Talend Data Integration Basics

Course Objectives

Talend Data Integration provides an extensible, highly scalable set of tools to access, transform, and integrate data from any business system. This course will teach you from basics to advanced features of Talend Data Integration as quickly as possible. Participants can work in teams on projects shared on a remote repository to monitor jobs and database changes. This three days course is designed for anyone who wants to perform data integration and manangement tasks.

What Will You Learn?

- Create a project
- Create and run a Job that reads, converts, and writes data
- Merge data from several sources within a Job
- Save a schema for repeated use
- Create and use metadata and context variables within Jobs
- Connect to, read from, and write to a database from a Job
- Access a web service from a Job
- Work with master Jobs and subJobs
- Build, export, and test run Jobs outside Studio
- Invoke basic error-handling techniques
- Use best practices for Job and component naming, hints, and documentation
- Start and connect Talend Studio to a remote repository
- Use SVN branches in Studio
- Run a Job in Studio on a remote Job server
- Monitor host CPU and JVM memory in real time during Job execution
- Use debugging features in Studio
- Configure a Talend project to capture statistics and logs, and monitor them from Activity Monitoring Console (AMC)
- Implement several methods of parallel execution in a Talend Job
- Create Joblets
- Create a unit test from a working Job
- Configure a database to monitor and log changes in a separate change data capture (CDC) database, and use it to perform incremental updates between the source and target

Target Audience

This course is suggested to Data Engineers, Dat Analyst, individuals who involved in ETL.

Training Outline

Getting Started

- Starting Talend Studio
- Creating a first Job
- Running a Job

Working with Files

- Reading an input file
- Transforming data
- Running a Job
- Combining columns
- Duplicating a Job

Joining Data Sources

- Creating metadata
- Joining data sources
- Capturing rejects
- Correcting a lookup

Filtering Data

- Filtering output data
- Using multiple filters

Using Context Variables

- Understanding and using context variables
- Using repository context variables

Error Handling

- Detecting and handling basic errors
- Raising a warning

Generic Schemas

- Setting up sales data files
- Creating customer metadata
- Creating product metadata

Working with Databases

- Creating database metadata
- Creating a customer table
- Creating a product table
- Setting up a sales table
- Joining data
- Finalizing a Job

Creating Master Jobs

- Controlling Job execution using a master Job

Working with Web Services

- Accessing a web service

Running Jobs Standalone

- Building a Job
- Modifying a Job

Documenting a Job

- Using best practices while documenting a Job

Connecting to a Remote Repository

- Creating a remote connection

SVN in Studio

- Copying a Job to a branch
- Comparing Jobs
- Resetting a branch

Remote Job Execution

- Creating and running a Job remotely

Resource Usage and Basic Debugging

- Using Memory Run to view real-time resource usage
- Debugging Jobs using Debug Run

Activity Monitoring Console (AMC)

- Configuring statistics and logging
- Using Activity Monitoring Console (AMC)

Parallel execution

- Writing large files
- Writing to databases
- Parallelizing automatically
- Partitioning

Joblets

- Creating a Joblet from an existing Job
- Creating a Joblet from scratch
- Triggering Joblets

Unit Test

- Creating a unit test

Change Data Capture

- Examining databases
- Configuring the CDC database

- Monitoring changes
- Updating a warehouse
- Resetting the databases

Prerequisite

Basic knowledge of computing, including familiarity with Java or another programming language, SQL, and general database concepts.