

# **Visual Flow**

## **User Guide**

*November, 2021*

*Version 0.7*

## Document Revisions

Date	Version Number	Document Changes
08/12/2020	0.1	Initial Draft
04/22/2021	0.2	Pipeline Operators
04/26/2021	0.2	Job Operators
05/07/2021	0.3	Project Name, Project Operations
05/25/2021	0.4	Project Name in document
09/07/2021	0.5	Pipeline Operators, Job Operations, Storages
10/24/2021	0.6	Jobs and Pipelines statuses, Custom container, Storages
11/13/2021	0.7	New Data Storages

## Table of Contents

1	Introduction .....	4
1.1	...Terminology .....	4
1.2	...Scope and Purpose.....	5
1.3	...Process Overview .....	5
2	Roles and Authorizations .....	6
3	Project Operations .....	7
3.1	...Create Project.....	7
3.2	...Project Overview .....	8
3.3	...Manage Project Settings .....	9
4	Job Operations .....	11
4.1	...Jobs Overview .....	11
4.2	...Create a Job.....	12
4.3	...Job Designer Functions Overview .....	18
4.4	...Job Execution .....	18
5	Pipeline Operations.....	20
5.1	...Pipelines Overview .....	20
5.2	...Create a Pipeline .....	21
5.3	...Pipeline Designer Functions Overview.....	25
5.4	...Pipeline Execution .....	25

# 1. Introduction

## 1.1. Terminology

**ETL** is an abbreviation for *extract, transform, load*, three database functions combined into one tool to pull data out of one database, transform it and place it into another database.

- **Extract** is the process of *reading data* from a database. In this stage, the data is collected, often from multiple and different types of sources.
- **Transform** is the process of *converting the extracted data* from its previous form into the form needed to place it into another database.
- **Load** is the process of *writing the data* into the target database.

**Job** is a chain of individual stages linked together. It describes the flow of data from a data source to a data target. Usually, a stage has a minimum of one data input and/or one data output. However, some stages can accept more than one data input and output to more than one stage.

In Visual Flow, various stages user can use are:

- Read
- Write
- Join
- Union
- Filter
- Group By
- Remove Duplicates
- Transformer
- Change Data Capture

**Pipeline** is a compound of multiple jobs and can be run. In Visual Flow, user can use such stages as:

- Job
- Notification
- Container

## 1.2. Scope and Purpose

Visual Flow web application is an ETL tool designed for effective data manipulation via convenient and user-friendly interface.

The tool has the following capabilities:

- Can integrate data from heterogeneous sources:
  - ✓ AWS S3
  - ✓ DB2
  - ✓ Cassandra
  - ✓ Elastic Search
  - ✓ IBM COS
  - ✓ Mongo
  - ✓ MSSQL
  - ✓ MySQL
  - ✓ Oracle
  - ✓ PostgreSQL
- Leverage direct connectivity to enterprise applications as sources and targets
- Perform data processing and transformation
- Leverage metadata for analysis and maintenance

## 1.3. Process Overview

Visual Flow jobs and pipelines exist within a certain namespace (project) so the first step in the application would be to create a project or enter an existing project. Then user needs to enter Job Designer to create a job.

*Job Designer* is a graphical design interface used to create, maintain, execute and analyze jobs. Each job determines the data sources, the required transformations and destination of the data.

*Pipeline designer* is a graphical design interface aimed for managing pipelines. Designing a pipeline is similar to designing a job.

Visual Flow key functions include, but not limited to

- ✓ Create project which serves as a namespace for jobs and/or pipelines
- ✓ Manage project settings
- ✓ User access management
- ✓ Run custom code
- ✓ Create/maintain a job in Job Designer
- ✓ Job execution and logs analysis
- ✓ Create/maintain a pipeline in Pipeline Designer
- ✓ Pipeline execution
- ✓ Import/Export jobs and pipelines

## 2. Roles and authorizations

The following roles are available in the application:

- ✓ Viewer
- ✓ Operator
- ✓ Editor
- ✓ Administrator

They can perform the below operations within the namespaces they are authorized to. Only a Super-admin user can create a workspace (project) and grant access to this project.

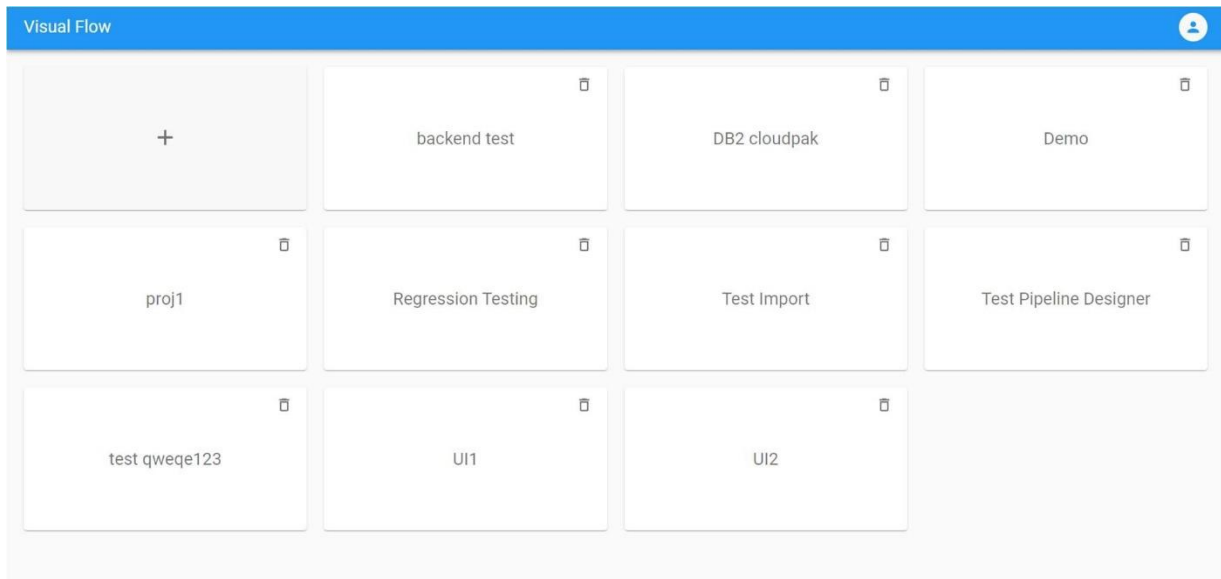
Role	Actions		
	Project Settings	Jobs	Pipelines
Viewer	View all	View all	View all
Operator	View all	View all / execute jobs	View all / execute pipelines
Editor	Edit all but Users and Roles	Edit / execute jobs	Edit / execute pipelines
Admin	Edit all	Edit / execute jobs	Edit / execute pipelines

### 3. Project operations

#### 3.1. Create a Project

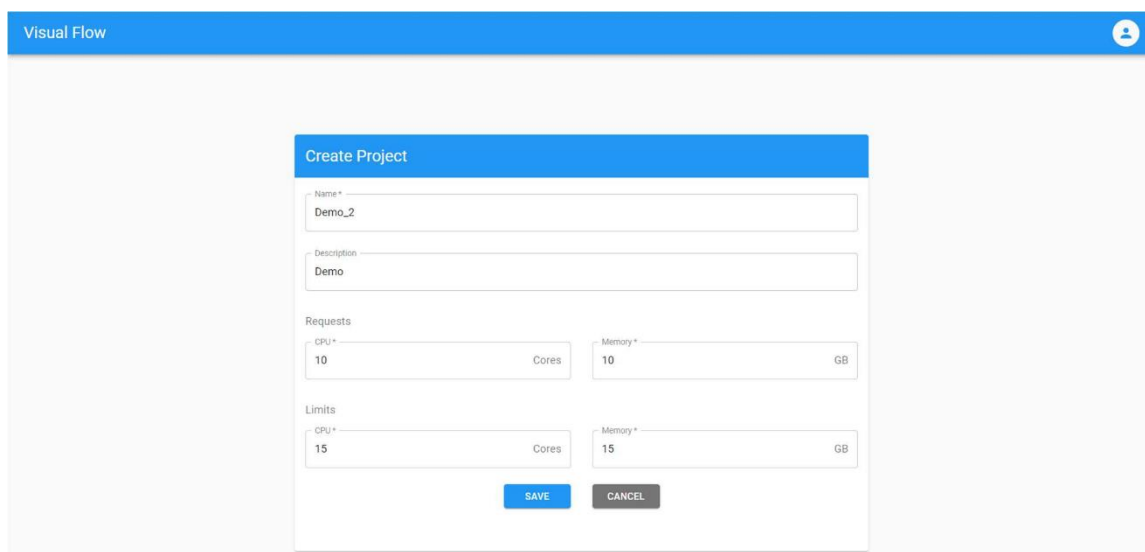
To create a project, user needs to push “+” button on the initial screen.

Note: this is an action of super-admin user only. The button is not visible for the application roles (Viewer, Operator, Editor, Admin).

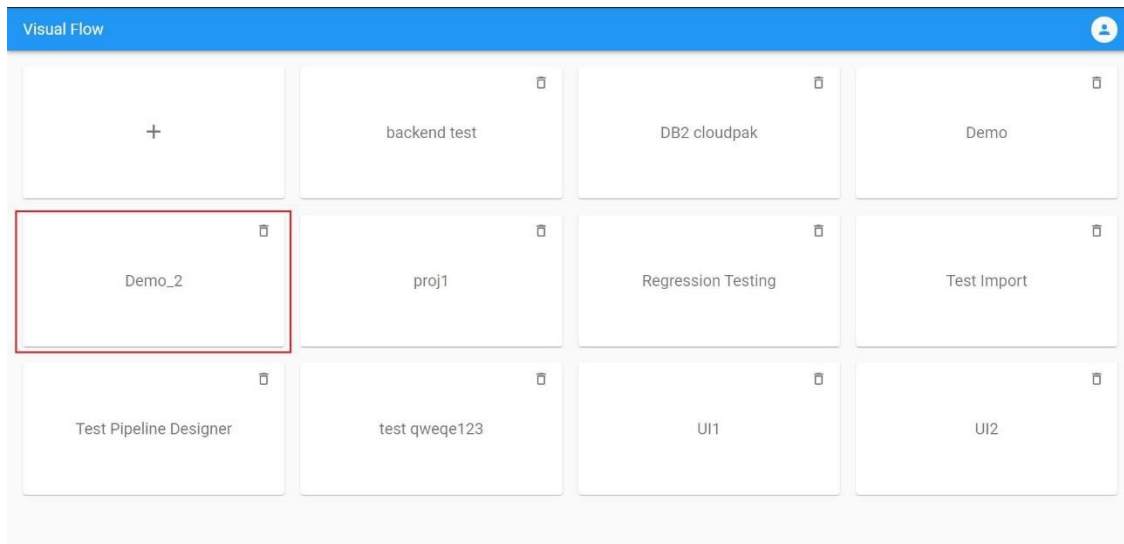


With “+” button pushed, user will get to *Create Project Form* to enter project basic settings:

- Project Name
- Project Description
- Requests (CPU/Memory)
- Limits (CPU/Memory)

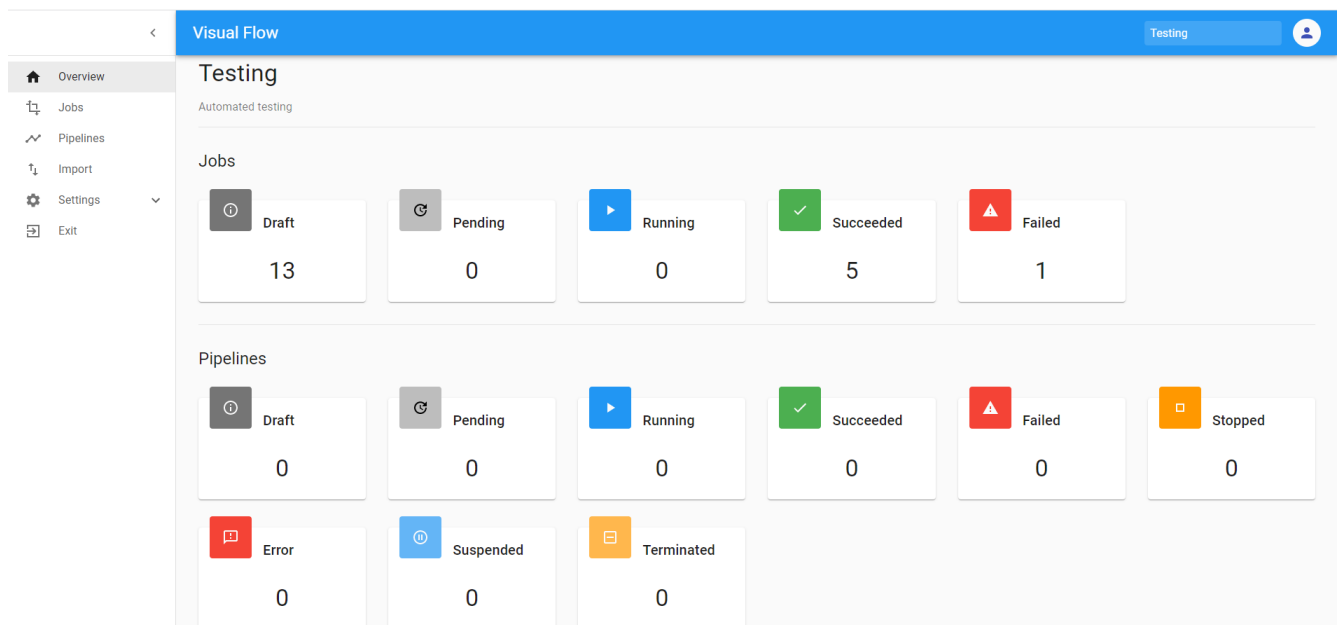
The screenshot shows the 'Create Project' form. It has a blue header bar with the text 'Visual Flow' on the left and a user profile icon on the right. The form itself is a white box with a blue header bar containing the text 'Create Project'. Inside the form, there are several input fields: 'Name \*' with the value 'Demo\_2', 'Description' with the value 'Demo', 'Requests' section with 'CPU \*' set to '10' (Cores) and 'Memory \*' set to '10' (GB), and 'Limits' section with 'CPU \*' set to '15' (Cores) and 'Memory \*' set to '15' (GB). At the bottom of the form are two buttons: 'SAVE' (blue) and 'CANCEL' (gray).

After saving *Create Project Form*, the project created under the given name and then can be found on the initial screen:



### 3.2. Project Overview

Click the project card to enter the newly created project, and user will get to the *ProjectOverview Screen*:



The screen contains project left menu and displays information about the project jobs, pipelines and their resource utilization (applicable for running jobs).



### 3.3. Manage Project Settings

*Settings* submenu contains:

- Basic
- Parameters
- Users and Roles

1) *Basic* is already there after project creation. *Edit* button turns on the edit mode for updates.

The screenshot shows the 'Visual Flow' application interface. On the left is a sidebar with a menu: Overview, Jobs, Pipelines, Import, Settings (with a sub-menu arrow), Basic, Parameters, Users/Roles, and Exit. The 'Settings' sub-menu is expanded, showing 'Basic' as the selected option. The main content area displays a 'View Project' dialog box. The dialog has a title bar with 'View Project' and an edit icon. It contains the following fields: 'Name' (value: Demo\_2), 'Description' (value: Demo), 'Requests' (CPU: 10 Cores, Memory: 10 GB), and 'Limits' (CPU: 15 Cores, Memory: 15 GB).

*Parameters* serve to store values required for the entire project, e.g. JDBC connection, DB2 credentials or table schemas can be the same for all jobs within the project and therefore stored at the project level. *Edit* button turns on the edit mode for updates.

The screenshot shows the 'Visual Flow' application interface. On the left is a sidebar with a menu: Overview, Jobs, Pipelines, Import, Settings (with a sub-menu arrow), Basic, Parameters, Users/Roles, and Exit. The 'Settings' sub-menu is expanded, showing 'Parameters' as the selected option. The main content area displays a 'View Project Parameters' dialog box. The dialog has a title bar with 'View Project Parameters' and an edit icon. It contains a search bar and a list of parameters with their values: 'accessKey' (1ae5ab46ec004860af18a9de3aa334c9), 'bucket' (big-data-education), 'endpoint' (s3.eu-de.cloud-object-storage.appdomain.cloud), 'index' (vsw-test), 'jdbc' (jdbc:db2://10.224.0.52:30100/EXAMPLE), 'nodes' (23434ed07a9405ca751a3a764027b69.us-east-1.aws.fo), and 'nodes1' (elastic.okd.comel.lba.bv).

2) *User and Roles* allows user access management or view user access depending on authorization.

The user cannot change his role, this operation can be done by an Admin or a Super-admin. If the user tries to change his role, the error will occur «You cannot change your role".

*Edit* button and therefore Edit mode is only available for admin within the project or super-admin.

The screenshot shows the 'Visual Flow' application interface. On the left is a sidebar menu with options: Overview, Jobs, Pipelines, Import, Settings (expanded), Basic, Parameters, Users/Roles (selected), and Exit. The main content area has a blue header 'Visual Flow' with a 'Demo' button and a user icon. Below this is a 'Users and Roles' modal window with a search bar and a table of users.

<input type="checkbox"/>	ID	Name	Role
<input type="checkbox"/>	ABandarenka	Бондаренко Антон Алексеевич	vf-admin
<input type="checkbox"/>	AHud	Гуд Алексей Сергеевич	vf-editor
<input type="checkbox"/>	AKrauchanka	Кравченко Олег Сергеевич	vf-admin
<input type="checkbox"/>	ASamoilenka	Самойленко Артём Павлович	vf-viewer

At the bottom of the table, it says 'Rows per page: 5' and '1-4 of 4'.

## 4. Job Operations

### 4.1. Jobs Overview

Clicking *Jobs* menu item will lead user to *Jobs Overview Screen*, which allows user to see a list of jobs existing within a project. Some of the jobs can be used in pipelines, this is indicated by the



icon.

Jobs Overview Screen displays the following information:

- Job Name
- Job Last run/Last finished/Last edit
- Resource Utilization (CPU/Memory)
- Available Actions (Run/Job Designer/Logs/Copy/Delete)

Job has a certain status at various phases of execution:

- Draft
- Pending
- Running
- Succeeded
- Failed
- Unknown (This status appears very rarely in the case of an undefined error)

Notes:

- The actions availability and therefore visibility is depending on user authorizations
- The user cannot delete job that is used in pipeline

NAME	LAST RUN	STATUS	CPU	Memory	Actions
Demo1_COS_trans	Last Run: N/A; Last Finished: N/A; Last Edit: 2021-03-22 08:56:47	Draft	0%	0%	Run, Copy, Delete
Demo2_union_TestOne	Last Run: N/A; Last Finished: N/A; Last Edit: 2021-02-26 11:24:55	Draft	0%	0%	Run, Copy, Delete
Demo2_union_TestOne	Last Run: 2021-04-02 10:52:45; Last Finished: 2021-04-02 10:53:11; Last Edit: 2021-02-26 11:24:55	Failed	0%	0%	Run, Copy, Delete
Job_CDC	Last Run: 2021-04-01 12:44:00; Last Finished: 2021-04-01 12:44:08; Last Edit: 2021-03-10 07:29:02	Failed	0%	0%	Run, Copy, Delete
Job_CDC	Last Run: 2021-04-01 15:15:21; Last Finished: 2021-04-01 15:16:04; Last Edit: 2021-03-10 07:29:02	Succeeded	0%	0%	Run, Copy, Delete

## 4.2. Create a Job

With *Add Job* button pushed, user will get to *Job Designer* for creating a new job.

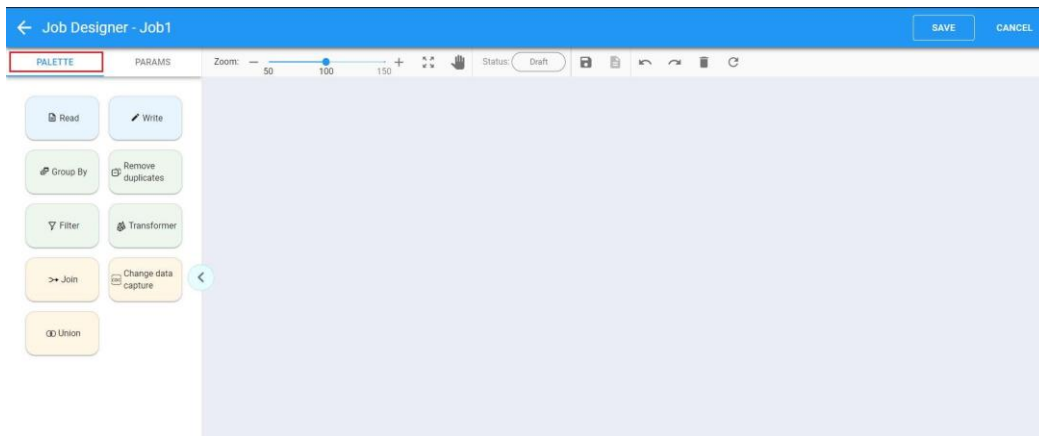
1) On the left configuration panel, user will need to give job a name, update parameters or keep their default values and then push *Confirm* on the panel:

The screenshot shows the 'Job Designer' interface with the header 'Please enter name and save params'. On the left, there is a configuration panel with the following fields: Name (Job1), Driver Request Cores (0,1), Driver Cores (1), Driver Memory (1 GB), Executor Request Cores (0,1), Executor Cores (1), Executor Memory (1 GB), Executor Instances (2), and Shuffle Partitions (10). At the bottom of the panel are 'CONFIRM' and 'DISCARD' buttons. The main area is a large empty canvas with a zoom slider (50, 100, 150) and a status 'Draft'.

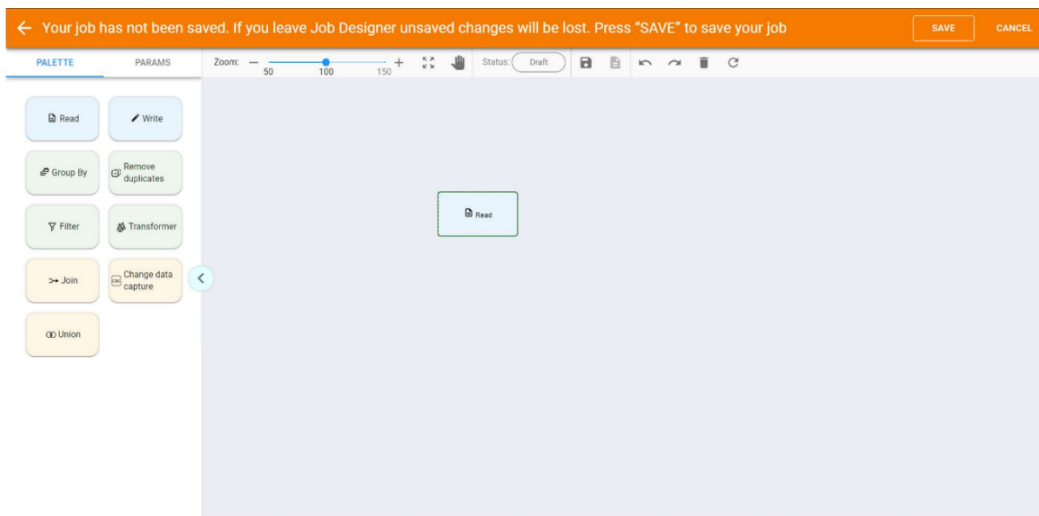
The screenshot shows the 'Job Designer' interface with the header 'Your job has not been saved. If you leave Job Designer unsaved changes will be lost. Press "SAVE" to save your job'. On the left, there is a configuration panel with the following fields: Name (Job1), Driver Request Cores (0,1), Driver Cores (1), Driver Memory (1 GB), Executor Request Cores (0,1), Executor Cores (1), Executor Memory (1 GB), Executor Instances (2), and Shuffle Partitions (10). At the bottom of the panel are 'CONFIRM' and 'DISCARD' buttons. The main area is a large empty canvas with a zoom slider (50, 100, 150) and a status 'Draft'.

2) Save the job by pushing *Save* button on the *Job Designer* header.

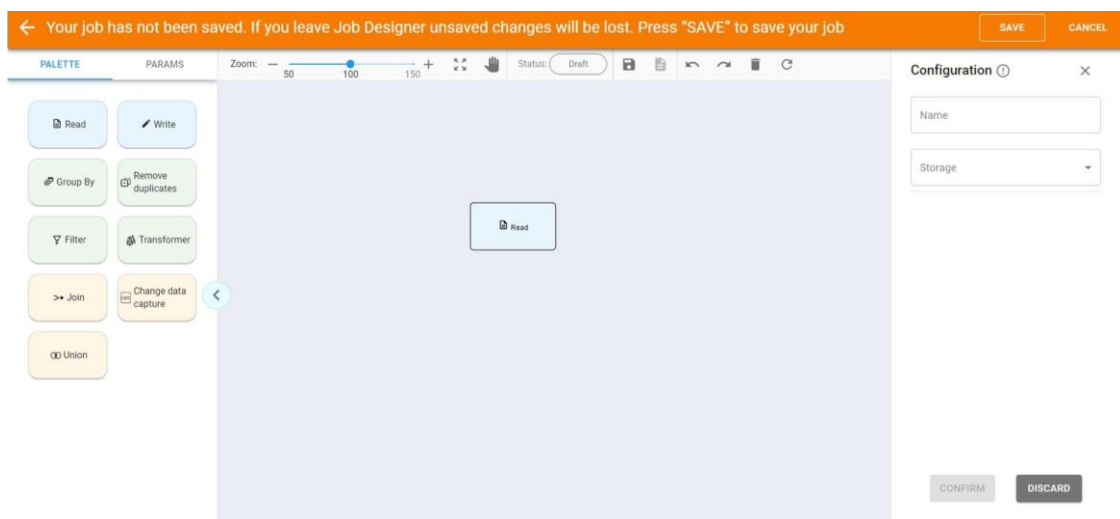
3) Go to *Palette* tab to see all available stages:



4) User can start creating a job by dragging a stage to the canvas, e.g. user can drag *Read* stage:



5) Double-click on the stage will open the configuration panel on the right:




Enter name for the stage and select *Storage* DB2 if user wants to read data from DB2 table.

The image shows a 'Configuration' dialog box with a title bar containing a close button (X) and a help icon (?). The dialog has several fields: 'Name' with the value 'Read\_stage\_DB2', 'Storage' with a dropdown menu showing 'DB2', 'JDBC URL' with a 'Parameters' icon and a close button (X), 'User' with a 'Parameters' icon and a close button (X), 'Password' with a 'Parameters' icon and a close button (X), 'Custom SQL' with a dropdown arrow, and 'CertData (optional)' with a 'Parameters' icon and a close button (X).

Available *Storage* values for Read stage are:

- ✓ AWS S3
- ✓ DB2
- ✓ Cassandra
- ✓ Elastic Search
- ✓ IBM COS
- ✓ Mongo
- ✓ MSSQL
- ✓ MySQL
- ✓ Oracle
- ✓ PostgreSQL

6) Fill required parameters for DB2 *Storage*.

Important: user can pick up a parameter value with *Parameters*  button on the right panel if user has it previously created as project parameters.

The image shows a 'Configuration' dialog box with a title bar containing a close button (X) and a help icon (?). The dialog has several fields: 'Name' with the value 'Read\_stage\_1', 'Storage' with a dropdown menu showing 'DB2', 'JDBC URL' with a 'Parameters' icon and a close button (X). The 'JDBC URL' field and its 'Parameters' icon are highlighted with a red rectangle.

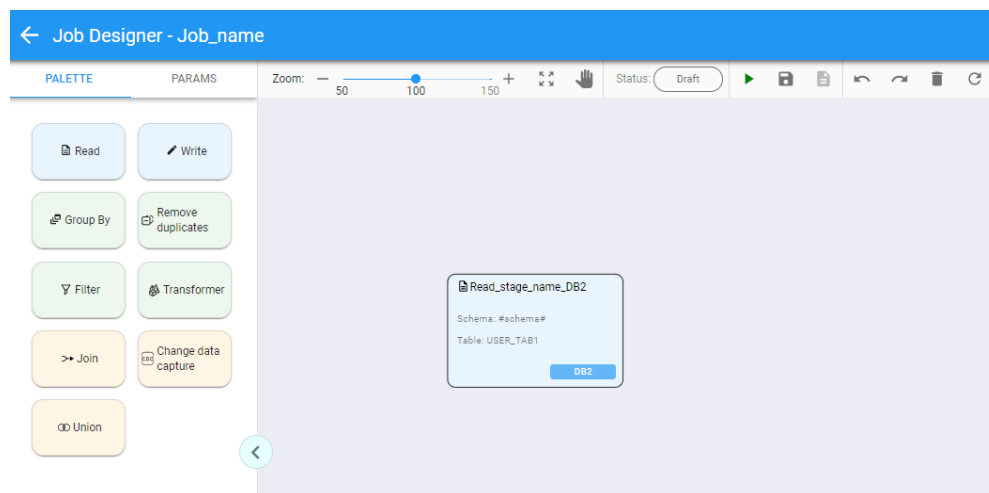
For the DB2 storage, user can use *Custom SQL* only Read stage (e.g. *select \* from table where field = value*). Displays the schema and the table fields, if user chooses false. If user chooses true, user will be able to write his own SQL code in the provided field.

Custom SQL

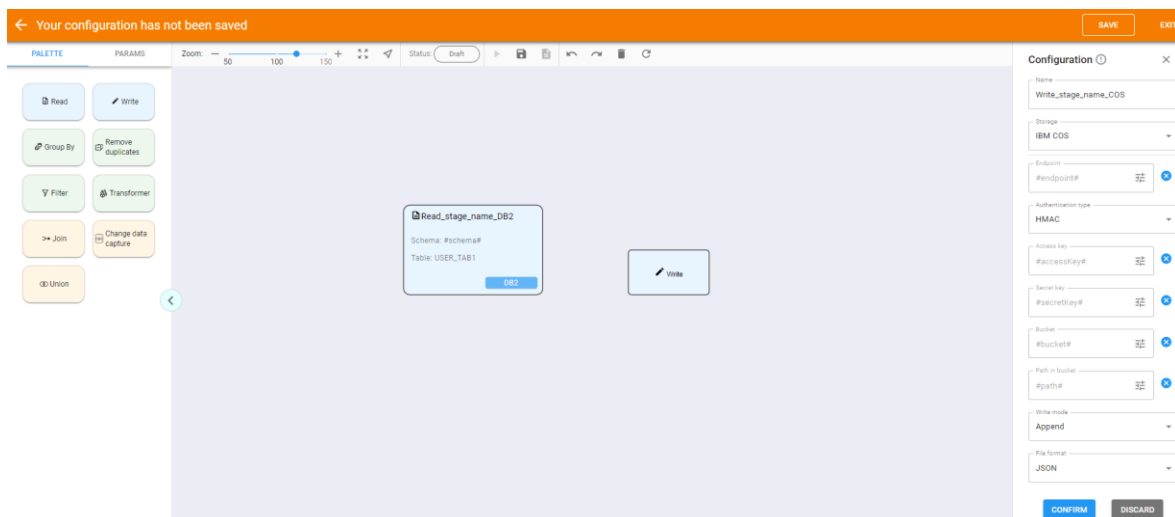
SQL statement

7) Save the stage by pushing Confirm button on the configuration panel. If user wants to save his job at this step, user should press *Save* button on the header.

User has configured the first stage of the job, and it now looks like this:



8) Now drag another stage, e.g. *Write* stage:



9) Enter a name for the stage and select *Storage* IBM COS if user wants to post data from the DB2 table to Cloud Object Storage file. Fill required parameters for IBM COS *Storage*.

Available *Storage* values for Write stage are:

- ✓ AWS S3
- ✓ DB2
- ✓ Cassandra
- ✓ Elastic Search
- ✓ IBM COS
- ✓ Mongo
- ✓ MSSQL
- ✓ MySQL
- ✓ Oracle
- ✓ PostgreSQL
- ✓ STDOUT

For *IBM COS Storage*, user can use *Authentication type*. Authentication type displays accessKey and secretKey, if user chooses HMAC, or iamApiKey and iamServiceId, if user chooses IAM.

The image shows two side-by-side screenshots of a configuration interface. Both screenshots feature a dropdown menu labeled 'Authentication type'. In the left screenshot, 'HMAC' is selected, and below it are two text input fields: 'Access key' and 'Secret key'. In the right screenshot, 'IAM' is selected, and below it are two text input fields: 'IAM api key' and 'IAM service id'. Each text input field includes a copy icon (three horizontal lines) and a clear icon (an 'x' in a circle).

In order to import table data to *Cassandra* source with *Write* stage from another database, at first, user needs to create a layout of the table in *Cassandra* that he wants to output. Create columns, define a key field, correctly specify the data type of the fields of the future table.

*Important:*

All the above points must match the imported table.

If the column names have uppercase characters in the imported table, when data is output to *Cassandra*, the job will be failed. The reason is that in *Cassandra*, the column names are stored only as lowercase characters. This problem can be solved using a *Transformer* stage.

*Important:*

*Write mode* field defines how data will be posted to its destination. Available values are:

- ✓ Overwrite
- ✓ Append
- ✓ Error if Exists



*File format* is to choose a format of destination file. Available formats are:

- ✓ CSV
- ✓ JSON
- ✓ Parquet
- ✓ ORC
- ✓ Text

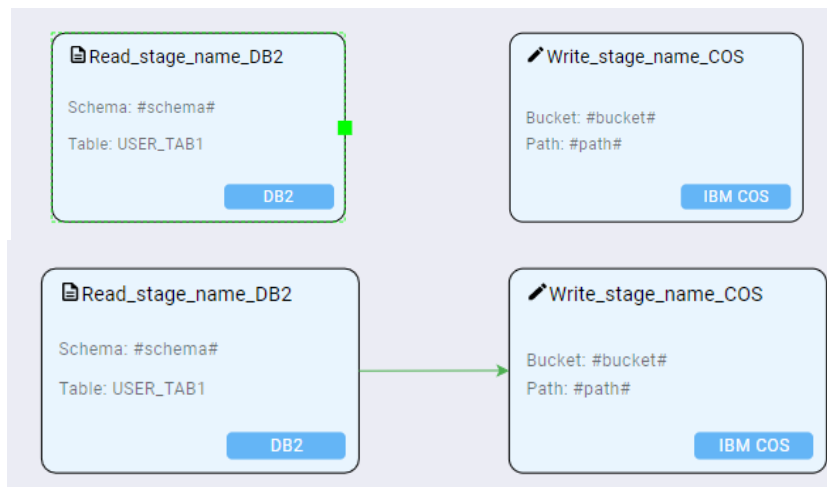
10) Save the stage by pushing *Save* on the panel.

11) Now user has two stages to connect to each other.



*Important:*

To connect stages, hover his mouse on a stage edge until user sees a green rectangle. Click it and drag it to the border of another stage and its green rectangle. When user reach it, a green arrow should appear.

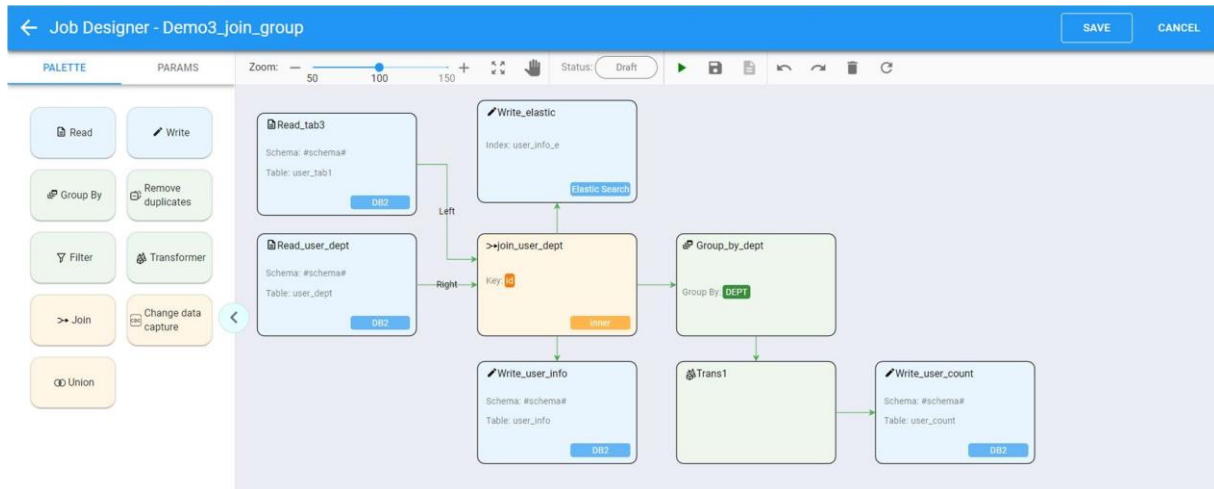


13) Save the job by pushing *Save* on the *Job Designer* header.

User has created a job reading data from the DB2 table and posting it to the IBM COS file. For newly created job, before he run it the status will be *Draft*:

Status: Draft

Drag other stages according to the flow of user job from source to destination. See the job with more stages as the example:



### 4.3. Job Designer functions overview

The following functions are available in *Job Designer*:

- ✓ Zoom operations:
- ✓ Show job status:
- ✓ Run job / Stop job (for running)
- ✓ Save job
- ✓ See job logs
- ✓ Undo / Redo operation on canvas
- ✓ Remove element from canvas
- ✓ Refresh


### 4.4. Job Execution

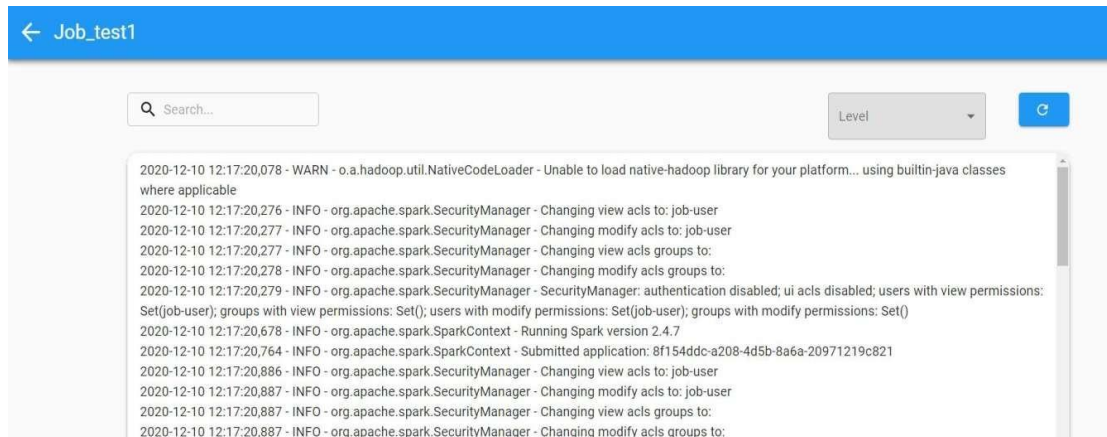
Push *Play* button to run the job:

User will see its status changed from *Draft* to *Pending*

Push *Refresh* to update the status. It should turn to *Running*

While running, it can be interrupted with *Stop* button. When job completed the status will be *Succeeded* or *Failed*

Use *Logs* button  to analyze job logs. User will get to *Logs Screen*:



*Logs Screen* has several levels:

- ✓ WARNING
- ✓ INFO
- ✓ ERROR
- ✓ DEBUG

## 5. Pipeline Operations

### 5.1. Pipelines Overview

Clicking *Pipelines* menu item will take user to *Pipelines Overview Screen*, which allows user to see a list of pipelines existing within a project.

It displays the following information:

- Pipeline Name
- Checkbox for deleting/exporting the pipeline
- Pipeline Last run/Last finished/Last edit
- Pipeline Status
- Pipeline Progress
- Available Actions (Run/Pipeline Designer/Copy/Delete)

Pipeline has a certain status at various phases of execution:

- Draft
- Running
- Succeeded
- Error (This status appears, e.g., due to incorrectly entered data)
- Terminated
- Suspended (This status can be reproduced via the API)
- Stopped
- Failed

Note: the actions availability and therefore visibility is depending on user authorizations.

Overview

Jobs

Pipelines

Import

Settings

Basic

Parameters

Users/Roles

Exit

Visual Flow

Test Pipeline Designer

Pipelines

Search...

+

ADD PIPELINE

☐

NAME

LAST RUN

STATUS

Status

Last Run

1-5 of 26

☐

Pipeline\_1

Last Run: 2021-08-22 11:52:45; Last Finished: 2021-08-22 11:53:05; Last Edit: 2021-07-23 18:45:57

Status

Terminated

Progress

100%

☐

Demo\_test1

Last Run: 2021-08-02 06:00:48; Last Finished: 2021-08-02 06:01:09; Last Edit: 2021-07-26 17:51:55

Status

Terminated

Progress

100%

☐

test\_pipe1

Last Run: 2021-09-02 11:56:18; Last Finished: 2021-09-02 11:56:28; Last Edit: 2021-09-02 11:55:36

Status

Succeeded

Progress

100%

☐

test\_pipe2

Last Run: 2021-09-02 12:13:38; Last Finished: 2021-09-02 12:24:55; Last Edit: 2021-09-02 12:12:59

Status

Succeeded

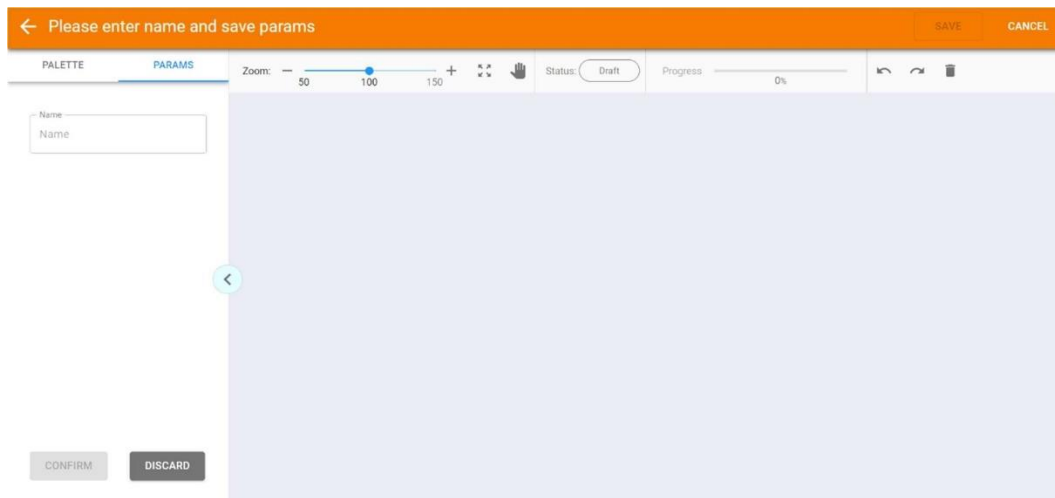
Progress

100%

## 5.2. Create a Pipeline

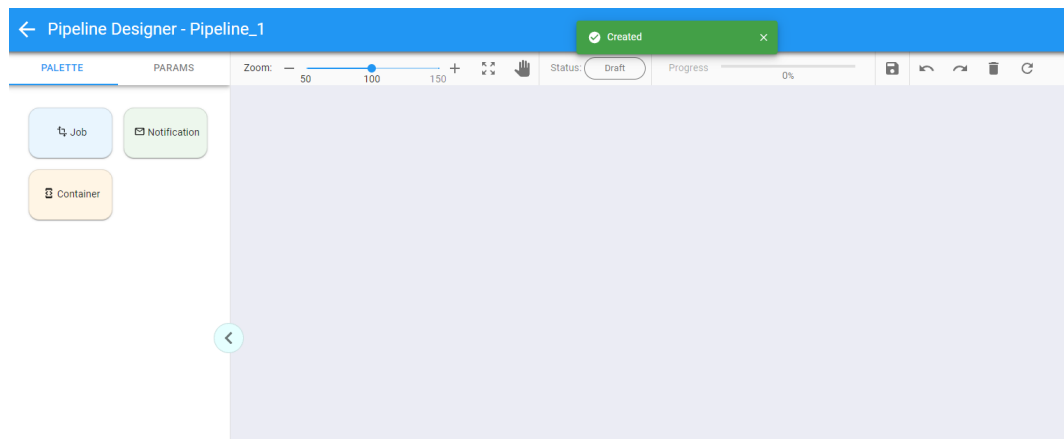
With *Add Pipeline* button pushed, user will get to *Pipeline Designer* for creating a pipeline.

1) On the left configuration panel *Params* tab is opened by default, user can enter pipeline name and push *Confirm* button on the panel:



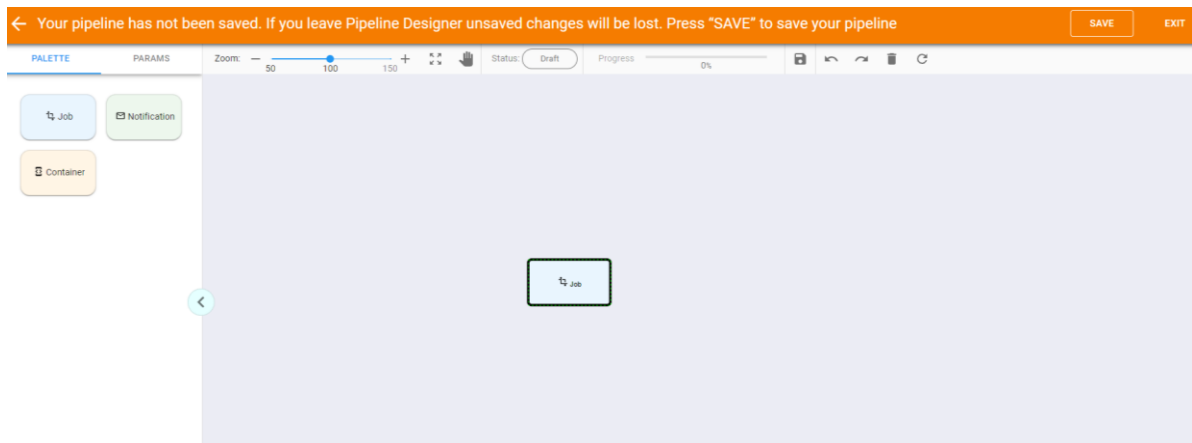
2) Save the pipeline by pushing *Save* button on the *Pipeline Designer* header.

3) After saving the pipeline, *Palette* tab is opened by default, at this tab user can see all available stages:

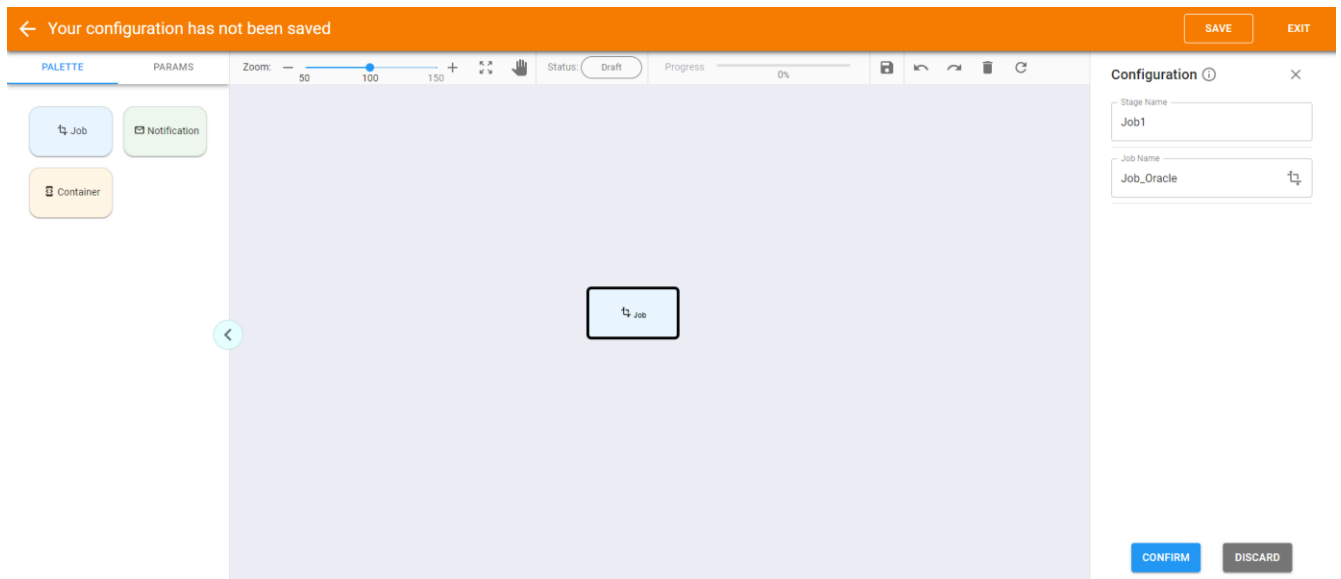



4) Pipeline is a combination of existing jobs stages and/or notification stages and container stages. Notification stage most often added to configuration in the case of job stage failure/success.

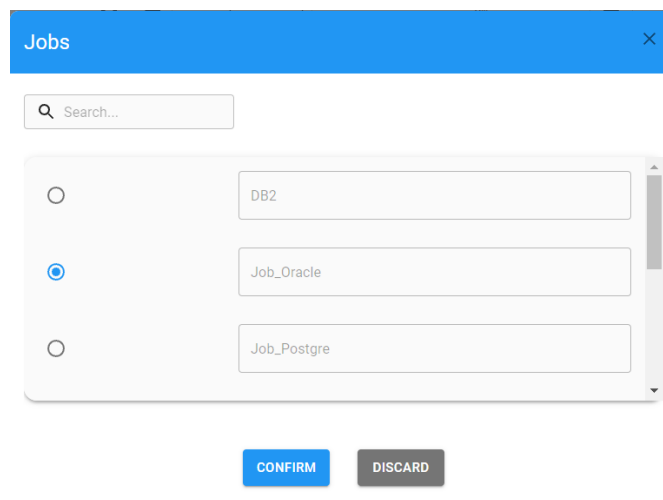
Start creating a pipeline by dragging *Job* stage to the canvas:



5) Double-click on the stage will open the configuration panel on the right:

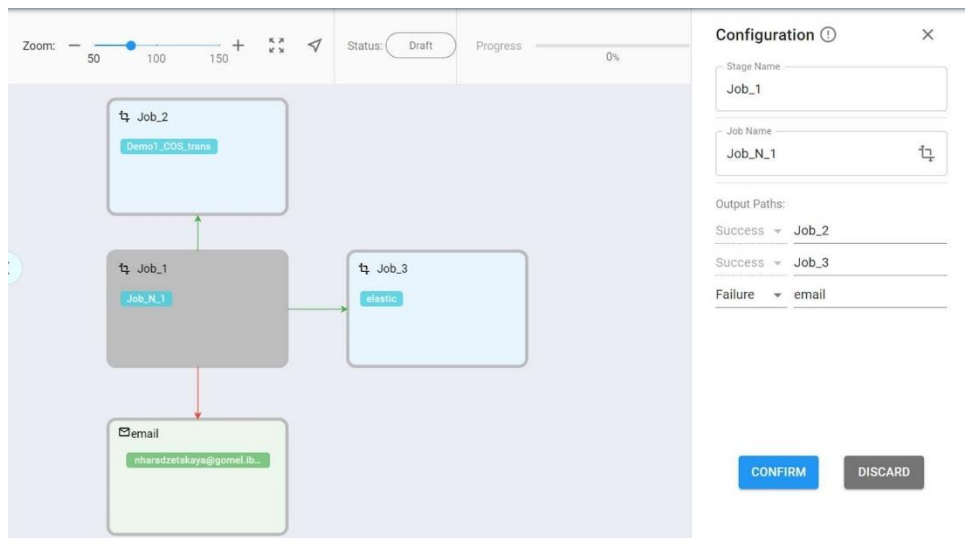


Enter a name for the stage and select a job from the list by pushing *Job* button. 



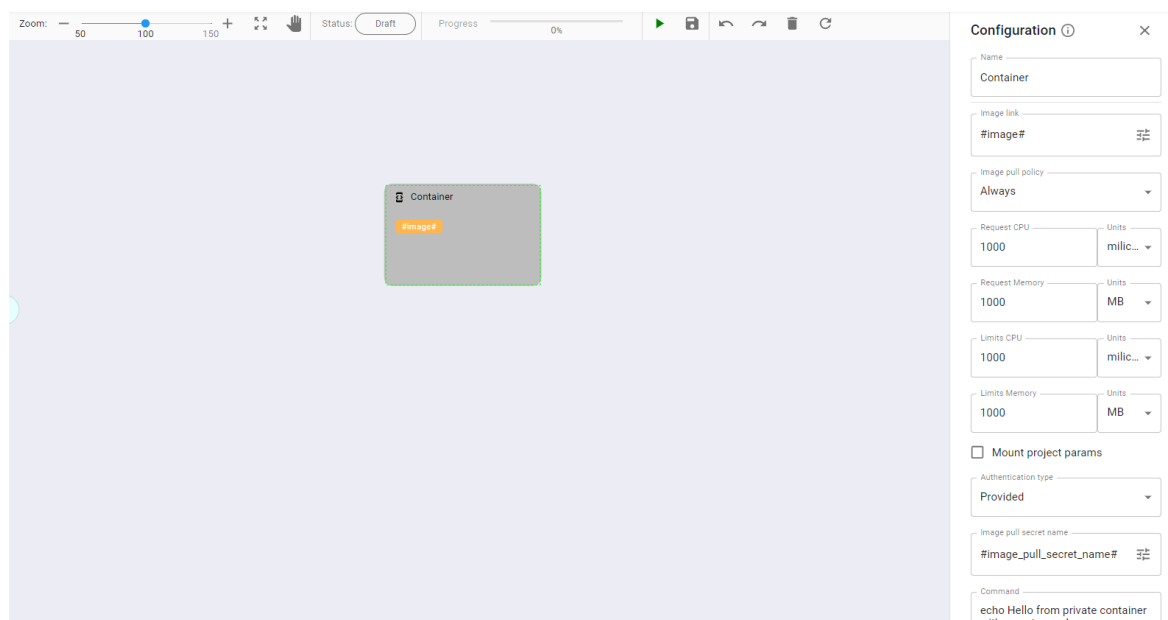
6) Save the stage by pushing *Confirm* button on the panel. If user wants to save his pipeline at this step, user should press *Save* button on the header.

7) Drag and configure other stages. Connect them with the same manner user did in Job Designer. User can link his stages based on the success or failure of each stage. After connecting stages between themselves, user can choose Success or Failure link on configuration panel. There can be only one connection for failure. See the example of configured pipeline:



A *Custom container* stage is required to run custom commands to execute any logic in the pipeline. Instead of custom commands, can use the created docker image.


1) Start creating a pipeline by dragging *Container* stage to the canvas and enter parameters in Configuration panel:

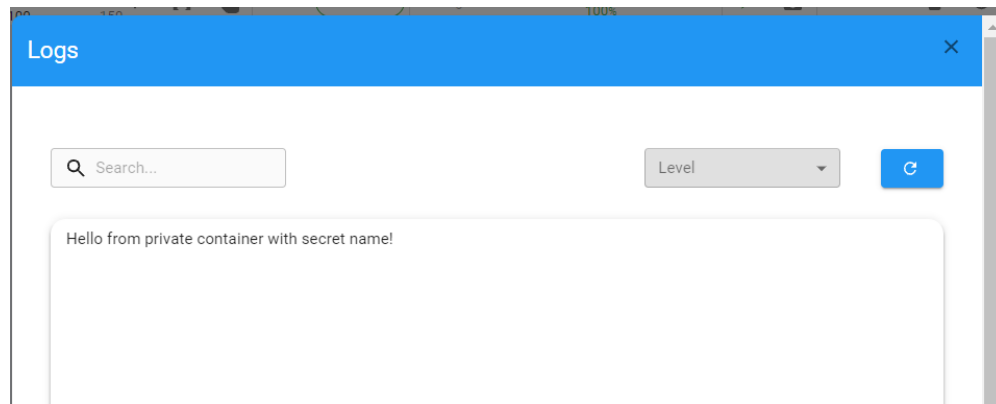



The Container stage has the following fields in the Configuration:

- ✓ Image link. Docker image path (Examples: mysql, mysql:latest, bitnami/argo-cd:2.1.2, localhost:5000/bitnami/argo-cd:2.1.2, registry.redhat.io/rhel7:latest.)
- ✓ Image pull policy. Defines when the image will be pulled(downloaded). Possible values:
  - *If not present* - download only if not exist locally;
  - *Always* - download before each start;
  - *Never* - do not download use only local copy.
- ✓ Requests and Limits CPU
- ✓ Requests and Limits memory
- ✓ Mount project params. Defines whether to mount all project params as environment variables inside the Pod.
- ✓ Authentication type
- ✓ Authentication mode that could be one of these:
  - *Not applicable* - image pull secrets are not needed, as the image is pulled from the public registry;
  - *New* - create a new image pull secret on the fly by providing all necessary information;
  - *Provided* - use existing image pull secret by providing it's name (Image pull secret name).
- ✓ Image pull secret name. Name of the secret to pull the image. Note that it must exist within the same k8s namespace as the current pipeline.
- ✓ Username
- ✓ Password
- ✓ Registry. Name of the registry for authentication.
- ✓ Command. Command that will be executed once Pod is be created.

*Important:*

Container stage has a *Logs* button . In Logs window, provided that the pipeline is successfully completed, the text of the command that was previously registered in the *Configuration* of Container stage will be displayed.














Before the first run or after updating, its status will be *Draft* . See each stage border painted in *Grey* color, which stands for *Draft*.



### 5.3. Pipeline Designer Functions Overview

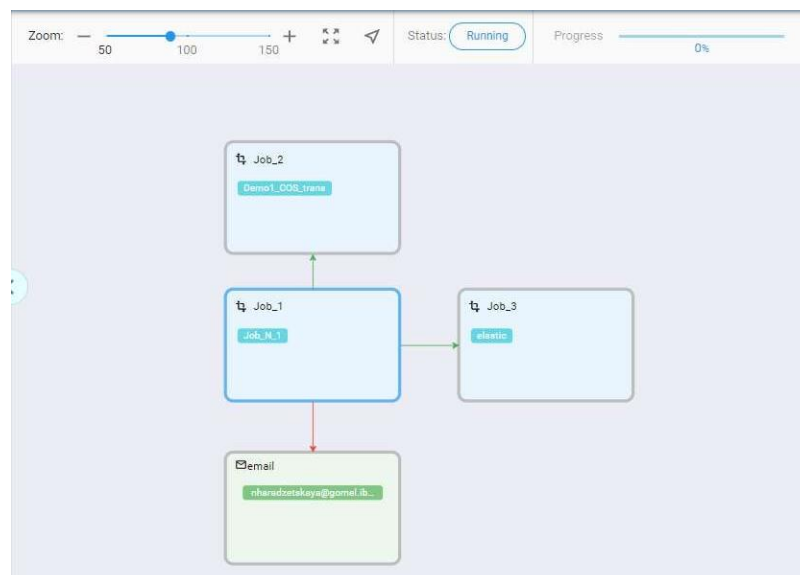
The following functions are available in *Pipeline Designer*:

- ✓ Zoom functions: 
- ✓ Move elements: 
- ✓ Move elements/screen: 
- ✓ Show pipeline status: 
- ✓ Show pipeline progress: 
- ✓ Run pipeline  / Stop pipeline  (for running)
- ✓ Save pipeline 
- ✓ Undo / Redo operation on canvas 
- ✓ Remove element from canvas 
- ✓ Refresh 

### 5.4. Pipeline Execution

If user runs a pipeline e.g. from the above example its status will change from *Draft* to *Pending* and then to *Running*. Push Refresh to update the status.

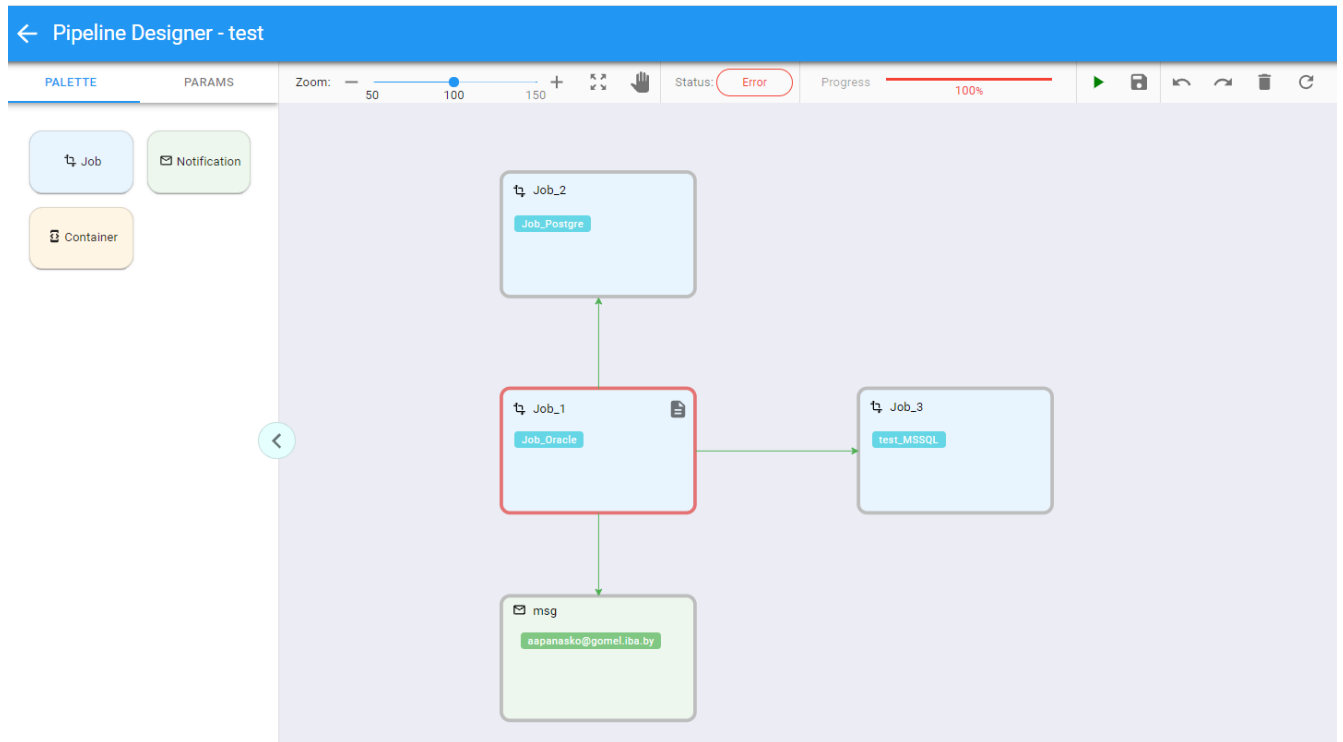
The border of the stage currently running will be painted in *Blue*:




If a pipeline succeeded, all completed stages will be painted in *Green* indicating success.


The ones configured for failure scenario (red arrow) of the previous stage will remain *Grey* as *Draft* as they have not been executed.

If a pipeline failed, then *Red* border will indicate the failed stage:



Failed pipeline can be re-run from the point of failure with button  located on the Pipelines Overview Screen.

*Important:*

*Job* stage has a *Logs* button  for analyzing logs of a certain job.