

Tombolo User Guide

HPCC Systems Solutions Lab

Contents

Introduction	4
Create an Application.....	5
Add a Cluster	6
Assets	6
Files	7
Files Details	7
File Layouts	8
License Restrictions for files.....	8
File Preview	9
Workflows	9
Indexes	10
Basic Info	10
Source File	10
Index fields	11
Payload fields	11
Dataflows	11
Queries	12
Input Fields.....	12
Output Fields.....	13
Functions.....	13
Input Files.....	14
Output files	14
Workflow Definitions	15
Designer Controls.....	16
Add a node to the designer.....	16
Add node details	16
Connecting nodes	16
Delete a node	16
Delete a connection	16
Move a node	16
Zoom in/Zoom out	16
Dataflow Instances.....	17

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?"

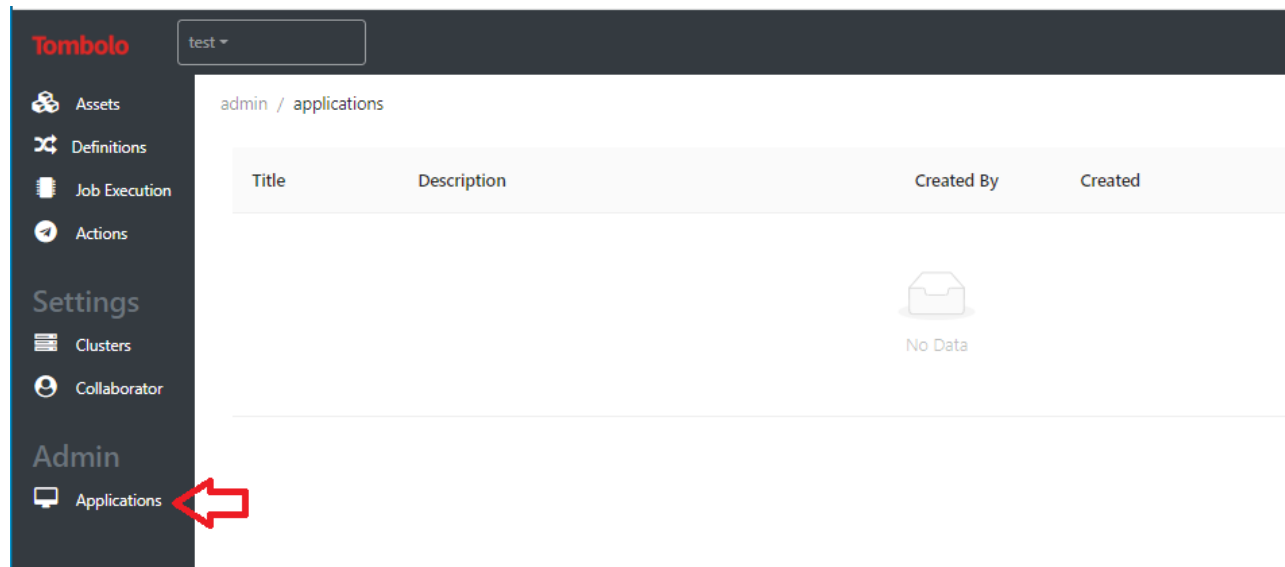
"What is the impact of using this data?"

"What happens if this data does not arrive on time?"

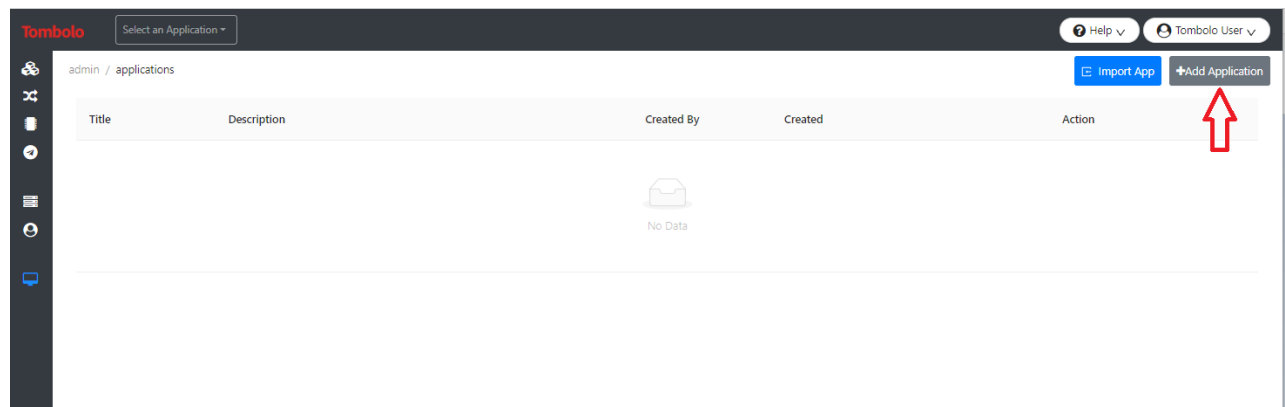
"What happens if the data is not used on time?"

Create an Application

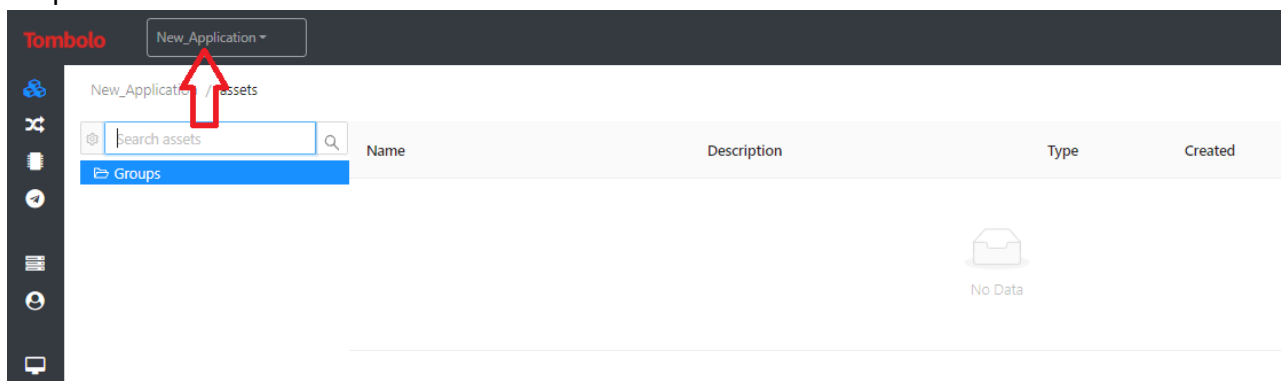
In order to start using Tombolo, an “Application” has to be created. Application is a way of grouping your assets within Tombolo. To create an application, click on the “Applications” link in the left nav. If you already have Applications, they will be listed in the Applications page



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



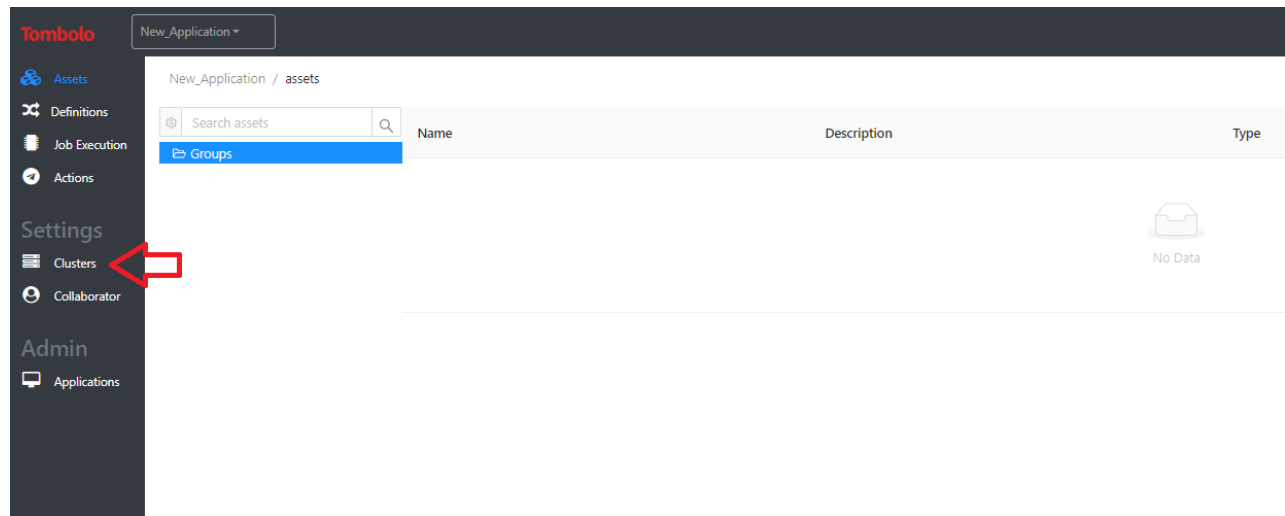
Click OK to create the Application. The Application should be now listed under the Applications dropdown.



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.



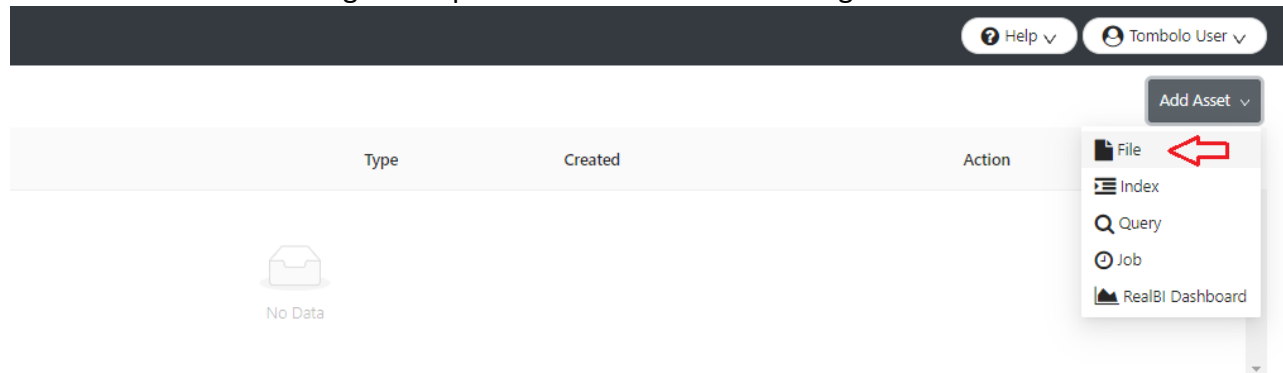
Assets

Tombolo currently supports tracking metadata for the following Asset types:

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

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 divided into several sections. At the top, there is a 'Basic' tab selected, with other tabs like 'Layout', 'Permissible Purpose', 'Validation Rules', and 'File Preview'. Below the tabs, there are buttons for 'View Changes', 'Delete', 'Cancel', and 'Save'. The form contains the following fields:

- Type:** Radio buttons for 'Thor File' (selected), 'CSV', 'JSON', and 'XML'.
- Cluster:** A dropdown menu showing '4-Way'.
- 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 showing 'Select a supplier'.
- Consumer:** A dropdown menu showing 'Select a consumer'.
- Owner:** A dropdown menu showing 'Select an Owner'.

Red annotations are present on the form:

- A red box highlights the 'Cluster', 'File', 'Title', 'Name', 'Scope', and 'Description' fields.
- A red arrow points to the 'File' field with the text: "Select a cluster and start typing in the file field to look up a file from a cluster".
- Red text states: "The metadata information is auto-populated from HPCC when you select the file you want to add."
- Red text states: "You can also manually enter these information if you wish to, instead of searching a file from a cluster".

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' configuration page in the Tombolo application. The 'Layout' tab is selected, displaying a table with columns: System Name, Name, Type, Description, and Action. The table contains seven rows of data, with the first three rows (field4, field5, field6) highlighted by a red box. Below the table, there are two buttons: 'Add a row' and 'Upload a sample file', also highlighted by a red box. A red arrow points from the text 'Layout info auto populated from cluster when you look up a file' to the 'Add a row' button.

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		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' configuration page in the Tombolo application. The 'Permissible Purpose' tab is selected, displaying a list of licenses. The list includes 'Creative Commons Attribution License' and 'U.S. Government Works'. The 'Permissible Purpose' tab is highlighted by 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 interface with the 'File Preview' tab selected. The interface includes a top header with the Tombolo logo, a 'Covid19' dropdown, and user information 'Yadhap Dahal'. A left sidebar contains various icons. The main content area is titled 'File : Sample File' and features a tabbed interface with 'Basic', 'Layout', 'Permissable Purpose', 'Validation Rules', 'File Preview' (highlighted with a red box), and 'Workflows'. The 'File Preview' tab displays 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 interface with the 'Workflows' tab selected. The interface includes the same top header and left sidebar as the previous screenshot. The main content area is titled 'File : Sample File' and features a tabbed interface with 'Basic', 'Layout', 'Permissable Purpose', 'Validation Rules', 'Workflows' (highlighted with a red box), and 'File Preview'. The 'Workflows' tab displays 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.

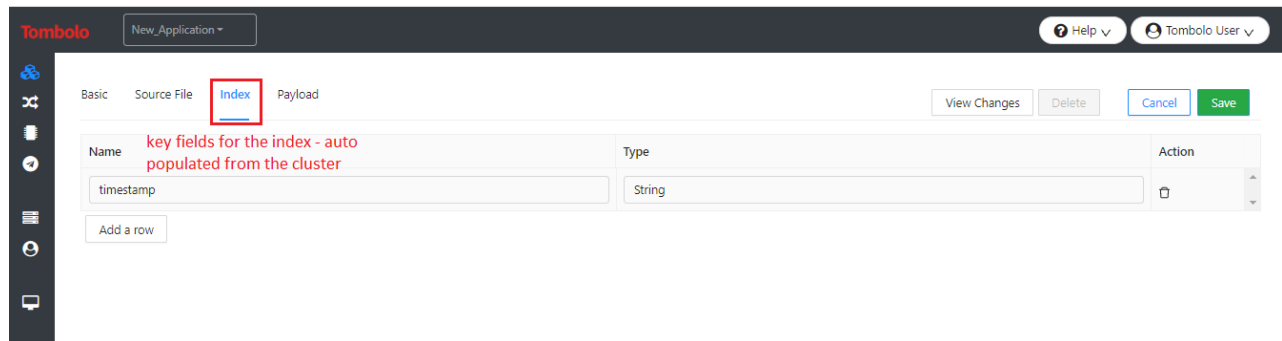
Basic Info

The screenshot shows the 'Basic Info' tab in the Tombolo application. The interface includes a top navigation bar with 'Tombolo' and a 'New_Application' dropdown. A left sidebar contains icons for various functions. The main content area has tabs for 'Basic', 'Source File', 'Index', and 'Payload'. The 'Basic' tab is active, showing fields for 'Cluster' (set to '4-Way-2'), 'Index' (containing 'drea:testpackagemap:20160224_133544_idx'), 'Name', 'Title', 'Description', 'Primary Service', 'Backup Service', and 'Path'. A red box highlights the 'Cluster' and 'Index' fields. A red text note on the right 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'. Action buttons 'View Changes', 'Delete', 'Cancel', and 'Save' are at the top right.

Source File

The screenshot shows the 'Source File' tab in the Tombolo application. The 'Source File' tab is active, displaying a dropdown menu with 'us_state_vaccinations.csv' selected. A red arrow points to this dropdown with the text 'Select source file used for this index'. The interface also shows the 'Basic', 'Index', and 'Payload' tabs, and the same top navigation and sidebar as the previous screenshot. Action buttons 'View Changes', 'Delete', 'Cancel', and 'Save' are present at the top right.

Index



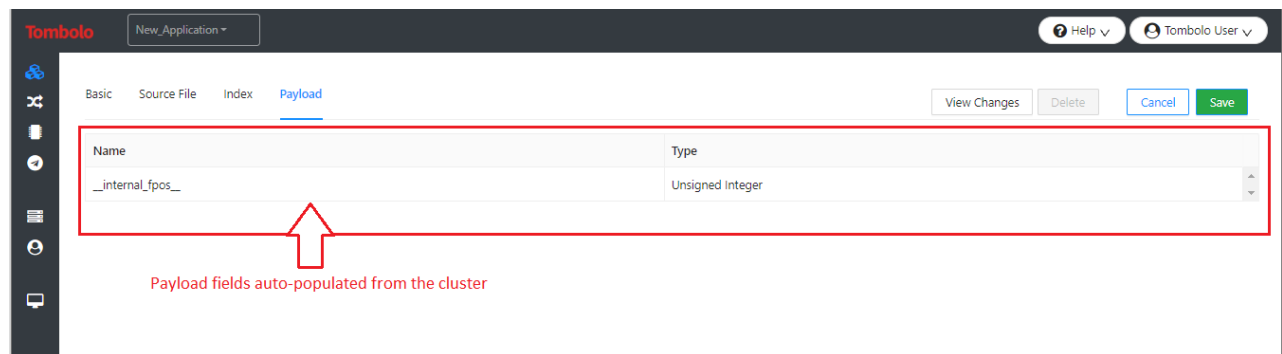
Tombolo New_Application ▾ Help ▾ Tombolo User ▾

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

Name	Type	Action
timestamp	String	

Add a row

Payload



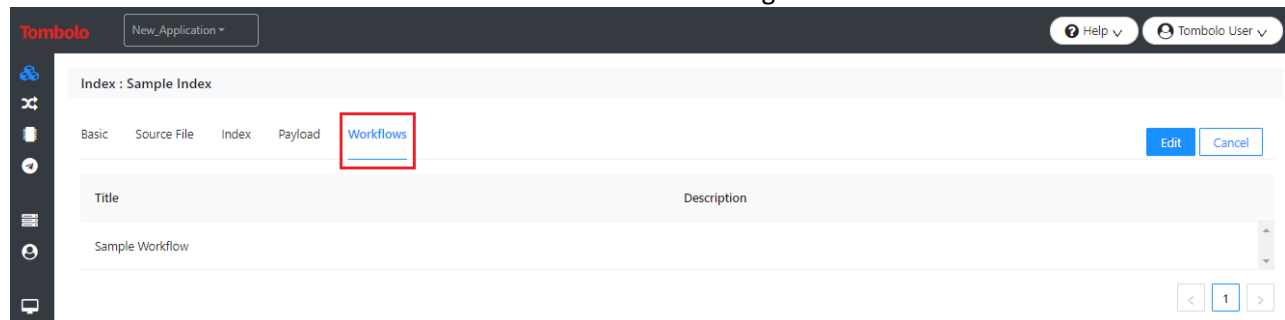
Tombolo New_Application ▾ Help ▾ Tombolo User ▾

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 ▾

BasicInput FieldsOutput Fields

View ChangesDeleteCancelSave

Type: ☒ Roxie Query ☐ API/Gateway

Cluster: 4-Way ▾

Query: Clear

Title:

Name:

Description:

URL:

Git Repo:

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 ▾

BasicInput FieldsOutput Fields

View ChangesDeleteCancelSave

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.

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. A table lists the output fields:

Name	Type	Possible Value	Value Description
result_count	number		

Below the table, a red text note states: "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 displays 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' selected. Action buttons 'Execute Job', 'View Changes', 'Cancel', and 'Save' are at the top right. The configuration form includes several fields:

- Job Type:** A dropdown menu with 'Job Type' selected.
- Cluster:** A dropdown menu with '4-Way' selected.
- Job:** A text input field containing 'Search jobs' and a 'Clear' button.
- * Name:** A text input field with 'Name'.
- * Title:** A text input field with 'Title'.
- Description:** A large text area.
- Git Repo:** A text input field with 'Git Repo'.
- Entry BWR:** A text input field with 'Entry BWR'.
- Contact Email:** A text input field with 'Contact'.
- Author:** A text input field with 'Author'.

Red boxes highlight the 'Job Type' and 'Cluster' dropdowns, the 'Git Repo' field, and the 'Contact Email' and 'Author' fields. Red text annotations provide context:

- Next to the Job Type and Cluster dropdowns: "Search for a job from the cluster to retrieve some metadata."
- Next to the Git Repo field: "If the job source resides in a GitHub repo, you can configure that as well."
- Next to the Contact Email and Author fields: "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' (which is 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, including 'hpccsystems::covid19::file::raw::johnhopkins::v2::temp' and a list of CSV files for various dates in 2020. 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

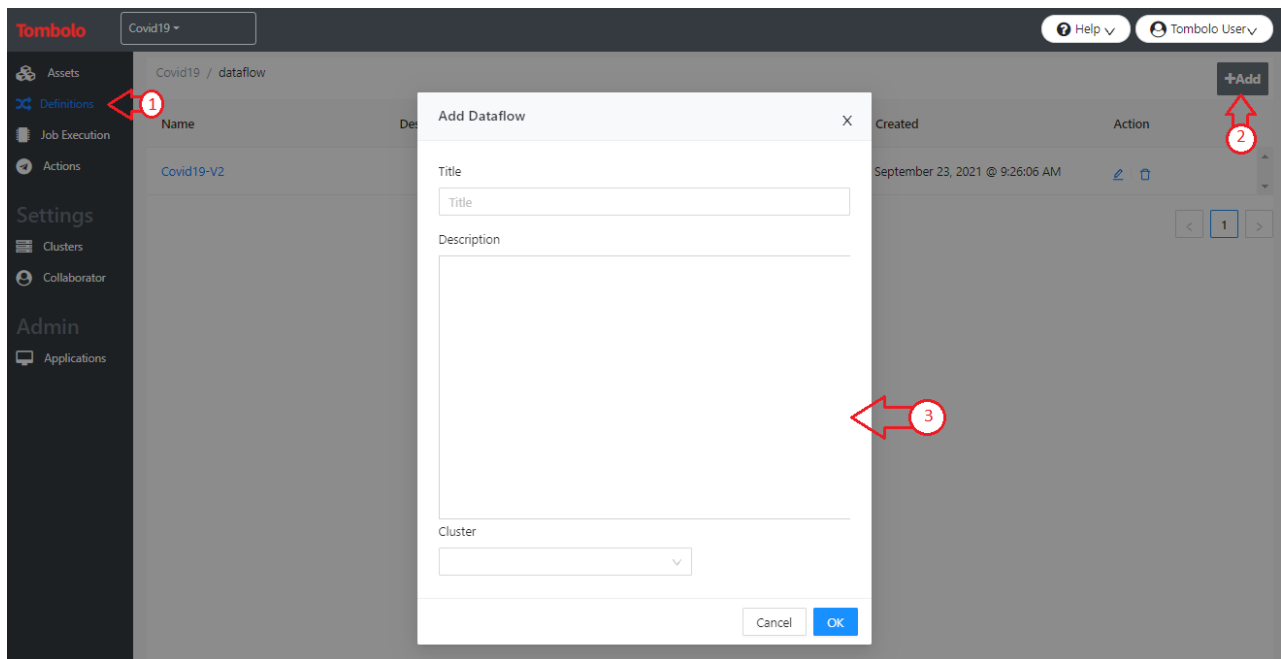
The screenshot shows the Tombolo web interface with the 'Output Files' tab selected. The header and navigation tabs are the same as in the previous screenshot. The 'Output Files' tab displays a table with two columns: 'Name' and 'Description'. The 'Name' column contains two file paths: 'hpccsystems::covid19::file::public::johnhopkins::us.flat' 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	

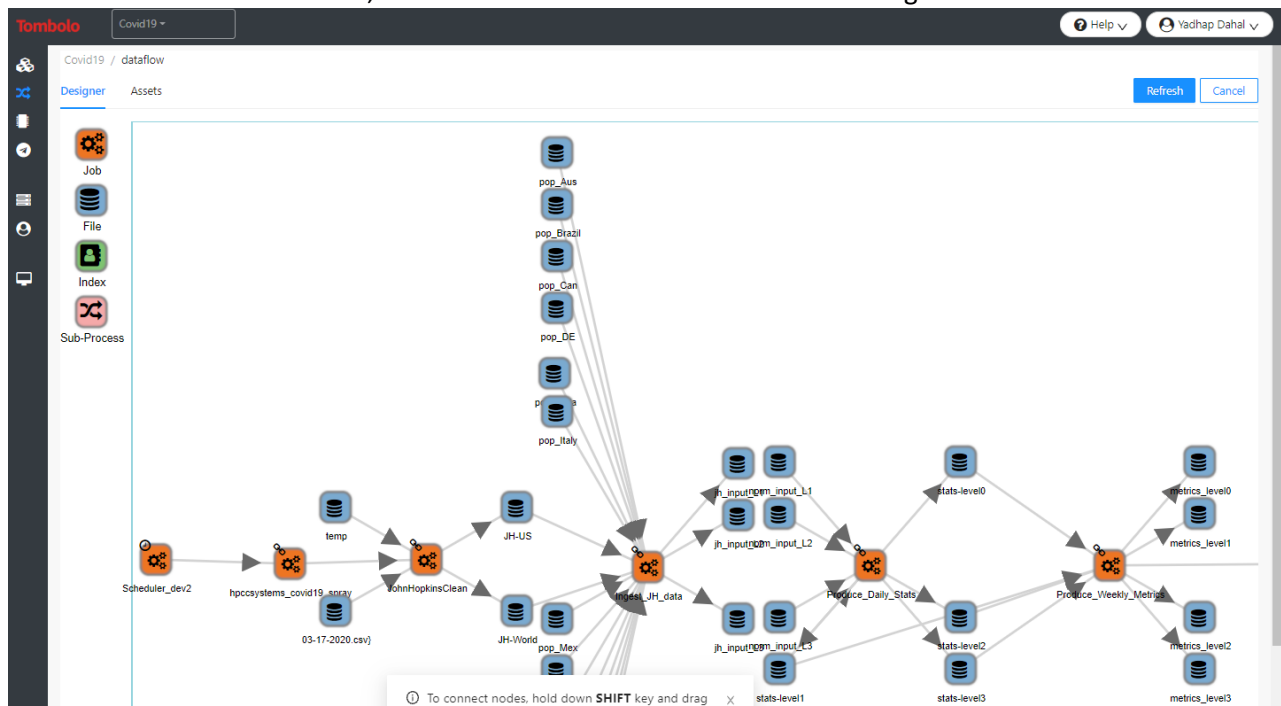
Workflow Definitions

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.

