CLOUD1

DIGITAALISUUDEN ARKKITEHDIT

Microsoft Partner

Microsoft

Gold Application Integration
Gold Cloud Platform
Silver Data Analytics
Silver Data Platform

Cloud1 Partnership

Architecture
Data Platform
Enterprise Integration

Custom Development
.NET & Azure PaaS

Staff 42
Revenue (2017) 4,5M

Clients:
Finnair, SGN Group
Vantaan Energia,
VR, VVO, Metsä Group

Java & Open Source
Custom Development
Solution Architects

CLOUD1

Agenda

Azure Data Warehousing

Overview

- Microsoft Azure
- Modern Data Estate
- Azure Data Platform

Data Platform Services

- Azure SQL Data Warehouse
- HDInsight
- Data Lake Store & Analytics
- Data Factory
- Additional Services

Practical Patterns

- Example Patterns
- Real-Life Case Example
- 3rd party Cloud & Tools

CLOUD1

Overview

Microsoft Azure

Modern Data Estate

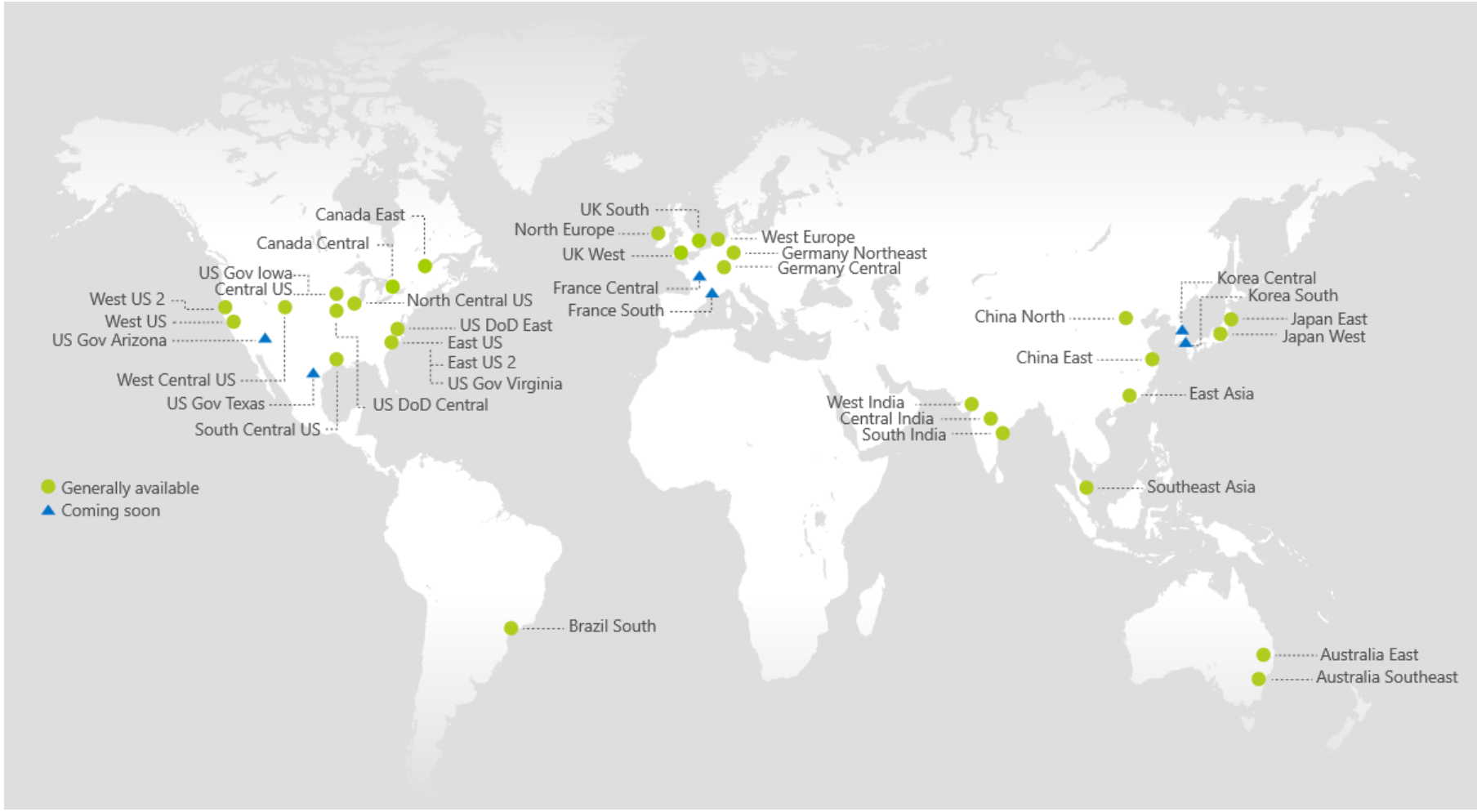
Azure Data Platform

CLOUD1

What is Azure?

CLOUD1

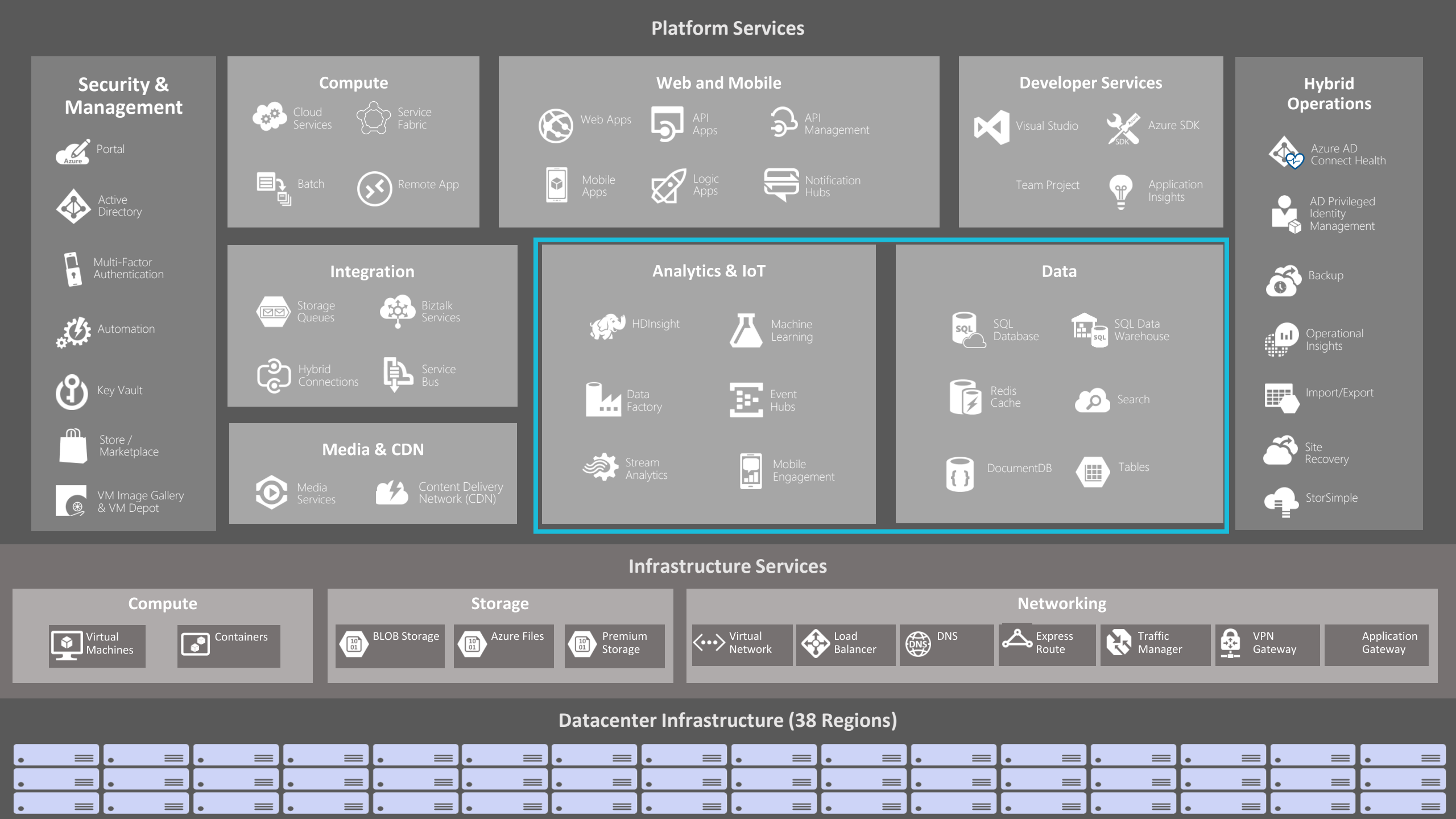
Azure Regions



Azure is generally available in 38 regions around the world.

Geographic expansion is a priority for Azure as it enables to achieve higher performance and it support customer requirements and preferences regarding data locations*

CLOUD1


Platform Services
Security & Management
Portal
Active Directory
Multi-Factor Authentication
Automation
Key Vault
Store / Marketplace
VM Image Gallery & VM Depot
Compute
Cloud Services
Service Fabric
Batch
Remote App
Web and Mobile
Web Apps
API Apps
API Management
Mobile Apps
Logic Apps
Notification Hubs
Developer Services
Visual Studio
Azure SDK
Team Project
Application Insights
Hybrid Operations
Azure AD Connect Health
AD Privileged Identity Management
Backup
Operational Insights
Import/Export
Site Recovery
StorSimple
Integration
Storage Queues
Biztalk Services
Hybrid Connections
Service Bus
Media & CDN
Media Services
Content Delivery Network (CDN)
Analytics & IoT
HDInsight
Machine Learning
Data Factory
Event Hubs
Stream Analytics
Mobile Engagement
Data
SQL Database
SQL Data Warehouse
Redis Cache
Search
DocumentDB
Tables
Infrastructure Services
Compute
Virtual Machines
Containers
Storage
BLOB Storage
Azure Files
Premium Storage
Networking
Virtual Network
Load Balancer
DNS
Express Route
Traffic Manager
VPN Gateway
Application Gateway
Datacenter Infrastructure (38 Regions)

Overview

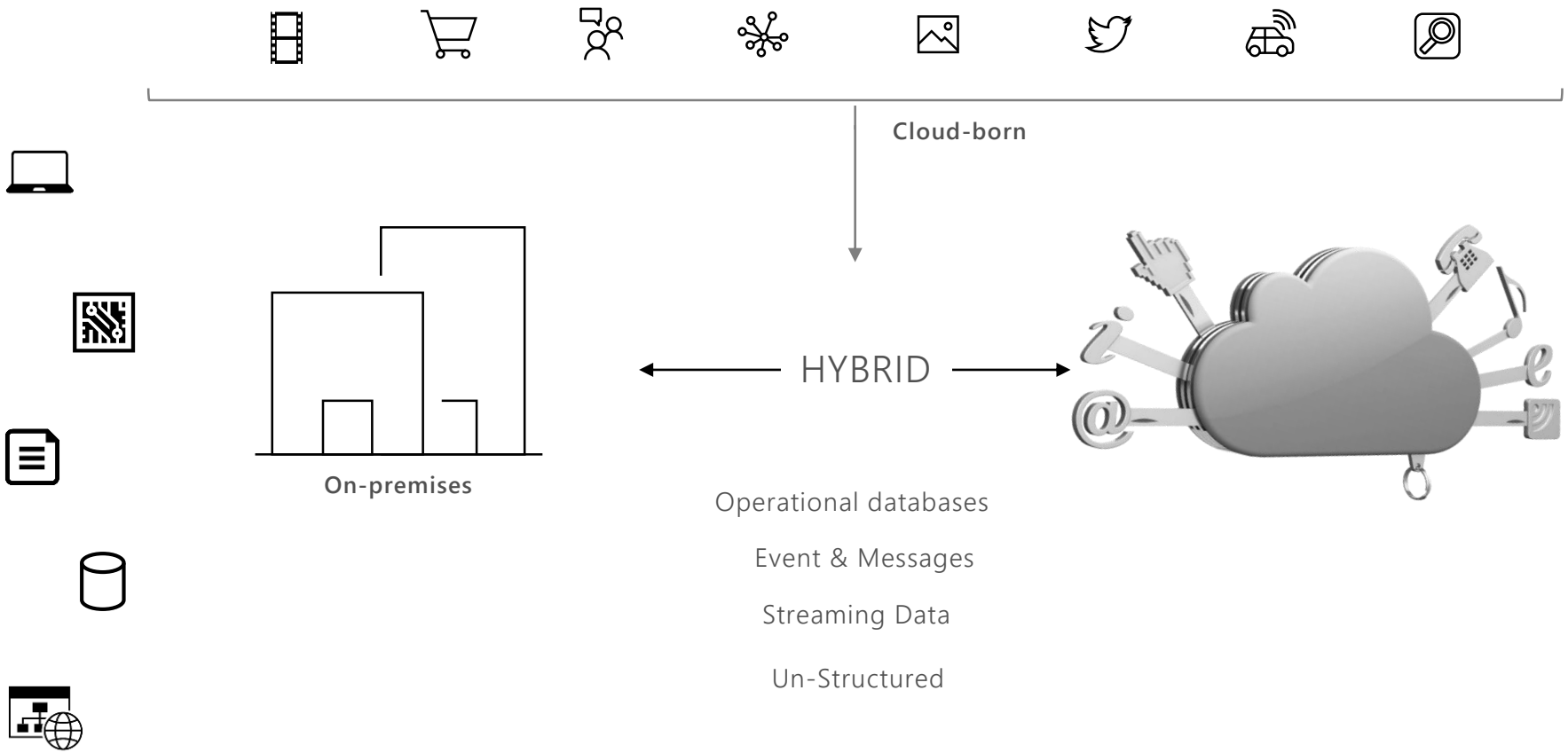
Microsoft Azure

Modern Data Estate

Azure Data Platform

CLOUD1

Modern Data Estate



Data Sources

- Websites, Servers
- Mobile & Sensors
- LOB applications
- Forms, Surveys
- Social, SaaS, PaaS …

Data Formats

- Relational, Text
- JSON/XML
- Images, Documents
- Video / Audio
- 3D, Spatial ….

CLOUD1

Roles in a Data Platform

Data Integration	• Hybrid and Cloud data movement & transform • Automation, orchestration, connectivity
Storage	• Raw data storage, staging and archiving
HaaS	• Transformation and processing (Big Data) • On-Demand provisioning • Advanced Analytics (ML, R etc.)
Data Warehouse	• Delivery for consumption • Massaged & enriched data for querying
Catalog	• Discoverability and management of data

CLOUD1

Overview

Microsoft Azure

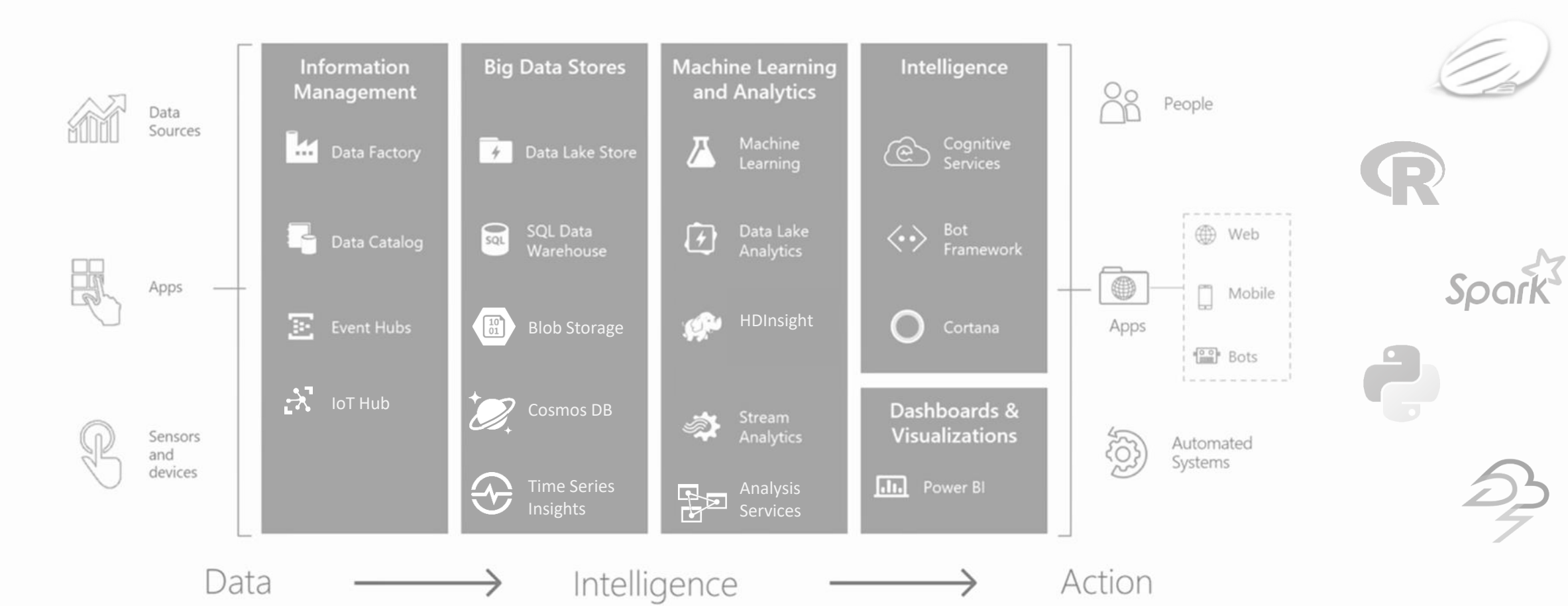
Modern Data Estate

Azure Data Platform

CLOUD1

Data Platform Overview

Flexibility of Options



Data Platform Services

Azure SQL Data Warehouse

HDInsight

Data Lake Store & Analytics

Data Factory

Additional Services

CLOUD1

Azure SQL Data Warehouse

Overview

- MPP DW Technology
- Scale-out distributed query engine
- De-coupled storage from compute
- Fully managed and Elastic PaaS-solution
- Few hundred Gigabytes to Terabytes and Petabyte scale
- Broad range of connectivity options
- Compressed Columnstore by default (5x on average)
- Azure Active Directory (AAD) and SQL logins
- Supports Data Encryption

Differences to normal SQL Server

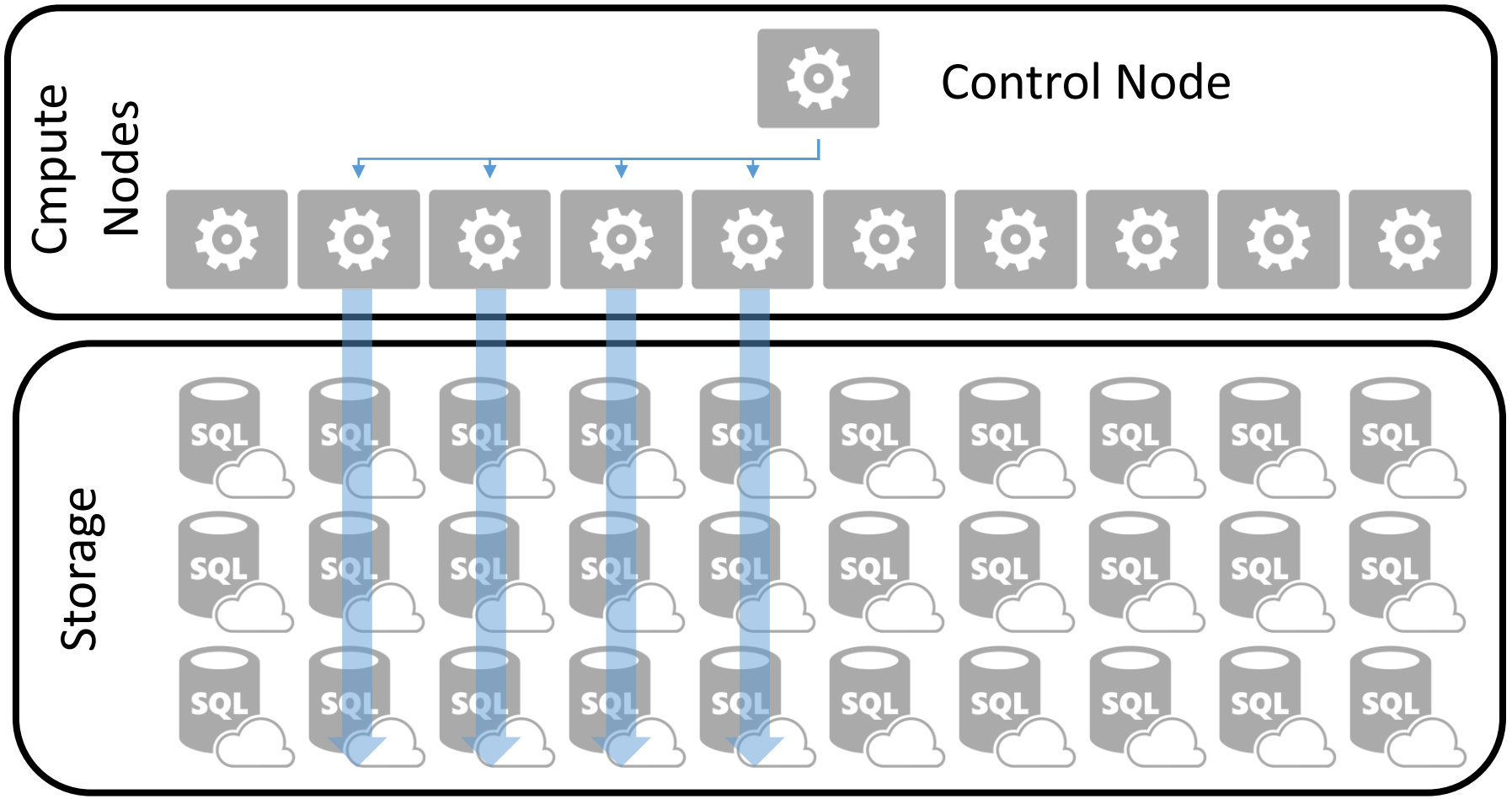
- Only single database (i.e. divide objects by schema)
- Procedure nesting level 8, no cursors (use temp tables, CTAS, CTE instead)
- Some datatype limitations (e.g. Timestamp, XML, Geometry, text, uuid)
- Isolation level always READ UNCOMMITTED
- No indexed views, no unique indexes
- No auto-generated statistics (use stored procedure)
- No JOIN in update/delete (utilize CTAS or EXISTS)
- No Primary Key, Foreign Key or Unique constraints

On-premises ≈ Azure SQL != Azure DW

CLOUD1

** Based on Data Allegro technology, acquired by Microsoft in 2008*

Azure DW / Massive Parallel Processing

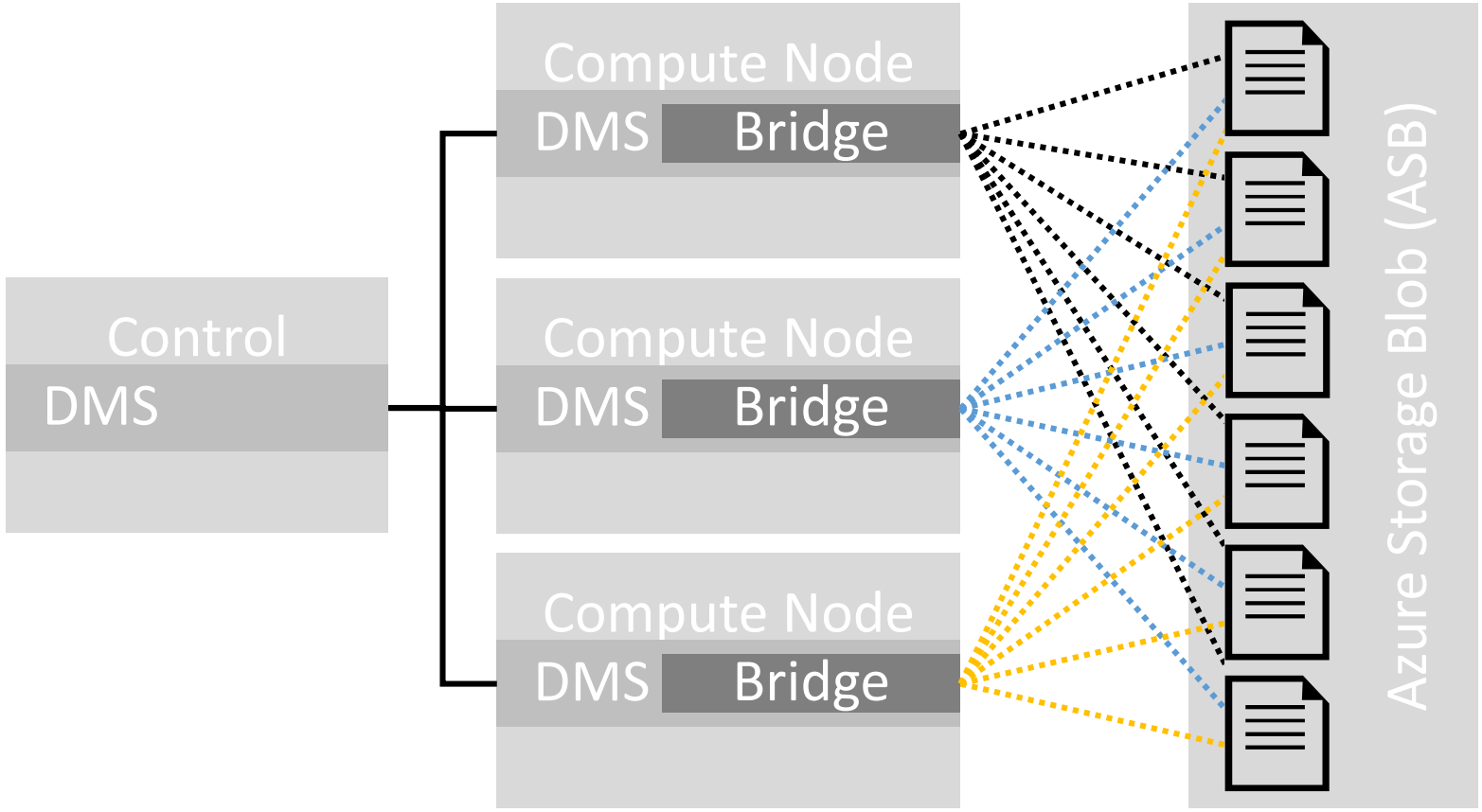


MPP-architecture

- Data is distributed to large number of storage units
- Storages are mounted up to 120 Compute nodes
- Query sent to Control Node, which forms sub-queries for Nodes
- Aggregated results are sent back to Control
- Data can be shuffled between Nodes if needed

CLOUD1

Azure DW / PolyBase



External Table Support

- A parallel operation to read data from ASB or HDFS / Data Lake
- Schema-on-Read
- Can persist result set with CTAS operation

CLOUD1

Azure SQL DW / Sizing

Date Warehouse Unit

- Database capacity
- Tempdb size and IO
- Concurrency & Memory
- Load
- Transaction size
- Memory management

} DWU

Current Limitations (as of 09/2016)

- Maximum transaction size limited by allocated DWUs
- Maximum of 1024 concurrent connections
- Maximum of 4 concurrent queries per 100 DWUs

Resource Classes

- Resource allocation is controlled per User by placing them in as specific **Resource Class**
- Two types of classes exist: Static and Dynamic
- *Running e.g. table creation bulk insert with too small Resource Class can create severly misaligned partitions!!*

"1TB / DWU100 is good place to start."*

CLOUD1

** With less than 4 TB (compressed), regular Azure SQL is an option*

Data Platform Services

Azure SQL Data Warehouse

HDInsight

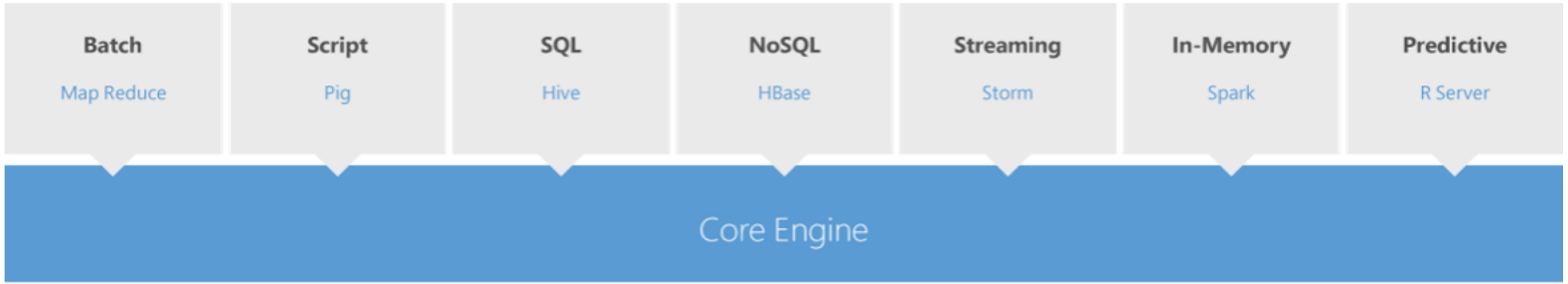
Data Lake Store & Analytics

Data Factory

Additional Services

CLOUD1

HDInsight



Distributed storage and processing of data

- HDFS-files system distributes data to multiple nodes
- Control node passes processing requests to compute nodes, which perform the operations on their local data and return results

Azure HDInsight = pre-packaged HADOOP

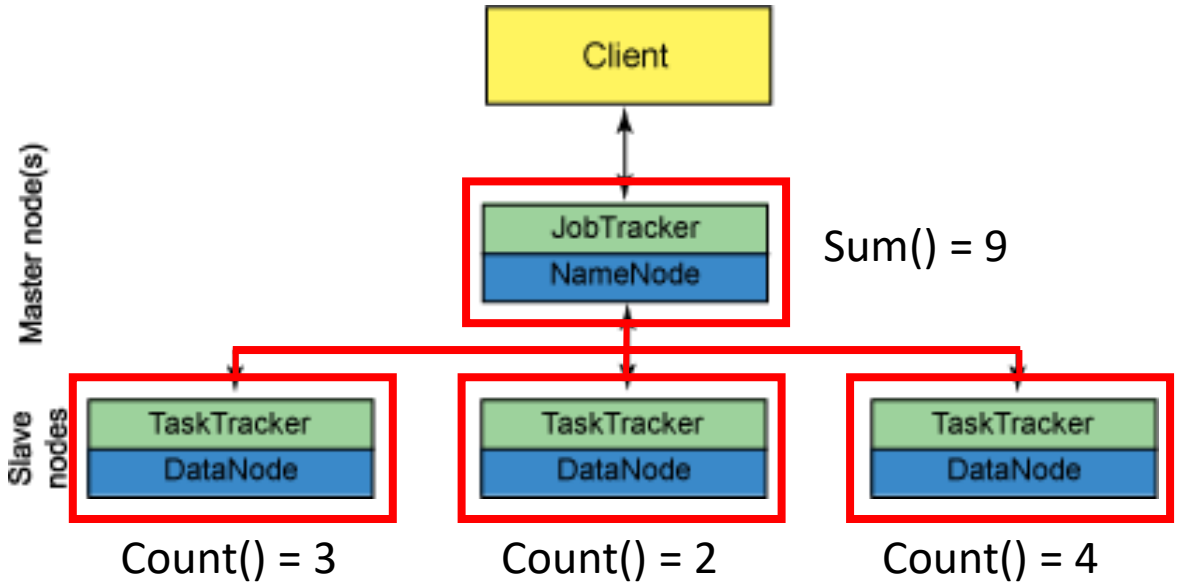
- Supports commons extensions and tools (Spark, Solr, R, Giraph etc.)
- In addition to standard HADOOP languages (Java, PHP, Python..) supports also .NET languages
- Integrates natively with following MS services and tools: WABS, ML, Power BI, Excel, Data Factory

Cofigurations

- Can be deployed as standard HADOOP, supporting a rich ecosystem of solutions
- Multiple pre-configured deployment options available, e.g. Hbase, Apache Storm, Apach Spark ...

CLOUD1

Map Reduce

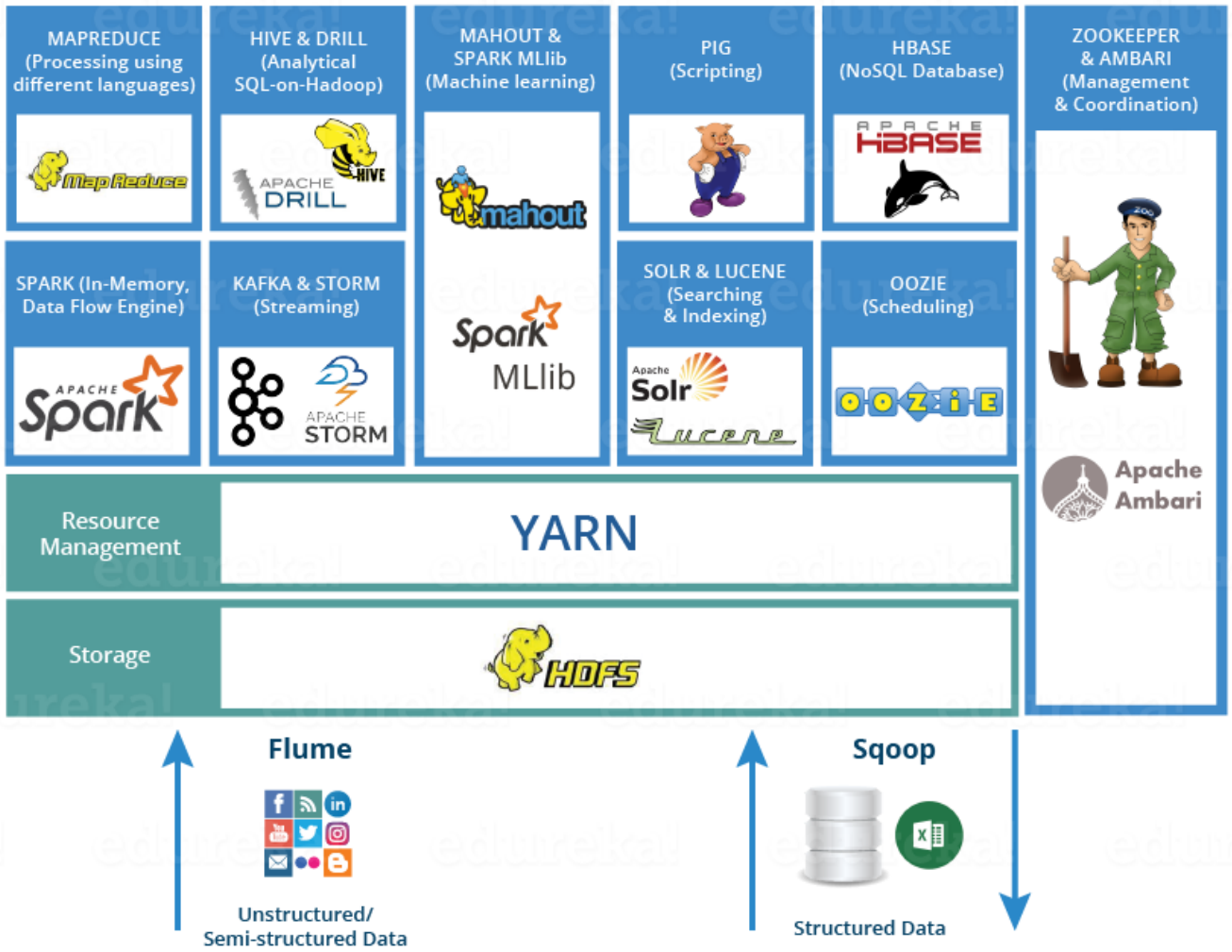


REDUCE: ”Sum Map results”

MAP: ”Count number of Toyotas”

CLOUD1

HADOOP Ecosystem



+ MORE....

CLOUD1

HaaS = HADOOP as a Service

Flexible usage model

- Provision Compute Capacity when needed and discard after processing
- Minimal management and maintenance overhead
- Built-in connectivity with other PaaS components

CLOUD1

Data Platform Services

Azure SQL Data Warehouse

HDInsight

Data Lake Store & Analytics

Data Factory

Additional Services

CLOUD1

Data Lake

The big data paradigm

- All data has potential value
- Data hoarding – store it all!
- 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

Azure Data Lake

- **Data Lake Analytics** for distributed parallel processing (pay-per-minute)
- **Data Lake Store** for hyperscalable HDFS-compliant storage system

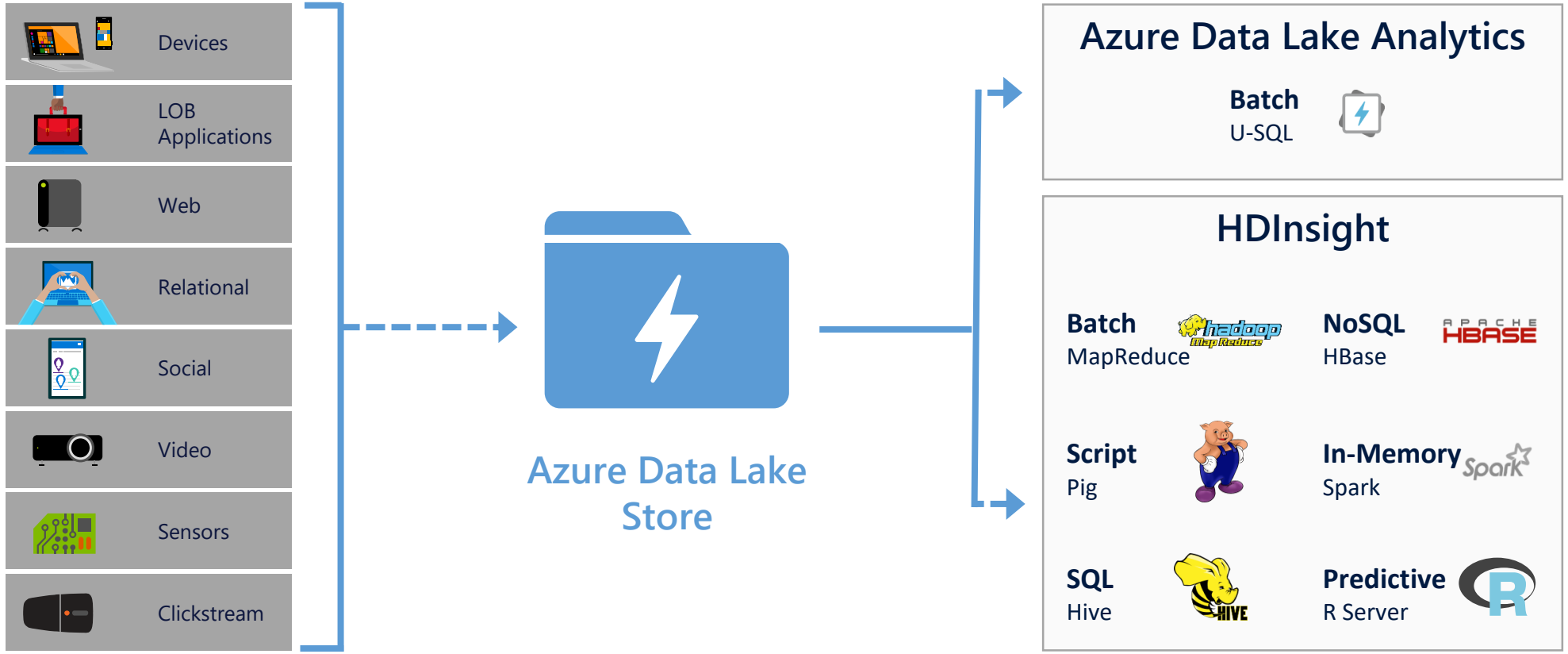
CLOUD1

Data Lake 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
 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 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.

CLOUD1

Azure Data Lake Store



CLOUD1

Azure Data Lake Store

Hyperscale

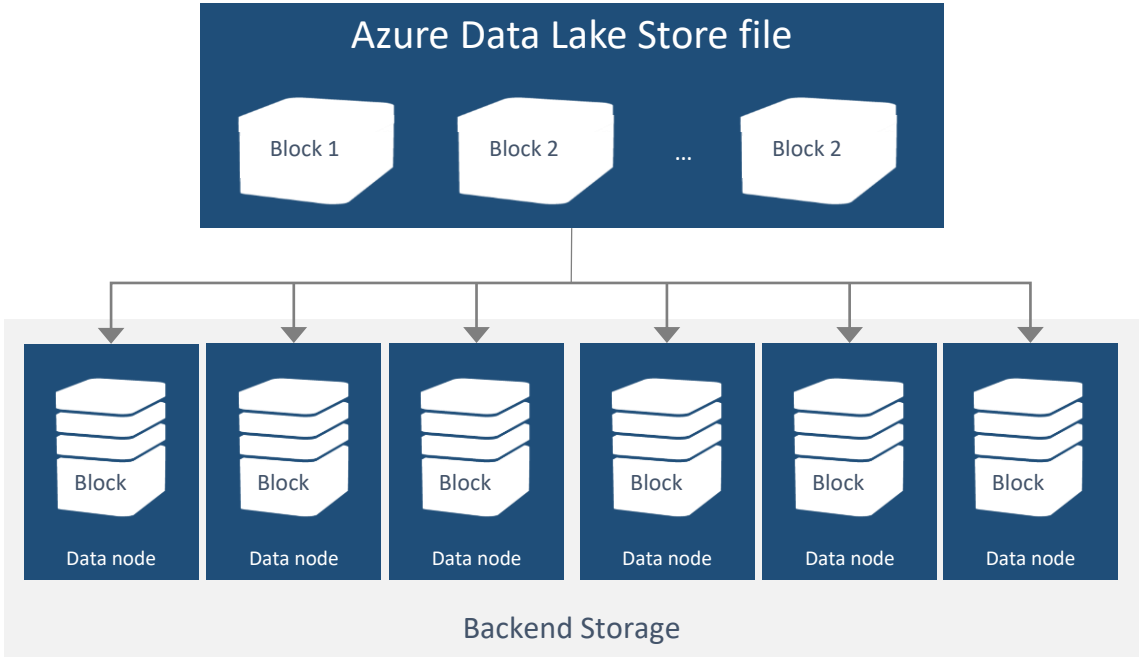
- Each file in ADL Store is sliced into blocks
- Blocks are distributed across multiple data nodes in the backend storage system (PaaS)
- With high number of nodes file sizes of 2 PB+
- Through read parallelism ADL Store provides massive throughput

Redundancy

- Azure maintains 3 replicas of each data object per region across three fault and upgrade domains
- Writes are committed to application only after all replicas are successfully updated

Connectivity

- Many Azure PaaS services already support natively
- HDFS-compliant as a data source

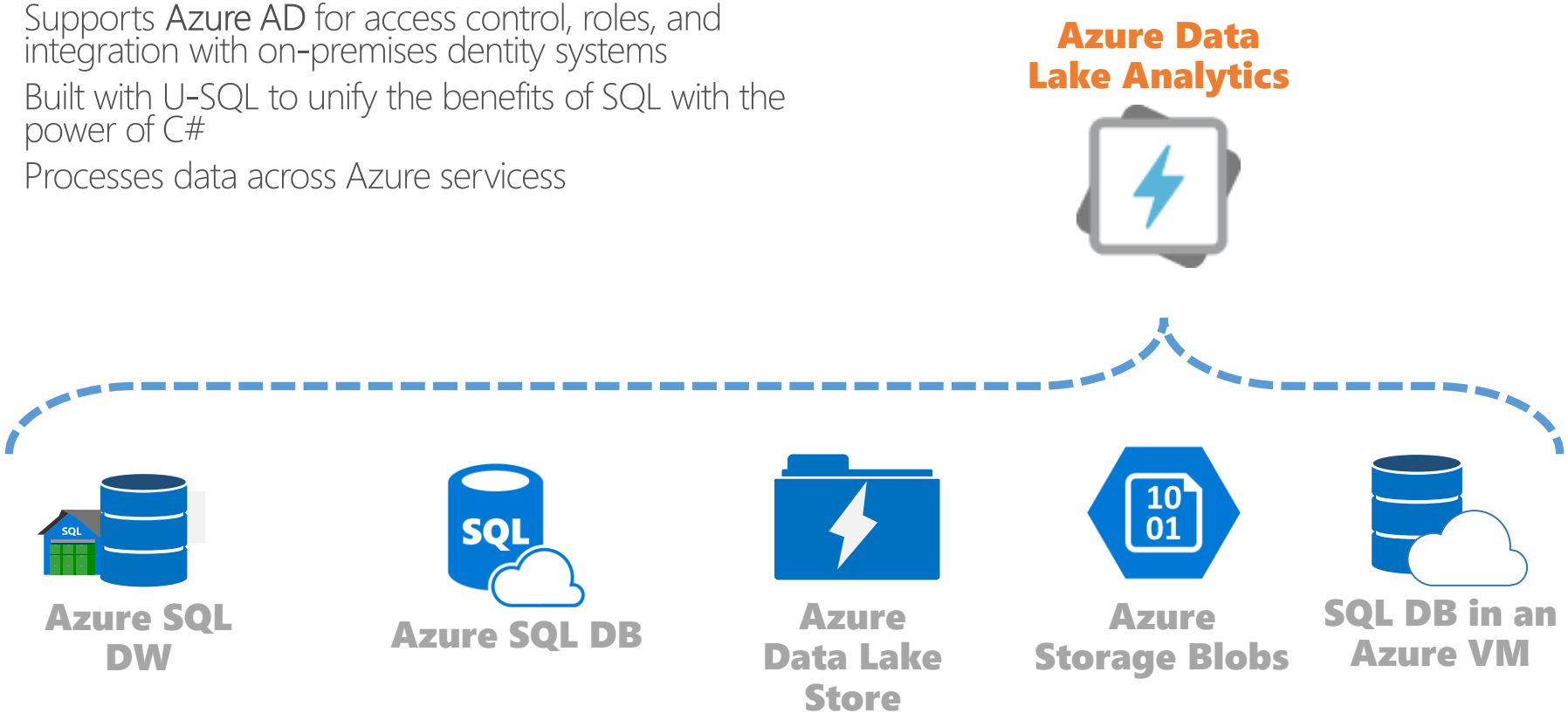


CLOUD1

Data Lake Analytics

Summary

- Distributed analytics service based on Apache YARN
- Dynamic Scaling and Pay-by-the-Query
- Supports Azure AD for access control, roles, and integration with on-premises dentity systems
- Built with U-SQL to unify the benefits of SQL with the power of C#
- Processes data across Azure servicess



CLOUD1

Data Lake Analytics / U-SQL

```
@rs1 =
    SELECT
        Region,
        SUM(Duration) AS TotalDuration
    FROM SearchLogDb.dbo.SearchLog2
GROUP BY Region;

@res =
    SELECT *
    FROM @rs1
    ORDER BY TotalDuration DESC
    FETCH 5 ROWS;

OUTPUT @res
    TO "/output/Searchlog-query-table.csv"
    ORDER BY TotalDuration DESC
    USING Outputters.Csv();
```

U-SQL example

Overview

- SQL-like language to process data
- Development with Visual Studio
- Extensible with C# based "User Defined Operators"

Use Cases

- Schematizing unstructured data (Load-Extract-Transform-Store) for analysis
- Cook data for other users (LETS & Share)
- Large-scale custom processing with custom code
- Augment big data with high-value data from where it lives

Sequence

- Each command produces a new dataset, which the following commands can access to read data
- No modifications are done on data sets
- Final results can be output to files or to DL tables

CLOUD1

Data Platform Services

Azure SQL Data Warehouse

HDInsight

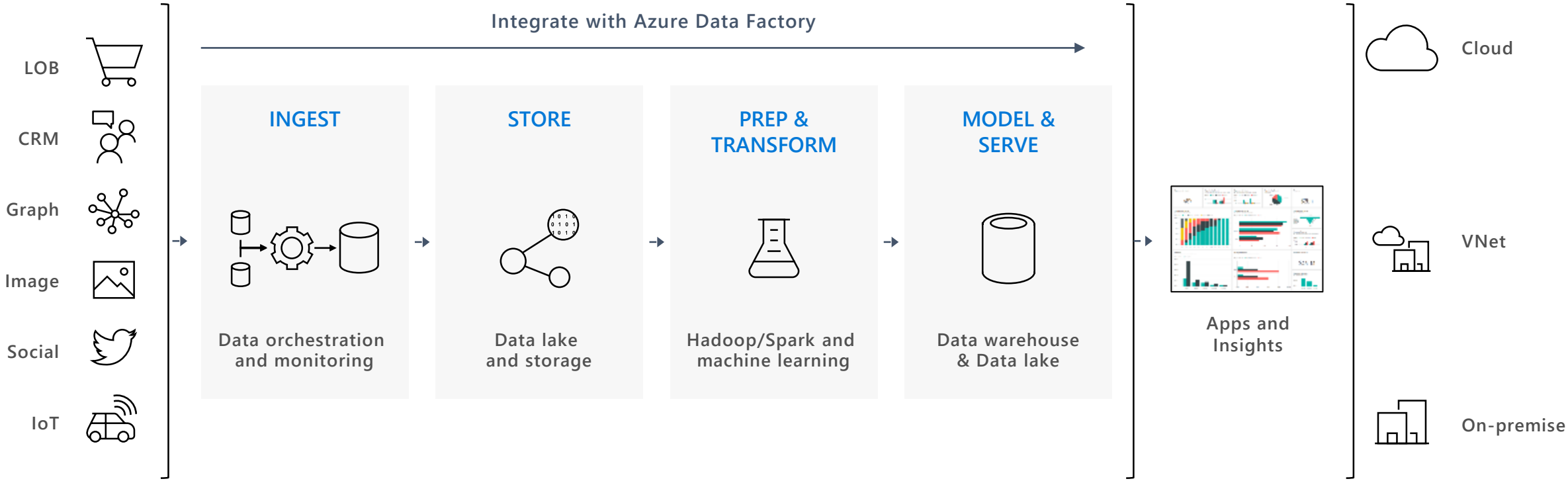
Data Lake Store & Analytics

Data Factory

Additional Services

CLOUD1

Data Factory



CLOUD1

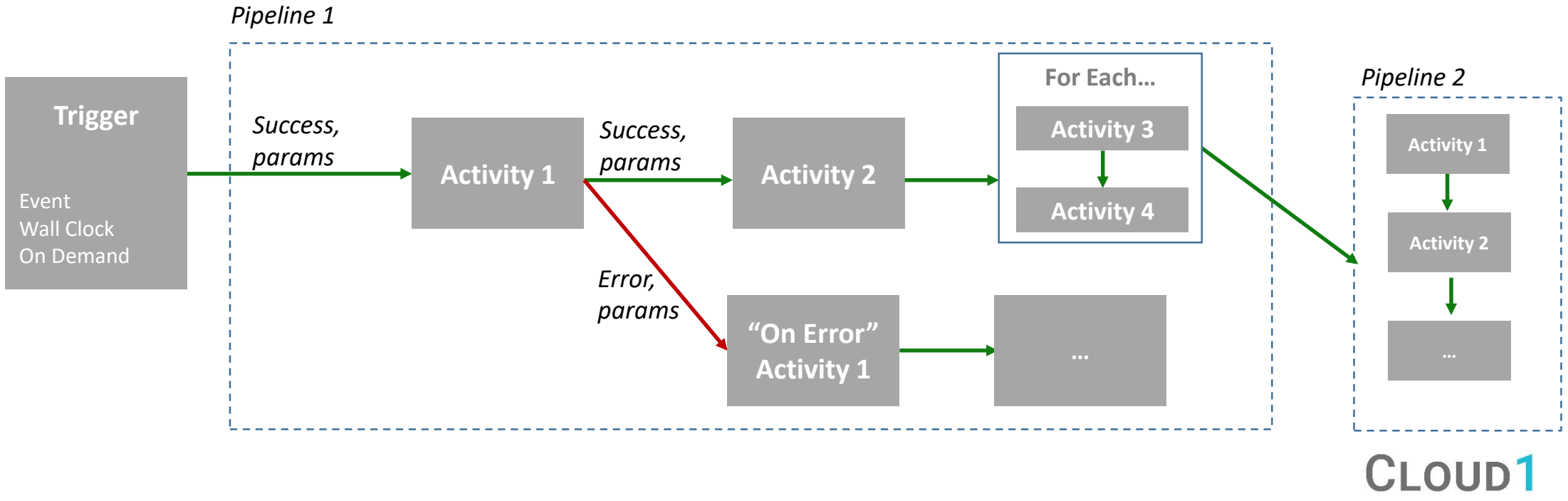
ADFv2 Control Flow

Basic Structure

- Coordinate activities into execution steps, supporting looping, conditionals and chaining
- Data load and transformations contained into Data Flows

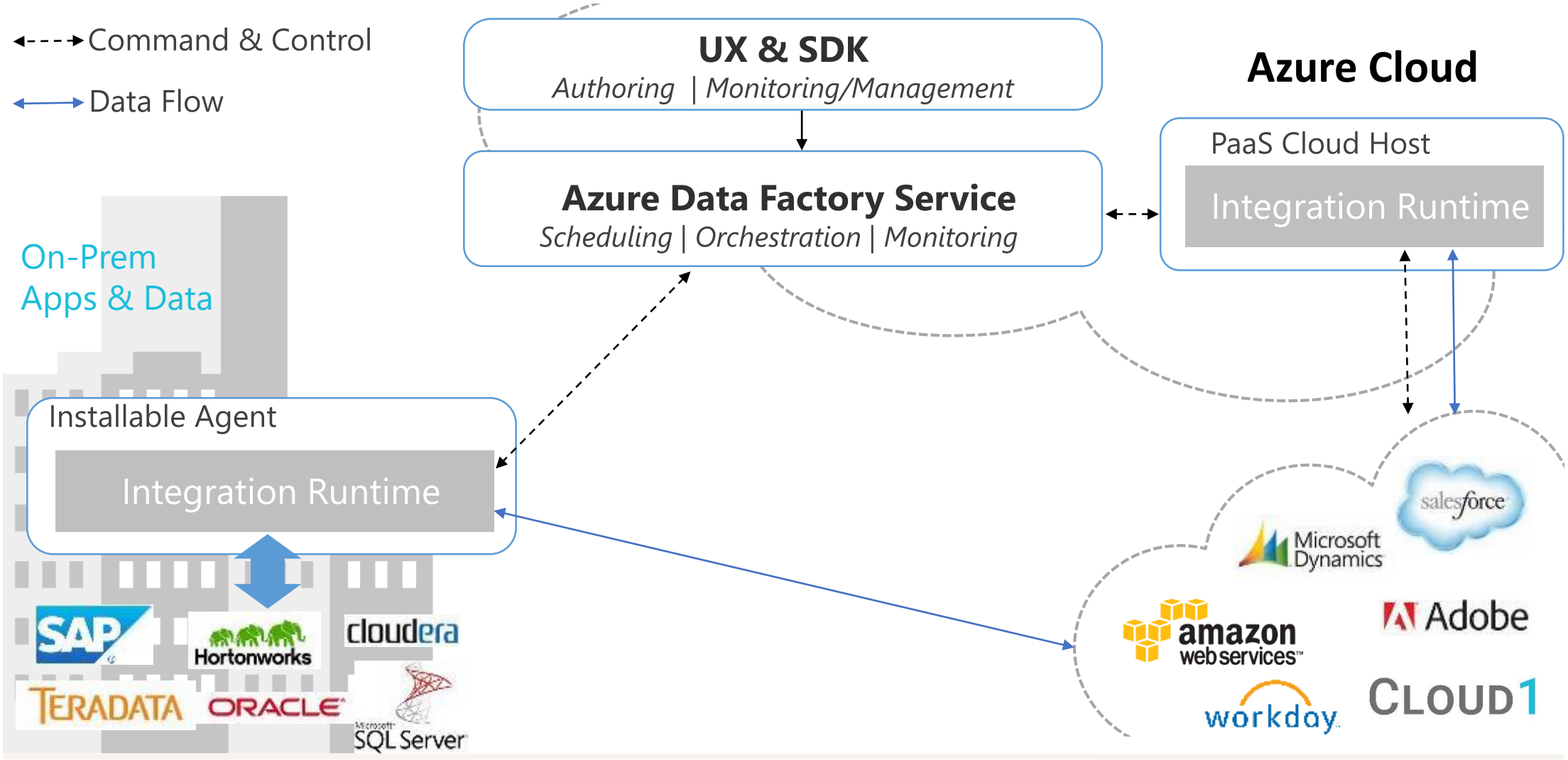
Integration Runtime

- ADFv2 isolated execution process enables also running SSIS packages
- Integration Runtime Agent can be installed on VMs or on-premises (also replaces DMG-functionality)



CLOUD1

ADFv2 Architecture



SSIS vs.Data Factory

SSIS

Benefits

- Very flexible and extensible tool
- Good for moving data on-premises and hybrid scenarios
- Great variety of connectors for Azure and other SaaS/PaaS (3rd party)
- Extensible with C# and custom components

Issues

- Offers just bare bones framework to work with, resilience of loading patterns needs to be hand-built (at least once)

Data Factory

Benefits

- Efficient for pure-Azure data movement
- Scales to large masses better than SSIS
- Built-in runtime resiliency (i.e. it is a PaaS solution)
- Has fixed method for big data delta loads
- Cost efficient (no SQL licenses)

Issues

- Limited flexibility and extensibility
- Limited amount of connectors
- Still requires on-premises server (DMG) for hybrid scenarios
- Has fixed method for big data delta loads

Generic Issue for Hybrid Scenarios

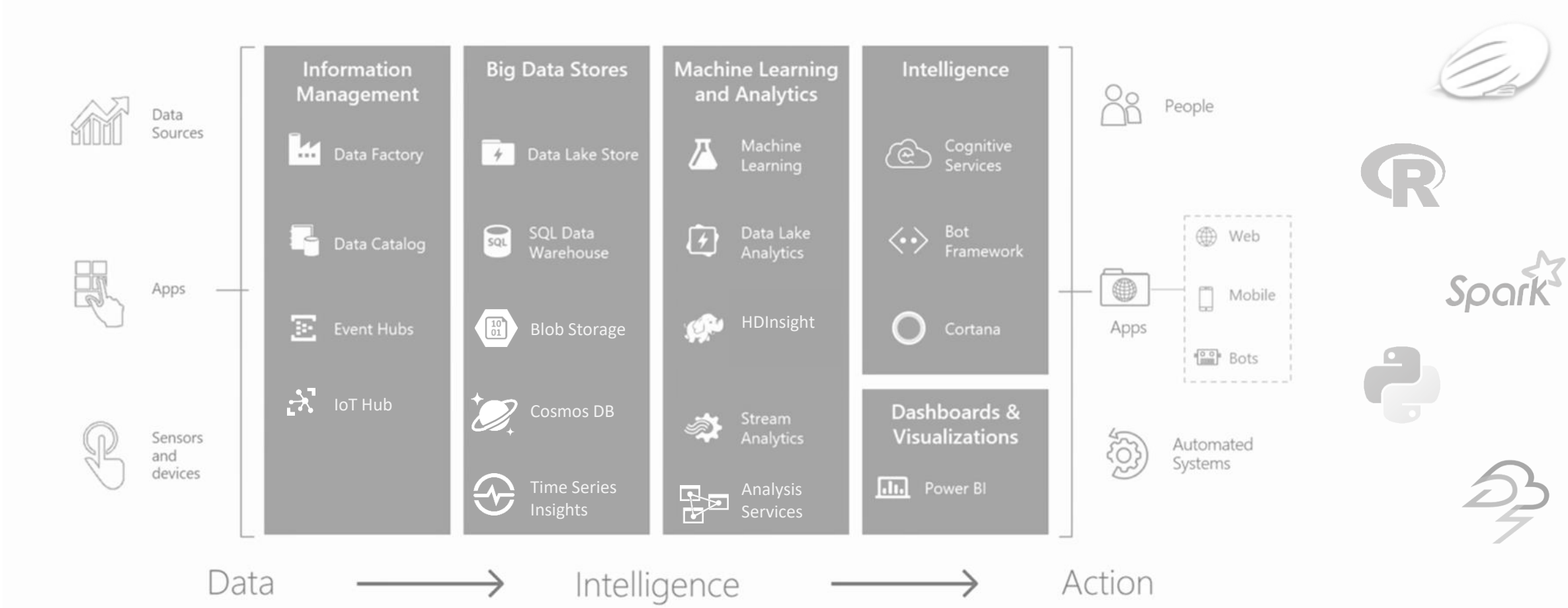
- Network will always be unreliable

➔ Prefer resumable loading patterns

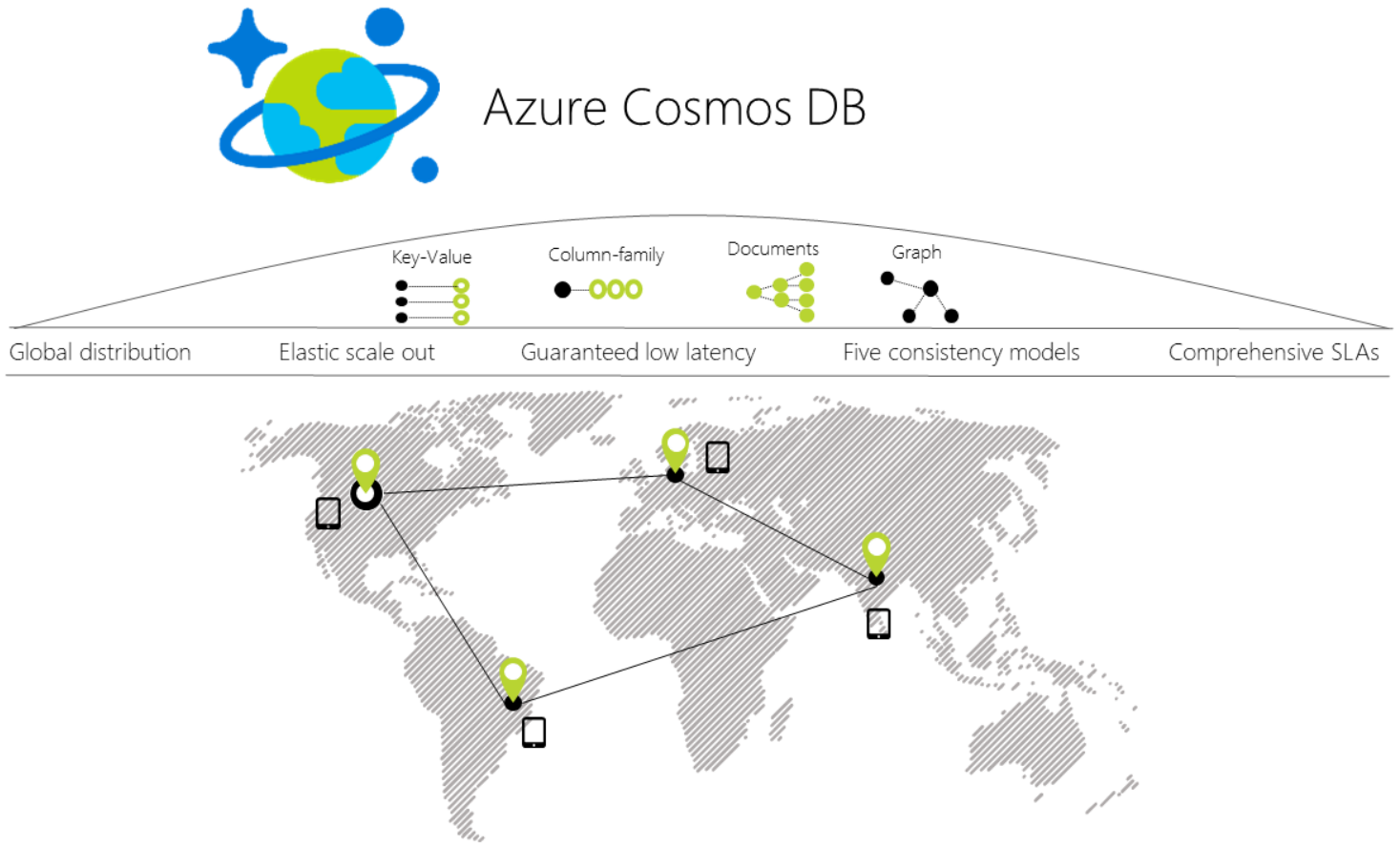
➔ Avoid combining data on far apart

Additional Services

Transform data into intelligent action



CosmosDB



Overview

- A geo-replicated and highly scalable multi-model **NoSQL** database
- Supports *document*, *graph*, and *key-value* storage
- Very High 99.99% SLA
- Guaranteed latencies (< 15 ms)
- Supports Encryption at rest
- Consistency models: strong >> eventual

Interfaces

- Accessed using REST APIs
- Extensible interfaces for Node.js, Java, .NET, .NET Core, Python
- MongoDB-compatible
- Queried with SQL or *Gremlin* (graph)
- Graph supports Apache TinkerPop

CLOUD1

Practical Patterns

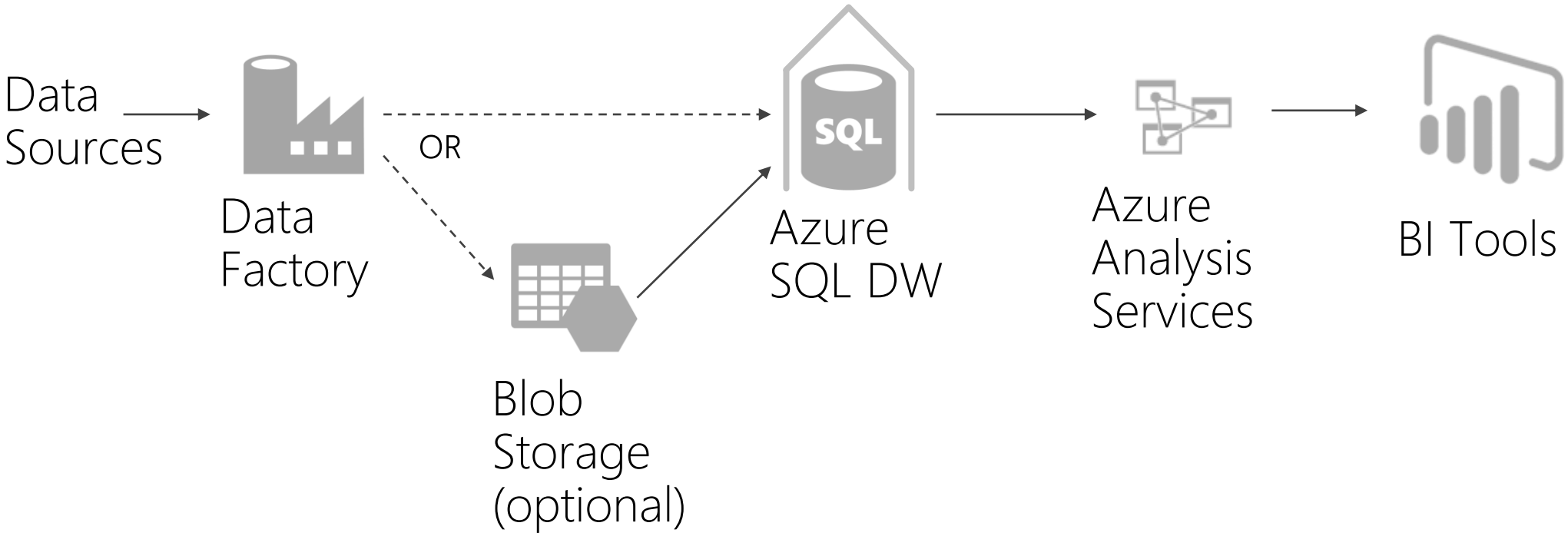
Example Patterns

Real-Life Examples

3rd party Cloud & Tools

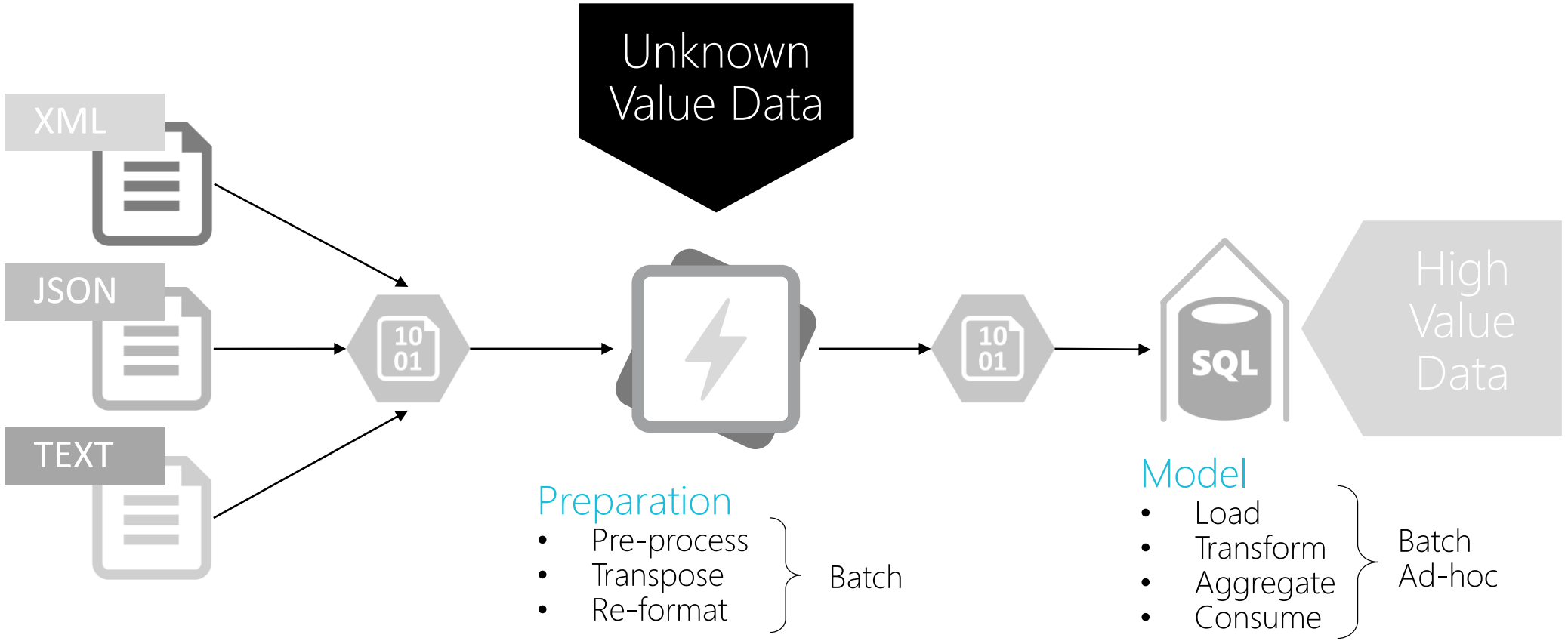
CLOUD1

Pattern: DW to Analysis Services



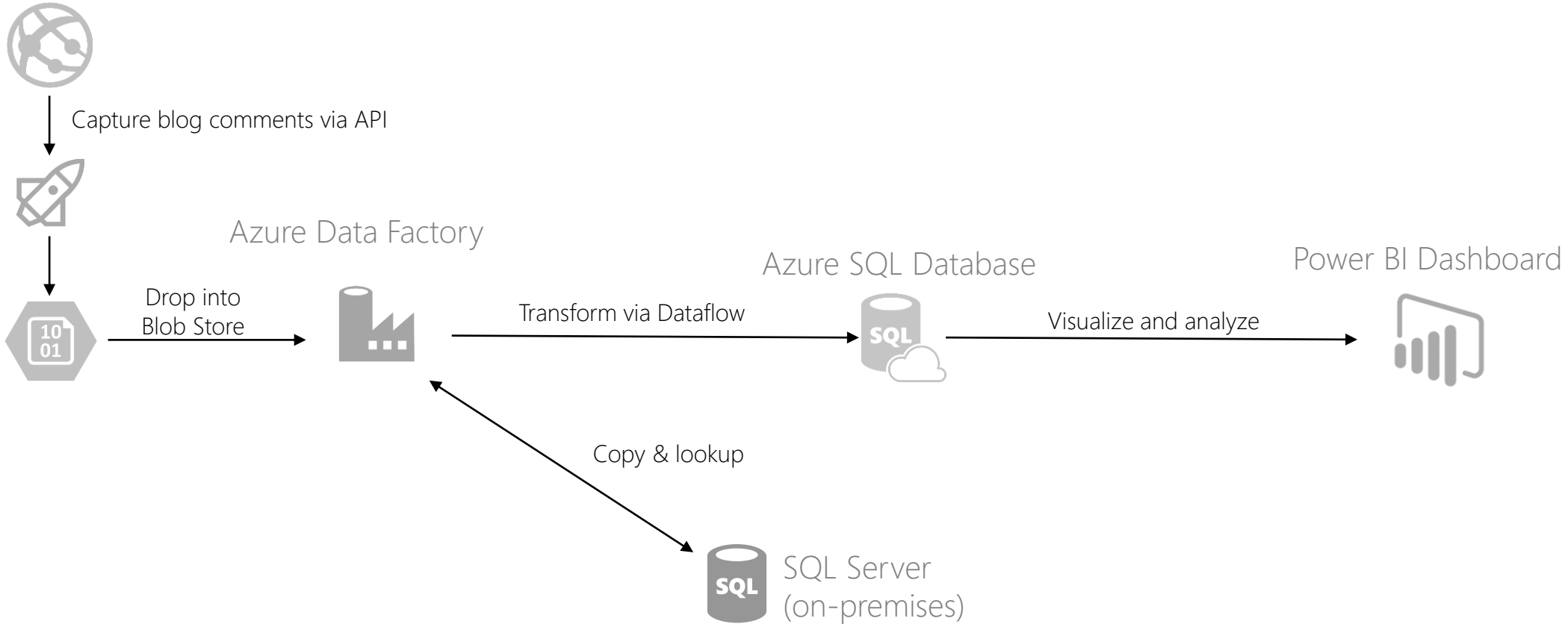
CLOUD1

Pattern: Data Lake to Azure DW



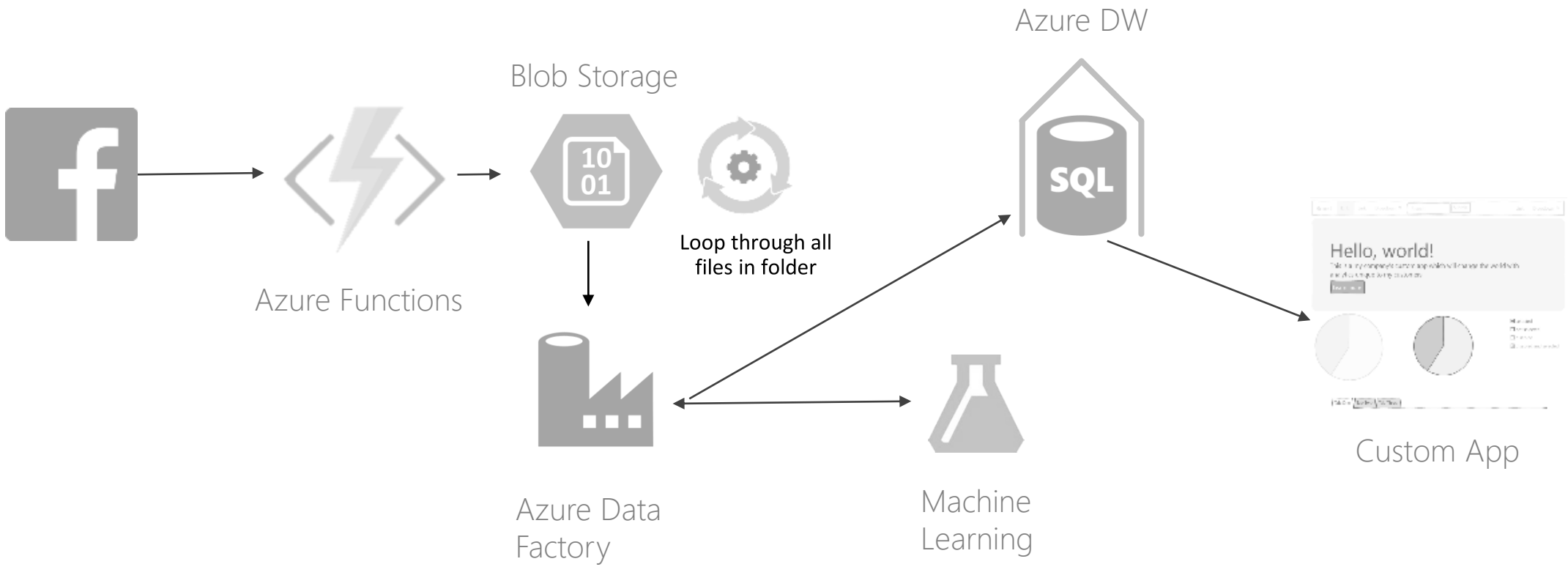
CLOUD1

Pattern: Analyze Blog Comments



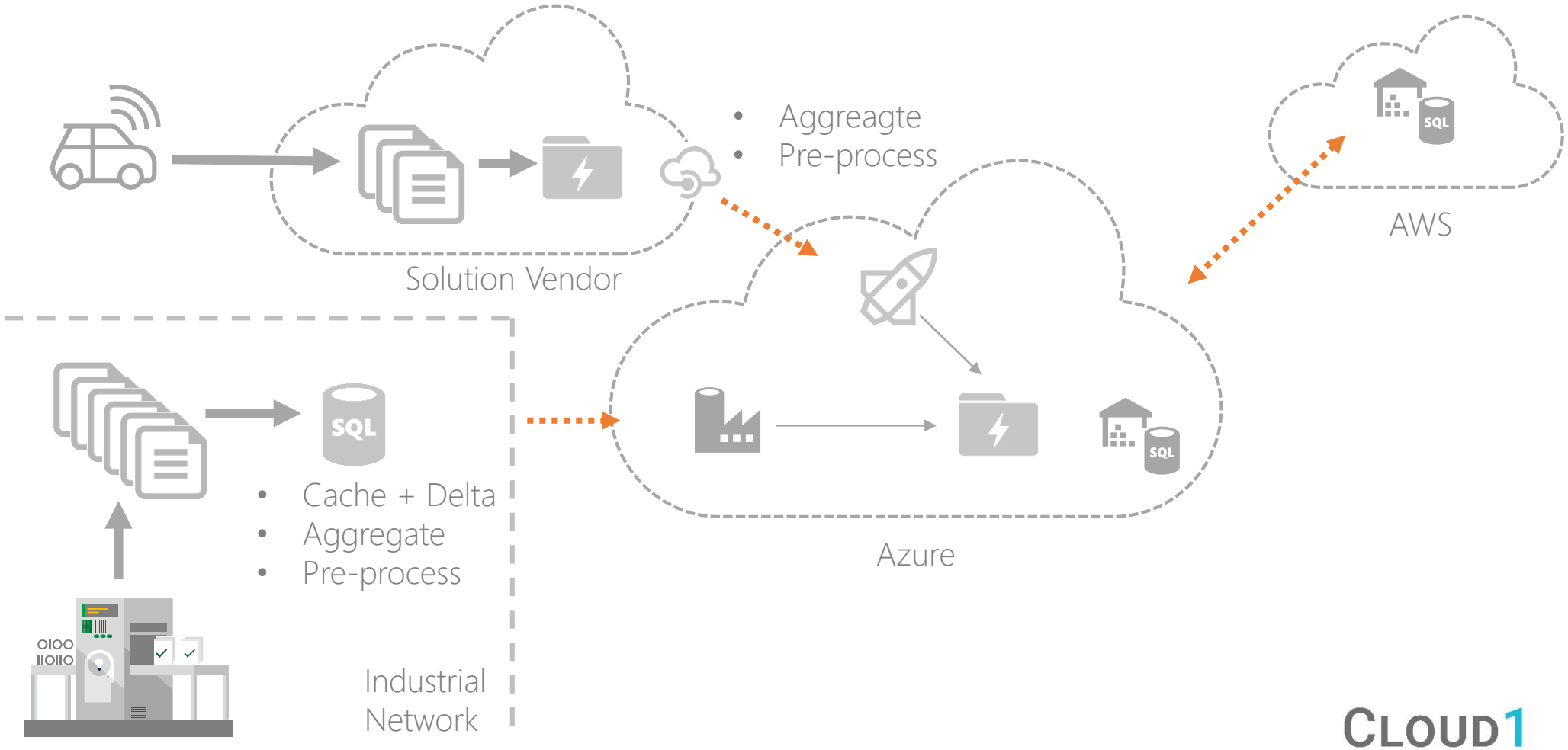
CLOUD1

Pattern: Sentiment Analysis



CLOUD1

Data Transfer Considerations



CLOUD1

Practical Patterns

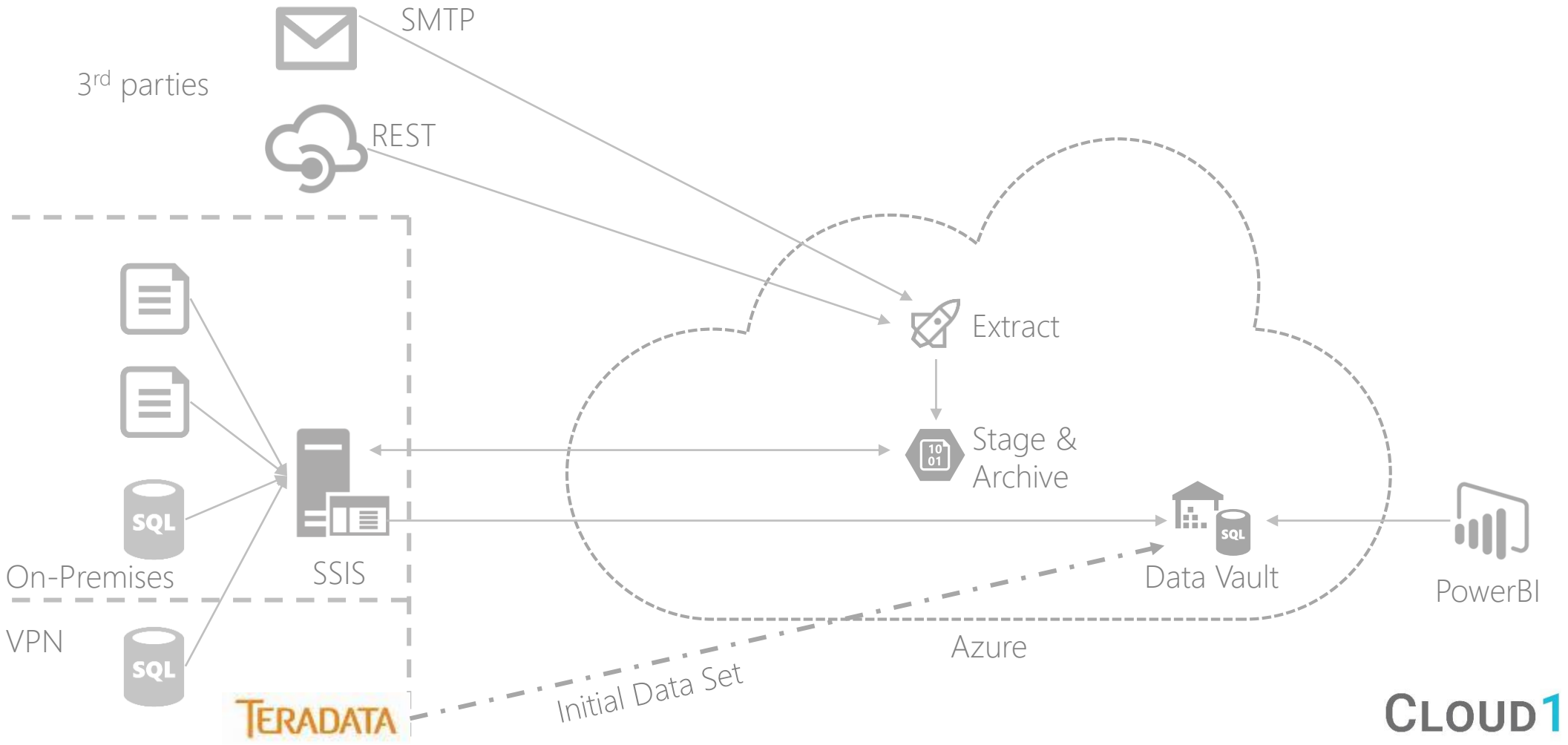
Example Patterns

Real-Life Case Example

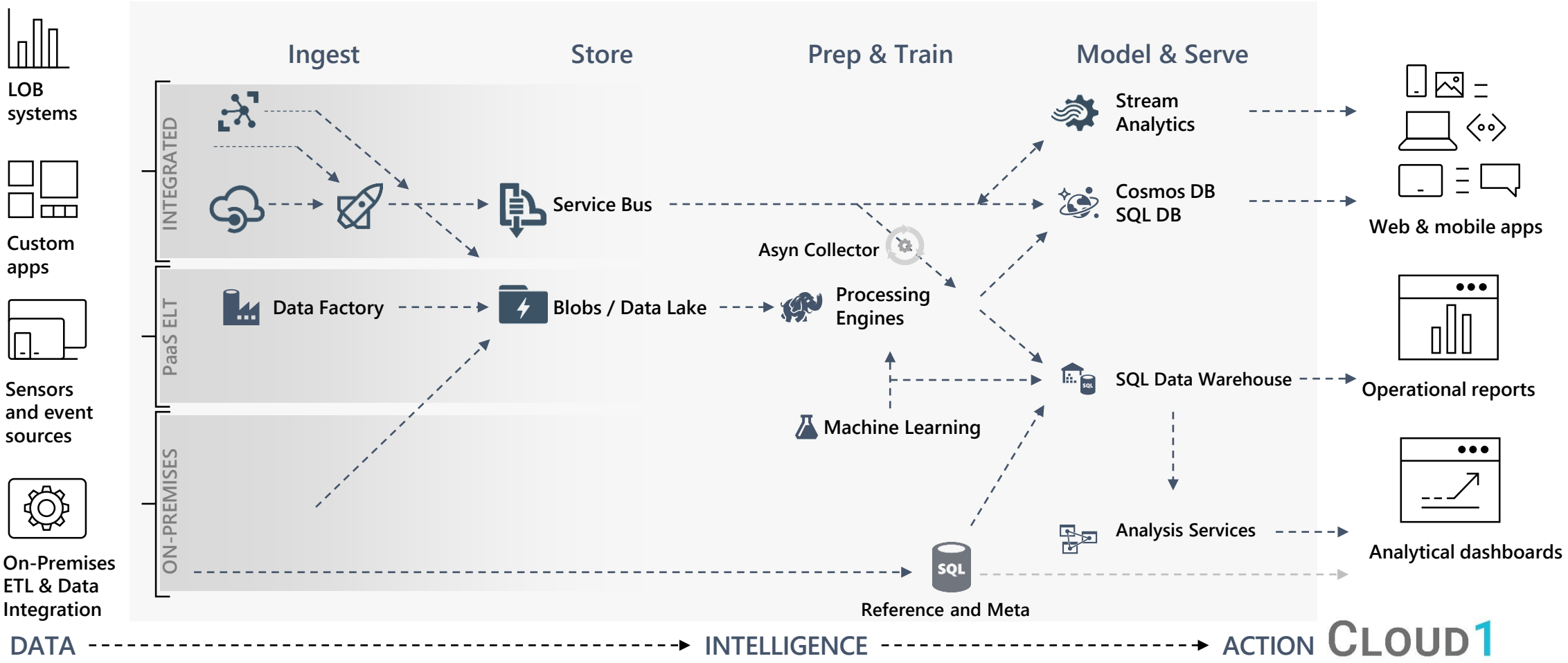
3rd party Cloud & Tools

CLOUD1

Simple Real-Life Architecture



Integrated Architecture



Practical Patterns

Example Patterns

Real-Life Case Example

3rd Party Cloud & Tools

CLOUD1

3rd Party Cloud & Tools

The Other Clouds

- *Data Factory* supports copying data ***from*** Amazon S3, Redshift, Salesforce and few others
- The rest can be done with SSIS (batch) or e.g. Logic Apps, Azure Functions and Web Jobs

3rd Party Analysis tools

- More and more of the tools are starting to support Azure DW and Azure SQL (e.g. Qlik, Tableau)
- HDFS-compliant tools are potentially able to use HDInsight and Data Lake

CLOUD1

DIGITAALISUUDEN ARKKITEHDIT

Microsoft Partner

Microsoft

Gold Application Integration
Gold Cloud Platform
Silver Data Analytics
Silver Data Platform