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

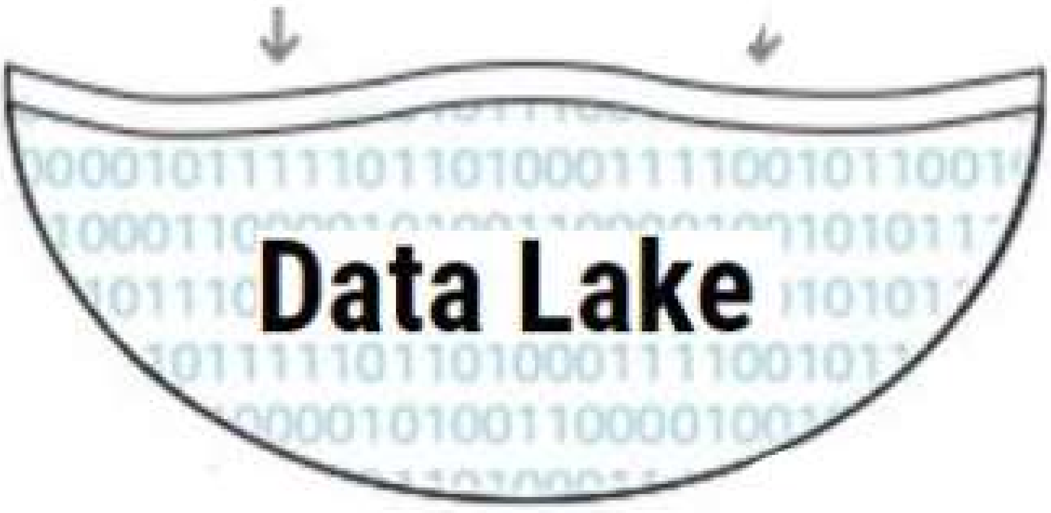
1990

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

― *Domo report (6th edition)*

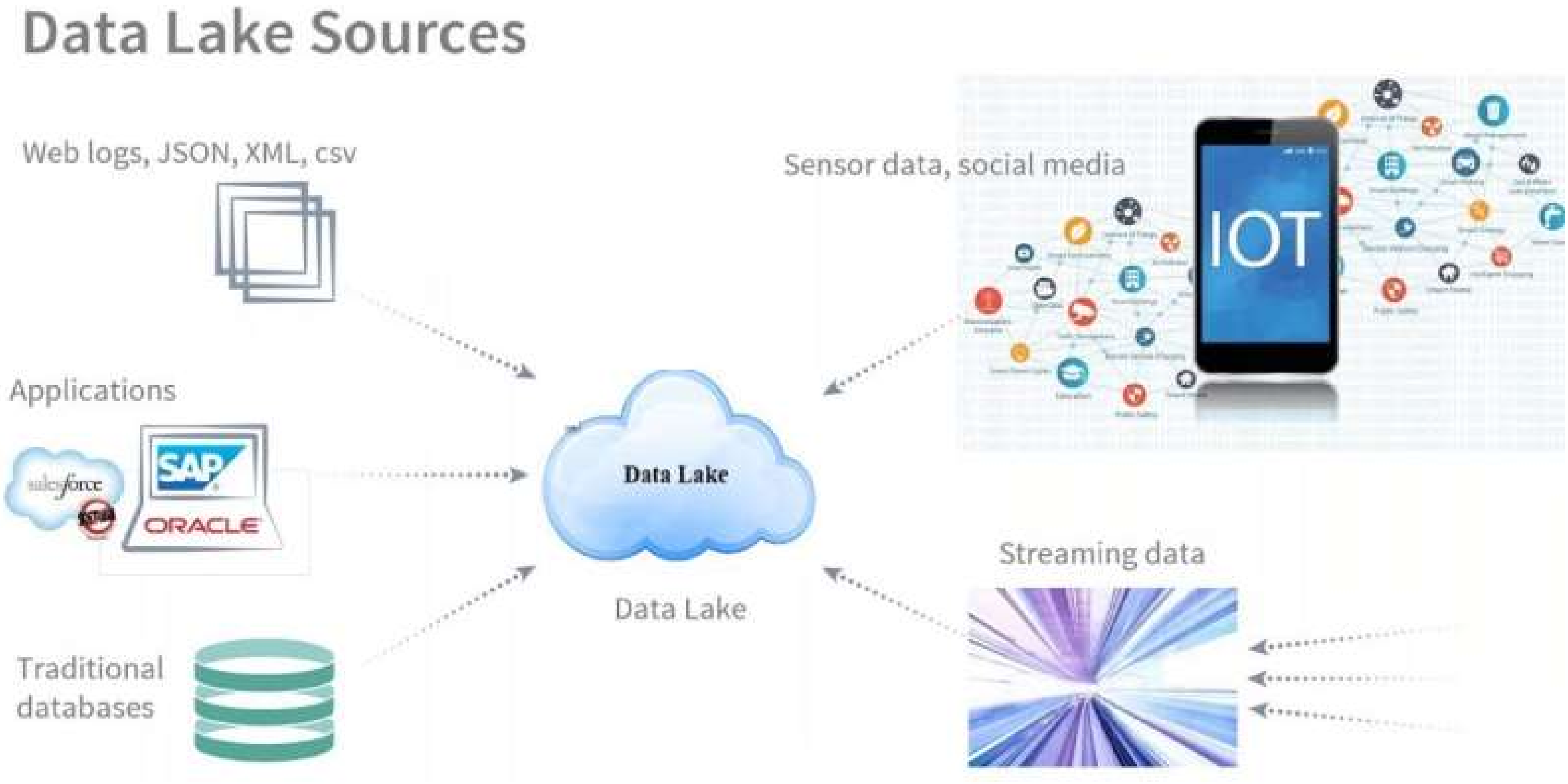
2020

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"

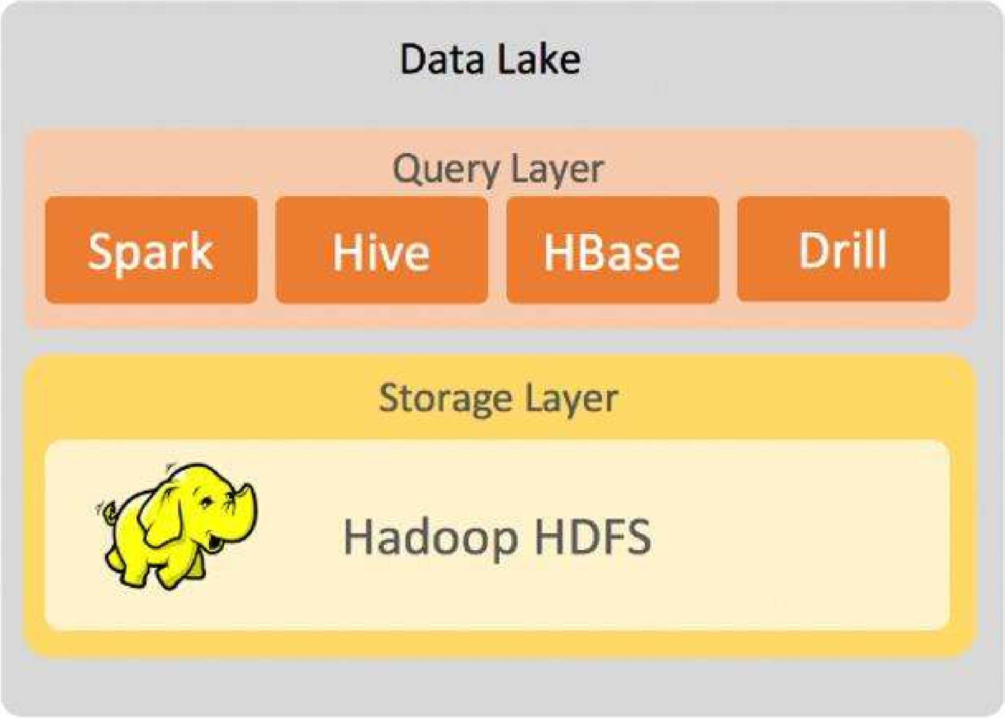
Data Warehouse

Data Lake

Azure Data Lake Gen1 evolution

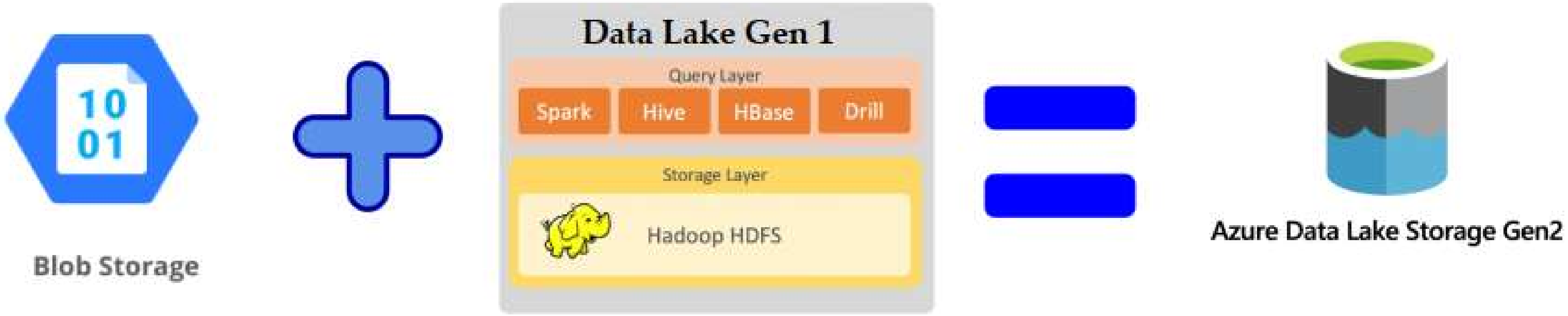
- HDFS in Cloud is nothing but Data Lake Gen1 in cloud.

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

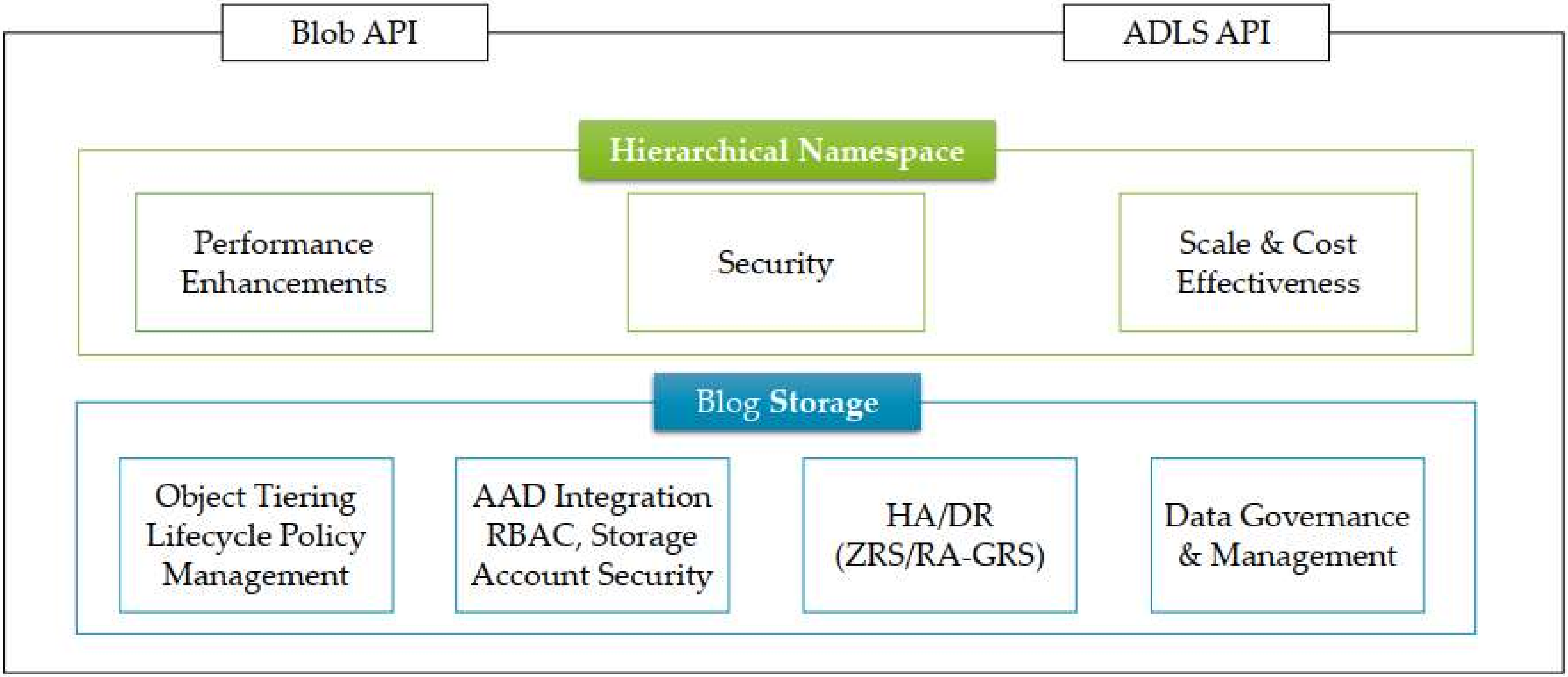


Azure Data lake Gen 2

- MICRSOFT RECOMMENDS: Data Lake Storage Gen2



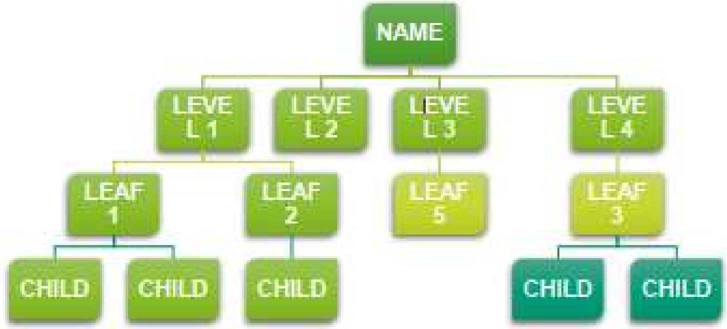
Data Lake Architecture

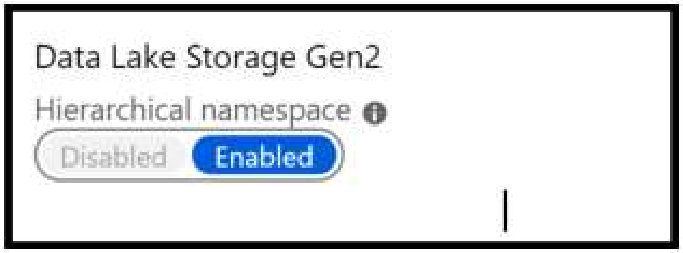

Blob API
ADLS API
Hierarchical Namespace
Performance Enhancements
Security
Scale & Cost Effectiveness
Blog Storage
Object Tiering Lifecycle Policy Management
AAD Integration RBAC, Storage Account Security
HA/DR (ZRS/RA-GRS)
Data Governance & Management

Hands-On: How to create ADLS Gen2?

- How to create ADLS Gen2?

Hierarchical namespace





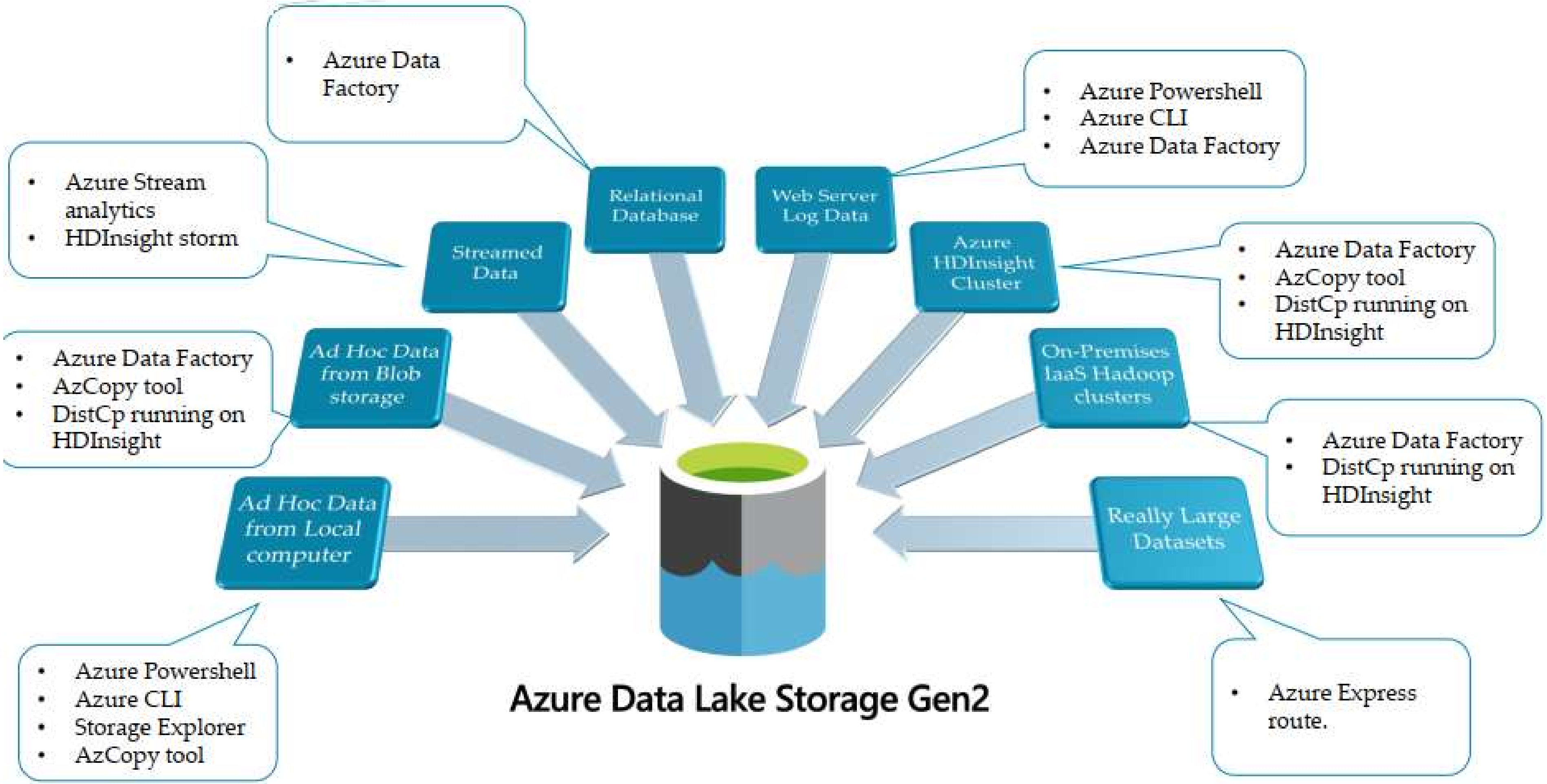
Data Lake Storage Gen2

Hierarchical namespace

○ Disabled ● Enabled

- 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

Data Ingestion



Relational Database

- Azure Data Factory

Streamed Data

- Azure Stream analytics
- HDInsight storm

Ad Hoc Data from Blob storage

- Azure Data Factory
- AzCopy tool
- DistCp running on HDInsight

Ad Hoc Data from Local computer

- Azure Powershell
- Azure CLI
- Storage Explorer
- AzCopy tool

Web Server Log Data

- Azure Powershell
- Azure CLI
- Azure Data Factory

Azure HDInsight Cluster

- Azure Data Factory
- AzCopy tool
- DistCp running on HDInsight

On-Premises IaaS Hadoop clusters

- Azure Data Factory
- DistCp running on HDInsight

Really Large Datasets

- Azure Express route.

Azure Data Lake Storage Gen2

Thanks