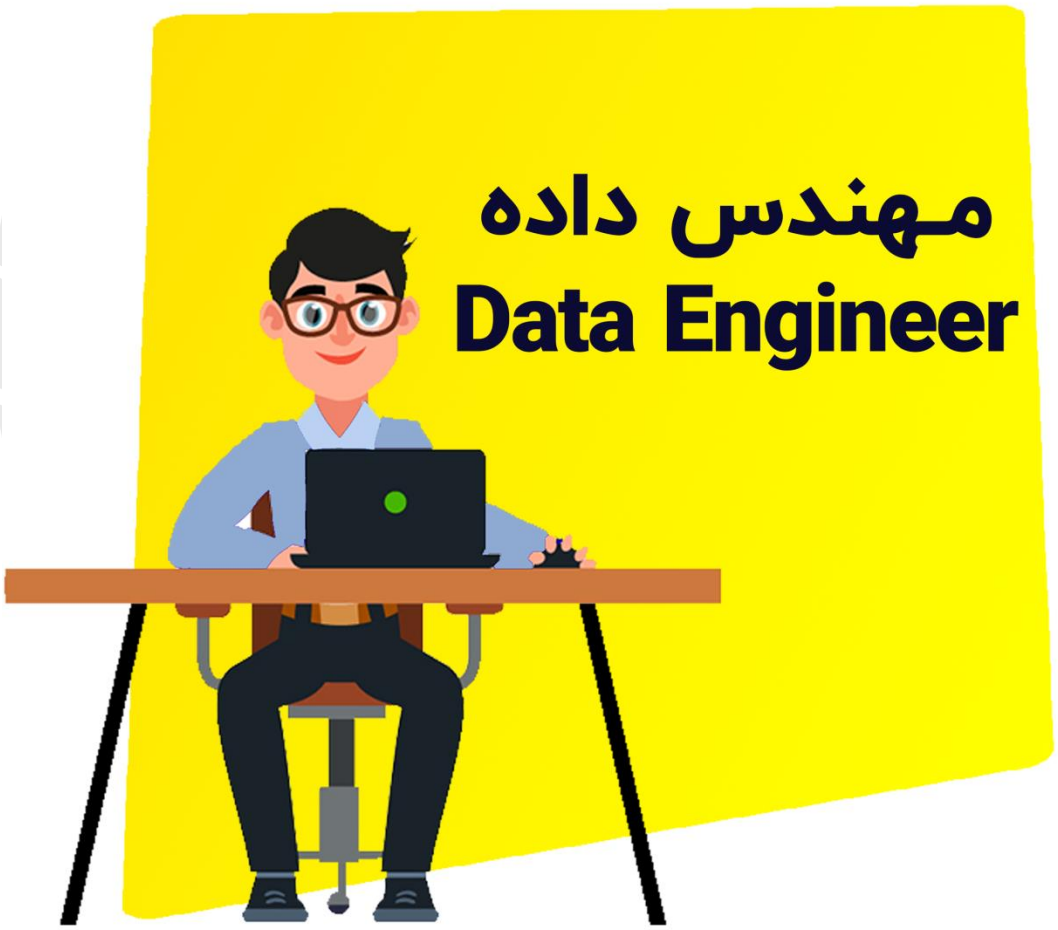
سازمان جهاد دانشگاهی صنعتی شریف

مرکز آموزش های تخصصی کاربردی



ریز سرفصل های دوره آموزشی مهندس داده به شرح ذیل می باشد:

۱-R For Data Engineering

Installation

Data Types

Control Flow in R

Vectorized Thinking

introduction to packages

Introduction to Date & Time manipulation

Introduction to String Manipulation

Importing Data

-Importing Data from Flat Files

-Importing Excel Files

-Importing Data from DBs

-Importing Data from Web and APIs

Data Manipulation

-Data Tidying

-Data Transformation

-Data Fusion

Date & Time Manipulation in depth

String Manipulation in depth

Introduction to dealing with missing Data

۲- Big Data کلان داده

معرفی و استفاده از سیستم پردازش و ذخیره سازی توزیع شده هادوپ جهت ذخیره سازی و پردازش کلان داده ها / Hadoop Ecosystem

-تعریف، توصیف و بررسی ویژگی های کلان داده ها (Big Data)

-داده های جریانی(Stream Data)، مشخصات و تولید کننده های آنها

-معرفی Hadoop به عنوان سکوی پردازش و ذخیره سازی داده در ابعاد کلان

-مدل پردازش توزیع شده نگاشت کاهش/ MapReduce

-سیستم فایل توزیع شده هادوپ / HDFS

-سیستم مدیریت منابع و وظایف در اکوسیستم هادوپ / YARN

-قابلیت ها و توانمندی های Hadoop

-نقاط قوت و ضعف هادوپ در مقایسه با سیستم های پردازشی موجود

-معرفی مدل پردازشی نگاشت کاهش

-مراحل انجام کار در اجرای وظایف نگاشت کاهش

-حل مسئله و توسعه نمونه برنامه های نگاشت کاهش

-مسائل قابل حل در مدل پردازشی نگاشت کاهش و مسائل سازگار با این مدل

-اجرای وظایف نگاشت کاهش در هادوپ و بررسی مراحل اجرا

-معرفی سیستم فایل توزیع شده در Hadoop

-ویژگی های HDFS و نحوه عملکرد HDFS

NameNode و DataNode و وظایف هر کدام

-ساختار داده ها و بلاک ها در HDFS

-معرفی YARN و وظیفه YARN

-اجزاء YARN

-منابع قابل مدیریت و نحوه مدیریت منابع و وظایف توسط YARN

-امکانات YARN برای مدیریت و نظارت بر وظایف

-حوزه کارکردی مناسب برای HDFS، YARN و MapReduce

-برنامه ریزی و منابع لازم برای ایجاد کلاستر هادوپ

-روش های ایجاد کلاستر

-نصب نرم افزار ها و پیش نیاز ها

-نصب، انجام پیکربندی و راه اندازی کلاستر

-فایل های پیکر بندی و انجام پیکربندی های تکمیلی برای عملکرد بهتر کلاستر پردازشی و ذخیره سازی

-تنظیمات Memory و نحوه تخصیص RAM در سرور به اجزاء کوچکتر و ایجاد Container

-تعامل و کار با سیستم فایل توزیع شده Hadoop و انجام اعمال کاربری، مدیریتی و نظارتی در HDFS

-تعامل و کار با YARN

-ارسال وظایف نگاشت کاهش و نظارت و مدیریت آنها به کمک YARN

-مدیریت کلاستر هادوپ

Apache Pig (استفاده از پیگ برای تعامل با داده های کلان)

-معرفی Pig

-ویژگی ها

-کاربرد ها

-اجزاء

-جایگاه Pig در سیستم هادوپ

-مدل اجرایی

-مزایا و معایب Pig

-راه اندازی Pig

-توسعه نمونه برنامه ها و استفاده از Pig جهت تعامل با داده های کلان و انجام پردازش دسته ای و عملیات پاک سازی داده

وارد کردن داده های کلان به سیستم ذخیره سازی توزیع شده هادوپ

-تزریق و ورود داده به سیستم فایل توزیع شده هادوپ

-نحوه نصب و راه اندازی سیستم های تزریق داده

-روش ها و ابزار های ورود داده غیر ساخت یافته

-معرفی Flume به عنوان ابزار ورود داده غیر ساخت یافته و نیمه ساخت یافته به سیستم ذخیره سازی توزیع شده هادوپ

-نصب و راه اندازی به صورت تک نود و کلاستر

-اجزاء و روش کارکرد

-معماری یک کلاستر Ingestion چند لایه برای کنترل و مدیریت ورود داده به کلاستر هادوپ

-ورود داده ساخت یافته از پایگاه داده های رابطه ای به سیستم فایل توزیع شده هادوپ

-معرفی Sqoop به عنوان ابزار ورود داده ساخت یافته

-اجزاء و روش کارکرد Sqoop

-نصب و راه اندازی

-اجرای سناریو های مختلف ورود داده

-امکانات پیش‌رفته Sqoop برای انجام روال های ورود داده

-بررسی و ارزیابی داده های وارد شده به سیستم

Apache Spark Overview (پردازش و تحلیل کلان داده و جریان داده توسط اسپارک)

-معرفی اسپارک و مدل پردازش توزیع شده در اسپارک

-نصب و راه اندازی کلاستر اسپارک

-مفاهیم کار با اسپارک و اسپارک کلاستر

-مدل داده ای RDD

-توسعه نمونه برنامه های اسپارک برای انجام فرآیند های پردازش دسته ای و ETL

-توسعه نمونه برنامه های اسپارک برای تحلیل برخط کلان داده

-کار با Spark SQL

-اتصال اسپارک به دیتابیس

-معرفی، ایجاد و کار با DataFrame

-معرفی و کار با Dataset

-معرفی MLlib جهت انجام فرآیند های یادگیری ماشینی در اسپارک

-توسعه و اجرای روال های تحلیل آماری

-توسعه و اجرای الگوریتم های یادگیری ماشینی در اسپارک

-معرفی Spark Streaming

-توسعه و استفاده از اسپارک برای پردازش جریان داده ای

-مقایسه اسپارک و سایر سکوهای پردازش جریان داده ای

Apache Hive

-معرفی هایو

-مدل اجرای وظایف در هایو

-اجرای سناریو های مختلف تحلیل داده در هایو

-معرفی و انجام پارتیشن بندی و باکت بندی در هایو

-روش های پیوند در هایو

-اتصال اسپارک به هایو

-یکپارچه سازی و استفاده از اسپارک به عنوان موتور اجرایی در هایو

-انجام تنظیمات پیش‌رفته در هایو

-انجام روال های بهبود کارایی در هایو

Data Science with Zeppelin (انجام عملیات تحلیل داده و فرآیند های علم داده ای توسط زپلین)

-معرفی قابلیت های Zeppelin

-کاربرد های Zeppelin برای مصور سازی نتایج

-کاربرد های Zeppelin برای انجام فرایند های علم داده ای

-بررسی و انجام روش های مختلف ارتباط با کلان داده ها توسط زپلین

-تعامل با کلان داده ها و مصور سازی نتایج تحلیل ها توسط زپلین

-مقایسه Zeppelin با سایر Data Science Notebook ها

Splunk for Big Data (استفاده از اسپلانک جهت انجام عملیات تحلیل داده های متنی غیر ساخت یافته/ کلان داده ها)

-معرفی اسپلانک / Splunk

-قابلیت های اسپلانک در حوزه تحلیل داده های ماشین

-نصب و راه اندازی

-منابع مورد نیاز برای راه اندازی اسپلانک به صورت تک نود یا کلاستر

-شاخص/Index گذاری روی داده های ماشین

-انواع ایندکس ها

-اجزاء و معماری اسپلانک

-مبانی جستجو در اسپلانک

-جستجو و گزارش از داده های شاخص گذاری شده

-ساخت گزارش و داشبورد

-مقایسه اسپلانک و Elastic (Splunk vs ELK)

-نحوه ارتباط Splunk به اکوسیستم هادوپ و سیستم ذخیره سازی هادوپ

Qlik View and Hadoop (ارتباط و تعامل با داده های ذخیره شده در اکوسیستم هادوپ به کمک کلیک ویو و کلیک Scence)

مقدمه ای بر نحوه ارتباط با سیستم ذخیره سازی هادوپ و مصور سازی نتایج به کمک ابزارهای Qlik