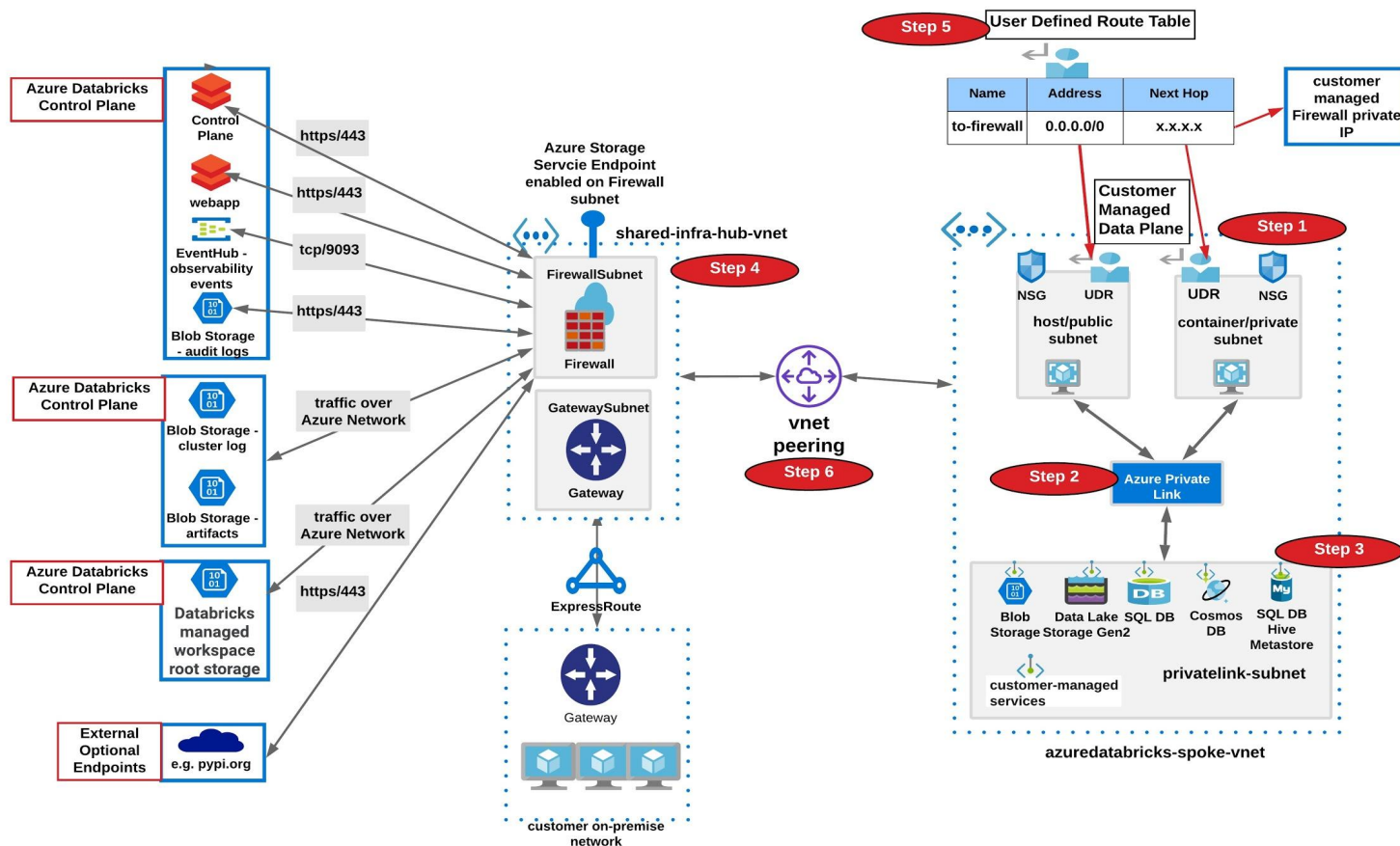


Azure Databricks

(Data Exfiltration Prevention)

- Restrict inbound communication to Azure Databricks clusters (data plane)
- Filter outbound communication using an egress appliance (Firewall)
- Traffic to Azure hosted services stays on Azure backbone

Deploy Azure Databricks to prevent data exfiltration



Prerequisites - Azure Databricks(ADB) Workspace

- Bring your own Virtual Network (VNET) i.e. create a brand new or use an existing one.
 - CIDR range between /16 to /24
- Network Security Group (dedicated to ADB subnets)
 - With default rules
- A pair of subnets, used by Azure Databricks only, 1 workspace requires 2 subnets.
 - CIDR range between /18 to /26
 - Subnet delegation set to Azure Databricks
 - Azure Databricks NSG associated with these subnets
- **Multiple ADB workspaces** could reside in the **same VNET**
- **Pair of subnet's** used by Azure Databricks can only be associated with **1 workspace**

Prerequisites - Egress firewall configuration

- Azure Firewall (you could also use any other Azure hosted egress appliance)
 - Deployed in a separate VNET or in the same VNET as ADB
 - Dedicated subnet called AzureFirewallSubnet with CIDR /26
 - Egress rules for Azure Databricks Control Plane services
 - Get your regional endpoints from [here](#)

Firewall Egress Rules - Checklist

Azure Databricks Control Plane Services	Type	Endpoint Address	Transport	Port
Relay (used by no public ip enabled workspaces)			https	443
artifacts (runtime images)	blob storage		https	443
logs (audit and cluster)	blob storage		https	443
health-check (observability)	eventhub		tcp	9093
webapp			https	443
dbfs (customer owned)	blob storage		https	443
managed-hive	mysql		tcp	3306
External lib dependency provider services required by application code				
python lib repo	public repo	*pypi.org , *pythonhosted.org	https	443
r package repo	public repo	cran.r-project.org , cran.rstudio.com	https	443
content delivery / required by Ganglia UI	cdn	cdnjs.cloudflare.com	https	443
Optional Services				
demo-datasets-mounts token service, to get temporary token from STS to access the databricks-datasets S3 bucket		sts.amazonaws.com	https	443
demo-datasets-mounts storage bucket, to access the /databricks-datasets folder on DBFS in ADB		databricks-datasets-oregon.s3.us-west-2.amazonaws.com/	https	443

Follow Along Video

Video [Link](#)

Deployment Sequence

- Make a list of [Azure Databricks regional control plane service endpoints](#), we'll need them to configure firewall egress rules. Please use [this](#) checklist template.
- For Azure Databricks Workspace
 - Create VNET and a pair of Subnets if not exists (spoke-vnet)
 - Enable Azure Active Directory service endpoint on both subnets
 - Create NSG if not exists
 - Associate NSG to Subnets
- Delegate Subnet to Azure Databricks
- Create Azure Databricks Workspace using VNET and Subnets created earlier ([ARM template](#))
 - Note down [DBFS storage account](#) address post workspace deployment
- For Azure Firewall
 - Create Firewall VNET and Subnet (hub-vnet)
 - Deploy Azure Firewall
- Configure Firewall egress rules ([ARM Template](#))
- Peer spoke and hub vnets
- Create a user defined route table
 - Add a default (0.0.0.0/0) route which forwards all of the traffic originating from Azure Databricks subnets to Azure Firewall private ip address
 - Attach route table to Azure Databricks Subnets
- Launch Workspace and create cluster (test deployment)

Further reading

- Understand [ADB security](#)
- Data exfiltration [strategy and guide](#)
- ARM [templates](#) used in this video

github.com/bhavink/adb4u