Data Mining

SPSS Clementine 12.0

2. Building Streams

Fall 2009

Instructor: Dr. Masoud Yaghini

Outline

- Stream-Building Overview
- Working with Nodes
- Working with Streams
- Executing Streams
- Saving Files
- Loading Files
- References

Stream-Building Overview

Stream-Building Overview

- The process of running data through a series of nodes, referred to as a stream.
- This series of nodes represents operations to be performed on the data, while links between the nodes indicate the direction of data flow.
- Typically, you use a data stream
  - to read data into Clementine,
  - run it through a series of manipulations, and then
  - send it to a destination, such as an SPSS file or the Clementine Solution Publisher.

Stream-Building Overview

- For example, suppose that you want
  - to open a data source,
  - add a new field,
  - select records based on values in the new field, and then
  - display the results in a table.
- In this case, your data stream would consist of four nodes:

A **Variable File node**, which you set up to read the data from the data source.

A **Derive node**, which you use to add the new, calculated field to the dataset.

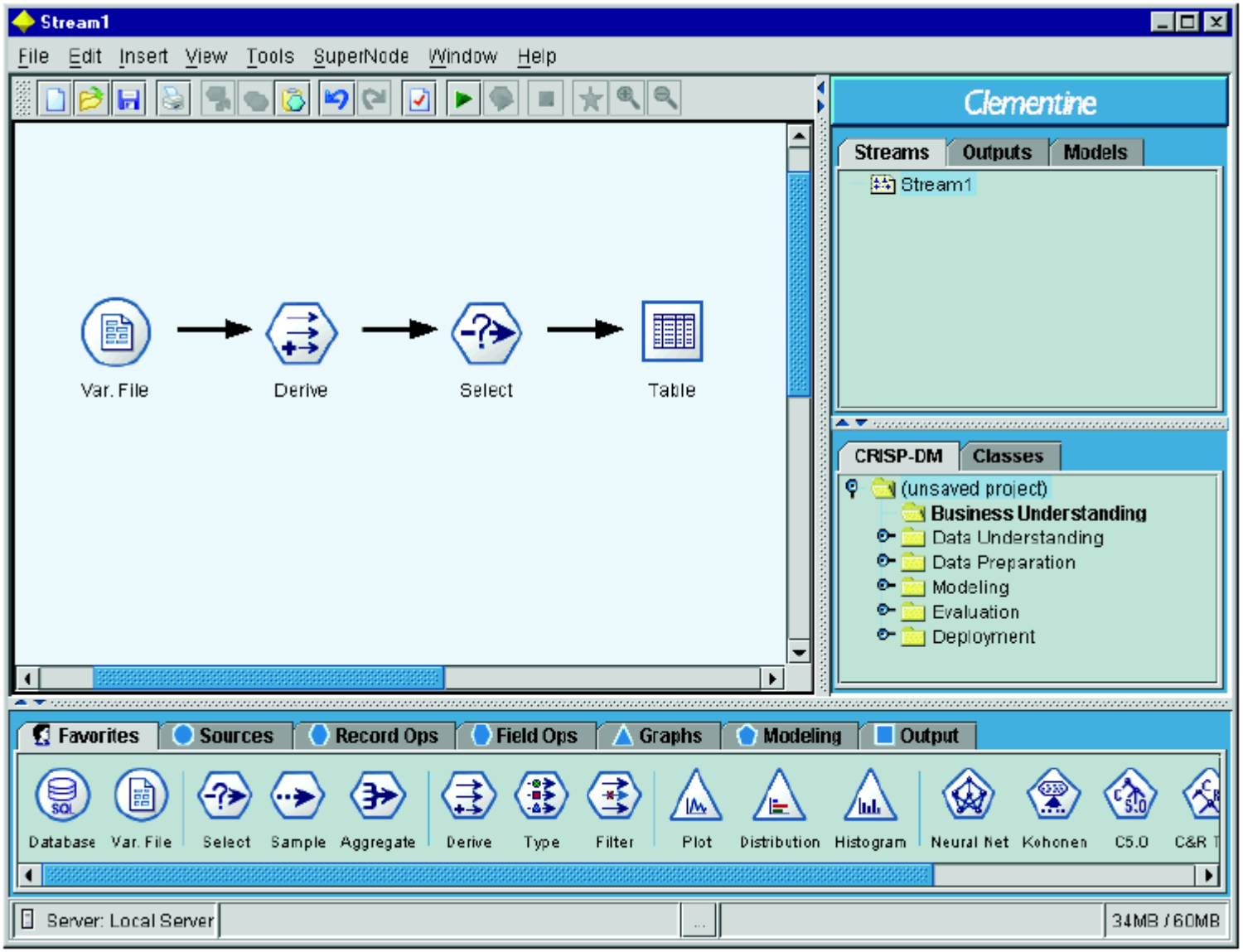
A **Select node**, which you use to set up selection criteria to exclude records from the data stream.

A **Table node**, which you use to display the results of your manipulations onscreen.

Building Data Streams

- You can build a data stream using the following steps:
  - Add nodes to the stream canvas.
  - Connect the nodes to form a stream.
  - Specify any node or stream options.
  - Execute the stream.

Completed stream on the stream canvas



Working with Nodes

Adding Nodes to a Stream

- There are three ways to add nodes to a stream from the nodes palette:
  - Double-click a node on the palette.
  - Drag and drop a node from the palette to the stream canvas.
  - Click a node on the palette, and then click on the stream canvas.
- Once you have added a node to the stream canvas, double-click the node to display its dialog box.
- The available options depend on the type of node that you are adding.
- For information about specific controls within the dialog box, click Help.

Removing Nodes

- To remove a node from the data stream, click it and press the Delete key,
- Or right-click and choose Delete from the context menu.

Connecting Nodes in a Stream

- Nodes added to the stream canvas do not form a data stream until they have been connected.
- Connections between the nodes indicate the direction of the data as it flows from one operation to the next.
- There are a number of ways to connect nodes to form a stream:
  - Double-clicking
  - Using the middle mouse button / Alt key
  - Manually

To Add and Connect Nodes by Double-Clicking

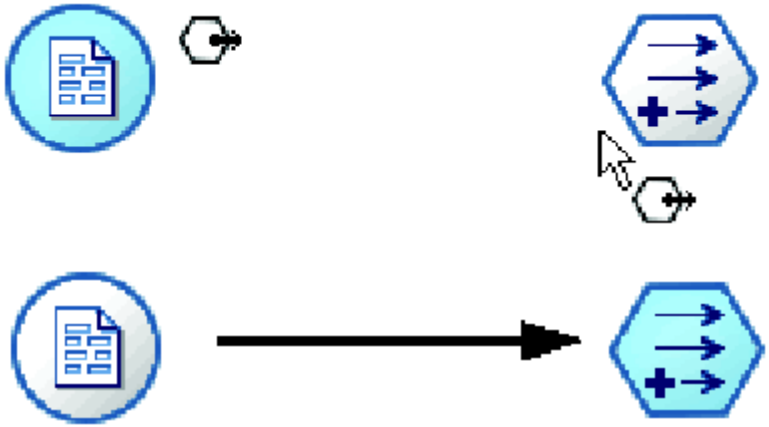
- The simplest way to form a stream is to double-click nodes on the palette.
- This method automatically connects the new node to the **selected node** on the stream canvas.

To Connect Nodes Using the Middle Mouse Button

- On the stream canvas, you can click and drag from one node to another using the middle mouse button.
- You can simulate this by pressing the **Alt key** while dragging with the mouse from one node to another.

To Manually Connect Nodes

- Select a node and right-click to open the context menu.
- From the menu, choose Connect.
- A connection icon will appear both on the start node and the cursor. Click on a second node on the canvas to connect the two nodes.

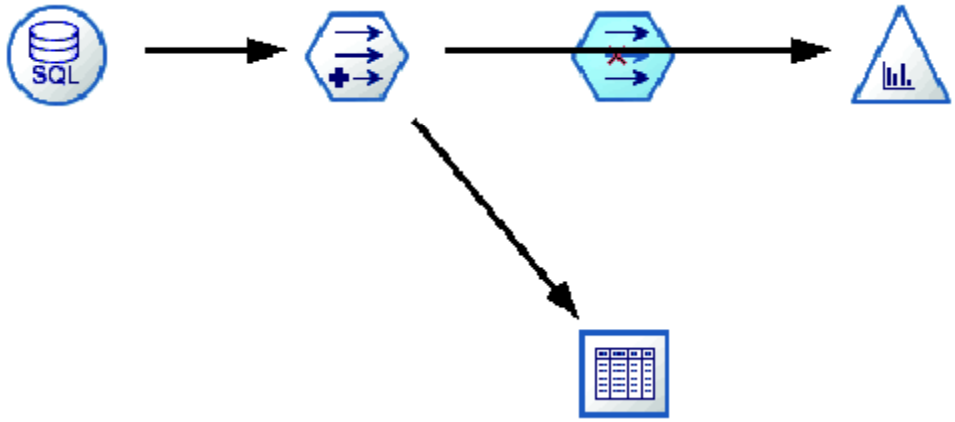


Error Message

- You will receive an error message if you attempt to make any of the following types of connections:
  - A connection leading to a source node
  - A connection leading from a terminal node
  - A node having more than its maximum number of input connections
  - Connecting two nodes that are already connected
  - Circularity (data returns to a node from which it has already flowed)

Bypassing Nodes in a Stream

- When you bypass a node in the data stream, all of its input and output connections are replaced by connections that lead directly from its input nodes to its output nodes.
- Example: Bypassing a previously connected Filter node



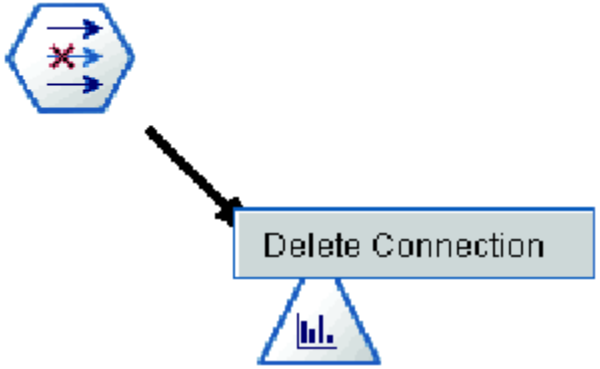
- On the stream canvas, use the **middle mouse button** to **double-click** the node that you want to bypass. Alternatively, you can use **Alt-double-click**.

Adding Nodes in Existing Connections

- You can add a new node between two connected nodes by dragging the arrow that connects the two nodes.
- With the middle mouse button, click and drag the connection arrow into which you want to insert the node.
- Alternatively, you can hold down the Alt key while clicking and dragging.

Deleting Connections between Nodes

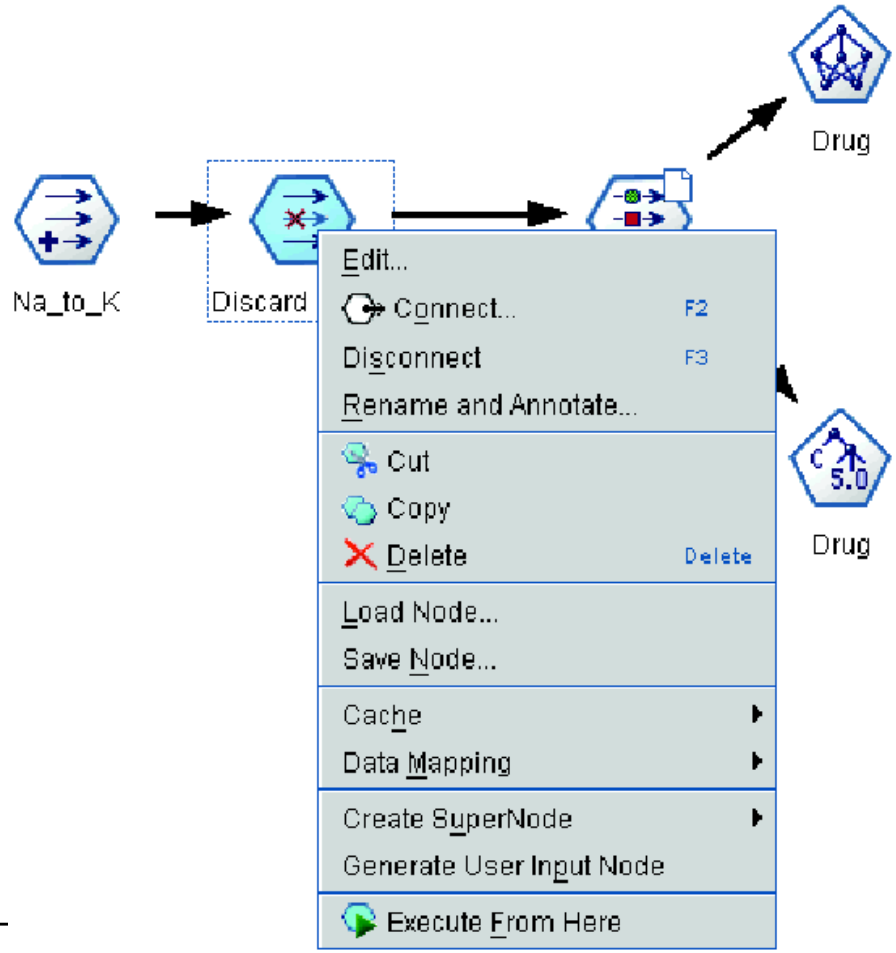
- You can delete the connection between nodes using two methods:
  - Press and hold down the **right mouse** button on the connection arrowhead. From the context menu, choose Delete Connection.



  - Or, you can delete a connection as follows:
    - Select a node and press the F3 key to delete all connections.
    - Select a node, and from the context menus choose Disconnect
    - Select a node, and from the main menus choose:
      - Edit > Node > Disconnect

Setting Options for Nodes

- There are several options for customizing nodes.
  - Right-click on a node and choose one of the menu options.



Setting Options for Nodes

- Context menu options for nodes
  - Choose **Edit** to open the dialog box for the selected node.
  - Choose **Connect** to manually connect one node to another.
  - Choose **Disconnect** to delete all links to and from the node.
  - Choose **Rename and Annotate** to open the Annotations tab of the editing dialog box.
  - Choose **Cut** or **Delete** to remove the selected node(s) from the stream canvas. Choosing Cut allows you to paste nodes, while Delete does not.
  - Choose **Copy** to make a copy of the node with no connections. This can be added to a new or existing stream.
  - Choose **Load Node** to open a previously saved node and load its options into the currently selected node.
  - Choose **Save Node** to save the node's details in a file. You can load node details only into another node of the same type.

Annotating Nodes

- Nodes, streams, and models can be annotated in a number of ways.
- You can add descriptive annotations and specify a custom name.
- These options are useful especially when generating reports for streams added to the projects tool.
- For nodes, you can also add ToolTip text to help distinguish between similar nodes on the stream canvas.
- To annotate a node, right-click on the node on the stream canvas and choose Rename and Annotate.
  - The editing dialog box opens with the Annotations tab visible.

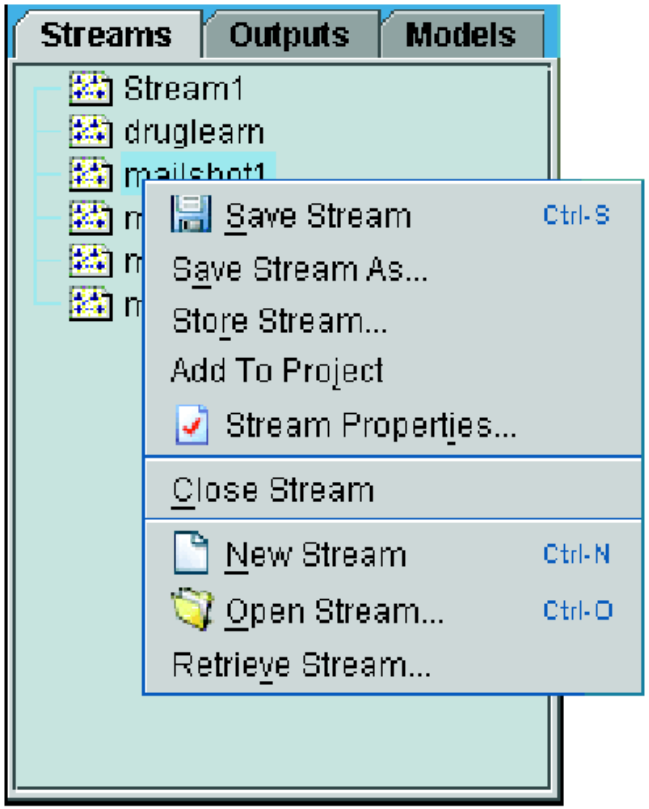
Working with Streams

Working with Streams

- Once you have connected source, process, and terminal nodes on the stream canvas, you have created a stream.
- As a collection of nodes, streams can be saved, annotated, and added to projects.
- In Clementine, you can use and modify more than one data stream at a time.
- The right side of the main window contains the managers tool, which helps you to navigate the streams currently open.
- To view the managers tool, choose Managers from the View menu.

Working with Streams

- Streams tab in the managers tool with context menu options:
- From this tab, you can:
  - Save streams
  - Save streams to the current project
  - Close streams
  - Open new streams



Setting Options for Streams

- For the current stream, you can specify a number of options.
- From the File menu, choose Stream Properties.
  - Alternatively, you can use the context menu on the Streams tab in the managers tool.
- **Streams options tabs:**
  - Options
    - For the current stream, you can specify a number of options.
  - Layout
    - you can specify a number of options regarding the display and use of the stream canvas.

Setting Options for Streams

- **Streams options tabs (cont.):**
  - Messages tab
    - Messages regarding stream operations, such as execution, optimization, and time elapsed for model building
    - Error messages are also reported in this table.
  - Parameters
    - Parameters can be defined for use in CLEM expressions and in scripting.
    - They are, in effect, user-defined variables that are saved and persisted with the current stream, session, or SuperNode and can be accessed from the user interface as well as through scripting.
  - Deployment
    - Specifies options for deploying the stream as a scenario within SPSS Predictive Enterprise Services

Setting Options for Streams

- **Streams options tabs (cont.):**
  - Globals
    - Using the Globals tab in the stream properties dialog box, you can view the global values set for the current stream.
    - Global values are created using a Set Globals node to determine statistics such as mean, sum, or standard deviation for selected fields.
  - Annotations
    - Using the Annotations tab in the stream properties dialog box, you can add descriptive annotations for a stream and create a custom name for the stream.

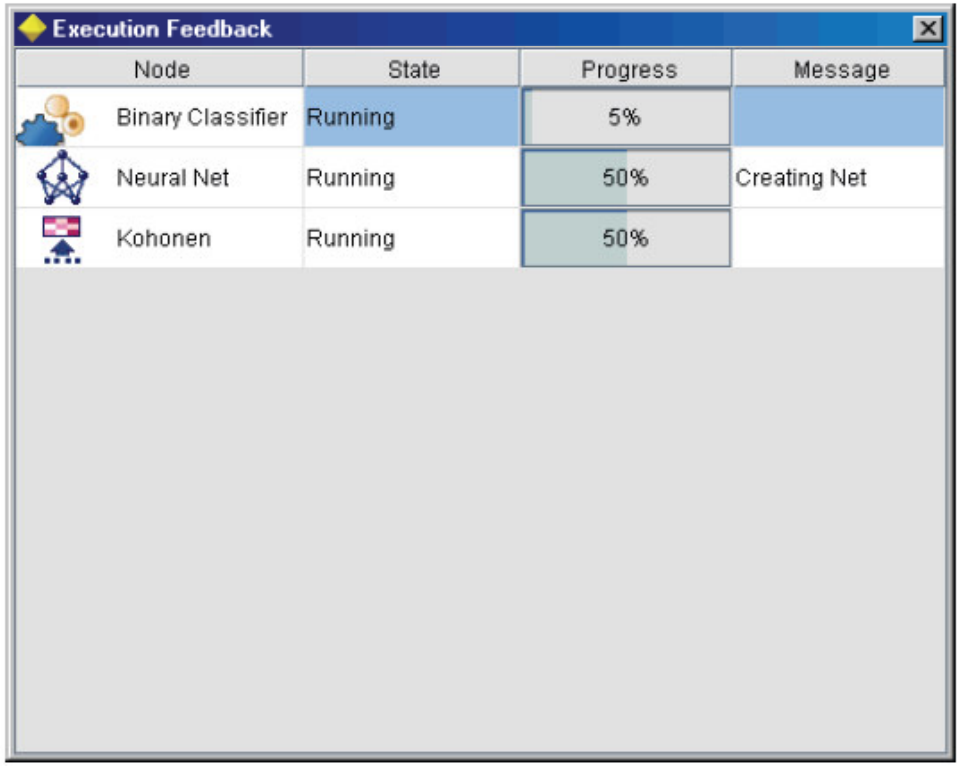
Executing Streams

Executing Streams

- There are several ways to execute a stream within Clementine:
  - You can choose Execute from the Tools menu.
  - You can click one of the execute buttons on the **toolbar**.
    - These buttons allow you to execute the entire stream or simply the selected terminal node.
  - You can execute a single data stream by right-clicking a terminal node and choosing Execute from the context menu.
  - You can execute part of a data stream by right-clicking any non-terminal node and choosing Execute From Here from the context menu, which executes all operations after the selected node.

Executing Streams

- To halt the execution of a stream in progress,
  - you can click the red stop button on the toolbar or
  - choose Stop Execution from the Tools menu.
- If any stream takes longer than three seconds to execute, the Execution Feedback dialog box is displayed to indicate the progress.



Saving Files

Saving Streams

- After you have created a stream, you can save it for future reuse.
- **To Save a Stream**
  - From the File menu, choose Save Stream or Save Stream As.
  - In the Save dialog box, browse to the folder in which you want to save the stream file.
  - Enter a name for the stream in the File Name text box.
  - Select Add to project if you would like to add the saved stream to the current project.

Saving Streams

- The stream is stored with the extension *.str in the specified directory.
- **Automatic backup files**
  - Each time a stream is saved, the previously saved version of the file is automatically preserved as a backup, with a hyphen appended to the filename
  - for example mystream.str-
  - To restore the backed-up version, simply delete the hyphen and reopen the file.

Saving States

- You can save **states**, which include the currently displayed stream diagram and any model nuggets that you have created (listed on the Models tab in the managers window).
- **To Save a State**
  - From the File menu, choose: State > Save State or Save State As
  - In the Save dialog box, browse to the folder in which you want to save the state file.
  - Clicking Save stores the state with the extension *.cst in the specified directory.

Saving Nodes

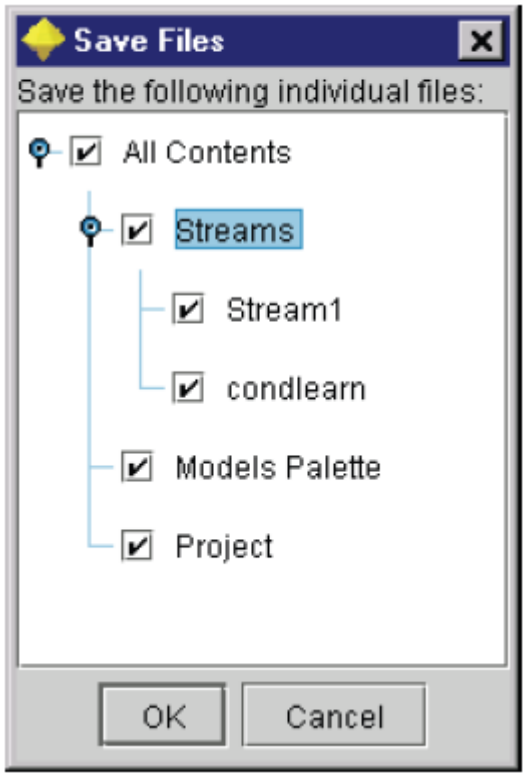
- You can also save an individual node by right-clicking the node on the stream canvas and choosing **Save Node** from the context menu.
- Use the file extension *.nod.

Saving Output

- Tables, graphs, and reports generated from Clementine output nodes can be saved in output object (*.cou) format.
- When viewing the output you want to save, from the output window menus choose: File > Save
- Specify a name and location for the output file.
- Optionally, select Add file to project in the Save dialog box to include the file in the current project.
- Alternatively, you can right-click on any output object listed in the managers window and select Save from the context menu.

Saving Multiple Stream Objects

- When you exit Clementine with multiple unsaved objects, such as streams, projects, or the model nuggets palette,
  - you will be prompted to save before completely closing the software.
  - If you choose to save items, a dialog box will open with options for saving each object.

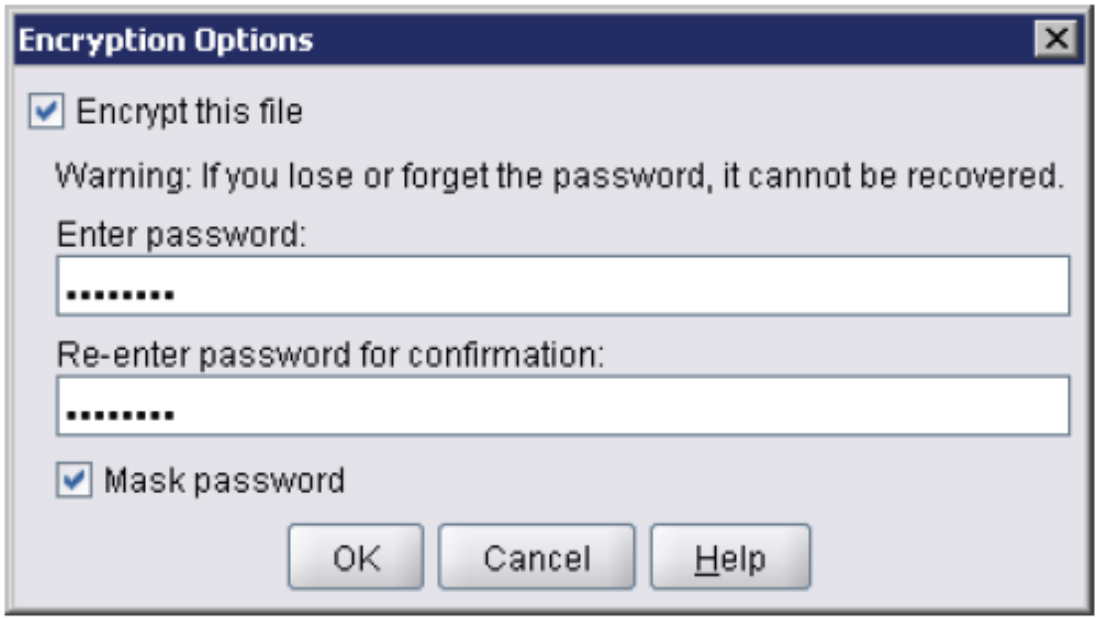


Encrypting and Decrypting Information

- When you save a stream, node, project, output file, or model nugget, you can **encrypt** it to prevent its unauthorized use.
- To do this, you select an extra option when saving, and add a password to the item being saved.
- When you try to open an encrypted item, you are prompted to enter the password.
- After you enter the correct password, the item is decrypted automatically and opens as usual.

Encrypting and Decrypting Information

- **To Encrypt a File or Model**
  - In the Save dialog box, for the item to be encrypted, click Options.
  - The Encryption Options dialog box opens.
  - Select Encrypt this file.
  - Enter the password.
  - Click OK to return to the Save dialog box.



Loading Files

Loading Files

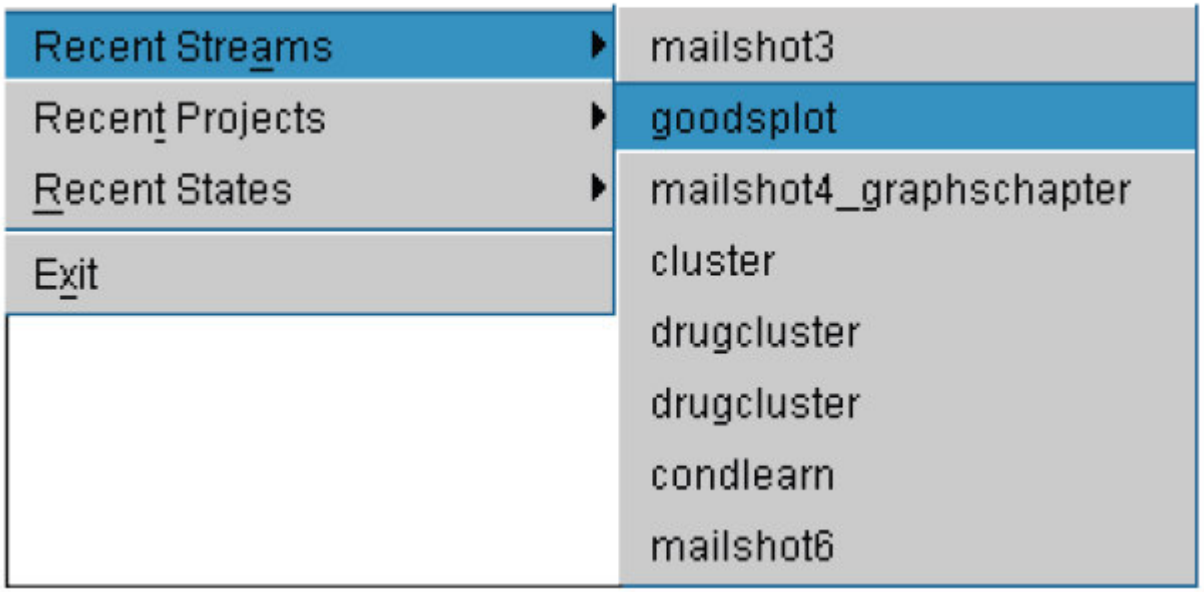
- You can reload a number of saved objects in Clementine:
  - Streams (*.str)*
  - States (*.cst)*
  - Models (*.gm)*
  - Models palette (*.gen)*
  - Nodes (*.nod)*
  - Output (*.cou)*
  - Projects (*.cpj)*

Loading Files

- **Opening New Files**
  - Streams can be loaded directly from the File menu.
    - From the File menu, choose Open Stream.
  - All other file types can be opened using the submenu items available on the File menu.
    - For example, to load a model, from the File menu, choose:
      - Models > Open Model or Load Models Palette

Loading Files

- **Opening Recently Used Files**
  - For quick loading of recently used files, you can use the options at the bottom of the File menu.
  - Select Recent Streams, Recent Projects, or Recent States to expand a list of recently used files.



References

References

- Integral Solutions Limited., **Clementine® 12.0 User's Guide**, 2007.

The end