Data Warehousing, Analysis and Processing: A Brief Overview

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

Copyright PR3 Systems, 2013

Topics to be covered

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

Copyright PR3 Systems, 2013

PR3 systems

What 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 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.

Copyright PR3 Systems, 2013

systems

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

Copyright PR3 Systems, 2013

systems

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

Copyright PR3 Systems, 2013

systems

Careers in this Domain

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

Copyright PR3 Systems, 2013

systems

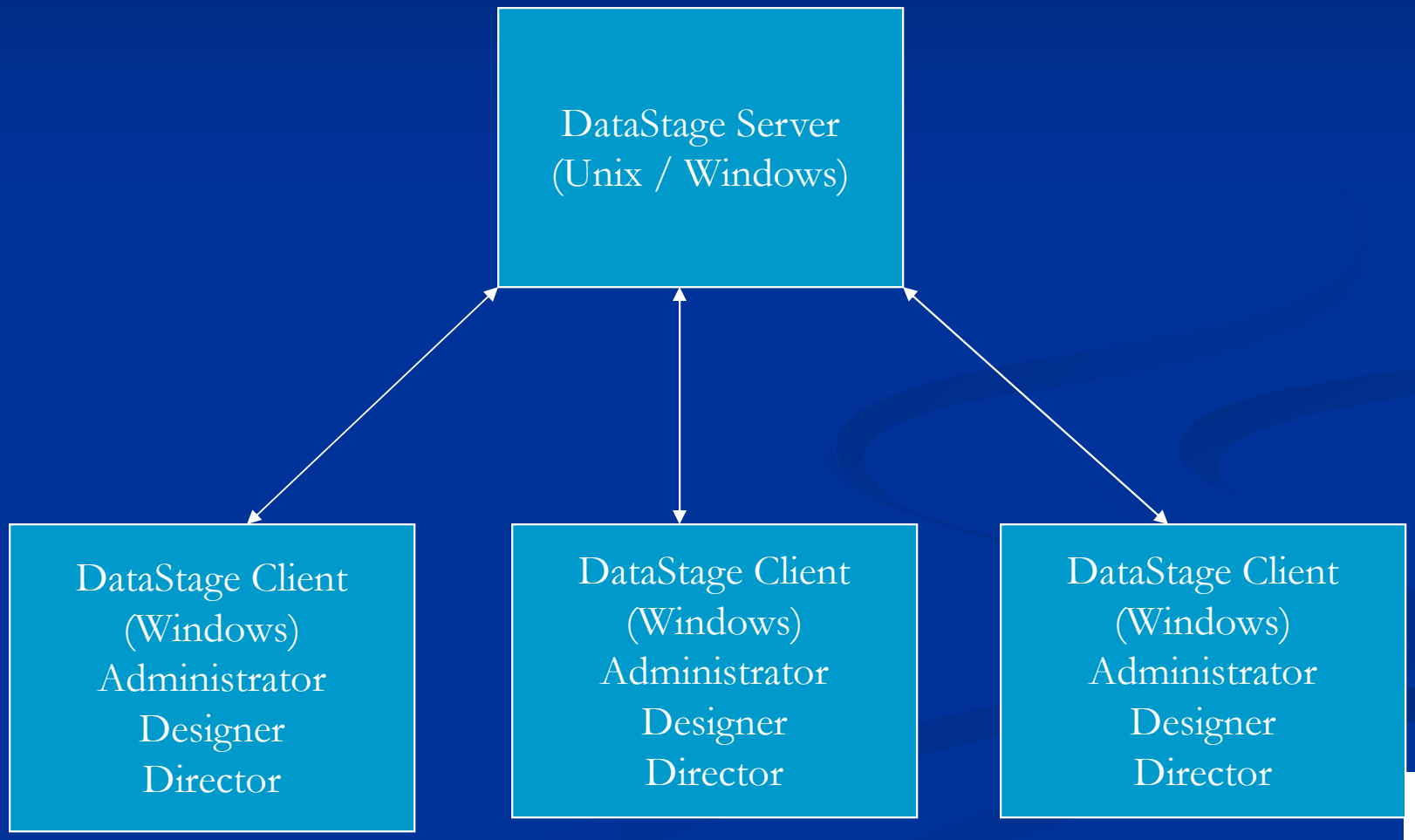
PR3 DataStage Training

- We specialize in providing education and consulting services for IBM's InfoSphere 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.

Copyright PR3 Systems, 2013

systems

DataStage Architecture



Copyright PR3 Systems, 2013

Product Overview

- DataStage is an IBM product used as the strategic ETL tool within many organizations.
- It can be used for multiple purposes:
  - Interfacing between multiple databases.
  - Changing data from one format to another. Eg. From a 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.

Copyright PR3 Systems, 2013

PR3 systems

DataStage Usage in Organizations

- DataStage has Windows Clients which connect to the server on the Unix or Windows platforms
- 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 / Windows servers

Copyright PR3 Systems, 2013

PR3 systems

Types of DataStage clients

- DataStage Administrator
- DataStage Designer
- DataStage Director

Copyright PR3 Systems, 2013

PR3 systems

DataStage Designer

- Used to design and compile DataStage jobs that extract, integrate, aggregate, load, and transform 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

Copyright PR3 Systems, 2013

PR3 systems

DataStage Designer

Use the 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 based on defined constraints

Copyright PR3 Systems, 2013

DataStage Designer

The Designer graphical interface lets you select:

- Stage icons, drop them onto the Designer work area, and add links. Then, you can 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.

Copyright PR3 Systems, 2013

PR3 systems

DataStage Terms and Concepts

Term	Description
Aggregator Stage	A stage that computes totals or other aggregation functions on sets of data.
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.
Connector Stage	A stage which offers the best performance and most functionality for accessing databases in a DataStage job.
Container	A reusable built-in DataStage component that represents a group of stages and links in a job design.
Data Browser	A tool used from within the DataStage Designer to view the contents of a table or file during design time.

Copyright PR3 Systems, 2013

PR3 systems

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 transformations that define how to extract, cleanse, transform, integrate, and load data into a target database.
Link Collector Stage	A server job stage that collects previously partitioned data.
Link Partitioner Stage	A server job stage that allows you to partition data so that it can be processed in parallel.
Metadata	Data about data. An example of metadata is a DataStage Table Definition which describes the structure of the table.

Copyright PR3 Systems, 2013

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 target database in a DataStage job.
Table Definition	A definition describing the data you want to integrate, including information about the data table and the columns associated with it. Also referred to as metadata.
Transformer Stage	A stage where data is transformed (converted) using transform functions.

Copyright PR3 Systems, 2013

PR3 systems

DataStage Client Login

1. Enter the name of your host in the **Host name of the services tier** field. This is the name of the system where the Application 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 first, you may only have one project installed on your system and this is displayed by default.

Copyright PR3 Systems, 2013

The DataStage Designer Window

The DataStage Designer window consists of the following parts:

- One or more **Job** windows where you design your jobs.
- The **Job Properties** 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 design components.

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

Copyright PR3 Systems, 2013

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 place frequently used components so you can access them quickly.
- You can also drag other components to the Favorites group from the Repository window, such as jobs and shared containers.

Copyright PR3 Systems, 2013

DataStage Director

- The DataStage Director is the client which is used to validate, run, schedule, and monitor jobs running on the DataStage Server.
- This is the starting point for most of the tasks a DataStage operator is required to do.

Copyright PR3 Systems, 2013

systems

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 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.

Copyright PR3 Systems, 2013

PR3 systems

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.

Copyright PR3 Systems, 2013

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. I
- **Help.** Invokes the help system. You can also get help from any screen or dialog box in the DataStage Director.

Copyright PR3 Systems, 2013

PR3 systems

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.

Copyright PR3 Systems, 2013

PR3 systems

Conclusions

- DataStage has proved to be an excellent ETL tool within the industry.
- The need for data integration and consolidation within organizations is fueling 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.

Copyright PR3 Systems, 2013

systems

Contact Us

For information about our training and consulting services, you can send an email to info@pr3systems.com or call 630-364-1469

Copyright PR3 Systems, 2013