.NET Conf

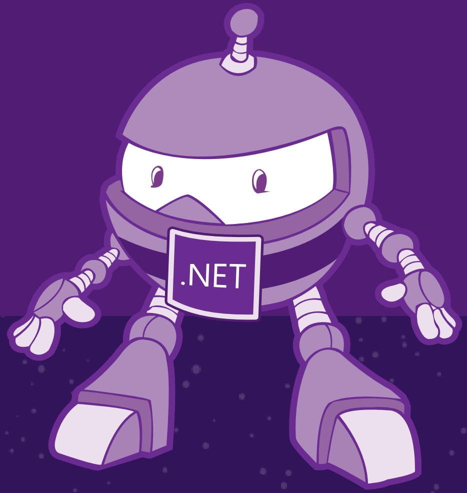
探索 .NET 新世界

Host by

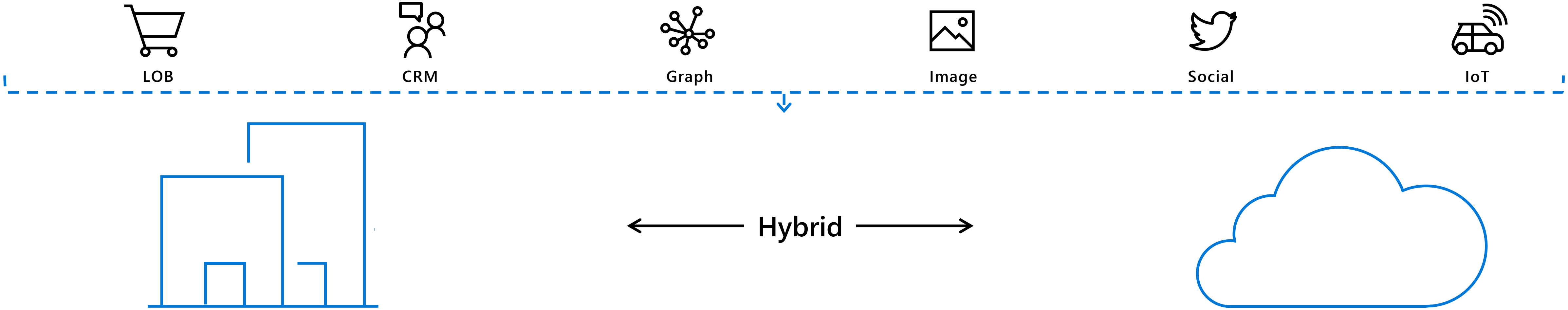
STUDY4

深入淺出 Modern Data Warehouse

John Chang
資深軟體開發協理
Microsoft



The modern data estate



LOB CRM Graph Image Social IoT

Hybrid

Operational databases

Data warehouses

Data lakes

Operational databases

Data warehouses

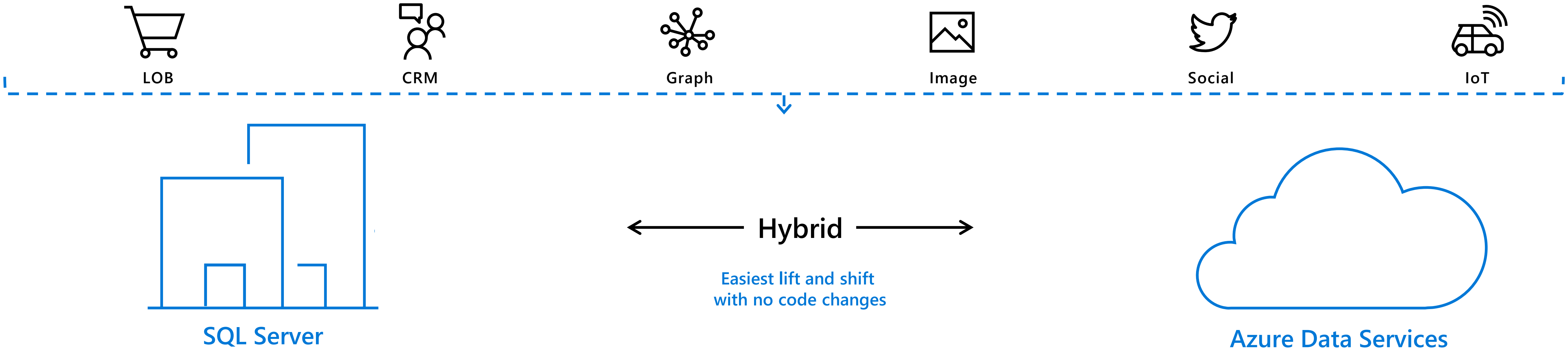
Data lakes

Reason over any data, anywhere

Flexibility of choice

Security and performance

The Microsoft offering



LOB CRM Graph Image Social IoT

SQL Server

Hybrid

Easiest lift and shift
with no code changes

Azure Data Services

Industry leader 4 years in a row	Operational databases
#1 TPC-H performance	Data warehouses
T-SQL query over any data	Data lakes

Operational databases	70% faster
Data warehouses	2x the global reach
Data lakes	99.9% SLA

AI built-in | Most secure | Lowest TCO

Reason over any data, anywhere

Flexibility of choice

Security and performance

Understanding the customer landscape for Big Data and advanced analytics

Our customers

Traditionalists

Have strong processes and practices, need prescriptive guidance, mature stack players

Early adopters

Have lean structures and are agile and flexible, usually on the latest-and-greatest tech

Understanding the Azure portfolio for Big Data and advanced analytics

The Azure data landscape

Azure Data Factory

Azure Import/Export service

Azure CLI

Azure SDK

Azure IoT Hub

Azure event hubs

Kafka on Azure HDInsight

Azure SQL DB

Azure Cosmos DB

Azure Blob Storage

Azure Data Lake Store

Azure Search

Azure Data Catalog

Azure SQL data warehouse

Azure Data Lake Analytics

Azure HDInsight

Azure Databricks

Azure Stream Analytics

Azure HDInsight

Azure Databricks

Azure Analysis Services

Power BI

Azure ML

ML Server

Azure Databricks

Bot service

Cognitive services

Azure ExpressRoute

Azure Active Directory

Azure network security groups

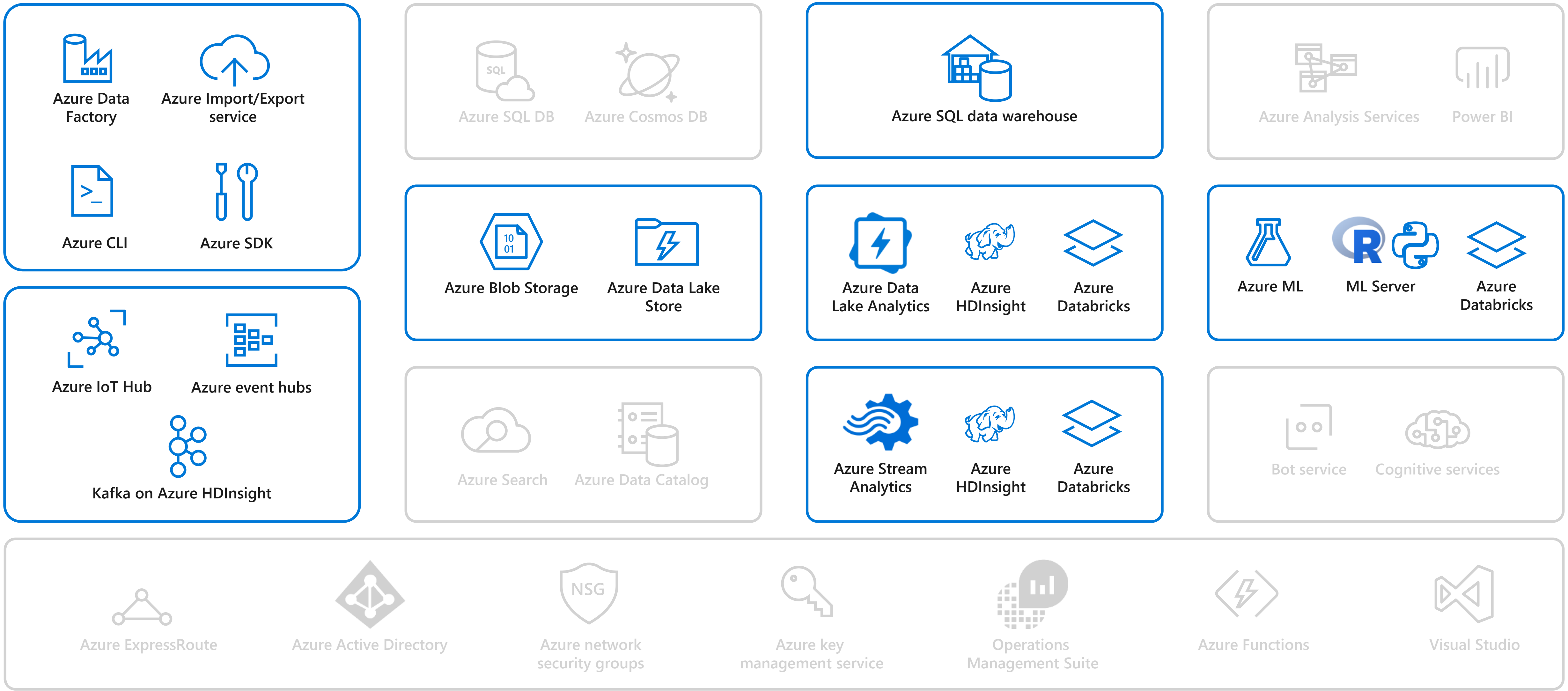
Azure key management service

Operations Management Suite

Azure Functions

Visual Studio

The Azure Big Data landscape



Azure Data Factory

Azure Import/Export service

Azure CLI

Azure SDK

Azure IoT Hub

Azure event hubs

Kafka on Azure HDInsight

Azure SQL DB

Azure Cosmos DB

Azure Blob Storage

Azure Data Lake Store

Azure Search

Azure Data Catalog

Azure SQL data warehouse

Azure Data Lake Analytics

Azure HDInsight

Azure Databricks

Azure Stream Analytics

Azure HDInsight

Azure Databricks

Azure Analysis Services

Power BI

Azure ML

ML Server

Azure Databricks

Bot service

Cognitive services

Azure ExpressRoute

Azure Active Directory

Azure network security groups

Azure key management service

Operations Management Suite

Azure Functions

Visual Studio

Common Big Data and advanced analytics scenarios

Solution scenarios

Big Data and advanced analytics

Modern data warehousing

“We want to integrate all our data—including Big Data—with our data warehouse”

Advanced analytics

“We’re trying to predict when our customers churn”

Real-time analytics

“We’re trying to get insights from our devices in real-time”

Things to note

- There are no right or wrong solutions, only optimal solutions
- We lead with certain solutions and customize based on customer scenarios
- Customer voice and product and service maturity govern lead solutions
- Consider price and performance, ease of use, and ecosystem acceptance as factors
- Competitor SWOT also plays a role
- Everything is fluid - a lead solution today might be non-optimal tomorrow, based on the factors above and new releases

Modern data warehousing

The modern data warehouse extends the scope of the data warehouse to serve Big Data that's prepared with techniques beyond relational ETL

SQL

Modern data warehousing

"We want to integrate all our data—including Big Data—with our data warehouse"

Advanced analytics

"We're trying to predict when our customers churn"

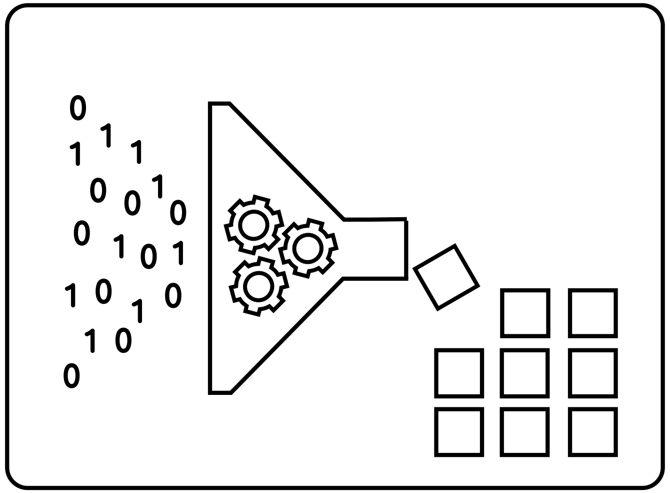
Real-time analytics

"We're trying to get insights from our devices in real-time"

Modern data warehousing

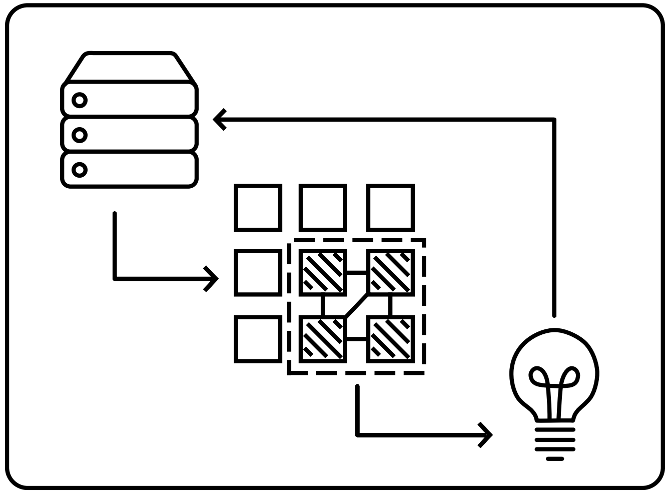
Canonical operations

Load and ingest



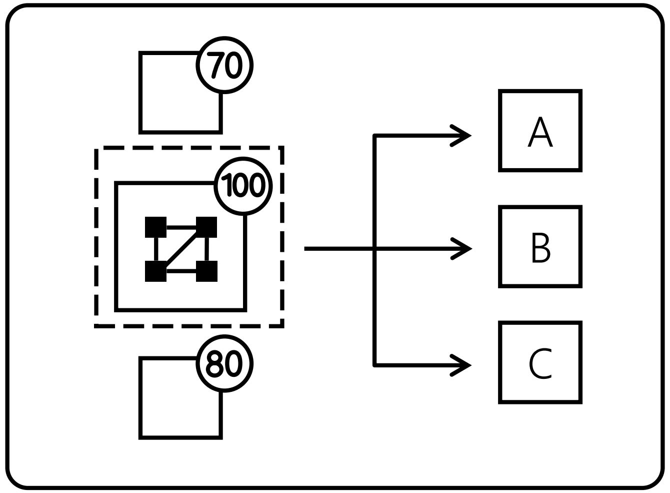
Transfer and store

Process



Process and clean

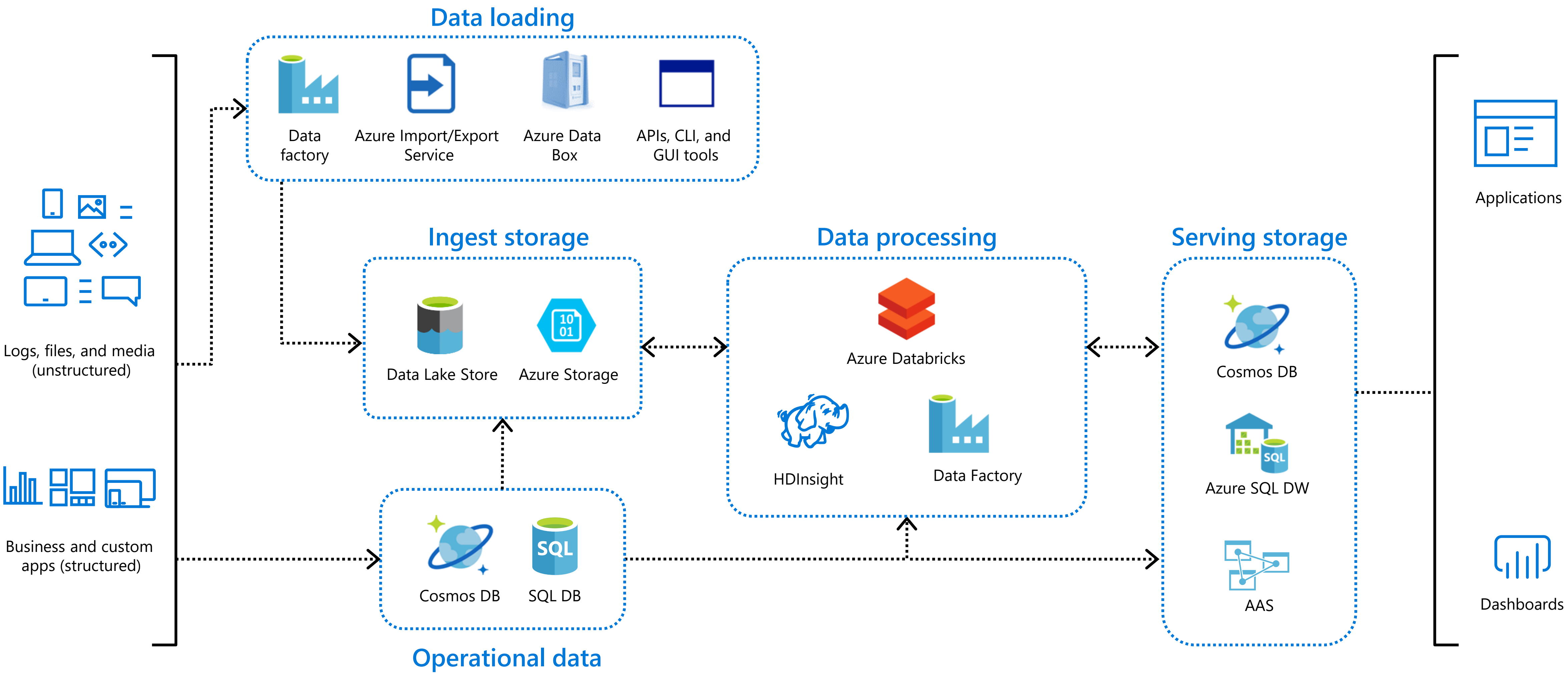
Serve



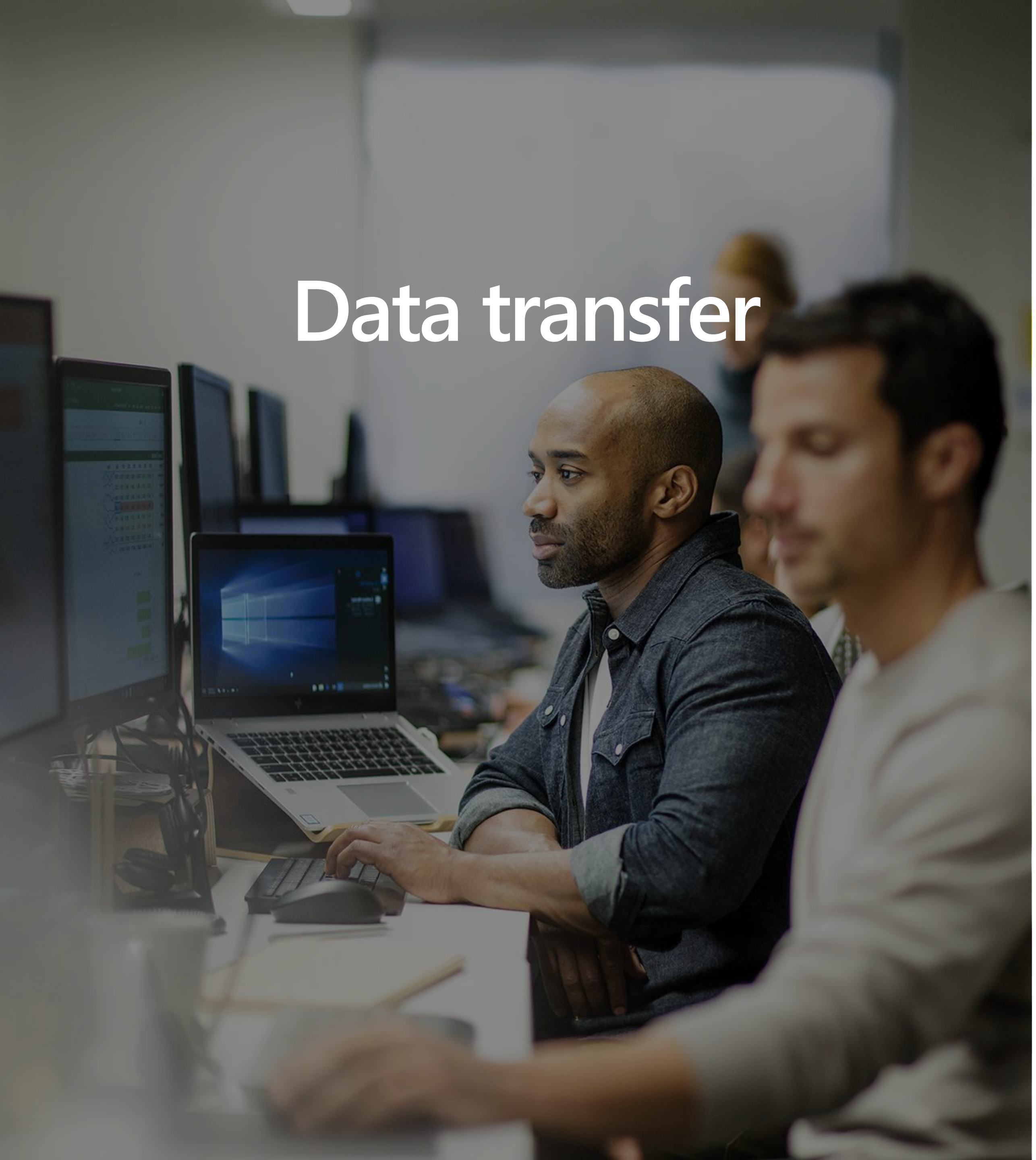
Serve and analyze

Data warehousing pattern in Azure

Loading and preparing data for analysis with a data warehouse



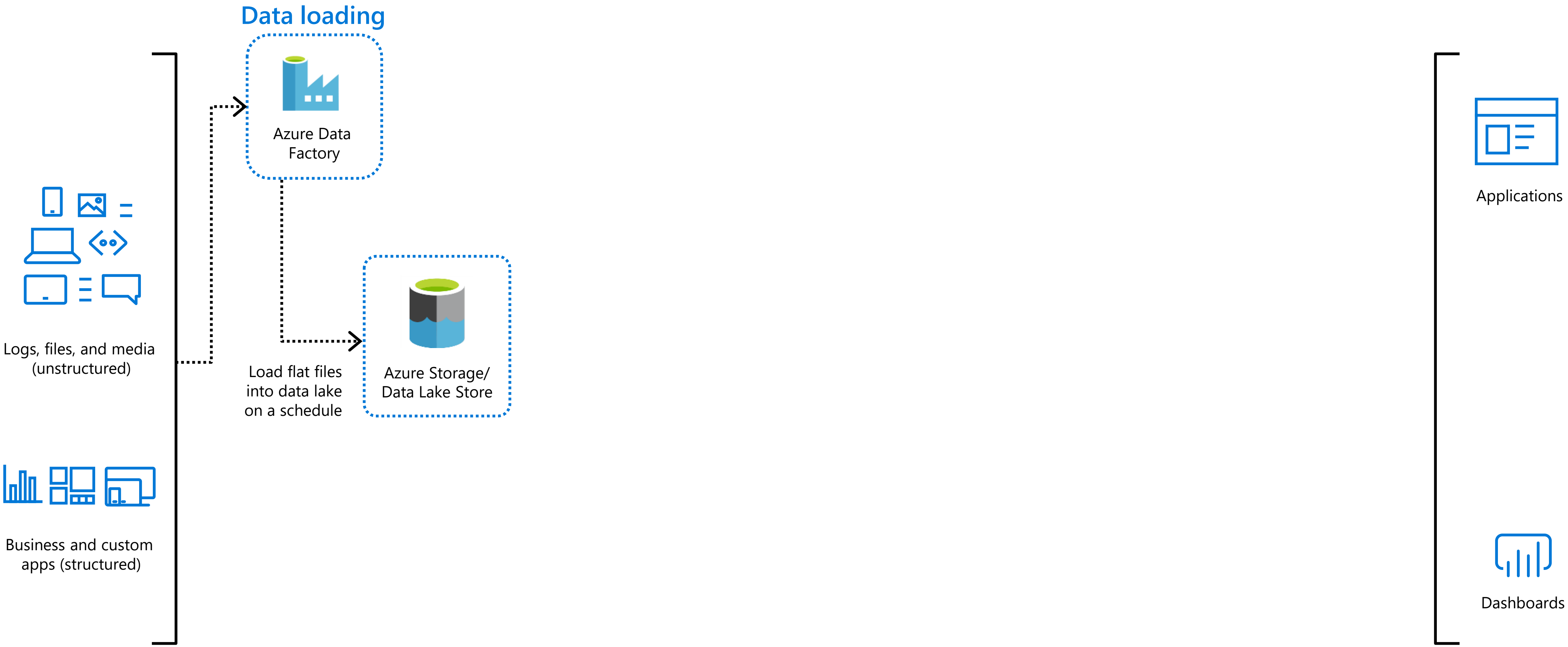
Data transfer



The process and tools used to move data from the source to the initial destination for processing

Data warehousing pattern in Azure

Loading data into ingest storage



Data loading

Logs, files, and media (unstructured)

Business and custom apps (structured)

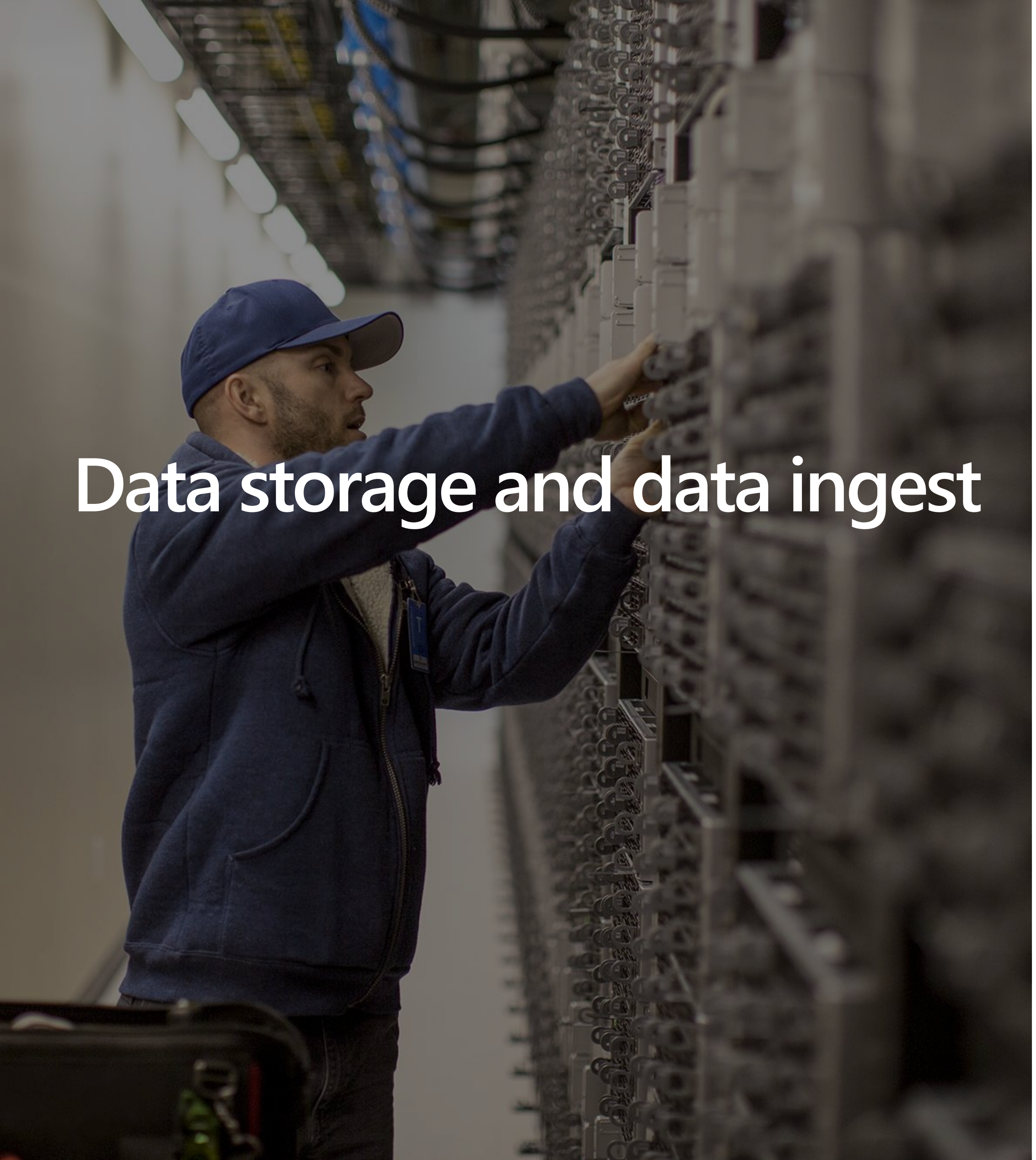
Azure Data Factory

Load flat files into data lake on a schedule

Azure Storage/ Data Lake Store

Applications

Dashboards



Data storage and data ingest

The storage that persists the transferred data and is consumed by subsequent processing

File storage requirements

Requirement	Comment
Capacity	Should be able to store terabytes or petabytes of data economically. File storage should be able to store any number of objects and associated metadata.
Performance	Should be able to store the incoming data as fast as it arrives. Should support high bandwidth, high throughput, and low-latency writes.
Multiple tiers	Should support storing data for extended periods of time— that is months or years—economically. Should have with multiple storage tiers (hot, cold, and archival).
Multiple object sizes	Objects can be stored individually—could be as small as few tens of bytes—or as data sets that create large objects—from several gigabytes to terabytes.
Replication	Should provide data replication that suits your needs for a combination of durability, bandwidth, and data governance requirements. Locally redundant storage provides the highest maximum bandwidth, with the least durability, whereas geo-redundancy options provide higher durability with possible asynchronous replication delay.

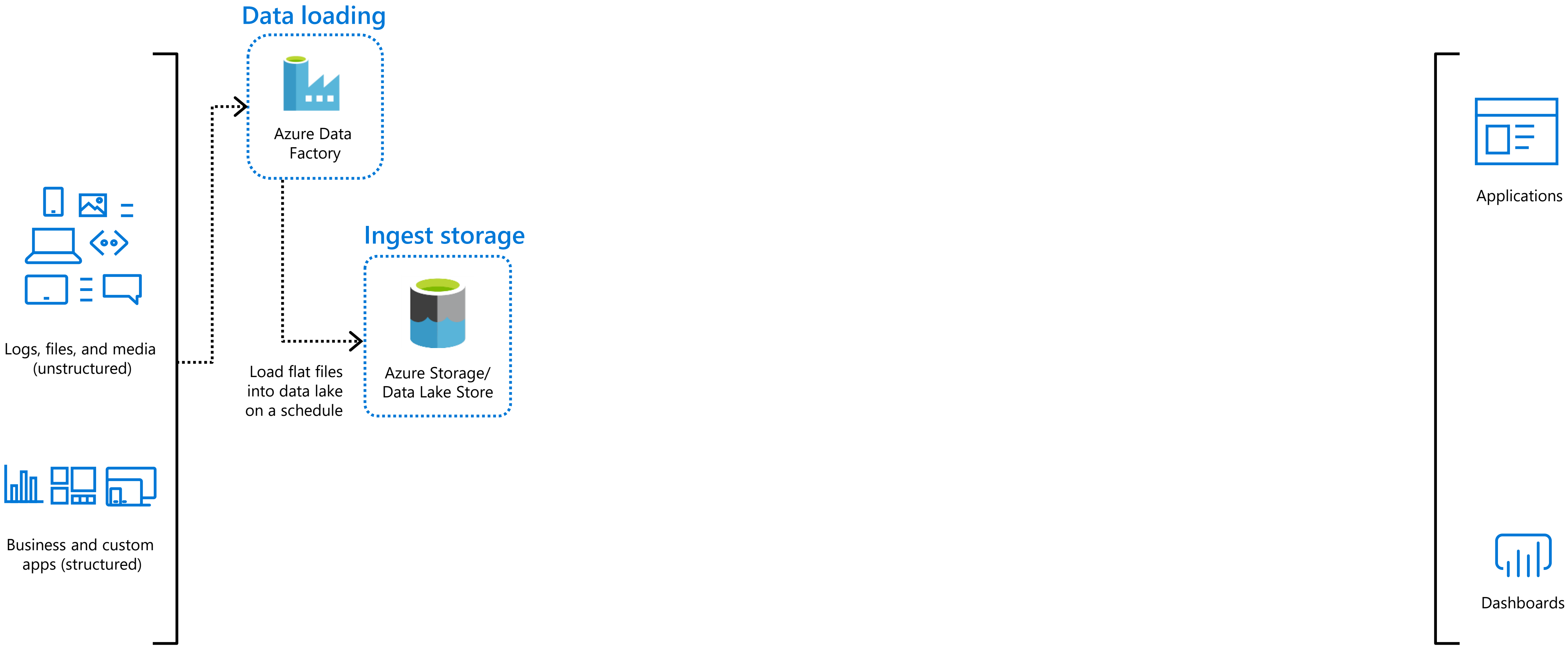
File storage

A side-by-side comparison of the capabilities and features

	Azure Data Lake Store	Azure Blob Storage containers
Purpose	Optimized storage for Big Data analytics workloads	General purpose object store for a wide variety of storage scenarios
Structure	Hierarchical file system	Object store with flat namespace
API	REST API over HTTPS	REST API over HTTP/HTTPS
Analytics workload performance	Optimized performance for parallel analytics workloads, high throughput and IOPS	Not optimized for analytics workloads
Size limits	No limits on account sizes, file sizes, or number of files	Max 500 TB per account and 4.75 TB per file
Geo-redundancy	Locally-redundant (multiple copies of data in one Azure region)	Locally redundant (LRS), globally redundant (GRS), and read-access globally redundant (RA-GRS). See Azure Storage replication for more information
Service state	Generally available	Generally available
Regional availability	Some regions	All regions

Data warehousing pattern in Azure

Loading data into ingest storage



Data loading

Azure Data Factory

Ingest storage

Azure Storage/ Data Lake Store

Load flat files into data lake on a schedule

Logs, files, and media (unstructured)

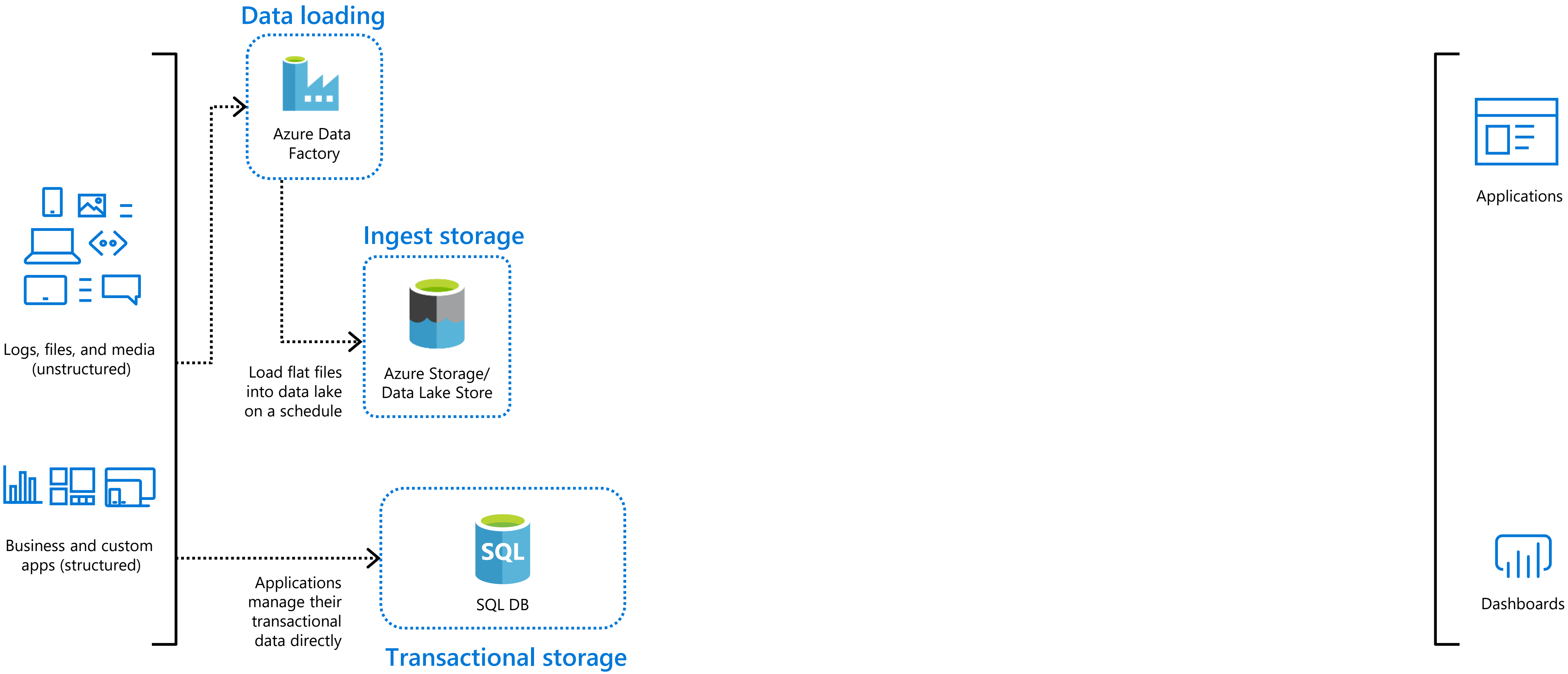
Business and custom apps (structured)

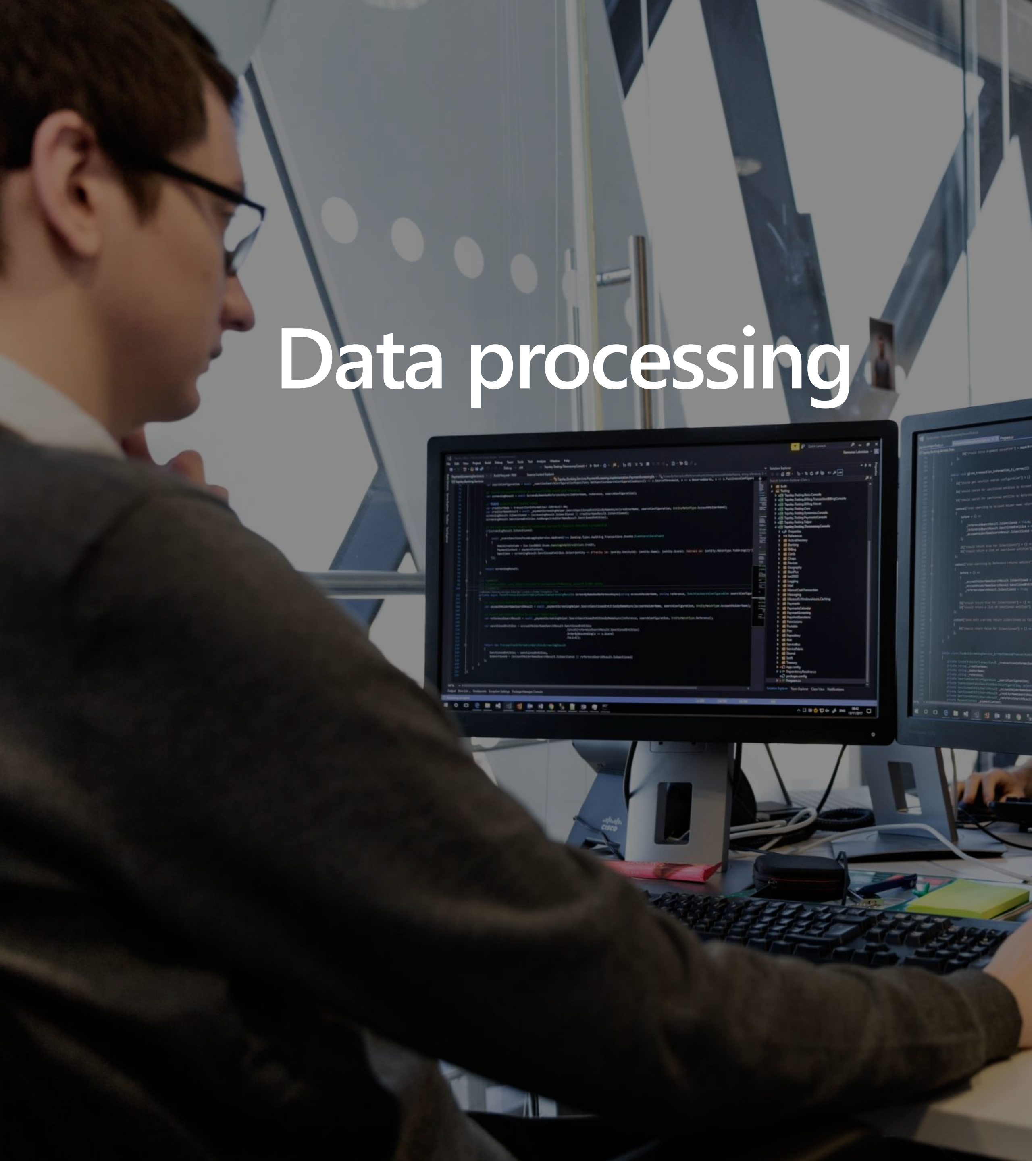
Applications

Dashboards

Data warehousing pattern in Azure

Load data into multiple source data stores


Data loading
Azure Data Factory
Ingest storage
Azure Storage/ Data Lake Store
Logs, files, and media (unstructured)
Load flat files into data lake on a schedule
Business and custom apps (structured)
Applications manage their transactional data directly
SQL DB
Transactional storage
Applications
Dashboards



Data processing

Is data cleansing, structuring, curation, and aggregation

In data warehousing, the data is batch processed in preparation for loading into a data warehouse

Data processing requirements

Requirement	Comment
Scalability	The amount of data that needs to be processed at once can vary widely on any given day, or can grow over time. The batch data processing technology should scale to meet your needs with the level of granularity and within an acceptable time range for your solution.
Choice of language	The batch data processing technology should provide a choice of languages with which to create batch operations, including Python, Java, U-SQL, HiveQL, and R.
Integration choices	The batch data processing technology should allow you to choose a cloud-based data source from which to query, such as Azure Storage or Azure Data Lake Store. Some options allow additional integration options by also allowing you to query external relational data stores, such as SQL Data Warehouse.

Batch data processing

A side-by-side comparison of general capabilities and features

	Azure Data Lake Analytics	HDInsight with Spark	HDInsight with Hive	HDInsight with Hive LLAP	SQL Data Warehouse	Azure Databricks
Is a managed service	Yes	Yes	Yes	Yes	Yes	Yes
Auto-scaling	No	No	No	No	No	Yes
Supports pausing compute	No	No	No	No	Yes	Yes
Programmability	U-SQL	Python, Scala, Java, R, SQL	HiveQL	HiveQL	T-SQL	Python, Scala, Java, SQL, R
Programming paradigm	Mixture of declarative and imperative	Mixture of declarative and imperative	Declarative	Declarative	Declarative	Mixture of declarative and imperative
Pricing model	Per job (by job run per hour times analytics unit used)	By cluster hour	By cluster hour	By cluster hour	By cluster hour	By cluster hour

Batch data processing

A side-by-side comparison of integration capabilities

	Azure Data Lake Analytics	HDInsight with Spark	HDInsight with Hive	HDInsight with Hive LLAP	SQL Data Warehouse	Azure Databricks
Access Azure Data Lake Store	Yes	Yes	Yes	Yes	Yes	Yes
Query Azure Storage	Yes	Yes	Yes	Yes	Yes	Yes
Query external relational stores (like Azure SQL Database, SQL Server in virtual machine, or Azure SQL Data Warehouse)	Yes	Yes	Yes	No	Yes	Yes

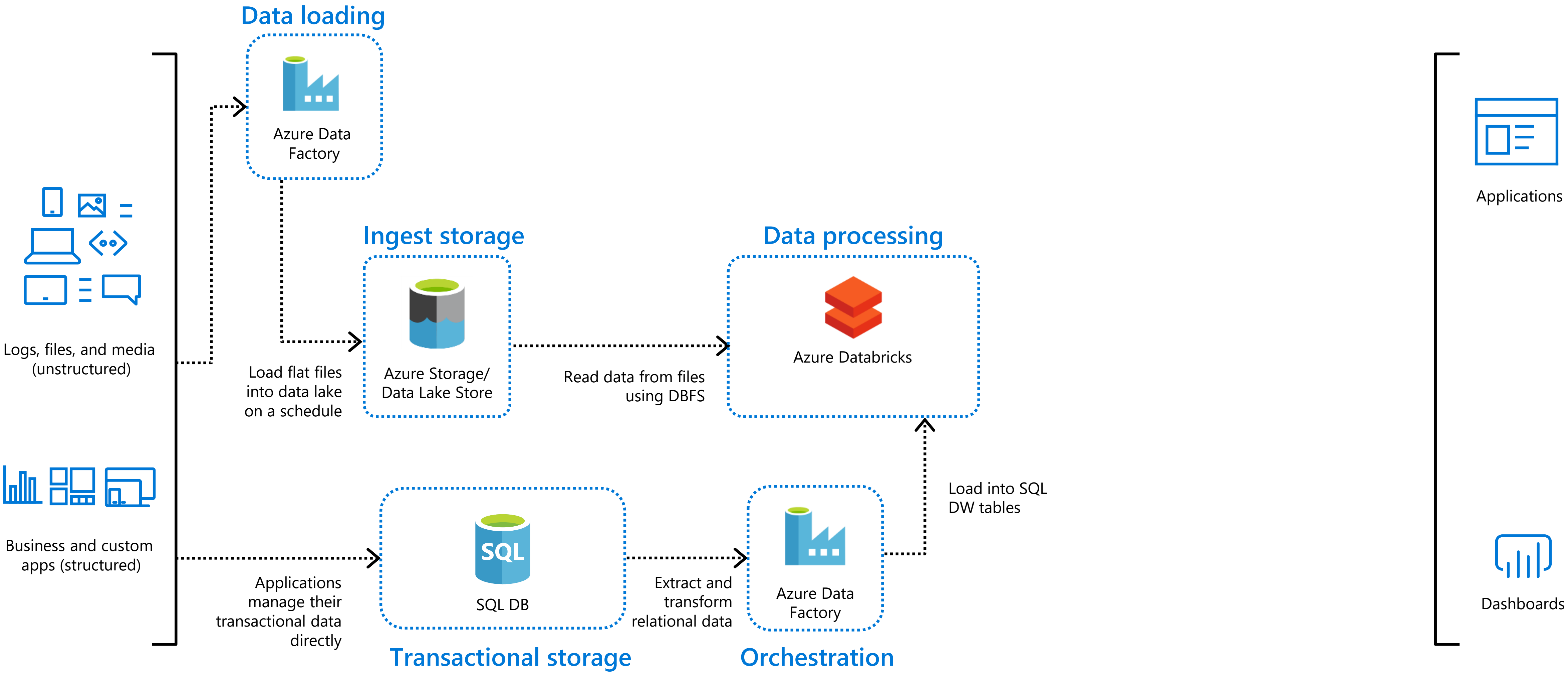
Batch data processing

A side-by-side comparison of scalability capabilities

	Azure Data Lake Analytics	HDInsight with Spark	HDInsight with Hive	HDInsight with Hive LLAP	SQL Data Warehouse	Azure Databricks
Scale-out granularity	Per job	Per cluster	Per cluster	Per cluster	Scale out by compute units (DWU)	Per cluster
Supports fast scale out (less than 1 minute)	Yes	No	No	No	No	Yes
Supports in-memory caching of data	No	Yes	No	Yes	Yes	Yes

Data warehousing pattern in Azure

Data processing with Azure Databricks


Data loading
Azure Data Factory
Logs, files, and media (unstructured)
Load flat files into data lake on a schedule
Ingest storage
Azure Storage/ Data Lake Store
Read data from files using DBFS
Data processing
Azure Databricks
Applications
Load into SQL DW tables
Business and custom apps (structured)
Applications manage their transactional data directly
SQL
SQL DB
Extract and transform relational data
Azure Data Factory
Dashboards
Transactional storage
Orchestration

Data serving

Processed data served by a data warehouse to analytic clients and reporting tools

The data warehouse provides increased query flexibility and reduced query latency in comparison to batch data processing options

Data serving

A side-by-side comparison of general capabilities and features

	SQL Database	SQL Data Warehouse	Azure Analysis Services
Is a managed service	Yes (Azure SQL Database)	Yes	Yes
Primary database model	Relational (columnar format when using columnstore indexes)	Relational tables with columnar storage	Tabular and MOLAP semantic models
SQL language support	Yes	Yes	No
Optimized for speed serving layer	Yes, using memory-optimized tables and hash or nonclustered indexes	No	No

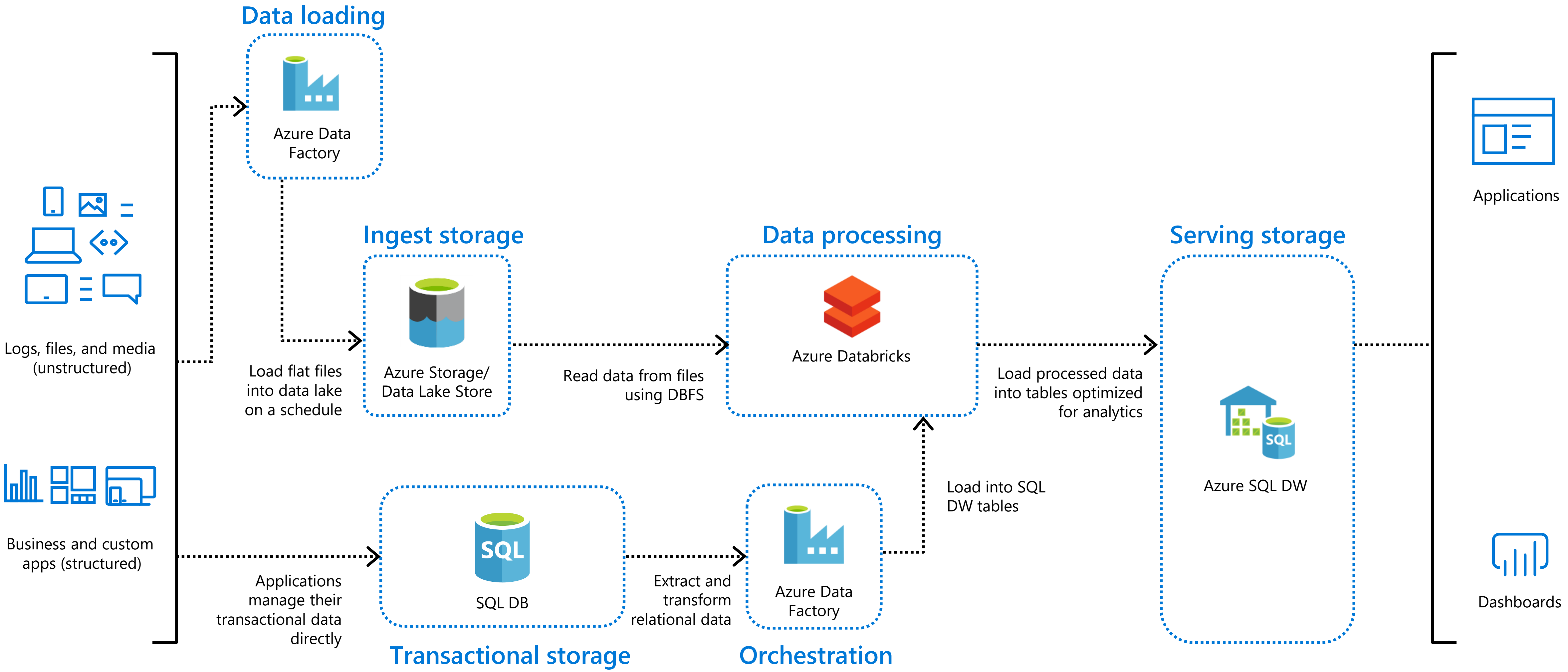
Data serving

A side-by-side comparison of scalability capabilities

	SQL Database	SQL Data Warehouse	Azure Analysis Services
Redundant regional servers for high availability	Yes (Azure SQL Database)	Yes	No
Supports query scale out	No	Yes	Yes
Dynamic scalability (scale up)	Yes (Azure SQL Database)	Yes	Yes
Supports in-memory caching of data	Yes	Yes	Yes

Data warehousing pattern in Azure

Data processing with Azure Databricks


Data loading
Azure Data Factory
Logs, files, and media (unstructured)
Load flat files into data lake on a schedule
Ingest storage
Azure Storage/ Data Lake Store
Read data from files using DBFS
Data processing
Azure Databricks
Load processed data into tables optimized for analytics
Serving storage
Azure SQL DW
Applications
Dashboards
Business and custom apps (structured)
Applications manage their transactional data directly
SQL
SQL DB
Transactional storage
Extract and transform relational data
Azure Data Factory
Orchestration
Load into SQL DW tables
SQL

Hybrid architectures

Enable the data storage, processing, and serving to span on-premises and cloud environments

Hybrid architectures

A side-by-side comparison of the connectivity options

	Dedicated low latency (5 to 10 ms)	Secure transfer	Reliability
Public internet		X	Good
VPN		X	Good
ExpressRoute	X	X	Best

Choosing Azure data factory

When Azure Data Factory can be a good option for your hybrid data pipelines

When you want…	Description
To orchestrate your data pipeline on-premises and in the cloud	Use Data Management Gateway and Azure Data Factory to move data between both cloud and self-hosted environments. Data is compressed and transferred in parallel and resilient to intermittent network issues through auto retry logic. You can connect on-premises data to cloud services to benefit from cloud services while keeping the business running with on-premises data.
To execute your SQL Server Integration Services (SSIS) packages in the cloud	When you provision an Azure and SSIS integration runtime (IR) in Azure Data Factory, you can deploy your SSIS packages to the runtime in Azure. Azure Data Factory orchestrates the SSIS package execution, which creates new opportunities for shifting existing on-premises data workflows to Azure.
To move your non-relational data to Azure for processing and transformation	Create and schedule data-driven workflows—pipelines—in Azure Data Factory that move your non-relational and unstructured data to Azure, then process and transform the data using compute services such as Azure HDInsight Hadoop, Spark, Azure Data Lake Analytics, and Azure Machine Learning.

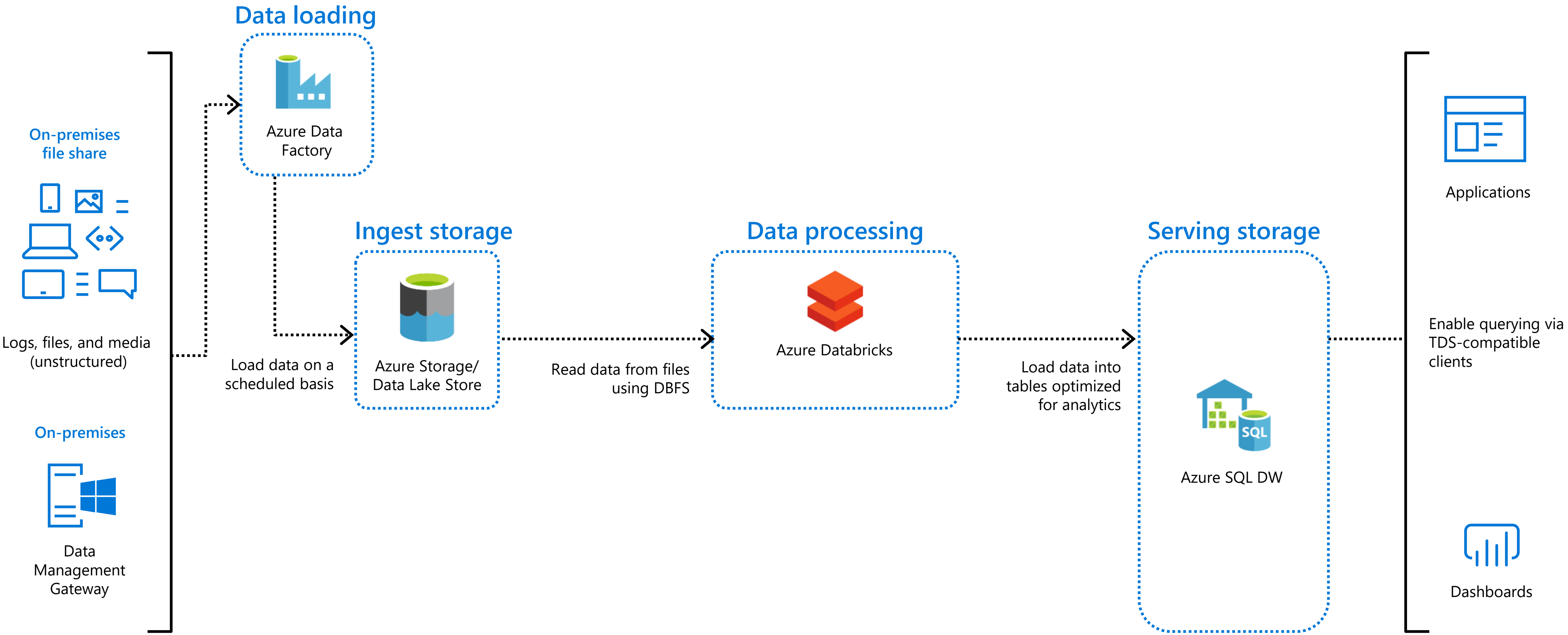
Choosing SQL server stretch database

When SQL Server Stretch Database can be a good option for your hybrid architecture

When you want…	Description
Cost-effective availability for cold data	Stretch warm and cold transactional data dynamically from your on-premises SQL Server to Microsoft Azure with Stretch Database. Unlike typical cold data storage, your data is always online and available to query. You can provide longer data retention timelines without breaking the bank for large tables like customer order history. Benefit from the low cost of Azure rather than scaling expensive, on-premises storage. You choose the pricing tier and configure settings in the Azure Portal to maintain control over price and costs.
Access to your SQL data, regardless of location, without changes to your queries or applications	Access your SQL Server data seamlessly regardless of whether it's on-premises or stretched to the cloud. You set the policy that determines where data is stored, and SQL Server handles the data movement in the background. The entire table is always online and query- able. Stretch Database doesn't require any changes to existing queries or applications—the location of the data is completely transparent to the application.
Streamlined on-premises data maintenance	Reduce on-premises maintenance and storage for your data. Backups for your on-premises data run faster and finish within the maintenance window. Backups for the cloud portion of your data run automatically. Your on-premises storage needs are greatly reduced. Azure storage can be 80 percent less expensive than adding to on-premises SSD.

Data warehousing pattern in Azure

Loading data from on-premises sources



Data loading
Azure Data Factory
On-premises file share
Logs, files, and media (unstructured)
On-premises
Data Management Gateway
Load data on a scheduled basis
Ingest storage
Azure Storage/ Data Lake Store
Read data from files using DBFS
Data processing
Azure Databricks
Load data into tables optimized for analytics
Serving storage
Azure SQL DW
SQL
Enable querying via TDS-compatible clients
Applications
Dashboards

Security

Enables the data warehouse to control access in order to protect sensitive data and maintain desired compliance

Data storage security

A side-by-side comparison of the capabilities and features

	Azure Data Lake Store	Azure Blob Storage containers
API	REST API over HTTPS	REST API over HTTP/HTTPS
Data operations: Authentication	Based on Azure Active Directory Identities	Based on shared secrets account access keys and shared access signature keys, and role-based access control (RBAC)
Data operations: Authorization	POSIX access control lists (ACLs). ACLs based on Azure Active Directory identities can be set at file and folder level	For account-level authorization use account access keys. For account, container, or blob authorization use shared access signature keys
Encryption data at rest	Transparent, server side With service-managed keys With customer-managed keys in Azure Key Vault	Transparent, server side With service-managed keys With customer-managed keys in Azure Key Vault (coming soon) Client-side encryption
Management operations (for example, account create)	Role-based access control (RBAC) provided by Azure for account management	Role-based access control (RBAC) provided by Azure for account management

Batch data processing security

A side-by-side comparison of the capabilities and features

	Azure Data Lake Analytics	HDInsight with Spark	Apache Hive on HDInsight	Hive LLAP on HDInsight	Azure Databricks
Authentication	Azure Active Directory	No	Local/Azure Active Directory *	Local/Azure Active Directory *	Azure Active Directory (native/built-in)
Authorization	Yes	No	Yes *	Yes *	Yes
Auditing	Yes	No	Yes *	Yes *	Yes
Data encryption at rest	Yes	Yes	Yes	Yes	Yes
Row-level security	No	No	Yes *	Yes *	No
Supports firewalls	Yes	Yes	Yes ***	Yes ***	Coming soon
Dynamic data masking	No	No	Yes *	Yes *	No

* Requires using a domain-joined HDInsight cluster
** Requires using Transparent Data Encryption (TDE) to encrypt and decrypt your data at rest
*** Supported when used within an Azure virtual network

Data serving security

A side-by-side comparison of the capabilities and features

	SQL Database	SQL Data Warehouse	Azure Analysis Services	Azure Cosmos DB
Authentication	SQL/Azure Active Directory	SQL/Azure Active Directory	Azure Active Directory	Database users and Azure Active Directory via access control (IAM)
Authorization	Yes	Yes	Yes	Yes (hash-based message authentication code (HMAC))
Auditing	Yes	Yes	Yes (when integrated with Azure Monitor resource diagnostic logs)	Yes (through audit logging and activity logs)
Data encryption at rest	Yes **	Yes **	Yes	Yes
Row-level security	Yes	No	Yes (through object-level security in model)	No
Supports firewalls	Yes	Yes	Yes	Yes
Dynamic data masking	Yes	No	No	No

** Requires using transparent data encryption (TDE) to encrypt and decrypt your data at rest

*** Supported when used within an Azure virtual network

Automation

Enables all components of the data warehouse solution to be controlled, deployed, and monitored programmatically

Choosing Azure automation

When Azure Automation can be a good option for cloud-based automation

When you want…	Description
Process automation	Automate frequent, time-consuming, and error-prone cloud management tasks by authoring runbooks in a graphical UI, in PowerShell, or in Python.
Configuration management	Manage your desired state configuration (DSC) resources and apply configurations to virtual or physical machines in Azure. Monitor and automatically update machine configuration across physical and virtual machines, Windows or Linux, in the cloud or on-premises. Collect inventory about in-guest resources and track changes across services, daemons, software, registry, and files.
Update management	Update Windows and Linux systems across hybrid environments. Gain visibility of update compliance across Azure, on-premises and in other clouds. Schedule deployments to orchestrate installation of updates within a defined maintenance window.
Build and deploy resources	Deploy Azure resources using Runbooks and Azure Resource Manager (ARM) templates. Integrate into development tools like Jenkins and Visual Studio Team Services, ensuring continuous delivery and operations automation.

Choosing Azure resource manager templates

When Azure Resource Manager (ARM) templates can be a good option for cloud-based automation

When you want...	Description
To consistently and repeatedly deploy resources	ARM templates are composed of a JavaScript Object Notation (JSON) file that defines one or more resources, including any dependencies between them. This adds the benefit of treating your resources for a solution as a single unit, rather than independent components, making it easier to consistently deploy and manage the resources to development, test, staging, and production environments.
To manage your infrastructure through declarative templates rather than scripts	Declarative templates make it easier to define your resource parameters, dependencies, and infrastructure, compared to executing a series of scripts. Furthermore, you can apply tags to resources to logically organize all the resources in your subscription.
To include your infrastructure definition as part of your app source code	The template can become part of the source code for your app. Check it in to your source code repository and update it as your app evolves. Simply edit the template through Visual Studio or your favorite IDE.
To ensure your resources are deployed in the correct order	Resource dependencies are declaratively expressed within the ARM template. This ensures that components add their dependencies when they are provisioned, and that the resources within the template are created in the proper order. For instance, an ARM template that creates a VM and a VNet to which it is added, first creates the VNet, then creates the VM and associates the two.

Monitoring

Provides insights into the status and health of the data warehouse solution

Choosing Azure for monitoring

When Azure Monitor can be a good option for your monitoring solution

When you want…	Description
To access base-level metrics and logs	Azure Monitor provides base-level infrastructure metrics and logs for most services in Microsoft Azure. Azure services that do not yet put their data into Azure Monitor will put it there in the future.
To discover, configure, and on-board Azure Monitor features	Provides a landing page that helps you understand the monitoring capabilities offered by Azure. This starting point for on-boarding platform and premium monitoring capabilities shows curated notable issues from different services, allowing you to navigate to them in context.
To view important monitoring events across a given subscription	In Azure Monitor, select a subscription and view the following across the components of the subscription: Triggered alerts and alert sources Activity log errors Azure Service Health data and alerts Application Insights KPIs (key performance indicators) If Log Analytics, Azure Alerts, or Application Insights haven't been configured, the page provides links to begin your on-boarding process.

Choosing Azure application insights

When Azure Application Insights can be a good option for your monitoring solution

When you want…	Description
To monitor your live web application	Azure Application Insights provides a rich monitoring solution optimized for collecting and visualizing performance metrics from both the server and client, identifying exceptions in the application, and monitoring application usage.
To track exceptions down to actual failing code	When you receive an alert or discover a problem, you can assess how many users are affected. Correlate failures with exceptions, dependency calls, and traces. To dig deeper, examine profiler, snapshots, stack dumps, and trace logs.
To write custom telemetry in your code	Use the Azure Application Insights core telemetry API to send custom events and metrics, and your own versions of standard telemetry. Create custom events that are relevant to your application, and that can provide more custom monitoring options than standard out-of-the-box telemetry. For example, your e-commerce site can send events like *item added to cart* and *coupon applied* to Azure Application Insights, where you can use the built-in visualization tools to aggregate and compare these events over a given timeframe.
To monitor web site availability and responsiveness	Create availability tests for any HTTP or HTTPS endpoint that is accessible from the public internet. Web requests will be sent to your application at regular intervals from points around the world and alert you if your application doesn't respond, or responds slowly. Use Visual Studio to record a multi-step web test scenario for more advanced test automation.

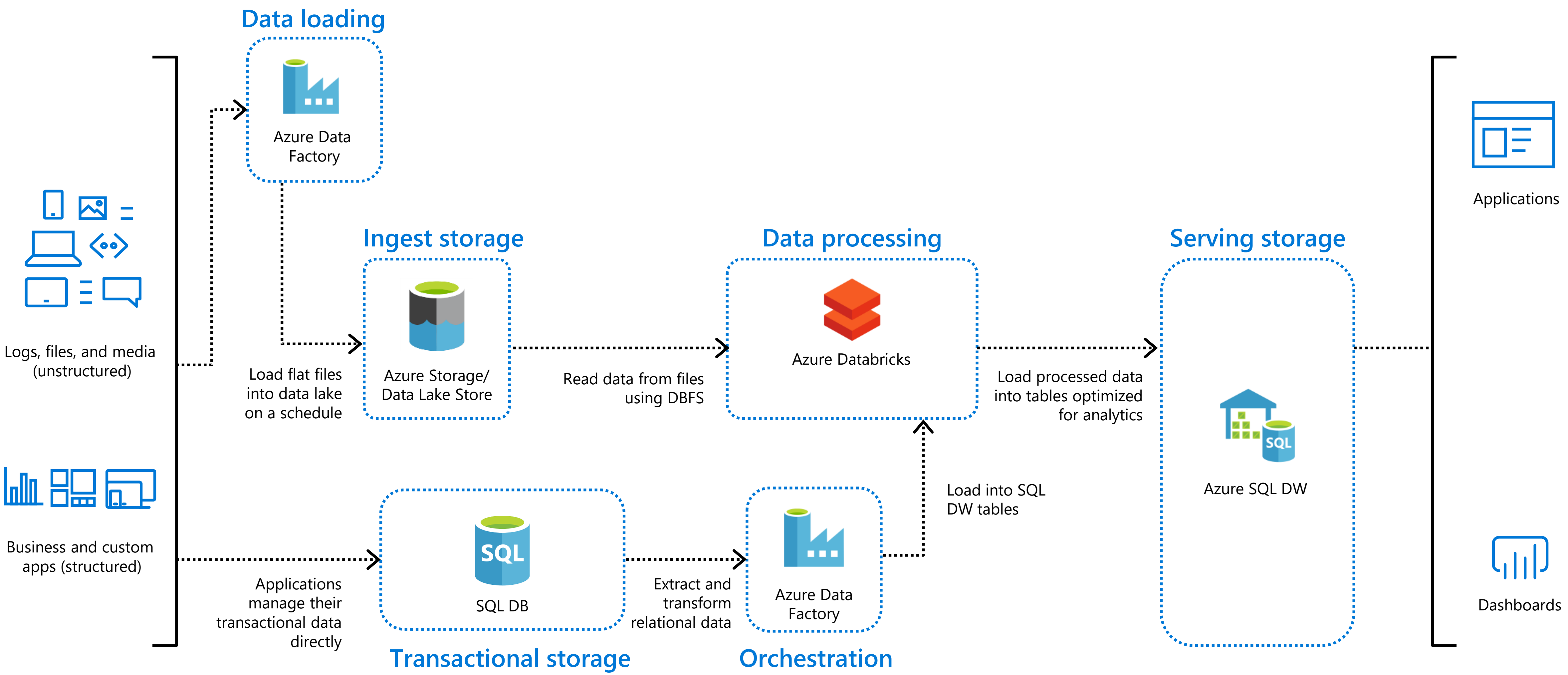
Choosing Azure Log Analytics

When Azure Log Analytics can be a good option for your monitoring solution

When you want...	Description
To be able to collect data generated by your cloud resources, on-premises environments, and other monitoring tools	Azure Log Analytics can collect data from multiple sources, including information sent directly from agents running on VMs, and other monitoring tools such as Azure Monitor, System Center Operations Manager, and Azure Application Insights. This allows you to correlate this data and have a single pane of glass through which to query and view logs pertinent to your cloud and on-premises environments.
Service and application-specific monitoring	Solutions are available for a variety of functions and additional solutions are constantly being added. You can easily browse available solutions and add them to your workspace from the Azure Marketplace. Many will be automatically deployed and start working immediately while others will require moderate configuration. Some examples include Logic Apps Management, Azure Search, HDInsight, SQL Health Check, and Azure Active Directory.
To create alert rules on log data	Alerts can be created through alert rules that automatically run log searches at regular intervals. If results of the log search match particular criteria, an alert record is created and can be configured to perform an automated response.

Modern data warehousing pattern in Azure

Data processing with Azure Databricks


Data loading
Azure Data Factory
Logs, files, and media (unstructured)
Load flat files into data lake on a schedule
Ingest storage
Azure Storage/ Data Lake Store
Read data from files using DBFS
Data processing
Azure Databricks
Load processed data into tables optimized for analytics
Serving storage
Azure SQL DW
Applications
Business and custom apps (structured)
Applications manage their transactional data directly
SQL
SQL DB
Transactional storage
Extract and transform relational data
Azure Data Factory
Orchestration
Load into SQL DW tables
Dashboards

It’s all on

Microsoft Azure

Microsoft

特別感謝

MVP Microsoft® Most Valuable Professional

SEA 台灣軟體工程學會 Taiwan

R-Ladies Taipei

TAIWAN

91APP

多奇·數位創意

Wishing-Soft

KKTIX

Geek&Nerd studios

HackMD

以及各位參與活動的你們

STUDY4.TW
為 學 習 而 生

.NET Conf