Tombolo User Guide

HPCC Systems Solutions Lab

Contents

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?”

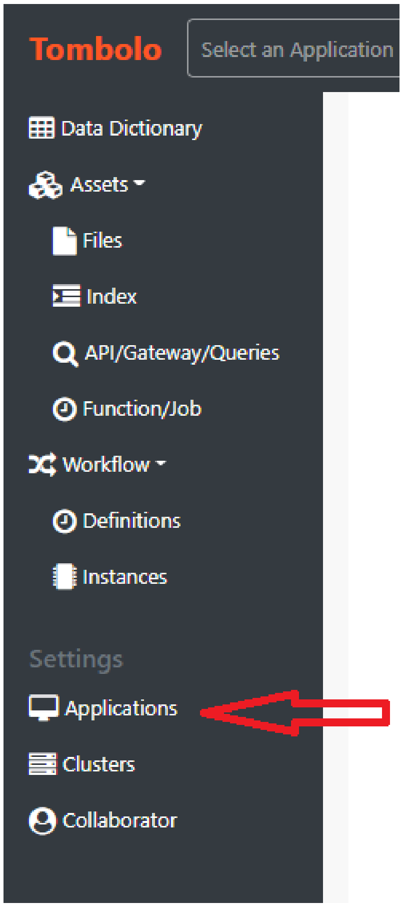
"What is the impact of using this data?”

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

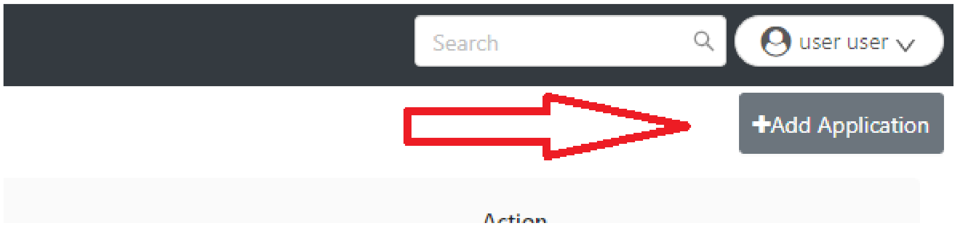
"What happens if the data is not used on time?”

Create an Application

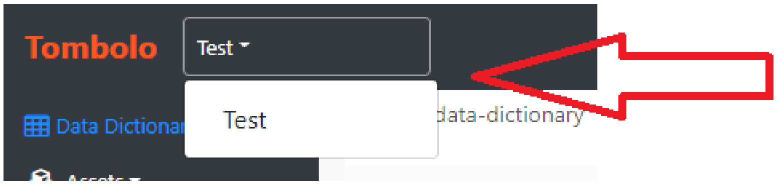
In order to start using Tombolo, an “Application” has to be created. Application is a way of grouping your assets with in Tombolo. To create an application, click on the “Applications” link in the left nav. If you already have Applications, they will be listed in the Applications page



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



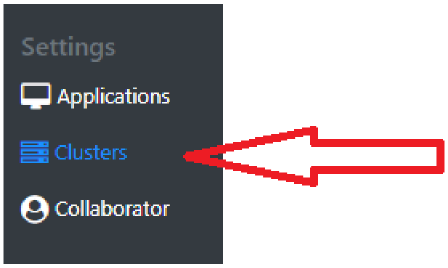
Click OK to create the Application. The Application should be now listed under the Applications dropdown.



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.



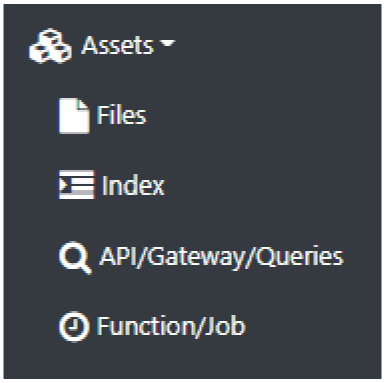
Assets

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)

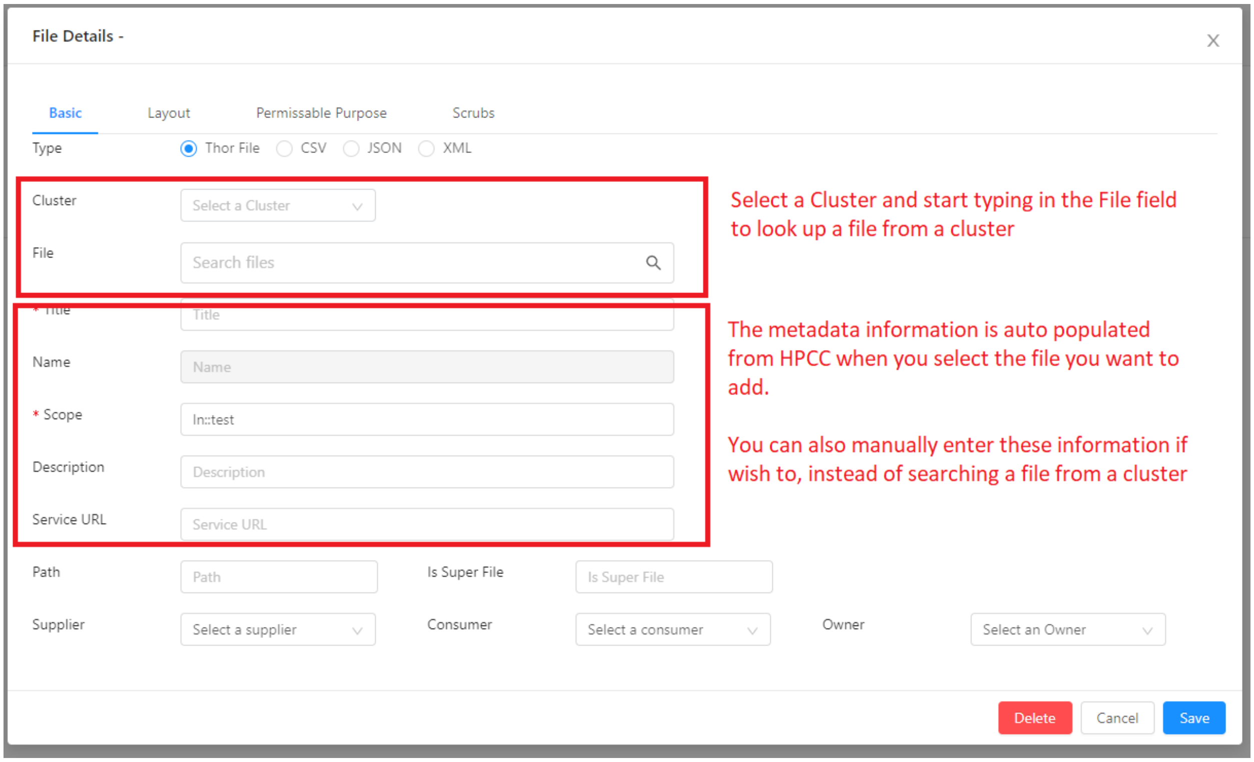
Files

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



File Details -

Type: ● Thor File ○ CSV ○ JSON ○ XML

Cluster: Select a Cluster

File: Search files

Select a Cluster and start typing in the File field to look up a file from a cluster

* Title: Title

Name: Name

* Scope: In::test

Description: Description

Service URL: Service URL

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

Path: Path

Is Super File: Is Super File

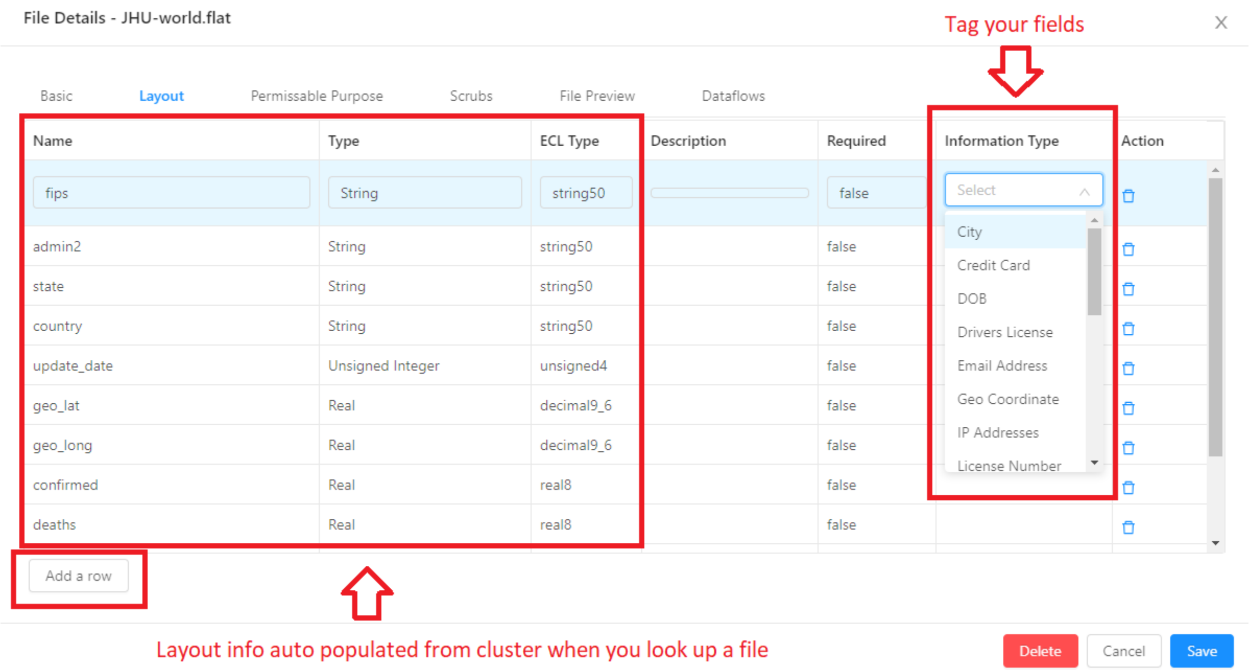
Supplier: Select a supplier

Consumer: Select a consumer

Owner: Select an Owner

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

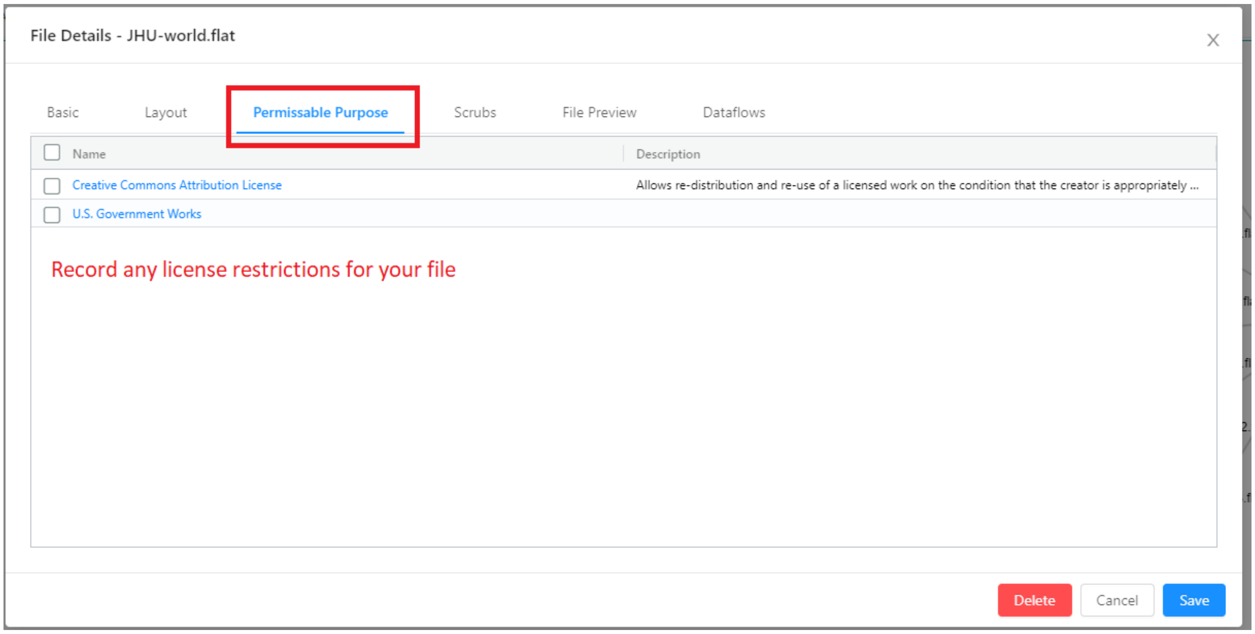
Tag your fields

Name	Type	ECL Type	Description	Required	Information Type	Action
fips	String	string50		false	Select	
admin2	String	string50		false		
state	String	string50		false		
country	String	string50		false		
update_date	Unsigned Integer	unsigned4		false		
geo_lat	Real	decimal9_6		false		
geo_long	Real	decimal9_6		false		
confirmed	Real	real8		false		
deaths	Real	real8		false		

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.



File Details - JHU-world.flat

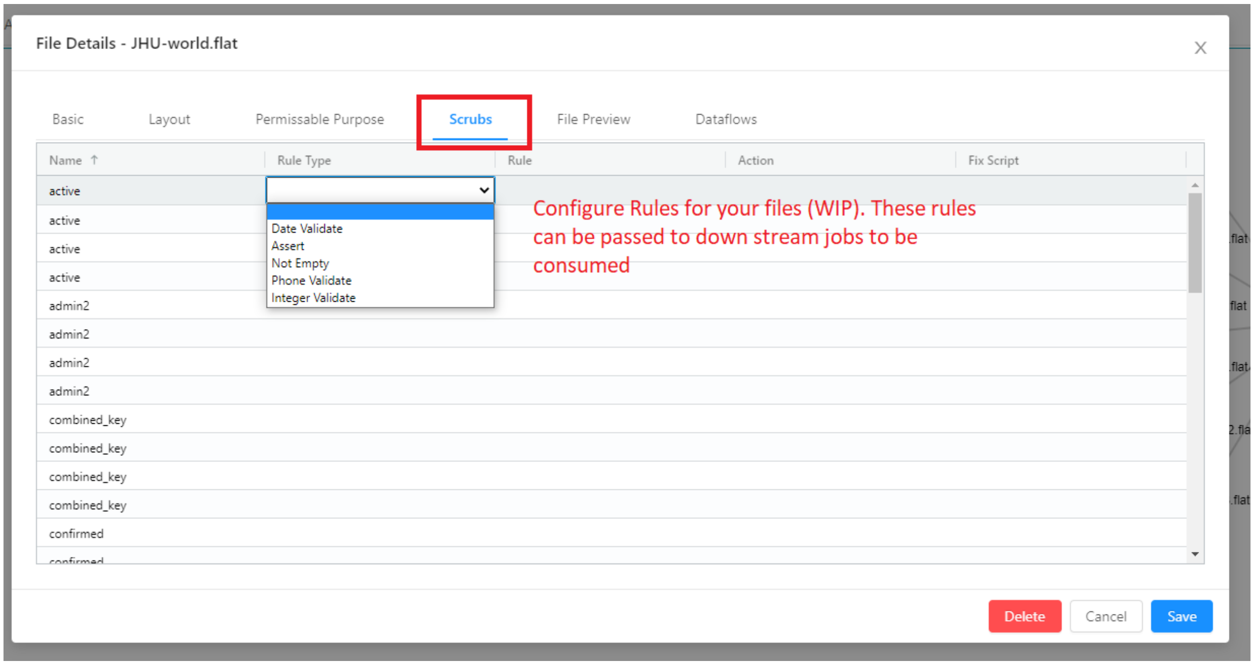
□ Name	Description
□ Creative Commons Attribution License	Allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately ...
□ U.S. Government Works	

Record any license restrictions for your file

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

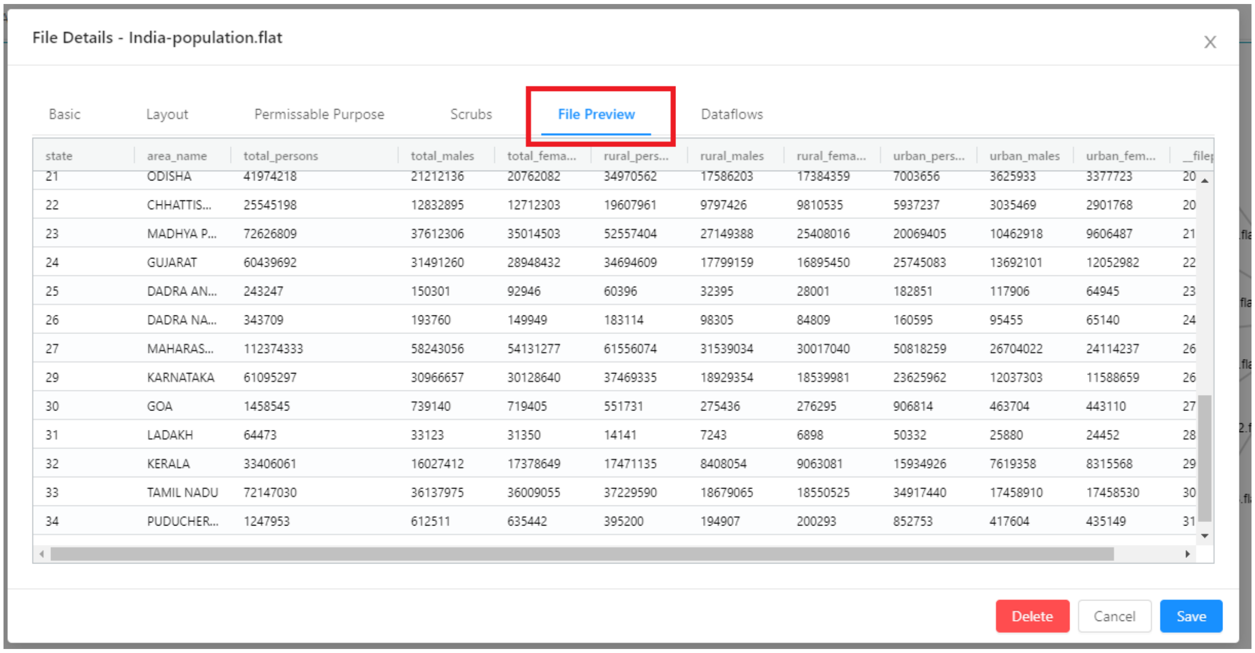


Name ↑	Rule Type	Rule	Action	Fix Script
active				
active				
active				
active				
admin2				
admin2				
admin2				
admin2				
combined_key				
combined_key				
combined_key				
combined_key				
confirmed				
confirmed				

Configure Rules for your files (WIP). These rules can be passed to down stream jobs to be consumed

File Preview

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



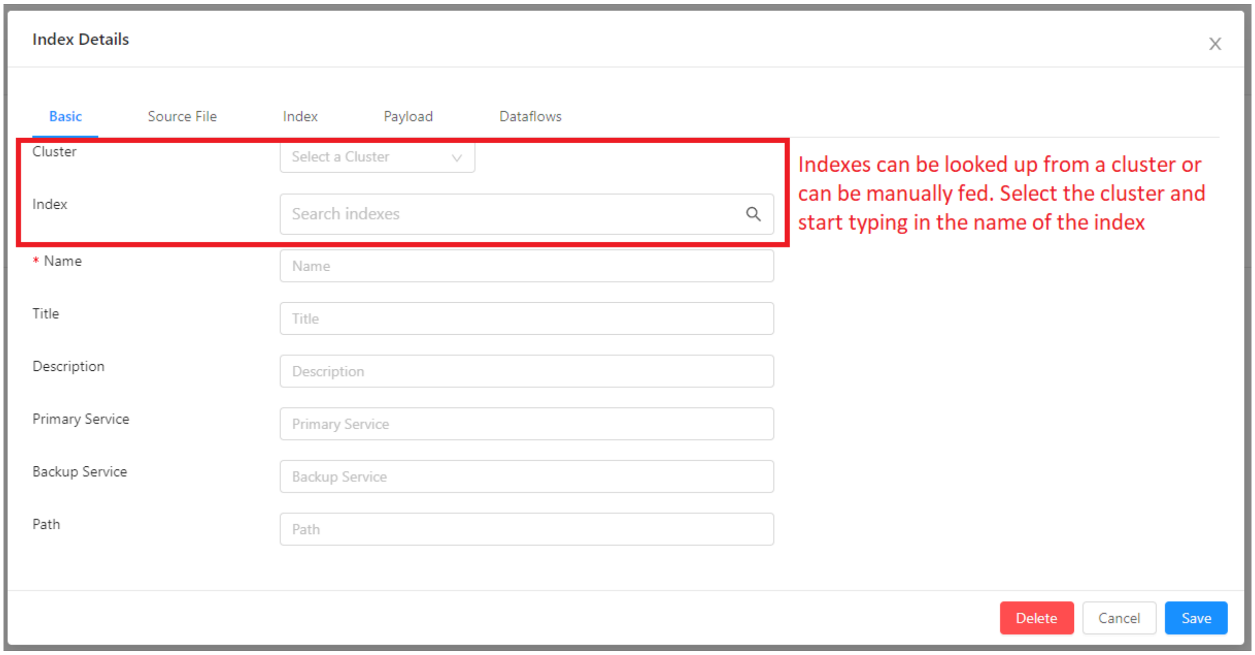
state	area_name	total_persons	total_males	total_fema...	rural_pers...	rural_males	rural_fema...	urban_pers...	urban_males	urban_fem...	__filep
21	ODISHA	41974218	21212136	20762082	34970562	17586203	17384359	7003656	3625933	3377723	20
22	CHHATTIS...	25545198	12832895	12712303	19607961	9797426	9810535	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

Dataflows – Shows the Tombolo Dataflows this file belongs to

Indexes

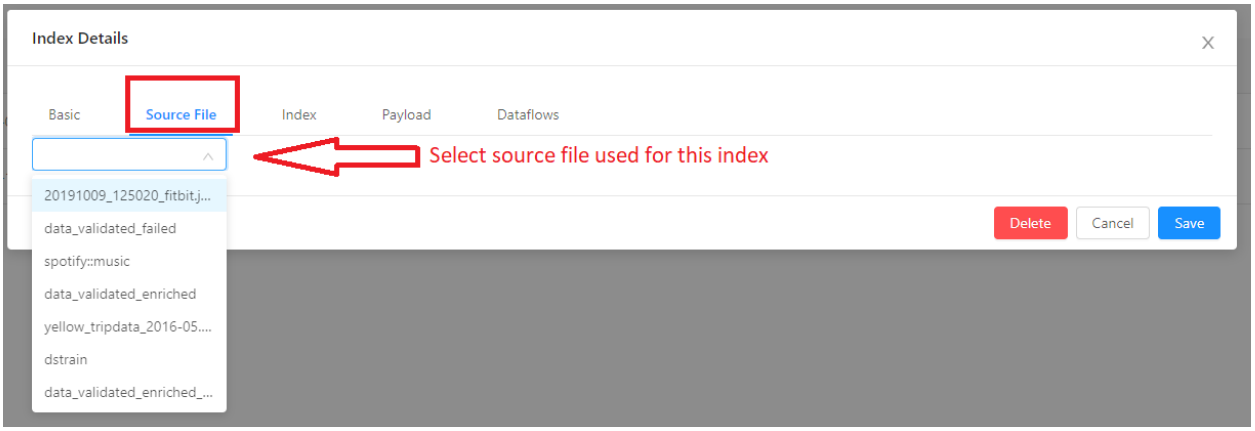
Click on the Index option on the left nav to view the Indexes that are already added to Tombolo. New Indexes can be added using Add button.

Basic Info



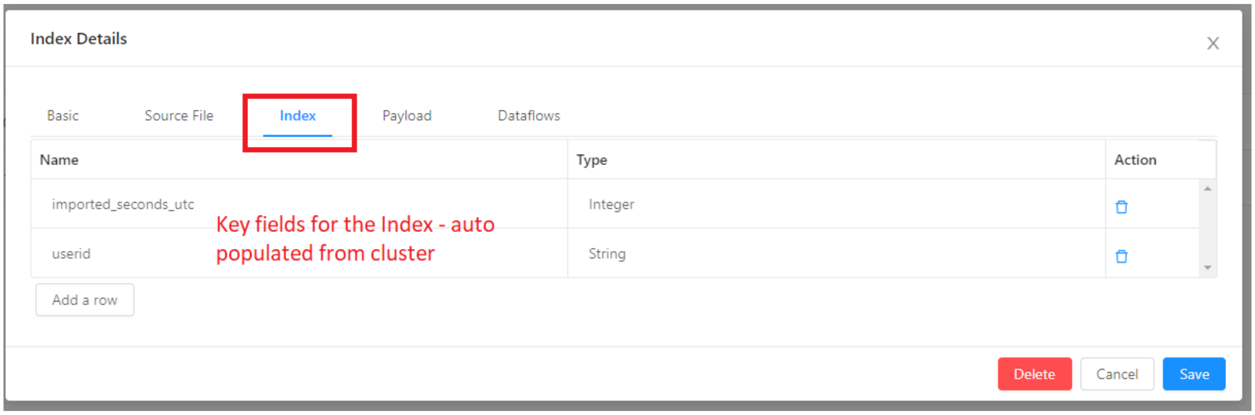
Indexes can be looked up from a cluster or can be manually fed. Select the cluster and start typing in the name of the index

Source File



Select source file used for this index

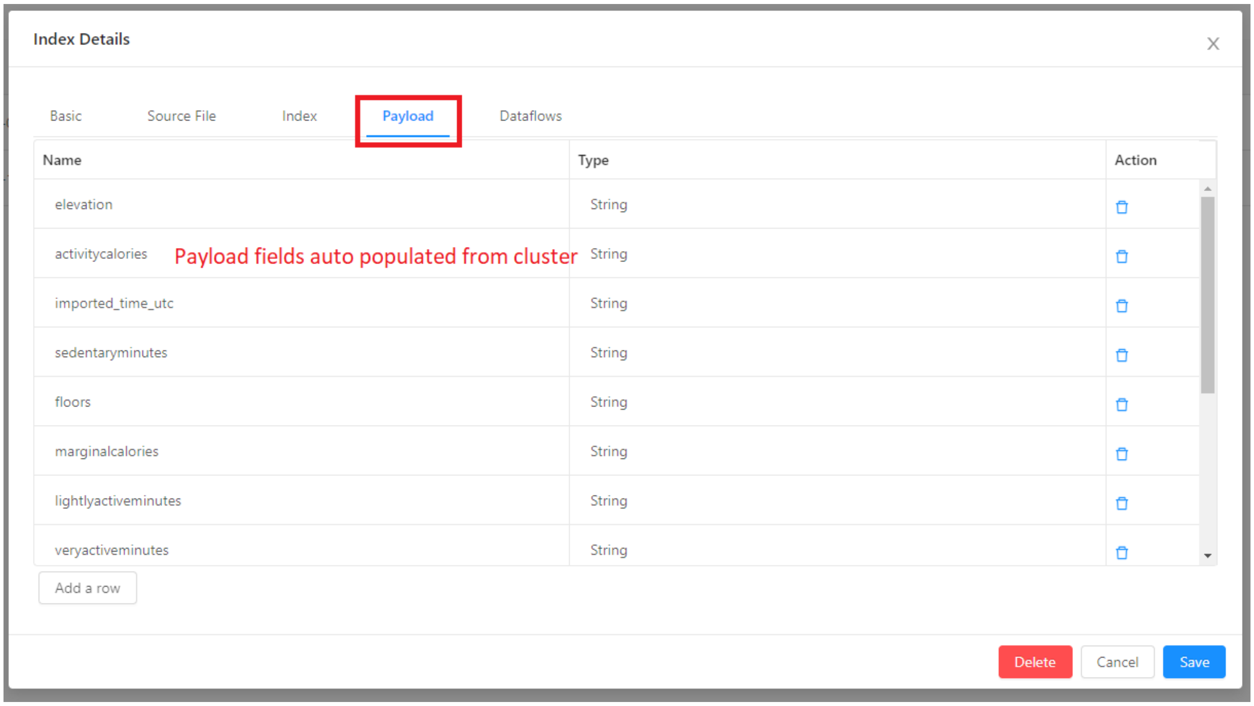
Index fields



Name	Type	Action
imported_seconds_utc	Integer	🗑
userid	String	🗑

Key fields for the Index - auto populated from cluster

Payload fields

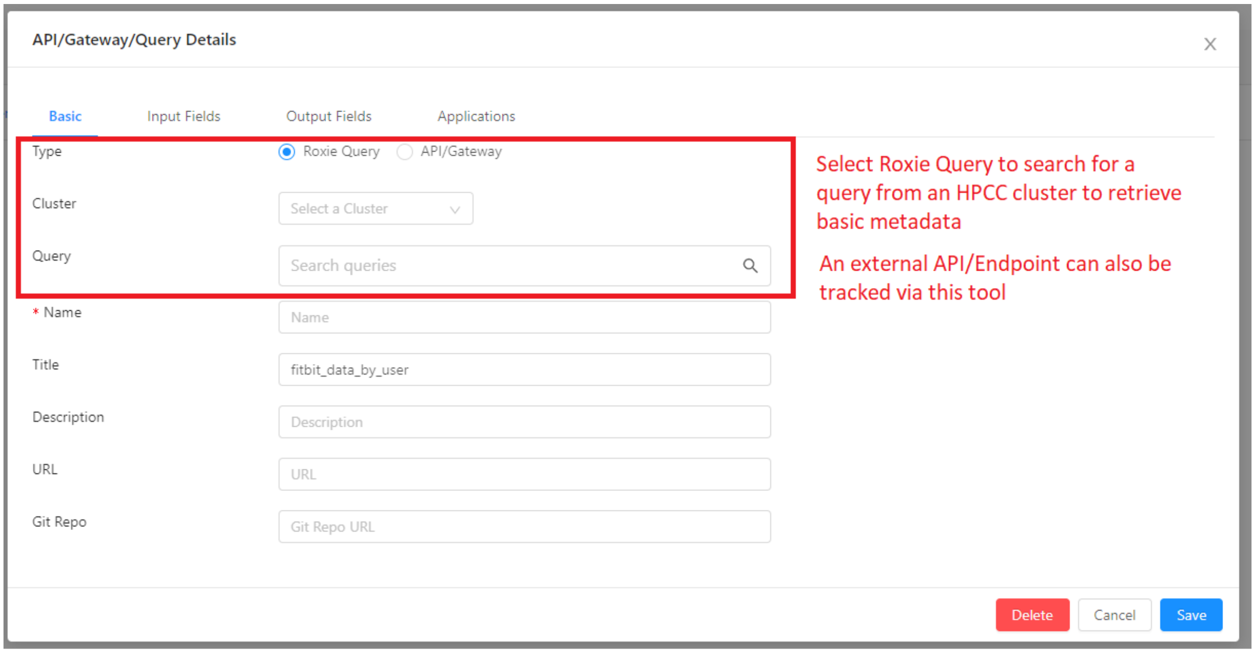


Name	Type	Action
elevation	String	🗑
activitycalories	String	🗑
imported_time_utc	String	🗑
sedentaryminutes	String	🗑
floors	String	🗑
marginalcalories	String	🗑
lightlyactiveminutes	String	🗑
veryactiveminutes	String	🗑

Payload fields auto populated from cluster

Dataflows – Shows the Tombolo 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

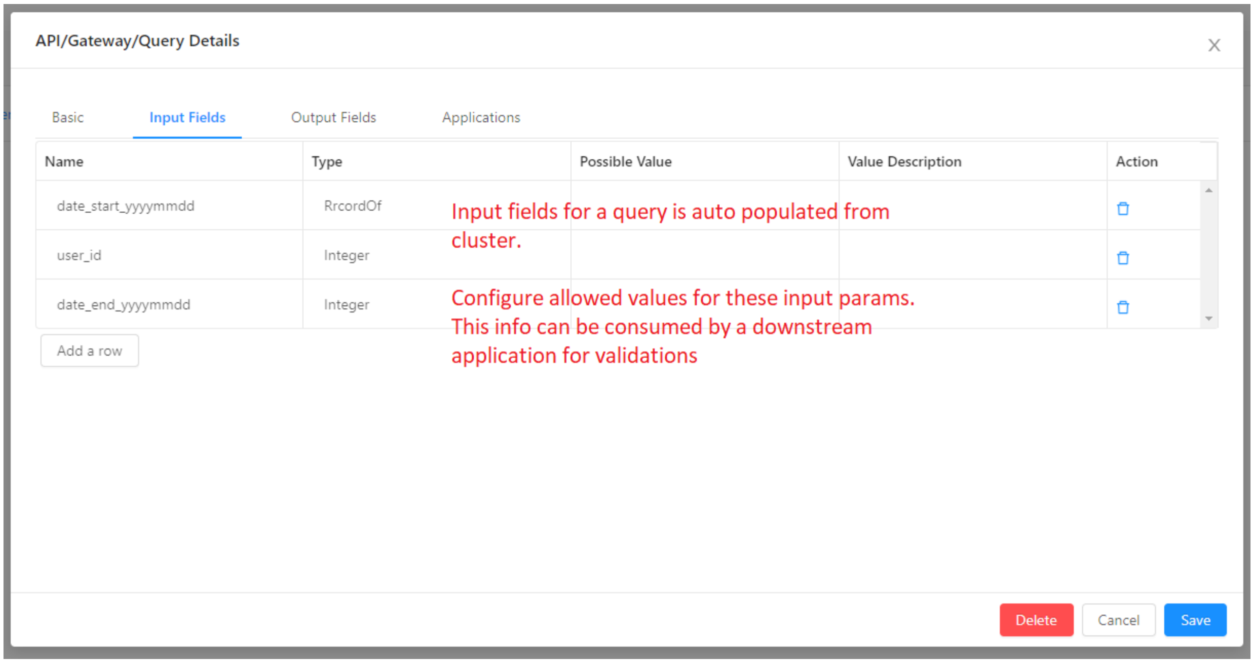
URL: URL

Git Repo: Git Repo URL

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

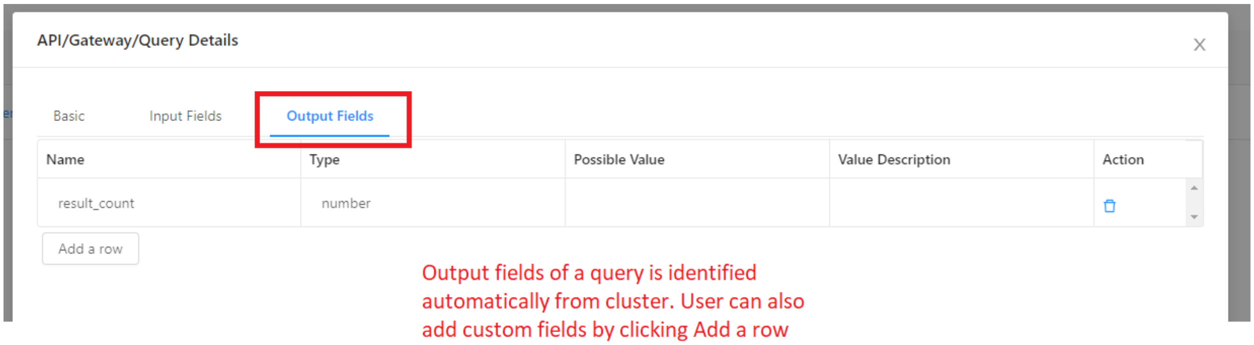
Basic | Input Fields | Output Fields | Applications

Name	Type	Possible Value	Value Description	Action
date_start_yyyymmdd	RrcordOf			🗑
user_id	Integer			🗑
date_end_yyyymmdd	Integer			🗑

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

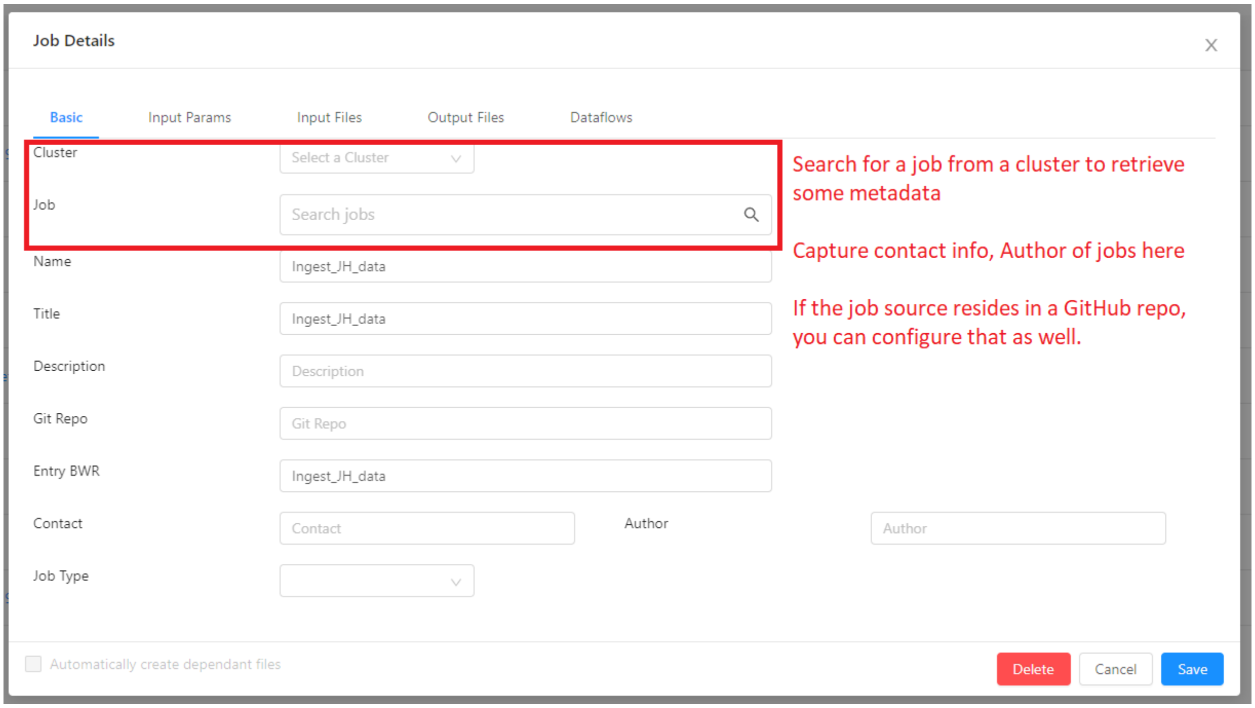
Output Fields



Name	Type	Possible Value	Value Description	Action
result_count	number			

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

Job/Functions



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	

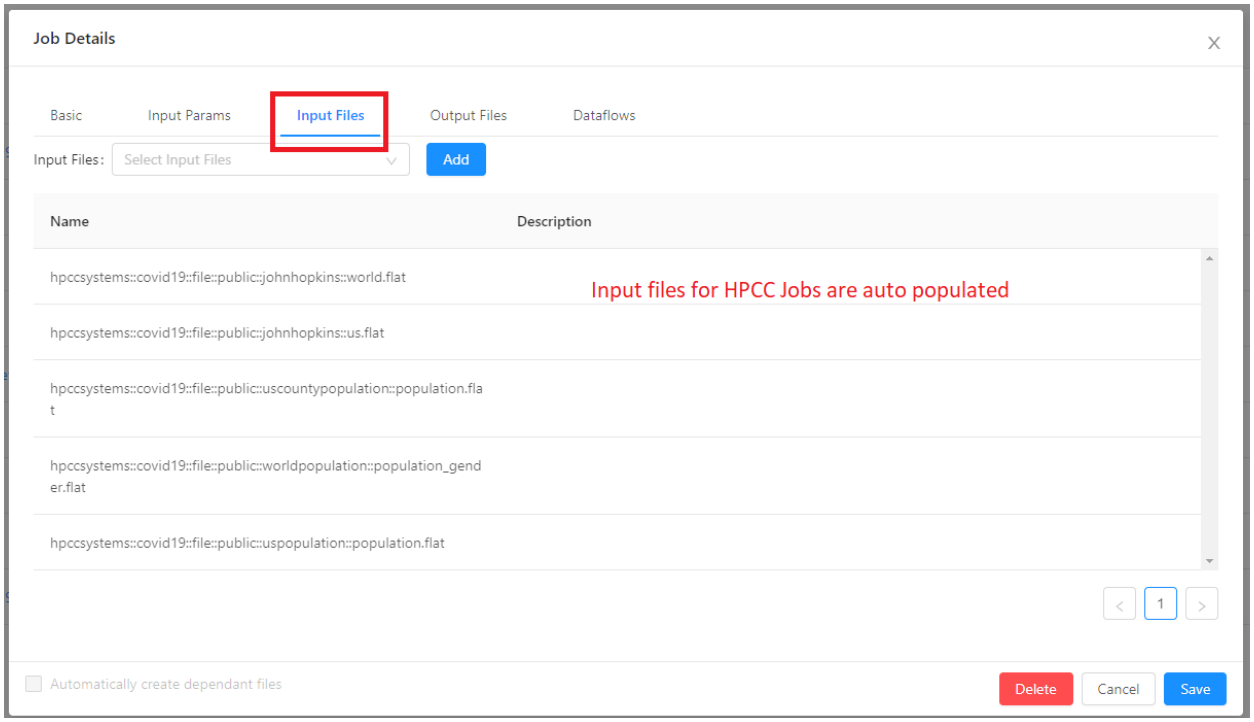
□ Automatically create dependant files

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



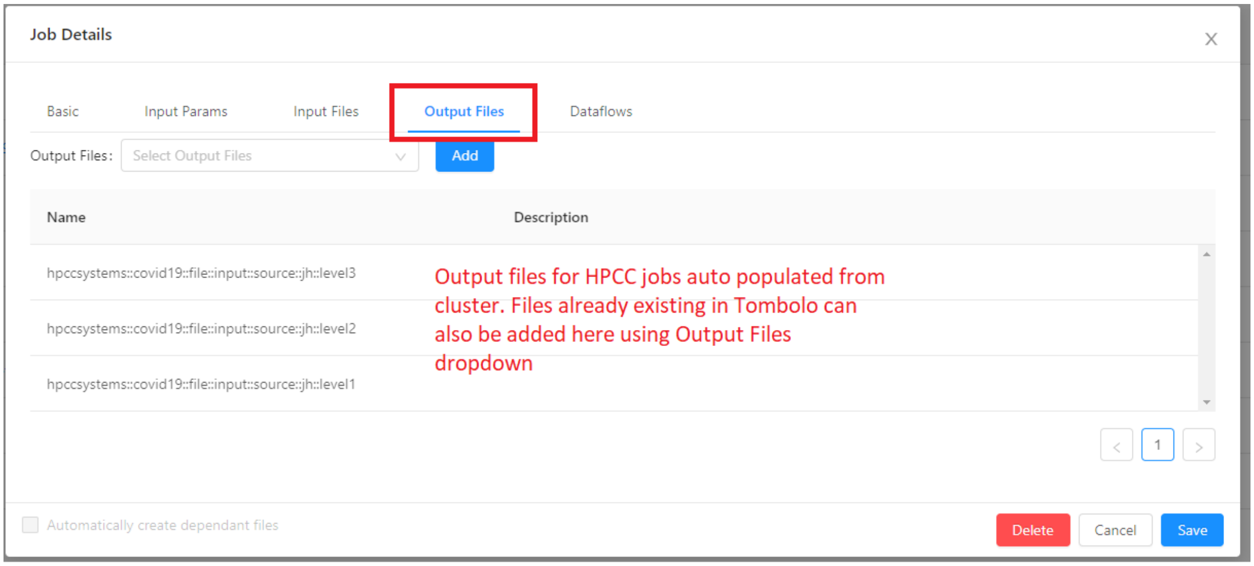
Input Files: Select Input Files

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

Input files for HPCC Jobs are auto populated

□ Automatically create dependant files

Output files



Output Files: Select Output Files

Name	Description
hpccsystems::covid19::file::input::source::jh::level3	
hpccsystems::covid19::file::input::source::jh::level2	
hpccsystems::covid19::file::input::source::jh::level1	

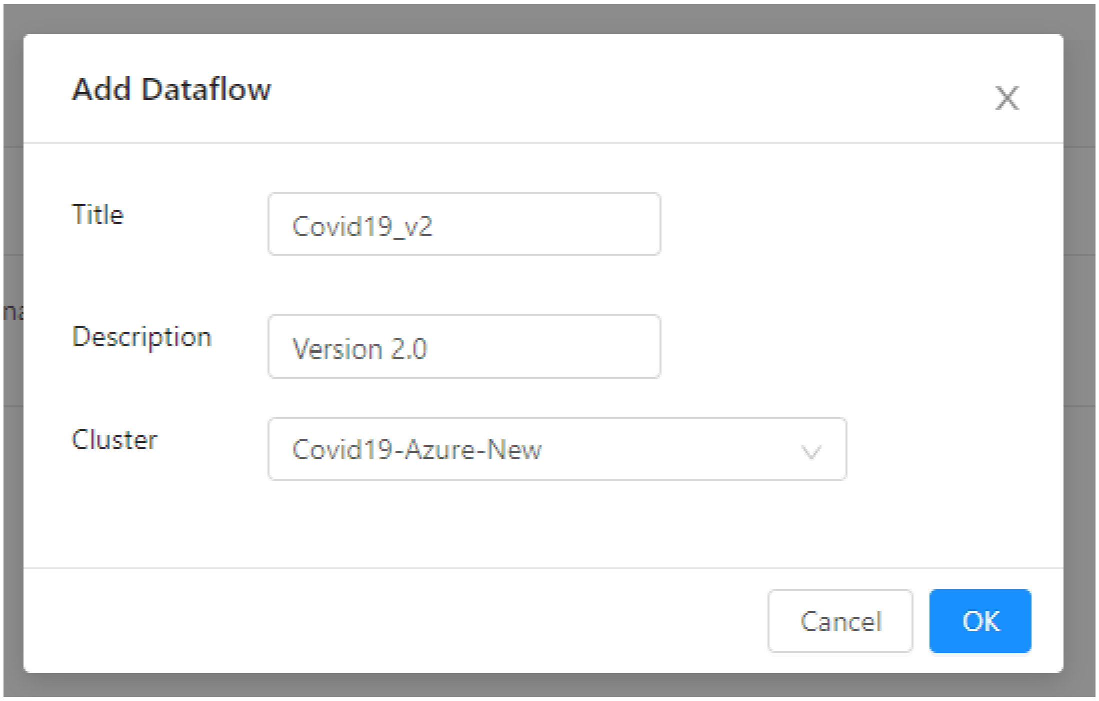
Output files for HPCC jobs auto populated from cluster. Files already existing in Tombolo can also be added here using Output Files dropdown

□ Automatically create dependant files

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.



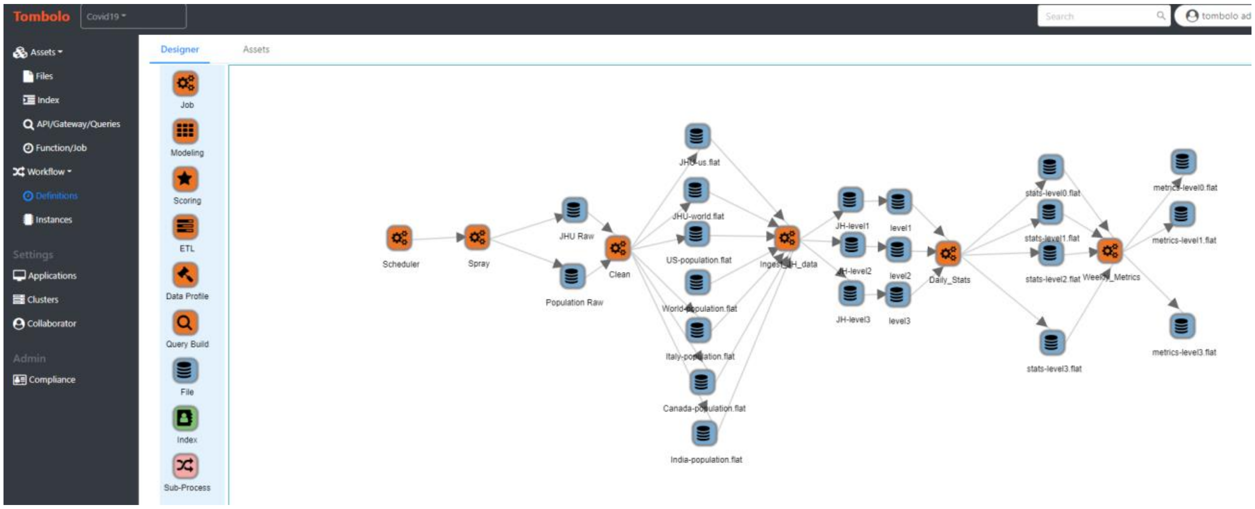
Add Dataflow

Title: Covid19_v2

Description: Version 2.0

Cluster: Covid19-Azure-New

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

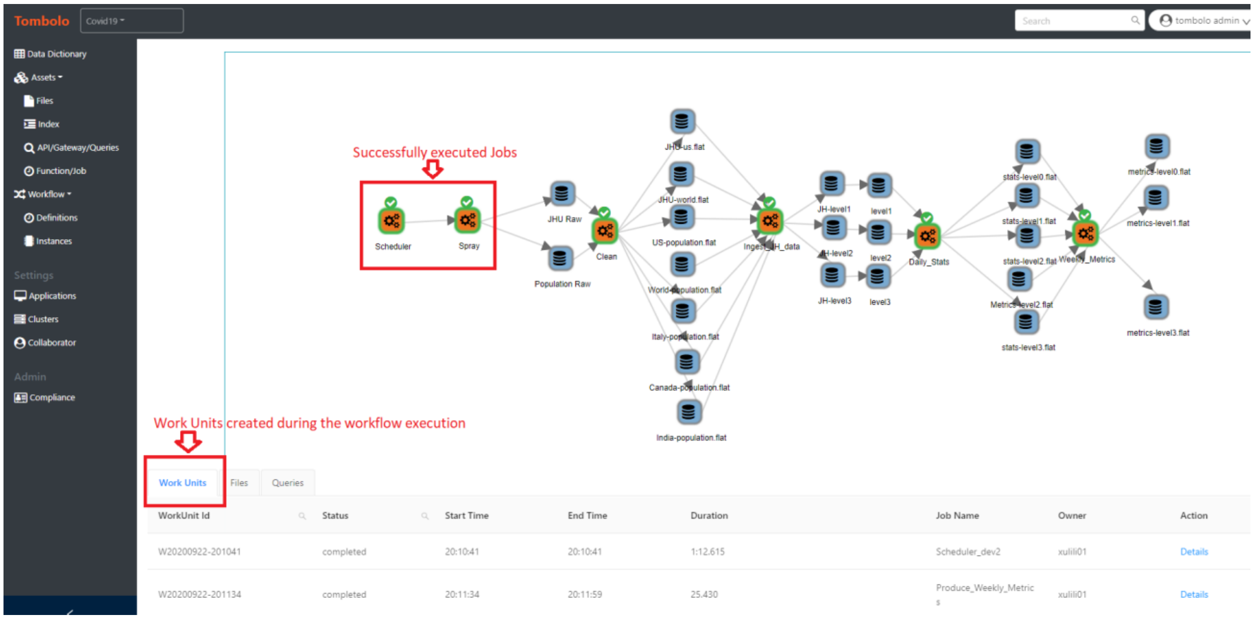
Delete a connection – select the connection and press Delete button

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.



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	xulili01	Details
W20200922-201134	completed	20:11:34	20:11:59	25.430	Produce_Weekly_Metrics	xulili01	Details