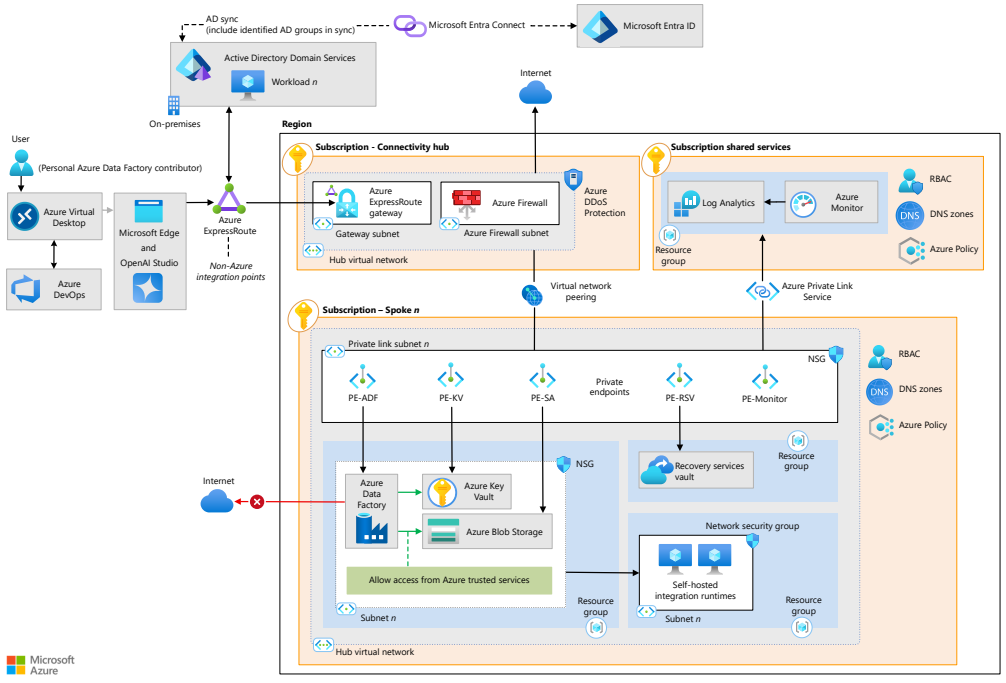
ARCHITECTURE GUIDE

Azure Databricks Zero-Trust Blueprint

Enterprise Network Isolation & Exfiltration Defense

Executive Summary: Standard VNet injection is insufficient for regulated industries. This blueprint details the exact configurations, routing decisions, and networking anti-patterns required to establish strict data residency, eliminate public inbound/outbound paths, and prevent identity-driven exfiltration across the Azure Data platform.

The Golden State Architecture



Hard Design Decisions (The "Why")

Decision 1: Bypassing the NVA for Storage Traffic

The Trade-off: Security teams typically mandate all inter-spoke traffic routes through a Hub firewall (Azure Firewall Premium).

The Reality: Azure Firewall Premium costs roughly $0.016/GB. Databricks shuffle operations can hit 100TB/day. Forcing internal Spoke-to-Spoke storage traffic through the Hub Firewall adds severe latency and catastrophic cost ($1,600+/day) for zero security benefit, as the traffic is already on the encrypted Microsoft backbone.

Implementation: Apply User Defined Routes (UDRs) directly to the Databricks subnets routing `10.x.x.x` (Storage Private Endpoints) as `NextHopType: VirtualNetwork`. Keep the traffic local.

Decision 2: Centralized Private DNS Resolution

The Reality: Spoke-local Private DNS zones create unmanageable "split-brain" routing where on-premise users or Hub VMs cannot accurately resolve the private endpoints inside the Spoke.

Implementation: Deploy the **Azure DNS Private Resolver** in the Hub VNet. Link central Private DNS Zones *only* to the Hub, and configure conditional forwarders from your corporate Active Directory to point to the Resolver's inbound endpoint.

Decision 3: Dedicated Browser Authentication Workspace

The Reality: Databricks uses a global authentication mechanism per region. Relying on the primary data-processing workspace for Auth means if that specific VNet experiences a routing failure, *all* Databricks workspaces in that entire Azure region will fail to authenticate users.

Implementation: Provision a dedicated, empty "Auth" Databricks workspace with its own Private Endpoint explicitly for SSO token exchange.

Mandatory Assumptions & Prerequisites

1. Identity & Governance Matrix

The organization must use **Entra ID (Azure AD)** paired with **Unity Catalog**. Relying on workspace-local groups, legacy Hive Metastore ACLs, or long-lived Service Principal secrets is incompatible with this lockdown design. All cluster access must utilize Azure Managed Identities.

2. IP Address Space Allocation

The organization has allocated a minimum of a `/24` subnet for the Databricks Public (Host) subnet and a `/24` for the Private (Container) subnet to handle node scale-out during burst processing. Attempting to cram these into a `/26` will cause job failures during peak load.

The Anti-Pattern Graveyard

Anti-Pattern 1: The "Lazy" Service Endpoint

What it is: Using VNet Service Endpoints to secure ADLS Gen2 instead of Azure Private Link.

Why it fails: Service Endpoints keep the storage account resolving to a public IP (it merely ACLs access to your VNet). This means data still transverses public endpoints (failing strict data-residency compliance audits) and does *not* prevent data exfiltration to other generic Azure storage accounts in the same region.

Anti-Pattern 2: Default Route (0.0.0.0/0) Forced Tunneling Without Exceptions

What it is: Forcing all outbound traffic from the Databricks subnets back on-premises or to a Hub firewall blindly.

Why it fails: Databricks control-plane traffic (Secure Cluster Connectivity relay, Azure Storage metrics) will be dropped by on-premise asymmetric routing or stateful firewalls, preventing clusters from ever starting. You *must* use Service Tags (e.g., `AzureDatabricks`, `Sql`, `Storage`) in your UDRs to allow required Azure control traffic to bypass the default route.

Anti-Pattern 3: Disabling Public Access Before DNS Validation

What it is: Flipping the storage firewall to "Deny All" immediately upon creating the Private Endpoint.

Why it fails: DNS propagation up to corporate forwarders takes time. If the Private DNS zone hasn't fully synced to your on-premise network, locking down the storage account public endpoint will instantly break production pipelines that are still resolving the public IP.

Use this blueprint to accelerate your cloud security review.