گروه کامپیوتر – دانشکده مهندسی – دانشگاه فردوسی مشهد

مستندات گزارش بلند

برای بخشی از فعالیت های کلاسی

درس روش پژوهش و شیوه ارائه مطالب علمی و فنی

موضوع:

علوم داده در کسب و کار

Data science in Business

ارائه دهنده:

حمید رضا زه تاب

استاد درس:

دکتر رضا منصفی

زمستان 1400

چکیده

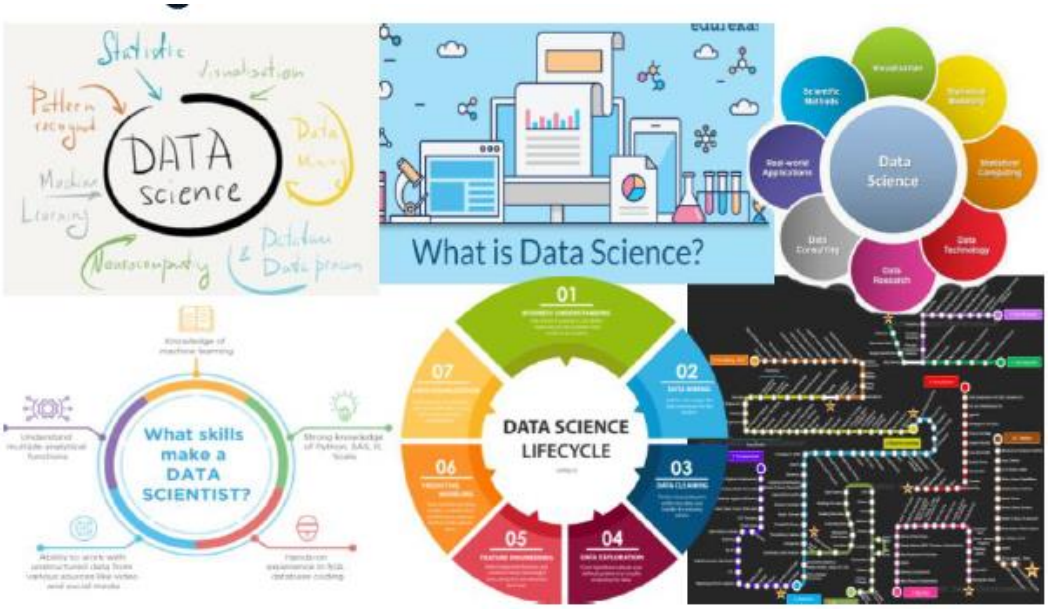
امروزه بیش از هرزمان دیگری داده داریم و با توجه به سریع تر شدن و قدرتمند شدن کامپیوتر ها و همچنین به لطف تکنولوژی ذخیره سازی اطلاعات می توانیم هزاران هزار ترابایت داده را مهار کنیم . به زبان ساده علم داده به ما کمک می کند تا این حجم گسترده از اطلاعات را برای خود قابل فهم کنیم . با کمک علم داده ، اطلاعات را طبقه بندی و دسته بندی میکنیم سپس آنها را مصور سازی می کنیم و سپس از داده ها نتیجه گیری می کنیم . نتیجه گیری نهایی می تواند در گستره بزرگی از شاخه ها من جمله کسب و کار به ما کمک کند . در این مقاله تلاش می کنیم تا ابتدا درک درستی از آنچه در علوم داده اتفاق می افتد ارایه کنیم و سپس به معرفی کاربرد های علوم داده در کسب و کار و نحوه به کار گیری آنها می پردازیم .

فهرست مطالب

۱– مقدمه

علم داده یک رشته پویا است . ابزار هایی که ما استفاده می کنیم و توانایی های تیم هایمان هر روز در حال تغییر است . در ادامه خواهید دانست که علم داده چیست و چگونه می توانید از آن برای تقویت سازمان خود استفاده کنید .

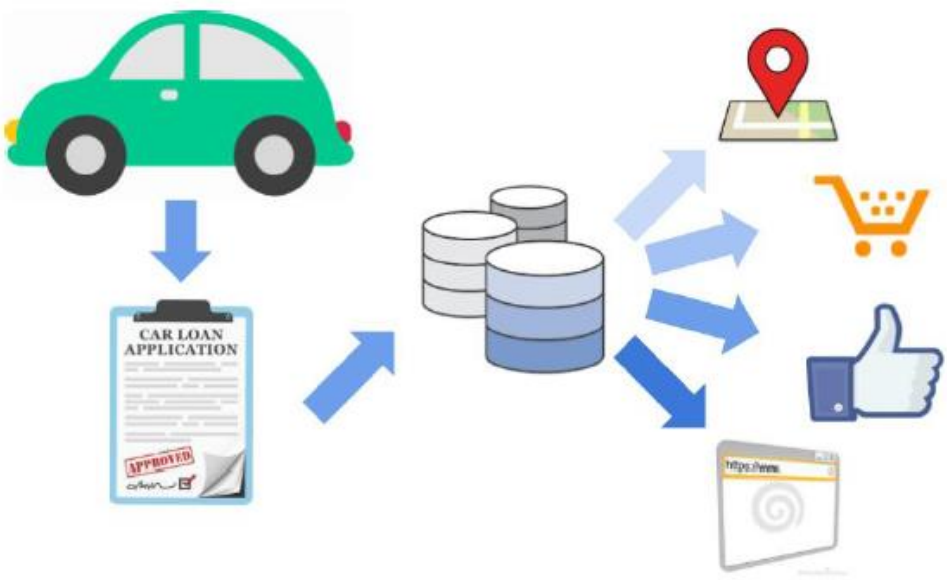
اگر عبارت "علوم داده چیست ؟" را گوگل کنیم ، با حجم انبوهی از اطلاعات سردرگم کننده روبرو می شویم ، اما علم داده در حقیقت بسیار ساده است . علم داده مجموعه ای از روش ها برای گرفتن هزاران شکل از داده هایی است که امروزه در دسترس ما هستند و از آنها برای نتیجه گیری معنادار استفاده می کنیم . داده ها در همه جا هستند . هر لایک ، کلیک ، ایمیل ، تراکنش بانکی یا توییت یک داده جدید است که می تواند برای توصیف بهتر زمان حال یا پیش بینی بهتر آینده استفاده شود .



سوال اینجاست که داده ها چه کاری می توانند برای ما انجام دهند ؟ داده ها می توانند وضعیت فعلی ما را توصیف کنند . این را می توان با داشبوردها یا هشدارها انجام داد و فرآیندهای زمانبر گزارشدهی را با فناوری های جدید ساده کرد. علم داده می تواند به تشخیص رویدادهای غیر عادی کمک کند. اگر داده‌هایی در مورد آنچه قبلاً اتفاق افتاده است داشته باشیم، می‌توانیم با شناسایی خودکار یک رویداد جدید که غیرمنتظره است، کارایی را افزایش دهیم. داده ها همچنین می توانند علل رویدادها و رفتارهای مشاهده شده را تشخیص دهند. به جای تعیین همبستگی بین تعداد کمی از رویدادها، تکنیک‌های علم داده به ما کمک می‌کنند تا سیستم‌های پیچیده را با دلایل احتمالی متعدد درک کنیم. در نهایت، داده ها می توانند رویدادهای آینده را پیش بینی

کنند. ما می توانیم از تکنیک های جدید برای در نظر گرفتن علل مختلف و پیش بینی نتایج بالقوه استفاده کنیم. علاوه بر این، ما می توانیم احتمال پیش بینی خود را به صورت ریاضی ارزیابی کنیم تا سطح عدم قطعیت خود را روشن کنیم.

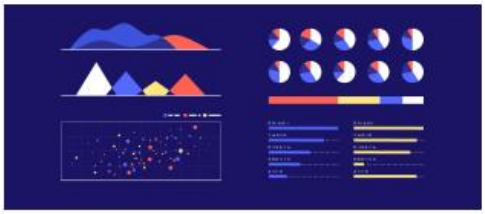
بنابراین اکنون می دانیم علم داده چیست. سوال بعدی این است که چرا اینقدر محبوب است؟ پاسخ بسیار ساده است: ما در حال جمع آوری داده های بیشتری نسبت به قبل هستیم. فرض کنید به یک نمایندگی خودرو مراجعه می کنید و اطلاعاتی را تکمیل می کنید. همه این داده ها به طور خودکار در رایانه وارد می شوند و با داده های صدها نمایندگی در یک پایگاه داده بزرگ ترکیب می شوند. هنگامی که این داده ها را در اختیار داریم، استفاده از آدرس ایمیلی که هنگام خرید آن خودرو ارائه کرده اید برای پیوند دادن داده های خرید خودرو با داده های شما از رسانه های اجتماعی یا مرور وب آسان است. ناگهان، ما یک تصویر بسیار کامل در مورد همه کسانی که در سال گذشته ماشین خریداری کرده اند داریم: سن آنها، علایق آنها، دوستان و خانواده آنها. این داده‌های اضافی را می‌توان برای پیش‌بینی اینکه چه قیمتی می‌توانید برای ماشین خود بپردازید، چه خریدهای دیگری انجام دهید، یا بهترین نحوه فروش بیمه آن ماشین جدید به شما ، استفاده می‌شود. داده ها در همه جا وجود دارد و برای کسب و کارها بسیار ارزشمند است.



بنابراین، چگونه شروع به استفاده از داده های خود کنیم؟ در علم داده، ما به طور کلی سه مرحله برای هر پروژه داریم. اول، ما داده‌ها را از منابع بسیاری، مانند نظرسنجی‌های مشتریان، نتایج ترافیک وب، ایمیل‌های بین نمایندگان فروش و مشتریان بالقوه، و تراکنش‌های مالی جمع‌آوری می‌کنیم. در مرحله بعد، داده های خود را بررسی و مصور (visualize)می کنیم. این می تواند شامل ساخت داشبوردهایی برای ردیابی اینکه چگونه داده های ما در طول زمان تغییر می کنند یا انجام مقایسه بین دو مجموعه داده باشد. در نهایت، ما با داده های خود پیش بینی می کنیم. به عنوان مثال، این می تواند شامل ساختن سیستمی باشد که مشتریان را تقسیم بندی می کند یا تصاویر انواع مختلف خودروها را طبقه بندی می کند.

Data collection

Exploration and visualization



Experimentation and prediction

2– کاربرد های علوم داده

قبلاً با تعریف علم داده و مراحل یک گردش کار علم داده آشنا شدید. در ادامه ، شما یاد خواهید گرفت که چگونه علم داده را برای مشکلات واقعی کسب و کار به کار ببرید.

بیایید به سه حوزه مهیج علم داده نگاهی بیندازیم: یادگیری ماشین ، اینترنت اشیا و یادگیری عمیق.

2–1– یادگیری ماشین

فرض کنید در یک بانک بزرگ در اداره کشف کلاهبرداری و تقلب (fraud detection)کار می کنید. می‌خواهید از داده‌ها برای تعیین احتمال جعلی بودن تراکنش استفاده کنید.

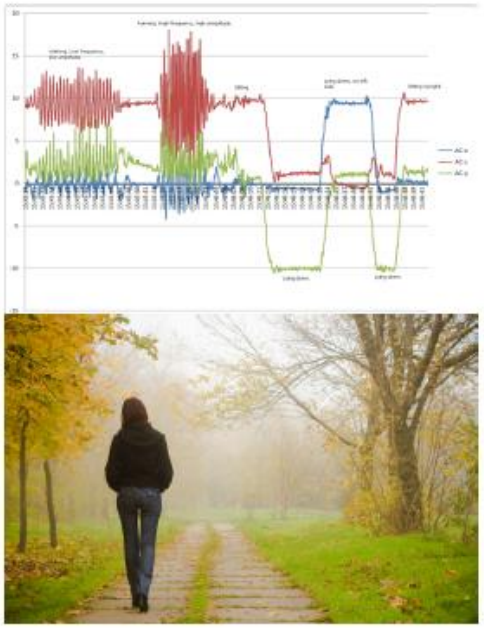
برای پاسخ به این سوال، ممکن است با جمع آوری اطلاعات در مورد هر خرید، مانند مبلغ، تاریخ، مکان، نوع خرید و آدرس دارنده کارت شروع کنید. شما به نمونه های زیادی از تراکنش ها، از جمله این اطلاعات، و همچنین برچسبی نیاز دارید که معتبر یا تقلبی بودن هر تراکنش را به شما بگوید . خوشبختانه، شما احتمالاً این اطلاعات را در یک پایگاه داده دارید. این رکوردها "داده های آموزشی(training data)" نامیده می شوند و برای ساخت یک الگوریتم استفاده می شوند. هر بار که یک تراکنش جدید رخ می دهد، اطلاعات الگوریتم خود مانند مبلغ و تاریخ را می دهید و به سوال اصلی پاسخ می دهد: احتمال تقلبی بودن(fraudulent) این تراکنش چقدر است؟

Amount	Date	Type	...
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

قبل از اینکه بتوانیم به این سوال پاسخ دهیم، بیایید مثال خود را مرور کنیم و آنچه را که برای یادگیری ماشینی نیاز داریم تا جادوی خود را انجام دهد برجسته کنیم. اول، یک مسئله علم داده با یک سوال کاملاً تعریف شده شروع می شود. سوال ما این بود که "احتمال تقلبی بودن این تراکنش چقدر است؟" در مرحله بعد، ما به برخی داده ها برای تجزیه و تحلیل نیاز داریم. ماه‌ها تراکنش‌های قدیمی کارت اعتباری و ابرداده‌های مرتبط داشتیم که قبلاً تقلبی یا معتبر شناخته شده بودند. در نهایت، هر بار که می‌خواهیم پیش‌بینی و قضاوت جدیدی انجام دهیم، به داده‌های اضافی نیاز داریم. ما باید در مورد هر خرید جدید اطلاعات یکسانی داشته باشیم تا بتوانیم آن را به عنوان "تقلبی" یا "معتبر" بدانیم.

2–2– اینترنت اشیاء

حال، فرض کنید قصد دارید یک ساعت هوشمند برای نظارت بر فعالیت بدنی بسازید. شما می خواهید بتوانید فعالیت های مختلف مانند راه رفتن یا دویدن را به طور خودکار تشخیص دهید. ساعت هوشمند شما مجهز به حسگر ویژه ای به نام «شتاب سنج» است که حرکت را به صورت سه بعدی کنترل می کند. داده های تولید شده توسط این سنسور اساس یادگیری ماشین شما است. می توانید از چندین داوطلب بخواهید که ساعت شما را بپوشند و فعالیت خود را هنگام دویدن یا راه رفتن ضبط کنند. سپس می توانید الگوریتمی ایجاد کنید که داده های شتاب سنج را به عنوان نماینده یکی از این دو حالت تشخیص می دهد: راه رفتن یا دویدن.

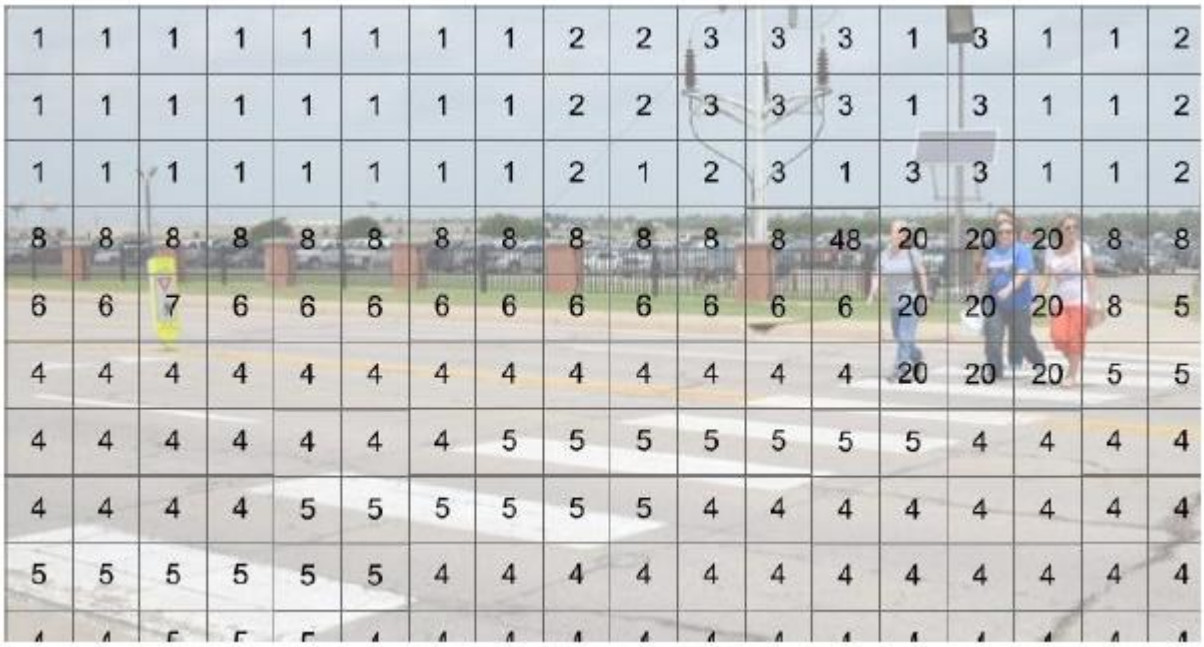


ساعت هوشمند شما بخشی از یک زمینه در حال رشد به نام "اینترنت اشیا" است که به عنوان IOT نیز شناخته می شود، که اغلب با علم داده ترکیب می شود. اینترنت اشیا به ابزارهایی اطلاق می‌شود که کامپیوتر های استاندارد نیستند، اما همچنان توانایی انتقال داده‌ها را دارند. این شامل ساعت‌های هوشمند، سیستم‌های امنیتی خانگی متصل به اینترنت، سیستم‌های جمع‌آوری عوارض الکترونیکی، سیستم‌های مدیریت انرژی ساختمان و بسیاری موارد دیگر می‌شود . داده های تولید شده توسط اینترنت اشیا منبع عالی برای پروژه های علم داده است!

2–3– یادگیری عمیق

بیایید به مثال دیگری بپردازیم. یکی از وظایف کلیدی برای خودروهای خودران، شناسایی زمانی است که تصویری حاوی یک انسان است. مجموعه داده برای این مشکل چه خواهد بود؟

ما می توانیم تصویر را به صورت ماتریسی از اعداد بیان کنیم که در آن هر عدد نشان دهنده یک پیکسل است. با این حال، اگر ماتریس را به یک مدل یادگیری ماشین وارد کنیم، احتمالاً این رویکرد شکست خواهد خورد. به سادگی داده های ورودی بیش از حد وجود دارد!



ما به الگوریتم پیشرفته تری نیاز داریم که به عنوان یادگیری عمیق شناخته می شود. در یادگیری عمیق ، چندین لایه از مینی الگوریتم‌ها به نام « نورون‌ها » با هم کار می‌کنند تا نتیجه‌گیری‌های پیچیده‌ای را به دست آورند. یادگیری عمیق نسبت به یک مدل یادگیری ماشینی به داده های آموزشی بسیار بسیار بیشتری نیاز دارد، اما همچنین می تواند روابطی را بیاموزد که مدل های سنتی نمی توانند. یادگیری عمیق برای حل مشکلات پیچیده ای مانند طبقه بندی تصویر یا درک زبان استفاده می شود.

3– ساخت تیم علوم داده

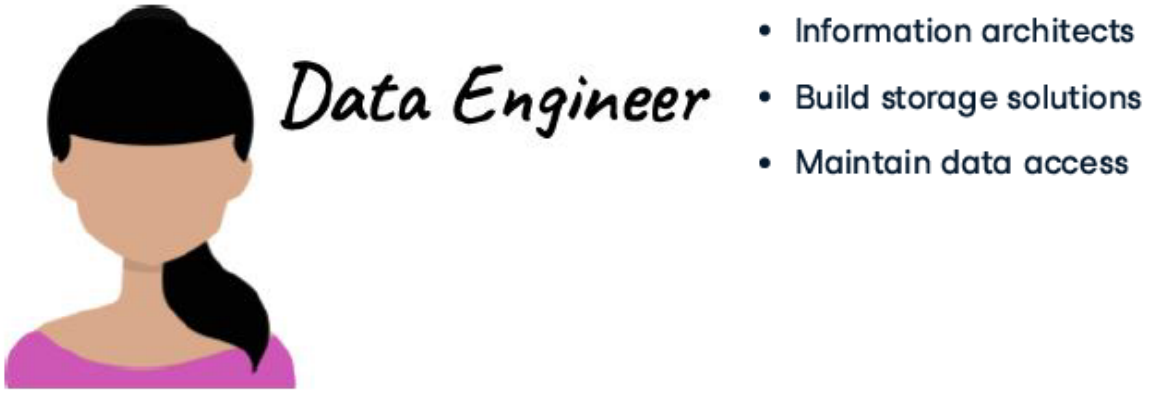
در ادامه شما یاد خواهید گرفت که چگونه تیم داده خود را برای برآوردن نیازهای سازمان خود بسازید و ساختار دهید.

3–1– اعضاء تیم

شاید تعجب کنید که بدانید "علم داده" یک رشته واحد نیست. این در واقع سه شغل مختلف است: مهندس داده، تحلیلگر داده، و دانشمند یادگیری ماشین. بیایید هر یک را بررسی کنیم.

3–1–1– مهندس داده

مهندسان داده جریان اطلاعات را کنترل می کنند: آنها سیستم های ذخیره سازی داده ها و زیرساخت های تخصصی ایجاد می کنند تا اطمینان حاصل کنند که داده ها به راحتی به دست می آیند و پردازش می شوند.



اکثر مهندسان داده با SQL آشنا هستند که از آن برای ذخیره و مدیریت کلان داده ها استفاده می کنند. آنها همچنین از یکی از زبان های برنامه نویسی مانند جاوا، اسکالا یا پایتون برای پردازش داده ها و خودکارسازی وظایف مربوطه استفاده می کنند.

3–1–2– تحلیلگر داده

تحلیلگران داده ، حال را از طریق داده ها توصیف می کنند. آنها این کار را با داشبورد، آزمون فرضیه ها(hypothesis tests) و تجسم(visualization) انجام می دهند. آنها اغلب پیشینه ای در زمینه آمار یا علوم کامپیوتر دارند، اما تمایل دارند تجربه مهندسی کمتری نسبت به مهندسان داده و تجربه ریاضی کمتری نسبت به دانشمندان داده داشته باشند.

تحلیلگران داده از spreadsheets برای انجام تجزیه و تحلیل های ساده بر روی مقادیر کم داده استفاده می کنند. آنها ازSQL ، همان زبانی که توسط مهندسان داده استفاده می شود، برای تحلیل های بزرگتر استفاده می کنند. در حالی که مهندسان داده راه حل های ذخیره سازی SQL را می سازند و پیکربندی می کنند، تحلیلگران داده از پایگاه های داده موجود برای مصرف و خلاصه کردن داده ها استفاده می کنند. تحلیلگران همچنین از هوش تجاری یا BI ، ابزارهایی مانند Tableau ، Power BI یا Looker برای ایجاد داشبورد و به اشتراک گذاری تحلیل های خود استفاده می کنند.

- Spreadsheets (Excel or Google Sheets)
  - Simple storage and analysis
- SQL
  - Large-scale analysis
- BI Tools (Tableau, Power BI, Looker)
  - Dashboarding and sharing information

3–1–3– دانشمند یادگیری ماشین

یادگیری ماشین شاید پرهیاهوترین بخش علم داده باشد. از آن برای تشخیص و حدس آنچه که احتمالاً درست است از آنچه قبلاً می دانیم استفاده می شود. این دانشمندان از داده های آموزشی برای طبقه بندی داده های بزرگتر و غیر طبقه بندی شده استفاده می کنند. یادگیری ماشینی می‌تواند به ما بگوید ارزش یک سهام در هفته آینده چقدر است، کدام تصاویر حاوی یک ماشین هستند یا چه احساساتی توسط مجموعه‌ای از توییت‌ها بیان می‌شوند.

- Predictions and extrapolations
- Classification
- Stock price prediction
- Image processing
- Automated text analysis

دانشمندان یادگیری ماشین از پایتون یا R برای ایجاد مدل های پیش بینی خود استفاده می کنند. هر دو زبان برنامه نویسی عالی برای علم داده هستند و فردی که یک زبان را می داند احتمالاً می تواند کد را در زبان دیگر بخواند. به یاد داشته باشید، یادگیری زبان های برنامه نویسی به اندازه زبان های گفتاری دشوار نیست. اگر کسی بلد باشد فرانسوی صحبت کند، ممکن است سالها طول بکشد تا زبان آلمانی را یاد بگیرد. زبان های برنامه نویسی بیشتر شبیه ابزارها هستند. اگر می دانید چگونه از دریل برقی استفاده کنید، لزوماً نحوه استفاده از اره برقی را بلد نیستید، اما احتمالاً می توانید با کمی آموزش یاد بگیرید!

- Python and R
  - Programming languages for creating predictive models

3–1–4– مرور

برای جمع‌بندی: مهندسان داده داده‌ها را ذخیره و نگهداری می‌کنند، تحلیلگران داده‌ها داده‌ها را تجسم(visualize) و توصیف می‌کنند، و دانشمندان یادگیری ماشین با داده‌ها مدل‌سازی و پیش‌بینی می‌کنند. هر موقعیت از مجموعه کمی متفاوت از ابزارها برای رسیدن به اهداف خود استفاده می کند.

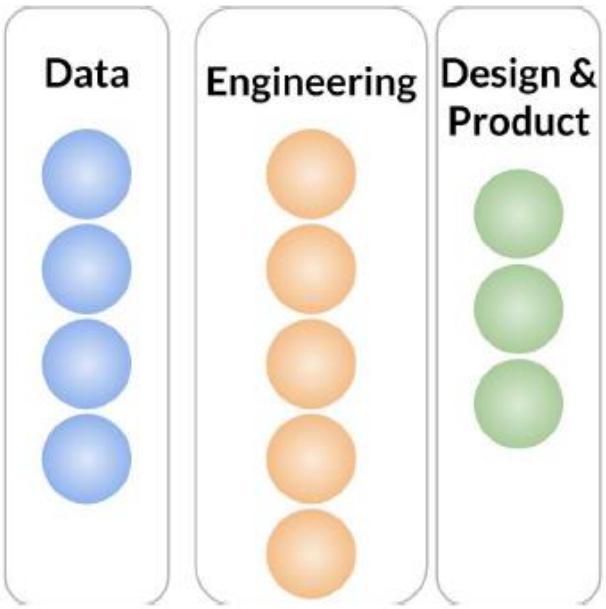
Data Engineer	Data Analyst	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Model and predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R

3–2– ساختار تیم علم داده

هنگامی که تعدادی متخصص داده را استخدام کردید، سه راه اصلی وجود دارد که می توانید ساختار تیم داده خود را سامان دهید: ایزوله شده ، تعبیه شده یا ترکیبی.

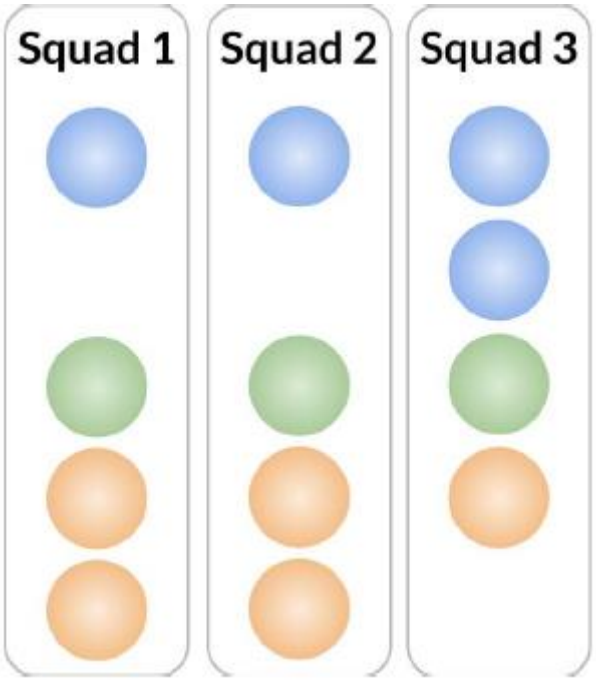
3–2–1– ایزوله شده (isolated)

یک تیم داده ایزوله می تواند شامل یک یا چند نوع کارمند داده بدون هیچ تیم دیگری مانند مهندسی یا محصول باشد. این یک ساختار عالی برای آموزش اعضای جدید تیم و تغییر سریع پروژه ای است که هر عضو روی آن کار می کند.



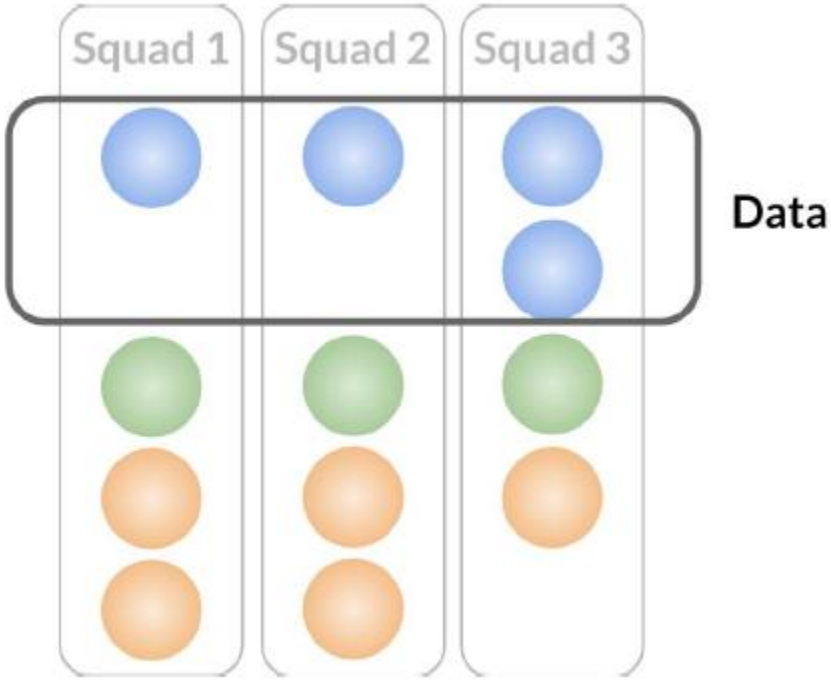
3–2–2–**تعبیه شده (** embedded **)**

روش دیگر، استفاده از یک مدل تعبیه شده است که در آن هر کارمند داده بخشی از یک تیم است، که شامل مهندسان و مدیران محصول نیز می شود، این روش می تواند مفید باشد. این مدل به هر کارمند داده‌ای اجازه می‌دهد تا در یک پروژه تجاری خاص تجربه کسب کند و تبدیل به یک متخصص ارزشمند شود.



3–2–3–**ترکیبی (** hybrid **)**

مدل هیبریدی شبیه به مدل تعبیه شده به نظر می رسد، اما با یک همگام سازی اضافی برای همه کارکنان داده در تمام گروه ها. این لایه اضافی سازمان، بدون در نظر گرفتن اینکه کارمند به کدام پروژه منصوب شده است، امکان فرآیندهای داده یکنواخت و توسعه شغلی را فراهم می کند.



4–منابع داده و خطرات

قبلاً با نحوه کار در علم داده آشنا شدید. در این درس، ما بر مرحله اول تمرکز خواهیم کرد: جمع آوری داده ها.

4–1– منابع مشترک داده ها

داده ها در همه جا وجود دارد و تقریباً هر فرآیند تجاری می تواند کوه هایی از داده را تولید کند. برخی از رایج ترین منابع داده، رویدادهای وب، داده های مشتری، داده های لجستیک و تراکنش های مالی هستند. این امکان وجود دارد که شرکت شما در حال حاضر تمام این اطلاعات را جمع آوری کرده باشد. بهتر است از مهندسان داده خود بپرسید که چه چیزی جمع آوری می شود و چه چیزی جمع آوری نمی شود و بر اهمیت شروع فرآیند جمع آوری داده های متمرکز زودتر از دیرتر تاکید کنید.

4–2– داده های وب

بیایید نگاهی عمیق تر به داده های وب بکنیم. هنگامی که کاربر از یک صفحه وب بازدید می کند یا روی یک پیوند کلیک می کند، ردیابی این اطلاعات به منظور محاسبه نرخ تبدیل (conversion rates) یا نظارت بر محبوبیت قسمت های مختلف محتوا می تواند مفید باشد. حداقل، می‌خواهید نام رویداد را جمع‌آوری کنید، که می‌تواند به معنای نشانی اینترنتی صفحه بازدید شده یا شناسه عنصری که روی آن کلیک شده، مهر زمانی رویداد(timestamp)، و یک شناسه برای کاربری باشد که این کار را انجام داده است.

Web data

- Events
- Timestamps
- User information

user_id	event_name	timestamp
1234	homepage_visit	2019-01-01 12:01:01

4–3– اطلاعات قابل شناسایی شخصی(PII)

فرض کنید علی محمدی مشتری است که از وب سایت شرکت شما بازدید می کند و یکی از محصولات شما را می پسندد. ممکن است انتخاب کنید نام او، مهر زمانی و شیئی که روی آن کلیک کرده است را ردیابی کنید. مهم است که به یاد داشته باشید که نام علی محمدی، اطلاعات قابل شناسایی شخصی(Personally Identifiable Information) یا PII است. PII شامل نام، موقعیت مکانی، آدرس ایمیل، و هر اطلاعات دیگری است که می تواند برای پیوند دادن یک رویداد وب به یک انسان واقعی استفاده شود . بایستی با PII با احتیاط زیادی برخورد کرد .

Timestamp	Object Clicked
2019-01-20 12:05:00	Like Button

4–4– نام مستعار داده ها (Data pseudonymization)

یکی از ساده‌ترین راه‌ها برای محافظت از هویت علی محمدی، تقسیم این اطلاعات به دو ورودی جداگانه است. ما می توانیم به علی یک شناسه کاربری، در این مورد 185477 اختصاص دهیم و آن اطلاعات را در جدول کاربران ذخیره کنیم. سپس می توانیم رویداد او را با استفاده از این شناسه شناسایی کنیم. ما داده‌های جدول رویدادها را با نام مستعار می‌خوانیم زیرا علی را نمی‌توان به تنهایی با آن جدول شناسایی کرد، اما اگر اطلاعات جدول کاربران را با جدول رویدادها ترکیب کنیم، می‌توان او را شناسایی کرد. برای محافظت از علی ، می‌خواهیم مطمئن شویم که دسترسی به جدول کاربران فقط به افرادی محدود می‌شود که

باید هویت علی را بدانند، مانند نمایندگان ارشد تیم خدمات مشتری یا اعضای تیم حقوقی. همچنین می‌خواهیم به‌طور دوره‌ای بازرسی کنیم که چه کسی به این داده‌ها دسترسی داشته است و چگونه از آن استفاده کرده است تا اطمینان حاصل شود که داده‌های علی امن هستند و به حریم شخصی او احترام گذاشته می شود .

user_id	Timestamp	Object Clicked
185477	2019-01-20 12:05:00	Like Button

4–5– ناشناس سازی داده ها(Data anonymization)

بهترین راه برای محافظت از حریم خصوصی علی این است که پس از اختصاص شناسه کاربری علی، اطلاعات جدول کاربران را از بین ببرید. بدون جدول کاربران، جدول رویدادها داده های کاملاً ناشناس است. برای بسیاری از اهداف تحلیل، داده های ناشناس کافی است. ما باید بدانیم که جین یک فرد منحصر به فرد است، اما نیازی نیست که نام او یا هر PII دیگری را بدانیم.

4–6– مقررات عمومی حفاظت از داده ها

ممکن است اخیراً اصطلاح "GDPR" را از تیم داده خود شنیده باشید. GDPR مخفف عبارت General Data Protection Regulation است و برای تمام داده های داخل اتحادیه اروپا اعمال می شود. هدف از GDPR این است که افراد بر داده های شخصی خود کنترل داشته باشند. از جمله موارد دیگر، GDPR مدت زمان ذخیره داده ها را تنظیم می کند، ناشناس سازی مناسب را الزامی می کند و نیاز به افشای جمع آوری داده ها و کسب رضایت دارد.

General Data Protection Regulation (GDPR)

- Applies to all data inside of the EU
- Give individuals control over their personal data
- Regulates how long data can be stored
- Mandates appropriate anonymization
- Disclose data collection and gain consent

5– داده های درخواستی (Solicited data)

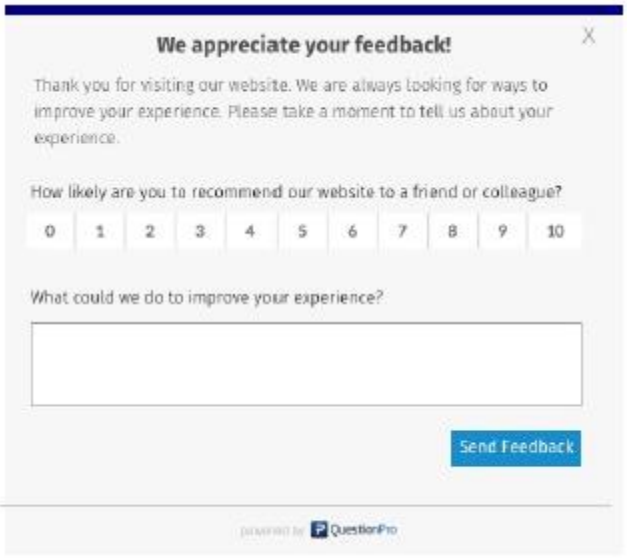
در چند صفحه گذشته، در مورد داده هایی که شرکت شما می تواند از طریق وب یا تراکنش های مالی جمع آوری کند، یاد گرفتیم. اکنون، ما داده‌هایی را پوشش می‌دهیم که می‌توانید با پرسیدن نظرات مشتریان خود به دست آورید، که به آن داده‌های درخواستی می‌گوییم.

5–1– چرا داده ها را درخواست می کنیم ؟

از داده های درخواستی می توان برای ایجاد وثیقه بازاریابی استفاده کرد، مانند پستی در مورد اینکه چند درصد از مشتریان ما از محصول ما راضی هستند. همچنین می‌توان از آن برای کاهش ریسک فرآیند تصمیم‌گیری استفاده کرد، مانند زمانی که از کاربران نظرسنجی می‌کنیم تا علاقه به محصول جدید را ارزیابی کنیم. در نهایت، می توان از آن برای نظارت بر کیفیت استفاده کرد.

5–2– انواع داده ای درخواستی؟

انواع متداول داده های درخواستی شامل نظرسنجی ها، بررسی های مشتریان، پرسشنامه های درون برنامه ای و گروه های تمرکز است. این تصویر نوع بسیار رایج بازخورد درخواستی را نشان می دهد: امتیاز خالص تبلیغ کننده یا NPS ، که از کاربر می پرسد چقدر احتمال دارد که کاربر یک محصول را به یک دوست یا همکار توصیه کند.



داده های درخواستی می تواند کیفی باشد، مانند مکالمات و سؤالات باز، یا کمی باشد، مانند سؤالات چند گزینه ای یا مقیاس های رتبه بندی. داده های کیفی بسیار ذهنی هستند و نیاز به تحلیل زیادی دارند. داده های کمی را می توان به راحتی در یک نمودار یا نمودار خلاصه کرد. به طور کلی، جمع آوری داده های کیفی در مقیاس کوچک برای ایجاد فرضیه خوب است. به عنوان مثال، یک گروه تمرکز ممکن است ایده هایی در مورد ویژگی هایی که ممکن است بخواهیم بسازیم ارائه دهد. برای تایید این فرضیه ها به مجموعه کمی در مقیاس بزرگتر نیاز است. به عنوان مثال، می‌توانیم از کاربران بخواهیم فهرستی از ویژگی‌ها را از مطلوب‌ترین به حداقل مطلوب رتبه‌بندی کنند.

5–3– ترجیحات آشکار و بیان شده(Revealed and stated preferences)

مهم است که به یاد داشته باشید که داده های درخواست شده به طور کلی اولویت اعلام شده کاربران را به ما می گوید. ترجیح اعلام شده چیزی است که کسی به ما می گوید که می خواهد یا معتقد است و تا حدودی فرضی است. هنگامی که کاربر واقعاً اقدامی مانند خرید یک محصول انجام می دهد، ترجیحات آشکار او را می آموزیم. ما امیدواریم که اولویت‌های اعلام‌شده کاربران ما، شاخص‌های خوبی برای اولویت آشکار آنها باشد، اما همیشه اینطور نیست. بسیاری از مردم ترجیح می دهند مرتباً به باشگاه بروند. با این حال، ترجیح بسیاری از همان افراد نشان داده شده تنها رفتن گهگاهی است. برخی از باشگاه‌ها یک مدل کسب‌وکار کامل دارند که بر اساس تفاوت مورد انتظار بین اولویت‌های اعلام شده و آشکار افراد برای ورزش است.

اکنون که انواع داده‌های درخواستی را می‌شناسیم و از مشکلات اولویت‌های آشکار شده در مقابل اولویت‌های اعلام‌شده آگاه هستیم، اجازه دهید برخی از بهترین شیوه‌ها را مرور کنیم. اول، ما باید سعی کنیم در هنگام پرسیدن سوال تا حد امکان دقیق باشیم. این ویژگی باید هم در مورد جمله بندی سؤال و هم در مورد پاسخ های بالقوه ای که ارائه می دهیم اعمال شود.

5–4– بهترین شیوه ها

در مرحله بعد، ما باید از زبان بارگذاری شده (loaded language) اجتناب کنیم، به خصوص اگر ممکن است پاسخ دهندگان را نسبت به یک انتخاب خاص سوگیری کند. به عنوان مثال، از استفاده از صفت برای توصیف گزینه های احتمالی خودداری کنید و سعی کنید عینی (objective) باشید.

در صورت امکان، نظرسنجی خود را با مقایسه با مقادیر شناخته شده کالیبره کنید. به عنوان مثال، به جای اینکه از پاسخ دهنده بپرسید که آیا به یک محصول جدید علاقه مند است، از او بخواهید که علاقه خود را به آن محصول با علاقه خود به یکی از پیشنهادات فعلی شما یا پیشنهادات یک رقیب شناخته شده مقایسه کند.

در نهایت، پرسیدن هر چه بیشتر سؤالات ممکن است وسوسه انگیز باشد. با این غریزه مبارزه کنید و در عوض اطمینان حاصل کنید که هر سؤالی که می‌پرسید به شما کمک می‌کند تا یک اقدام قاطع انجام دهید. به عنوان مثال، اگر مشتریان ترجیحی برای ویژگی A نسبت به ویژگی B نشان دادند، شما باید ابتدا متعهد شوید که ویژگی A را بسازید.

6– جمع آوری داده های اضافه (additional data)

در حالی که جمع‌آوری داده‌های داخلی برای برخی از پروژه‌های علم داده مفید است، اما تنها یک تکه از این پازل است. اغلب، شما نیاز به جمع آوری داده ها از منابع خارجی نیز دارید.

6–1– حتی داده های بیشتر

راه های زیادی وجود دارد که می توانید داده های اضافی را برای سازمان خود جمع آوری کنید. چند راه متداول شامل APIها، رکوردهای عمومی و Mechanical Turk است که در ادامه به همه آنها خواهیم پرداخت.

6–1–1– API های داده

بیایید با API ها شروع کنیم .API مخفف Application Programming Interface است. این یک راه آسان برای درخواست داده از شخص ثالث از طریق اینترنت است. بسیاری از شرکت‌ها API دارند تا به تیم شما اجازه دسترسی به داده‌هایشان را بدهند. برخی ازAPI های قابل توجه عبارتند ازTwitter ، Wikipedia، Yahoo! ، و نقشه های گوگل، اما بسیاری از موارد دیگر وجود دارد. اگر با یک شریک کار می کنید و فکر می کنید که آنها ممکن است داده های مفیدی داشته باشند، یک جستجوی سریع در وب انجام دهید و ببینید آیا یک API وجود دارد یا خیر!

بیایید به نمونه ای از API توییتر نگاه کنیم. فرض کنید می‌خواهیم توییت‌ها را با #Example دنبال کنیم. ما می‌توانیم از API توییتر برای درخواست همه توییت‌های دارای این هشتگ استفاده کنیم. در این مرحله، ما گزینه های زیادی برای تجزیه و تحلیل داریم. ما می‌توانیم روی متن هر توییت یک تحلیل احساسی انجام دهیم و درباره نظرات و احساسات مردم ایده بگیریم. ما به سادگی می‌توانیم ردیابی کنیم که هر هفته چند بار هشتگ Example ظاهر می‌شود.

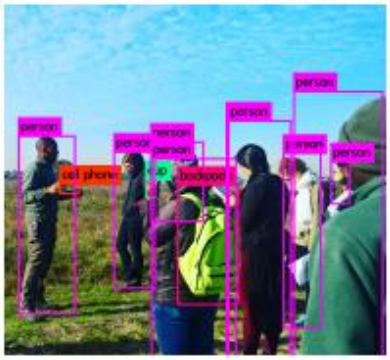
6–1–2– سوابق عمومی

سوابق عمومی یکی دیگر از راه های عالی برای جمع آوری داده های اضافی است. در ایالات متحده، data.gov داده های بهداشت، آموزش و تجارت را برای دانلود رایگان در دسترس دارد. در اتحادیه اروپا، data.europa.eu داده های مشابهی دارد. اینها می توانند منابع خوبی برای درک روندهای سطح جمعیت(population-level trends) یا جمع آوری داده های اقتصادی باشند.

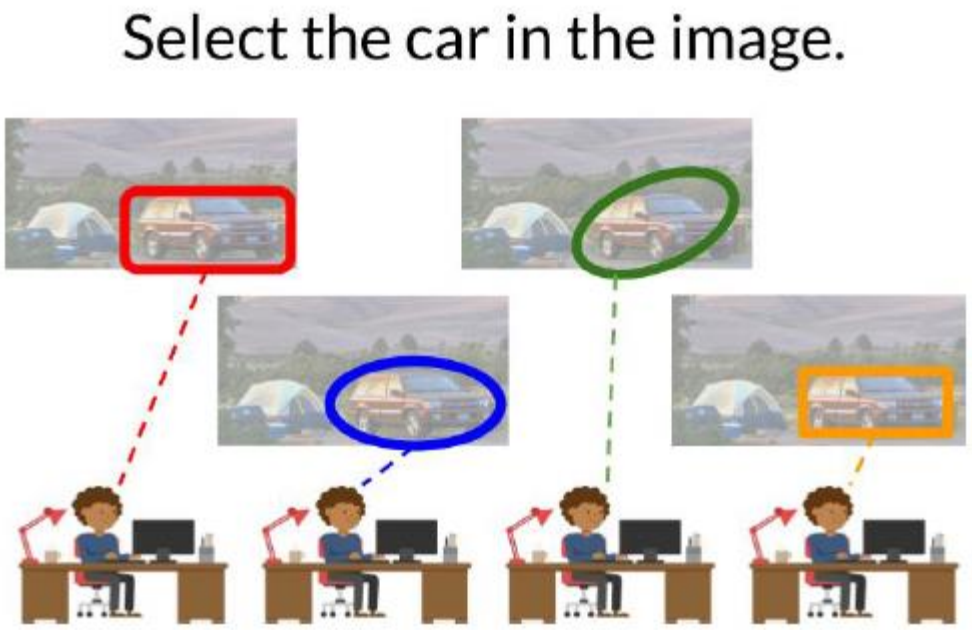
.

6–1–3– ترک مکانیکی

قبلا ، ما در مورد تشخیص تصویر به عنوان یک نوع مشکل علم داده بحث کردیم. برای ساختن یک الگوریتم تشخیص تصویر خوب، به مجموعه ای از تصاویر نیاز داریم که در آن تصاویر قبلاً برچسب گذاری شده باشند که به آن مجموعه آموزشی (training set) می گویند. اما ما فقط به یک یا دو عکس نیاز نداریم. ما به صدها یا هزاران عکس نیاز داریم. دریافت این تصاویر برچسب‌گذاری شده می‌تواند واقعاً سخت و زمان‌بر باشد و فقدان مجموعه آموزشی اغلب باعث می‌شود پروژه‌های علمی داده خوب تکمیل نشوند.



بسته به نوع مجموعه آموزشی مورد نیاز، Mechanical Turk که MTurk نیز نامیده می شود، می تواند یک گزینه عالی باشد. ترک مکانیکی به معنای درخواست از انسان برای تکمیل کاری است که ما در نهایت برای کامپیوتری کردن آن عمل برنامه ریزی می کنیم. در مثال قبلی ما، این به معنای برچسب زدن تعدادی عکس برای ایجاد یک مجموعه آموزشی برای تشخیص تصویر است. به جای اینکه از یک نفر بخواهیم هزاران تصویر را برچسب گذاری کند، هزاران نفر را استخدام می کنیم و به هر یک از آنها برای برچسب زدن چند تصویر پول می دهیم. برای اطمینان از کیفیت، ممکن است از دو یا سه نفر بخواهیم تصویر مشابه را مرور کنند و سپس رایج ترین پاسخ را در نظر می گیریم .



پلتفرم‌های زیادی برای کمک به Mechanical Turkو استخدام کمک‌کنندگان مانند AWS MTurk وجود دارد. ترک مکانیکی فقط برای تشخیص تصویر نیست. همچنین می توانید از آن برای برچسب زدن نظرات مشتریان به عنوان مثبت یا منفی، استخراج متن از فرم یا برجسته کردن کلمات کلیدی در یک جمله استفاده کنید. در مثال سمت راست، از کاربران خواسته می شود که مشخص کنند کدام بخش از تصویر دارای علامت خیابان است.

7– ذخیره و بازیابی داده ها

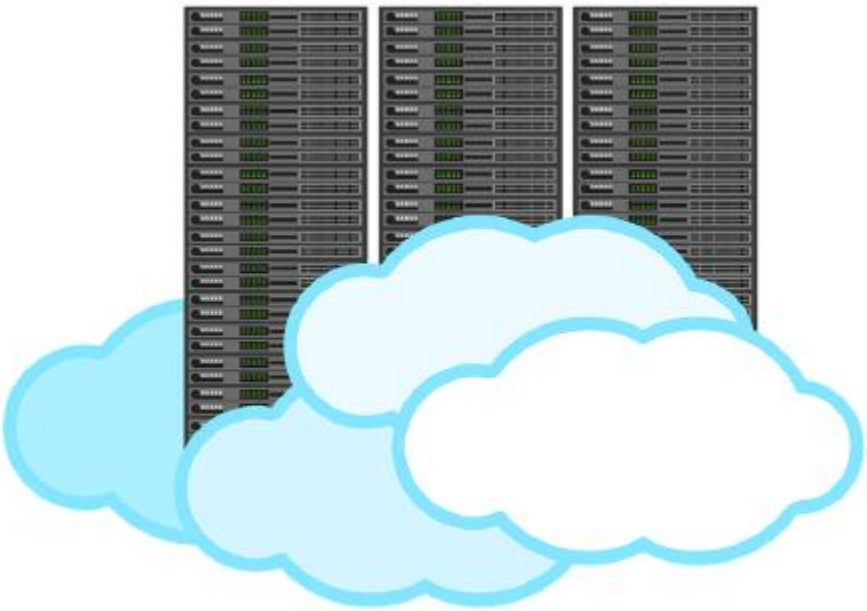
تا اینجا، با بسیاری از منابع داده آشنا شدید. اکنون، بیایید روش های کارآمد ذخیره سازی داده هایی را که سازمان شما ممکن است جمع آوری کند، مورد بحث قرار دهیم.

7–1– راه حل های ذخیره سازی موازی

کسب‌وکار شما احتمالاً داده‌های بسیار بیشتری از آنچه که حتی در یک رایانه ذخیره می‌شود تولید می‌کند. برای اطمینان از اینکه همه داده‌ها ذخیره شده و دسترسی به آنها آسان است، باید آن‌ها را در رایانه‌های مختلف ذخیره کنید. شرکت شما ممکن است مجموعه ای از رایانه های ذخیره سازی خود را به نام "خوشه" یا "سرور" در محل داشته باشد.

7–2– ابر

از طرف دیگر، ممکن است شرکت شما به شرکت دیگری پول بدهد تا داده ها را برای شما ذخیره کند. این به عنوان "ذخیره سازی ابری" نامیده می شود. ارائه دهندگان فضای ابری رایج عبارتند از Microsoft Azure ، Amazon Web Services یا AWS و Google Cloud . این خدمات بیش از ذخیره سازی داده ها را ارائه می دهند. آنها همچنین می توانند به سازمان شما در تجزیه و تحلیل داده ها، یادگیری ماشینی و یادگیری عمیق کمک کنند. در حال حاضر، ما فقط بر روی ذخیره سازی داده ها تمرکز می کنیم.



7–3– انواع ذخیره سازی داده ها

انواع مختلف داده ها به راه حل های ذخیره سازی متفاوتی نیاز دارند. برخی از داده‌ها مانند ایمیل، متن، فایل‌های ویدیویی و صوتی، صفحات وب و پیام‌های رسانه‌های اجتماعی بدون ساختار هستند. این نوع داده ها اغلب در نوعی پایگاه داده به نام پایگاه داده اسناد ذخیره می شوند.

Unstructured

- Email
- Text
- Video and audio files
- Web pages
- Social media

Document Database

معمولاً داده ها را می توان به صورت جداول اطلاعات بیان کرد، مانند آنچه ممکن است در یک صفحه گسترده بیابید. پایگاه داده ای که اطلاعات را در جداول ذخیره می کند، پایگاه داده رابطه ای نامیده می شود. هر دوی این نوع پایگاه‌های داده را می‌توان در ارائه‌دهندگان ذخیره‌سازی ابری که قبلاً ذکر شد، یافت.

Tabular

Customer Name	Customer Address	...
Jane Doe	123 Maple St.	...

Relational Database

7–4– پرس و جو داده ها

هنگامی که داده ها در یک پایگاه داده سند یا یک پایگاه داده رابطه ای ذخیره شدند، باید به آن دسترسی داشته باشیم. در سطح پایه، ما می‌خواهیم بتوانیم یک داده خاص مانند «همه تصاویری که در 3 بهمن ماه ایجاد شده‌اند» یا «همه آدرس‌های مشتریان در شهر تهران» را درخواست کنیم. علاوه بر این، حتی ممکن است بخواهیم تحلیل هایی مانند جمع، شمارش یا میانگین گیری داده ها را انجام دهیم.

هر نوع پایگاه داده زبان پرس و جو خود را دارد. پایگاه‌های داده اسناد عمدتاً از NoSQL استفاده می‌کنند، در حالی که پایگاه‌های داده رابطه‌ای عمدتاً از SQL استفاده می‌کنند.

Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

7–5– کنار هم قرار دادن همه : مکان

ذخیره داده های شرکت شما مانند ساخت یک کتابخانه است. ابتدا باید تصمیم بگیرید که کجا کتابخانه خود را بسازید. این مربوط به انتخاب یک ابر است: یا یک خوشه داخلی یا یکی از ارائه دهندگانی که قبلاً در مورد آن صحبت کردیمAzure ، AWS یا Google Cloud .

7–6– کنار هم قرار دادن همه : نوع داده ها

در مرحله بعد، باید تصمیم بگیرید که چه نوع قفسه هایی را برای نگهداری کتاب های خود نصب کنید. انواع قفسه ها به انواع کتاب ها بستگی دارد.

این مشابه انتخاب بین یک پایگاه داده سند برای داده های بدون ساختار یا یک پایگاه داده رابطه ای برای داده های جدولی است. درست مانند یک کتابخانه که ممکن است چندین نوع قفسه داشته باشد، ممکن است نیاز باشد که برخی از داده ها در یک پایگاه داده اسناد و سایر داده ها در یک پایگاه داده رابطه ای ذخیره شوند.

Data Type	Storage Solution
Unstructured	Document Database
Tabular	Relational Database

7–7– کنار هم قرار دادن همه : پرس و جو

در نهایت، به سیستمی برای ارجاع و بررسی کتاب ها نیاز دارید. نحوه مکان یابی و بازیابی هر کتاب به نحوه ذخیره آن کتاب بستگی دارد.

به طور مشابه، برای صحبت با پایگاه داده به یک زبان پرس و جو نیاز دارید. برای پایگاه‌های داده اسناد، ما معمولاً از NoSQLو برای پایگاه‌های داده رابطه‌ای، معمولاً از SQL استفاده می‌کنیم.

Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

منابع

1. Data Science for Business and Decision Making by Luiz Favero and Patrícia Belfiore
2. Harvard Business School at online.hbs.edu
3. Harvard Business Review at hbr.org
4. wikipedia.org
5. ibm.com/cloud/learn/data-science-introduction
6. coursera.org/browse/data-science
7. http://cloudscaling.com/blog/cloud-computing
8. https://www.redhat.com/en/technologies/management
9. https://github.com
10. https://www.darkreading.com/cloud