

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
Scrubs (WIP).....	9
File Preview	9
Dataflows	9
Indexes	10
Basic Info	10
Source File.....	10
Index fields	11
Payload fields	11
Dataflows	11
API/Gateway/Queries	12
Input Fields.....	12
Output Fields.....	13
Job/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?"

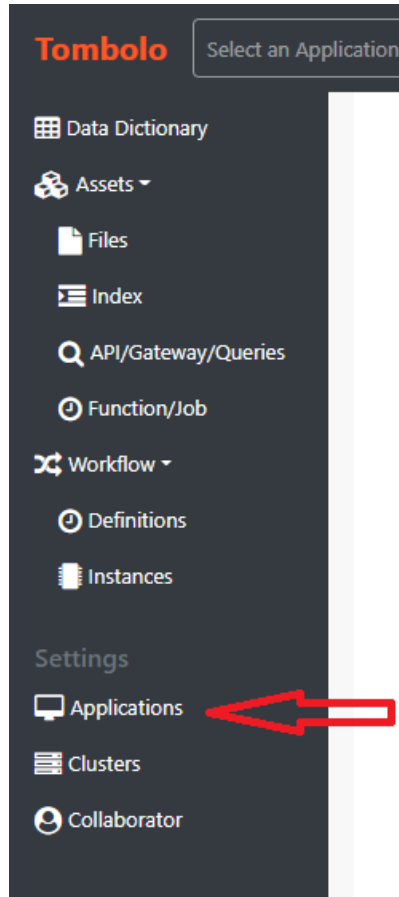
"What is the impact of using this data?"

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

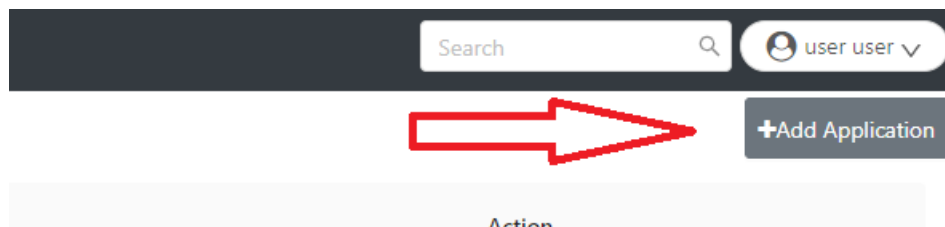
"What happens if the data is not used on time?"

Create an Application

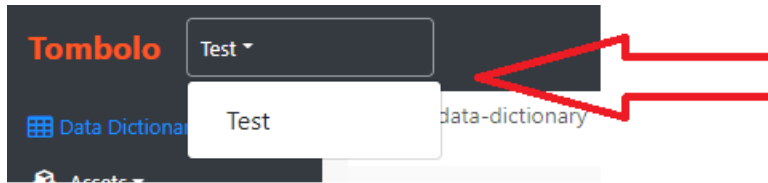
In order to start using Tombolo, an “Application” has to be created. Application is a way of grouping your assets with in Tombolo. To create an application, click on the “Applications” link in the left nav. If you already have Applications, they will be listed in the Applications page



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



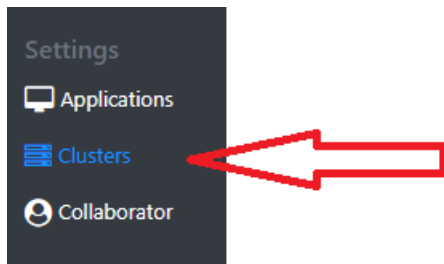
Click OK to create the Application. The Application should be now listed under the Applications dropdown.



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.



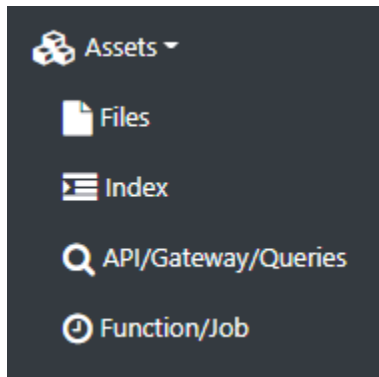
Assets

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)

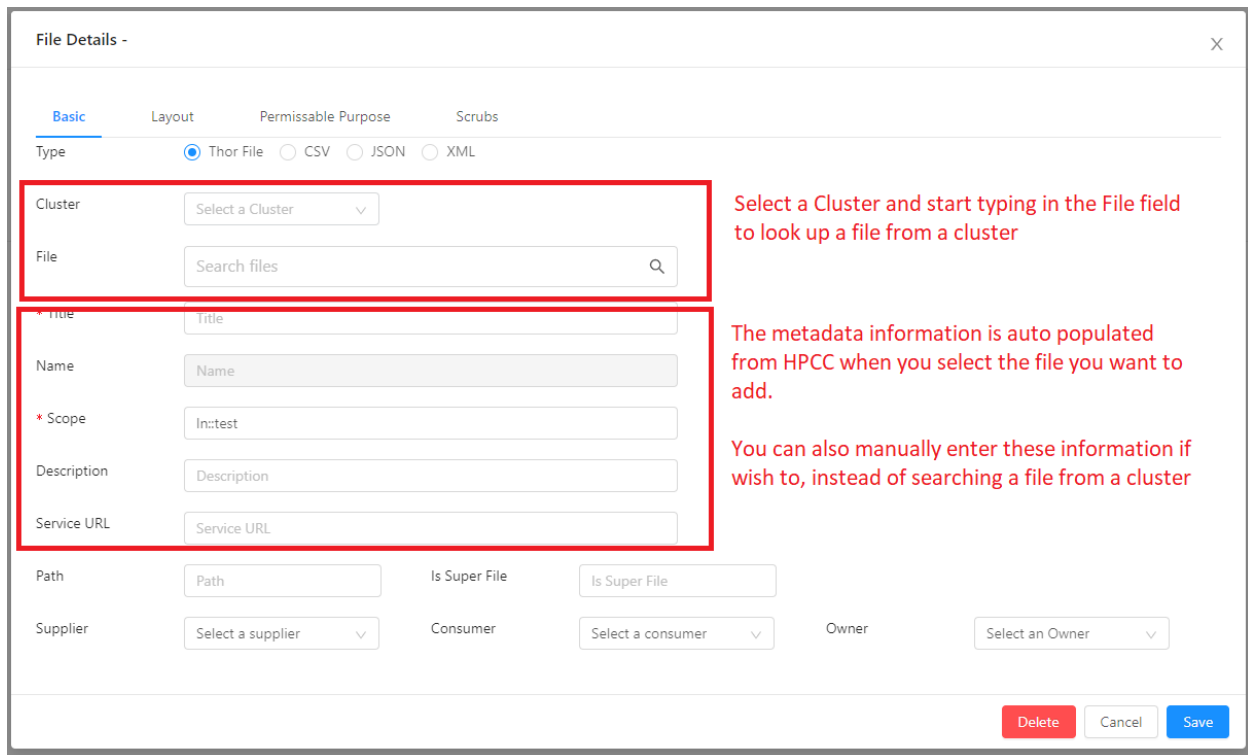
Files

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



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

Files Details

The 'File Details' form is shown with several annotations. A red box highlights the 'Cluster' dropdown and the 'File' search field, with a note: 'Select a Cluster and start typing in the File field to look up a file from a cluster'. Another red box highlights the metadata fields: 'Title', 'Name', '* Scope' (pre-filled with 'In::test'), 'Description', and 'Service URL', with a note: '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 wish to, instead of searching a file from a cluster'. The form includes tabs for 'Basic', 'Layout', 'Permissible Purpose', and 'Scrubs'. At the bottom are 'Delete', 'Cancel', and 'Save' buttons.

File Details -

Basic Layout Permissible Purpose Scrubs

Type ☒ Thor File ☐ CSV ☐ JSON ☐ XML

Cluster

File

* Title

Name

* Scope

Description

Service URL

Path Is Super File

Supplier Consumer Owner

Delete Cancel Save

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 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.

File Details - JHU-world.flat

Basic **Layout** Permissible Purpose Scrubs File Preview Dataflows

Name	Type	ECL Type	Description	Required	Information Type	Action
fips	String	string50		false	Select	
admin2	String	string50		false	City	
state	String	string50		false	Credit Card	
country	String	string50		false	DOB	
update_date	Unsigned Integer	unsigned4		false	Drivers License	
geo_lat	Real	decimal9_6		false	Email Address	
geo_long	Real	decimal9_6		false	Geo Coordinate	
confirmed	Real	real8		false	IP Addresses	
deaths	Real	real8		false	License Number	

Add a row

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

Delete Cancel Save

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.

File Details - JHU-world.flat

Basic Layout **Permissible Purpose** Scrubs File Preview Dataflows

Name	Description
<input type="checkbox"/> Creative Commons Attribution License	Allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately ...
<input type="checkbox"/> U.S. Government Works	

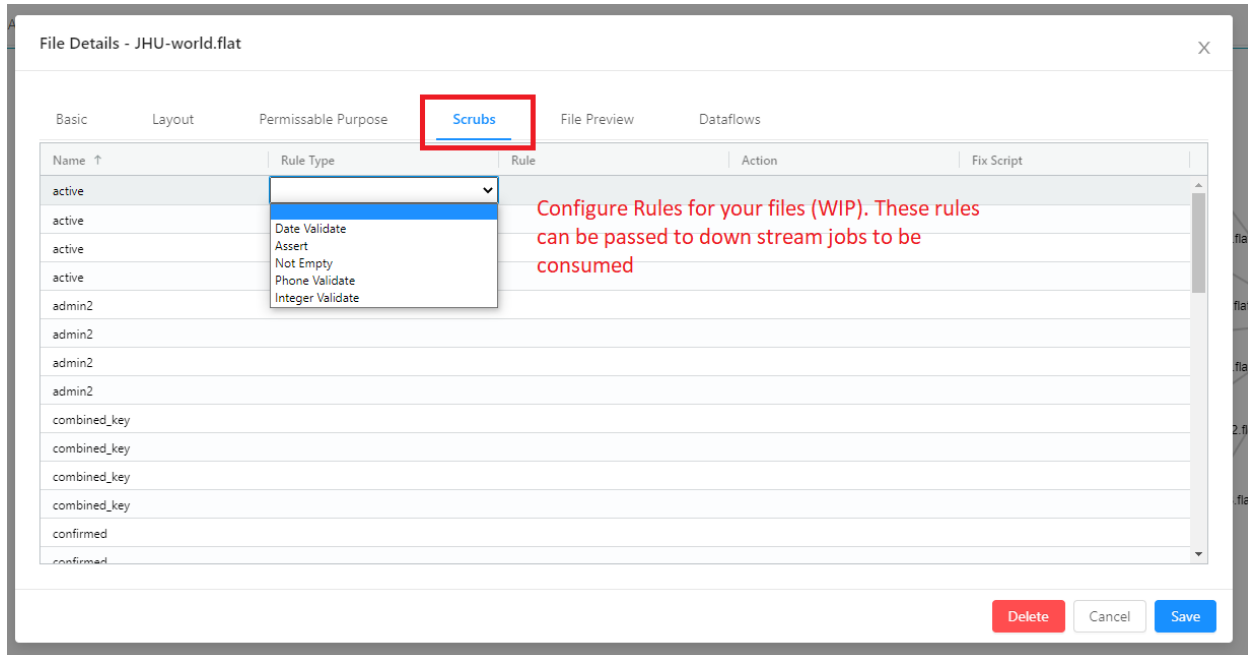
Record any license restrictions for your file

Delete Cancel Save

Scrubs (WIP)

You can configure Rules for each fields in your files to be then consumed by a downstream application/job.

PS - This feature is currently a work in Progress



File Preview

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

File Details - India-population.flat

Basic Layout Permissible Purpose Scrubs **File Preview** Dataflows

state	area_name	total_persons	total_males	total_fema...	rural_pers...	rural_males	rural_fema...	urban_pers...	urban_males	urban_fem...	_file
21	ODISHA	41974218	21212136	20762082	34970562	17586203	17384359	7003656	3625933	3377723	20
22	CHHATTIS...	25545198	12832895	12712303	19607961	9797426	9810535	5937237	3035469	2901768	20
23	MADHYA P...	72626809	37612306	35014503	52557404	27149388	25408016	20069405	10462918	9606487	21
24	GUJARAT	60439692	31491260	28948432	34694609	17799159	16895450	25745083	13692101	12052982	22
25	DADRA AN...	243247	150301	92946	60396	32395	28001	182851	117906	64945	23
26	DADRA NA...	343709	193760	149949	183114	98305	84809	160595	95455	65140	24
27	MAHARAS...	112374333	58243056	54131277	61556074	31539034	30017040	50818259	26704022	24114237	26
29	KARNATAKA	61095297	30966657	30128640	37469335	18929354	18539981	23625962	12037303	11588659	26
30	GOA	1458545	739140	719405	551731	275436	276295	906814	463704	443110	27
31	LADAKH	64473	33123	31350	14141	7243	6898	50332	25880	24452	28
32	KERALA	33406061	16027412	17378649	17471135	8408054	9063081	15934926	7619358	8315568	29
33	TAMIL NADU	72147030	36137975	36009055	37229590	18679065	18550525	34917440	17458910	17458530	30
34	PUDUCHER...	1247953	612511	635442	395200	194907	200293	852753	417604	435149	31

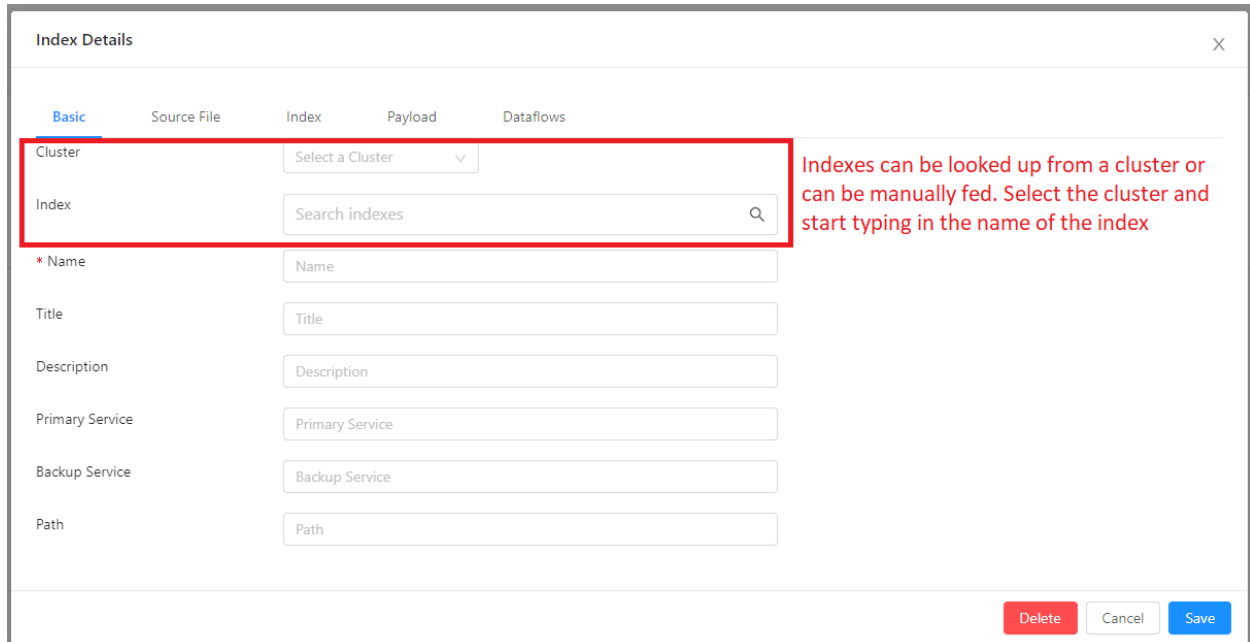
Delete Cancel Save

Dataflows – Shows the Tombolo Dataflows this file belongs to

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 'Index Details' form is shown with the 'Basic' tab selected. A red box highlights the 'Cluster' dropdown menu and the 'Index' search input field. A red text annotation points to these fields, stating: 'Indexes can be looked up from a cluster or can be manually fed. Select the cluster and start typing in the name of the index'. Below the highlighted fields are input fields for Name, Title, Description, Primary Service, Backup Service, and Path. At the bottom right are 'Delete', 'Cancel', and 'Save' buttons.

Index Details

Basic Source File Index Payload Dataflows

Cluster Select a Cluster

Index Search indexes

* Name Name

Title Title

Description Description

Primary Service Primary Service

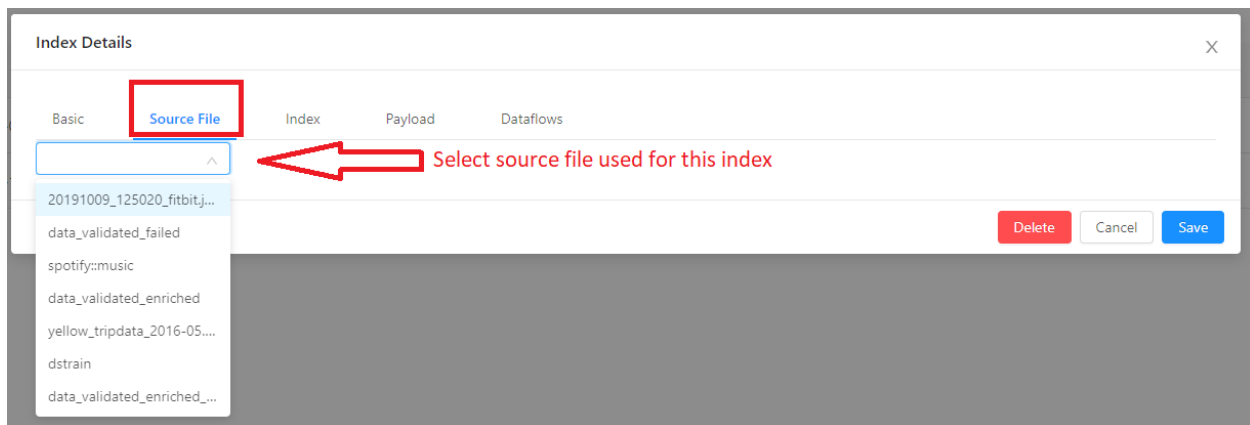
Backup Service Backup Service

Path Path

Delete Cancel Save

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

Source File



The 'Index Details' form is shown with the 'Source File' tab selected. A red box highlights the 'Source File' tab and the dropdown menu below it. A red arrow points from the text 'Select source file used for this index' to the dropdown menu. The dropdown menu lists several source files: 20191009_125020_fitbitj..., data_validated_failed, spotify::music, data_validated_enriched, yellow_tripdata_2016-05..., dstrain, and data_validated_enriched_... At the bottom right are 'Delete', 'Cancel', and 'Save' buttons.

Index Details

Basic Source File Index Payload Dataflows

20191009_125020_fitbitj...

data_validated_failed

spotify::music

data_validated_enriched

yellow_tripdata_2016-05...

dstrain

data_validated_enriched_...

Select source file used for this index

Delete Cancel Save

Index fields

Index Details

Basic

Source File

Index

Payload

Dataflows

Name	Type	Action
imported_seconds_utc	Integer	
userid	String	

Add a row

Delete

Cancel

Save

Key fields for the Index - auto populated from cluster

Payload fields

Index Details

Basic

Source File

Index

Payload

Dataflows

Name	Type	Action
elevation	String	
activitycalories	String	
imported_time_utc	String	
sedentaryminutes	String	
floors	String	
marginalcalories	String	
lightlyactiveminutes	String	
veryactiveminutes	String	

Add a row

Delete

Cancel

Save

Payload fields auto populated from cluster

Dataflows – Shows the Tombole Dataflows this Index belongs to

API/Gateway/Queries

API/Gateway/Query Details

Basic

Input Fields

Output Fields

Applications

Type

☒ Roxie Query ☐ API/Gateway

Cluster

Select a Cluster

Query

Search queries

* Name

Name

Title

fitbit_data_by_user

Description

Description

URL

URL

Git Repo

Git Repo URL

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

API/Gateway/Query Details

Basic

Input Fields

Output Fields

Applications

Name	Type	Possible Value	Value Description	Action
date_start_YYYYMMDD	RrcordOf			
user_id	Integer			
date_end_YYYYMMDD	Integer			

Add a row

Delete

Cancel

Save

Input fields for a query is auto populated from cluster.

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

Output Fields

API/Gateway/Query Details

Basic Input Fields **Output Fields**

Name	Type	Possible Value	Value Description	Action
result_count	number			

Add a row

Output fields of a query is identified automatically from cluster. User can also add custom fields by clicking Add a row

Job/Functions

Job Details

Basic Input Params Input Files Output Files Dataflows

Cluster Select a Cluster

Job Search jobs

Name Ingest_JH_data

Title Ingest_JH_data

Description Description

Git Repo Git Repo

Entry BWR Ingest_JH_data

Contact Contact Author Author

Job Type

☐ Automatically create dependant files

Delete Cancel Save

Search for a job from a cluster to retrieve some metadata

Capture contact info, Author of jobs here

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

Input Files

Job Details

Basic

Input Params

Input Files

Output Files

Dataflows

Input Files:

Select Input Files

Add

Name	Description
hpccsystems::covid19::file::public::johnhopkins::world.flat	Input files for HPCC Jobs are auto populated
hpccsystems::covid19::file::public::johnhopkins::us.flat	
hpccsystems::covid19::file::public::uscountypopulation::population.flat	
hpccsystems::covid19::file::public::worldpopulation::population_gender.flat	
hpccsystems::covid19::file::public::uspopulation::population.flat	

<

1

>

☐ Automatically create dependant files

Delete

Cancel

Save

Output files

Job Details

Basic

Input Params

Input Files

Output Files

Dataflows

Output Files:

Select Output Files

Add

Name	Description
hpccsystems::covid19::file::input::source::jhl::level3	Output files for HPCC jobs auto populated from cluster. Files already existing in Tombolo can also be added here using Output Files dropdown
hpccsystems::covid19::file::input::source::jhl::level2	
hpccsystems::covid19::file::input::source::jhl::level1	

<

1

>

☐ Automatically create dependant files

Delete

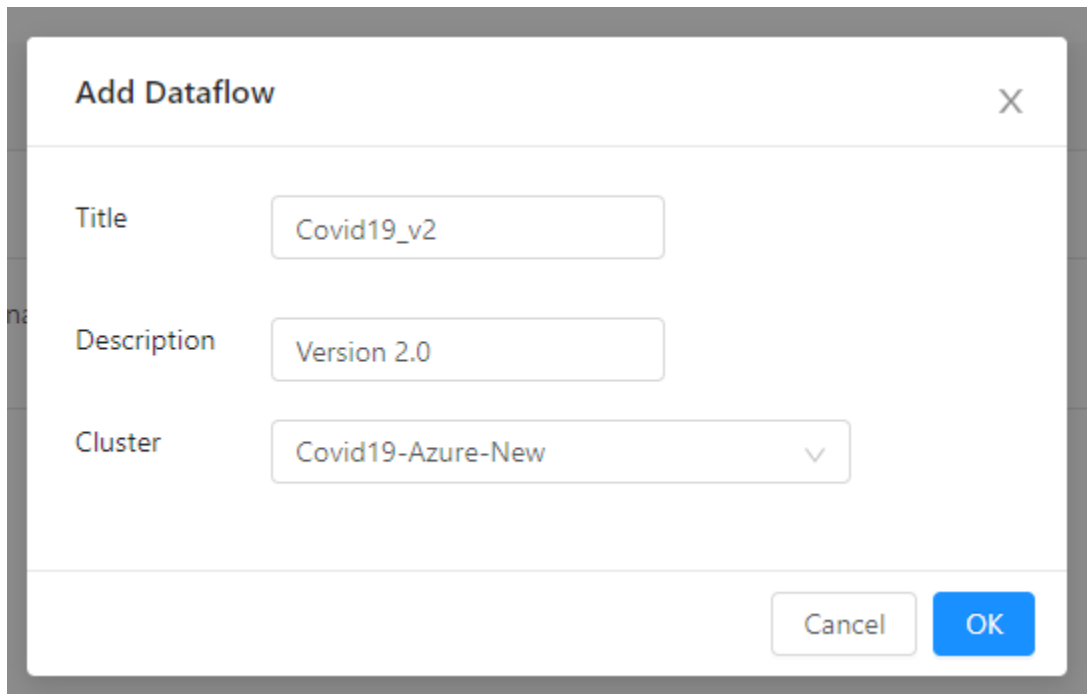
Cancel

Save

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.

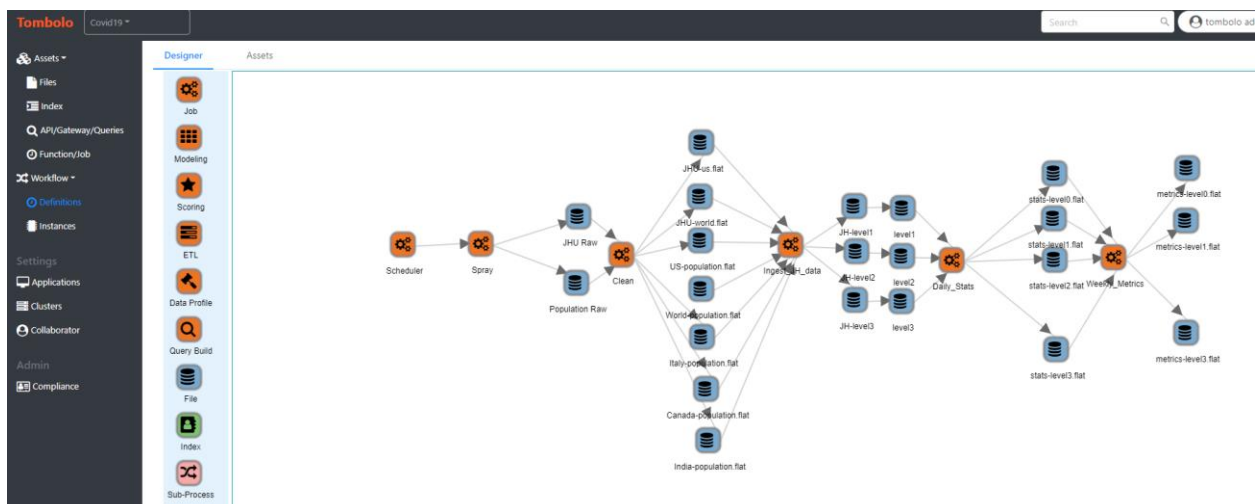


The 'Add Dataflow' dialog box is shown with the following fields:

- Title:** Covid19_v2
- Description:** Version 2.0
- Cluster:** Covid19-Azure-New (with a dropdown arrow)

At the bottom right, there are two buttons: 'Cancel' and 'OK'.

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](#) – Click on the node and press Delete button or use the delete icon that shows up on the node

[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.

The screenshot displays the Tombolo workflow management interface. The top navigation bar includes the 'Tombolo' logo, a 'Covid19' filter, a search bar, and a user profile 'tombolo admin'. The left sidebar contains a menu with categories: Data Dictionary, Assets (Files, Index, API/Gateway/Queries, Function/Job), Workflow (Definitions, Instances), Settings (Applications, Clusters, Collaborator), and Admin (Compliance).

The main area features a dataflow graph. A red box highlights the 'Scheduler' and 'Spray' jobs, with a red arrow pointing to them and the text 'Successfully executed Jobs'. Below the graph, a red arrow points to the 'Work Units' tab in the table, with the text 'Work Units created during the workflow execution'.

The 'Work Units' table lists the following data:

WorkUnit Id	Status	Start Time	End Time	Duration	Job Name	Owner	Action
W20200922-201041	completed	20:10:41	20:10:41	1:12.615	Scheduler_dev2	xulll01	Details
W20200922-201134	completed	20:11:34	20:11:59	25.430	Produce_Weekly_Metric	xulll01	Details