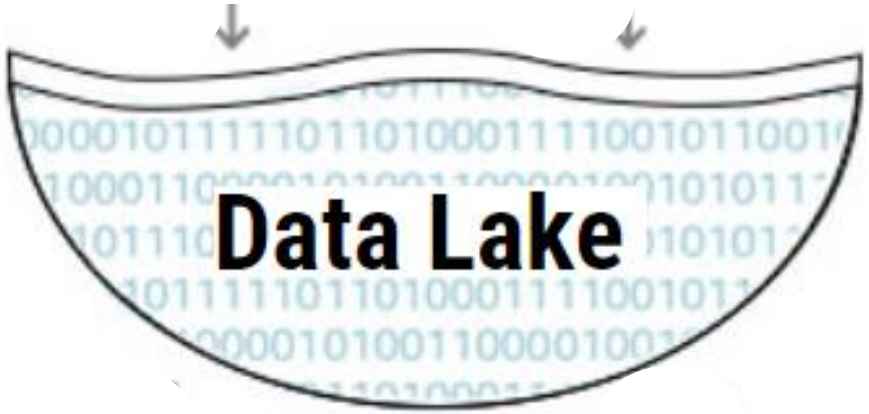
Azure Data Lake Introduction

Eshant Garg

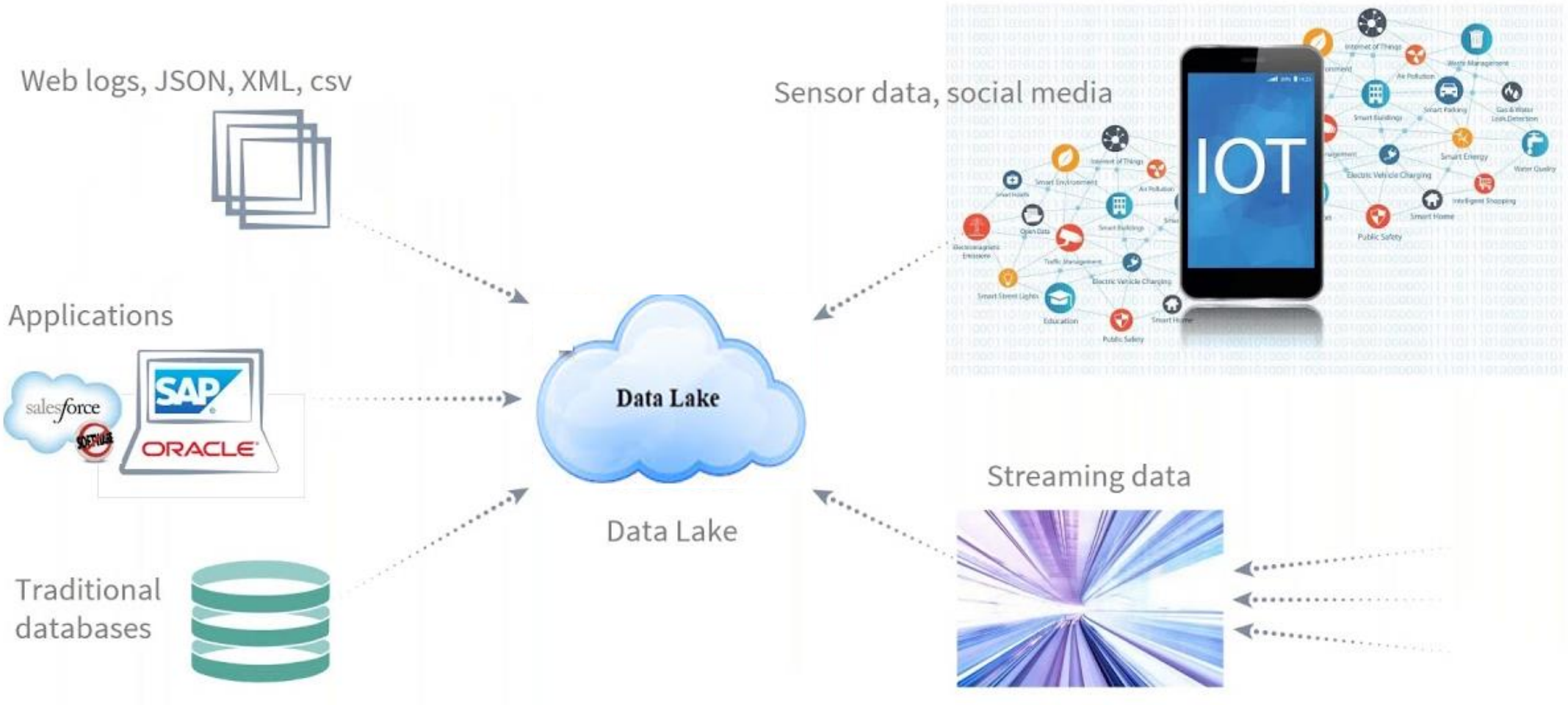
Azure Data Engineer, Architect, Advisor

eshant.garg@gmail.com



Data Lake is a big container to store data.

Data Lake Sources



What is Data Lake?

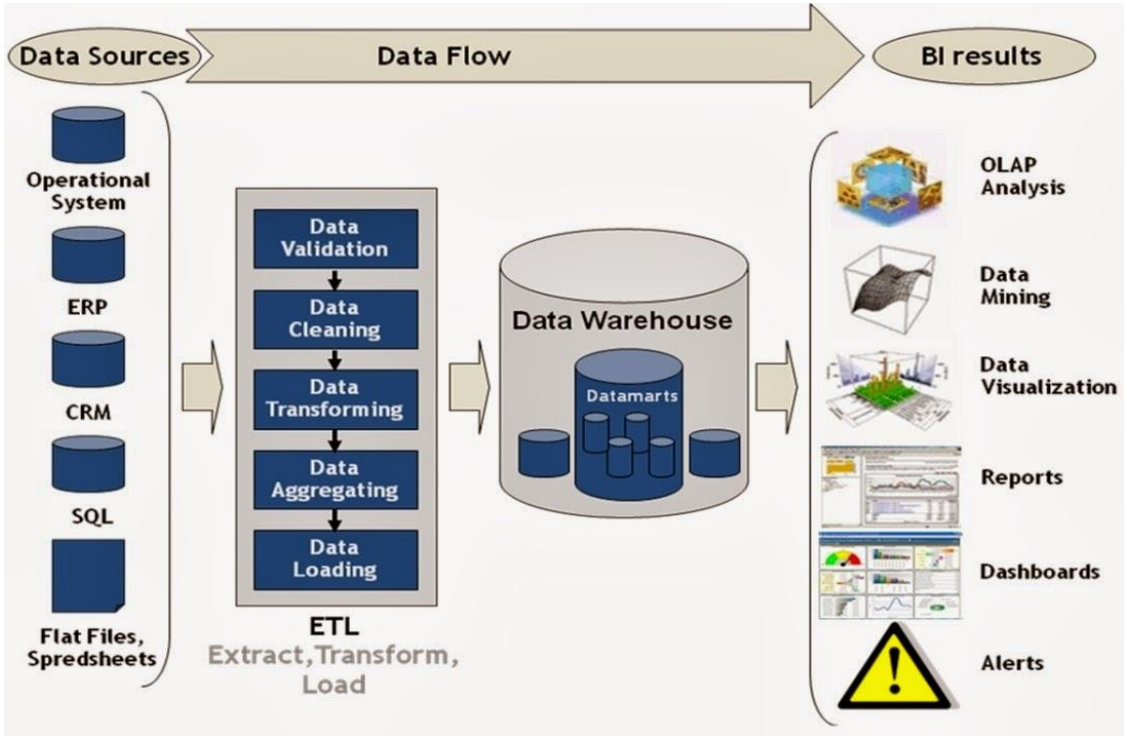
"If you think of a DataMart as a store of bottled water – clean and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."

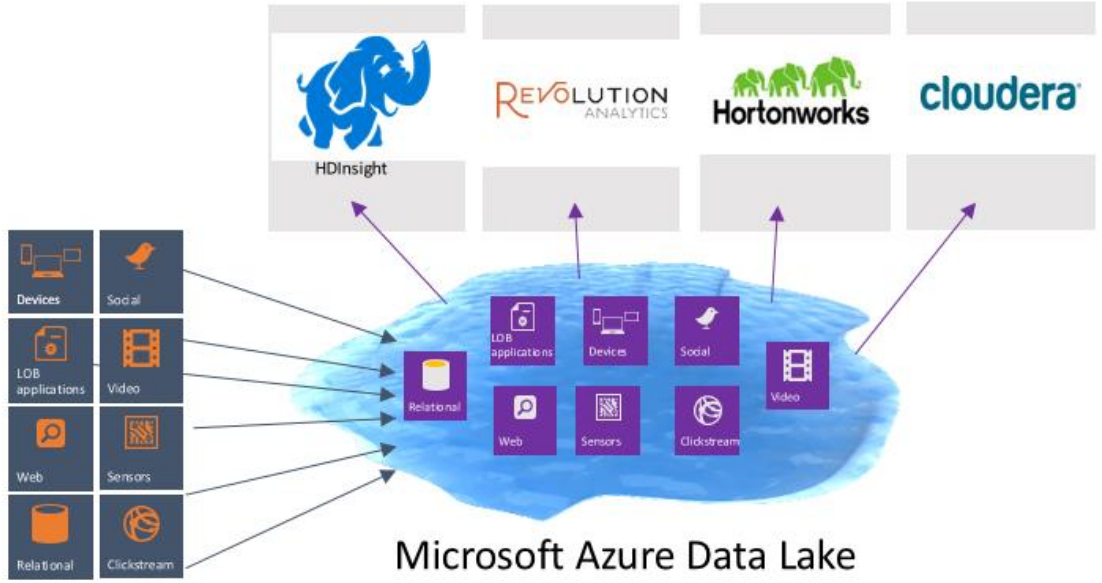
James Dixon
CTO, Pentaho

Data Warehouse

Data Lake

Data Warehouse vs Data Lake





Data Warehouse

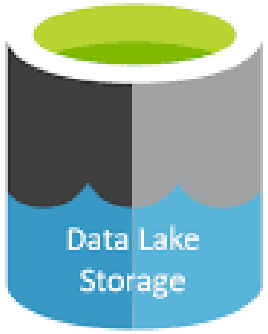
Data Lake

LearnCloud.Info

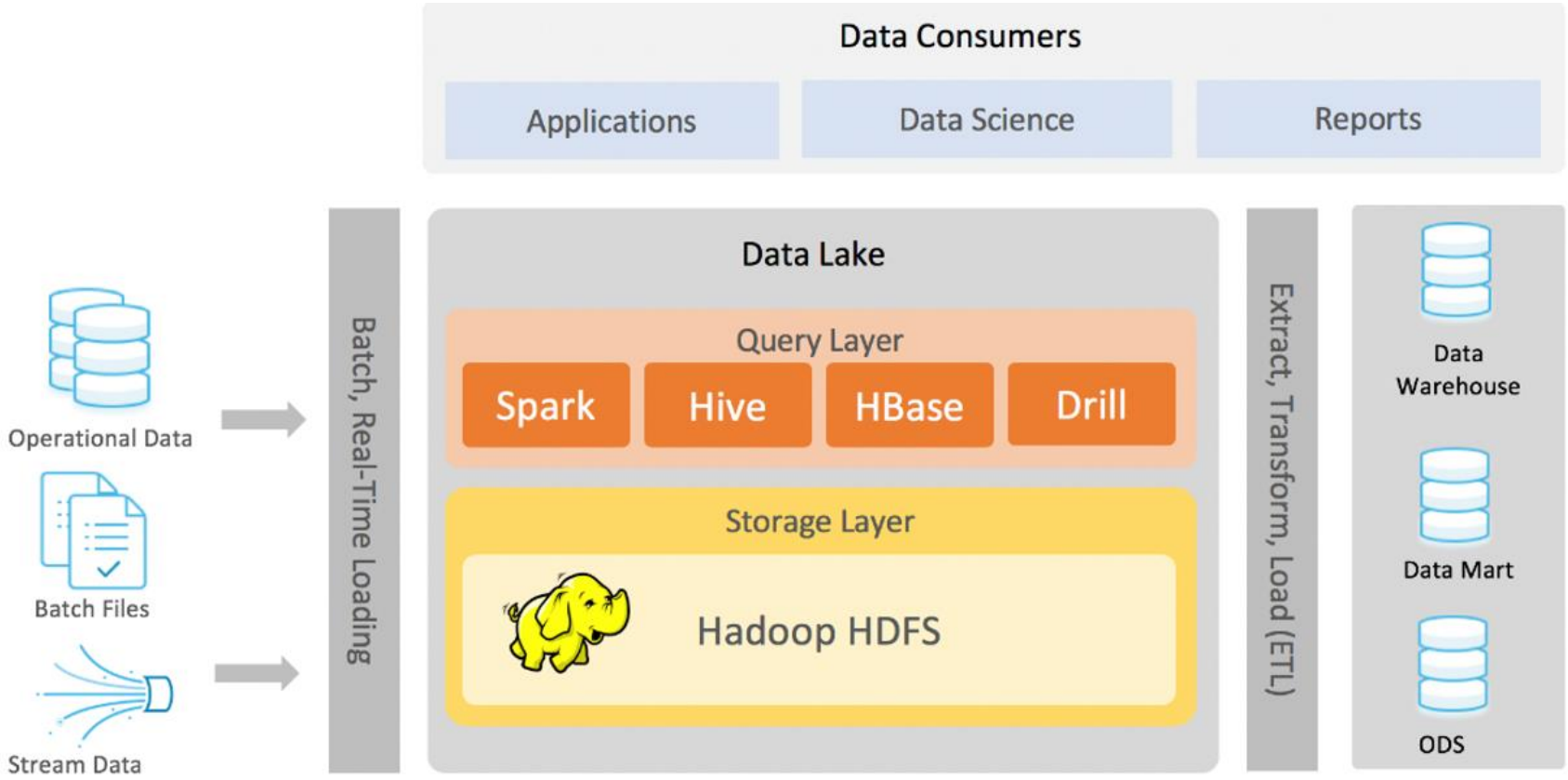

LearnCloud.Info

Data lake vs Hadoop

VS



Data lake vs Hadoop

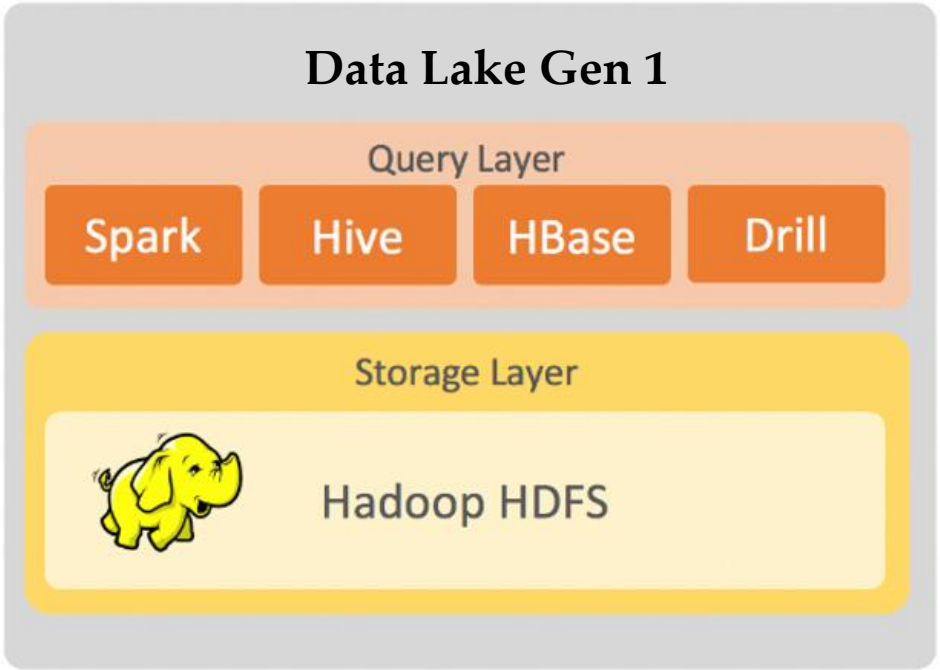


Data Lake can use:

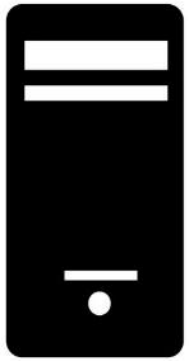

LearnCloud.Info

Azure Data Lake Gen1 evolution

- Fault tolerant file system
- Runs on commodity hardware
- MapReduce, Pig, Hive, Spark etc.
- HDFS in Cloud -> Data Lake Storage Gen1



Cloud storage challenge



Processing

- Easy to optimize processing by increasing vCPU and Ram

Storage

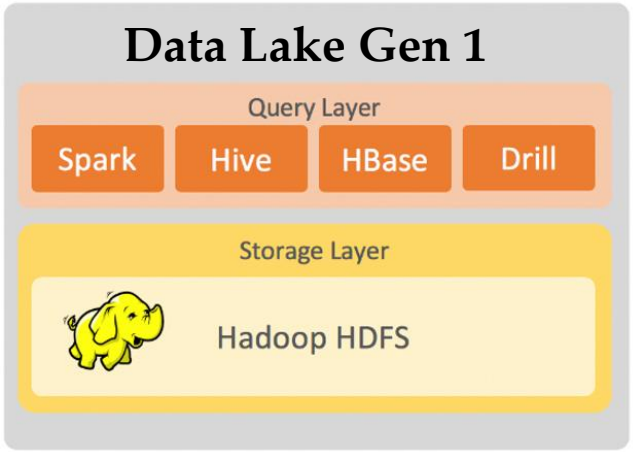
- Different requirements
- No direct solution

Azure Blob Storage

- Large object storage in cloud
- Optimized for storing massive amounts of unstructured data
  - Text or Binary Data
- General purpose object storage
- Cost efficient
- Provide multiple Tiers

Azure Data lake Gen 2

Blob Storage



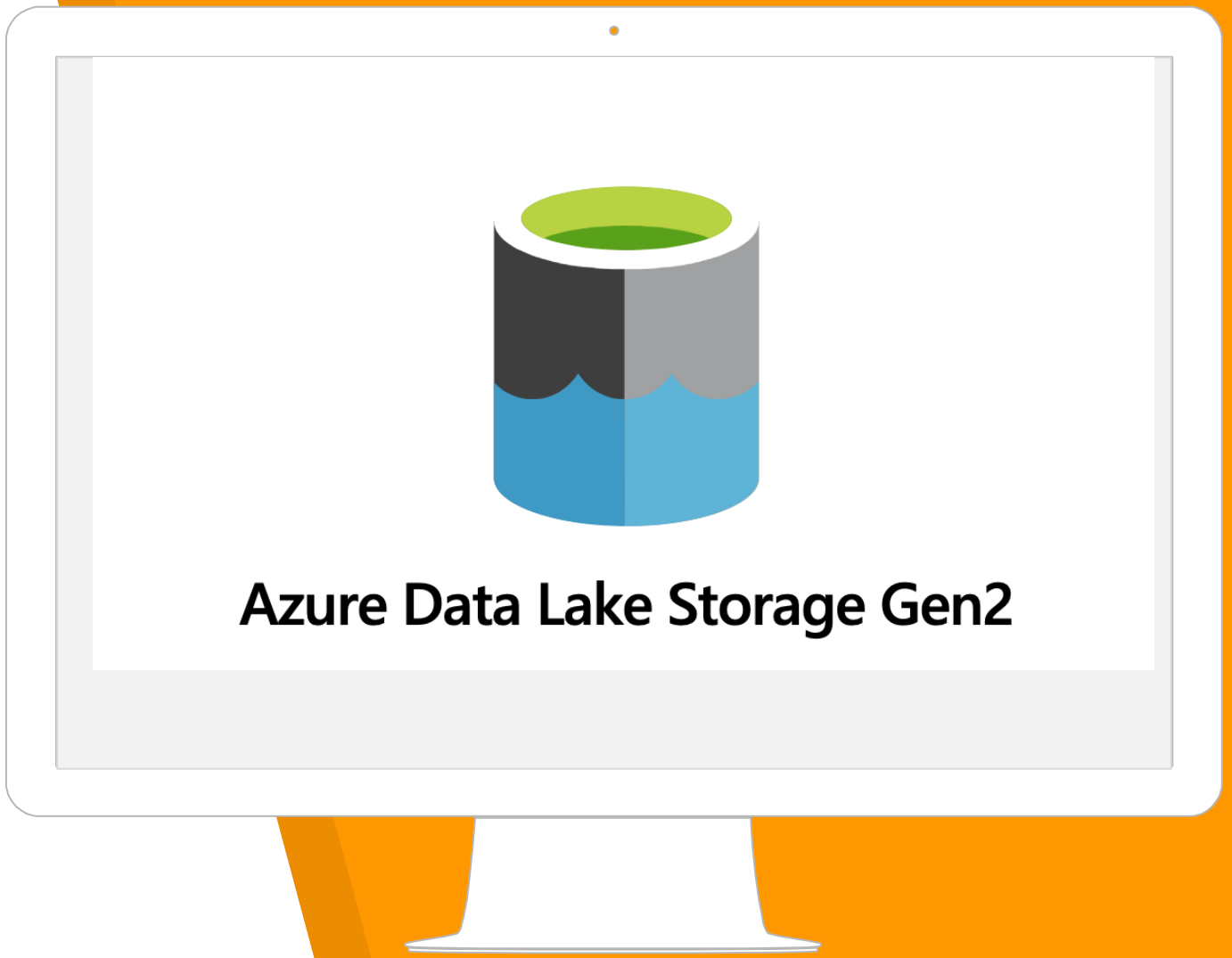
Azure Data Lake Storage Gen2

LearnCloud.Info

MICRSOFT RECOMMENDS

Data Lake Storage Gen2 for your big data storage needs.

Note: USQL currently not supported in Gen 2



LearnCloud.Info

Blob Storage vs Data Lake Storage

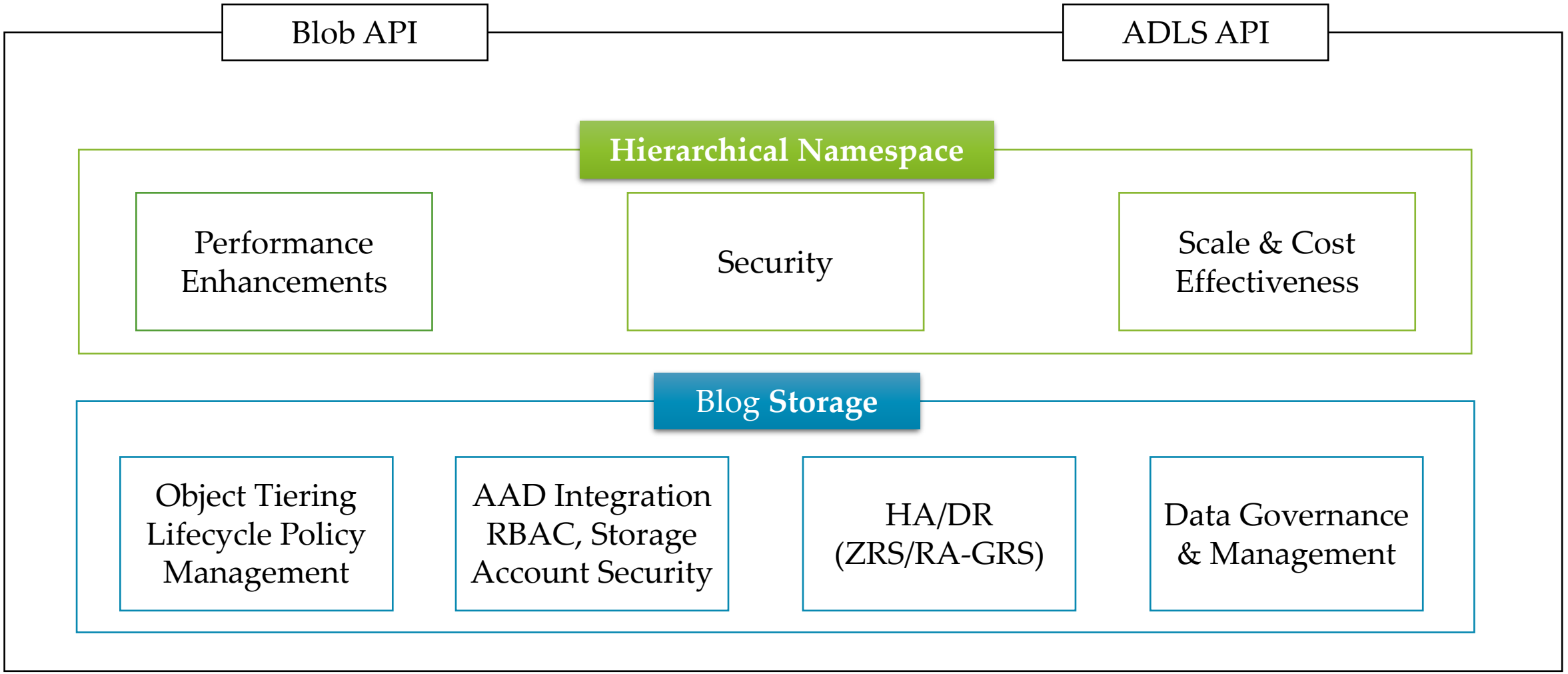
Azure Blob Storage

- General purpose data storage
- Container based object storage
- Available in every Azure region
- Local and global redundancy
- Processing performance limit

Azure Data Lake Storage (Gen 2)

- Optimized for big data analytics
- Hierarchical namespace on Blob Storage
- Available in every Azure region
- Local and global redundancy
- Supports a subset of Blob storage features
- Supports multiple Azure integrations
- Compatible with Hadoop

Data Lake Architecture



LearnCloud.Info