Accessing private resources with Oracle Digital Assistant.

With the introduction of Oracle Digital Assistant release 23.04 it is now possible for a Digital Assistant instance to access resources which are not publicly accessible from the Internet. Such resources can be located in an Oracle Cloud Infrastructure Virtual Cloud Network (VCN) or running in an on-premise environment or within other data centers.

Like other public services running in Oracle Cloud Infrastructure, Oracle Digital Assistant now supports the private endpoint feature. At this moment Digital Assistant provides outbound connectivity to private resources. However, unlike other services supporting the private endpoint feature, exposing a Digital Assistant instance itself as the private resource in a VCN for inbound calls is not yet supported.

About ODA Private Endpoints

An ODA private endpoint is a private resource provisioned within your VCN and represents the Oracle Digital Assistant in this VCN. The Digital Assistant service sets up the private endpoint in a subnet of your choice within the VCN.

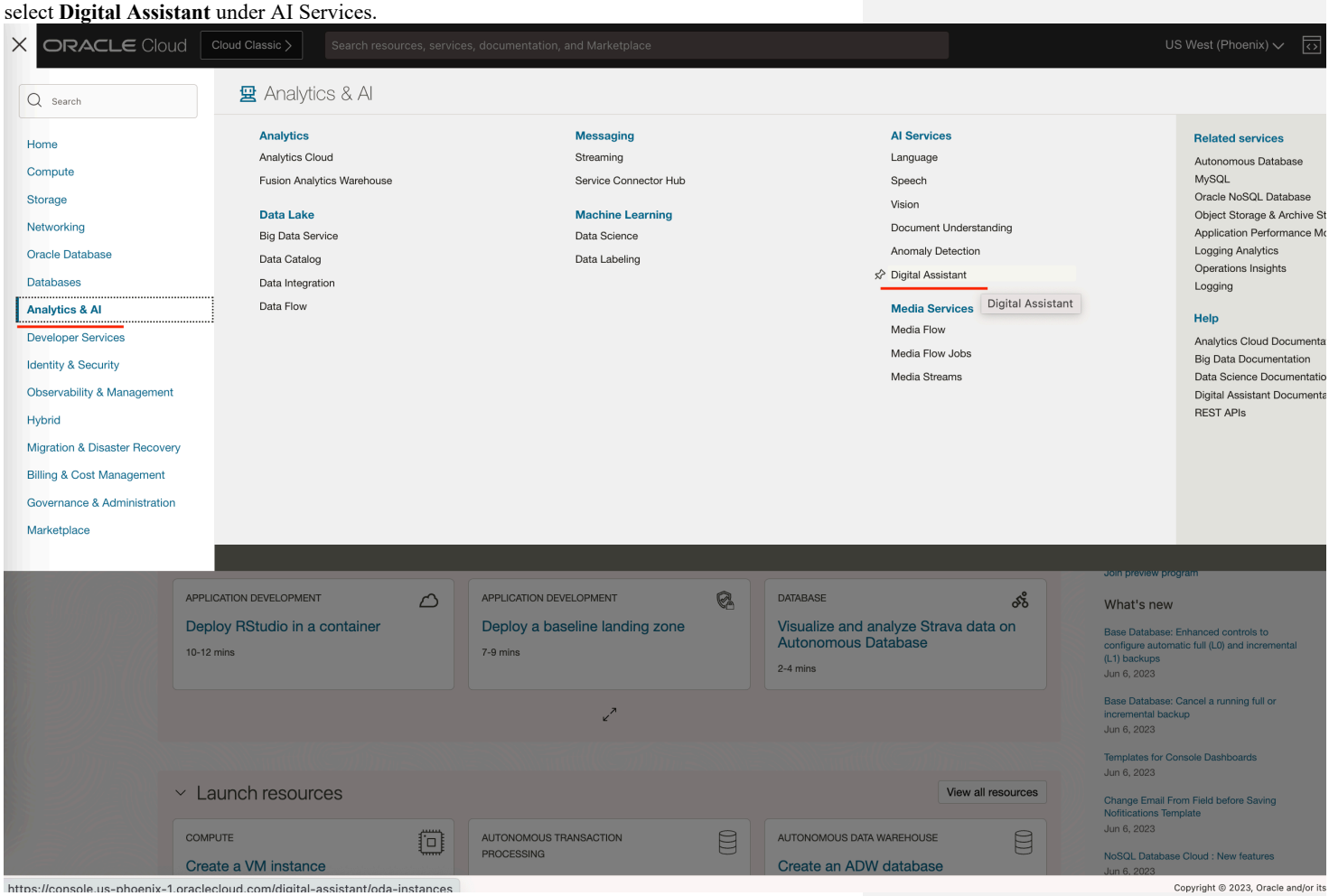
You can think of the private endpoint as just another VNIC in your VCN.
You can control the access to it like you would for any other VNIC: by using security rules.
However, the service sets up this VNIC and maintains its availability on your behalf. You only need to maintain the subnet and the security rules.
ODA Private Endpoint also supports specifying Network Security Groups when creating private endpoint.

Currently the Digital Assistant service supports private endpoint access to on-premises databases or databases running in an Oracle Cloud Infrastructure VCN when you use SQL Dialog skills.
You can also use a private endpoint to connect to REST services running on-premises or within Oracle Cloud Infrastructure VCN using built-in Call REST Service component.

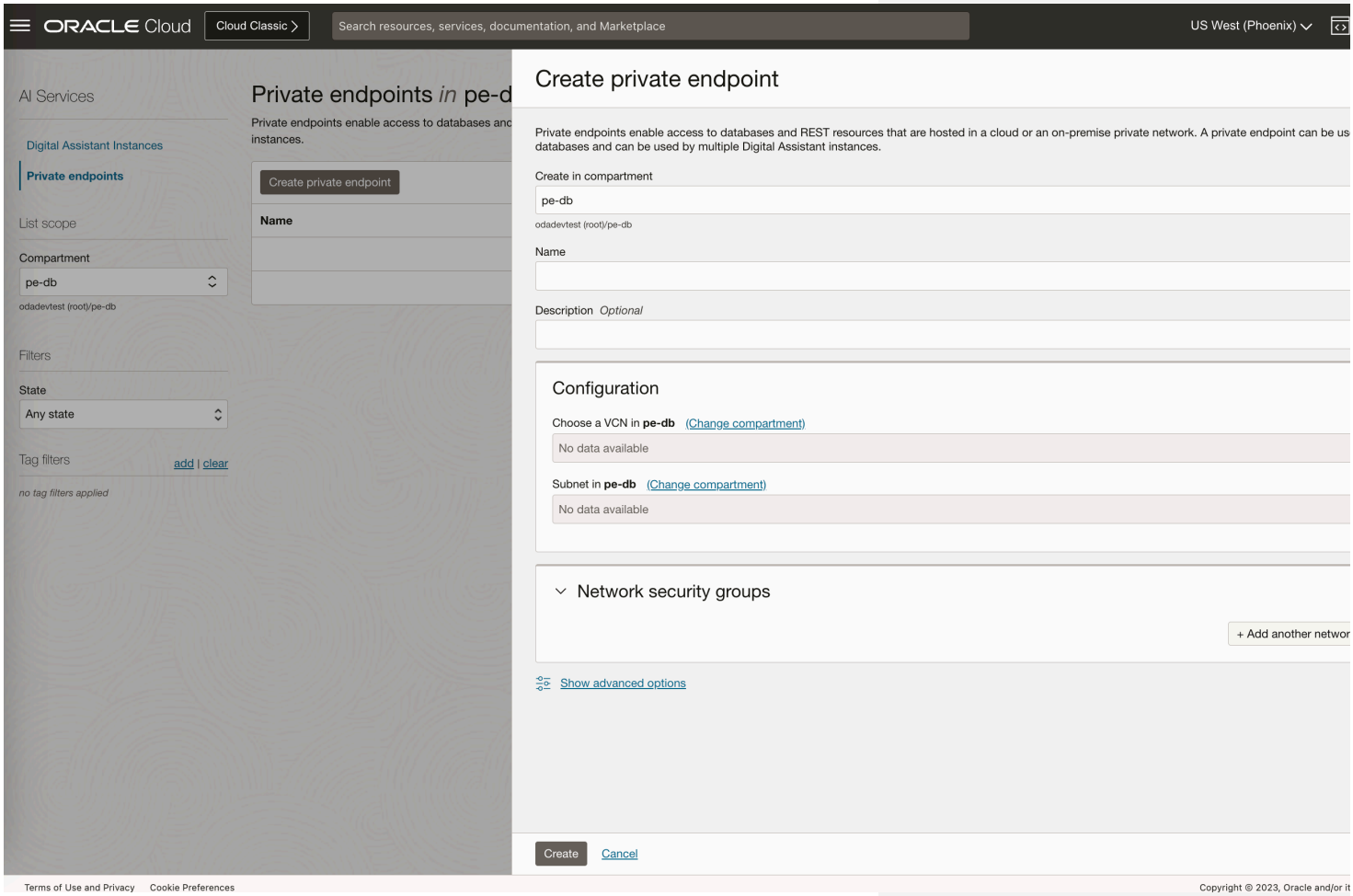
You can create an ODA Private Endpoint, associate it to one or multiple ODA instances and create scan proxies by means of any of the following:
- OCI Console
- One of the available OCI SDKs - there are SDKs for Java, Python, and a few other languages
- OCI CLI (Command Line Interface)
- Terraform using OCI Provider

Creating an ODA Private Endpoint using OCI Console

To create an ODA private endpoint, open the hamburger menu, select **Analytics & AI**, then select **Digital Assistant** under AI Services.

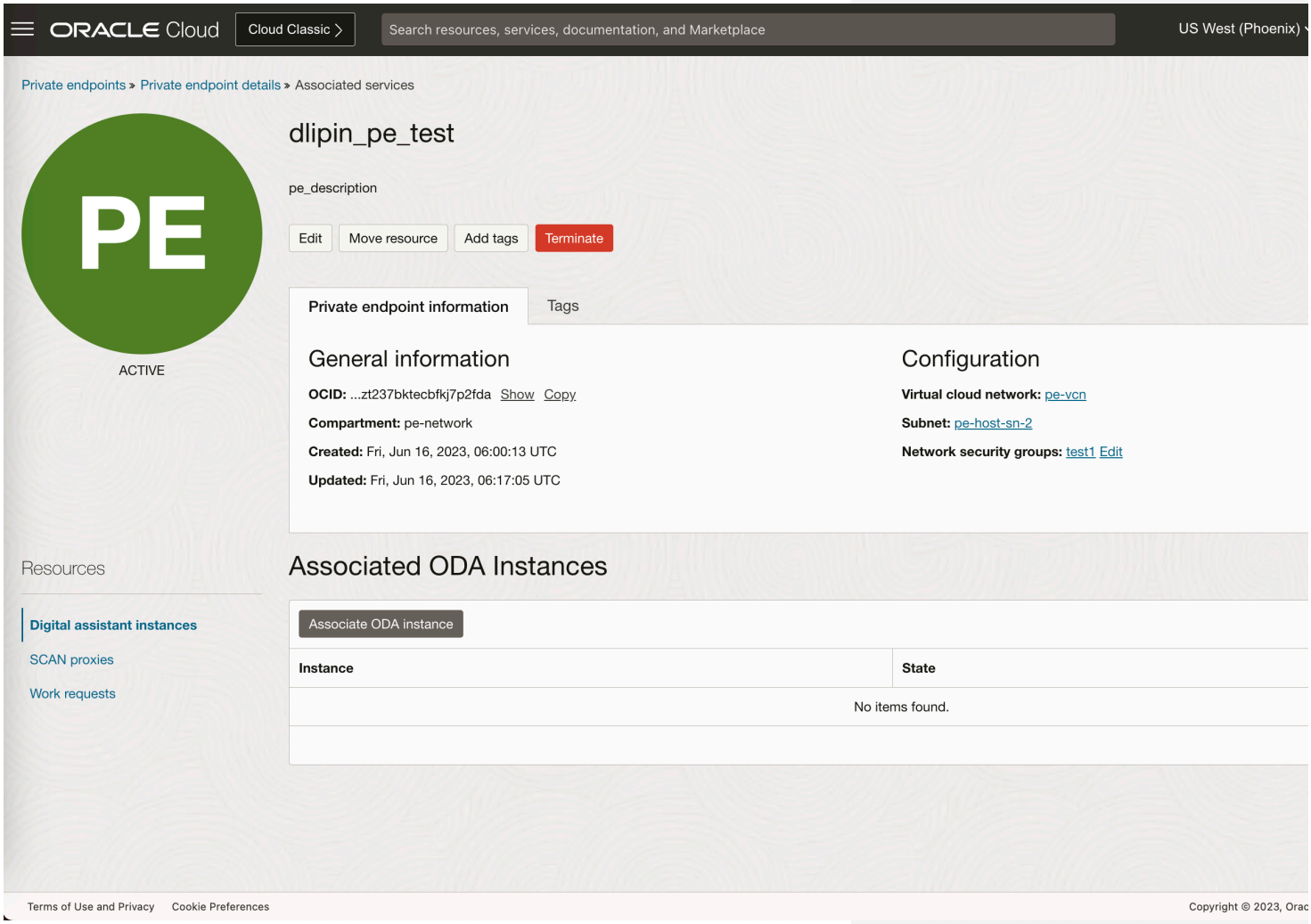


Once the Digital Assistant Instances page is opened, click “Private endpoints” and select “Create private endpoint”:

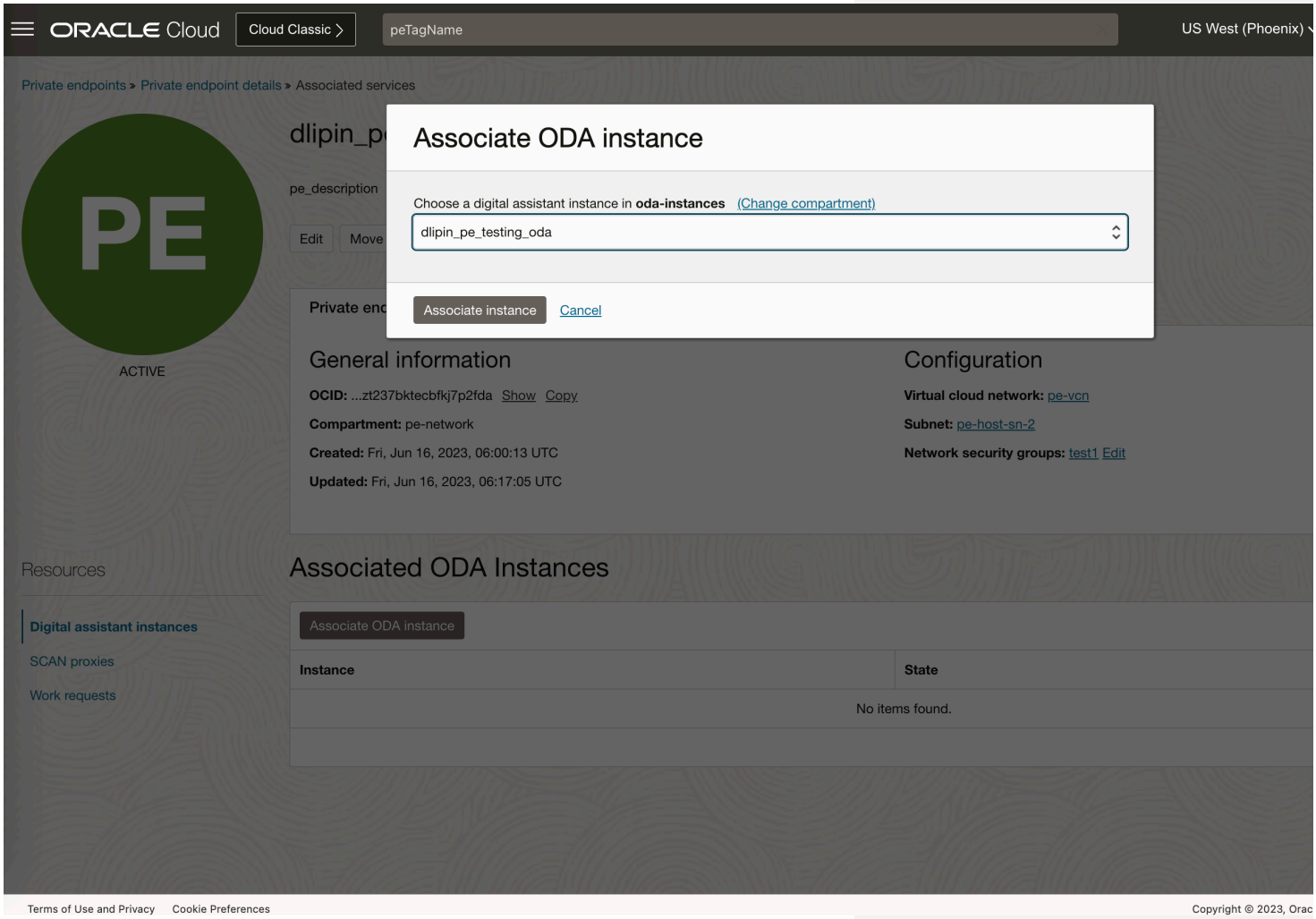


In the opened dialog you will need to select the compartment in which the ODA private endpoint should be managed, the virtual cloud network (VCN) and subnet in which the private endpoint should be created and the name of the private endpoint. Optionally you can specify the description, network security groups and tags.

Once the ODA private endpoint is created, you can view its details.

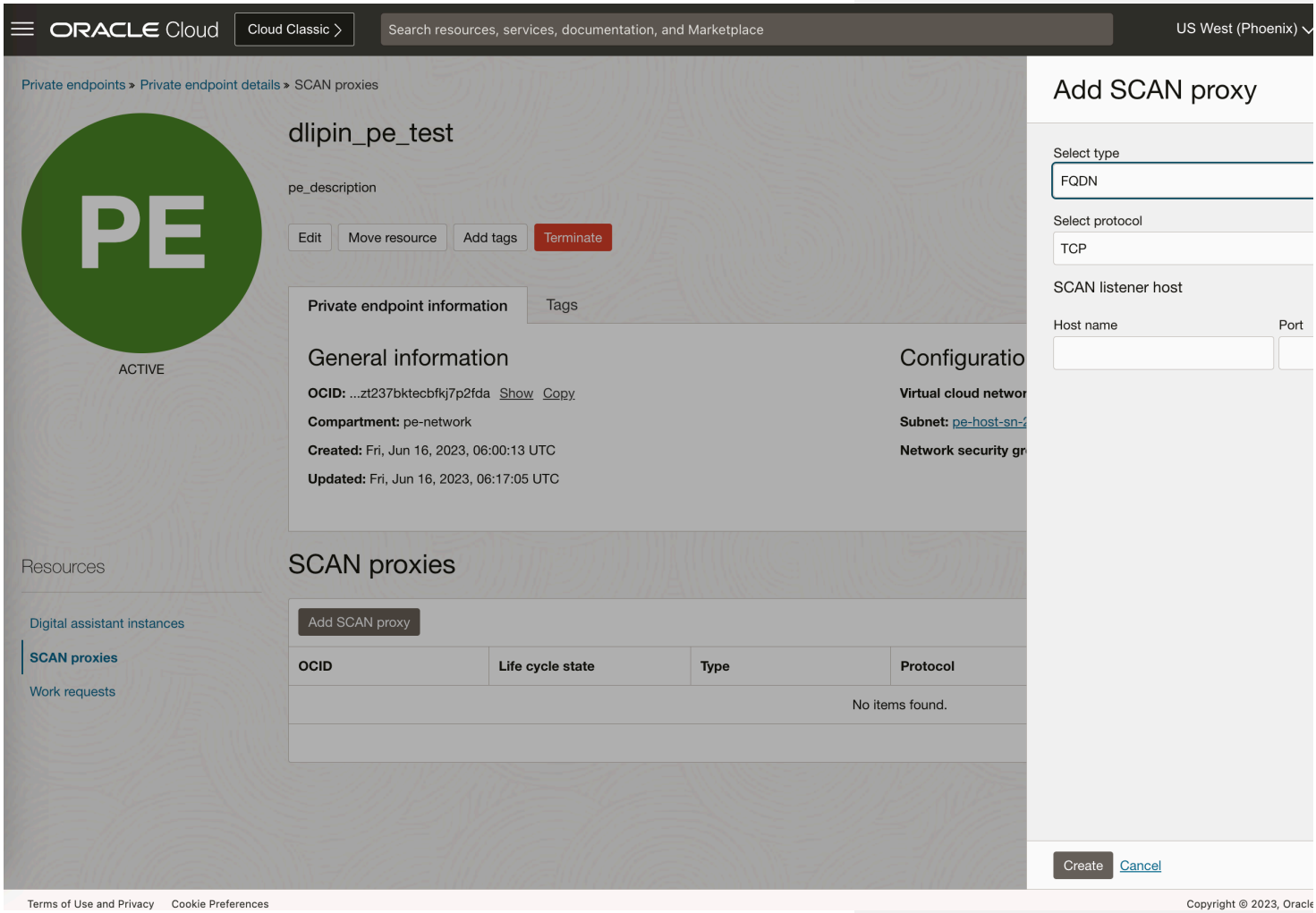


In order to enable an ODA private endpoint to be used in Digital Assistant, you'll need to associate this private endpoint with the Digital Assistant instance by clicking the "Associate ODA instance" button:



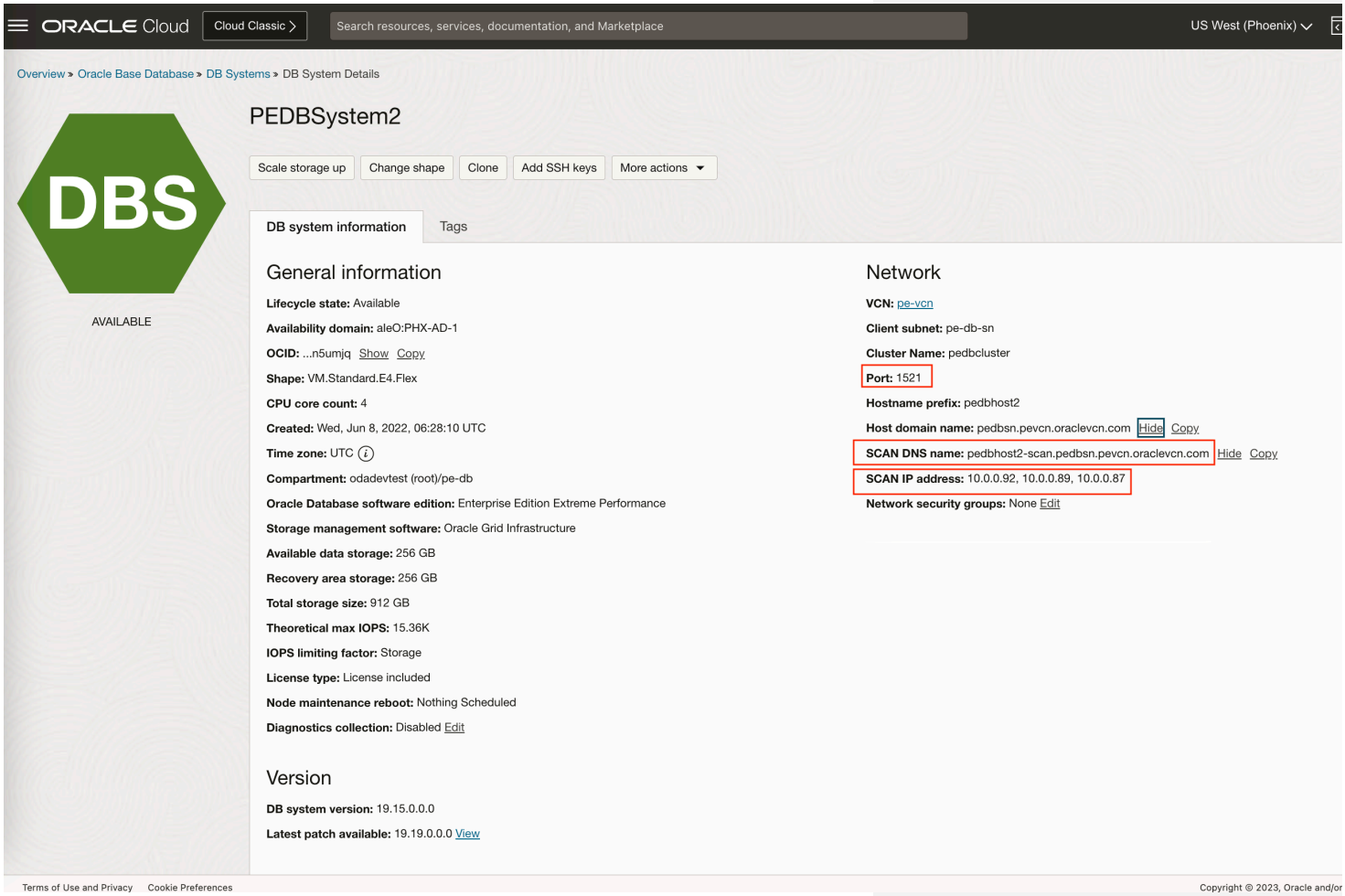
SCAN Proxies
If you plan to use SQL Dialogs Skills to connect to a RAC-enabled database using the private endpoint, you will need to configure a SCAN proxy for this database pointing to the database SCAN Listener.

Open the created private endpoint, click on “SCAN proxies” in the left panel under the Resources section and then click “Add SCAN proxy” and provide the required details.



When creating a new SCAN proxy, you can either select the FQDN type, if you have the DNS name of SCAN Listener configured for your RAC-enabled database, or the IP type, if you have the list of IPs for your SCAN Listener. If your SCAN listener has an associated DNS name, then it is recommended to use the FQDN type when creating the SCAN Proxy and later when you are registering the data service in Digital Assistant; otherwise, you can register IP addresses in the SCAN proxy and use those addresses when configuring the data service.

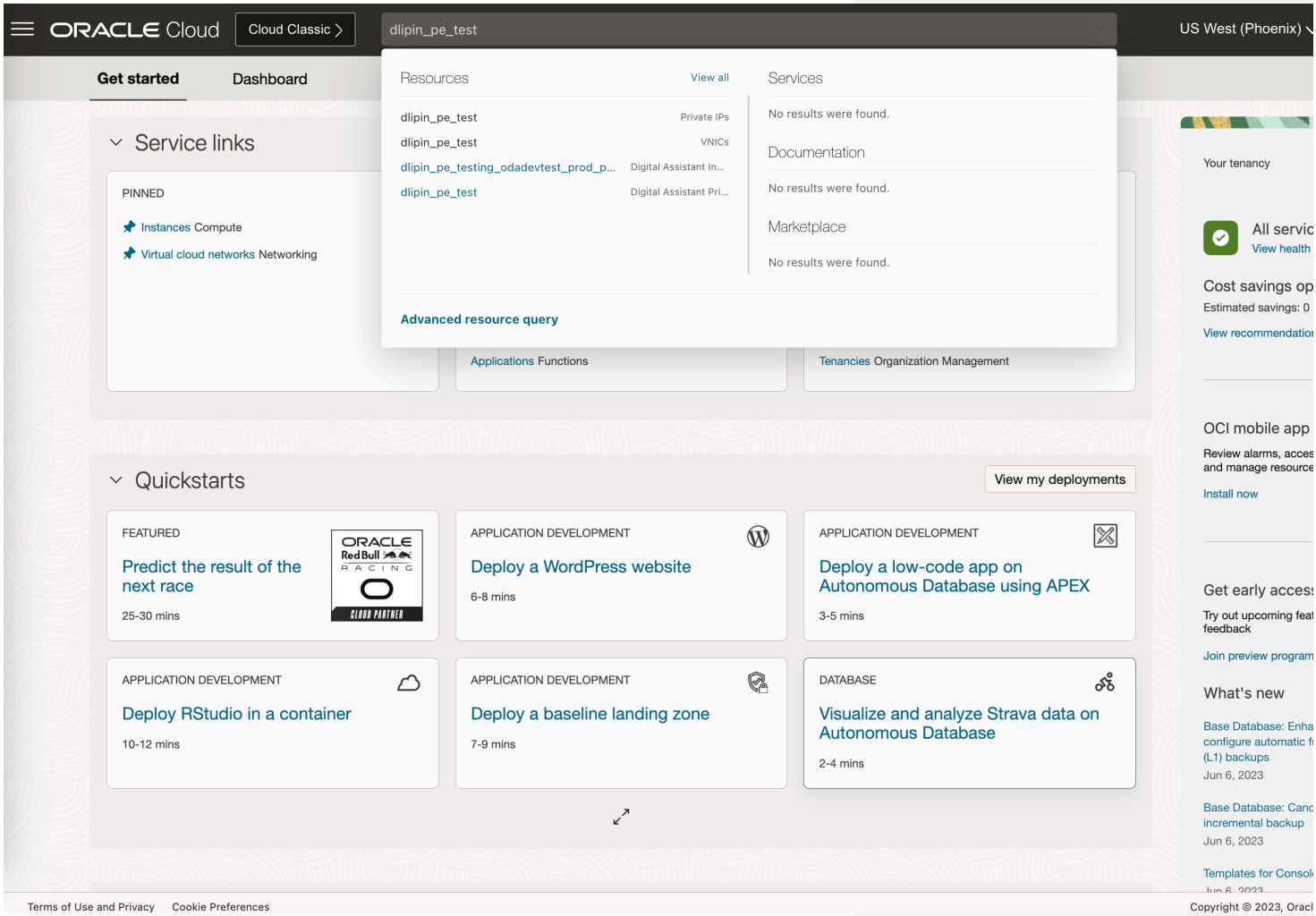
If you use Oracle Base Database provisioned in Oracle Cloud Infrastructure, you can find the FQDN name and port in the OCI Console UI by searching for the “SCAN DNS name” value. You would need to specify this value as the Hostname while creating the SCAN proxy and, later on, the data service in Digital Assistant.



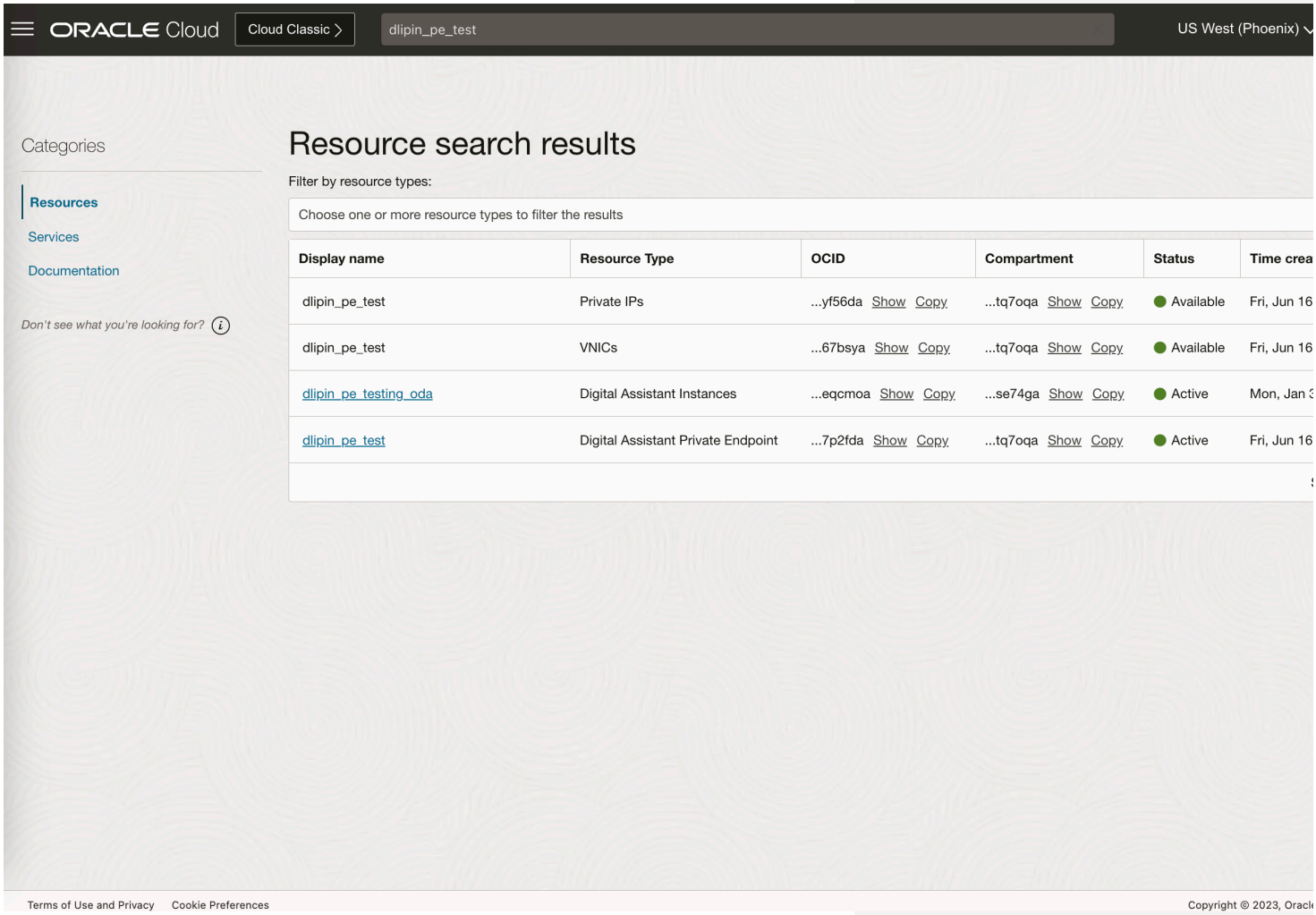
Connecting to an ATP/ADW instance on shared infrastructure does not require a SCAN proxy to be created. It is also not required when you connect to a single-node database which does not have a SCAN listener.

Search for ODA Private Endpoints

You can always search for ODA private endpoints using the OCI console's search. Currently you can search by OCID, name, description, OCIDs of network security groups, OCID of subnet, SCAN proxy OCID, SCAN proxy FQDN/IP Addresses and Port, tag names and values.



If you click "View All" in that search box, you'll be able see the details about all search results:



Identifying ODA Private Endpoint details

An ODA private endpoint is represented as a VNIC with private IPs in the subnet where you create it.

To create an ODA private endpoint, you need at least 3 available IPs in the subnet that you selected. If less than 3 IPs are available, then your provisioning will fail.

To find out how many private IPs are already used in the subnet, you can run the following command using the CLI (assuming you have the OCI CLI and the jq tool already installed):

```
oci network private-ip list --subnet-id ocid1.subnet.oc1.phx.xxx --all | jq '.data | length'
```

To find out what is the configured the CIDR range for the subnet, you can run the following command using the CLI (or you can also find this information in OCI Console):

```
oci network subnet get --subnet-id ocid1.subnet.oc1.phx.aaabbbcc  | jq -r '.data."cidr-block"'
```

Once you have identified the CIDR range, you can find out how many IP addresses are in this range. For reference, here is a CIDR calculator: http://www.ipaddressguide.com/cidr

The total number of used private IPs plus 3 should be less than the total number of IPs in the CIDR range.

If you have less than 3 IPs available in the existing subnet, then the solution is either to set up a new subnet, or to expand the CIDR range for the existing subnet (make sure there's no intersection with other subnets in the same VCN) or remove unused resources consuming private IPs in that subnet.

Once the ODA private endpoint is created, it will consume 3 IP addresses in your subnet. You might need those IPs to set up narrow security lists. How can you find out those addresses?

Currently IPs consumed by ODA private endpoint are not returned in its metadata, but you can figure them out if you run the following command:
oci network private-ip --subnet-id <subnet-ocid> --all

For each private endpoint there would be 3 private IPs in the output of that command (maybe among other IPs not related to the ODA private endpoint). They will be created very close in time – usually within a few minutes of each other.

Here's an example of this command output:


```
{
  "availability-domain": "aIeO:PHX-AD-2",
  "compartment-id": "ocid1.compartment.oc1..aaabbbccc",
  "defined-tags": {},
  "display-name": "test_oda_pe",
  "freeform-tags": {},
  "hostname-label": "test_oda_pe",
  "id": "ocid1.privateip.oc1.phx.abcabcabc",
  "ip-address": "10.0.0.44",
  "is-primary": true,
  "subnet-id": "ocid1.subnet.oc1.phx.aabbcc",
  "time-created": "2023-04-20T21:31:51.126000+00:00",
  "vlan-id": null,
```

```
  "vnic-id": "ocid1.vnic.oc1.phx.aaaaaaaaa"
 },
 {
  "availability-domain": null,
  "compartment-id": "ocid1.compartment.oc1.. aaabbbccc ",
  "defined-tags": {
   "Oracle-Tags": {
    "CreatedBy": "rce",
    "CreatedOn": "2023-04-20T21:32:46.073Z"
   }
  },
  "display-name": "RCE_PE_SEC_IP",
  "freeform-tags": {},
  "hostname-label": null,
  "id": "ocid1.privateip.oc1.phx.bcabcabca",
  "ip-address": "10.0.0.57",
  "is-primary": false,
  "subnet-id": "ocid1.subnet.oc1.phx.aabbcc",
  "time-created": "2023-04-20T21:32:46.113000+00:00",
  "vlan-id": null,
  "vnic-id": "ocid1.vnic.oc1.phx.aaaaaaaaa"
 },
 {
  "availability-domain": null,
  "compartment-id": "ocid1.compartment.oc1.. aaabbbcccx",
  "defined-tags": {},
  "display-name": "ReverseConnectionIp",
  "freeform-tags": {},
  "hostname-label": null,
  "id": "ocid1.privateip.oc1.phx.cbacbacba",
  "ip-address": "10.0.0.58",
  "is-primary": false,
  "subnet-id": "ocid1.subnet.oc1.phx.aabbcc",
  "time-created": "2023-04-20T21:32:17.131000+00:00",
  "vlan-id": null,
  "vnic-id": "ocid1.vnic.oc1.phx.aaaaaaaaa"
 }
```


And here's what to look for in that output:

1) One object in that array will have the same **display-name** value as the name of the ODA private endpoint you created. In the example above it has id "ocid1.privateip.oc1.phx.abcabcabc" and IP "10.0.0.44". Currently this IP is not used – it is reserved for future use.
2) Another one will be with display name ReverseConnectionIp. In the example above it has the id "ocid1.privateip.oc1.phx.cbacbacba" and IP "10.0.0.58".

This IP address is used as the source of the traffic coming from ODA instance into the subnet in which you created this private endpoint. You can use this IP address in your ingress/egress rule as a source/destination address.
3) The third one has the “RCE_PE_SEC_IP” as its display name. In the example above it has the ID “ocid1.privateip.oc1.phx.bcabcabca” and IP “10.0.0.57”
This IP address is used to originate DNS traffic from the ODA instance into your VCN.

Troubleshooting

Sometimes your request to create an ODA private endpoint or SCAN proxy can fail.
How can you debug why did it fail and what needs to be fixed?

First of all, you need to identify the OCID of the work request which has failed. You can do this either in the OCI console (by opening the Work Requests tab under the Resources panel) or using the OCI CLI. The following command will return the list of all work requests submitted for the given ODA private endpoint and sorted by submission time in ascending order so the last submitted work requests will be at the end of the list:

oci oda work-request list --compartment <private-endpoint-compartment-ocid> --resource-id <oda-private-endpoint-ocid> --sort-by TIME_ACCEPTED --sort-order ASC

Get the summary of the specific work request:
oci oda work-request get --work-request-id ocid1.coreservicesworkrequest.oc1.phx.aaabbbccc

Get log entries (usually informational messages) about the work request:
oci oda work-request-log-entry list --work-request-id
ocid1.coreservicesworkrequest.oc1.phx.aaabbbccc

Get errors associated with specific work request:
oci oda work-request-error list --work-request-id
ocid1.coreservicesworkrequest.oc1.phx.aaabbbccc

Typical reason for failures while creating a private endpoint:

1) Missing IAM policies
To create a private endpoint you should have “manage” permission for oda-family or oda-private-endpoints
allow group <group-name> to manage oda-private-endpoints
The same policy is required if you want to create or delete SCAN proxies.

If you want to use a private endpoint from the ODA UI, you need to associate the private endpoint with the ODA instance. To do that you should have “manage” permission for oda-private-endpoint-attachment (oda-family also works):

allow group <group-name> to manage oda-private-endpoint-attachments

If you need to get the list of work request IDs or read the specific one you would need, at least,
allow group <group-name> to inspect oda-instances (to get list) in compartment <abc>
allow group <group-name> to read oda-instances (to read specific work request by ID) in compartment <abc>
"oda-family" should also work instead of "oda-instances" in the policy above.

Since the process whereby ODA creates private endpoints requires making changes in a subnet, you'll need to create a corresponding policy:

allow group <group-name> to manage virtual-network-family in compartment <pe-compartment>

If your private endpoint is created in a compartment different from the subnet compartment, then additionally you should create another policy:
allow group <group-name> to manage virtual-network-family in compartment <subnet-compartment>

If you prefer to define more granular policies, you can use the following:

allow group <group> to manage private-ips in compartment <pe-compartment>
allow group <group> to manage vnics in compartment <pe-compartment>
allow group <group> to use subnets in compartment <subnet-compartment>
allow group <group> to use network-security-groups in compartment <pe-compartment>

2) Not enough private IPs in the private endpoint subnet
You need to make sure you have at least 3 unused IPs in the subnet in which you provision the ODA private endpoint.

3) Creating a private endpoint in a subnet which has a IPv6 prefix block assigned. Currently IPv6 is not supported with ODA private endpoint. Learn more about IPv6 support in Oracle Cloud Infrastructure here: https://docs.oracle.com/en-us/iaas/Content/Network/Concepts/ipv6.htm

Commented [PK1]: Is there anything else to say about items 3 and 4? E.g. where is the failure if you try to create a PE when there's AN IPv6 CIDR block? And when might a subnet be unavailable?

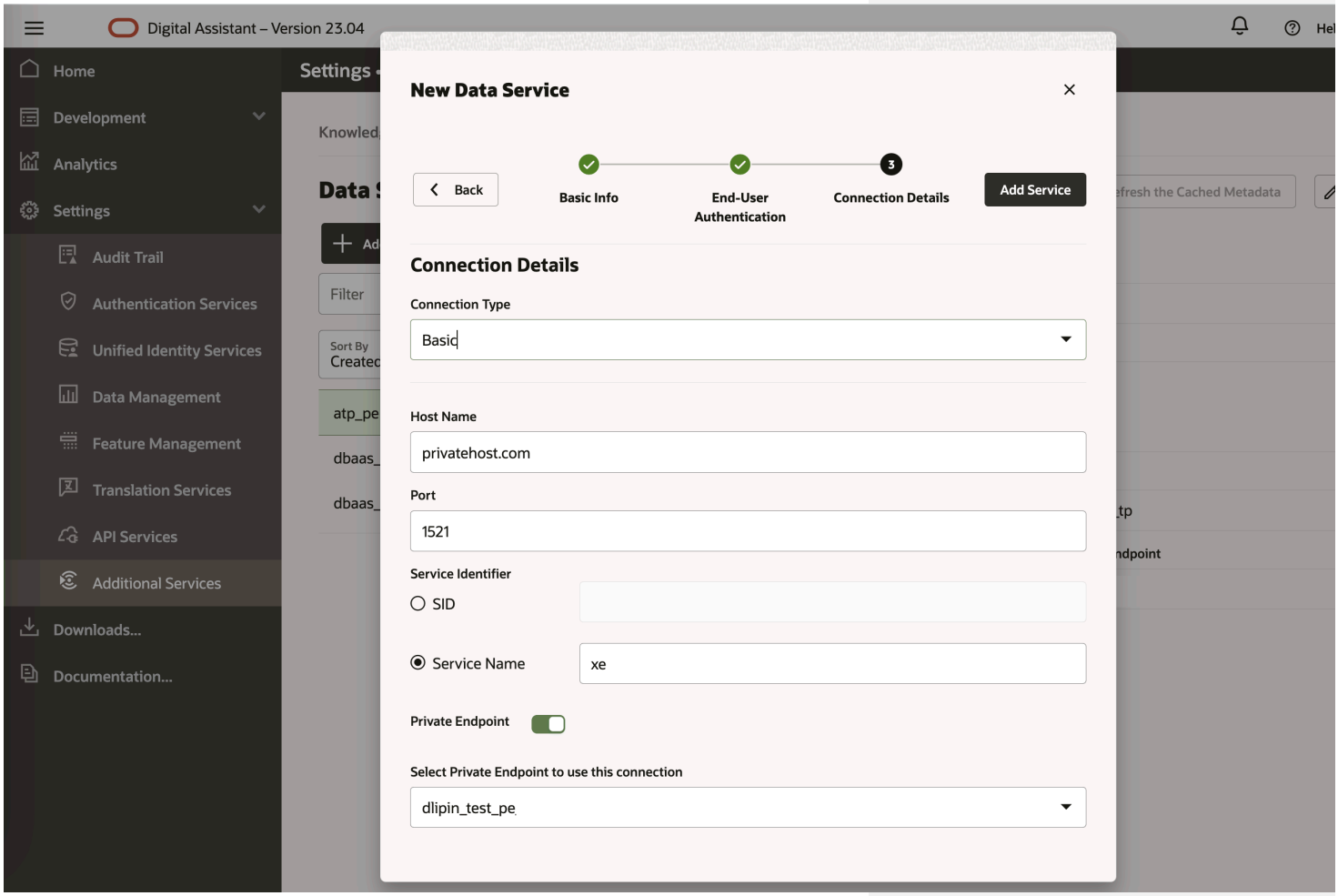
4) Subnet is not in Available state, e.g. if it is still being provisioned or updated. You can check the subnet state in OCI Console.

Using ODA Private Endpoint on the ODA Service Instance UI

Once you have created an ODA private endpoint and have associated it with an ODA instance, you can utilize this private endpoint to make connections to your private resources.
Currently the use of private endpoints is supported for SQL Dialogs (data services) and REST services.

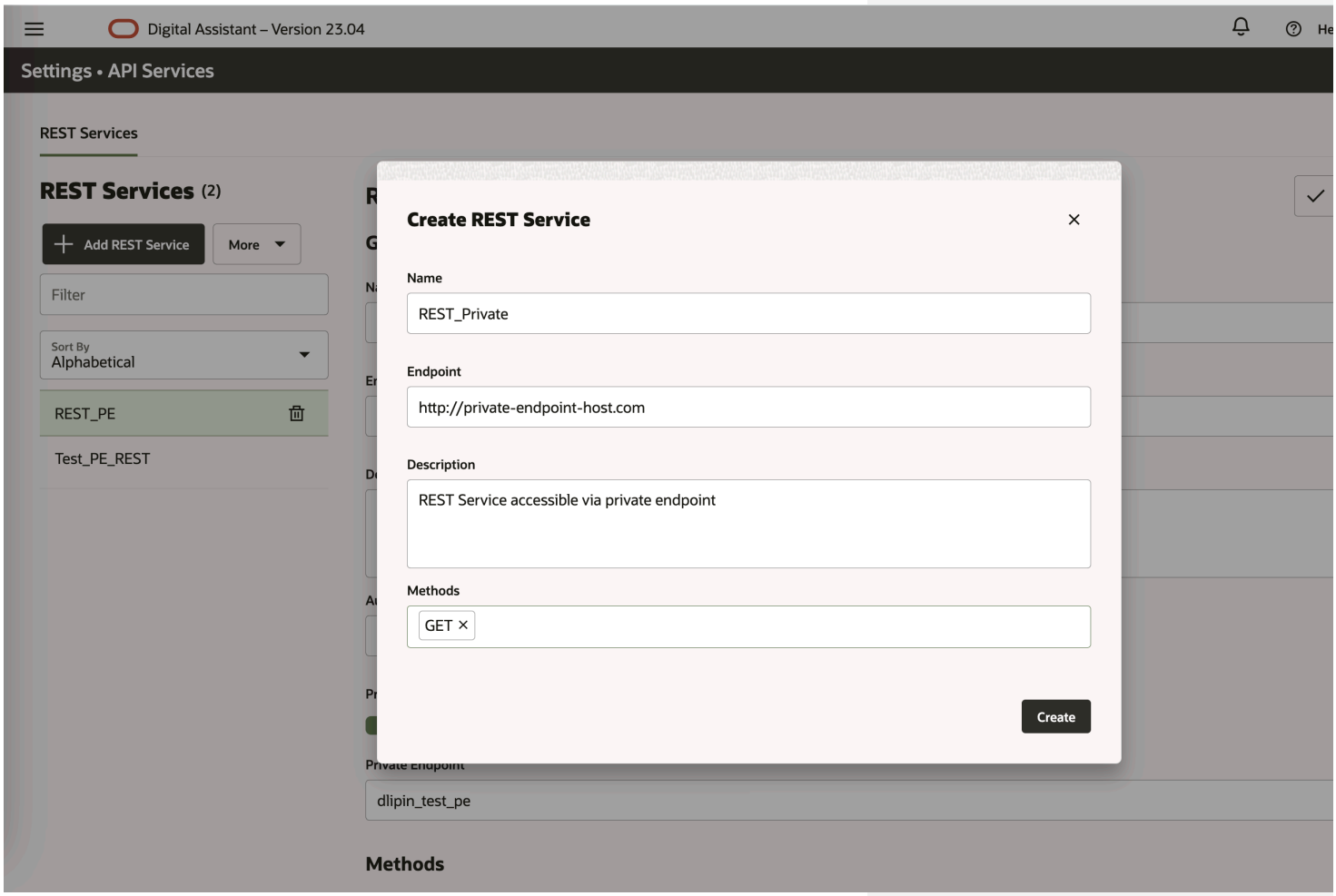
Configure SQL Dialogs with Private Endpoint

To create a data service using a private endpoint, you need to specify the details for the database connection, then enable the “Private Endpoint” option in the last step of the wizard and select one of the attached private endpoints. If you have multiple private endpoints, make sure you select the right one.

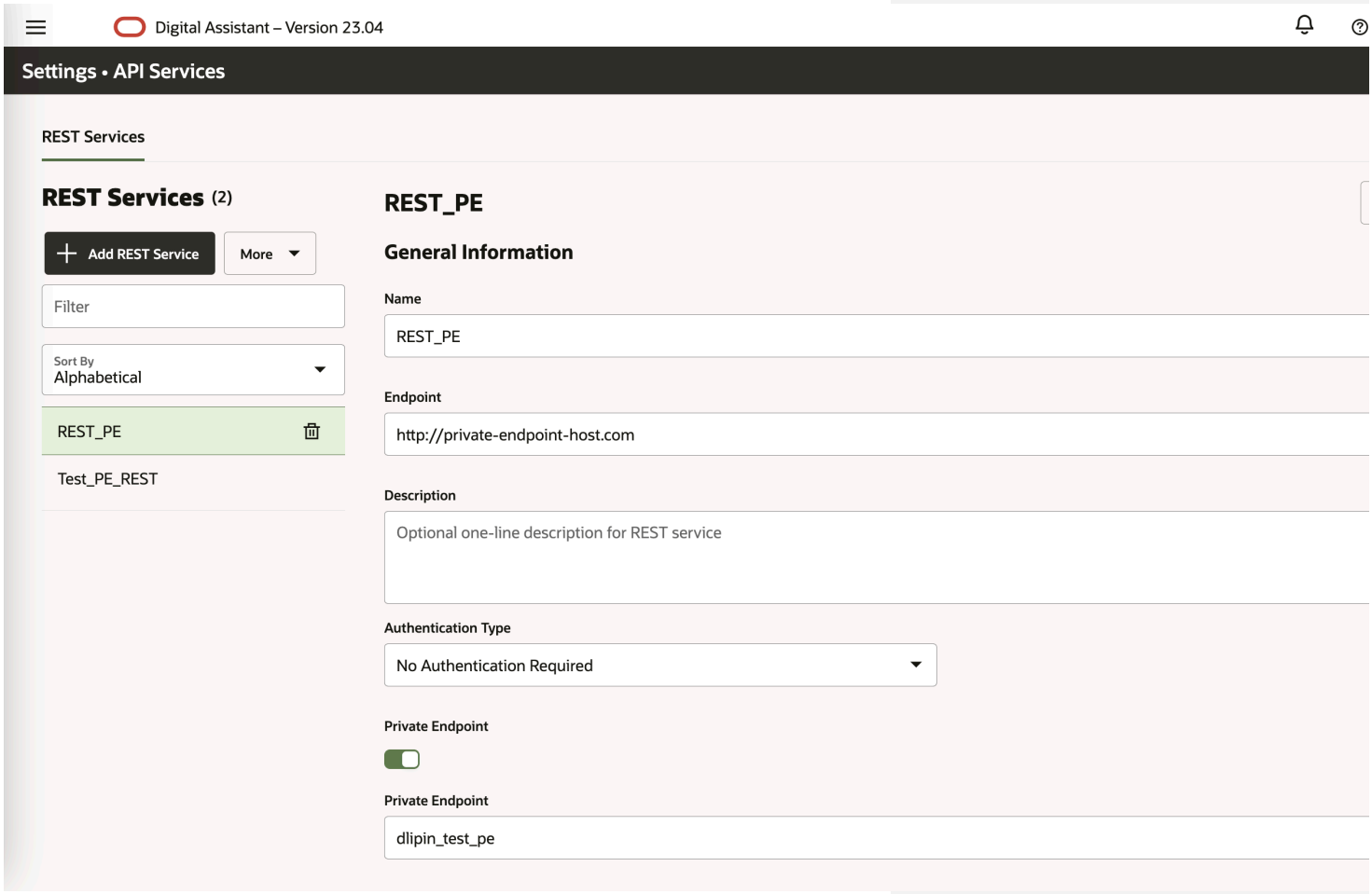


Configure API Service (Call REST Service component) with Private Endpoint

To create a REST service using a private endpoint, you first add a REST service and specify the URL in the Endpoint field.



After you have created the REST service, you will see an option to enable Private Endpoint for this REST Service and to select one of the available private endpoints.
If you don’t see the Private Endpoint enablement option or list of available private endpoints is empty, then make sure you have associated the private endpoint with that ODA instance.



Private Endpoint Limitations and Restrictions

You can assign a maximum of 5 network security groups to a given ODA private endpoint.

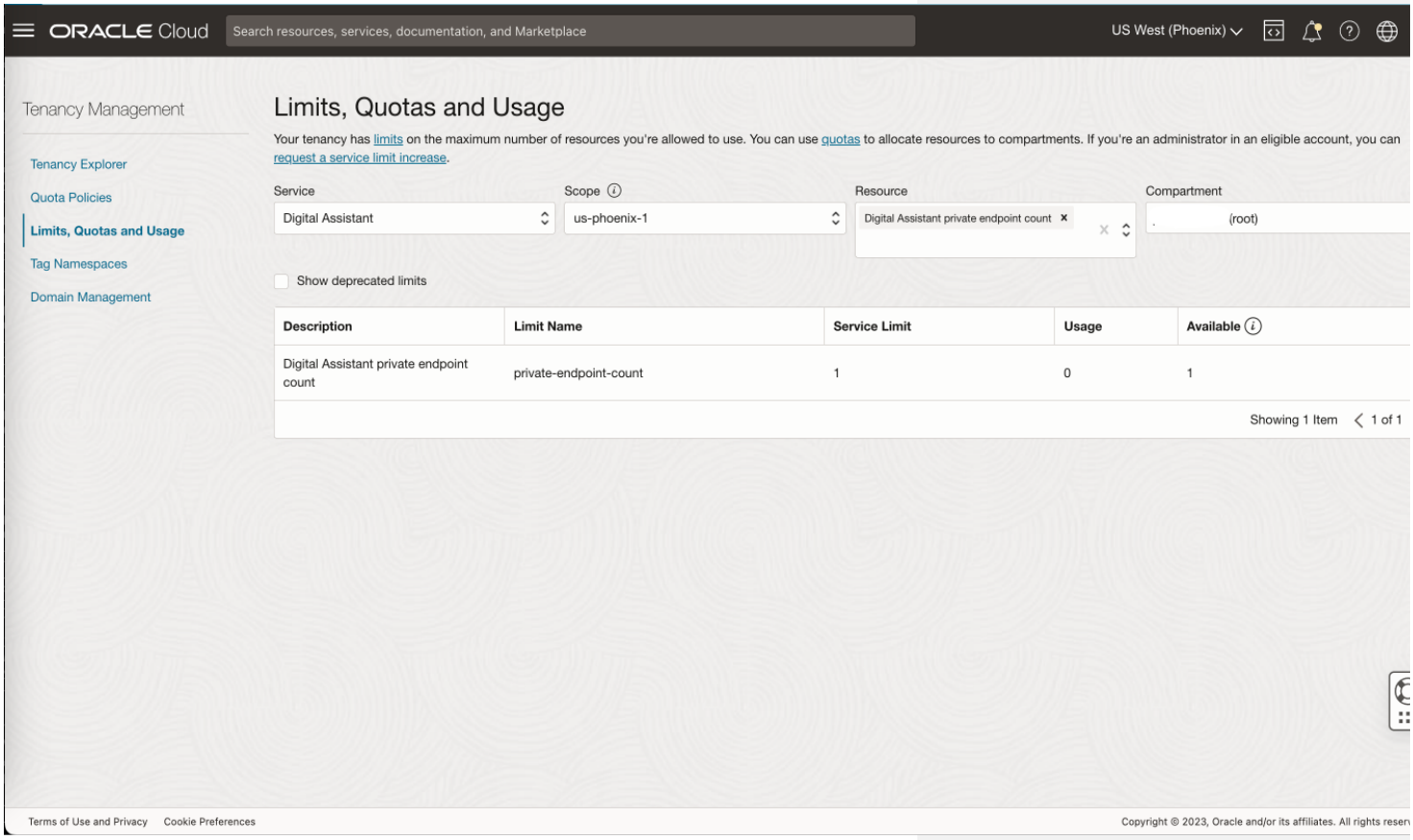
You can create a maximum of 15 SCAN proxies (for a RAC database) for the given ODA private endpoint. If you need to access more that 15 RAC databases and thus need to create more than 15 SCAN proxies, create an additional ODA private endpoint. Make sure you associate this

private endpoint with the ODA instance in order to use it in SQL Dialog or Call REST Service components.

When you create a SCAN proxy, you need to specify the type of the proxy. And to make this choice, you need to decide whether you want to use IP addresses or the fully qualified domain name of the SCAN Listener to access it. For an IP-based SCAN proxy you can specify up to 3 IP addresses for each SCAN proxy and for a FQDN-based SCAN proxy, you can specify only one fully qualified domain name.

Technically, there is no limit on how many ODA private endpoints you can associate with one ODA instance, or how many ODA instances can use the same ODA private endpoint. However, there is a limit on how many ODA private endpoints can be created in your tenancy. By default, this limit is set to 1 for all customers. You can check your limits in the OCI Console on the Governance and Administration -> Limits, Quotas and Usage page:

Select “Digital Assistant” as the service name, the region of your choice and “Digital Assistant private endpoint count”, leave compartment at the default root (tenancy) compartment. If you need to increase the limit, please file a service request.



When registering a data service in Digital Assistant with private endpoints, there is an existing limitation for using IP addresses as a database hostname.
Currently you can specify the IP address only for SCAN Listener of the RAC database and only if you have already created a corresponding SCAN proxy for it.
Specifying an IP address for non-RAC database is currently not supported. FQDN addresses work just fine for non-RAC databases.

When creating a new API Service (REST service) with a private endpoint, you can only use the fully qualified domain name at this moment. IPv4/6 addresses are not currently supported.

Conclusion

ODA private endpoint is a new powerful feature which enhances ODA capabilities to connect to customers resources.