

Copyright, Disclaimer of Warranties and Limitation of Liability

© Copyright IBM Corporation 2007

IBM Software Group
One Rogers Street
Cambridge, MA 02142

All rights reserved. Printed in the United States.

IBM and the IBM logo are registered trademarks of International Business Machines Corporation.

The following are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

| | | |
|---|-----------------------------------|---------------------------------|
| AnswersOnLine | DynamicServer, WorkgroupEdition | RedBrick Decision Server |
| AIX | Enterprise Storage Server | RedBrickMineBuilder |
| APPN | FFST/2 | RedBrickDecisionscape |
| AS/400 | Foundation.2000 | RedBrickReady |
| BookMaster | Illustra | RedBrickSystems |
| C-ISAM | Informix | RelyonRedBrick |
| Client SDK | InformixGL | S/390 |
| Cloudscape | InformixExtendedParallelServer | Sequent |
| Connection Services | InformixInternet Foundation.2000 | SP |
| Database Architecture | Informix RedBrick Decision Server | System View |
| DataBlade | JFoundation | Tivoli |
| DataJoiner | MaxConnect | TME |
| DataPropagator | MVS | UniData |
| DB2 | MVS/ESA | UniData&Design |
| DB2 Connect | Net.Data | UniversalDataWarehouseBlueprint |
| DB2 Extenders | NUMA-Q | UniversalDatabaseComponents |
| DB2 Universal Database | ON-Bar | UniversalWebConnect |
| Distributed Database | OnLineDynamicServer | UniVerse |
| Distributed Relational | OS/2 | VirtualTableInterface |
| DPI | OS/2 WARP | Visionary |
| DRDA | OS/390 | VisualAge |
| DynamicScalableArchitecture | OS/400 | WebIntegrationSuite |
| DynamicServer | PTX | WebSphere |
| DynamicServer.2000 | QBIC | |
| DynamicServer with Advanced DecisionSupportOption | QMF | |
| DynamicServer with Extended ParallelOption | RAMAC | |
| DynamicServer with UniversalDataOption | RedBrickDesign | |
| DynamicServer with WebIntegrationOption | RedBrickDataMine | |

Microsoft, Windows, Window NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java, JDBC, and all Java-based trademarks are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

All other product or brand names may be trademarks of their respective companies.

All information contained in this document has not been submitted to any formal IBM test and is distributed on an "as is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will result elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk. The original repository material for this course has been certified as being Year 2000 compliant.

This document may not be reproduced in whole or in part without the priori written permission of IBM.

Note to U.S. Government Users – Documentation related to restricted rights – Use, duplication, or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

Course Contents

| | |
|---|-----|
| • Mod 01: Introduction..... | 05 |
| • Mod 02: Deployment | 27 |
| • Mod 03: Administering DataStage..... | 41 |
| • Mod 04: DataStage Designer..... | 67 |
| • Mod 05: Creating Parallel Jobs..... | 91 |
| • Mod 06: Accessing Sequential Data..... | 127 |
| • Mod 07: Platform Architecture..... | 157 |
| • Mod 08: Combining Data..... | 195 |
| • Mod 09: Sorting and Aggregating Data | 235 |
| • Mod 10: Transforming Data | 259 |
| • Mod 11: Repository Functions | 287 |
| • Mod 12: Working With Relational Data..... | 311 |
| • Mod 13: Metadata in the Parallel Framework..... | 357 |
| • Mod 14: Job Control | 379 |

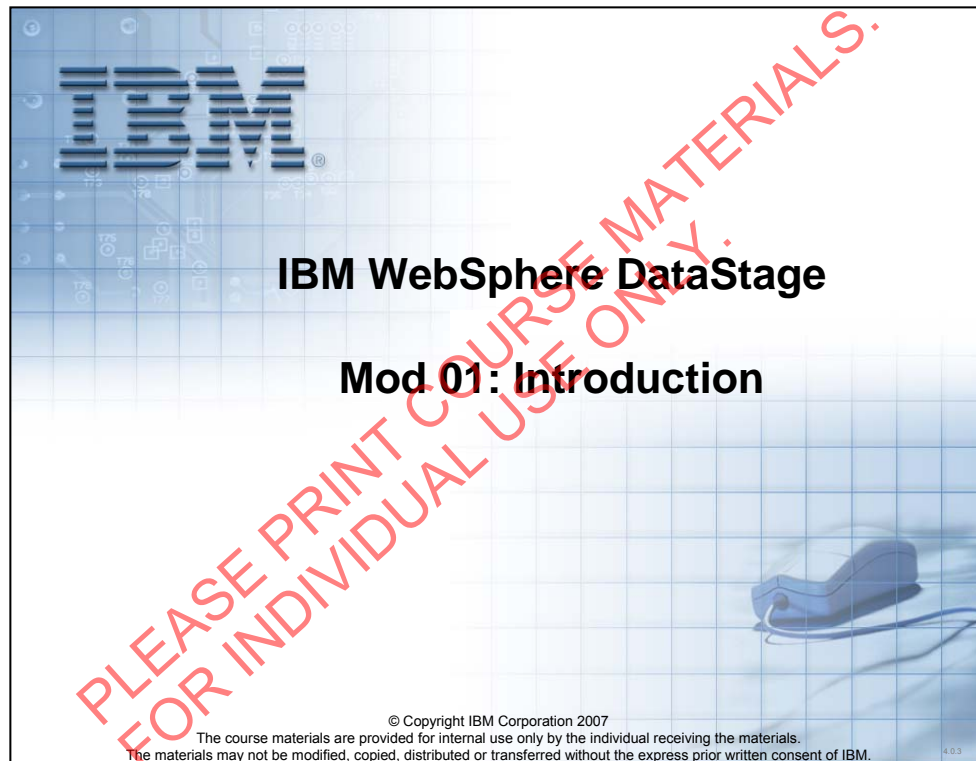
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

4



Unit objectives

After completing this unit, you should be able to:

- List and describe the uses of DataStage
- List and describe the DataStage clients
- Describe the DataStage workflow
- List and compare the different types of DataStage jobs
- Describe the two types of parallelism exhibited by DataStage parallel jobs

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

6

Notes:

What is IBM WebSphere DataStage?

- Design jobs for Extraction, Transformation, and Loading (ETL)
- Ideal tool for data integration projects – such as, data warehouses, data marts, and system migrations
- Import, export, create, and manage metadata for use within jobs
- Schedule, run, and monitor jobs, all within DataStage
- Administer your DataStage development and execution environments
- Create batch (controlling) jobs

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

7

DataStage is a comprehensive tool for the fast, easy creation and maintenance of data marts and data warehouses. It provides the tools you need to build, manage, and expand them. With DataStage, you can build solutions faster and give users access to the data and reports they need.

With DataStage you can:

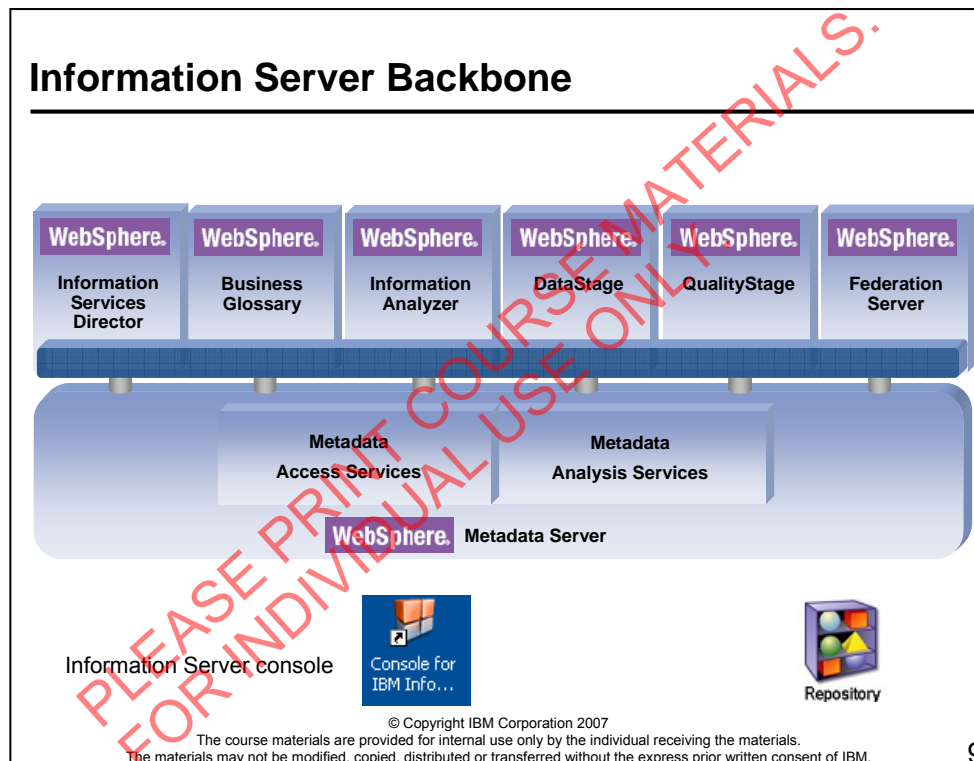
- Design the jobs that extract, integrate, aggregate, load, and transform the data for your data warehouse or data mart.
- Create and reuse metadata and job components.
- Run, monitor, and schedule these jobs.
- Administer your development and execution environments.

IBM Information Server

- Suite of applications, including DataStage, that:
 - [Share a common repository](#)
 - DB2, by default
 - [Share a common set of application services and functionality](#)
 - Provided by Metadata Server components hosted by an application server
 - [IBM WebSphere Application Server](#)
 - Provided services include:
 - [Security](#)
 - [Repository](#)
 - [Logging and reporting](#)
 - [Metadata management](#)
- Managed using web console clients
 - [Administration console](#)
 - [Reporting console](#)

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.



Information Server Administration Console

IBM Information Server - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Print

Address <http://hawkvm:9080/index.jsp> Go Links

IBM Information Server

Home Administration Glossary Information Services Catalog Reporting Help About Change Password Log Out

Navigation

- Contents
- Domain Management
- Session Management
- Users and Groups
 - Users
 - Groups
- Log Management
- Scheduling Management

Users

Select Users to Work With

Search First Name Search Last Name Find Clear Search

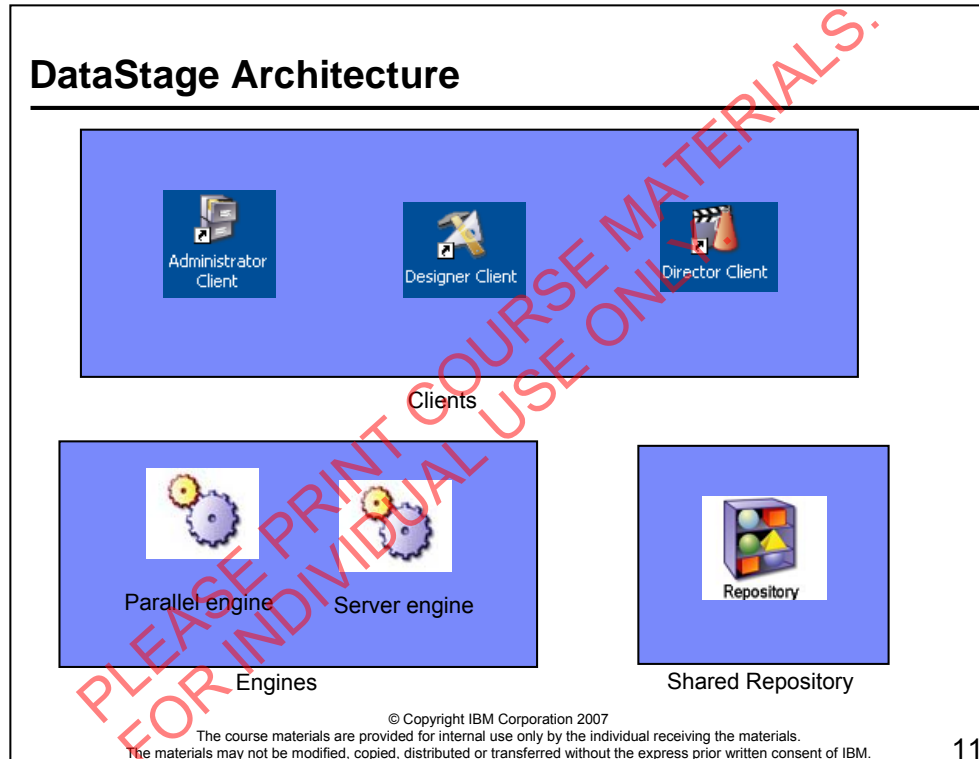
Additional Search Criteria

Item 1 of 2 out of 2 Items per Page: 25 50 100

| | Last Name | First Name | User Name | Title | Business Phone | Location | |
|--------------------------|-----------|------------|-----------|-------|----------------|----------|------------------|
| <input type="checkbox"/> | admin | admin | admin | | | | New Assign Roles |
| <input type="checkbox"/> | appserv | appserv | appserv | | | | Open |
| | | | | | | | Delete |

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.



The top half displays the Clients. Below there are two engines: The Server engine that runs DataStage server jobs and the parallel engine that runs parallel jobs. Our focus in this course is on Parallel jobs.

The DataStage client components are:

Administrator

Administers DataStage projects and conducts housekeeping on the server

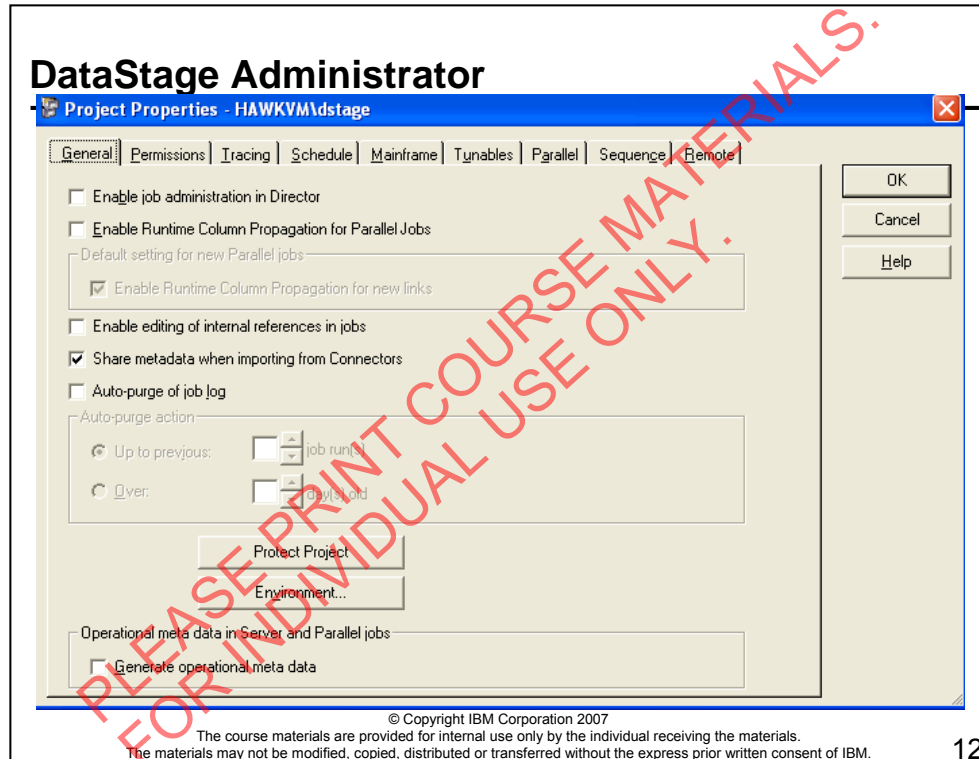
Designer

Creates DataStage jobs that are compiled into executable programs

Director

Used to run and monitor the DataStage jobs

The Repository is used to store DataStage objects. The Repository is shared with other applications in the Suite.

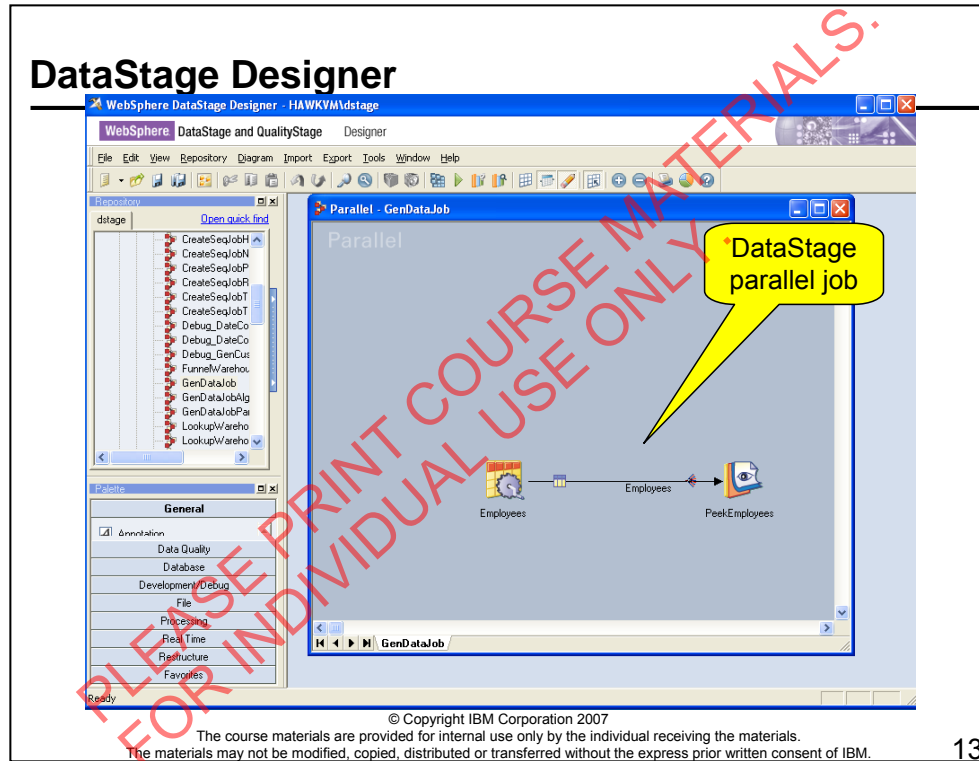


12

Use the Administrator to specify general server defaults, add and delete projects, and to set project properties. The Administrator also provides a command interface to the DataStage repository.

Use the Administrator **Project Properties** window to:

- Set job monitoring limits and other Director defaults on the **General** tab.
- Set user and group privileges on the **Permissions** tab.
- Enable or disable server-side tracing on the **Tracing** tab.
- Specify a user name and password for scheduling jobs on the **Schedule** tab.
- Specify parallel job defaults on the **Parallel** tab.
Specify Job Sequencer defaults on the **Sequence** tab.



The **DataStage Designer** allows you to use familiar graphical point-and-click techniques to develop job flows for extracting, cleansing, transforming, integrating and loading data into target files and tables.

The Designer provides a “visual data flow” method to easily interconnect and configure reusable components.

DataStage Director

WebSphere DataStage Director - HAWKVMdstage

Project View Search Job Tools Help

| > Occurred | > On date | Type | Event |
|------------|------------|---------|---|
| 2:56:09 PM | 11/10/2006 | Control | Starting Job GenDataJob. |
| 2:56:28 PM | 11/10/2006 | Info | Environment variable settings: (...) |
| 2:56:28 PM | 11/10/2006 | Info | Parallel job initiated |
| 2:56:45 PM | 11/10/2006 | Info | main_program: IBM WebSphere DataStage Enterprise Edition 8.0.0 (...) |
| 2:57:10 PM | 11/10/2006 | Info | main_program: orchgeneral: loaded (...) |
| 2:57:27 PM | 11/10/2006 | Info | main_program: APT configuration file: C:/IBM/InformationServer/Server/... |
| 2:57:27 PM | 11/10/2006 | Info | main_program: Step execution finished with status = OK. |
| 2:57:28 PM | 11/10/2006 | Info | main_program: Startup time, 0:42; production run time, 0:00. |
| 2:57:29 PM | 11/10/2006 | Info | Contents of phantom output file (...) |
| 2:57:29 PM | 11/10/2006 | Info | Parallel job reports successful completion |
| 2:57:30 PM | 11/10/2006 | Control | Finished Job GenDataJob. |

Log for job: GenDataJob 12 entries (filtered) Server time: 11/10/2006 02:57 PM

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Use the Director to validate, run, schedule, and monitor your DataStage jobs. You can also gather statistics as the job runs.

Developing in DataStage

- Define global and project properties in Administrator
- Import metadata into the Repository
- Build job in Designer
- Compile job in Designer
- Run and monitor job log messages in Director
 - Jobs can also be run in Designer, but job log messages cannot be viewed in Designer

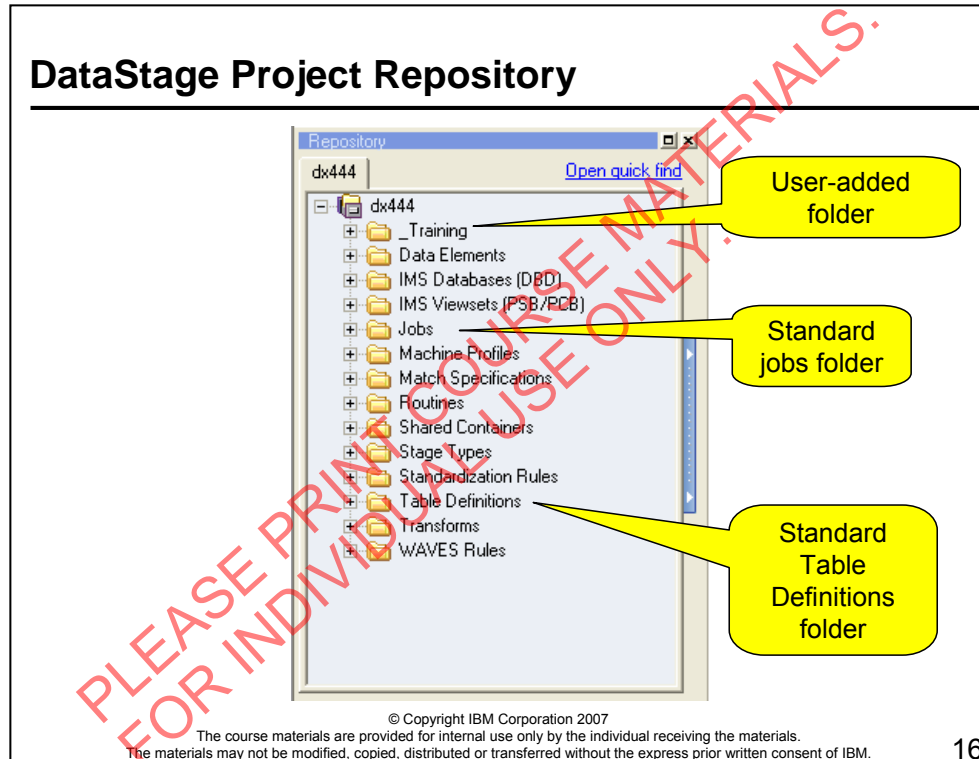
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

15

Development Workflow:

- Define your project's properties: Administrator
- Open (attach to) your project
- Import metadata that defines the format of data stores your jobs will read from or write to
- Design the job: Designer
 - Define data extractions (reads)
 - Define data flows
 - Define data integration
 - Define data transformations
 - Define data constraints
 - Define data loads (writes)
 - Define data aggregations
- Compile and debug the job: Designer
- Run and monitor the job: Director



All your work is stored in a DataStage *project*. Before you can do anything, other than some general administration, you must open (attach to) a project.

Projects are created during and after the installation process. You can add projects after installation on the **Projects** tab of Administrator.

A project is associated with a *directory*. The project directory is used by DataStage to store your jobs and other DataStage objects and metadata on your server.

You must open (attach to) a project before you can do any work in it.

Projects are self-contained. Although multiple projects can be open at the same time, they are separate environments. You can, however, *import* and *export* objects between them.

Multiple users can be working in the same project at the same time. However, DataStage will prevent multiple users from editing the same DataStage object (job, Table Definition, etc.) at the same time.

Types of DataStage Jobs

- Parallel jobs
 - Executed by the DataStage parallel engine
 - Built-in functionality for pipeline and partition parallelism
 - Compiled into OSH (Orchestrate Scripting Language)
 - OSH executes Operators
 - Executable C++ class instances
 - Runtime monitoring in DataStage Director
- Job sequences (batch jobs, controlling jobs)
 - Master Server jobs that kick-off jobs and other activities
 - Can kick-off Server or Parallel jobs
 - Runtime monitoring in DataStage Director
 - Executed by the Server engine
- Server jobs
 - Executed by the DataStage server engine
 - Compiled into Basic
 - Runtime monitoring in DataStage Director

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

17

This course focuses on parallel jobs and job sequences that control parallel jobs. However, job sequences can include both server and parallel jobs in the same sequence.

Design Elements of Parallel Jobs

- Stages
 - Implemented as OSH operators (pre-built components)
 - Passive stages (E and L of ETL)
 - Read data
 - Write data
 - E.g., Sequential File, DB2, Oracle, Peek stages
 - Processor (active) stages (T of ETL)
 - Transform data
 - Filter data
 - Aggregate data
 - Generate data
 - Split / Merge data
 - E.g., Transformer, Aggregator, Join, Sort stages
- Links
 - “Pipes” through which the data moves from stage to stage

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

18

Pipeline Parallelism



- Transform, clean, load processes execute simultaneously
- Like a conveyor belt moving rows from process to process
 - Start downstream process while upstream process is running
- Advantages:
 - Reduces disk usage for staging areas
 - Keeps processors busy
- Still has limits on scalability

© Copyright IBM Corporation 2007
 The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Partition Parallelism

- Divide the incoming stream of data into subsets to be separately processed by an operator
 - Subsets are called partitions
- Each partition of data is processed by the same operator
 - E.g., if operation is Filter, each partition will be filtered in exactly the same way
- Facilitates near-linear scalability
 - 8 times faster on 8 processors
 - 24 times faster on 24 processors
 - This assumes the data is evenly distributed

© Copyright IBM Corporation 2007

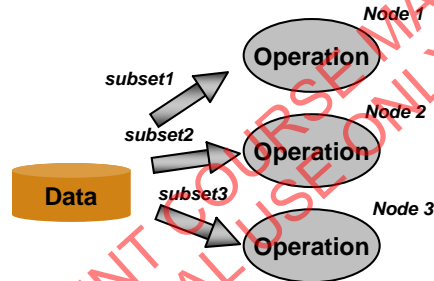
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

20

Partitioning breaks a dataset into smaller sets. This is a key to scalability. However, the data needs to be evenly distributed across the partitions; otherwise, the benefits of partitioning are reduced.

It is important to note that what is done to each partition of data is the same. How the data is processed or transformed is the same.

Three-Node Partitioning



- Here the data is partitioned into three partitions
- The operation is performed on each partition of data separately and in parallel
- If the data is evenly distributed, the data will be processed three times faster

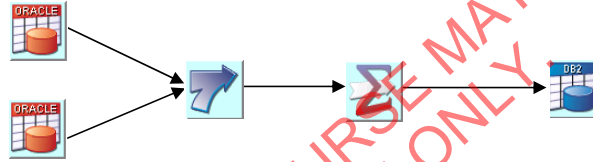
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

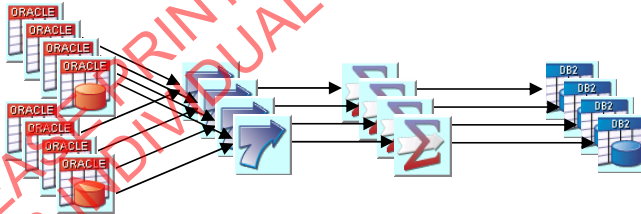
21

Job Design v. Execution

User designs the flow in DataStage Designer



... at runtime, this job runs in parallel for any configuration or partitions (called nodes)



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

22

Much of the parallel processing paradigm is hidden from the programmer. The programmer simply designates the process flow, as shown in the upper portion of this diagram. The parallel engine, using definitions in a configuration file, will actually execute processes that are partitioned and parallelized, as illustrated in the bottom portion.

Checkpoint

1. True or false: DataStage Designer is used to build and compile your ETL jobs
2. True or false: User Director to monitor your job during execution
3. True or false: Administrator is used to set global and project properties

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

23

Notes:

Write down your answers here:

1.

2.

3.

Checkpoint solutions

1. True.
2. True.
3. True.

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

24

Unit summary

Having completed this unit, you should be able to:

- List and describe the uses of DataStage
- List and describe the DataStage clients
- Describe the DataStage workflow
- List and compare the different types of DataStage jobs
- Describe the two types of parallelism exhibited by DataStage parallel jobs

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

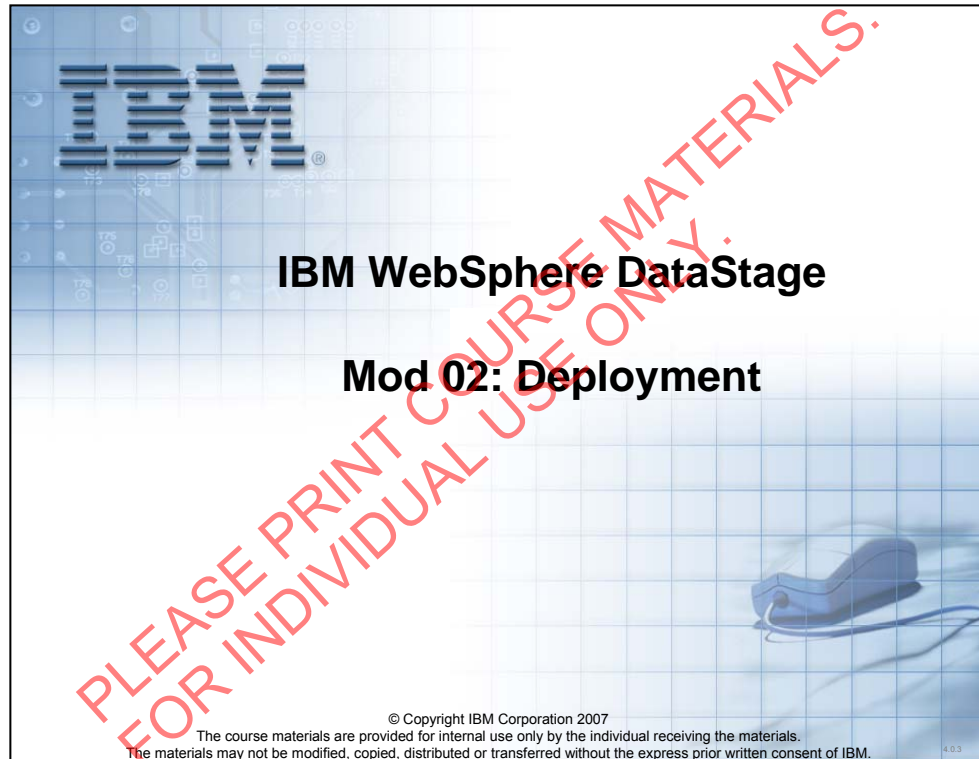
25

Notes:

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

26



Unit objectives

After completing this unit, you should be able to:

- Identify the components of Information Server that need to be installed
- Describe what a deployment domain consists of
- Describe different domain deployment options
- Describe the installation process
- Start the Information Server

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

28

Notes:

What Gets Deployed

An Information Server domain, consisting of the following:

- Metadata Server, hosted by an IBM WebSphere Application Server instance
- One or more DataStage servers
 - DataStage server includes both the parallel and server engines
- One DB2 UDB instance containing the Repository database
- Information Server clients
 - Administration console
 - Reporting console
 - DataStage clients
 - Administrator
 - Designer
 - Director
- Additional Information Server applications
 - Information Analyzer
 - Business Glossary
 - Rational Data Architect
 - Information Services Director
 - Federation Server

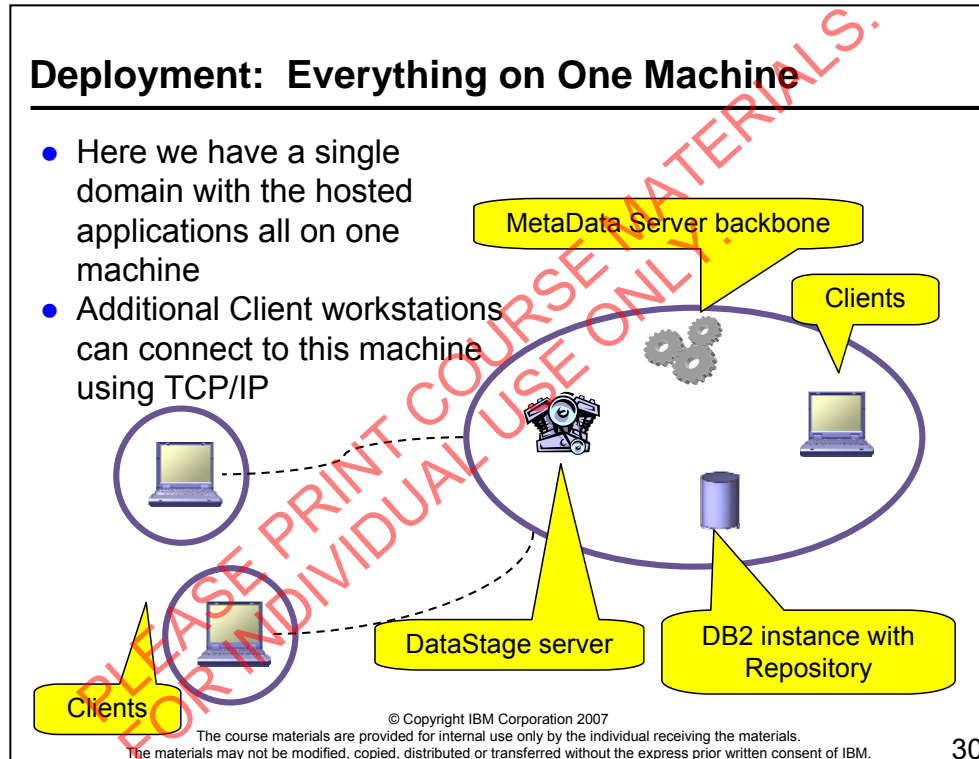
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

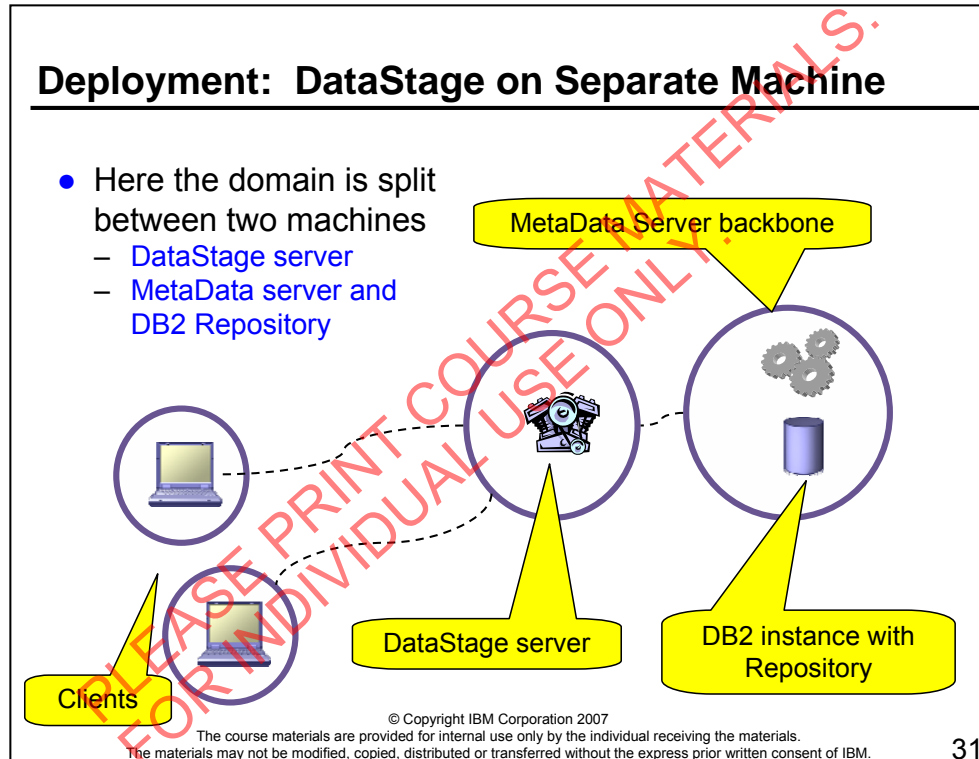
29

Database servers other than DB2 can be configured and used. DB2, however, is available on the installation image.

Which additional applications are available for installation depends on which have been licensed.

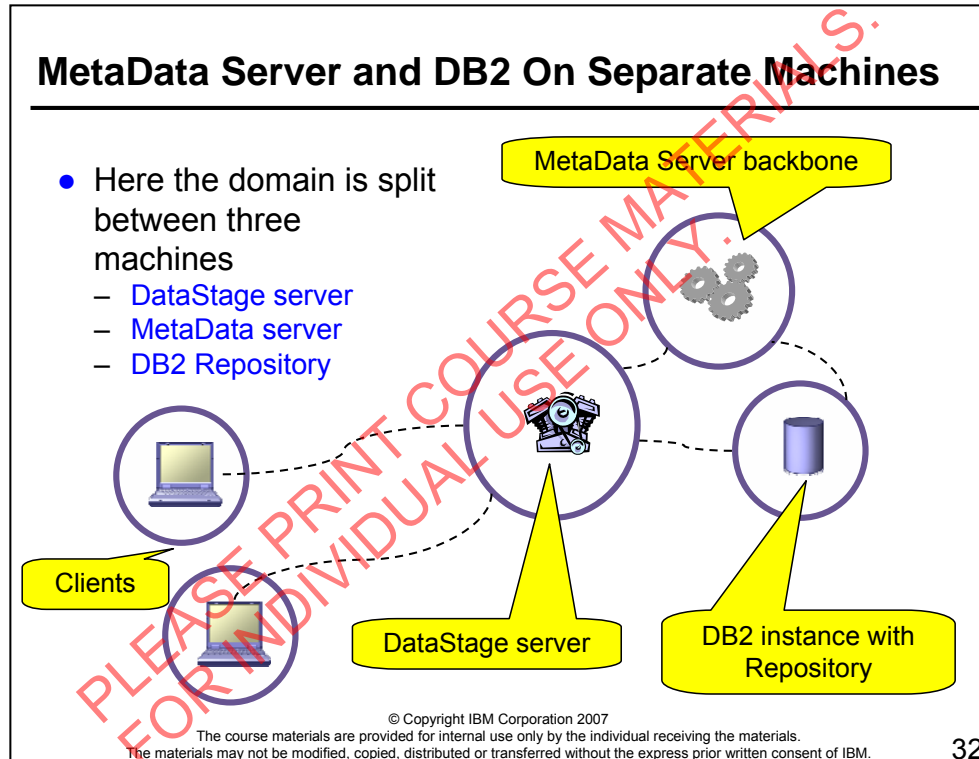


Not shown here and in later slides is the possibility of additional DataStage player-node machines connected to the DataStage server machine using a high-speed network. For simplicity in the diagram only the conductor node machine is displayed.



Additional DataStage servers can be part of this domain, but they would have to be on separate machines.

Here the Database is depicted as DB2, but additional databases are supported.



Additional DataStage Servers can be part of this domain, but they would have to be separate from one another.

Information Server Installation

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

33

Installation Configuration Layers

- Configuration layers include:
 - [Client](#)
 - DataStage and Information Server clients
 - [Engine](#)
 - DataStage and other application engines
 - [Domain](#)
 - MetaData Server and hosted metadata server components
 - Installed products domain components
 - [Repository](#)
 - Repository database server and database
 - [Documentation](#)
- Selected layers are installed on the machine local to the installation
- Already existing components can be configured and used
 - E.g., DB2, WebSphere Application Server

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

34

Selected configuration layers determine what gets installed on the machine local to the installation. If your domain is spread across multiple machines, you will select just the configuration types that are to be installed on the machine you are currently installing on.

Information Server Start-Up

- **Start the MetaData Server**
 - From Windows Start menu, click “Start the Server” after the profile to be used (e.g., “default”)
 - From the command line, open the profile bin directory
 - Enter “startup server1”
 - > server1 is the default name of the application server hosting the MetaData Server
- **Start the ASB agent**
 - From Windows Start menu, click “Start the agent” after selecting the Information Server folder
 - Only required if DataStage and the MetaData Server are on different machines
- **To begin work in DataStage, double-click on a DataStage client icon**
- **To begin work in the Administration and Reporting consoles, double-click on the Web Console for Information Server icon**

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

35

By default upon installation in Windows, the application server and ASB agent are set to start upon system startup.

Starting the MetaData Server Backbone

Application Server Profiles folder

Profile

Start the Server

Profile

Profile \bin directory

```
D:\Hawk_Install>cd D:\IBM\WebSphere\AppServer\profiles\default\bin
D:\IBM\WebSphere\AppServer\profiles\default\bin>startServer server1
ADMU0116I: Tool information is being logged in file
D:\IBM\WebSphere\AppServer\profiles\default\logs\startServer.log
ADMU0128I: Starting tool with the default profile
ADMU3100I: Reading configuration for server: server1
ADMU3200I: Server launched. Waiting for initialization status.
ADMU3000I: Server server1 open for e-business; process id is 2832
D:\IBM\WebSphere\AppServer\profiles\default\bin>
```

Startup command

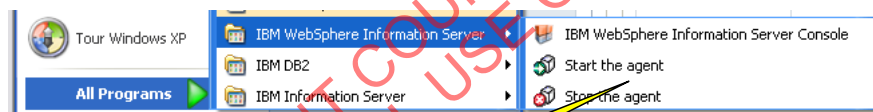
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

36

Starting the ASB Agent



Start the agent

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

37

Checkpoint

1. What application components make up a domain?
2. Can a domain contain multiple DataStage servers?
3. Does the DB2 instance and the Repository database need to be on the same machine as the Application Server?
4. Suppose DataStage is on a separate machine from the Application Server. What two components need to be running before you log onto DataStage?

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

38

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. Metadata server hosted by the Application Server. One or more DataStage servers. One DB2/UDB instance containing the Suite Repository database.
2. Yes. The DataStage servers must be on separate machines. They can be on different platforms, e.g., one server running on Windows and another running on Linux
3. No. The DB2 instance with the Repository can reside on a separate machine/platform than the Application Server
4. The Application Server and the ASB agent

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

39

Unit summary

Having completed this unit, you should be able to:

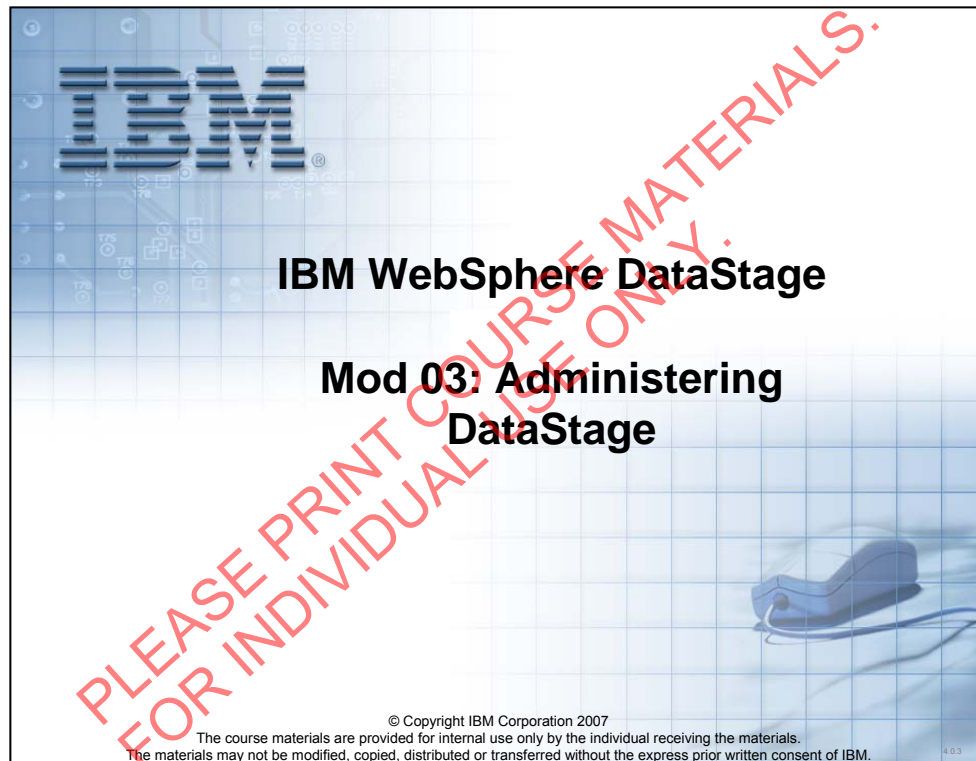
- Identify the components of Information Server that need to be installed
- Describe what a deployment domain consists of
- Describe different domain deployment options
- Describe the installation process
- Start the Information Server

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

40

Notes:



Unit objectives

After completing this unit, you should be able to:

- Open the Administrative console
- Create new users and groups
- Assign Suite roles and Product roles to users and groups
- Give users DataStage credentials
- Log onto DataStage Administrator
- Add a DataStage user on the Permissions tab and specify the user's role
- Specify DataStage global and project defaults
- List and describe important environment variables

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

42

Notes:

Information Server Administration Web Console

- Web application for administering Information Server
- Use for:
 - Domain management
 - Session management
 - Management of users and groups
 - Logging management
 - Scheduling management

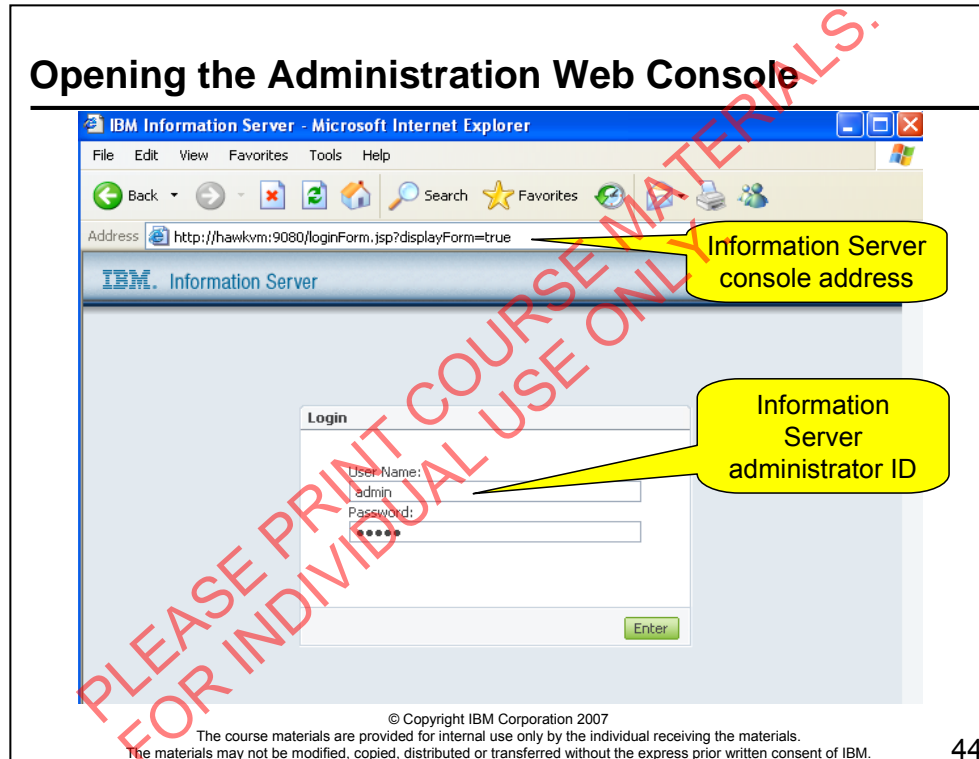
PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

43

Our focus in this course is on the management of users and groups and domain management.



To open the Administrative Web Console, select Web Console for IBM Information Server Web Console.

The console address is of the form: `http://machine:nnnn`.

Here **machine** is the host name of the machine running the application server that hosts MetaData Server.

nnnn is the port address of the console. By default, it is 9080.

The Information Server administrator ID and password is specified during installation. After installation, new administrator IDs can be specified.

Users and Group Management

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

45

User and Group Management

- Suite authorizations can be provided to users or groups
 - Users that are members of a group acquire the authorizations of the group
- Authorizations are provided in the form of roles
 - Two types of roles
 - Suite roles: Apply to the Suite
 - Suite Component roles: Apply to a specific product or component of Information Server, e.g., DataStage
- Suite roles
 - Administrator
 - Perform user and group management tasks
 - Includes all the privileges of the Suite User role
 - User
 - Create views of scheduled tasks and logged messages
 - Create and run reports
- Suite Component roles
 - DataStage
 - DataStage user
 - Permissions are assigned within DataStage
 - > Developer, Operator, Super Operator, Production Manager
 - DataStage administrator
 - Full permissions to work in DataStage Administrator, Designer, and Director
 - And so on, for all products in the Suite

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

46

A user ID that is assigned Suite roles can immediately log onto the Information Server console.

What about a user ID that is assigned a DataStage Suite Component role? If the user ID is assigned the DataStage Administrator role, then the user will immediately acquire the DataStage Administrator permission for all projects.

If the user ID is assigned the DataStage User role, one more step is required for any DataStage project that user will log onto. A DataStage Administrator must assign a Developer role to that user ID on the Permissions tab. This is done within DataStage, as we will see later.

Creating a DataStage User ID

Administration console

Create new user

Users

| | Last Name | First Name | User Name | Title | Business Phone | Location | |
|--------------------------|-----------|------------|-----------|-------|----------------|----------|---------------------------------------|
| <input type="checkbox"/> | admin | admin | admin | | | | New Assign Roles Open Delete |
| <input type="checkbox"/> | appserv | appserv | appserv | | | | |
| <input type="checkbox"/> | demohawk | demohawk | demohawk | | | | |

© Copyright IBM Corporation 2007
 The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

47

The process of creating a new group is similar to creating a new user. Users assigned to a group inherit the authorizations assigned to the group.

Assigning DataStage Roles

Users

User ID

Assign Suite Administrator role

Assign Suite User role

Assign DataStage Administrator role

| Role | Inherited |
|--|------------------------------------|
| <input type="checkbox"/> Suite Administrator | <input type="checkbox"/> |
| <input type="checkbox"/> Suite User | <input type="checkbox"/> |
| Suite Component | |
| <input type="checkbox"/> Role | <input type="checkbox"/> Inherited |
| <input type="checkbox"/> Business Glossary Administrator | <input type="checkbox"/> |
| <input type="checkbox"/> Business Glossary Author | <input type="checkbox"/> |
| <input type="checkbox"/> Business Glossary User | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> DataStage and QualityStage Administrator | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> DataStage and QualityStage User | <input type="checkbox"/> |
| <input type="checkbox"/> Information Analyzer Data Administrator | <input type="checkbox"/> |

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

48

DataStage Credential Mapping

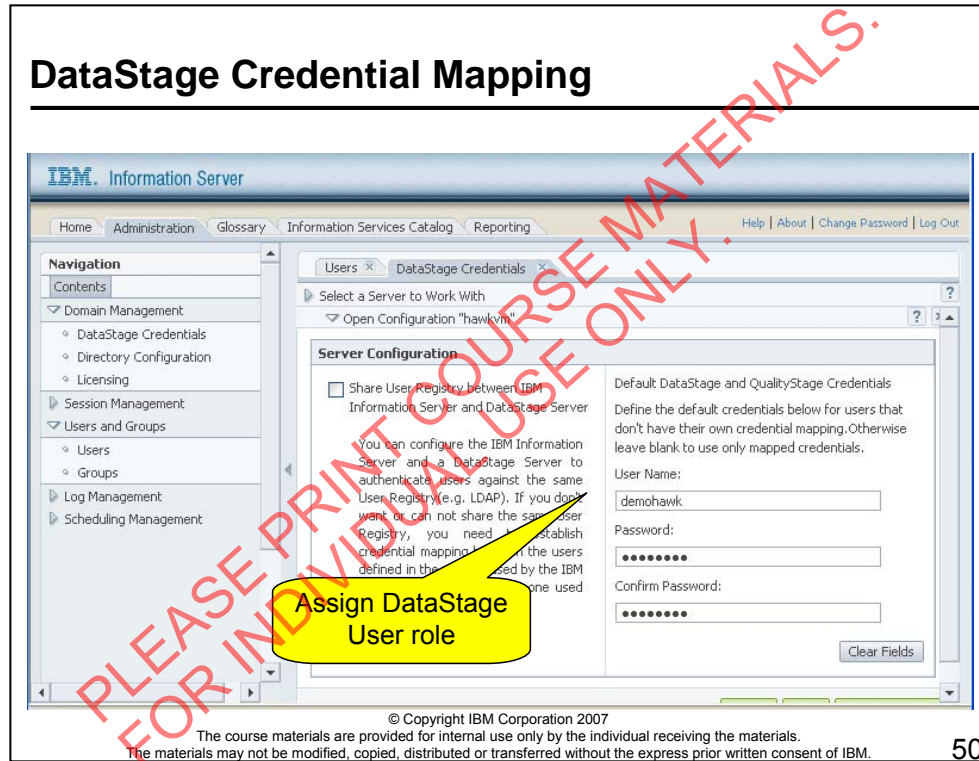
- DataStage credentials
 - Required by DataStage in order to log onto a DataStage client
 - Managed by the DataStage Server operating system or LDAP
- Users given DataStage Suite roles in the Suite Administration console do not automatically receive DataStage credentials
 - Users need to be mapped to a user who has DataStage credentials on the DataStage Server machine
 - This DataStage user must have file access permission to the DataStage engine/project files or Administrator rights on the operating system
 - Even user IDs that have DataStage credentials must be so mapped
 - Map the user ID to itself

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

49

This assumes that when the DataStage Server was installed, the style of user registry selected for the installation was “Internal User Registry.”



50

All Suite users without their own DataStage credentials will be mapped to this user ID and password. Here the user name and password are demohawk / demohawk. demohawk is assumed to be a valid user on the DataStage Server machine and has file permissions on the DataStage engine and projects directories.

Suite users can also be mapped individually to specific users.

Note that demohawk need not be a Suite administrator or user.

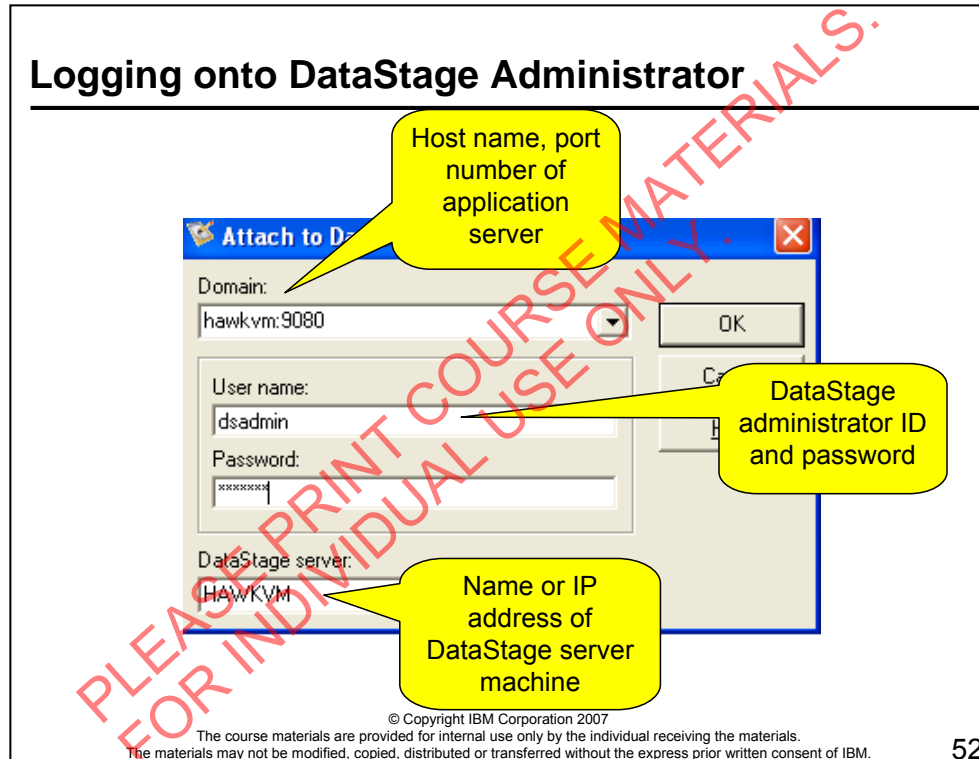
DataStage Administrator

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

51

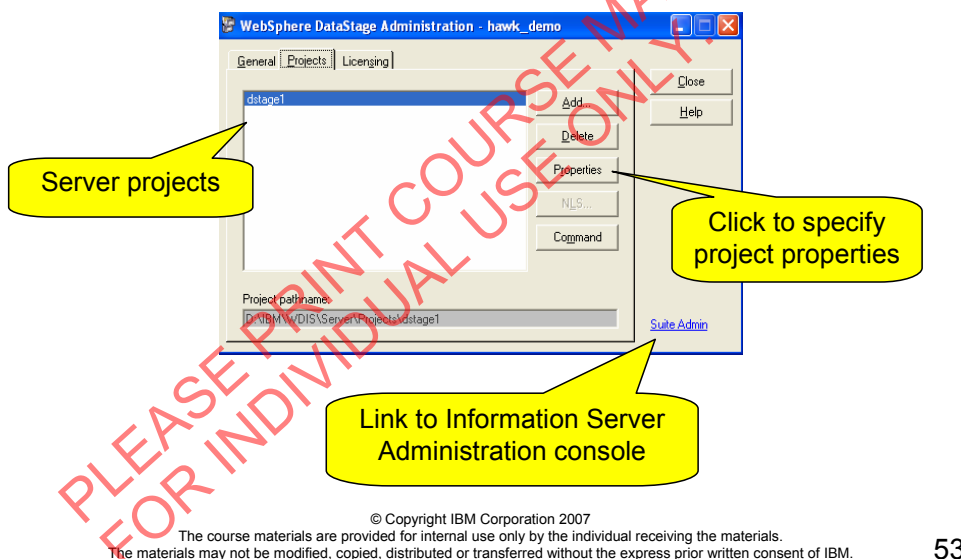


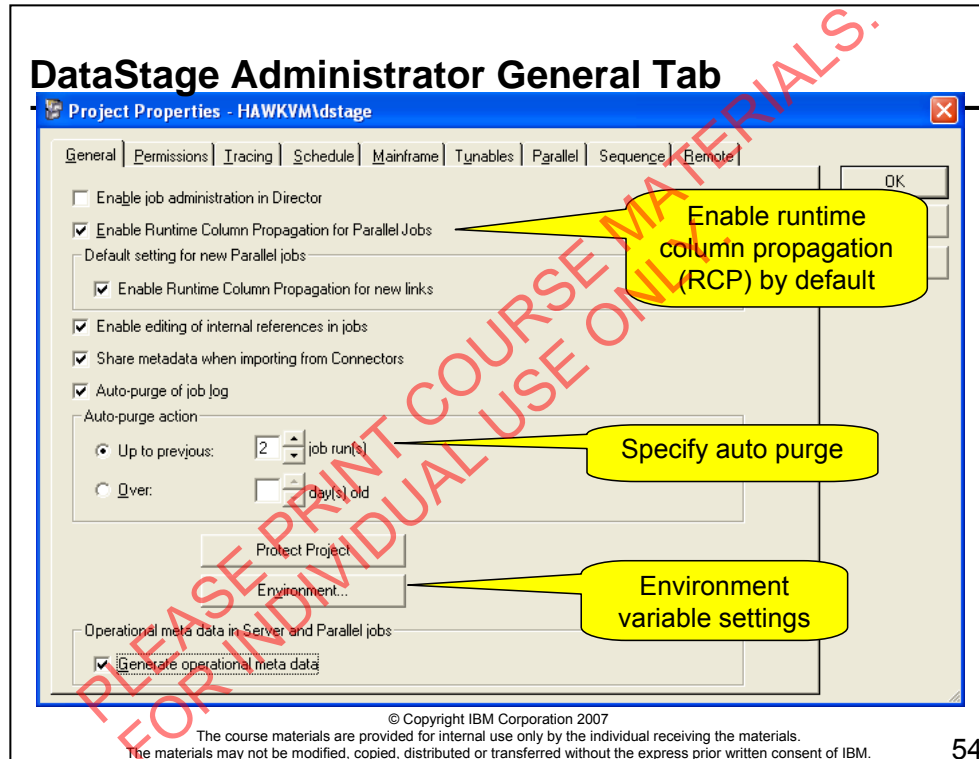
This port is also used when logging into the Information Server Administration console.

Recall that multiple DataStage servers can exist in a domain, although they must be on different machines. Here you select the server that has the DataStage projects you want to administer.

The user ID requires a DataStage Administrator or DataStage User product role. There are some limits as to what a DataStage User role provides. For example, a DataStage User cannot delete projects and cannot set permissions.

DataStage Administrator Projects Tab



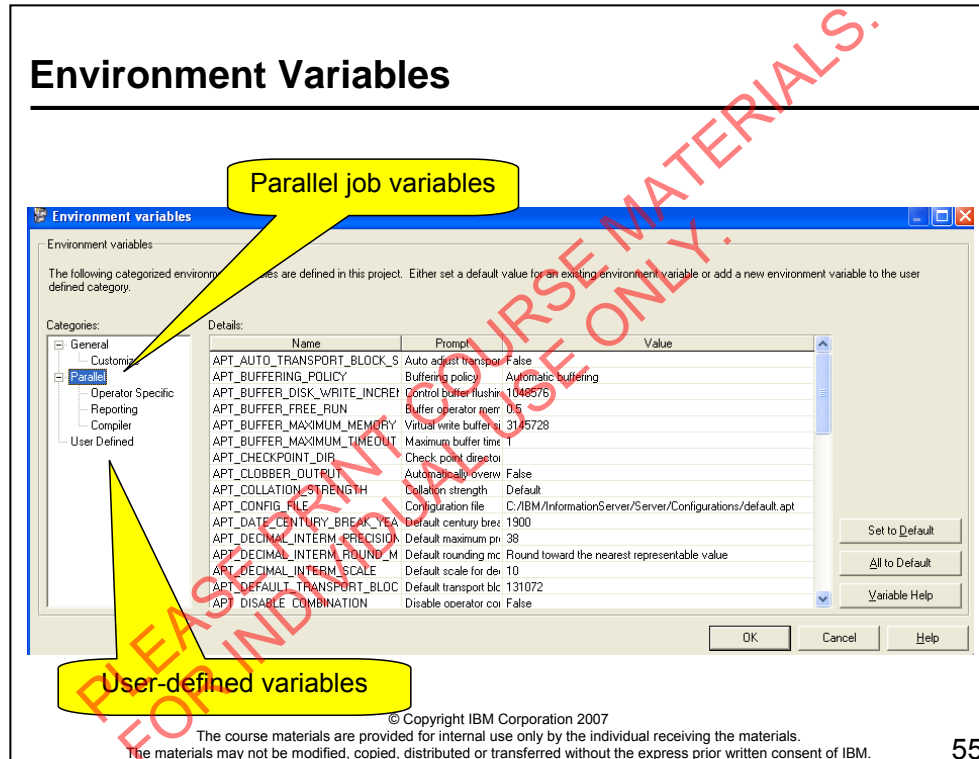


54

Runtime Column Propagation (RCP) is discussed in a later module.

A running job generates many messages in the Director job log. Here you can specify that you want old messages purged.

Click the Environment button to display environment variables settings.

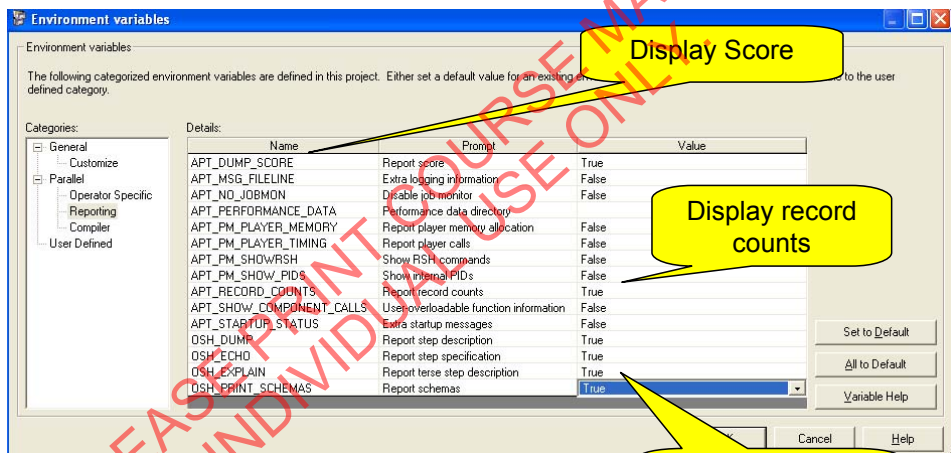


Click the Environment button on the General tab to specify environment variables. The variables listed under the Parallel branch apply to Parallel jobs.

You can also specify your own environment variables under the User Defined branch. These variables can be passed to jobs through their job parameters to provide project level job defaults.

There are also other environment variables that are hidden from the GUI. See the Parallel Job Advanced Developers Guide documentation for a list of the environment variables.

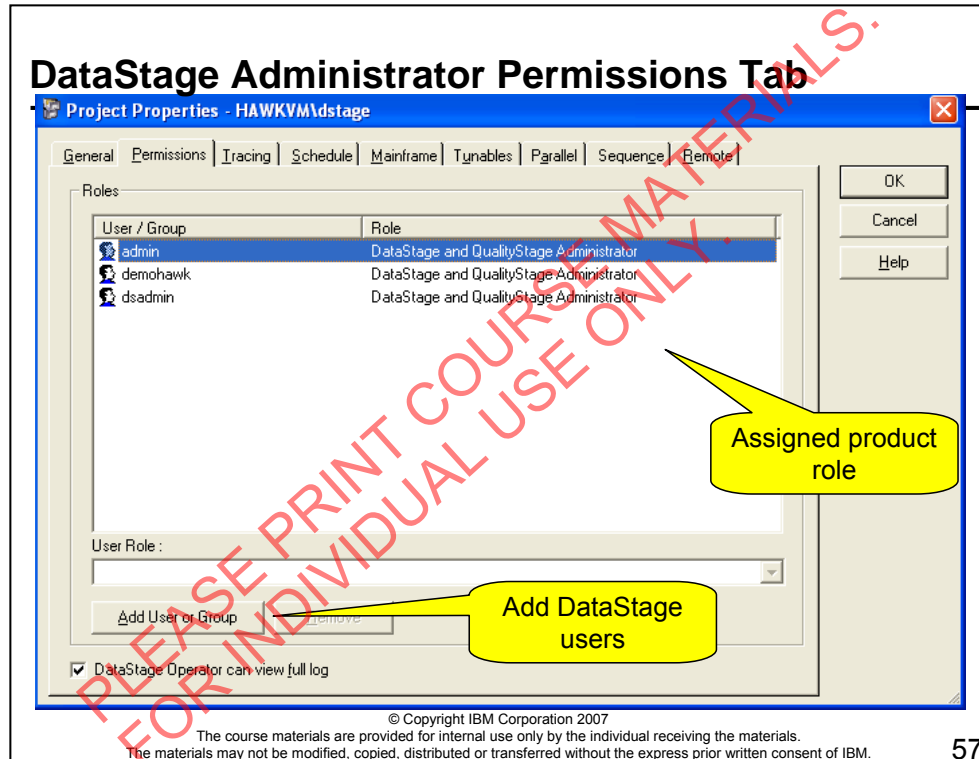
Environment Reporting Variables



© Copyright IBM Corporation 2007
 The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

56

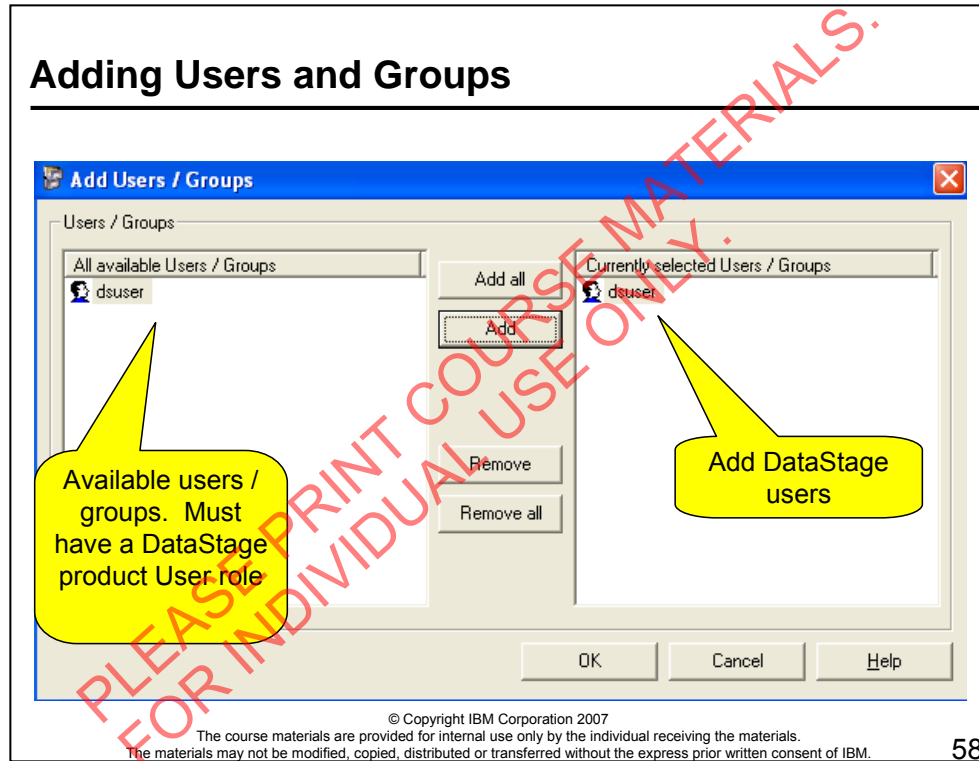
The Score and OSH are discussed in a later module.



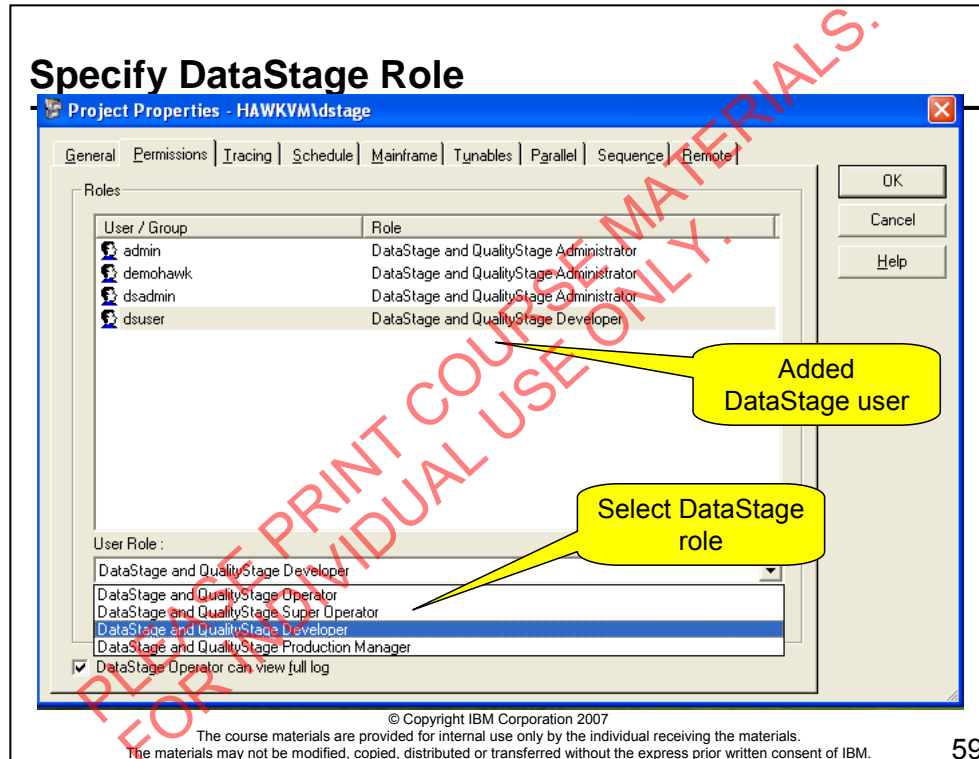
This lists Suite users and groups that have a DataStage User or Administrator role.

When Suite users or groups that have a DataStage Administrator role are added, they automatically they are automatically entered here and assigned the role of "DataStage Administrator".

Suite users or groups that have a DataStage User role need to be manually added. To accomplish this, click the Add User or Group button (which is hidden in this screen shot). Then you need to select the DataStage user role (Operator, Super Operator, Developer, Production Manager) that this user ID is to have.



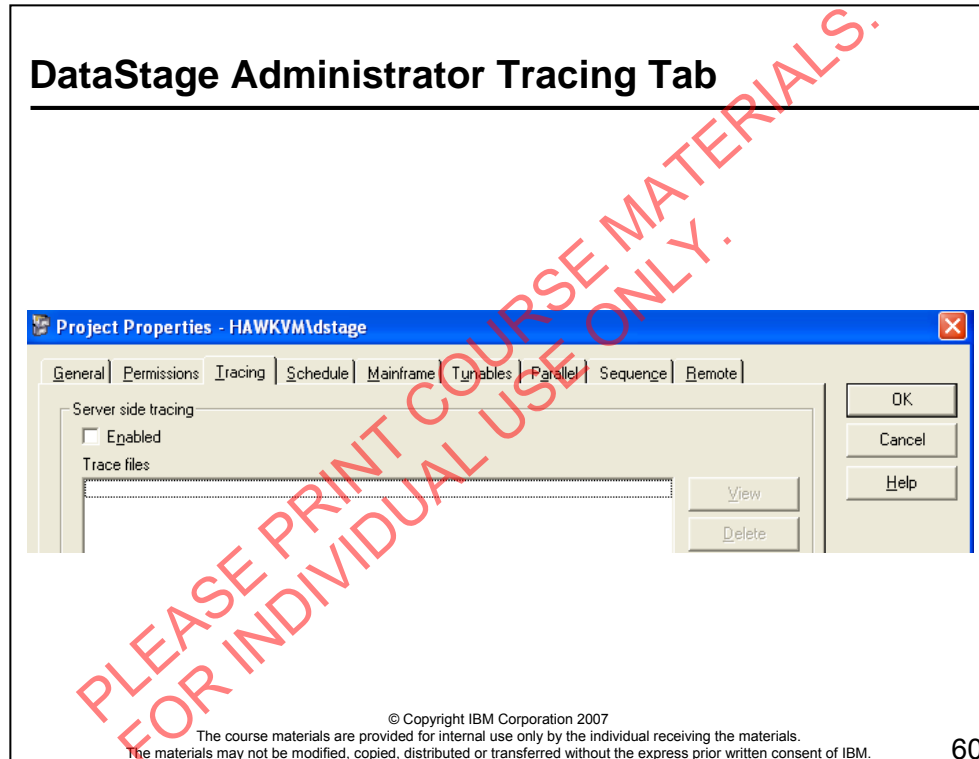
Click the “Add User or Group” button to open this window. On the left are Information Server users and groups that have been given a DataStage User role.



Use this page to set user group permissions for accessing and using DataStage. All DataStage users must belong to a recognized *user role* before they can log on to DataStage. This helps to prevent unauthorized access to DataStage projects.

There are four roles that can be assigned to a DataStage user:

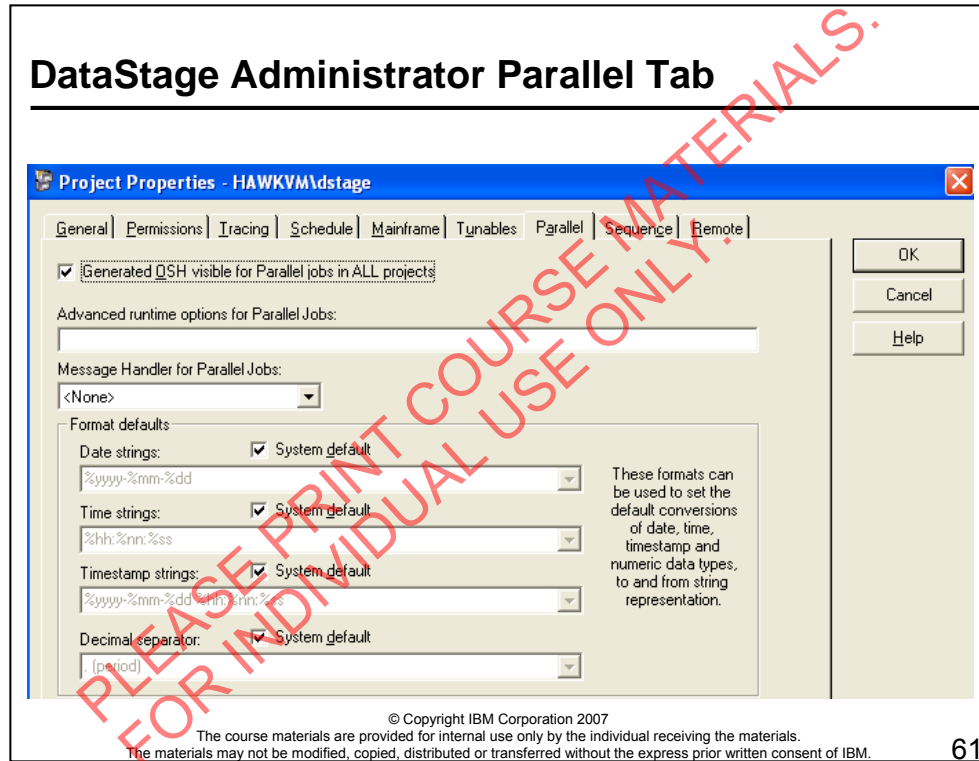
- *DataStage Developer*, who has full access to all areas of a DataStage project.
- *DataStage Operator*, who can run and manage released DataStage jobs.
- *DataStage Super Operator*, who can open Designer and view the Repository in a read-only mode.
- *DataStage Production Manager*, who has create and manipulate protected projects.



This tab is used to enable and disable server-side tracing.

The default is for server-side tracing to be disabled. When you enable it, information about server activity is recorded for any clients that subsequently attach to the project. This information is written to trace files. Users with in-depth knowledge of the system software can use it to help identify the cause of a client problem. If tracing is enabled, users receive a warning message whenever they invoke a DataStage client.

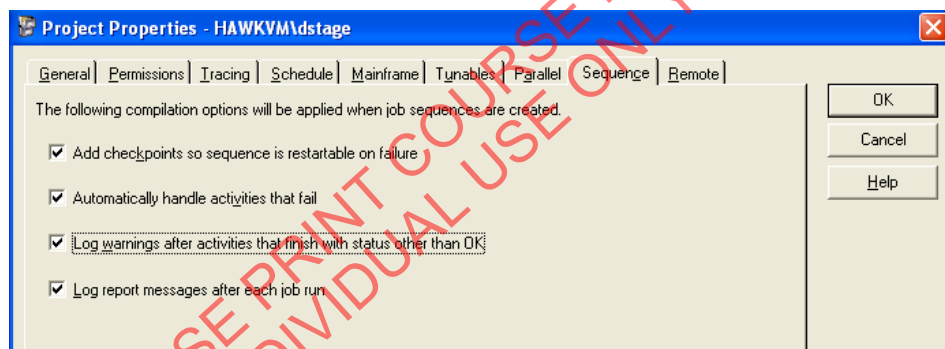
Warning: Tracing causes a lot of server system overhead. This should only be used to diagnose serious problems with the help of IBM DataStage customer support.



Use this tab to specify Parallel job defaults. In addition to displaying the OSH generated by DataStage from Parallel jobs, you specify default formats for dates and times.

In general, you should choose to display the OSH. This provides useful information about how your job works.

DataStage Administrator Sequence Tab



© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Checkpoint

1. Authorizations can be assigned to what two items?
2. What two types of authorization roles can be assigned to a user or group?
3. In addition to Suite authorization to log onto DataStage what else does a DataStage developer require to work in DataStage

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

63

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. Users and groups. Members of a group acquire the authorizations of the group.
2. Suite roles and Product roles.
3. Must be mapped to a user with DataStage credentials.

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

64

Unit objectives

Having completing this unit, you should be able to:

- Open the Administrative console
- Create new users and groups
- Assign Suite roles and Product roles to users and groups
- Give users DataStage credentials
- Log onto DataStage Administrator
- Add a DataStage user on the Permissions tab and specify the user's role
- Specify DataStage global and project defaults
- List and describe important environment variables

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

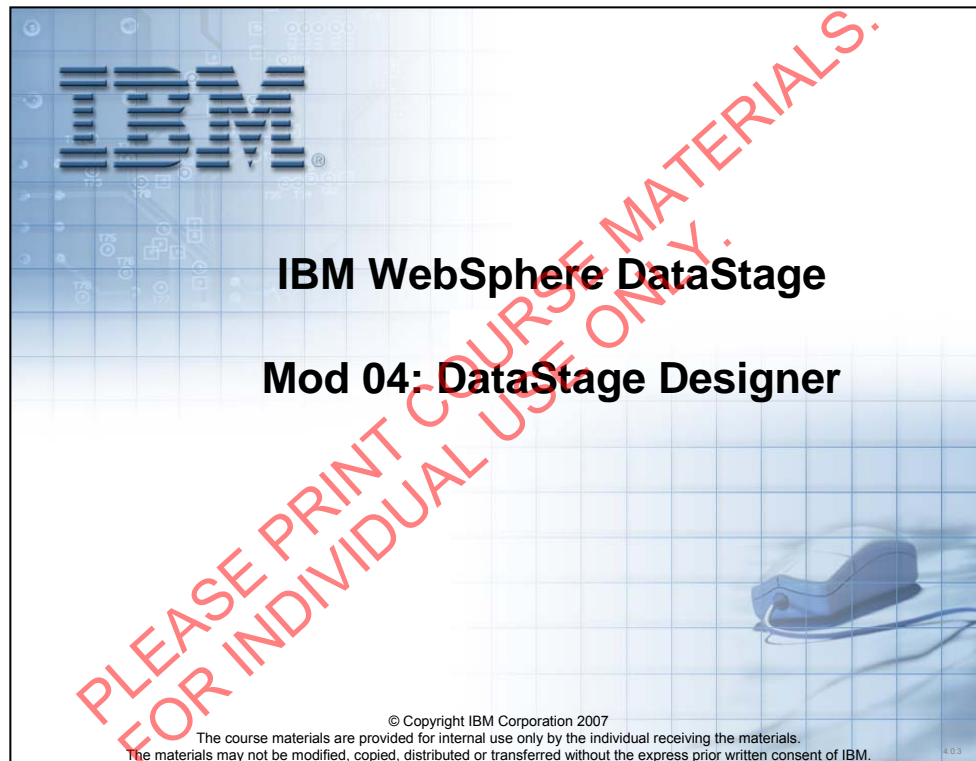
65

Notes:

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

66



Unit objectives

After completing this unit, you should be able to:

- Log onto DataStage
- Navigate around DataStage Designer
- Import and export DataStage objects to a file
- Import a Table Definition for a sequential file

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

68

Notes:

Logging onto DataStage Designer

Host name, port number of application server

DataStage server machine / project

Attach to Project

Domain: hawk_demo:9080

User name: super

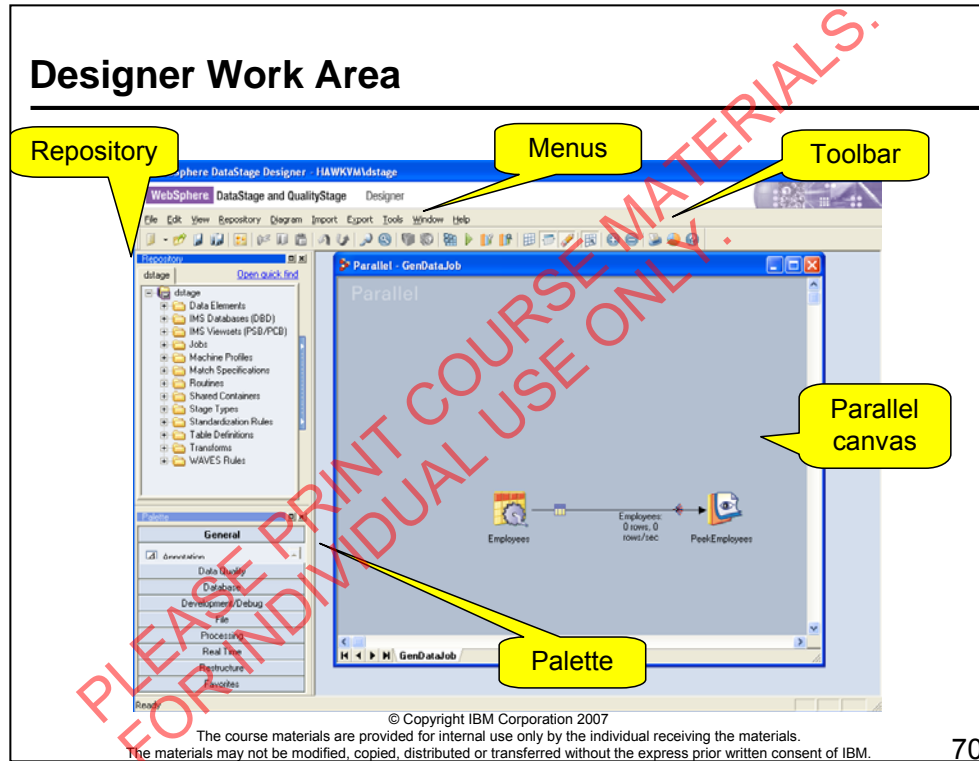
Password: [REDACTED]

Project: hawk_demo/dstage1

OK Cancel Help

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.



70

The appearance of the designer work space is configurable; the graphic shown here is only one example of how you might arrange components.

In the right center is the **Designer canvas**, where you create stages and links. On the top left is the **Repository** window. Items in the Repository, such as jobs and Table Definitions can be dragged to the canvas area. On the bottom left is the Tools Palette, which contains stages you can add to the canvas.

Click **View>Repository** to display the **Repository** window. Click **View>Palette** to display the **Palette** window.

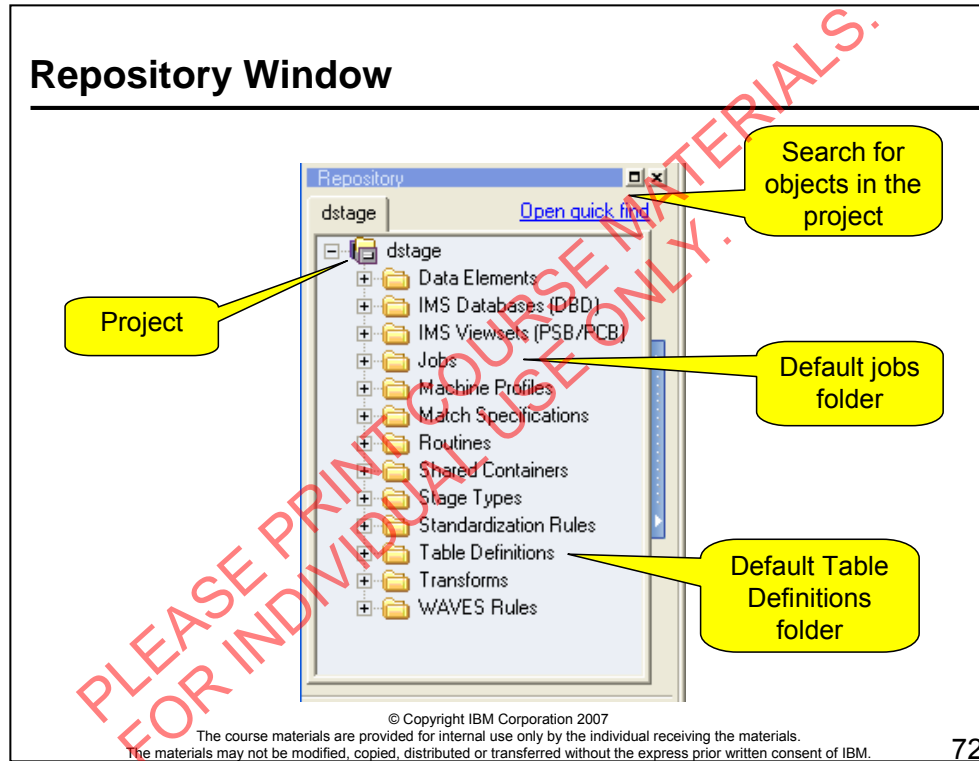
Importing and Exporting DataStage Objects

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

71



The Repository window displays the folders of objects stored in the repository for the DataStage project logged into.

New folders can be created at any level in which to store repository objects.

Import and Export

- Any object in the repository window can be exported to a file
- Can export whole projects
- Use for backup
- Sometimes used for version control
- Can be used to move DataStage objects from one project to another
- Use to share DataStage jobs and projects with other developers

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

73

Any set of DataStage objects, including whole projects, which are stored in the Repository, can be exported to a file. This export file can then be imported back into DataStage.

Import and export can be used for many purposes, including:

- Backing up jobs and projects.
- Maintaining different versions of a job or project.
- Moving DataStage objects from one project to another. Just export the objects, move to the other project, then re-import them into the new project.
- Sharing jobs and projects between developers. The export files, when zipped, are small and can be easily emailed from one developer to another.

Export Procedure

- Click “Export>DataStage Components”
- Select DataStage objects for export
- Specify type of export:
 - DSX: Default format
 - XML: Enables processing of export file by XML applications, e.g., for generating reports
- Specify file path on client machine

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

74

Click **Export>DataStage Components** to begin the export process.

Any object in Manager can be exported to a file. Use this procedure to *backup* your work or to move DataStage objects from one project to another.

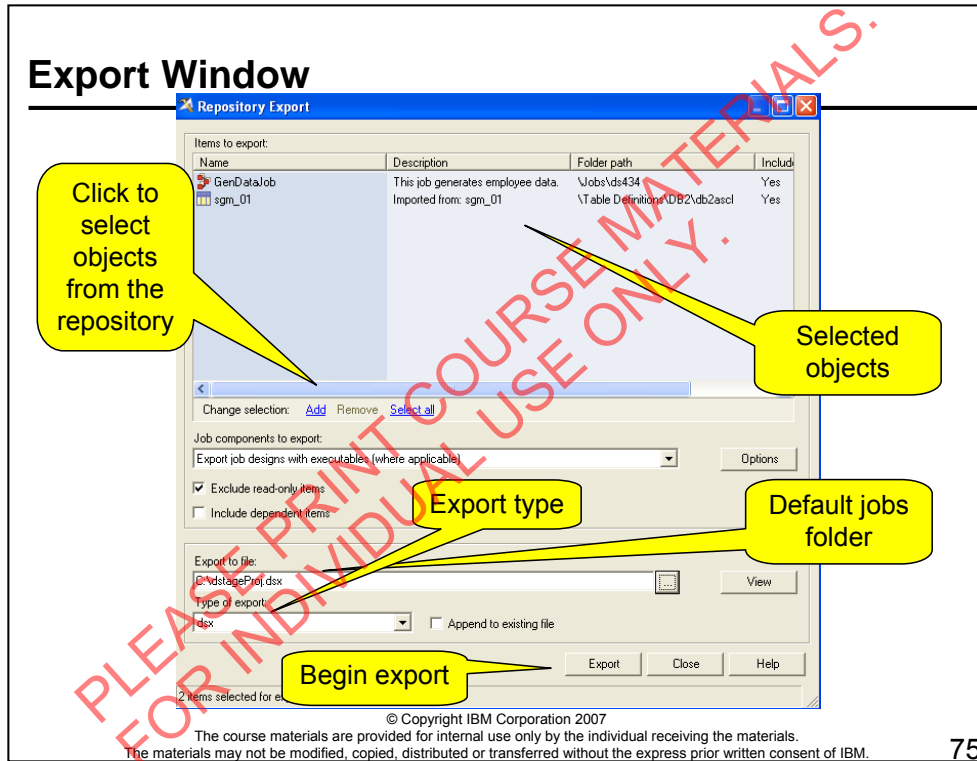
Select the types of components to export. You can select either the whole project or select a portion of the objects in the project.

Specify the name and path of the file to export to. By default, objects are exported to a text file in a special format. By default, the extension is **dsx**. Alternatively, you can export the objects to an XML document.

The directory you export to is on the DataStage *client*, not the server.

Objects can also be exported from the list of found objects of a search. This procedure is discussed later in the course.

Export Window



75

Import Procedure

- Click “Import>DataStage Components”
 - Or “Import>DataStage Components (XML)” if you are importing an XML-format export file
- Select DataStage objects for import

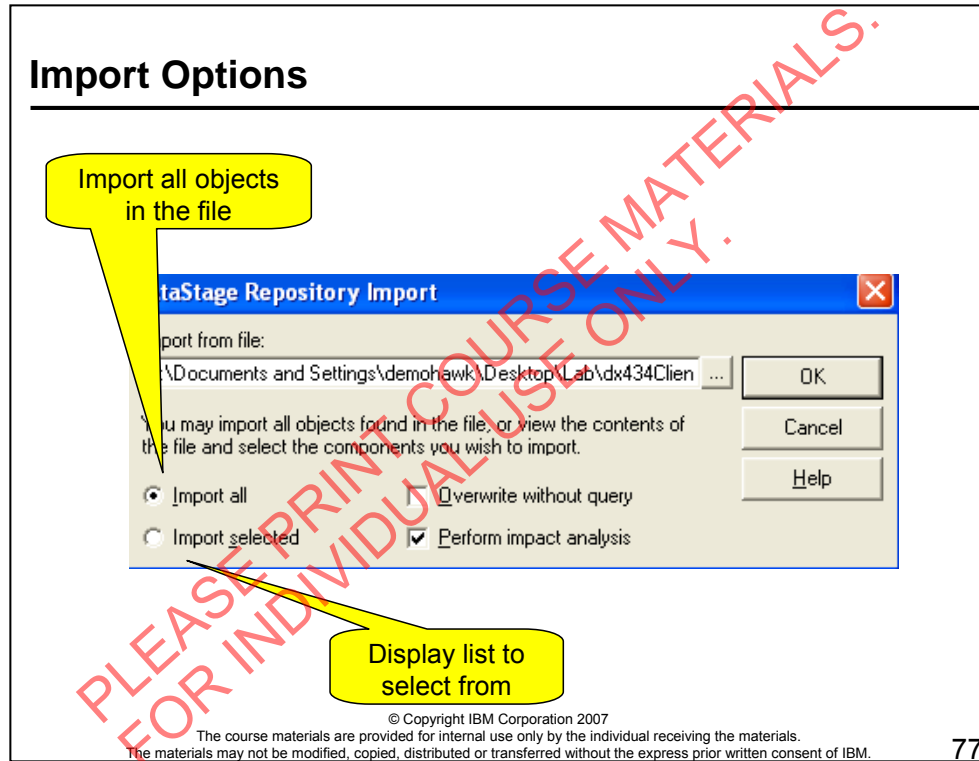
PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

76

To import DataStage components, click **Import>DataStage Components**. Select the file to import. Click **Import all** to begin the import process or **Import selected** to view a list of the objects in the import file. You can import selected objects from the list. Select the **Overwrite without query** button to overwrite objects with the same name without warning.



For large imports, you may want to disable “Perform impact analysis.” This adds overhead to the import process.

Importing Table Definitions

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

78

Importing Table Definitions

- Table Definitions describe the format and columns of files and tables
- You can import Table Definitions for:
 - Sequential files
 - Relational tables
 - COBOL files
 - Many other things
- Table Definitions can be loaded into job stages

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

79

Table Definitions define the formats of a variety of data files and tables. These definitions can then be used and reused in your jobs to specify the formats of data stores.

For example, you can import the format and column definitions of the **Customers.txt** file. You can then load this into the sequential source stage of a job that extracts data from the **Customers.txt** file.

You can load this same metadata into other stages that access data with the same format. In this sense the metadata is *reusable*. It can be used with any file or data store with the same format.

If the column definitions are similar to what you need you can modify the definitions and save the Table Definition under a new name.

You can import and define several different kinds of Table Definitions including: Sequential files and ODBC data sources.

Sequential File Import Procedure

- Click Import>Table Definitions>Sequential File Definitions
- Select directory containing sequential file and then the file
- Select a repository folder to store the Table Definition in
- Examine format and column definitions and edit as necessary

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

80

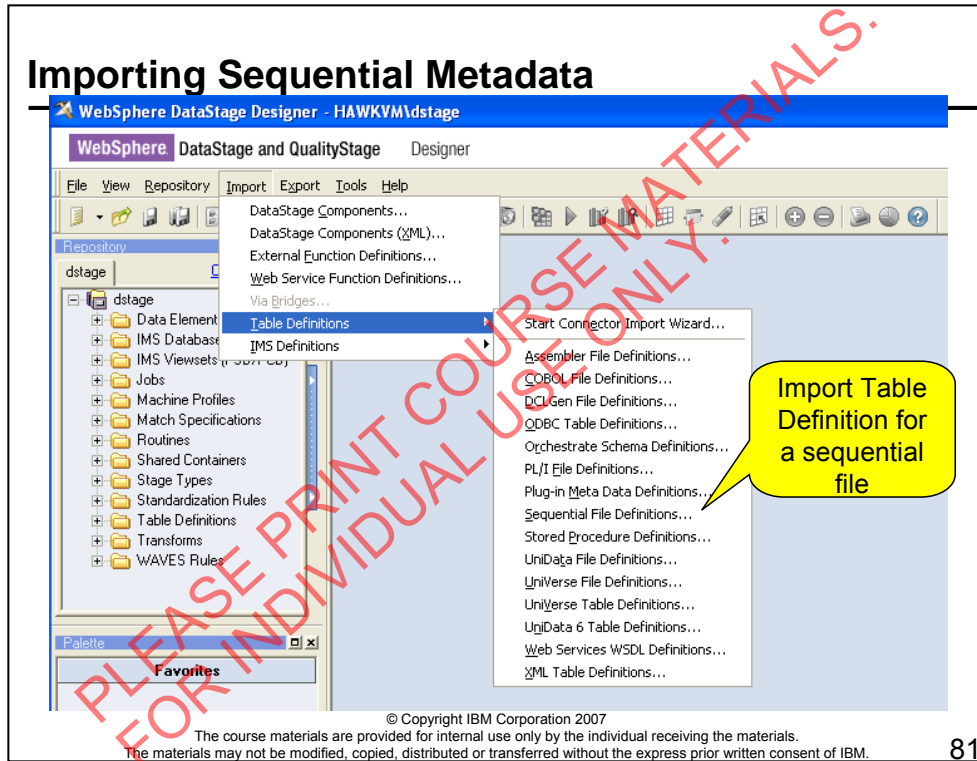
To start the import, click **Import>Table Definitions>Sequential File Definitions**. The **Import Meta Data (Sequential)** window is displayed.

Select the directory containing the sequential files. The **Files** box is then populated with the files you can import.

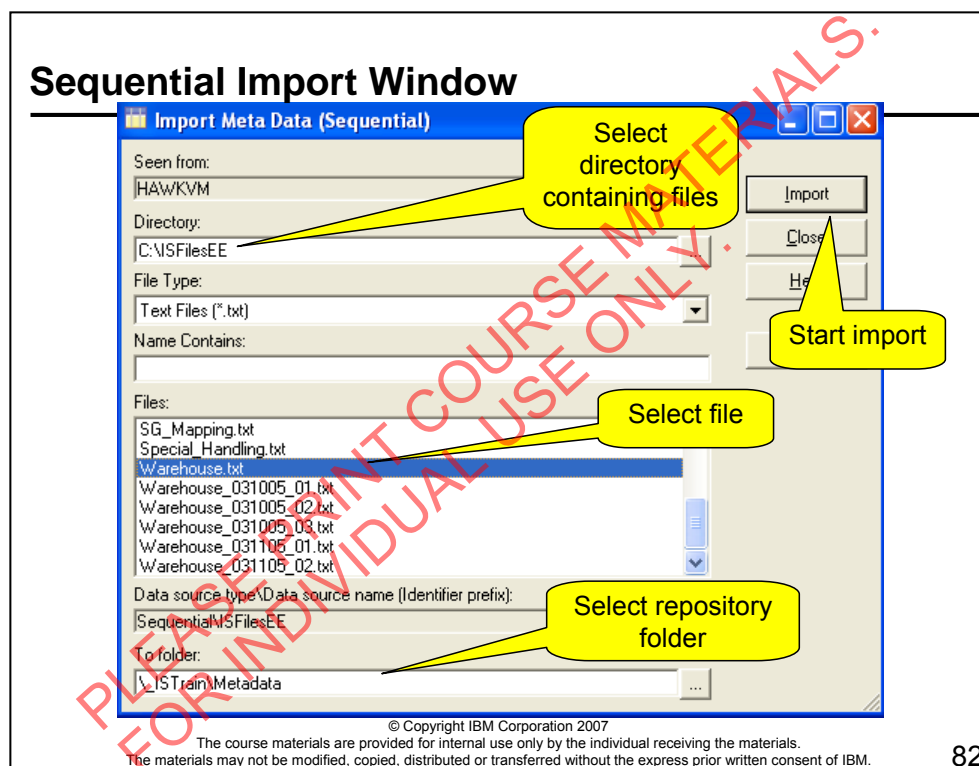
Select the file to import.

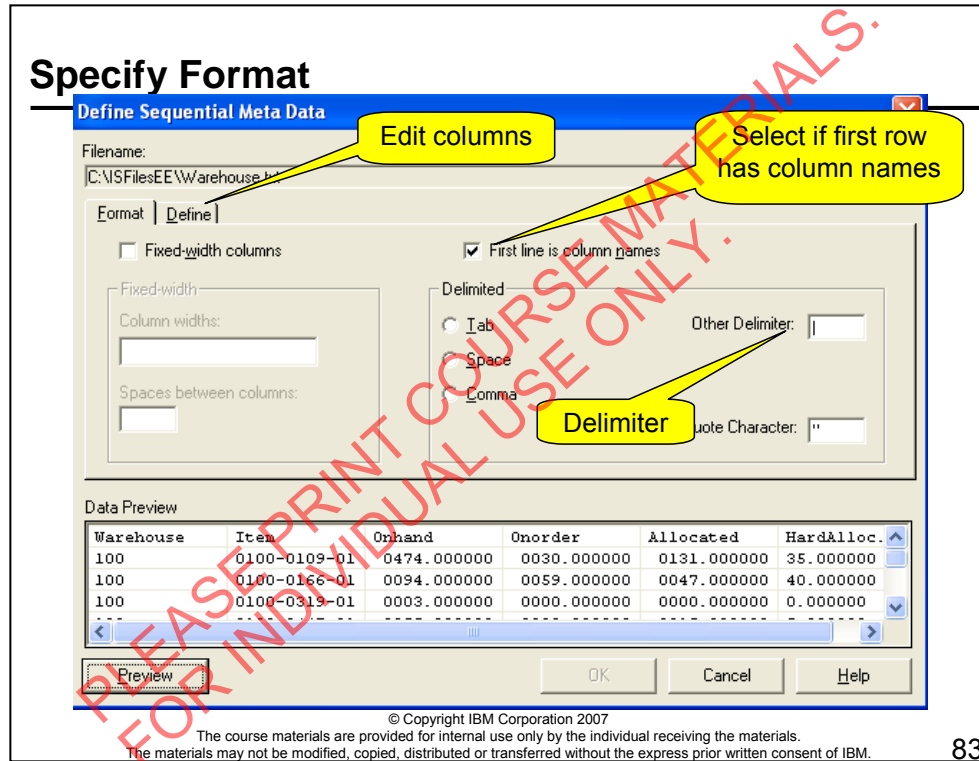
Select or specify a folder to import into.

Importing Sequential Metadata



81





Edit Column Names and Types

Define Sequential Meta Data

Filename: C:\SFilesEE\Warehouse.txt

Format: [Define...](#)

Double-click to define extended properties

| | Column name | Key | SQL type | Length | Scale | Nullable | Display | Data element | escriptic |
|---|---------------|--------------------------|----------|--------|-------|----------|---------|--------------|-----------|
| 1 | Warehouse | <input type="checkbox"/> | Integer | 10 | | No | 3 | | |
| 2 | Item | <input type="checkbox"/> | VarChar | 255 | | No | 14 | | |
| 3 | Onhand | <input type="checkbox"/> | Numeric | 10 | | No | 12 | | |
| 4 | Onorder | <input type="checkbox"/> | Numeric | 10 | | No | 12 | | |
| 5 | Allocated | <input type="checkbox"/> | Numeric | 10 | | No | 12 | | |
| 6 | HardAllocated | <input type="checkbox"/> | Numeric | 10 | | No | 9 | | |

Data Preview

| | Warehouse | Item | Onhand | Onorder | Allocated | HardAlloc. |
|-----|-----------|--------------|-------------|-------------|-------------|------------|
| 100 | | 0100-0109-01 | 0474.000000 | 0030.000000 | 0131.000000 | 35.000000 |
| 100 | | 0100-0166-01 | 0094.000000 | 0059.000000 | 0047.000000 | 40.000000 |
| 100 | | 0100-0319-01 | 0003.000000 | 0000.000000 | 0000.000000 | 0.000000 |

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

84

You can also add “Extended Properties”. Double-click on the number to the left of the column name to open up a window in which you specify these extended properties. Extended properties are discussed later in this course.

Extended Properties window

Edit Column Meta Data

Column name: Warehouse Key: No

Native type: SQL type: Integer Length: 10

Scale: 0 Nullable: No Date format:

Description:

Parallel properties

Server: COBOL Parallel Analytical information

Field type: int32

☐ Extended (Unsigned)

Vector occurs:

☐ Variable

Level number:

Property categories

Properties:

- Field level
- Integer type
- Generator

Properties that apply to fields as a whole:

Available properties to add:

- Bytes to skip
- Drop on input
- Delimiter
- Generate on c
- Delimiter string
- Prefix bytes

Available properties

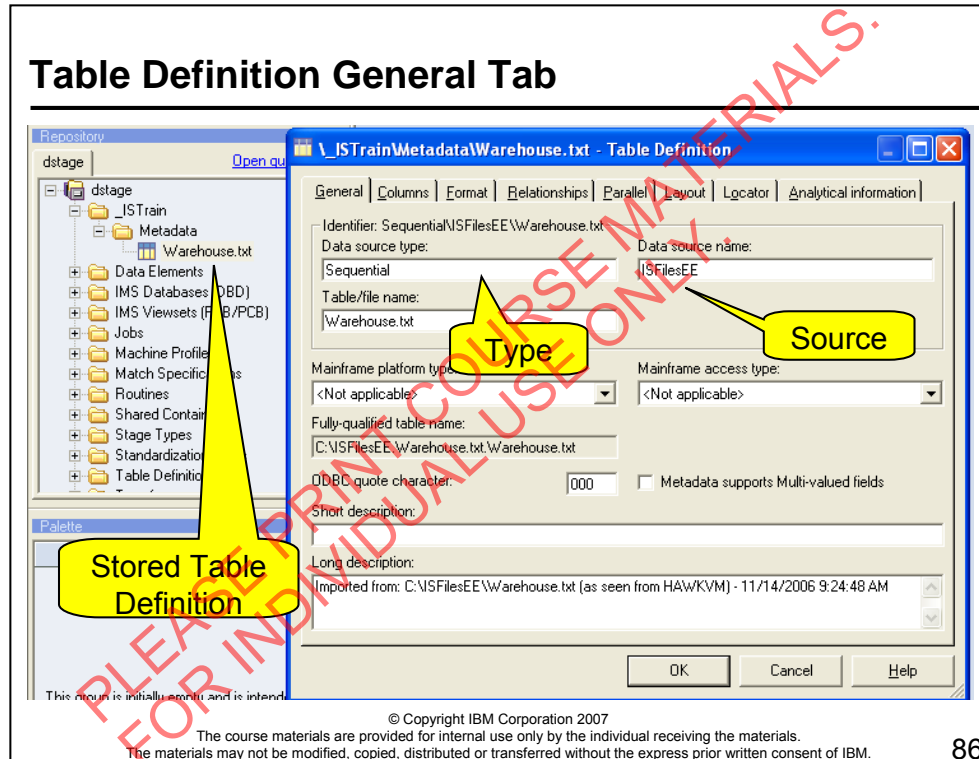
< Back Next >

Close Apply Reset Help

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

85



In the Repository window, select the folder that contains the Table Definition. Double-click the Table Definition to open the **Table Definition** window.

Click the **Columns** tab to view and modify any column definitions. Select the **Format** tab to edit the file format specification. Select the **Parallel** tab to specify parallel format properties.

Checkpoint

- True or False? The directory to which you export is on the DataStage client machine, not on the DataStage server machine.
- Can you import Table Definitions for sequential files with fixed-length record formats?

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

87

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. True.
2. Yes. Record lengths are determined by the lengths of the individual columns.

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

88

Unit summary

Having completed this unit, you should be able to:

- Log onto DataStage
- Navigate around DataStage Designer
- Import and export DataStage objects to a file
- Import a Table Definition for a sequential file

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

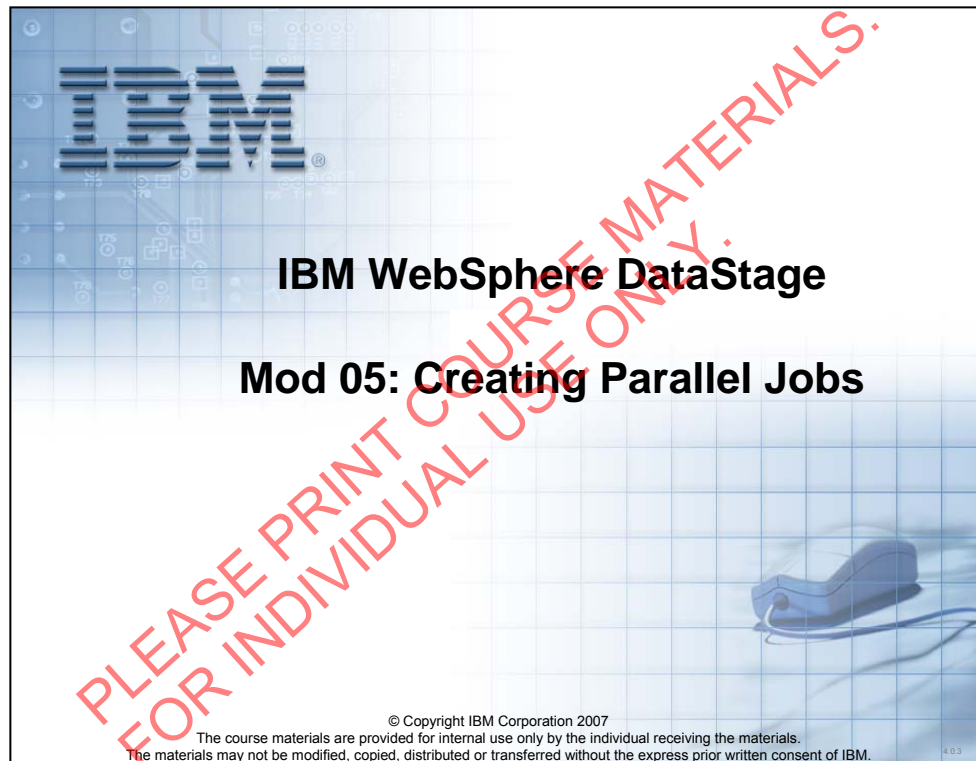
89

Notes:

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

90



Unit objectives

After completing this unit, you should be able to:

- Design a simple Parallel job in Designer
- Define a job parameter
- Use the Row Generator, Peek, and Annotation stages in a job
- Compile your job
- Run your job in Director
- View the job log

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

92

Notes:

What Is a Parallel Job?

- Executable DataStage program
- Created in DataStage Designer
 - Can use components from Repository
- Built using a graphical user interface
- Compiles into Orchestrate script language (OSH) and object code (from generated C++)

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

93

A *job* is an executable DataStage program. In DataStage, you can design and run jobs that perform many useful data integration tasks, including data extraction, data conversion, data aggregation, data loading, etc.

DataStage jobs are:

- Designed and built in Designer.
- Scheduled, invoked, and monitored in Director.
- Executed under the control of DataStage.

Job Development Overview

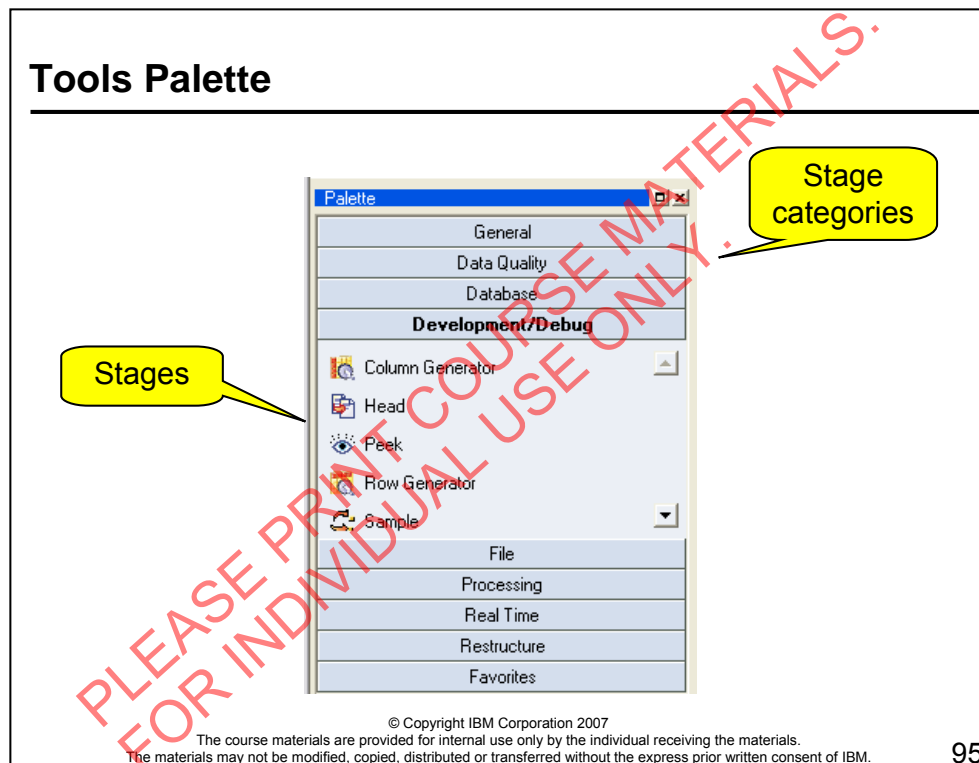
- Import metadata defining sources and targets
 - Done within Designer using import process
- In Designer, add stages defining data extractions and loads
- Add processing stages to define data transformations
- Add links defining the flow of data from sources to targets
- Compile the job
- In Director, validate, run, and monitor your job
 - Can also run the job in Designer
 - Can only view the job log in Director

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

94

In this module, you will go through the whole process with a simple job, except for the first bullet. In this module you will manually define the metadata.



The tool palette contains icons that represent the components you can add to your job design.

Adding Stages and Links

- Drag stages from the Tools Palette to the diagram
 - Can also be dragged from Stage Type branch to the diagram
- Draw links from source to target stage
 - Right mouse over source stage
 - Release mouse button over target stage

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

96

Job Creation Example Sequence

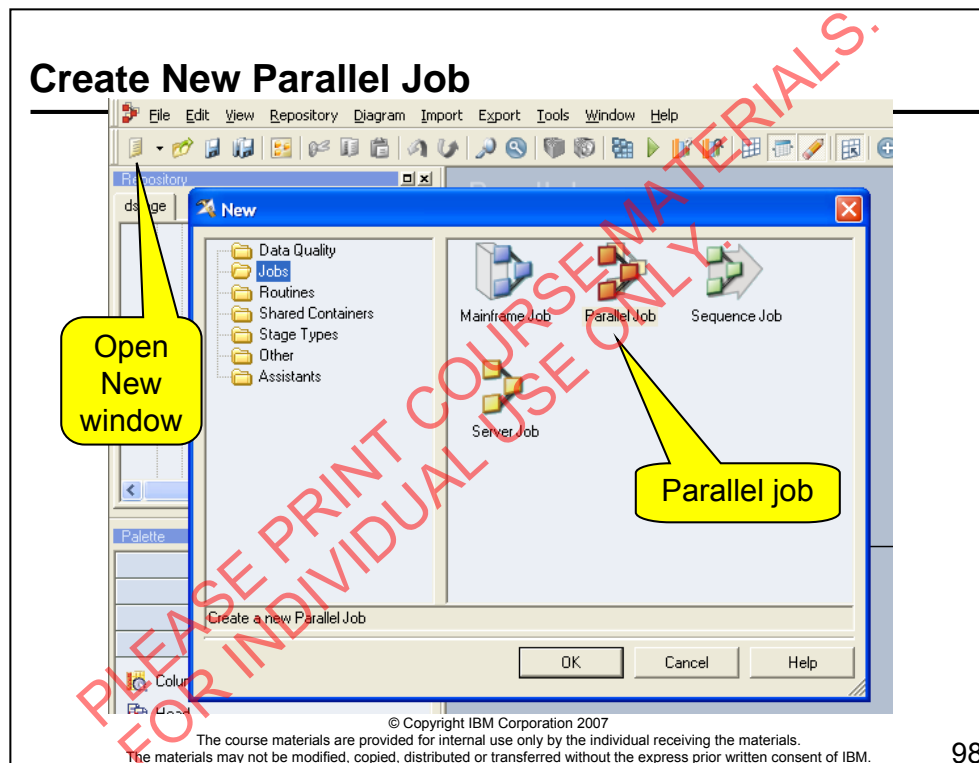
- Brief walkthrough of procedure
- Assumes Table Definition of source already exists in the repository

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

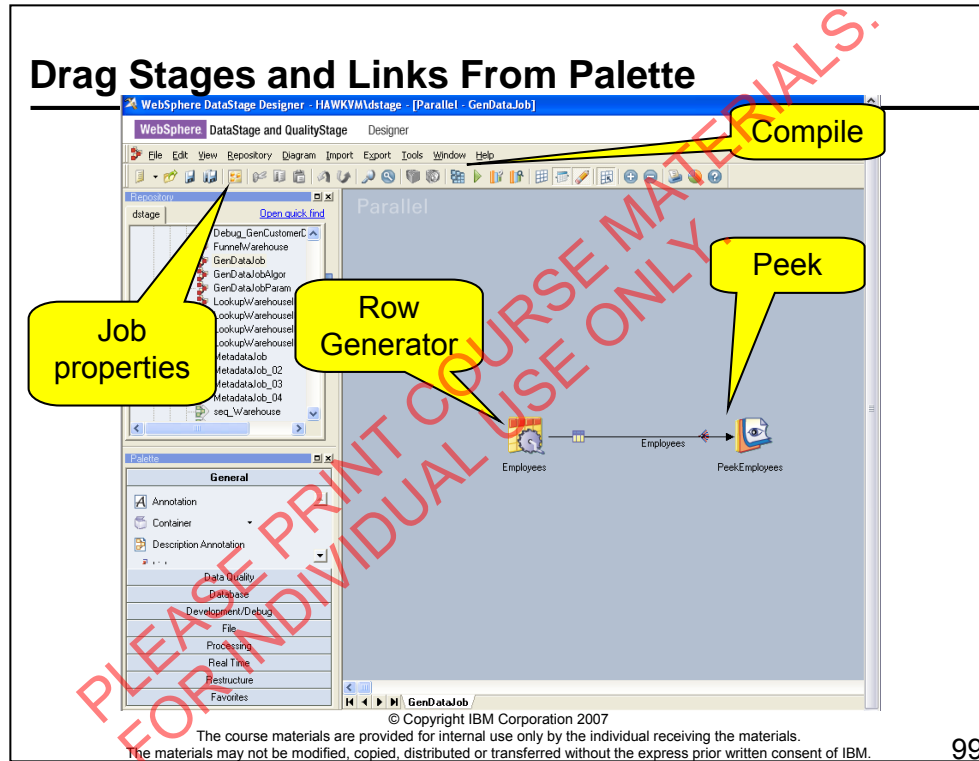
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

97



98

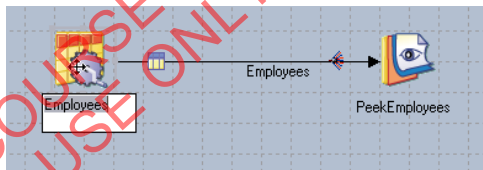
Click the New button in the toolbar to open the New window. Click on the Parallel Job icon to create a new parallel job (the focus of this course).



The tools palette may be shown as a floating dock or placed along a border. Alternatively, it may be hidden and the developer may choose to pull needed stages from the repository onto the design work area.

Renaming Links and Stages

- Click on a stage or link to rename it
- Meaningful names have many benefits
 - Documentation
 - Clarity
 - Fewer development errors



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

100

RowGenerator Stage

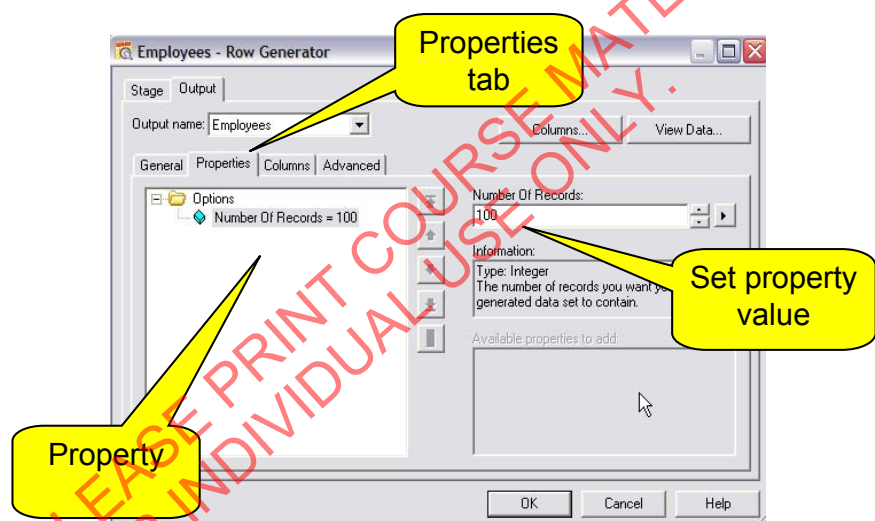
- Produces mock data for specified columns
- No inputs link; single output link
- On Properties tab, specify number of rows
- On Columns tab, load or specify column definitions
 - Open Extended Properties window to specify the values to be generated for that column
 - A number of algorithms for generating values are available depending on the data type
- Algorithms for Integer type
 - Random: seed, limit
 - Cycle: Initial value, increment
- Algorithms for string type: Cycle , alphabet
- Algorithms for date type: Random, cycle

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

101

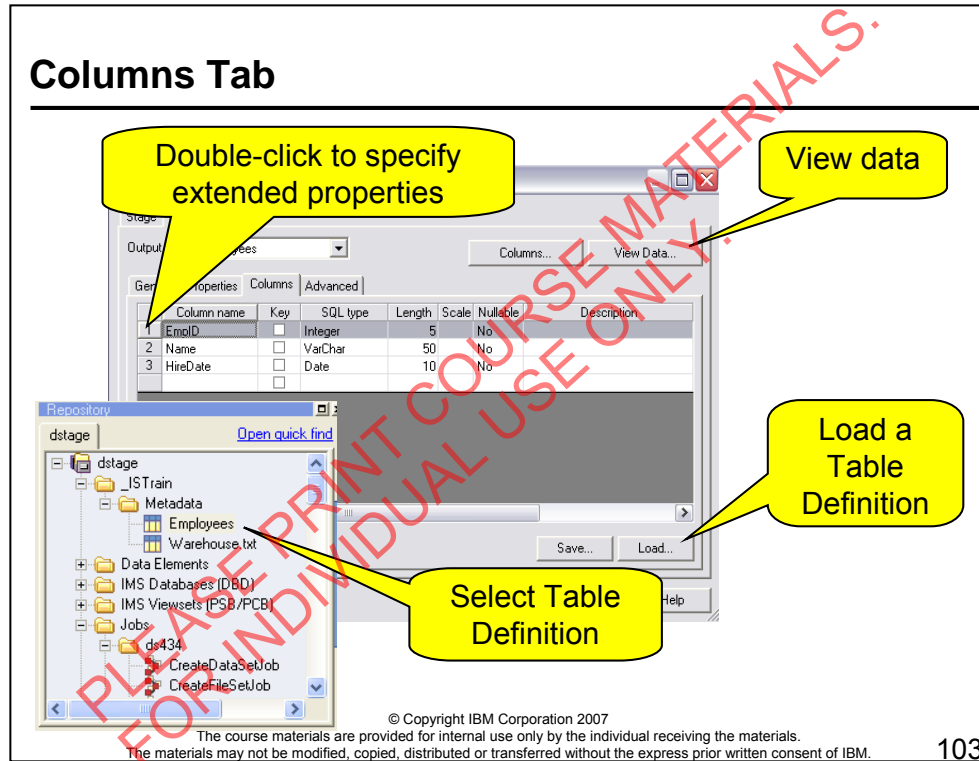
Inside the Row Generator Stage



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

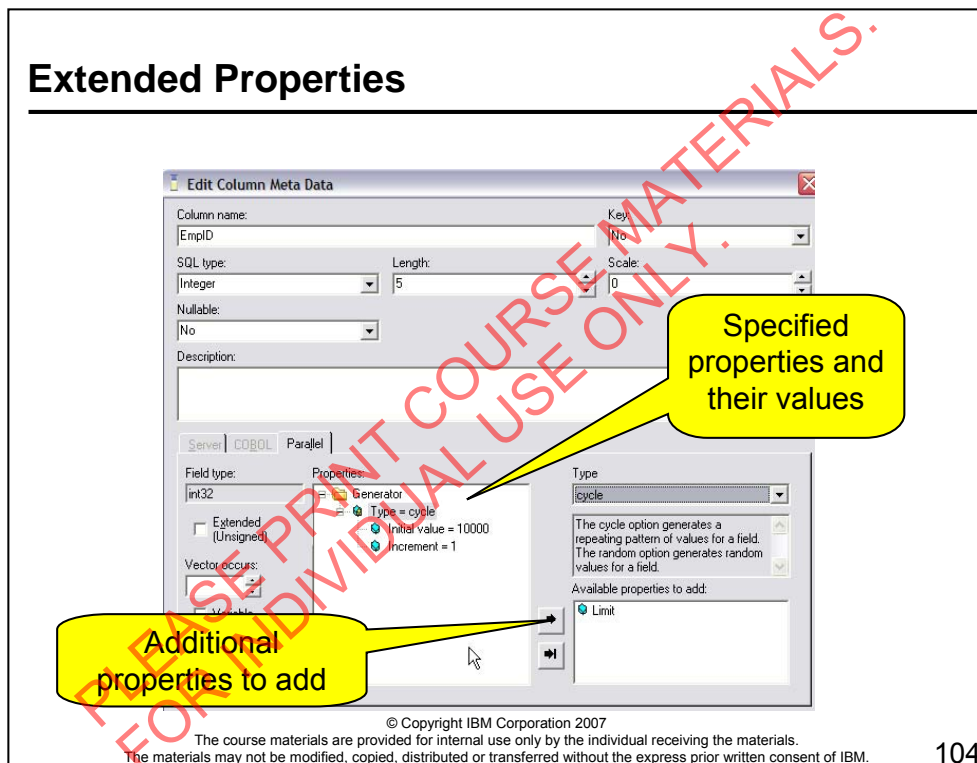
102



103

On the Columns tab, define the column definitions. Either manually specify the columns or load the column definitions from a Table Definition. A Table Definition can either be loaded, as shown here, or dragged from the Repository and dropped on the link.

Extended Properties



104

Double-click on the column number to define the extended properties for the column

Peek Stage

- Displays field values
 - Displayed in job log or sent to a file
 - Skip records option
 - Can control number of records to be displayed
 - Shows data in each partition, labeled 0, 1, 2, ...
- Useful stub stage for iterative job development
 - Develop job to a stopping point and check the data

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

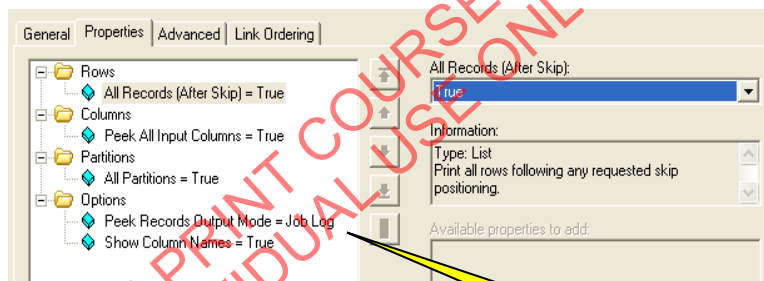
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

105

The peek stage will display column values in a job's output messages log.

Peek Stage Properties



Output to
job log

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

106

You can also output from the Peek stage to a file.

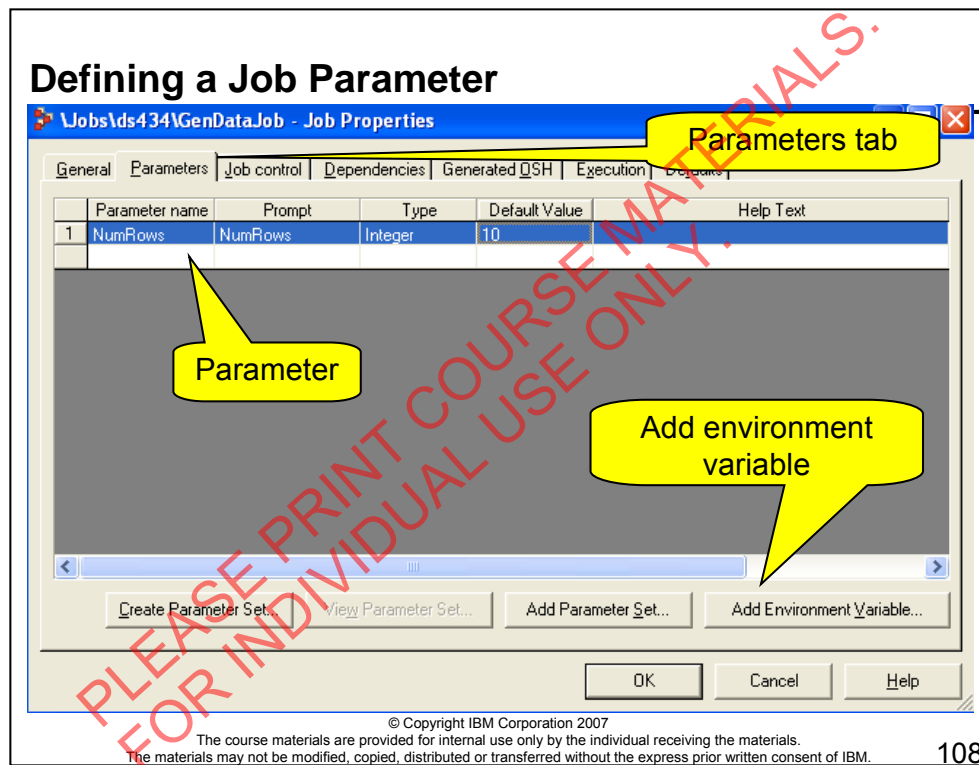
Job Parameters

- Defined in Job Properties window
- Makes the job more flexible
- Parameters can be:
 - Used in directory and file names
 - Used to specify property values
 - Used in constraints and derivations
- Parameter values are determined at run time
- When used for directory and files names and property values, they are surrounded with pound signs (#)
 - E.g., #NumRows#
- Job parameters can reference DataStage environment variables
 - Prefaced by \$, e.g., \$APT_CONFIG_FILE

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

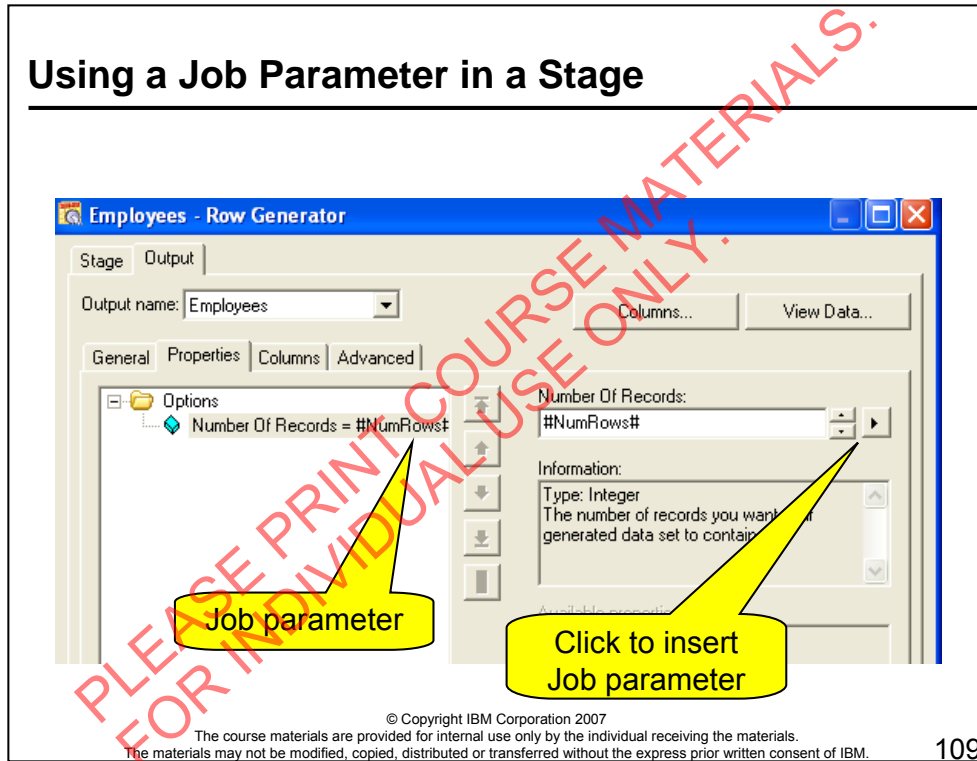
107



108

Click the Job Properties icon to open this window.

Using a Job Parameter in a Stage



109

Select the property. Then enter the value in the text box. Click the button at the right of the text box to display a menu for selecting a job parameter.

Adding Job Documentation

- Job Properties
 - Short and long descriptions
- Annotation stage
 - Added from the Tools Palette
 - Displays formatted text descriptions on diagram

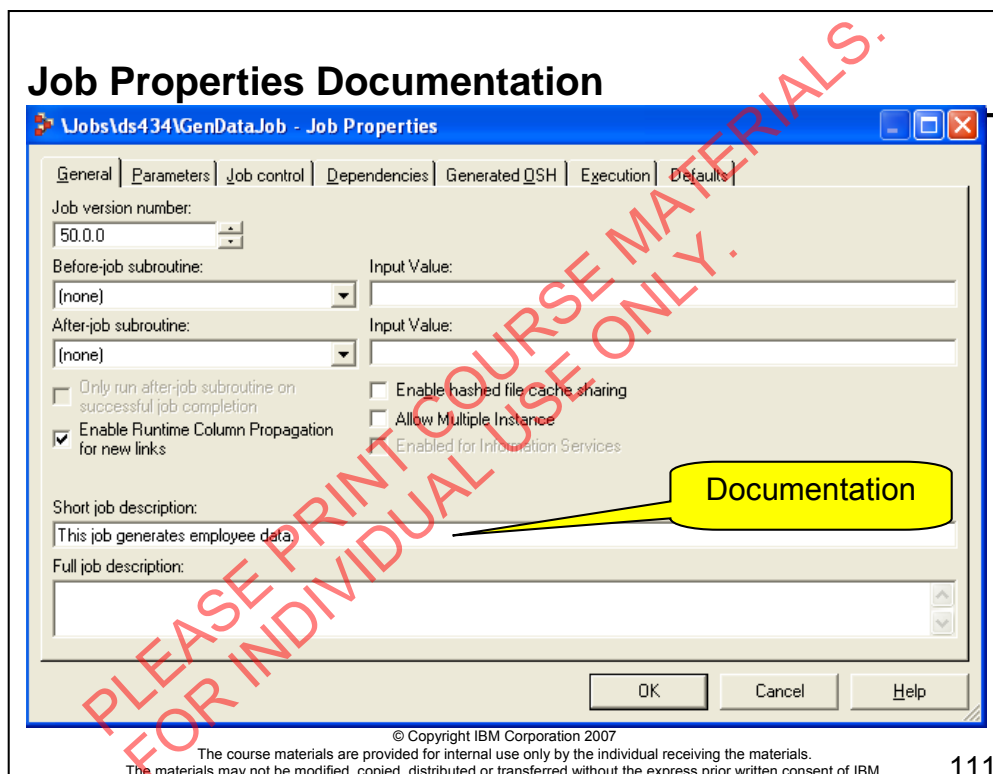
PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

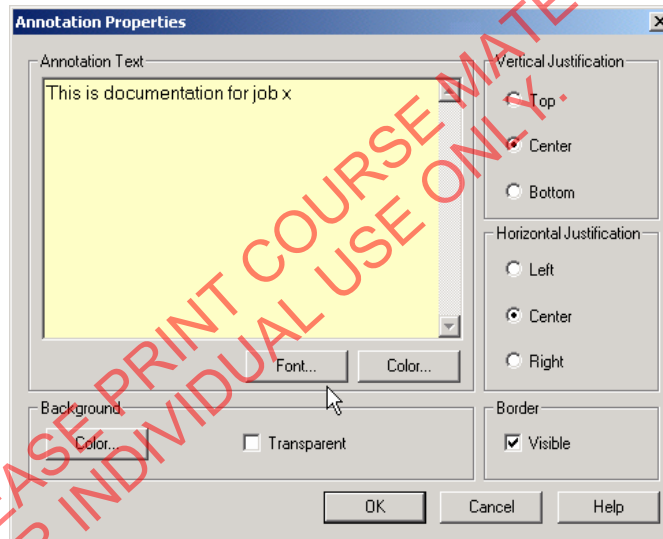
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

110

This documentation is displayed in Manager and Director in addition to the job diagram.



Annotation Stage Properties



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

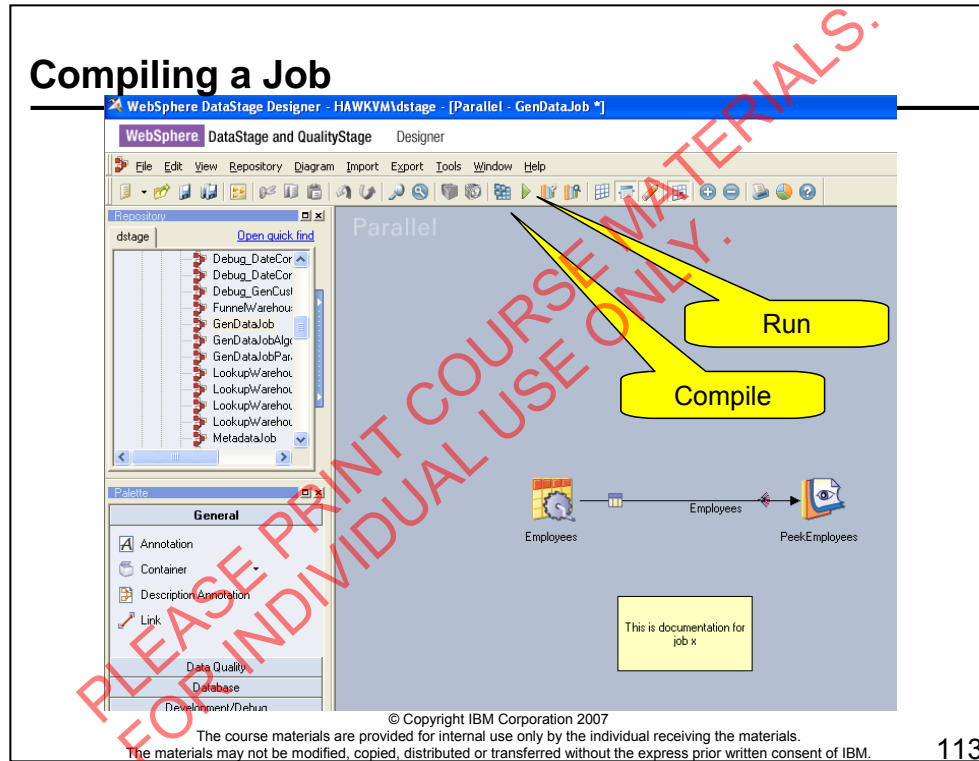
112

You can type in whatever you want; the default text comes from the short description of the jobs properties you entered, if any.

Add one or more Annotation stages to the canvas to document your job.

An Annotation stage works like a text box with various formatting options. You can optionally show or hide the Annotation stages by pressing a button on the toolbar.

There are two Annotation stages. The Description Annotation stage correlates its text with the Descriptions specified as part of the job properties.

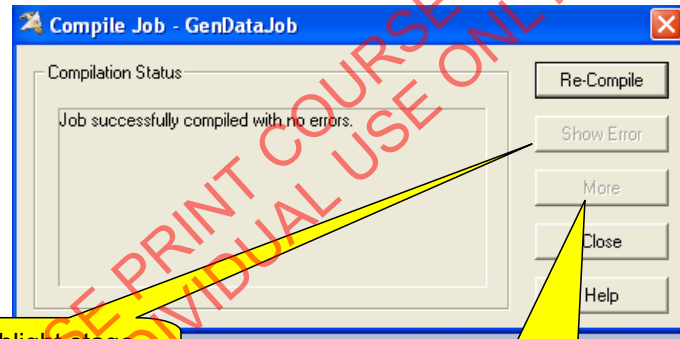


113

Before you can run your job, you must compile it. To compile it, click **File>Compile** or click the **Compile** button on the toolbar. The **Compile Job** window displays the status of the compile.

A compile will generate OSH.

Errors or Successful Message



Highlight stage
with error

Click for more info

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

114

If an error occurs:

Click **Show Error** to identify the stage where the error occurred. This will highlight the stage in error.

Click **More** to retrieve more information about the error. This can be lengthy for parallel jobs.

Many errors also show up on the diagram if "Show Stage Validation Errors" is turned on.

Running Jobs and Viewing the Job Log in Designer

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

115

DataStage Director

- Use to run and schedule jobs
- View runtime messages
- Can invoke directly from Designer
 - [Tools > Run Director](#)

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

116

As you know, you run your jobs in Director. You can open Director from within Designer by clicking Tools>Run Director.

In a similar way, you can move between Director and Designer.

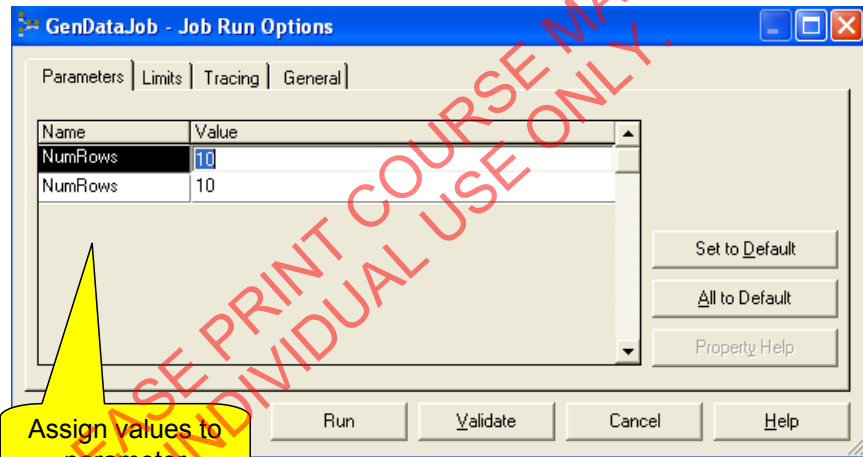
There are two methods for running a job:

- Run it immediately.
- Schedule it to run at a later time or date.

To run a job immediately:

- Select the job in the Job Status view. The job must have been compiled.
- Click Job>Run Now or click the Run Now button in the toolbar. The Job Run Options window is displayed.

Run Options



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

117

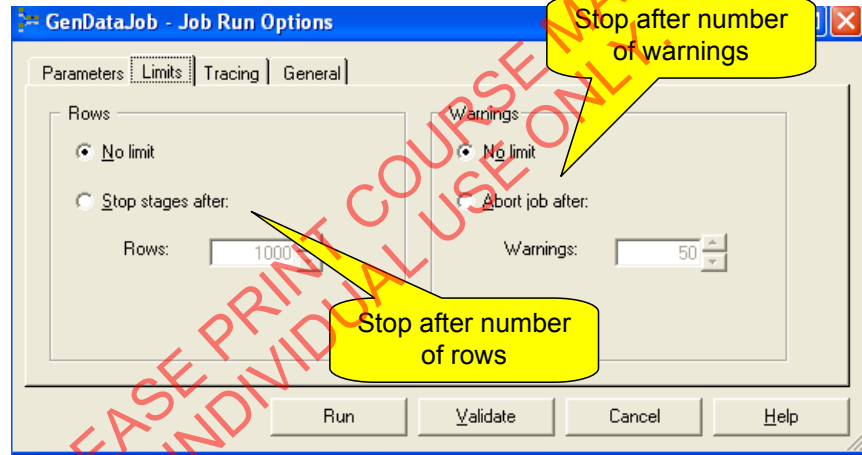
The Job Run Options window is displayed when you click Job>Run Now.

This window allows you to stop the job after:

- A certain number of rows.
- A certain number of warning messages.

Click Run to run the job after it is validated. The Status column displays the status of the job run.

Run Options



© Copyright IBM Corporation 2007
 The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

118

The Job Run Options window is displayed when you click Job>Run Now.

This window allows you to stop the job after:

- A certain number of rows.
- A certain number of warning messages.

Click Run to run the job after it is validated. The Status column displays the status of the job run.

Director Status View

Status view

Schedule view

Log view

| Job name | Status | Started | On date | Last ran | On... | Elapsed... | Description |
|------------|----------|----------|------------|----------|--------|------------|-----------------------------------|
| GenDataJob | Finished | 01:39 PM | 11/14/2006 | 01:41 PM | 11/... | 00:02:03 | This job generates employee data. |

Status of jobs: 1 entries

Server time: 11/14/2006 02:09 PM

Select job to view messages in the log view

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Director Log View

Click the open book icon to view log messages

Peek messages

| > Occurred | > On date | Type | Event |
|------------|------------|---------|---|
| 1:39:20 PM | 11/14/2006 | Control | Starting Job GenDataJob. (...) |
| 1:40:20 PM | 11/14/2006 | Info | Environment variable settings: (...) |
| 1:40:21 PM | 11/14/2006 | Info | Parallel job initiated |
| 1:40:21 PM | 11/14/2006 | Info | OSH script (...) |
| 1:40:31 PM | 11/14/2006 | Info | main_program: IBM WebSphere DataStage Engine 8.0.0 (...) |
| 1:40:37 PM | 11/14/2006 | Info | main_program: orchgeneral: loaded (...) |
| 1:40:37 PM | 11/14/2006 | Info | main_program: Echo (...) |
| 1:40:39 PM | 11/14/2006 | Info | main_program: Explanation: (...) |
| 1:40:47 PM | 11/14/2006 | Info | main_program: Dump: (...) |
| 1:40:47 PM | 11/14/2006 | Info | main_program: APT configuration file: C:/IBM/InformationServer/Server/Conf... |
| 1:40:47 PM | 11/14/2006 | Info | main_program: This step has no datasets: (...) |
| 1:41:06 PM | 11/14/2006 | Info | main_program: Schemas: (...) |
| 1:41:06 PM | 11/14/2006 | Info | PeekEmployees,0: EmplD:10000 Name:aaaaa HireDate:1960-01-01 (...) |
| 1:41:06 PM | 11/14/2006 | Info | Employees,0: Output 0 produced 10 records. |
| 1:41:06 PM | 11/14/2006 | Info | PeekEmployees,0: EmplD:10009 Name: HireDate:1960-01-10 |
| 1:41:06 PM | 11/14/2006 | Info | main_program: Step execution finished with status = OK. |
| 1:41:06 PM | 11/14/2006 | Info | main_program: Startup time, 0:33; production run time, 0:01. |
| 1:41:18 PM | 11/14/2006 | Info | Parallel job reports successful completion |
| 1:41:38 PM | 11/14/2006 | Control | Finished Job GenDataJob. |

Log for job: GenDataJob 19 entries (filtered) Server time: 11/14/2006 02:15 PM

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Click the **Log** button in the toolbar to view the job log. The job log records events that occur during the execution of a job.

These events include *control events*, such as the starting, finishing, and aborting of a job; informational messages; warning messages; error messages; and program-generated messages.

Message Details

Event Detail

Server: [HAWKVM] Project: [dstage] User: [HAWKVM\demohawk]

Job No: [18] Job name: [GenDataJob] Invocation: []

Event Number: [12] Event type: [Info] Timestamp: [11/14/2006 1:41:06 PM]

Message Id: [IIS-DSEE.TOPK-00003]

Message:

```
PeekEmployees,0: EmpID:10000 Name:aadea HireDate:1960-01-01
PeekEmployees,0: EmpID:10001 Name:bbbbb HireDate:1960-01-02
PeekEmployees,0: EmpID:10002 Name:cccccccccccccccccccc HireDate:1960-01-03
PeekEmployees,0: EmpID:10003 Name:dd HireDate:1960-01-04
PeekEmployees,0: EmpID:10004 Name:eeeeeeeeeeeeeeeeeeee HireDate:1960-01-05
PeekEmployees,0: EmpID:10005 Name:ffffffffffff HireDate:1960-01-06
PeekEmployees,0: EmpID:10006 Name:gggggggggggggggggggggggggggg HireDate:1960-01-07
PeekEmployees,0: EmpID:10007 Name:h h h h h h h h h h h HireDate:1960-01-08
PeekEmployees,0: EmpID:10008 Name:i i i i i i i i i i i HireDate:1960-01-09
```

Close

Next

Previous

Copy

Help

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Double-click on a message to open it up and read the details.

Other Director Functions

- Schedule job to run on a particular date/time
- Clear job log of messages
- Set job log purging conditions
- Set Director options
 - Row limits
 - Abort after x warnings

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

122

Running Jobs from Command Line

- `dsjob -run -param numRows=10 dx444 GenDataJob`
 - Runs a job
 - Use `-run` to run the job
 - Use `-param` to specify parameters
 - In this example, `dx444` is the name of the project
 - In this example, `GenDataJob` is the name of the job
- `dsjob -logsum dx444 GenDataJob`
 - Displays a job's messages in the log
- Documented in "Parallel Job Advanced Developer's Guide"

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

123

Checkpoint

1. Which stage can be used to display output data in the job log?
2. Which stage is used for documenting your job on the job canvas?
3. What command is used to run jobs from the operating system command line?

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

124

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. Peek stage
2. Annotation stage
3. dsjob -run

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

125

Unit summary

Having completed this unit, you should be able to:

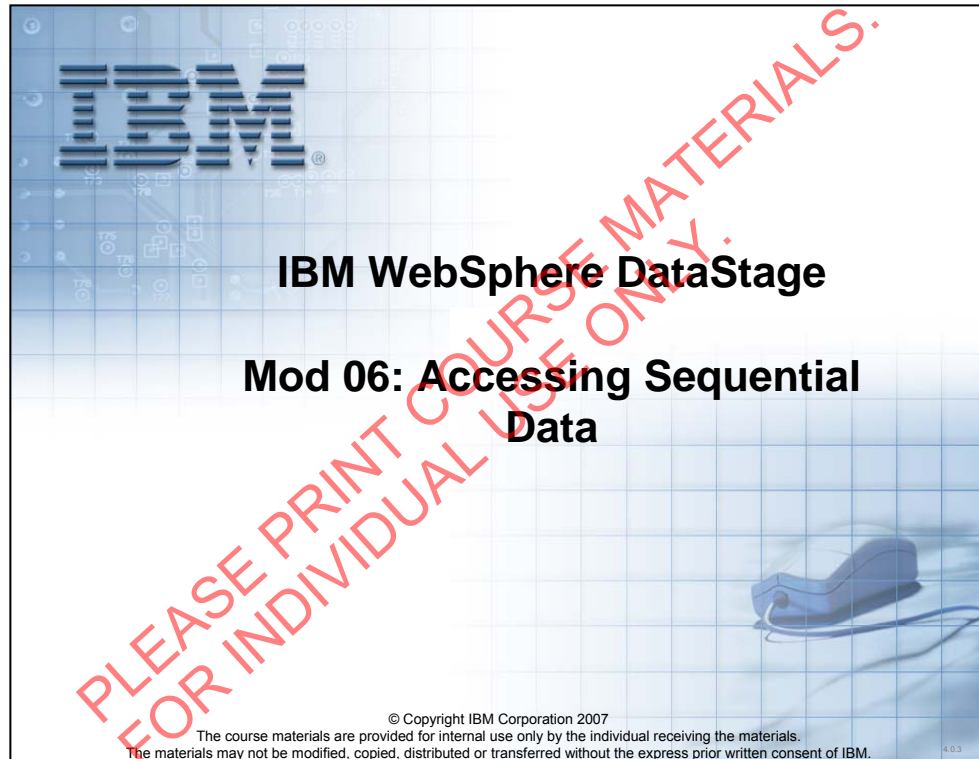
- Design a simple Parallel job in Designer
- Define a job parameter
- Use the Row Generator, Peek, and Annotation stages in a job
- Compile your job
- Run your job in Director
- View the job log

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

126

Notes:



Unit objectives

After completing this unit, you should be able to:

- Understand the stages for accessing different kinds of file data
- Sequential File stage
- Data Set stage
- Create jobs that read from and write to sequential files
- Create Reject links
- Work with NULLs in sequential files
- Read from multiple files using file patterns
- Use multiple readers

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

128

Notes:

Types of File Data

- Sequential
 - Fixed or variable length
- Data Set
- Complex flat file

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

129

Several stages handle sequential data. Each stage has both advantages and differences from the other stages that handle sequential data.

Sequential data can come in a variety of types -- including both fixed length and variable length.

How Sequential Data is Handled

- Import and export operators are generated
 - Stages get translated into operators during the compile
- Import operators convert data from the external format, as described by the Table Definition, to the framework internal format
 - Internally, the format of data is described by schemas
- Export operators reverse the process
- Messages in the job log use the “import” / “export” terminology
 - E.g., “100 records imported successfully; 2 rejected”
 - E.g., “100 records exported successfully; 0 rejected”
 - Records get rejected when they cannot be converted correctly during the import or export



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

130

Using the Sequential File Stage

Both import and export of sequential files (text, binary) can be performed by the **SequentialFile** Stage.

- Data import:  → Internal format
- Data export: Internal format → 

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

131

When data is imported the imported operator translates that data into the framework internal format. The export operator performs the reverse action.

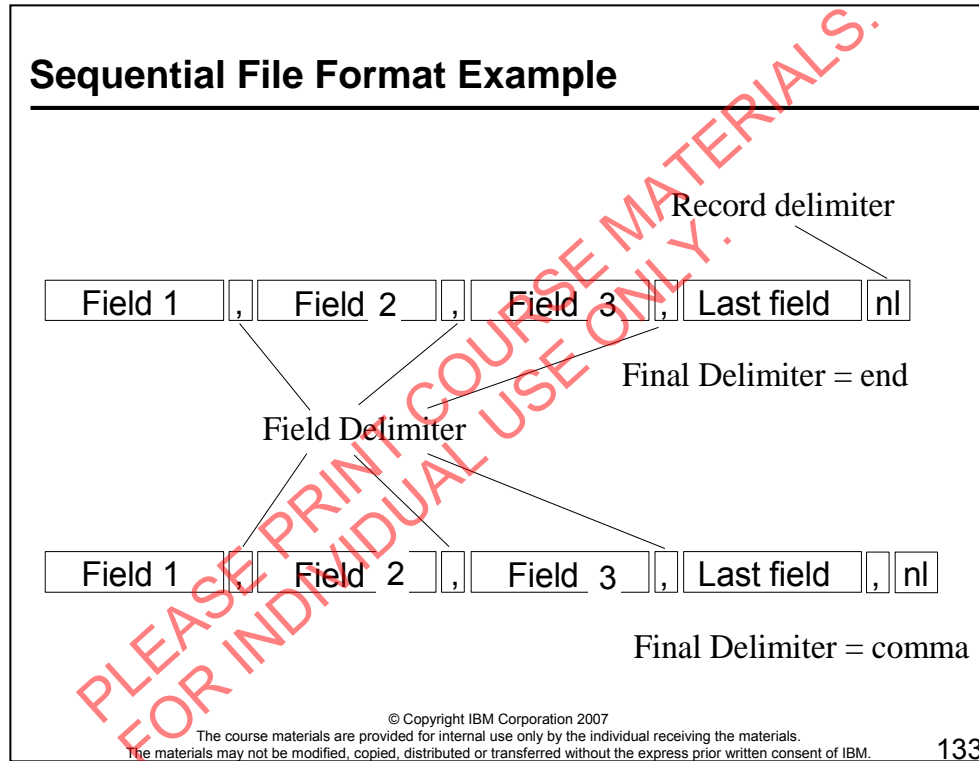
Features of Sequential File Stage

- Normally executes in sequential mode
- Executes in parallel when reading multiple files
- Can use multiple readers within a node
 - Reads chunks of a single file in parallel
- The stage needs to be told:
 - How file is divided into rows (record format)
 - How row is divided into columns (column format)

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

132



133

This shows the format of one type of sequential file. Delimiters separate columns. Fields all columns are defined my delimiters. Similarly, records are defined by terminating characters.

Sequential File Stage Rules

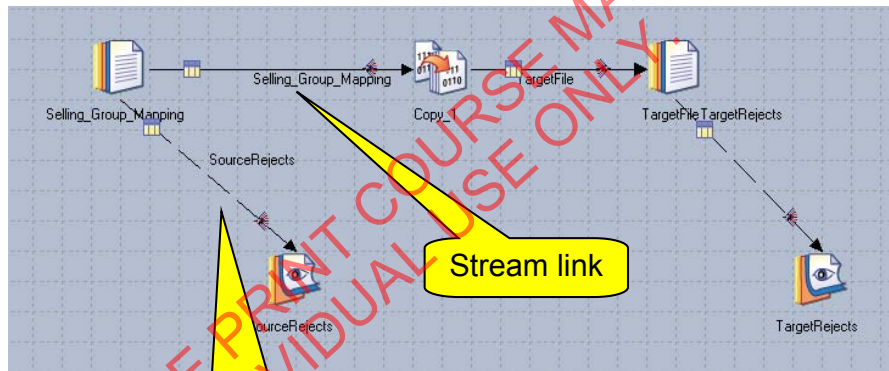
- One input link
- One stream output link
- Optionally, one reject link
 - Will reject any records not matching metadata in the column definitions
 - Example: You specify three columns separated by commas, but the row that's read had no commas in it
 - Example: The second column is designated a decimal in the Table Definition, but contains alphabetic characters

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

134

Job Design Using Sequential Stages

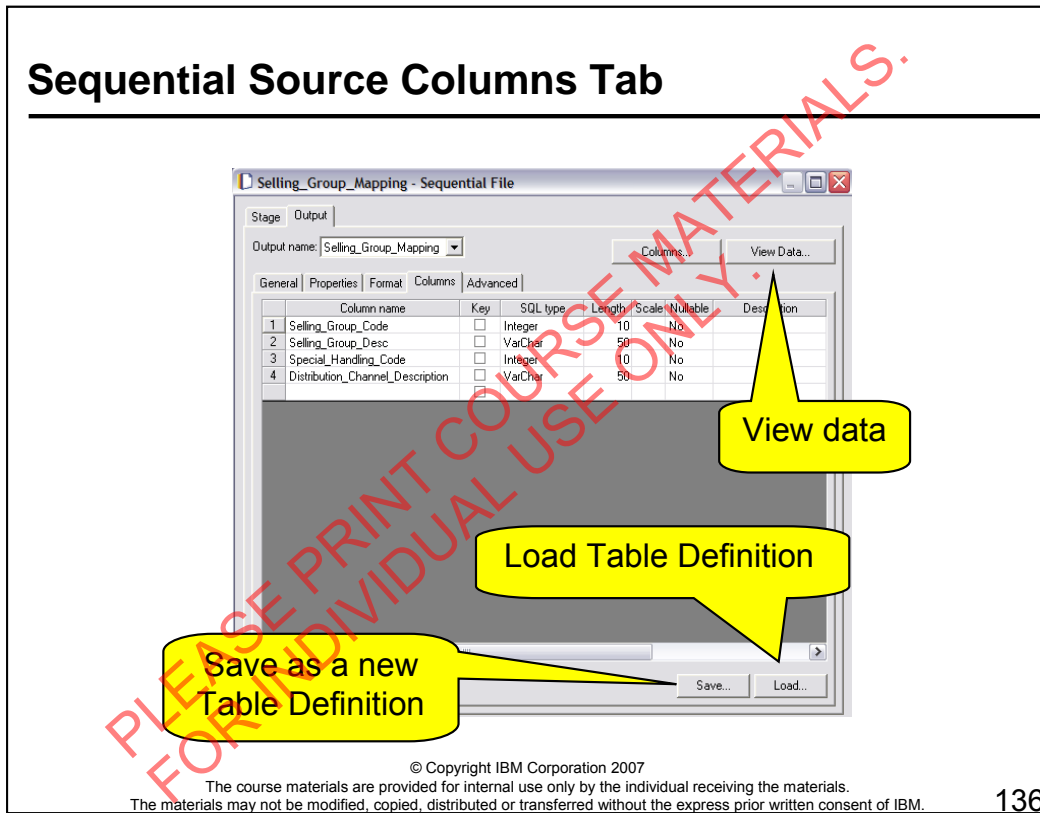


© Copyright IBM Corporation 2007
 The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

135

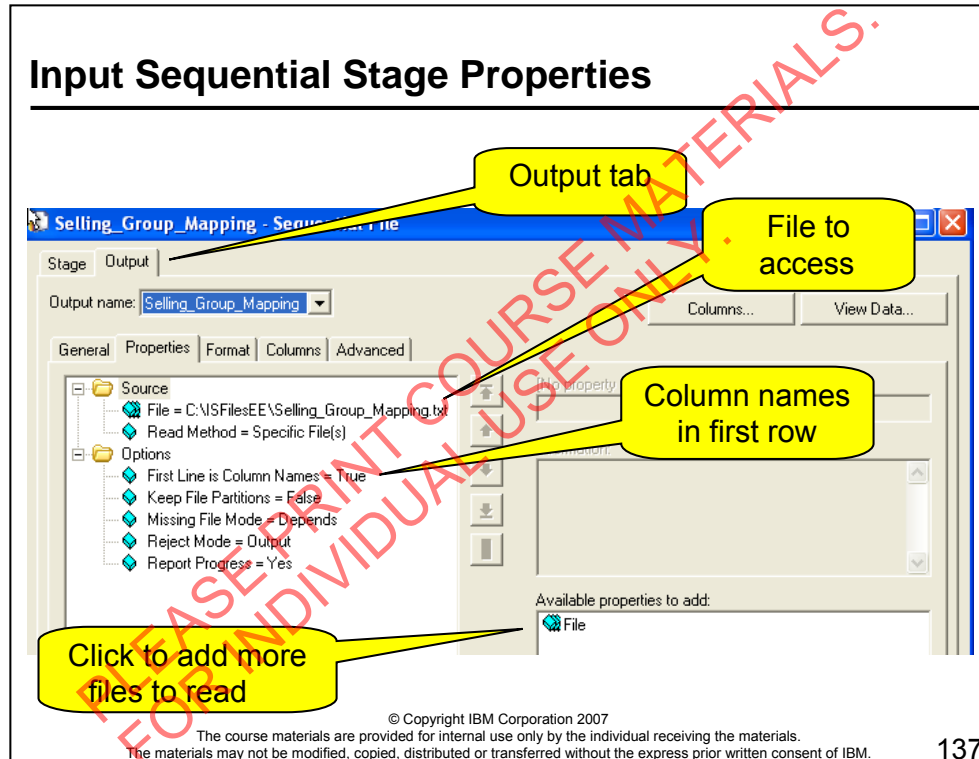
The source Sequential stage has a stream output link and a reject output link. The target Sequential stage has an input link and a reject, output link.

Sequential Source Columns Tab



136

The Columns tab has these features: Column definitions, View Data, Load button. The Save button is used to save the column definitions into the Repository as a new Table Definition.

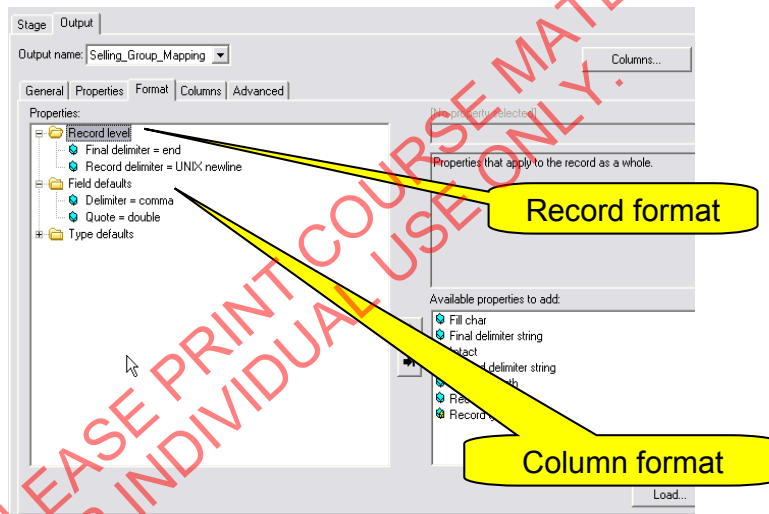


If specified individually, you can make a list of files that are unrelated in name.

If you select “read method” and choose file pattern, you effectively select an undetermined number of files.

Set the Report Progress property to Yes to show percent of processing during the job run.

Format Tab



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

138

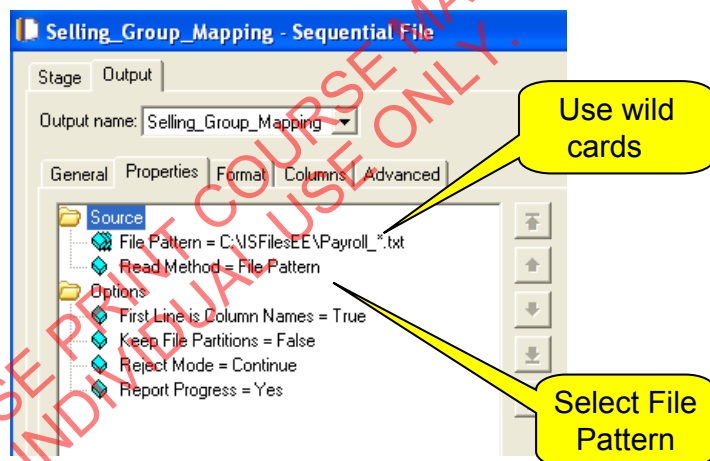
DataStage needs to know:

How a file is divided into rows

How a row is divided into columns

Column properties set on this tab are defaults for each column; they can be overridden at the column level (from columns tab).

Reading Using a File Pattern

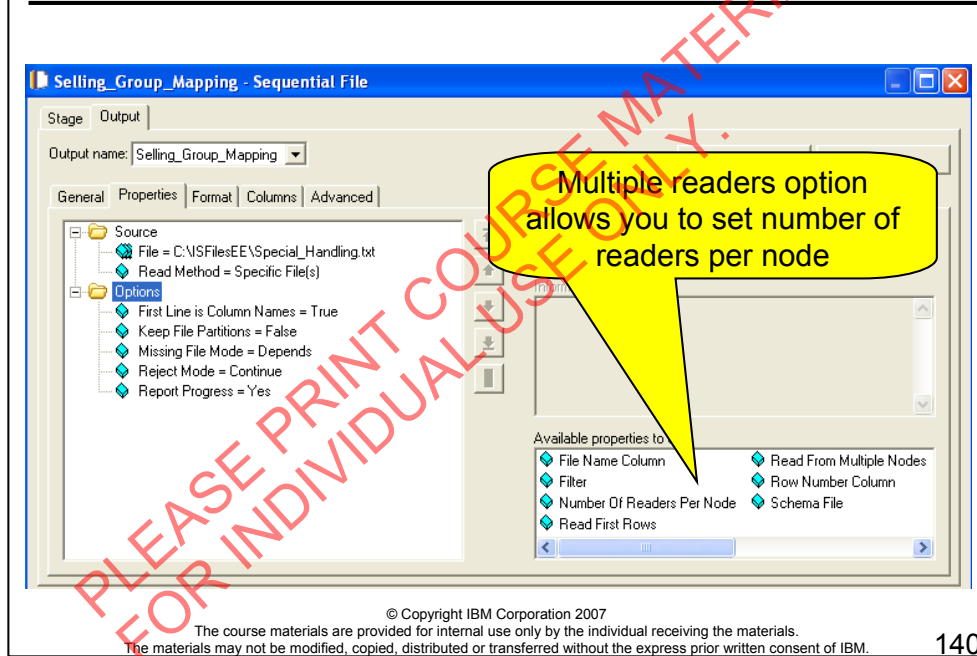


© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

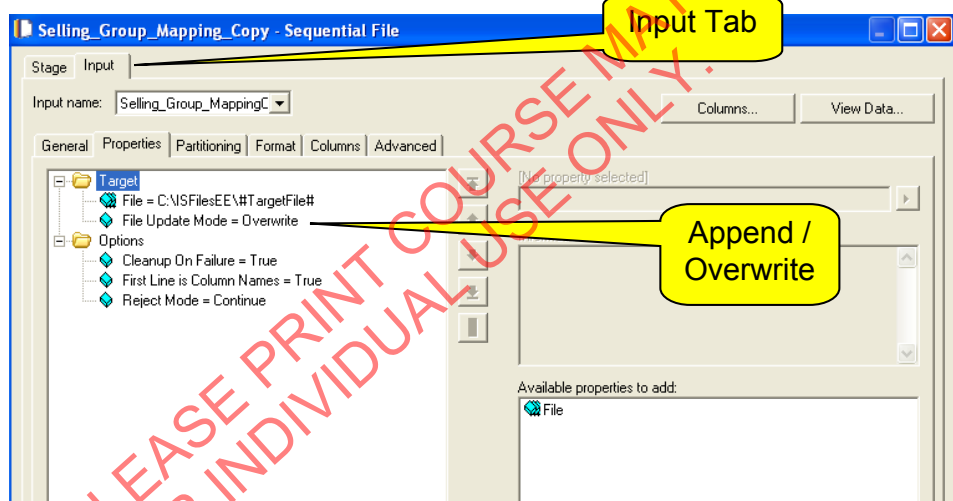
139

Properties - Multiple Readers



This option allows you to read a single sequential file in parallel. However, the row order is not maintained. Therefore if input rows need to be identified, this option can only be used if the data itself provides a unique identifier. This works for both fixed-length and variable-length records.

Sequential Stage As a Target



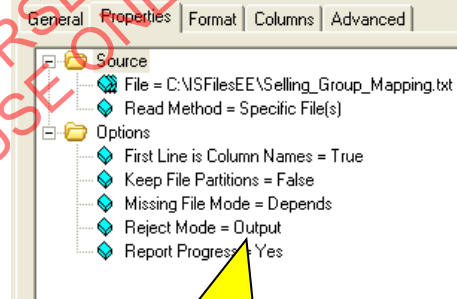
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

141

Reject Link

- Reject mode =
 - Continue: Continue reading records
 - Fail: Abort job
 - Output: Send down output link
- In a source stage
 - All records not matching the metadata (column definitions) are rejected
- In a target stage
 - All records that fail to be written for any reason
- Rejected records consist of one column, datatype = raw



Reject mode property

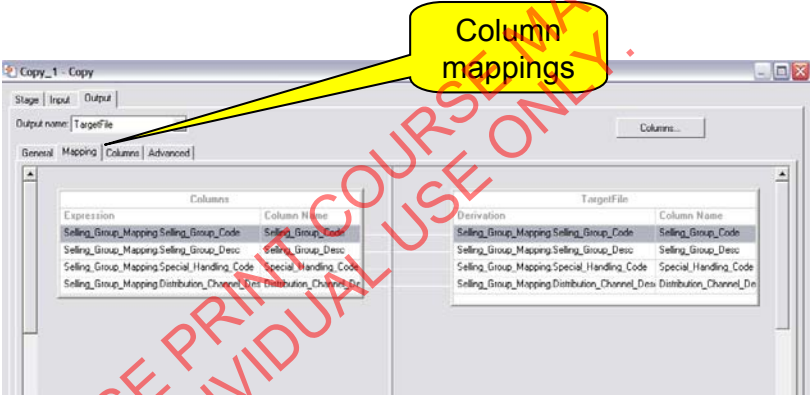
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

142

The sequential stage can have a single reject link. This is typically used when you are writing to a file and provides a location where records that have failed to be written to a file for some reason can be sent. When you are reading files, you can use a reject link as a destination for rows that do not match the expected column definitions.

Inside the Copy Stage



PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Reading and Writing NULL Values to a Sequential File

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

144

Working with NULLs

- Internally, NULL is represented by a special value outside the range of any existing, legitimate values
- If NULL is written to a non-nullable column, the job will abort
- Columns can be specified as nullable
 - NULLs can be written to nullable columns
- You must “handle” NULLs written to nullable columns in a Sequential File stage
 - You need to tell DataStage what value to write to the file
 - Unhandled rows are rejected
- In a Sequential File source stage, you can specify values you want DataStage to convert to NULLs

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

145

Specifying a Value for NULL

Column name:
Special_Handling_Code

SQL type: Integer Length: 10

Nullable: Yes

Description:

Server: Parallel

Field type: int32
Extended (Unsigned)

Vector occurs:

Properties:

- Field level
 - Quote = none
- Integer type
- Nullable
 - Null field value = 1

Nullable column

Added property

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

146

DataSet Stage

DataSet Stage

PLEASE PRINT COURSE MATERIALS
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

147

Data Set

- Binary data file
- Preserves partitioning
 - Component dataset files are written to each partition
- Suffixed by .ds
- Referred to by a header file
- Managed by Data Set Management utility from GUI (Manager, Designer, Director)
- Represents persistent data
- Key to good performance in set of linked jobs
 - No import / export conversions are needed
 - No repartitioning needed
- Accessed using DataSet stage
- Implemented with two types of components:
 - Descriptor file:
 - contains metadata, data location, but NOT the data itself
 - Data file(s)
 - contain the data
 - multiple files, one per partition (node)

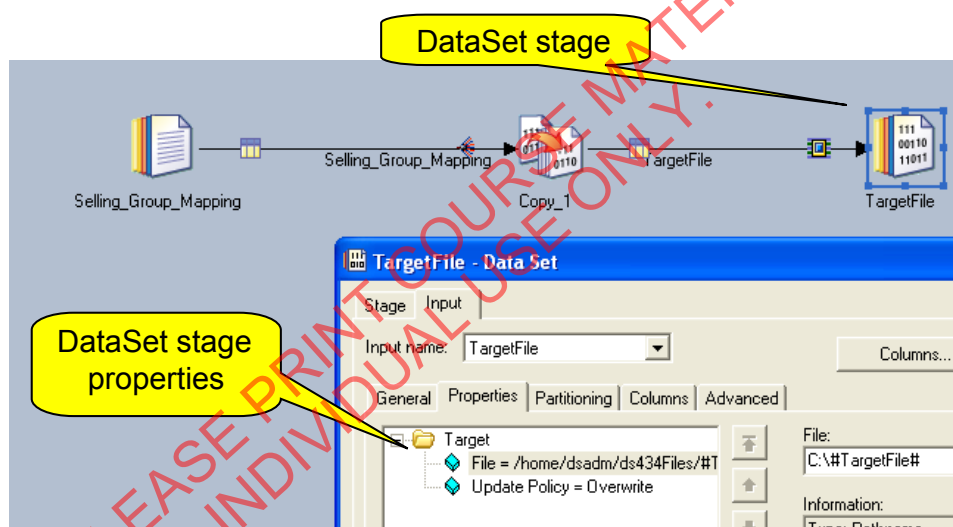


© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

148

Data sets represent persistent data maintained in the internal format.

Job With DataSet Stage



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

149

Data Set Management Utility

Data Set Properties

| | | | |
|----------|--------------------------------------|-------------------|------|
| Name: | TargetDataSet.ds | Total Records: | 47 |
| Path: | C:\SFFilesEE | Total 32K Blocks: | 2 |
| Version: | ORCHESTRATE V8.0.0 DM Block Format 6 | Total Bytes: | 5640 |
| Created: | 11/15/2006 12:06:14 | | |

Partitions:

| # | Node | Records | Blocks | Bytes |
|---|-------|---------|--------|-------|
| 0 | node1 | 24 | 1 | 2880 |
| 1 | node2 | 23 | 1 | 2760 |

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

The window is available (data sets management) from Designer and Director. In Designer, click Tools>Data Set Management to open this window.

Data and Schema Displayed

Data viewer

Record Schema

```
record
{
  Selling_Group_Code: int32;
  Selling_Group_Desc: string(max=50);
  Special_Handling_Code: int32;
  Distribution_Channel_Description: string(max=50);
}
```

Schema describing the format of the data

| Selling_Group... | Selling_Group_Desc | Special_Handling_Code | Distribution_Channel_Des |
|------------------|--------------------------|-----------------------|--------------------------|
| 100000 | SG005 FRESHNESS-MILLE | 6 | Other |
| 550000 | SG055 LIVE SWINE | 6 | Other |
| 870000 | SG087 FREEZERS | 6 | Other |
| 1150000 | SG115 BURGER KING | 2 | Food Service |
| 1200000 | SG120 RETAIL P/P | 1 | Retail |
| 1230000 | SG123 STORE DODOR | 2 | Food Service |
| 1380000 | SG138 MCDONALDS | 2 | Food Service |
| 1390000 | SG139 SPECIALTY FOODS | 4 | Specialty Products |
| 1400000 | SG140 CLUB STORES | 1 | Retail |
| 1420000 | SG142 GREATER CHINA | 3 | International |
| 1440000 | SG144 JAPAN | 3 | International |
| 1480000 | SG148 R.E.M.A. | 3 | International |
| 3110000 | SG311 INTERNATIONAL... | 6 | Other |
| 4010000 | SG401 KITCH INDUSTRIES | 2 | Food Service |
| 4080000 | SG408 HOSPITALITY SPE... | 6 | Other |
| 4230000 | SG423 MEXICAN ORIGIN... | 6 | Other |
| 4430000 | SG443 MALLARD'S SALES | 6 | Other |
| 5230000 | SG523 ENTREE | 1 | Retail |
| 5950000 | SG595 COBB VANTRESS | 6 | Retail |
| 6110000 | SG611 REFRIGERATED P | 1 | Retail |

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

File Set Stage

- Use to read and write to filesets
- Files suffixed by .fs
- Files are similar to a dataset
 - Partitioned
 - Implemented with header file and data files
- How filesets differ from datasets
 - Data files are text files
 - Hence readable by external applications
 - Datasets have a proprietary data format which may change in future DataStage versions



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

152

Number of raw data files depends on: the configuration file – more on configuration files later.

Checkpoint

1. List three types of file data
2. What makes datasets perform better than other types of files in parallel jobs
3. What is the difference between a data set and a file set?

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

153

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. Sequential, dataset, complex flat files
2. They are partitioned and they store data in the native parallel format
3. Both are partitioned. Data sets store data in a binary format not readable by user applications. File sets are readable.

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

154

Unit summary

Having completed this unit, you should be able to:

- Understand the stages for accessing different kinds of file data
- Sequential File stage
- Data Set stage
- Create jobs that read from and write to sequential files
- Create Reject links
- Work with NULLs in sequential files
- Read from multiple files using file patterns
- Use multiple readers

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

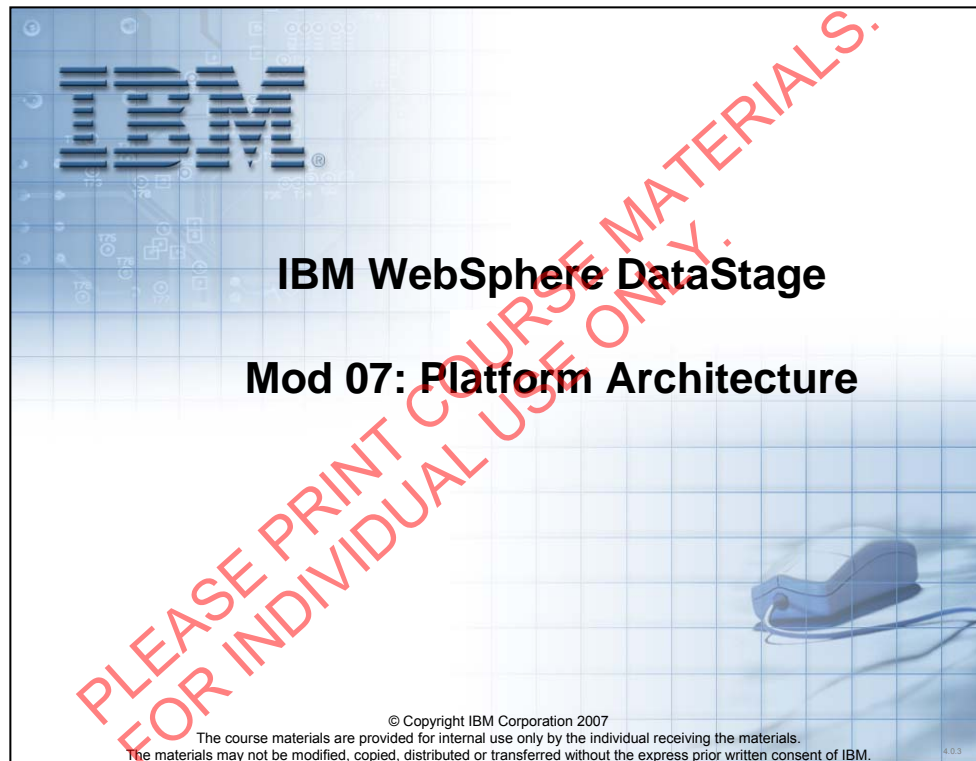
155

Notes:

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

156



Unit objectives

After completing this unit, you should be able to:

- Describe parallel processing architecture
- Describe pipeline parallelism
- Describe partition parallelism
- List and describe partitioning and collecting algorithms
- Describe configuration files
- Describe the parallel job compilation process
- Explain OSH
- Explain the Score

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

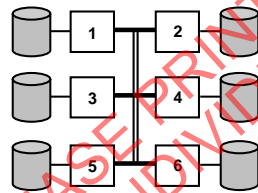
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

158

Notes:

Key Parallel Job Concepts

- Parallel processing:
 - Executing the job on multiple CPUs
- Scalable processing:
 - Add more resources (CPUs and disks) to increase system performance



- Example system: 6 CPUs (processing nodes) and disks
- Scale up by adding more CPUs
- Add CPUs as individual nodes or to an SMP system

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

159

Parallel processing is the key to building jobs that are highly scalable.

The parallel engine uses the processing node concept. A processing node is a CPU or an SMP, or a board on an MPP.

Scalable Hardware Environments

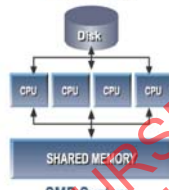
Dedicated Disk



Uniprocessor

- Single CPU
- Dedicated memory & disk

Shared Disk

SMP System
(Symmetric Multiprocessor)

- SMP
- Multi-CPU (2-64+)
- Shared memory & disk

Shared Nothing

MPP, Clustered Systems
(Massively Parallel Processing)

- GRID / Clusters
 - Multiple, multi-CPU systems
 - Dedicated memory per node
 - Typically SAN-based shared storage
- MPP
 - Multiple nodes with dedicated memory, storage
- 2 – 1000's of CPUs

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

160

DataStage parallel jobs are designed to be platform-independent – a single job, if properly designed, can run across resources within a single machine (SMP) or multiple machines (cluster, GRID, or MPP architectures).

While parallel jobs can run on a single-CPU environment, it is designed to take advantage of parallel platforms.

Pipeline Parallelism



- Transform, clean, load processes execute simultaneously
- Like a conveyor belt moving rows from process to process
 - Start downstream process while upstream process is running
- Advantages:
 - Reduces disk usage for staging areas
 - Keeps processors busy
- Still has limits on scalability

© Copyright IBM Corporation 2007
 The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Partition Parallelism

- Divide the incoming stream of data into subsets to be separately processed by an operation
 - Subsets are called partitions (nodes)
- Each partition of data is processed by the same operation
 - E.g., if operation is Filter, each partition will be filtered in exactly the same way
- Facilitates near-linear scalability
 - 8 times faster on 8 processors
 - 24 times faster on 24 processors
 - This assumes the data is evenly distributed

© Copyright IBM Corporation 2007

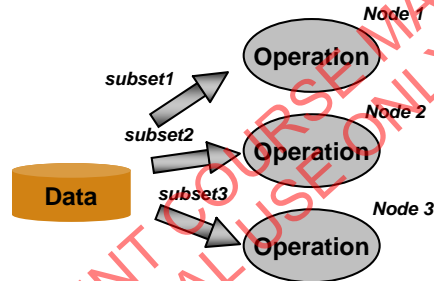
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

162

Partitioning breaks a dataset into smaller sets. This is a key to scalability. However, the data needs to be evenly distributed across the partitions; otherwise, the benefits of partitioning are reduced.

It is important to note that what is done to each partition of data is the same. How the data is processed or transformed is the same.

Three-Node Partitioning



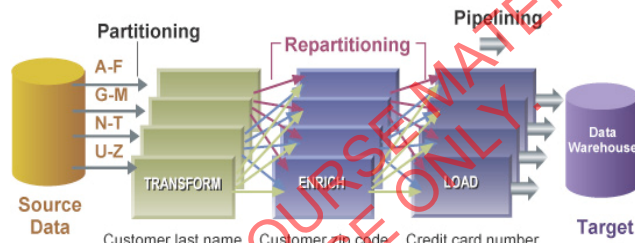
- Here the data is partitioned into three partitions
- The operation is performed on each partition of data separately and in parallel
- If the data is evenly distributed, the data will be processed three times faster

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

163

Parallel Jobs Combine Partitioning and Pipelining



- Within parallel jobs pipelining, partitioning, and repartitioning are automatic
- Job developer only identifies:
 - Sequential vs. parallel mode (by stage)
 - Default for most stages is parallel mode
 - Method of data partitioning
 - The method to use to distribute the data into the available nodes
 - Method of data collecting
 - The method to use to collect the data from the nodes into a single node
 - Configuration file
 - Specifies the nodes and resources

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

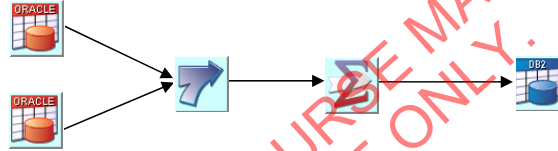
164

By combining both pipelining and partitioning DataStage creates jobs with higher volume throughput.

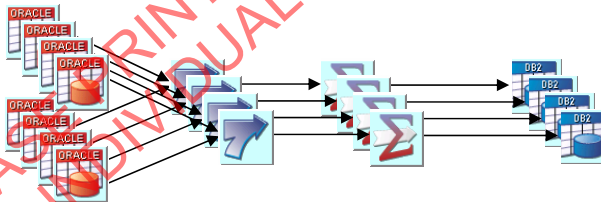
The configuration file drives the parallelism by specifying the number of partitions.

Job Design v. Execution

User assembles the flow using DataStage Designer



... at runtime, this job runs in parallel for any configuration
(1 node, 4 nodes, N nodes)



No need to modify or recompile the job design!

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

165

Much of the parallel processing paradigm is hidden from the programmer. The programmer simply designates the process flow, as shown in the upper portion of this diagram. The parallel engine, using the definitions in the configuration file, will actually execute processes that are partitioned and parallelized, as illustrated in the bottom portion.

Configuration File

- Configuration file separates configuration (hardware / software) from job design
 - Specified per job at runtime by \$APT_CONFIG_FILE
 - Change hardware and resources without changing job design
- Defines nodes with their resources (need not match physical CPUs)
 - Dataset, Scratch, Buffer disk (file systems)
 - Optional resources (Database, SAS, etc.)
 - Advanced resource optimizations
 - “Pools” (named subsets of nodes)
- Multiple configuration files can be used different occasions of job execution
 - Optimizes overall throughput and matches job characteristics to overall hardware resources
 - Allows runtime constraints on resource usage on a per job basis

© Copyright IBM Corporation 2007

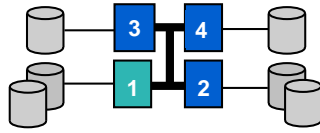
The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

166

DataStage jobs can point to different configuration files by using job parameters. Thus, a job can utilize different hardware architectures without being recompiled. It can pay to have a 4-node configuration file running on a 2 processor box, for example, if the job is “resource bound.” We can spread disk I/O among more controllers.

Example Configuration File



Key points:

1. Number of nodes defined
2. Resources assigned to each node. Their order is significant.
3. Advanced resource optimizations and configuration (named pools, database, SAS)

```
{
  node "n1" {
    fastname "s1"
    pool "" "n1" "s1" "app2" "sort"
    resource disk "/orch/n1/d1" {}
    resource disk "/orch/n1/d2" {"bigdata"}
    resource scratchdisk "/temp" {"sort"}
  }
  node "n2" {
    fastname "s2"
    pool "" "n2" "s2" "appl"
    resource disk "/orch/n2/d1" {}
    resource disk "/orch/n2/d2" {"bigdata"}
    resource scratchdisk "/temp" {}
  }
  node "n3" {
    fastname "s3"
    pool "" "n3" "s3" "appl"
    resource disk "/orch/n3/d1" {}
    resource scratchdisk "/temp" {}
  }
  node "n4" {
    fastname "s4"
    pool "" "n4" "s4" "appl"
    resource disk "/orch/n4/d1" {}
    resource scratchdisk "/temp" {}
  }
}
```

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

167

This example shows a typical configuration file. Pools can be applied to nodes or other resources. Note the curly braces following some disk resources.

Following the keyword “node” is the name of the node (logical processing unit).

The order of resources is significant. The first disk is used before the second, and so on.

Keywords, such as “sort” and “bigdata”, when used, restrict the signified processes to the use of the resources that are identified. For example, “sort” restricts sorting to node pools and scratchdisk resources labeled “sort”.

Database resources (not shown here) can also be created that restrict database access to certain nodes.

Question: Can objects be constrained to CPUs? No, a request is made to the operating system and the operating system chooses the CPU.

Partitioning and Collecting

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

168

Partitioning and Collecting

- Partitioning breaks incoming rows into multiple streams of rows (one for each node)
- Each partition of rows is processed separately by the stage/operator
- Collecting returns partitioned data back to a single stream
- Partitioning / Collecting is specified on stage input links

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

169

Partitioning / Collecting Algorithms

- Partitioning algorithms include:
 - Round robin
 - Random
 - Hash: Determine partition based on key value
 - Requires key specification
 - Modulus
 - Entire: Send all rows down all partitions
 - Same: Preserve the same partitioning
 - Auto: Let DataStage choose the algorithm
- Collecting algorithms include:
 - Round robin
 - Auto
 - Collect first available record
 - Sort Merge
 - Read in by key
 - Presumes data is sorted by the key in each partition
 - Builds a single sorted stream based on the key
 - Ordered
 - Read all records from first partition, then second, ...

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Keyless V. Keyed Partitioning Algorithms

- Keyless: Rows are distributed independently of data values
 - Round Robin
 - Random
 - Entire
 - Same
- Keyed: Rows are distributed based on values in the specified key
 - Hash: Partition based on key
 - Example: Key is State. All “CA” rows go into the same partition; all “MA” rows go in the same partition. Two rows of the same state never go into different partitions
 - Modulus: Partition based on modulus of key divided by the number of partitions. Key is a numeric type.
 - Example: Key is OrderNumber (numeric type). Rows with the same order number will all go into the same partition.
 - DB2: Matches DB2 EEE partitioning

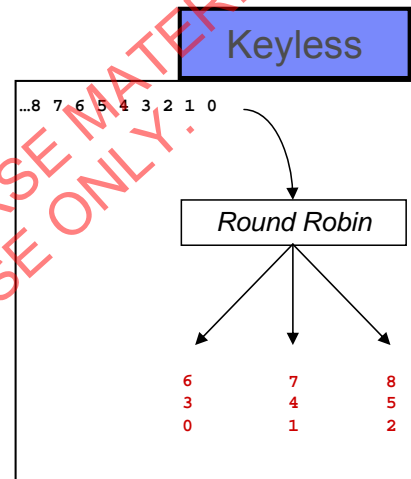
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

171

Round Robin and Random Partitioning

- Keyless partitioning methods
- Rows are evenly distributed across partitions
 - Good for initial import of data if no other partitioning is needed
 - Useful for redistributing data
- Fairly low overhead
- Round Robin assigns rows to partitions like dealing cards
 - Row/Partition assignment will be the same for a given \$APT_CONFIG_FILE
- Random has slightly higher overhead, but assigns rows in a non-deterministic fashion between job runs



© Copyright IBM Corporation 2007

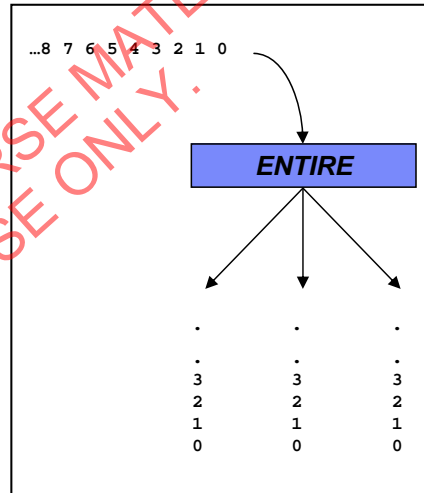
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

172

ENTIRE Partitioning

- Each partition gets a complete copy of the data
 - Useful for distributing **lookup** and **reference data**
 - May have performance impact in MPP / clustered environments
 - On SMP platforms, Lookup stage (only) uses shared memory instead of duplicating ENTIRE reference data
 - On MPP platforms, each server uses shared memory for a single local copy
- ENTIRE is the default partitioning for Lookup reference links with "Auto" partitioning
 - On SMP platforms, it is a good practice to set this explicitly on the *Normal/*Lookup reference link(s)

keyless



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

173

HASH Partitioning

Keyed

- Keyed partitioning method
- Rows are distributed according to the values in key columns
 - Guarantees that rows with same key values go into the same partition
 - Needed to prevent matching rows from "hiding" in other partitions
 - E.g. Join, Merge, Remove Duplicates, ...
 - Partition distribution is relatively equal if the data across the source key column(s) is evenly distributed

Values of key column

...0 3 2 1 0 2 3 2 1 1

| | | |
|---|---|---|
| 0 | 1 | 2 |
| 3 | 1 | 2 |
| 0 | 1 | 2 |
| 3 | | 2 |

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

174

For certain stages (Remove Duplicates, Join, Merge) to work correctly in parallel, the user must override the default (Auto) to Hash or a variant: Range, Modulus.

Here the numbers are no longer row IDs, but values of key column.

- Hash guarantees that all the rows with key value 3 end up in the same partition.

- Hash does not guarantee "continuity": here, 3s are bunched with 0s, not with neighboring value 2.

This is an expensive version of Hash, "Range," that guarantees continuity.

- Hash does not guarantee load balance.

Make sure key column(s) takes on enough values to distribute data across available partitions.

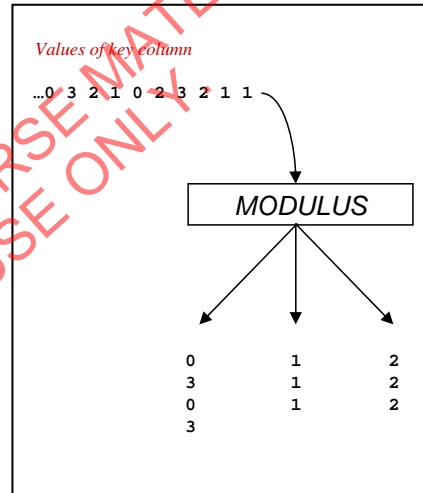
("gender" would be a poor choice of key...)

Modulus Partitioning

Keyed

- Keyed partitioning method
- Rows are distributed according to the values in one integer key column
 - Uses modulus

$$\text{partition} = \text{MOD}(\text{key_value} / \text{\#partitions})$$
- Faster than HASH
- Guarantees that rows with identical key values go in the same partition
- Partition size is relatively equal if the data within the key column is evenly distributed



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
 The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

175

Auto Partitioning

- DataStage inserts partition components as necessary to ensure correct results
 - Before any stage with “Auto” partitioning
 - Generally chooses *ROUND-ROBIN* or *SAME*
 - Inserts *HASH* on stages that require matched key values (e.g. Join, Merge, Remove Duplicates)
 - Inserts *ENTIRE* on Normal (not Sparse) Lookup reference links
 - NOT always appropriate for MPP/clusters
- Since DataStage has limited awareness of your data and business rules, explicitly specify *HASH* partitioning when needed
 - DataStage has no visibility into Transformer logic
 - Hash is required before Sort and Aggregator stages
 - DataStage sometimes inserts un-needed partitioning
 - Check the log

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

176

Auto generally chooses Round Robin when going from sequential to parallel. It generally chooses Same when going from parallel to parallel.

Since DataStage has limited awareness of your data and business rules, explicitly specify *HASH* partitioning when needed, that is, when processing requires groups of related records.

Partitioning Requirements for Related Records

- Misplaced records
 - Using Aggregator stage to sum customer sales by customer number
 - If there are 25 customers, 25 records should be output
 - But suppose records with the same customer numbers are spread across partitions
 - This will produce more than 25 groups (records)
 - Solution: Use hash partitioning algorithm
- Partition imbalances
 - If all the records are going down only one of the nodes, then the job is in effect running sequentially

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

177

Unequal Distribution Example

- Same key values are assigned to the same partition
- Hash on UName, with 2-node config file

| Source Data | ID | UName | PName | Address |
|-------------|----|-------|---------|---------------------|
| | 1 | Ford | Henry | 66 Edison Avenue |
| | 2 | Ford | Clara | 66 Edison Avenue |
| | 3 | Ford | Edsel | 7900 Jefferson |
| | 4 | Ford | Eleanor | 7900 Jefferson |
| | 5 | Dodge | Horace | 17840 Jefferson |
| | 6 | Dodge | John | 75 Boston Boulevard |
| | 7 | Ford | Henry | 4901 Evergreen |
| | 8 | Ford | Clara | 4901 Evergreen |
| | 9 | Ford | Edsel | 1100 Lakeshore |
| | 10 | Ford | Eleanor | 1100 Lakeshore |

| Part 0 | ID | UName | PName | Address |
|--------|----|-------|--------|---------------------|
| | 5 | Dodge | Horace | 17840 Jefferson |
| | 6 | Dodge | John | 75 Boston Boulevard |

| Partition 1 | ID | UName | PName | Address |
|-------------|----|-------|---------|------------------|
| | 1 | Ford | Henry | 66 Edison Avenue |
| | 2 | Ford | Clara | 66 Edison Avenue |
| | 3 | Ford | Edsel | 7900 Jefferson |
| | 4 | Ford | Eleanor | 7900 Jefferson |
| | 7 | Ford | Henry | 4901 Evergreen |
| | 8 | Ford | Clara | 4901 Evergreen |
| | 9 | Ford | Edsel | 1100 Lakeshore |
| | 10 | Ford | Eleanor | 1100 Lakeshore |
| | | | | |

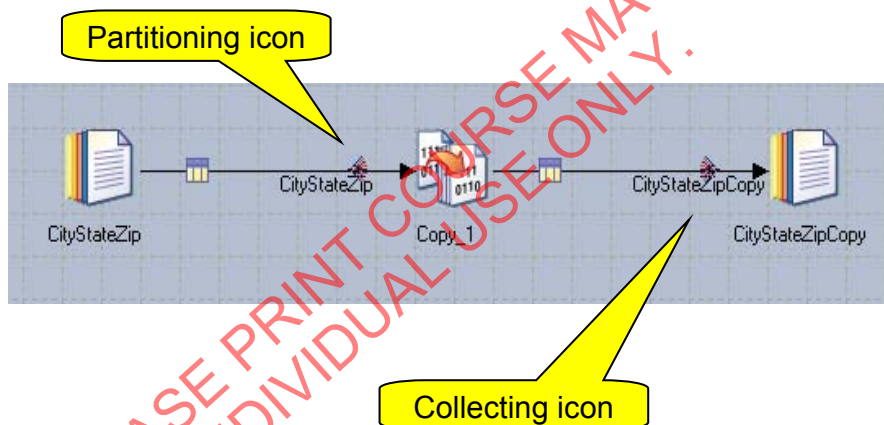
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

178

This is an example of unequal distribution of rows down the different partitions. Partition distribution matches source data distribution. In this example, number of distinct hash key values limits parallelism!

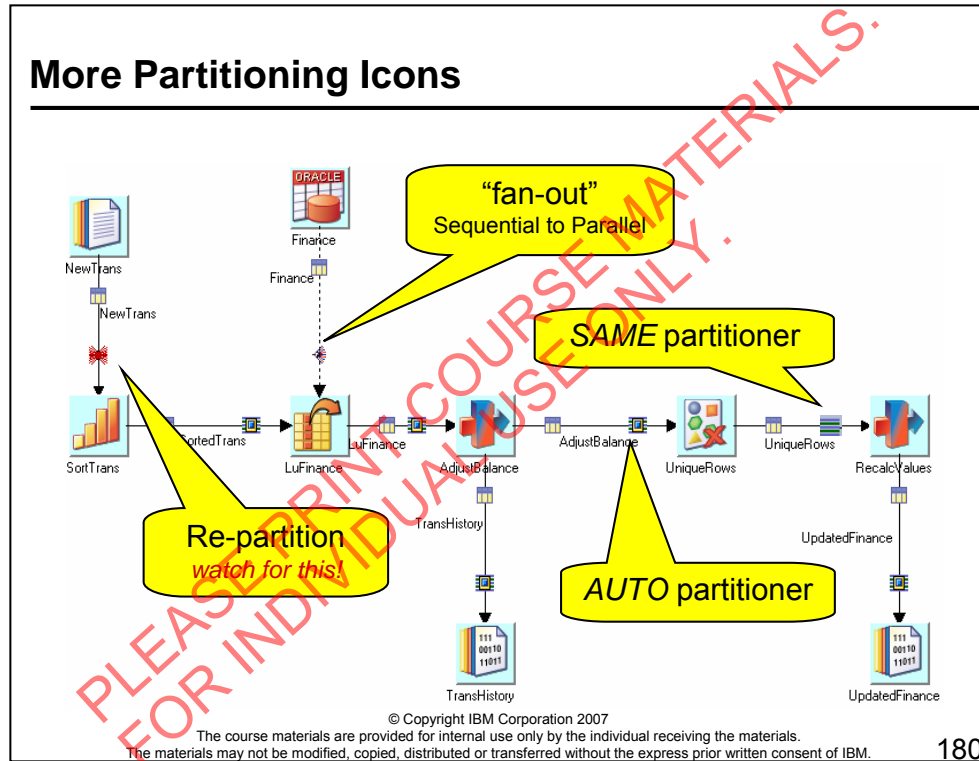
Partitioning / Collecting Link Icons



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

179



Partitioners and collectors have no stage nor icons of their own.
 They live live on input links of stages running in parallel / sequentially, respectively.
 Link markings indicate their presence.

S----->S (no Marking)
 S----(fan out)--->P (partitioner)
 P----(fan in) ---->S (collector)
 P----(box)----->P (no reshuffling: partitioner using "SAME" method)
 P----(bow tie)--->P (reshuffling: partitioner using another method)

Collectors = inverse partitioners

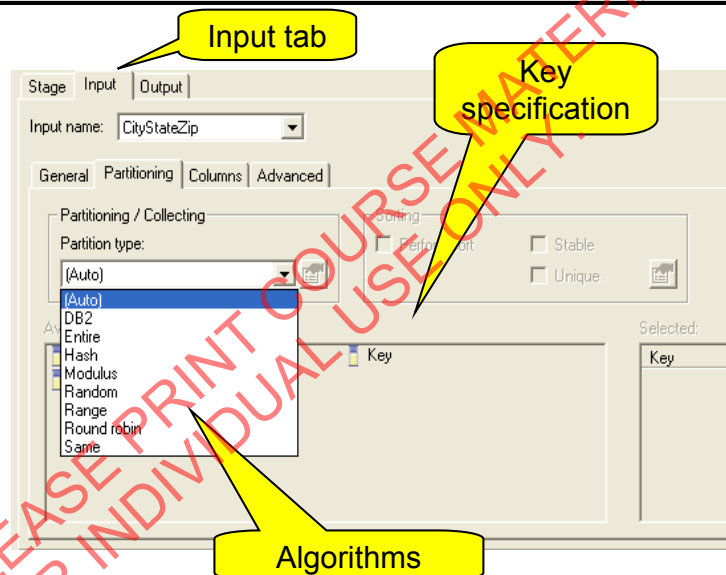
recollect rows from partitions into a single input stream to a sequential stage

They are responsible for some surprising behavior:

The default (Auto) is "eager" to output rows and typically causes non-determinism:
 row order may vary from run to run with identical input.

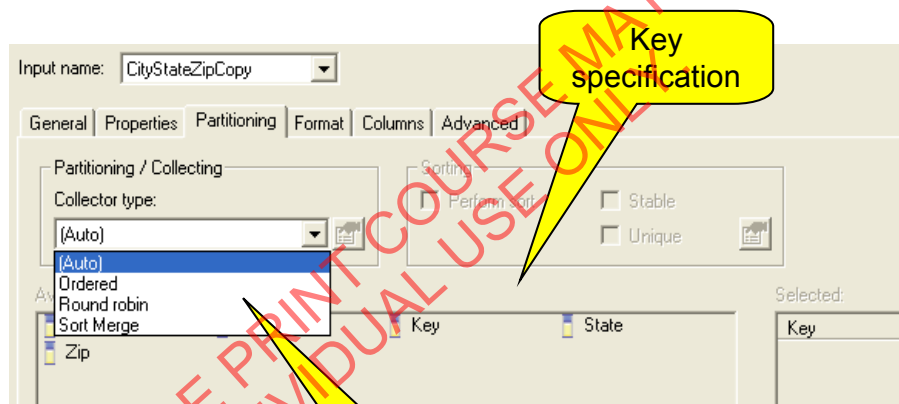
In the example of repartitioning, the NewTrans file is using multiple readers, so the data is being read into multiple partitions. Therefore, a reshuffling occurs at the Sort stage.

Partitioning Tab



© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Collecting Specification



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

182

Runtime Architecture

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

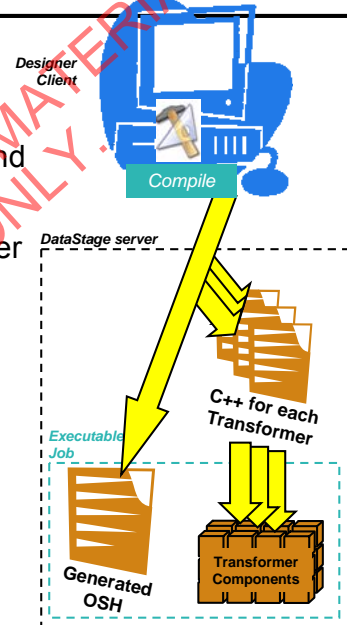
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

183

Parallel Job Compilation

What gets generated:

- OSH: A kind of script
- OSH represents the design data flow and stages
 - Stages become OSH operators
- Transform operator for each Transformer
 - A custom operator built during the compile
 - Compiled into C++ and then to corresponding native operators
 - Thus a C++ compiler is needed to compile jobs with a Transformer stage



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

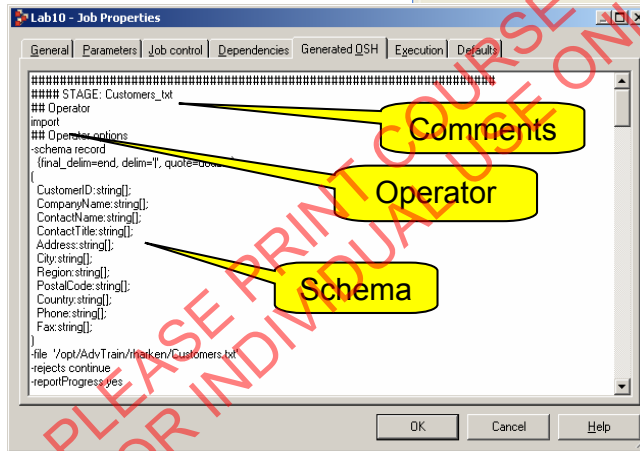
184

Compilation generates OSH (Orchestrate Shell Script) and also C++ code for any Transformer stages used.

For each Transformer, DataStage builds a C++ Framework operator that is then compiled. So when a job contains a Transformer, it takes longer to compile (but not to run).

Generated OSH

Enable viewing of generated OSH in Administrator:



OSH is visible in:

- Job properties
- Job run log
- View Data
- Table Defs

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

185

You can view generated OSH through DataStage Designer. This provides an overview of the OSH that will be generated. It is important to note, however, that this OSH will go through some changes for optimization and execution.

Stage to Operator Mapping Examples

- Sequential File
 - Source: import
 - Target: export
- DataSet: copy
- Sort: tsort
- Aggregator: group
- Row Generator, Column Generator, Surrogate Key Generator: generator
- Oracle
 - Source: oraread
 - Sparse Lookup: oralookup
 - Target Load: orawrite
 - Target Upsert: oraupsert

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

186

The stages on the diagram do not necessarily map one-to-one to operators. For example, the Sequential stage when used as a source is mapped to the import operator. When used as a target it is mapped to the export operator.

The converse is also true. Different stages can be mapped to a single operator. For example, Row Generator and Column Generator are both mapped to the generator operator.

Generated OSH Primer

Generated OSH for first 2 stages

- Comment blocks introduce each operator
 - Operator order is determined by the order stages were added to the canvas
- Syntax
 - Operator name
 - Schema
 - Operator options (“-name value” format)
 - Input (indicated by n< where n is the input #)
 - Output (indicated by n> where n is the output #)
 - may include modify
- For every operator, input and/or output datasets are numbered sequentially starting from 0. E.g.:
 - op1 0> dst
 - op1 1< src
- Virtual datasets are generated to connect operators



```
#####
### STAGE: Row_Generator_0
### Operator
### Operator options
-scheme record
-a int32;
-b string(max=12);
-c nullable decimal(10,2) (nulls=10);
-records 50000

### General options
[ident('Row_Generator_0'); jobmon_id('Row_Generator_0')]
### Inputs
0< [] 'Row_Generator_0.lnk_gen.v'
### Outputs
0> [] 'Row_Generator_0.lnk_gen.v'

#####
### STAGE: SortSt
### Operator
### Operator options
-key 'a'
-asc

### General options
[ident('SortSt'); jobmon_id('SortSt'); par]
### Inputs
0< 'Row_Generator_0.lnk_gen.v'
### Outputs
0> [modify (
  keep
  a,b,c;
)] 'SortSt.lnk_sorted.v'
```

Virtual dataset is used to connect output of one operator to input of another

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

187

The actual execution order of operators is dictated by input/output designators, not by placement on the diagram.

The datasets connect the osh operators. These are “virtual datasets”, that is, in-memory data flows.

Link names are used in dataset names. So good practice is to name links meaningfully.

Job Score

- Generated from OSH and configuration file
- Think of “Score” as in musical score, not game score
- Assigns nodes for each operator
- Inserts sorts and partitioners as needed
- Defines connection topology (virtual datasets) between adjacent operators
- Inserts buffer operators to prevent deadlocks
- Defines the actual processes
 - Where possible, multiple operators are combined into a single process

© Copyright IBM Corporation 2007

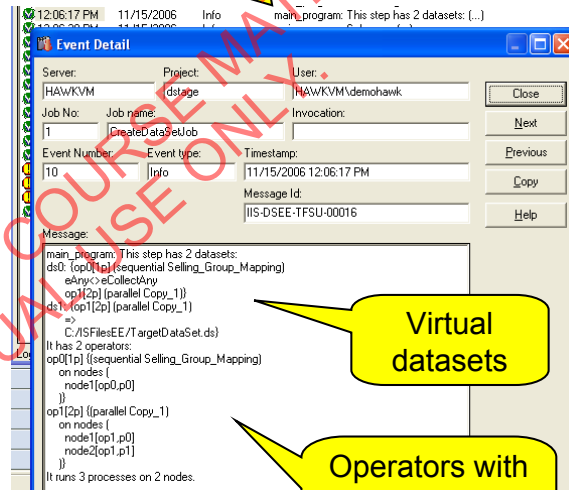
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

188

Viewing the Job Score

- Set \$APT_DUMP_SCORE to output the Score to the job log
- To identify the Score dump, look for “main program: This step has 2 datasets: This step ...”
 - You don't see anywhere the word 'Score'

Score message in job log



Virtual datasets

Operators with node assignments

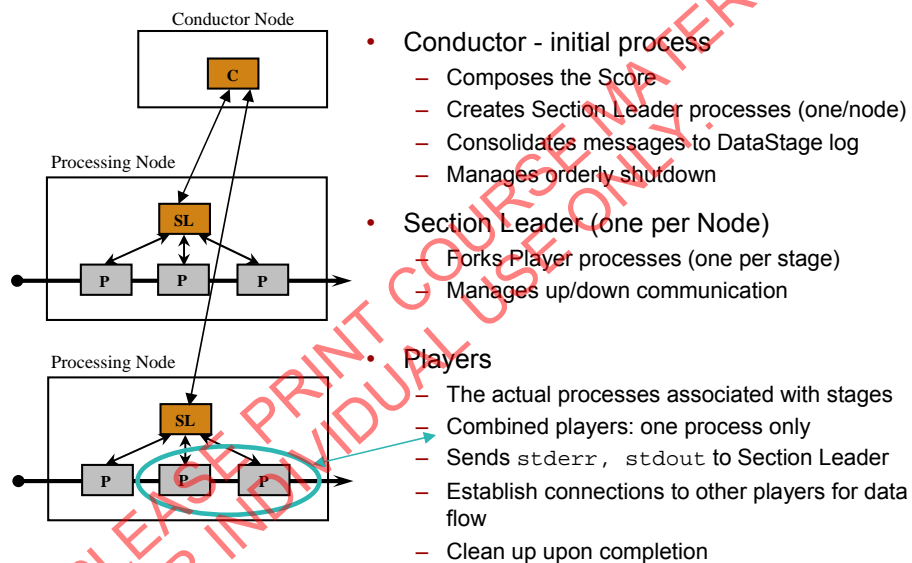
© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

189

Job Execution: The Orchestra



© Copyright IBM Corporation 2007

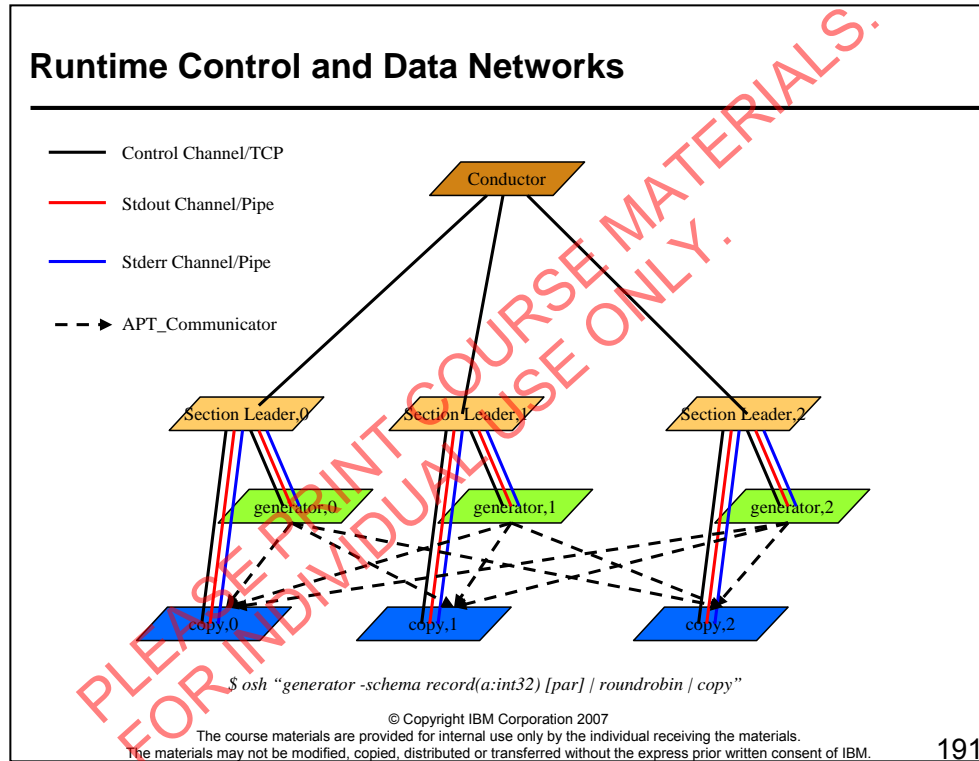
The course materials are provided for internal use only by the individual receiving the materials.

The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

190

The conductor node has the start-up process. Creates the score. Starts up section leaders.

Section leaders communicate with the conductor only. The Conductor communicates with the players.



Every player has to be able to communicate with every other player. There are separate communication channels (pathways) for control, messages, errors, and data. Note that the data channel does not go through the section leader/conductor, as this would limit scalability. Data flows directly from upstream operator to downstream operator using APT_Communicator class.

Checkpoint

1. What file defines the degree of parallelism a job runs under?
2. Name two partitioning algorithms that partition based on key values?
3. What partitioning algorithm produces the most even distribution of data in the various partitions?

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

192

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. Configuration file
2. Hash, Modulus, DB2
3. Round robin. Or Entire

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

193

Unit summary

Having completed this unit, you should be able to:

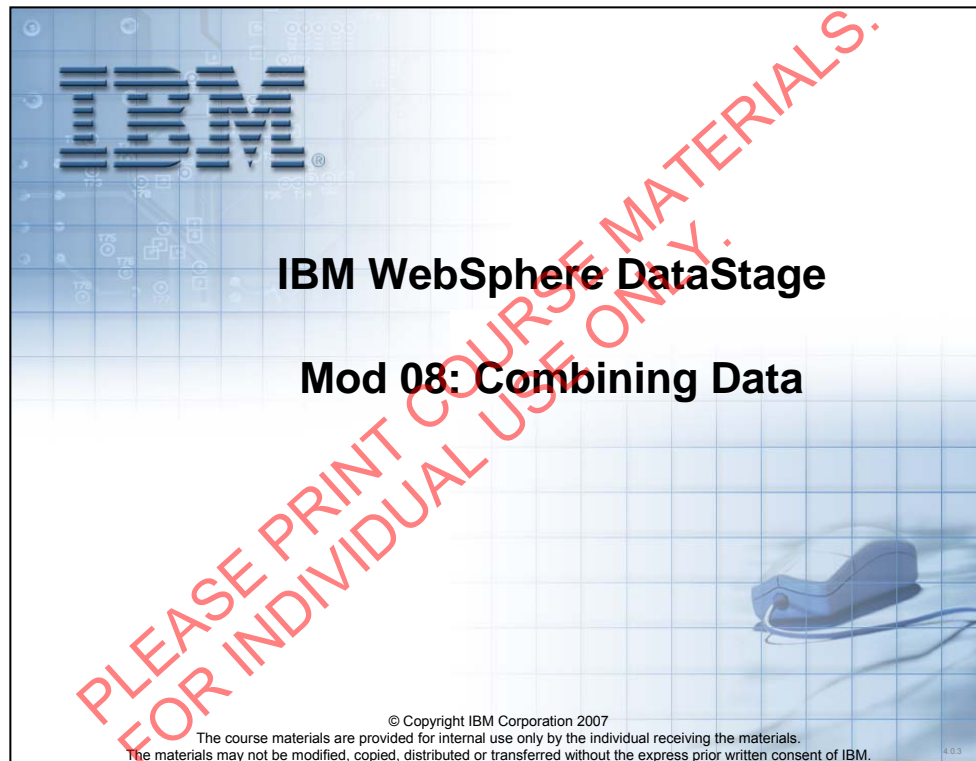
- Describe parallel processing architecture
- Describe pipeline parallelism
- Describe partition parallelism
- List and describe partitioning and collecting algorithms
- Describe configuration files
- Describe the parallel job compilation process
- Explain OSH
- Explain the Score

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

194

Notes:



Unit objectives

After completing this unit, you should be able to:

- Combine data using the Lookup stage
- Define range lookups
- Combine data using Merge stage
- Combine data using the Join stage
- Combine data using the Funnel stage

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

196

Notes:

Combining Data

Ways to combine data:

- Horizontally:
 - Multiple input links
 - One output link made of columns from different input links.
 - Joins
 - Lookup
 - Merge
- Vertically:
 - One input link, one output link combining groups of related records into a single record
 - Aggregator
 - Remove Duplicates
- Funneling: Multiple input streams funneled into a single output stream
 - Funnel stage

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

197

This module discusses the horizontal way of merging data. The next module discusses the vertical way.

Lookup, Merge, Join Stages

- These stages combine two or more input links
 - Data is combined by designated "key" column(s)
- These stages differ mainly in:
 - Memory usage
 - Treatment of rows with unmatched key values
 - Input requirements (sorted, de-duplicated)

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

198

Lookup Stage

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

199

Lookup Features

- One stream input link (source link)
- Multiple reference links
- One output link
- Lookup failure options
 - Continue, Drop, Fail, Reject
- Reject link
 - Only available with Reject lookup failure option
- Can return multiple matching rows
- Hash file is built in memory from the lookup files
 - Indexed by key
 - Should be small enough to fit into physical memory



PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

200

Lookup Types

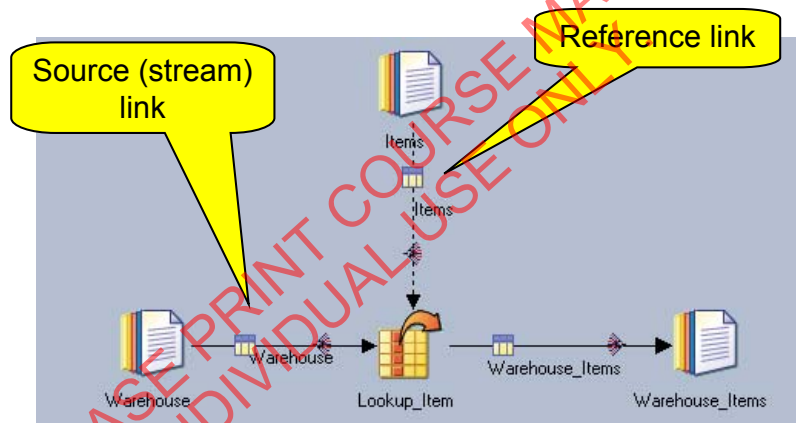
- Equality match
 - Match exactly values in the lookup key column of the reference link to selected values in the source row
 - Return row or rows (if multiple matches are to be returned) that match
- Caseless match
 - Like an equality match except that it's caseless
 - E.g., "abc" matches "AbC"
- Range on the reference link
 - Two columns on the reference link define the range
 - A match occurs when a selected value in the source row is within the range
- Range on the source link
 - Two columns on the source link define the range
 - A match occurs when a selected value in the reference link is within the range

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

201

Lookup Example



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

202

Lookup Stage With an Equality Match

Source link columns

Lookup constraints

Mappings to output columns

Reference key column and source mapping

Column definitions

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

203

For an equality or caseless match lookup, one or more columns in the reference link are selected as keys (see lower left panel). Columns from the source link are matched to the key columns using drag and drop. To specify an equality match, select the equal sign (=) from the Key Type box of the reference link panel. To specify a caseless match, select the Caseless from the Key Type box of the reference link panel.

Output columns are specified in the top, right panel. Columns from the source and reference link are dragged to the front of these columns to specify the values to be mapped to the output columns.

Lookup Failure Actions

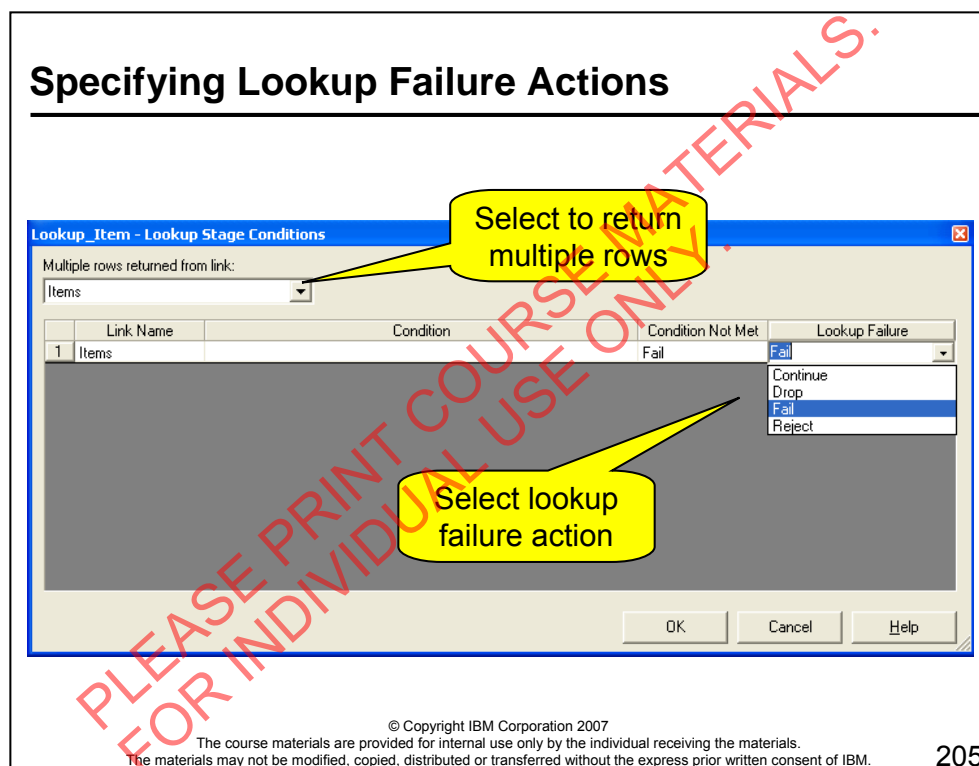
If the lookup fails to find a matching key column, one of several actions can be taken:

- Fail (Default)
 - Stage reports an error and the job fails immediately
- Drop
 - Input row is dropped
- Continue
 - Input row is transferred to the output. Reference link columns are filled with null or default values
- Reject
 - Input row sent to a reject link
 - This requires that a reject link has been created for the stage

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

204



Click the Constraints icon (top, second from left) to open this window.

Lookup Stage Behavior

Source link

| <u>Revolution</u> | <u>Citizen</u> |
|-------------------|----------------|
| 1789 | Lefty |
| 1776 | M_B_Dextrous |

Reference link

| <u>Citizen</u> | <u>Exchange</u> |
|----------------|-----------------|
| M_B_Dextrous | Nasdaq |
| Righty | NYSE |

Lookup key
column

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

206

Lookup Stage

Output of Lookup with *Continue* option

| Revolution | Citizen | Exchange |
|------------|--------------|----------|
| 1789 | Lefty | |
| 1776 | M_B_Dextrous | Nasdaq |

Empty string
or NULL

Output of Lookup with *Drop* option

| Revolution | Citizen | Exchange |
|------------|--------------|----------|
| 1776 | M_B_Dextrous | Nasdaq |

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

207

For the first source row (1789), the lookup fails to find a match. Since Continue is the lookup failure option, the row is output. The Exchange column is populated with NULL (if the column is nullable) or a default value, empty string (if the column is not nullable).

For the second source row (1776), the lookup finds a match, so the Exchange column gets a value from the lookup file.

When Drop is the lookup failure action, the first, unmatched row is dropped.

Designing a Range Lookup Job

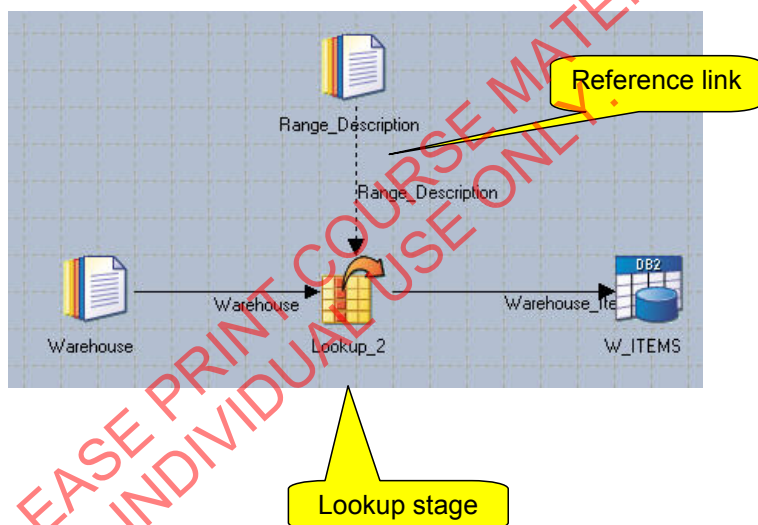
PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

208

Range Lookup Job



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

209

Range on Reference Link

Reference range values

| StartItem | EndItem | Description |
|--------------|--------------|---------------|
| 0100-0109-01 | 0100-0166-01 | Description A |
| 0100-0166-01 | 0100-0319-01 | Description B |
| 0100-0319-01 | 0100-0447-01 | Description C |

Retrieve description

Source values

| Warehouse | Item | Onhand | Onorder | Allocated | HardAllocated |
|-----------|--------------|-------------|-------------|-------------|---------------|
| 100 | 0100-0109-01 | 0474.000000 | 0030.000000 | 0131.000000 | 35.000000 |
| 100 | 0100-0166-01 | 0094.000000 | 0059.000000 | 0047.000000 | 40.000000 |
| 100 | 0100-0319-01 | 0003.000000 | 0000.000000 | 0000.000000 | 0.000000 |

Diagram illustrating a Range on Reference Link configuration in DataStage. The diagram shows a Warehouse icon connected to a LookUp_2 icon, which is then connected to a Warehouse icon labeled W_ITEMS. A Range_Description icon is shown above the LookUp_2 icon. A yellow callout points to the Range_Description icon, labeled 'Reference range values'. Another yellow callout points to the LookUp_2 icon, labeled 'Retrieve description'. A third yellow callout points to the Warehouse icon, labeled 'Source values'. The diagram also includes two data tables: one for 'LookupItemDescription..Range_Description.Range_D' and another for 'LookupItemDescription..Warehouse.Warehouse - Data Browser'.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials. The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Selecting the Stream Column

Double-click to specify range

Source link

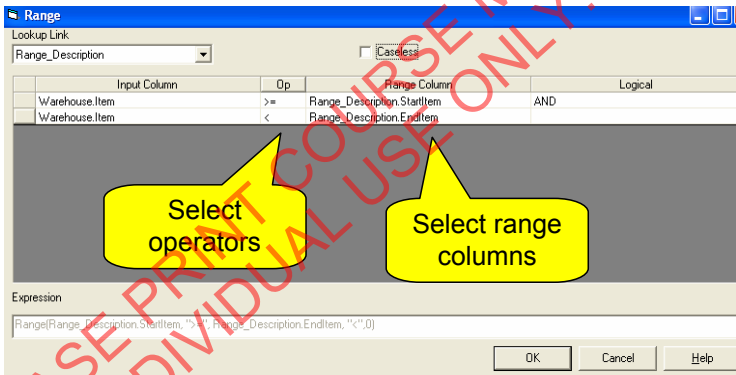
Reference link

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

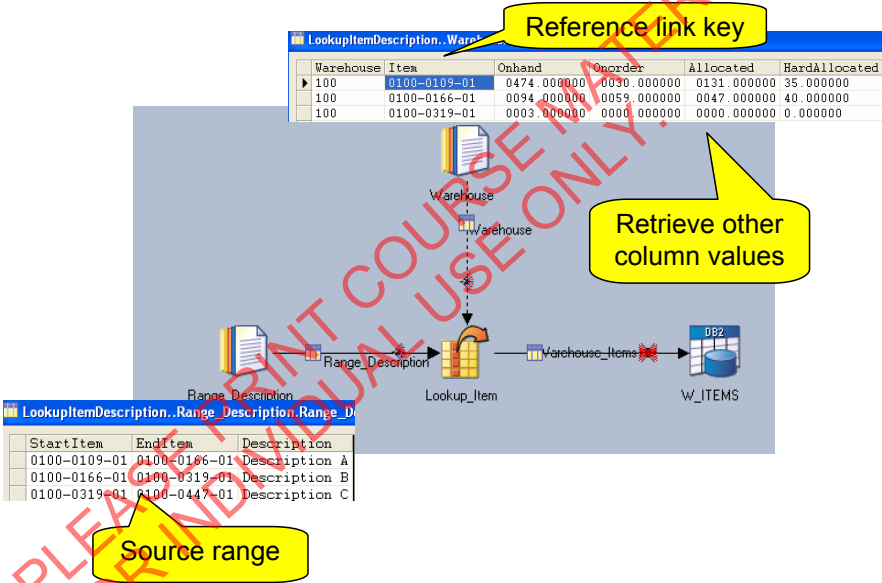
211

Select the column in the source link that contains the value to match to the range on the reference link.

Range Expression Editor



Range on Stream Link



© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Specifying the Range Lookup

Range_Description

| Column name | Key | SQL type | Length | Scale | Nullable |
|-------------|-------------------------------------|----------|--------|-------|----------|
| 1 Warehouse | <input type="checkbox"/> | Integer | 10 | | No |
| 2 Item | <input checked="" type="checkbox"/> | VarChar | 50 | | No |
| 3 Onhand | <input type="checkbox"/> | VarChar | 15 | | No |
| 4 Onorder | <input type="checkbox"/> | VarChar | 15 | | No |

Warehouse_Items

| Column name | Key | SQL type | Length | Scale | Nullable |
|-------------|-------------------------------------|----------|--------|-------|----------|
| 1 Warehouse | <input type="checkbox"/> | Integer | 10 | | No |
| 2 Item | <input checked="" type="checkbox"/> | VarChar | 50 | | No |
| 3 Onhand | <input type="checkbox"/> | VarChar | 15 | | No |
| 4 Onorder | <input type="checkbox"/> | VarChar | 15 | | No |

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Range Expression Editor

Range

Lookup Link

Caseless

| Input Column | Op | Range Column | Logical |
|----------------|----|-----------------------------|---------|
| Warehouse.Item | >= | Range_Description.StartItem | AND |
| Warehouse.Item | <= | Range_Description.EndItem | |

Select range columns

Expression

Range(Range_Description.StartItem,">=",Range_Description.EndItem,"<=","")

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

Join Stage

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

216

Join Stage

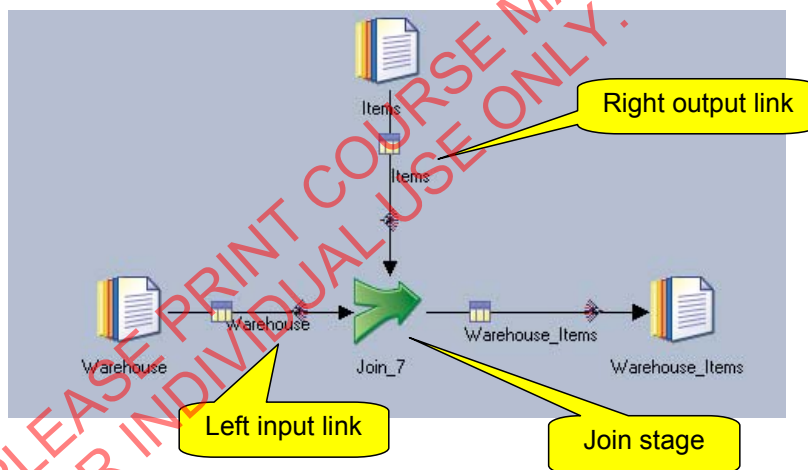
- Four types of joins
 - Inner
 - Left outer
 - Right outer
 - Full outer
- Input links must be sorted
 - Left link and a right link
 - Supports additional "intermediate" links
- Light-weight
 - Little memory required, because of the sort requirement
- Join key column or columns
 - Names for each input link must match

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

217

Job With Join Stage

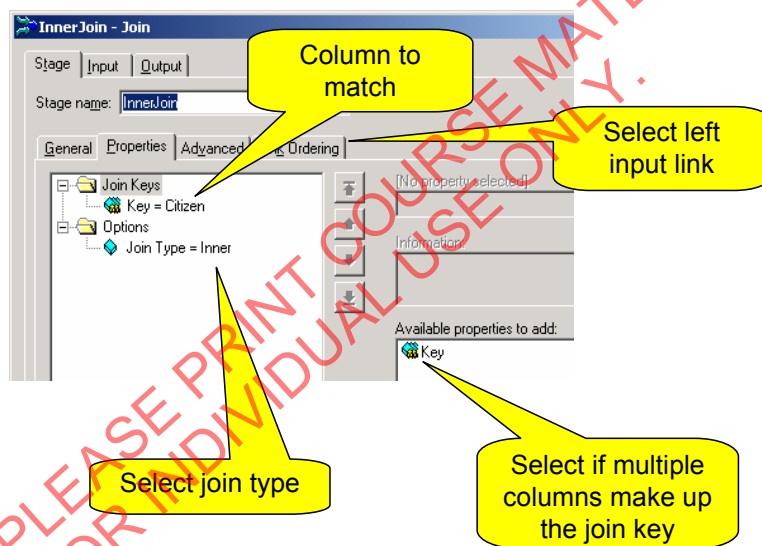


© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

218

Join Stage Editor



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

219

Join Stage Behavior

Left link (primary input)

| <u>Revolution</u> | <u>Citizen</u> |
|-------------------|----------------|
| 1789 | Lefty |
| 1776 | M_B_Dextrous |

Right link (secondary input)

| <u>Citizen</u> | <u>Exchange</u> |
|----------------|-----------------|
| M_B_Dextrous | Nasdaq |
| Righty | NYSE |

Join key
column

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

220

Inner Join

- Transfers rows from both data sets whose key columns have matching values

Output of inner join on key Citizen

| Revolution | Citizen | Exchange |
|------------|----------------|----------|
| 1776 | M. B. Dextrous | Nasdaq |

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

221

Left Outer Join

- Transfers *all* values from the left link and transfers values from the right link only where key columns match.

| <u>Revolution</u> | <u>Citizen</u> | <u>Exchange</u> |
|-------------------|----------------|-----------------|
| 1789 | Lefty | |
| 1776 | M_B_Dextrous | Nasdaq |

Null or default
value

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

222

Right Outer Join

- Transfers *all* values from the right link and transfers values from the left link only where key columns match.

| <u>Revolution</u> | <u>Citizen</u> | <u>Exchange</u> |
|-------------------|----------------|-----------------|
| 1776 | M_B_Dextrous | Nasdaq |
| | Righty | NYSE |

Null or default
value

© Copyright IBM Corporation 2007
The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

223

Merge Stage

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

224

Merge Stage

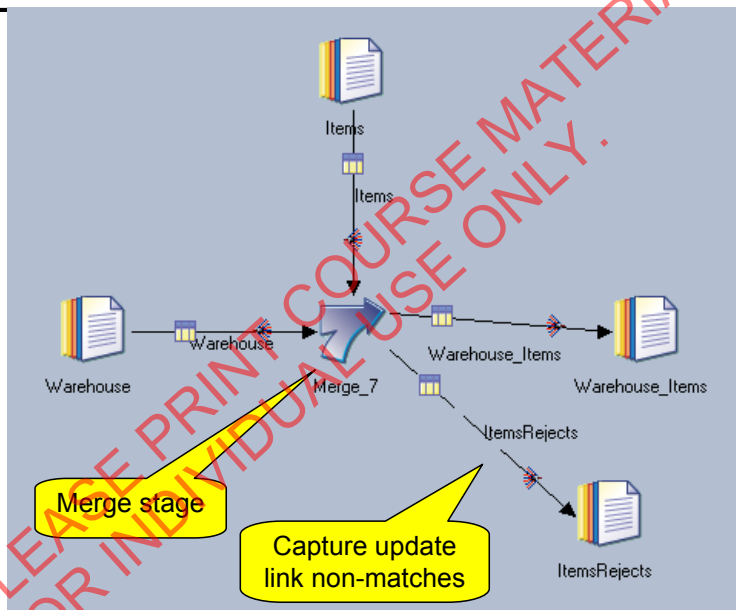
- Similar to Join stage
- Input links must be sorted
 - Master link and one or more secondary links
 - Master must be duplicate-free
- Light-weight
 - Little memory required, because of the sort requirement
- Unmatched master rows can be kept or dropped
- Unmatched secondary links can be captured in a reject link

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

225

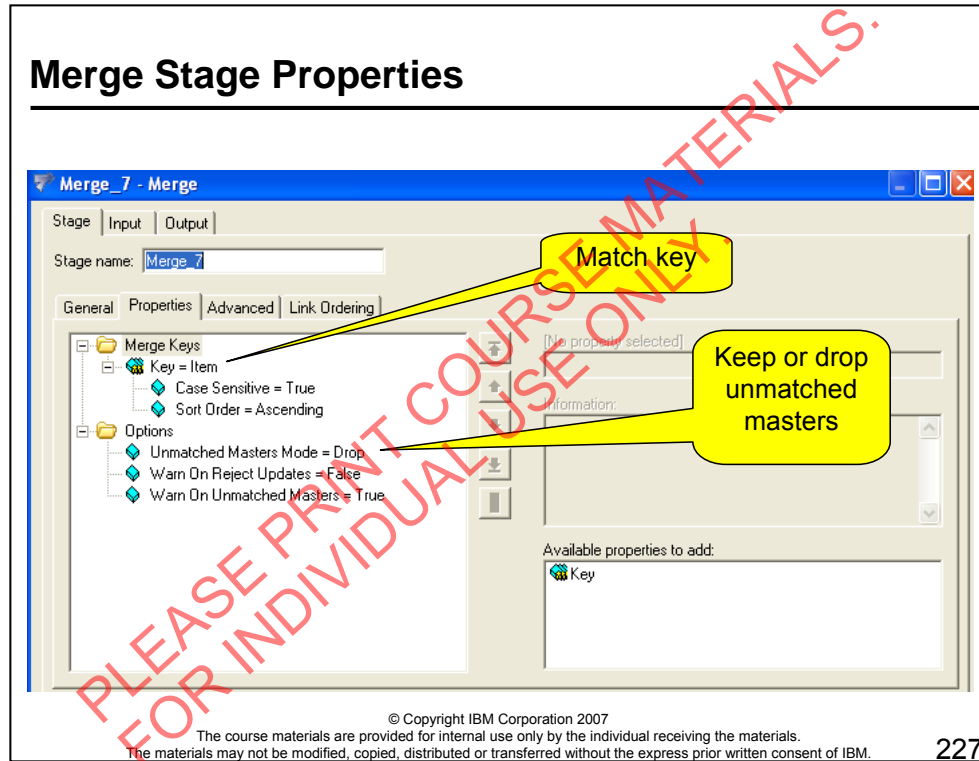
Merge Stage Job



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

226



Unmatched updates are captured by adding additional reject links (one for each update link).

Comparison: Joins, Lookup, Merge

| | Joins | Lookup | Merge |
|----------------------------------|----------------------------------|-----------------------------------|-----------------------------|
| Model | RDBMS-style relational | Source - in RAM LU Table | Master -Update(s) |
| Memory usage | light | heavy | light |
| # and names of Inputs | 2 or more: left, right | 1 Source, N LU Tables | 1 Master, N Update(s) |
| Mandatory Input Sort | all inputs | no | all inputs |
| Duplicates in primary input | OK | OK | Warning! |
| Duplicates in secondary input(s) | OK | Warning! | OK only when N = 1 |
| Options on unmatched primary | Keep (left outer), Drop (Inner) | [fail] continue drop reject | [keep] drop |
| Options on unmatched secondary | Keep (right outer), Drop (Inner) | NONE | capture in reject set(s) |
| On match, secondary entries are | captured | captured | consumed |
| # Outputs | 1 | 1 out, (1 reject) | 1 out, (N rejects) |
| Captured in reject set(s) | Nothing (N/A) | unmatched primary entries | unmatched secondary entries |

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

Funnel Stage

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

229

What is a Funnel Stage?

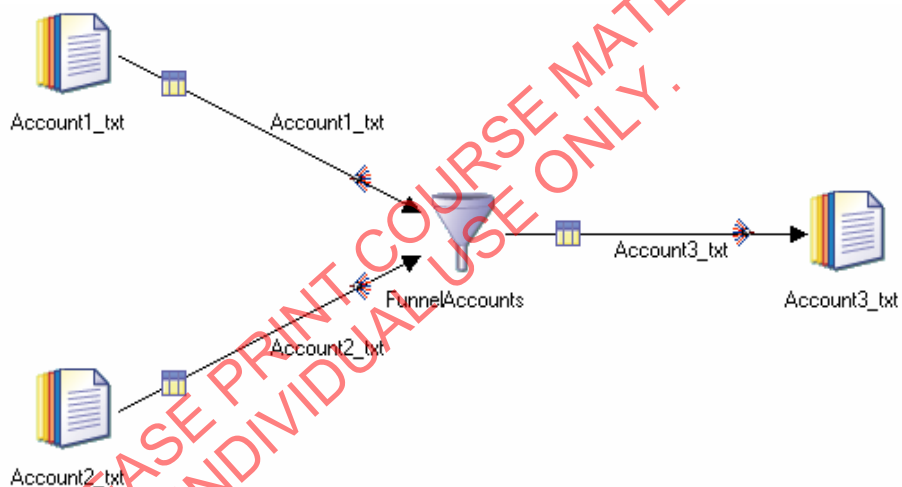
- Combines data from multiple input links to a single output link
- All sources must have identical metadata
- Three modes
 - Continuous
 - Records are combined in no particular order
 - First available record
 - Sort Funnel
 - Combines the input records in the order defined by a key
 - Produces sorted output if the input links are all sorted by the same key
 - Sequence:
 - Outputs all records from the first input link, then all from the second input link, and so on

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

230

Funnel Stage Example



© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

231

Checkpoint

1. Name three stages that horizontally join data?
2. Which stage uses the least amount of memory? Join or Lookup?
3. Which stage requires that the input data is sorted? Join or Lookup?

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

232

Notes:

Write down your answers here:

1.

2.

Checkpoint solutions

1. Lookup, Merge, Join
2. Lookup
3. Join

PLEASE PRINT COURSE MATERIALS.
FOR INDIVIDUAL USE ONLY.

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

233

Unit summary

Having completed this unit, you should be able to:

- Combine data using the Lookup stage
- Define range lookups
- Combine data using Merge stage
- Combine data using the Join stage
- Combine data using the Funnel stage

© Copyright IBM Corporation 2007

The course materials are provided for internal use only by the individual receiving the materials.
The materials may not be modified, copied, distributed or transferred without the express prior written consent of IBM.

234

Notes: