ETL Solution using IBM DataStage

Course description
This course covers the basic of Data Warehousing fundamentals, ETL Fundamentals and focuses on implementing Extract, Transform and Load strategy using IBM DataStage. Learn the real work Data Transformation rules, Data Quality Check, Basic and Advance Business Rules, Data Loading Strategies and optimized ETL solutions. The training is full of real time case studies, Problem/Solutions based model and based on current job requirements.

Student Take away
Study Material
Learning stuff
Sample project for practice

ETL Solution using IBM DataStage web based Online Training Curriculum
IBM Info sphere Data Stage and Quality Stage 11.3

Data Warehouse Fundamentals
- An introduction to Data Warehousing
- Purpose of Data Warehouse
- Data Warehouse Architecture
- Operational Data Store OLTP Vs Warehouse Applications

Data Marts
- Data marts Vs Data Warehouses
- Data Warehouse Life cycle

Data Modeling
- Introduction to Data Modeling
- Entity Relationship model (E-R model)
- Data Modeling for Data Warehouse
- Normalization process
- Dimensions and fact tables
- Star Schema and Snowflake Schemas

ETL Design Process
- Introduction to Extraction
- Transformation & Loading
- Types of ETL Tools
- Key tools

Introduction to Data stage latest version
- DataStage introduction
- IBM information Server architecture
- DataStage components
- DataStage main functions
- Client components

Data Stage Designer
- Introduction to Data stage Designer
- Importance of Parallelism
- Pipeline Parallelism
- Partition Parallelism
- Partitioning and collecting Symmetric
- Multi Processing (SMP)
- Massively Parallel Processing (MPP)
- Partition techniques
- Data stage Repository
- Palette Passive and Active stages
- Job design overview
- Designer work area Annotations
- Creating jobs importing flat file
- Definitions
- Managing the Metadata environment
- Dataset management
- Deletion of Dataset Routines Arguments

Working with Parallel Job Stages
- Database Stages
- Oracle connector Teradata Connector ODBC

File Stages
- Sequential file Dataset File set Lookup file set-Complex Flat File Stage

Processing Stages
- Copy Filter Funnel Sort
- Remove duplicate
- Aggregator Modify Compress Expand Decode Encode
- Switch Pivot stage
- Lookup Join Merge look-up
- Join and merge change capture Change apply
- Compare Difference Surrogate key generator Transformer

Debug Stages
- Head Tail Peek Column generator
- Row generator
- Write Range Map Stage

Routines creation

> **Advanced Stages in Parallel Jobs (Version 11.3)**

Range Look process
- Surrogate key generator stage
- Slowly changing dimension stage
- I-way stage
- FTP stage-Pivot Enterprise Job performance analysis
- Resource estimation- Performance Optimizer
- Slowly Changing Dimensions implementation
- Transformer stage looping condition
- Transformer stage Last Row handling

DataStage Director
- Introduction to Data stage Director
- Validating Data stage Jobs
- Executing Data stage jobs
- Job execution status monitoring a job
- Job log view
- Job scheduling
- Creating batches scheduling batches

DATASTAGE Administrator
- Data stage project Administration
- Editing projects
- Adding a Project
- Deleting projects
- Cleansing up project files
- Environmental Variables
- Environment management
- Auto purging Runtime
- Column Propagation (RCP)
- Add checkpoints for sequencer
- NLS configuration
- Generated OSH (Orchestra Engine) System formats like data, timestamp Project protect version details

Job Sequencers

- Arrange job activities in Sequencer
- Triggers in Sequencer
- Re-stability
- Re-cover ability
- Notification activity
- Terminator activity
- Wait for file activity
- Start Look activity
- Execute Command activity
- Nested Condition activity
- Exception handling activity
- User Variable activity
- End Loop activity
- Adding Checkpoints

Info sphere Quality Stage

Why Data Quality?
Data Quality Challenges
Types of Data Quality Tools Provided by IBM
Differences between IA and QS
Quality stage Architecture

➢ **Data stage Quality Stages**

Investigate Stage

- Default Class Descriptions
- Word Investigation
- Character Discrete Investigation
- Character concatenate investigation

Standardize Stage

- Standardize Process
- Domain Specific Rule sets
- Domain Pre processing Rule sets
- Creation of Custom Rule sets with Examples
- SEP LIST
- STRIP LIST
- Classification file
- Dictionary Files
- Pattern Action File
- Look-up Tables
- Over ride Tables
- Introduction to Pattern Action language
- Types of Patterns (Conditional and Unconditional)
- Build customized Action statements using PAL with Examples
- Standardize Quality Assessment Report (SQA)

Match Stage
- Match Process
- Creation of Match Passes
- Match Frequency Stage & Reports
- Un duplicate Match Stage with Examples
- Reference Match Stage with Examples

Survive Stage
Importance of Survive Stage
Build Survive Process
Implementation of Survive Rules

- ➢ **Explanation about the entire Data Quality Life Cycle**

IBM Information Server Administration
- ➢ IBM Info sphere Data Stage administration
- ➢ Opening the IBM Information Server Web console
- ➢ Setting up a project ion
- ➢ Console Customizing
- ➢ Project dashboard
- ➢ Setting up security
- ➢ Creating users in the console
- ➢ Assigning security roles to users and groups
- ➢ Managing licenses
- ➢ Managing active sessions
- ➢ Managing logs
- ➢ Managing schedules
- ➢ Backing up and restoring IBM Information Server

DataStage 11.3 Real time project

- ➢ **Project architecture and BRD discussion**

- ➢ **Dimensional tables and fact tables with modeling**

- ➢ **Flow of subject area discussion**

- ➢ **Design of HLDS and LLDS for a project**

- ➢ **Project flow-job design process with ETL Documents**

- ➢ **Complex jobs discussion-unit testing process**

- ➢ **System & User Acceptance**

- ➢ **Regression & End-to-End Testing**

- ➢ **Deployment Process of code to different phases**

DataStage 11.3 Real time project continues…

- ➢ **Creation of job design documents or overview docs, tech specs**

- ➢ **Production support process**

- ➢ **UNIX scripting for automation of code**

- ➢ **Discussion of scheduling process with Control-M/Auto system**

- ➢ **Fixing of Defects and Problem Tickets and Incidents**

Additional Features

- ➢ **Data stage project on Banking & Insurance & Health Care domain.**

- ➢ **Data stage Certification Guidance.**

- ➢ **Performance Tuning of Parallel Jobs**

- ➢ **DataStage Installation process and setup**

- ➢ **Well Versed Materials Which Covers Data warehousing Basics, DataStage Concepts UNIX Commands, Shall Script, Databases.**