Thought for 3m 50s

Gotcha—here’s a clean, end-to-end way to connect Azure Data Factory (ADF) to an HDFS cluster privately using a **Managed Private Endpoint (MPE)** to a **Private Link Service (PLS)**. This pattern works whether HDFS runs on Azure VMs or on-prem (reachable via VPN/ExpressRoute).

**What you’ll build (at a glance)**

ADF (Managed VNet IR)

│

└─► Managed Private Endpoint ──► Private Link Service ──► \*Internal\* Standard Load Balancer

└─► HDFS NameNode (WebHDFS TCP : 9870 or 50070)

(via direct backend or via a tiny “forwarder” VM to on-prem)

* ADF talks only over private IPs inside its **managed VNet**.
* PLS fronts an internal **Standard** Load Balancer that sends TCP to your WebHDFS endpoint. (PLS requires Standard LB.) [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)
* ADF’s HDFS connector uses **WebHDFS** (HTTP/HTTPS). Typical ports: Hadoop 3.x → 9870, older Hadoop 2.x → 50070. Use the WebHDFS URL in the linked service. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)
* You’ll approve the MPE request on the PLS side before traffic flows. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint?utm_source=chatgpt.com)

**Prereqs**

* An HDFS cluster with **WebHDFS** enabled (HTTP/HTTPS). Default WebHDFS ports are **9870** (Hadoop 3) or **50070** (Hadoop 2). If using Kerberos/WebHDFS over HTTPS, ensure TLS is set up as usual. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)
* If HDFS is **on-prem**, set up **VPN/ExpressRoute** from an Azure VNet (“provider VNet”) to your on-prem network.
* An ADF instance where you can enable **Managed Virtual Network** + **Managed Private Endpoints**. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)

**Step 1 — Expose WebHDFS behind an *internal* Standard Load Balancer**

1. In your **provider VNet**, create a **Standard Internal Load Balancer** (ILB).
2. Backend choices:
   * **HDFS on Azure VMs**: put the **NameNode VM(s)** in the backend pool, open WebHDFS on the VM(s).
   * **HDFS on-prem**: add a tiny Linux “**forwarder**” VM to the backend pool. It will DNAT traffic to your on-prem NameNode over your VPN/ER.
3. Create:
   * **Health probe** on the WebHDFS TCP port (9870 or 50070).
   * **LB rule**: Frontend :9870 → Backend :9870 (or :50070 if Hadoop 2).

**CLI snippets (examples)**

# Create internal Standard LB

az network lb create \

-g rg-net \

-n hdfs-ilb \

--sku Standard \

--vnet-name vnet-provider \

--subnet subnet-pls \

--frontend-ip-name fe \

--backend-pool-name be \

--private-ip-address 10.10.1.10 \

--frontend-ip-zone 1 2 3

# Health probe on 9870

az network lb probe create \

-g rg-net \

--lb-name hdfs-ilb \

-n hp-9870 \

--protocol Tcp \

--port 9870

# LB rule 9870->9870

az network lb rule create \

-g rg-net \

--lb-name hdfs-ilb \

-n rule-9870 \

--protocol Tcp \

--frontend-port 9870 \

--backend-port 9870 \

--frontend-ip-name fe \

--backend-pool-name be \

--probe-name hp-9870

(Reference quickstarts and CLI commands for internal LBs, probes, and rules.) [Microsoft Learn+2Microsoft Learn+2](https://learn.microsoft.com/en-us/azure/load-balancer/quickstart-load-balancer-standard-internal-cli?utm_source=chatgpt.com)

**If HDFS is on-prem: forwarder VM (DNAT) example**

On the forwarder VM (Linux), enable IP forwarding and DNAT 9870 to your on-prem NameNode:

# enable IP forwarding

sudo sysctl -w net.ipv4.ip\_forward=1

sudo sh -c 'echo "net.ipv4.ip\_forward = 1" >> /etc/sysctl.conf'

# DNAT 9870 -> on-prem NameNode:9870

sudo iptables -t nat -A PREROUTING -p tcp --dport 9870 -j DNAT --to-destination 172.16.10.50:9870

sudo iptables -t nat -A POSTROUTING -p tcp -d 172.16.10.50 --dport 9870 -j MASQUERADE

This is the same pattern Microsoft shows when fronting an on-prem DB via PLS (we’re just using the WebHDFS port instead of SQL). [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/tutorial-managed-virtual-network-on-premise-sql-server)

**Step 2 — Create a Private Link Service (PLS) pointing at the ILB**

Create the PLS in the same subnet as the ILB (PLS requires a Standard LB frontend). Save its **service alias** (you’ll paste it in ADF). [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)

az network private-link-service create \

-g rg-net \

-n hdfs-pls \

--vnet-name vnet-provider \

--subnet subnet-pls \

--lb-name hdfs-ilb \

--lb-frontend-ip-configs fe \

-l eastus

# Get/share the PLS service alias for consumers (ADF will use this)

az network private-link-service show \

-g rg-net \

-n hdfs-pls \

--query "alias" -o tsv

(Official PLS CLI docs / quickstart for LB+PLS.) [Microsoft Learn+1](https://learn.microsoft.com/en-us/cli/azure/network/private-link-service?view=azure-cli-latest&utm_source=chatgpt.com)

Tip: Configure **auto-approval** for your subscription/tenant if you control both sides, or be ready to approve the request manually in **Private Link Center**. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/private-link/manage-private-endpoint?utm_source=chatgpt.com)

**Step 3 — Enable Managed VNet in ADF and create a Managed Private Endpoint**

1. In ADF Studio → **Manage** → **Integration runtimes** → create/enable **Managed Virtual Network** on your IR. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)
2. In ADF Studio → **Manage** → **Security** → **Managed private endpoints** → **+ New** → **Private Link Service**.
3. Paste the **service alias** from Step 2 and submit. The MPE will show **Pending**.
4. Go to the PLS (provider) and **approve** the connection. After approval, ADF will show **Approved** and the MPE will have a **private IP**. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint?utm_source=chatgpt.com)

DNS note (important): ADF Managed VNet doesn’t let you bring your own DNS resolver. The simplest approach is to use the **MPE private IP** directly in your HDFS linked service URL (next step). This avoids any custom DNS plumbing. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)

**Step 4 — Create the HDFS linked service (via WebHDFS over the MPE)**

In ADF Studio → **Manage** → **Linked services** → **HDFS**:

* **URL**: http://<MPE-PRIVATE-IP>:9870/webhdfs/v1/  
  (Use :50070 for Hadoop 2.x, or https://... if you front WebHDFS with TLS.)
* **Authentication**: choose what your cluster uses (Anonymous/Basic/Windows/Kerberos).
* **Connect via Integration Runtime**: choose your **Managed VNet IR**. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)

**Example (JSON)**

{

"name": "ls\_hdfs\_webhdfs",

"type": "LinkedService",

"properties": {

"type": "Hdfs",

"typeProperties": {

"url": "http://10.20.0.5:9870/webhdfs/v1/",

"authenticationType": "Anonymous"

},

"connectVia": { "referenceName": "AutoResolveIntegrationRuntime", "type": "IntegrationRuntimeReference" }

}

}

Test connection should succeed only **after** the MPE shows **Approved**. If you use Kerberos or Basic, supply the relevant properties/certificates supported by the HDFS connector. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)

**Step 5 — Use it in a pipeline (quick check)**

Create a **Copy activity**:

* Source: HDFS linked service, a path like /data/incoming/.
* Sink: Azure Blob/ADLS Gen2 linked service (private, ideally via its own MPE).  
  Run a pipeline and verify data moves over private IPs only. (ADF MPEs route traffic within Microsoft’s backbone.) [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)

**Variations & tips**

* **HDFS HA (2 NameNodes)**:  
  Put both NameNodes in the ILB backend pool. WebHDFS on the standby typically redirects; ensure your health probe targets a port that reflects “active” health or use a small reverse proxy that follows redirects. Keep the LB rule on the WebHDFS TCP port. (General pattern follows the same PLS → ILB approach.) [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)
* **HTTPS/TLS**:  
  If you terminate TLS on the NameNode(s), expose **:9871** (or your chosen port) on the ILB/PLS and set the ADF URL to https://<MPE-IP>:9871/webhdfs/v1/. The HDFS connector supports HTTPS. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)
* **On-prem routing**:  
  The forwarder VM approach is a Microsoft-documented pattern for PLS to on-prem (they show it for SQL; the idea is identical here with WebHDFS). Keep firewall rules open from the forwarder to your on-prem NameNode WebHDFS port. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/tutorial-managed-virtual-network-on-premise-sql-server)
* **Why not SHIR?**  
  You can avoid deploying a Self-Hosted IR by using **Managed VNet + MPE + PLS**—supported for WebHDFS/HDFS. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)
* **Security**:  
  Use NSGs on the ILB subnet; allow the health probe and WebHDFS port; least-privileged on the forwarder VM; restrict PLS visibility/auto-approve to your subscriptions. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)

**Quick checklist (troubleshooting)**

* PLS shows **Healthy** and bound to the **Standard ILB** frontend. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)
* MPE in ADF shows **Approved**. If not, approve in **Private Link Center**. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/private-link/manage-private-endpoint?utm_source=chatgpt.com)
* ILB **Health probe** is green (verify port 9870/50070 reachable from ILB to the backend). [Microsoft Learn](https://learn.microsoft.com/en-us/azure/private-link/create-private-link-service-cli?utm_source=chatgpt.com)
* From the **forwarder VM** (if used), curl http://<onprem-nn>:9870/webhdfs/v1/?op=LISTSTATUS succeeds.
* In ADF, the HDFS **linked service URL** uses **MPE private IP** and correct port/path. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)

**References**

* **ADF managed VNet & Managed Private Endpoints** (concepts, approvals). [Microsoft Learn+1](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)
* **Private Link Service** (requires Standard LB, creation steps). [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/managed-virtual-network-private-endpoint)
* **HDFS connector (WebHDFS)**: protocol, ports, auth, URL shape. [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/connector-hdfs)
* **CLI** for LB + probe + rule + PLS creation. [Microsoft Learn+3Microsoft Learn+3Microsoft Learn+3](https://learn.microsoft.com/en-us/azure/load-balancer/quickstart-load-balancer-standard-internal-cli?utm_source=chatgpt.com)
* **On-prem via PLS pattern** (tutorial uses SQL; same flow with your WebHDFS port). [Microsoft Learn](https://learn.microsoft.com/en-us/azure/data-factory/tutorial-managed-virtual-network-on-premise-sql-server)