Microsoft Azure Developer: Implementing Data Lake Storage Gen2

GETTING STARTED WITH AZURE DATA LAKE STORE GEN2

Xavier Morera
PASSIONATE ABOUT ENTERPRISE SEARCH AND BIG DATA
@xmorera www.xaviermorera.com

Azure

But why do we need data?

One might wonder...

“Without data, you're just another person with an opinion”

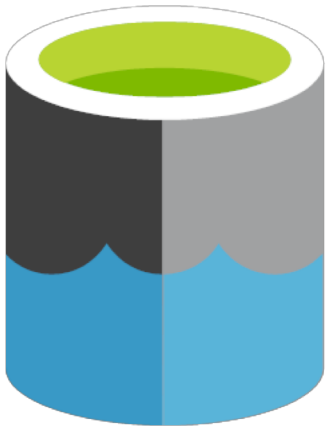
W. Edwards Deming

What Data Lake Storage Is All About

Azure Data Lake Store Gen2

Secure, manageable, performant, scalable, cost effective, and integration ready

Azure Data Lake Storage Gen2

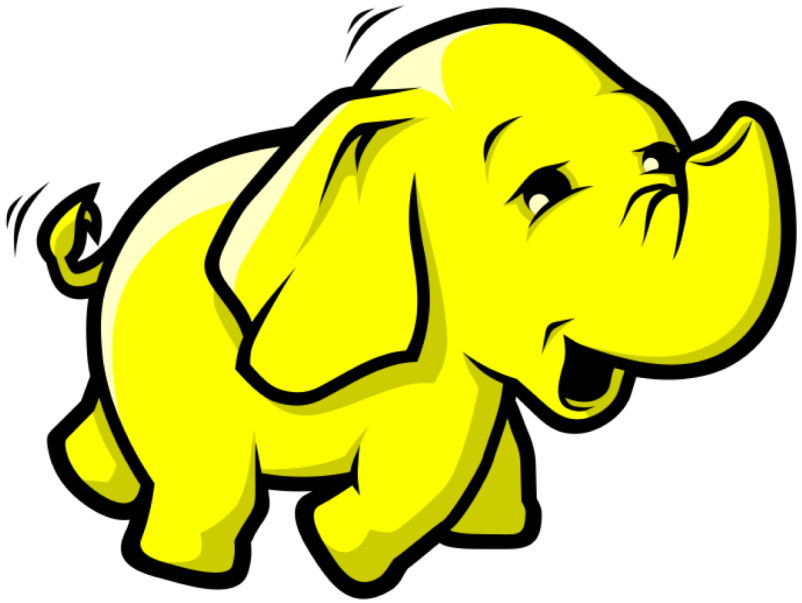


- Secure
- Manageable
- Fast (Atomic operations)
- Scale
- Integration ready
- Cost effective

Why 2?

Putting things in perspective

HDFS: A Widely Used Data Lake



Hadoop Distributed File System

Fault-tolerant file system

Runs on commodity hardware

MapReduce, Pig, Hive, Solr, Spark

HDFS in the cloud

- Data Lake Store Gen1

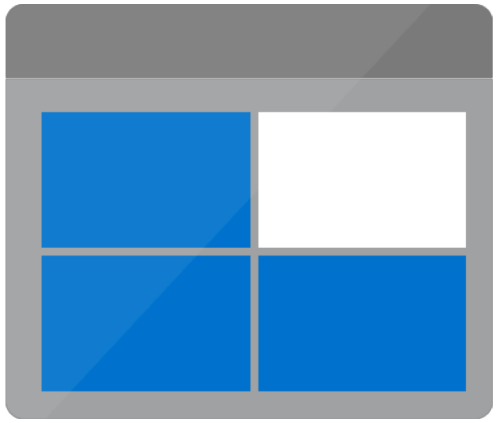
Object Store in the Cloud

Move to the cloud

- Compute

Storage requirements are broader

Azure Blob Storage



Large object storage in the cloud

Optimized for storing massive amounts of unstructured data

- Text or binary data

General purpose object storage

Cost efficient, tiered

Blob Storage

Massively scalable object storage for unstructured data

With exabytes of capacity and massive scalability, Blob Storage stores from hundreds to billions of objects in hot, cool, or archive tiers, depending on how often data access is needed. Store any type of unstructured data—images, videos, audio, documents, and more—easily and cost-effectively.

Configuring and Using Microsoft Azure Blob Storage

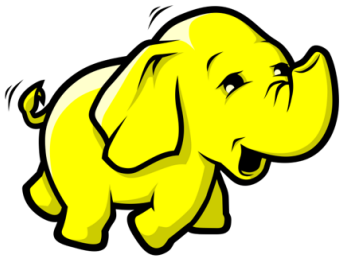
by Neil Morrissey

This course will teach you how to upload and access data securely using a variety of tools and code, and how to integrate with Azure Search and the Content Delivery Network.

A Tale of Two Storage Options

HDFS

HDFS: A Widely Used Data Lake



Hadoop Distributed File System

Fault-tolerant file system

Runs on commodity hardware

MapReduce, Pig, Hive, Solr, Spark

HDFS in the cloud

- Data Lake Store Gen1

Blob Storage

Azure Blob Storage

Large object storage in the cloud

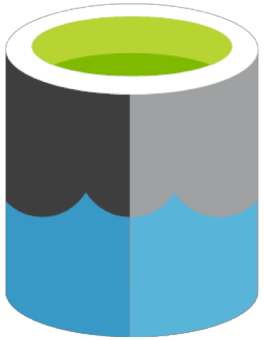
Optimized for storing massive amounts of unstructured data

- Text or binary data

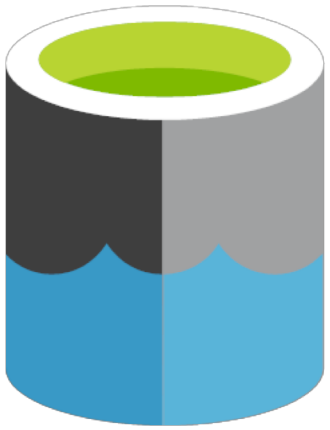
General purpose object storage

Cost efficient, tiered

Important Details
Azure Data Lake Storage Gen2



Azure Data Lake Storage Gen2



Designed for enterprise big data analytics

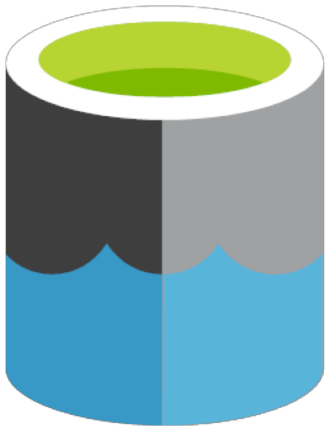
- Data lake on Azure

High throughput, enormous scale

Hierarchical namespace

- Superset of POSIX permissions

Features



Hadoop compatible access

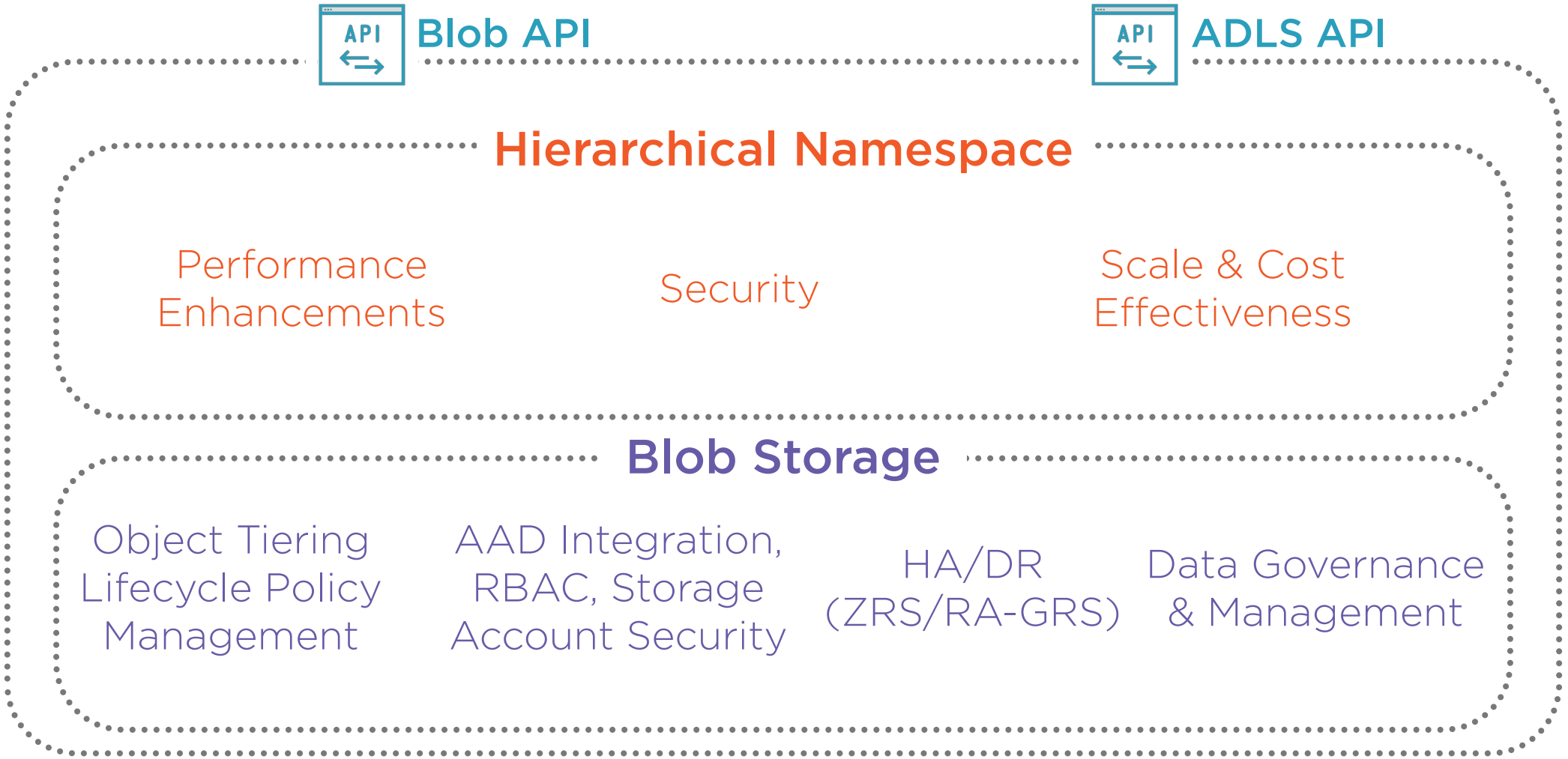
- The Azure Blob File System driver
- Blob compatible

Cost effective

Hierarchical Namespace

Enabled	Disabled
Atomic directory manipulation	Some workloads may not benefit
Familiar file system	Backups
Organized datasets	Images
Nearly linear scalability	Organization data stored separately

Architecture



Azure Storage Documentation

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. Learn how to leverage Azure Storage in your applications with our quickstarts and tutorials.

5-Minute Quickstarts

Learn to work with Azure Storage by exploring these quickstarts.

							
Create storage account							
Blob quickstarts							
Data Lake Storage Gen2							

Demo

Creating an Azure Data Lake Gen2

Using Portal

Create a resource
All services
FAVORITES
Dashboard
All resources
Resource groups
Storage accounts
Virtual machines
Virtual networks
Network security groups
Azure Active Directory
App Services
SQL databases
Data Lake Storage Gen1
Data Lake Analytics
Azure Cosmos DB
Load balancers
Security Center
Cost Management + Billing
Help + support
Monitor
Advisor

Storage accounts
N/A

Add | Edit columns | More

Filter by name...

NAME

- bigdataforscdiag965
- bigdataforscdisks330
- cs2251d961b545ex4b26xabd
- maxdiag717

Create storage account

Basics | Advanced | Tags | Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more

PROJECT DETAILS

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

* Subscription: Visual Studio Enterprise
* Resource group: bigdata-for-sc
Create new

INSTANCE DETAILS

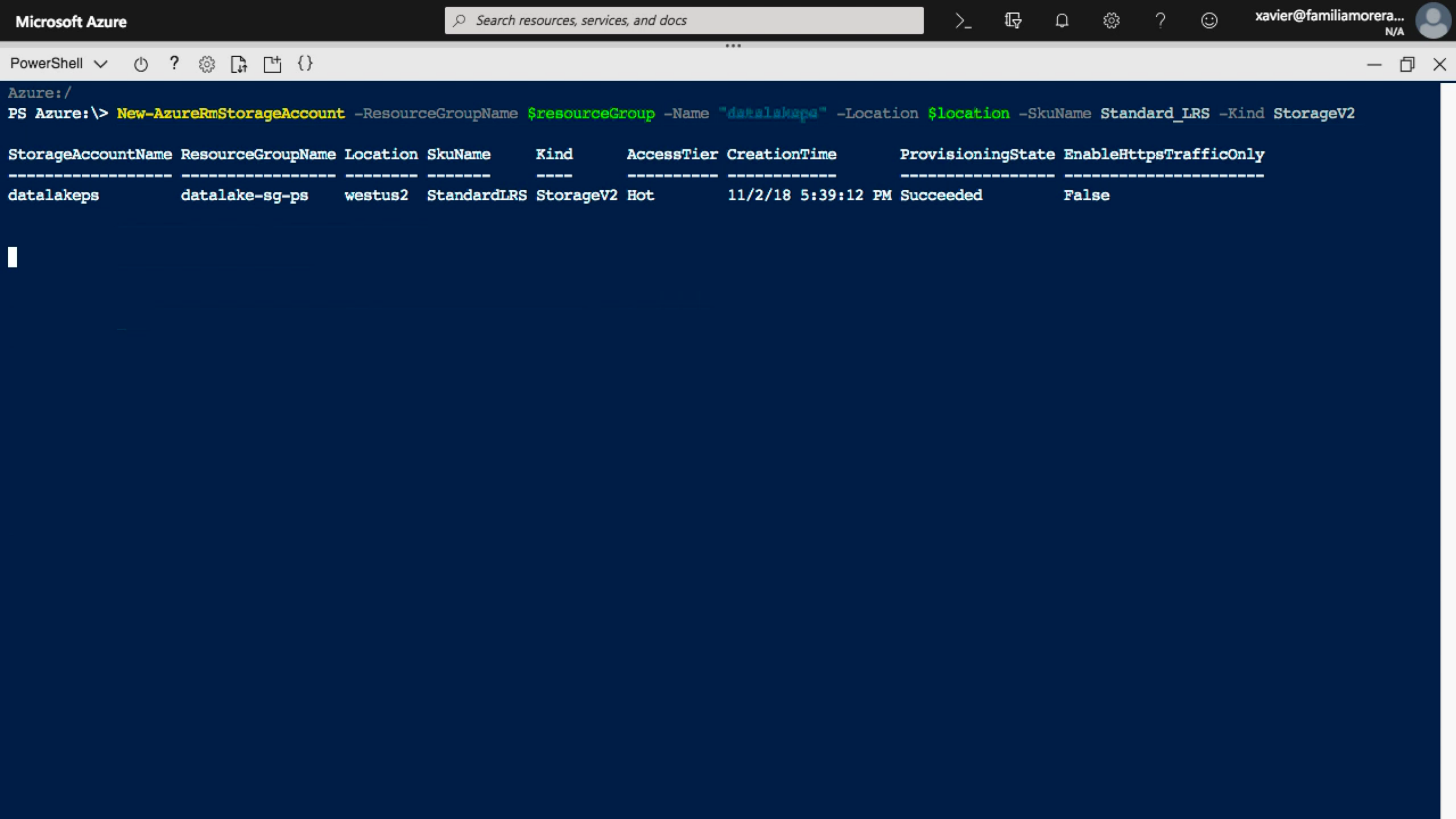
The default deployment model is Resource Manager, which supports the latest Azure features. You may choose to deploy using the classic deployment model instead. Choose classic deployment model

* Storage account name:
* Location: East US
Performance: (•) Standard ( ) Premium
Account kind: StorageV2 (general purpose v2)
Replication: Read-access geo-redundant storage (RA-GRS)
Access tier (default): ( ) Cool (•) Hot

Review + create | Previous | Next : Advanced >

Demo

Creating and Deleting an Azure Data Lake Store Gen2 with PowerShell


Microsoft Azure
Search resources, services, and docs
xavier@familiamorera...
N/A
PowerShell
Azure:/
PS Azure:\> New-AzureRmStorageAccount -ResourceGroupName $resourceGroup -Name "datalakeps" -Location $location -SkuName Standard_LRS -Kind StorageV2
StorageAccountName ResourceGroupName Location SkuName Kind AccessTier CreationTime ProvisioningState EnableHttpsTrafficOnly
datalakeps datalake-sg-ps westus2 StandardLRS StorageV2 Hot 11/2/18 5:39:12 PM Succeeded False

Takeaway

Data is key, the more the better

- HDFS

Data in the cloud

- Blob Storage

HDFS plus Blob Storage

Azure Data Lake Store Gen2

Takeaway

All the benefits of both worlds

- Hierarchical namespace
- Massive scale
- Performance
- Low cost
- Integration ready

Easy to create