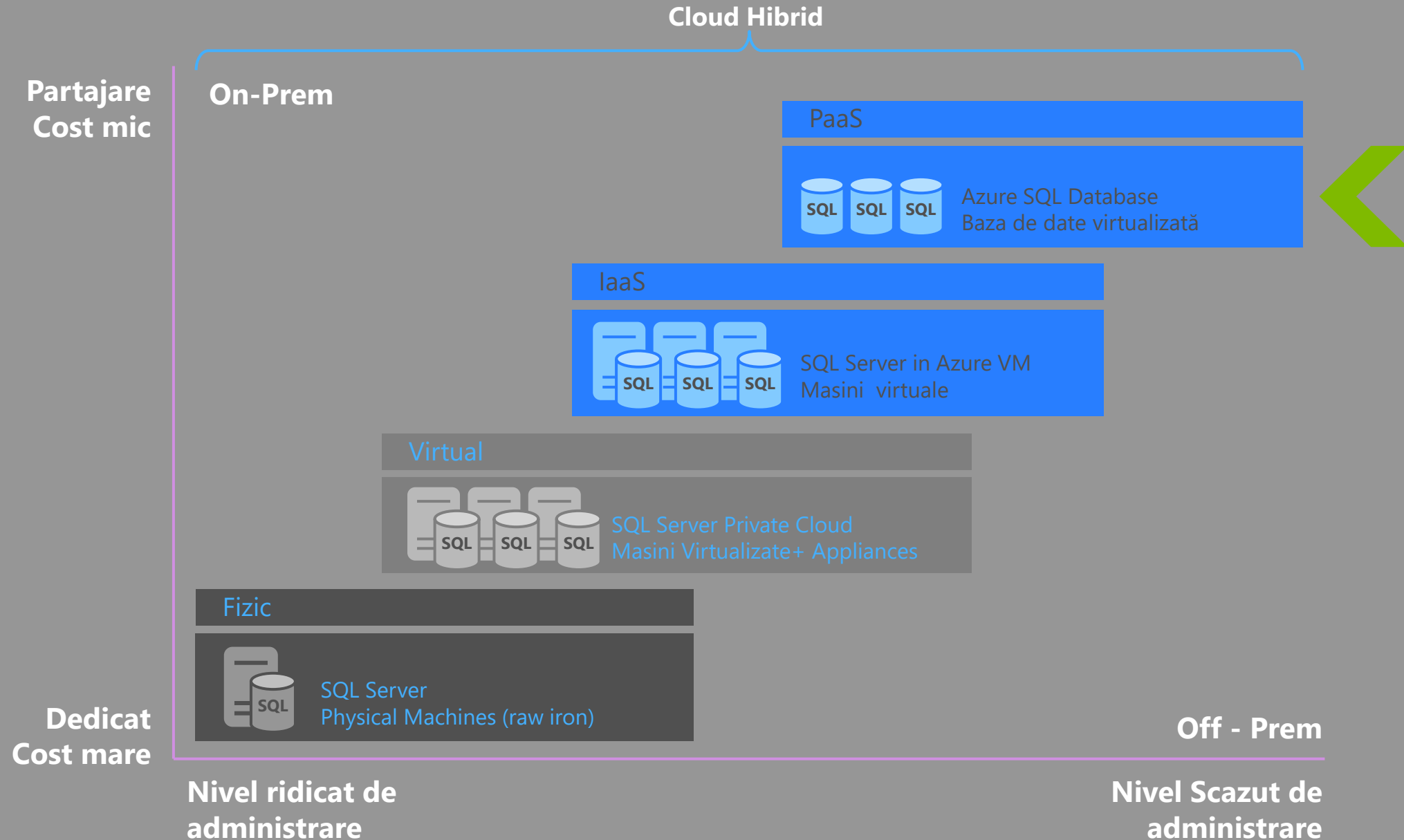


# Azure SQL database, Azure Data Warehouse și Azure Data Lake

Conf.dr. Cristian Kevorchian  
Facultatea de Matematică și Informatică  
ck@fmi.unibuc.ro

# Microsoft SQL - Baza de date ca serviciu



# Azure SQL Database

Baza-de-date livrată ca serviciu, complet gestionată de Microsoft  
Proiectată pentru aplicații în cloud cu efort minim de administrare

## Scalabilitate

Nivel de performanță predict.  
Scale up/down & out/in  
Vizualizare prin dashboard-uri a  
metricilor BD

## Business Continuity și Protecția datelor

Self-service restore  
Disaster recovery  
Compliance-enabled

## Self-managed and Familiar

Self-managed  
API-uri programabile  
Instrumente de lucru și  
limbaje familiare(e)

# Platformă de baze de date enterprise

# Scalabilitate

Basic, Standard și Premium furnizează nivelele de performanță și implicit de preț.

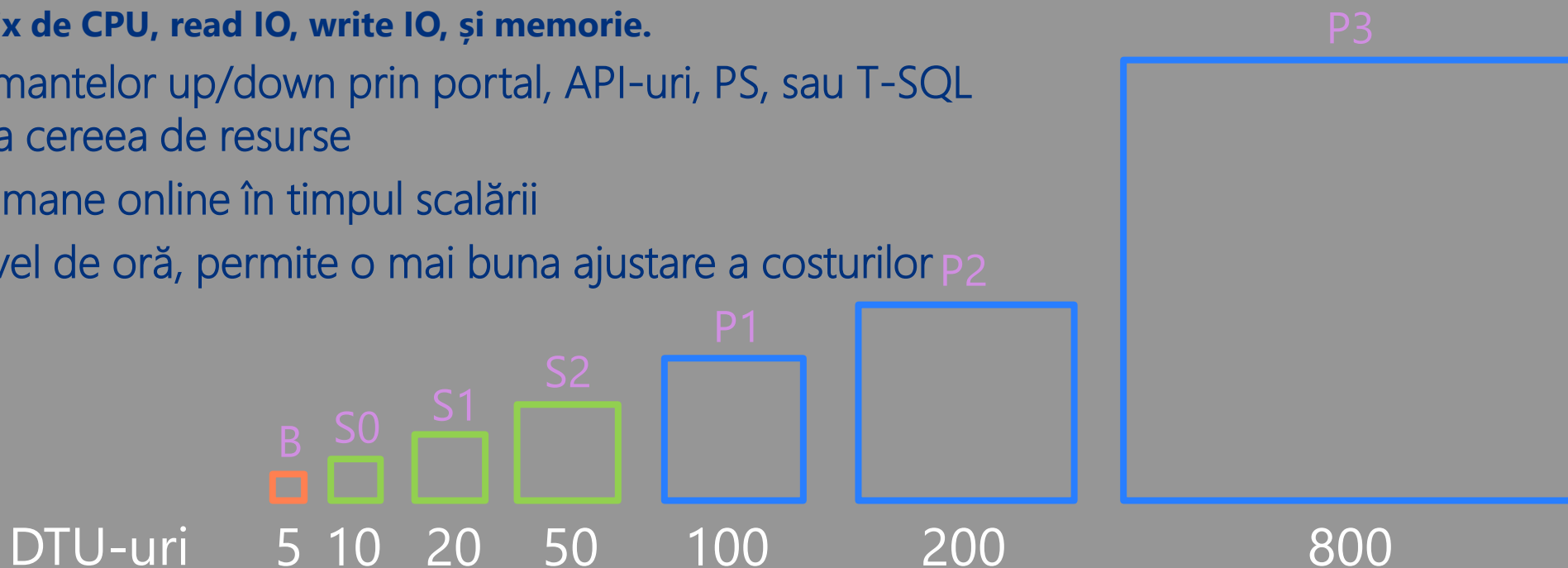
Performanța este exprimată prin DTU(Database Throughput Unit)

**DTU este un mix de CPU, read IO, write IO, și memorie.**

Scalarea performanțelor up/down prin portal, API-uri, PS, sau T-SQL pentru a reflecta cererea de resurse

Baza de date ramane online în timpul scalării

Facturarea la nivel de oră, permite o mai buna ajustare a costurilor



Performante usor scalabile în funcție de cerințele de business

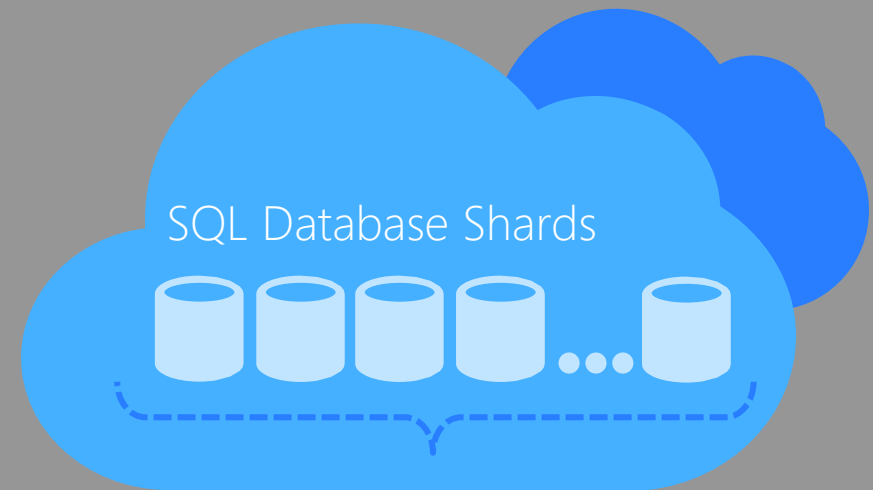
# Scalare elastică

Scalarea a mii de baze de date utilizând modelul bazelor de date partitionate(sharded)

Suportă adăugarea, împărțirea și reunirea partițiilor în funcție de mișcarea datelor

Clienții pot reuni rezultatele interogării din mai multe partiții.

Operațiuni de gestiune asincronă (întreținerea indexului, DDL, DML)



# Geo-replicare Asincronă

## Geo-replicare standard (standard și premium)

Opțional generarea unei replici secundare(non-readable) într-o regiune pereche

Replica este taxată la costuri reduse

Activarea replicii se face de Microsoft la apariția unei catastrofe.

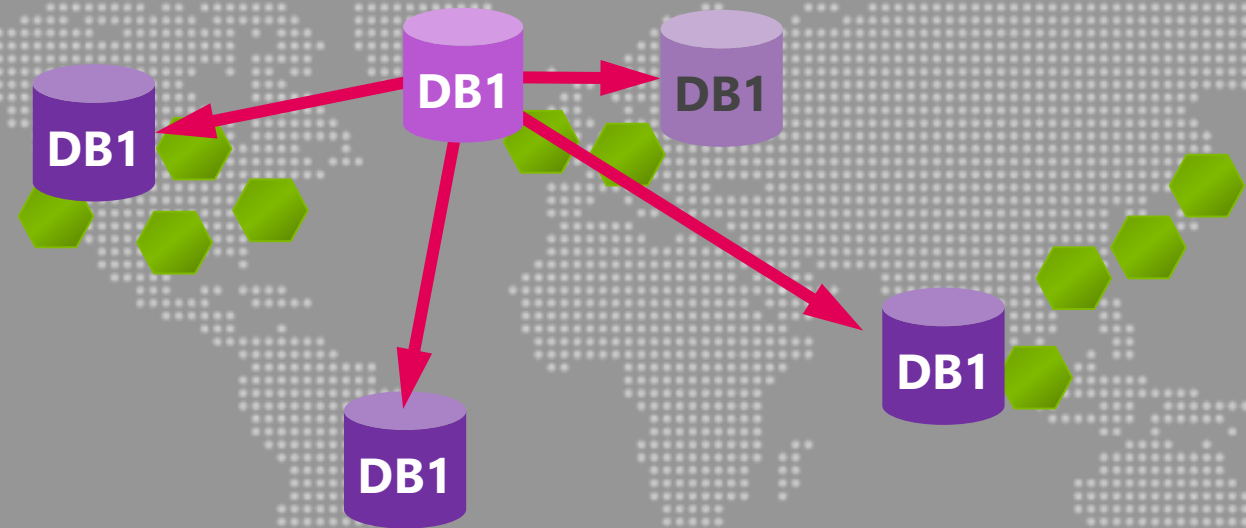
## Geo-replicare activă (Premium)

Până la patru replici în zone secundare

Control complet peste locația secundară

Suportă load balancing, actualizarea aplicațiilor și scenarii de relocare

Poate fi combinată cu o replica din zona secundară



Geo-replicarea minimizează discontinuitatea business-ului

# "Zero-admin" și "Self-managed"

## Zero-administrare

Infrastructura virtualizată elimină aproape tot efortul de întreținere, inclusiv aplicarea patch-urilor.

Platforma HA tolerantă-la-erori nu necesită monitorizare

Generarea de back-ul automat

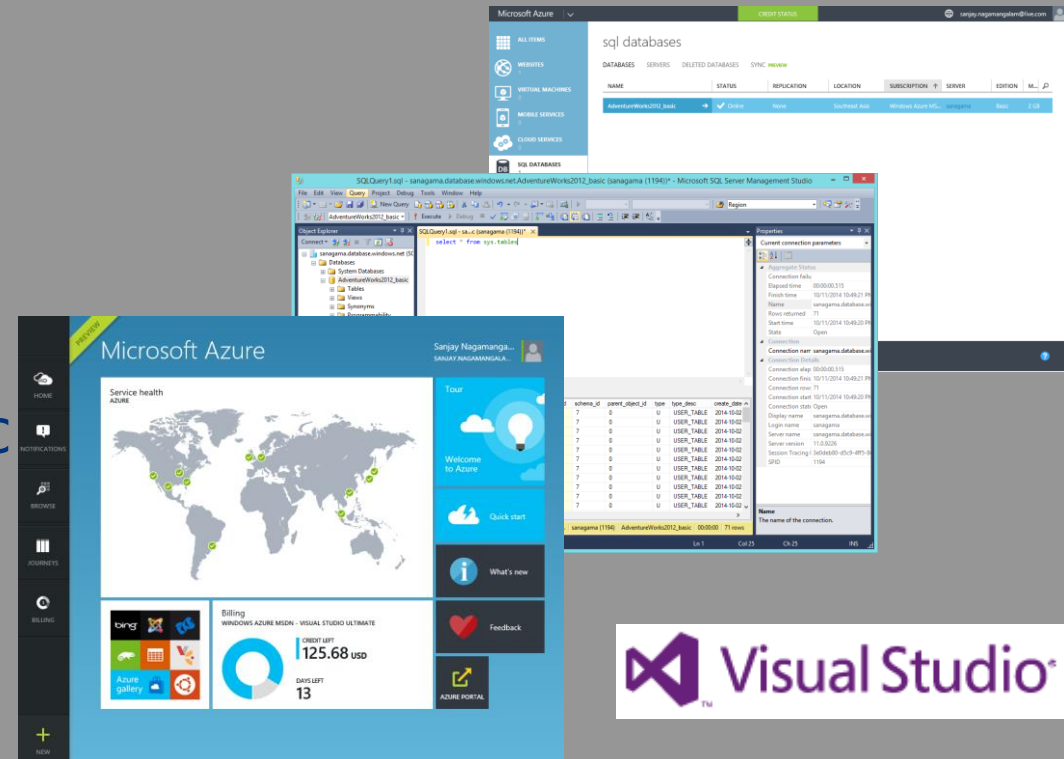
## Self-service management

Realizat prin Azure management portal, T-SQL, REST APIs, PowerShell  
Provision, copy, delete, restore, configure geo-replication, auditare,  
export/import, etc.

Mentenanță și toleranță-la-erori prin servicii built-in.

# Dezvoltarea de aplicații

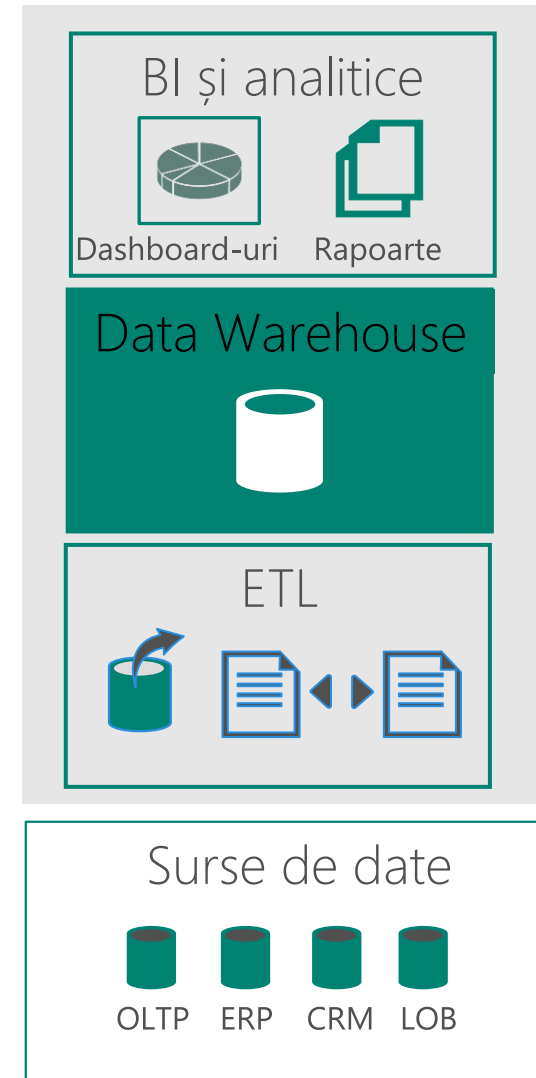
T-SQL, REST API-uri, PowerShell  
SQL Server Management Studio (SSMS),  
Visual Studio  
Suporta platforme și tehnologii dintre care  
amintim .NET, Java, Ruby on Rails, Node.js etc  
Azure Machine Learning Services





# Data Warehouse Tradițional

... data warehouse a atins cel mai important punct al evoluției sale. Cel mai elaborat sistem de gestiune a datelor din industria IT care se transformă de o manieră radicală.



# Azure Synapse Analytics(SQL Data Warehouse)

- Serviciu dedicat încărcării de volume mari de date, cu posibilități de scalare "up and down" si expus "full managed".

	Azure SQL Database	Azure SQL DW
Tipul de date	Relațional	Relațional
Data Masking dinamic	DA	NU
Data Encryption	DA	DA
Interogări T-SQL Polybase	DA	DA
MPP(Masive Parallel Processing)	NU	DA
Cant. de date /baza de date	4TB	1PB
Max. sesiuni concurente	30000	1024
Max. interogări concurente	6400	32

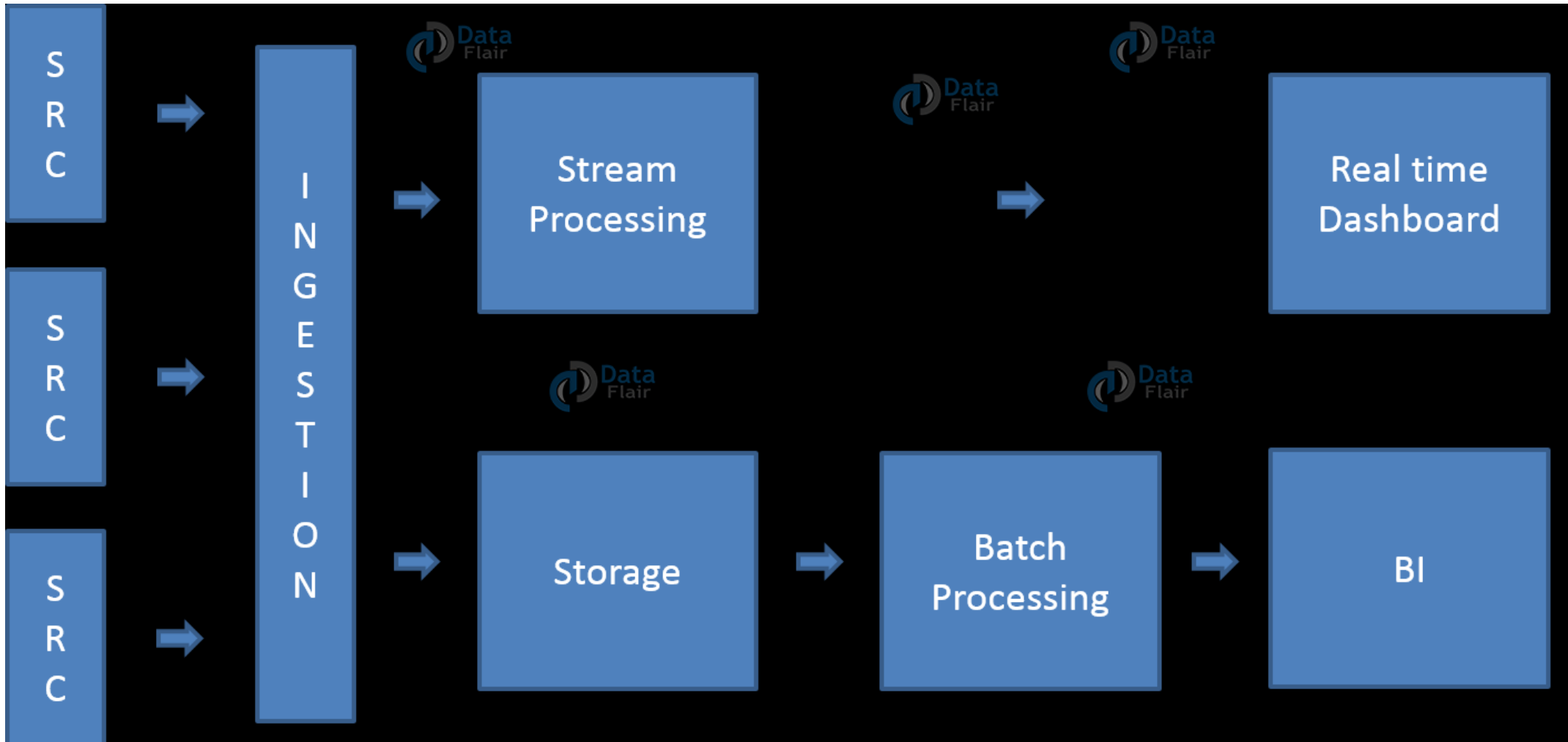
# Big Data este motorul transformării

“Big data reprezintă o familie de active informaționale care se caracterizează prin “high-volume”, “high-velocity” și/sau “high-variety” care implică costuri-efective, forme inovative de procesare a informației care conduc la o cunoaștere îmbunătățită a proceselor interne, a modului de luare a deciziilor, și a automatizării proceselor.”

Gartner, Big Data Definition\*

\* Gartner, Big Data (Stamford, CT.: Gartner, 2016), URL: <http://www.gartner.com/it-glossary/big-data/>

# Arhitecturi Big Data



# Big Data implică modificări radicale



Caracteristicile datelor

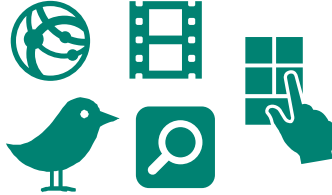



Costuri

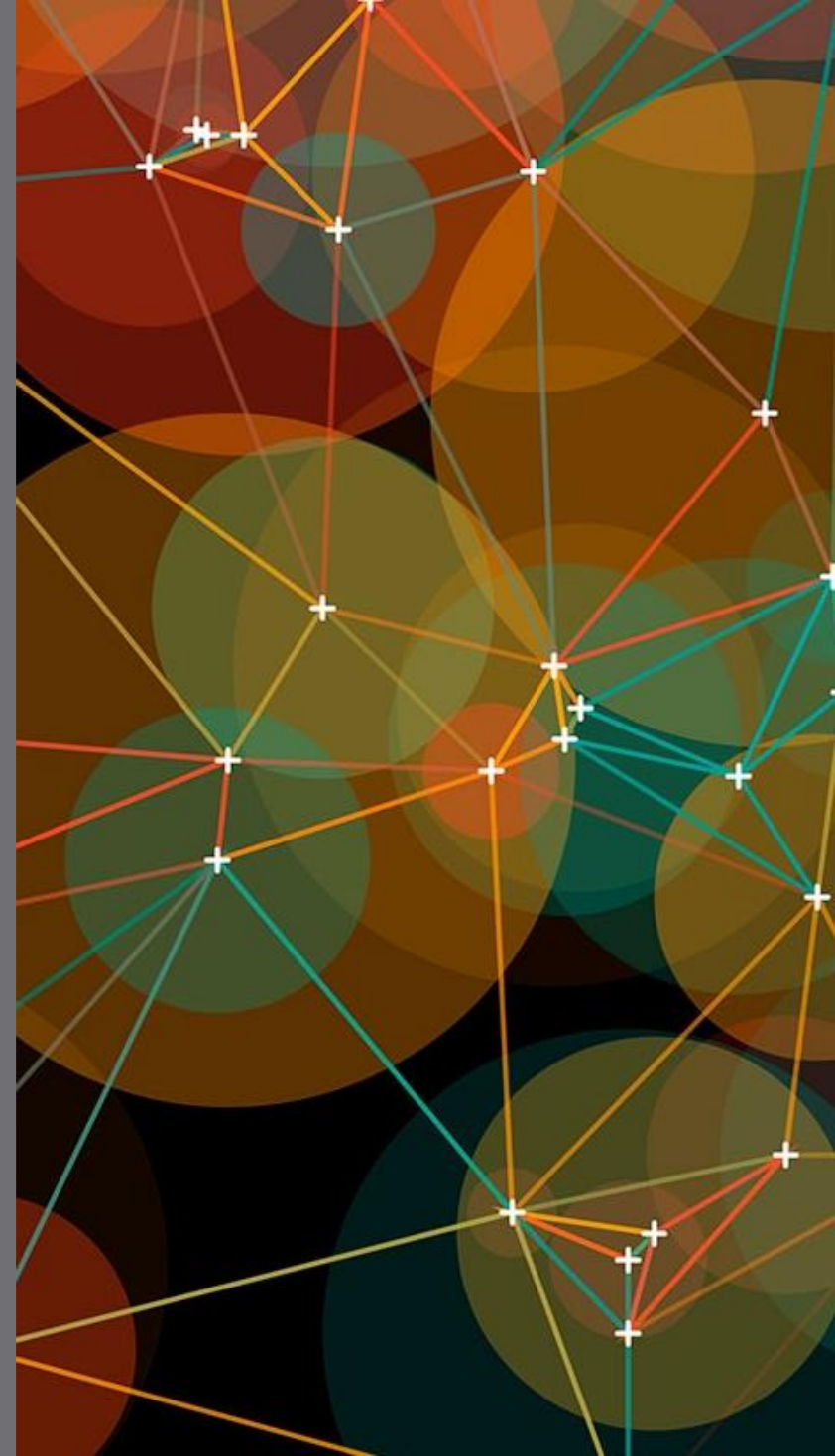


Cultură

# Big Data induce modificări radicale

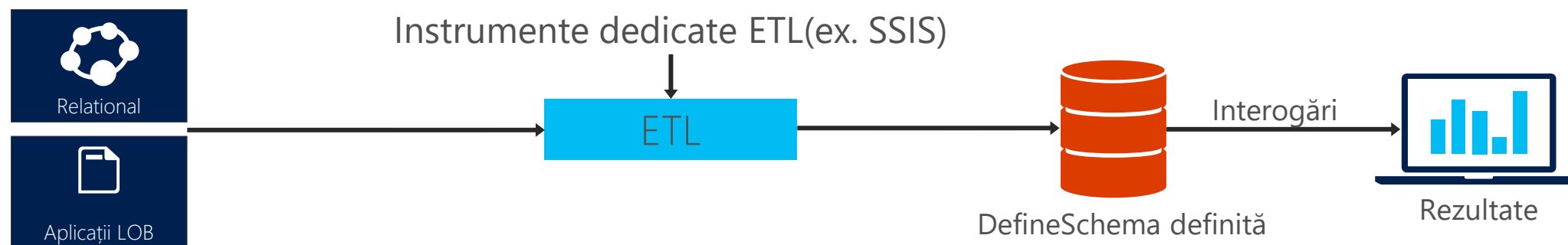
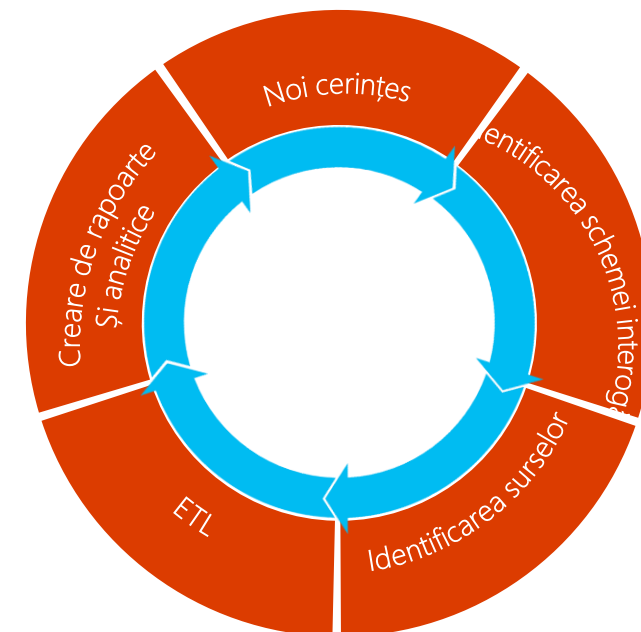
	Tradițional	Big Data
Caracteristicile datelor	 <p>Relațional (cu dependență de schemă)</p>	 <p>Date (cu schemă adaptivă)</p>
Costuri	 <p>Costuri mari (capacitate mare de calcul și stocare)</p>	 <p>ieftine (stocarea și capacitate de calcul)</p>
Cultură	 <p>Vizualizare rapoarte (utilizarea algebrei relaționale)</p>	 <p>Acțiuni inteligente (utilizând alg. relațională și ML)</p>

# Data Lake



# Analitice asociate business-ului tradițional

1. Se pornește de la cerințele end-user-ului pentru a identifica rapoartele și analizele dorite.
2. Definirea corespunzătoare a schemei bazei de date și interogărilor.
3. Identificarea surselor de date
4. Crearea unui ETL(Extract-Transform-Load) pentru extragerea datelor cerute și transformarea acestora schemei țintă ('*schema-on-write*')
5. Crearea de rapoarte și analize.

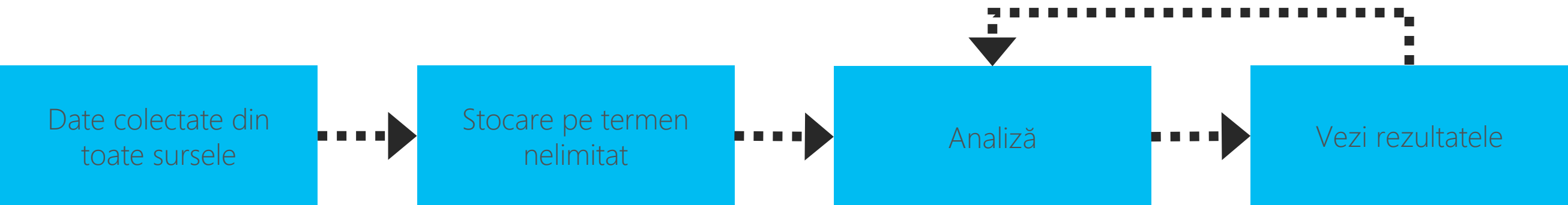


Toate datele care nu sunt necesare imediat sunt eliminate sau arhivate

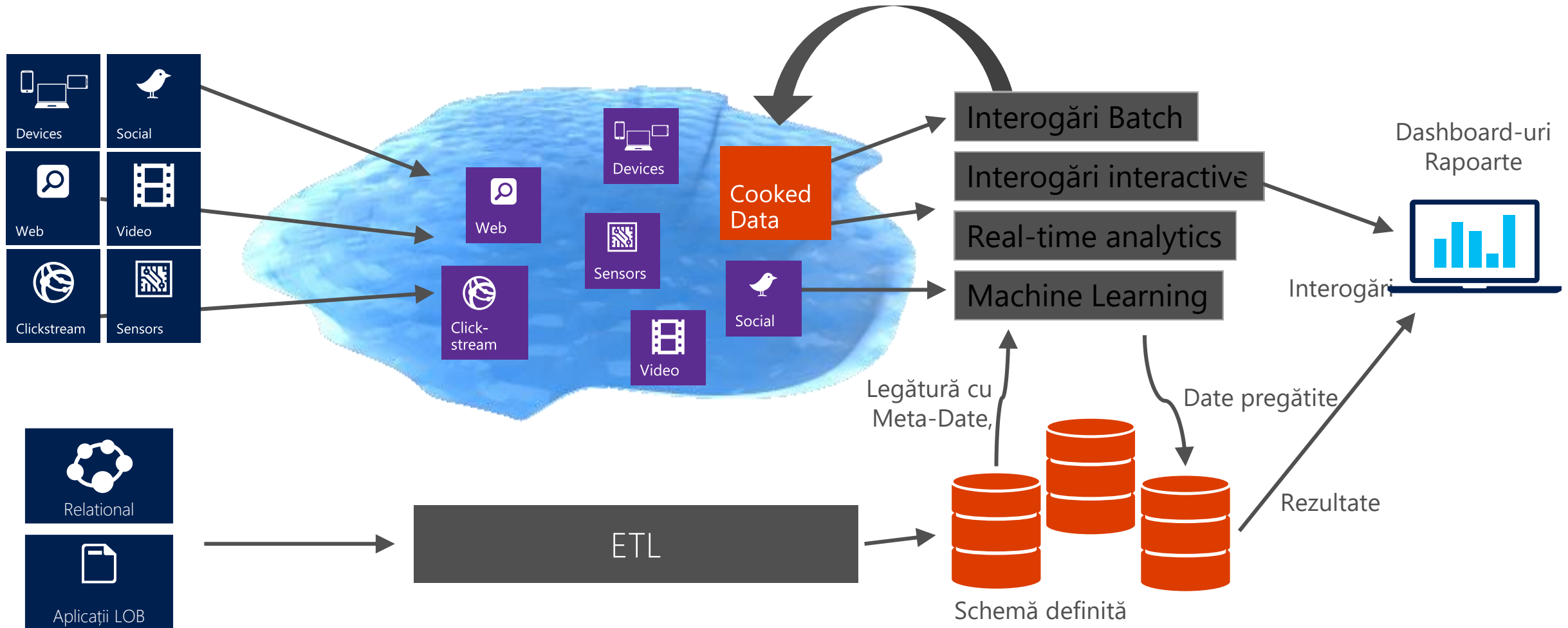


# Big Data: Toate datele sunt valoroase

- Toate datele au o valoare potențială
- Date tezaurizate
- Fără schemă definită—încărcate în format nativ
- Schema este impusă și transformările se fac la momentul interogării(*schema-on-read*).
- Aplicațiile și utilizatorii interpretează datele așa cum consideră de cuviință



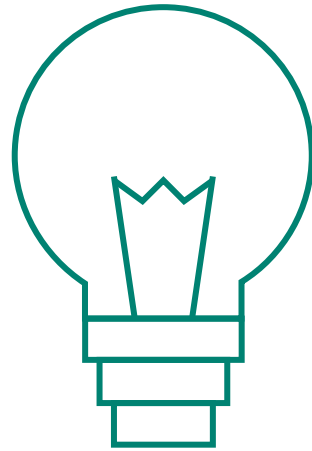
# Data lake și warehouse



# ... Big Data nu este un lucru banal



Obținerea de  
competențe

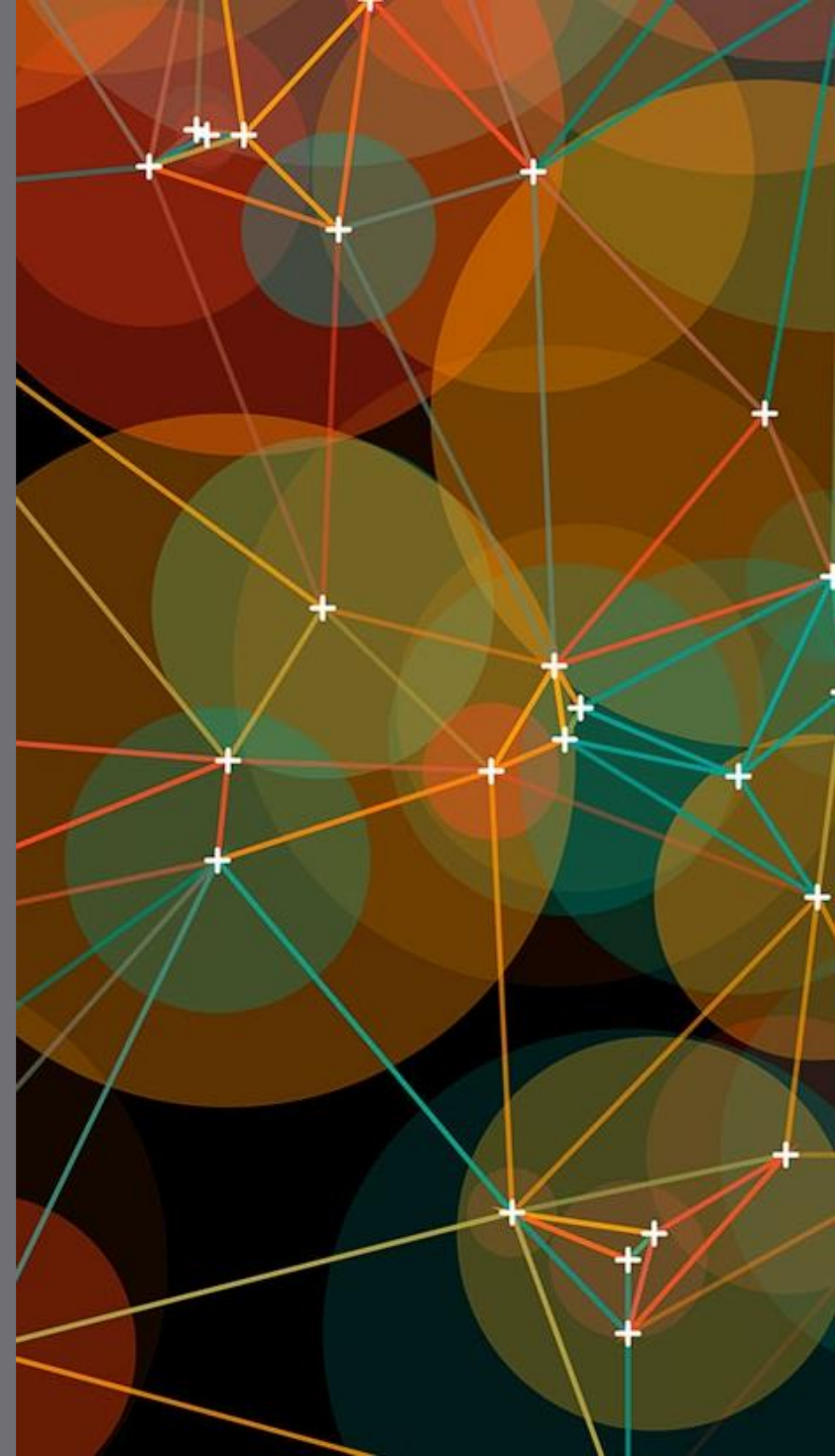


Determină modul  
în care putem obține  
valoare

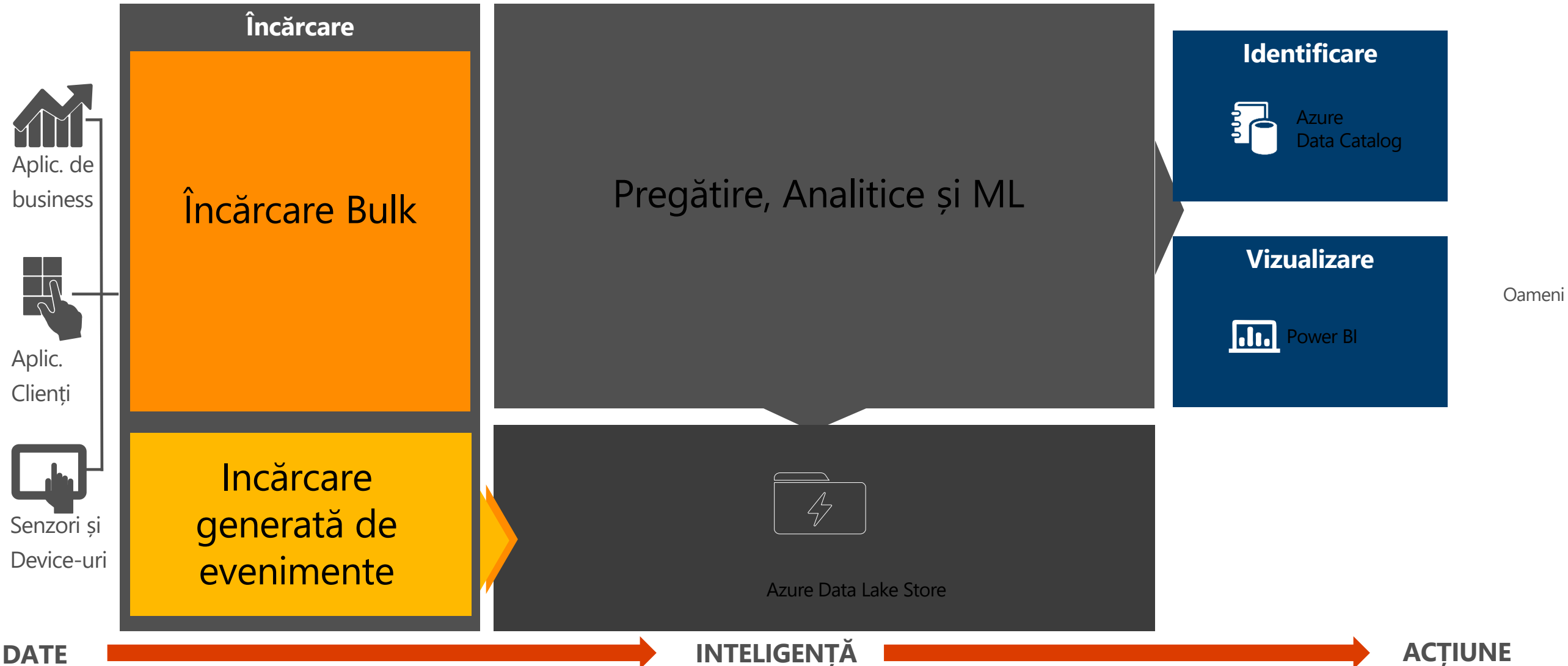


Integrarea cu  
Investițiile IT existente

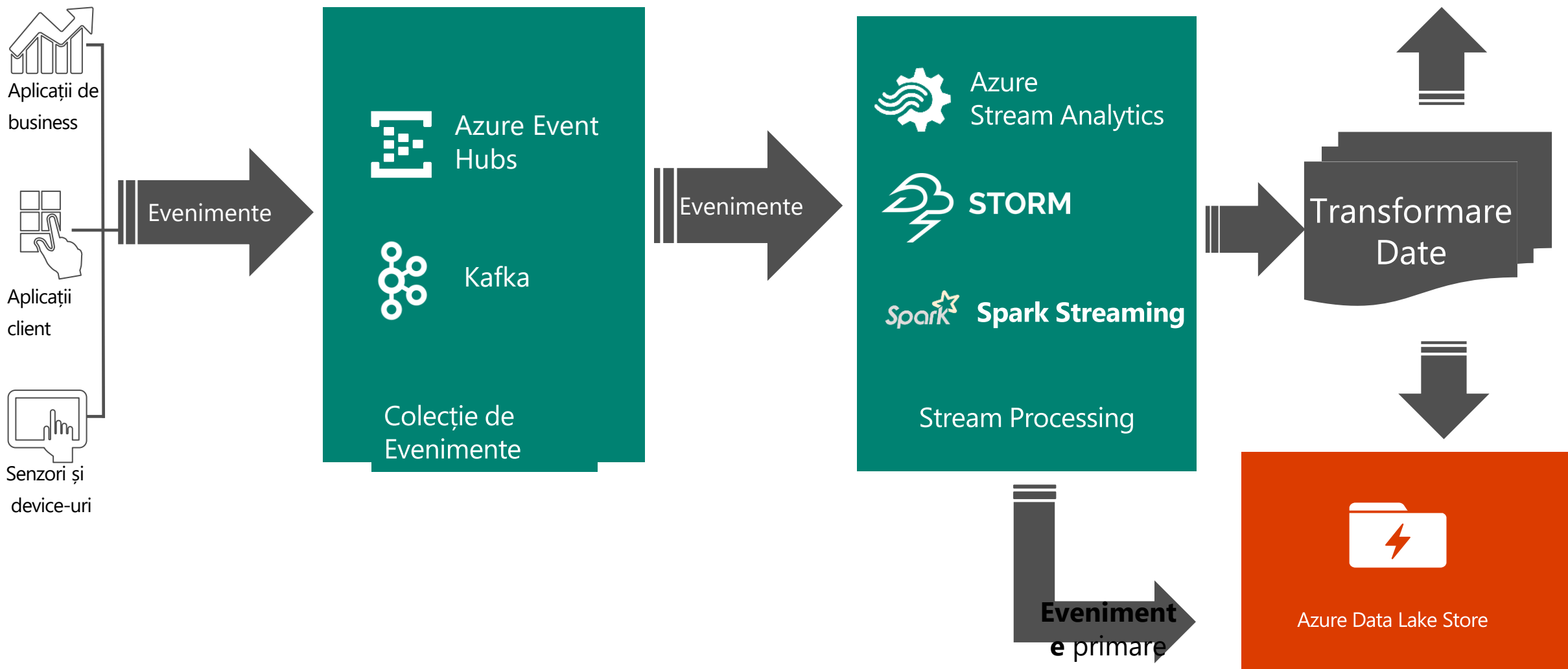
# Pattern-uri pentru Big Data



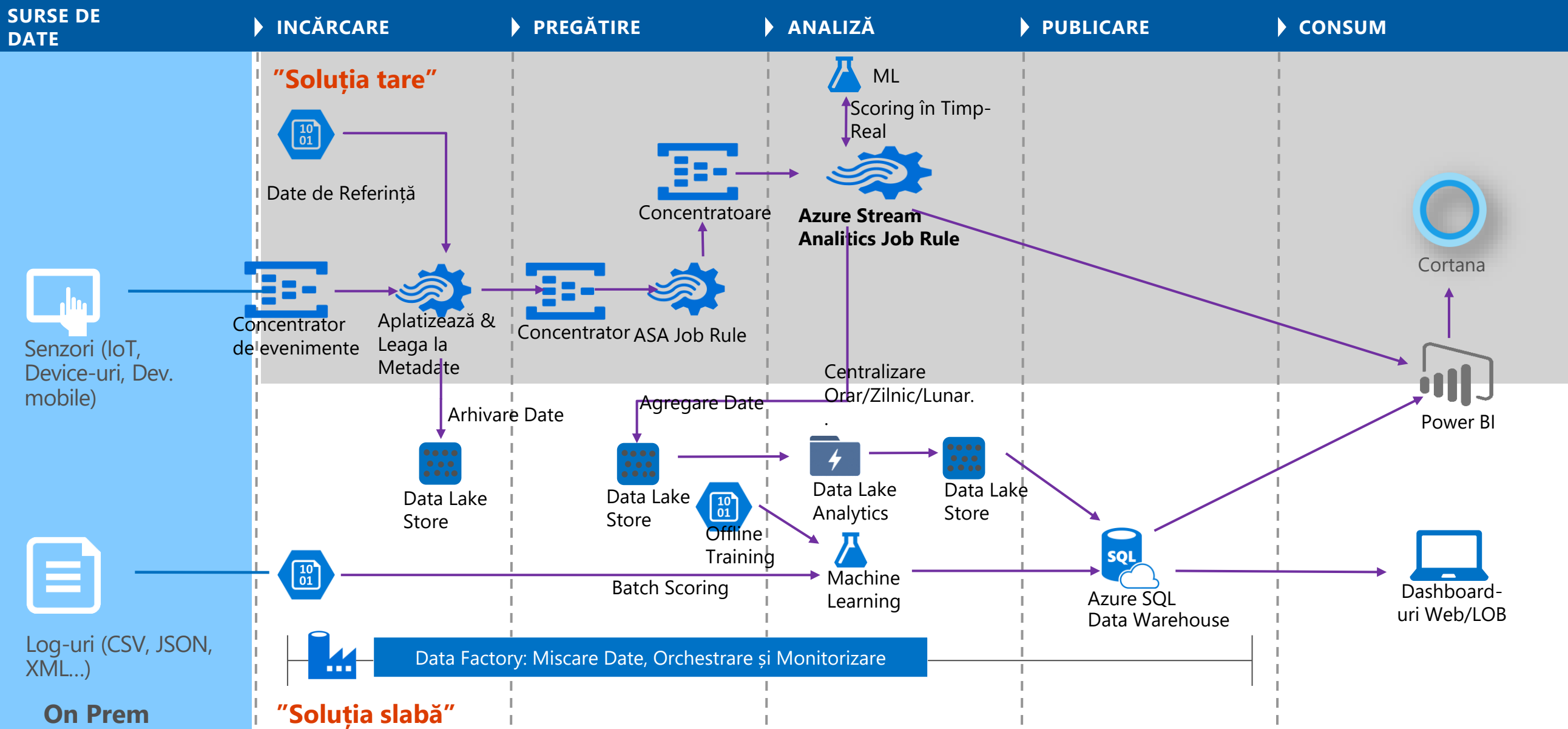
# Big Data Analytics – Fluxuri de date



# Pattern-uri de Încărcare Date furnizate de Evenimente



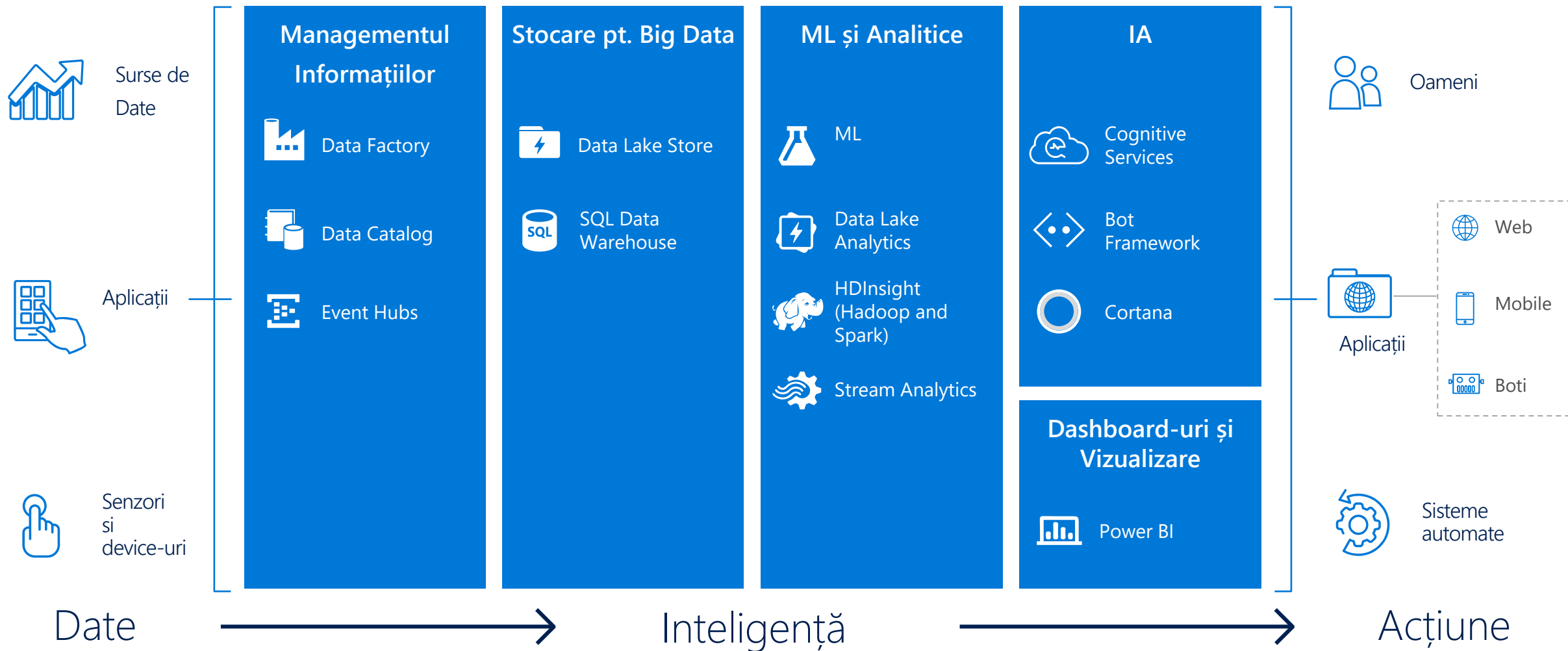
# Arhitectură Lambda



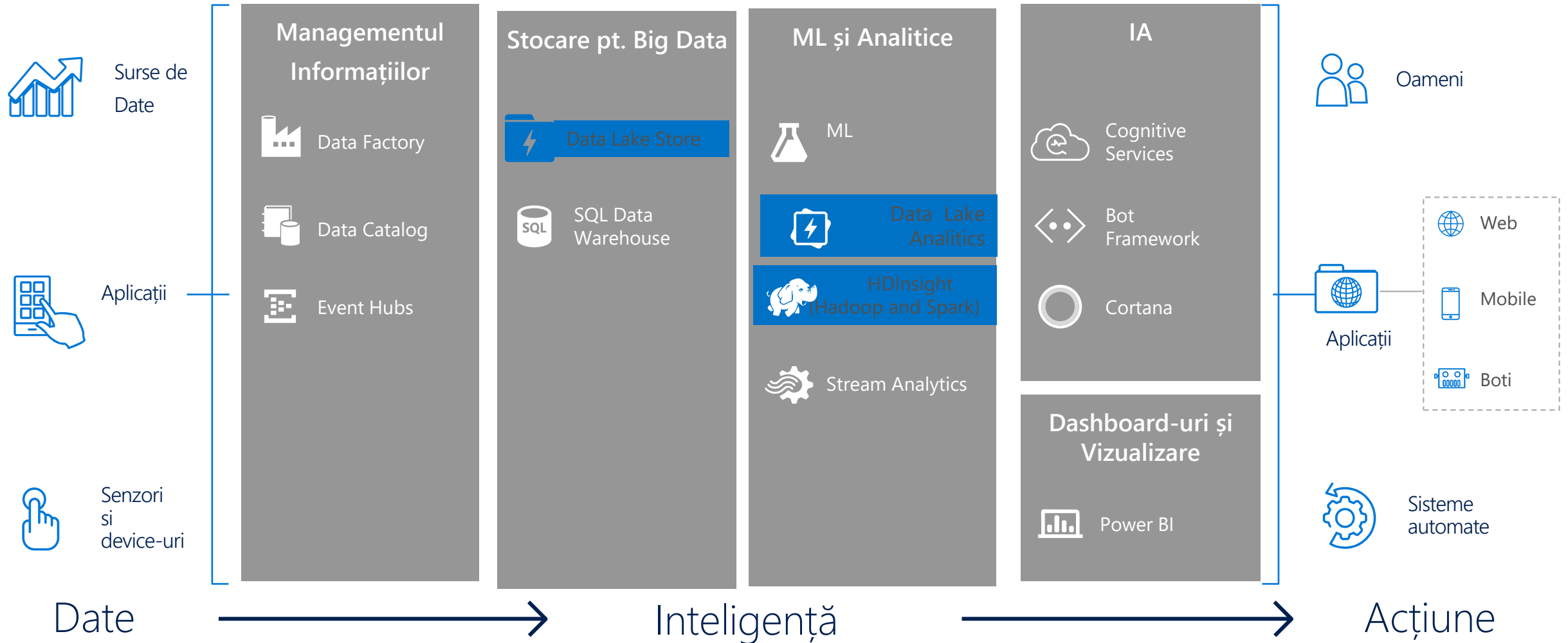
# Trecerea la big data – o cale dificilă

- Selectarea unei stive open source
- Selectați mai multe componente din zecile care sunt disponibile în baza volumului de lucru.
- Achiziție de hardware, networking, etc
- Instalare și management, de ex. cu Ambari
- Instalarea componentelor Big Data
- Instalarea serviciilor opționale big data
- Adauga serviciile de securizare și autentificare
- Configurare și testarea cluster
- <câte și mai câte ...>

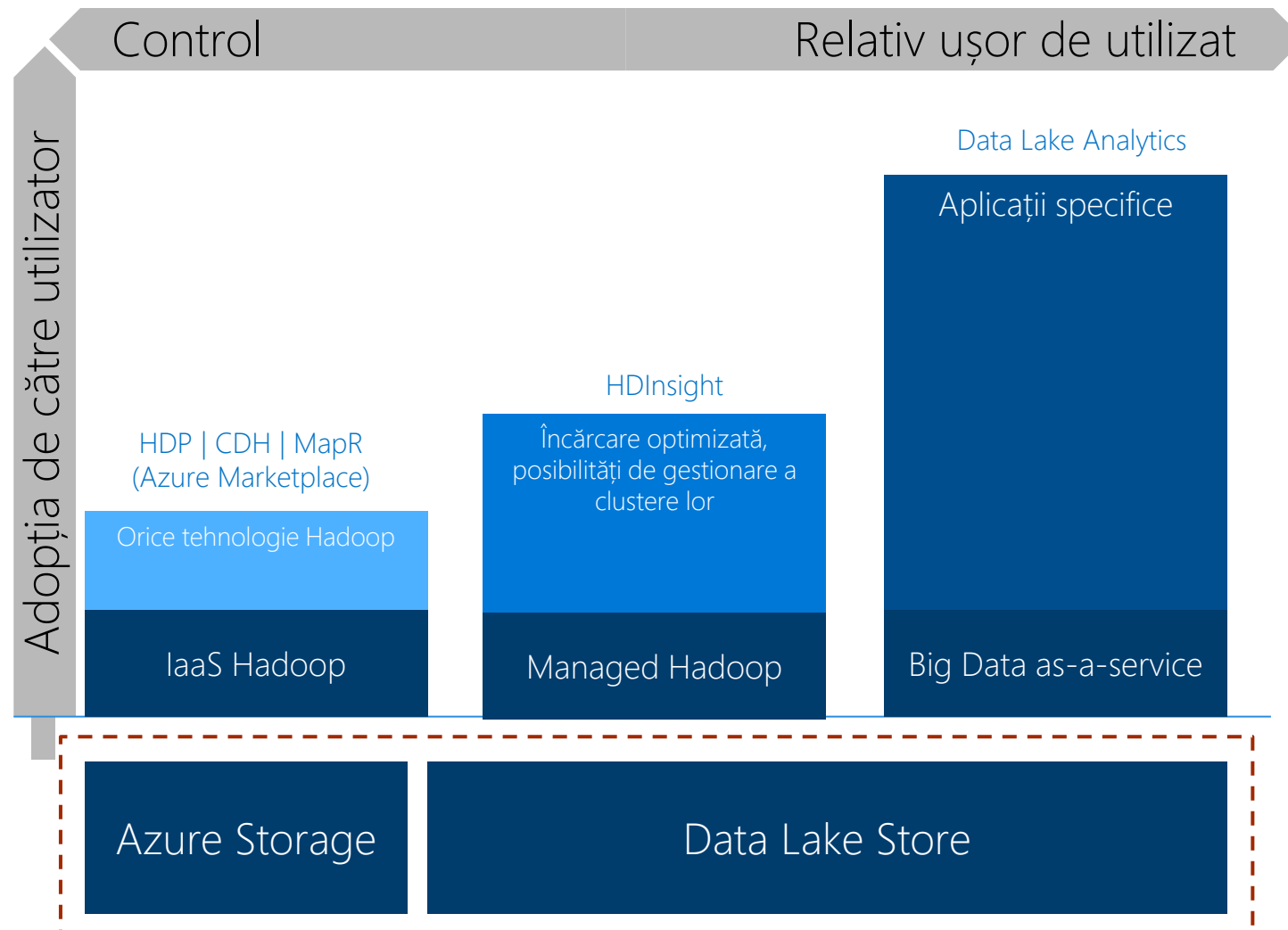




# Unde Big Data este o soluție fundamentală



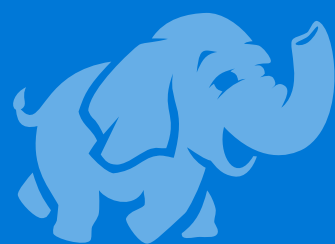
# Democratizarea Big Data



**Construit pentru cloud pentru a accelera ritmul inovării pe o platformă calcul modernă**

# Azure HDInsight

Hadoop și Spark  
ca Serviciu în Azure



**Fully-managed** Hadoop and Spark  
pentru cloud

**100% Open Source** - Platforma de date de  
**la** Hortonworks

Clusters disponibil **relativ ușor**

Familie **de instrumente BI pentru analiză**,  
sau medii open source pentru **data science**  
**interactiv**

**63% costuri mai mici ale TCO** decât  
utilizare Hadoop on-prem

# Azure Data Lake Store

Un hyper-scalabil  
storage pentru analitice  
Big Data



Hadoop File System (HDFS) adaptat la cloud

**Scalare** fără limite

Încarcă **orice date** în format nativ.

Optimizat pentru lucrul cu analitice  
performante

# Azure Data Lake Analytics

Serviciu distribuit de analitice



**Serviciu distribuit de analitice** bazat pe  
Apache YARN

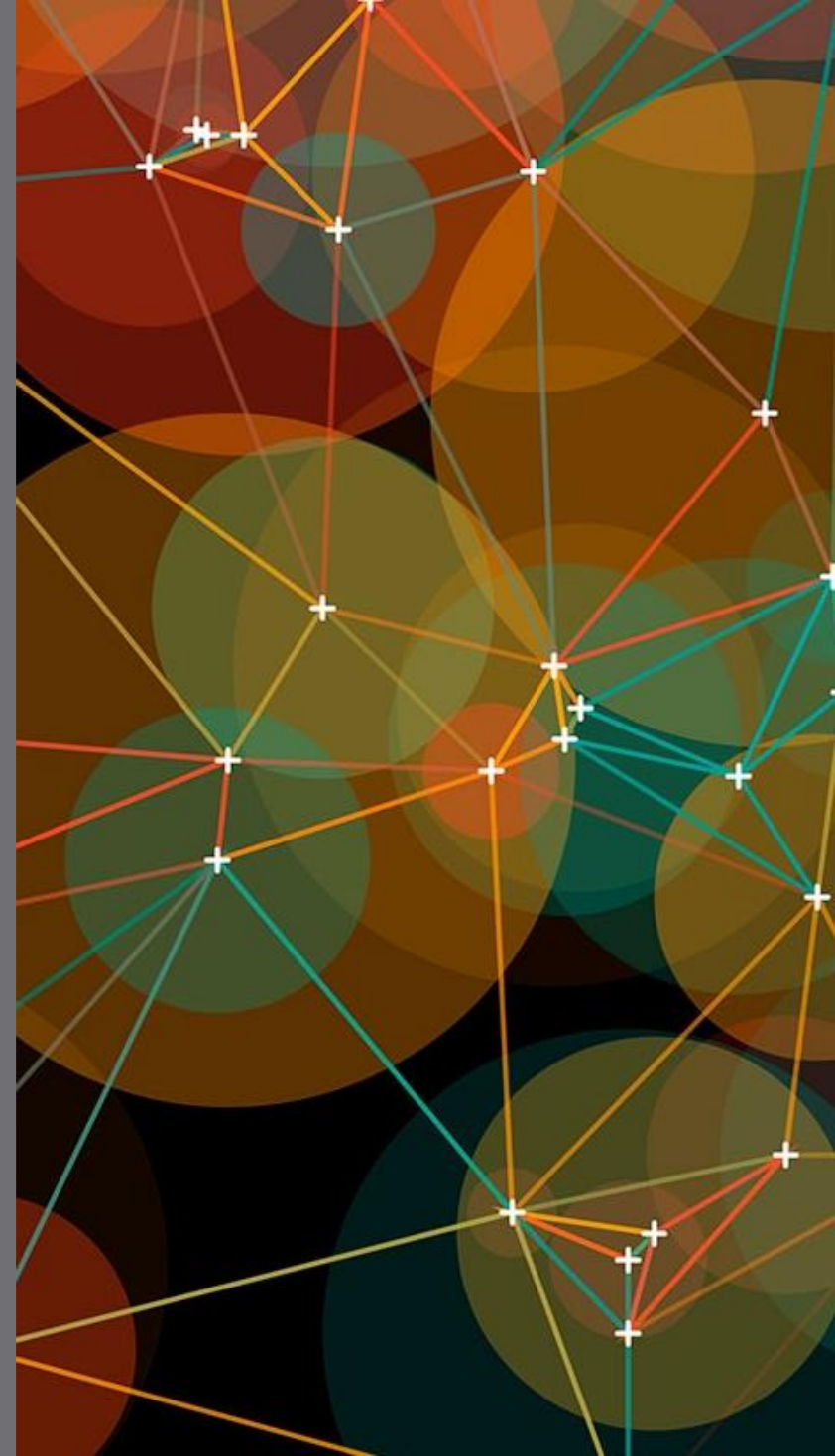
**Sistem elastic de interogare** permite  
utilizatorilor să se concentreze pe  
problematika de business —nu pe  
configurări de mediu

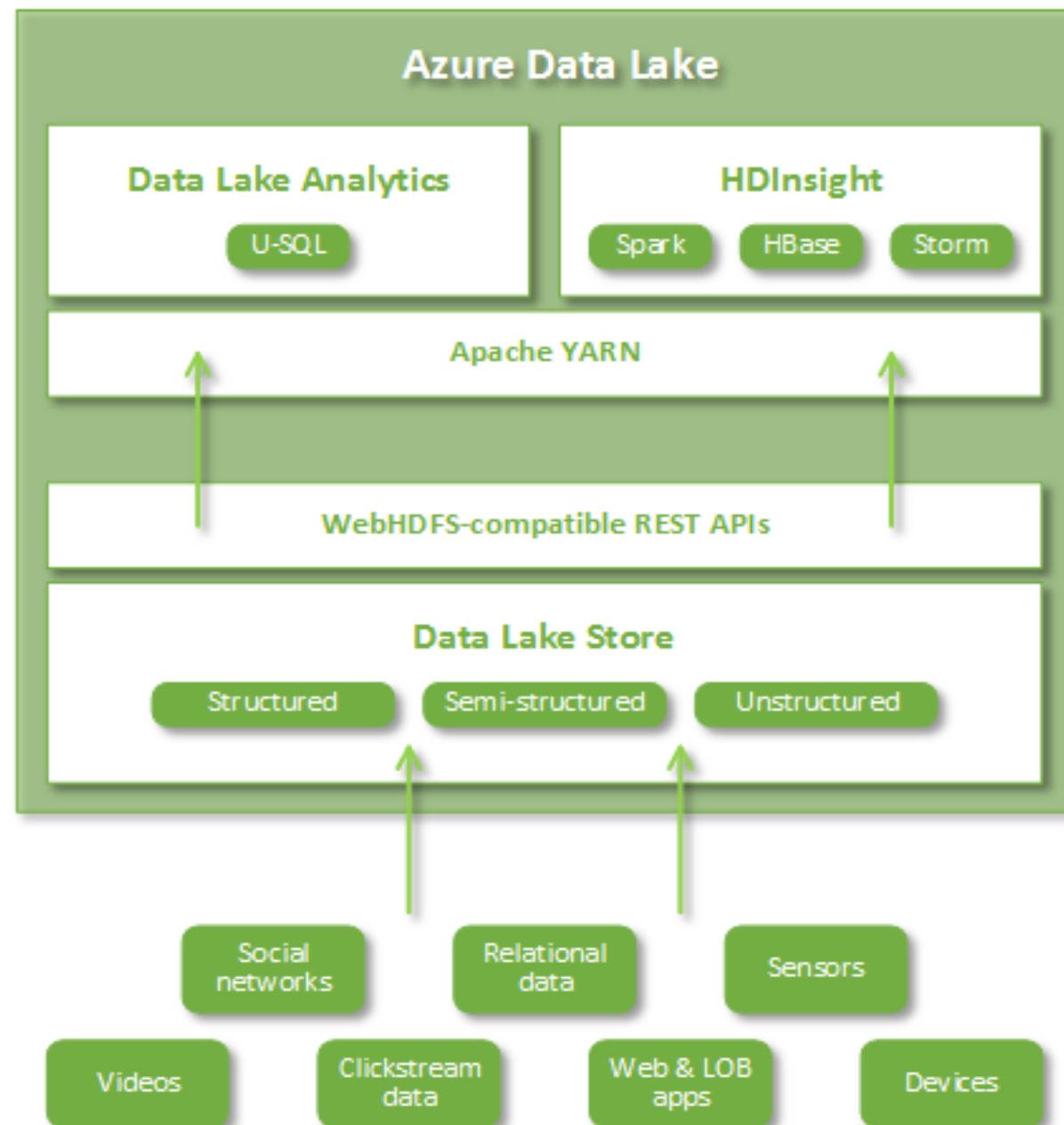
Include U-SQL—un limbaj care unifică  
**beneficiile SQL puterea C#**

**Integrat cu Visual Studio** pentru  
dezvoltare, debug, și optimizare de cod

**Interogări federalizate** dealungul mai  
multor surse de date din Azure.

# Sintetic







DEMO

