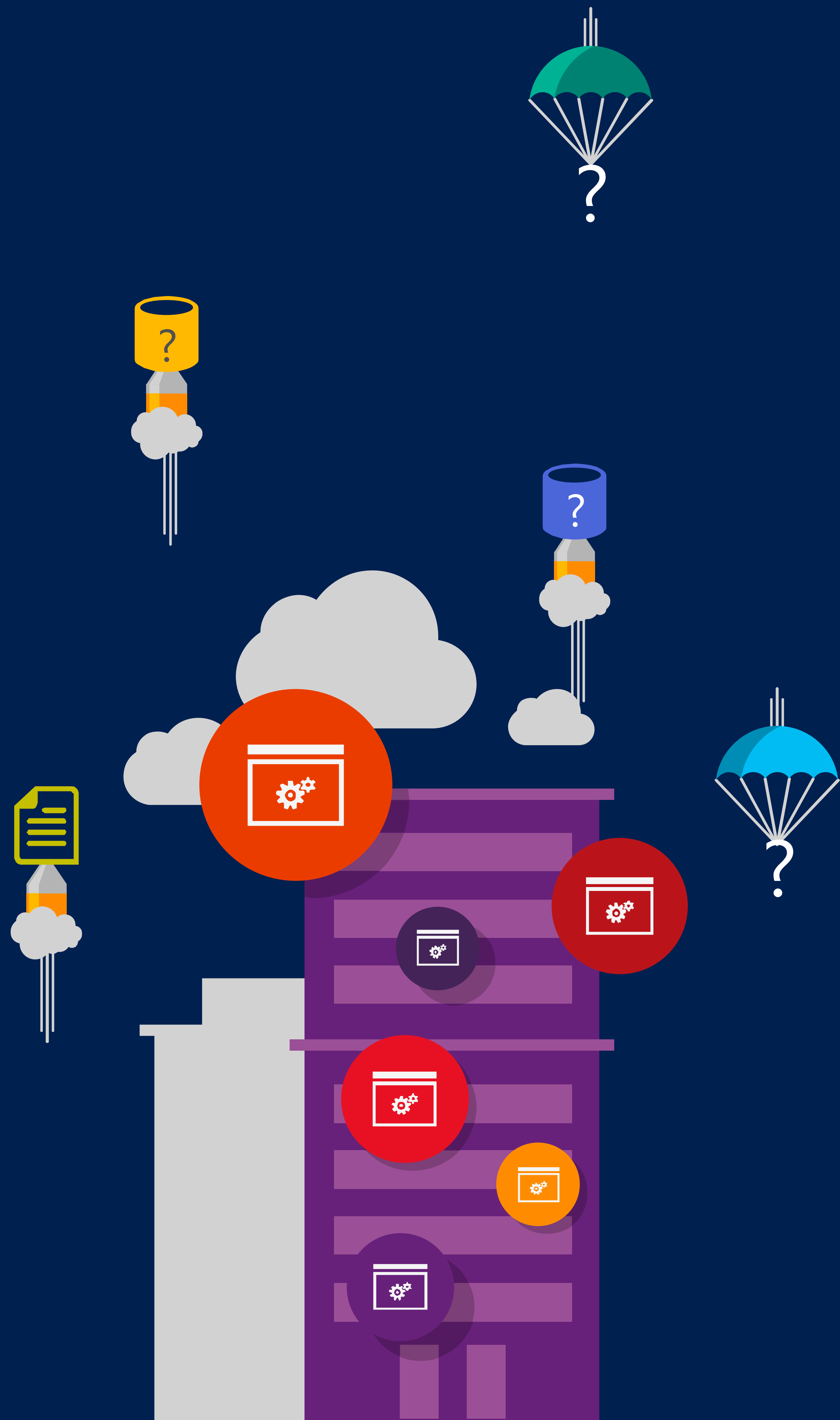
Transactions by API name

EntityGroupTransaction adlsgen2new	BlobPreflightRequest adlsgen2new	ListFilesystems adlsgen2new	ListFilesystemDir adlsgen2new	GetBlobServiceProper... adlsgen2new	ListShares adlsgen2new
48	8	3	3	2	1

Availability

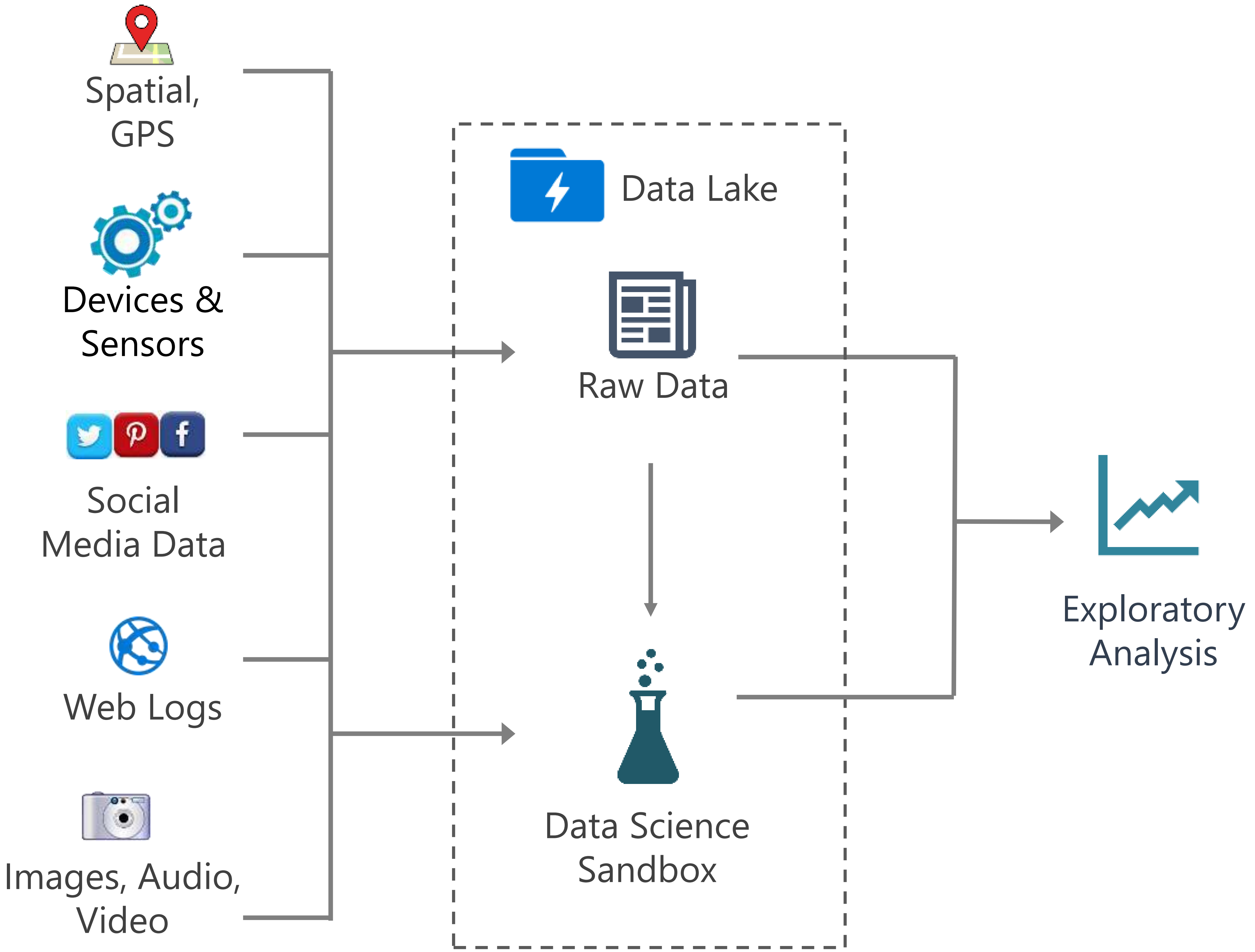
Used capacity

Data Lake Use Cases



Data Lake Use Cases

Ingestion of New File Types

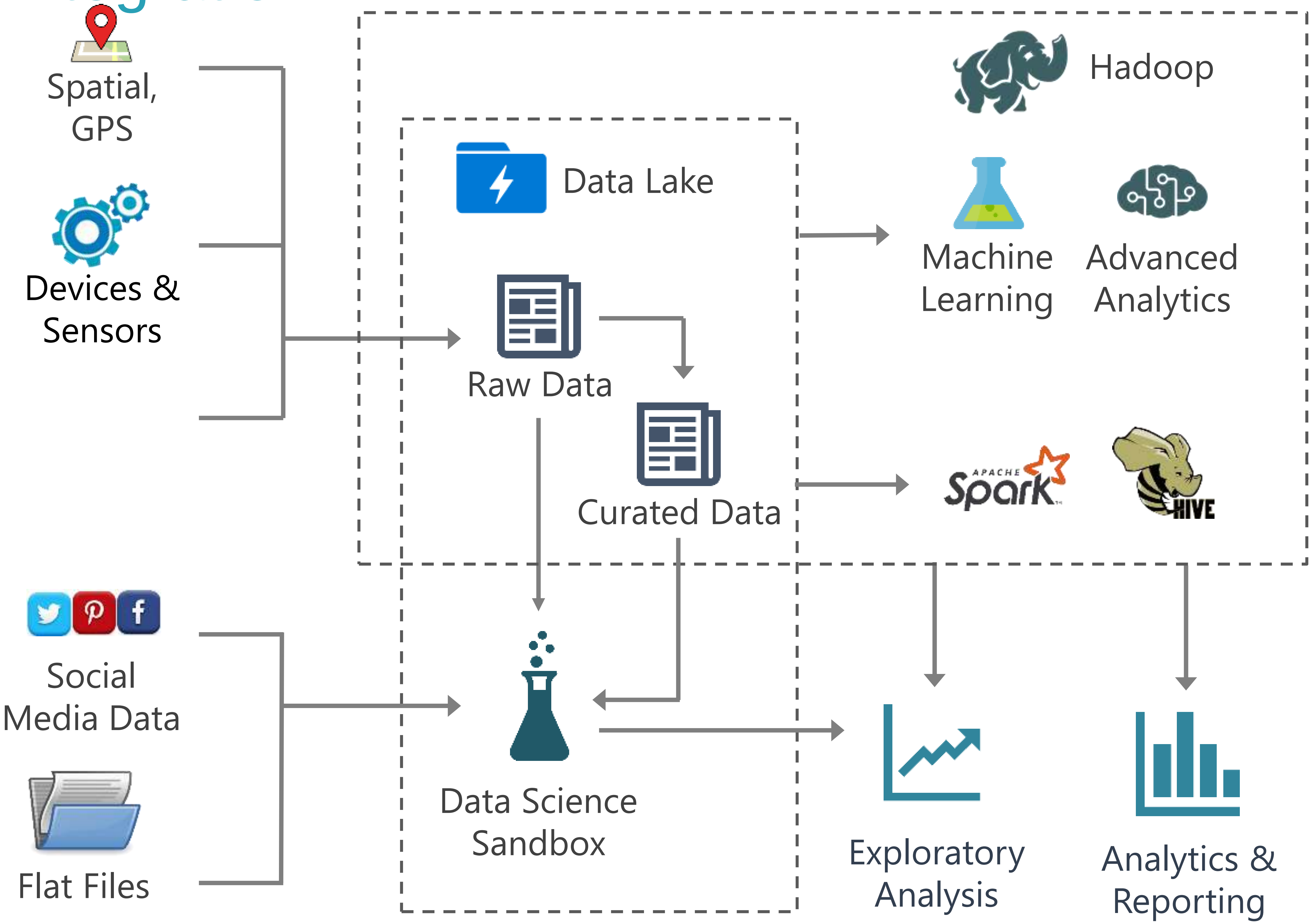


- ✓ Preparatory file storage for multi-structured data
- ✓ Exploratory analysis + POCs to determine value of new data types & sources
- ✓ Affords additional time for longer-term planning while accumulating data or handling an influx of data

Thanks to Melissa Coates,
www.CoatesDataStrategies.com

Data Lake Use Cases

Data Science Experimentation | Hadoop Integration

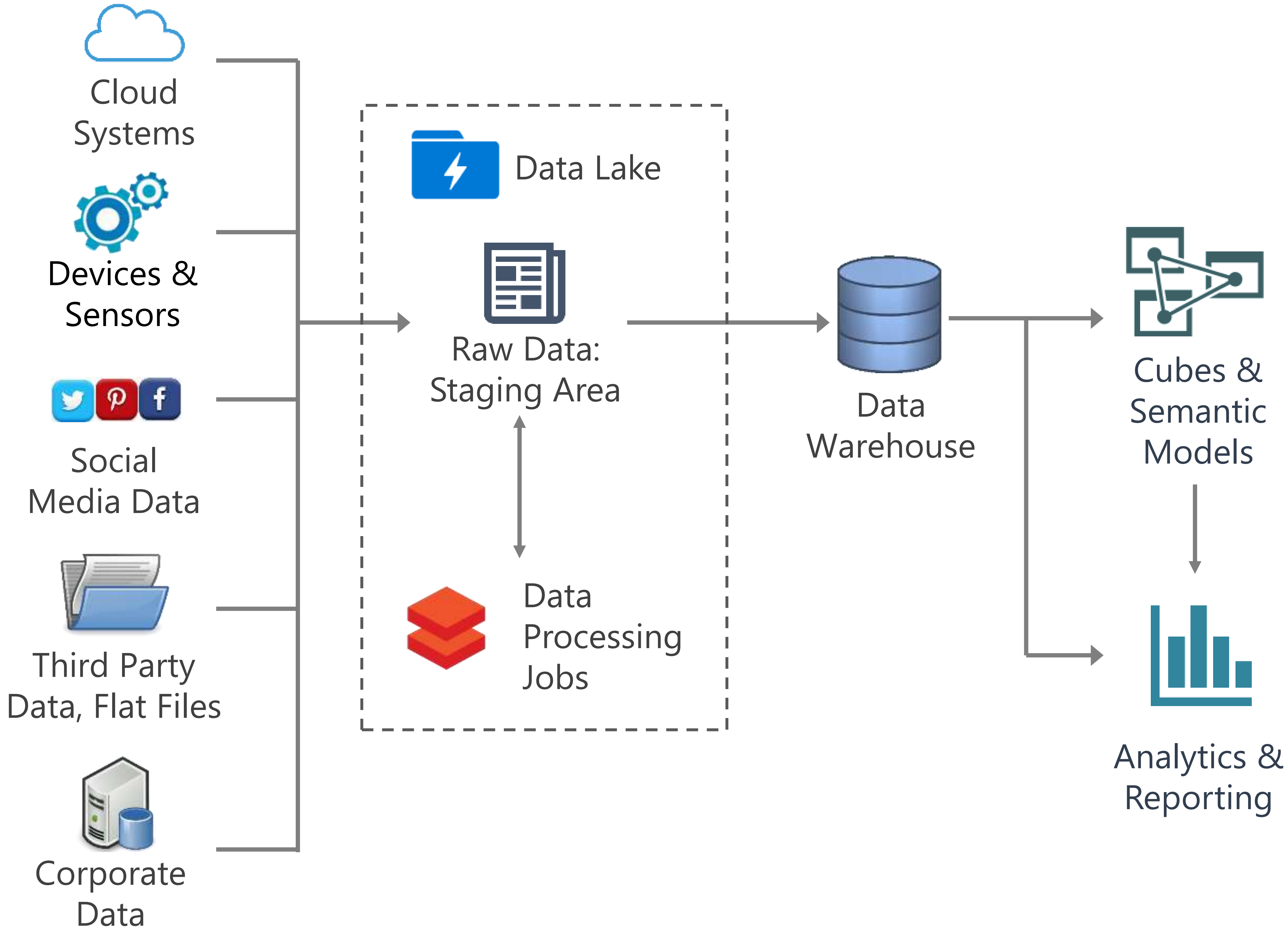


- ✓ Sandbox solutions for initial data prep, experimentation, and analysis
- ✓ Migrate from proof of concept to operationalized solution
- ✓ Integrate with open source projects such as Hive, Pig, Spark, Storm, etc.
- ✓ Big data clusters
- ✓ SQL-on-Hadoop solutions

Thanks to Melissa Coates, www.CoatesDataStrategies.com

Data Lake Use Cases

Data Warehouse Staging Area

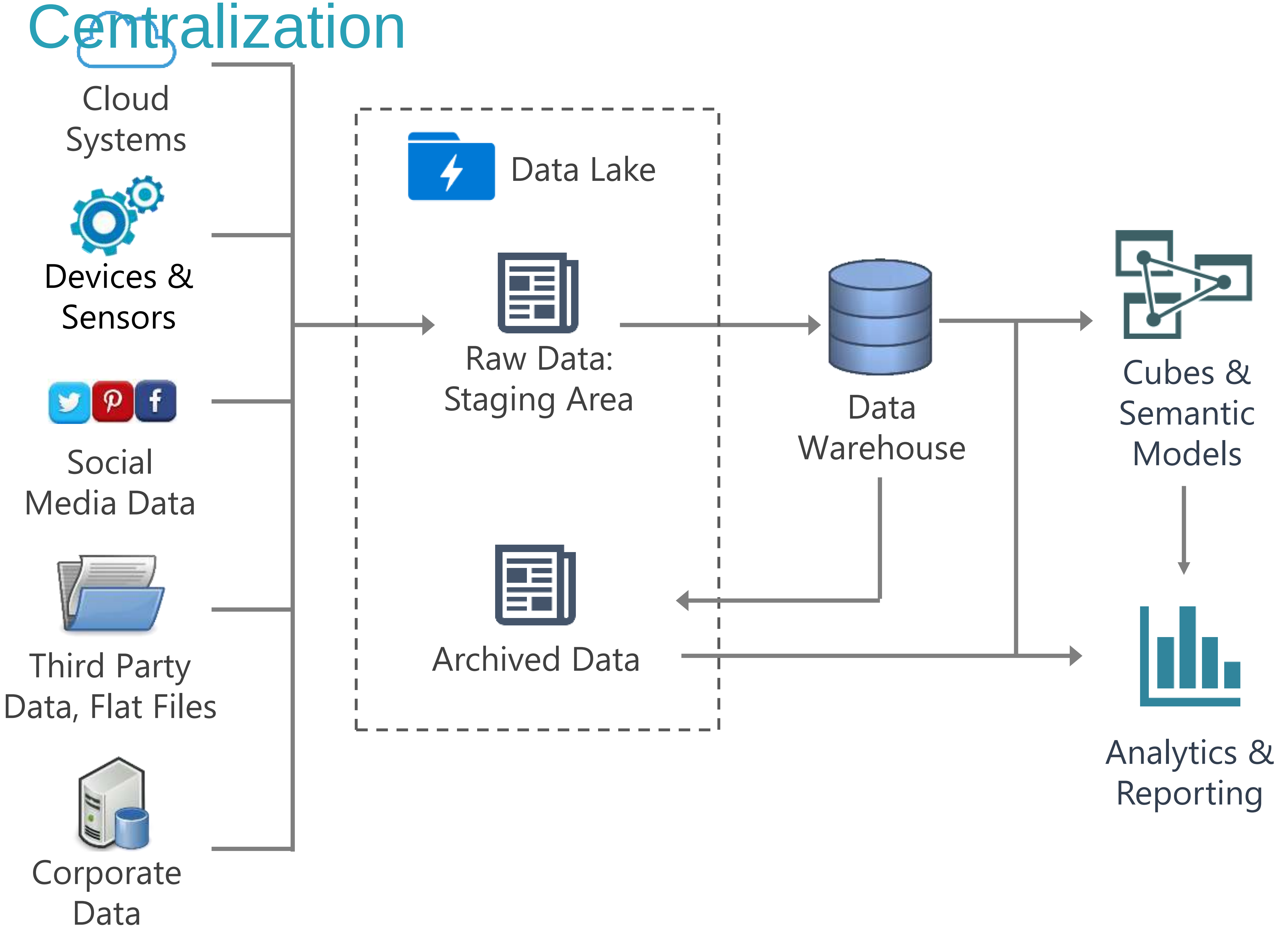


- ✓ ELT strategy
- ✓ Reduce storage needs in relational platform by using the data lake as landing area
- ✓ Practical use for data stored in the data lake
- ✓ Potentially also handle transformations in the data lake

Thanks to Melissa Coates,
www.CoatesDataStrategies.com

Data Lake Use Cases

Integration with DW | Data Archival | Centralization

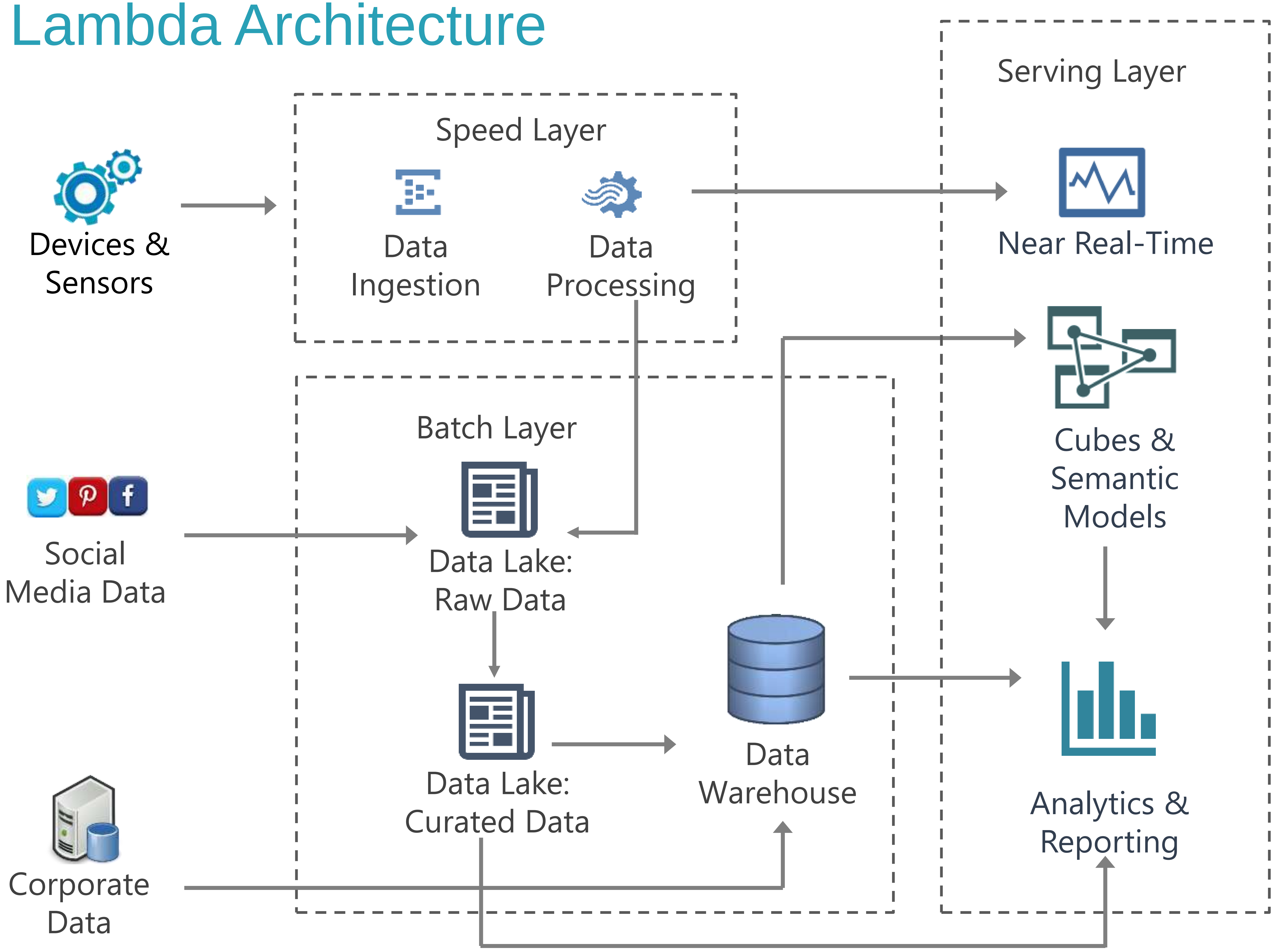


- ✓ Grow around existing DW
- ✓ Aged data available for querying when needed
- ✓ Complement to the DW via data virtualization
- ✓ Federated queries to access current data (relational DB) + archive (data lake)

Thanks to Melissa Coates, www.CoatesDataStrategies.com

Data Lake Use Cases

Lambda Architecture



- ✓ Support for low-latency, high-velocity data in near real time
- ✓ Support for batch-oriented operations

Thanks to Melissa Coates,
www.CoatesDataStrategies.com

Q & A

James Serra, Big Data Evangelist

Email me at: JamesSerra3@gmail.com

Follow me at: @JamesSerra

Link to me at: www.linkedin.com/in/JamesSerra

Visit my blog at: JamesSerra.com (where this slide deck is posted under the “Presentations” tab)