

IBM WebSphere DataStage : A Brief Overview

Rajeev Priyadarshi,
President,
PR3 Systems
rpriyadarshi@pr3systems.com



Copyright PR3 Systems, 2005



Topics to be covered

- What is Data Warehousing, ETL and Business Intelligence?
- Product Overview of DataStage
- Types of DataStage Clients
- DataStage Administrator
- DataStage Manager
- DataStage Designer
- DataStage Director

Why is Data Warehousing?

- A data warehouse is a collection of data gathered and organized so that it can easily be analyzed, extracted, synthesized, and otherwise be used for the purposes of further understanding the data. It may be contrasted with data that is gathered to meet immediate business objectives such as order and payment transactions, although this data would also usually become part of a data warehouse.

What is Data ETL?

- A process of gathering, converting and storing data, often from many locations. The data is often converted from one format to another in the process. ETL is an abbreviation for "Extract, Transform and Load"
Examples : IBM DataStage, Informatica

What is BI?

- Business intelligence (BI) is a broad category of application programs and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining.

Examples : BusinessObjects :
www.businessobjects.com

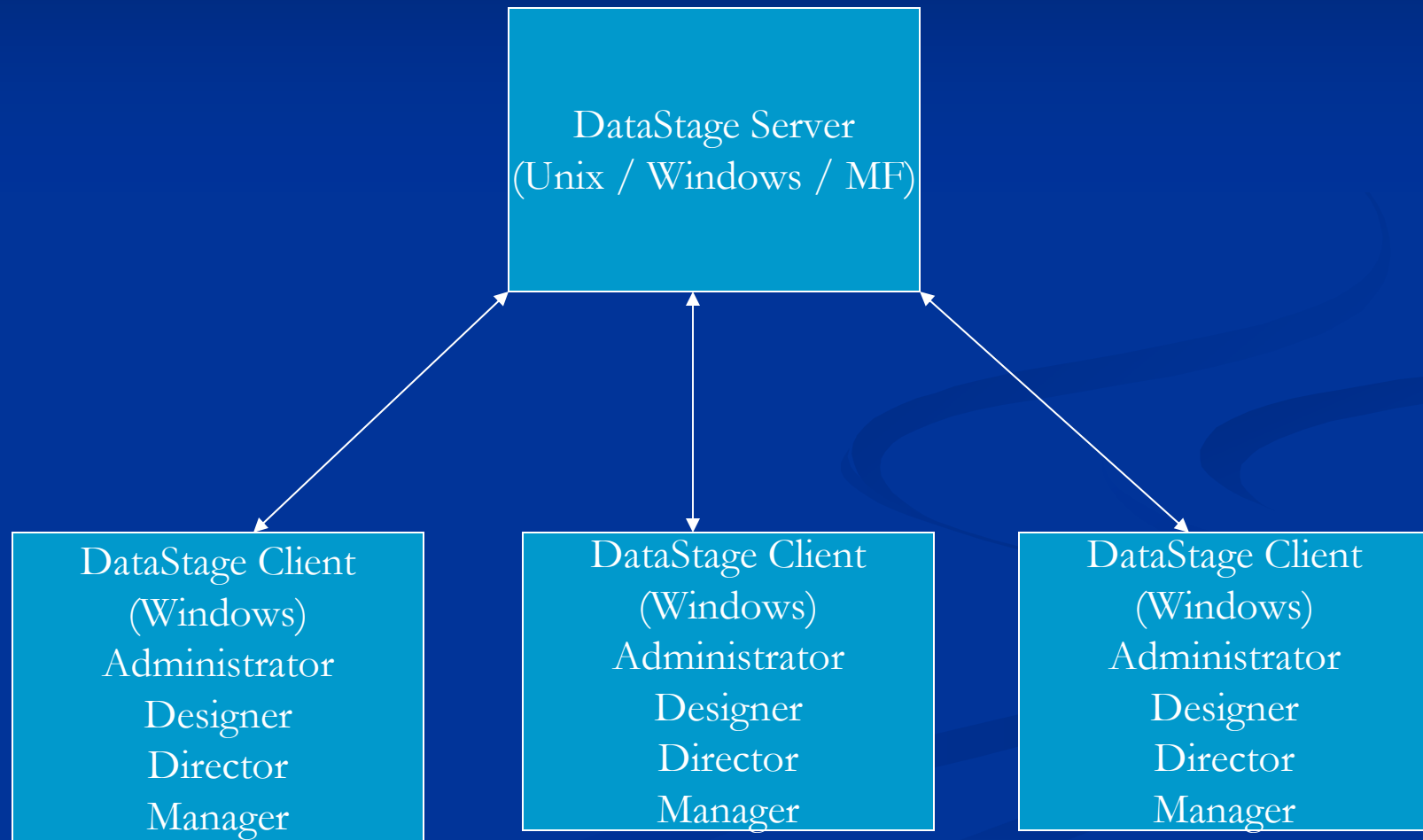
Careers in this Domain

- Much easier to pick up than software languages.
- New technologies, hence requires new resources.
- Return on Investment high. Salary / Training Effort Ratio.
- Less Competitive than mainstream software development.

PR3 DataStage Training Course

- We specialize in providing education and consulting services for IBM's WebSphere DataStage Products.
- We have completed several successful DataStage projects for Fortune 500 companies.
- We have come up with an unique approach of ETL project development [PR3 RUSK Framework] enhancing the re-usability, scalability and High-Availability of the processing framework.

DataStage Architecture



Product Overview

- DataStage is a product from IBM being used as the strategic ETL tool within many organizations.
- It can be used for multiple purposes:
 - Interfacing between multiple databases.
 - Changing of data from one format to another. Eg. From database to flat files, XML files, etc.
 - Fast access to data that doesn't change often.
 - Interacts with WebSphere MQ to provide real time processing capabilities triggered by external messages.

Usage of DataStage within organizations

- DataStage has Windows Clients which connect to the Server on the Unix / Windows or Mainframe platform.
- The clients can be used to develop, deploy and run datastage jobs.
- In a deployment environment, the jobs can be kicked off through scripts directly on Unix servers

Types of DataStage clients

- DataStage Administrator
- DataStage Designer
- Datastage Manager
- Datastage Director

DataStage Administrator

- Most DataStage configuration tasks are carried out using the DataStage Administrator, a client program provided with DataStage.
- To access the DataStage Administrator:
 1. From the Ascential DataStage program folder, choose **DataStage Administrator**.
 2. Log on to the server. If you do so as an Administrator (for Windows NT servers), or as dsadm (for UNIX servers), you have unlimited administrative rights; otherwise your rights are restricted as described in the previous section.
 3. The DataStage Administration window appears: The **General** page lets you set server-wide properties. It is enabled only when at least one project exists. The controls and buttons on this page are enabled only if you logged on as an administrator

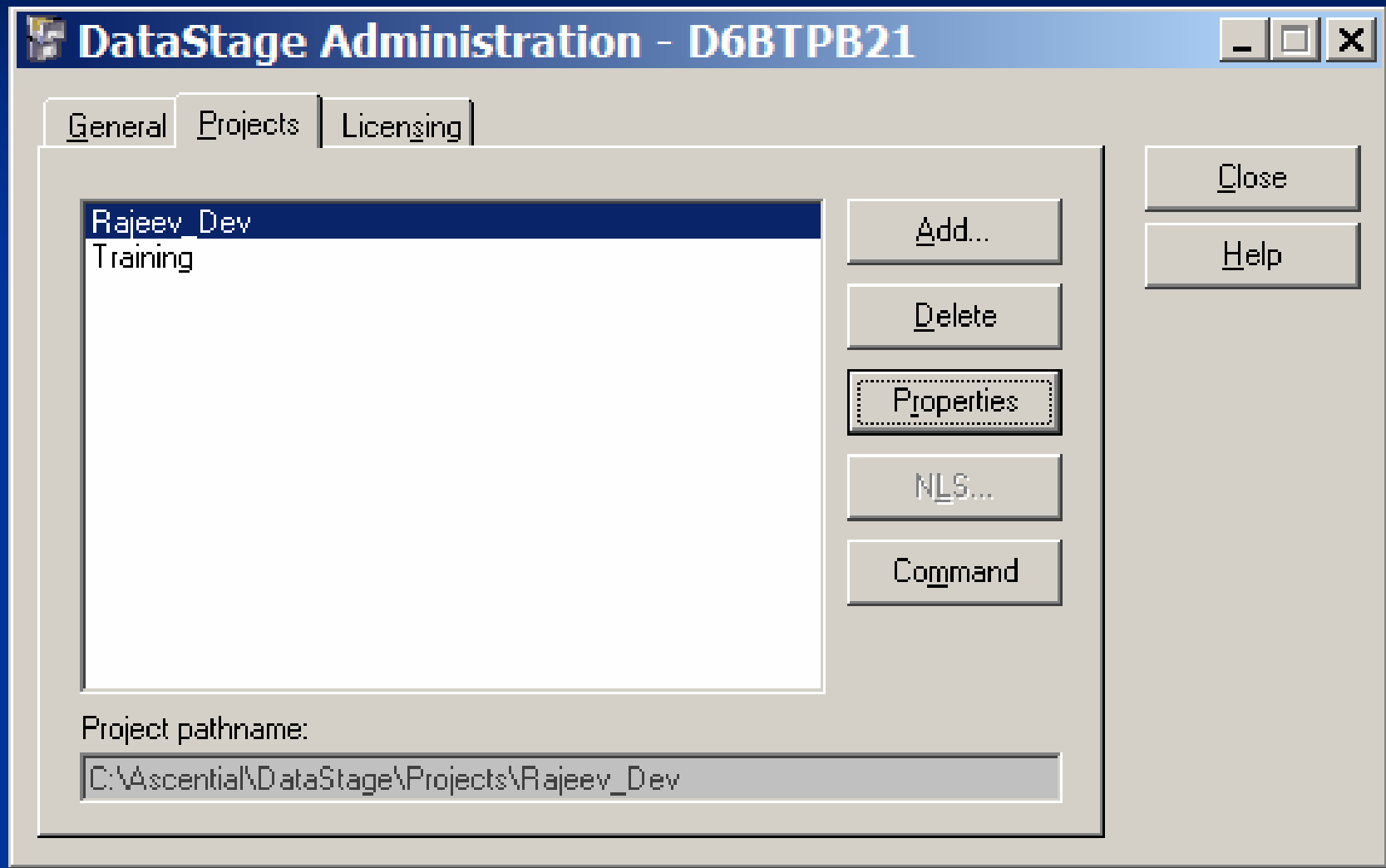
Administrator Interface

The screenshot shows a window titled "DataStage Administration - D6BTPB21" with three tabs: "General", "Projects", and "Licensing". The "Licensing" tab is active. It contains two main sections: "License Details" and "Client License".

| Field | Value |
|-----------------|-----------------|
| Serial # | 85979 |
| CPU Count | 4 (3 Available) |
| Expiration Date | 12/31/2005 |
| Serial # | 75717-DSDDES |
| User Limit | 7 |
| Expiration Date | 1/01/2500 |

Buttons: "Change..." (under License Details), "Upgrades..." and "Change" (under Client License), "Close" and "Help" (on the right).

Administrator Interface



Administrator Interface

Project Properties - D6BTPB21\Rajeev_Dev [X]

General | Permissions | Tracing | Schedule | Mainframe | Tunables | Parallel | Sequence | Remote

Enable job administration in Director

Enable Runtime Column Propagation for Parallel Jobs

Default setting for new Parallel jobs:

Enable Runtime Column Propagation for new links

Enable editing of internal references in jobs

Auto-purge of job log

Auto-purge action:

Up to previous: [] job run(s)

Over: [] day(s) old

Protect Project

Environment...

OK

Cancel

Help

Project Properties Screen

Environment variables

The following categorized environment variables are defined in this project. Either set a default value for an existing environment variable or add a new environment variable to the user defined category.

Categories:

- General
- Customize
- Parallel
 - Compiler
 - Operator Specific
 - Reporting
 - User Defined

Details:

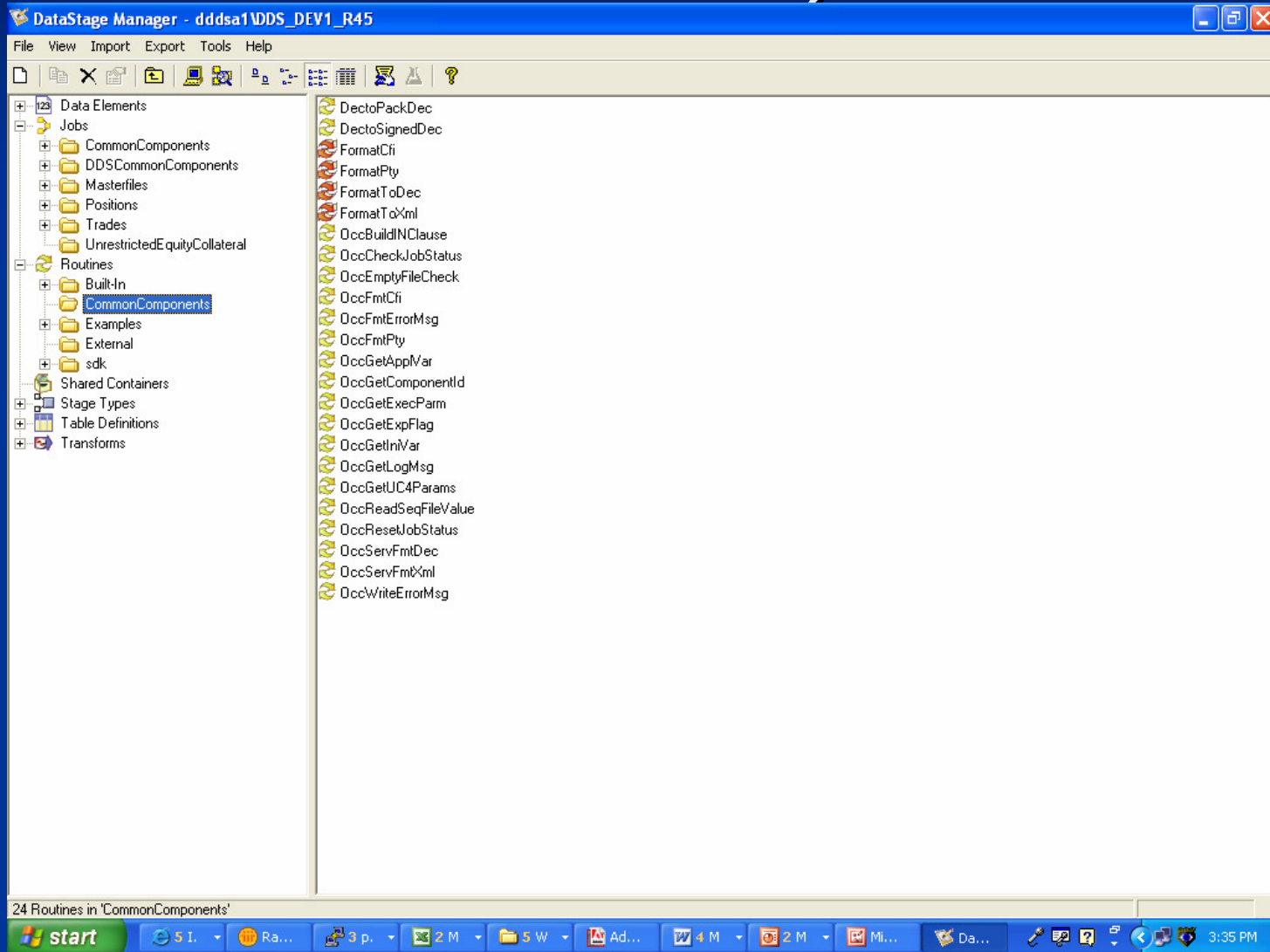
| Name | Prompt | Value |
|-----------------|---------------------|---|
| LD_LIBRARY_PATH | LD library path | /usr/lib:/lib:/apps/SUNW/spro/WS6U2/lib |
| PATH | Shell search path | /bin:/apps/SUNW/spro/WS6U2/bin:/usr/bin |
| TMPDIR | Temporary directory | |

Buttons: Set to Default, All to Default, Variable Help, OK, Cancel, Help

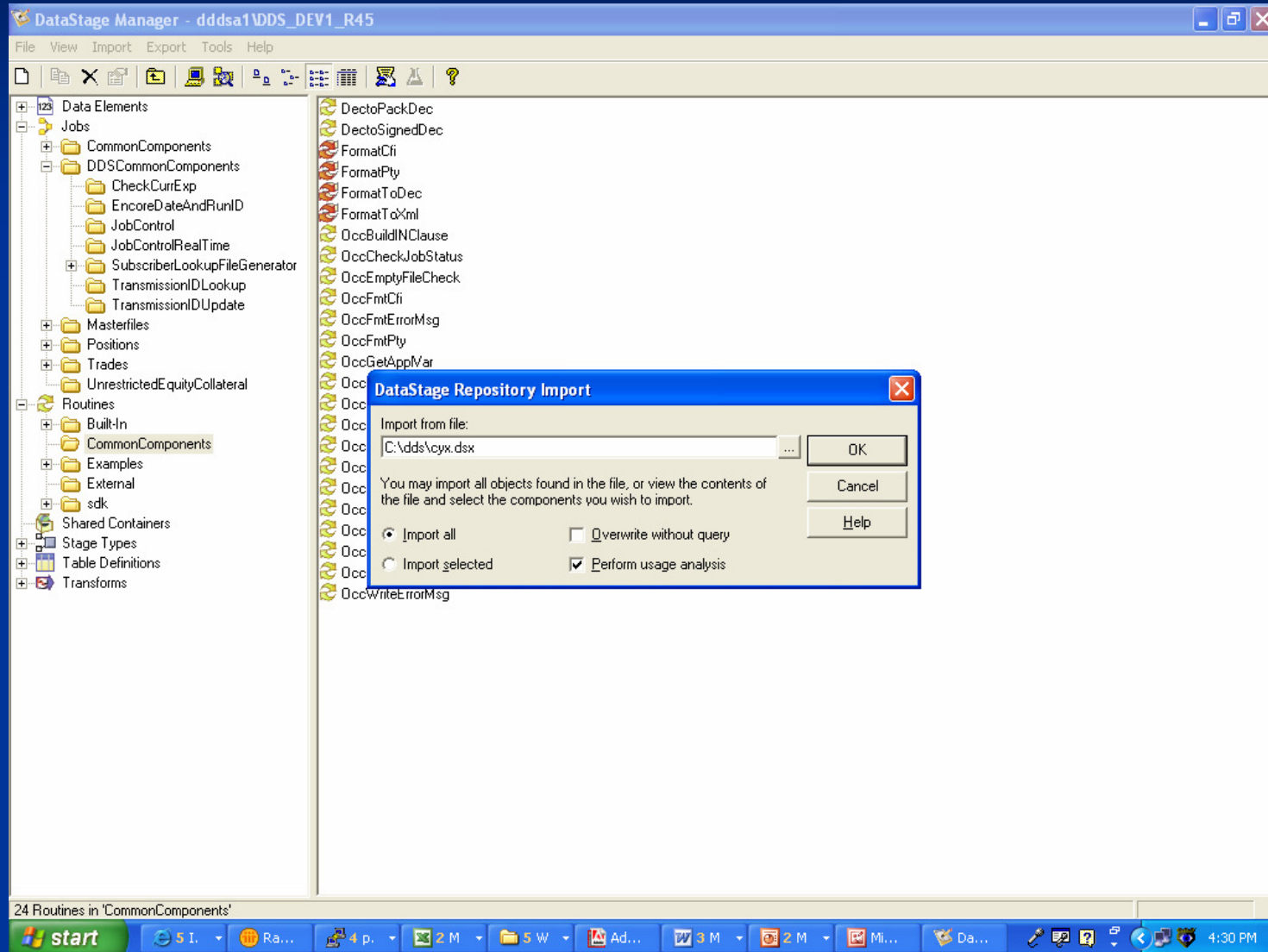
DataStage Manager

- Used to store and manage re-usable metadata for the jobs.
- Used to import and export components from file-system to Datastage projects.
- Primary interface to the DataStage Repository.
- Custom routines and transforms can also be created in the Manager

DataStage Routines (Manager window)



Importing / Exporting a Project



DataStage Designer

DataStage Designer is used to :

- Create DataStage Jobs that are compiled into executable programs.
- Design the jobs that extract, integrate, aggregate, load, and transform the data.
- Create and reuse metadata and job components
- Allows you to use familiar graphical point-and-click techniques to develop processes for extracting, cleansing, transforming, Integrating and loading data.

DataStage Designer

Use Designer to:

- Specify how data is extracted.
- Specify data transformations.
- Decode data going into the target tables using reference lookups
- Aggregate Data.
- Split data into multiple outputs on the basis of defined constraints

DataStage Designer

The Designer graphical interface lets you select

- Stage icons, drop them onto the Designer work area, and add links. Then, still working in the Designer, you define the required actions and processes for each stage and link.
- A job created with the Designer is easily scalable. This means that you can easily create a simple job, get it working, then insert further processing, additional data sources, and so on.

DataStage Terms and Concepts

| Term | Description |
|-------------------|---|
| Aggregator stage | A stage type that computes totals or other functions of sets of data. |
| BCPLoad stage | A plug-in stage supplied with DataStage that bulk loads data into a Microsoft SQL Server or Sybase table. |
| CFD | COBOL File Description. A text file that describes the format of a file in COBOL terms. |
| Column definition | Defines the columns contained in a data table. Includes the column name and the type of data contained in the column. |
| Container stage | A built-in stage type that represents a group of stages and links in a job design. |
| Data Browser | A tool used from within the DataStage Manager or DataStage Designer to view the content of a table or file. |

| | |
|------------------------|---|
| hashed file | A file that uses a hashing algorithm for distributing records in one or more groups on disk. |
| Hashed File stage | A stage that extracts data from or loads data into a database that contains hashed files. |
| job | A collection of linked stages, data elements, and transforms that define how to extract, cleanse, transform, integrate, and load data into a target database. See also mainframe job and server job . |
| Link Collector stage | A server job stage that collects previously partitioned data together. |
| Link Partitioner stage | A server job stage that allows you to partition data so that it can be processed in parallel on an SMP system. |
| meta data | Data about data. A table definition which describes the structure of the table is an example of meta data. |

| | |
|-----------------------|---|
| ODBC stage | A stage that extracts data from or loads data into a database that implements the industry standard Open Database Connectivity API. Used to represent a data source, an aggregation step, or a target data table. |
| plug-in stage | A stage that performs specific processing that is not supported by the Aggregator, Hashed File, ODBC, UniVerse, UniData, Sequential File, and Transformer stages. |
| Repository | A DataStage area where projects and jobs are stored as well as definitions for all standard and user-defined data elements, transforms, and stages. |
| Sequential File stage | A stage that extracts data from, or writes data to, a text file. |
| stage | A component that represents a data source, a processing step, or a data mart in a DataStage job. |
| table definition | A definition describing the data you want including information about the data table and the columns associated with it. Also referred to as meta data. |
| Transformer stage | A stage where data is transformed (converted) using transform functions. |

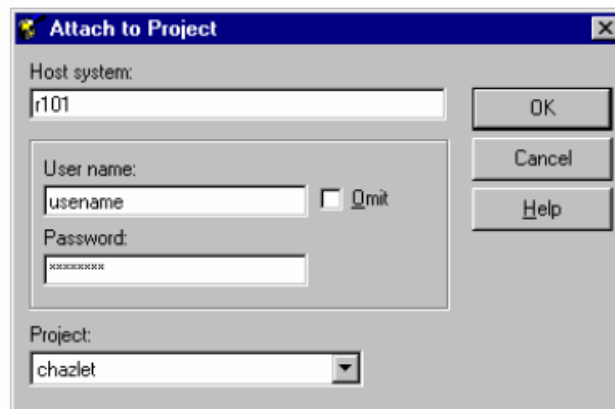
DataStage Client Login

1. Enter the name of your host in the **Host system** field. This is the name of the system where the DataStage server components are installed.
2. Enter your user name in the **User name** field. This is your user name on the server system.
3. Enter your password in the **Password** field.
4. Choose the project to connect to from the **Project** list. This list box displays all the projects installed on your DataStage server. At this point, you may only have one project installed on your system and this is displayed by default.
5. Select the **Save settings** check box to save your logon settings

DataStage Designer

Starting the DataStage Designer

Choose **Start** ► **Programs** ► **Ascential DataStage** ► **DataStage Designer** to run the DataStage Designer. The **Attach to Project** dialog box appears:



The screenshot shows a standard Windows-style dialog box titled "Attach to Project". It features a close button (X) in the top right corner. The dialog is organized into several sections:

- Host system:** A text input field containing "r101".
- User name:** A text input field containing "username", followed by an unchecked checkbox labeled "omit".
- Password:** A text input field containing "XXXXXXXX".
- Project:** A dropdown menu currently displaying "chazlet".

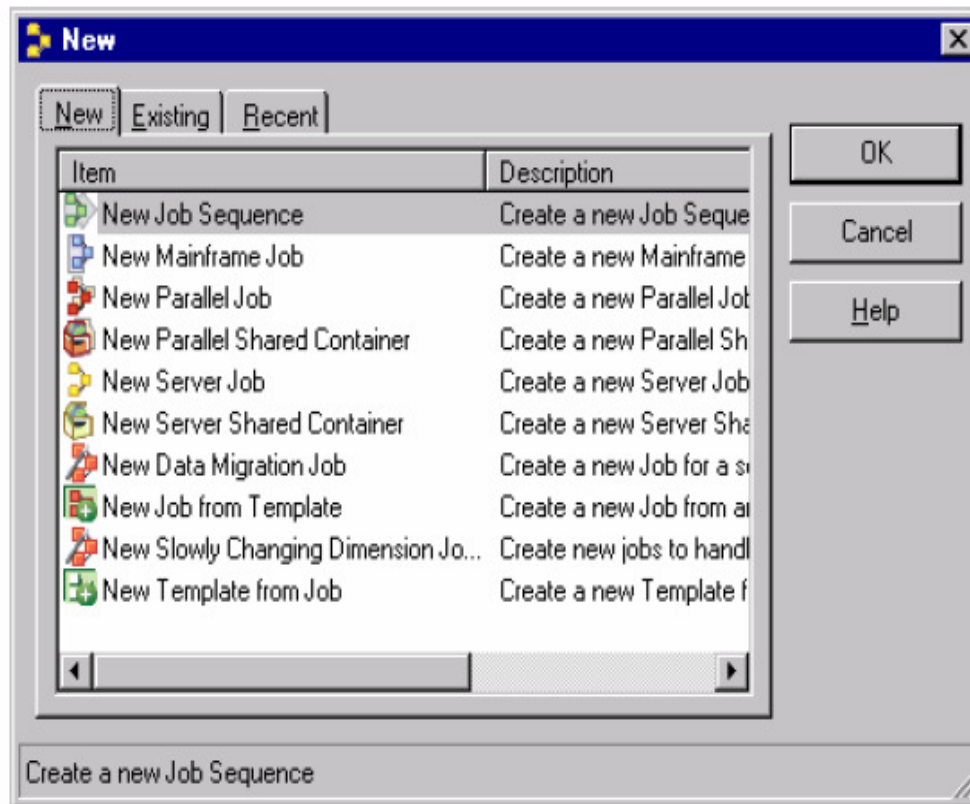
On the right side of the dialog, there are three buttons: "OK", "Cancel", and "Help".

Note: This dialog box appears when you start the DataStage Manager, Designer, or Director client components from the DataStage program folder. In all cases, you must attach to a project by entering your logon details.

To attach to a project:

Creating a New DataStage Job

6. Click **OK**. The **New** dialog box appears:



7. Click the **New Server Job** item, and click **OK**. The DataStage Designer window appears with a blank canvas:

The DataStage Designer Window

The DataStage Designer window consists of the following parts:

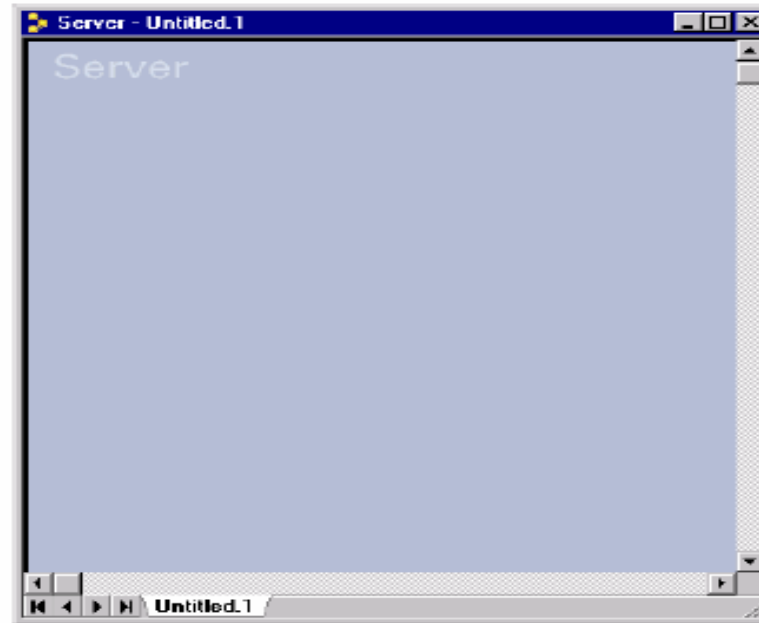
- One or more **Job** windows where you design your jobs.
- The **Property Browser** window where you view the properties of the selected job.
- The **Repository** window where you view components in a projects.
- A **Toolbar** from where you select Designer functions.
- A **Tool Palette** from which you select job components.
- A **Debug Toolbar** from where you select debug functions.
- A **Status Bar** which displays one-line help for the window components, and information on the current state of job operations, for example, compilation.

For full information about the Designer window, including the functions of the pull-down and shortcut menus, refer to the *DataStage Designer Guide*.

New Job Window Screen

The Job Window

The Job window is the main window where you design your DataStage jobs:

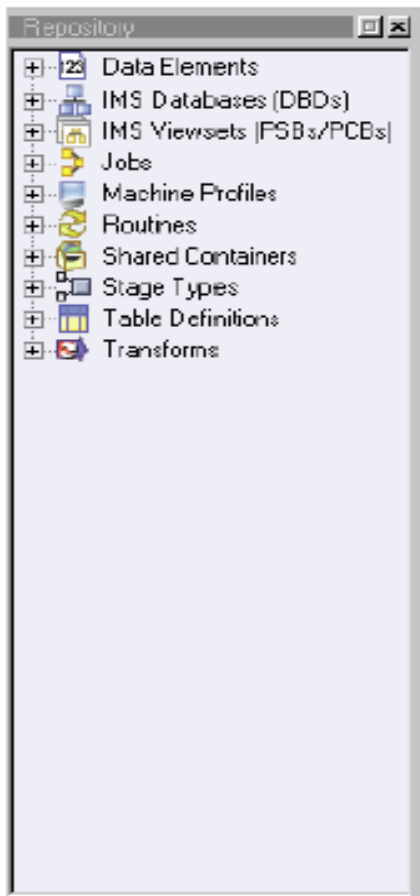


- Individual jobs are shown in separate windows, and you can have multiple jobs open at the same time.
- The title bar displays the job name and category.
- Grid lines in the window allow you to position stages precisely. You can turn the grid lines on or off, and you can have objects snap to the grid to make alignment easy.
- The tab(s) at the bottom of the window allows you to select a job when you are working with a Container Stage.

The Repository Window

The Repository Window

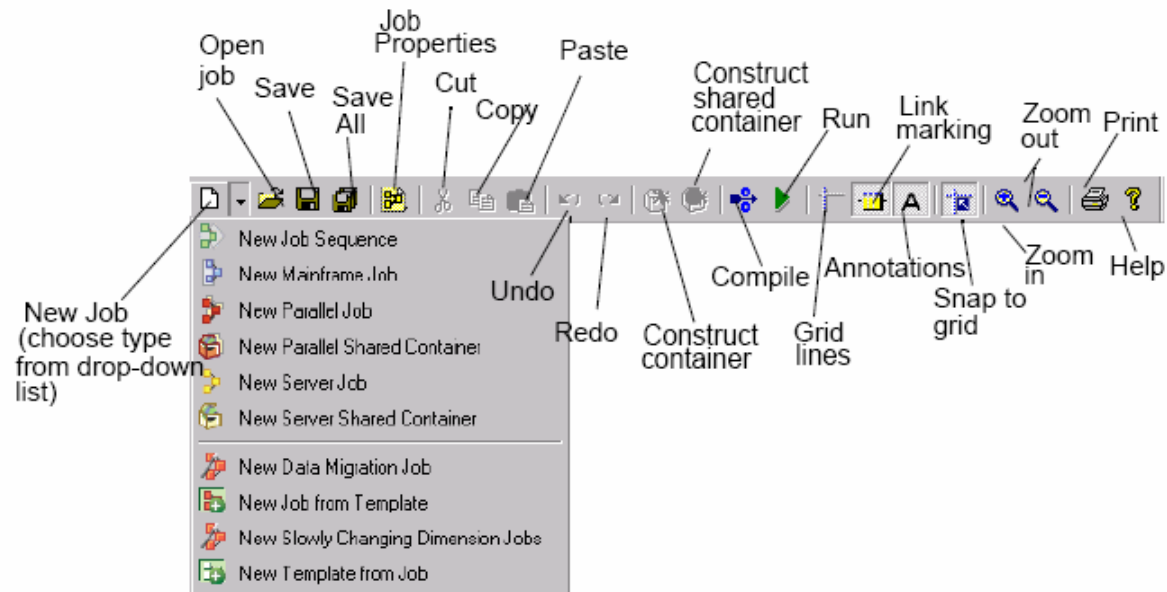
The Repository window displays a hierarchical tree of all the components of your project:



Designer Toolbar

The Designer Toolbar

The Designer toolbar contains the following buttons:



You can display ToolTips for the toolbar by letting the cursor rest on a button in the toolbar. The status bar then also displays an expanded description of that button's function.

The toolbar appears under the menu bar by default, but you can drag and drop it anywhere on the screen. If you move the toolbar to the edge of the Designer window, it attaches to the side of the window.

A Simple job

The screenshot displays the DataStage Designer interface. The main workspace shows a job named 'GetAllDates' with a simple flow: a 'PROCESS' icon connected to an 'ExtractOut' icon, which is then connected to a 'Sequential_File_10' icon. The 'ExtractOut' icon displays the text '1 rows, 0 rows/sec'. A floating window titled 'Server - GetAllDates (Multiple Instance)' shows the same flow on a 'Server' background. The left-hand 'Repository' pane shows a tree structure with folders like 'Jobs', 'CommonComponents', 'PORTFOLIOREVAL', 'Acquisition', 'AGGREGATION', 'MARGINABLEINTERFACE', and 'STARTOFDAY'. Below the repository is a 'Palette' with various tool icons under categories like 'Processing' and 'Real Time'. The bottom status bar shows 'Ready' and a taskbar with several open applications.

The Designer Tool Palette

- The tool palette contains buttons that represent the components you can add to your job designs.
- The palette has different groups to organize the tools available. Click the group title to open the group.
- The Favorites group allows you to drag frequently used tools there so you can access them quickly.
- You can also drag other items there from the Repository window, such as jobs and server shared containers:

DataStage Director

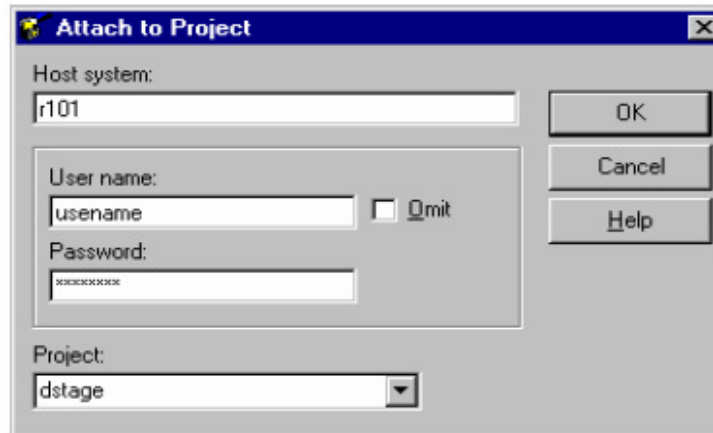
- The DataStage Director is the client component that validates, runs, schedules, and monitors jobs run by the DataStage Server.
- It is the starting point for most of the tasks a DataStage operator needs to do in respect of DataStage jobs.

DataStage Director

Starting the DataStage Director

To start the DataStage Director:

1. Choose **Start** ► **Programs** ► **Ascential DataStage** ► **DataStage Director**, or choose the appropriate program folder if you installed DataStage elsewhere. The **Attach to Project** dialog box appears:

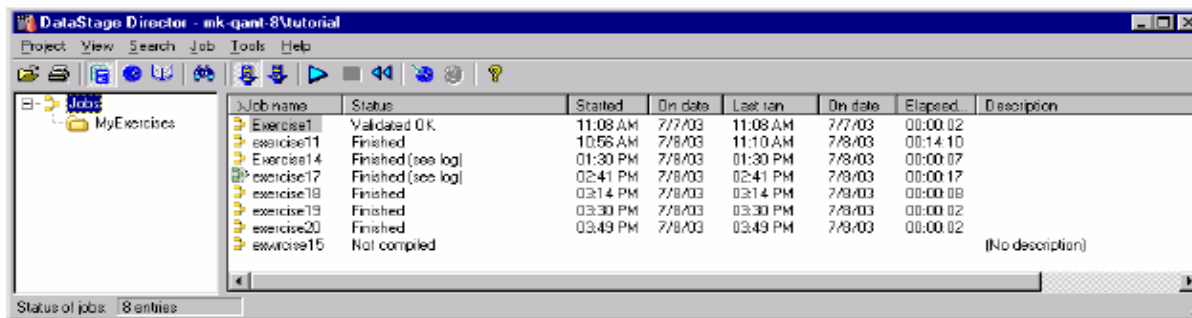


2. Enter the name of your host in the **Host system** field. This is the name of the system where the DataStage Server is installed.
3. Enter your user name in the **User name** field. This is your user name on the server system.
4. Enter your password in the **Password** field. If you are connecting to a Windows NT server via LAN Manager, you can select the **Omit** check box. The **User name** and **Password** fields gray out and you log on to the server using your current Windows account details.

DataStage Director [Contd.]

The DataStage Director Window

The DataStage Director window appears when you start the Director:



This section describes the features of the DataStage Director window including:

- The job category pane
- The display area
- The menu bar
- The toolbar
- The status bar

Job Category Pane

The left pane of the DataStage Director window is the job category pane. It displays the job category tree, which lists job categories and subcategories that contain server jobs. The jobs in the currently selected category are listed in the display area. You can hide the job category pane by choosing **View ► Show Categories**.

Display Area

The display area is the main part of the DataStage Director window. There are three views:

- **Job Status.** The default view, which appears in the right pane of the DataStage Director window. It displays the status of all jobs in the category currently selected in the job category tree. If you hide the job category pane, the Job Status view includes a Category column, and displays the status of all server jobs in the current project, regardless of their category.
- **Job Schedule.** Displays a summary of scheduled jobs and batches in the currently selected job category. If the job category pane is hidden, the display area shows all scheduled jobs and batches, regardless of their category.
- **Job Log.** Displays the log file for a job chosen from the Job Status view or the Job Schedule view.

Menu Bar

The menu bar has six pull-down menus that give access to all the functions of the Director:

- **Project.** Opens an alternative project and sets up printing.
- **View.** Displays or hides the toolbar, status bar, buttons, or job category pane, specifies the sorting order, changes the view, filters entries, shows further details of entries, and refreshes the screen.
- **Search.** Starts a text search dialog box.

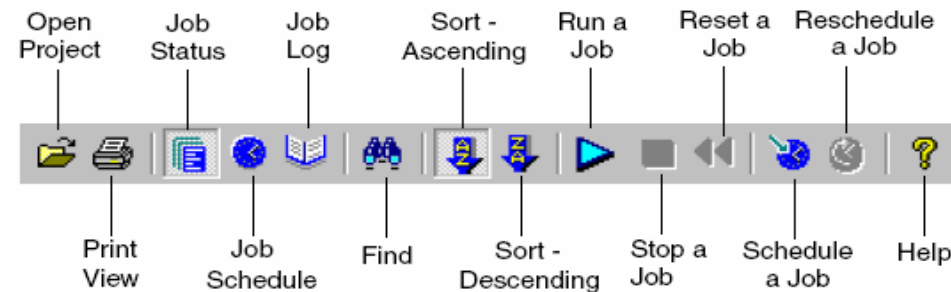
Menu Bar [Contd.]

- **Job.** Validates, runs, schedules, stops, and resets jobs, purges old entries from the job log file, deletes unwanted jobs, cleans up job resources (if the administrator has enabled this option), and allows you to set default job parameter values.
- **Tools.** Monitors running jobs, manages job batches, and starts the DataStage Designer and DataStage Manager. It also starts MetaStage Explorer and Quality Manager, if these components are installed on the system, and custom software.
- **Help.** Invokes the Help system. You can also get help from anyscreen or dialog box in the DataStage Director.

DataStage Director [Contd.]

Toolbar

The toolbar gives quick access to the main functions of the DataStage Director.



The toolbar is displayed by default, but can be hidden by choosing **View > Toolbar** or by changing the Director options. See "[Director Options](#)" on page 2-19 for more details. To display ToolTips, let the cursor rest on a button in the toolbar.

Status Bar

The status bar appears at the bottom of the DataStage Director window and displays the following information:

- The name of a job (if you are displaying the Job Log view).
- The number of entries in the display. If you look at the Job Status or Job Schedule view and use the **Filter Entries...** command, this panel specifies the number of lines that meet the filter criteria. If you have set a filter then (filtered) or (limited) is displayed.
- The date and time on the DataStage server.

Job States within Director

| Job State | Description |
|--------------------|---|
| Compiled | The job has been compiled but has not been validated or run since compilation. |
| Not compiled | The job is under development and has not been compiled successfully. |
| Running | The job is currently being run, reset, or validated. |
| Finished | The job has finished. |
| Finished (see log) | The job has finished but warning messages were generated or rows were rejected. View the log file for more details. |
| Stopped | The job was stopped by the operator. |
| Aborted | The job finished prematurely. |
| Validated OK | The job has been validated with no errors. |
| Has been reset | The job has been reset with no errors. |

Conclusions

- DataStage has proved to be an excellent ETL tool within the industry.
- The need for Data Integration and Consolidation within organizations is fuelling the need for DataStage.
- Data Transfer format landscape is gradually moving towards XML in every industry.
- PR3 Systems provides detailed training and consulting services for DataStage, Best Practices in DataStage Project Design and Helping organizations to consolidate their Data and Information Network.

Contact Us

For information about our training and consulting services, you can send an email to info@pr3systems.com or call 630-452-9883.