Top 50 Datastage Interview Questions & Answers

1) Define Data Stage?

A data stage is basically a tool that is used to design, develop and execute various applications to fill multiple tables in data warehouse or data marts. It is a program for Windows servers that extracts data from databases and change them into data warehouses. It has become an essential part of IBM WebSphere Data Integration suite.

2) Explain how a source file is populated?

We can populate a source file in many ways such as by creating a SQL query in Oracle, or by using row generator extract tool etc.

3) Name the command line functions to import and export the DS jobs?

To import the DS jobs, dsimport.exe is used and to export the DS jobs, dsexport.exe is used.

4) What is the difference between Datastage 7.5 and 7.0?

In Datastage 7.5 many new stages are added for more robustness and smooth performance, such as Procedure Stage, Command Stage, Generate Report etc.

5) In Datastage, how you can fix the truncated data error?

The truncated data error can be fixed by using ENVIRONMENT VARIABLE ' IMPORT_REJECT_STRING_FIELD_OVERRUN'.

6) Define Merge?

Merge means to join two or more tables. The two tables are joined on the basis of Primary key

columns in both the tables.

7) Differentiate between data file and descriptor file?

As the name implies, data files contains the data and the descriptor file contains the description/information about the data in the data files.

8) Differentiate between datastage and informatica?

In datastage, there is a concept of partition, parallelism for node configuration. While, there is no concept of partition and parallelism in informatica for node configuration. Also, Informatica is more scalable than Datastage. Datastage is more user-friendly as compared to Informatica.

9) Define Routines and their types?

Routines are basically collection of functions that is defined by DS manager. It can be called via transformer stage. There are three types of routines such as, parallel routines, main frame routines and server routines.

10) How can you write parallel routines in datastage PX?

We can write parallel routines in C or C++ compiler. Such routines are also created in DS manager and can be called from transformer stage.

11) What is the method of removing duplicates, without the remove duplicate stage?

Duplicates can be removed by using Sort stage. We can use the option, as allow duplicate = false.

12) What steps should be taken to improve Datastage jobs?

In order to improve performance of Datastage jobs, we have to first establish the baselines. Secondly, we should not use only one flow for performance testing. Thirdly, we should work in increment. Then, we should evaluate data skews. Then we should isolate and solve the problems, one by one. After that, we should distribute the file systems to remove bottlenecks, if any. Also, we should not include RDBMS in start of testing phase. Last but not the least, we should understand and assess the available tuning knobs.

13) Differentiate between Join, Merge and Lookup stage?

All the three concepts are different from each other in the way they use the memory storage, compare input requirements and how they treat various records. Join and Merge needs less memory as compared to the Lookup stage.

14) Explain Quality stage?

Quality stage is also known as Integrity stage. It assists in integrating different types of data from various sources.

15) Define Job control?

Job control can be best performed by using Job Control Language (JCL). This tool is used to execute multiple jobs simultaneously, without using any kind of loop.

16) Differentiate between Symmetric Multiprocessing and Massive Parallel Processing?

In Symmetric Multiprocessing, the hardware resources are shared by processor. The processor has one operating system and it communicates through shared memory. While in Massive Parallel processing, the processor access the hardware resources exclusively. This type of processing is also known as Shared Nothing, since nothing is shared in this. It is faster than the Symmetric Multiprocessing.

17) What are the steps required to kill the job in Datastage?

To kill the job in Datasatge, we have to kill the respective processing ID.

18) Differentiate between validated and Compiled in the Datastage?

In Datastage, validating a job means, executing a job. While validating, the Datastage engine verifies whether all the required properties are provided or not. In other case, while compiling a job, the Datastage engine verifies that whether all the given properties are valid or not.

19) How to manage date conversion in Datastage?

We can use date conversion function for this purpose i.e. Oconv(Iconv(Filedname,"Existing Date Format"),"Another Date Format").

20) Why do we use exception activity in Datastage?

All the stages after the exception activity in Datastage are executed in case of any unknown error occurs while executing the job sequencer.

21) Define APT_CONFIG in Datastage?

It is the environment variable that is used to identify the *.apt file in Datastage. It is also used to store the node information, disk storage information and scratch information.

22) Name the different types of Lookups in Datastage?

There are two types of Lookups in Datastage i.e. Normal lkp and Sparse lkp. In Normal lkp, the data is saved in the memory first and then the lookup is performed. In Sparse lkp, the data is directly saved in the database. Therefore, the Sparse lkp is faster than the Normal lkp.

23) How a server job can be converted to a parallel job?

We can convert a server job in to a parallel job by using IPC stage and Link Collector.

24) Define Repository tables in Datastage?

In Datastage, the Repository is another name for a data warehouse. It can be centralized as well as distributed.

25) Define OConv () and IConv () functions in Datastage?

In Datastage, OConv () and IConv() functions are used to convert formats from one format to another i.e. conversions of roman numbers, time, date, radix, numeral ASCII etc. IConv () is basically used to convert formats for system to understand. While, OConv () is used to convert formats for users to understand.

26) Explain Usage Analysis in Datastage?

In Datastage, Usage Analysis is performed within few clicks. Launch Datastage Manager and right click the job. Then, select Usage Analysis and that's it.

27) How do you find the number of rows in a sequential file?

To find rows in sequential file, we can use the System variable @INROWNUM.

28) Differentiate between Hash file and Sequential file?

The only difference between the Hash file and Sequential file is that the Hash file saves data on hash algorithm and on a hash key value, while sequential file doesn't have any key value to save the data. Basis on this hash key feature, searching in Hash file is faster than in sequential file.

29) How to clean the Datastage repository?

We can clean the Datastage repository by using the Clean Up Resources functionality in the Datastage Manager.

30) How a routine is called in Datastage job?

In Datastage, routines are of two types i.e. Before Sub Routines and After Sub Routines. We can call a routine from the transformer stage in Datastage.

31) Differentiate between Operational Datastage (ODS) and Data warehouse?

We can say, ODS is a mini data warehouse. An ODS doesn't contain information for more than 1 year while a data warehouse contains detailed information regarding the entire business.

32) NLS stands for what in Datastage?

NLS means National Language Support. It can be used to incorporate other languages such as French, German, and Spanish etc. in the data, required for processing by data warehouse. These languages have same scripts as English language.

33) Can you explain how could anyone drop the index before loading the data in target in Datastage?

In Datastage, we can drop the index before loading the data in target by using the Direct Load functionality of SQL Loaded Utility.

34) Does Datastage support slowly changing dimensions ?

Yes. Version 8.5 + supports this feature

35) How can one find bugs in job sequence?

We can find bugs in job sequence by using DataStage Director.

36) How complex jobs are implemented in Datstage to improve performance?

In order to improve performance in Datastage, it is recommended, not to use more than 20 stages in every job. If you need to use more than 20 stages then it is better to use another job for those stages.

37) Name the third party tools that can be used in Datastage?

The third party tools that can be used in Datastage, are Autosys, TNG and Event Co-ordinator. I have worked with these tools and possess hands on experience of working with these third party tools.

38) Define Project in Datastage?

Whenever we launch the Datastage client, we are asked to connect to a Datastage project. A Datastage project contains Datastage jobs, built-in components and Datastage Designer or User-Defined components.

39) How many types of hash files are there?

There are two types of hash files in DataStage i.e. Static Hash File and Dynamic Hash File. The static hash file is used when limited amount of data is to be loaded in the target database. The dynamic hash file is used when we don’t know the amount of data from the source file.

40) Define Meta Stage?

In Datastage, MetaStage is used to save metadata that is helpful for data lineage and data analysis.

41) Have you have ever worked in UNIX environment and why it is useful in Datastage?

Yes, I have worked in UNIX environment. This knowledge is useful in Datastage because sometimes one has to write UNIX programs such as batch programs to invoke batch processing etc.

42) Differentiate between Datastage and Datastage TX?

Datastage is a tool from ETL (Extract, Transform and Load) and Datastage TX is a tool from EAI (Enterprise Application Integration).

43) What is size of a transaction and an array means in a Datastage?

Transaction size means the number of row written before committing the records in a table. An array size means the number of rows written/read to or from the table respectively.

44) How many types of views are there in a Datastage Director?

There are three types of views in a Datastage Director i.e. Job View, Log View and Status View.

45) Why we use surrogate key?

In Datastage, we use Surrogate Key instead of unique key. Surrogate key is mostly used for retrieving data faster. It uses Index to perform the retrieval operation.

46) How rejected rows are managed in Datastage?

In the Datastage, the rejected rows are managed through constraints in transformer. We can either place the rejected rows in the properties of a transformer or we can create a temporary storage for rejected rows with the help of REJECTED command.

47) Differentiate between ODBC and DRS stage?

DRS stage is faster than the ODBC stage because it uses native databases for connectivity.

48) Define Orabulk and BCP stages?

Orabulk stage is used to load large amount of data in one target table of Oracle database. The BCP stage is used to load large amount of data in one target table of Microsoft SQL Server.

49) Define DS Designer?

The DS Designer is used to design work area and add various links to it.

50) Why do we use Link Partitioner and Link Collector in Datastage?

In Datastage, Link Partitioner is used to divide data into different parts through certain partitioning methods. Link Collector is used to gather data from various partitions/segments to a single data and save it in the target table.