

Contents

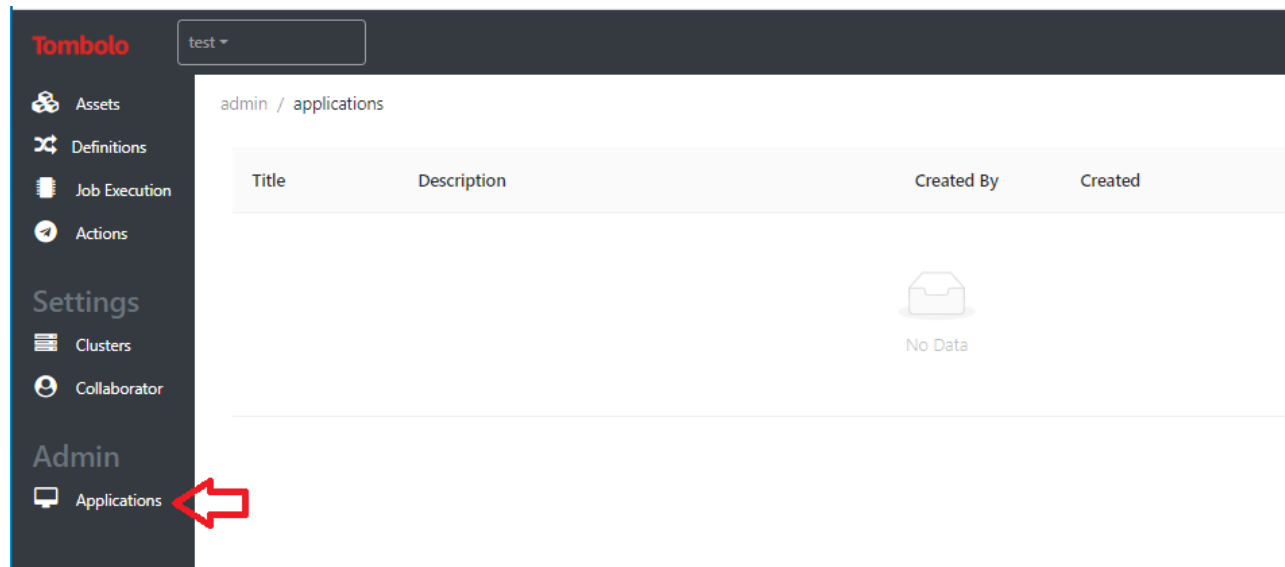
Introduction	2
Create an Application	3
Add a Cluster	4
Assets	4
Files	5
Files Details	5
File Layouts:	6
License Restrictions for files.	6
File Preview	7
Workflows – Shows the Tombolo Dataflows this file belongs to	7
Indexes	8
Basic Info	8
Source File	8
Index	9
Payload	9
Queries	10
Input Fields	10
Output Fields	11
Job	11
Input Files	12
Output files	12
RealBI-Dashboards	13
Workflow Definitions	14
Designer Controls	16
Dataflow Instances	17

Tombolo is a metadata tracking tool for HPCC Data Lake solution. It tracks the metadata around how every asset is used in a Data Lake, and the process flow as to how these assets evolve.

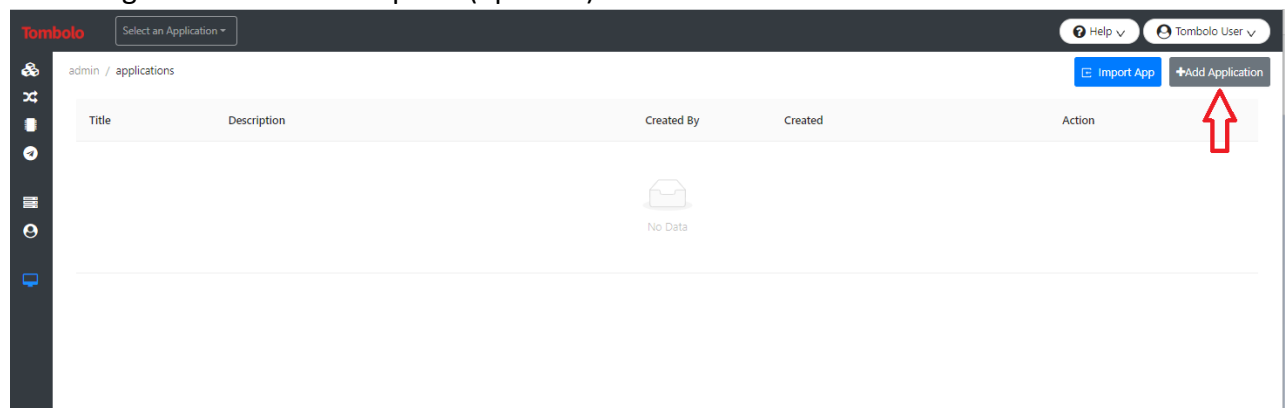
Tombolo helps you answer the following questions in a Data Lake environment.

- "Who is the owner of xyz data?"
- "What is the source of xyz data?" "What does the data contain?"
- "What are the compliance rules around xyz data?" "Who approved the usage of this data?"
- "When was this data last used?"
- "Can you show me how this data is being used?" "Is this data being handled securely?"
- "What is the impact of using this data?"
- "What happens if this data does not arrive on time?" "What happens if the data is not used on time?"

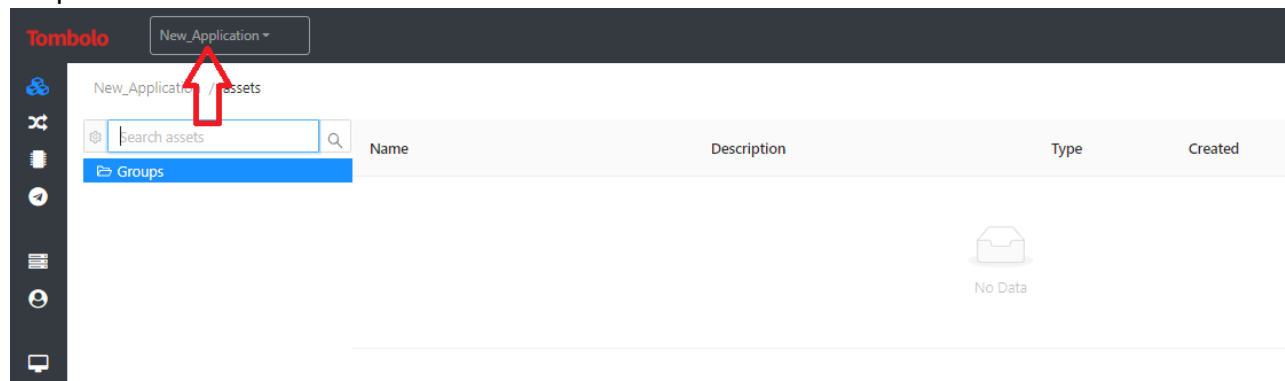
In order to start using Tombolo, an “Application” has to be created. Application is a way of grouping your assets within Tombolo. To create an application, click on the “Applications” link in the left nav. If you already have Applications, they will be listed in the Applications page



To create a new Application, click on Add Application button. Give the Application a meaningful name and description (optional)

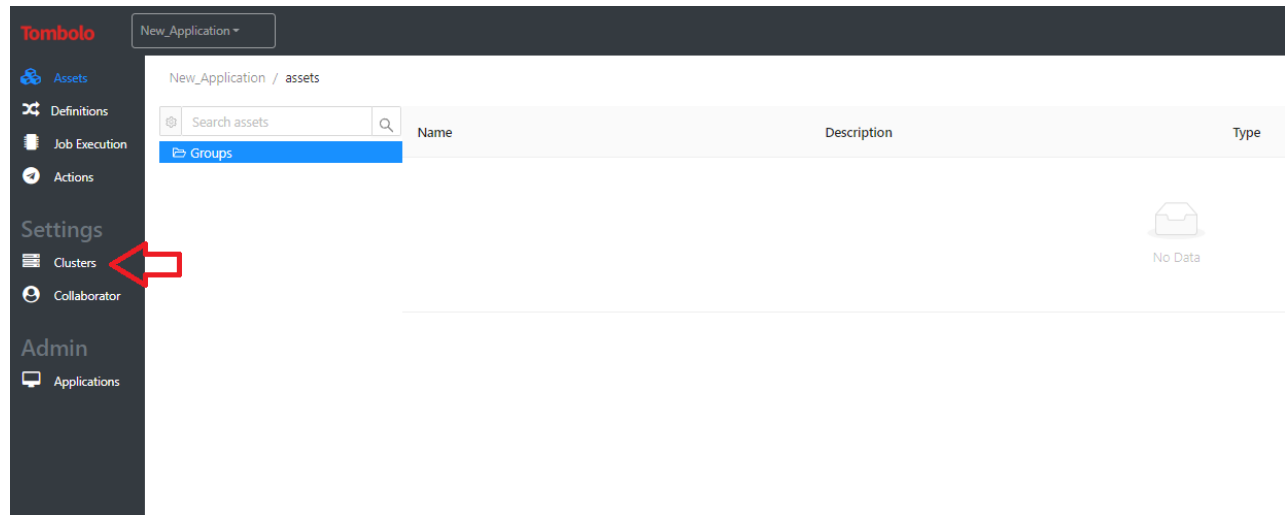


Click OK to create the Application. The Application should be now listed under the Applications dropdown.



Tombolo gives you the ability to lookup your assets directly from an HPCC cluster. You can add Clusters through the Clusters options in the navigation.

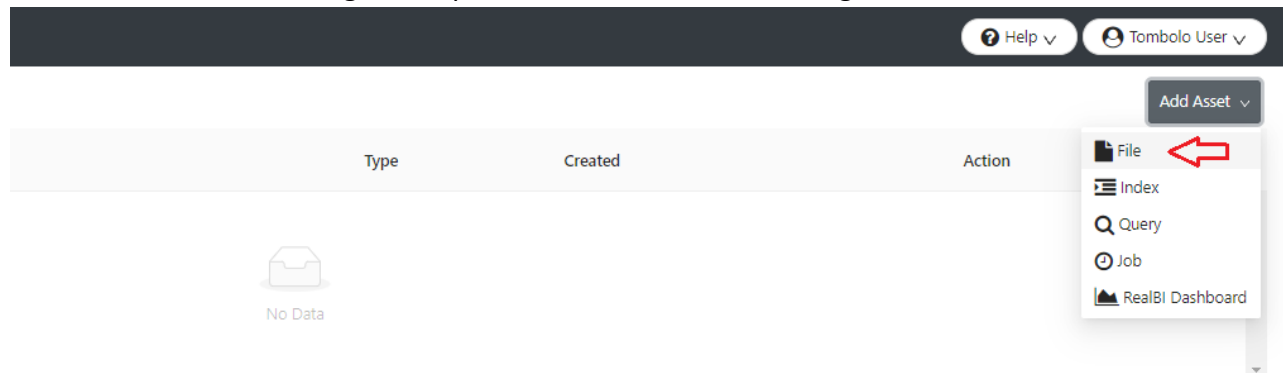
PS: The system will allow you to add only the pre-configured clusters. If you need other clusters to be added, please let us know.



Tombolo currently supports tracking metadata for the following Asset types:

- Files (Thor, CSV, JSON, XML)
- Index (HPCC)
- API/Queries (Roxie queries/other API's)
- Job (HPCC Jobs/other jobs)
- Dashboards (Visulaization)

Files can be added through File option under Assets in the navigation.



Click Add button under each asset type to add respective asset.

The screenshot shows the 'Add Asset' form in the Tombolo application. The form is titled 'New Application' and has tabs for 'Basic', 'Layout', 'Permissible Purpose', 'Validation Rules', and 'File Preview'. The 'Basic' tab is active. The form contains the following fields:

- Type:** Radio buttons for 'Thor File' (selected), 'CSV', 'JSON', and 'XML'.
- Cluster:** A dropdown menu with '4-Way' selected.
- File:** A text input field containing 'thor::test_fileopinternal:edits:part_1' and a 'Clear' button.
- * Title:** A text input field containing 'thor::test_fileopinternal:edits:part_1'.
- * Name:** A text input field containing 'thor::test_fileopinternal:edits:part_1'.
- * Scope:** A text input field containing 'thor::test_fileopinternal:edits'.
- Description:** A large text area.
- Service URL:** A text input field containing 'Service URL'.
- Path:** A text input field containing 'part_1_\$PS_of_4'.
- Is Super File:** A checkbox.
- Supplier:** A dropdown menu with 'Select a supplier'.
- Consumer:** A dropdown menu with 'Select a consumer'.
- Owner:** A dropdown menu with 'Select an Owner'.

On the right side of the form, there are two red text annotations:

- Select a cluster and start typing in the file field to look up a file from a cluster**
- The metadata information is auto-populated from HPCC when you select the file you want to add.**
- You can also manually enter these information if you wish to, instead of searching a file from a cluster**

Layouts for any files that is looked up directly from cluster will be auto populated. But you can also manually add Layout information for a file using 'Add a row' option.

The screenshot shows the Tombolo application interface. At the top, there's a header with the Tombolo logo, a 'New_Application' dropdown, and user information 'Tombolo User'. The main content area is titled 'File : Sample file'. Below this, there are tabs: 'Basic', 'Layout' (selected), 'Permissible Purpose', 'Validation Rules', 'File Preview', and 'Workflows'. To the right of these tabs are buttons: 'View Changes', 'Delete', 'Cancel', and 'Save'. The 'Layout' tab contains a table with the following data:

System Name	Name	Type	Description	Action
field4	field4	String	City	
field5	field5	String	Credit Card	
field6	field6	String	DOB	
field7	field7	String	Driver License	
field8	field8	String	E-mail	

Below the table, there's a text input field with the placeholder text 'Upload a sample file'. To the right of this field, there's a text box containing the text: 'Layouts for any files that is looked up directly from cluster will be auto populated. But you can also manually add Layout information for a file using 'Add a row' option.' At the bottom right, there's a button labeled 'Add a row'.

If you have any licensing restrictions for your files, record them here. The list of licenses are configurable in the system.

The screenshot shows the Tombolo application interface. At the top, there's a header with the Tombolo logo, a 'Covid19' dropdown, and user information 'Yadhap Dahal'. The main content area is titled 'File: Test File'. Below this, there are tabs: 'Basic', 'Layout', 'Permissible Purpose' (selected), 'Validation Rules', and 'Workflows'. To the right of these tabs are buttons: 'Edit' and 'Cancel'. The 'Permissible Purpose' tab contains a table with the following data:

Name
Creative Commons Attribution License
U.S. Government Works

A preview of data. This tab will be shown only if your Tombolo Role has access to see the file data

Tombolo Covid19 ▾ Help ▾ Yadhap Dahal ▾

File : Sample File

Basic Layout Permissible Purpose Validation Rules **File Preview** Workflows Edit Cancel

fips	country	level2	level3	date	cumcases	cumdeaths	cumhosp	tested	positive	negative
000000	AUSTRALIA	AUSTRALIA		20210924	849	3	0	0	0	0
0	AUSTRALIA	AUSTRALIA...		20210923			817	3	0	0
0	AUSTRALIA	AUSTRALIA...		20210922			798	3	0	0
0	AUSTRALIA	AUSTRALIA...		20210921			782	3	0	0
0	AUSTRALIA	AUSTRALIA...		20210919			749	3	0	0
0	AUSTRALIA	AUSTRALIA...		20210917			725	3	0	0
0	AUSTRALIA	AUSTRALIA...		20210915			680	3	0	0
0	AUSTRALIA	AUSTRALIA...		20210913			652	3	0	0

Tombolo Covid19 ▾ Help ▾ Yadhap Dahal ▾

File : Sample File

Basic Layout Permissible Purpose Validation Rules **Workflows** Edit Cancel

Title	Description
Sample Workflow	Sample workflow description

< 1 >

Click on the Index option on the left nav to view the Indexes that are already added to Tombolo. New Indexes can be added using Add button.

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Basic Source File Index Payload View Changes Delete Cancel Save

Cluster: 4-Way-2 ▾

Index: drea:testpackagemap:20160224_133544_idx Clear

* Name: drea:testpackagemap:20160224_133544_idx

* Title: 20160224_133544_idx

Description:

Primary Service: Primary Service

Backup Service: Backup Service

Path: 20160224_133544_idx_1_of_1

Indexes can be looked up from a cluster or can be manually fed. Select a cluster and start typing in the name of the index

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Basic Source File Index Payload View Changes Delete Cancel Save

Source File: us_state_vaccinations.csv

Select source file used for this index

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Basic Source File **Index** Payload View Changes Delete Cancel Save

Name	Type	Action
timestamp	String	

Add a row

key fields for the index - auto populated from the cluster

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Basic Source File Index **Payload** View Changes Delete Cancel Save

Name	Type
__internal_fpos__	Unsigned Integer

Payload fields auto-populated from the cluster

– Shows the Tombolo Dataflows this Index belongs to

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Index : Sample Index

Basic Source File Index Payload **Workflows** Edit Cancel

Title	Description
Sample Workflow	

< 1 >

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Basic Input Fields Output Fields View Changes Delete Cancel Save

Type: ☒ Roxie Query ☐ API/Gateway

Cluster: 4-Way ▾

Query: Search queries Clear

Title: Title

Name: Name

Description:

URL: URL

Git Repo: Git Repo URL

Select Roxie Query to search for a query from an HPCC cluster to retrieve basic metadata.

An external API/Endpoint can also be tracked via this tool

Tombolo New_Application ▾ Help ▾ Tombolo User ▾

Basic Input Fields Output Fields View Changes Delete Cancel Save

Name	Type	Possible Value	Value Description	Action
structure_id	string			
date_start_YYYYMMDD	number			
date_end_YYYYMMDD	number			
tz_offset_minutes	number			

Add a row

Input fields for a query are auto-populated from a cluster.

Configure allowed values for these input params. This info can be consumed by a downstream application for validation.

Tombolo

New_Application ▾

Help ▾

Tombolo User ▾

Basic

Input Fields

Output Fields

View Changes

Delete

Cancel

Save

Name	Type	Possible Value	Value Description
result_count	number		

Output fields of a query are identified automatically from the cluster. Users can also add custom fields by clicking Add a Row

Tombolo

New_Application ▾

Help ▾

Tombolo User ▾

Basic

ECL

Input Params

Input Files

Output Files

Execute Job

View Changes

Cancel

Save

Job Type:

Job Type ▾

Cluster:

4-Way ▾

Job:

Search jobs

Clear

* Name:

Name

* Title:

Title

Description:

Git Repo:

Git Repo

Entry BWR:

Entry BWR

Contact Email:

Contact

Author:

Author

Search for a job from the cluster to retrieve some metadata.

If the job source resides in a GitHub repo, you can configure that as well.

Capture contact info, author of jobs here

TomboloCovid19

HelpTombolo User

Job: Sample Job

BasicECLInput ParamsInput FilesOutput FilesWorkflows

Execute JobEditCancel

Name	Description
hpccsystems::covid19::file::raw::johnhopkins::v2::temp	
(hpccsystems::covid19::file::raw::johnhopkins::v1::03-21-2020.csv, hpccsystems::covid19::file::raw::johnhopkins::v1::03-20-2020.csv, hpccsystems::covid19::file::raw::johnhopkins::v1::03-19-2020.csv, hpccsystems::covid19::file::raw::johnhopkins::v1::03-18-2020.csv, hpccsystems::covid19::file::raw::johnhopkins::v1::03-17-2020.csv)	Input files for HPCC Jobs are auto-populated

TomboloCovid19

HelpTombolo User

Job: Sample Job

BasicECLInput ParamsInput FilesOutput FilesWorkflows

Execute JobEditCancel

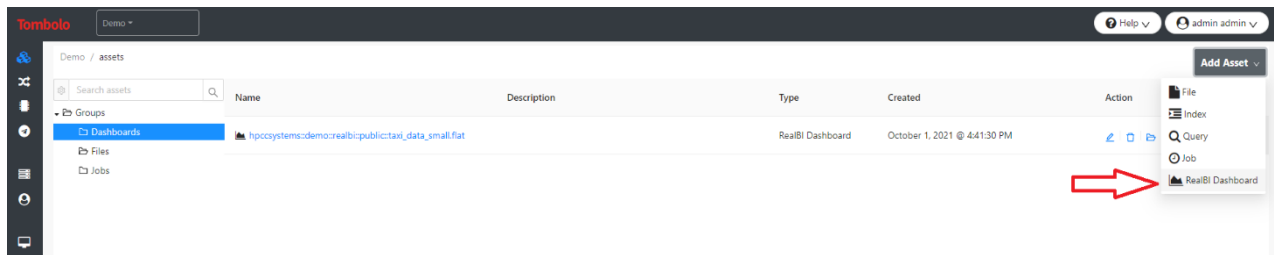
Name	Description
hpccsystems::covid19::file::public::johnhopkins::us.flat	
hpccsystems::covid19::file::public::johnhopkins::world.flat	

Output files for HPCC jobs auto-populated from the cluster. Files already existing in Tombolo can also be added here using Files dropdown

<1>

RealBI is a data visualization tool used to create Dashboards and Charts from HPCC. RealBI enables you to create data visualizations without moving your data out of HPCC. Tombolo has been integrated with RealBI to provide the ability to create RealBI Dashboards directly from assets (logical file) in Tombolo.

To create a RealBI dashboard from Tombolo, click on RealBI Dashboard option under Add Asset



To select a logical file to be used in the Dashboard, please type in the name of the file in File field, which will show a list of logical file assets stored in Tombolo that matches the name. The name will be prepopulated automatically. Once you click on Save, Tombolo passes this information to RealBI which creates an empty dashboard.

The screenshot shows the 'Basic' form in Tombolo for creating a RealBI dashboard. It includes a 'File' field with a 'Search files' button and a 'Clear' button. Below it is a '* Name' field with a 'Name' placeholder. At the bottom is a 'Description' field with a large text area.

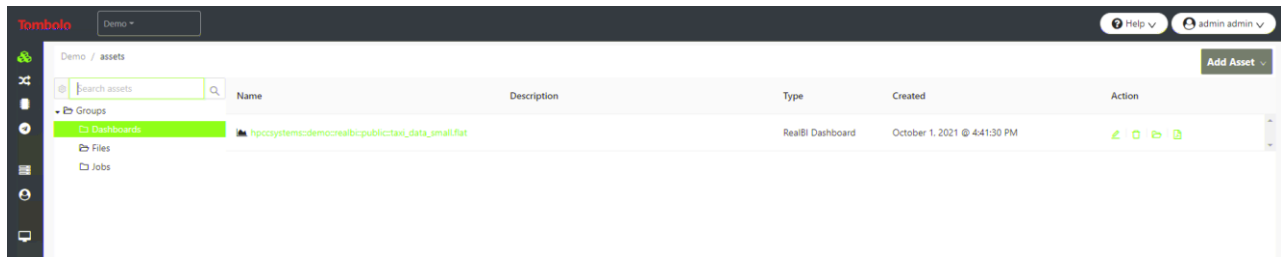
Basic

File:

* Name:

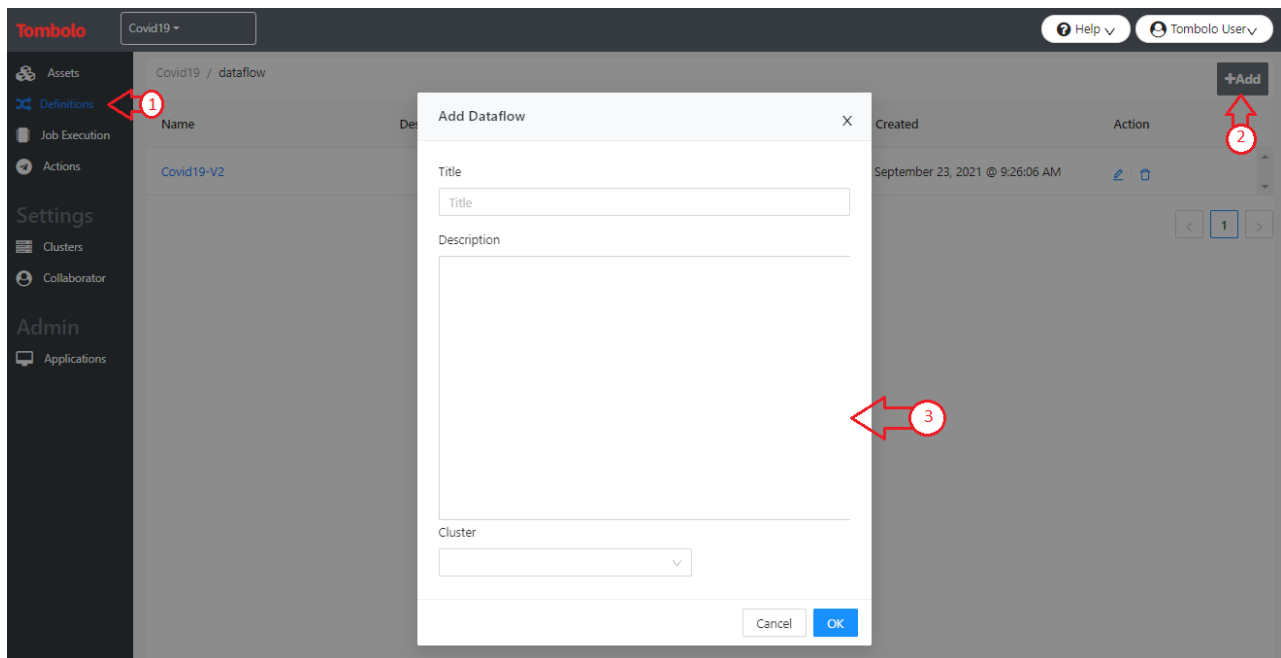
Description:

After the RealBI dashboard is created, Tombolo will show the Dashboard as an asset under the Group you selected while creating the Dashboard. Clicking on the dashboard name will directly open RealBI application in your browser.

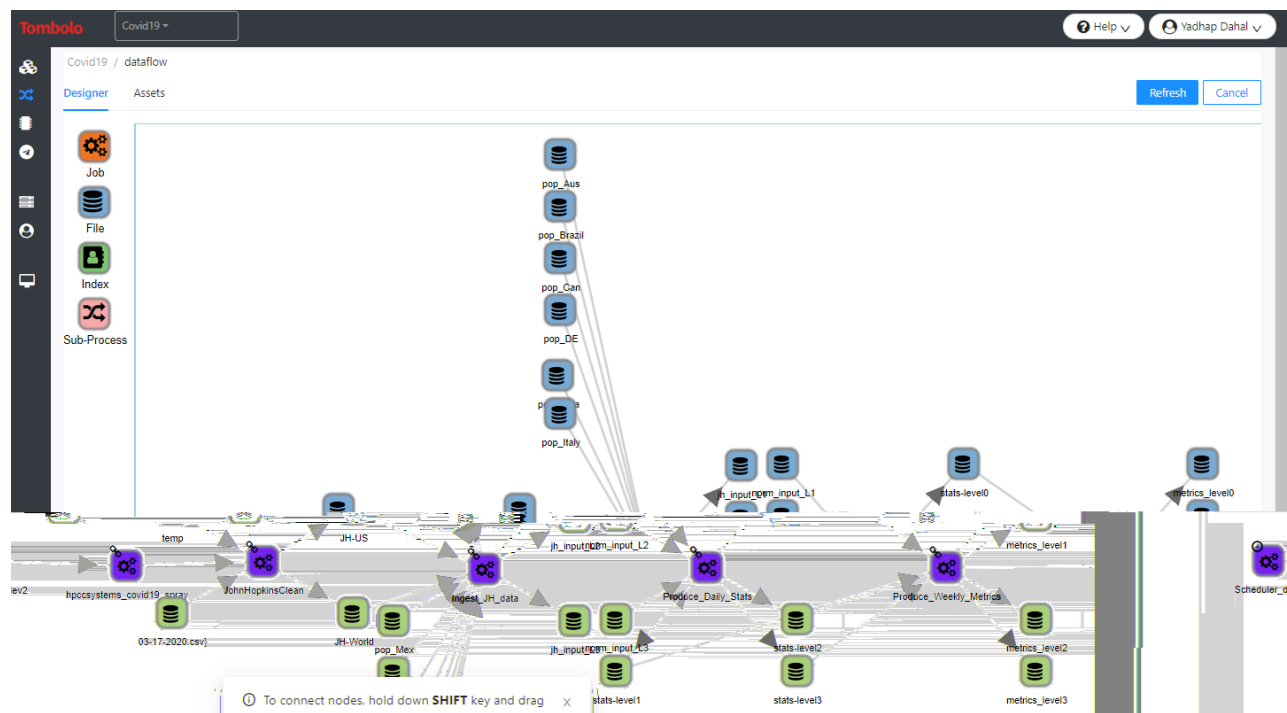


Capturing Data Lineage of a Data Lake is a key feature of Tombolo.

To create a Dataflow, click on Definitions under Workflow in the navigation. Dataflows that are already created will be listed. Click on Add and select a Cluster to which you want to point the dataflow. The cluster selection will be used later for automating tracking of workflows.



Once the Dataflow is created, click on the Dataflow name to view the Designer.



The Palette contains various nodes that are supported currently. Even though all the Jobs captures the same metadata, the idea is to capture job specific metadata in the future.

- Job – Any ECL Job
- Modelling – ML Modelling job
- Scoring – ML Scoring Job
- ETL – Any ETL job
- Data Profile – To run a Data Profile job
- Query Build – A job that builds and publishes roxie query
- File – Logical File/CSV/JSON/XML
- Index – An HPCC Index
- Sub-Process – A sub-process (child Dataflows within the main dataflow)

To use a node in the Dataflow, click on the node in the left pallet and drag it to the Designer.

The nodes can be associated with any of the asset (File/Index/Job/Query) by double clicking on it. It will then open the same Details dialog where you can either lookup an asset from a cluster or manually add the metadata.

- select the node from palette and drop to the designer
- Double click on a node
 - Keep holding Shift key and drag mouse from Source node to target node
- Hover over the node and click on trash icon
 - select the connection and press Delete button
- select the node and drag the mouse to where the node needs to be moved.
 - Place mouse on the designer and roll the scroll control on the mouse up/down

Tombolo has live workflow support to track what is happening in your workflow. Workflow tracking is done using Kafka as the integrator. This would mean that your ECL jobs will have to integrate with Kafka.

