Workflow Environment Perspective

User Guide

world programming

Version: 4.0.3

www.worldprogramming.com

Contents

Introduction

The Workflow Environment is the collective name for a set of features to help with data mining, predictive modelling tasks, and a range of Machine Learning capabilities.

The Workflow Environment features operate in a completely different way to how WPS Workbench traditionally handles SAS language program development, execution, log and result output handling. For this reason, WPS is supplied with one perspective called the SAS Language Environment, dedicated to 'classic' SAS language programming and output handling, and a second perspective called the *Workflow Environment*, dedicated to the **Workflow Editor** view and **Data Profiler** view.

Note:

The current release of the Workflow Environment is a small subset of capabilities from the overall vision of the module. The subset of capabilities in this current release is aimed at users who want to prepare and explore their data and then perform modelling tasks using decision trees, regression, scorecard and clustering/segmentation analysis. Additional modelling and data handling facilities will be added to future WPS releases.

The **Workflow Editor** view provides a palette of drag-and-drop interactive blocks that you combine to create workflows that connect to a data source, filter and manipulate the data into a smaller subset using the **Data Preparation** blocks, and save to an external dataset for downstream use. You can then use the Machine Learning capabilities available in the **Model Training** blocks to discover predictive relationships in your data.

Once created, a workflow is re-usable, so can be re-used with different input datasets to generate similarly filtered or manipulated output datasets.

The **Data Profiler** view is a graphical tool that enables you to explore WPS datasets used in a workflow, or external to a workflow. You can use the **Data Profiler** view to interact and explore your data through graphical views and predictive insights.

Many of the Data Science features accessible through the **Workflow Editor** view are enabled by World Programming's built-in SAS language capabilities. Although coding is not a prerequisite for creating a workflow, those in the team who have the requisite programming skills can carry out more advanced tasks in a workflow, using not only the SAS language code block, but also R, Python and soon, SQL.

Who should use the Workflow Environment

If you are familiar with the Cross-Industry Standard Process for Data Mining (CRISP-DM), you can use workflows to follow this process in the preparation and modelling of data sources of all sizes. The Workflow Environment tools enable anyone with differing skill sets to work on the same data in a collaborative environment. For example, a single workflow could be created that enables:

- Data Analysts to blend the prepared data through joining, transformation, partitioning, and so on, to create raw datasets of varying sizes.

- Data Scientists to use machine learning algorithms to build, explore and validate reproducible predictive models, including scorecards, from proven datasets.
- Auto-generation of error-free code from the models, ready for deployment and execution in production.

Workflow Environment Perspective

The Workflow Environment Perspective contains the views and tools required to create a manage workflows.

Workflow files are created and managed in projects using the **Project Explorer** view, or on a host to which you have access using the **File Explorer** view.

The **Workflow Editor** view is used to create and run workflows; and the **Data Profiler** view to inspect the contents of a dataset in the workflow.

The **Workflow Link Explorer** view and **File Explorer** view enable you to connect to a remote host and specify a default Workflow Engine on a remote host to run your workflows.

The perspective also enables you to view tasks or bookmarks attached to a workflow. Other views and functionality available in WPS Workbench are visible in this release, but are not currently supported by the Workflow Environment Perspective.

The SAS language library feature for creating datasets, such as the temporary `WORK` library, is not available in a workflow. *Working datasets* are created as part of a workflow, and the **Data Profiler** view is used to explore the dataset in the workflow.

To open the Workflow Environment Perspective:

Click the **Window** menu, click **Perspective**, click **Open Perspective** and click **Other**. In the **Open Perspective** dialog box, select **Workflow Environment** and click **OK**.

Project Explorer view

The **Project Explorer** view is used to edit and manage projects and their associated workflows and datasets.

The **Project Explorer** view only displays projects that are in your current workspace. You can create several projects in a workspace, and use each project for a specific task.

The file that describes a workflow can only be created in a project, and that project is always stored on the workstation on which WPS Workbench is installed (the local host).

If you change the default Workflow Engine to a remote host, the location of the projects and workflow files does not change but remains on local host. Any references to datasets in the workspace or project cannot be used with a remote Workflow Engine, and must be moved to the remote host using the **File Explorer** view, and any references in the workflow changed to the appropriate filepath for the remote host.

Selecting an item in the **Project Explorer** view displays information about that item in the **Properties** view.

Displaying the view

To display the **Project Explorer** view, click the **Window** menu, click **Show View** and then click **Project Explorer**.

Create a new project

A project is a Workbench folder that contains workflows and related datasets. Projects are accessed through a workspace. You can only have one workspace open at a time, but a workspace can contain multiple projects.

To create a new project

1. Click the **File** menu, click **New** and click **Project**.
2. In the **New** dialog box, expand the **General** group, select **Project** and click **Next**.
3. In the **Project** pane, specify a **Project name** and click **Finish**.

File Explorer view

The **File Explorer** view is used to access the local file system, and the file systems of connected remote hosts.

The **File Explorer** view can be used to manage workflows and associated datasets on remote and local file systems. Because the **File Explorer** view does not manage content through projects, you cannot use the local history, not use Workbench functionality to import or export projects or archives.

The **File Explorer** view enables access to all files that are available on your local file system, and remote hosts with connections visible in the **Workflow Link Explorer** view

Displaying the view

To display the **File Explorer** view:

From the **Window** menu, click **Show View** and then click **File Explorer**.

Connections and shortcuts

It is only possible to view files on remote systems when you have enabled the required remote host connection. The remote host connection is created and enabled in the **Workflow Link Explorer** view

Working with files and folders

It is possible to perform the following operations on files and folders within the **File Explorer**:

- Create a new directory shortcut, directory or file.
- Selected files and directories to be moved, copied or deleted.
- Rename a selected file or directory.

File explorer preferences

To display hidden files and folders in the **File Explorer** view:

1. From the **Window** menu, click **Preferences**.
2. In the **Preferences** dialog box, expand the **WPS** group and select **File Explorer**.
3. In the **File Explorer** view, select **Show hidden files and folders** and click **OK**.

Workflow Editor view

The **Workflow Editor** view provides a visual programming canvas enabling operations such as data preparation and model training.

A workflow is defined by a single file, identified with a `.workflow` extension, and these files are managed in either the **Project Explorer** view or **File Explorer** view.

The **Workflow Editor** view canvas has a pallet of related operations, such as data import tasks, data preparation tasks, machine learning operations, coding blocks, scoring, and so on. Hover over the group in this palette to select the required operation, and drag the corresponding block onto the canvas where it is connected with other blocks already on the canvas to create a workflow. The workflow canvas has built-in intelligence to automatically indicate which blocks are suitable to connect together.

Nearly all blocks that are added to a workflow will have one or more properties that will need some sort of user input or action. Properties are set in the configuration dialog box that can be accessed through the shortcut menu for a block. To view or specify block properties, right-click the required block in the **Workflow Editor** view and click **Configure**.

By default, the workflow automatically runs as blocks are added or edited. You can also right click on a block and choose to run the workflow up to that position. Working datasets created as part of the workflow are deleted when the **Workflow Editor** view is closed. If you automatically run a workflow, reopening a saved workflow automatically recreates all working datasets in the workflow.

Workflow navigation

The **Workflow Editor** view enables you to zoom in and out of a workflow, to relocate blocks in a workflow and move the workflow within the editor view.

Zoom-in and zoom-out features in the **Workflow Editor** view can be controlled from the **Zoom** toolbar, keyboard, or mouse.

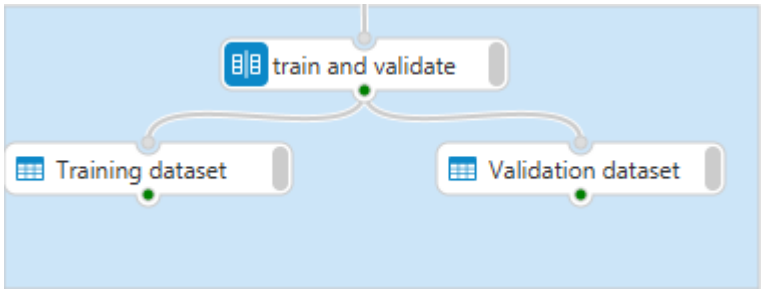
To change the scale of the workflow view:

- Click **Zoom in** to increase the scale of the workflow view by 10% of the current scale. Alternatively, press **Ctrl**+**Equals Sign**.
- Click **Zoom out** to decrease the scale of the workflow view by 10% of the current scale. Alternatively, press **Ctrl**+ **Minus Sign**.
- In the zoom list, click **Page** to scale the workflow view to fit the available screen area in the **Workflow Editor**.
- Press **Ctrl**+**0** (zero) to set the scale of the workflow view to 100%
- In the zoom list, click **Height** to scale the workflow view to fit the height in the available screen area in the **Workflow Editor**. If the workflow design is too wide for the scale, some blocks may be partly- or not visible after scaling.
- In the zoom list, click **Width** to scale the workflow view to fit the width in the available screen area in the **Workflow Editor** view. If the workflow design is too tall for the scale, some blocks may be partly- or not visible after scaling.

Relocate workflow blocks

The position of multiple blocks can be changed in the workflow view. If blocks are connected to other blocks in the workflow, the connections are maintained between the block. To relocate multiple items in the workflow:

1. Click the **Workflow Editor** view.
2. Press **Ctrl** and drag the pointer to cover all the blocks:



3. Click one of the highlighted blocks, and move the group of blocks to the required location in the workflow.

   After dragging, only the block you selected remains highlighted.

Move the visible workflow

You can move the visible area of the workflow to enable you to focus on a specific part of the workflow. To move the workflow within the editor view, press the Space Bar, click in the **Workflow Editor** view, and drag the pointer to the required location.

Save a workflow

After modifying, save the workflow to ensure no loss of data.

A workflow can be saved in a project in the active workspace, or to a location on a local or remote host. If you use the **Project Explorer** view to manage workflow files, the workflow is saved to a project folder in the current workspace. If you use the **File Explorer** view to manage workflow files, the workflow can be saved to any local or remote host location in which you have permission to create files.

To save a workflow, click the **File** menu and then click **Save**. The workflow file is updated with any changes you have made.

If you use the **Project Explorer** view to manage files, a version of the workflow is created every time you save. To view all saved versions, right-click the workflow in the **Project Explorer** view, click **Compare With** and click **Local history**.

The **History** view contains a list of all saved versions of the specified workflow. To view a version of a workflow in the **Workflow Editor** view, right-click the required version in the list and click **Open** in the shortcut menu.

Save a workflow to a different file

You can save a workflow to an alternative file either in the current workspace or to a folder on any host to which you have access. To save the workflow to a different file:

1. Select the view you use to manage files:
   - If you use the **Project Explorer** view, the **Save File** dialog box displays projects in the current workspace
2. Select the workflow in the **Workflow Editor** view.
3. Click the **File** menu and then click **Save as**.
4. In the **Save File** dialog box, select the required folder, enter and new **Filename** and click **Finish**.
   - If you use the **Project Explorer** view, the **Save File** dialog box displays projects in the current workspace.
   - If you use the **File Explorer** view to manage files, the **Save File** dialog box displays the available host connections and folder structure that you can access on each host.

Workflow Link Explorer view

The **Workflow Link Explorer** view enables you to create remote host connections, a new Workflow Engine instance and set options for the connection.

A Workflow Engine is an instance of the WPS processing engine running on a local or remote host. You can have multiple connections to remote hosts active in the **Workflow Link Explorer** view, and one Workflow Engine defined per connection. You are, however, only able to have one active Workflow Engine in the Workflow Environment Perspective.

Displaying the view

To display the **Workflow Link Explorer** view:

From the **Window** menu, click **Show View** and then click **Workflow Link Explorer**.

Objects displayed

There are two node types visible in this view:

1. A *Connection node* – represents a connection to a host machine.
2. A *Workflow Engine* – represents the WPS installation that will run the currently-active workflow.

Managing the connections and servers

From this view you can:

- Open or close a connection.
- Specify a default Workflow Engine.
- Define a new connections and Workflow Engine.
- Specify system options for a Workflow Engine.

Create a new remote host connection

A remote host connection enables you to access a remote host's file system using the **File Explorer** view and to link to the Workflow Engine installed on remote host to run workflows.

Before creating a new connection, you will need SSH access to the server machine that has a licensed WPS server installation. You may need to contact the administrator to obtain this information, and to ensure that you have access.

To create a new remote host connection:

1. In the **Workflow Link Explorer** view click **Create a new remote host connection**.

2. In the **New Server Connection** dialog, select **New SSH Connection (3.2 and later -- UNIX, MacOS, Windows)**
3. Click **Next** and complete the **New Remote Host Connection (3.2 and later)** dialog box:
   a. In **Hostname**, enter the name or IP address of the remote server machine.
   b. Unless modified by your system administrator, the default **Port** value (*22*) should be left unchanged.
   c. In **Connection name** enter a unique name to be displayed in **Workflow Link Explorer** view for this connection.
   d. In **User name**, enter your user ID for the remote host.

   Select the required option:

Option	Description
Enable compression	Controls whether or not data sent between the Workbench and the remote connection is compressed.
Verify hostname	Confirms whether the host you have specified in the **Hostname** entry exists.
Open the connection now	Whether the connection is automatically opened immediately
Open the connection automatically on Workbench startup	Controls whether the connection is automatically opened when the Workbench is started.

4. Click **Next** to define the connection directory shortcuts:
   1. Click **Add** to display the **Directory Shortcut** dialog box.
   2. In Directory Name enter the displayed shortcut name, and in **Directory Path** the full path of the target directory. Click **OK** to save the changes.

   Enter the displayed shortcut name in **Directory Name**, and the full path in **Directory Path** and click **OK**.
5. Click **Finish** to save the changes.

   If you have not previously validated the authenticity of the remote host, an **SSH2 Message** dialog box is displayed. If the identity of the host is correct, click **Yes**.
6. In the **Password Required** dialog enter your **password** for the remote machine and click **OK**.

Define a new Workflow Engine

Workflows are run using a Workflow Engine, you are able to define a new engine.

Before creating a new Workflow Engine, you need to know the installation directory path for the licenced WPS installation on the remote host.

To define a new server:

1. In the **Workflow Link Explorer** view, right-click the required remote host connection node and click **New Engine** in the shortcut menu.
2. In the **New Remote Engine** dialog box:
   a. In **Engine Name**, enter a name to display in the **Workflow Link Explorer** view.
   b. In **Base WPS install directory**, enter the absolute path to the WPS installation directory on the remote host.

Specify a default Workflow Engine

Specifying a new Workflow Engine to use when running workflows.

The pre-defined default Workflow Engine is the engine on the local host connection. Changing the default engine to a different host may require changes to the workflow to enable the workflow to run on a different Workflow Engine.

Specifying a default engine requires a WPS Workbench restart. You should save your workflow before specifying a different Workflow Engine.

1. Save and close any open workflows in the **Workflow Editor** view.
2. In the **Workflow Link Explorer** view, right-click the required Workflow Engine and click **Set as Default Engine** in the shortcut menu.
3. In the **Restart required** dialog box, click **Yes**. When WPS Workbench restarts, the specified default Workflow Engine is shown in bold in the **Workflow Link Explorer** view.

After changing the default Workflow Engine, you need to check the remote WPS engine is able to support the current workflow, for example:

- Ensure that file paths used to import datasets can be accessed from the remote host.
- If you use the **Database Import** block, ensure that the remote WPS engine has the required database client software installed.

Specify startup options for a Workflow Engine

Modifying the startup options enables you to specify the settings applied to a Workflow Engine.

The engine startup options (WPS system options) enable you to control items such as system resource usage or language settings. Not all system options can be set as startup options, and a list of system options that can be set at engine startup, their effect and supported values can be found in the *WPS Reference for Language Elements*.

All startup options are set in WPS Workbench in a similar manner. For example, to define the `ENCODING` startup option for a Workflow Engine:

1. In the **Workflow Link Explorer** view, right-click the required Workflow Engine and click **Properties**.
2. In the **Properties for Workflow Engine** dialog box, expand **Startup** and then click **System Options**.
3. On the **System Options** panel, click **Add** to display the **Startup Option** dialog box.
4. Enter `ENCODING` in the **Name** field, and enter `UTF-8` in the **Value** field.

   If you do not know the system option name, click **Select** and select the required option in the **Select Startup Option** dialog box.
5. Restart the Workflow Engine for the changes to take effect.

Data Profiler view

The **Data Profiler** view enables you to view content and details about a dataset. The view is opened automatically when you click **open** in the shortcut menu of an imported dataset or working dataset. The view contains the following tabs:

Summary View
: The Summary view tab lists dataset variables and their characteristics such as data type, length applied informat and applied formats. Information about the data content such as whether a variable contains missing values or the frequency distribution is also shown.

Data
: The Data tab lists all observations in the dataset. This is a read-only view and the data cannot be edited.

Univariate View
: The Univariate view tab provides univariate statistics for all categorical variables, including (missing values, min, max, mean, standard deviation, variance, kurtosis, skewness and range). Additional variables such as percentiles can be displayed using the **Calculate Statistics** wizard. To open the wizard, in the view's toolbar click **Configure Statistics**.

Univariate Charts
: The Univariate charts tab enables you to view the frequency of values in a dataset variable using histograms, line charts or pie charts.

Predictive Power
: The Predictive power tab enables you to segment a categorical variable and overlay frequency distribution for each of the categories on a single chart. It also ranks the predictive power of variables in relation to the segmented variable.

Bookmarks view and Tasks view

The **Bookmarks** view lists bookmarks added to a workflow; the **Tasks** view lists tasks added to a workflow. Both bookmarks and tasks can be used to identify parts of the workflow that require further investigation.

To add a bookmark:

1. In the **Project Explorer** view, select the required workflow.
2. Click the **Edit** menu and then click **Add Bookmark** and enter a **Description** in the **Bookmark Properties** dialog box.
3. Click **OK** and the new bookmark is displayed in the **Bookmarks** view.

To add a task:

1. In the **Project Explorer** view, select the required workflow.
2. Click the **Edit** menu and then click **Add Task** and enter a **Description** in the **Properties** dialog box.
3. Click **OK** and the new task is displayed in the **Tasks** view.

You can double click on a bookmark in the **Bookmarks** view, or a task in the **Tasks** view to open the workflow in the **Workflow Editor** view.

Workflow Environment preferences

The Workflow Environment preferences in the **WPS** section of the **Preferences** dialog box specify defaults and preferences when using the Workflow Environment Perspective.

Data Profiler panel

The **Data Profiler** panel is used to specify whether frequency summary charts are displayed and the number of variables displayed on an entropy chart.

Data panel

The **Data** panel is used to specify the number of unique values in a variable above which the variable is treated as continuous and the maximum number of cells in a dataset used to profile the data in the **Data Profiler** view or a decision tree.

Workflow panel

The **Workflow** panel is used to specify whether the workflow is automatically run when opened or run when you select either **Run to here** or **Force run to here** options in a block shortcut menu. Automatically running a workflow recreates all working datasets in the workflow.

The panel also contains options used to specify the location for temporary resources, whether the resources are deleted when the perspective is closed, and the location of the Workflow Engine used by a workflow. In addition, the following panels can be accessed from the **Workflow** panel:

Binning panel

Specifies the default number of bins created, and the maximum number of bins that can be created.

Decision Tree panel

Specify the growth algorithm and associated options that determine the depth of the tree, minimum node size, treatment of missing values and so on.

Create a new workflow

A workflow is a program sequence visually represented by connected, colour-coded graphical elements (blocks) through which the background code runs from start to end.

A workflow enables collaboration between the project stakeholders; it provides a project framework so the projects can be easily replicated or audited and can be used as a framework to carry out projects in standardised manner.

Any datasets you need to use with your workflow must be located on the same host as the Workflow Engine you will use to run the workflow. If your workflow requires a connection to a database, you must have the appropriate database client software installed on the same host as the Workflow Engine.

To create a new workflow

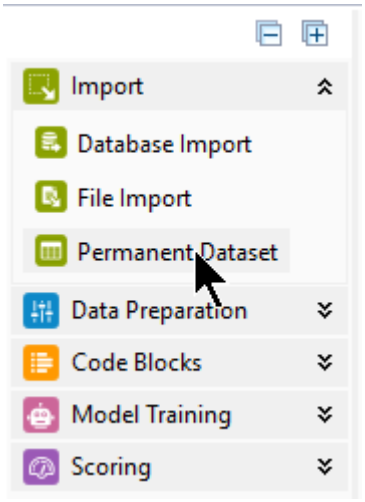
1. Select the required view for the new workflow.
    - If you select **Project Explorer** view, the folder structure of the current workspace is displayed in the **Workflow** dialog box. The new workflow can be created in any folder in the workspace, and the file is saved on your local workstation.
    - If you select **File Explorer** view, the **Workflow** dialog box displays the folder structure of the local workstation, any remote host, and any network drives available through either local or remote hosts. The new workflow can be created in any displayed folder in which you have permission to create files.
2. Click the **File** menu, click **New** and click **Workflow**.
3. In the **Workflow** pane, select the required folder for the new workflow, specify a **File name** and click **Finish**.

Add blocks to a workflow

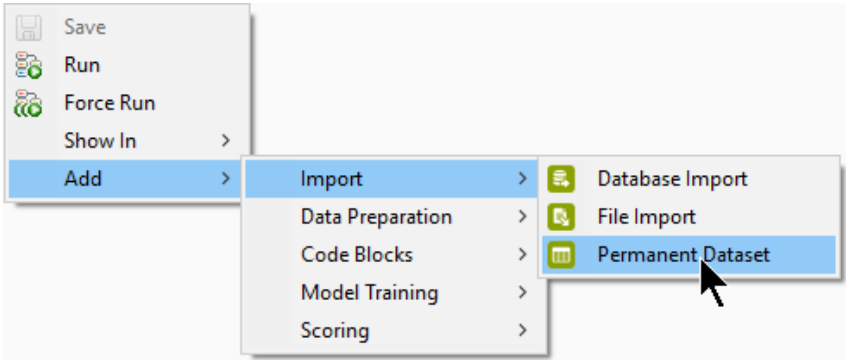
A workflow visually represents a program using block visuals, including colours, icons labels and connections between blocks to identify program steps and specific operations within each step. Workflows enable collaboration between the project stakeholders and provide a project framework so the projects can be easily replicated or audited.

Most workflows begin with a **Permanent Dataset** block. To add a **Permanent Dataset** block to a workflow:

1. In the **Workflow Editor** view, expand **Import** in the group pallet.
2. Click **Permanent Dataset** and drag the label into the working area of the editor.

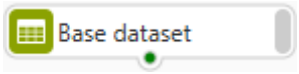


Alternatively, right-click the **Workflow Editor** view, in the shortcut menu click **Add**, click **Import**, and then click **Permanent Dataset**:



3. Right-click the **Permanent Dataset** block and click **Rename Block** in the shortcut menu.
4. In the **Rename** dialog box, enter a **Block label** to describe the dataset, for example *Base dataset* and click **OK**. The name entered is displayed as the label of the block on the **Workflow Editor** view.
5. Right-click the **Permanent Dataset** block and click **Configure** in the shortcut menu.
6. In the **Configure Permanent Dataset** dialog box, select a file location, either **Workspace** or **External**, and click **Browse** to locate an existing WPD-format dataset.
7. Click **OK** to save the changes.

If the dataset is successfully loaded, the Execution status in the **Output** port of the block is green:



At this point, you can add other blocks to the editor and connect them together to create a workflow.

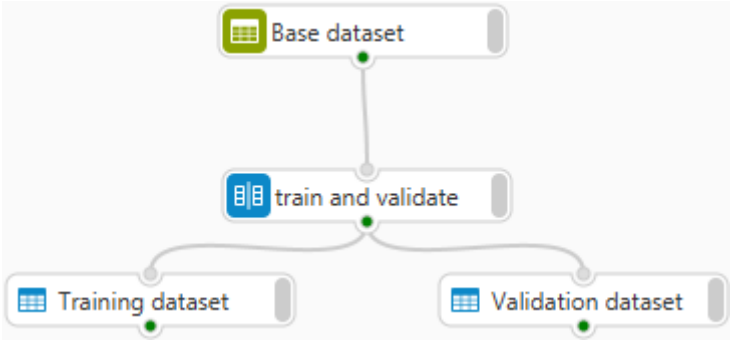
Connect blocks in a workflow

Blocks can be connected together using connectors to create a workflow. Connectors are linked to the block at either the **Input** port located at the top of a block, or an **Output** port located at the bottom of the block. You can connect the output from the **Output** port of one block to the **Input** ports of one or more other blocks.

You can randomly divide a dataset into two or more datasets using a **Partition** block. To partition the *Base dataset*:

1. In the **Workflow Editor** view, click the **Data Preparation** group in the group pallet.
2. Select the **Partition** block, and drag onto the working area. By default, two partitions are created and the **Partition** block has two working datasets:

3. Select the **Partition** block and click **Rename Block** in the shortcut menu. In the **Rename** dialog box, enter a **Block label** to describe the partition, for example *Train and validate* and click **OK**.
4. Click the **base dataset** block **Output** port and drag towards the **Input** port of the **train and validate** block.
5. When the **Base dataset** block is connected to the **train and validate** block:
   a. Select the **Partition1** dataset and change the **Block label** to *Training dataset*.
   b. Select the **Partition2** dataset and change the **Block label** to *Validation dataset*.



More blocks can be connected to the output datasets following the same approach to create a workflow that produces, for example, a credit risk scorecard.

Remove blocks from a workflow

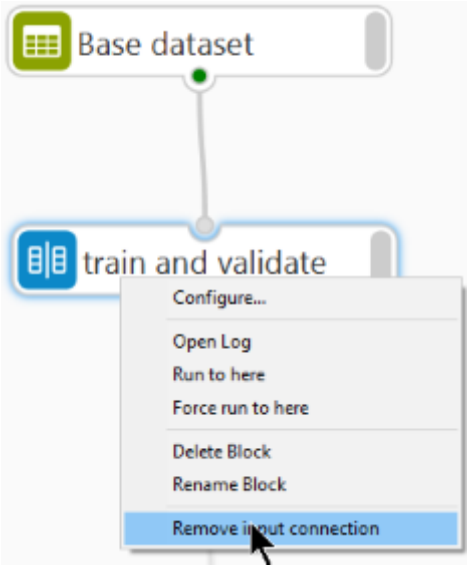
Blocks can be excluded from a workflow path by deleting the connectors leading to the **Input** port of the block.

Connections between blocks can be removed from the block at either end of the connection using the shortcut menu for the block. Where multiple connectors are linked to the **Input** port of a block, the shortcut menu displays all connectors identified by the block label.

Connectors leading to the **Input** port of a working dataset cannot be removed.

To exclude the **train and validate** block from the workflow:

1. In the **Workflow Editor** view, right-click the **train and validate** block.



2. Click **Remove input connection** in the shortcut menu.

Removing blocks enables to you alter the workflow, for example to attach a different dataset to the **train and validate** partition block to create new training and validation datasets for use in the rest of the workflow.

Delete workflow block

Blocks and any associated working datasets can be deleted from a workflow. To delete a block, select the required block, right-click and click **Delete Block** in the shortcut menu.

Deleting a block deletes any associated working datasets created by the block; any connectors leading to either the **Input** port; any connectors leading from the **Output** port of the block; or if the block automatically creates working datasets, any connectors leading from the **Output** port for every working dataset created by the block.

Working datasets cannot be deleted; to delete a working dataset you must either delete the block that created the dataset, or change the number of working datasets created by the block. For example, to remove a partition working dataset created using the **Partition** block:

1. Right-click the **Partition** block and in the shortcut menu click **Configure**.
2. In the **Configure** dialog box, select the partition to be removed and click **Remove Partition**.

   You cannot delete the default partitions created when you drag a **Partition** block onto the **Workflow Editor** view.

Workflow execution

A workflow may be automatically or manually run as specified in the Workflow preferences in the **Workflow** panel of the **Preferences** dialog box.

If the workflow is automatically run, additions to the workflow are evaluated as when you connect new blocks to a workflow. If the work flow is manually run, select either **Run to here** or **Force run to here** in the block shortcut menu to run the workflow.

- **Run to here** evaluates the additions to the workflow, but uses the output from any previously-run blocks in the workflow.
- **Force run to here** reruns the whole workflow, evaluating all blocks in the workflow and recreating all working datasets.

Workflow block reference

This section provides a guide to blocks currently supported by the **Workflow Editor** view.

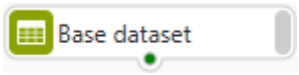
Blocks

Blocks are the individual items that make up a workflow.

Blocks are used to represent:

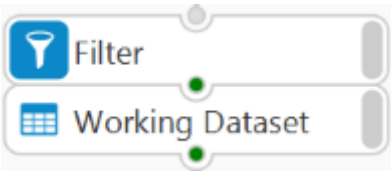
- Datasets, either imported into the workflow or created by the workflow.
- Programming code such as Python, R or the SAS language.
- Functionality that manipulates data to prepare a dataset for analysis.
- Modelling operations.

Blocks can be connected together to create a workflow using connectors. Blocks typically contain an **Output** port and an **Input** port. These ports enable you to link the output from one block to become the input to one or more other blocks. For example, **Permanent Dataset** blocks have an **Output** port located at the bottom of the block:



Block input and output ports

Blocks that manipulate a dataset have an **Input** port located at the top of a block, and an **Output** port located at the bottom of either the block, or the automatically-created output dataset of the block:

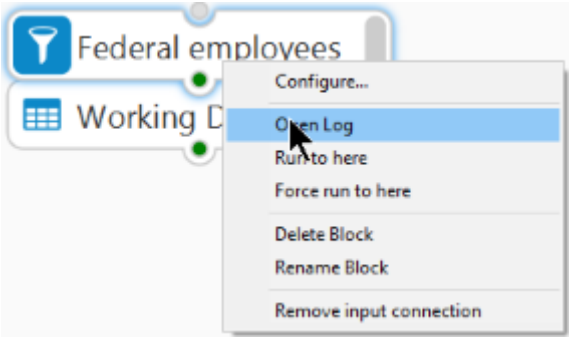


Block execution status

The Execution status in the **Output** port of the block indicates the state when the workflow has been run.

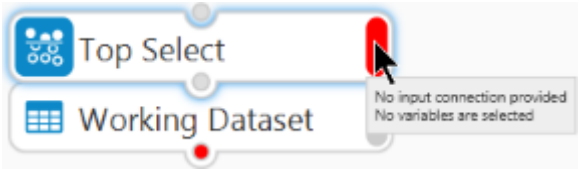
- If there are no errors, the Execution status is green.
- If the Execution status is grey, there is either an error in a connected upstream block, or the block does not yet have an input.

- If there are errors, the Execution status is red and you can review the error by reading the log output. Each contains its own log file; to view the log, right click the block and click **Open Log** on the shortcut menu:



Block configuration status

The Configuration status – the right-hand bar of a block – indicates whether all the required details have been specified for the block. If the Configuration status is grey, the block is complete and can be used in a workflow. If the Configuration status is red, some details are missing; hover over the bar to see what information is required to complete the block:



Block shortcut menu

Configuring, modifying and deleting blocks functionality is accessible from the block's shortcut menu. To access the block functionality, right-click the required block and choose the appropriate option:

- **Delete Block** enables you to remove the block from the workflow. All connectors leading to and from the deleted block are also removed.
- **Rename Block** enables you to enter a new display label for the block.
- **Remove output connection** enables you to delete a connector linked to the **Output** port of the block.
- **Remove input connection** enables you to delete a connector linked to the **Input** port of the block.
- **Configure** enables you to specify the details of the for example, dataset location for **Import** blocks

Where a block has multiple input or output connections, each connection is identified by the block to which it connects; the connector is removed when you click the name of the block in the list.

Some blocks have additions to the shortcut menu. For example, the **Permanent Dataset** block and working datasets have **Open** and **Open with** options that enable you to view the dataset content with either the **Dataset File viewer** or in the **Data Profiler** view. You specify which viewer opens when you click **Open** on the shortcut menu using the **File Associations** panel in the **Preferences** dialog box, and all viewers specified appear as options when you click **Open with** on the shortcut menu.

Where the block runs some SAS language code, for example the **Partition** block, the menu contains an **Open Log** option that enables you to view the log created by the Workflow Engine.

Block group palette

The palette provides tools to import and prepare datasets for analysis, and then use those datasets to create and train models for use in, for example, scoring applications. Blocks are arranged in the following groups:

- **Import** – Contains blocks that enable you to import data into your workflow.
- **Data Preparation** – Contains blocks that enable you to modify the datasets in your workflow.
- **Code Blocks** – Contains blocks that enable you to include an existing program, or add a new program to the workflow.
- **Model Training** – Contains blocks that enable you to discover predictive relationships in your data.
- **Scoring** – Contains blocks that enable you build a predictive model.

Import group

Contains blocks that enable you to import data into your workflow.

Database Import block

Connect to a database server to import views and tables into a workflow. You can use the **Database Import** block to import multiple tables into a workflow in a single step. You can also use a view to pre-process tables in the database, and import the pre-processed data as a dataset into a workflow.

The databases you can currently connect to are:

- DB2 – supports the selection of tables and views from a database or schema.
- Oracle – supports the selection of tables and views from a database or schema.
- MySQL – supports the selection of tables and views from a database.
- PostgreSQL – supports the selection of tables and views from a database or schema.
- SQL Server – supports the selection of tables and views from a database or schema.

To connect to a database, the database client connection software must be installed on the same host as the Workflow Engine you use. You must enter a valid **Username** and **Password** to connect to a database. The username is stored as part of the block information, but the password is not. You must re-enter the password when you re-open a closed workflow.

File Import block

Specify a `.csv` file to import into the workflow, which can be imported from a project in the current workspace, or from a location on the file system.

When imported, a delimiter is selected, which can be changed in the **Configure** dialog box. You can specify other options such as using whether the first row in the file is treated as headers and the formats of columns in the file when imported.

Permanent Dataset block

Specify a permanent dataset to import into the workflow. The dataset must be in WPD format, and can be imported from a project in the current workspace, or from a location on the file system. Dataset properties and contents are viewed in the **Data Profiler** view.

The **File Import** block and **Permanent Dataset** block only display datasets that can be accessed through the connection associated with the selected Workflow Engine.

When a connection to the local host is specified:

- Selecting the **Workspace** file location enables you to select a dataset saved in any of the projects in the current workspace.
- Selecting the **External** file location enables you to select a dataset stored on the file system of the local host, or any mapped drives.

The Workspace file location is not supported when a Workflow Engine on a remote host connection is specified.

Data Preparation group

Contains blocks that enable you to modify the datasets in your workflow.

Filter block

Enables you to select observations based on the value of one or more variables. The **Filter** block may be used for example, to reduce the number of observations in a large dataset but keep all the variables in the dataset, or to find all customers in a dataset who are over 65.

Join block

Enables you to combine variables from two datasets into a single dataset. Observations can be joined using one or more variables, for example if all datasets have an ID variable, the variables for the same value of ID in each dataset can be joined together in the same observation. Where variables have the same name in both datasets, the variables in the Right dataset are dropped.

Mutate block

Enables you to create new variables based on the existing data in a dataset. The new variable name, type and format are set and the content based on an SQL expression. You can for example use the expression to duplicate the content of a variable into the new variable, or apply a function to transform the existing variable into the new variable

Partition block

Enables you to randomly divide the observations in a dataset into two or more datasets, for example to create a *training* and *validation* dataset for model training. A **Partition** block creates a miniumum of two working datasets, and all working datasets contain the same variables as the input dataset. By default, two datasets are created each containing 50% of the input dataset observations. You can add or remove output datasets in the **Configure** dialog box, alter the weighting to reflect the relative importance of the output datasets created, and introduce a seed to replicate the split of datasets.

Select block

Enables you to create a new dataset containing specific variables from an input dataset. The variables to be kept are selected in the **Configure** dialog box, and all observations are kept for the selected variables in the output dataset.

Top Select block

Enables you to create a new dataset from an input dataset containing a dependent variable and select the most influential independent variables of that dependent variable. Independent variables are selected in the **Configure** dialog box, and are displayed in the panel by entropy variance order.

Impute block

Enables you to assign a value to a fill in missing values in a variable. Values are determined by the content of the other values in the dataset variable, and the selected algorithm. The available algorithms depend on the variable type, and can be selected in the **Configure** dialog box of the block.

Code Blocks group

Contains blocks that enable you to include an existing program, or add a new program to the workflow.

Although programming is not a prerequisite for creating a workflow, if you have the requisite programming skills you can carry out more advanced tasks in a workflow, using the SAS language, R, and Python blocks.

SAS Language block

Enables the use of SAS language programs in the workflow. The **SAS Language** block provides a self-contained environment that supports all of the SAS language syntax available in the WPS engine, including creating working datasets.

Input and output datasets must be referred to in your SAS language program using macro syntax rather than the names displayed in the **Manage Inputs** dialog box or **Manage Outputs** dialog box. For example, to use an input dataset named `Input_1`, you must refer to the dataset as `&Input_1`.

Python Code block

Enables the use of Python language programs in the workflow. To use this block, you must have a Python interpreter installed and correctly configured on the machine running the Workflow Engine.

Input and output datasets are referred to in your Python language program by the case-sensitive names displayed in the **Manage Inputs** dialog box or **Manage Outputs** dialog box.

R code block

Enables the use of R language programs in the workflow. To use this block, you must have an R interpreter installed and correctly configured on the machine running the Workflow Engine.

Input and output datasets are referred to in your R language program by the case-sensitive names displayed in the **Manage Inputs** dialog box or **Manage Outputs** dialog box.

Model Training group

Contains blocks that enable you to discover predictive relationships in your data.

Decision Tree block

Provides an interface to create a decision tree for classification purposes. Trees can be grown using one of the algorithms specified in the **Decision Tree** panel of the **Preferences** dialog box.

A decision tree is edited through the Decision Tree editor. To open the editor, right-click the **Decision Tree** block and click **Edit Decision Tree**.

The Decision Tree editor displays the tree nodes, node properties and information. You can use the editor to grow the tree manually based on the entropy variance of the selected independent variables, automatically to create a full tree, or automatically one level of the tree at a time.

For information about the available settings for decision tree, see the `DECISIONTREE` procedure in the *machine learning* section of the *WPS Reference for Language Elements*.

Logistic Regression block

Fits a logistic model between a set of independent variables and a categorical dependent variable. Automatic variable selection of the independent variables can be made using one of the supported model-selection algorithms.

The **Logistic Regression** block generates a report and deployment code. The report contains details of the Model fit statistics, Maximum likelihood estimate analysis, odds ratio estimates, and predicted probability and observed responses. The generated deployment code is a `DATA` step that can be copied into a SAS language program for testing or production use.

Scorecard Model block

The scorecard model generates a standard scorecard, based on the output from both a **Logistic Regression** block and a **WoE Transform** block. The **Scorecard Model** block performs automatic scaling of the model parameters into equivalent points based on base points, base odds and points to double the odds. The higher scores can be assigned to either good or bad category, and deployment code is generated as a `DATA` step that can be copied into a SAS language program for testing or production use.

K Means Clustering block

Finds cluster centres from unlabelled data using the K-means clustering technique. This node works on numeric variable inputs only. The node allows both the pre-standardisation of the input variables and the imputation of missing values, based on average values.

WoE Transform block

Enables you to measure the influence of an independent variable on a dependent variable. You can use the **WoE Transform** block transformation to measure risk, for example to test hypothesis such as the risk of loan default based on area of residence if your dataset is grouped by geographic areas.

A WoE Transform is edited through the WoE Transform editor. To open the editor, right-click the **WoE Transform** block and click **Edit WoE Transform**.

The **WoE Transform** block editor can be used to select dependent variable and the independent variables, the data can be binned and the output transformation code be deployed in either SQL or SAS language.

Scoring group

Contains blocks that enable you build a predictive model.

Analyse Model block

Enables you to test the same scored model against multiple datasets to ensure the model has not been over-fitted to the training data. To test the same model with multiple datasets, create one **Analyse Model** block for each dataset, and connect the same model output to each block.

The **Analyse Model** block generates a report showing the Gains chart, K-S Chart, ROC chart and Lift chart for the model. To display the report, select the required block in the **Workflow Editor** view and in the **Configure** dialog box, click **Open Report**.

Score block

Enables you to tests multiple models against the same dataset, or to test the validity of a model on holdout (test) samples. To test multiple models, create one **Score** block for each model you want to test, and connect the same dataset to each **Score** block. The output from the multiple blocks will enable you to identify the preferred model to deploy.

Legal Notices

Trademarks

WPS and World Programming are registered trademarks or trademarks of World Programming Limited in the European Union and other countries. (r) or ® indicates a Community trademark.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

All other trademarks are the property of their respective owner.

General Notices

World Programming Limited is not associated in any way with the SAS Institute.

WPS is not the SAS System.

The phrases "SAS", "SAS language", and "language of SAS" used in this document are used to refer to the computer programming language often referred to in any of these ways.

The phrases "program", "SAS program", and "SAS language program" used in this document are used to refer to programs written in the SAS language. These may also be referred to as "scripts", "SAS scripts", or "SAS language scripts".

The phrases "IML", "IML language", "IML syntax", "Interactive Matrix Language", and "language of IML" used in this document are used to refer to the computer programming language often referred to in any of these ways.

WPS includes software developed by third parties. More information can be found in the THANKS or acknowledgments.txt file included in the WPS installation.