KM423G: IBM InfoSphere DataStage v11.5 - Advanced Data Processing

Course Code: KM423G
Duration: 2 days
Instructor-led Training (ILT) | Virtual Instructor-led Training (VILT)

OVERVIEW

This course is designed to introduce you to advanced parallel job data processing techniques in DataStage v11.5. In this course you will develop data techniques for processing different types of complex data resources including relational data, unstructured data (Excel spreadsheets), and XML data. In addition, you will learn advanced techniques for processing data, including techniques for masking data and techniques for validating data using data rules. Finally, you will learn techniques for updating data in a star schema data warehouse using the DataStage SCD (Slowly Changing Dimensions) stage. Even if you are not working with all of these specific types of data, you will benefit from this course by learning advanced DataStage job design techniques, techniques that go beyond those utilized in the DataStage Essentials course..

SKILLS COVERED

- Use Connector stages to read from and write to database tables
- Handle SQL errors in Connector stages
- Use Connector stages with multiple input links
- Use the File Connector stage to access Hadoop HDFS data
- Optimize jobs that write to database tables
- Use the Unstructured Data stage to extract data from Excel spreadsheets
- Use the Data Masking stage to mask sensitive data processed within a DataStage job
- Use the Hierarchical stage to parse, compose, and transform XML data
- Use the Schema Library Manager to import and manage XML schemas
- Use the Data Rules stage to validate fields of data within a DataStage job
- Create custom data rules for validating data
- Design a job that processes a star schema data warehouse with Type 1 and Type 2 slowly changing dimensions

WHO SHOULD ATTEND?

Experienced DataStage developers seeking training in more advanced DataStage job techniques and who seek techniques for working with complex types of data resources.

PREREQUISITES

DataStage Essentials course or equivalent.

MODULES

Module 1: Accessing Databases

Topic 1: Connector stage overview
• Use Connector stages to read from and write to relational tables
• Working with the Connector stage properties
Topic 2: Connector stage functionality
• Before / After SQL
• Sparse lookups
• Optimize insert/update performance
Topic 3: Error handling in Connector stages
• Reject links
• Reject conditions
Topic 4: Multiple input links
• Designing jobs using Connector stages with multiple input links
• Ordering records across multiple input links

Topic 5: File Connector stage
• Read and write data to Hadoop file systems
Demonstration 1: Handling database errors
Demonstration 2: Parallel jobs with multiple Connector input links
Demonstration 3: Using the File Connector stage to read and write HDFS files

Module 2: Processing Unstructured Data

Topic 1: Using the Unstructured Data stage in DataStage jobs
• Extract data from an Excel spreadsheet
• Specify a data range for data extraction in an Unstructured Data stage
• Specify document properties for data extraction.
Demonstration 1: Processing unstructured data

Module 3: Data Masking

Topic 1: Using the Data Masking stage in DataStage jobs
• Data masking techniques
• Data masking policies
• Applying policies for masquerading context-aware data types
• Applying policies for masquerading generic data types
• Repeatable replacement
• Using reference tables
• Creating custom reference tables
Demonstration 1: Data masking

Module 4: Using Data Rules

Topic 1: Introduction to data rules
• Using the Data Rules Editor
• Selecting data rules
• Binding data rule variables
• Output link constraints
• Adding statistics and attributes to the output information
Topic 2: Use the Data Rules stage to valid foreign key references in source data
Topic 3: Create custom data rules
Demonstration 1: Using data rules

Module 5: Processing XML Data

Topic 1: Introduction to the Hierarchical stage
• Hierarchical stage Assembly editor
• Use the Schema Library Manager to import and manage XML schemas
Topic 2: Composing XML data
• Using the HJoin step to create parent-child relationships between input lists
• Using the Composer step
Topic 3: Writing Hierarchical data to a relational table
Topic 4: Using the Regroup step
Topic 5: Consuming XML data
• Using the XML Parser step
• Propagating columns
Topic 6: Transforming XML data
• Using the Aggregate step
• Using the Sort step
• Using the Switch step
• Using the H-Pivot step
Demonstration 1: Importing XML schemas
Demonstration 2: Compose hierarchical data
Demonstration 3: Consume hierarchical data
Demonstration 4: Transform hierarchical data

Module 6: Updating a Star Schema Database

Topic 1: Surrogate keys
• Design a job that creates and updates a surrogate key source key file from a dimension table
Topic 2: Slowly Changing Dimensions (SCD) stage
• Star schema databases
• SCD stage Fast Path pages
• Specifying purpose codes
• Dimension update specification
• Design a job that processes a star schema database with Type 1 and Type 2 slowly changing dimensions
Demonstration 1: Build a parallel job that

updates a star schema database with two dimensions

END OF PAGE