CBTU presents a course on **Big data and Hadoop**

Module 1: BIG DATA

Section 1.6: Data Lake

Data Lake

- A Data Lake is a method of storing data within a system or repository, in its natural format, that facilitates the collocation of data in various forms.
- Data lake is a single store of all data in the enterprise ranging from raw data to transformed data which is used for various tasks including reporting, visualization, analytics and machine learning.
- *James Dixon* coined the term “Data Lake” and promoted.

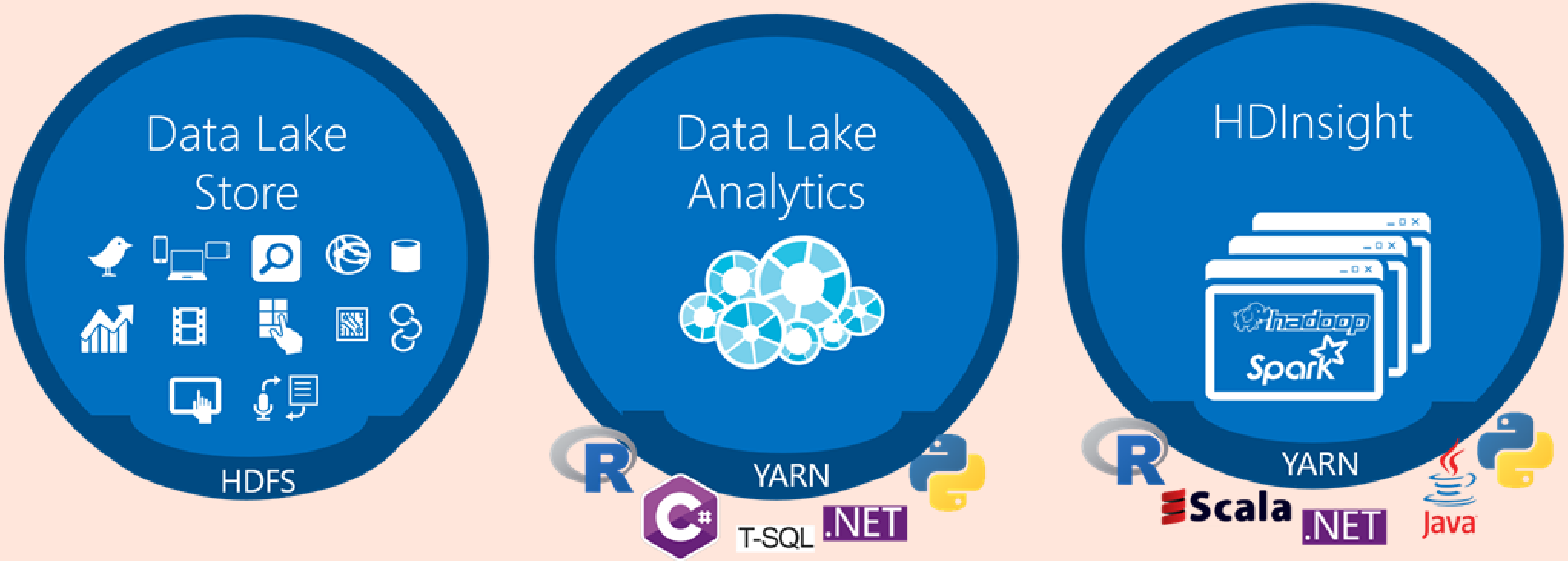
Data Lake

The data lake includes all forms of data:

- Structured data from relational databases
- Semi-structured data
- Unstructured data

A <u>data swamp</u> is a deteriorated data lake, that is inaccessible to its intended users and provides little value.

Azure Data Lake



Data Lake Store

HDFS

Data Lake Analytics

YARN

HDInsight

YARN

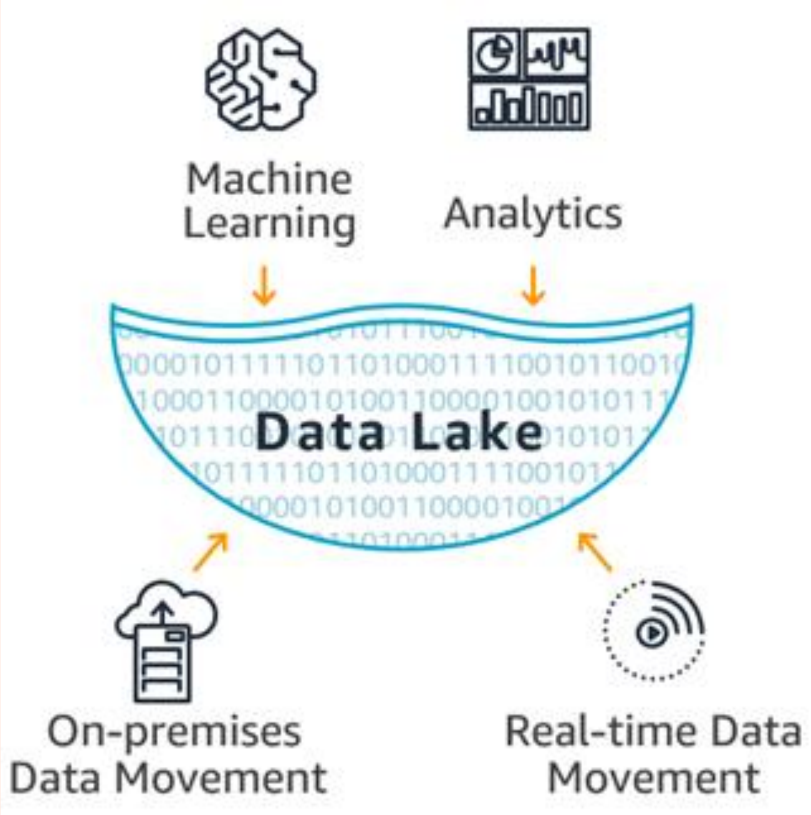
No limits Data Lake

Analytics job service

Managed Clusters

Source: Microsoft Azure

AWS Data lake



Players in Data lake

- – Hortonworks
- – Google
- – Microsoft
- – Zaloni
- – Teradata
- – Cloudera
- – Amazon

Data Lake in use

- The distributed file system used in Apache Hadoop is Data lake.
- Cloud storage services such as Azure Data Lake and Amazon S3.
- Personal Data Lake aims at managing big data of individual users by providing a single point of collecting, organizing, and sharing personal data.

Hadoop and the Data Lake

Hadoop 2.0 made it possible - new processing paradigms like streaming, interactive, on-line to be available via Hadoop and the Data Lake.

Flipside

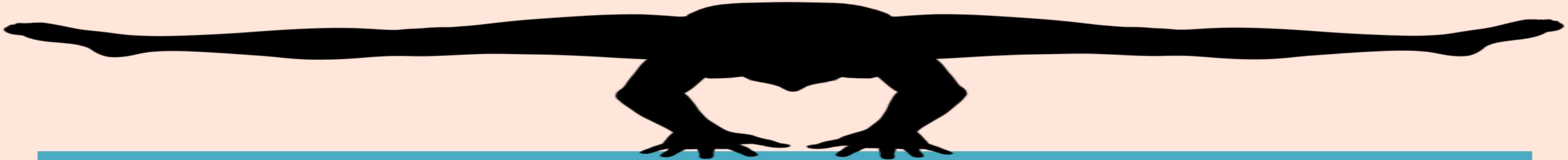
The main challenge is not creating a data lake, but taking advantage of the opportunities it presents.

Customers creating big data lakes, dumping everything into HDFS, lose track of what’s there.

Companies must build successful data lakes, gradually maturing their lake as they figure out which data and metadata are important to the organization.

Data lake is ETL hub?

- The data lake has been labeled as a raw data reservoir or a hub for ETL (extract, transform, and load) offload.
- The data lake has been defined as a central hub for self-service analytics.



Thanks for watching

CBTUniversity.com

✍ learnq@CBTUniversity.com

☎ +91 963 246 5599

/CBTUniversity