Tombolo User Guide

HPCC Systems Solutions Lab

Contents

Introduction

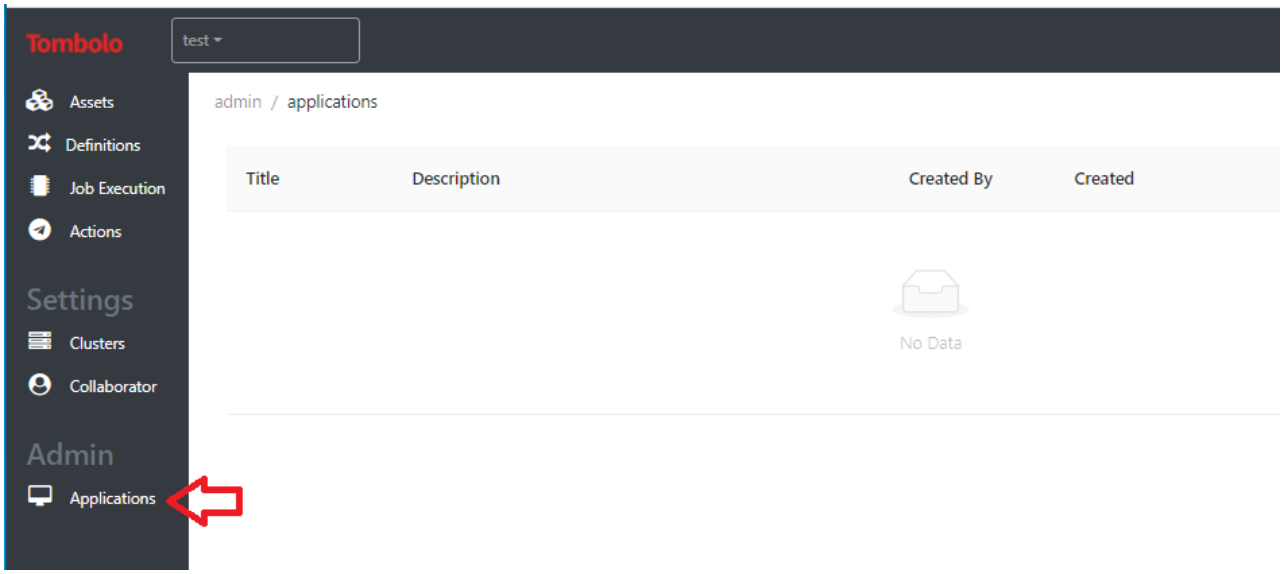
Tombolo is a metadata tracking tool for HPCC Data Lake solution. It tracks the metadata around how every asset is used in a Data Lake, and the process flow as to how these assets evolve.

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

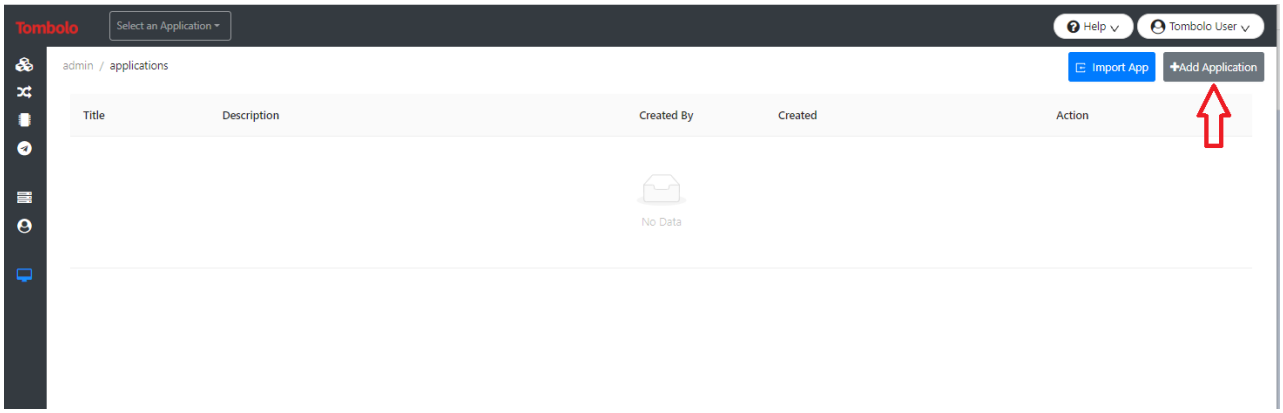
- "Who is the owner of xyz data?"
- "What is the source of xyz data?" "What does the data contain?"
- "What are the compliance rules around xyz data?" "Who approved the usage of this data?"
- "When was this data last used?"
- "Can you show me how this data is being used?" "Is this data being handled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

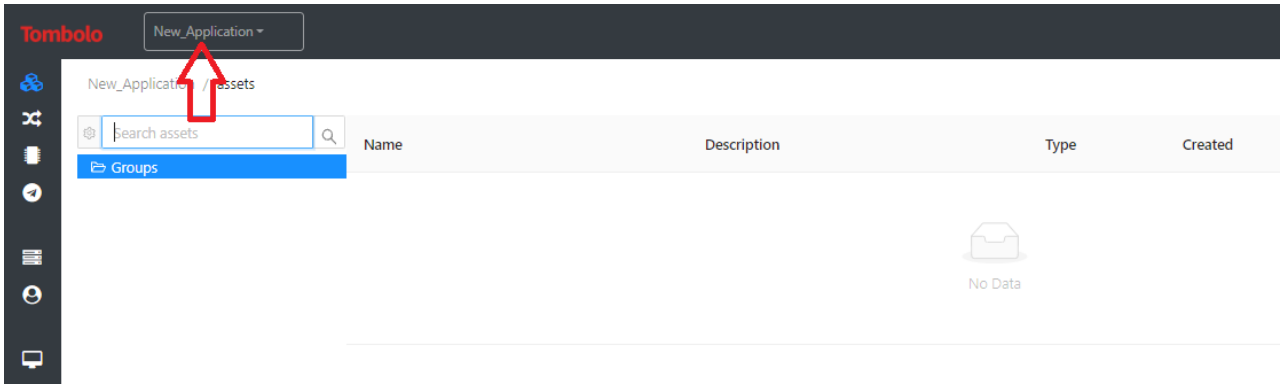
In order to start using Tombolo, an “Application” has to be created. Application is a way of grouping your assets within Tombolo. To create an application, click on the “Applications” link in the left nav. If you already have Applications, they will be listed in the Applications page



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



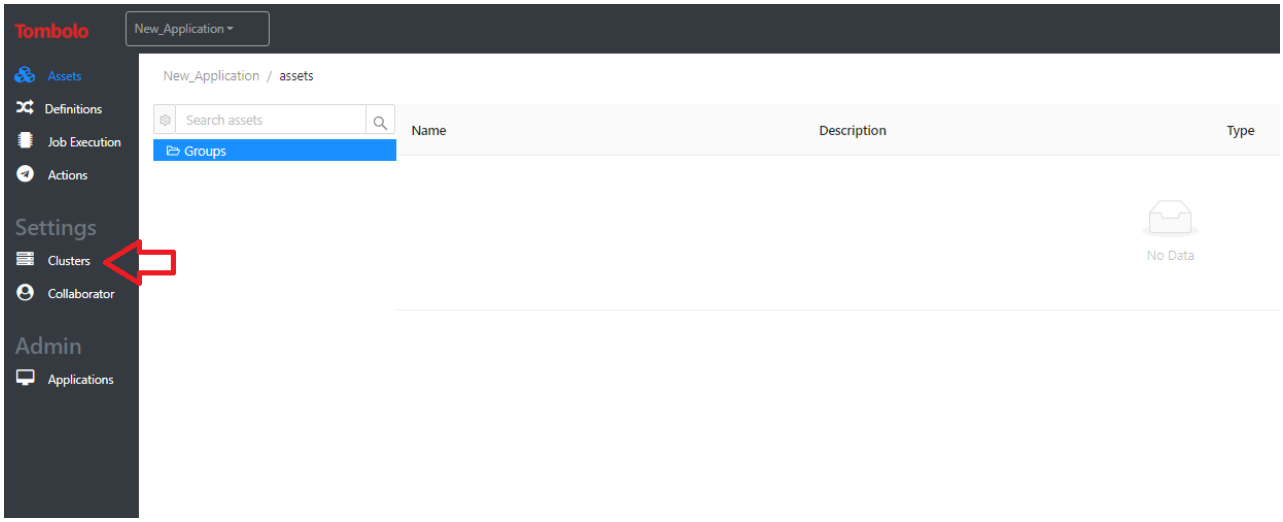
Click OK to create the Application. The Application should be now listed under the Applications dropdown.



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.



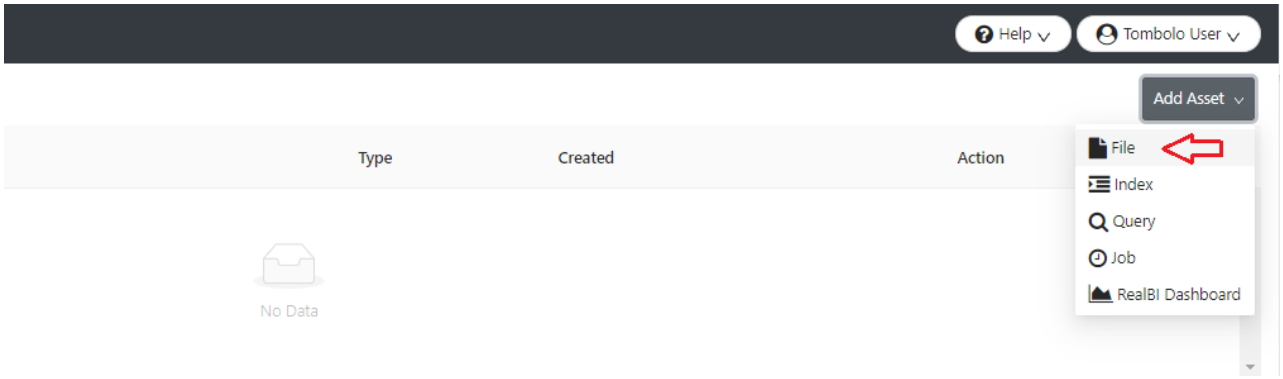
Assets

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)

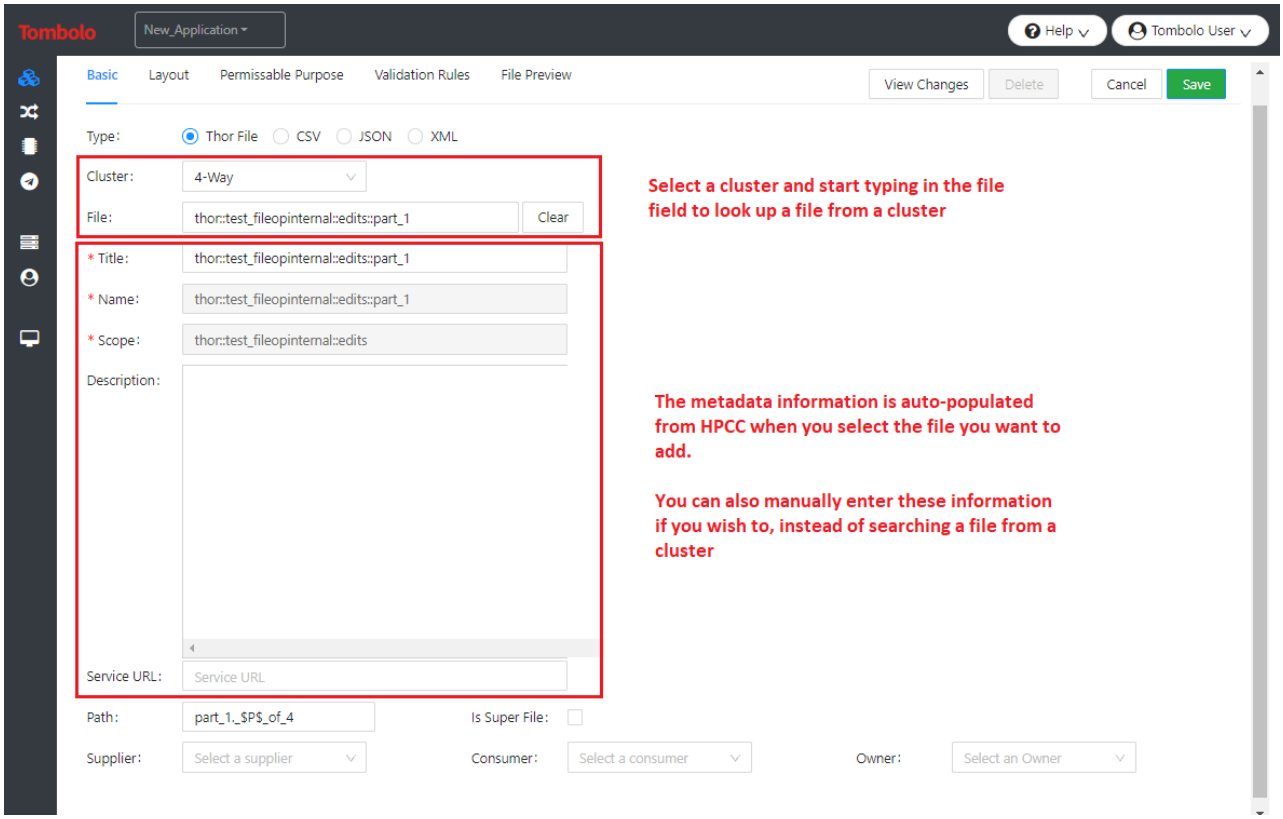
Files

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



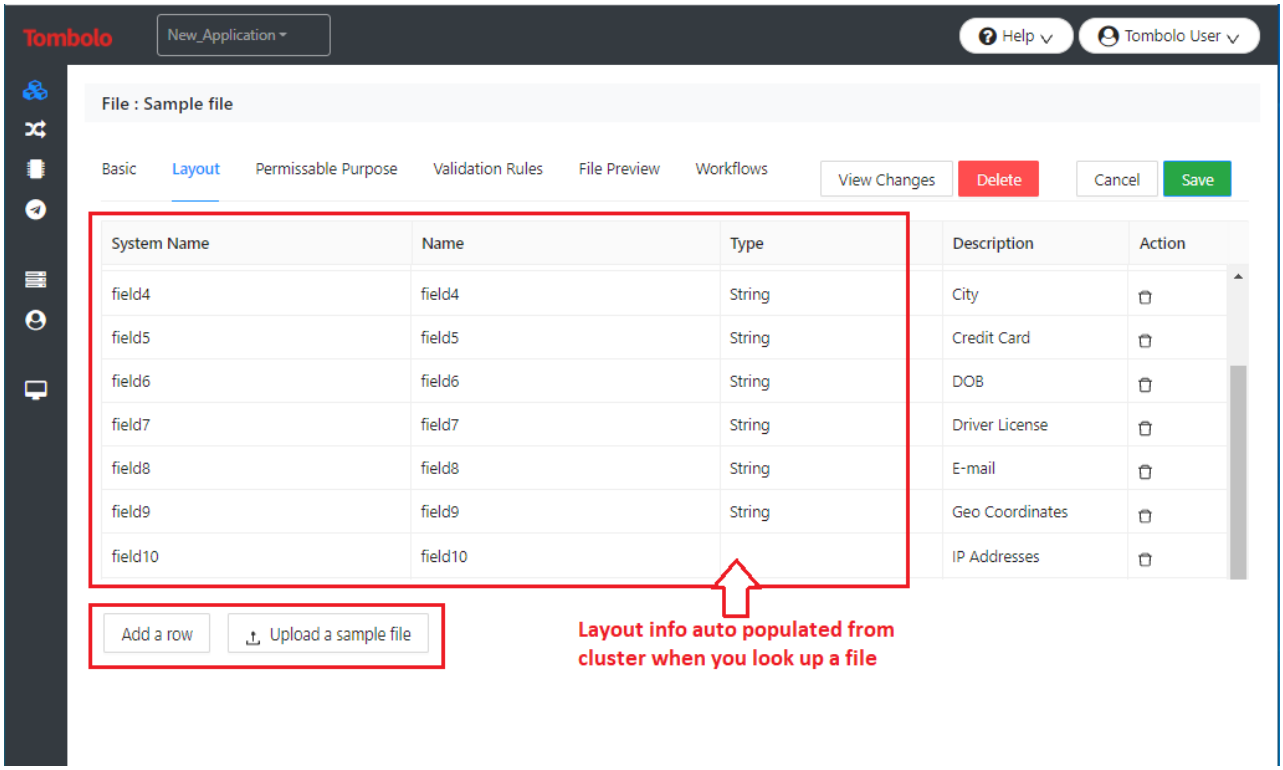
Click Add button under each asset type to add respective asset.

Files Details



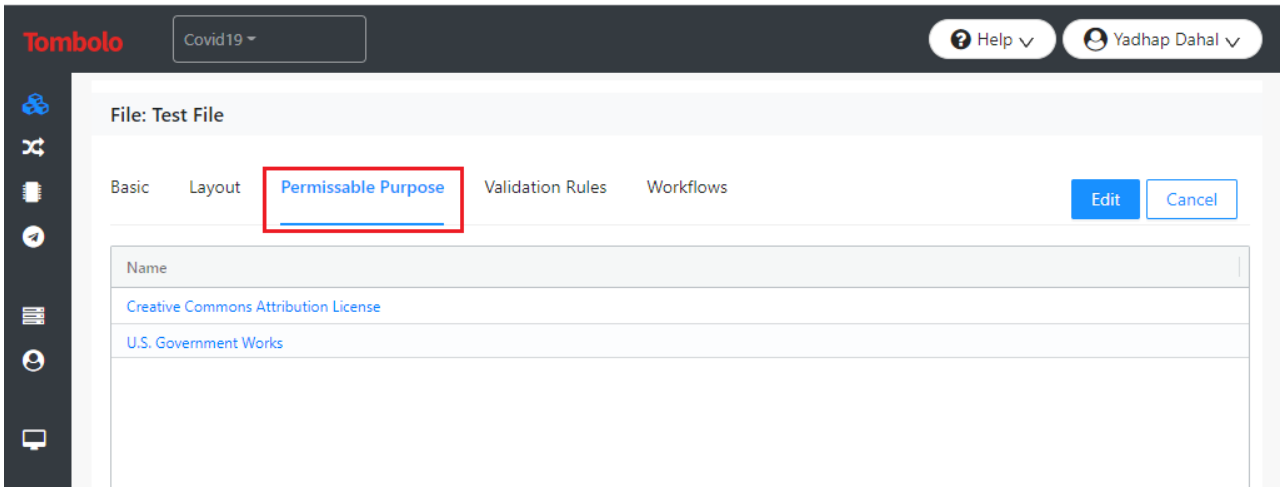
File Layouts:

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.



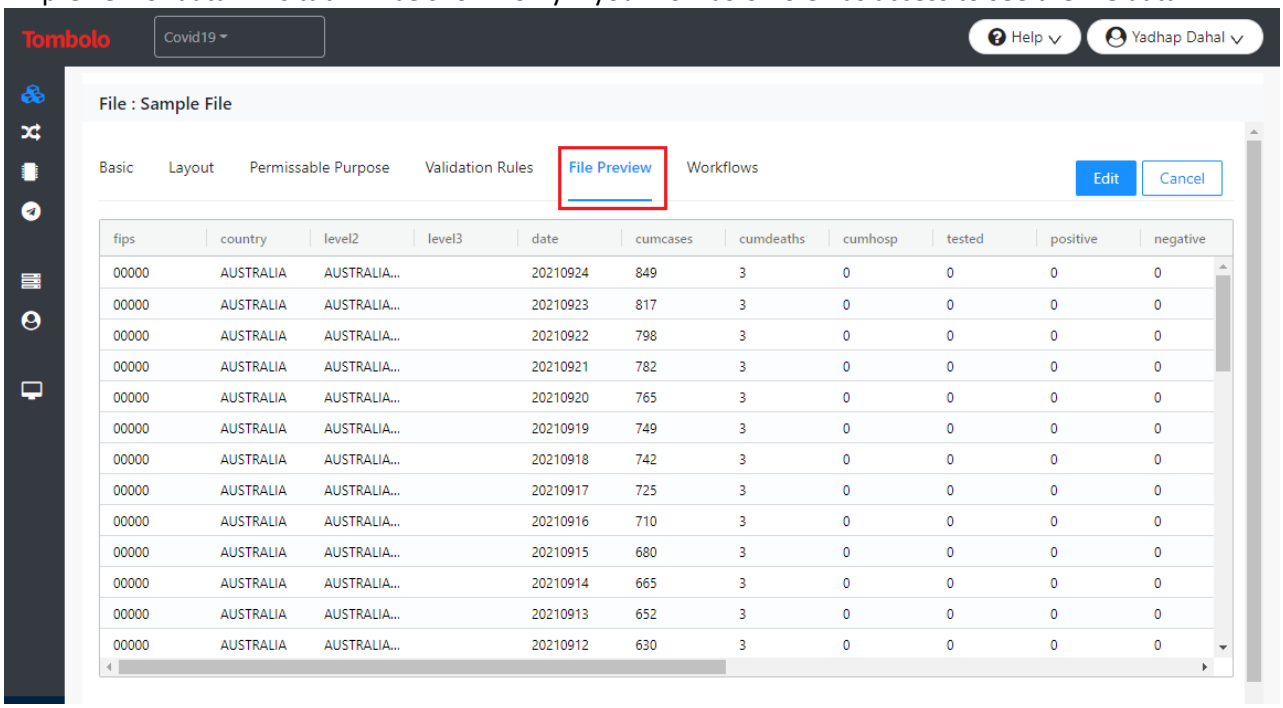
License Restrictions for files.

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

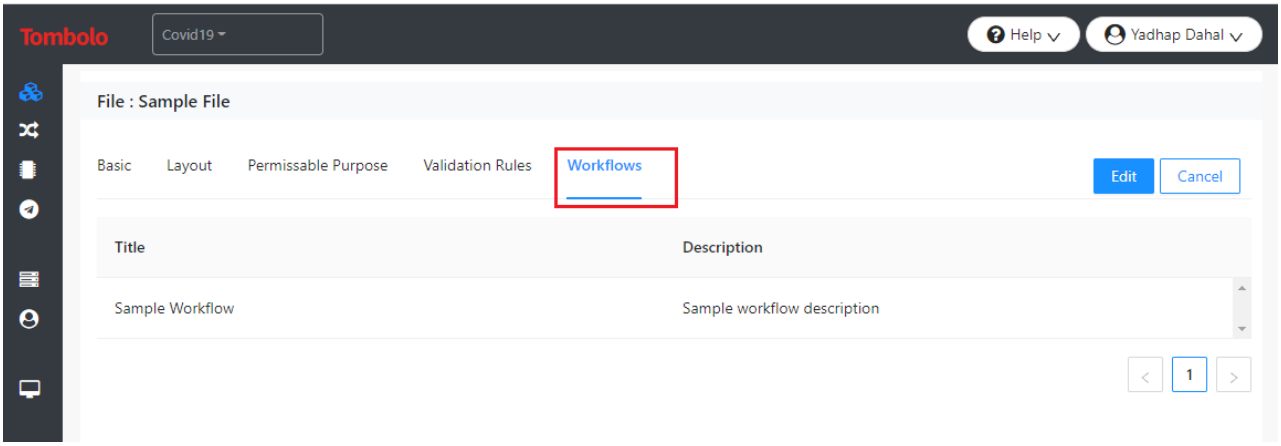


File Preview

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



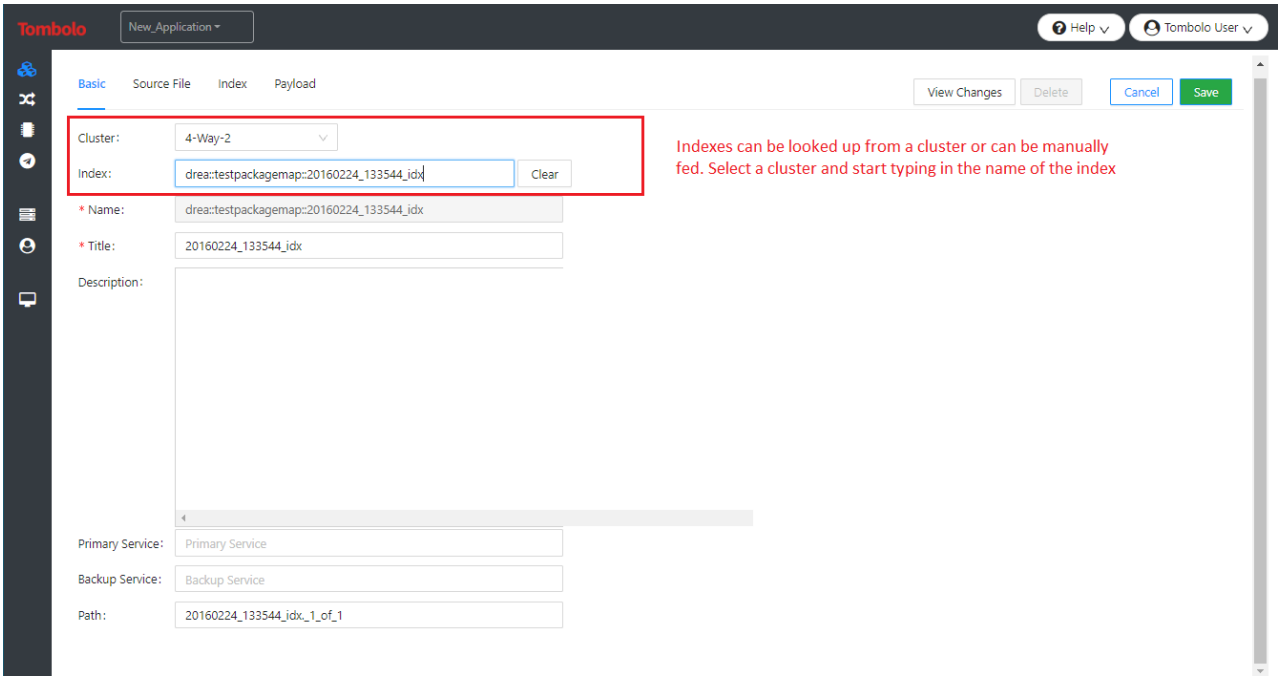
Workflows – Shows the Tombolo Dataflows this file belongs to



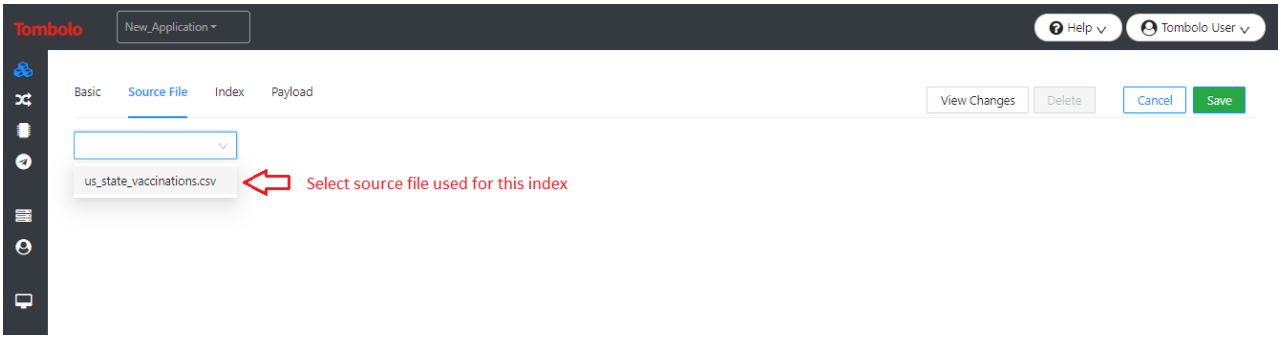
Indexes

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.

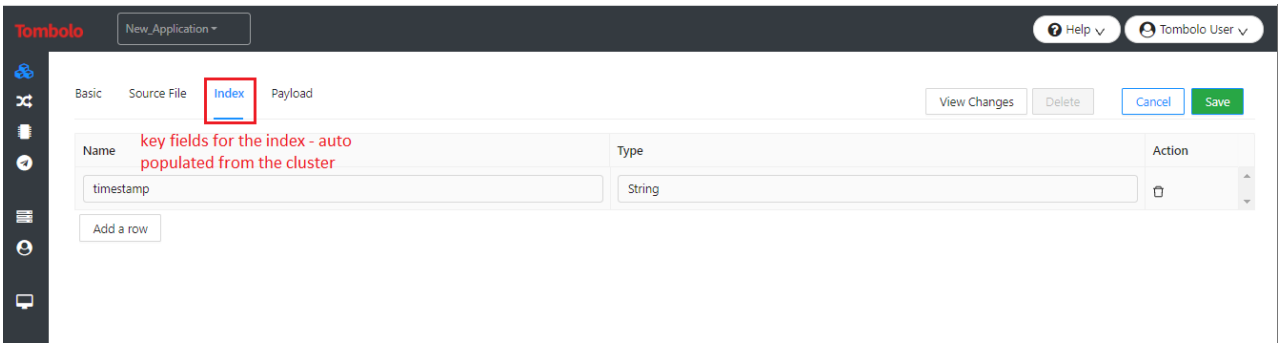
Basic Info



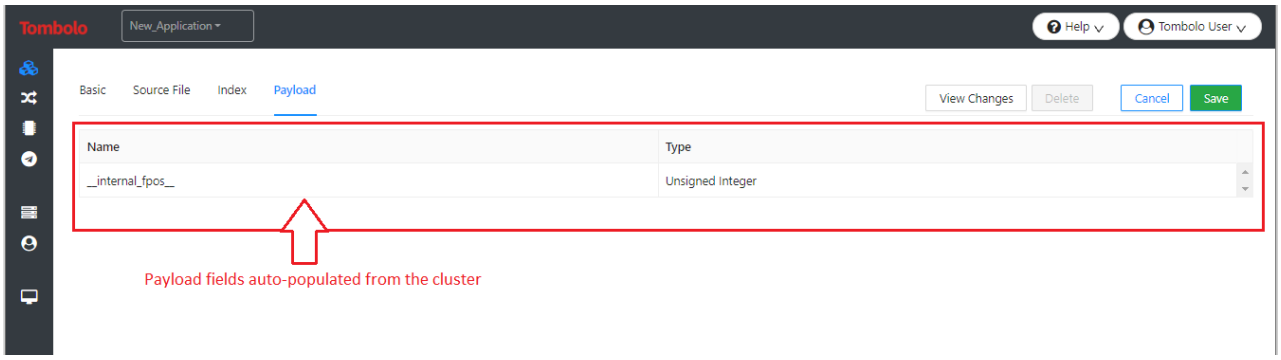
Source File



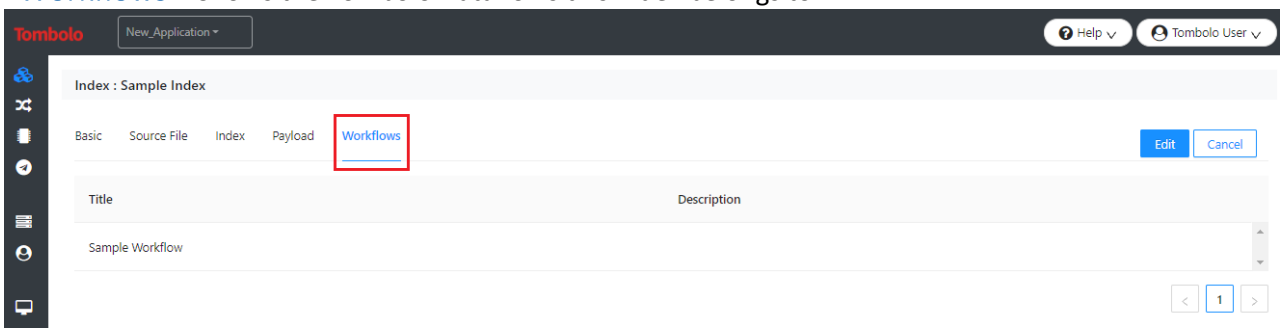
Index



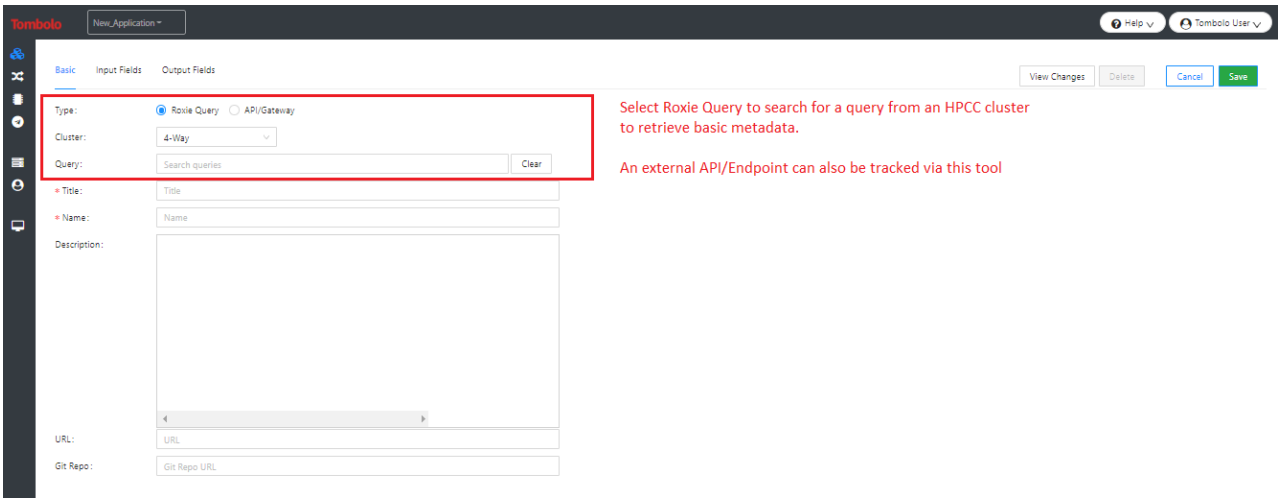
Payload



Workflows – Shows the Tombolo Dataflows this Index belongs to



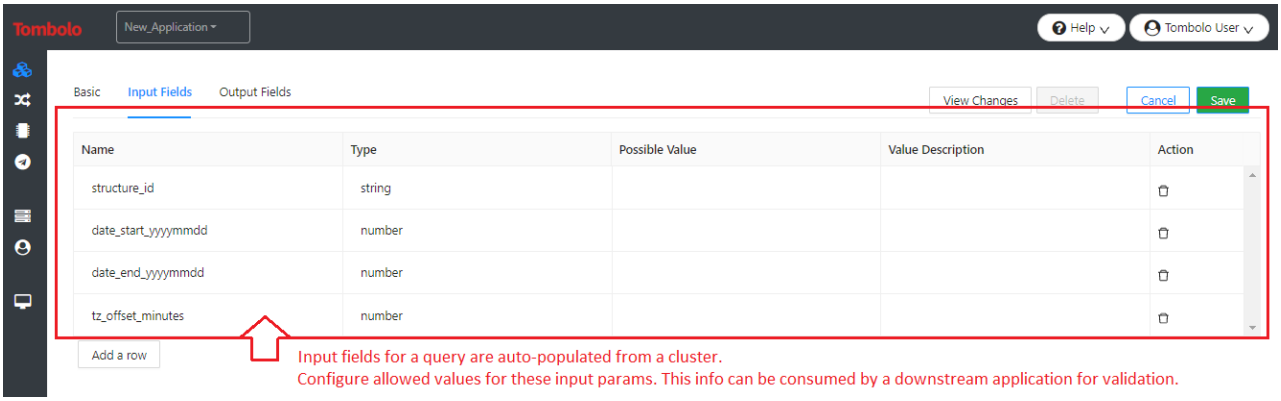
Queries



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

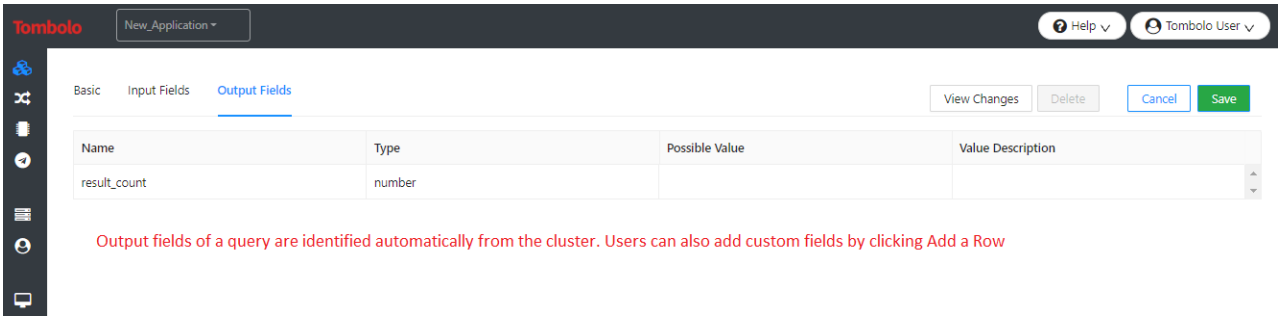
An external API/Endpoint can also be tracked via this tool

Input Fields

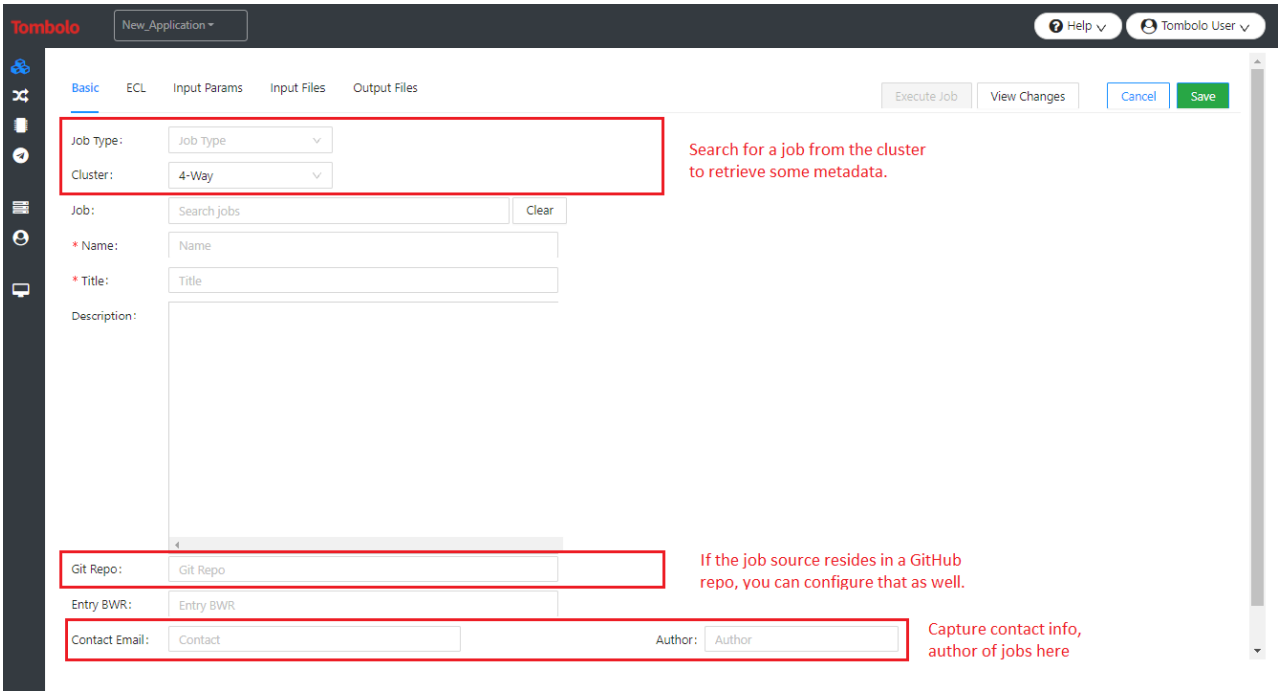


Input fields for a query are auto-populated from a cluster.
Configure allowed values for these input params. This info can be consumed by a downstream application for validation.

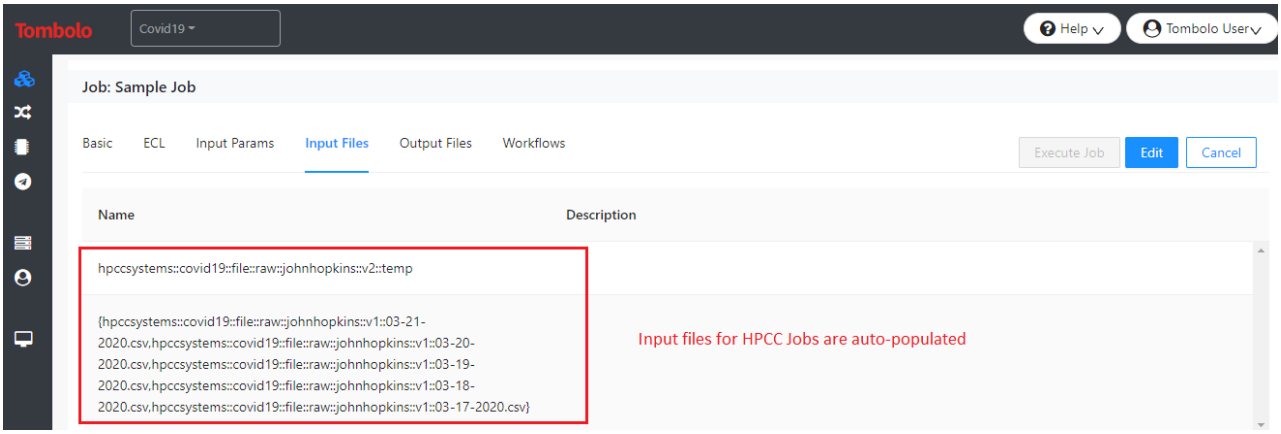
Output Fields



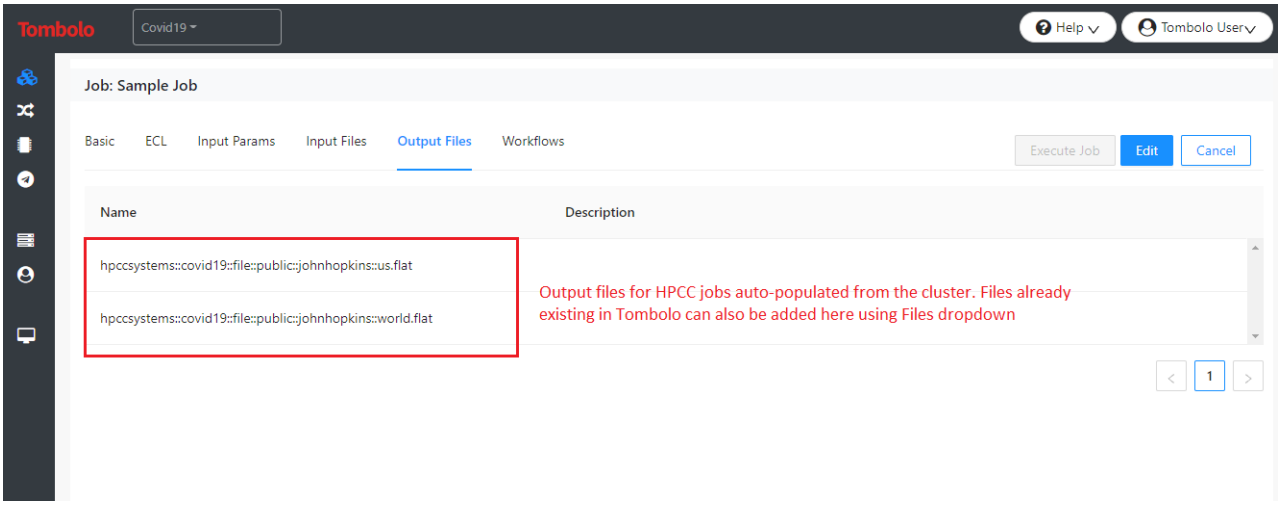
Job



Input Files



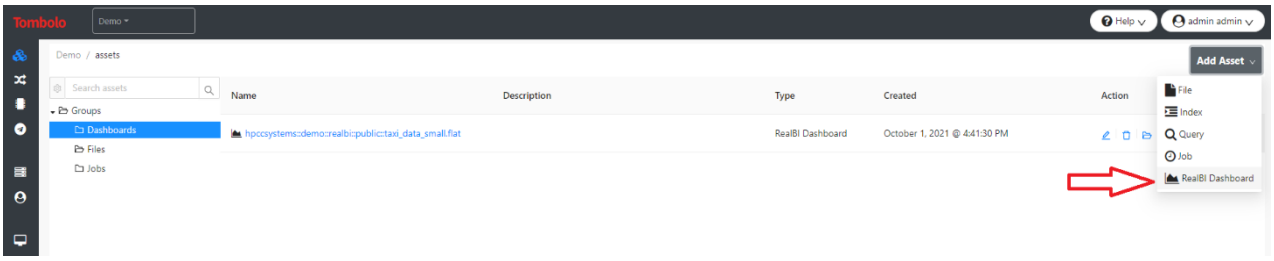
Output files



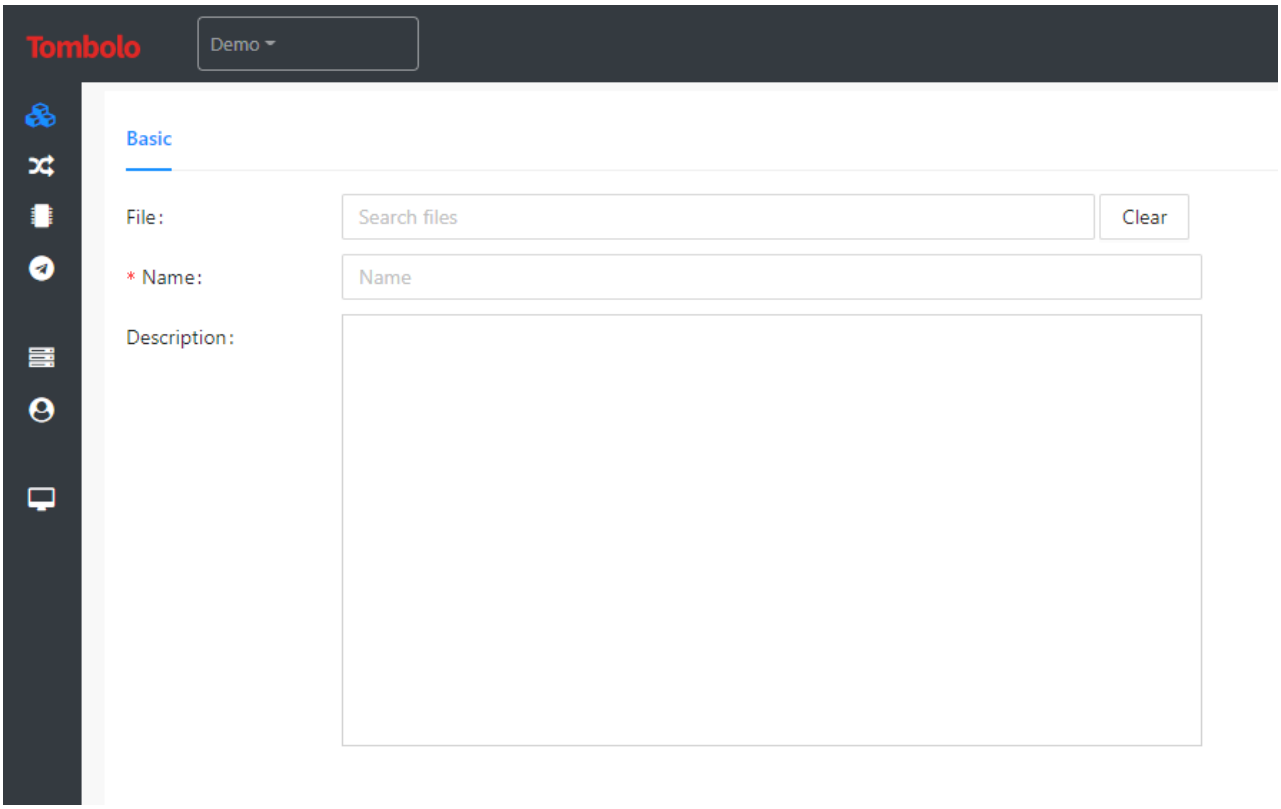
RealBI-Dashboards

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.

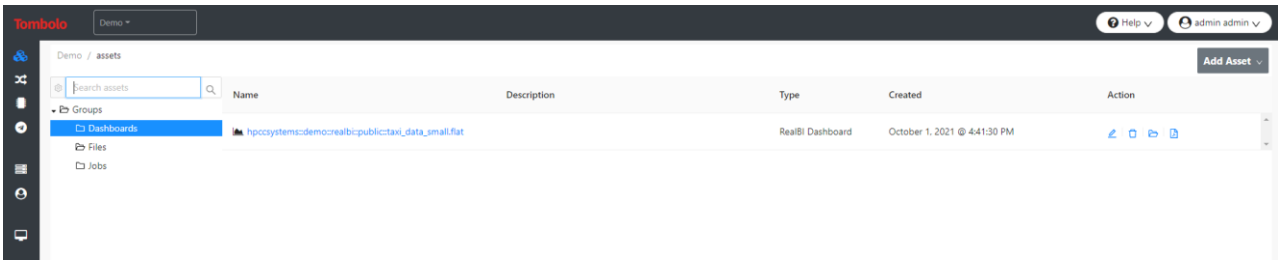
To create a RealBI dashboard from Tombolo, click on RealBI Dashboard option under Add Asset



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.



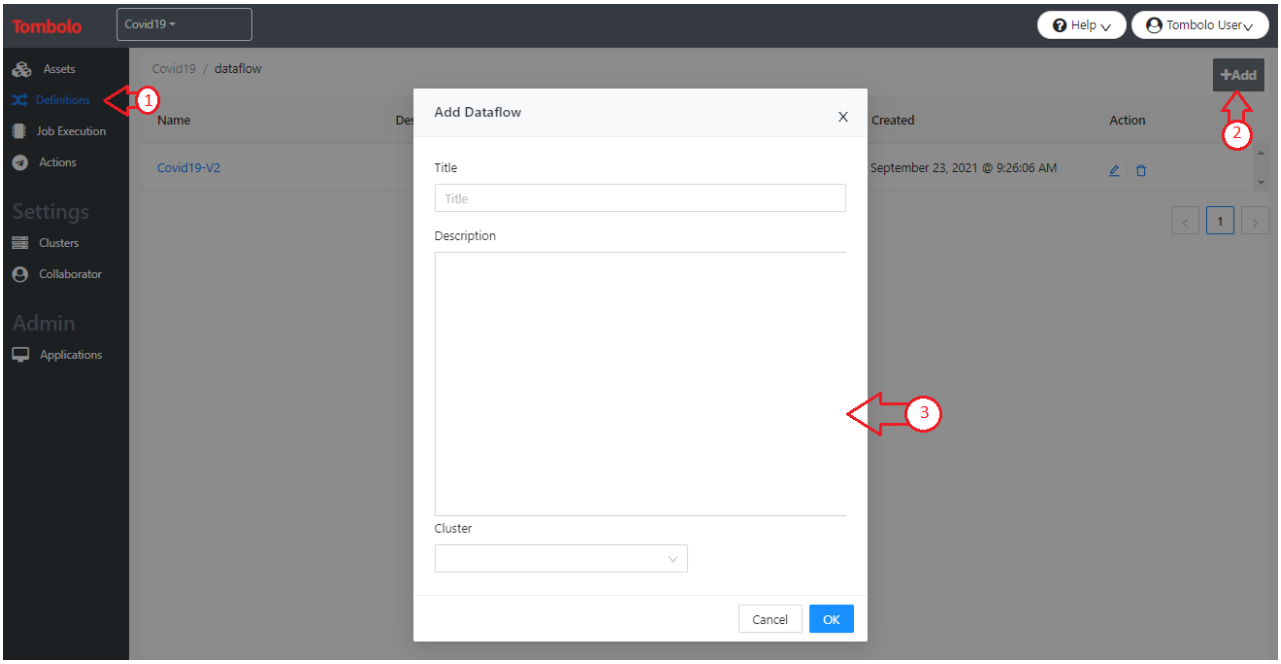
After the RealBI dashboard is created, Tombolo will show the Dashboard as an asset under the Group you selected while creating the Dashboard. Clicking on the dashboard name will directly open RealBI application in your browser.



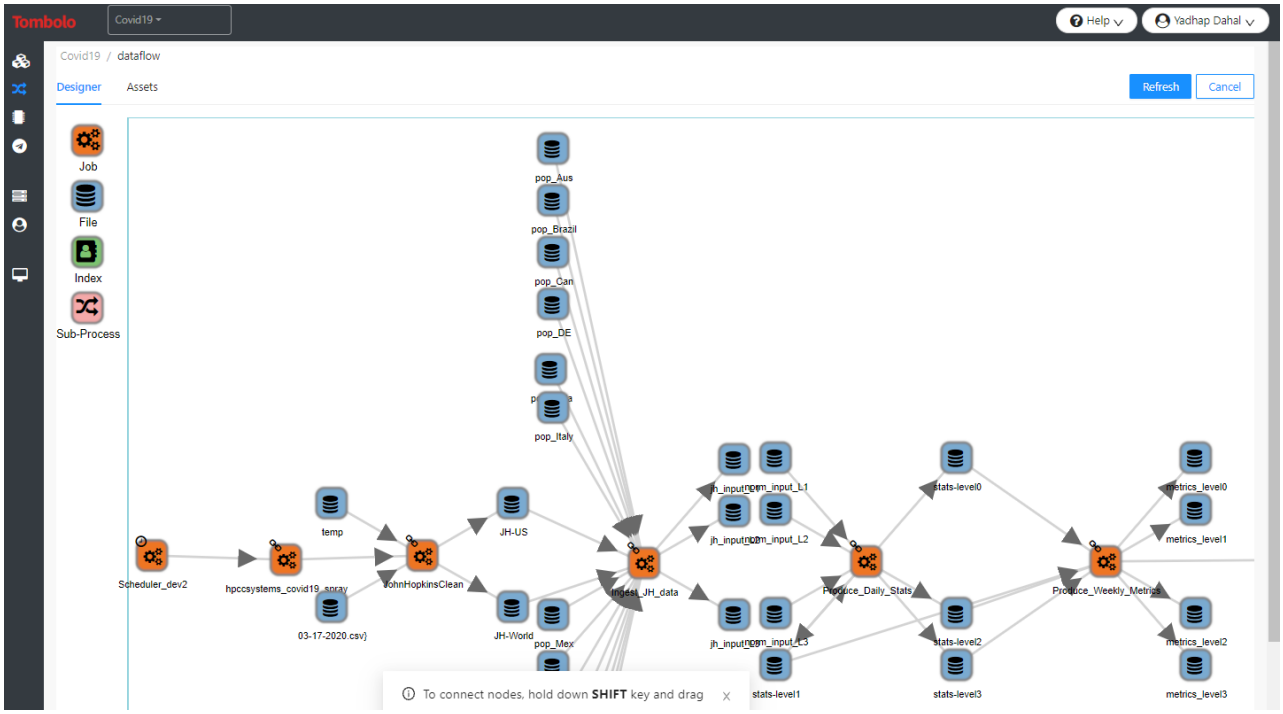
Workflow Definitions

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.


Tombolo
Covid19
Help
Yadhap Dahal
Covid19 / dataflow
Designer
Assets
Refresh
Cancel
Job
File
Index
Sub-Process
pop_Aus
pop_Brazil
pop_Can
pop_DE
pop_Italy
temp
JH-US
Scheduler_dev2
hpccsystems_covid19_spray
JohnHopkinsClean
03-17-2020.csv)
JH-World
pop_Mex
Ingest_JH_data
jh_input_com_input_L1
jh_input_com_input_L2
jh_input_com_input_L3
Produce_Daily_Stats
stats-level0
stats-level1
stats-level2
stats-level3
Produce_Weekly_Metrics
metrics_level0
metrics_level1
metrics_level2
metrics_level3
To connect nodes, hold down SHIFT key and drag

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.