Training: IBM
IBM InfoSphere Advanced DataStage - Parallel Framework

TRAINING GOALS:

This course is designed to introduce advanced parallel job development techniques in DataStage v11.5. In this course you will develop a deeper understanding of the DataStage architecture, including a deeper understanding of the DataStage development and runtime environments. This will enable you to design parallel jobs that are robust, less subject to errors, reusable, and optimized for better performance.

Please refer to course overview

Experienced DataStage developers seeking training in more advanced DataStage job techniques and who seek an understanding of the parallel framework architecture.

CONSPECT:

1: Introduction to the parallel framework architecture
• Describe the parallel processing architecture
• Describe pipeline and partition parallelism
• Describe the role of the configuration file
• Design a job that creates robust test data

2: Compiling and executing jobs
• Describe the main parts of the configuration file
• Describe the compile process and the OSH that the compilation process generates
• Describe the role and the main parts of the Score
• Describe the job execution process

3: Partitioning and collecting data
• Understand how partitioning works in the Framework
• Viewing partitioners in the Score
• Selecting partitioning algorithms
• Generate sequences of numbers (surrogate keys) in a partitioned, parallel environment

4: Sorting data
• Sort data in the parallel framework
• Find inserted sorts in the Score
• Reduce the number of inserted sorts
• Optimize Fork-Join jobs
• Use Sort stages to determine the last row in a group

Compendium Education Center Ltd.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Bielska 17, 02-394 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

• Describe sort key and partitioner key logic in the parallel framework

5: Buffering in parallel jobs
• Describe how buffering works in parallel jobs
• Tune buffers in parallel jobs
• Avoid buffer contentions

6: Parallel framework data types
• Describe virtual data sets
• Describe schemas
• Describe data type mappings and conversions
• Describe how external data is processed
• Handle nulls
• Work with complex data

7: Reusable components
• Create a schema file
• Read a sequential file using a schema
• Describe Runtime Column Propagation (RCP)
• Enable and disable RCP
• Create and use shared containers

8: Balanced Optimization
• Enable Balanced Optimization functionality in Designer
• Describe the Balanced Optimization workflow
• List the different Balanced Optimization options.
• Push stage processing to a data source
• Push stage processing to a data target
• Optimize a job accessing Hadoop HDFS file system
• Understand the limitations of Balanced Optimizations

REQUIREMENTS:

IBM InfoSphere DataStage Essentials course or equivalent and at least one year of experience developing parallel jobs using DataStage.

Difficulty level

Compendium Education Center Ltd.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Bielska 17, 02-394 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007