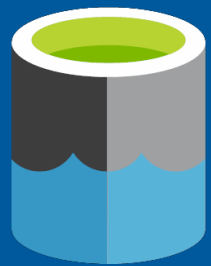


Data Lake și U-SQL



Conf.dr. Cristian KEVORCHIAN
Universitatea din București



Data Lake – Generalități

Azure Data Lake – Definiție

Un singur punct de stocare a datelor , în care plecând de la date brute (care provin printr-o preluare dintr-o sursă dată) care prin extragere, analiză și organizare a unor volume mari de date rezultă analize profunde într-o formă acceptabilă, utilă și benefică pentru o anumită utilitate de natură socială, economică, științifică etc.

Livrate de o manieră open source

Azure Data Lake

analytics service

U-SQL

Analytics



Spark

clusters (HDInsight)



YARN

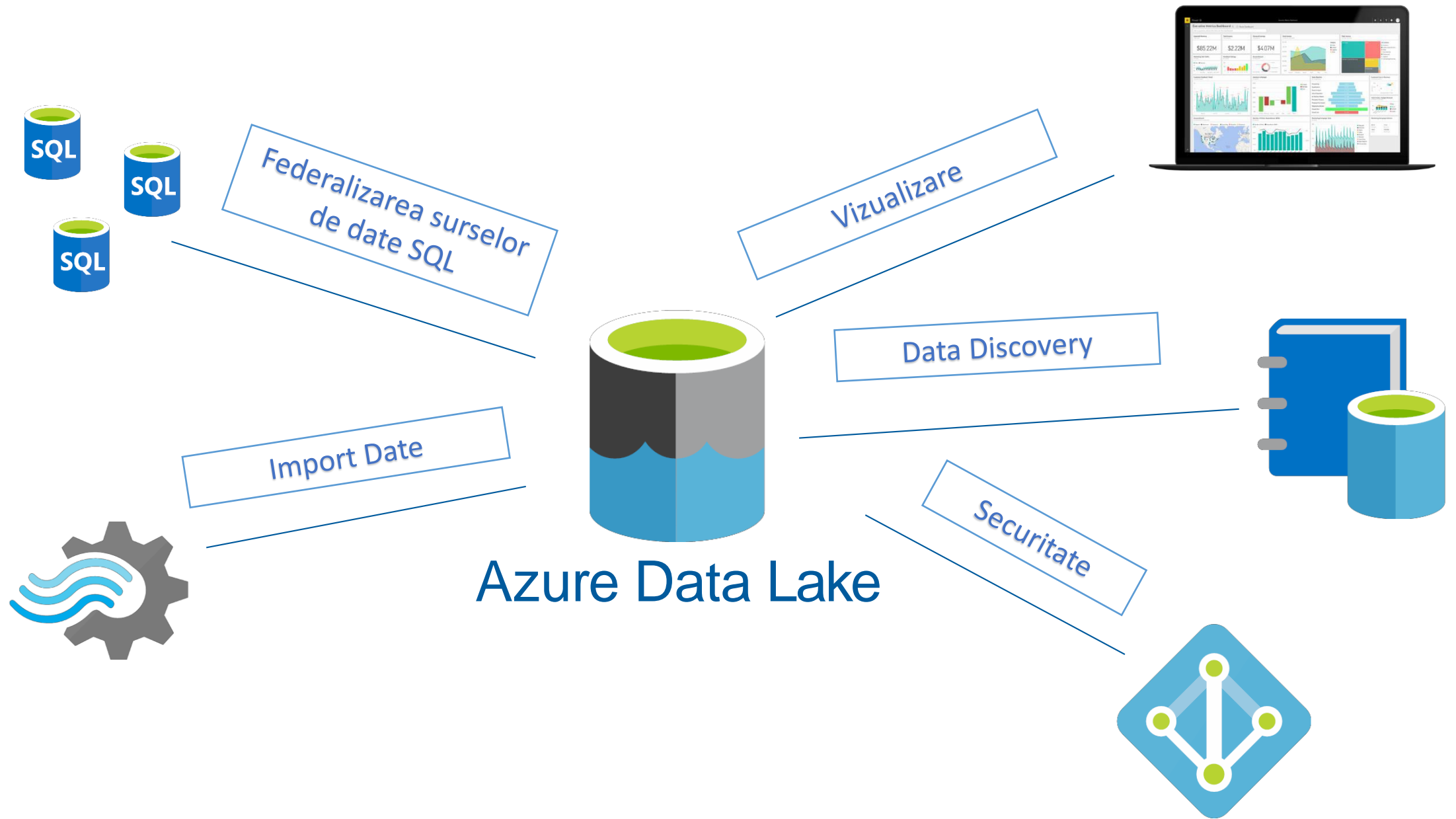
WebHDFS

Store

unstructured

semi-structured

structured

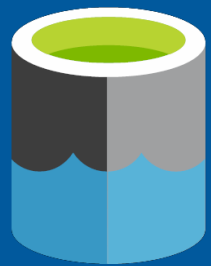


Funcționalități oferite de Data Lake

- Analitice (Data Lake Analytics)
- Hadoop Cluster as a Service(HDInsight)
- Stocare(Data Lake Store)
- Dezvoltare, debugging și optimizare de aplicații Big Data
- Stocare și analiză colecții de date de dimensiuni mari (petabyte) și trilioane de entități
- Securitate, audit și asistență la nivel de "enterprise"

Data Lake vs Data Warehouse

DATA WAREHOUSE	vs.	DATA LAKE
Procesare eminamente structurată	DATA	Structurat, SemiStructurat, NeStructurat, Brut
Schema-on-Write	PROCESARE	Schema-on-Read
Costuri mari pentru volume mari de date	STORAGE	Proiectate pentru democratizarea costurilor de stocare
Agilitate scăzută Configurare fixă	AGILITATE	Agilitate ridicată Configurarea și reconfigurarea un "must"
Matură	SECURITATE	Matură
Specialisti BI	UTILIZATORI	Specialisti în Data Scientist

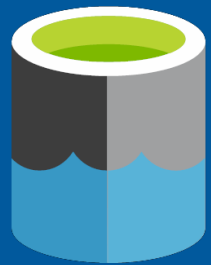


Data Lake Store

Data Lake Store – Definitie

- Un repository fault-tolerant și distribuit global de nivel înalt destinat furnizării de analitice big data de mare complexitate.
 - O structură Data Lake permite captarea și stocarea de date fără o limită de volum(deocamdată), tip, și viteză de ingestie într-un singur punct în scopul furnizării de analitice operaționale și exploratorii.
- Poate fi accesat din Hadoop (disponibil prin cluster-ul HDInsight) utilizând WebHDFS-compatibil REST.
- Conceput special pentru a permite dezvoltarea de analitice peste datele stocate fiind optimizat pentru implementarea de scenarii de analiză complex a datelor.
- Include "out of the box", toate standardele de nivel enterprise
 - securitate, management intuitiv, scalabilitate, încredere, și disponibilitate

DATA LAKE STORE	vs.	BLOB STORAGE
Stocare Optimizată pentru sarcini de dezvoltare de analitice big data	SCOP	Stocare de nivel general pentru o varietate largă de scenarii de stocare
Batch, interactiv, analitice în timp-real și date destinate ML cum ar fi fișiere log, date în IoT, stream-uri de click-uri, seturi de date de volume mari	UTILIZARE	Orice tip de date text sau binare livrate de back-end, backup-uri, stocări media pentru streaming și date de interes general
Un cont de Data Lake Store este organizat pe foldere care conțin fișiere de date stocate.	CONCEPTE	Un cont de Storage conține containere, care conțin date ca blob-uri
Sistem de fișiere ierarhic constituite	STRUCTURĂ	Obiecte stocate și identificate prin namespace
Bazat pe managementul identității furnizat de Azure Active Directory	SECURITATE	Bazat pe - Account Access Keys și Shared Access Signature Keys.



Analitice Data Lake

Analitice Data Lake – Definiție

- Conceput special pentru a permite analiza datelor stocate fiind optimizat pentru dezvoltarea de scenarii complexe de analiză a datelor.
- Realizează cu prioritate scrierea, rularea și gestionarea job-urilor, lăsând pe planul doi operarea infrastructurii distribuite.
- Poate gestiona execuția de job-uri la orice scară, dimensionând resursele necesare.
- Se plătește pentru acel job care se execută., fiind cost-effective.
- Serviciul de analiză acceptă Azure Active Directory, fapt ce permite gestionarea accesului și rolurilor, integrate cu sistemul de identitate local.

Analytics Data Lake

Funcționalități cheie

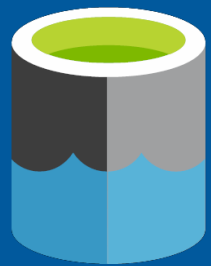
- Scalare dinamică
- Dezvoltare rapidă, debugging și optimizare prin utilizarea unor familii de instrumente familiar.
- U-SQL: simplu și extensibil
- Integrare largă de componente IT ce operează cu date din Azure.



HDInsight

Azure HDInsight

- O solutie cloud full-maged pentru Apache Hadoop
- Furnizează analitice open-source optimizate la nivelul clusterului pentru
 - Spark,
 - Hive,
 - MapReduce,
 - HBase,
 - Storm,
 - Kafka,
 - Microsoft R Server
- Un SLA 99.9%
- Instalarea acestei tehnologii Big data și a aplicațiilor ISV(independent Software Vendor) în contextul "managed cluster" cu Securitate livrată la nivel enterprise.



U-SQL

U-SQL

Este un limbaj de interogare destinate soluțiilor Big Data(serviciile Data Lake Analytics)

A evoluat ca proiect intern al Microsoft

SCOPE : Easy and Efficient Parallel Processing of Massive Data Sets

*de Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, Bill Ramsey,
Darren Shakib, Simon Weaver, Jingren Zhou*

<http://www.vldb.org/pvldb/vol1/1454166.pdf>

U-SQL combină

- limbaj declarativ
- Tipuri si expresii C#
- Asemanari cu Hive si LINQ
- Concepte de procesare Big Data cum ar fi “schema on reads”,

Interogheaza
si cmbină date
dintr-o
varietate de
surse

- Azure Data Lake Storage,
- Azure Blob Storage,
- Azure SQL DB, Azure SQL Data Warehouse,
- Instante SQL Server executate pe masini virtuale.

Nu este ANSI SQL

- SELECT trebuie scris cu majuscule
- Expresiile din cadrul clauzei SELECT predicatele WHERE etc sunt C#.
- Asta înseamnă că operatorii de comparare sunt sintactic cei din C# (`a == "foo"`),
- Utilizează semantica null din C# care este una cu 2-valori și nu una cu 3-valori ca în ANSI SQL.

- Azure Data Lake Analytics provides U-SQL for batch processing.
- U-SQL is written and executed in form of a batch script.
- U-SQL also supports data definition statements such as `CREATE TABLE` to create metadata artifacts either in separate scripts or sometimes even in combination with the transformation scripts.
- U-SQL Scripts can be submitted in a variety of ways.
 - Directly from within the Azure Data Lake Tools for Visual Studio,
 - From the Azure Portal
 - Programmatically via the Azure Data Lake SDK job submission API
 - Azure Powershell extension's job submission command

Script-uri U-SQL și procesarea datelor

Se aplică următorul pattern de procesare:

- **Preluarea datelor din locațiile stocate drept familii de linii**
 - Locațiile de stocare pot fi fișiere care vor fi "schematizate" "on read" cu EXTRACT
 - Locațiile de stocare pot fi tabele U-SQL care sunt încărcate într-un format "schematizat"
 - Locațiile de stocare pot fi tabele furnizate din alte surse cum ar fi baze de date Azure SQL
- **Transformarea familiilor de linii**
 - Mai multe transformări peste seturile de linii pot fi tratate într-un format de flux de date
- **Stocarea datelor care au fost supuse transformării**
 - Incarcarea într-un fișier cu aluza OUTPUT, sau
 - Incarcarea într-o tabelă U-SQL cu ajutorul clauzei INSERT

Script U-SQL

```
DECLARE @in  string = "/Samples/Data/SearchLog.tsv";
DECLARE @out string = "/output/result.tsv";

@searchlog =  EXTRACT UserId int, Start DateTime, Region string, Query string,
               Duration int?, Urls string, ClickedUrls string
               FROM @in  USING Extractors.Tsv();

@rs1 = SELECT Start, Region, Duration FROM @searchlog WHERE Region == "en-gb";

@rs1 = SELECT Start, Region, Duration FROM @rs1
        WHERE Start >= DateTime.Parse("2012/02/16");

OUTPUT @rs1
TO @out
USING Outputters.Tsv();
```

DEMO

- Creare Data Lake Store
- Creare cont Data Lake Analytics și conectare cu Data Lake Stores
- Import date în Data Lake Stores
- Execuție job-uri U-SQL în Azure Data Lake Analytics

