Databricks Workspaces with Front End Private Link

This page covers the operational layout for using Databricks Workspaces with Private Access enabled within the Jemena environment.

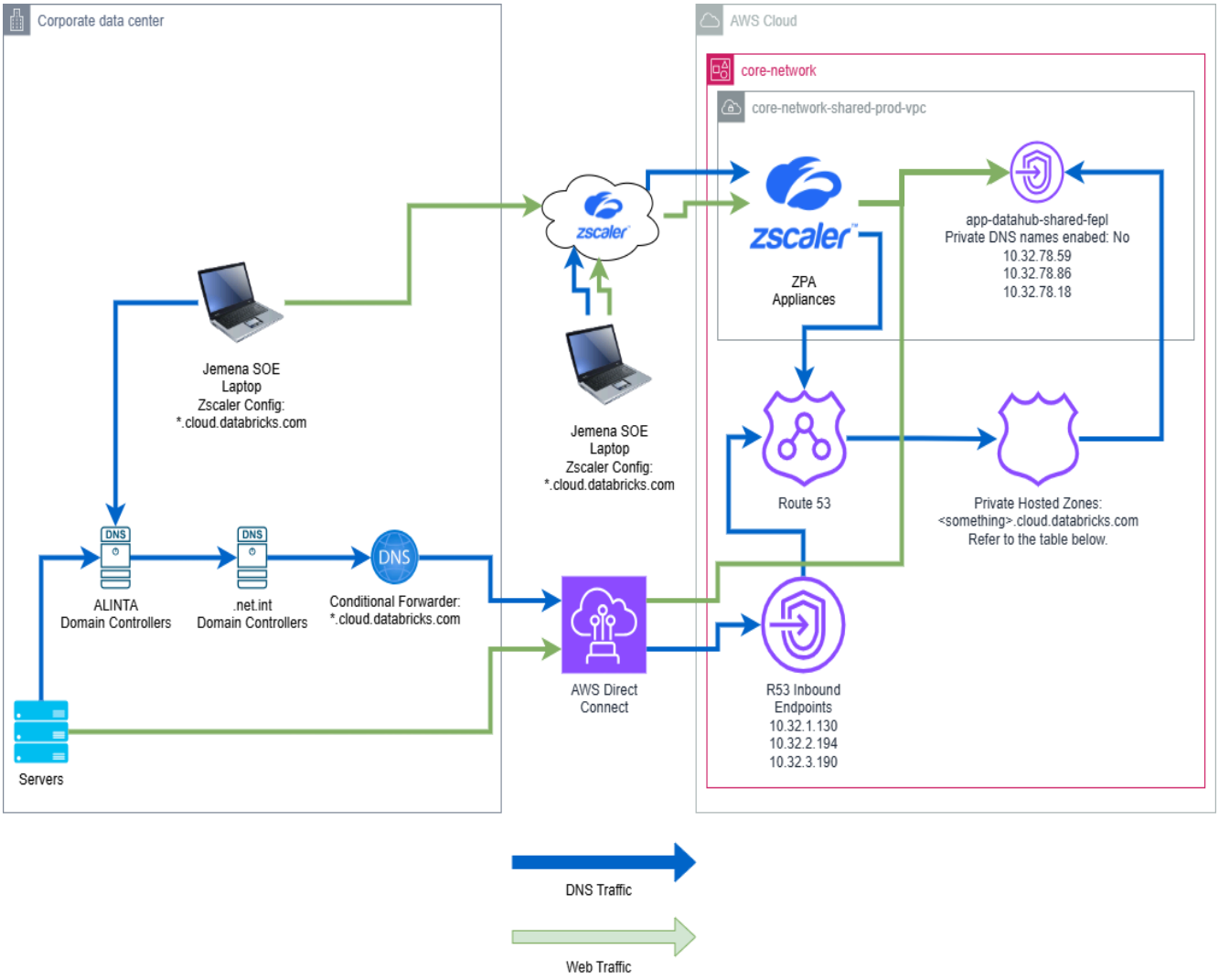
TL;DR: To publish a newly created Databricks Workspace to Jemena Laptops and Servers it needs to be added as a Private Hosted Zone in Route 53. The code repository to manage the Private Hosted Zones can be found here:

https://gitlab.com/jemena/platforms/core-network-databricks-workspaces-phz Connect your GitLab account

Restricting access to Databricks Workspaces from only Jemena corporate network and assets is a Security Requirement. Further information on Private Link enablement for Jemena's Databricks Instance can be found here.

Page Contents

Network Configuration & Flow



There are three system components where configuration has been implemented:

1. **Route 53** **Private Hosted Zones** - these zones direct requests to the Databricks Workspace addresses to the Front End Private Link
2. **Zscaler** - a rule has been put in place to "catch" requests for `*.cloud.databricks.com` and forward them through the Zscaler Appliances located in AWS
3. **On Prem DNS** - a conditional forwarder has been configured on the **.net.int Domain Controllers** to "catch" requests for `*.cloud.databricks.com` and forward those requests to Route 53 via the Inbound Endpoints, which will return the Front End Private Link private IP addresses for internal servers to connect to

Further information can be found below.

> The implemented solution includes several technical workarounds. See the Issues section below.

Private Hosted Zones

Private Hosted Zones have been created for each Databricks Workspace address. In each PHZ an alias A record is created that points to the Front End Private Link VPC Endpoint.

⚠ These Private Hosted Zones in R53 are managed by the code repository found in the link below. Refer to the repository for the up to date list of deployed zones.

Use the repository to **add or remove** hosted zones.

https://gitlab.com/jemena/platforms/core-network-databricks-workspaces-phz Connect your GitLab account

Workspace Address	Workspace Name
`dbc-fecbb5ff-7592.cloud.databricks.com`	digital-field-workspace-nonprod
`dbc-eaba2339-eb1e.cloud.databricks.com`	digital-lab-workspace-nonprod
`dbc-94129c9d-8f32.cloud.databricks.com`	elec-network-field-workspace-nonprod
`dbc-de6c0ca1-0e35.cloud.databricks.com`	elec-network-lab-workspace-nonprod
https://jemena-digital-field.cloud.databricks.com	digital-field-workspace-prod
https://jemena-digital-lab.cloud.databricks.com	digital-lab-workspace-prod
https://jemena-elec-network-field.cloud.databricks.com	elec-network-field-workspace-prod

https://jemena-elec-network-lab.cloud.databricks.com	elec-network-lab-workspace-prod

Zscaler

Zscaler operates on all Jemena SOE endpoints (Laptops) and is responsible for directing traffic intended for Jemena's internal network through its tunnel.

Configuration has been added to Zscaler clients to catch requests for `*.cloud.databricks.com` and direct them through the Zscaler appliances in AWS.

When a Databricks Workspace address has been configured as a PHZ in Route 53 the Zscaler appliances will look up and connect to the Workspace through the Front End Private Link.

Any other requests for `*.cloud.databricks.com` will be directed to the Internet through the Zscaler appliances.

On Prem DNS

For on prem servers to be able to connect to Databricks Workspaces they need to know what the IP addresses of the Front End Private Link endpoint are.

A conditional forwarder has been added to the **.net.int Domain Controllers** to catch DNS requests for `*.cloud.databricks.com` and forward them to the AWS Route 53 Inbound Endpoints.

When a Databricks Workspace address has been configured as a PHZ in Route 53 the private IP addresses for the Front End Private Link are returned in the DNS request. The requesting server will then connect to the Databricks Workspace using a private IP address over our Direct Connect private link.

Any other DNS requests for `*.cloud.databricks.com` will be directed to the Internet through Route 53.

Preferred (but not implemented) DNS Configuration

When a Databricks Workspace has Private Access enabled Databricks will update the DNS record for that workspace address to CNAME `sydney.privatelink.cloud.databricks.com`. It is then up to the Databricks Customers to configure their DNS system to direct requests to their Private Link Endpoint.

When a Databricks Private VPC endpoint is deployed in AWS, and that endpoint has Private DNS names enabled, it is automatically given DNS name `sydney.privatelink.cloud.databricks.com`. Any systems deployed within the same VPC will automatically resolve `sydney.privatelink.cloud.databricks.com` to the private IP addresses of that endpoint.

For on prem servers to resolve `sydney.privatelink.cloud.databricks.com` to the VPC endpoint IP addresses the recommended approach is to set up a conditional forwarder on the on prem DNS servers. The conditional forwarder will catch DNS requests for `sydney.privatelink.cloud.databricks.com` and forward them to the AWS Route 53 Inbound endpoints, which in turn will return the private IP addresses for the VPC endpoint.

An in-depth look at this architecture can be found here:

Configuring DNS resolution for Private Databricks Workspaces (AWS)

This approach has the advantage of directing access to Databricks Workspaces that are configured for Private Access automatically through the VPC endpoint. No further configuration is required.

> There were multiple technical issues encountered when trying to implement this preferred DNS solution. Refer to the next section.

Issues

The following section details the list of technical issues that were encountered during deployment.

DNS conditional forwarding on Active Directory

A conditional forwarder was put in place to forward DNS requests for `sydney.privatelink.cloud.databricks.com` to our Route 53 Inbound Endpoints. This appeared to work initially, however we soon discovered that the DNS query result would flip between returning the correct result (the endpoint private IP addresses) and the internet based IP addresses (incorrect).

This meant that the end user would get inconsistent results when trying to log into a Databricks Workspace with Private Access enabled.

There is an ongoing investigation with Microsoft and Databricks into this issue.

WORKAROUND: the conditional forwarder is set to `*.cloud.databricks.com`. This means all DNS requests for Databricks are sent through to Route 53.

Front End & Back End Endpoint Conflict

When you create a Databricks VPC endpoint with **Private DNS names** enabled it will automatically set the Private DNS name to `sydney.privatelink.cloud.databricks.com`.

You cannot create a 2nd Databricks VPC endpoint in the same VPC with Private DNS names enabled as this results in a conflict - both endpoints can't use the same private DNS name.

In accordance with the Databricks Private Link Implementation the VPC endpoint for the "Back End" private link and the "Front End" private link must be separate. As the "Back End" VPC endpoint had been created first and is critical for Databricks Cluster functionality the options for creating the "Front End" VPC endpoint are:

1. Create the endpoint in the same VPC without Private DNS names enabled, or
2. Create the endpoint in a different VPC with Private DNS names enabled

Option 1 is what was implemented (see WORKAROUND below for more detail).

Option 2 would present further technical challenges:

- Do we have another VPC in our infrastructure suitable?
- If not, how much effort is there to stand up another VPC just for this purpose?
- How do we direct Zscaler requests through to a different VPC?

WORKAROUND: the Front End VPC endpoint was created without Private DNS names enabled. For DNS requests to the Databricks Workspace addresses to be answered correctly a Private Hosted Zone has been created for reach Workspace address. In each PHZ, an Alias A record has been created that points to the Front End VPC endpoint.

Zscaler with CNAMES

Zscaler was initially configured to catch requests for `sydney.privatelink.cloud.databricks.com` and direct them through the Zscaler Appliances. However, when you would request the workspace via the workspace address the CNAME wouldn't be resolved via Zscaler and the public IP addresses were always returned.

WORKAROUND: Zscaler configuration now catches `*.cloud.databricks.com` to direct all requests through Zscaler (and thus the Jemena network).