Azure Data Lake Storage

Introduction to ADLS

Why Data Lake?

Why Data Lake?

Why Data Warehouse is failing today?

Once upon a time

Total Capacity – 1.44 MB

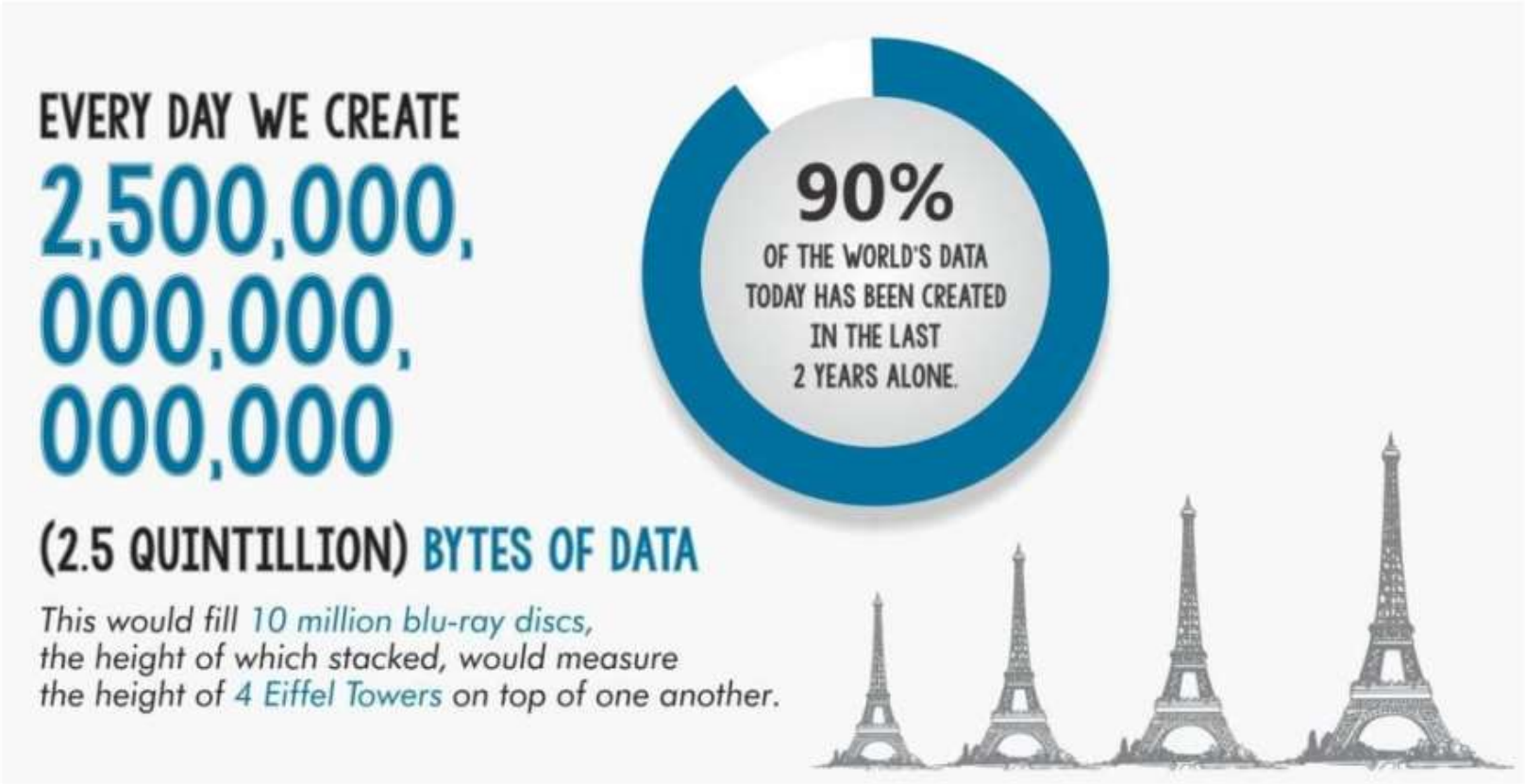
1990

By 2020, it's estimated that 1.7MB of data will be created every second for every person on earth."

— Domo report (6th edition)

2020

How much data we create daily?



1 Quintillion = 1 million terabyte

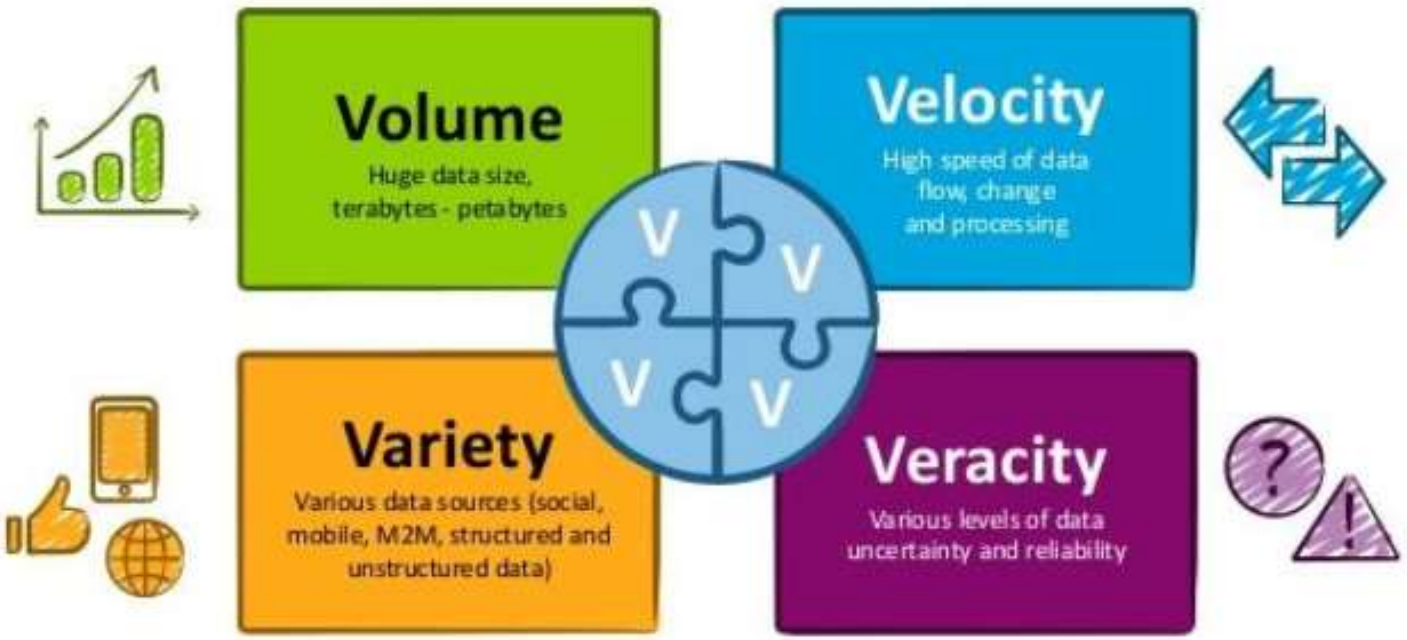
Data is now new oil

We need to find it, extract it, refine it, distribute it and monetize it

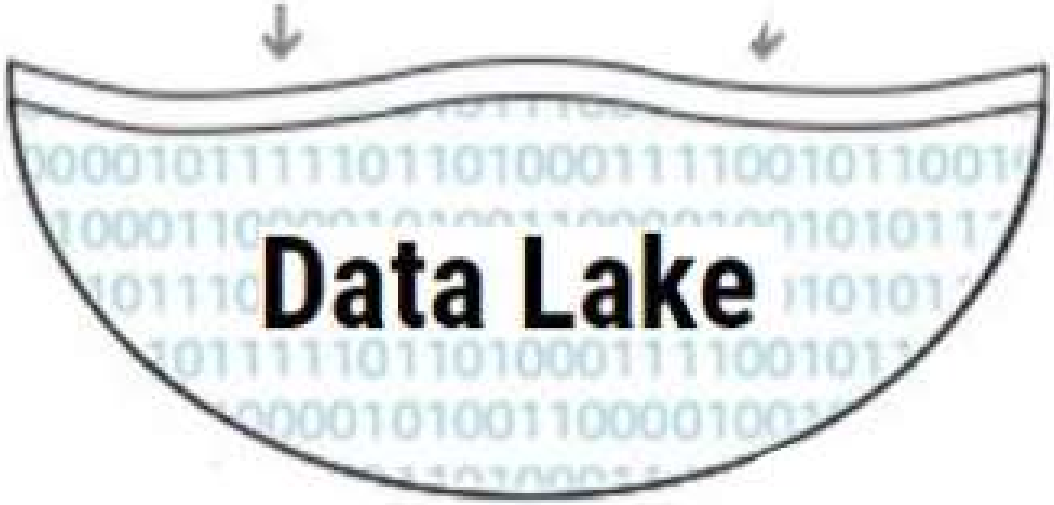
– David Buckinham, Big data expert

Problem statement

- Need a solution which can handle below 4 V's of data.

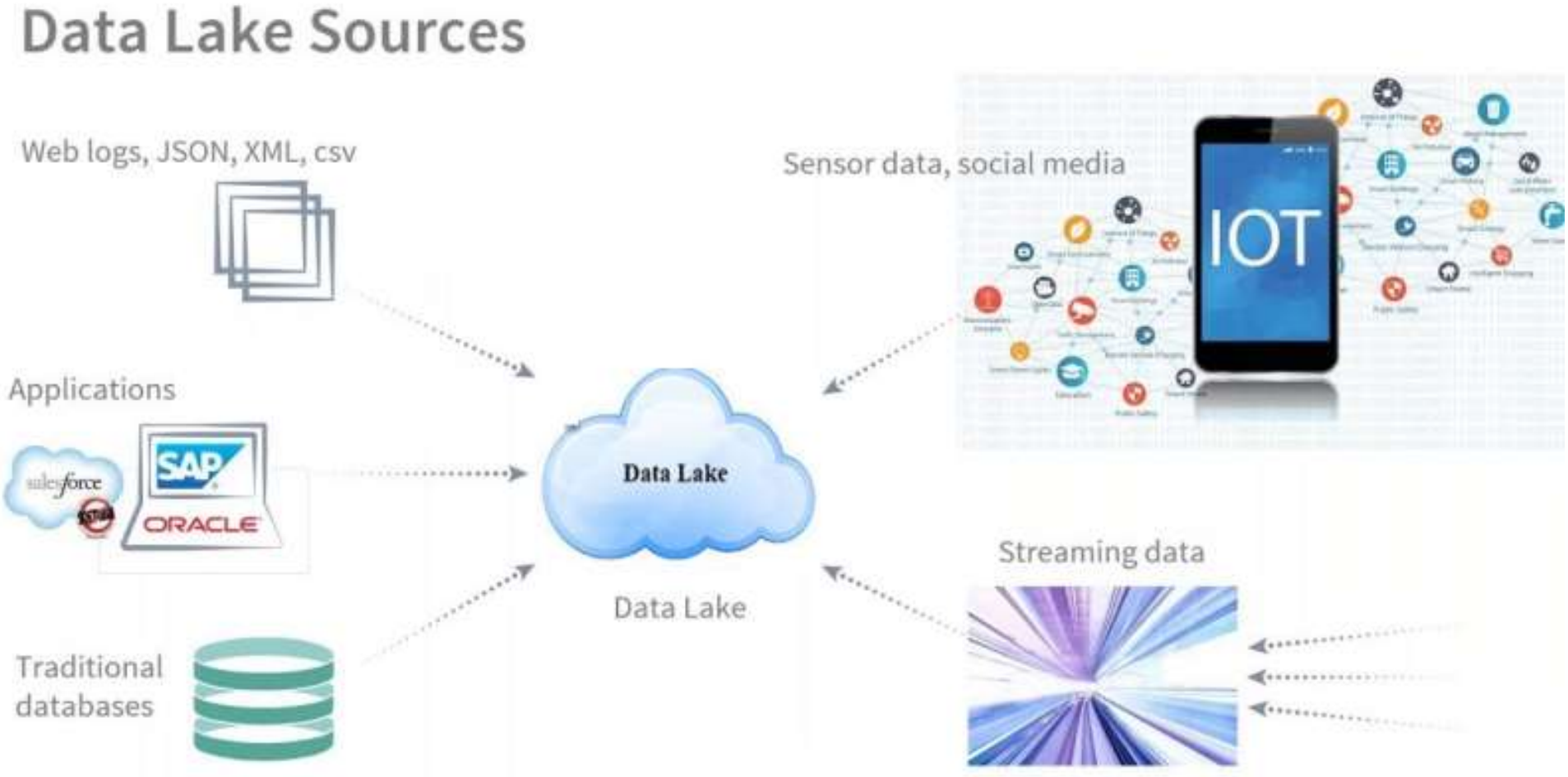


Introduction to Data Lake



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."

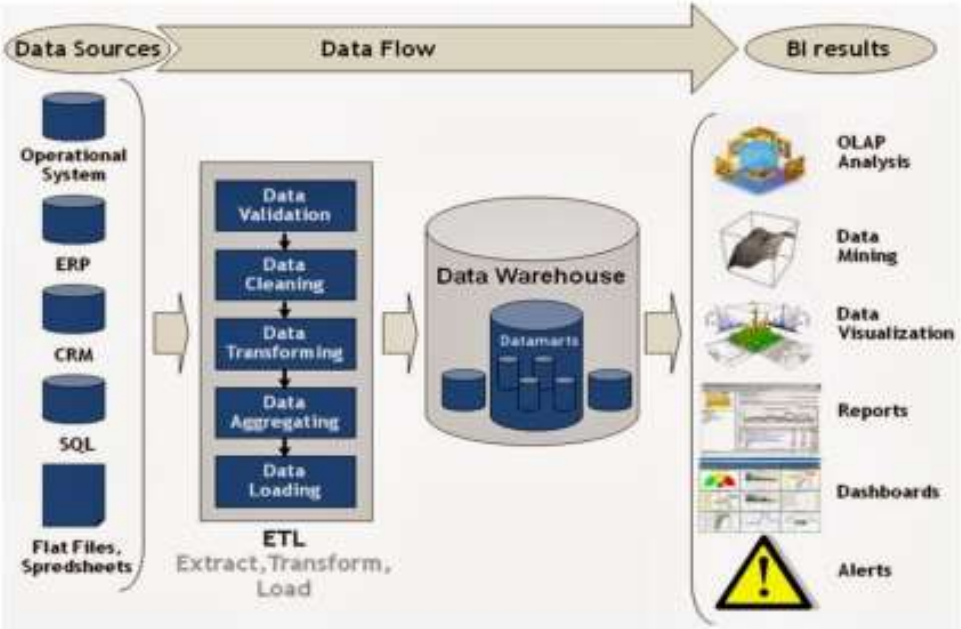
James Dixon
CTO, Pentaho

He coined terminology – "Data Lake"

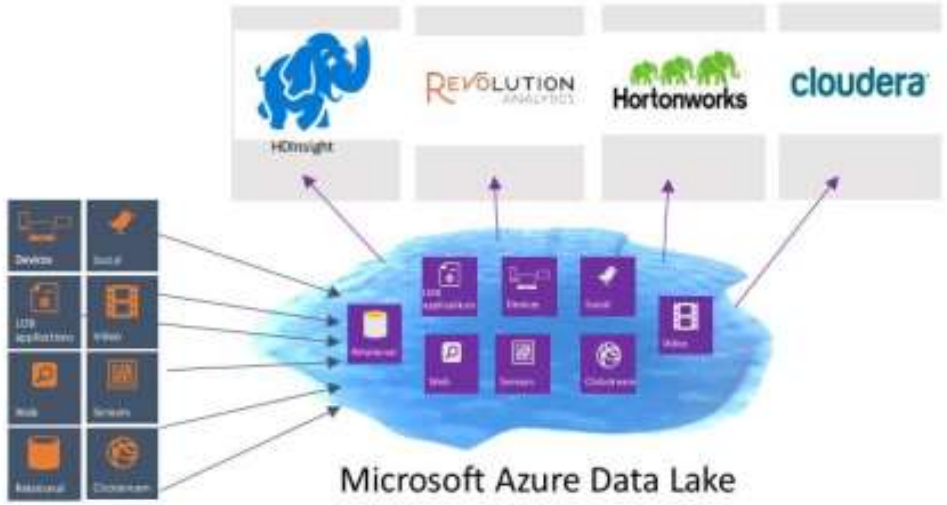
Data Warehouse

Data Lake

Data Warehouse vs Data Lake



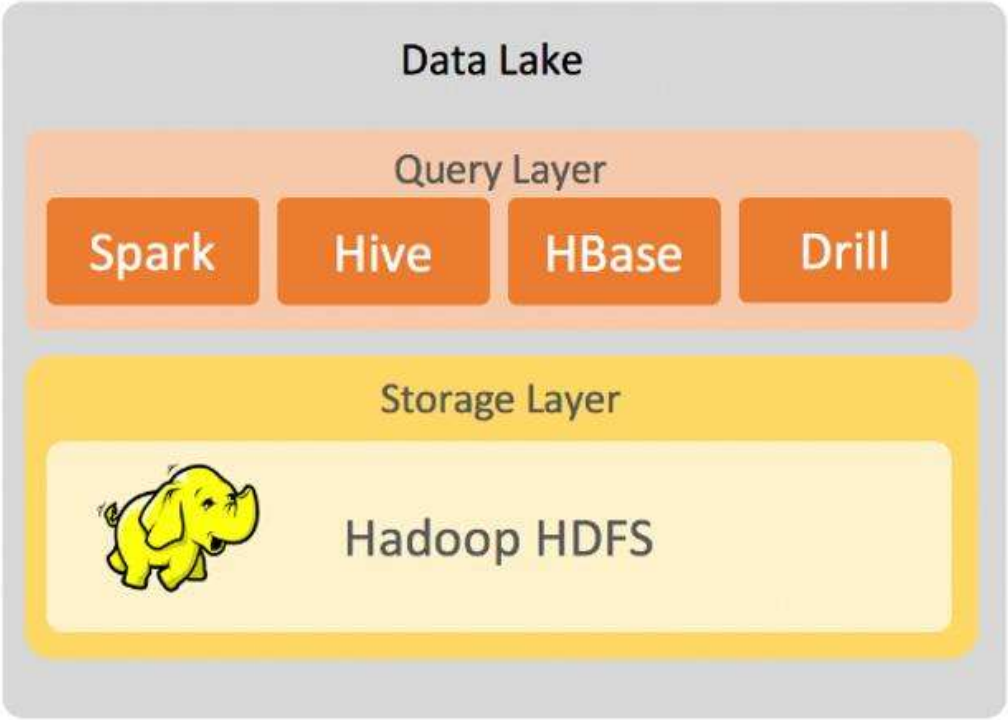
Data warehouse stores data only if it's usage is first identified



Data is just dumped in Data Lake. Processing is done later on to identify the required data

Azure Data Lake Gen1 evolution

- HDFS in Cloud is nothing but Data Lake Gen1 in cloud.



- Fault tolerant file system
- Runs on commodity hardware
- MapReduce, Pig, Hive, Spark etc.

Cloud storage challenge

Processing

- Easy to optimize processing by increasing vCPU and Ram

Storage

- Different requirements
- No direct solution

Microsoft Azure
Blob Storage

Large object storage in cloud

Optimized for storing massive amounts of unstructured data

- Text or Binary Data

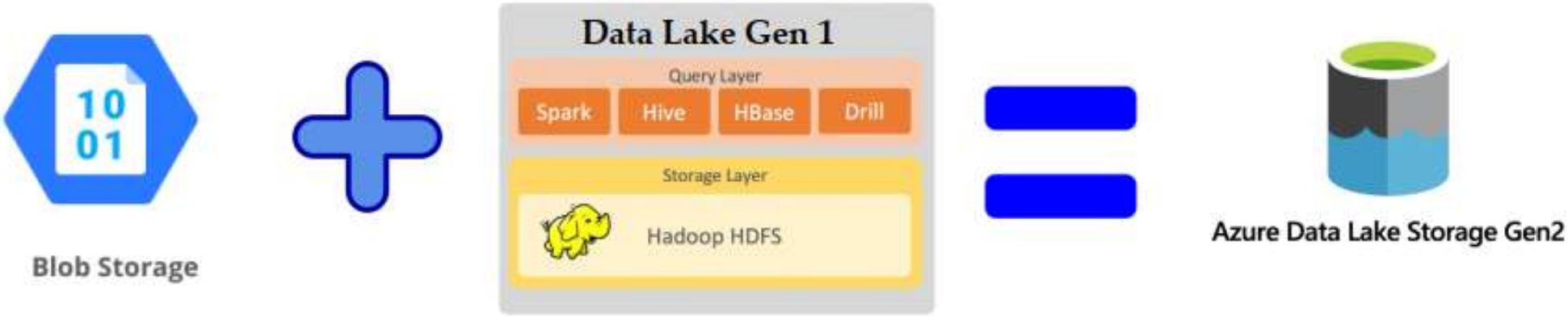
General purpose object storage

Cost efficient

Provide multiple Tiers

Azure Data lake Gen 2

- MICRSOFT RECOMMENDS: Data Lake Storage Gen2

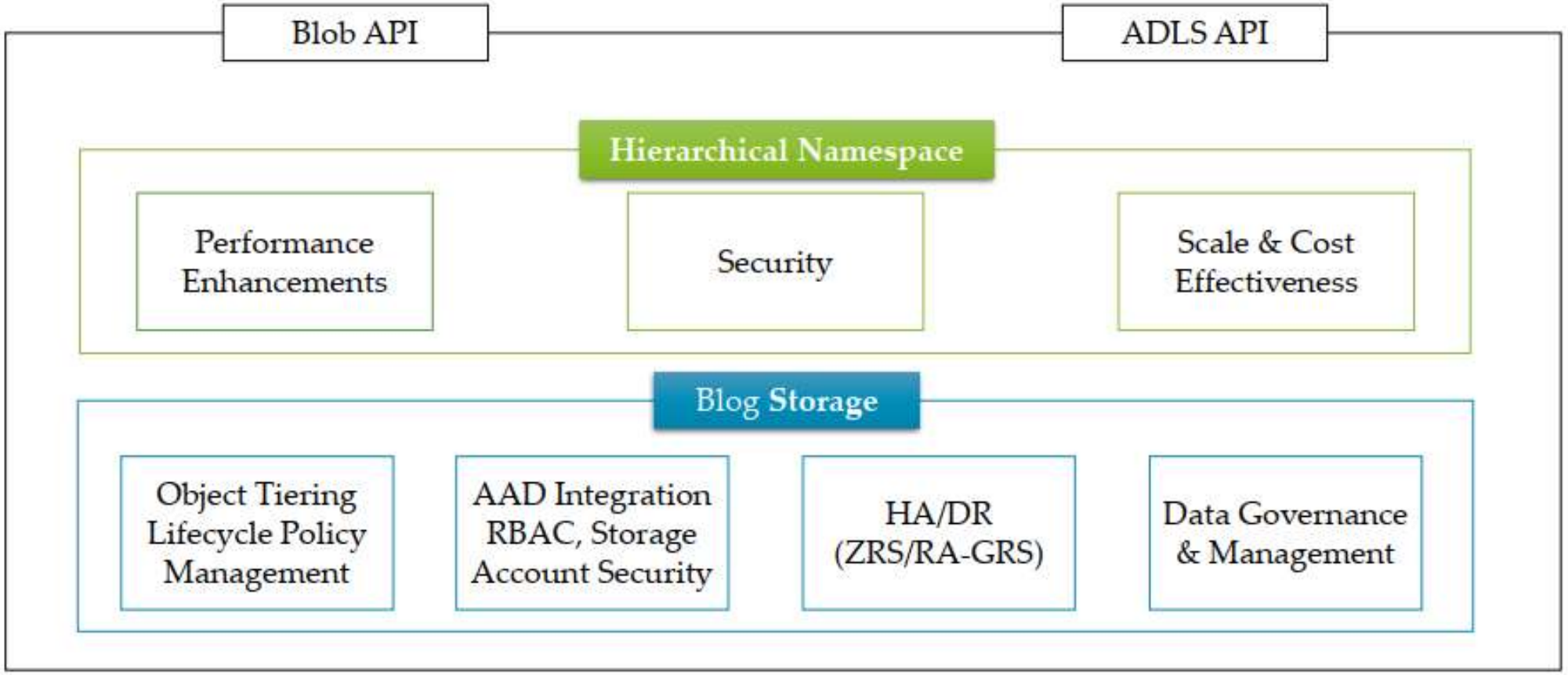


Blob Storage vs Data Lake Storage

Azure Blob Storage	Azure Data Lake Storage (Gen 2)
• General purpose data storage	• Optimized for big data analytics
• Container based object storage	• Hierarchical namespace on Blob Storage
• Available in every Azure region	• Available in every Azure region
• Local and global redundancy	• Local and global redundancy
• Processing performance limit	• Supports a subset of Blob storage features
	• Supports multiple Azure integrations
	• Compatible with Hadoop

Use Gen1 only if USQL is to be used. USQL is currently being supported only in Gen1. USQL is a special Query language for Big Data.

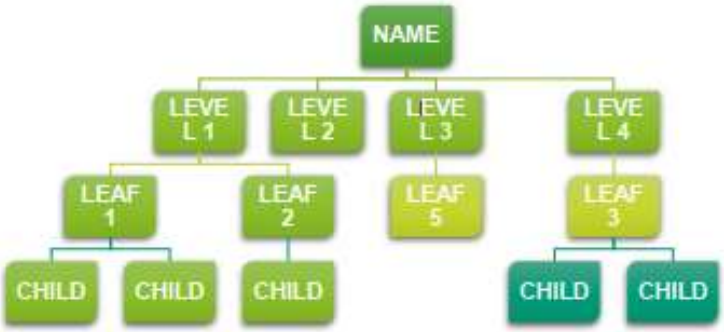
Data Lake Architecture



Hands-On: How to create ADLS Gen2?

- How to create ADLS Gen2?

Hierarchical namespace



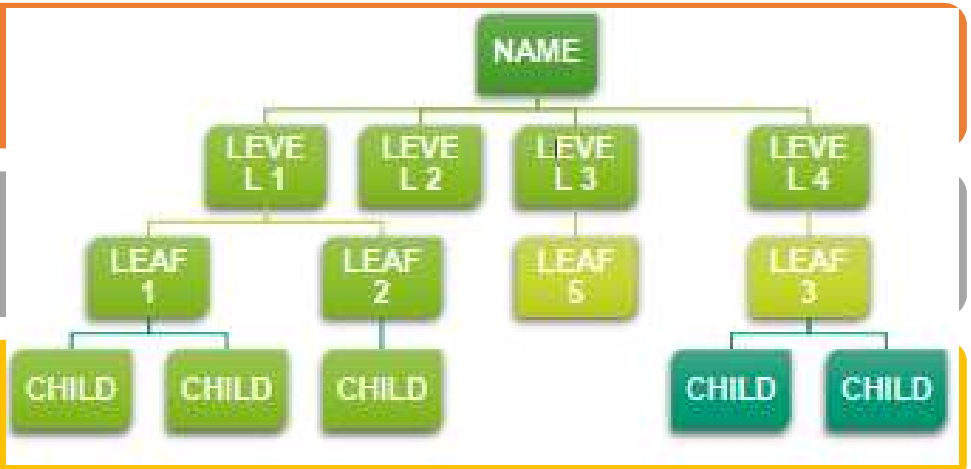
- Hierarchical namespace organizes objects/files into a hierarchy of directories for efficient data access.
- Blob storage is not hierarchical namespace
- Blob can’t integrate with Hadoop

Hierarchical namespace

Manageable

Performance

Cost

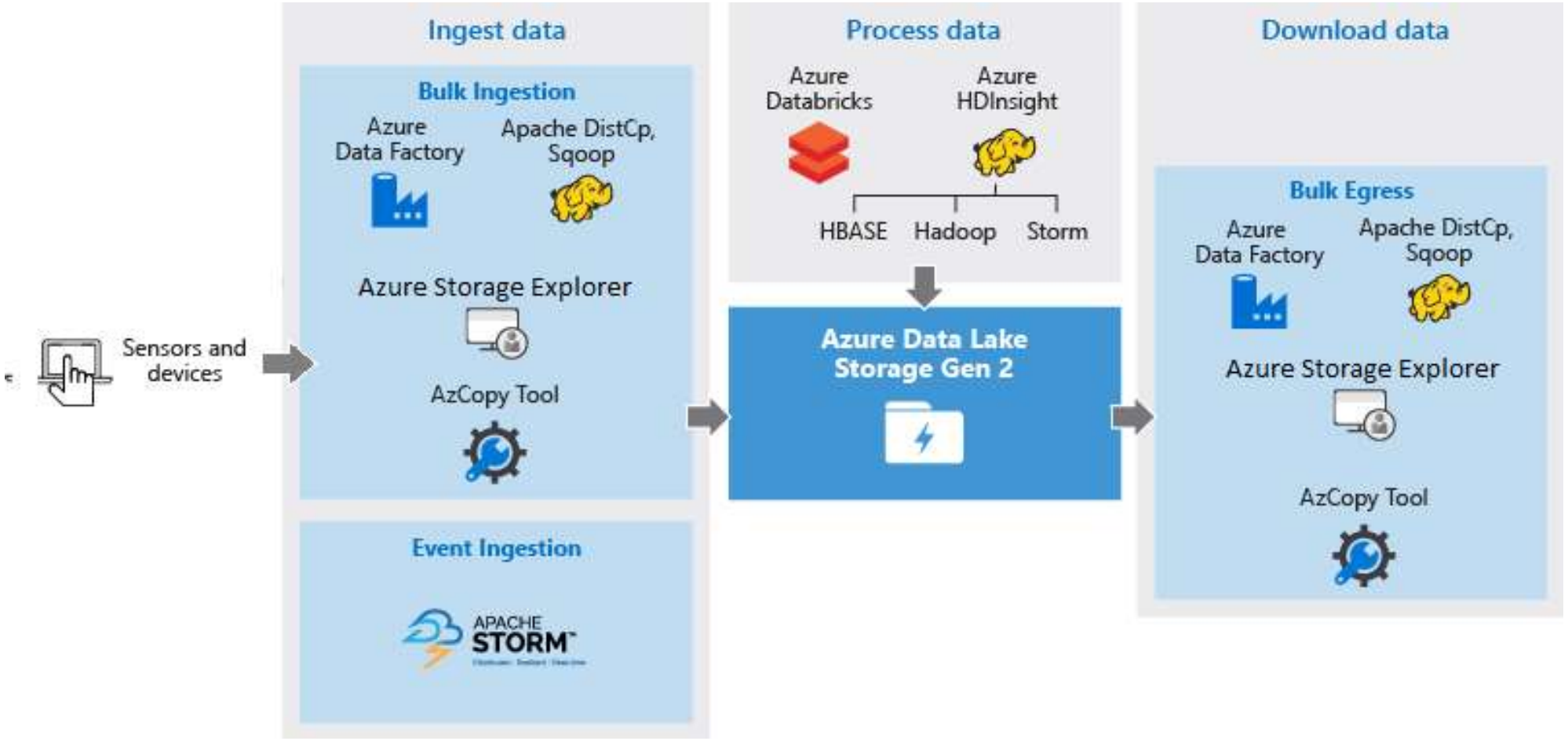


Familiar interface style

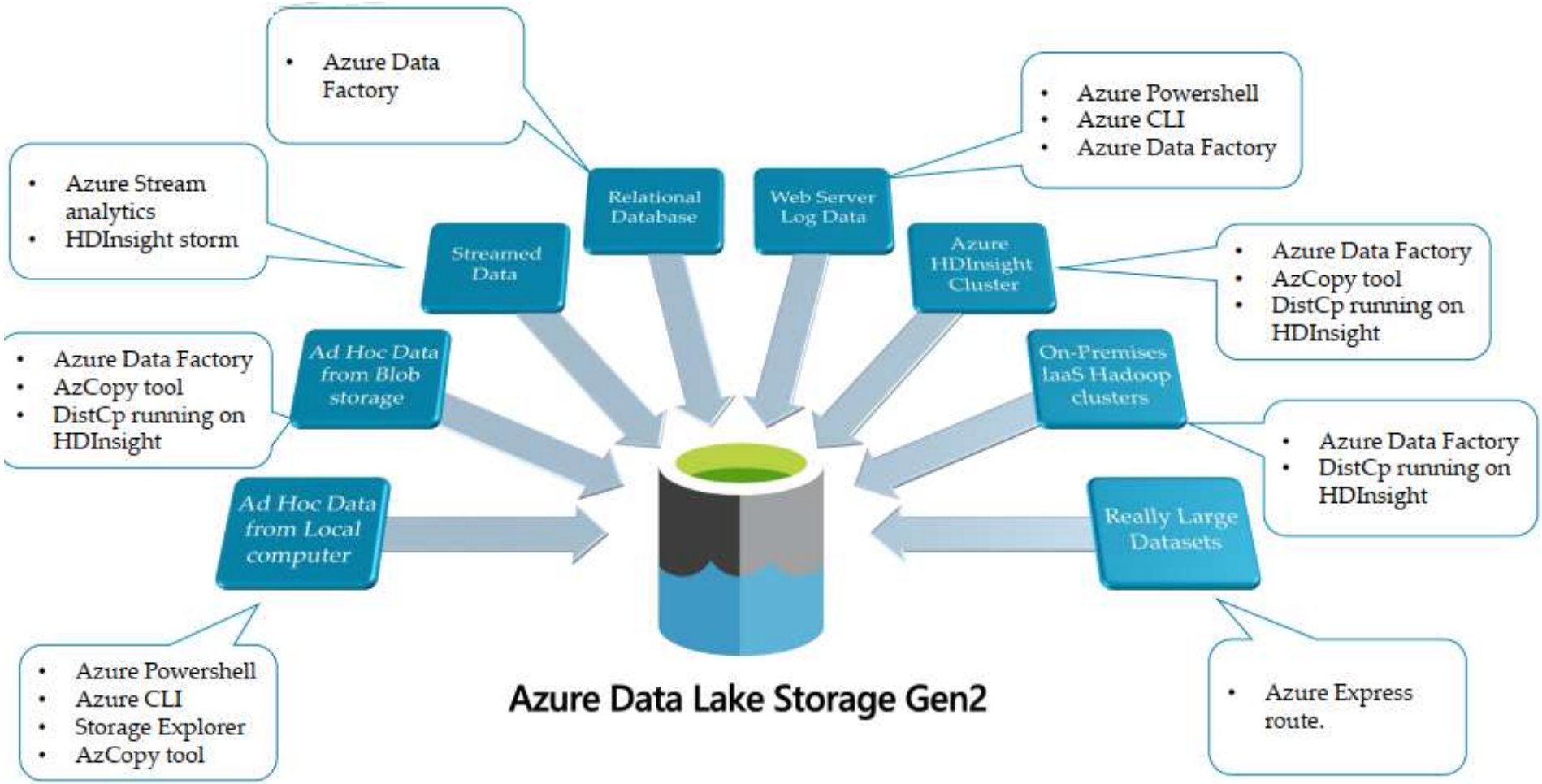
When not to use?

- Where object organization stored separately
- E.g.: Backups, image storage, etc.

Data Lake Architecture



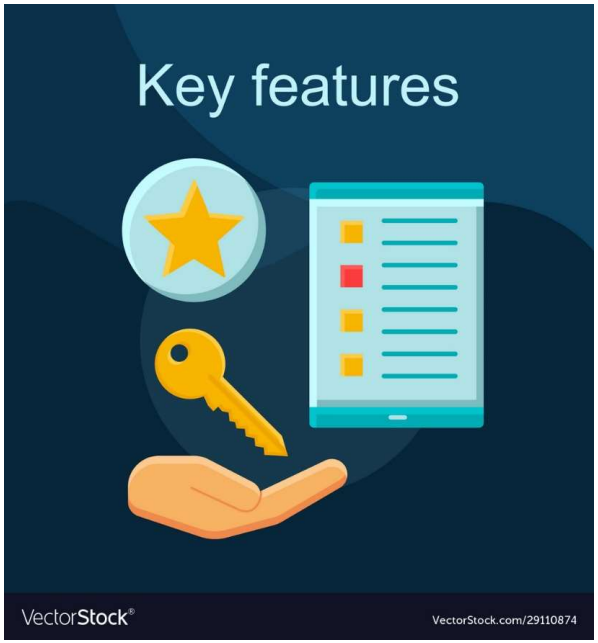
Data Ingestion



Key Features of ADLS Gen2

Important features of Data Lake Gen2

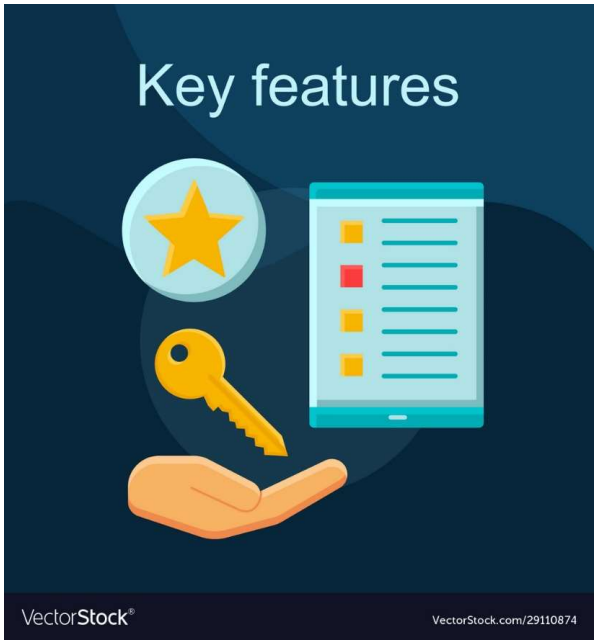
- Integration
  - POSIX complaint
  - Hadoop Integration (use ABFS driver)
  - Other Azure services integration
- Scalability
- Cost effective
  - Build on top of Azure low cost blob storage
  - No need to move data
  - Hierarchical namespace -> less compute -> save cost
- Performance – optimized for high speed throughput



Important features of Data Lake Gen2

- Security
- Fault tolerant/ High availability/ Disaster recovery
- Global footprint

- Challenges
  - Hard to query unstructured data
  - Hard to manage data quality



Hands-on: Security Layers in Data Lake


thanks