Welcome to ineuron.ai

Create A Data Pipeline based on Messaging Using PySpark and Airflow

Description:

In this Project, we will learn how to Build a Big Data pipeline on AWS at scale. You will be using the Covid-19 dataset. This will be streamed in real time from an external API using NiFi. The complex JSON data will be parsed into CSV format using NiFi and the result will be stored in HDFS. Then this data will be sent to Kafka for data processing using PySpark. The processed data will then be consumed from Spark and stored in HDFS. Then a Hive external table is created on top of HDFS. Finally the cleaned, transformed data is stored in the data lake and deployed. Visualization is then done using Tableau and AWS QuickSight.

Start Date: 3rd Jan'23

Doubt Clear Time:

Course Time: Flexible

Features:

Do Everything In Industry Grade Lab

Learn As Per Your Timeline

Hands-On Industry Real-Time Projects.

Self Paced Learning

Dashboard Access

What we learn:

Real Time Projects

Create A Data Pipeline based on Messaging Using PySpark and Airflow

Build End to End Datapipeline

How to Extract Streaming Data into NFFI

Data Encryption

Data processing using pyspark

Build Dashboards

Requirements:

System with minimum i3 processor or better

At least 4 GB of RAM

Working internet connection

Dedication to learn

Instructor:

Name:

MD Imran

Description:

Working as Data Scientist with experience in solving real world business problems across different domains.

>Welcome to the Course:

>>Course Overview

>>Dashboard Introduction

>Project :- Create A Data Pipeline based on Messaging Using PySpark and Airflow:

>>Introduction of Instructor

>>Introduction to Data Pipeline

>>What is Data Engineering

>>Project Overview

>>End Notes

>>Problem Description

>>Understand the application scope

>>Tour to existing solution

>>End Notes

>>Data Infrastructure: Components used

>>Nifi

>>Hdfs

>>Kafka

>>Hive

>>Airflow

>>Pyspark

>>Aws services

>>Data Visualization Tools

>>End Notes

>>Solution Description

>>Data Architecture

>>Tour to Architecture diagram

>>Cost Involved

>>End Notes

>>system Requirements

>>Create EC2 Instance

>>SSH into EC2 Instance

>>Envirnoment setup with docker

>>Copy Important folder from local to ec2 and give required permissions

>>To connect to different services locally after port forwarding

>>To get into bash shell of different containers

>>Data Extraction with Nifi

>>Data encryption parsing

>>Data sources hdfs kafka

>>streaming data from kafka to pyspark

>>pyspark streaming output kafka nifi hdfs

>>Move Data HDFS to hive Table

>>Dataflow Orchestration with Airflow

>>Connecting with Data Visualization Tool

>>Building Dahbaord and Report

>>End Notes

>>Conclude the project

>>Assignments & External Resources