My Datastage Notes

Pages

- Home
- Different Versions
- **All Datastage Stages**
- Configuration File
- Sequentional_Stage
- Dataset
- Transformer Stage
- Sort Stage
- Aggregator_Stage
- Join Stage
- Lookup_Stage
- Merge_Stage
- Filter_Stage
- Copy Stage
- Funnel_Stage
- Column Generator
- Surrogate_Key_Stage
- SCD
- Pivot_Enterprise_Stage
- Sequence_Activities
- Datastage Study Material/Interview Questions
- Datastage Errors and Resolution
- Datastage Scenarios and solutions
- Unix Shell Scripting
- SQL/Database
- Datawarehousing Concepts

Blog Archive

▼ 2014 (46)
▼ September (40)

All Datastage Stages

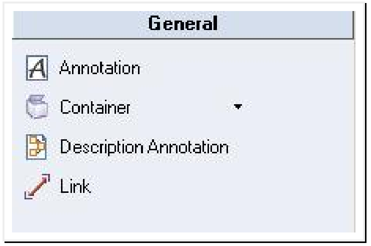
Datastage parallel stages groups

DataStage and QualityStage stages are grouped into the following logical sections:

- General objects
- Data Quality Stages
- Database connectors
- Development and Debug stages
- File stages
- Processing stages
- Real Time stages
- Restructure Stages
- Sequence activities

Please refer to the list below for a description of the stages used in DataStage and QualityStage.
We classified all stages in order of importancy and frequency of use in real-life deployments (and also on certification exams). Also, the most widely used stages are marked bold or there is a link to a subpage available with a detailed description with examples.

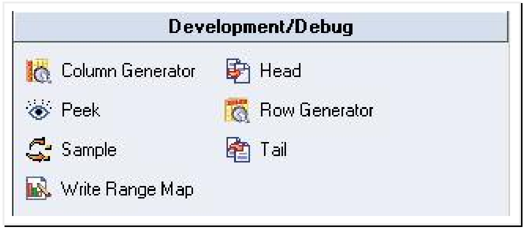
DataStage and QualityStage parallel stages and activities



General elements

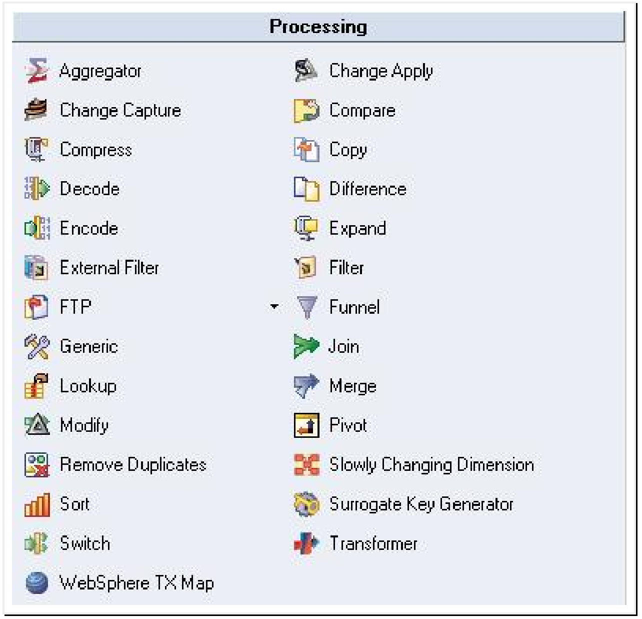
- **Link** indicates a flow of the data. There are three main types of links in Datastage: stream, reference and lookup.

- **Container** (can be private or shared) - the main outcome of having containers is to simplify visually a complex datastage job design and keep the design easy to understand.
- **Annotation** is used for adding floating datastage job notes and descriptions on a job canvas. Annotations provide a great way to document the ETL process and help understand what a given job does.
- Description Annotation shows the contents of a job description field. One description annotation is allowed in a datastage job.



Debug and development stages

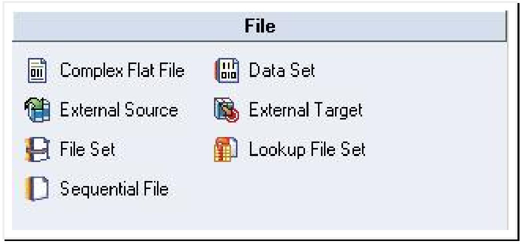
- **Row generator** produces a set of test data which fits the specified metadata (can be random or cycled through a specified list of values). Useful for testing and development. Click here for more..
- **Column generator** adds one or more column to the incoming flow and generates test data for this column.
- **Peek** stage prints record column values to the job log which can be viewed in Director. It can have a single input link and multiple output links.Click here for more..
- Sample stage samples an input data set. Operates in two modes: percent mode and period mode.
- Head selects the first N rows from each partition of an input data set and copies them to an output data set.
- Tail is similiar to the Head stage. It select the last N rows from each partition.
- Write Range Map writes a data set in a form usable by the range partitioning method.



Processing stages

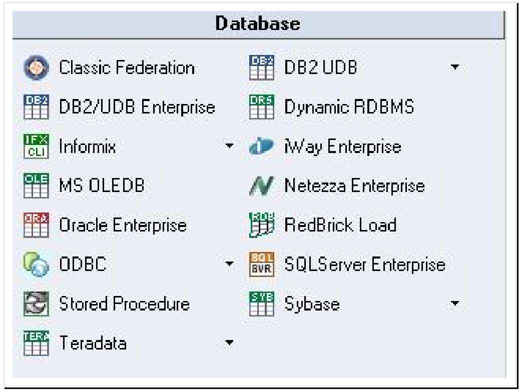
- **Aggregator** joins data vertically by grouping incoming data stream and calculating summaries (sum, count, min, max, variance, etc.) for each group. The data can be grouped using two methods: hash table or pre-sort. Click here for more..
- **Copy** - copies input data (a single stream) to one or more output data flows
- **FTP** stage uses FTP protocol to transfer data to a remote machine
- **Filter** filters out records that do not meet specified requirements.Click here for more..
- **Funnel** combines mulitple streams into one. Click here for more..
- **Join** combines two or more inputs according to values of a key column(s). Similiar concept to relational DBMS SQL join (ability to perform inner, left, right and full outer joins). Can have 1 left and multiple right inputs (all need to be sorted) and produces single output stream (no reject link). Click here for more..
- **Lookup** combines two or more inputs according to values of a key column(s). Lookup stage can have 1 source and multiple lookup tables. Records don't need to be sorted and produces single output stream and a reject link. Click here for more..
- **Merge** combines one master input with multiple update inputs according to values of a key column(s). All inputs need to be sorted and unmatched secondary entries can be captured in multiple reject links. Click here for more..
- **Modify** stage alters the record schema of its input dataset. Useful for renaming columns, non-default data type conversions and null handling
- **Remove duplicates** stage needs a single sorted data set as input. It removes all duplicate records according to a specification and writes to a single output
- **Slowly Changing Dimension** automates the process of updating dimension tables, where the data changes in time. It supports SCD type 1 and SCD type 2.Click here for more..
- **Sort** sorts input columns.Click here for more..
- **Transformer** stage handles extracted data, performs data validation, conversions and lookups.Click here for more..

- Change Capture - captures before and after state of two input data sets and outputs a single data set whose records represent the changes made.
- Change Apply - applies the change operations to a before data set to compute an after data set. It gets data from a Change Capture stage
- Difference stage performs a record-by-record comparison of two input data sets and outputs a single data set whose records represent the difference between them. Similiar to Change Capture stage.
- Checksum - generates checksum from the specified columns in a row and adds it to the stream. Used to determine if there are differencies between records.
- Compare performs a column-by-column comparison of records in two presorted input data sets. It can have two input links and one output link.
- Encode encodes data with an encoding command, such as gzip.
- Decode decodes a data set previously encoded with the Encode Stage.
- External Filter permits speicifying an operating system command that acts as a filter on the processed data
- Generic stage allows users to call an OSH operator from within DataStage stage with options as required.
- Pivot Enterprise is used for horizontal pivoting. It maps multiple columns in an input row to a single column in multiple output rows. Pivoting data results in obtaining a dataset with fewer number of columns but more rows.
- Surrogate Key Generator generates surrogate key for a column and manages the key source.
- Switch stage assigns each input row to an output link based on the value of a selector field. Provides a similiar concept to the switch statement in most programming languages.
- Compress - packs a data set using a GZIP utility (or compress command on LINUX/UNIX)
- Expand extracts a previously compressed data set back into raw binary data.



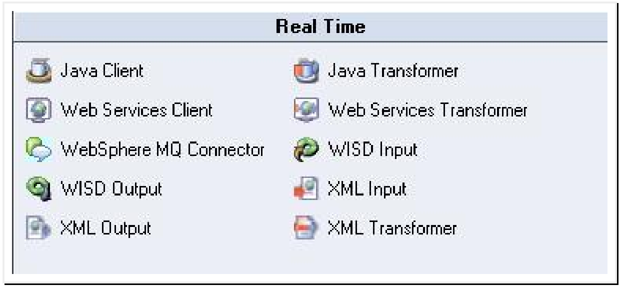
File stage types

- **Sequential file** is used to read data from or write data to one or more flat (sequential) files.Click here for more..
- **Data Set** stage allows users to read data from or write data to a dataset. Datasets are operating system files, each of which has a control file (.ds extension by default) and one or more data files (unreadable by other applications). Click here for more info
- **File Set** stage allows users to read data from or write data to a fileset. Filesets are operating system files, each of which has a control file (.fs extension) and data files. Unlike datasets, filesets preserve formatting and are readable by other applications.
- **Complex flat file** allows reading from complex file structures on a mainframe machine, such as MVS data sets, header and trailer structured files, files that contain multiple record types, QSAM and VSAM files.Click here for more info.
- External Source - permits reading data that is output from multiple source programs.
- External Target - permits writing data to one or more programs.
- Lookup File Set is similiar to FileSet stage. It is a partitioned hashed file which can be used for lookups.



Database stages

- **Oracle Enterprise** allows reading data from and writing data to an Oracle database (database version from 9.x to 10g are supported).
- **ODBC Enterprise** permits reading data from and writing data to a database defined as an ODBC source. In most cases it is used for processing data from or to Microsoft Access databases and Microsoft Excel spreadsheets.
- **DB2/UDB Enterprise** permits reading data from and writing data to a DB2 database.
- **Teradata** permits reading data from and writing data to a Teradata data warehouse. Three Teradata stages are available: Teradata connector, Teradata Enterprise and Teradata Multiload
- **SQLServer Enterprise** permits reading data from and writing data to Microsoft SQLl Server 2005 amd 2008 database.
- **Sybase** permits reading data from and writing data to Sybase databases.
- Stored procedure stage supports Oracle, DB2, Sybase, Teradata and Microsoft SQL Server. The Stored Procedure stage can be used as a source (returns a rowset), as a target (pass a row to a stored procedure to write) or a transform (to invoke procedure processing within the database).
- MS OLEDB helps retrieve information from any type of information repository, such as a relational source, an ISAM file, a personal database, or a spreadsheet.
- Dynamic Relational Stage (Dynamic DBMS, DRS stage) is used for reading from or writing to a number of different supported relational DB engines using native interfaces, such as Oracle, Microsoft SQL Server, DB2, Informix and Sybase.
- Informix (CLI or Load)
- DB2 UDB (API or Load)
- Classic federation
- RedBrick Load
- Netezza Enterprise
- iWay Enterprise

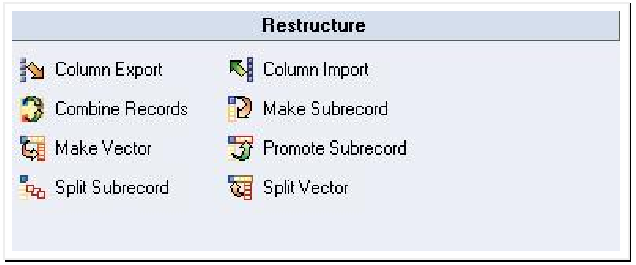


Real Time

Java Client
Java Transformer
Web Services Client
Web Services Transformer
WebSphere MQ Connector
WISD Input
WISD Output
XML Input
XML Output
XML Transformer

Real Time stages

- **XML Input** stage makes it possible to transform hierarchical XML data to flat relational data sets
- **XML Output** writes tabular data (relational tables, sequential files or any datastage data streams) to XML structures
- **XML Transformer** converts XML documents using an XSLT stylesheet
- **Websphere MQ** stages provide a collection of connectivity options to access IBM WebSphere MQ enterprise messaging systems. There are two MQ stage types available in DataStage and QualityStage: WebSphere MQ connector and WebSphere MQ plug-in stage.
- Web services client
- Web services transformer
- Java client stage can be used as a source stage, as a target and as a lookup. The java package consists of three public classes: com.ascentialsoftware.jds.Column, com.ascentialsoftware.jds.Row, com.ascentialsoftware.jds.Stage
- Java transformer stage supports three links: input, output and reject.
- WISD Input - Information Services Input stage
- WISD Output - Information Services Output stage



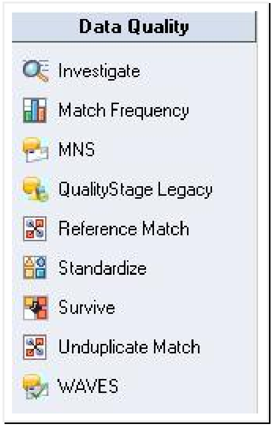
Restructure

Column Export
Column Import
Combine Records
Make Subrecord
Make Vector
Promote Subrecord
Split Subrecord
Split Vector

Restructure stages

- Column export stage exports data from a number of columns of different data types into a single column of data type ustring, string, or binary. It can have one input link, one output link and a rejects link. Click here for more..
- Column import complementary to the Column Export stage. Typically used to divide data arriving in a single column into multiple columns.
- Combine records stage combines rows which have identical keys, into vectors of subrecords.
- Make subrecord combines specified input vectors into a vector of subrecords whose columns have the same names and data types as the original vectors.
- Make vector joins specified input columns into a vector of columns
- Promote subrecord - promotes input subrecord columns to top-level columns
- Split subrecord - separates an input subrecord field into a set of top-level vector columns

- Split vector promotes the elements of a fixed-length vector to a set of top-level columns



Data Quality

Investigate

Match Frequency

MNS

QualityStage Legacy

Reference Match

Standardize

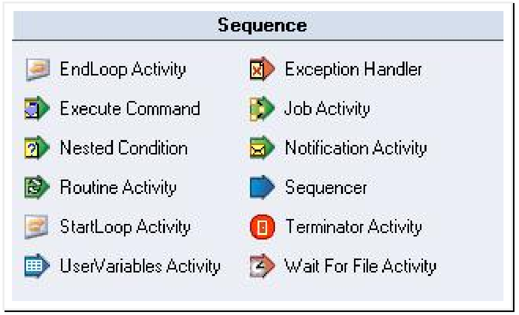
Survive

Unduplicate Match

WAVES

Data quality QualityStage stages

- Investigate stage analyzes data content of specified columns of each record from the source file. Provides character and word investigation methods.
- Match frequency stage takes input from a file, database or processing stages and generates a frequence distribution report.
- MNS - multinational address standarization.
- QualityStage Legacy
- Reference Match
- Standarize
- Survive
- Unduplicate Match
- WAVES - worldwide address verification and enhancement system.



Sequence

EndLoop Activity

Exception Handler

Execute Command

Job Activity

Nested Condition

Notification Activity

Routine Activity

Sequencer

StartLoop Activity

Terminator Activity

UserVariables Activity

Wait For File Activity

Sequence activity stage types

- **Job Activity** specifies a Datastage server or parallel job to execute.
- **Notification Activity** - used for sending emails to user defined recipients from within Datastage
- **Sequencer** used for synchronization of a control flow of multiple activities in a job sequence.

- **Terminator Activity** permits shutting down the whole sequence once a certain situation occurs.
- **Wait for file Activity** - waits for a specific file to appear or disappear and launches the processing.
- EndLoop Activity
- Exception Handler
- Execute Command
- Nested Condition
- Routine Activity
- StartLoop Activity
- UserVariables Activity

No comments:

Post a Comment

Home

Subscribe to: Posts (Atom)