

## **DATAWAREHOUSING CONCEPTS**

- Datawarehouse Concepts
- Slowly Changing Dimension Concept with examples
- Dimensional Data Model and ETL Model difference
- Different types of Dimensions and Facts

## **INTRODUCTION TO DATASTAGE**

- Brief history of DataStage
- Various versions available
- Introduction to DataStage server components
  - Repository
  - DataStage server
  - DataStage package installer
- Introduction to DataStage client components
  - DataStage administrator
  - DataStage designer
  - DataStage director
  - DataStage manager (removed in latest version)

**Comparison between different Versions of DataStage( Both Architectural and Technical )**

## **OVERVIEW OF IBM WEB SPHERE DATASTAGE AND QUALITY STAGE DESIGNER**

- How to connect to a project
- IBM information server repository
- Developing a job
- Introduction to job properties
- Introduction to job parameters
- Introduction to table definitions
- Importing and exporting from the repository

## **PARALLEL PROCESSING IN DATASTAGE**

- Client server architecture for data warehouse
  - Various server hardware available
  - SMP (symmetric multiprocessing)
  - Clusters
  - MPP ( massively parallel processing)

## **TYPES OF PARALLEL PROCESSING IN DATASTAGE**

- pipeline parallelism
- partition parallelism
- Combining pipeline and partition parallelism
- Repartitioning data
- Parallel processing environments
- The configuration file

## **TYPES OF PARTITIONING TECHNIQUES IN DATASTAGE**

- round robin
- random
- same
- entire
- hash by field
- modulus
- range
- DB2
- Auto

## **TYPE OF COLLECTING TECHNIQUES IN DATASTAGE**

- round robin
- ordered
- sorted merge
- auto

## **MECHANICS OF PARTITIONING AND COLLECTING**

- Difference between a passive stage and active stage
- Concepts about the usage of partitioning and collecting methods together in a job .

### **Handling metadata in DataStage**

- Run Time column propagation (RPC)
- Table definitions
- Schema files and partial schemas
- Data types
- Data and time formats
- Complex data types

### **Handling oracle enterprise stage in parallel jobs**

- loading tables
- type conversions writing to oracle
- updating an oracle database
- deleting rows from an oracle database
- inserting data to an oracle database
- reading an oracle database
- performing a direct lookup on an oracle database table

- using SQL builder

### **Handling oracle enterprise stage in parallel jobs**

- handling special characters(# and \$)
- loading tables
- type conversions writing to oracle
- updating an oracle database
- deleting rows from an oracle database
- leading an oracle database
- reading an oracle database
- performing a direct lookup on an oracle database table
- using SQL builder

### **Handling transformer stage in parallel jobs**

- how it is different from server transformer stage
- creating and deleting columns
- handling null values
- defining constraints and handling otherwise links
- specifying link order
- defining local stage variables
- what is a BASIC transformer stage
- transformer functions
- To perform different operations like Remove Duplicate , Copying data , sorting ,Filter, Aggregation , performing arithmetic operations and other significant functionalities using transformer stage .

### **Combining data in DataStage parallel jobs**

- horizontal and vertical combining
- Join stage
  - inner
  - Left outer
  - Right outer
  - Full order
- Look up stage
- Merge stage
- Comparison between join /merge/look up stage
- Partitioning in reference links
- Aggregator stage
- Funnel stage
  - Continuous Funnel
  - Sort funnel
  - Sequence Funnel

## **MOST USEFUL STAGES IN DATASTAGE PARALLEL JOBS**

- Sort stage

- sequential sort
- Parallel sort
- Unique Sort
- Stable Sort
- Partitioning requirement
- Different scenarios to use sort stage .
- Remove duplicates stage
  - Requirements for using remove duplicate stage
  - How to remove duplicate without using remove duplicate stage
  - Different real time scenarios for Remove duplicate stage
- Modify stage
  - Dropping and keeping columns
  - Changing data type
  - Null handling
  - Different scenarios to use Modify stage .
- Pivot stage
  - Horizontal Pivot
  - Vertical Pivot
  - Limitations in Pivot Stage
  - Different real time scenarios for Pivot stage
- Copy stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Filter stage ( Concepts , Scenarios, Real Time Implementation Examples )
- External filter stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Switch stage ( Concepts , Scenarios, Real Time Implementation Examples , Difference between Filter and Switch )
- Compress stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Expand stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Encode stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Decode stage ( Concepts , Scenarios, Real Time Implementation Examples )
- FTP enterprise stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Generic stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Surrogate key generator stage( Concepts , Scenarios, Real Time Implementation Examples )
- Aggregator Stage ( Concepts , Scenarios, Real Time Implementation Examples )

## **CAPTURING CHANGES IN DATA STAGE PARALLEL JOBS**

- Change capture stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Change apply stage (Concepts , Scenarios, Real Time Implementation Examples )
- Difference stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Compare stage ( Concepts , Scenarios, Real Time Implementation Examples )
- Slowly changing dimension stage (SCD1 ,SCD2 , SCD 3 )

## **DATABASE STAGES**

- Oracle Enterprise Stage
- DB2 Stage
- Netezza Stage
- ODBC Stage

- Teradata Stage
- Informix Stage

## **HANDLING DEVELOP / DEBUG STAGES IN DATASTAGE PARALLEL JOBS**

- Head stage
  - Head stage
  - Head stage default behavior
  - Skipping data
- Tail stage
- Sample stage
- Peek stage
- Row generator stage
  - How to specify data to be generated
  - Generating data in parallel
- Column generator stage
- Write range map stage

## **HANDLING DIFFERENT TYPES OF FILES IN DATASTAGE**

- Sequential File stage
- DataSet Stage
- Lookup File Set Stage
- Complex Flat File Stage
- File Set Stage
- Difference among the above file stages

## **HANDLING RESTRUCTURE STAGES IN DATASTAGE PARALLEL JOBS**

- column import stage
- Column export stage
- Make sub record stage
- Split sub record stage
- Combine records stage
- Promote sub record stage
- Make vector stage
- Split vector stage

## **HANDLING XML FILE IN DATASTAGE PARALLEL JOBS**

- Introduction to XML files
- Using the XML meta data importer
- Using xml input stage
  - Validating documents and schemas
  - Processing namespaces
  - Supported x path expressions

## **USING XML OUTPUT STAGE**

- Processing names spaces
- Supported x path expressions
- Aggregating input rows on output
- Writing output to your file system
- Processing NULLS and empty values
- How repetition paths work

## **USING XML TRANSFORMER STAGE**

- Optimizing performance in server and parallel jobs

## **DATABASE SPARSE LOOKUP VS. JOIN IMPROVING PERFORMANCE IN PARALLEL JOBS**

- Understanding a flow
- Performance monitoring
- Resolving bottlenecks
- Ensuring data is evenly partitioned

## **INTRODUCTION TO PROGRAMMING COMPONENTS/ROUTINES**

- Transform functions
- Before /after subroutines
- Custom universe functions
- Subroutines
- Creating a routine
- Defining custom transforms

## **CREATING A JOB SEQUENCE**

- Overview of activity stages
- Triggers
- Expressions
- Job activity properties
- Routine activity properties

- Email notification activity properties
- Wait for file activity properties
- exception activity properties
- Nested condition activity properties
- Start loop activity properties
- End loop activity properties
- User variables activity properties
- Compiling and restarting the job sequence
- Creating a job sequence
- Handling Errors in sequence
- Creating Custom triggers
- Passing Sequence job parameters to Parallel job Activities
- How to control the flow of the jobs when there is a failure
- How to find the point of failure in a sequence
- Batch Jobs

## ADVANCED CONCEPTS IN DATASTAGE

- Achieving reusability in DataStage using containers
  - Types of containers
  - Local containers
  - Server shared containers
  - Parallel shared containers
- Creating a shared containers
- Using shared containers in DataStage jobs
- Converting shared containers to local containers
- Deconstruction of shared containers
- Usage of administrator client in datastage
  - Adding environment variables
  - Setting job parameters default values
  - Changing license details
  - Handling projects
- Multiple instances of jobs in DataStage
- DataStage job control utility
- Jobs – compilation execution and checking of logs using DataStage tool
- Handling multilingual data in DataStage
- How to enable NLS on DataStage
- Orchestrate architecture and commands
- Orchestrate parallel processing framework in datastage
- Orchestrate utility in DataStage
- Surrogate key generation using DataStage
- Version control in DataStage
- Difference between DataStage Parallel and Server Jobs .

## UNIX CONCEPTS

- Basic Unix Commands
- How to use these unix commands in Jobs/Sequences
- How to call scripts/run commands in Sequence Job

- How to write a shell Script

## SQL

- How to write SQL Commands
- How to convert a DataStage design into SQL
- How to convert an SQL to a DataStage design
- How to do unit /integration testing using SQL and Jobs .