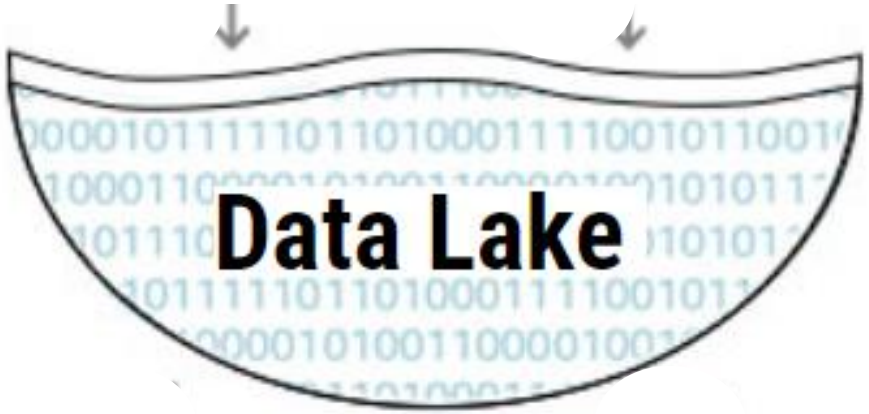
Azure Data Lake Introduction

Eshant Garg

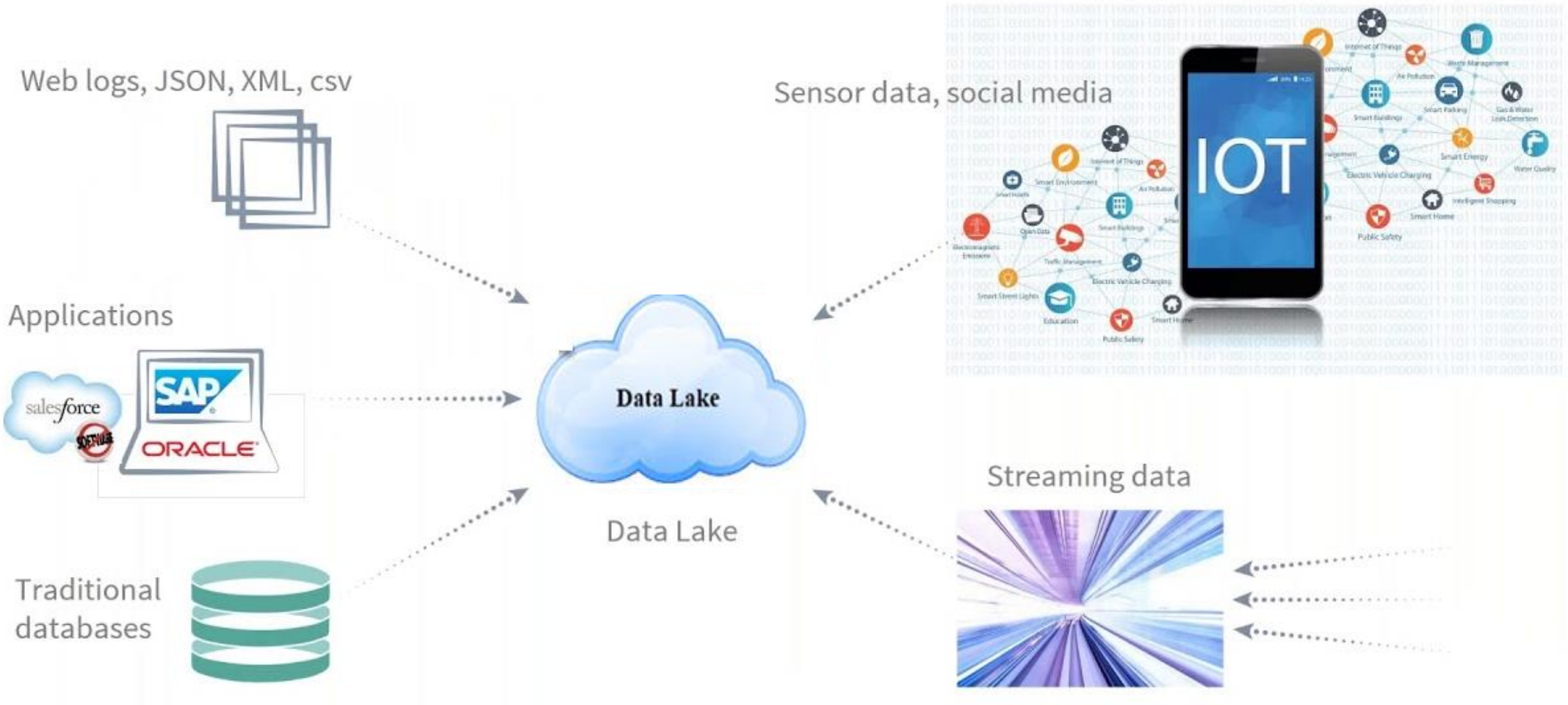
Azure Data Engineer, Architect, Advisor

eshant.garg@gmail.com



Data Lake is a big container to store data.

Data Lake Sources



What is Data Lake?

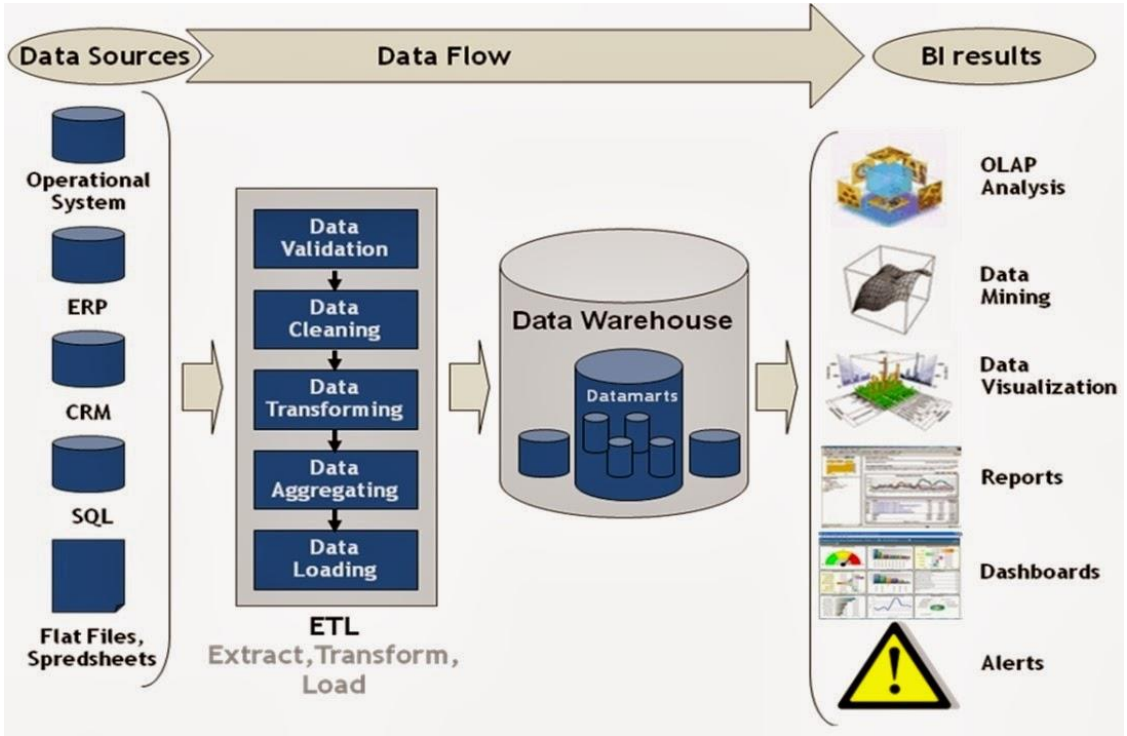
"If you think of a DataMart as a store of bottled water – clean and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine,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

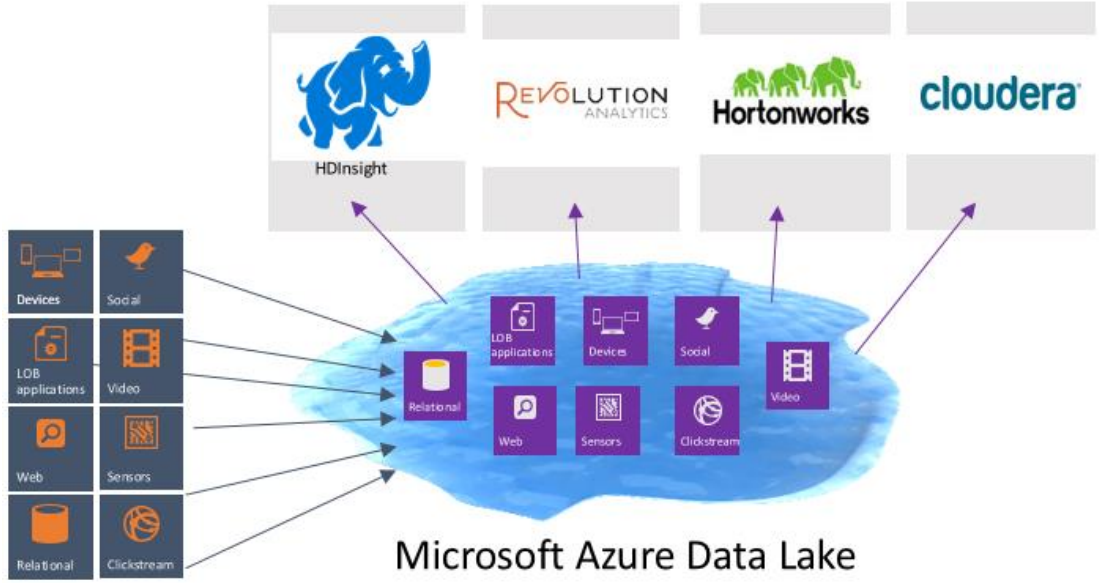
James Dixon
CTO, Pentaho

Data Warehouse

Data Lake

Data Warehouse vs Data Lake





Data Warehouse

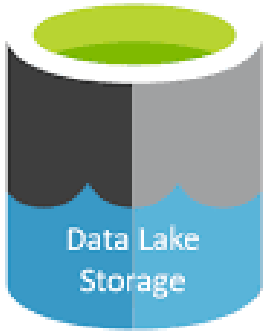
Data Lake

LearnCloud.Info

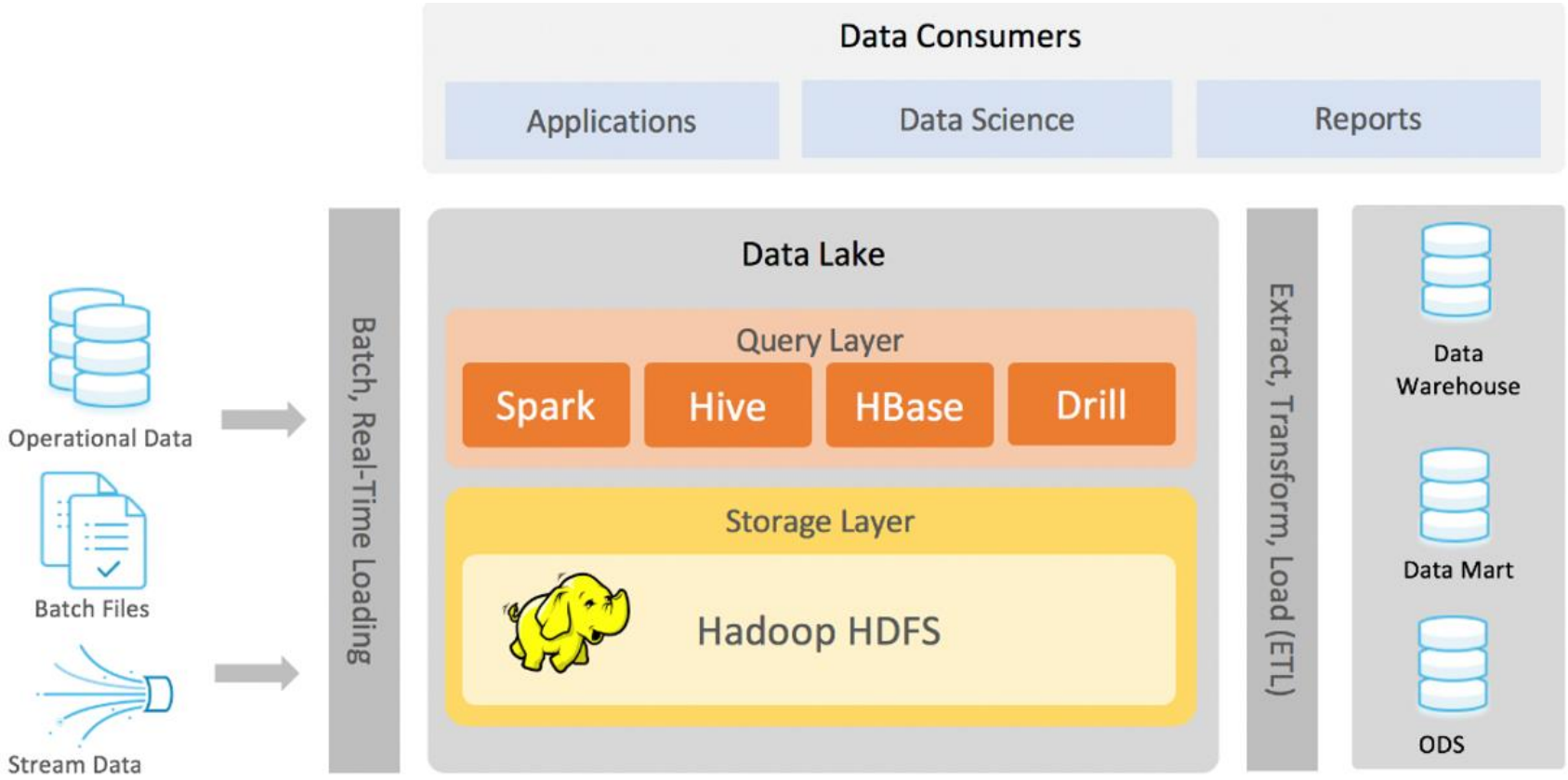

LearnCloud.Info

Data lake vs Hadoop

VS



Data lake vs Hadoop

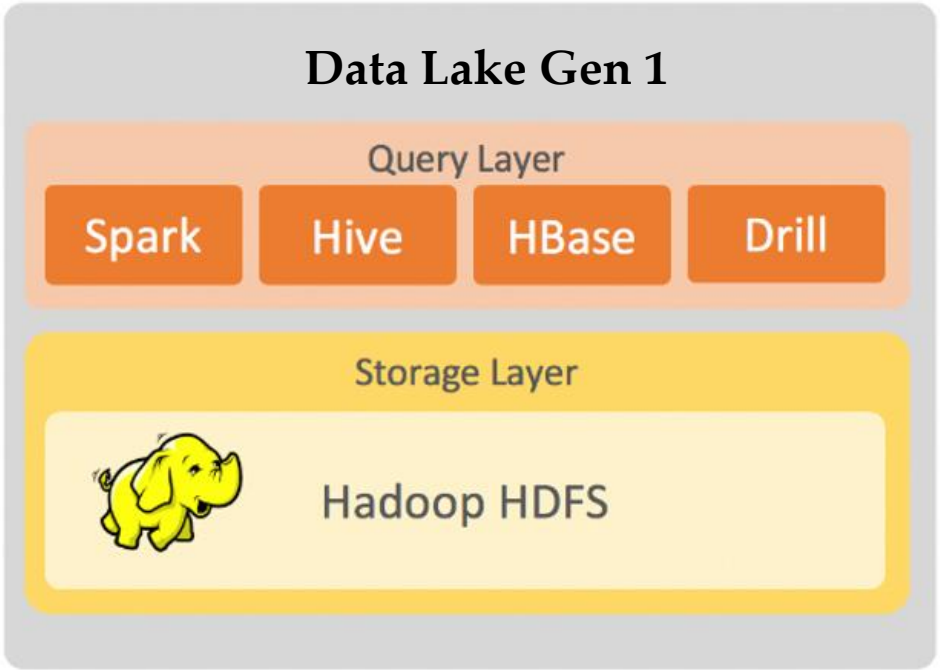


Data Lake can use:

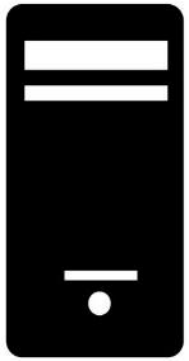

LearnCloud.Info

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

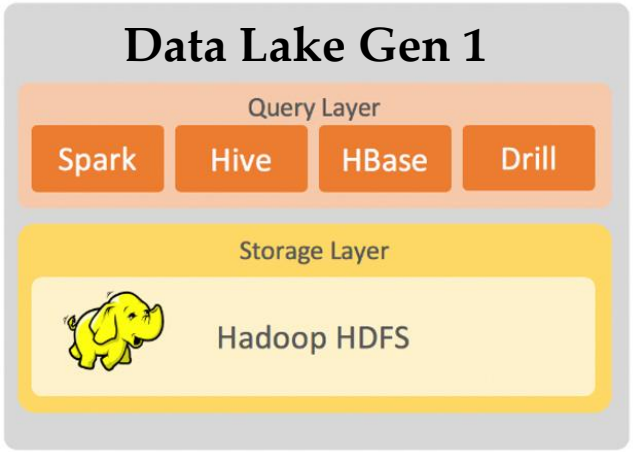
- Different requirements
- No direct solution

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

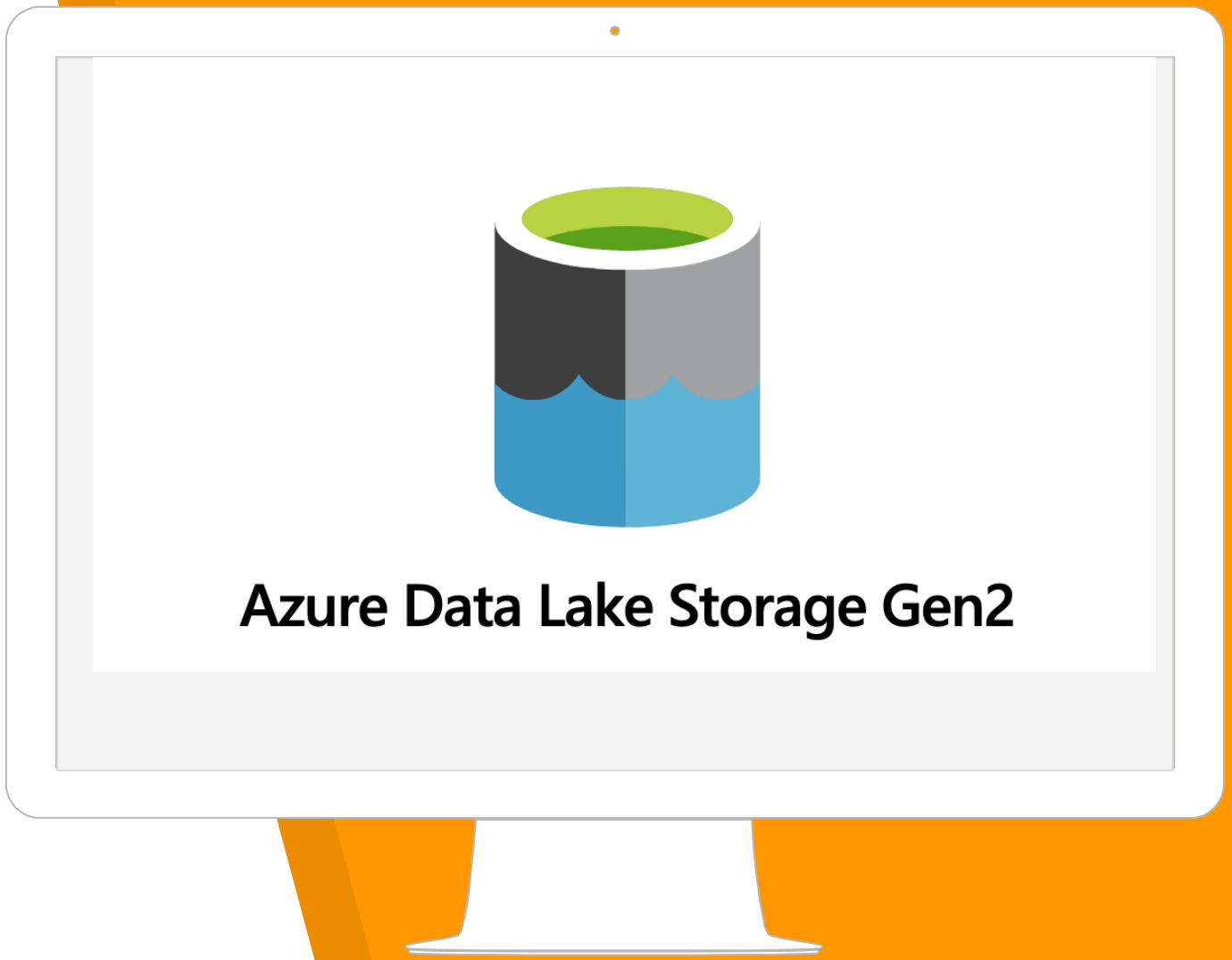


Azure Data Lake Storage Gen2

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

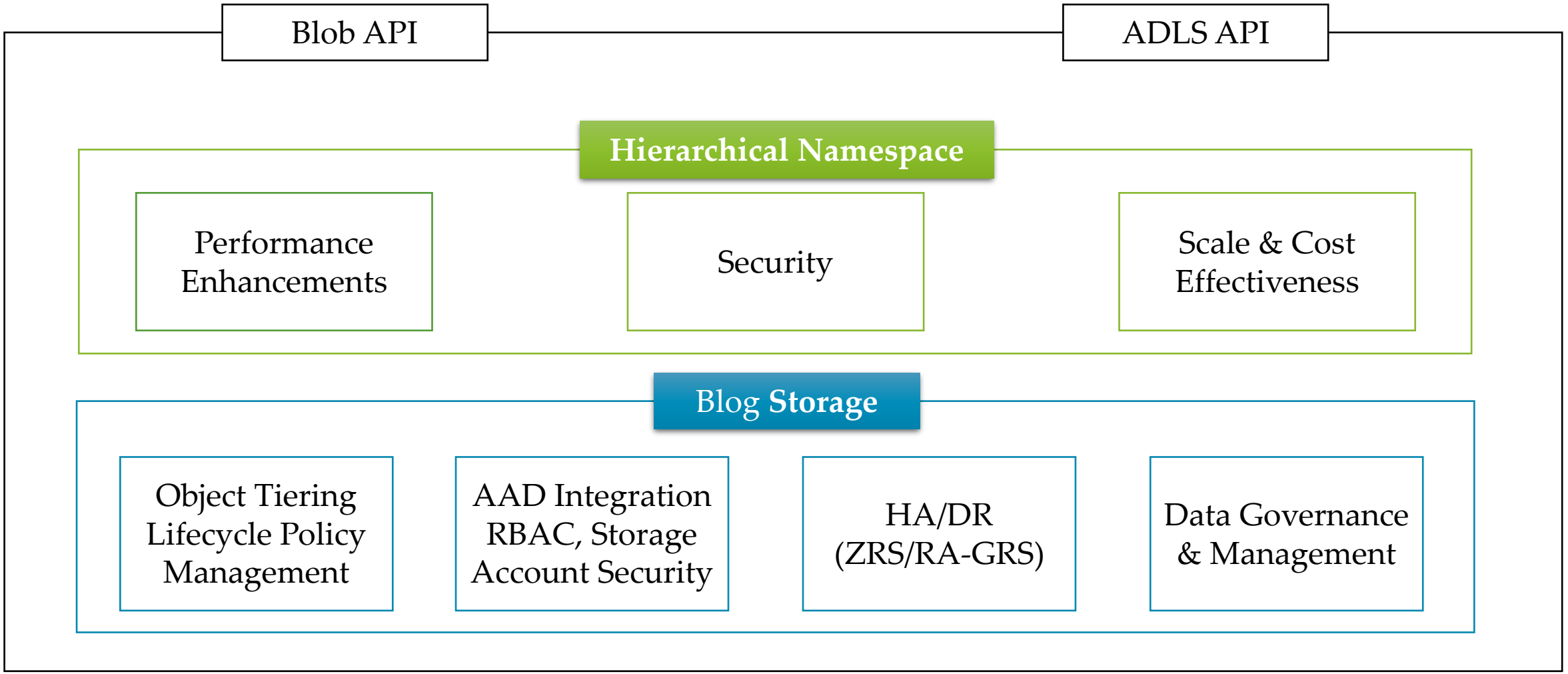
Azure Blob Storage

- 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



Learning objective

- **Authentication**
  - Storage Account keys
  - Shared access signature (SAS)
  - Azure Active Directory (Azure AD)
- **Access Control**
  - Role based access control (RBAC)
  - Access control list (ACL)
- **Network access**
  - Firewall and virtual network
- **Data Protection**
  - Data encryption in transit
  - Data encryption at rest
- **Advanced threat Protection**

Storage Account Access Keys

Authentication

Shared Access Signature (SAS)

Authentication

Shared Access Signature (SAS)

Shared Access Signature

Security token string

“SAS Token”

Contains permission like start and end time

Azure doesn’t track SAS after creation

To invalidate, regenerate storage account key used to sign SAS

Stored access policy

Stored access policy

Reused by multiple SAS

Defined on a resource container

Permissions + validity period

Service level SAS only

Stored access policy can be revoked

Azure Active Directory

Authentication

Azure Active Directory

Azure Active Directory

Azure Active Directory (AD)

- Grand access to Azure Active directory (AD) Identities
- AD is an enterprise identity provider, Identity as a Service (IDaaS)
- Globally available from virtually any device
- Identities – user, group or application principle
- Assign role at Subscription, RG, Storage account, container level.
- No longer need to store credentials with application config files
- Similar to IIS Application pool identity approach

Azure Active Directory

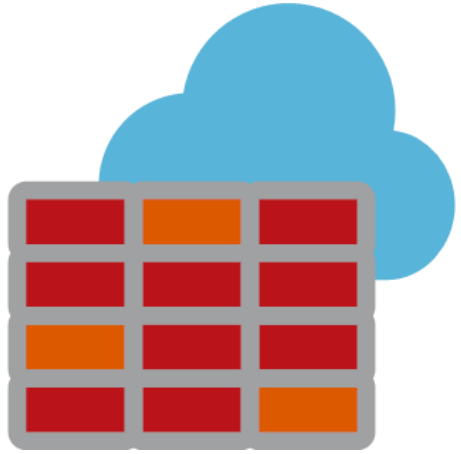
Role based access control (RBAC)

Access Control

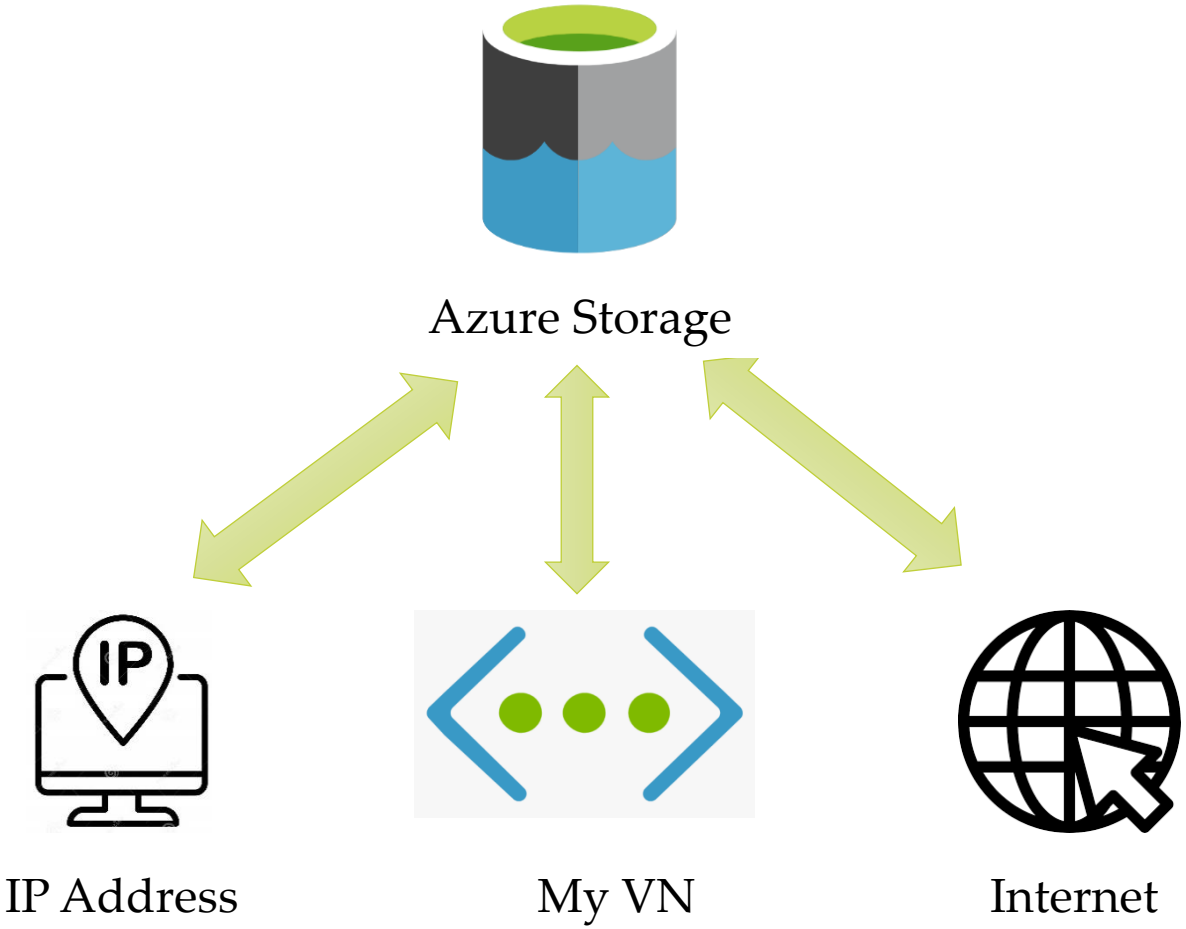
Role based access control (RBAC)

Access control

Firewalls and Virtual Networks



Firewalls and Virtual Networks



Storage Account Access Keys

Authentication

Shared Access Signature (SAS)

Authentication

Shared Access Signature (SAS)

Azure doesn't track SAS after creation

To invalidate, regenerate storage account key used to sign SAS

Stored access policy

Stored access policy

Reused by multiple SAS

Defined on a resource container

Permissions + validity period

Service level SAS only

Stored access policy can be revoked

Azure Active Directory

Authentication

Access Control List (ACL)

Access control

Azure Active Directory (AD)

- Grand access to Azure Active directory (AD) Identities
- AD is an enterprise identity provider, Identity as a Service (IDaaS)
- Globally available from virtually any device
- Identities – user, group or application principle
- Assign role at Subscription, RG, Storage account, container level.
- No longer need to store credentials with application config files
- Similar to IIS Application pool identity approach

Azure Active Directory

Encrypting Data in Transit

Data Protection

Encrypting Data in Transit – Advance

- Site-to-site VPN
- Point-to-site VPN
- Azure ExpressRoute

Client-side Encryption

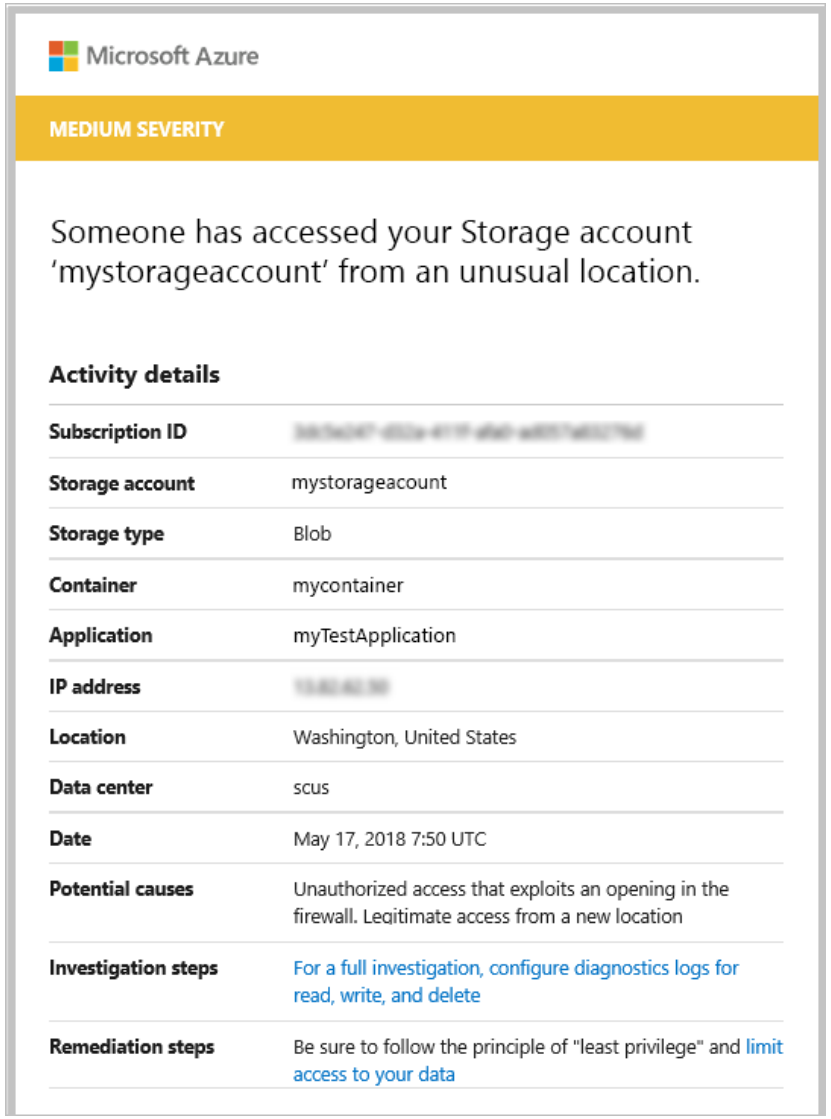
- Encrypt data within application
- Data is encrypted in transit and at rest
- Application decrypt data when retrieved
- HTTPS has integrity checks built-in
- .Net and Java storage client libraries
- Can leverage Azure Key Vault to generate and/or store encryption keys

Encrypting Data at Rest

Data Protection

Encrypting Data at Rest

- Encryption enabled by default
- Can't be disabled
- Storage Service Encryption (SSE)
  - Automatically encrypt and decrypt while writing and reading
  - It's free, no charge
  - Applied to both standard and premium tiers
  - 256 bit AES Encryption
- Option: Use your own encryption keys
  - Blobs and files only



Advanced threat protection

LearnCloud.Info