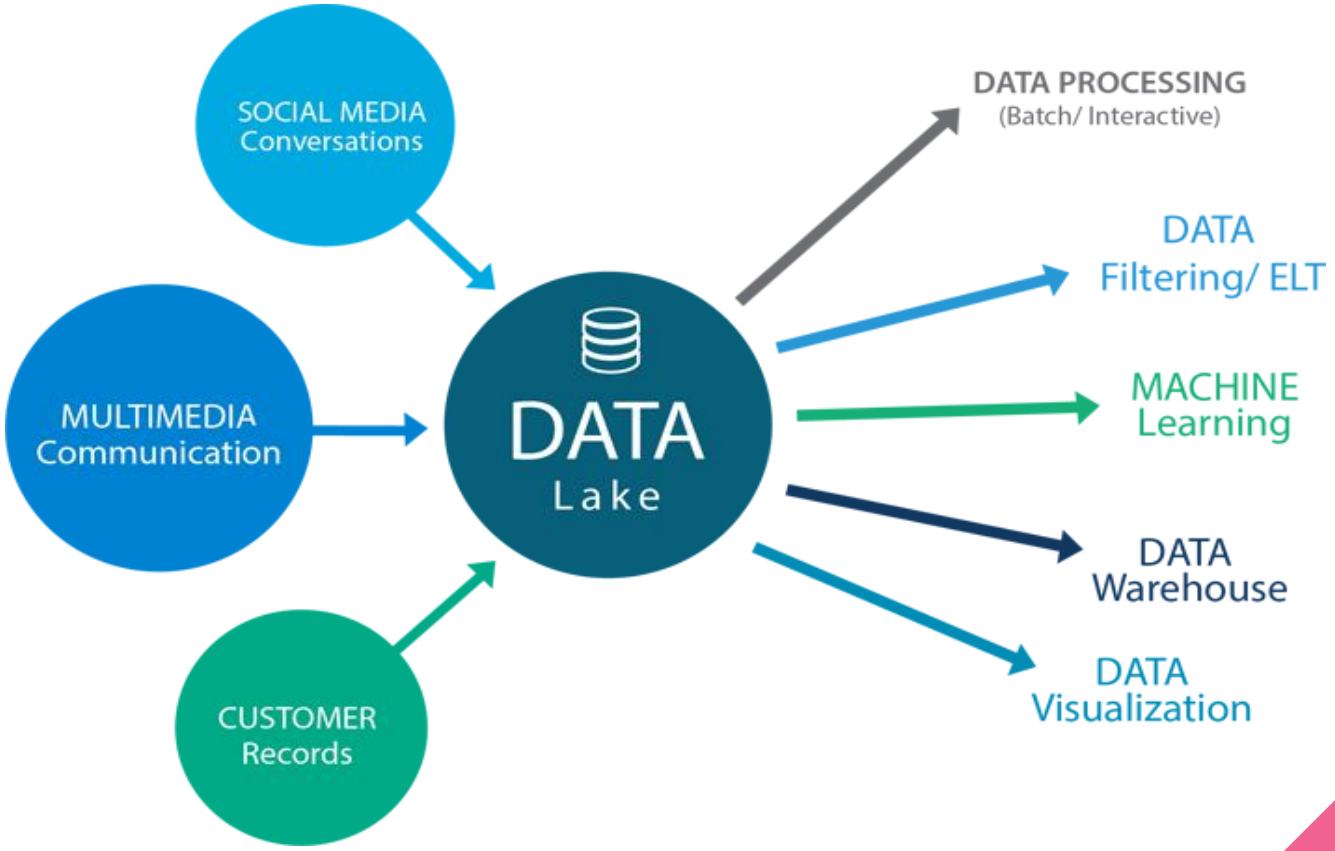
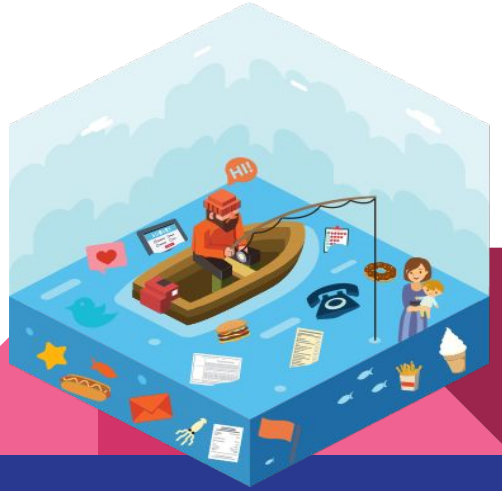
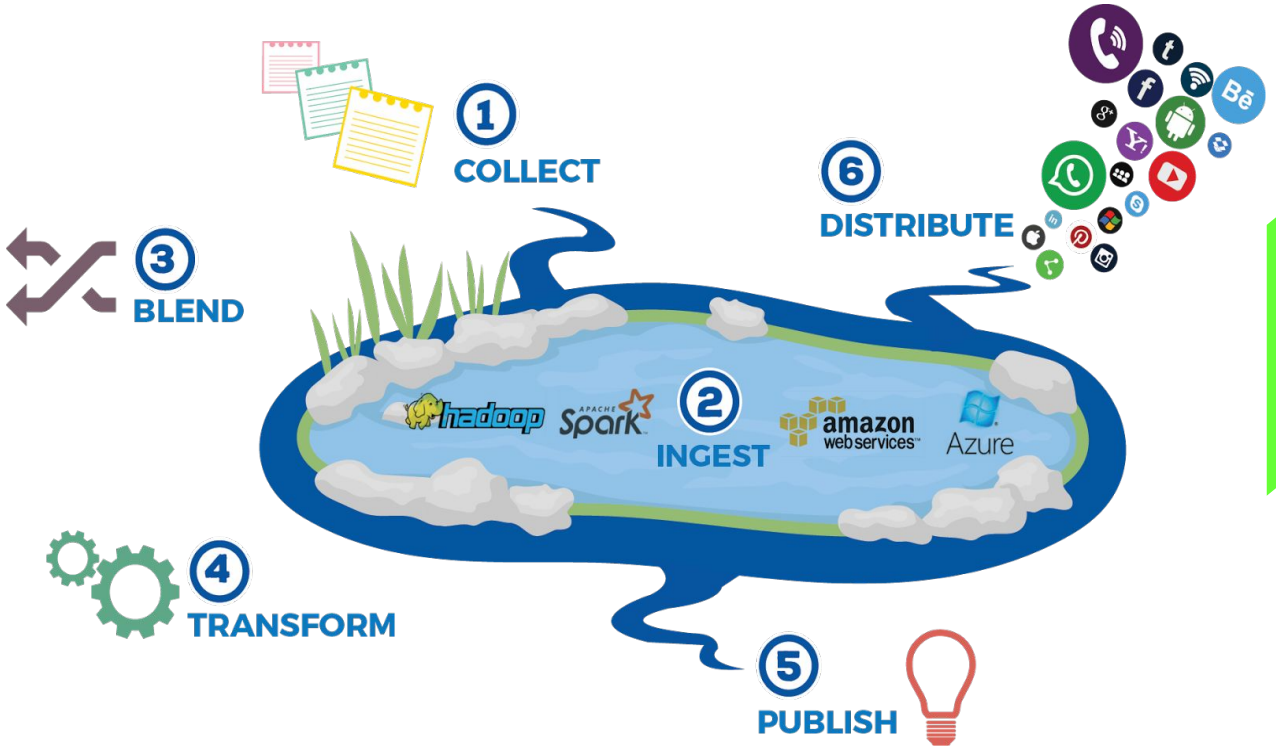
Microsoft Azure Data Lakes

What is Data Lake?

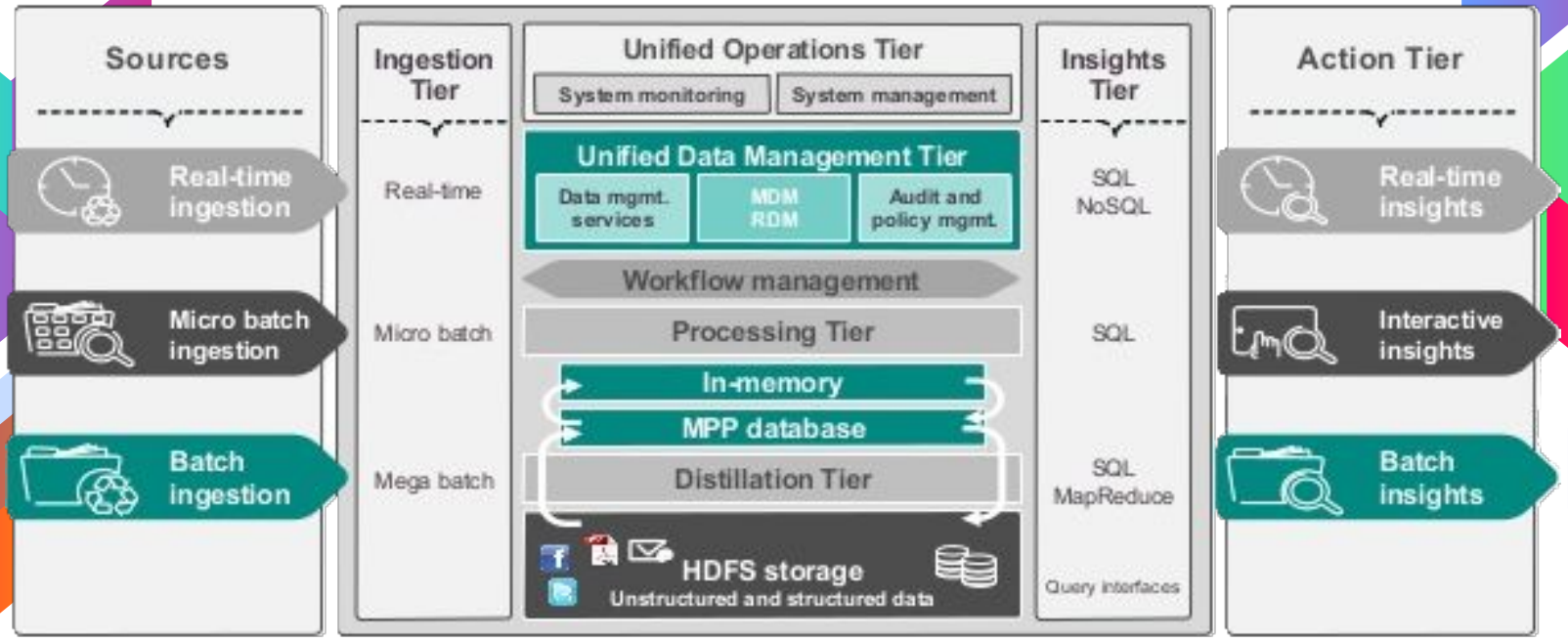




How Data Lake works?



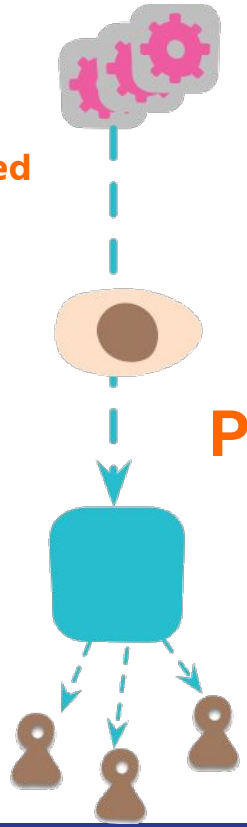
Data Lake Architecture



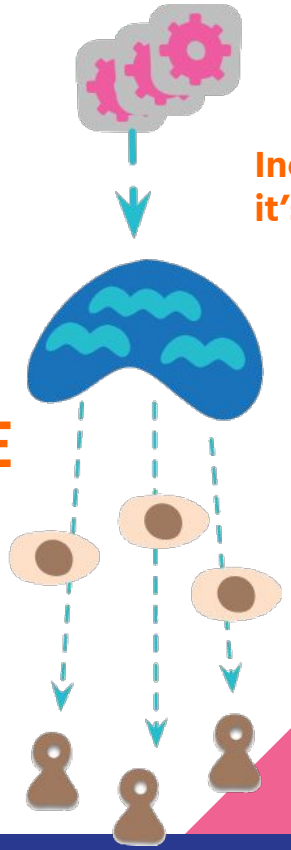
Data Warehouse vs. Data Lake

Incoming data is cleaned and organized into a single consistent schema

Analysis is done directly on the curated warehouse data



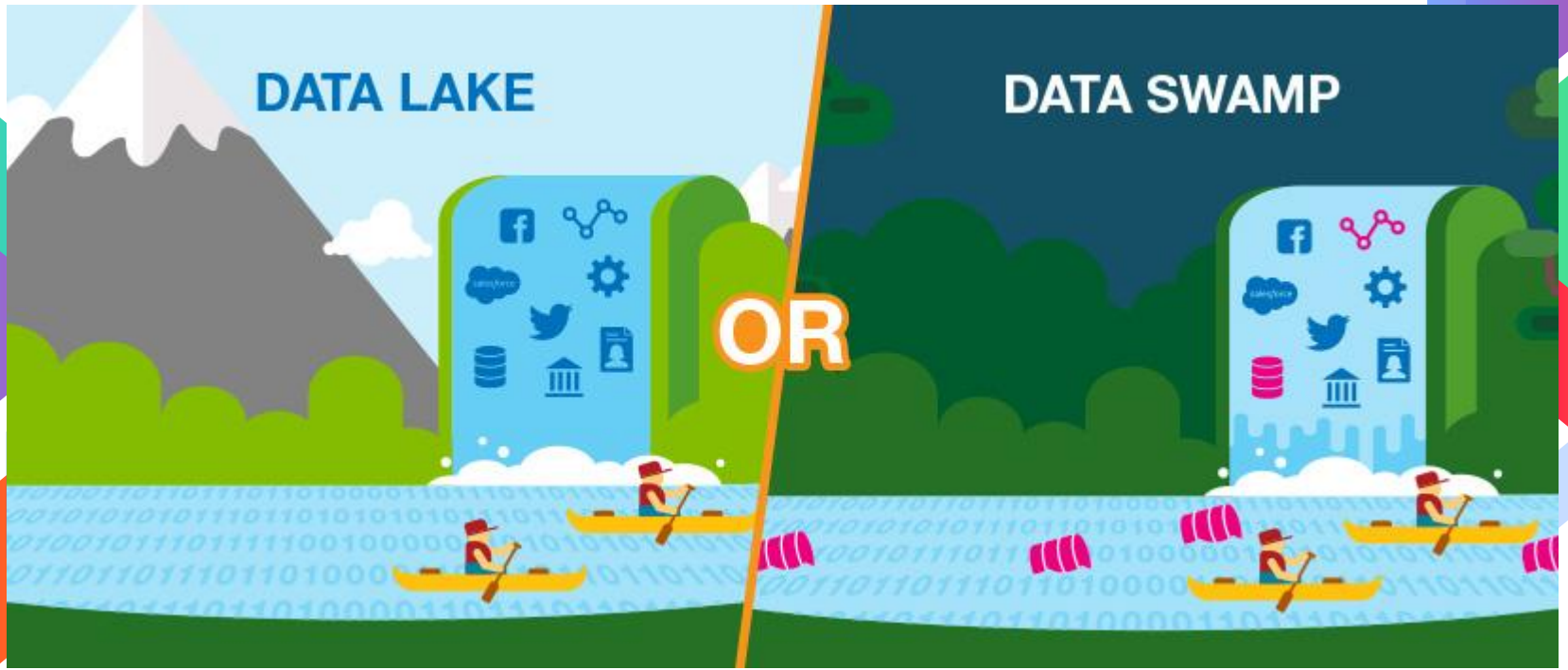
STORAGE
PROCESSING
AGILITY
PERFORMANCE
SECURITY
USERS



Incoming data comes in it's raw format

Data is selected and organized for each need

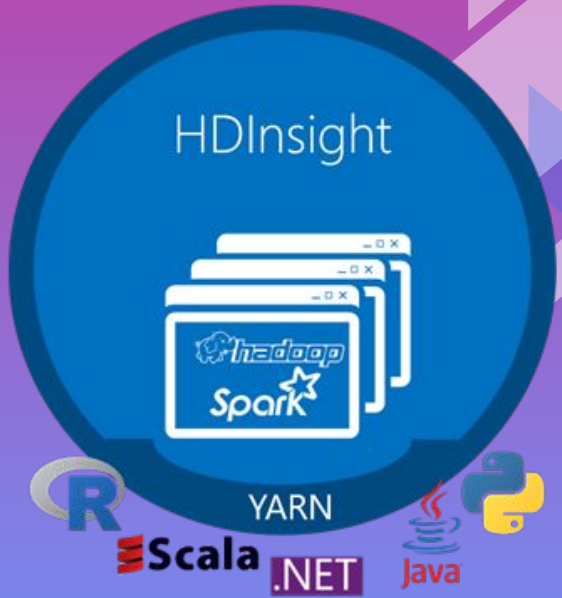
Don’t let Data Lake become Data Swamp !!



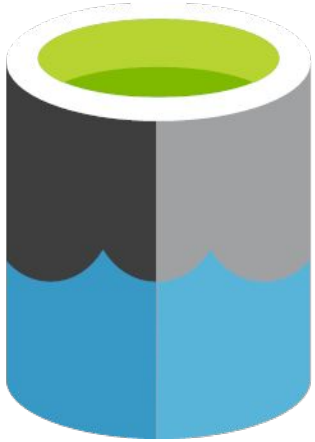
Microsoft Azure Data Lake

No limits Data Lake

Analytics job service



Managed Clusters



Azure Data Lake Store

What is Data Lake Store ?

- ➔ *Cloud Storage*
- ➔ *Stores ANY data in native format*
- ➔ *No limit to SCALE*
- ➔ *HDFS for the cloud*
- ➔ *ENTERPRISE READY access control*
- ➔ *Encryption at rest*
- ➔ *Optimized for analytic workload*
- ➔ *PERFORMANCE*

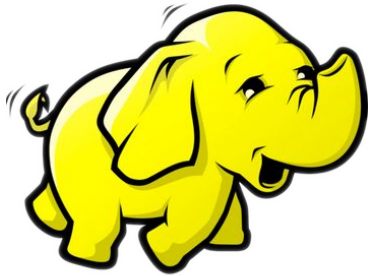
Spark + 

Why Azure Data Lake Store?

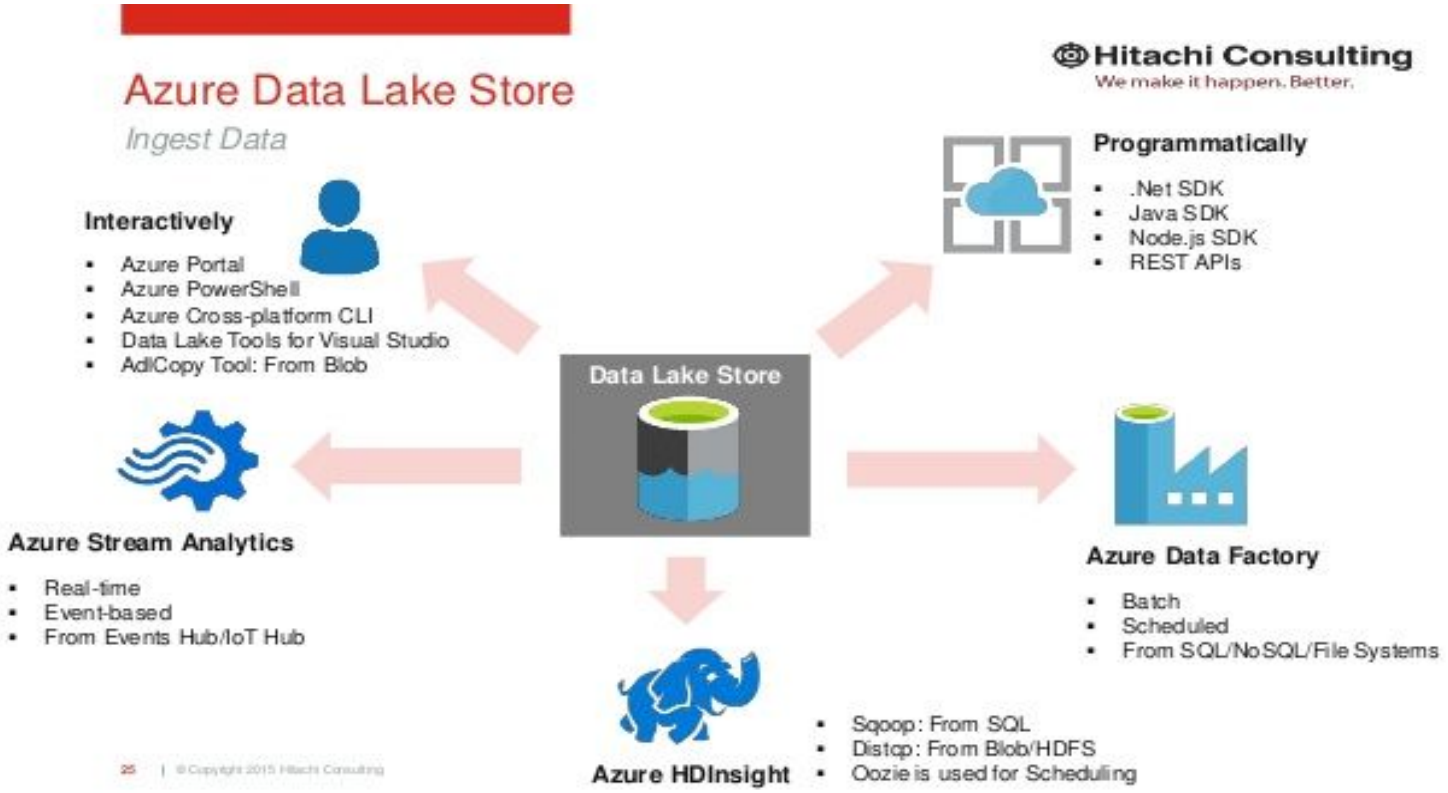
- ➔ Highly Scalable
- ➔ Reliable and Available
- ➔ Not expensive
- ➔ Handles more than petabytes of memory.
- ➔ Easily used with other azure services
- ➔ Security

Applications compatible with Azure Data Lake Store

Open Source Software	Distribution
Apache Sqoop	HDInsight 3.2, 3.4, 3.5, and 3.6
MapReduce	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Storm	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Hive	HDInsight 3.2, 3.4, 3.5, and 3.6
HCatalog	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Mahout	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Pig/Pig Latin	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Oozie	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Zookeeper	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Tez	HDInsight 3.2, 3.4, 3.5, and 3.6
Apache Spark	HDInsight 3.4, 3.5, and 3.6



Ingesting Data



Securing data in Azure Data Lake Store

Azure Data Lake Store uses ***Azure Active Directory*** for authentication and access control lists (ACLs) to manage access to the data

- ➔ **Authentication**
- ➔ **Access control**
- ➔ **Encryption**

Pricing

Data Lake Store		
	Capacity	Cost (In Euros)
Storage	100 TB	0.0329 /GB
Write Transactions	10000	0.043
Read Transactions	10000	0.034

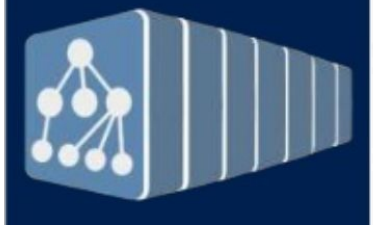
Azure Data Lake Analytics

All data

Productivity

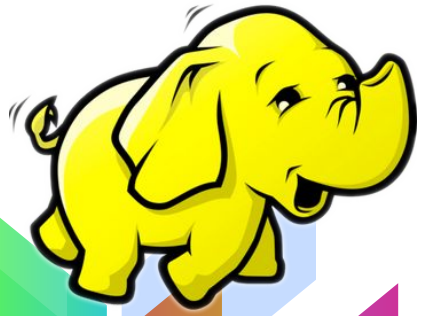
Easy and powerful data preparation

Limitless scale



Enterprise-grade

How Data Lake Analytics works?



- Built on YARN
- Scales Dynamically with the turn of a dial
- Built with U-SQL
- Processes data across Azure
- Payment by query

U-SQL

Query language of the Azure Data Lake Analytics service

Combines:
SQL-like declarative language
C# types and the C# expression language
Big data processing concepts such as "schema on reads", custom processors and reducers.

Provides the ability to query and combine data from a variety of data sources, including Azure Data Lake Storage, Azure Blob Storage, and Azure SQL DB, Azure SQL Data Warehouse, and SQL Server instances running in Azure VMs.

Submitting jobs to Data Lake Analytics

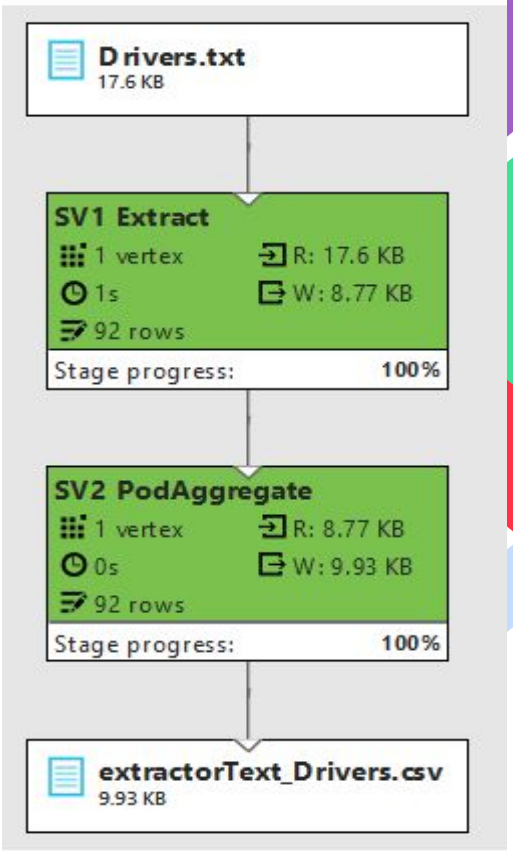
- Azure Data Lake Tools *in* ***Visual Studio*** to submit jobs directly
- ***Azure Portal*** to submit jobs *via the* ***Data Lake Analytics account***
- Data Lake ***SDK job submission API*** to submit jobs programmatically (SDK .NET, SDK for Python, SDK for Java)
- ***Azure PowerShell*** extensions to submit jobs programmatically

Example

```
@Drivers =
    EXTRACT driver_id   int,
        name            string,
        street          string,
        city            string,
        region          string,
        zipcode         string,
        country         string,
        phone_numbers   string
    FROM "/sample_data/Drivers.txt"
    USING Extractors.Text(delimiter: '\t', encoding:Encoding.Unicode);

@print =
    SELECT * FROM @Drivers;

OUTPUT @print
TO "/Output/ReferenceGuide/BuiltIn/UDOs/extractorText_Drivers.csv"
USING Outputters.Csv();
```



Pricing

Pay-As-You-Go	
Analytical Units	**Cost per month**
100	**1.687 / hour**

INCLUDED ANALYTICS UNIT HOURS	PRICE/MONTH	SAVINGS OVER PAY-AS-YOU-GO
100	€97	43%
500	€421	50%
1,000	€759	55%
5,000	€3,542	58%
10,000	€6,324	63%
50,000	€28,250	67%
100,000	€50,598	70%
> 100,000	Contact Us	

Overage on Analytics Unit Hour will be billed at €1.265/Hour.

THANK YOU

TEAM: 1208

Jayanthi Kambayatughar - jayanthi.kambayatughar@est.fib.upc.edu

Anastasiia Zavolozhina -anastasiia.zavolozhina@est.fib.upc.edu

https://github.com/jayanthi456/CLOUD-COMPUTING-CLASS-2018