

Tombolo User Guide

HPCC Systems Solutions Lab

Contents

Introduction	3
Create an Application	4
Add a Cluster	5
Assets	5
Files	6
Files Details	6
File Layouts:	7
License Restrictions for files.	7
File Preview	8
Workflows – Shows the Tombolo Dataflows this file belongs to	8
Indexes	9
Basic Info	9
Source File	9
Index	10
Payload	10
Queries	11
Input Fields	11
Output Fields	12
Job	12
Input Files	13
Output files	13
RealBI-Dashboards	14
Workflow Definitions	15
Designer Controls	17
Dataflow Instances	18

Introduction

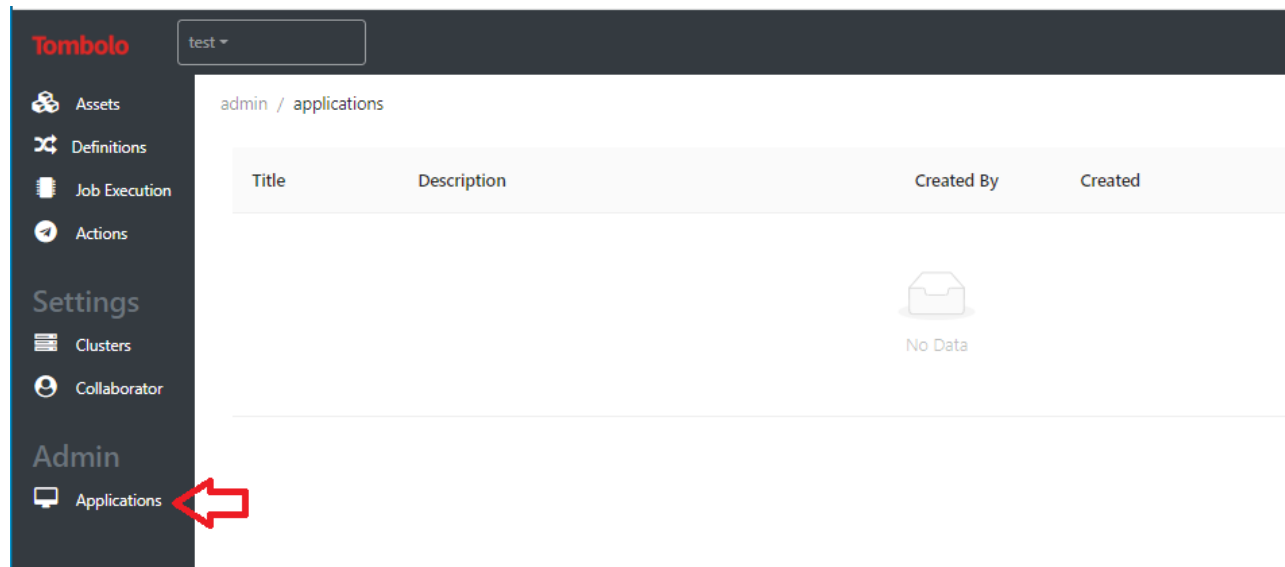
Tombolo is a metadata tracking tool for HPCC Data Lake solution. It tracks the metadata around how every asset is used in a Data Lake, and the process flow as to how these assets evolve.

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

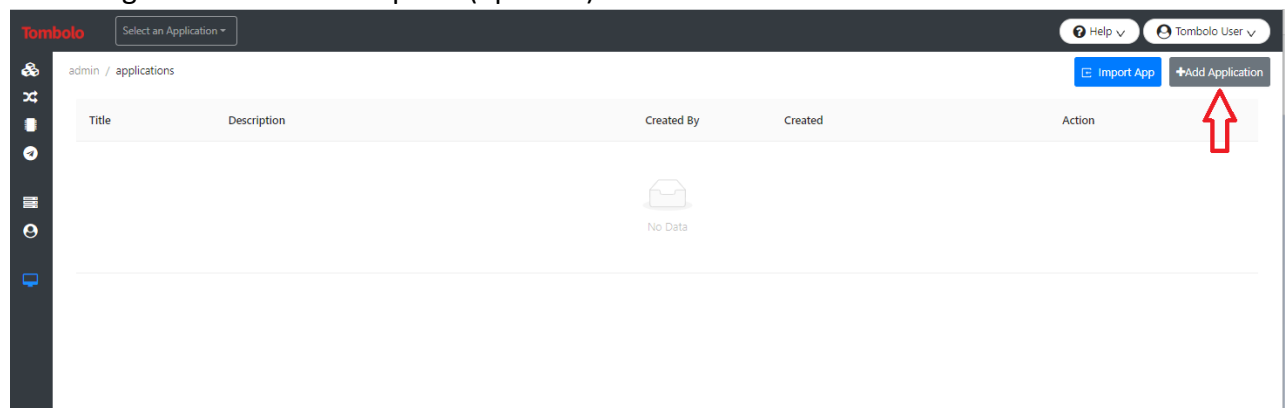
- "Who is the owner of xyz data?"
- "What is the source of xyz data?" "What does the data contain?"
- "What are the compliance rules around xyz data?" "Who approved the usage of this data?"
- "When was this data last used?"
- "Can you show me how this data is being used?" "Is this data being handled 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?"

Create an Application

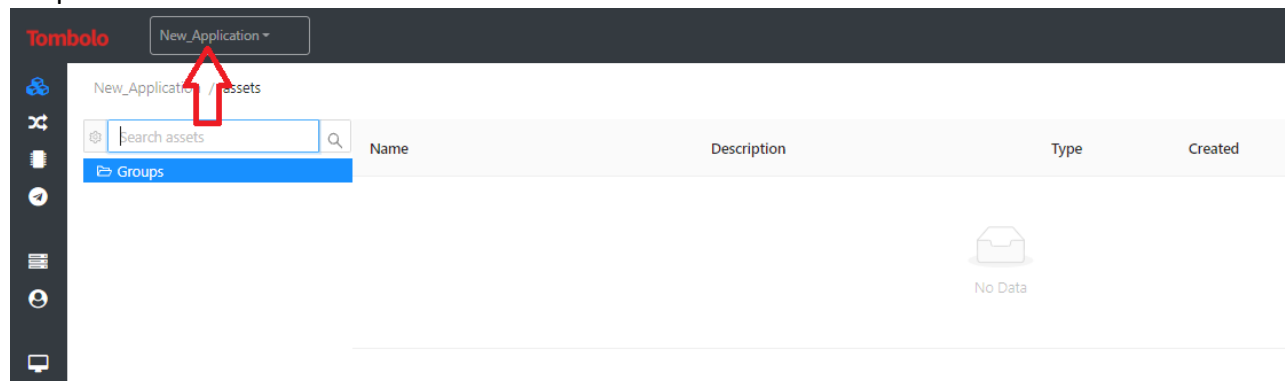
In order to start using Tombolo, an “Application” has to be created. Application is a way of grouping your assets within Tombolo. To create an application, click on the “Applications” link in the left nav. If you already have Applications, they will be listed in the Applications page



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



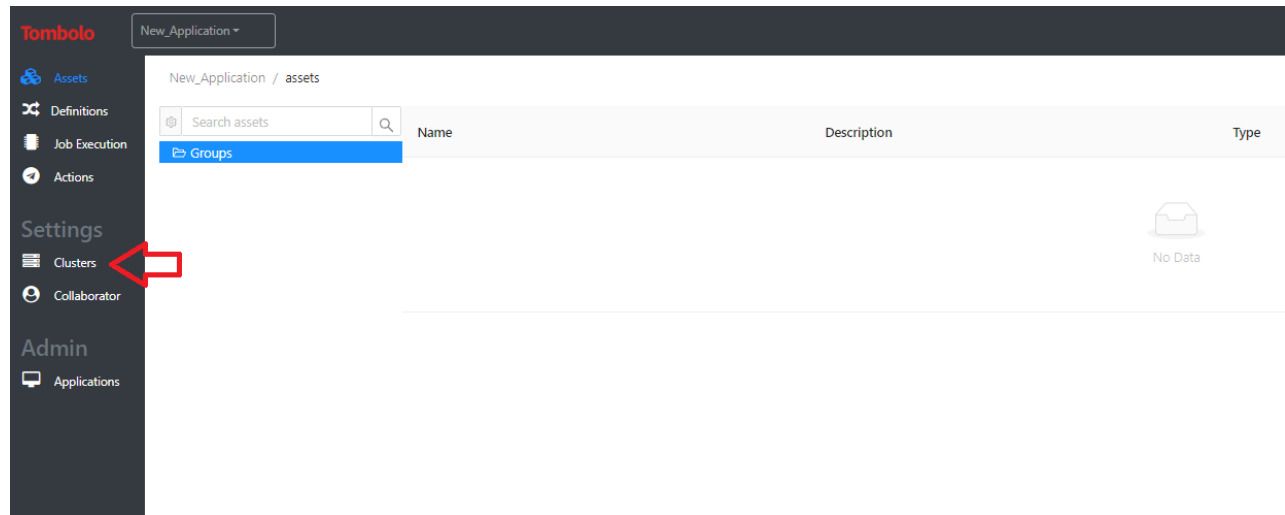
Click OK to create the Application. The Application should be now listed under the Applications dropdown.



Add a Cluster

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.



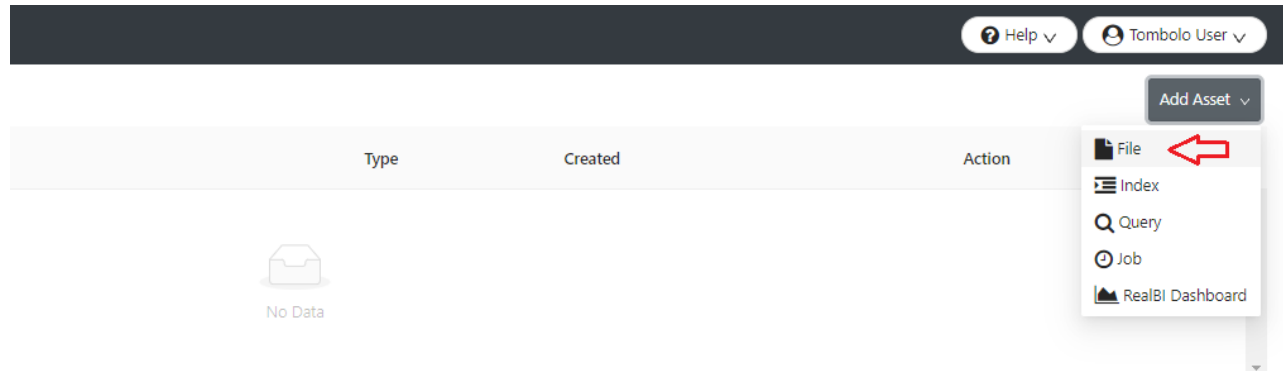
Assets

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

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



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

Files Details

The screenshot shows the 'Files Details' 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 selected. The form contains several fields: 'Type' (radio buttons for Thor File, CSV, JSON, XML), 'Cluster' (dropdown menu), 'File' (text input with a 'Clear' button), 'Title' (text input), 'Name' (text input), 'Scope' (text input), 'Description' (text area), 'Service URL' (text input), 'Path' (text input), 'Is Super File' (checkbox), 'Supplier' (dropdown menu), 'Consumer' (dropdown menu), and 'Owner' (dropdown menu). A red box highlights the 'Cluster', 'File', 'Title', 'Name', 'Scope', and 'Description' fields. To the right 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' and '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'. At the bottom right, there are buttons for 'View Changes', 'Delete', 'Cancel', and 'Save'.

File Layouts:

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 'File : Sample file' interface in the Tombolo application. The 'Layout' tab is selected, displaying a table with columns: System Name, Name, Type, Description, and Action. The table contains 7 rows of data, all with 'String' types. A red box highlights the first three columns (System Name, Name, Type) for the first six rows. Below the table, there are two buttons: 'Add a row' and 'Upload a sample file'. A red arrow points from the text 'Layout info auto populated from cluster when you look up a file' to the 'Type' column of the last row in the table.

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	
field9	field9	String	Geo Coordinates	
field10	field10	String	IP Addresses	

Layout info auto populated from cluster when you look up a file

License Restrictions for files.

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 'File: Test File' interface in the Tombolo application. The 'Permissible Purpose' tab is selected, displaying a table with a single column: Name. The table contains two rows of data: 'Creative Commons Attribution License' and 'U.S. Government Works'. The 'Permissible Purpose' tab is highlighted with a red box.

Name
Creative Commons Attribution License
U.S. Government Works

File Preview

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

The screenshot shows the Tombolo application interface. At the top, there's a header with the Tombolo logo, a dropdown menu set to 'Covid19', and user information 'Yadhap Dahal'. A sidebar on the left contains various icons. The main content area is titled 'File : Sample File' and has several tabs: 'Basic', 'Layout', 'Permissable Purpose', 'Validation Rules', 'File Preview' (which is highlighted with a red box), and 'Workflows'. To the right of the tabs are 'Edit' and 'Cancel' buttons. Below the tabs is a table with the following data:

fips	country	level2	level3	date	cumcases	cumdeaths	cumhosp	tested	positive	negative
00000	AUSTRALIA	AUSTRALIA...		20210924	849	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210923	817	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210922	798	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210921	782	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210920	765	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210919	749	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210918	742	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210917	725	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210916	710	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210915	680	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210914	665	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210913	652	3	0	0	0	0
00000	AUSTRALIA	AUSTRALIA...		20210912	630	3	0	0	0	0

Workflows – Shows the Tombolo Dataflows this file belongs to

The screenshot shows the Tombolo application interface with the 'Workflows' tab selected (highlighted with a red box). The main content area is titled 'File : Sample File' and has tabs: 'Basic', 'Layout', 'Permissable Purpose', 'Validation Rules', 'Workflows', and 'File Preview'. To the right of the tabs are 'Edit' and 'Cancel' buttons. Below the tabs is a table with the following data:

Title	Description
Sample Workflow	Sample workflow description

At the bottom right of the table, there are navigation buttons: '<', '1', and '>'.

Indexes

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.

Basic Info

The screenshot shows the 'Basic Info' tab in the Tombolo application. The interface includes a top header with the 'Tombolo' logo, a 'New_Application' dropdown, and user controls for 'Help' and 'Tombolo User'. A left sidebar contains navigation icons. The main content area has tabs for 'Basic', 'Source File', 'Index', and 'Payload'. The 'Basic' tab is active, displaying a form for creating an index. A red box highlights the 'Cluster' dropdown (set to '4-Way-2') and the 'Index' text input field (containing 'drea:testpackagemap:20160224_133544_idx'). To the right of this box, a red text note states: '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'. Below the highlighted fields, there are input fields for 'Name' (auto-filled with 'drea:testpackagemap:20160224_133544_idx'), 'Title' (containing '20160224_133544_idx'), and a large 'Description' text area. At the bottom, there are fields for 'Primary Service' (set to 'Primary Service'), 'Backup Service' (set to 'Backup Service'), and 'Path' (containing '20160224_133544_idx_1_of_1'). Action buttons 'View Changes', 'Delete', 'Cancel', and 'Save' are located in the top right of the form area.

Source File

The screenshot shows the 'Source File' tab in the Tombolo application. The interface is similar to the previous one, with the 'Source File' tab selected. A dropdown menu is open, showing a list of source files, with 'us_state_vaccinations.csv' selected. A red arrow points from the text 'Select source file used for this index' to the selected file in the dropdown. The 'Basic' tab is also visible in the background, showing the same form as before. The top header and left sidebar are consistent with the previous screenshot.

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

Payload

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

Workflows – 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 >

Queries

Tombolo

New_Application ▾

Help ▾

Tombolo User ▾

Basic

Input Fields

Output Fields

Type: ☒ Roxie Query ☐ API/Gateway

Cluster: 4-Way ▾

Query:

Title:

Name:

Description:

URL:

Git Repo:

View Changes

Delete

Cancel

Save

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

Input Fields

Tombolo

New_Application ▾

Help ▾

Tombolo User ▾

Basic

Input Fields

Output Fields

Name	Type	Possible Value	Value Description	Action
structure_id	string			<input type="button" value="✕"/>
date_start_YYYYMMDD	number			<input type="button" value="✕"/>
date_end_YYYYMMDD	number			<input type="button" value="✕"/>
tz_offset_minutes	number			<input type="button" value="✕"/>

View Changes

Delete

Cancel

Save

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.

Output Fields

The screenshot shows the 'Output Fields' tab in the Tombolo application. The top navigation bar includes the 'Tombolo' logo, a 'New_Application' dropdown, and user controls for 'Help' and 'Tombolo User'. The left sidebar contains icons for various application features. The main content area has tabs for 'Basic', 'Input Fields', and 'Output Fields', with the latter being selected. Action buttons 'View Changes', 'Delete', 'Cancel', and 'Save' are located at the top right of the main area. Below these is a table with four columns: 'Name', 'Type', 'Possible Value', and 'Value Description'. The table contains one row with 'result_count' as the name and 'number' as the type. A red text note below the table states: 'Output fields of a query are identified automatically from the cluster. Users can also add custom fields by clicking Add a Row'.

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

Job

The screenshot shows the 'Job' configuration page in the Tombolo application. The top navigation bar and left sidebar are consistent with the previous screenshot. The main content area has tabs for 'Basic', 'ECL', 'Input Params', 'Input Files', and 'Output Files', with 'Basic' being selected. Action buttons 'Execute Job', 'View Changes', 'Cancel', and 'Save' are at the top right. The form contains several fields: 'Job Type' (dropdown), 'Cluster' (dropdown), 'Job' (text input with a 'Clear' button), '* Name' (text input), '* Title' (text input), and 'Description' (text area). Below these are 'Git Repo' (text input), 'Entry BWR' (text input), 'Contact Email' (text input), and 'Author' (text input). Red boxes highlight the 'Job Type' and 'Cluster' dropdowns, the 'Git Repo' input, and the 'Contact Email' and 'Author' inputs. Red text annotations provide context: 'Search for a job from the cluster to retrieve some metadata.' points to the 'Job Type' and 'Cluster' dropdowns; 'If the job source resides in a GitHub repo, you can configure that as well.' points to the 'Git Repo' input; and 'Capture contact info, author of jobs here' points to the 'Contact Email' and 'Author' inputs.

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

Input Files

The screenshot shows the Tombolo web interface. At the top, there's a header with the Tombolo logo, a 'Covid19' dropdown, and user controls for 'Help' and 'Tombolo User'. The main area is titled 'Job: Sample Job' and has tabs for 'Basic', 'ECL', 'Input Params', 'Input Files' (selected), 'Output Files', and 'Workflows'. On the right, there are buttons for 'Execute Job', 'Edit', and 'Cancel'. The 'Input Files' tab displays a table with two columns: 'Name' and 'Description'. The 'Name' column contains a list of file paths: 'hpccsystems::covid19::file::raw::johnhopkins::v2::temp' and a multi-line path for 'hpccsystems::covid19::file::raw::johnhopkins::v1::03-21-2020.csv'. The 'Description' column contains the text 'Input files for HPCC Jobs are auto-populated'. A red box highlights the file paths in the 'Name' column.

Name	Description
hpccsystems::covid19::file::raw::johnhopkins::v2::temp	Input files for HPCC Jobs are auto-populated
(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)	

Output files

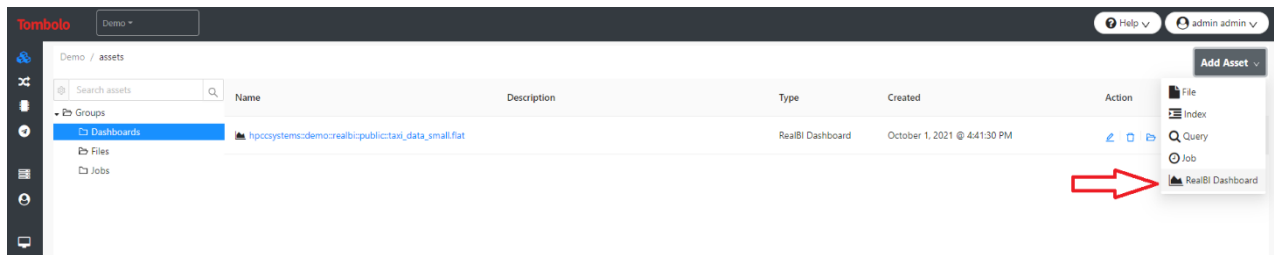
The screenshot shows the Tombolo web interface with the 'Output Files' tab selected. The layout is similar to the previous screenshot, but the 'Output Files' tab is active. The table in the 'Output Files' tab has two columns: 'Name' and 'Description'. The 'Name' column contains two file paths: 'hpccsystems::covid19::file::public::johnhopkins::us.flat' and 'hpccsystems::covid19::file::public::johnhopkins::world.flat'. The 'Description' column contains the text 'Output files for HPCC jobs auto-populated from the cluster. Files already existing in Tombolo can also be added here using Files dropdown'. A red box highlights the file paths in the 'Name' column. At the bottom right, there are navigation buttons: '<', '1', and '>'.

Name	Description
hpccsystems::covid19::file::public::johnhopkins::us.flat	Output files for HPCC jobs auto-populated from the cluster. Files already existing in Tombolo can also be added here using Files dropdown
hpccsystems::covid19::file::public::johnhopkins::world.flat	

RealBI-Dashboards

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' configuration page in Tombolo. It has three main input fields: 'File' with a search bar and a 'Clear' button, '* Name' with a text input, and 'Description' with a large text area.

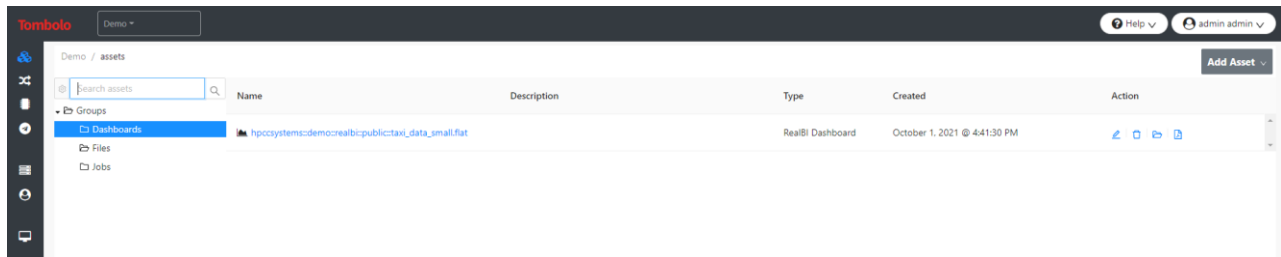
Basic

File:

* Name:

Description:

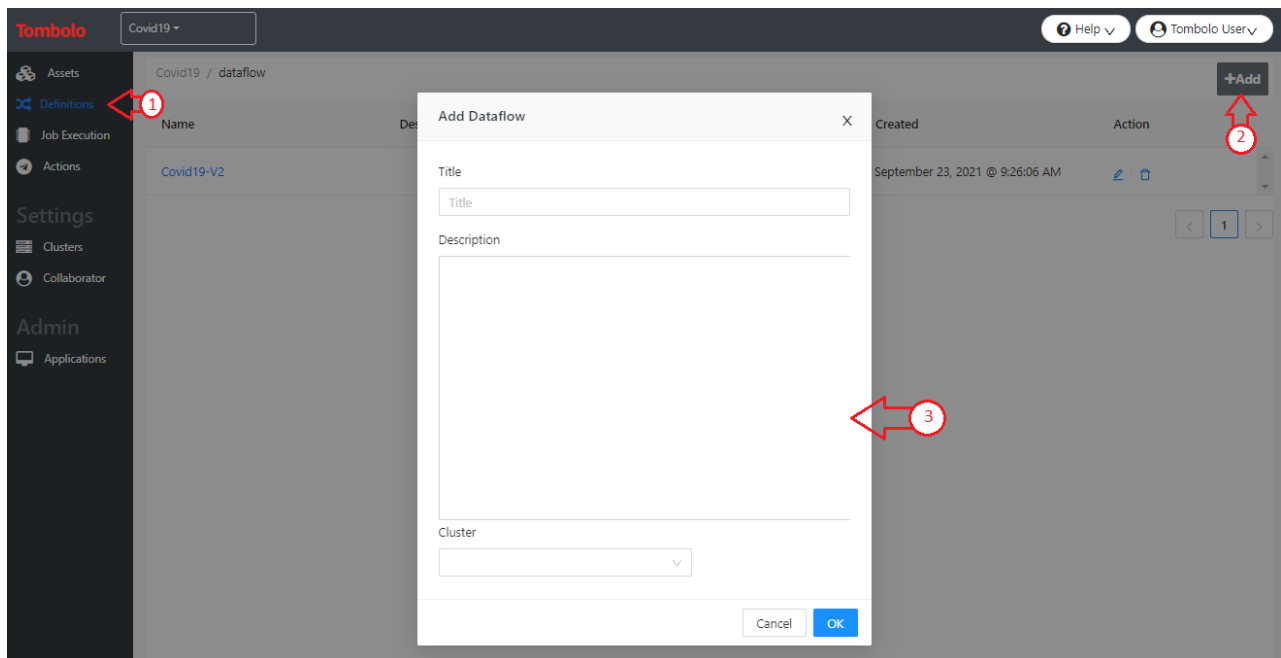
After the RealBI dashboard is created, Tombolo will show the Dashboard as an asset under the Group you selected while creating the Dashboard. Clicking on the dashboard name will directly open RealBI application in your browser.



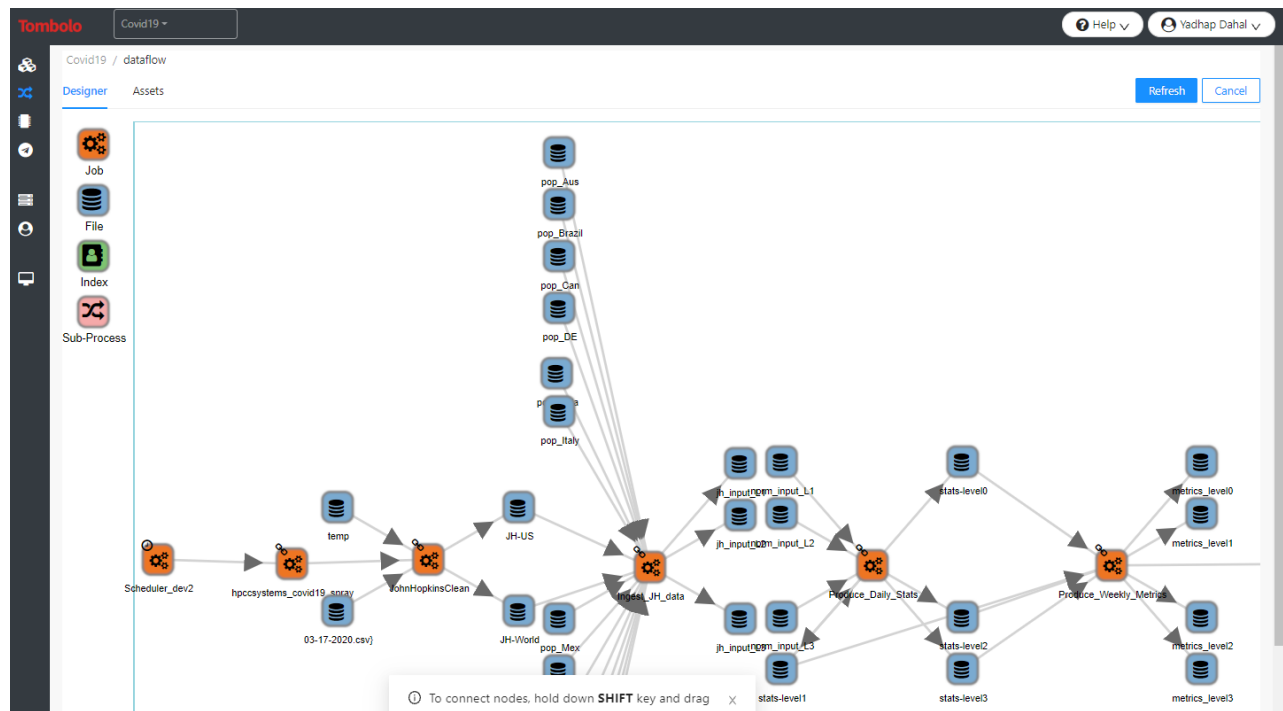
Workflow Definitions

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.

Designer Controls

[Add a node to the designer](#) – select the node from palette and drop to the designer

[Add node details](#) – Double click on a node

[Connecting nodes](#) – Keep holding Shift key and drag mouse from Source node to target node

[Delete a node](#) – Hover over the node and click on trash icon

[Delete a connection](#) – select the connection and press Delete button

[Move a node](#) – select the node and drag the mouse to where the node needs to be moved.

[Zoom in/Zoom out](#) – Place mouse on the designer and roll the scroll control on the mouse up/down

Dataflow Instances

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.

