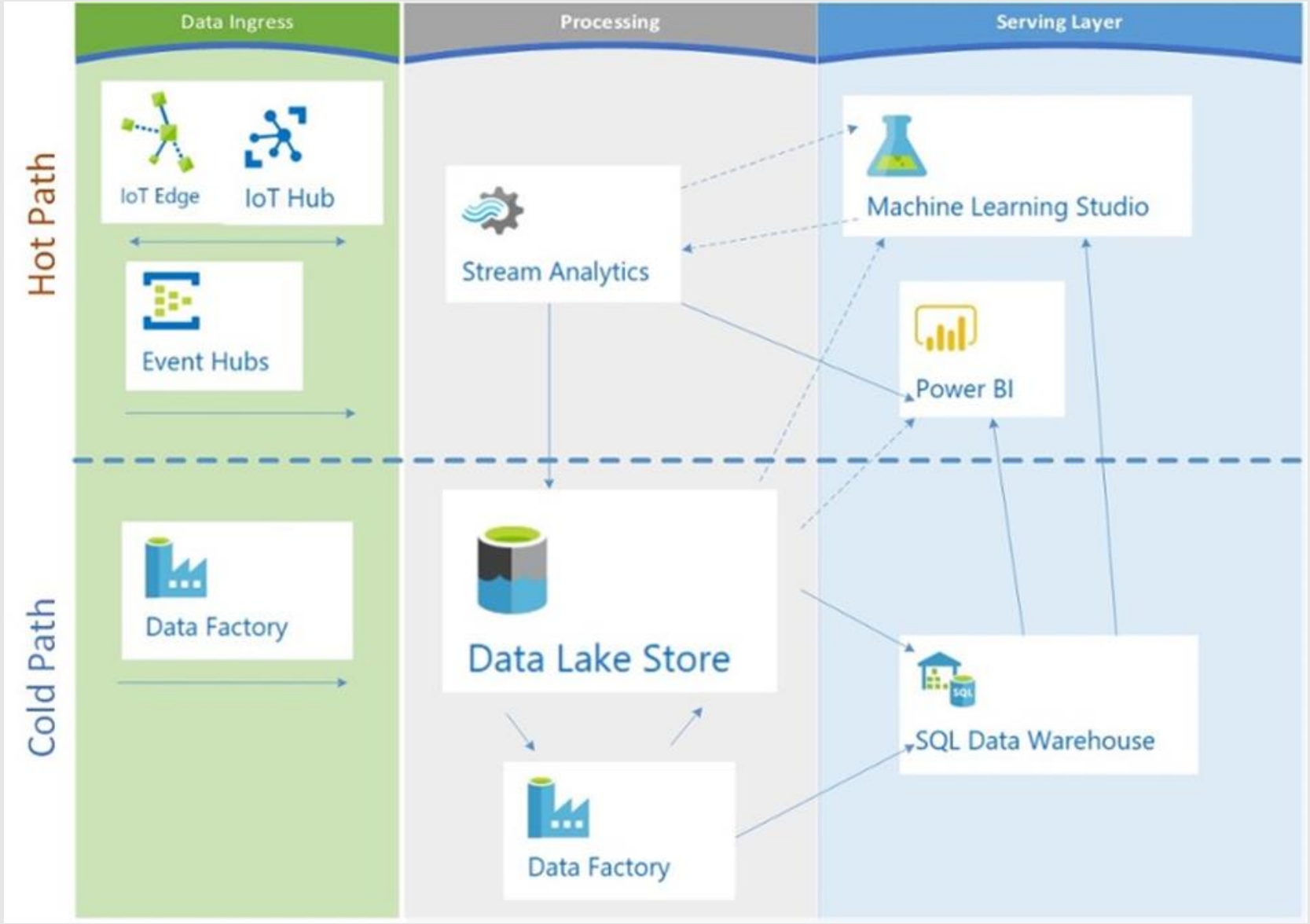
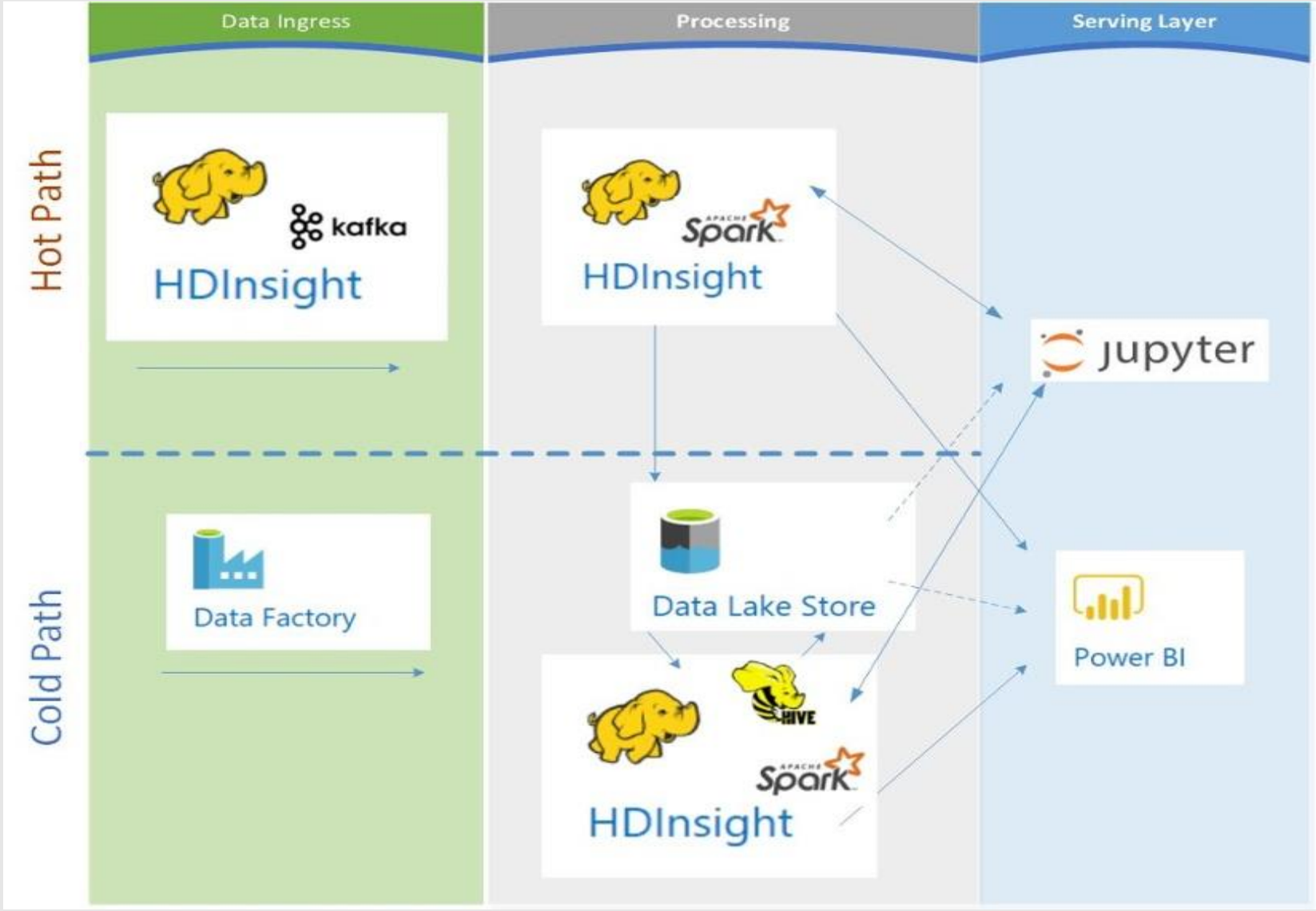
ETL Modernizat cu ADL: Scalare, formate-multiple, platforme-multiple, și analitice

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

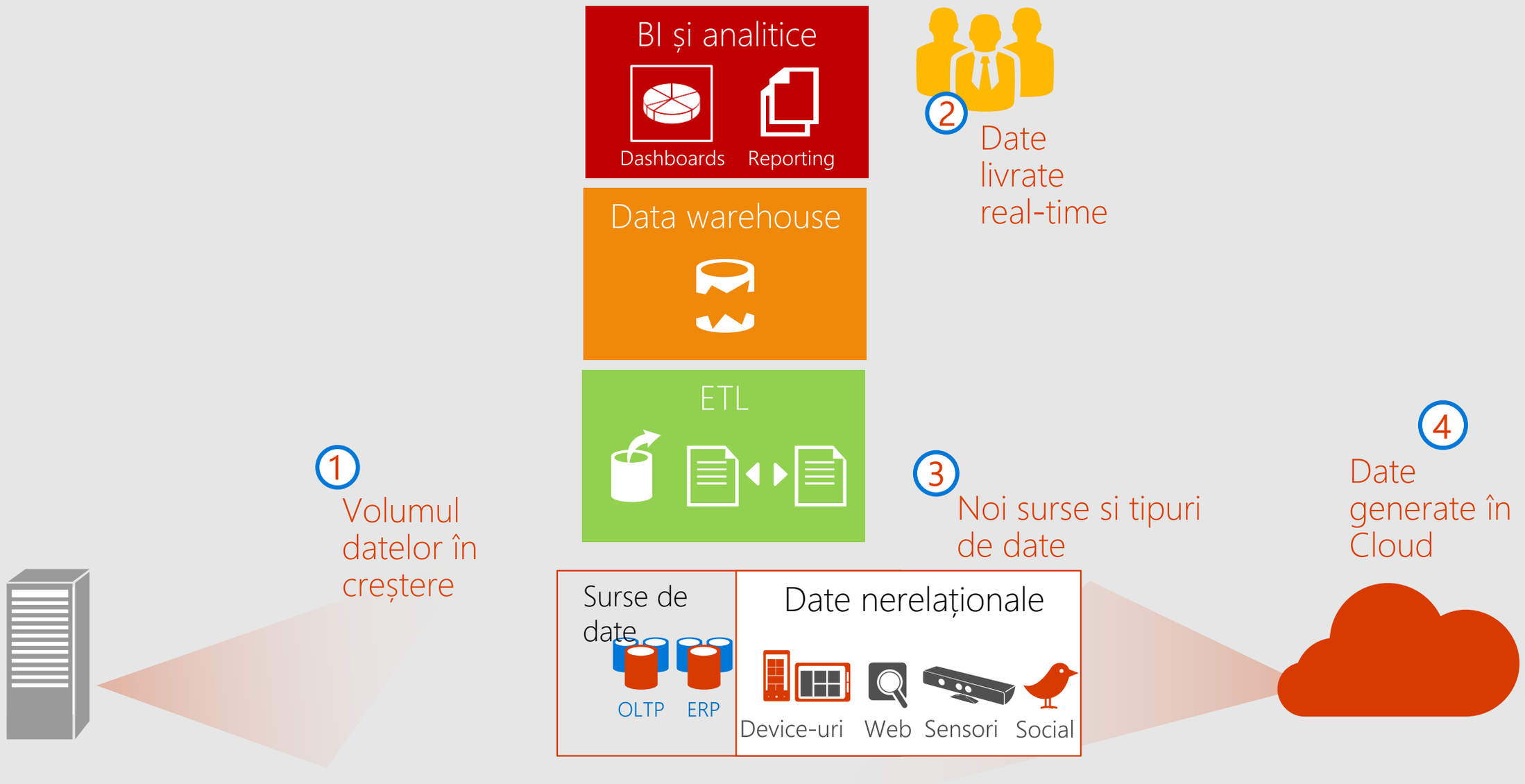
Utilizare Servicii Azure în Tehnologii Lambda



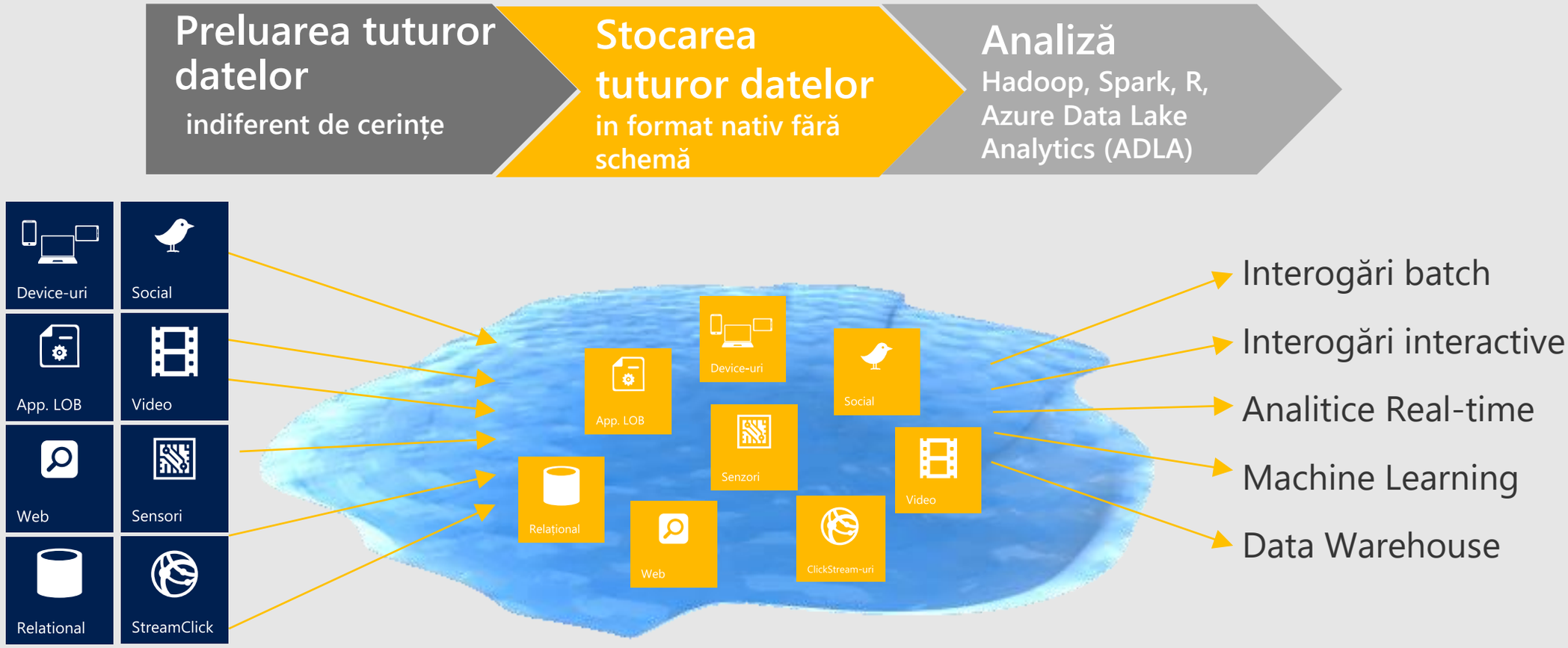
Arhitecturi de referință pentru HDInsight



Data Warehouse Tradițional



Abordare bazată pe Data Lake

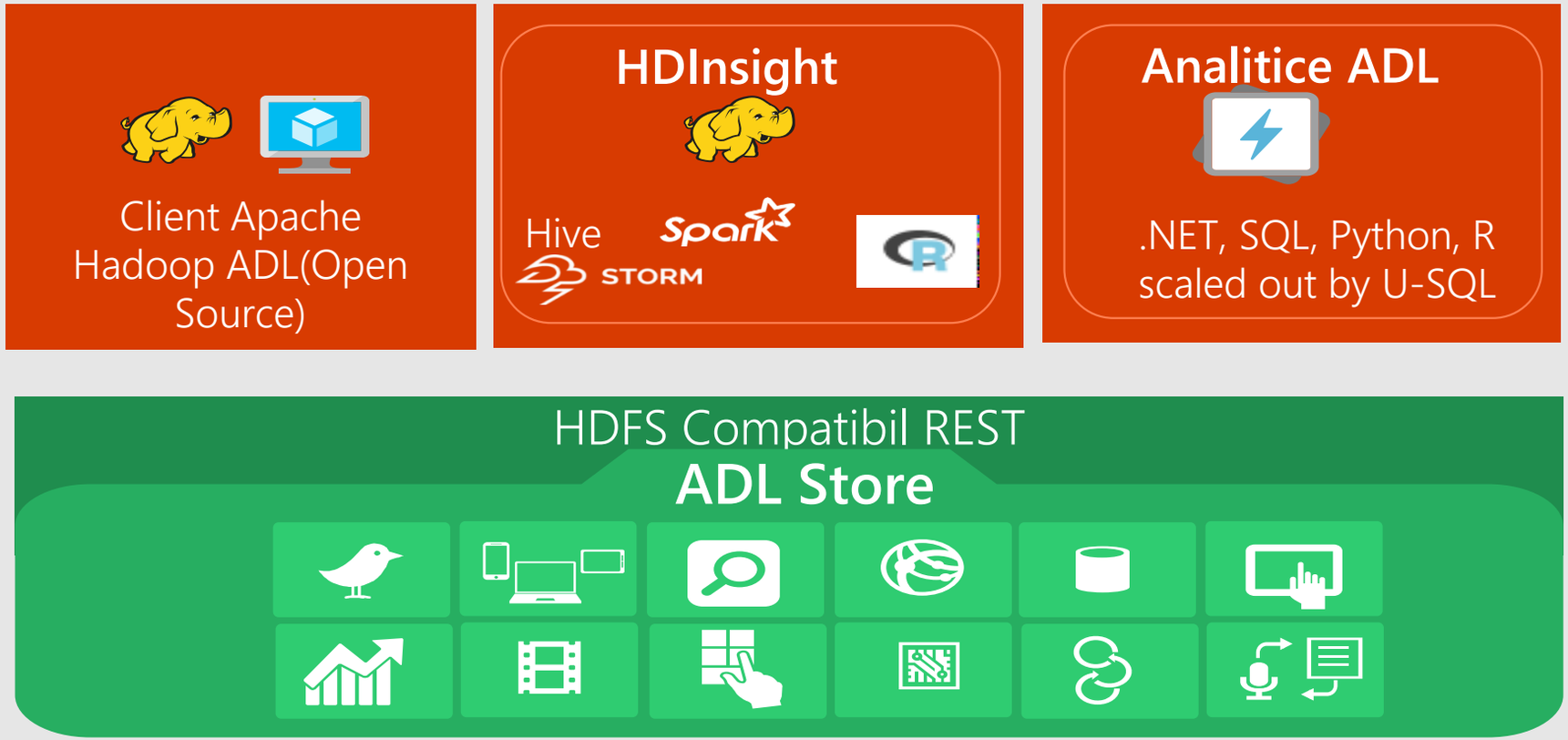


PROIECTAT PENTRU ÎNTREBĂRI PE CARE NU LE ȘTIM ÎNCĂ!

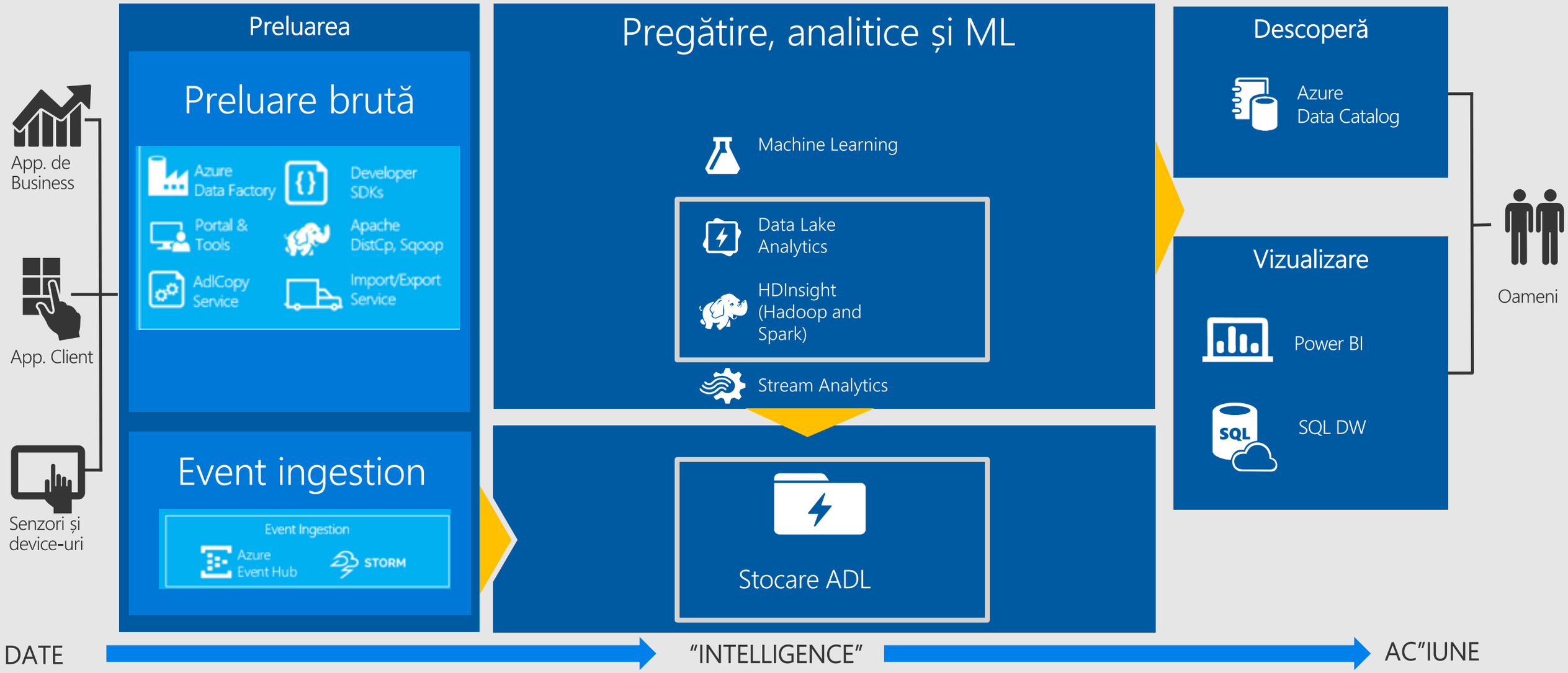
Azure Data Lake

An on-demand, real-time stream processing service with no-limits data lake built to support massively parallel analytics

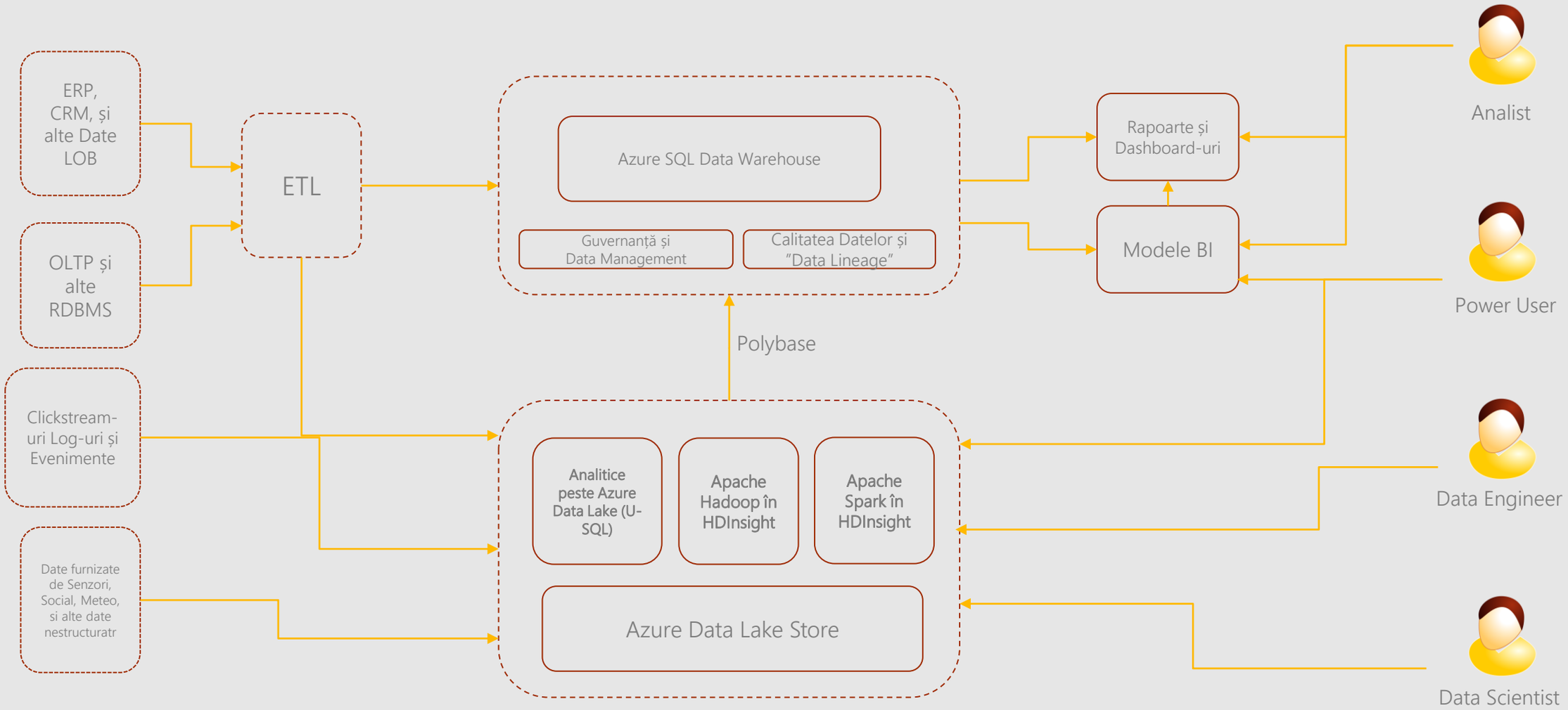
- Performanță în scalare
- Optimizat pentru analitice
- Multiple motoare de analitice
- Un singur repository partajat



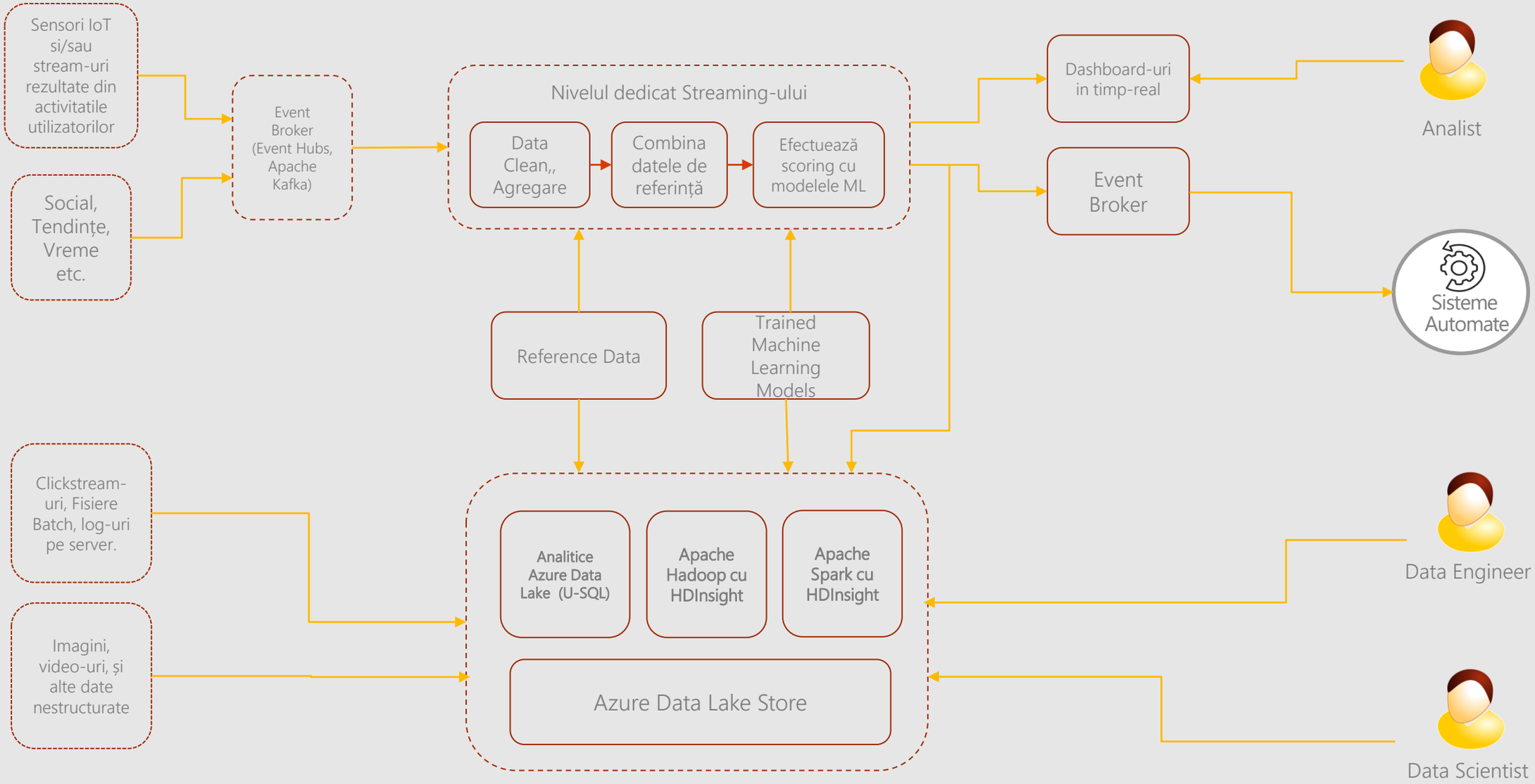
Canalele Big data și fluxurile de date în Azure



Big Data Warehouse



Procesare în Timp-Real cu Arhitectura Lambda



Load-Extract-Transform-Store & Share

Schematizând lucrul cu date nestructurare în procesele de analiza avem:
Load—Extract—Transform—Store

Pregătirea datelor pentru alti utilizatori(LETS & Share)

- Ca date nestructurate
- Ca date structurate

Procesare personalizată cu cod personalizat

Procesare Cognitiva în Cloud

Extinderea potențialului big data cu date de valoare ridicată de acolo de unde provin

Analitice ADL

Data Lake Analytics Workloads

BATCH este ideal pentru incarcarea unor volume mari de date în Data Lake Analytics:

- **Transformarea și pregătirea** datelor pentru utilizare în alte sisteme.
- **Analitice** peste volume mari de date

- **Programe masiv paralele** scrise în .NET, Python și R, scalate cu ajutorul scripturilor U-SQL

- Implementarea de procese cognitive la scalare pe colecții mari de date

U-SQL

Un framework pentru Big Data

Utilizarea codului scris in .NET, Python, R peste Data Lake

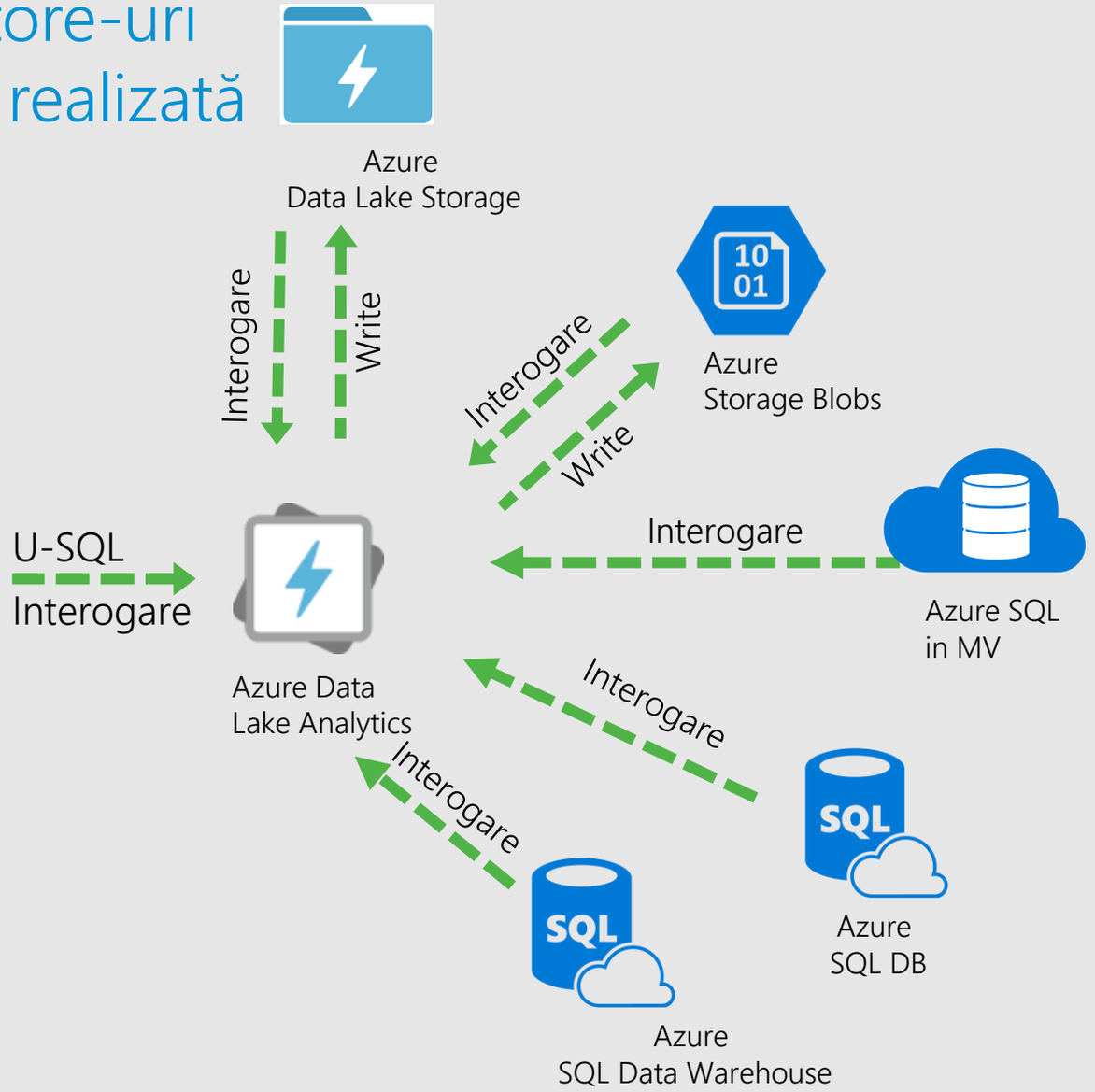
Sintaxă familiară dezvoltatorilor de SQL & .NET
Unifică:

- Natura declarativă a SQL cu puterea imperativă a C#
- Procesează date structurate, semi-structurate și nestructurate.
- Interoghează multiple Surse de Date din Azure (Interogări Federalizate)
- Procesează în mod Batch, Interactiv, Streaming & ML intr-un singur limbaj

Interoghează date în locațiile unde au fost stocate

Interogarea datelor în multiple data store-uri fără mutarea într-un singur store este realizată cu U-SQL

Unde este posibil, transformarea datelor este mutată în zona motorului de interogare la distanță pentru a minimiza transferul de date și pentru a maximiza performanța.



IA Integrat

Integreaza **D**eep **N**eural **N**etworks (DNN)

6 Functii Cognitive integrate

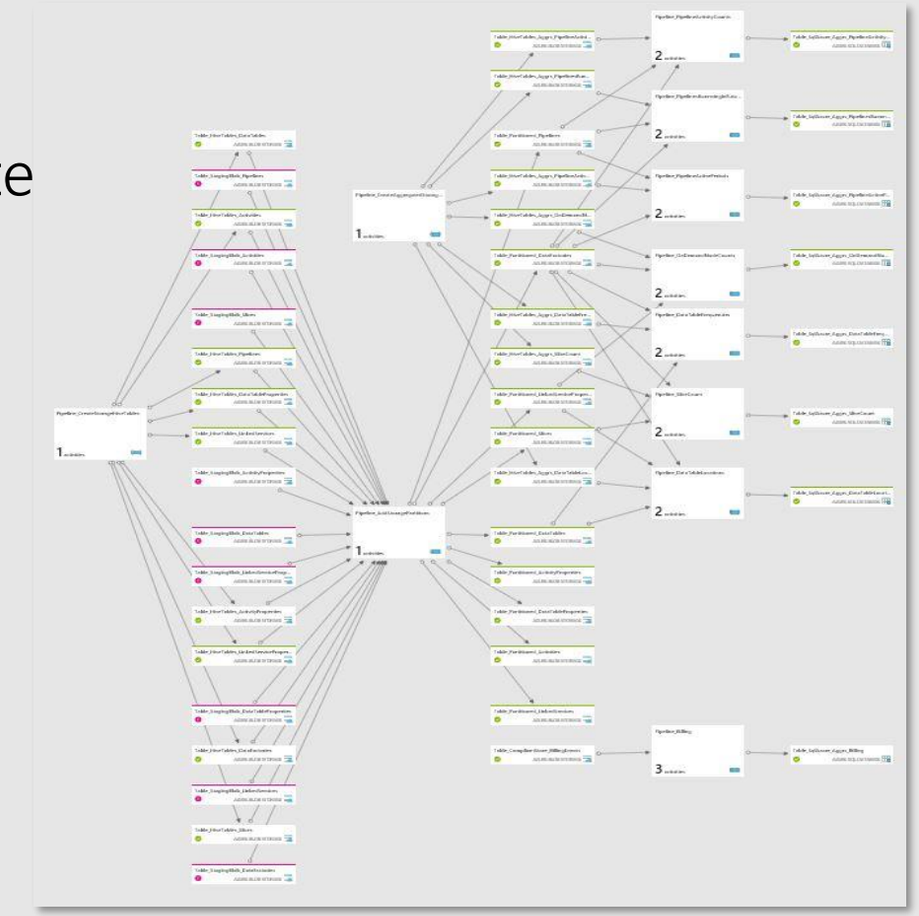
- Face API
- Image Tagging
- Analiza Emoțiilor
- OCR
- Text Key Phrase Extraction
- Text Sentiment Analysis

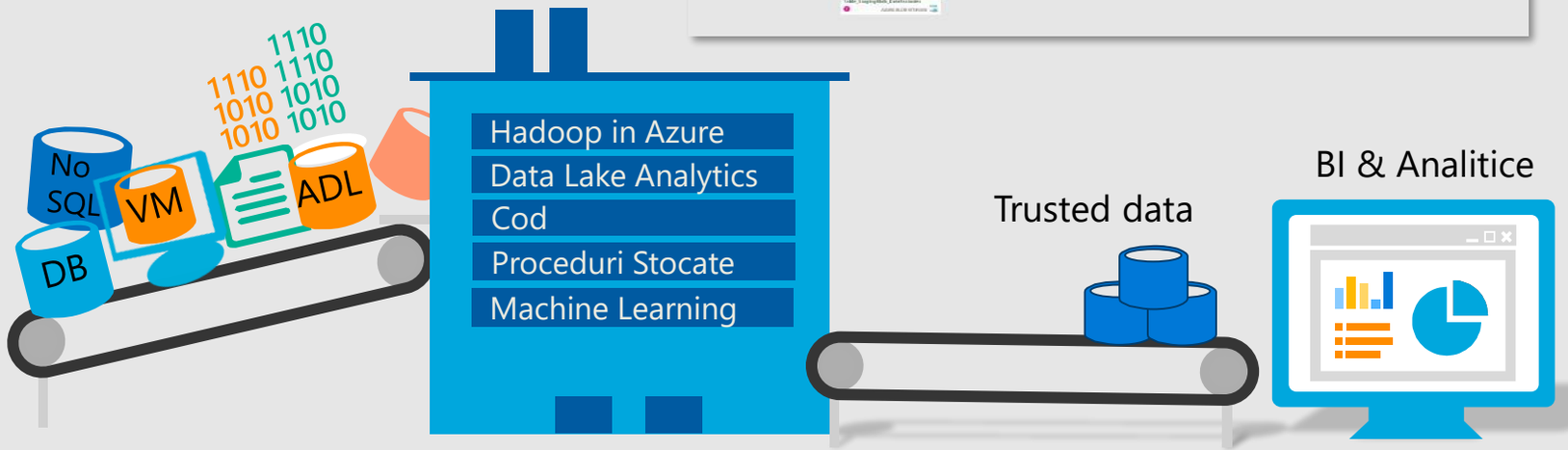
`

Azure Data Factory

Compune, orchestrează și monitorizează servicii de date

- Servicii cu management complet pentru a sustine orchestrarea, transformarea și mutarea datelor
- Conectarea la baze de date relaționale și non-relaționale care operează on-prem sau în cloud
- Un singur tablou de monitorizare și management al canalelor de procesare a datelor(pipe-uri)





Azure Data Factory

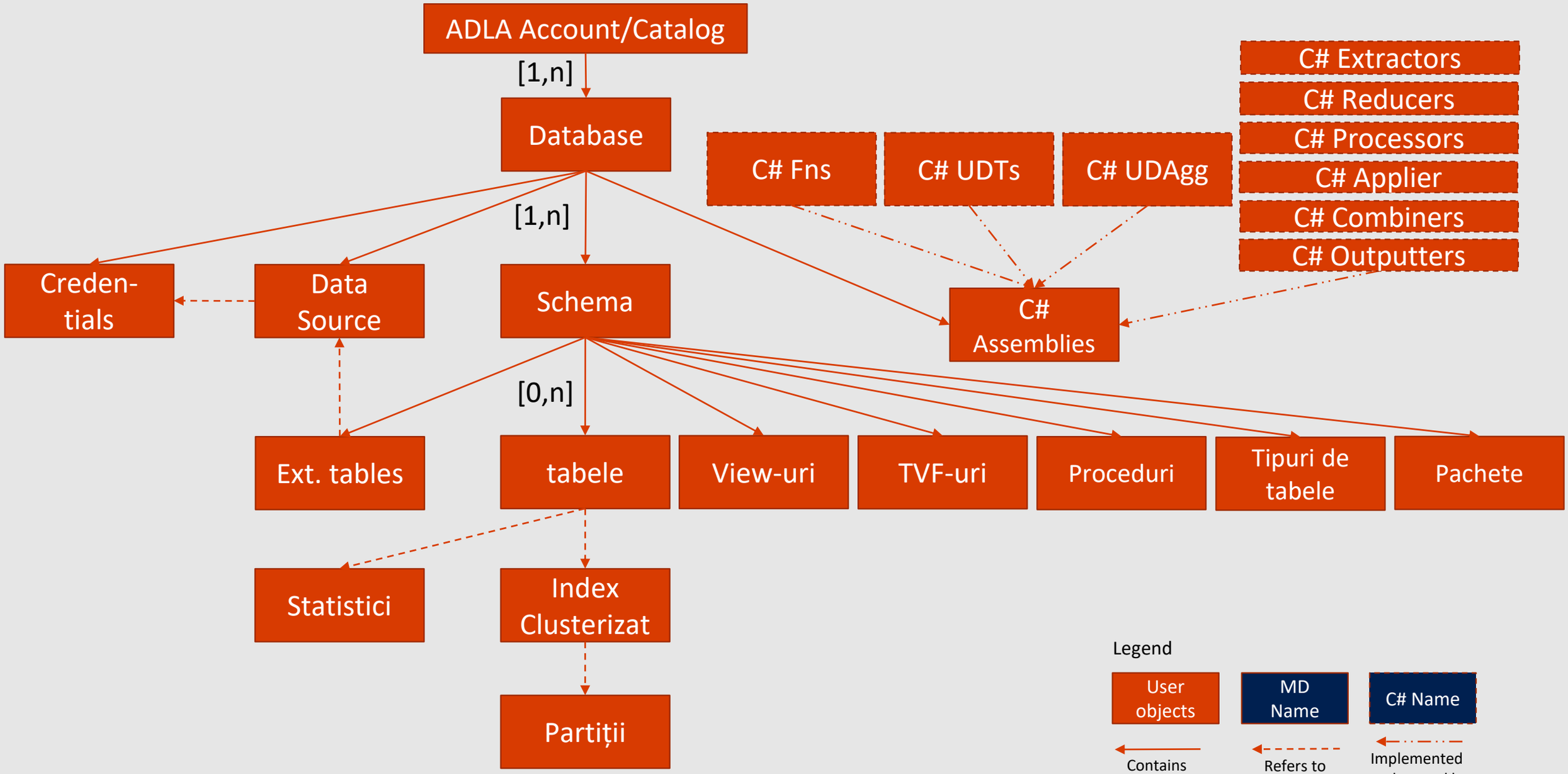
Conectează ADL Store cu toate depozitele de date

Categorii	Depozit de date	Acceptat ca sursa	Accepta primirea datelor
Azure	Azure Data Lake Store	✓	✓
	Azure Blob storage	✓	✓
	Azure SQL Database	✓	✓
	Azure SQL Data Warehouse	✓	✓
	Azure Table storage	✓	✓
	Azure DocumentDB	✓	✓
Baze de date	SQL Server*	✓	✓
	Oracle*	✓	✓
	MySQL*	✓	
	DB2*	✓	
	Teradata*	✓	
	PostgreSQL*	✓	
	Sybase*	✓	
	Cassandra*	✓	
	MongoDB*	✓	
	Amazon Redshift	✓	
Fisiere	File System*	✓	✓
	HDFS*	✓	
	Amazon S3	✓	
Altele	Salesforce	✓	
	Generic ODBC*	✓	
	Generic OData	✓	
	Web Table (table from HTML)	✓	
	GE Historian*	✓	

* Pot fi on-prem sau in Azure IaaS, permit utilizare de Data Management Gateway

Pregătirea Datelor și partajarea

Modelul obiectului Meta-data



Catalogul U-SQL

- Numire
- Identificare
- Partajare
- Securitate

Numire

- Baza implicită si contextul Schema: master.dbo
- Preluarea identificatorilor cu[]: `[tabela ck]`
- Incarca date in ADL Storage `/catalog` Director

Identificare

- Visual Studio Server Explorer
- Azure Data Lake Analytics Portal
- SDK-uri si comenzi Azure Powershell.

Partajare

- In contextul unui cont de Azure Data Lake Analytics
- Contul ADLA partajează același Azure Active Directory:

Securing

- Securitate cu AAD

VIEW-uri și TVF-uri

- View-uri pentru cazuri simple
- TVF-uri pentru parametrizare

View-uri

```
CREATE VIEW V AS EXTRACT…
CREATE VIEW V AS SELECT …
```

- Nu pot conține user-defined objects (e.g. UDF or UDOs)!
- Va fi "inline"

Table-Valued Functions (TVF-uri)

```
CREATE FUNCTION F (@arg string = "default")
RETURNS @res [TABLE ( … )]
AS BEGIN … @res = … END;
```

- Furnizeaza parametrizare
- Unu sau mai multe rezultate
- Poate contine mai multe instrucțiuni
- Poate contine cod utilizator(necesiat referinte assembly)
- Întotdeauna va fi inline
- Infera schema sau verifica conform schemei specificate

Proceduri

Permite incapsularea scripturilor U-SQL

```
CREATE PROCEDURE P (@arg string = "default“) AS
BEGIN
 …;
 OUTPUT @res TO …;
 INSERT INTO T …;
END;
```

- Furnizează parametrizare
- Scrie intr-un fisier sau tabela
- Poate contine multiple inregistrari
- Poate conține cod utilizator(este nevoie de referinte assembly)
- Va fi ”inline”
- Poate contine DDL (dar nu CREATE, DROP FUNCTION/PROCEDURE)

Tabele

- CREATE TABLE
- CREATE TABLE AS SELECT

```
CREATE TABLE T (col1 int
       , col2 string
       , col3 SQL.MAP<string,string>
       , INDEX idx CLUSTERED (col2 ASC)
         PARTITION BY (col1)
         DISTRIBUTED BY HASH (driver_id)
 );
```

- Date structurate, Tipuri de date integrate (nu UDTuri)
- Inex clusterizat
- Distribuție granulara(necesar a fi specificata):
  - HASH, DIRECT HASH, RANGE, ROUND ROBIN

```
CREATE TABLE T (INDEX idx CLUSTERED …) AS SELECT …;
CREATE TABLE T (INDEX idx CLUSTERED …) AS EXTRACT…;
CREATE TABLE T (INDEX idx CLUSTERED …) AS myTVF(DEFAULT);
```

- Infera schema din interogare
- Încă necesită indexare și distribuție (nu acceptă partiționare)

Folosirea tabelelor

Beneficiile clustering-ului si distribuției tabelelor

- Căutare mai rapidă a datelor furnizate prin distribuție și clustering în cazul în care se selectează distribuția / grupul drept
- Distribuția datelor oferă o scară mai bună localizată
- Folosit pentru filtre, îmbinare și grupare

• Beneficiile partitionarii tabelelor

- Permite managementul ciclului de viată(“expira” partitiile vechi)
- Re-calculare partiala a datelor la nivel de partitie.
- Predicatele de interogare pot furniza eliminarea partitiilor

Nu se utilizeaza…

- Fara filtre, join-uri si grupări
- Nu reutilizati datele pentru interogari ulterioare.

Daca sunteti in dubiu : utilizati sampling (e.g., **`SAMPLE ANY(x)`**) si testati.

Bibliografie

- Blog-uri si pagini ale unor comunități
  - http://usql.io (U-SQL Github)
  - http://blogs.msdn.microsoft.com/azuredatalake/
  - http://blogs.msdn.microsoft.com/mrys/
  - https://channel9.msdn.com/Search?term=U-SQL#ch9Search
- Articole și prezentări:
  - http://aka.ms/usql_reference
  - https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-programmability-guide
  - https://docs.microsoft.com/en-us/azure/data-lake-analytics/
  - https://msdn.microsoft.com/en-us/magazine/mt614251
  - https://msdn.microsoft.com/magazine/mt790200
  - http://www.slideshare.com/MichaelRys
  - Getting Started with R in U-SQL
  - https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-python-extensions
- ADL forums and feedback
  - https://social.msdn.microsoft.com/Forums/azure/en-US/home?forum=AzureDataLake
  - http://stackoverflow.com/questions/tagged/u-sql
  - http://aka.ms/adlfeedback