شناسنامه دوره آموزشی بیگ داده ها و تکنیک‌های کدنویسی

نام دوره: بیگ دیتا و تکنیکهای کدنویسی

سطح دوره: تخصصی

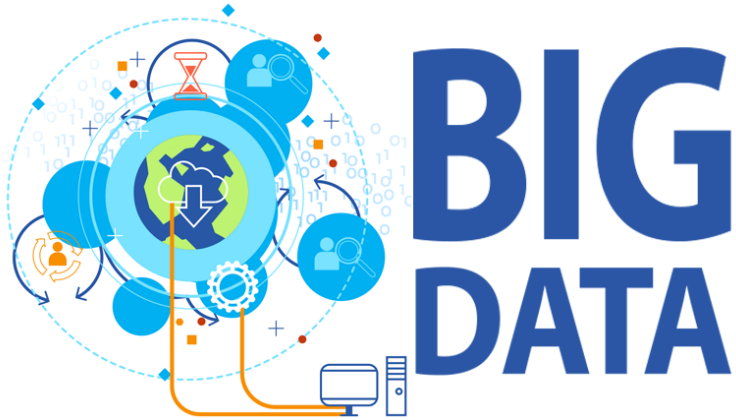
مخاطبین دوره: کارشناسان داده

نوع دوره: کارگاهی

مدت دوره: ۴۰ ساعت

پیش نیاز دوره: بدون پیش نیاز

نحوه برگزاری دوره: ۵ جلسه ۸ ساعته



معرفی دوره:

گسترش استفاده از فناوری اطلاعات در بخش های مختلف کسب و کار، باعث افزایش منبع ارزشمندی به نام داده شده است. هر چند در گذشته نیز سازمان ها این منبع را در اختیار داشتند، اما حجم ، تنوع و سرعت تولید این داده ها به مراتب کمتر بوده است. علم داده به عنوان علمی کاملا کاربردی می تواند پاسخی مناسب به این داده های عظیم تولید شده باشد. به منظور استفاده از این منابع ارزشمند وجود نیروی ماهر بسیار ضروری است. متاسفانه اکثر صاحبان صنایع در دنیا از کمبود نیروی ماهر در این حوزه شکایت دارند.

هدف برگزاری دوره علم داده و بیگ دیتا، توانمندسازی و تسهیل تصمیم‌گیری است. سازمان‌هایی که بر علم داده سرمایه‌گذاری می‌کنند، می‌توانند از شواهد قابل سنجش و مبتنی بر داده برای تصمیم‌سازی در کسب‌وکار خود استفاده کنند. تصمیم‌های داده‌محور می‌تواند منجر به افزایش سود و بهبود بهره‌وری عملیاتی، کارایی کسب‌وکار و جریان‌های کاری بشود. در سازمان‌هایی که با ارباب رجوع سر و کار دارند، علم داده به شناسایی و جلب مخاطبان هدف کمک می‌کند. این دانش همچنین می‌تواند به سازمان‌ها در استخدام نیروهایشان کمک کند. علم داده با پردازش داخلی کاربردها و آزمون‌های احراز صلاحیت داده‌محور، می‌تواند به واحد منابع انسانی سازمان‌ها در انجام انتخاب‌های صحیح‌تر و سریع‌تر در طول فرآیند استخدام کمک کند.

مخاطب این دوره افرادی می باشند که علاقه زیادی به حل مساله با رویکرد داده محور داشته و حوزه علم داده را به عنوان حیطه تخصصی برای خود در نظر گرفته اند و آینده شغلی خود را متخصص علوم داده می بینند.پیش بینی فرایندها، تحلیل سری زمانی،متن کاوی، تحلیل شبکه های اجتماعی و یادگیری عمیق از جمله مسائلی هستند که در این حوزه مطرح می باشند.

متخصصین علوم داده می توانند با استفاده از متدهای یادگیری ماشین با ناظر و بدون ناظر، به دانش پنهان موجود در داده ها دست یابند و آن را آشکار سازند. آموزش مدل های ریاضی به آنها این امکان را می دهد تا بتوانند الگوها را شناسائی کرده و به پیش بینی دقیقتری از آینده برسند. به نوعی می توان گفت که یک دانشمند داده متخصص آماری است که بیشتر از یک آماری کامپیوتر می داند و متخصص کامپیوتری است که بیشتر از یک کامپیوتری به آمار مسلط است.

هادوپ و اسپارک، ابزارهای مهم متن‌باز برای ذخیره و پردازش کارای داده‌های عظیم به صورت توزیع‌شده هستند. در حال حاضر، خانواده‌ای از فناوری‌ها در اطراف هادوپ شکل گرفته‌اند و امکانات مختلفی را در زمینه داده‌های عظیم ارائه می‌کنند. این خانواده که به اکوسیستم هادوپ معروف هستند، در کنار هم امکاناتی کارا و مقیاس‌پذیر برای ذخیره سریع، بازیابی با بار زیاد و پردازش توزیع‌شده را فراهم می‌سازند. در این درس، مخطبان با فناوری هادوپ و اسپارک و امکانات پیرامون آن آشنا می‌شوند و به صورت عملی یک سناریوی فرضی ذخیره و پردازشی با کمک هادوپ پیاده‌سازی می‌شود. همچنین با کاربردها و ابزارهای جدید این خانواده و جایگاه آن‌ها آشنا می‌شویم، و بایدها و نبایدهای استفاده صحیح از این فناوری‌ها را در چارچوب بیان تجارب موفق مرور می‌کنیم.

اهداف دوره:

تربیت مهندس داده و داده کاو به منظور استخراج دانش از داده های موجود سازمان

سرفصل دوره:

- معرفی Big Data و ویژگی‌های آن
- نحوه ی ارزش آفرینی Big Data
- مثال‌هایی از کاربردهای موفق Big Data
- منابع تولید Big Data و ساختار داده‌های تولید شده
- نگرانی‌ها و چالش‌های اصلی در مواجهه با Big Data
- معرفی مدل‌های برنامه‌نویسی و پردازش توزیع شده
- آشنایی با اجزای تشکیل دهنده Hadoop شامل HDFS و MapReduce
- آموزش تنظیم محیط برنامه نویسی هادوپ
- آموزش کارکردن با فایل سیستم هادوپ
- آموزش ایجاد کردن محیط لازم برای کار بر روی هادوپ
- آموزش اجرا و دنبال کردن Job های هادوپ
- آموزش بهینه سازی MapReduce
- آموزش کار با Hive و HBase
- آشنایی با Spark و آموزش کار با آن
- آشنایی با کتابخانه یادگیری ماشین در اسپارک شامل MLlib
- آموزش مصور سازی داده های خروجی گرفته شده از هادوپ
- بررسی مباحث پیش رفته در ایجاد و تعامل با RDD
- کار با Spark SQL
- اتصال اسپارک به دیتابیس
- معرفی، ایجاد و کار با DataFrame

- معرفی و کار با Dataset
- معرفی MLlib جهت انجام فرایند 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
- نحوه ی استفاده از اسپارک و کامپوننت های آن در انجام سناریو های مختلف پالایش و تحلیل داده
- آشنایی و ساخت انباره داده در Spark Delta Lake
- تعریف Cluster Sizing
- بررسی بهترین شیوه ها (Best Practice) در طرح ریزی ایجاد یک کلاستر هادوپ
- ملاحظات یک طرح ریزی مناسب
- نیازسنجی در زمینه حجم داده و میزان درخواست های پردازشی و تحیلی
- مثال و مشخصات Storage / HDD مورد نیاز برای نیازسنجی انجام شده و ملاحظات آن
- نحوه تخصیص منابع RAM و CPU مورد نیاز و ملاحظاتی که باید در نظر گرفت
- سایر منابع مورد نیاز و بهترین شیوه های تقسیم بندی منابع در ایجاد یک کلاستر
- انجام محاسبات و جزئیات کلاستربندی و مقدار دهی پارامترهای هر چارچوب در کلاستر هادوپ
- نصب و راه اندازی کلاستر Hadoop
- نصب و راه اندازی کلاستر Spark

1	**Introduction**	**Course Introduction**
2	Introduction to Hadoop and the Hadoop Ecosystem	Introduction to Hadoop
3	Hadoop Architecture and HDFS	
4	Importing Relational Data with Apache Sqoop	Importing and Modeling Structured Data
5	Introduction to Impala and Hive	
6	Modeling and Managing Data with Impala and Hive	
7	Data Formats	
8	Data Partitioning	
9	Capturing Data with Apache Flume	Ingesting Streaming Data
10	Spark Basics	Distributed Data Processing with Spark
11	Working with RDDs in Spark	
12	Aggregating Data with Pair RDDs	
13	Writing and Deploying Spark Applications	
14	Parallel Processing in Spark	
15	Spark RDD Persistence	
16	Common Patterns in Spark Data Processing	
17	Spark SQL and DataFrames	
18	Conclusion	Course Conclusion

منابع درسی:

- Hadoop- The Definitive Guide, 4th Edition-2015
- Advanced Analytics with Spark-Patterns for Learning from Data
- Machine Learning with Spark Create scalable machine learning applications to power a modern data-driven business using Spark