[Enter Client Name]

Cloud Pak for Data – Getting Started Guide

This page is left intentionally blank

**Document Data**

Client: [Enter Client Name]

Project: Cloud Pak for Data

Title: System Administration

Revision: 1

**Revision History**

|  |  |  |
| --- | --- | --- |
| Date | Revision | Notes |
|  |  |  |
|  |  |  |
|  |  |  |

**Contributors**

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Company | Role | Contact Information |
|  |  |  |  |

Cloud Pak for Data – Getting Started Guide

**Table of Contents**

[Introduction 7](#_Toc16542061)

[Purpose 7](#_Toc16542062)

[Objectives for the engagement 7](#_Toc16542063)

[Hardware/Software Environment 8](#_Toc16542064)

[Cloud Pak for Data software product and add-ons: 8](#_Toc16542065)

[Operating System 8](#_Toc16542066)

[Cloud Pak for Data hardware setup 8](#_Toc16542067)

[Hardware Configuration and IP Addresses 8](#_Toc16542068)

[Cloud Pak for Data URL to access 8](#_Toc16542069)

[Cloud Pak for Data Architectural Overview 9](#_Toc16542070)

[Cloud Pak for Data Node configuration 10](#_Toc16542071)

[Add-on components: 10](#_Toc16542072)

[Administration of Cloud Pak for Data environment 11](#_Toc16542073)

[Manage user access to Cloud Pak for Data 12](#_Toc16542074)

[Add User 12](#_Toc16542075)

[Monitor the health and utilization of your cluster nodes 15](#_Toc16542076)

[Manage the nodes in your cluster 18](#_Toc16542077)

[Monitor the health of pods where services are running 19](#_Toc16542078)

[Kubernetes commands 20](#_Toc16542079)

[Access system alerts 20](#_Toc16542080)

[Setup SMTP for notifications 21](#_Toc16542081)

[Connecting to data sources 23](#_Toc16542082)

[Organize Data Assets 24](#_Toc16542083)

[Business Glossary, Policies and Rules 24](#_Toc16542084)

[Analytics Projects and Assets 26](#_Toc16542085)

[Add collaborators to project 27](#_Toc16542086)

[Manage collaborators 27](#_Toc16542087)

[Data Access and Analysis Tools 28](#_Toc16542088)

[Load or Retrieve Data 28](#_Toc16542089)

[Quick access to data sets 29](#_Toc16542090)

[Analyze data using - Jupyter notebooks 30](#_Toc16542091)

[Manage Deployments 33](#_Toc16542092)

[Installing global libraries and packages on the cluster 34](#_Toc16542093)

[Installing a global Python library 34](#_Toc16542094)

[Troubleshooting Cloud Pak for Data Platform 35](#_Toc16542095)

[Assessing Health via CLI Tools 35](#_Toc16542096)

[Health assessment via serviceability tool 36](#_Toc16542097)

[Health assessment via platform tools 37](#_Toc16542098)

[Additional resources for Cloud Pak for Data and ML 41](#_Toc16542099)

[Product documentation: 41](#_Toc16542100)

[Community resources: 41](#_Toc16542101)

[Learn more about Kubernetes: 41](#_Toc16542102)

[Learn more about Machine Learning: 41](#_Toc16542103)

# Introduction

## Purpose

This document provides details of Cloud Pak for Data Platform software installation, system administration and setup for **<client name>** performed during **<engagement timeframe (month and year)>**. The scope of this document in context to the project is limited to:

* Install and configuration of Cloud Pak for Data
* General administration and features of the product

## Objectives for the engagement

The objectives for this project included:

* Installation of Cloud Pak for Data platform
* Setup connectivity and show features of the platform

# Hardware/Software Environment

This section provides details of the hardware and software used for Cloud Pak for Data installation.

### Cloud Pak for Data software product and add-ons:

Cloud Pak for Data version **<version installed>**

**<list of add-ons installed>**

### Operating System

Redhat **<operating system version>**

### Cloud Pak for Data hardware setup

Server nodes were created as virtual machines.

Total number of nodes – **<total number of nodes installed on>**

Master Nodes – **<number of master nodes>**

Worker Nodes - **<number of worker nodes>**

### Hardware Configuration and IP Addresses

**<environment deployed on – for example Azure/AWS/IBM Cloud/Prem>**

**<infrastructure deployed on – for example ICP/OpenShift/OpenStack>**

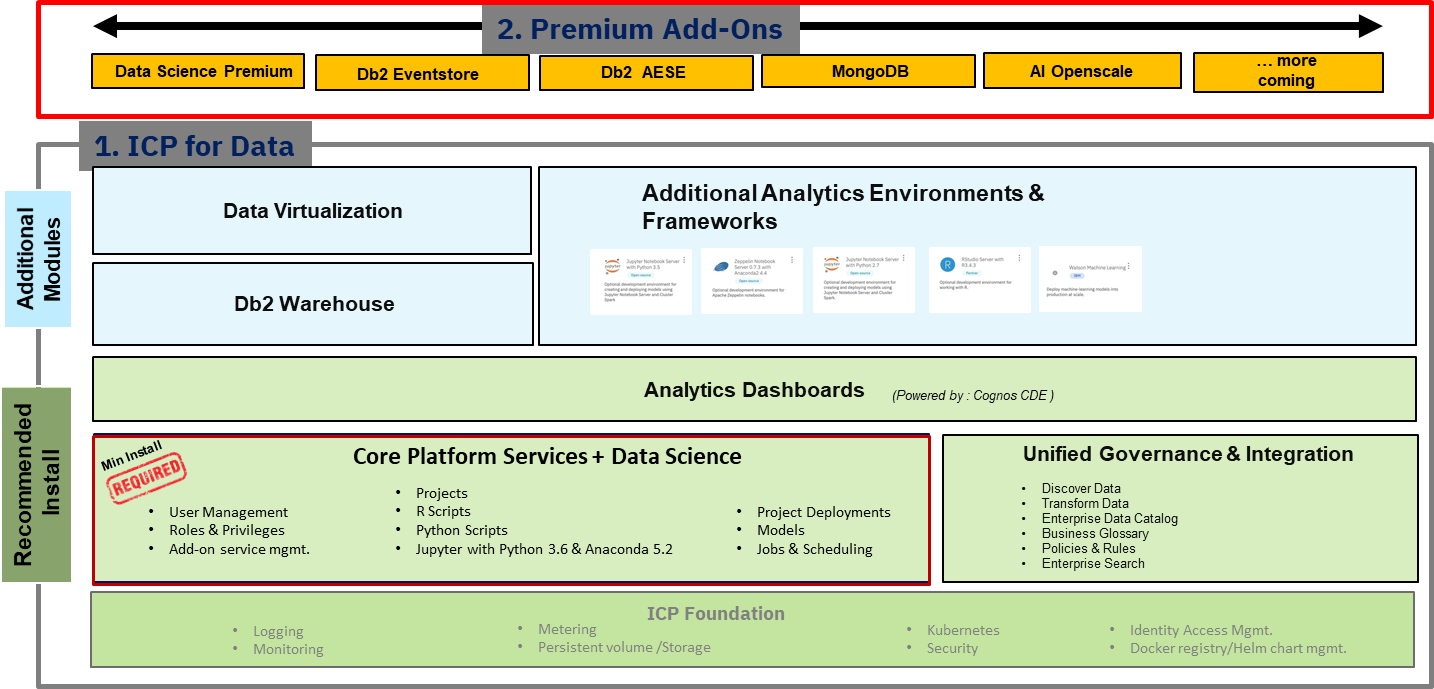
**<client hardware configuration and IP Adresses>**

### Cloud Pak for Data URL to access

Client has created an alias **<URL alias for the cluster>**

# Cloud Pak for Data Architectural Overview

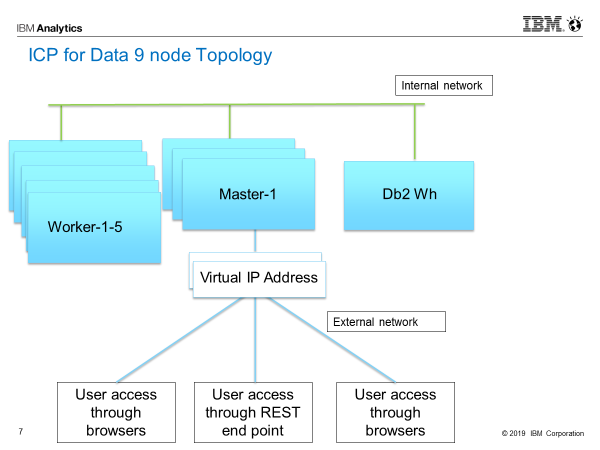
The platform provides a modern data and analytics architecture that is elastic and reliable. In addition, it provides an end-to-end architecture that enables you to catalog, govern, transform, and analyze your data so that you can spend less time managing your data and more time using it to grow your business.



Your data isn't static. Your machine learning models shouldn't be static either. As data is added to your on-premises and cloud data sources, you need to continually test and tune your machine learning models to ensure that they give you valuable insight. But you need to make sure that you're working with high-quality data, which is where the data governance and data manipulation features of IBM Cloud Private for Data come in.

# Cloud Pak for Data Node configuration

The diagram below depicts a small configuration where 6 worker nodes and master nodes were carved out for the install. (use diagram from initial pre-req deck ICP4D\_install\_prep\_xxx.ppt)



We have 9 node configuration with 1 db2wh node and 3 master and 5 worker nodes.

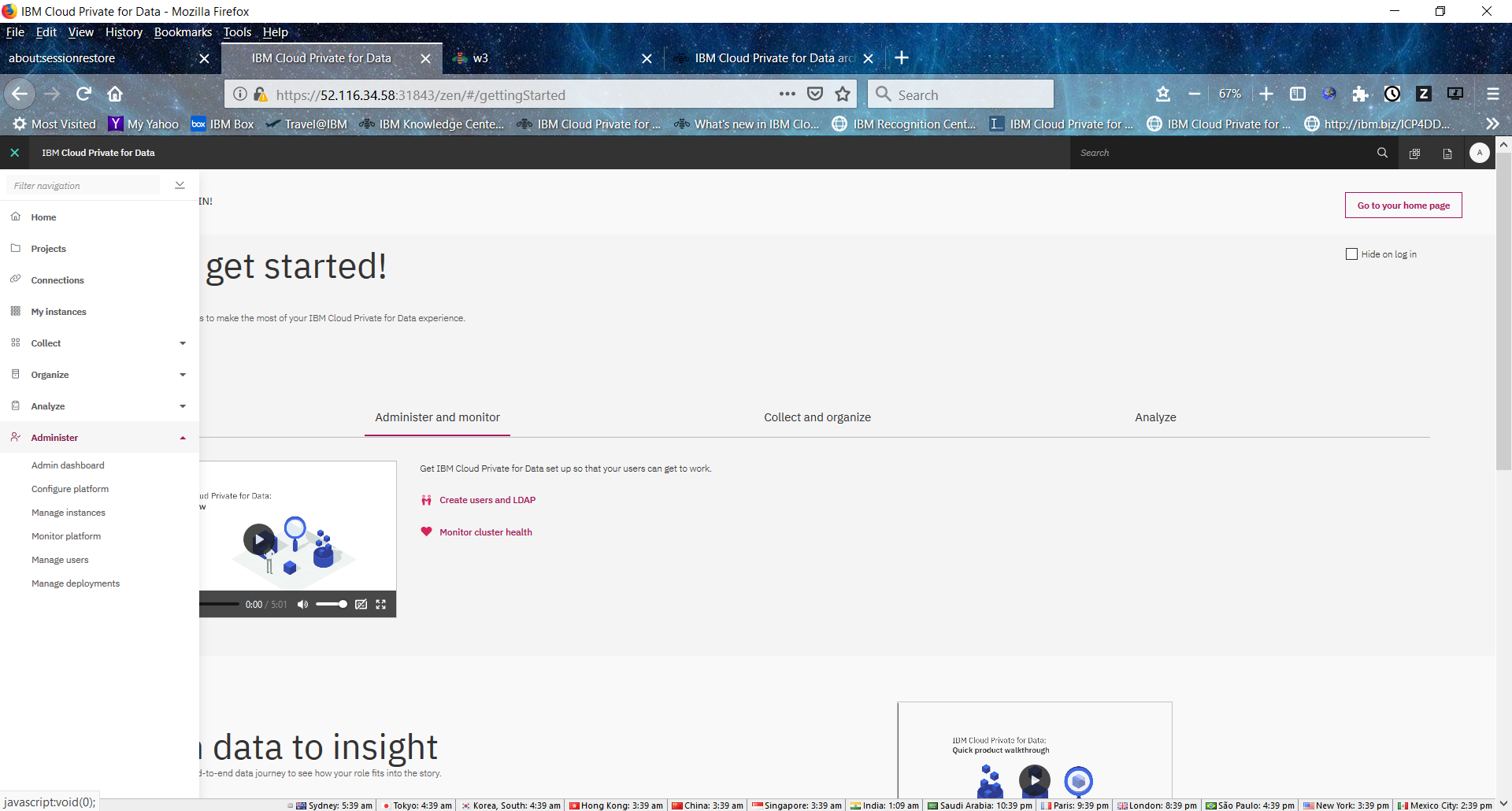
### Add-on components:

Data Science Premium add-on

Watson OpenScale add-on

# Administration of Cloud Pak for Data environment

Cloud Pak for Data provides the embedded administration tools to complete all administration tasks, such as adding users and monitoring the health of the application and the deployment environment.



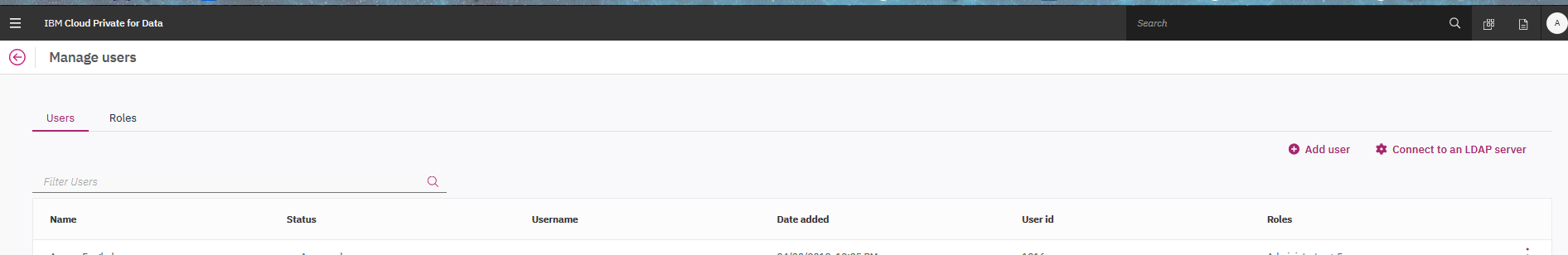
From this menu administrators can access menu of options from left side of menu bar:

Here are some key options that administrators will use:

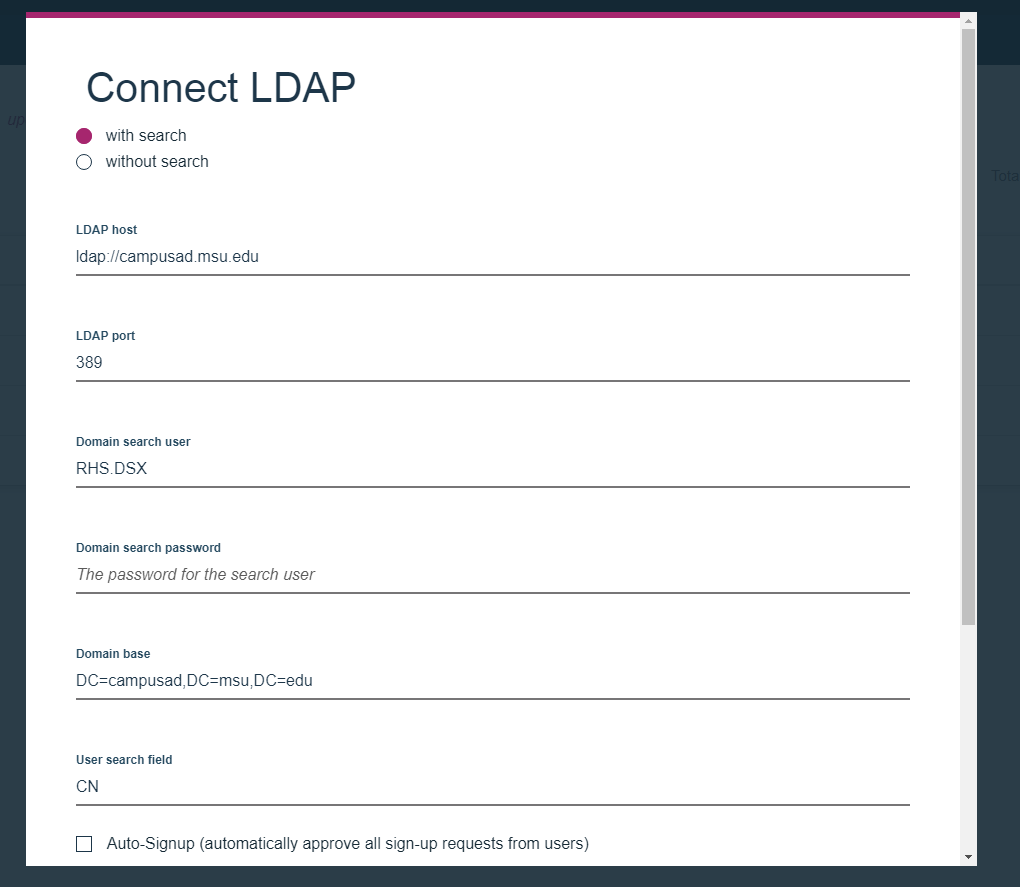
* Manage user access to Cloud Pak for Data
* Monitor the health and utilization of your cluster nodes
* Manage the nodes in your cluster
* Monitor the health of pods where services are running
* Access system alerts

### Manage user access to Cloud Pak for Data

The platform provides options to add directly to the platform or configure with enterprise LDAP registry.



Following is an example for Active Directory setup :

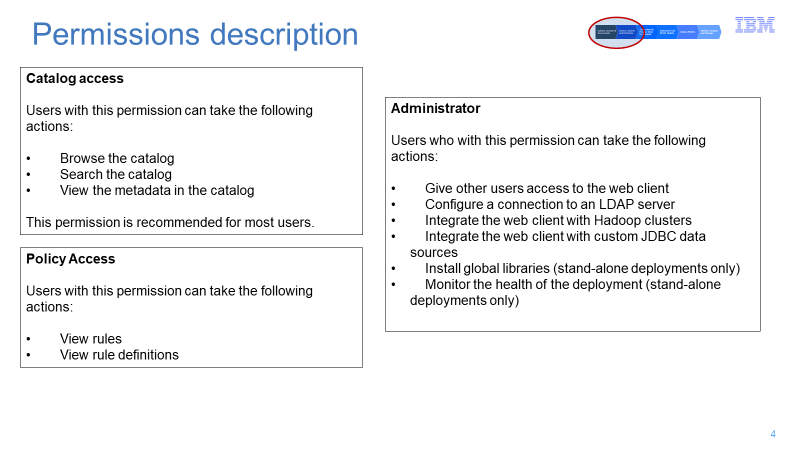


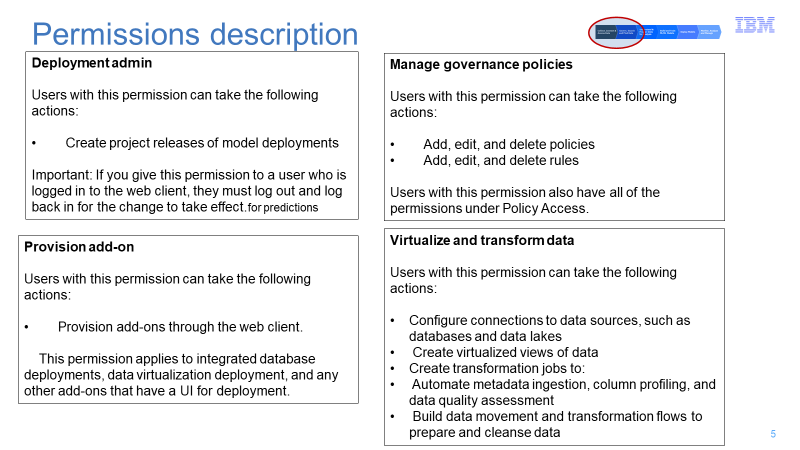
### Add User

If you are adding LDAP users, password management is handled by the LDAP server. If you are adding users to the internal repository database, you must provide the user their temporary password after you add them. With LDAP or without, “Administrator”, need to either search for user from Active Directory or add each userID in Cloud Pak for Data platform with proper role assignment.

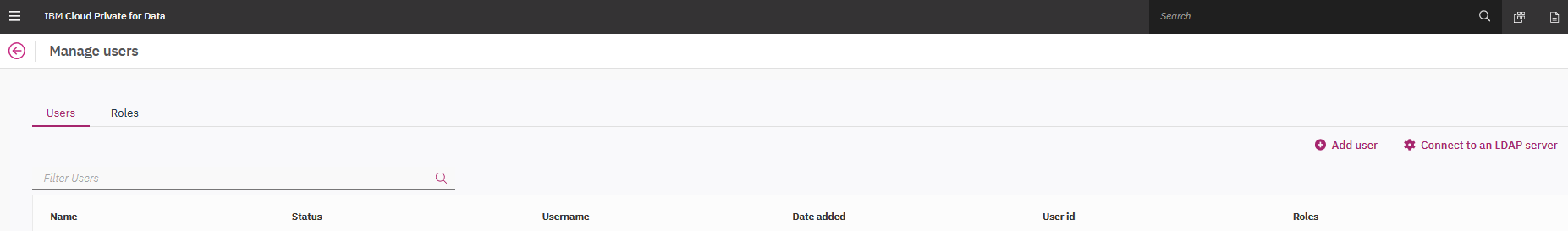
#### **USER Roles**

|  |  |
| --- | --- |
| **Role** | **Permissions** |
| **Administrator** | Administrator |
| Deployment Admin |
|  |
| **Business Analyst** | Catalog Access |
|
| **Data Engineer** | Catalog Access |
| Provision Databases |
| Virtualize and Transform Data |
|  |
| **Data Scientist** | Catalog Access |
|
| **Data Steward** | Catalog Access |
| Policy Access |
| Manage Catalog |
| Manage Governance Policies |
|  |

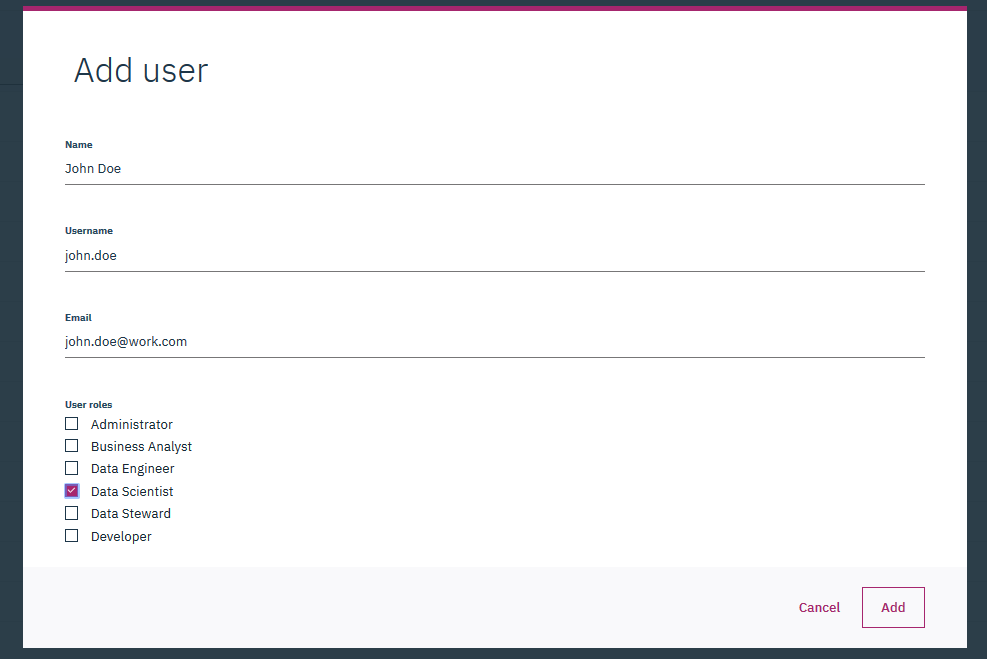




To add new user, Click on Add user link :



This will bring following screen :



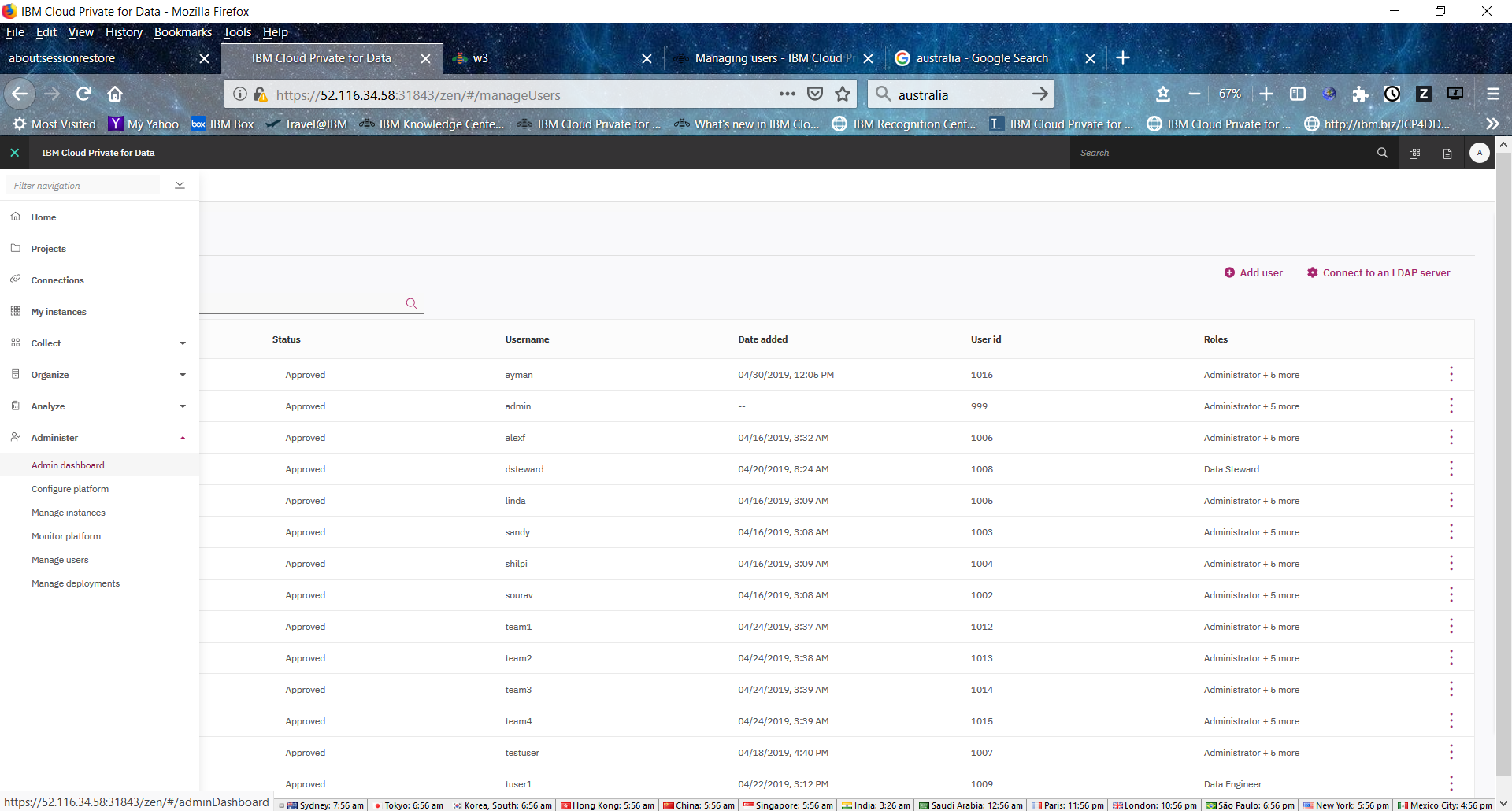
The Access Level option allows to make user as part of particular role. For details on the option, please see : <https://docs-icpdata.mybluemix.net/installadmin/com.ibm.icpdata.doc/zen/admin/users.html>

Once a user is added a system generated password will be flashed for quick access. The admin may choose to select a password through menu option “Edit” option.

[Show an example of customer’s user ID]

### Monitor the health and utilization of your cluster nodes

Admin console Dashboard shows overall health of the cluster resource utilization:

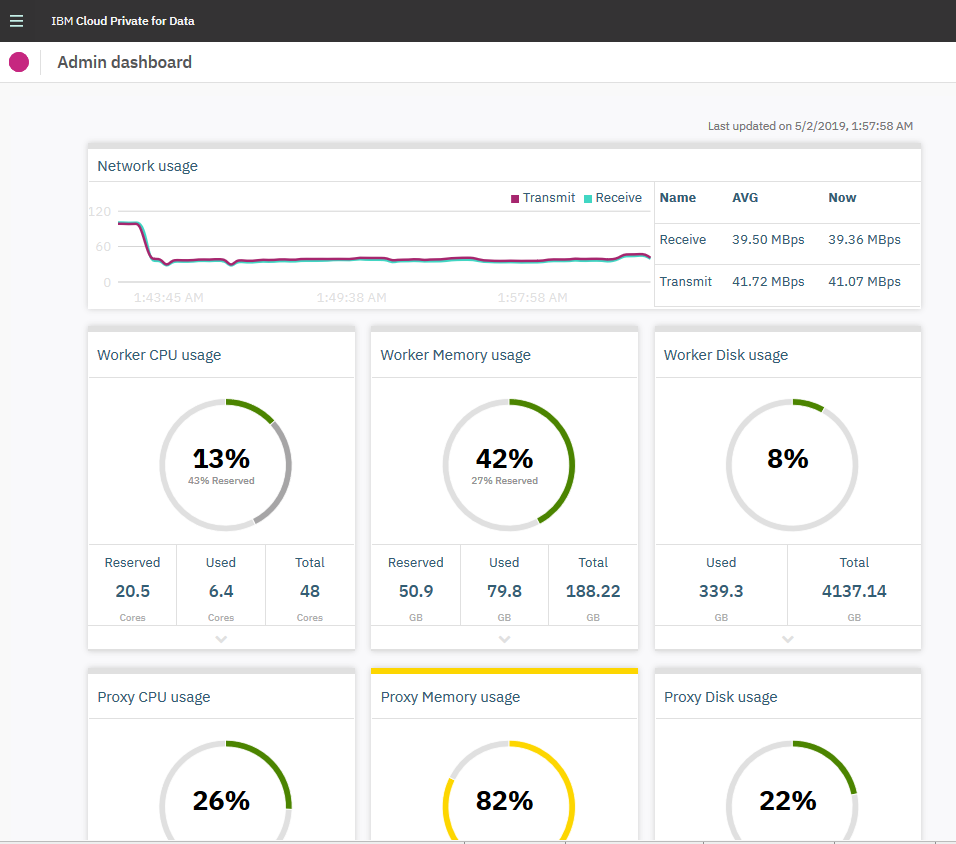


If you want a high-level overview of the status of your cluster, you can monitor the health of your cluster nodes from the Admin dashboard page. You can access the page by selecting Administer > Admin dashboard from the menu.

Specifically, you can monitor:

CPU usage

Memory usage



Each card on the Admin dashboard page shows the average usage across all of the nodes.

Contact IBM Software Support if you notice that all of the nodes in a group are overloaded for an extended period of time. Nodes are overloaded when they run above 90% usage.

Nodes can become overloaded when:

* You have more users than your cluster configuration can handle. For example, your cluster doesn't have sufficient CPU, memory, or storage.
* A node fails and other nodes need to handle requests that would normally be handled by that node.

More info : <https://docs-icpdata.mybluemix.net/installadmin/com.ibm.icpdata.doc/zen/admin/monitornodes.html>

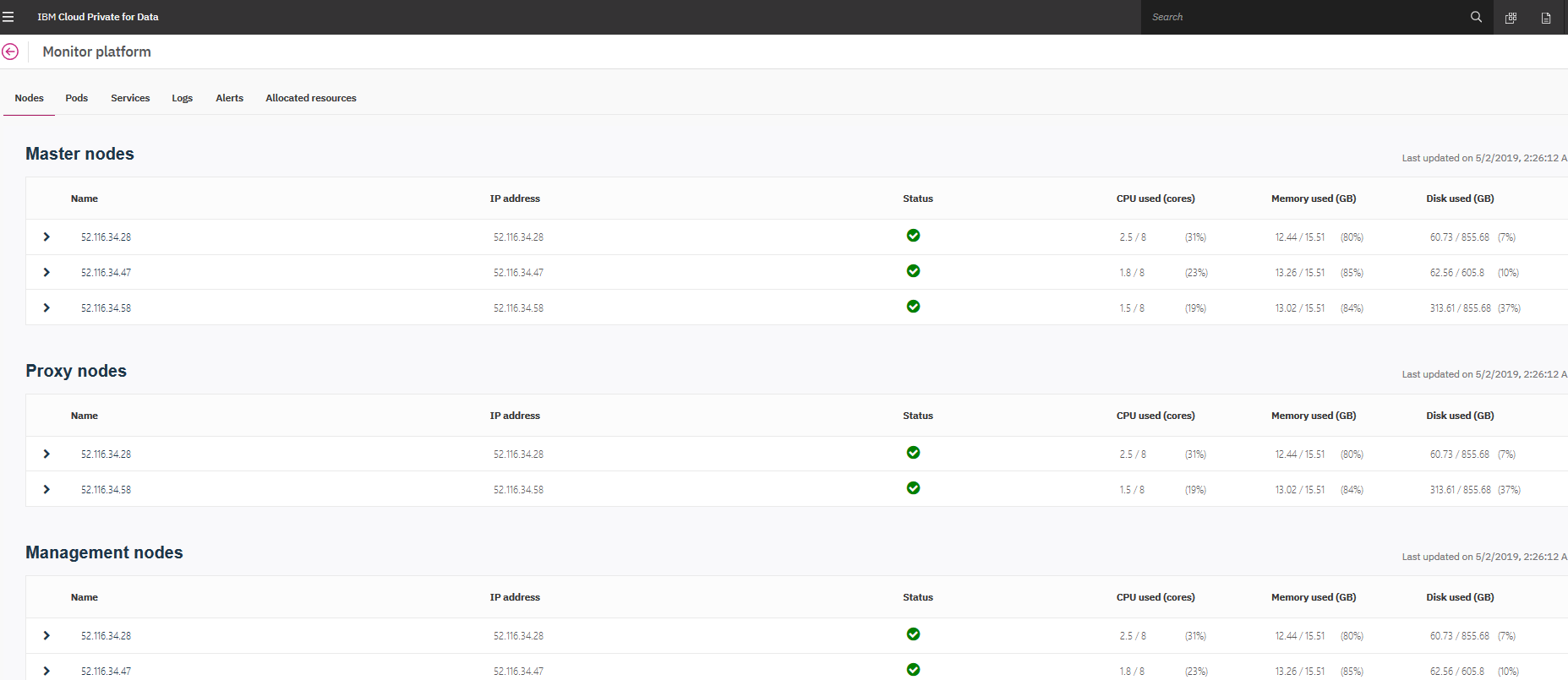
### Manage the nodes in your cluster

This option shows details of each physical node’s resource utilization such as disk, memory and CPU.

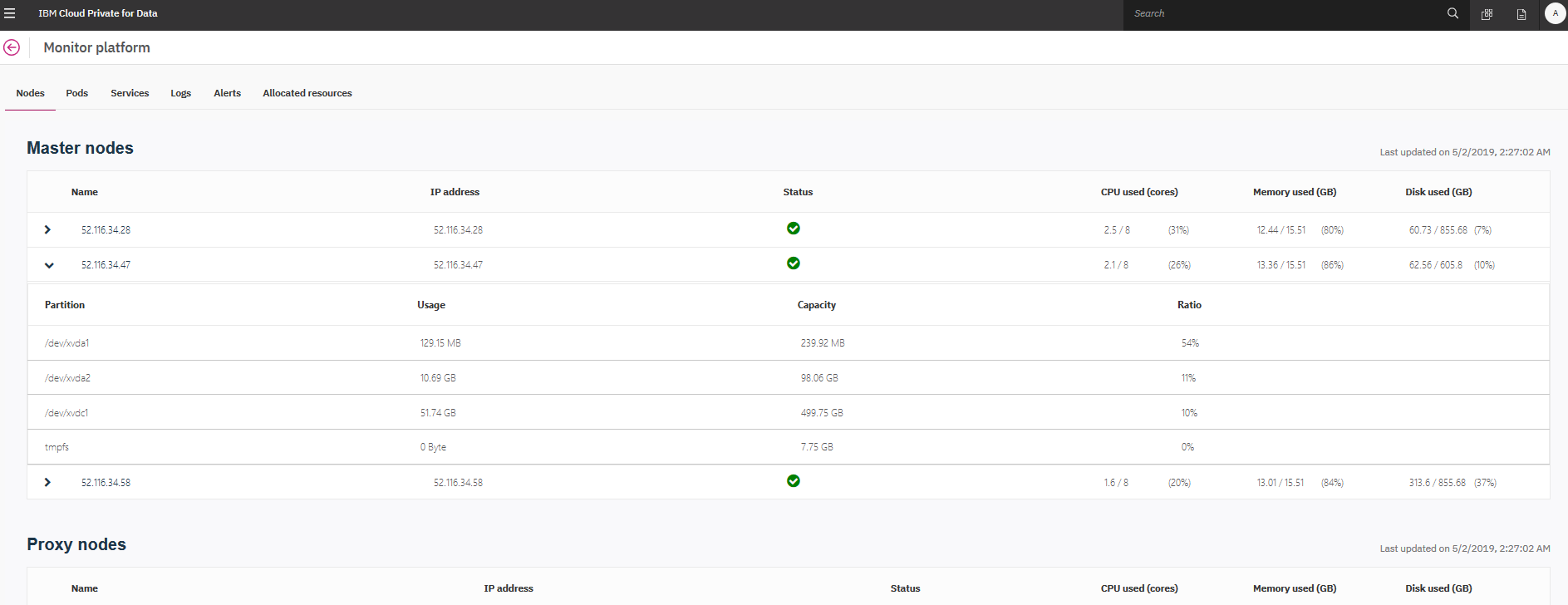
To access the Nodes page, select Administer > Monitor platform from the menu. The Nodes page is open by default.

For each node in the cluster, you can see:

* The IP address of the node
* Whether the node is running or down
* How many CPU cores the node is using
* The amount of memory (GB) the node is using
* The amount of disk storage (GB) the node is using

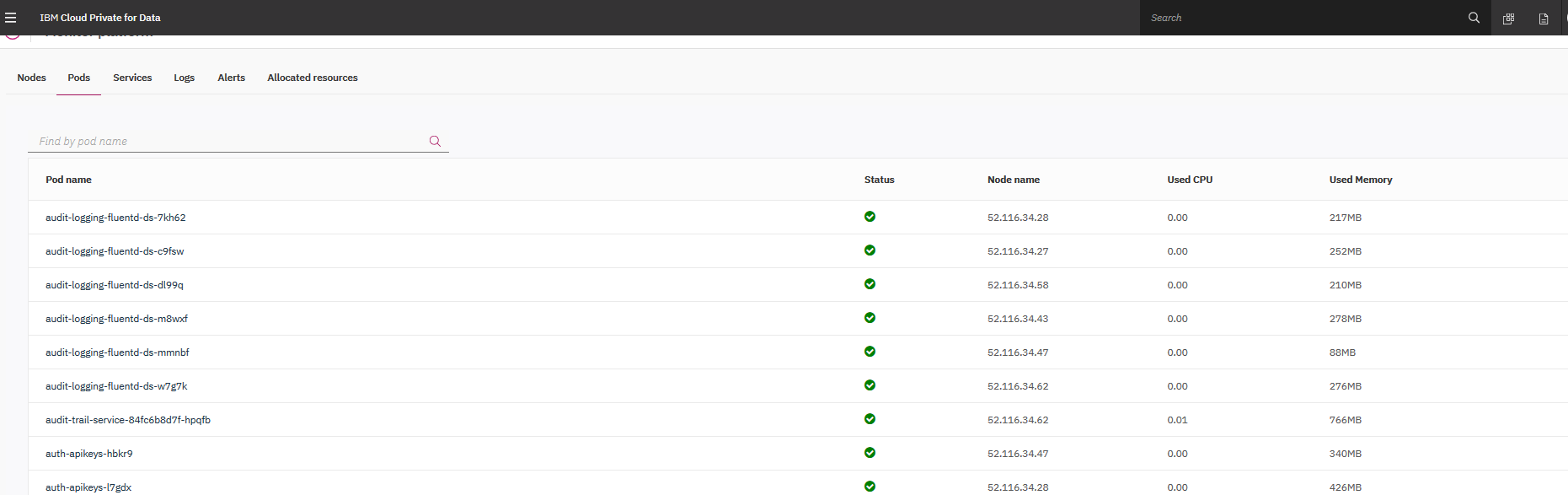


Click on side arrows for detailed view of each node:



### Monitor the health of pods where services are running

The POD options from admin menu shows current location of a kubernetes pod where it is executing.



Additionally, you can click the name of each pod to see the following information:

* The IP address of the node where the pod is deployed
* The name of the service that is running on the pod
* The amount of CPU the pod is using and a graph of the CPU usage over the last 15 minutes
* The amount of memory that the pod is using and a graph of the memory usage over the last 15 minutes
* The amount of disk space that the pod is using and a graph of the disk space usage over the last 15 minutes

For additional info : <https://docs-icpdata.mybluemix.net/installadmin/com.ibm.icpdata.doc/zen/admin/pods.html>

### Kubernetes commands

The root user can login to master1 node and use Kubernetes command line utilities to diagnose issues further. Some additional kubernetes commands that may help to diagnose issues with PODS

To get list of all pods :

kubectl get po --all-namespaces

To get a log dump of a pod :

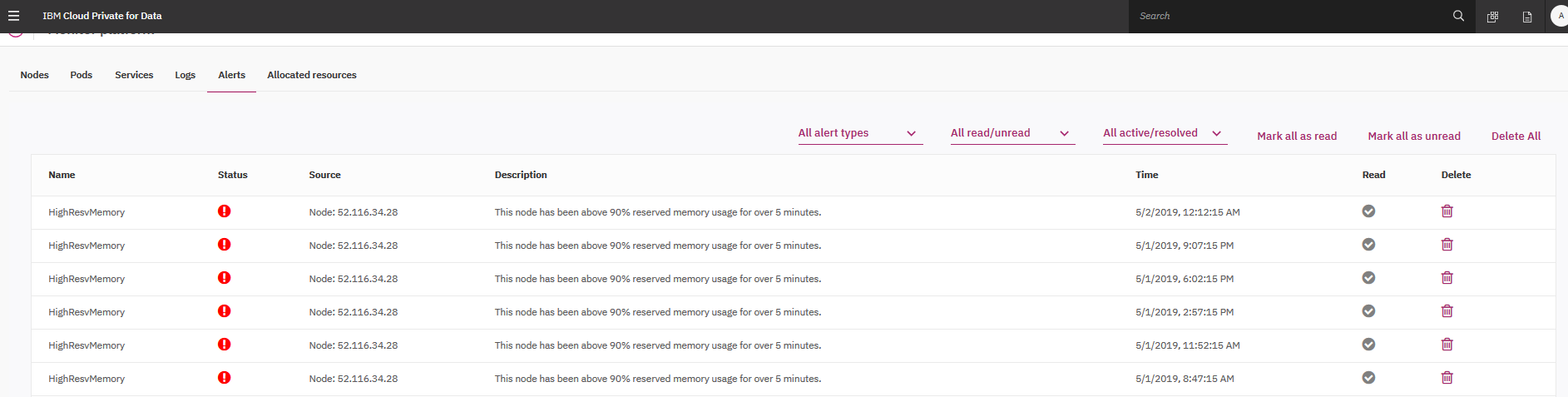
kubectl logs -n zen jupyter-server-1525194671935-1003-b79659b7-tpg84

To get detailed configuration of a pod :

kubectl describe po -n zen jupyter-server-1525194671935-1003-b79659b7-tpg84

### Access system alerts

Cloud Pak for Data platform provides alert for errors and warnings. These can be accessed from admin menu “ALERT”

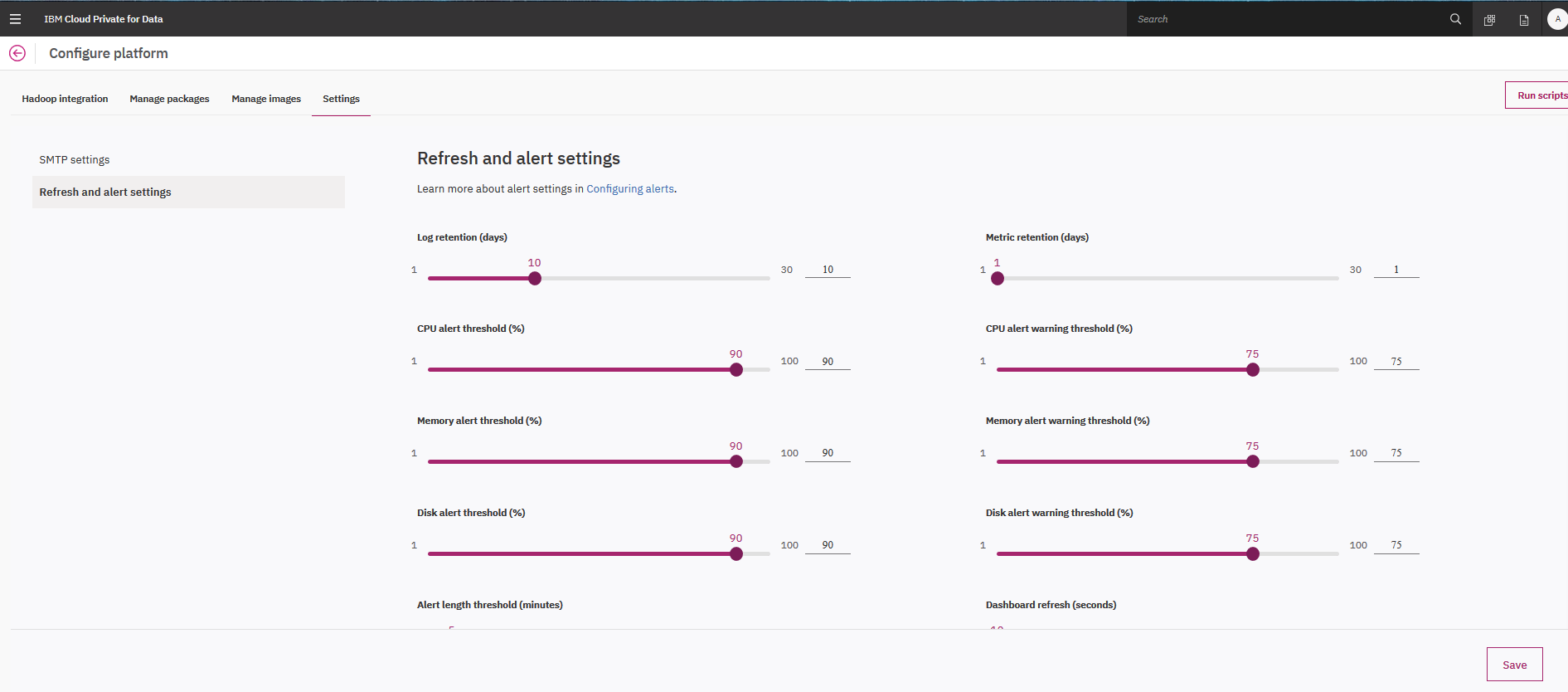


Use last column “Delete” to permanently delete alert. Make sure to keep number of alerts to minimum as these are saved in monitoring database.

The thresholds for alerts can be setup under from:

To set alert thresholds:

1. Log in to the web client as an administrator.
2. From the menu, select Administer > Configure platform.
3. Select Settings.
4. On the Refresh and alert settings page,

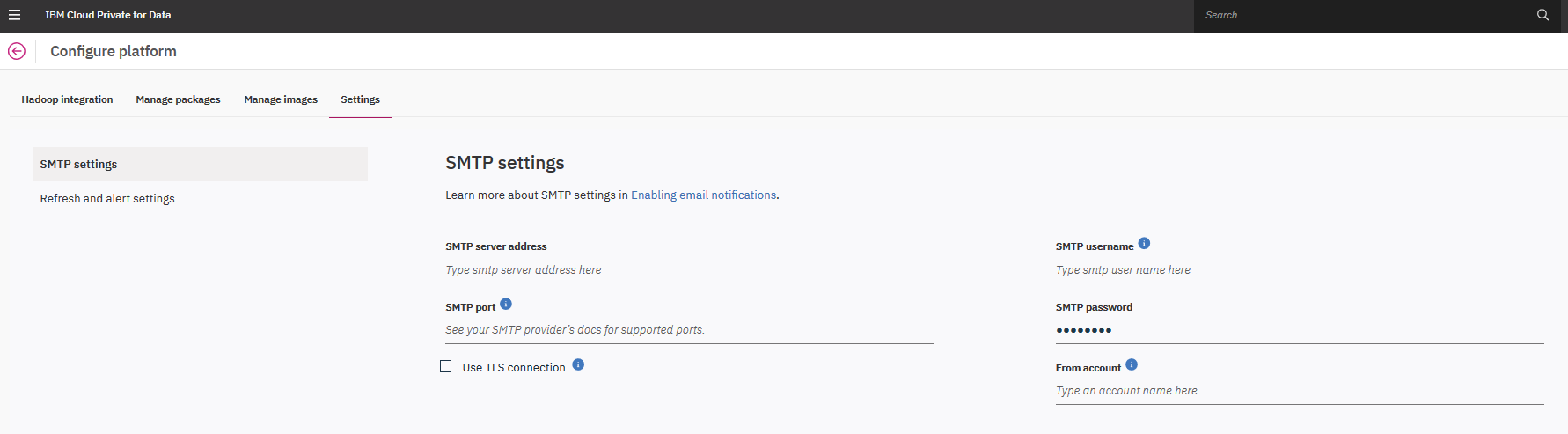


More info: <https://docs-icpdata.mybluemix.net/installadmin/com.ibm.icpdata.doc/zen/install/alert-config.html>

### Setup SMTP for notifications

To enable IBM Cloud Private for Data to send email:

1. Log in to the web client as an administrator.
2. From the menu, select Administer > Configure platform.
3. Select Settings.
4. On the SMTP settings page, specify the following information:
   * Your SMTP mail server address.
   * The port number of your SMTP server.



This screen also allows to setup SMTP server to forward alerts for further integration with enterprise alerting systems.

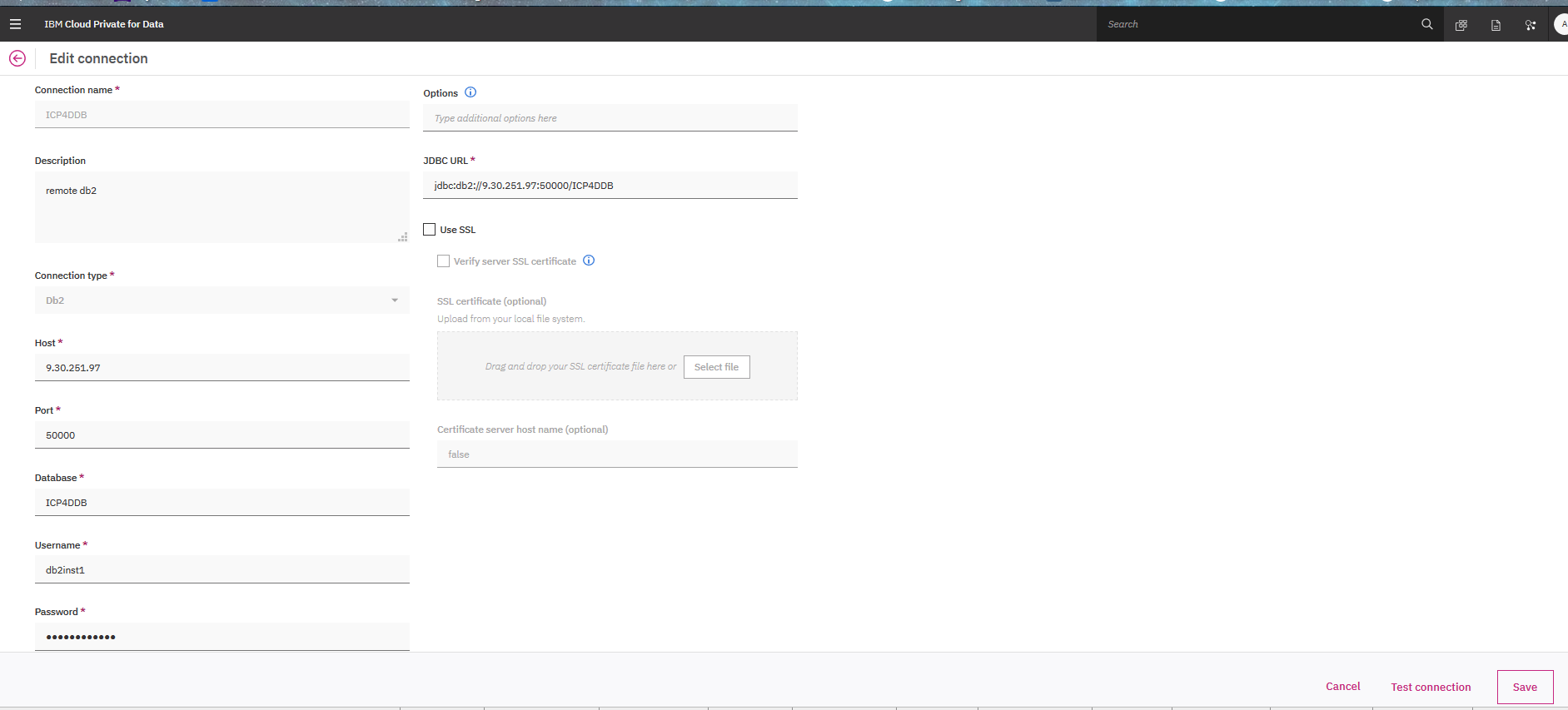
More info : <https://docs-icpdata.mybluemix.net/installadmin/com.ibm.icpdata.doc/zen/install/smpt-config.html>

# Connecting to data sources

Create data connections and browse the data connections that you can use to add data to the enterprise governance catalog, analyze data, virtualize data, and more.

1. From the main menu, click Connections to open the Connections page.
2. If you need to upload a driver for a connection, such as for a Teradata data source, click Upload driver and follow the instructions.
3. From the Connections page, click Add connection.
4. In the Add connection window, specify the information for your connection.

You need to specify the name of the connection and a description, and then select the connection type. Depending on the connection type you specify, you will need to provide a specific set of additional information; for example, a host name, port, and credentials.



You can test the connection using “**Test Connection**” button at the right bottom of the screen.

More info : <https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ_current/com.ibm.icpdata.doc/igc/t_connect_data_sources.html>

# Organize Data Assets

IBM® Cloud Private for Data enables you to structure your enterprise information in a logical way, discover relationships between assets, and keep your data always up-to-date.

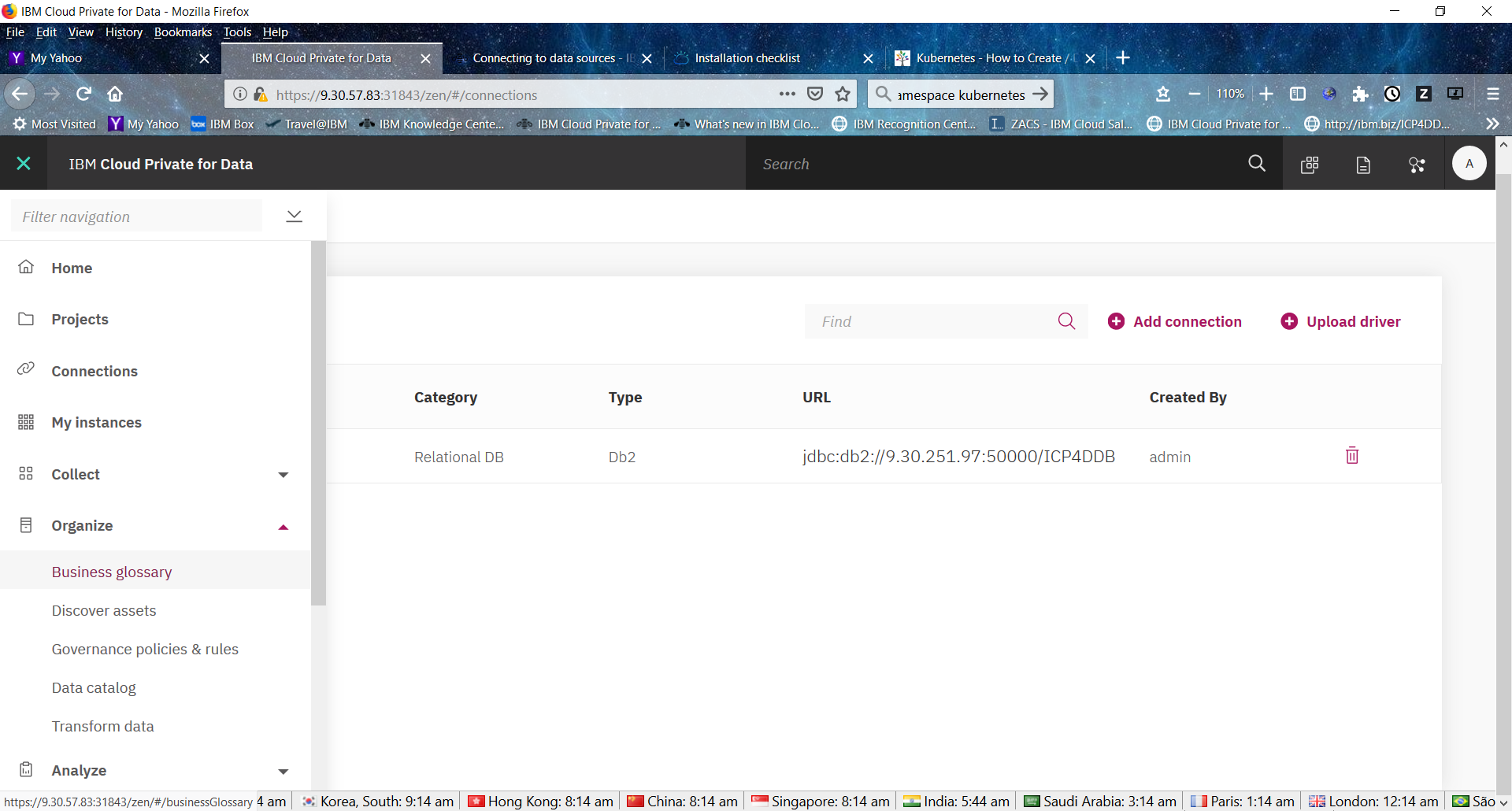
Your enterprise has a lot of data, many assets are related to one another but not in an obvious way. And this data changes, every day. IBM Cloud Private for Data helps you structure your data in a logical way. You can create a data dictionary with common business vocabulary which helps define all important aspects of your enterprise. To ensure compliance with business objectives, you can create information governance rules and policies.

### Business Glossary, Policies and Rules

Business glossary is a catalog of assets that defines the characteristics of your enterprise. It contains terms which you can add to categories to form a logical structure of your data. A common language increases trust and confidence in the information of an organization.

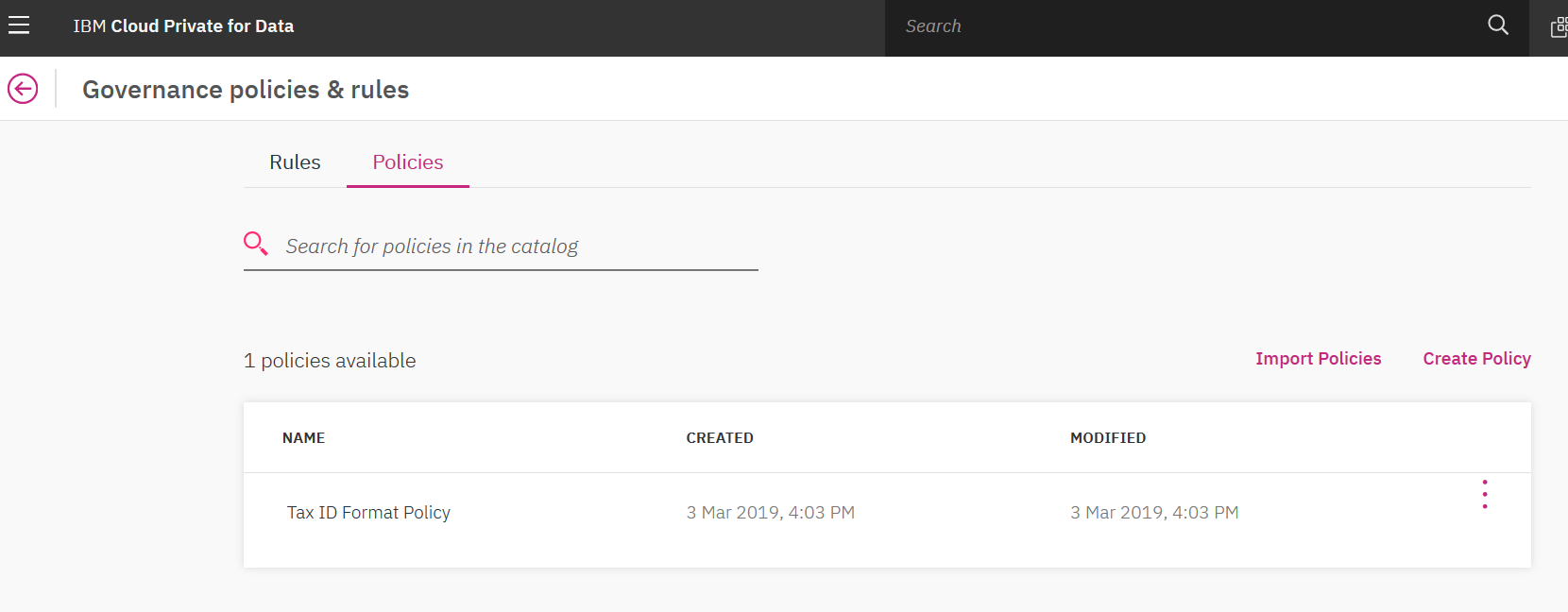
Go to Home menu and click on ‘Organize’

Choose ‘**Business Glossary’**



Return to the ‘**Organize**’ menu

Click on ‘**Governance Policies and Rules**’ and select tab called ‘**Policies**’

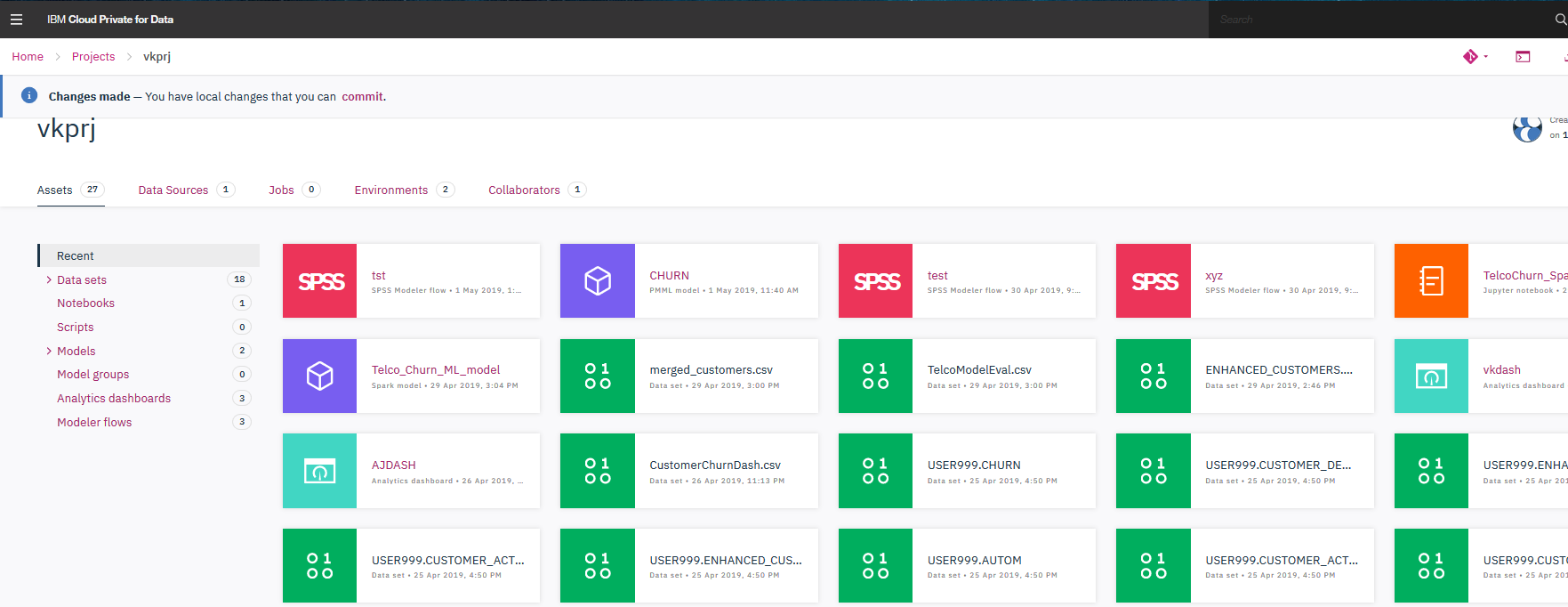


More info : https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ\_current/com.ibm.icpdata.doc/igc/c\_create\_dictionary.html

# Analytics Projects and Assets

A analytics project is how you organize your assets to achieve a particular data analysis goal. Your project assets can include:

* Notebooks
* RStudio files
* Models
* Data sets (local files and remote data sets)
* Scripts



You can also export or import a project as a ZIP or TAR.GZ file.

More info : <https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ_current/com.ibm.icpdata.doc/dsx/analyze-data.html>

# Add collaborators to project

### Manage collaborators

If you have Admin permissions for a project, you can add collaborators, change collaborator permissions, or remove collaborators from that project on its Collaborators page.

The collaborator permissions are:

Viewer

Can view the project, accept changes, and commit changes to their own local copy of the project.

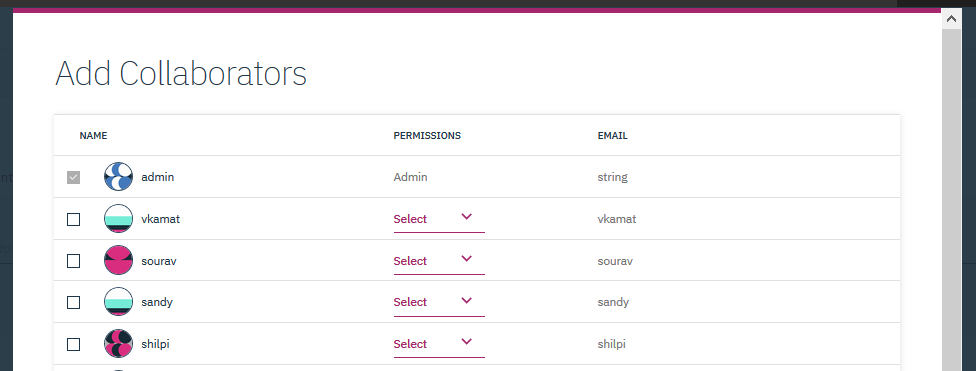
Editor

Can control project assets. Can accept, commit, and push changes.

Admin

Can control project assets, collaborators, and settings. Can accept, commit, and push changes.

From Project’s collaborators menu, access following screen to invite other users :



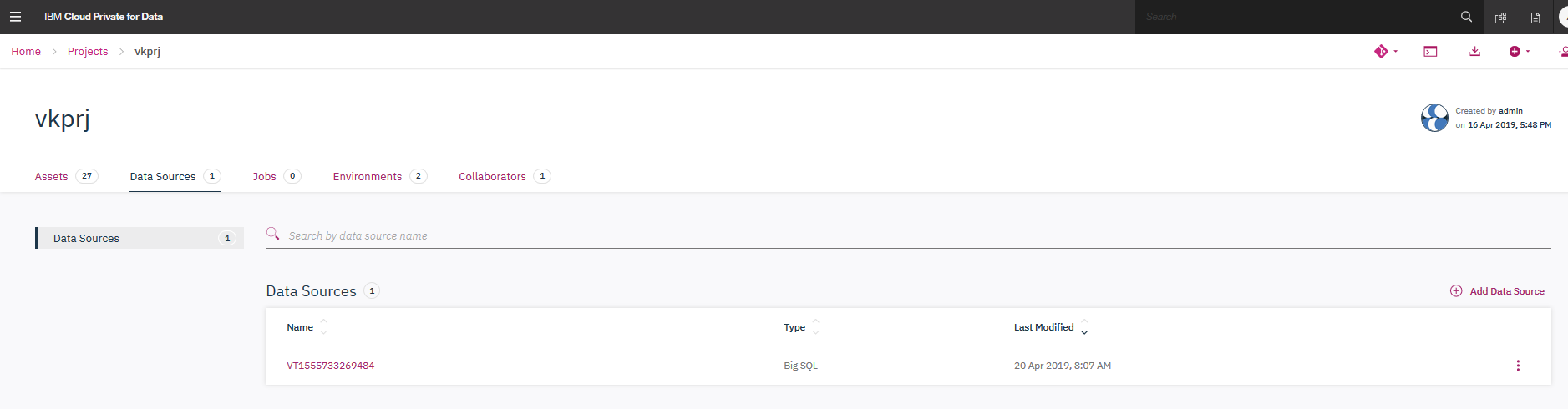
More info : <https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ_current/com.ibm.icpdata.doc/dsx/projects.html#projects__managecollabs>

# Data Access and Analysis Tools

There multiple tools to access data sets and analysis tool within Cloud Pak for Data.

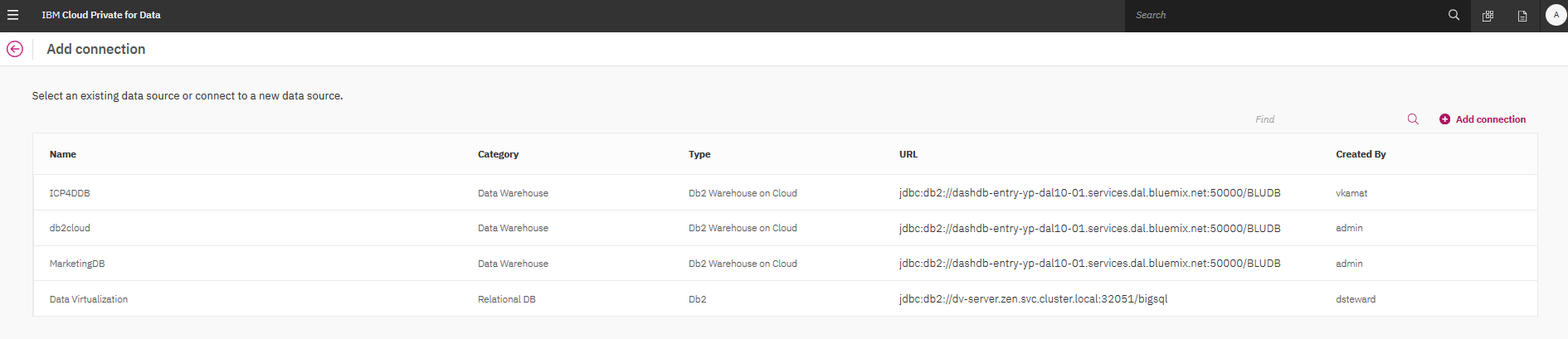
### Load or Retrieve Data

Within a project, administrators can create data sources.

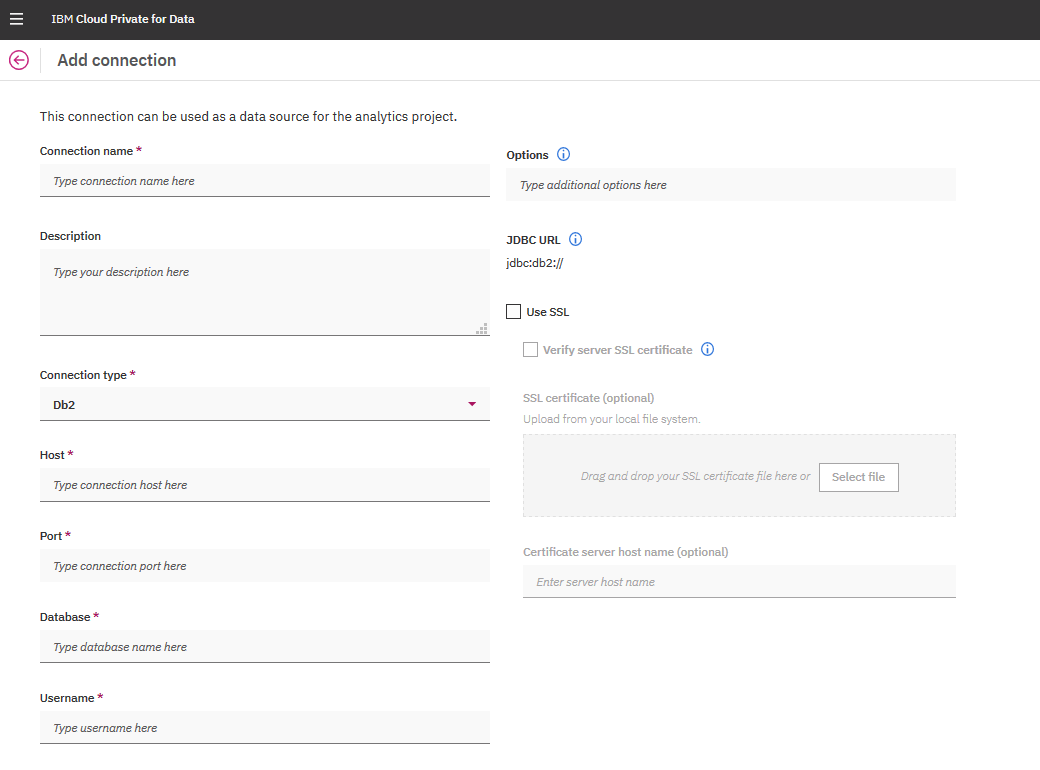


A data source could be a relational database.

To add new source, click on add source, you can use one of the existing connection in the platform or create your own :

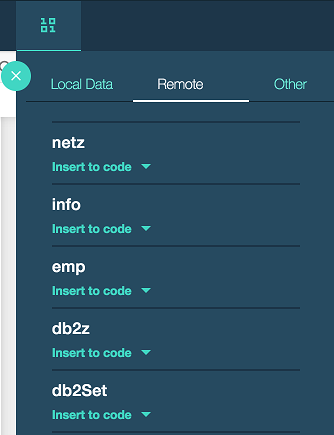


Select type of database connection required and enter driver specific details.



### Quick access to data sets

To analyze data from relational source, Cloud Pak for Data provides shortcuts that makes data available in dataframe APIs.

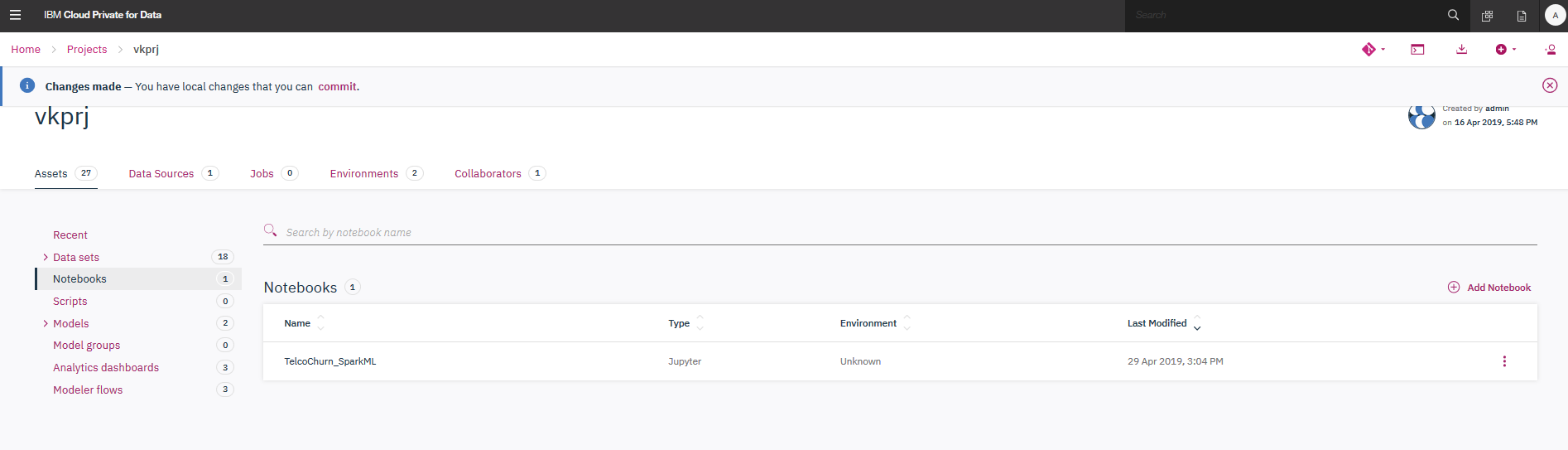


Click on desired data that has been defined in “Data source” and click on “insert to code” . This option will add a cell in Notebook interface to build analysis further.

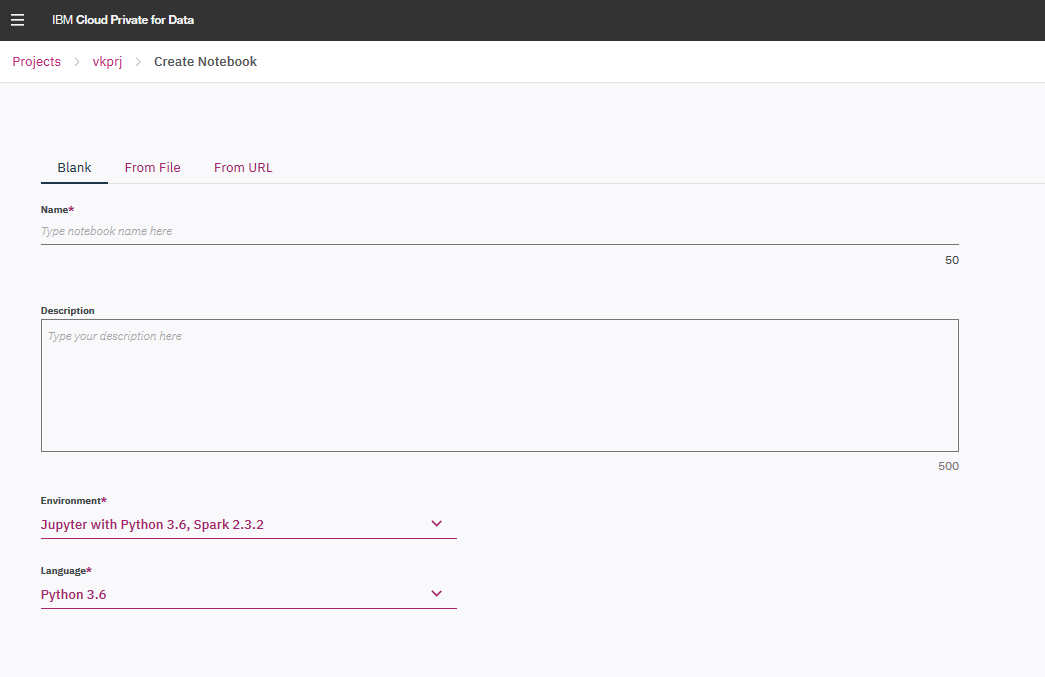
### Analyze data using - Jupyter notebooks

Jupyter notebook is an opensource REPL (read-eval-print-loop) client-server tool used by data scientists /analyst to build iteratively and interactively an analytical model. The notebook interface is accessed via web browser.

Within in Cloud Pak for Data Analytics Project, click on Notebook option :



Select “Add Notebook” to create a new blank notebook or from an example file or from open source repository using URL :



The “Create Notebook” allows you to select an environment and programming language to execute the commands within a notebook.

For details on navigation and options within Jupyter Notebook, please refer to:

<http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html>

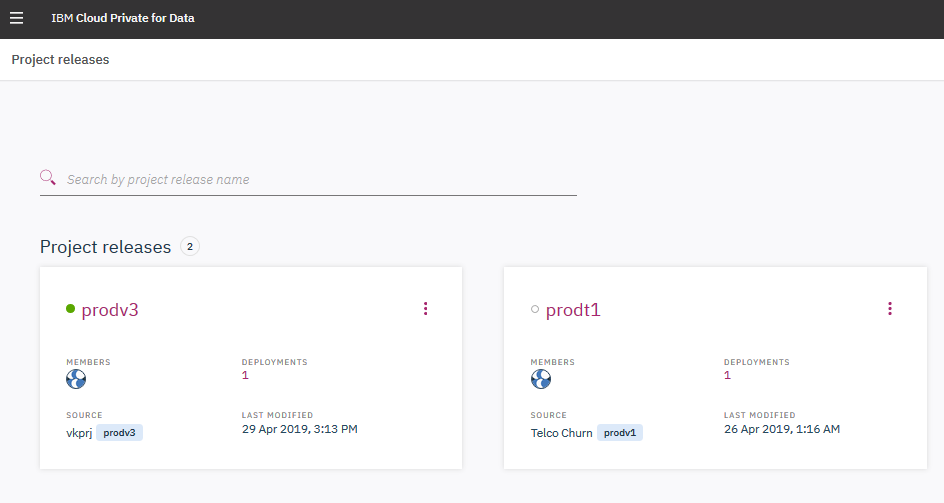
Also following blog has quick guide to Jupyter notebooks as well

<https://medium.com/ibm-data-science-experience/back-to-basics-jupyter-notebooks-dfcdc19c54bc>

Note that Cloud Pak for Data provide quick and easy access to generate code to access data assets defined within a “project”. (see “Load or Retrieve Data” section of this document)

# Manage Deployments

To expose a checkpoint of assets to outside users, a Cloud Pak for Data administrator can create a project release and deploy the assets within it. A project release represents a project tag that can be launched as a production environment within Cloud Pak for Data.



For additional info : <https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ_current/com.ibm.icpdata.doc/dsx/deployments.html>

# Installing global libraries and packages on the cluster

An IBM® Cloud Private for Data administrator can install Python or R packages in global directories so that the packages are available to all users of the cluster.

## Installing a global Python library

1. Log in as an IBM Cloud Private for Data administrator and create a Python notebook in a new or existing analytics project.
2. Use the Python pip package installer command to install Python libraries to your notebook. For example, run the following command in a code cell to install the prettyplotlib library to the Python 2.7 distribution.:

!pip install --target /user-home/\_global\_/python-2.7 prettyplotlib

The installed packages can be used by all notebook users that use the same Python version in the Spark service. Notebook users can use the Python import command to import the library components. For example, users can run the following command in a code cell:

import prettyplotlib as ppl

Installing in “air-gap” (disconnected from internet) environments use this : <https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ_current/com.ibm.icpdata.doc/zen/admin/admin-libraries.html#admin-libraries__installLibrary>

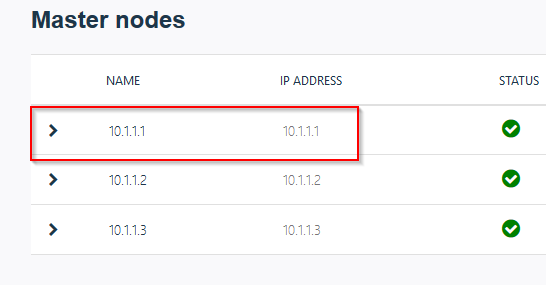
# Troubleshooting Cloud Pak for Data Platform

### Assessing Health via CLI Tools

In this exercise we would troubleshoot a problem using a mix of strategies. We would introduce a problem in the cluster by running a command and then run health check commands to see the status of the cluster. We would also use Kibana Dashboards to view and filter down to the relevant entries.

* 1. Login to Server

1. From the Cloud Pak for Data console, figure out the IP of a master node. This can be done by browsing to Administer->Nodes and looking under the Master nodes.



1. Login to master1 node with root credentials.

[ibm@ibmhost-1 ~] su –

Password: <passw0rd>

[root@ibmhost-1~]

### Health assessment via serviceability tool

1. Change Directory to Cloud Pak for Data tools and execute Cloud Pak for Data\_tools with help option. Review the various options that are available to you.

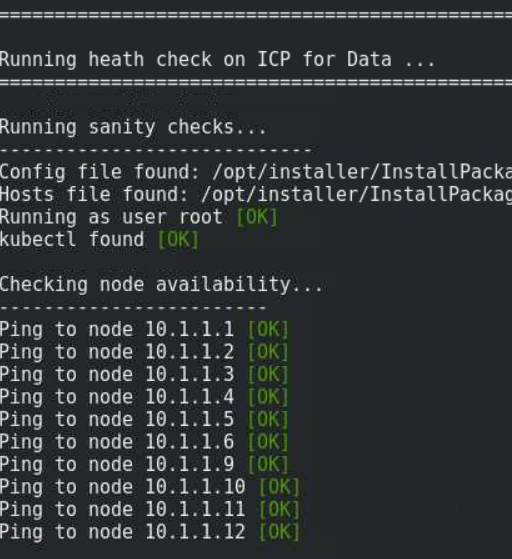
[root@ibmhost-1~] cd /ibm/InstallPackage/utils/Cloud Pak for Data-Support-Tools/

[root@ibmhost-1~] ./Cloud Pak for Data\_tools.sh --help

1. Execute the command with –health option to run a quick health check. Notice that most of the indicators are green right now.

[root@ibmhost-1~] ./Cloud Pak for Data\_tools.sh --health

The health check probes key infrastructure of Cloud Pak for Data such as gluster, docker, memory, cpu, unhealthy pods and many operating system level checks. When finished, it does generate a log in the current directory (health\_check.log)



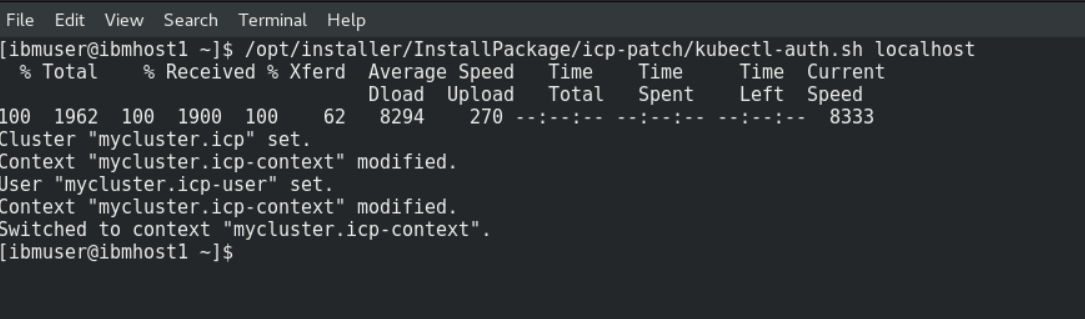
### Health assessment via platform tools

In this section we would run few commonly used commands to check the overall health of the system.

1. kubectl – kubectl is a go-to command for querying and changing the kubernetes objects such as Pods, Nodes, PVCs, Services etc.
2. Authenticate kubectl

Before we execute “kubectl” commands, we would need to authenticate kubectl by typing the below command.

/opt/installer/InstallPackage/icp-patch/kubectl-auth.sh localhost



1. Get all pods from all namespaces and from zen only.

kubectl get pods –all-namespaces

kubectl get pods -n zen

1. Get nodes which are not healthy

kubectl get po --all-namespaces | grep -Ev '1/1|2/2|3/3|4/4' | grep -v 'Completed'

1. Get logs from a pod

kubectl logs <pod-name> -n kube-system

For example,

kubectl logs icp-mongodb-1 -n kube-system

1. Describe Pod to see indications of any problems.

kubectl describe pod <pod-name> -n zen

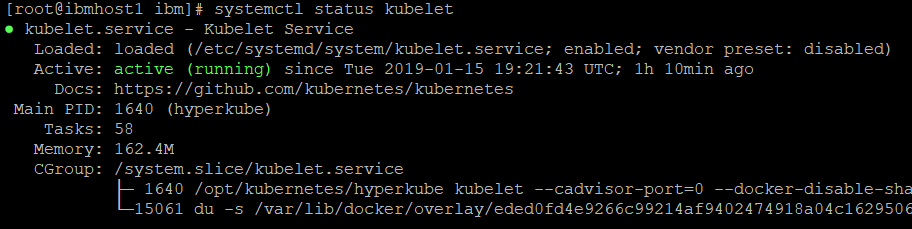
For example,

kubectl describe pod zen-metastoredb-0 -n zen

1. kubelet status

The kubelet is the primary “node agent” that runs on each node. Its one of the critical service which lets nodes communicate with each other. The command below should be in active state all the time. This can be identified by a green dot in the output.

systemctl status kubelet



1. gluster – GlusterFS is scalable network highly available filesystem used in Cloud Pak for Data. “gluster” command is the window to the glusterfs. In Cloud Pak for Data, worker nodes typically host the gluster filesystem where volumes are replicated across each node for high-availability and reliability. Troubleshooting gluster filesystem related problems would need you to run this command
2. Volume Status

Provides details of each volume and its status.

gluster volume status

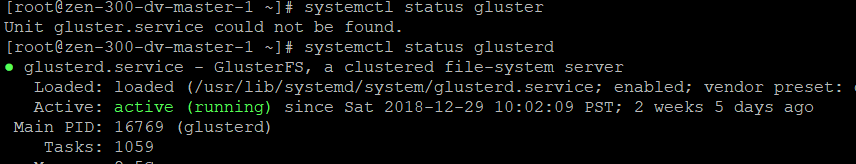
Grepping for ‘ N ‘ should not return any results. If it does, there are volumes that have problems.

gluster volume status | grep ' N '

1. Service Status

Like kubelet, Gluster volume management daemon should be up and running on each gluster node. The status should be active for the below command. This can be identified by a green dot in the output.

systemctl status glusterd



# Additional resources for Cloud Pak for Data and ML

### Product documentation:

<https://docs-icpdata.mybluemix.net/docs/content/SSQNUZ_current/com.ibm.icpdata.doc/zen/overview/overview.html>

### Community resources:

<https://community.ibm.com/community/user/icpfordata/home>

<https://medium.com/inside-machine-learning>

<https://samrose3.github.io/algorithm-explorer/>

### Learn more about Kubernetes:

<https://kubernetes.io/docs/home/?path=browse>

### Learn more about Machine Learning:

<https://cognitiveclass.ai/>