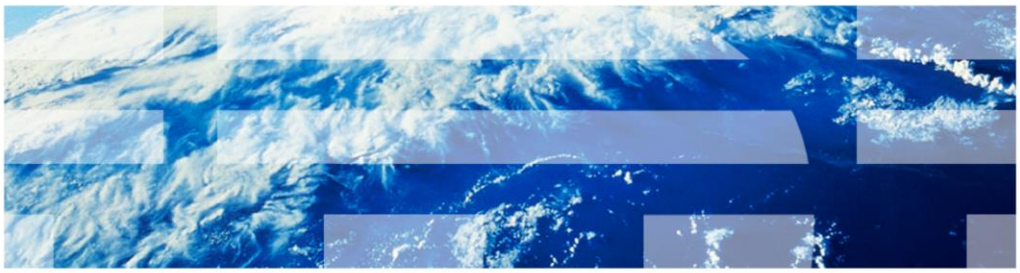
IBM

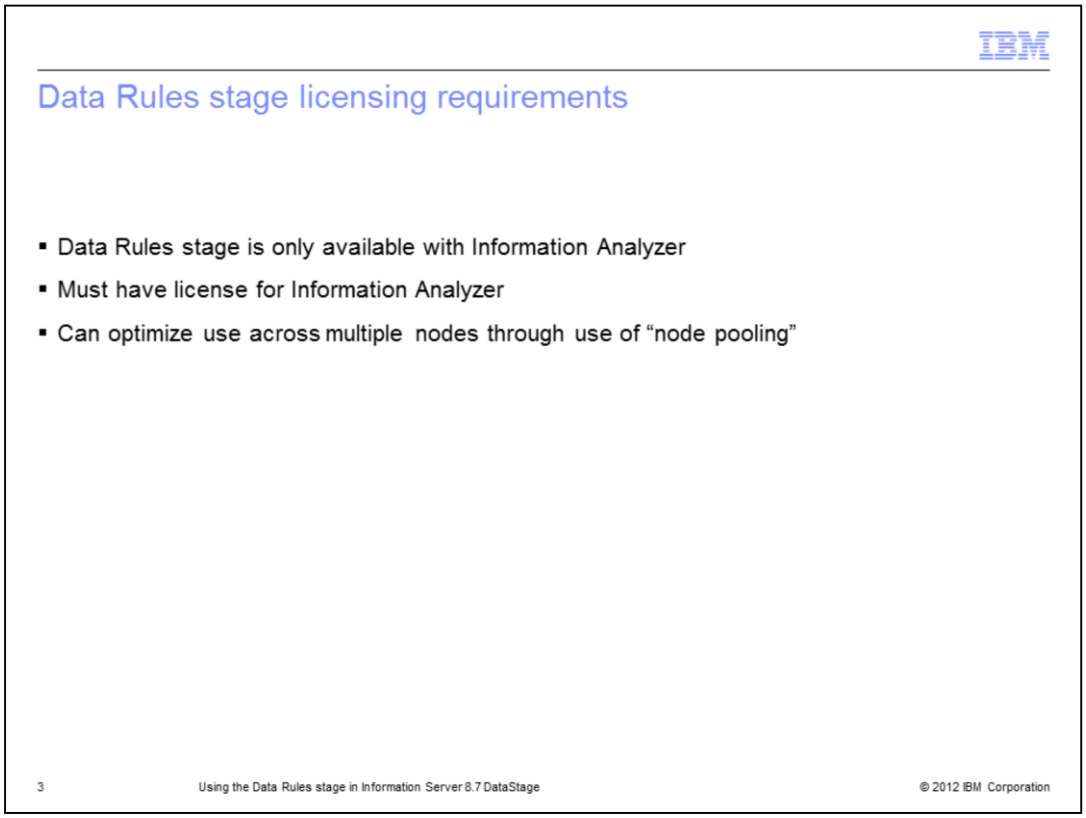
InfoSphere Information Server

Using the Data Rules stage in Information Server 8.7 DataStage



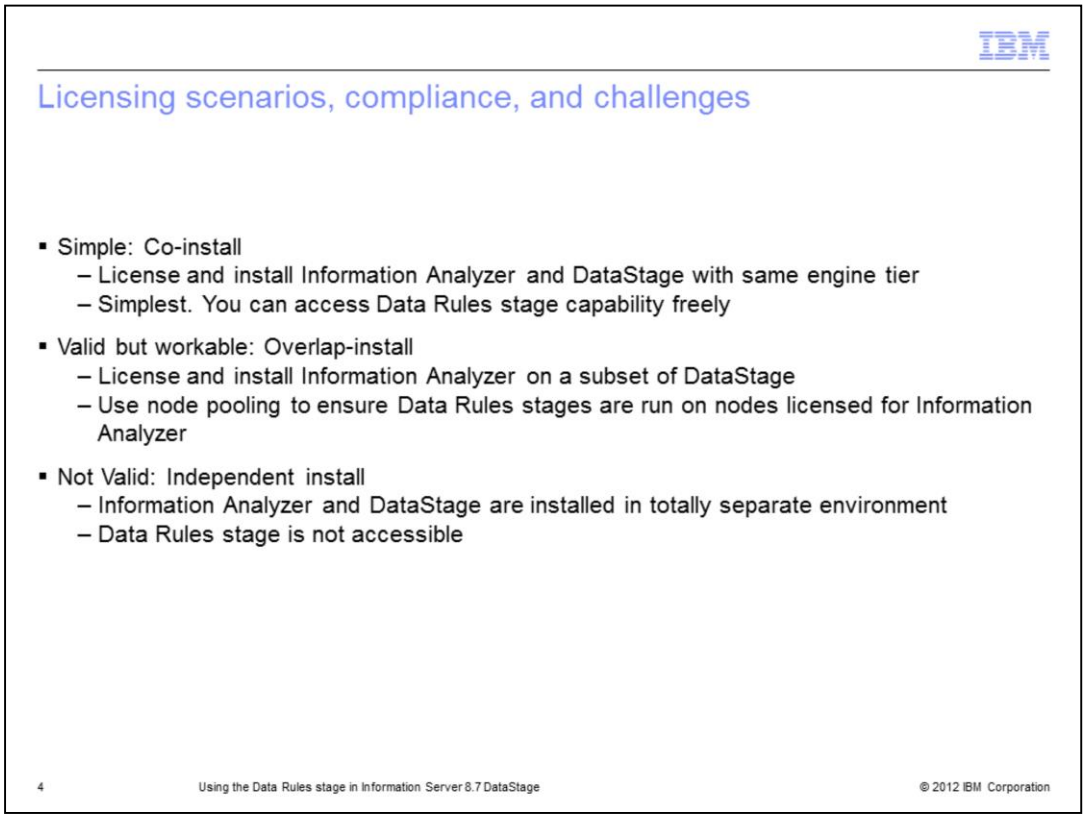
This presentation describes how to setup your environment to use the Data Rules stage in DataStage® version 8.7.

The objectives of this presentation are to understand the InfoSphere® Information Server version 8.7 licensing requirements for the Data Rules stage and to understand how to set up to use the Data Rules stage. This presentation also reviews a simple example job using the Data Rules stage.



The Data Rules stage is not part of the stages that are provided by default with DataStage. The Data Rules stage is a stage that is available with the Information Analyzer product. When Information Analyzer is installed as part of the InfoSphere Information Server install or installed later, the stage becomes available.

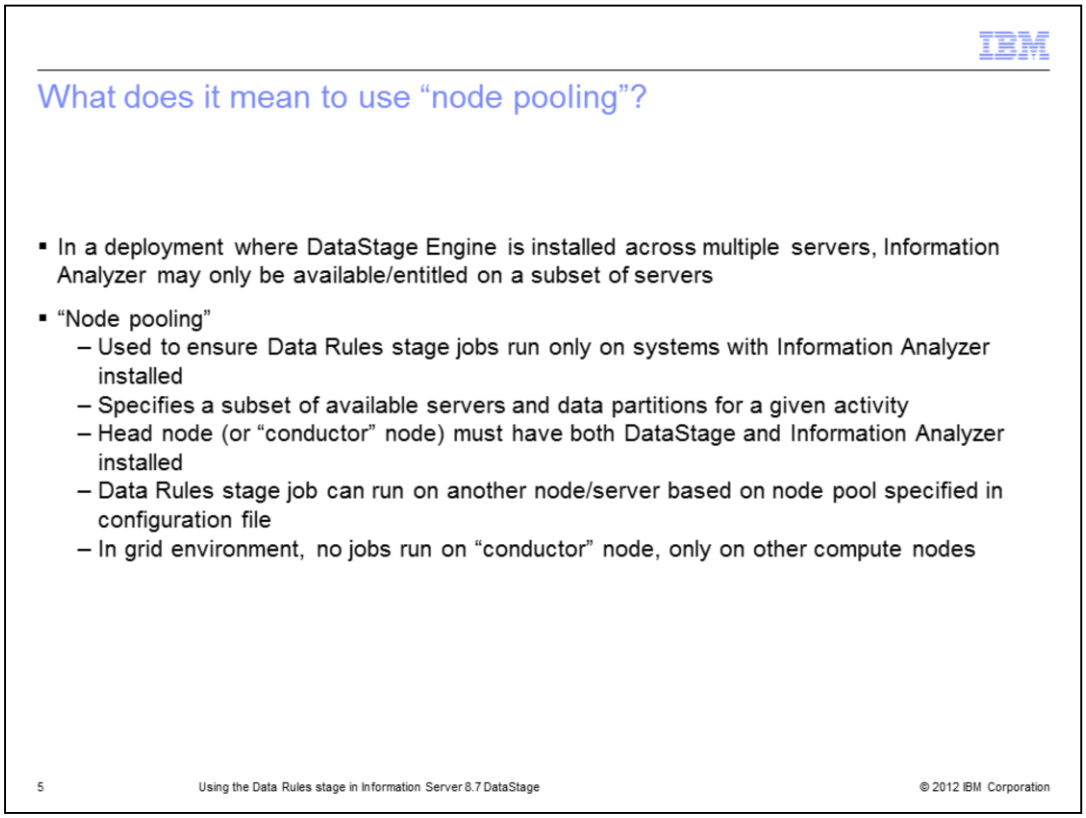
A means of optimizing the use of the Data Rules stage is to use node pooling to allow job execution of the Data Rules stage on multiple nodes that have Information Analyzer installed. There are some limitations and these are discussed in upcoming slides.



Three different licensing scenarios are discussed. The first scenario is called a co-installation. With co-installation, you have licenses for Information Server DataStage and Information Analyzer and you install the products on the same engine tier. This is the simplest way to have the Data Rules stage available within the DataStage Designer.

The next scenario is the overlapping installation. With overlapping installation you may have a limited number of Information Analyzer licenses. You will make use of node pooling to make sure that the jobs that have a Data Rules stage only run on those nodes licensed for Information Analyzer. The next slide discusses what node pooling is and later slides discuss how to set up for it.

The final scenario is the independent install. In this scenario, Information Analyzer and DataStage are installed in totally separate environments. The result of this is that the Data Rules stage is not available for use with DataStage.

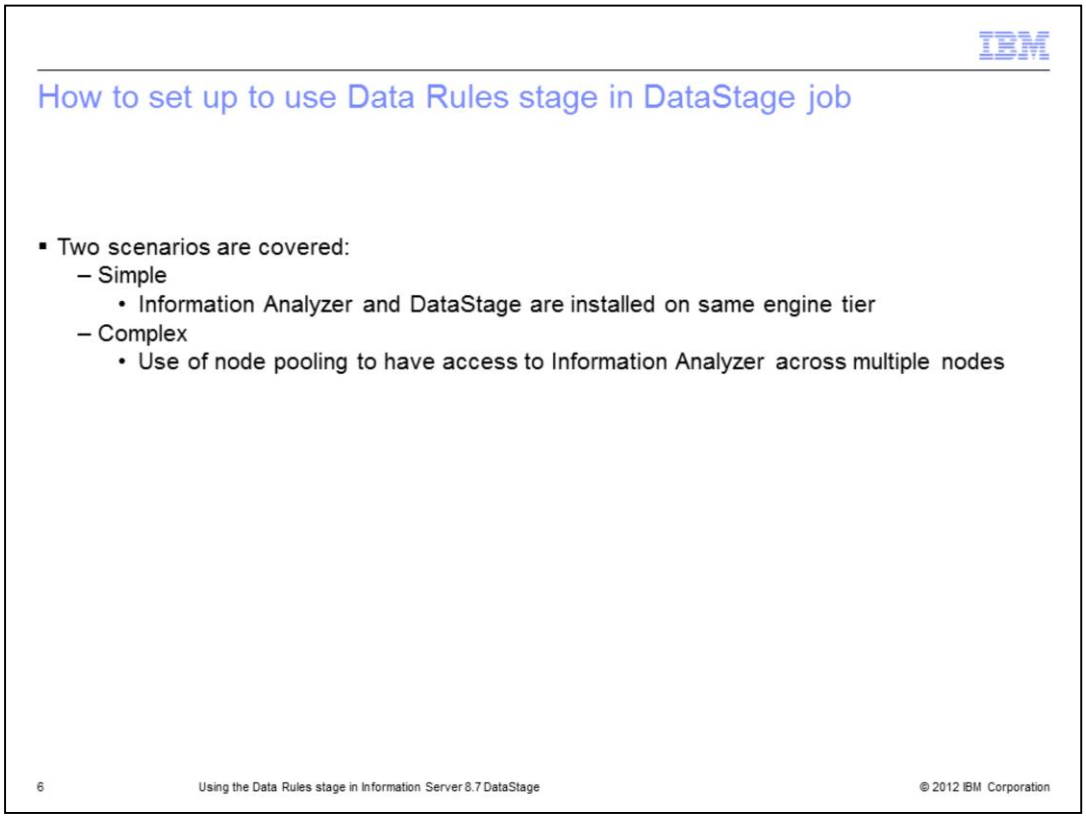


The InfoSphere Information Server parallel engine makes use of configuration files to determine what processing and storage resources belong to your system. You define each processing node on which the parallel engine runs jobs and classify its characteristics. You can use this classification capability to define nodes that have Information Analyzer installed to run your jobs with the Data Rules stage.

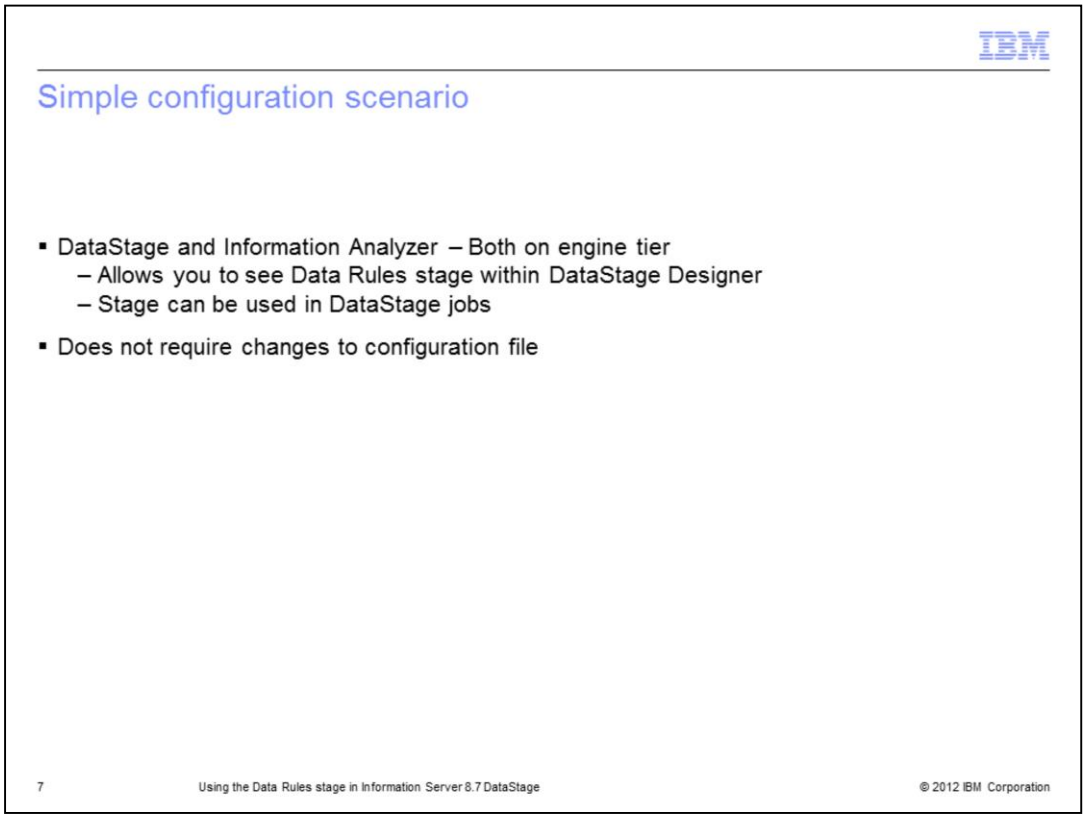
In order for node pooling to work in this case, a head node or conductor node must be designated. The conductor node drives the execution of the job across the other nodes and it must have both DataStage and Information Analyzer installed on it. Your job using the Data Rules stage may run on that node or another node server based on what is specified in the node pool configuration.

Other nodes are only required to have Information Analyzer installed to be used in your Data Rules stage node pool.

NOTE: In a grid environment, no job runs on the conductor node, but solely on the other compute nodes.

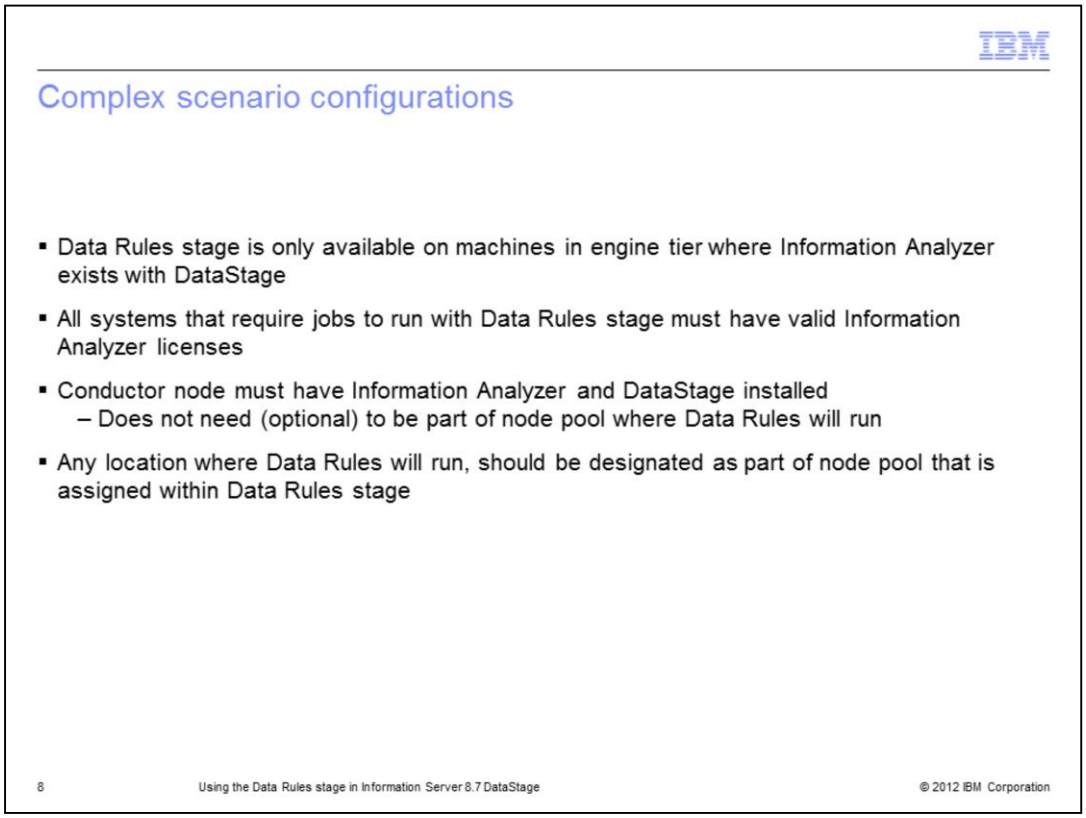


The next few slides review two scenarios for the set up of the Data Rules stage in a DataStage job; the simple and the complex scenario. By simple, this means that both DataStage and Information Analyzer are installed on the same engine tier. The complex scenario makes use of node pooling to have access to Information Analyzer across multiple nodes.



In the simple configuration scenario, DataStage and Information Analyzer are installed on the same machines that comprise the engine tier. With this configuration, when you launch DataStage Designer, the Data Rules stage automatically becomes available for use within jobs.

Configuration options are no different than for any other DataStage job. No changes to the configuration file is necessary to support this scenario.



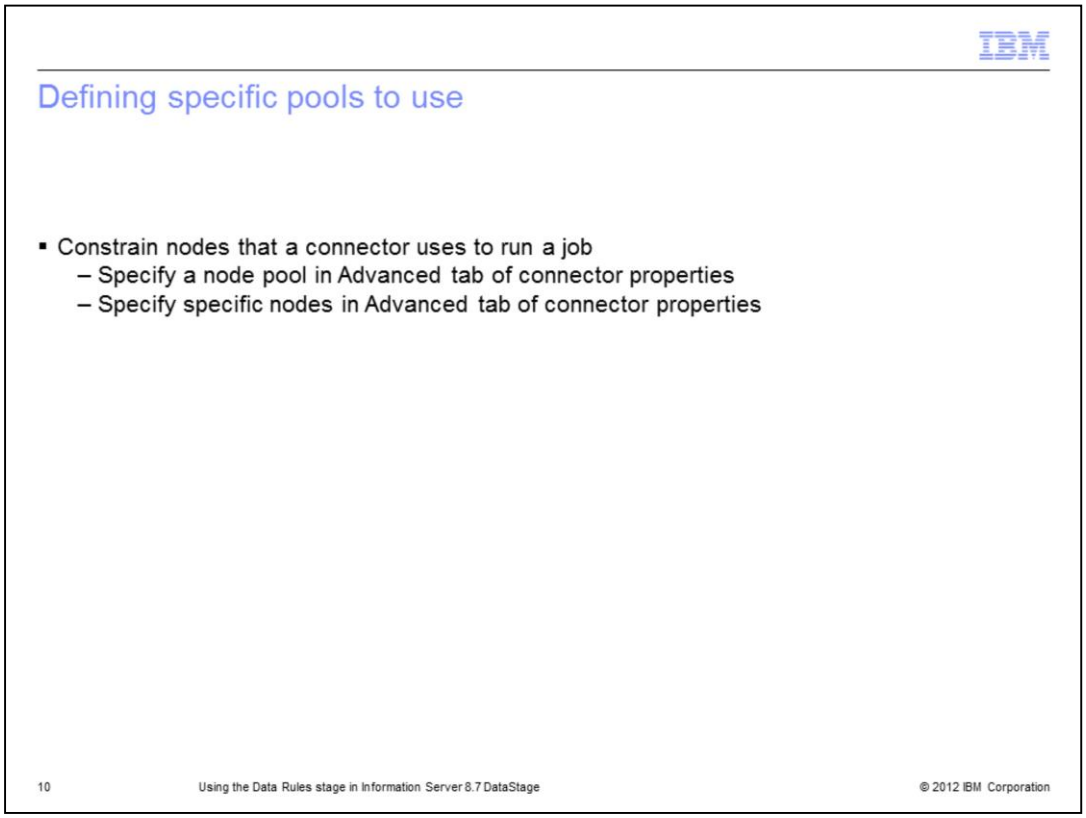
The complex scenario configuration can also be identified as an overlapping install. This means that you make use of node pooling to optimize the Information Analyzer licenses that are available. As discussed earlier, there are restrictions to the use of the Data Rules stage. The Data Rules stage will only be available on the machines in the engine tier where Information Analyzer exists with DataStage. Additionally, all systems that run jobs with the Data Rules stage must have valid licenses.

In order to accomplish this, the parallel engine's configuration file can be set up using node pools that point to nodes that have Information Analyzer installed. Additionally, at least one node, designated as the conductor node, must exist that has both DataStage and Information Analyzer installed on an engine tier machine. The other nodes in the pool, the non-conductor nodes, need only to have Information Analyzer installed.

In order to use node pools, the configuration file needs to be set up properly. The following example parallel configuration file defines four nodes: node1, node2, node3, and node4. Node1 is defined as the conductor node. The file defines three node pools: group1, group2, and group3. It also defines the default "" pool. The APT_CONFIG_FILE environment variable points to this parallel configuration file.

In this example, node1 has both Information Analyzer and DataStage installed on it, that is why it has been designated the conductor node. The nodes, node2 and node4, have Information Analyzer installed on them. The plan is to use group2 as the node pool for jobs containing the Data Rules stage.

Notice that you cannot use the node pool group1 or group3 as these node pools contain node3, which does not have Information Analyzer installed on it.

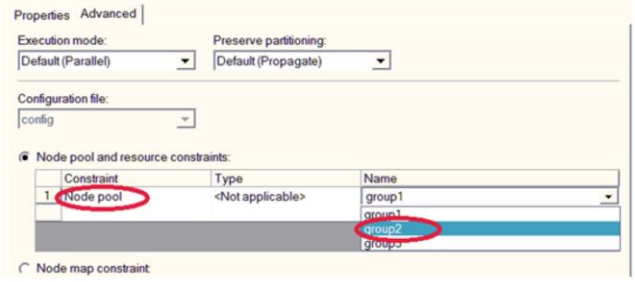


In the parallel configuration file, you specify nodes and node pools as discussed previously. Then you can use one of two methods to configure the connector used in the RuleStage job to run only a subset of the nodes that are specified in the parallel configuration file and have Information Analyzer licenses for your Data Rules stage.

There are two methods to do this. Method one is to define a node pool in the configuration file. Method two is to have the connector only run on specific nodes.

Specifying the node pool

- Specify node pool
  - On Advanced tab of stage properties, select 'Node pool and resource constraints'
  - In Constraint field, select 'Node Pool'
  - In Name field, select 'group2'

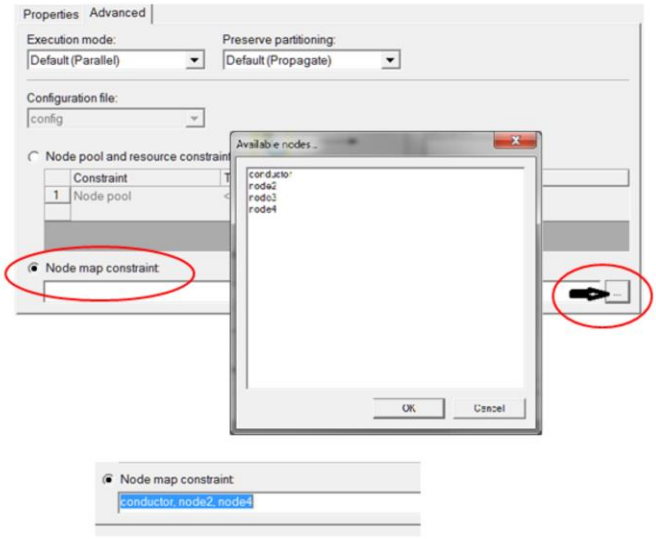


The first method to specify the pool to use is to define a node pool in the configuration file. In the previous example, the node pool for Data Rules stage used group2. On the Advanced tab of the connector properties, select that node pool. The connector will then only run the job on the nodes that are members of that node pool.

In the diagram displayed on this slide, and as referenced in the example, if 'group2‘ is selected, the job will run on 'node1', 'node2' and 'node4'.

Selecting specific nodes to use

- Specify specific nodes
  - On Advanced tab of stage properties, select 'Node map constraint'
  - Select 'conductor' and 'node2' and 'node4'

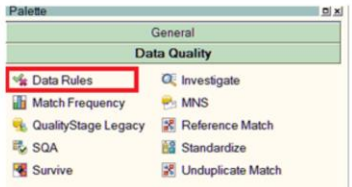


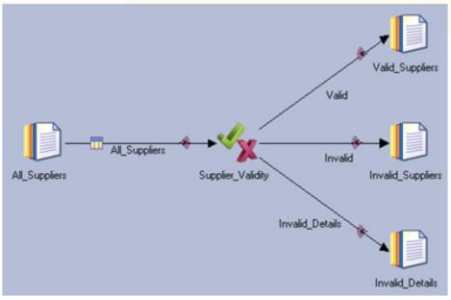
The second method is to have the connector only run on specific nodes. To do this, go to the Advanced tab of the connector properties and select the specific nodes that you want the connector to run.

From the diagram displayed on this slide, and as referenced in the example, select 'node1', which is the conductor node, 'node2' and 'node4'. Do not select 'node3' as it does not have Information Analyzer installed on it.

Using the Data Rules stage

- DataStage Designer, go to palette, Data Quality and you will see Data Rules stage
- Data Rules stage job that checks for data – completeness





IBM InfoSphere Information Server 8.7 Information Center has a topic on Information Analyzer Data Rules stage that provides more details and examples:
http://publib.boulder.ibm.com/infocenter/iisinfsv/v8r7/topic/com.ibm.swg.im.iis.ia.dsrules.doc/topics/dr_data_rules_stage.html

13 Using the Data Rules stage in Information Server 8.7 DataStage

Once you have completed the setup and have access to the Data Rules stage, you can begin creating jobs that check data quality anywhere in the flow of the job using the Data Rules stage.

You can find the Data Rules stage under the Data Quality section of the palette as shown with the red square in the image on the slide. To use this in a job, drag the Data Rules stage onto your job canvas. You need to add input and output links to the Data Rules stage and configure the links.

The example displayed on this slide is a Data Rules stage job that checks for data completeness.

The IBM InfoSphere Information Server version 8.7 Information Center documentation, contains information on the Information Analyzer Data Rules stage and provides additional details on the stage and how to use it in a job. It also has three example Data Rules stage jobs that can be examined. The link to this documentation is displayed on this slide.

Trademarks, disclaimer, and copyright information