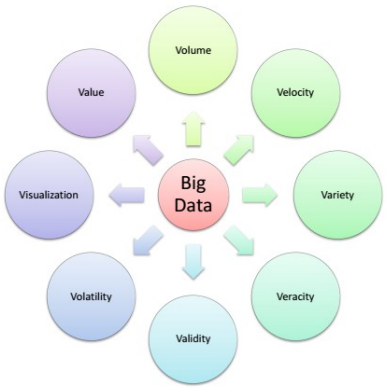
سرفصل درس تحلیل داده های حجیم

(دکتر عباس عکاسی)

مقدمه:

داده های حجیم به داده هایی اطلاق می شود که فراتر از حجم متداول ذخیره سازی، ظرفیت پردازش و محاسبات بانک های اطلاعاتی و یا الگوریتم های تحلیل داده ای باشند. امروزه با گسترش روز افزون تکنولوژی و ابزارهای آن (بانک های اطلاعاتی بزرگ، شبکه های اجتماعی، وب و اینترنت) با انبوهی از داده های تولید شده توسط این ابزارها مواجه هستیم. در واقع در داده ها غرق شده و از کمبود دانش رنج می بریم. کلان داده ها امروزه تبدیل به عنصری جدا ناپذیر در حوزه فناوری ارتباطات گردیده است علی الخصوص با توسعه مرتب شبکه های اجتماعی در دنیای امروز بصورت روزانه حجم بسیار بزرگی از داده ها تولید می گردد که عالوه بر نیاز به ذخیره سازی نیازمند پردازش و مدیریت می باشند. فناوریهای کلاسیک معمول در مواجهه با حجم بسیار بزرگ داده ها یا بسیار کند عمل می کنند و یا اینکه توانایی مدیریت حجم وسیعی از داده ها با انواع گوناگون را نخواهند داشت.آنچه که کلان داده ها و بدنبال آن فناوریهای مرتبط با آن را از مفاهیم متداول در حوزه ی داده مانند انبار داده، هوش تجاری،... متمایز میسازد، امکان پاسخگویی به چالشهایی است که تا کنون یا وجود نداشته اند و یا امکان پاسخگویی به آنها وجود نداشته است. شکل زیر چالشهای مهم مطرح در حوزه ی کلان داده ها را نشان میدهد. سه مورد Volume, Velocity و Variety به عنوان سه چالش اصلی حوزه ی کلان داده ها مطرح میشوند و بقیه ی موارد در طول سالیانی که محققین بر روی این حوزه کار میکردند مورد توجه و همیت قرار گرفته اند.



- حجم داده (Volume): حجم داده ها به صورت نمایی در حال رشد می باشد. منابع مختلفی نظیر شبکه های اجتماعی، لاگ سرورهای وب، جریان های ترافیک، تصاویر ماهواره ای، جریان های صوتی، تراکنش های بانکی، محتوای صفحات وب، اسناد دولتی و ... وجود دارد که حجم داده بسیار زیادی تولید می کنند.
- نرخ تولید(Velocity): داده ها از طریق برنامه های کاربردی و سنسورهای بسیار زیادی که در محیط وجود دارند با سرعت بسیار زیاد و به صورت بلادرنگ تولید می شوند. بسیاری از کاربردها نیاز دارند به محض ورود داده به درخواست کاربر پاسخ دهند. ممکن است در برخی موارد نتوانیم به اندازه کافی صبر کنیم تا مثلا یک گزارش در سیستم برای مدت طولانی پردازش شود.
- تنوع(Variety): انواع منابع داده و تنوع در نوع داده بسیار زیاد می باشد که در نتیجه ساختارهای داده ای بسیار زیادی وجود دارد. مثلا در وب، افراد از نرم افزارها و مرورگرهای مختلفی برای ارسال اطلاعات استفاده می کنند. بسیاری از اطلاعات مستقیما از انسان دریافت میشود و بنابراین وجود خطا اجتناب ناپذیر است. این تنوع سبب میشود جامعیت داده تحت تاثیر قرار بگیرد. زیرا هرچه تنوع بیشتری وجود داشته باشد، احتمال بروز خطای بیشتری نیز وجود خواهد داشت.
- صحت (Veracity): با توجه به اینکه داده ها از منابع مختلف دریافت میشوند، ممکن است نتوان به همه آنها اعتماد کرد. مثلا در یک شبکه اجتماعی، ممکن است نظرهای زیادی در خصوص یک موضوع خاص ارائه شود. اما اینکه آیا همه آنها صحیح و قابل اطمینان هستند، موضوعی است که نمیتوان به سادگی از کنار آن در حجم بسیار زیادی از اطلاعات گذشت. البته بعضی از تحقیقات این چالش را به معنای حفظ همه مشخصه های داده اصلی بیان کرده اند که باید حفظ شود تا بتوان کیفیت و صحت داده را تضمین کرد. البته تعریف دوم درمولدهای کلان داده صدق میکند تا بتوان داده ای تولید کرد که نشان دهنده ویژگی های داده اصلی باشد.
- اعتبار(Validity): با فرض اینکه دیتا صحیح باشد، ممکن است برای برخی کاربردها مناسب نباشد یا به عبارت دیگر از اعتبار کافی برای استفاده در برخی از کاربردها برخوردار نباشد.
- نوسان(Volatility): سرعت تغییر ارزش داده های مختلف در طول زمان میتواند متفاوت باشد. در یک سیستم معمولی تجارت الکترونیک، سرعت نوسان داده ها زیاد نیست و ممکن است داده های موجود مثلا برای یک سال ارزش خود را حفظ کنند، اما در کاربردهایی نظیر تحلیل ارز و بورس، داده با نوسان زیادی مواجه هستند و داده ها به سرعت ارزش خود را از دست میدهند و مقادیر جدیدی به خود می گیرند. اگرچه نگهداری اطلاعات در زمان طولانی به منظور تحلیل تغییرات و نوسان داده ها حائز اهمیت است. افزایش دوره نگهداری اطلاعات، مسلما هزینه های پیاده سازی زیادی را دربر خواهد داشت که باید در نظر گرفته شود.
- نمایش (Visualization): یکی از کارهای مشکل در حوزه کلان داده، نمایش اطلاعات است. اینکه بخواهیم کاری کنیم که حجم عظیم اطلاعات با ارتباطات پیچیده، به خوبی قابل فهم و قابل مطالعه باشد از طریق روش های تحلیلی و بصری سازی مناسب اطلاعات امکان پذیری است.
- ارزش(Value): این موضوع دلالت بر این دارد که از نظر اطلاعاتی برای تصمیم گیری چقدر داده حائز ارزش است. بعبارت دیگر آیا هزینه ای که برای نگهداری داده و پردازش آنها میشود، ارزش آن را از نظر تصمیم گیری دارد یا نه. معمولا داده ها میتوانند در لایه های مختلف جابجا شوند. لایه های بالاتر به معنای ارزش بیشتر داده می باشند. بنابراین برخی از سازمانها میتوانند هزینه بالای نگهداری مربوط به لایه های بالاتر راقبول کنند.

مدیریت داده های حجیم شامل سه بخش مهم است: ذخیره سازی، پردازش و **تحلیل داده های حجیم**. مورد آخیر، موضوع اصلی این درس میباشد که کلیدی ترین مرحله پیاده سازی داده های عظیم در سازمان است. در این مرحله، ارزش واقعی بر اساس تحلیل پیشرفته و داده کاوی از داده های آماده شده خلق می شود.در این مرحله، خدمات مختلفی می تواند به سازمان ارائه شود:

- گزارشات آماری
- بصری سازی داده ها
- تحلیل های پیشرفته و یادگیری ماشین
- طراحی داشبوردهای اختصاصی

ارزش افزوده اصلی در این بخش، رساندن سازمان به توانایی انجام تحلیل های پیشرفته، ورای آنچه تا امروز شاهد آن بوده است، از داده های موجود می باشد. آنچه که در این بخش انجام می شود، شناساسی الگوریتم ها و ایجاد مدل های تحلیلی برای پاسخگویی به نیازهای سازمان خواهد بود. گوشه ای از خروجی های ناشی از تحلیل پیشرفته داده ها را می توان به شرح زیر بیان کرد:

- تشخیص و پیش بینی خروج مشتریان سازمان
- افزایش رضایت مشتریان با ارتقاء تجربه مشتری
- کشف تقلب در حوزه های مختلف
- مبارزه با پولشویی
- تحلیل وضعیت بازار
- طراحی محصولات و خدمات جدید و یا بهبود محصولات فعلی
- بهینه سازی عملیات شعب و کاهش هزینه ها
- پایش وضعیت شبکه ارتباطی و پیش بینی خرابی
- تشخیص رفتارهای غیر متعارف در شبکه
- تغییر رفتار مشتریان/پذیرندگان

نکته: با فرض آشنایی کسانی که این درس را گرفته اند با مباحث مرتبط با داده کاوی و همچنین در نظر گرفتن ارزش و اهمیت پـایتون در حـوزه ی علـوم داده ای، محوریت این درس بر روی استفاده از **کتابخانه های پایتون** برای سرفصل دروس خواهد بود.

رئوس مطالب:

۱. مقدمه ای بر پایتون
۲. آشنایی با ساختارهای داده ای پیشفرض پایتون و کار با فایلها
۳. مفاهیم پایه ای Numpy
۴. شروع کار با Pandas
۵. بارگذاری، ذخیره سازی و فرمت بندی فایلها
۶. پاک سازی و پیش پردازش داده
۷. تبدیل داده
۸. مصور سازی داده
۹. تجمیع داده و اپراتورهای گروهی
۱۰. سریهای زمانی
۱۱. مفاهیم پیشرفته در pandas
۱۲. مفاهیم پیشرفته در Numpy

نکته: کلاسها به شکل عملی برگزار خواهد شد فلذا بهتر است لپتاپ متصل به اینترنت همراه داشته باشید(با یک انشعاب برق!)

زمان برگزاری: جمعه ها ساعت ۱۷ تا ۲۰ (امکان تغییر و یا تکثیر کلاس با هماهنگی با دانشجویان وجود دارد)

نحوه ی ارزیابی: ۱۰ نمره پروژه (مهلت تحویل تا روز امتحان) + ۵ نمره (پایانترم) + ۵ نمره فعالیت کلاسی

منابع کمکی:

- Big Data For Dummies - WSU EECS (book)
- Data science and Big Data Analytics (book)
- Big data Analytics (book)
- All the matreials for Pandas and Numpy

راه ارتباطی با مدرس: از طریق ایمیل **abbas .akkasi@gmail.com**