

What is the difference between DataStage and Informatica?

- DataStage support parallel processing which informatica doesn't.
- Links are object in the DataStage ,in Informatica it's a port to port connectivity .
- In Informatica its easy to implement Slowly Changing Dimensions which is little bit complex in DataStage.
- DataStage doesn't support complete error handling.

<http://questionsdoor.com/datastage-interview-questions/>

What are the components of Ascential Data Stage?

- Client Components - Administrator, Director, Manager, and Designer.
- Server Components - Repository, Server and Plug-ins.

What are System variables?

- System variables comprise of a set of variables which are used to get system information and they can be accessed from a transformer or a routine. They are read only and start with an @.

What are the enhancements made in datastage 7.5 compare with 7.0?

- Many new stages were introduced compared to datastage version 7.0. In server jobs we have stored procedure stage, CDC Stage ,command stage and generate report option was there in file tab. In job sequence many stages like startloop activity, end loop activity,terminate loop activity and user variables activities were introduced. In parallel jobs surrogate key stage, stored procedure stage were introduced.

What is DataStage?

- Design jobs for Extraction, Transformation and Loading(ETL).
- Ideal tool data integration projects such as data warehouses, data marts and system migrations.
- Import,export,create and managed metadata for use within jobs.
- Schedule, run and monitor jobs all within DataStage
- Adminsters your Datastage development and execution environments.

What is a merge?

- Merge is a stage that is available in both parallel and server jobs.
- The merge stage is used to join two tables(server/parallel) or two tables/datasets(parallel).
- Merge requires that the master table/dataset and the update table/dataset to be sorted.
- Merge is performed on a key field, and the key field is mandatory in the master and update dataset/table.

What are Sequencers?

- A sequencer allows you to synchronize the control flow of multiple activities in a job sequence. It can have multiple input triggers as well as multiple output triggers.

What is version Control?

- Version Control stores different versions of DS jobs runs different versions of same job reverts to previous version of a job view version histories.

What is Active and Passive Stage?

- Passive Stages are used for data extraction and loading ,while Active Stage are used to implements and process the business rules.

What are the main features of datastage?

- DataStage has the following features to aid the design and processing required to build a data warehouse :
- Uses graphical design tools. With simple point and click techniques you can draw a scheme to represent your processing requirements.
- Extracts data from any number or types of database.
- Handles all the metadata definitions required to define your data warehouse. You can view and modify the table definitions at any point during the design of your application.
- Aggregates data. You can modify SQL SELECT statements used to extract data.
- Transforms data. DataStage has a set of predefined transforms and functions you can use to convert your data. You can easily extend the functionality by defining your own transforms to use.
- Loads the data warehouse.

Define data aggregation?

- An operational data source usually contains records of individual transactions such as product sales. If the user of a data warehouse only needs a summed total, you can reduce records to a more manageable number by aggregating the data.

How the IPC Stage work?

- If we used the IPC Stage between source and target .One process will handling the communication from sequential file stage to IPC stage, and other will handling communication from IPC stage to ODBC stage. As soon as the Sequential File stage has opened its output link, the IPC stage can start passing data to the ODBC stage.

What are stage variables?

- An intermediate processing variable that retains value during read and doesn't pass the value into target column.

What are the types of Containers?

- There are Two types of containers namely
- Local Container
- Shared Container

Where the DataStage stored his repository?

- DataStage stored his repository in IBM Universe Database.

What is Staging Variable?

- These are the temporary variables created in transformer for calculation.

How do you generate Sequence number in Datastage?

- Using the Routine
- KeyMgtGetNextVal
- KeyMgtGetNextValConn
- They can also be done by Oracle Sequence.

What is the difference between Server Job and Parallel Jobs?

- Server Jobs work in sequential way while parallel jobs work in parallel fashion (Parallel Extender work on the principle of pipeline and partition) for Input/Output processing.

What is the difference between Account and Directory options ?

- Account option create the hashfile in the current project itself using condition and for Directory option we have to specify the path of the directory.

What are the types of jobs we have in Datastage?

- Server Jobs
- Parallel Jobs
- Mainframe Jobs

What are Routines?

- Routines are the functions which we develop in BASIC Code for required tasks, which we Datastage is not fully supported (Complex).

What are the Job parameters?

- These Parameters are used to provide Administrative access and change run time values of the job. **EDIT > JOBPARAMETERS** In that Parameters Tab we can define the name, prompt, type, value.

What is the difference between Hashfile and Sequential File?

- Hash file stores the data based on hash algorithm and on a key value. A sequential file is just a file with no key column. Hash file used as a reference for look up. Sequential file cannot.

What is the use of datastage designer?

- A design interface used to create DataStage applications(known as jobs). Each job specifies the data sources, the transforms required, and the destination of the data. Jobs are compiled to create executables that are scheduled by the Director and run by the Server.

Explain Datastage architecture?

- Datastage contains two components :

Client Components :

- Datastage Designer
 - Datastage Director
 - Datastage Manager
 - Datastage Administrator
- ### **Server components :**
- Repository
 - Datastage Server
 - Datastage Package Installer

How can we improve the performance in DataStage?

- In server canvas we can improve performance in two ways
- Firstly we can increase the memory by enabling interprocess row buffering in job properties
- Secondly by inserting an IPC stage we break a process into two processes. We can use this stage to connect two passive stages or two active stages.

What is orabulk Stage?

- This Stage is used to Bulk Load the Oracle Target Database.

What is APT_CONFIG in datastage?

- Datastage understands the architecture of the system through this file(APT_CONFIG_FILE). For example this file consists information of node names, disk storage information etc.
APT_CONFIG is just an environment variable used to identify the *.apt file. Dont confuse that with *.apt file that has the node's information and Configuration of SMP/MMP server.

What are orabulk and bcp stages?

- These are called as plug-in stages orabulk is used when v have bulk data in oracle then v go for orabulk for other than oracle database we go for bcp stages.

Why fact table is in normal form?

- A fact table consists of measurements of business requirements and foreign keys of dimensions tables as per business rules.

What is Merge Stage?

- The Merge stage combines a sorted master data set with one or more sorted update data sets. The columns from the records in the master and update data sets are merged so that the output record contains all the columns from the master record plus any additional columns from each update record.

How do you schedule or monitoring the job?

- Using the DataStage Director we can schedule or monitor the job.

What is Container and then types?

- Containers are the reusable set of stages.
There are two types of container are :
- Local Container
- Shared Container

What is the difference between the local container and shared container?

- **Local Container** is local to the particular job in which we developed the container. **Shared Container** is can be used in any other jobs also.

What all are the different way to run a job?

- Using DataStage Director and DataStage Administration Client.

How and where you used HASH file?

- HASH file is used for Lookup while populating fact table ,basically it keeps the data in buffer and use Hashing algorithm which improves the performance of the job.It uses **GENERAL** or **SEQ.NUM.** algorithm.

What are the Repository Tables in DataStage?

- A datawarehouse is a repository(centralized as well as distributed) of Data, able to answer any adhoc, analytical, historical or complex queries.
- Metadata is data about data.
- The repository environment encompasses all corporate metadata resources: database catalogs, data dictionaries, and navigation services. Metadata includes things like the name, length, valid values, and description of a data element.
- Metadata is stored in a data dictionary and repository.

How one source columns or rows to be loaded in to two different tables?

- For Columns - We can directly map the single source columns to two different targets.

For Rows - We have to put some constraint (condition).

What are the types of views in Datastage Director?

- There are three types of views in Datastage Director :
Job view :
 - Dates of Jobs Compiled.
 - Log view :
 - Status of Jobs last Run.
 - Status view :
- Warning messages, Event messages, Program Generated messages.

What are Sequencers?

- Sequencers are job control programs that execute other jobs with preset job parameters.

What can we do with DataStage Director?

- Validating
- Scheduling
- Executing
- Monitoring Jobs (server Jobs).

Which algorithm you used for your hashfile?

- It uses GENERAL or SEQ.NUM. algorithm

What is a folder?

- Folder contains repository objects such as sources, targets, mappings, transformation which are helps logically organize our data warehouse.

What are the difference types of stages?

- **Active Stage** : In which Transformation, Aggregation etc are done. Ex: Transformer, Aggregator.
- **Passive Stage** : In which Read/Write Process is done. Ex: ODBC, Sequential, Hash File.

What is Datastage Engine?

- It is a JAVA engine running at the background.

What all the types of Jobs you developed?

- Server Job
- Parallel Job
- Sequencer Job
- Container Job.

How do you register plug-ins?

- Using DataStage Manager.

What is orabulk Stage?

- Stage is used to Bulk Load the Oracle Target Database.

What are iConv and oConv?

- These are the Date functions, which we use to convert the Dates from internal format to External format
- iConv - External to Internal
- oConv- Internal to External

What are the different plug-ins Stages used in your projects?

- Oracle bulk
- BCP Load

Whats difference between operational data stage (ODS) and data warehouse?

- A dataware house is a decision support database for organisational needs. It is subject oriented, non volatile, integrated ,time varient collect of data.
- ODS(Operational Data Source) is a integrated collection of related information . it contains maximum 90 days information.
- ODS is nothing but operational data store is the part of transactional database. this db keeps integrated data from different tdb and allow common operations across organisation. eg: banking transaction.
- In simple terms ODS is dynamic data.

How do u convert the columns to rows in DataStage?

- Using Pivot Stage.

What is environment variables?

- Basically Environment variable is predefined variable those we can use while creating DS job. We can set either as Project level or Job level. Once we set specific variable that variable will be available into the project/job.

What are Stage Variables, Derivations and Constants?

- **Stage Variable** - An intermediate processing variable that retains value during read and does not pass the value into target column.
- **Derivation** - Expression that specifies value to be passed on to the target column.
- **Constant** - Conditions that are either true or false that specifies flow of data with a link.

What is OCI?

- If you mean by Oracle Call Interface (OCI), it is a set of low-level APIs used to interact with Oracle databases. It allows one to use operations like logon, execute, parss etc. using a C or C++ program.

What are an Entity, Attribute and Relationship?

- An entity represents a chunk of information. In relational databases, an entity often maps to a table.
- An attribute is a component of an entity and helps define the uniqueness of the entity. In relational databases, an attribute maps to a column.
- The entities are linked together using relationships.

How to join the multiple sources?

- An entity represents a chunk of information. In relational databases, an entity often maps to a table.
- An attribute is a component of an entity and helps define the uniqueness of the entity. In relational databases, an attribute maps to a column.
- The entities are linked together using relationships.

What is the difference between Server Job and Parallel Jobs?

- Server Jobs works in sequential way while parallel jobs work in parallel fashion (Parallel Extender work on the principal of pipeline and partition) for I/O processing.

Define dynamic arrays?

- Dynamic arrays map the structure of DataStage file records to character string data. Any character string can be a dynamic array. A dynamic array is a character string containing elements that are substrings separated by delimiters.

Where the DataStage stored his repository?

- DataStage stored his repository in IBM Universe Database.

What is transformer stage?

- Transformer stages do not extract data or write data to a target database. They are used to handle extracted data, perform any conversions required, and pass data to another Transformer stage or a stage that writes data to a target data table.

How do you schedule or monitoring the job?

- Using the DataStage Director we can schedule or monitor the job.

Define universe stage?

- A stage that extracts data from or loads data into a Universe database using SQL. Used to represent a data source, an aggregation step, or a target data table.

What are the types of stages?

- Built-in stages: Supplied with DataStage and used for extracting, aggregating, transforming, or writing data.
- Plug-in stages : Additional stages defined in the DataStage Manager to perform tasks that the built-in stages do not support.

What are the advantages of data warehousing?

- Capitalizes on the potential value of the organization's information.
- Improves the quality and accessibility of data.
- Combines valuable archive data with the latest data in operational sources.
- Increases the amount of information available to users.
- Reduces the requirement of users to access operational data.
- Reduces the strain on IT departments, as they can produce one database to serve all user groups.
- Allows new reports and studies to be introduced without disrupting operational systems.
- Promotes users to be self sufficient.

What is merge stage?

- The Merge stage combines a sorted master data set with one or more sorted update data sets. The columns from the records in the master and update data sets are merged so that the output record contains all the columns from the master record plus any additional columns from each update record.

What is the difference between datastage manager and datastage administrator?

- **DataStage Manager** : A user interface used to view and edit the contents of the Repository.
- DataStage Administrator** : A user interface used to set up DataStage users, create and move projects, and set up purging criteria.

How we can reuse the components?

- Using the Shared and Local Containers.

What can we do with DataStage Director?

- Validating
- Scheduling
- Executing and Monitoring Jobs (server Jobs)

What are the Types of Jobs?

- Server Jobs : Designed, developed and compiled are done in Designer. Validated, Scheduled, Executed and Monitoring are done in Director. Win NT/ Unix system Server.
- Parallel Jobs : Designed, developed and compiled are done in Designer Validated, Scheduled, Executed and Monitoring are done in Director. Server should be on Unix System Only.
- Mainframe Jobs : Designed and developed are done in Designer (win) Compiled, Validated, Scheduled, Executed and Monitored in Mainframe Systems Only.

<http://questionsdoor.com/datastage-interview-questions/>