Streaming and Transporting Data into Apache Hadoop

Course 3576 – 8 Hours

Overview

Over the last few years, many organizations have made a strategic decision to turn into big data. At the heart of this challenge, there is the process of extracting data from many sources, transforming it, and then load it into your Data Warehouse for subsequent analysis. A process known as "Extract, Transform & Load" (ETL).

Apache Hadoop is one of the most common platforms for managing big data, and in this course, we'll introduce you with three common methods of transporting and streaming your data into your Hadoop Data File System (HDFS):

- Data transfer between Hadoop and relational databases using **Apache Sqoop**
- Collecting, aggregating, and moving large amounts of streaming data into Hadoop using **Apache Flume** and **Apache Kafka**

Hand on exercises are included.

Who Should Attend

This course is mainly intended for System Administrators, Developers, Business Intelligence professionals, and other roles responsible for transferring data into Hadoop.

Prerequisites

- Working experience with Databases
- Prior knowledge of Hadoop, working experience with HDFS in particular.
- Basic understanding of Apache Hive

Course Contents

Working with Apache Sqoop

- Introduction to Sqoop
- Import Architecture
- Transferring RDBMS tables into the HDFS
- Integrating with Hive
- Incremental Import
- Export Architecture
- Exporting data from HDFS into RDBMS

Working with Apache Flume

- Introduction to Flume
- Flume Architecture
- Setting up Flume Agents

Working with Apache Kafka

- Introduction to Kafka
- Use Cases
- Kafka in the Enterprise
- Topics and Partitions
- Brokers
- Topic Replication Factor
- Producers
- Consumers
- Zookeeper
- Kafka Basic Configuration