

Differentiate Big Data vs Data Warehouse use cases for a cloud solution

James Serra

Big Data Evangelist

Microsoft

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"



Agenda

Data Lake Driven Analytics

Compute technologies

Patterns

Data Lake Driven Analytics

Two Approaches to Information Management for Analytics: Top-Down + Bottom-Up



Data Warehousing Uses A Top-Down Approach

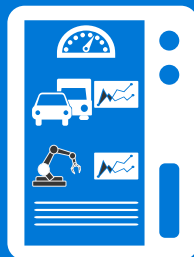
Understand
Corporate
Strategy

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



ETL



Data sources



OLTP



ERP

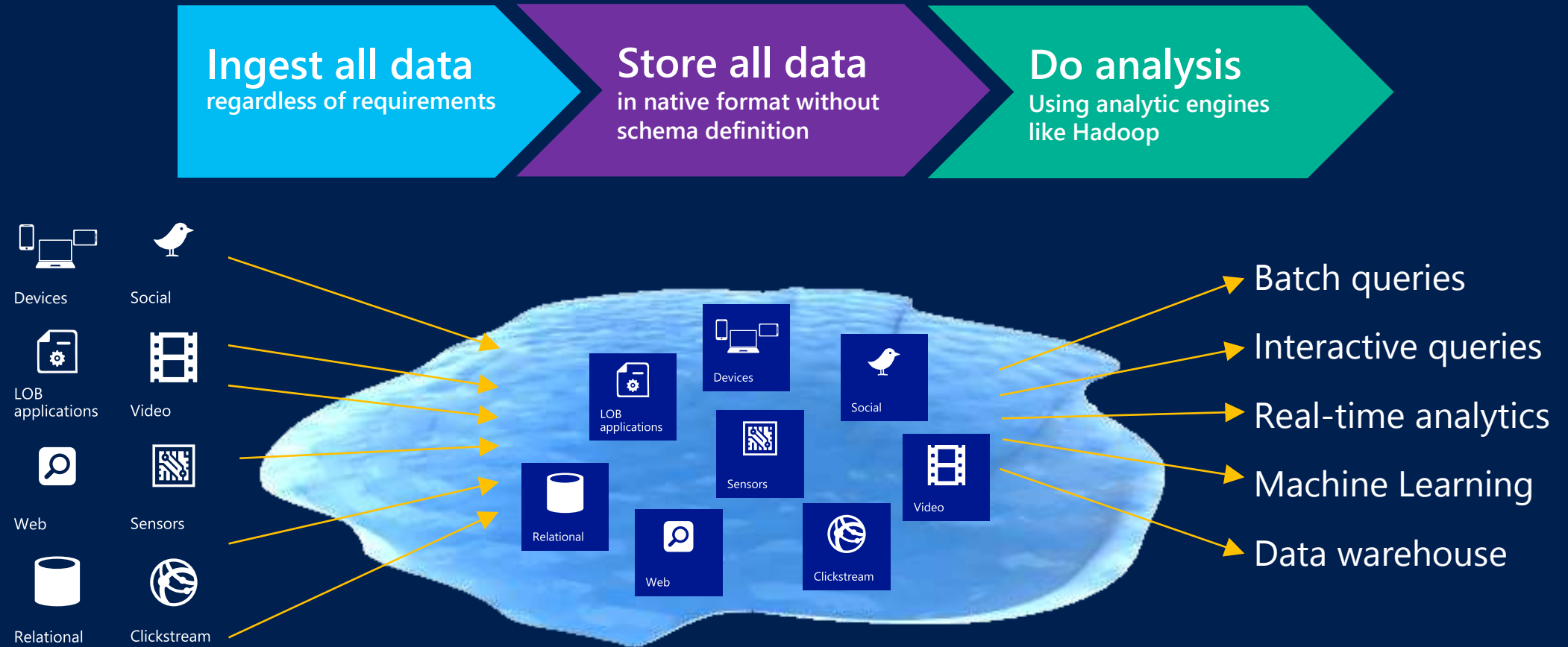


CRM



LOB

The "Data Lake" Uses A Bottom-Up Approach



The central store for analytical solutions

Vast
Capacity

Low
Cost

Raw
Data

Data Lake Layers



The diagram consists of four colored squares arranged horizontally. From left to right, the colors are dark blue, medium blue, bright blue, and light blue. Each square contains text representing a layer of a data lake. The text is white and centered within each square.

Raw
Data Layer

Cleansed
Data Layer

Application
Data Layer

Sandbox
Data Layer

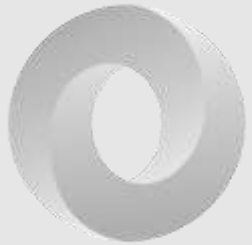
*Needs data governance so your data lake does not turn
into a data swamp!*

Considering Data Types



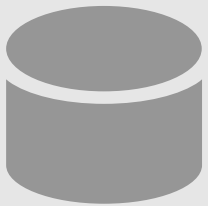
Unstructured

Audio, video, images. Meaningless without adding some structure



Semi-Structured

JSON, XML, sensor data, social media, device data, web logs. Flexible data model structure



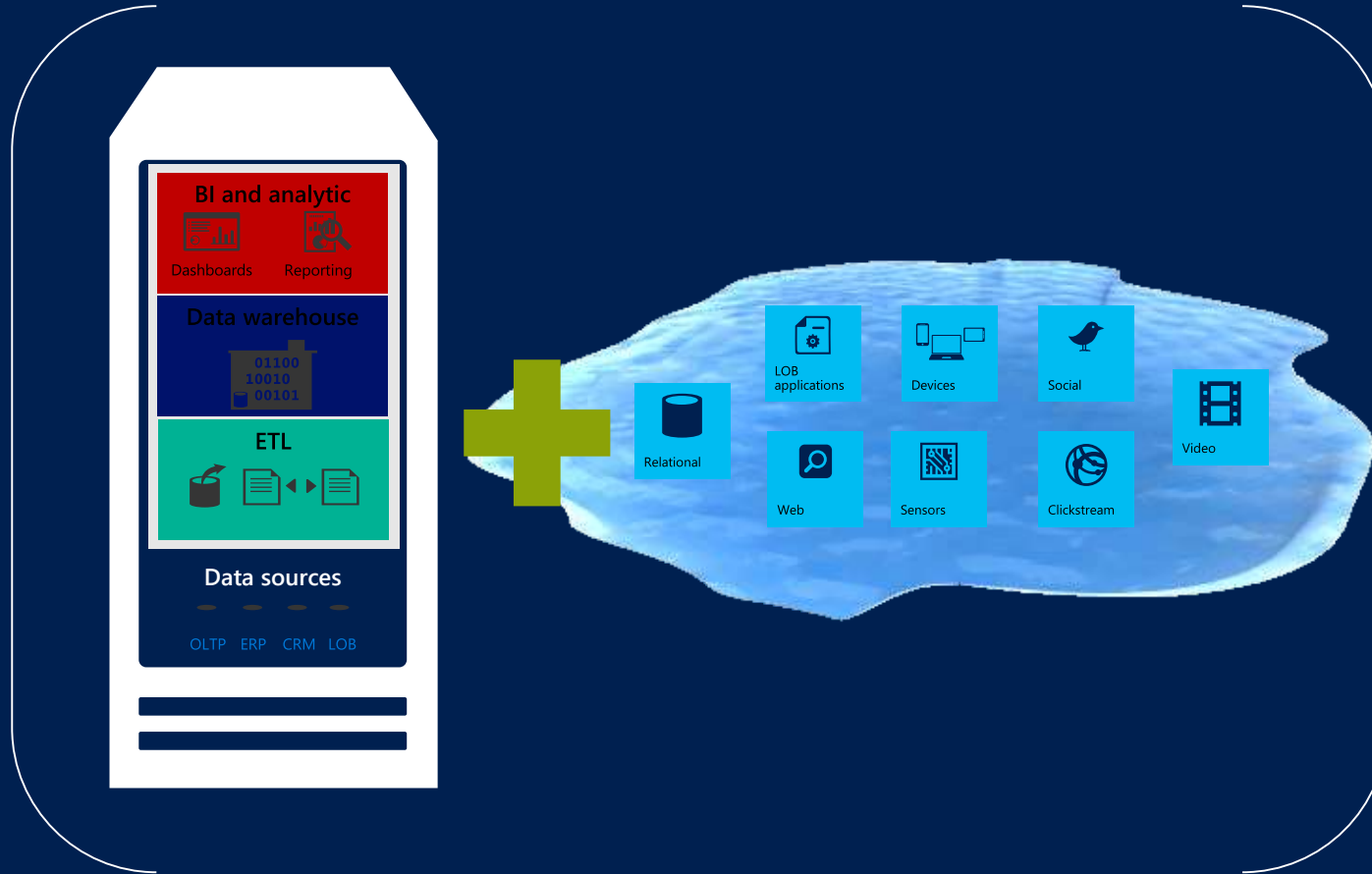
Structured

CSV, Columnar Storage (Parquet, ORC). Strict data model structure

Relational data and non-relational data are *data models*, describing how data is organized. Structured, semi-structured, and unstructured data are *data types*

Data Lake + Data Warehouse Better Together

What happened?
What is happening?
Why did it happen?
What are key relationships?



What will happen?
What if?
How risky is it?
What should happen?
What is the best option?
How can I optimize?

Data Lake and Data Warehouse Summary

Data Lake	Data Warehouse
Schema-on-read	Schema-on-write
Physical collection of uncurated data	Data of common meaning
System of Insight: Unknown data to do experimentation / data discovery	System of Record: Well-understood data to do operational reporting
Any type of data	Limited set of data types (ie. relational)
Skills are limited	Skills mostly available
All workloads – batch, interactive, streaming, machine learning	Optimized for interactive querying
Complementary to DW	Can be sourced from Data Lake

Both are needed!

Data Lake with DW use cases

Data Lake

- Staging & preparation
- Batch processing
- Data refinement/cleaning
- ETL workloads
- Store historical data
- Sandbox for data exploration
- One-time reports
- Data scientist workloads
- Quick results

Data Warehouse

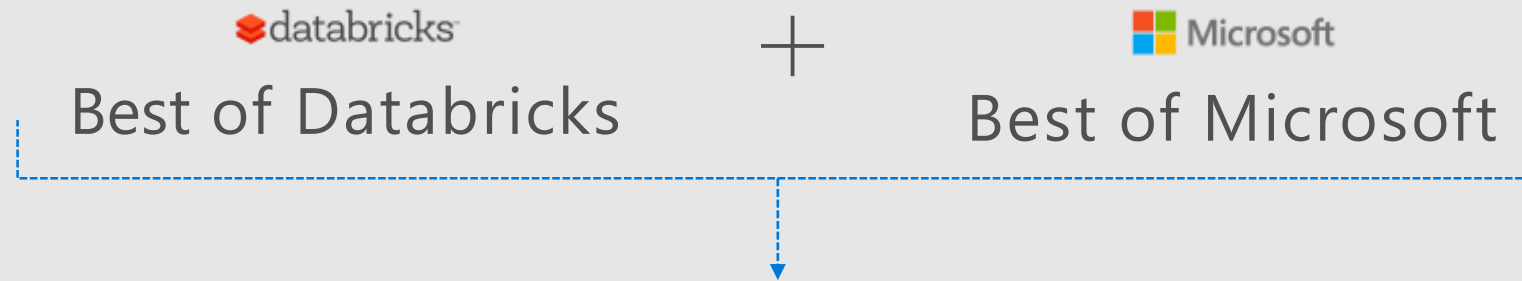
- Serving, Security & Compliance
- Low latency
- Interactive ad-hoc query
- High number of users
- Additional security
- Large support for tools
- Easily create reports (Self-service BI)


A data lake is just a glorified file folder with data files in it – how many end-users can accurately create reports from it?

Compute Technologies

What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



 Designed in collaboration with the founders of Apache Spark



One-click set up; streamlined workflows



Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.



Native integration with Azure services (Power BI, SQL DW, Cosmos DB, Blob Storage)



Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

Azure HDInsight

Hadoop and Spark
as a Service on Azure



Fully-managed Hadoop and Spark
for the cloud

100% Open Source Hortonworks
data platform

Clusters up and **running in minutes**

Managed, monitored and supported
by Microsoft with the **industry's best SLA**

Familiar **BI tools for analysis**, or open source
notebooks for **interactive data science**

63% lower TCO than deploy your own
Hadoop on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

Azure Data Lake Analytics

A new distributed
analytics service



Distributed analytics service built on
Apache YARN

Elastic scale per query lets users focus on
business goals—not configuring hardware

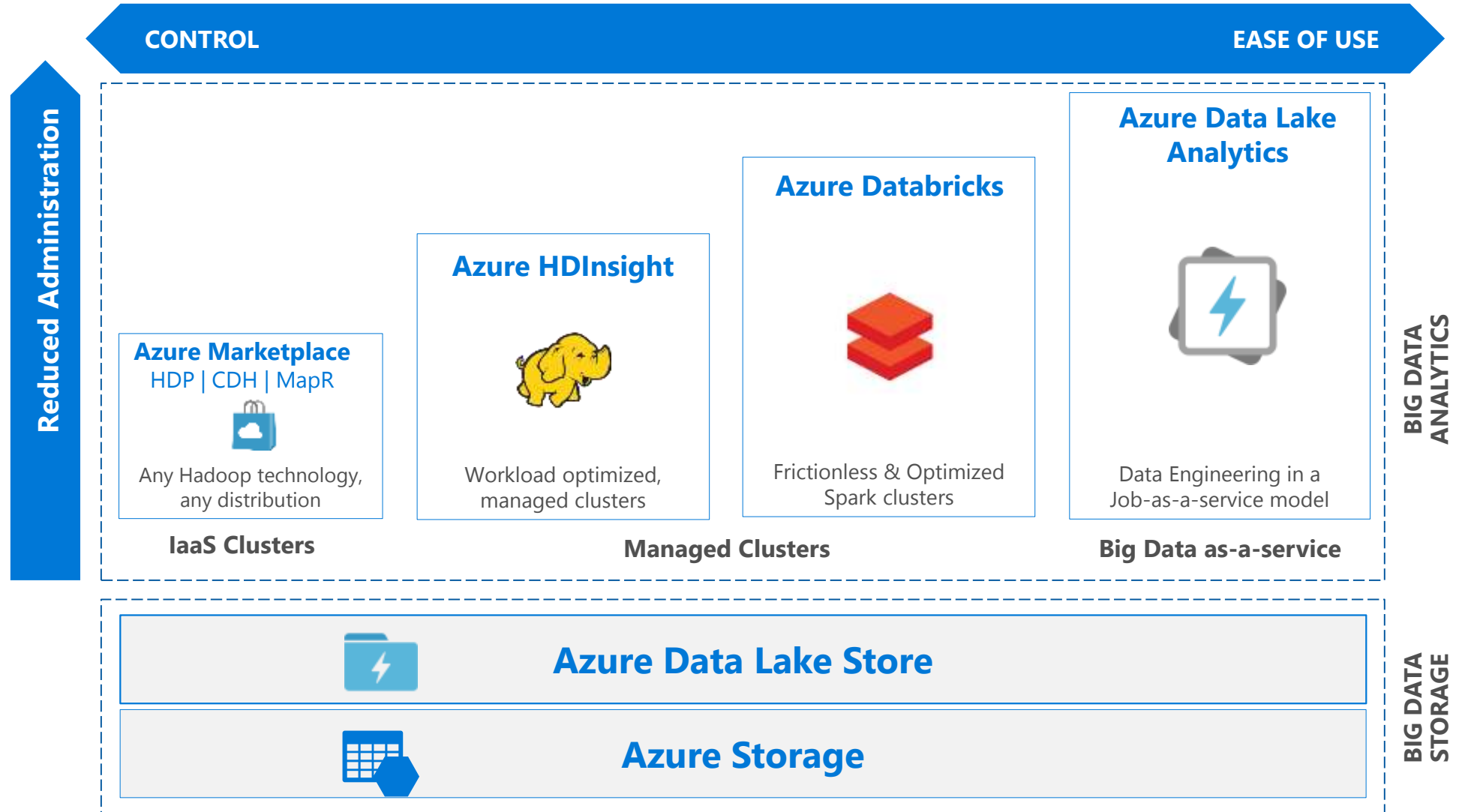
Includes U-SQL—a language that unifies the
**benefits of SQL with the expressive
power of C#**

Integrates with Visual Studio to develop,
debug, and tune code faster

Federated query across Azure data sources

Enterprise-grade **role based access control**

KNOWING THE VARIOUS BIG DATA SOLUTIONS



Azure SQL Data Warehouse

A relational **data warehouse-as-a-service**, fully managed by Microsoft.

Industries first **elastic** cloud data warehouse with **enterprise-grade** capabilities.

Support your **smallest to your largest** data storage needs while handling queries up to **100x faster**.

Elastic scale & performance

Scales to petabytes of data

Massively Parallel Processing

Instant-on compute scales in seconds

Query Relational / Non-Relational

Powered by the Cloud

Get started in minutes

Integrated with Azure ML, PowerBI & ADF

Enterprise Ready

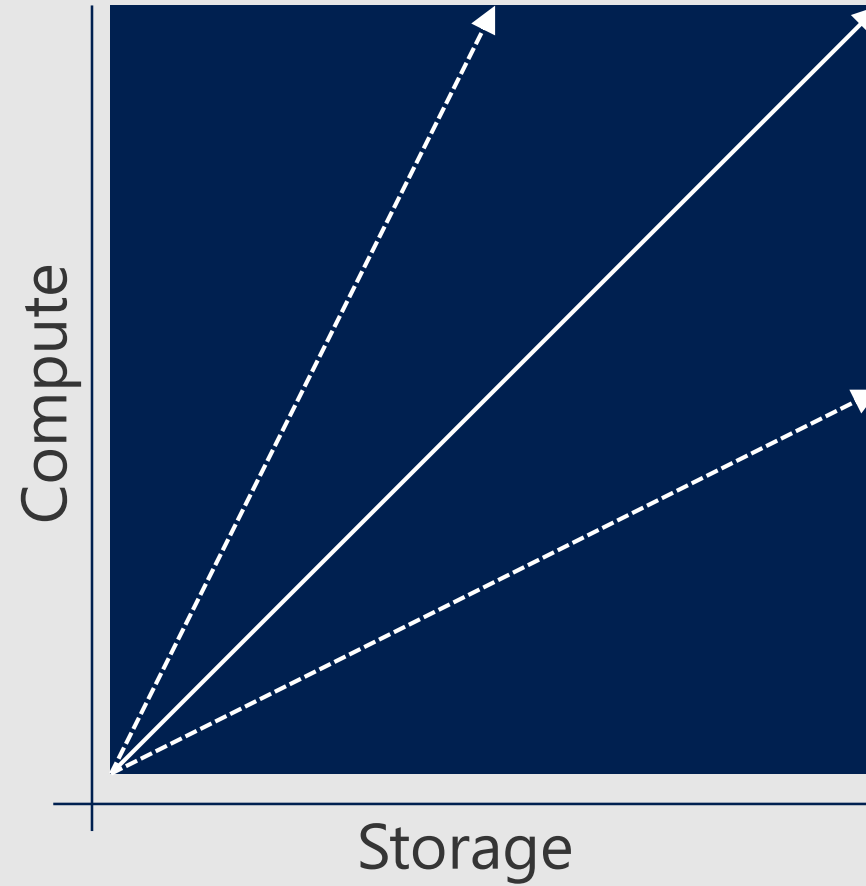
Market Leading Price & Performance

Simple billing compute & storage

Pay for what you need, when you need it with dynamic pause

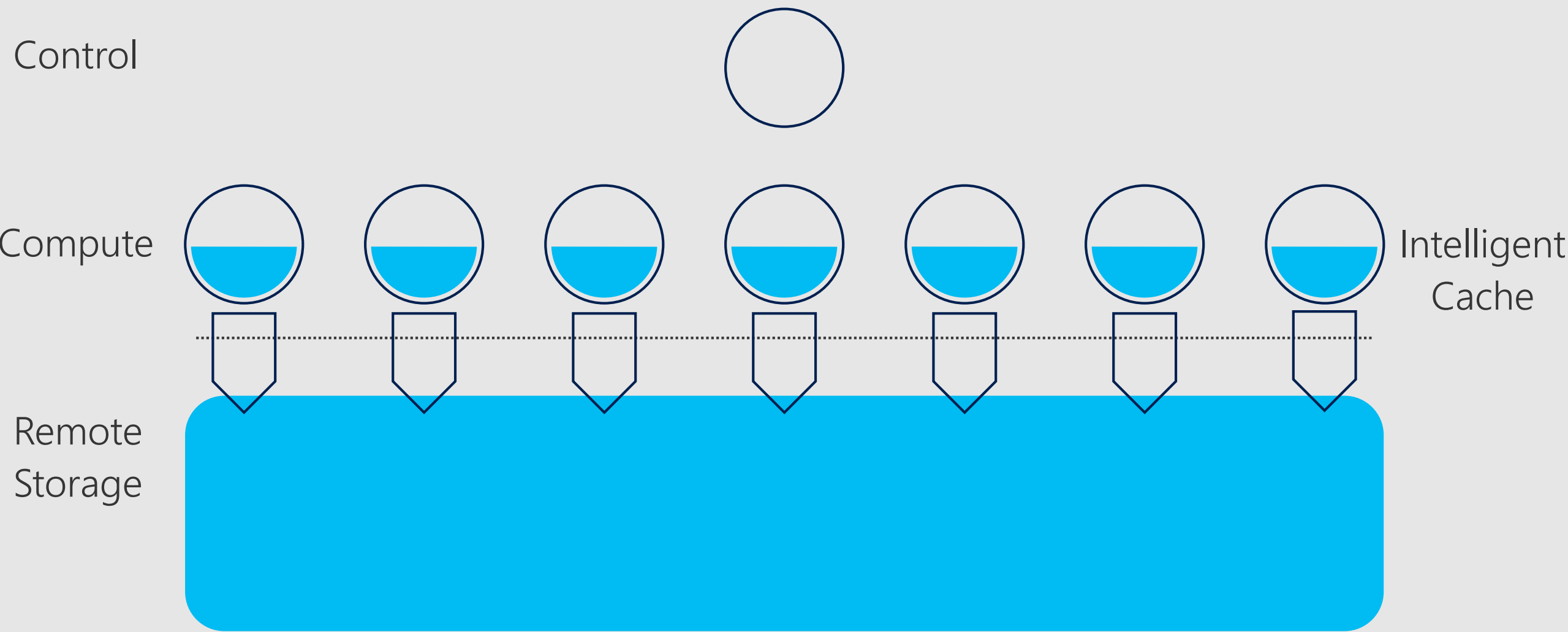
Bring DW to the Cloud without rewriting

Legacy: tightly coupled compute and storage



\$\$\$

Cloud: Separated Compute and Storage



Azure Analysis Services

All the benefits of the cloud in your analytics engine

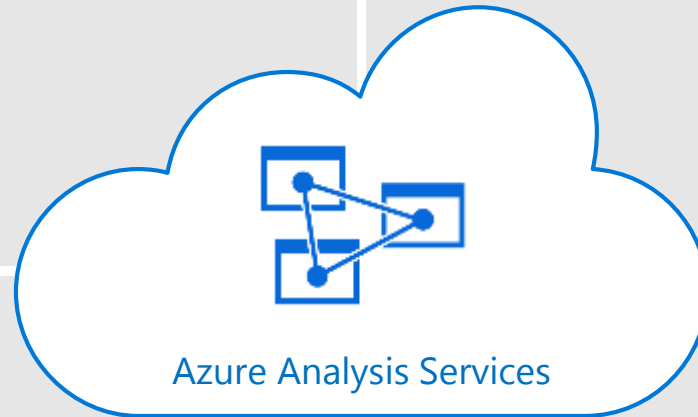


Get started quickly

Spin up a new server in seconds without managing the infrastructure

Provide secured access

From virtually anywhere



Access data when you need it

99.9% availability

Scale up, down, and pause

Only pay for what you need

Rely on Microsoft's experience running trusted enterprise cloud services

Why extend the data warehouse?

Semantic layer

Handle many concurrent users

Aggregating data for performance

Multidimensional analysis

No joins or relationships

Hierarchies, KPI's

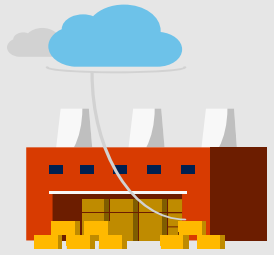
Row-level security

Advanced time-calculations

Slowly Changing Dimensions (SCD)

Big Data Technologies

Trade-offs and Considerations



	Azure SQL DW	HDInsight Hive LLAP	HDInsight Spark	ADLS/ADLA	SQL Server (IaaS)
Volume	Petabytes	Petabytes	Petabytes	Petabytes	Terabytes
Security	Encryption, TD, Audit	ADLS / Apache Ranger	ADLS	AAD Security Groups (data)	Encryption, TD Audit
Languages	T-SQL	HiveQL	SparkSQL, HiveQL, Scala, Java, Python, R	U-SQL	T-SQL
Extensibility	No	Yes, .NET/SerDe	Yes, Packages	Yes, .NET	Yes, .NET CLR
External File Types	ORC, TXT, Parquet, RCFile	ORC, CSV, Parquet + others	Parquet, JSON, Hive + others	Many	ORC, TXT, Parquet, RCFile
Admin	Low-Medium	Medium-High	Medium-High	Low	High
Cost Model	DWU	Nodes & VM	Nodes & VM	Units/Jobs	VM
Schema Definition	Schema on Write / Polybase	Schema on Read	Schema on Read	Schema on Read	Schema on Write / Polybase
Max DB Size	Unlimited CCI 240TB Comp (5X = 1PB) index/heaps			Unlimited	256TB (64 4TB drives)

Microsoft Products vs Hadoop/OSS Products

Microsoft Product	Hadoop/Open Source Software Product
Office365/Excel	OpenOffice/Calc
Cosmos DB	MongoDB, MarkLogic, HBase, Cassandra
SQL Database	SQLite, MySQL, PostgreSQL, MariaDB, Apache Ignite
Azure Data Lake Analytics/YARN	None
Azure VM/IaaS	OpenStack
Blob Storage	HDFS, Ceph
Azure HBase	Apache HBase (Azure HBase is a service wrapped around Apache HBase), Apache Trafodion
Event Hub	Apache Kafka
Azure Stream Analytics	Apache Storm, Apache Spark Streaming, Apache Flink, Apache Beam, Twitter Heron
Power BI	Apache Zeppelin, Apache Jupyter, Airbnb Caravel, Kibana
HDInsight	Hortonworks (pay), Cloudera (pay), MapR (pay)
Azure ML (Machine Learning)	Apache Mahout, Apache Spark MLlib, Apache PredictionIO
Microsoft R Open	R
SQL Data Warehouse/Interactive queries	Apache Hive LLAP, Presto, Apache Spark SQL, Apache Drill, Apache Impala
IoT Hub	Apache NiFi
Azure Data Factory	Apache Falcon, Airbnb Airflow, Apache Oozie, Apache Azkaban
Azure Data Lake Storage/WebHDFS	HDFS Ozone
Azure Analysis Services/SSAS	Apache Kylin, Apache Druid, AtScale (pay)
SQL Server Reporting Services	None
Hadoop Indexes	Jethro Data (pay)
Azure Data Catalog	Apache Atlas
PolyBase	Apache Drill
Azure Search	Apache Solr, Apache Elasticsearch (Azure Search build on ES)
SQL Server Integration Services (SSIS)	Talend Open Studio, Pentaho Data Integration
Others	Apache Ambari (manage Hadoop clusters), Apache Ranger (data security such as row/column-level security), Apache Knox (secure entry point for Hadoop clusters), Apache Flume (collecting log data)

Note: Many of the Hadoop/OSS products are available in Azure

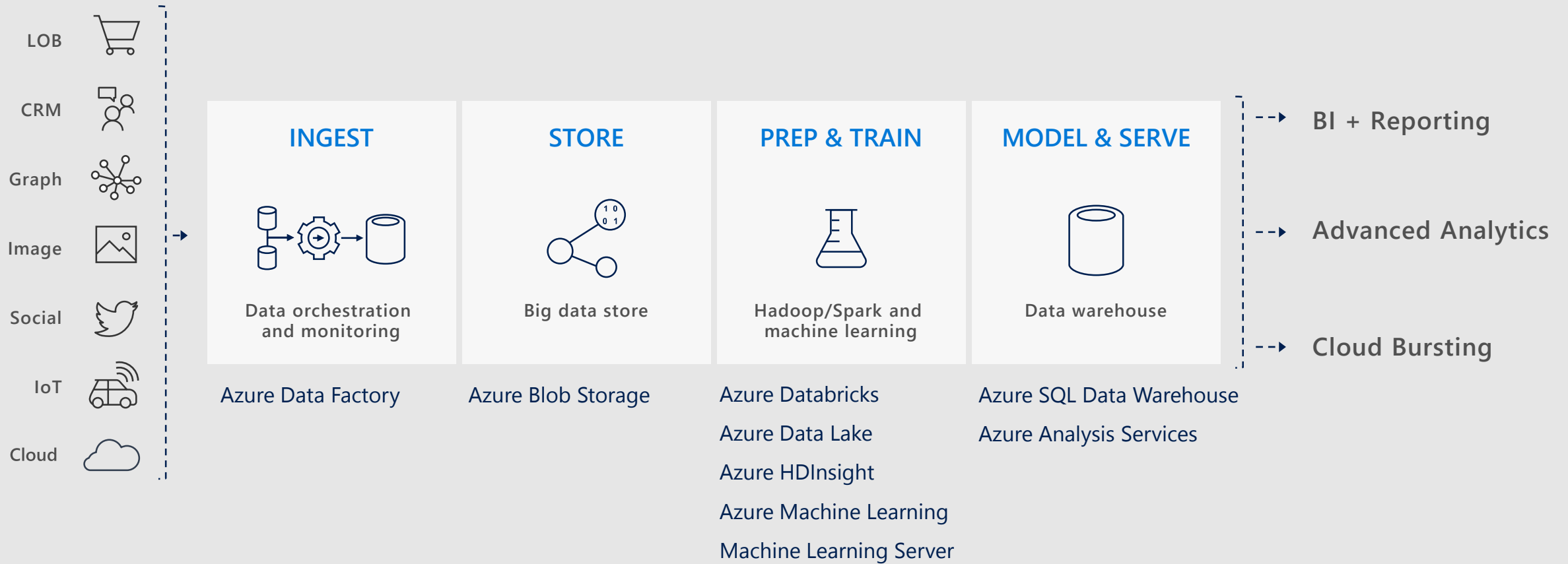
Patterns

Questions to ask client

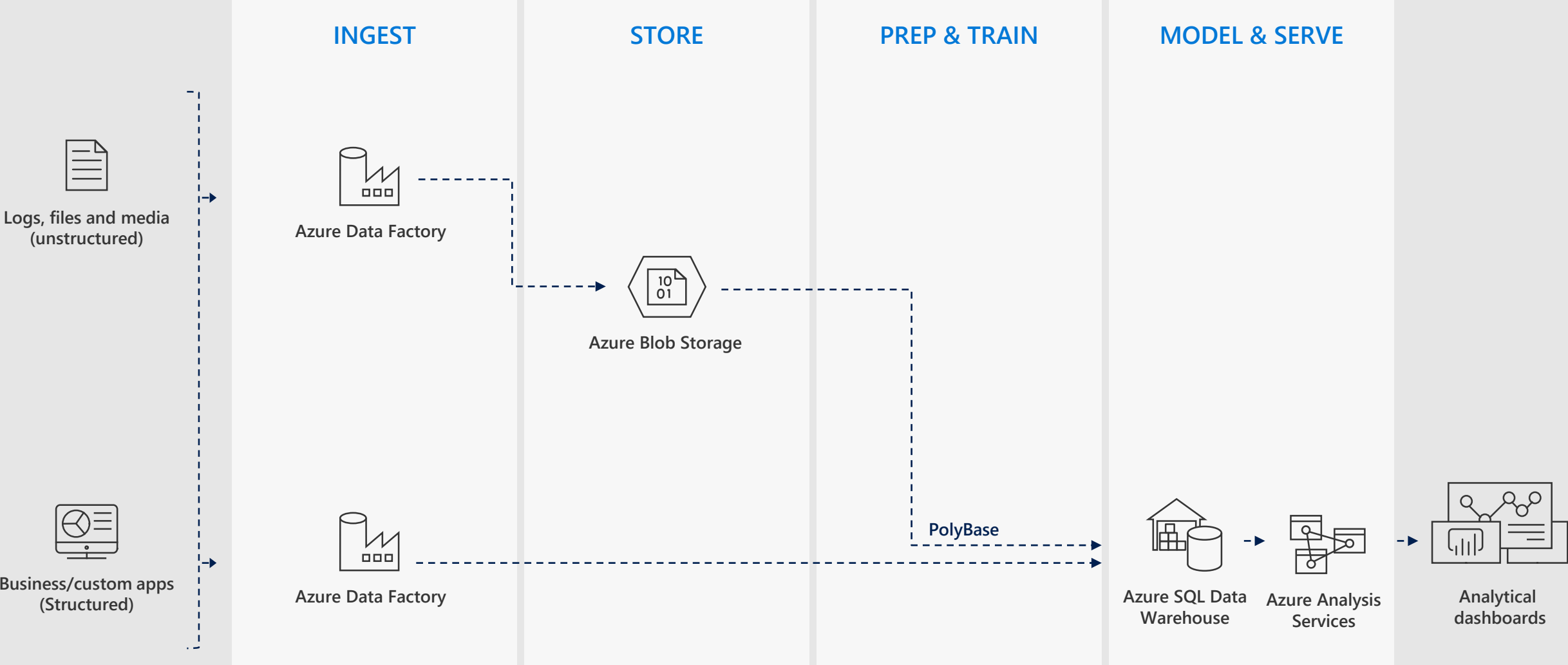
- Can you use the cloud?
- Is this a new solution or a migration?
- Do the developers have Hadoop skills?
- Will you use non-relational data (variety)?
- How much data do you need to store (volume)?
- Is this an OLTP or OLAP/DW solution?
- Will you have streaming data (velocity)?
- Will you use dashboards?
- How fast do the operational reports need to run?
- Will you do predictive analytics?
- Do you want to use Microsoft tools or open source?
- What are your high availability and/or disaster recovery requirements?
- Do you need to master the data (MDM)?
- Are there any security limitations with storing data in the cloud?
- Does this solution require 24/7 client access?
- How many concurrent users will be accessing the solution at peak-time and on average?
- What is the skill level of the end users?
- What is your budget and timeline?
- Is the source data cloud-born and/or on-prem born?
- How much daily data needs to be imported into the solution?
- What are your current pain points or obstacles (performance, scale, storage, concurrency, query times, etc)?
- Are you ok with using products that are in preview?

Microsoft Big Data & Data Warehouse

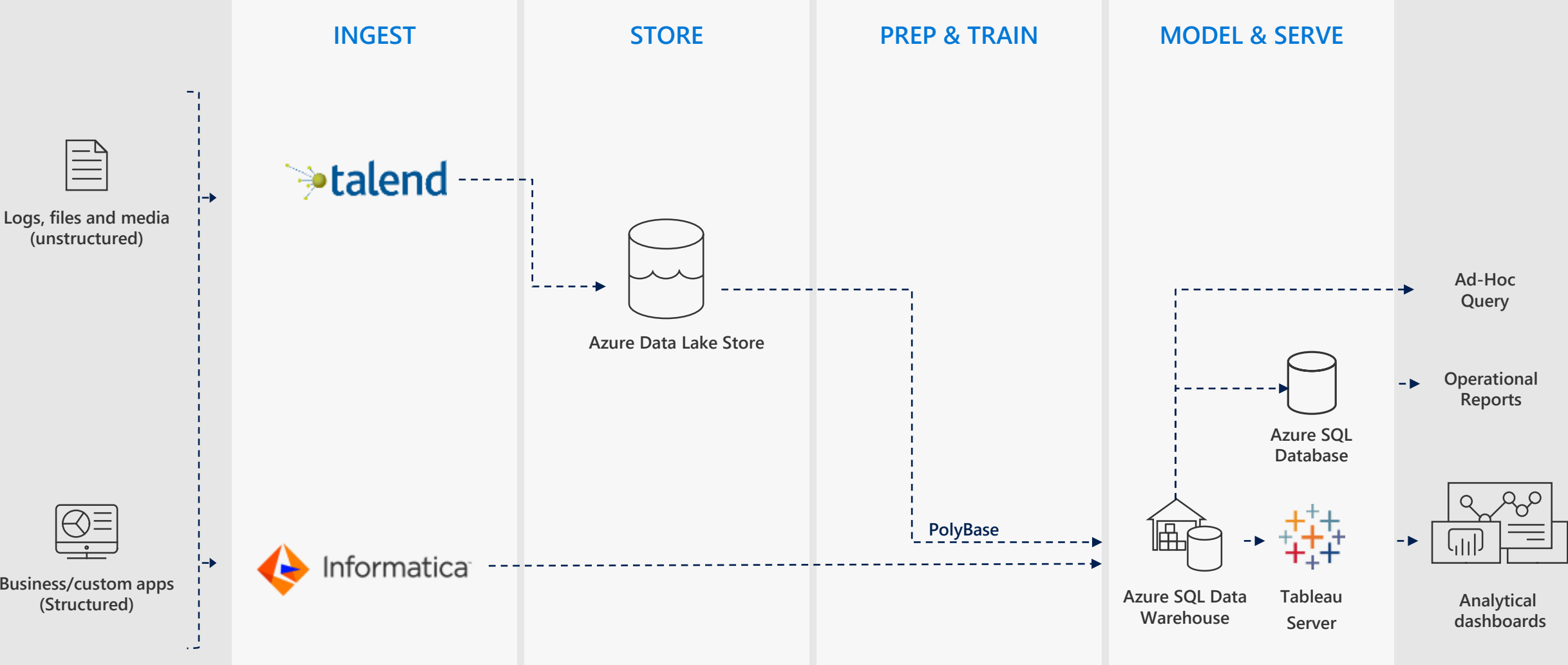
Your data hub for analytics



Cloud Data Warehouse



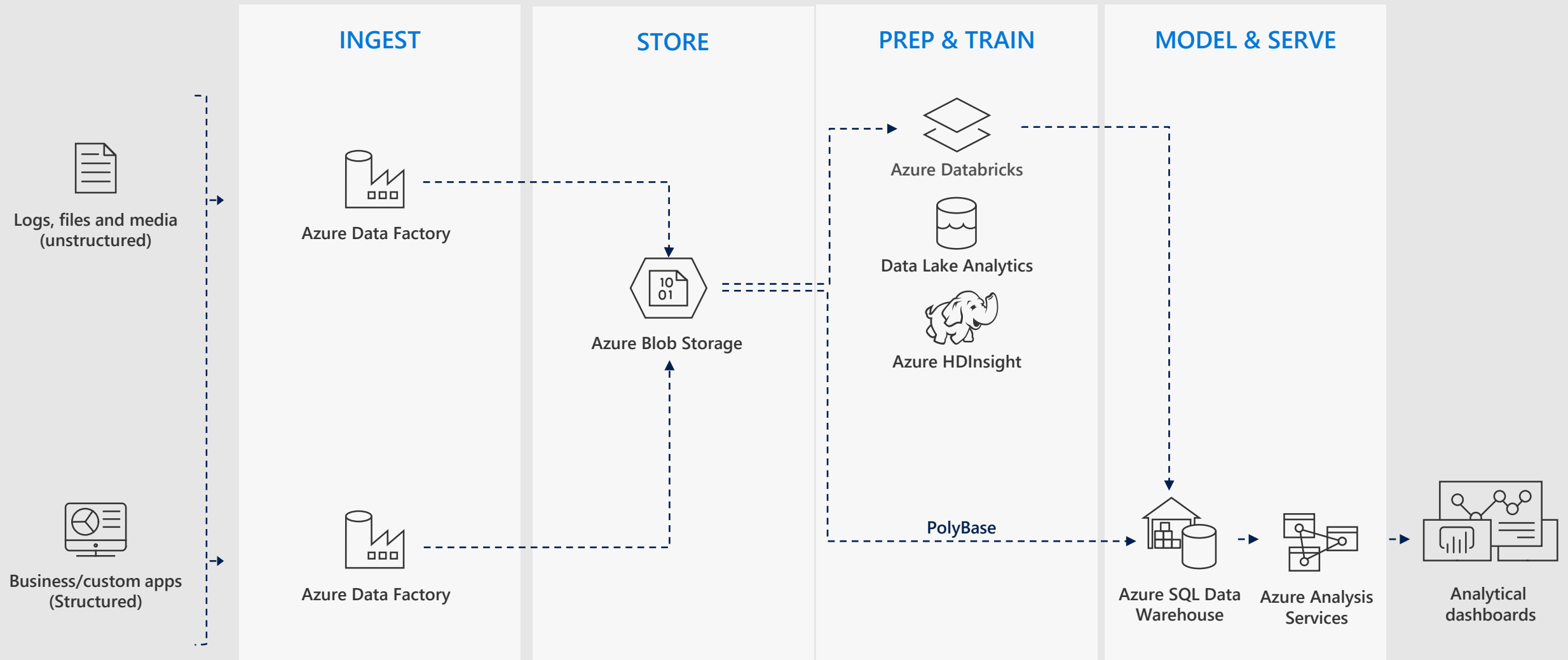
Cloud Data Warehouse variations



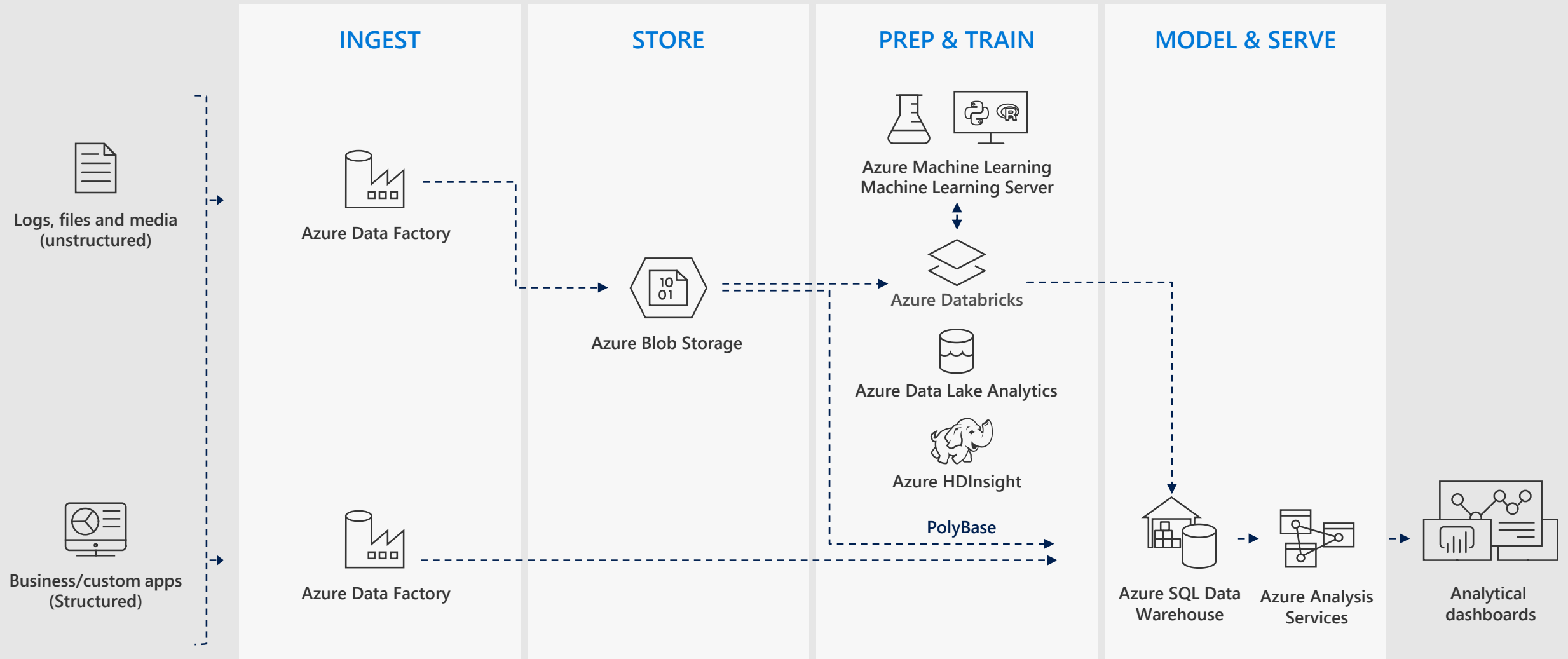
ADL Store and Blob Store

	Azure Data Lake Store	Azure Blob Storage
Purpose	Optimized storage for big data analytics workloads	General purpose object storage for a wide variety of storage scenarios including big data analytics
Use Cases	Batch, interactive, streaming analytics and machine learning data such as log files, IoT data, click streams, large datasets	Any type of text or binary data, such as application back end, backup data, media storage for streaming and general purpose data as well as big data analytics
Units of Storage	Accounts / Folders / Files	Accounts / Containers / Blobs
Structure	Hierarchical File System	Object Store with flat namespace
REST API	WebHDFS-compatible	Azure Blob Storage, compatible HDFS via WASB driver
Security	Azure Active Directory (AAD)	Shared Access Signature (SAS) keys
Authorization	POSIX Access Control Lists (ACLs)	Account-level: Account Access Keys; Account, container, or blob authorization: SAS keys
Account/File Size Limits	No limits on account size or file size	5PB account/4.75TB file
Single Object/Account Throughput Limit	Extremely high	2GB/s, or 50k tps (now stripe across multiple hard drives)/50GBs bandwidth
Geo-Replications	LRS	LRS, ZRS, GRS, RA-GRS
Cost/Month [1PB, East US 2]	No tiering: \$39k + Transactions	Tiering: Hot \$18k, Cool \$10k, Archive \$2k + Trans
Product integration/Tooling	Check	Check
Region Availability	Two US regions (East, Central) & North Europe	All Azure Regions

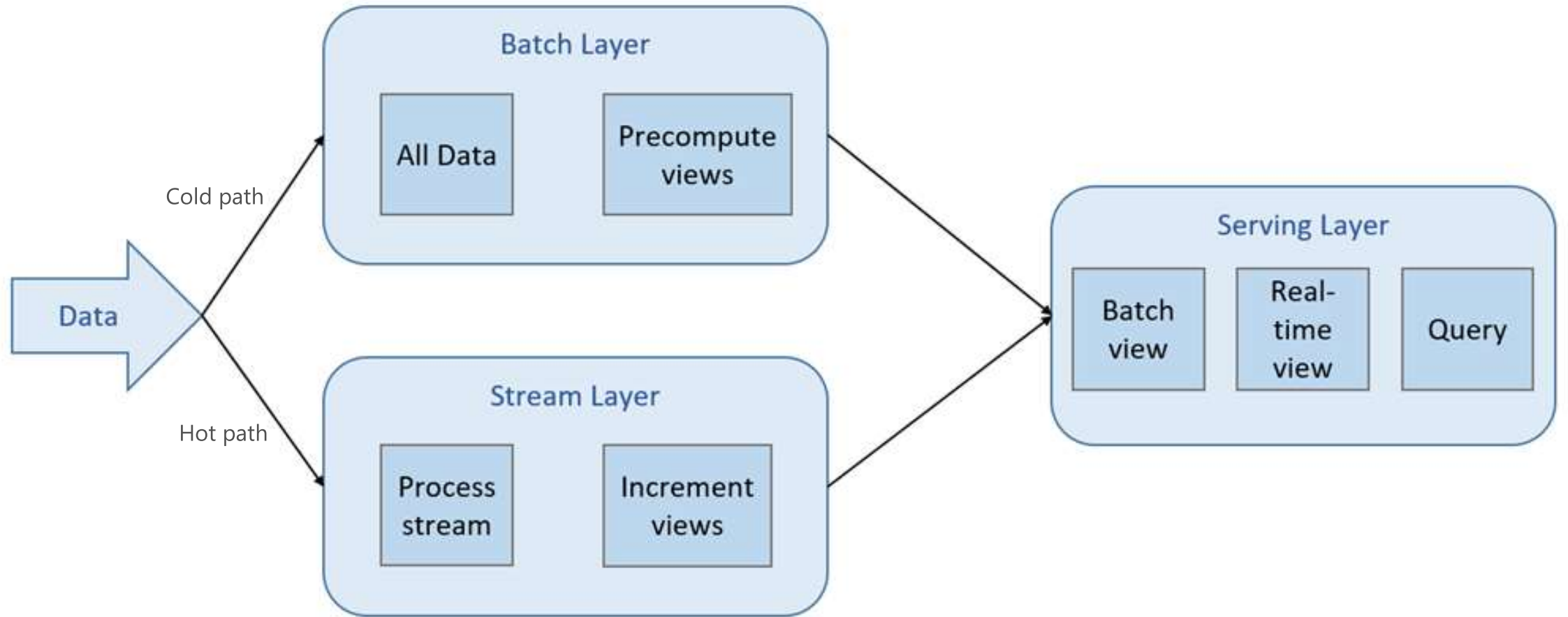
Modern Data Warehouse



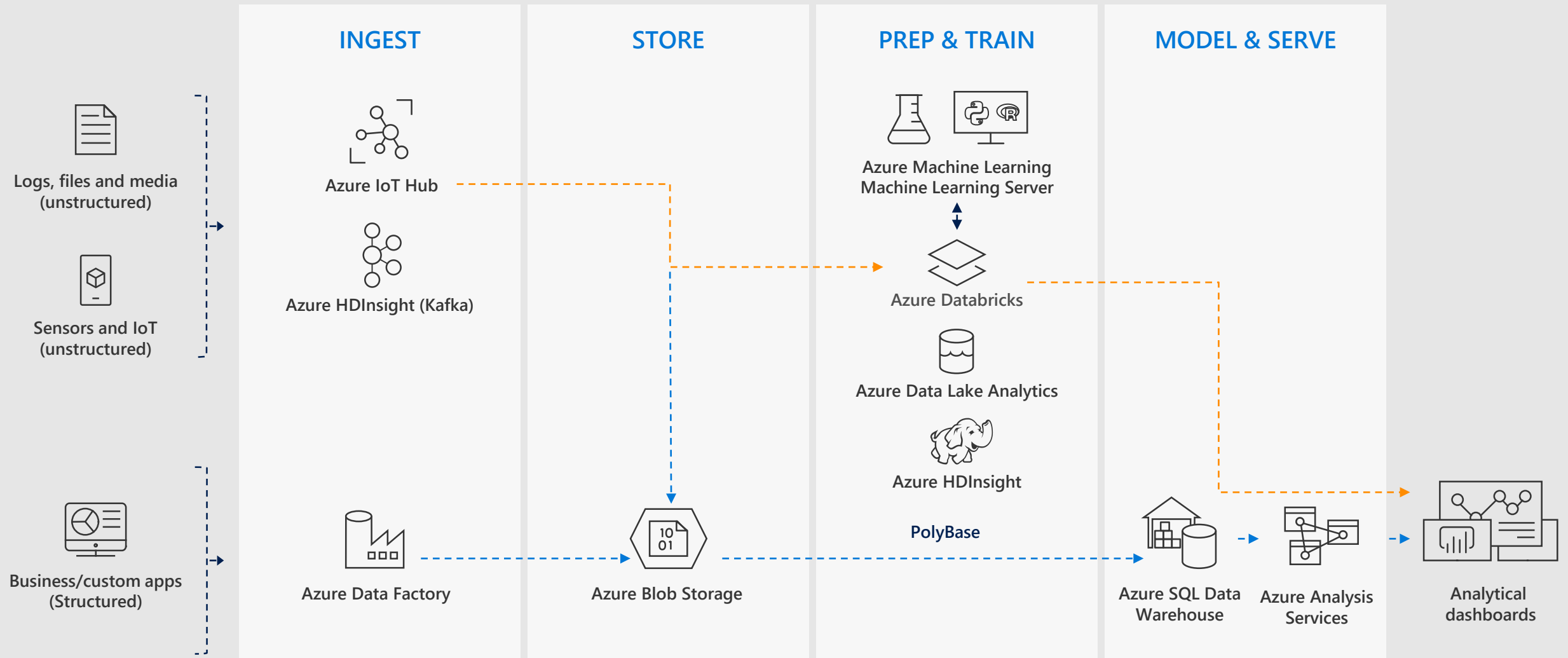
Advanced Analytics on Big Data



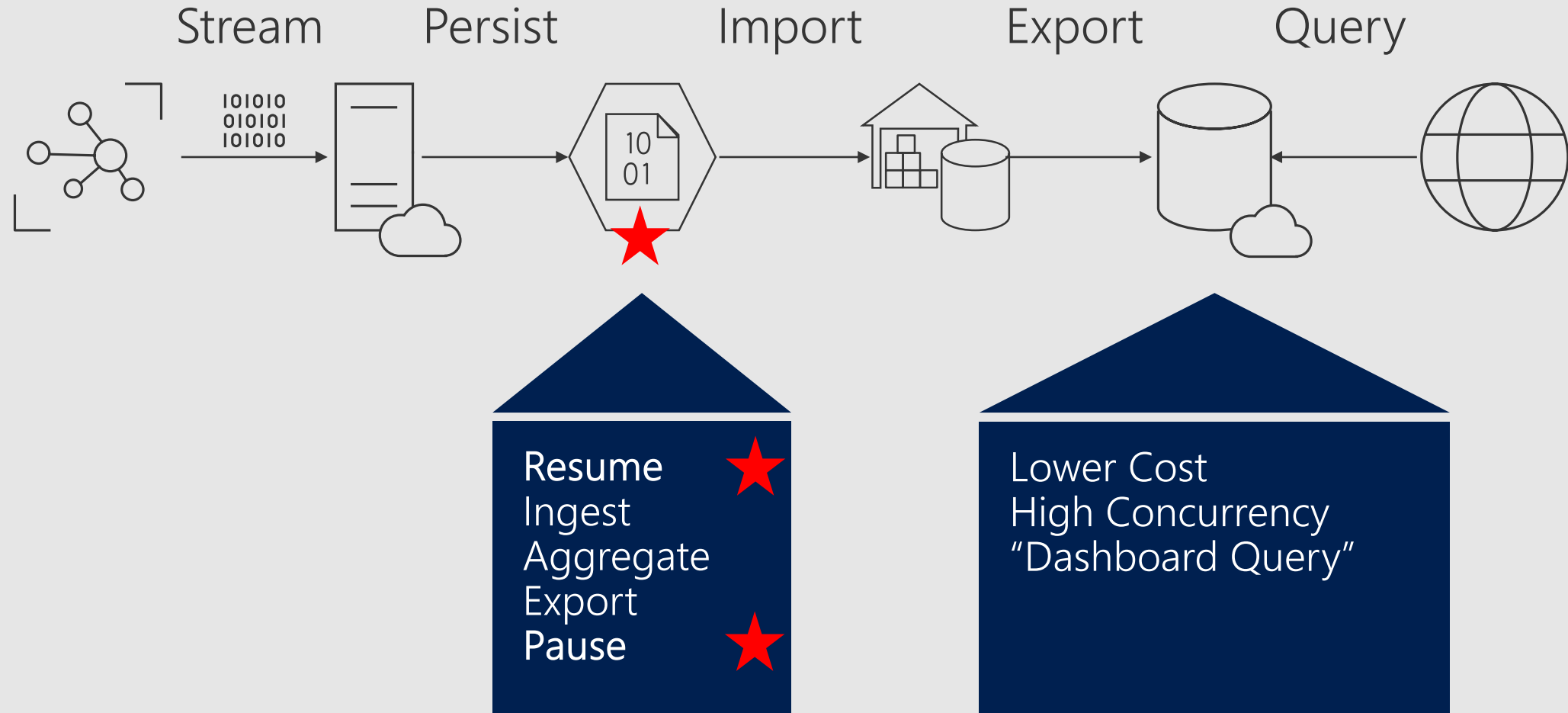
Lambda Architecture



Real time analytics



Cloud Economics



Wrapping up

Azure Data Architecture Guide

<https://aka.ms/ADAG>

Traditional RDBMS

Concepts

- Relational data
- Transactional data
- Semantic modeling

Scenarios

- Online analytical processing (OLAP)
- Online transaction processing (OLTP)
- Data warehousing and data marts
- ETL

Big data and NoSQL

Concepts

- Non-relational data stores
- Working with CSV and JSON files
- Big data architectures
- Advanced analytics
- Machine learning at scale

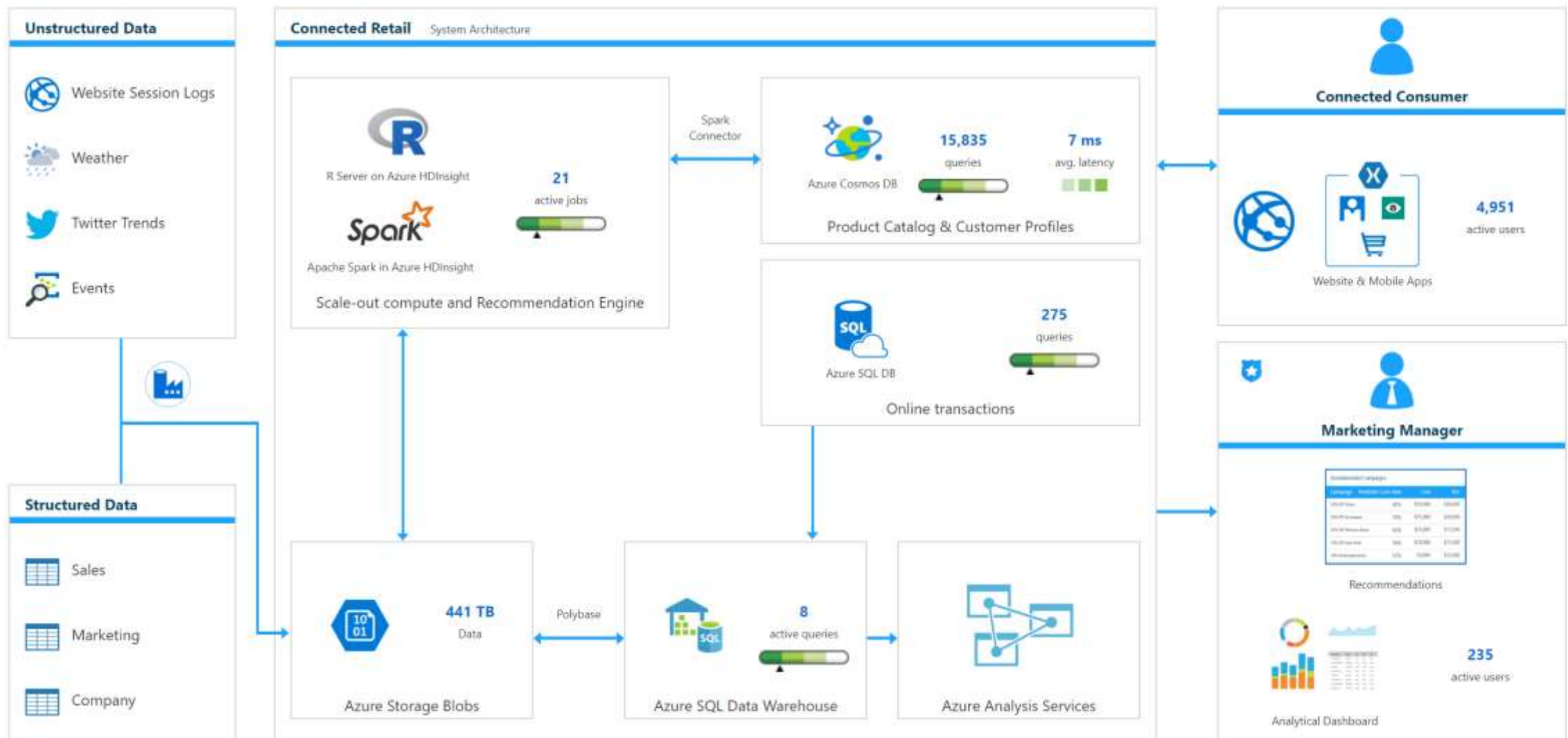
Scenarios

- Batch processing
- Real time processing
- Free-form text search
- Interactive data exploration
- Natural language processing
- Time series solutions

Cross-cutting concerns

- Data transfer
- Extending on-premises data solutions to the cloud
- Securing data solutions

CONNECTED RETAIL - OPERATIONS DASHBOARD



Q & A



James Serra, Big Data Evangelist

Email me at: JamesSerra3@gmail.com

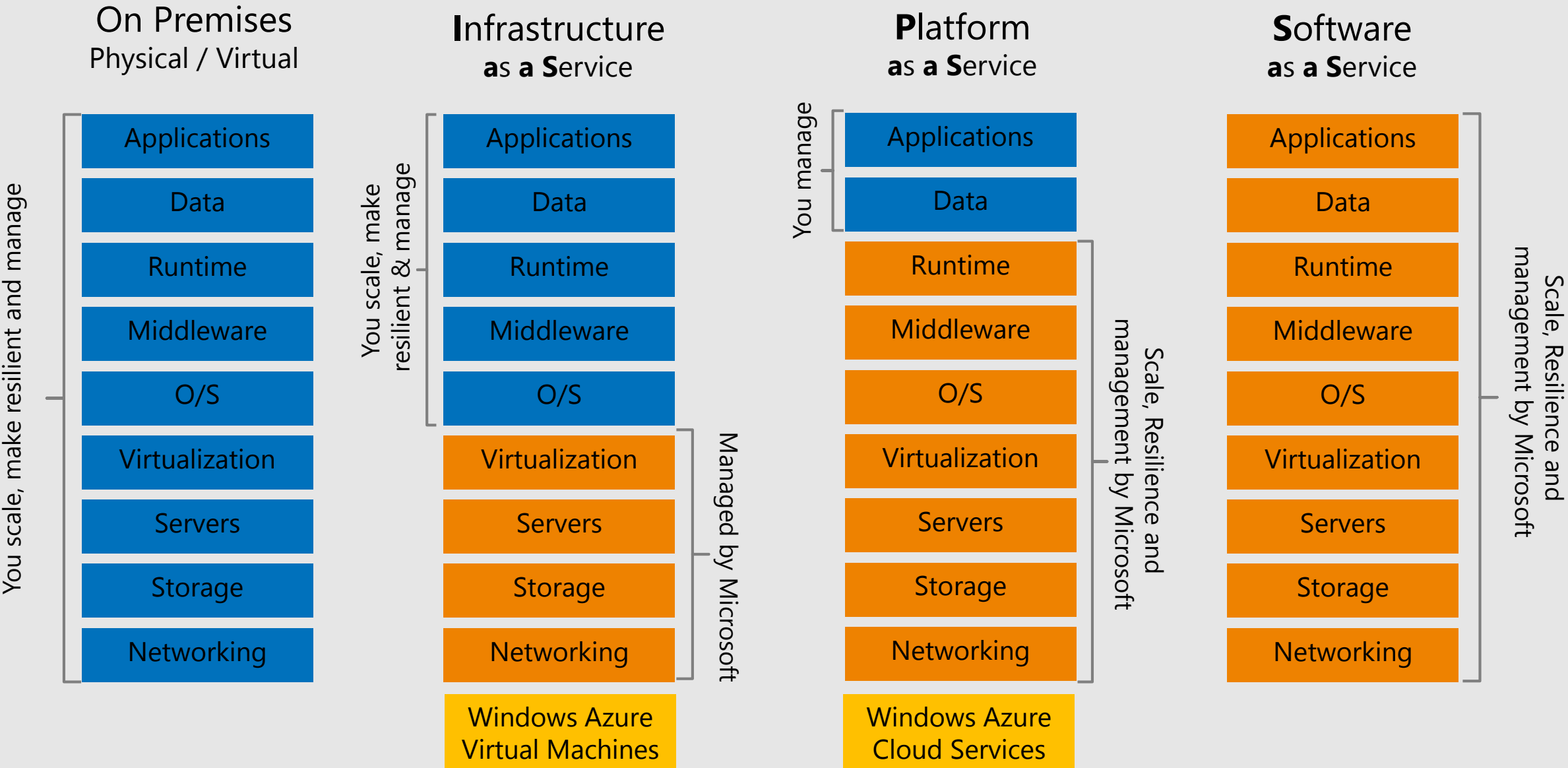
Follow me at: @JamesSerra

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)

APPENDIX

Who manages what?



Data management for analytics at any stage

STAGE 1:

Traditional

Query historical, relational data from a variety of sources

STAGE 2:

Operational

Gain real-time insights without impacting performance

STAGE 3:

Logical

Ask questions of big data—all types, volumes and locations

STAGE 4:

Free-form



Establish enterprise-wide data lake and run advanced analytics and deep learning on unstructured data that arrives in real-time

Microsoft data platform solutions (partial list)

Product	Category	Description	More Info
SQL Server 2017	RDBMS	Earned top spot in Gartner's Operational Database magic quadrant. JSON support. Linux support	https://www.microsoft.com/en-us/server-cloud/products/sql-server-2017/
SQL Database	RDBMS/DBaaS	Cloud-based service that is provisioned and scaled quickly. Has built-in high availability and disaster recovery. JSON support.	https://azure.microsoft.com/en-us/services/sql-database/
SQL Data Warehouse	MPP RDBMS/DBaaS	Cloud-based service that handles relational big data. Provision and scale quickly. Can pause service to reduce cost	https://azure.microsoft.com/en-us/services/sql-data-warehouse/
Azure Data Lake Store/Blob Storage	Hadoop storage	Removes the complexities of ingesting and storing all of your data while making it faster to get up and running with batch, streaming, and interactive analytics	https://azure.microsoft.com/en-us/services/data-lake-store/ https://azure.microsoft.com/en-us/services/storage/blobs
HDInsight	PaaS Hadoop compute/Hadoop clusters-as-a-service	A managed Apache Hadoop, Spark, R Server, HBase, Kafka, Interactive Query (Hive LLAP) and Storm cloud service made easy	https://azure.microsoft.com/en-us/services/hdinsight/
Azure Databricks	PaaS Spark clusters	A fast, easy, and collaborative Apache Spark based analytics platform optimized for Azure	https://databricks.com/azure
Azure Data Lake Analytics	On-demand analytics job service/Big Data-as-a-service	Cloud-based service that dynamically provisions resources so you can run queries on exabytes of data. Includes U-SQL, a new big data query language	https://azure.microsoft.com/en-us/services/data-lake-analytics/
Azure Analysis Services	Online analytical processing/PaaS	OLAP engine used in decision support and business analytics, providing the analytical data for business reports and client apps	https://azure.microsoft.com/en-us/services/analysis-services/
Azure Cosmos DB	PaaS NoSQL: Key-value, Column-family, Document, Graph	Globally distributed, massively scalable, multi-model, multi-API, low latency data service – which can be used as an operational database or a hot data lake	https://azure.microsoft.com/en-us/services/cosmos-db/
Azure Database for PostgreSQL, MySQL, and MariaDB	RDBMS/DBaaS	A fully managed database service for app developers	https://azure.microsoft.com/en-us/services/postgresql

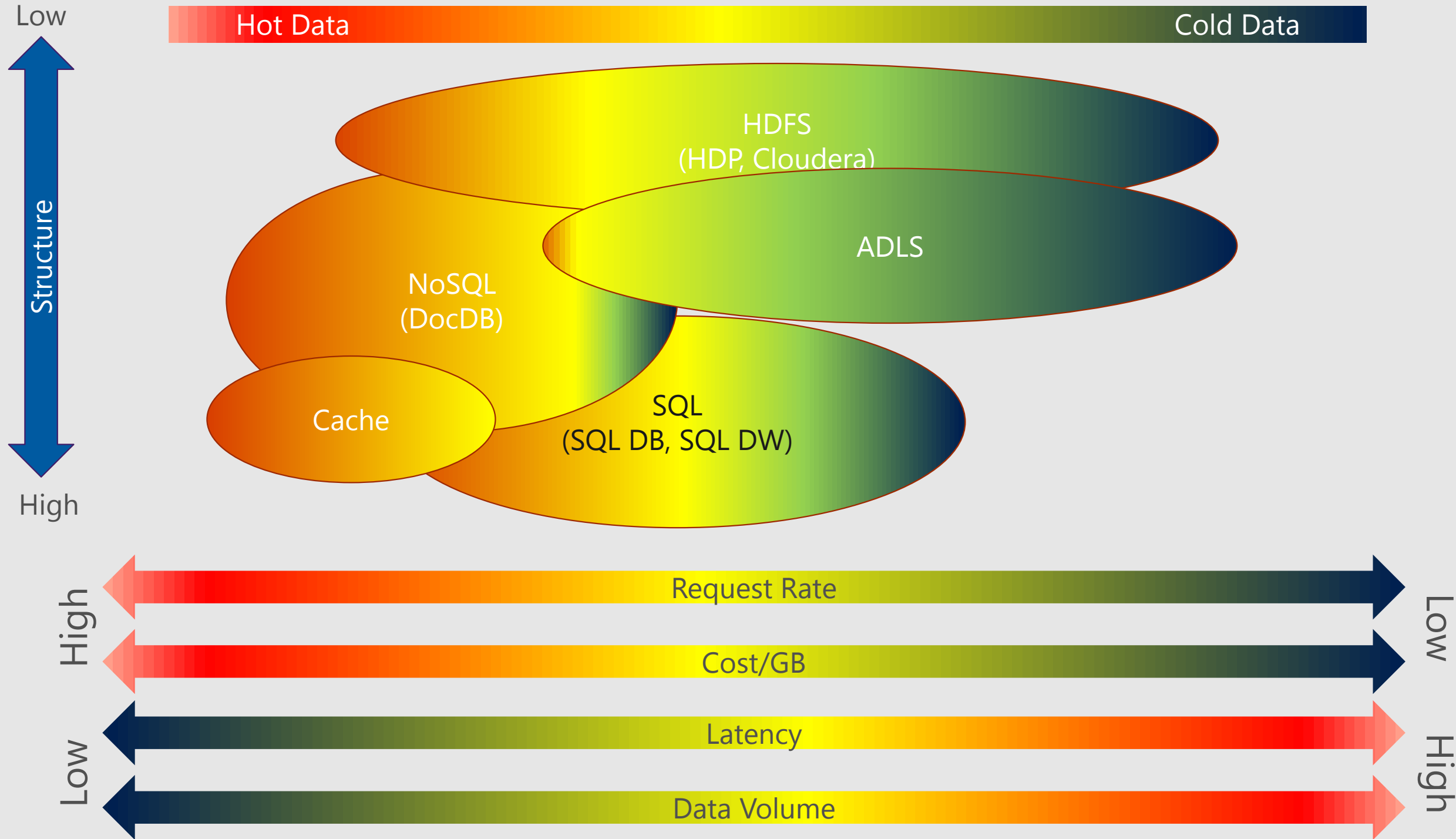
ExpressRoute

Metered Data plan







All inbound data transfer is free of charge, and all outbound data transfer is charged based on a pre-determined rate (listed below). Users are also charged a fixed monthly port fee (based on High Availability dual ports).

PORT SPEED		PRICE PER MONTH	PRICE PER MONTH WITH PREMIUM ADD-ON	INBOUND DATA TRANSFER INCLUDED	OUTBOUND DATA TRANSFER INCLUDED
50 Mbps		\$55	\$130	Unlimited	None
100 Mbps		\$100	\$200	Unlimited	None
200 Mbps		\$145	\$295	Unlimited	None
500 Mbps		\$290	\$690	Unlimited	None
1 Gbps	(3.6TB/hour)	\$436	\$1,186	Unlimited	None
2 Gbps	(7.2TB/hour)	\$872	\$2,372	Unlimited	None
5 Gbps	(18TB/hour)	\$2,180	\$5,180	Unlimited	None
10 Gbps	(36TB/hour)	\$5,000	\$8,000	Unlimited	None

We guarantee 99.95% ExpressRoute dedicated circuit availability



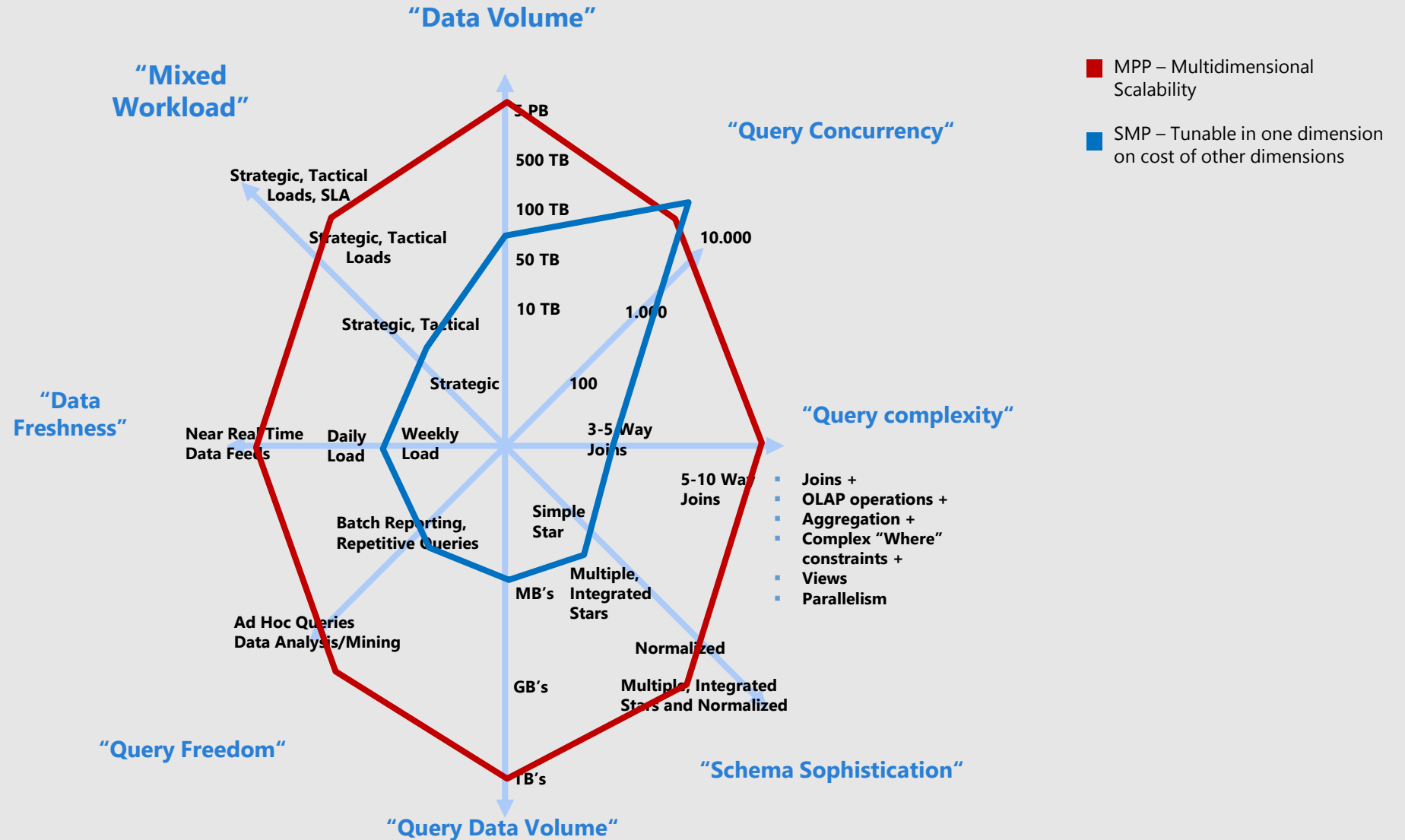
Data Lake Store: Technical Requirements

	Secure	Must be highly secure to prevent unauthorized access (especially as all data is in one place).
	Scalable	Must be highly scalable. When storing all data indefinitely, data volumes can quickly add up
	Reliable	Must be highly available and reliable (no permanent loss of data).
	Throughput	Must have high throughput for massively parallel processing via frameworks such as Hadoop and Spark
	Low latency	Must have low latency for high-frequency operations.
	Details	Must be able to store data with all details; aggregation may lead to loss of details.
	Native format	Must permit data to be stored in its 'native format' to track lineage & for data provenance.
	All sources	Must be able ingest data from a variety of sources-LOB/ERP, Logs, Devices, Social NWs etc.
	Multiple analytic frameworks	Must support multiple analytic frameworks —Batch, Real-time, Streaming, ML etc. No one analytic framework can work for all data and all types of analysis.

DW SCALABILITY SPIDER CHART

The spiderweb depicts important attributes to consider when evaluating Data Warehousing options.

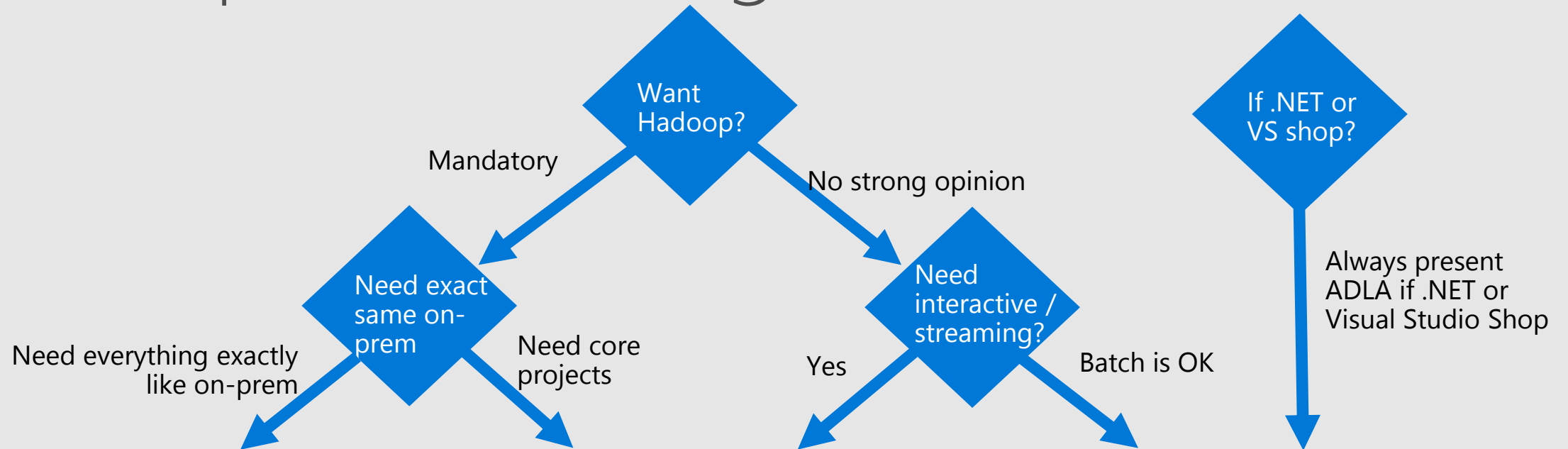
Big Data support is newest dimension.



HiveQL vs. T-SQL

	HiveQL	T-SQL
Updates	INSERT	UPDATE, INSERT, DELETE
Transactions	Supported	Supported
Index Types	Compact, Bitmap	Clustered, Non-Clustered, Columnar
Latency	Minutes	Seconds
Data Types	+ map, struct, array	Int, float, boolean, char, binary..
Functions	Dozens of built-in functions	Hundreds of built-in functions
Multi-Table Inserts	Supported	Not Supported
Partitioning	Supported	Supported
Multi-Level Partitioning	Supported	Not Supported
Views	Read-Only	Read-Only and Updateable
Extensibility	UDF's, MapReduce Scripts	UDF's, Stored Procedures

When to position the big data solutions



Azure Marketplace (IaaS)

- Need all workloads exactly like on-premises
- Need 100% Hortonworks/Cloudera/MapR

Azure HDInsight

- Most Hadoop workloads
- Fully managed by Microsoft
- Sell HDI + ADLS
- Stickier to Microsoft than VMs
- Can do interactive (Spark) and streaming (Storm/Spark)

Azure Data Lake Analytics

- Easiest experience for admin: no sense of clusters, instant scale per job
- Easiest experience for developers: Visual Studio/U-SQL (C#+SQL)
- Sell ADLA + ADLS
- Batch workloads only

HDInsight vs HDP on Azure VM

HDInsight	HDP on Azure VM
PaaS (setup, scale, manage, patch, etc)	IaaS
Managed by Microsoft	Managed by customer
Storage separate (Blob or ADLS)	Storage in VM (local disk), but can also have storage in Azure blob or ADLS
Delete VM keeps data	Delete VM deletes data (unless external)
Up to 30-days behind latest HDP version	Latest HDP Version
Limited Hadoop projects	Unlimited Hadoop projects
Microsoft supports VM and Hadoop	Microsoft: VM, HDP: Hadoop
No on-prem version	On-prem version

PolyBase use cases

Load data

Use Hadoop as an ETL tool to cleanse data before loading to data warehouse with PolyBase

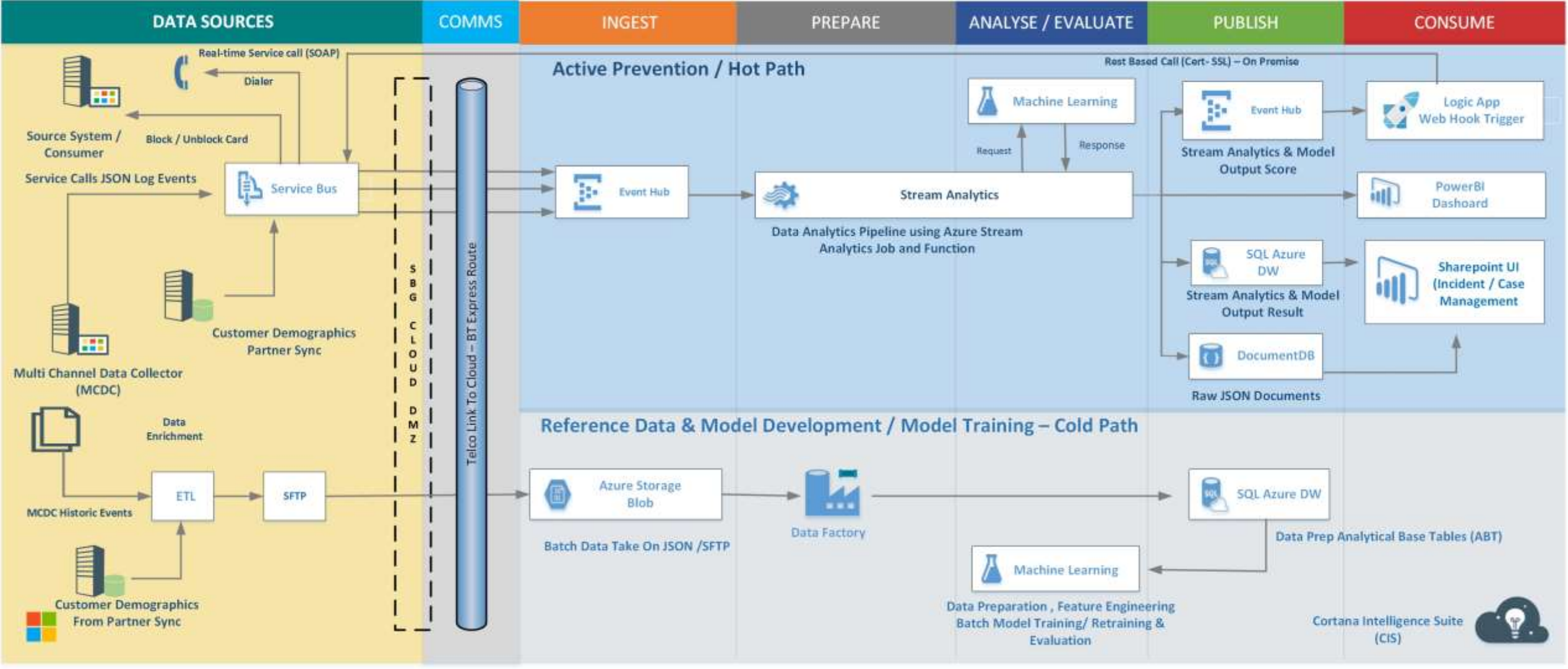
Interactively query

Analyze relational data with semi-structured data using split-based query processing (Federated query/ logical data warehouse)

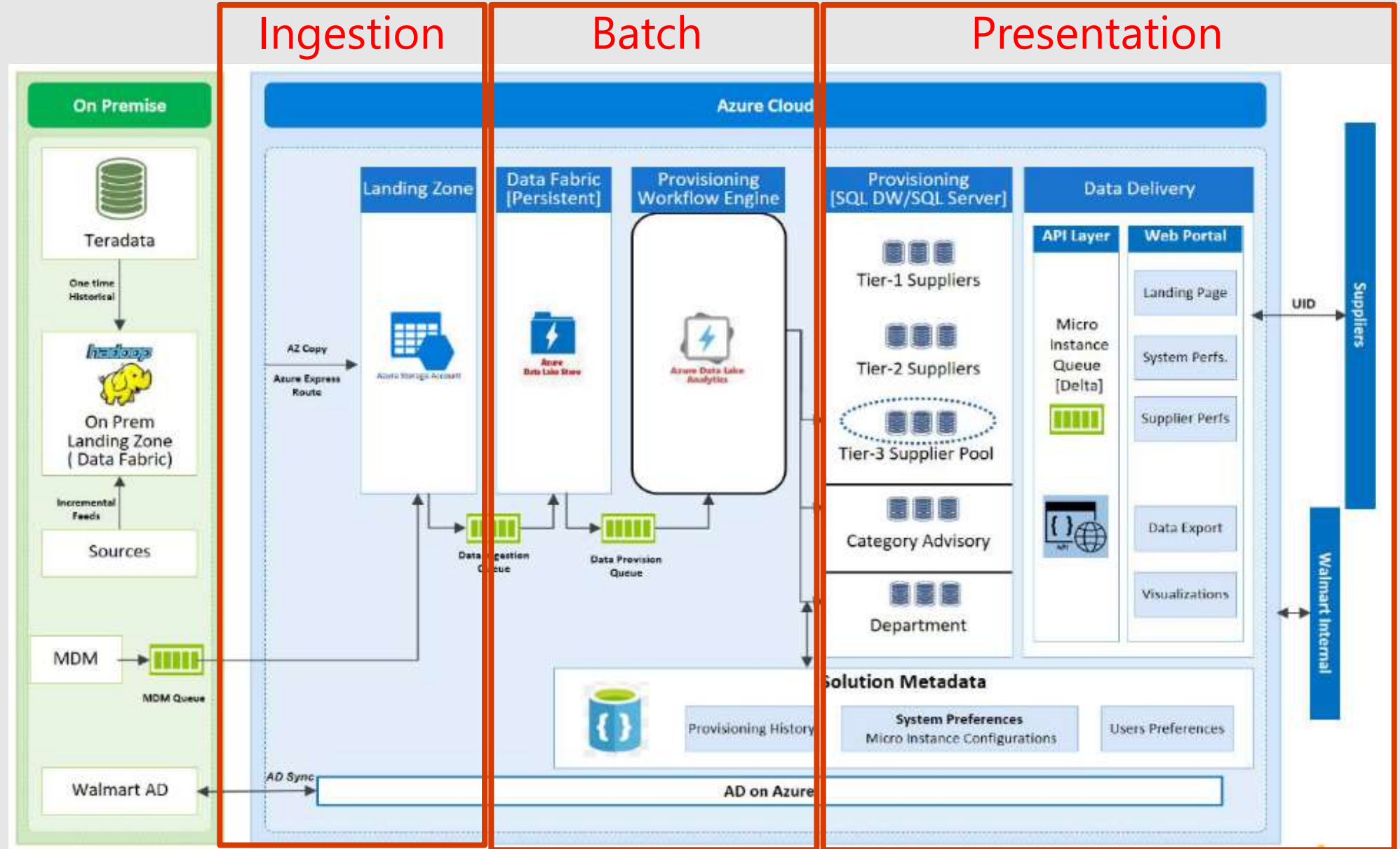
Age out data

Age out data to HDFS and use it as "cold" but queryable storage

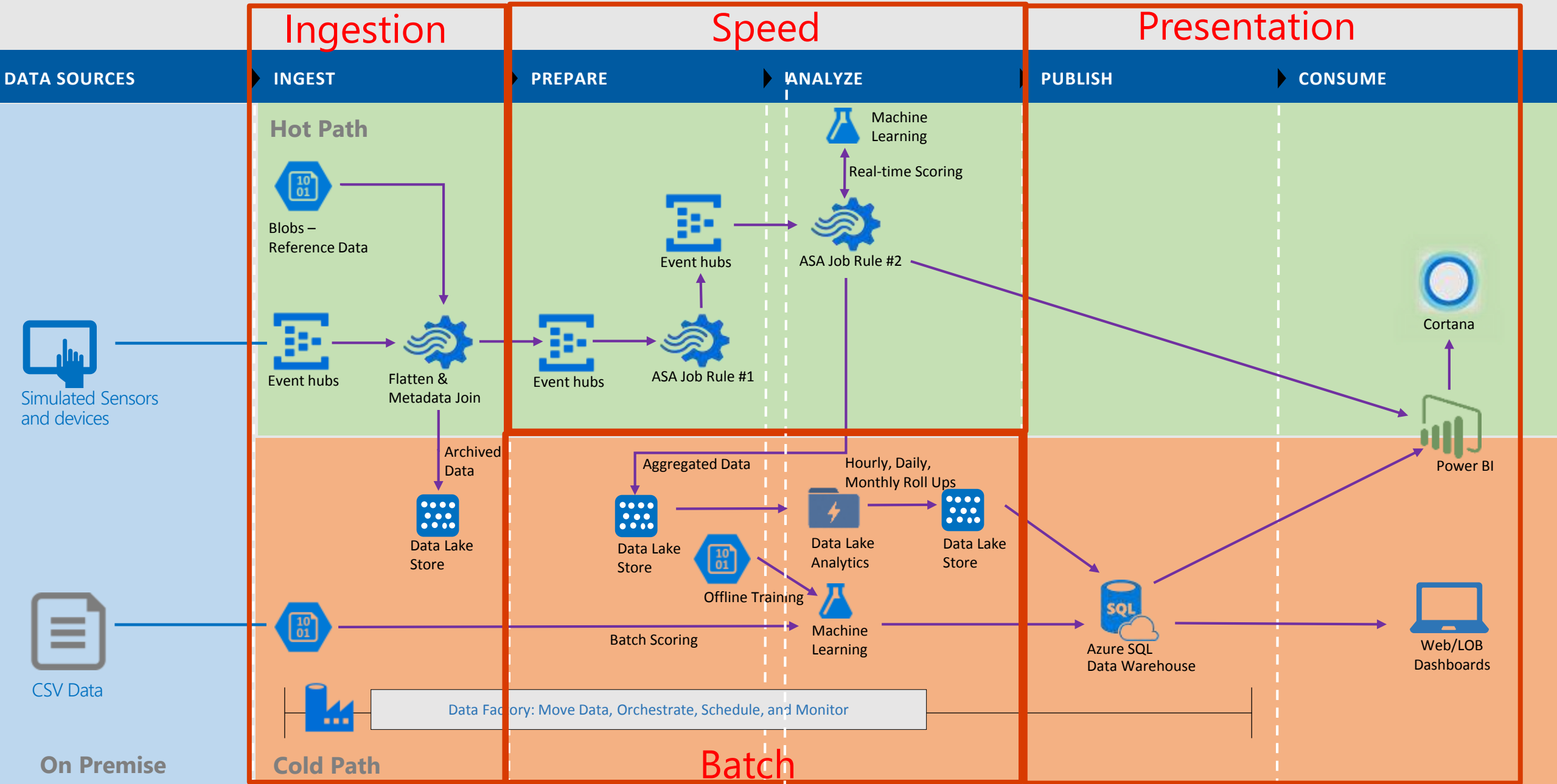
Company A



Company B

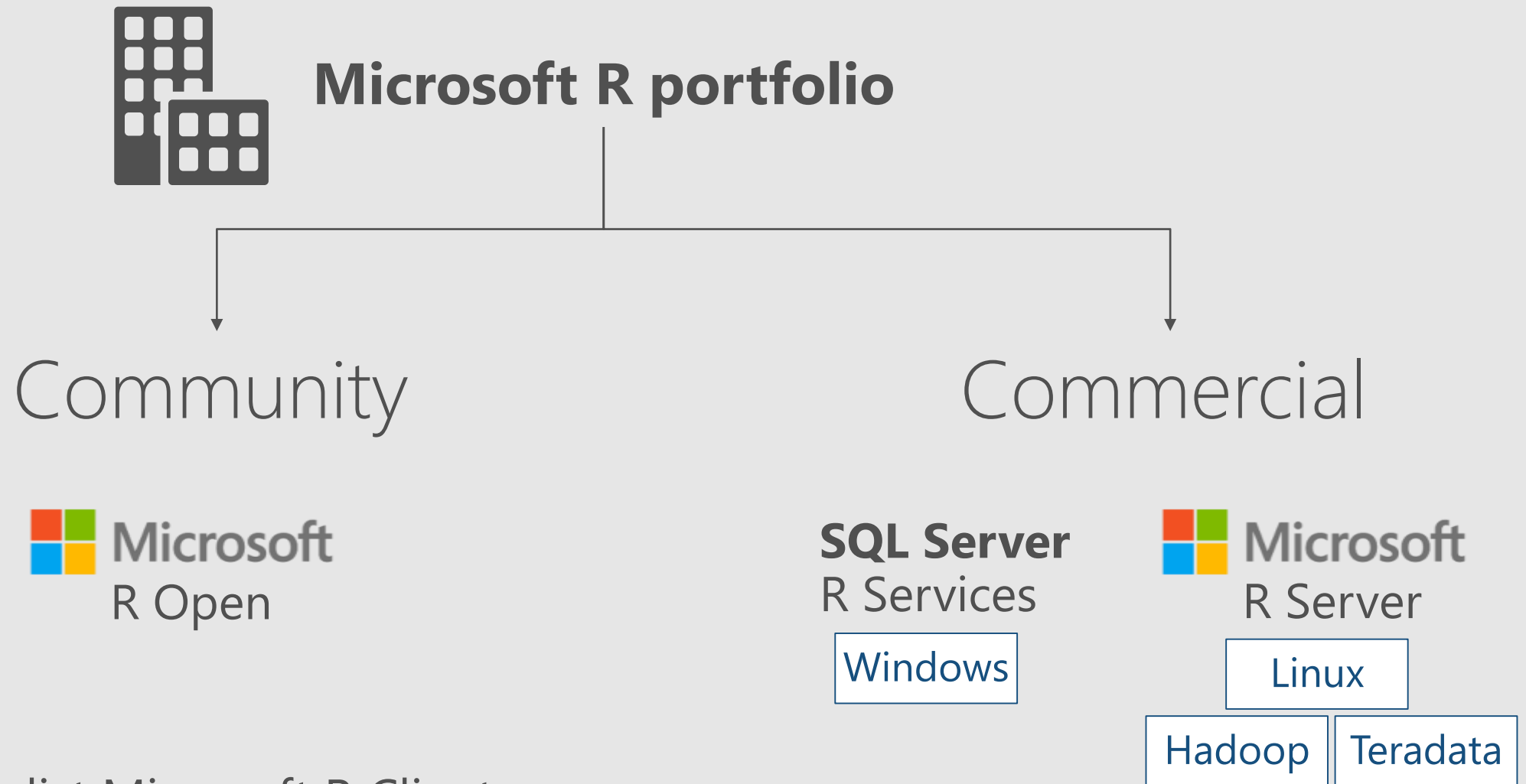


Company C



Microsoft R

Microsoft R portfolio



vs Azure ML, list Microsoft R Client

R vs Azure ML

R Open is Microsoft's open source version of R that (I believe) adds a few additional capabilities but largely mirrors the existing open source R.

Microsoft R Server (MRS) adds additional capability that is not available in open source R include R Scale for large scale deployment of R jobs on clusters (i.e. HDInsight) and Microsoft ML (MML) which is a library of Microsoft's best-in-breed ML algorithms available from within MRS only.

SQL Server R is a SQL product that adds the ability to apply R functions/algorithms to data operations performed from within SQL server.

Azure ML is a GUI-based product for building ML experiments and web services which includes access to most of the same underlying algorithms available programmatically in MML inside MRS.

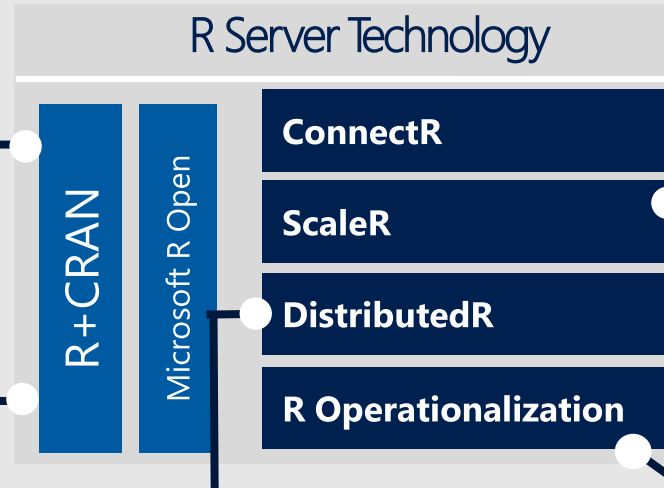
The Microsoft R Server Platform

R+CRAN

- Open source R interpreter
 - R 3.3.2
- Freely-available huge range of R algorithms
- Algorithms callable by MSR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages

RevoR

- Performance enhanced R interpreter
- Based on open source R
- Adds high-performance math library to speed up linear algebra functions



ScaleR

- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Range of predictive functions
- User tools for distributing customized R algorithms across nodes
- Wide data sets supported – thousands of variables

ConnectR

- High-speed & direct connectors

Available for:

- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS (text & XDF)
- Hive
- Parquet
- Teradata Database (TPT)
- EDWs and ADWs
- ODBC

DistributedR

- Distributed computing framework
- Delivers cross-platform portability

Microsoft R product comparison

	R Open	R Client	Microsoft R Server
Big Data	In-memory bound	In-memory bound Operates on large volumes when connected to R Server	Hybrid memory & disk scalability Operates on bigger volumes & factors
Speed of Analysis	Multiple-threaded when MKL installed	Same as R Open for non-ScaleR functions; Up to 2 threads for ScaleR functions when compute locally;	Parallel threading and Parallel Processing
Enterprise Readiness	Community support	Commercial support	Commercial support
Analytic Breadth & Depth	7000+ innovative analytic packages	Leverage and optimize open source packages plus Big Data ready packages (ScaleR APIs)	Leverage and optimize open source packages plus Multi threaded and Big Data ready packages
Commercial Viability	Risk of deployment of open source	Free for everyone	Commercial licenses

CRAN, MRO, MRS Comparison



Datasize	In-memory	In-memory	In-Memory or Disk Based
Speed of Analysis	Single threaded	Multi-threaded	Multi-threaded, parallel processing 1:N servers
Support	Community	Community	Community + Commercial
Analytic Breadth & Depth	10k+ innovative analytic packages	10k+ innovative analytic packages	10k+ innovative packages + commercial parallel high-speed functions
License	Open Source	Open Source	Commercial license. Supported release with indemnity

Why R Server with Spark?

