# Chapter 3: Data acquisition

## Introduction

Contents for this lesson include an introduction and focuses on the Altair Analytics Workbench Import group and blocks available to read from various data sources.

Each block is introduced with settings highlighted and special attention is focused on importing from databases, creating connections and database references. This is followed by a demonstration and then a summary.

## Fundamental

Acquiring data is a foundational and underlying aspect to conducting any analytics, whether for etl, reporting, profiling, modelling or simple data investigation.

Few analytics endeavours rely on one data source and predominately a variety of sources are required and a tool enabling connections to all of these is a must. Altair Analytics Workbench provides functionality to connect to data sources locally, on servers, mainframes and in the cloud whether that be flat files such as *.csv*, *.txt*, *.dat*, *.tab* and others. Proprietary file formats such as Microsoft Excel or databases, of which there are many.

A comprehensive list of the Altair Analytics Workbench data engine modules outlining connectivity capabilities can

be found by navigating to: [www.worldprogramming.com/information/sas-language/modules/data\_engine](http://www.worldprogramming.com/information/sas-language/modules/data_engine)

## The Import group

Altair Analytics Workbench Workflow perspective functionality to connect to and import data from a variety of sources is available via blocks found in the import group.

Figure 1: Import group blocks

A screenshot of a group block

Description automatically generated

Five blocks are available: Database Import, Excel import, Permanent dataset, Text file import and ALTAIR Dataset Import.

Databases references can be configured using the Database Import block or by right-clicking the Workbench node from the Database Explorer view and selecting New Database.

The Excel Import block can be used to read data from Microsoft Excel files only. Flat files including *.csv*, *.txt*, *.dat* and *.tab* can be imported using the Text File Import block.

The Permanent Dataset and ALTAIR Dataset Import blocks are used to read ALTAIR proprietary file format: *.wpd* only. The Permanent Dataset block is a legacy block and the ALTAIR Dataset Import block is a later addition providing more import options.

Note that help is available from any dialog by clicking the help icon. This will be illustrated when demonstrating.

### Excel Import block

The Excel Import block, as previously mentioned, is used to import data from Microsoft Excel files. The block configuration dialog has four pages: File, Column Selection, Column Properties and Errors.

Figure 2: Excel Import block

A screenshot of a computer

Description automatically generated

From the File page, the file to import is specified and two options are available: Workspace or External with the ability to browse to and select the file to import. If Workspace is selected, only local Engine datasets are listed when Browse is selected. Choosing the option External, enables the ability to import an Excel file from any location regardless of whether using a Local or Remote Engine.

The Data Format area provides options to specify the locale, select sheets or named ranges and use first rows as headers. If locale specifications are evident in the file being imported, this field is populated automatically. If there are locale specific variables in the file and no locale specifications are evident, a guess is made and this is picked from a list but the locale field will not be populated. Bear in mind that the locale can be selected manually also.

The Column Selection page allows filtering of variables and those to be imported are simply moved from the Unselected Columns list to the Selected Columns list.

The Column Properties page provides the facility to set and modify variable properties. The original variable name is listed with the ability to change this and the variable label as well as modify other properties via drop-down.

The Errors page contains one option: Remove rows with errors. This will exclude any rows deemed to be erroneous.

The Preview displays the data as per options specified and is evident across all pages.

### Text file Import block

The Text File Import block is used to import various flat file formats. The dialog layout is similar to that of the Excel Import block and the pages: Column Selection, Column Properties and Errors are identical in that they provide the ability to select and specify variables and their properties as well as remove rows with errors with a Preview available across all pages.

Figure 3: Text File Import block

A screenshot of a computer

Description automatically generated

The file page also has similarities in that the file to import can be either Workspace or External. The Text File Import block also includes an additional option: URL. This enables the import of files located on any accessible web site. Only files specified using the HTTP or HTTPS protocols can be imported.

This section also provides the ability to set the file encoding which is detected automatically but can be changed via drop-down.

The Data Format section provides the ability to specify the file delimiter, locale and read the first row of data as variable names. The Delimiter drop-down provides common options and also the ability to select Other and specify a delimiter if not available from the drop-down. The locale option here functions in the same way as the locale option on the Excel Import block and bear in mind this can be set manually.

### Permanent Dataset / ALTAIR Dataset Import block

The Permanent Dataset and ALTAIR Dataset Import blocks are identical in that they are used to import ALTAIR proprietary file format: WPD.

Figure 4: Permanent Dataset / ALTAIR Dataset Import blocks

A screenshot of a computer

Description automatically generated

The Permanent Dataset block is a legacy block and has been superseded by the ALTAIR Dataset Import block which provides more options.

Both blocks provide the facility to import from either a Workspace or an External location. In addition to this the ALTAIR Dataset Import block provides the facility to specify the Import Policy as either import once or re-import if input dataset is out of date.

This option relates to a file that has already been imported to a Workflow, if the data in the file has changed setting this option ensures that the file is re-imported automatically when the Workflow is opened.

### Database Import block

The database import block enables the use of already existing database references via a dropdown but also enables the creation of new database references by clicking the icon of the same name.

Once clicked, the Add Database wizard opens. The database type is selected from the dropdown and then it is necessary to proceed through additional pages to configure the connection with particulars such as the data source name and authentication details.

Figure 5: Database import

A screenshot of a computer

Description automatically generated

Once the connection has been created, all tables and views are accessible by clicking the appropriate tab and one or more tables and/or views can then be selected.

Rather than use the Database Import block to create a database reference, the Database Explorer view can be used to define a new database reference and also import and export references.

Figure 6: Database Explorer

A screenshot of a computer

Description automatically generated

The host nodes define the location of the database reference and there are two: Altair SmartWorks Hub and Workbench. Altair SmartWorks Analytics Hub is a cloud-based solution enabling governance of users and data sources and deployment of programs as APIs.

Users can log into the Hub from the Workbench. Once logged in all references defined and stored there are pushed to the Workbench and appear in an expandable list below the Hub node.

Options to create, import or export references are available using the icons or by right-clicking the Workbench node. Clicking the option New Database opens the same wizard that appears when creating a database reference using the Database Import block.

Additionally, right clicking any database node in the Database Explorer view provides access to options to edit, delete, export the definition, rename, and test the connection.

### Database references

Databases defined using the Database Import block or from the Database Explorer view are references.

Figure 7: Database references

A screenshot of a computer

Description automatically generated

A database reference is a connection to a database listing all tables and views contained therein. Any table or view dragged into the Workflow canvas exist as pointers. A pointer can be thought of as a signpost to the data, it points to it and provides information about it but the data is not imported.

Data is only imported when the tables and views are connected to an executable block and run. As databases tables can be large and when accessing information from them, there is generally a need to get information from more than one table.

Importing all tables individually and then applying, for example a join operation can be quite inefficient, having pointers means that tables can be used in a Workflow, a join operation set-up, and only at the point of the join being run, is the data imported.

Database references and pointers ensure efficiency and time saving when accessing data contained in databases.

One question then arises is what is the difference between creating a database reference using the Database import block in comparison to the using the Database Explorer view?

Figure 8: Database references

A screenshot of a computer

Description automatically generated

Database references defined using the Database Import block are saved with the Workflow they are defined from and are only accessible to it, whereas database references defined from the Database Explorer view are saved with the Workspace and are available and accessible from any Workflow.

### Settings tab

The settings tab lists all database references and options are available to add, edit, rename, remove, and test any connection and all databases defined using the Database import block are automatically added to this view.

Database references defined using the Database Explorer appear when a view or a table is used in a Workflow. Note that all connections evident here appear in the Database Import block database connections drop-down.

Figure 9: Settings tab

A screenshot of a computer

Description automatically generated

## Demonstration

So onto a demonstration. This demonstration uses the Workflow created in the previous lesson with the Workflow canvas cleared.

### Excel Import

Expanding the import group reveals the six import blocks. Dragging an Excel Import to the canvas outputs an empty dataset and the block configuration status is red, hovering displays a message: Import not yet configured.

Either double-clicking provides access to its configuration dialog. With the Excel Import block, data can be import from files in a current Workspace when using Local engine. As this is the case here, this option is selected and Browse clicked.

Note that the dialog is blank and the message: No files found is evident. To illustrate this option more clearly, a new empty project called temp\_ is created in the Project Explorer and a Microsoft Excel file copied to it.

Returning to the **Excel Import block configuration** dialog and **choosing the same option** it can be seen that the new project is listed and when **expanded** the Microsoft Excel file copied previously is available. This option works in the same way for the Permanent Dataset, ALTAIR Dataset and Text File import blocks in that If Workspace is chosen, only projects containing files of that format are listed, and only if Local Engine is in use.

The **created project is deleted** and the Excel Import block configuration dialog accessed. **Selecting External** allows browsing of the file system the engine is installed on. The current engine is local, meaning a desktop/laptop installation and therefore its files system is visible and navigable.

Here, the data folder in the ALTAIR eLearning **directory is expanded** and all Microsoft Excel files are listed. The **file *data\_.xlsx* is selected** and **Finish clicked.**

From the data format area sheets is selected by default and the **dropdown lists the first sheet**: data\_. Clicking the dropdown it can be seen that this is the only sheet available, but if there were others they would be listed here. Note if there is interest in importing data from multiple sheets, multiple import blocks must be used.

The Locale is left blank and adopted from the operating system. As a result of Use first row as headers being selected by default, variable names are visible in the Preview area.

From the **Column Selection page** it can be seen that all variables are by defaulted selected and evident in the **Selected Columns list**. If there is interest in not importing variables they should be moved from the Selected Columns list to the **Unselected Columns** list.

The **Column Properties page** provides the capability to specify variable properties. Note the first column contains the original variable names. To change either the name or the Label is a matter of **clicking the appropriate column** for any variable and typing a new name or label.

Other variable properties can be modified by selecting an appropriate option from a dropdown, for example, to change the variable type, the Type column drop-down is clicked and the type to apply to the variable can be easily selected.

Note also that on this page is a %Errors column. This relays the number of observations with errors for that particular variable. As can be seen, all variables have a value of 0, meaning there are no observations with errors.

From the Errors page, there is only one option: Remove rows with errors. This is not selected here as there are no observations with errors.

Clicking the Help icon provides additional detail for all available options and also lists related content, clicking the Help icon again, closes the help page and clicking OK, runs the process and imports the data.

Notice that the execution status of the block is green and hovering displays the message Completed successfully and up to date.

The data can be viewed by double-clicking, this opens the Dataset File Viewer with data visible. Returning to the Workflow, right-clicking the imported dataset and selecting the option Open With provides two options: Dataset File Viewer and Data Profiler. Double-clicking the block has the same effect as selecting the option Dataset File Viewer in that both open the data with the Dataset File Viewer.

Right-clicking the Excel Import block and selecting the option: Configure Outputs opens the Excel Import block Outputs dialog.

Note that the Selected Output columns lists the created dataset and the Unselected Output list contains an entry named Errors. Moving this to the Selected Output list and clicking OK, an additional dataset is evident on the canvas named Errors.

Opening this dataset reveals that it contains variable names but is empty, this is not surprising because as was evident from the Column Properties page, there were no observations with errors as denoted by the 0 for all variables in the %Errors column.

If there were, these observations would be evident in the Errors dataset.

### Text File Import

To import a text file, a Text File Import block is used - again dragging onto the canvas and accessing configuration settings reveals a lot of similarity with the Excel Import block. Here, again, the option External is selected and the data folder in the ALTAIR eLearning directory navigated to.

Considerably more files are evident, including Microsoft Excel files. However, the Text File Import block cannot be used to read this format. Here, the file: *data\_.csv* is selected. The delimiter is automatically detected and set as Comma.

Other delimiters can be selected and if the required delimiter is not listed, selecting Other provides the facility to specify any. Note that it may sometimes be necessary to open the file in a text reader to gain this information.

The Locale is left blank and the option: Use first row as headers is automatically selected and as a result variable names are evident in the Preview along with the data.

Options available from the Column Selection, Column Properties and Errors pages are identical to those available from the Excel Import block and are not modified here.

Clicking Help, again provides access to help for the Text File Import block and related content and clicking OK imports the data.

Right clicking the block and choosing Open With again provides options to open the data with the Dataset File Viewer or the Data Profiler, here the Dataset File Viewer is chosen and the data is revealed.

### Permanent Dataset / ALTAIR Dataset Import

The Dataset Import block is used to import ALTAIR’ proprietary file format: *.wpd*.

Accessing configuration settings reveals one page with options to read from a Workspace or an External location.

Here External is chosen, the same folder is accessed via the browse button and the file *data\_.wpd* selected. This has the effect of moving all variables into the Selected Variables list. Clicking the Output Name for any variable provides the facility to change it if desired.

The order of variables can also be modified by selecting one or more variables and using the options to the right-hand side to move as required.

The Import Policy enables an already imported *.wpd* dataset to be re-read if the data has changed. This happens automatically when the Workflow is opened if this option is selected.

Once complete, OK is clicked and the file imported and double-clicking opens with the Dataset File Viewer.

### Database Import block

The Database Import block enables database references to be created in a Workflow. Dragging the block onto the canvas and accessing its configuration dialog automatically opens the add database wizard.

Closing this momentarily reveals why: currently no databases are available for selection and must be added. Clicking the Create a new database icon opens the Add Database wizard as before. From the drop-down the database type must be selected, here, ODBC is chosen and clicking next navigates to the ODBC Settings page.

Figure 10:ODBC Settings

A screenshot of a computer

Description automatically generated

The path through the wizard is slightly different depending on the database type selected. Here the database to connect to must be supplied via Data source name and the method of authentication selected.

Depending on the authentication selected, a username and password may be necessary. Here, credentials are selected and a username and password supplied. The connection must be tested and successful prior to being able to proceed.

The next page allows the selection of the schema, if this is not known, it can be left blank and the database default will be used. Clicking Next> displays options to provide a name to reference the database.

Here databaseA is entered and once Finish is clicked all database tables are listed by default, selecting the Views tab lists all available views.

One or more tables and/or views can be selected and moved from the unselected list to the selected list by either using the chevrons or double-clicking.

Figure 11: Database tables

A screenshot of a computer

Description automatically generated

Clicking OK results in the database and table references being evident on the Workflow canvas. Note that the variable and observation count are visible as a result of the option Show/hide observation count being selected. Notice there is a question mark for the number of observations.

Figure 12: Database tables in Workflow

A screenshot of a computer

Description automatically generated

One thing to bear in mind here is that these are currently references or pointers to the database tables. The data is not imported to the Workflow until the table references are connected to and run through an executable block.

This is further illustrated when viewing results with the Data Profiler. From the Summary View tab it can be seen that the number of variables is evident but the Number of observations is as yet unknown.

A Select block is added from the Data Preparation group and connected to the Customers table reference. All variables are retained and the resulting dataset is again opened with the Data Profiler.

As can be seen both the number of variables and observations are now evident. As a result of running the table through an executable block, the table has been imported and its data read.

## Settings tab

Clicking the settings tab provides access to all database references used in the Workflow and notice the recently added database appears here.

Figure 13: Settings tab

A screenshot of a computer

Description automatically generated

Selecting any database provides options to edit, rename, remove and test the connection. Databases can also be added via Add Database.

This opens the Add Database wizard as before. Note the help option, again, clicking this provides access to help. This includes links to additional Help pages to assist when adding a specific database type.

The Database Explorer view provides an additional mechanism to add databases by either using the Create a new database icon or by right-clicking the Workbench node and selecting the option New Database.

This opens the same Add database wizard to that accessed from the Database Import block. Here, the the same database is configured and named databaseB.

Figure 14: New Database...

A screenshot of a computer

Description automatically generated

The Workbench node is now expandable and lists all added databases - here there is only one: DatabaseB. Expanding it reveals all tables and views in one list.

Figure 15: Database in Workbench

A screenshot of a computer

Description automatically generated

Notice that the settings tab does not list this database currently, this is because only databases used in a workflow are listed here.

All the tables and views in databaseB are, again, pointers. These can be dragged onto the Workflow

Canvas. As a result of this databaseB now appears in settings.

Right clicking any added database node from the Database Explorer view provides options to edit, delete, rename, export definition, and test the connection.

Right clicking the host node, provides options to create a new database, import definitions and export all definitions. The Export All Definitions option saves all defined databases to a *.dsx* file. This can be shared for easy configuration. Here the option Export All Definitions is selected, the name All\_ applied and the *.dsx* file saved to the data folder used previously.

Right clicking the Workbench host node and choosing the option Import Definitions and navigating to and selecting the file previously exported automatically creates new databases. Notice that the previously exported definition is replicated with an additional character.

Bear in mind this duplicates the already existing database reference databaseB but serves to illustrate the export and import of database definitions.

One final point to note is that there is a difference between defining a database reference using a Database Import block in comparison to using the Database Explorer view to do the same.

Database references defined from the Database Explorer ensures that the database reference is available to any Workflow – this is illustrated by creating a new Workflow, notice that the databases defined and imported using the Database Explorer are available to it.

The database defined using the Database Import block in the previous Workflow – databaseA is not, this is because database references defined using the Database Import block are saved with the Workflow they are created from and are only available to it.

## Summary

This lesson focused on the Altair Analytics Workbench Import group and blocks available to read from various data sources.

Each block was introduced and settings highlighted, additionally the creation and management of database connections was focused on prior to a demonstration of capabilities.