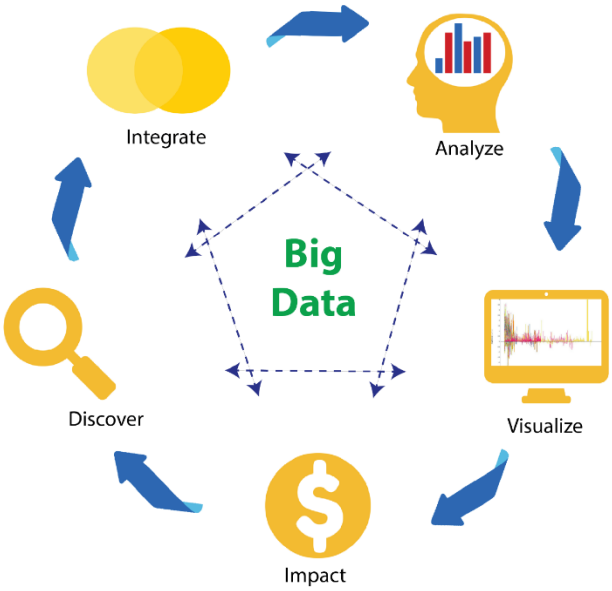
جلسه‌ی اول

منظور از big data چیست؟



BigData LifeCycle

به طور کلی bigData یک چرخه‌ی حیات هست . اول باید دیتا رو شناسایی کرد (discover) ، ممکنه نیاز باشه این دیتا از چندین source جمع‌آوری بشه. بعدش باید دیتاها رو با هم ادغام کرد (integrate) ، بعد دیتاها رو باید پردازش و تحلیل کرد (analyze) و بعد نتایج رو باید visualize کرد که از این نتایج باید بشه یه استنتاج‌هایی به دست آورد و در واقع خلق ارزش کرد (impact). به کل این فرآیندها bigData وارد میشه در نتیجه یه سری تکنولوژی‌هایی توی فرآیند discovery، یه سری تکنولوژی‌ها توی فرآیند integration یه سری توی فرآیند ذخیره‌سازی اطلاعات یه سری توی پردازش و تحلیل و یه سری توی visualize کردن داده‌های big و بالاخره یه سری هم توی استخراج دانش به ما کمک میکنند.

bigData characteristics

به صورت کلی bigData را بر اساس یه سری ویژگی و صفات تعریف میکنند.

- Volume: حجم دیتا زیاد باشه
- Velocity: سرعت ورود دیتا زیاد باشه.
- Variety: تنوع دیتا زیاد باشه
- Veracity: کیفیت دیتا کامل نباشه و دیتا حاوی نویز و عدم قطعیت باشه.
- Value: از یه دیتایی با همچین ویژگی‌هایی- 4 تا V بالایی- بشه خلق ارزش کرد.

ویژگی volume

Volume یعنی حجم دیتا از یه حدی فراتر میره که نمیشه با سیستم معمولی باهاش کار کرد مثلا از لحاظ ذخیره‌سازی و یا از لحاظ پردازشی ممکنه چندین ماه و یا چندین سال طول بکشه تا توی یه سیستم معمولی انجام بشه و یا انقدر دیتا زیاد باشه که نشه آن را visualize کرد.

مثلا ده گیگ دیتا رو نمیشه گفت bigData ولی مثلا ده ترابایت یا هزار ترابایت دیتا bigData حساب میشه. به عنوان مثال توی facebook به طور متوسط بالغ بر 15 ترابایت داده روزانه تولید میشه یعنی بعد از یک سال 5475 ترابایت داده خواهیم داشت. (طبق آمار و ارقام سال 2016).

همچنین وقتی میگیم volume معنیش این نیست که با یه داده‌ی static سر و کار داریم، معنیش این است که حجم دیتا روز به روز افزایش پیدا میکنه.

ویژگی velocity

Velocity یعنی سرعت ورود اطلاعات به حدی است که با سیستم‌های معمولی نمیشه handle شون کرد یعنی نمیشه توی سیستم همزمان هم ذخیره کرد هم پردازش کرد هم visualize کرد و .... مثلا توی یک دقیقه دو میلیون سرچ‌کوئری برای گوگل میاد. (کوئری باید دریافت بشه، پردازش بشه ، نتیجه آماده بشه و ارسال بشه به علاوه این کوئری باید تحلیل بشه و برای هر نفر مدل و پروفایل ساخته بشه تا بشه suggestion بهش داد و ....) یا مثلا توی youtube داره 1.3 میلیون ویدئو در دقیقه بازدید میشه . فارغ از اینکه ویدئوها باید به کاربر نمایش داده بشن، به ازای هر ویدئو که داره بازدید میشه باید یه سری آمار و ارقام نگهداری شوند (تعداد viewهای هر ویدئو باید آپدیت بشن ، به ازای هر view باید به صاحب محتوا پول داده بشه ، به ازای هر ویدئو باید مشخص شه که چه تبلیغاتی نشون داده بشه، به ازای هر کاربر باید پروفایل طرف آپدیت بشه و علاقه‌مندی‌هاش به دست بیاد ، .....)

ویژگی variety

قبلاًها اغلب دیتاهایی که تولید میشدند، دیتاهای ساخت‌یافته بودند یعنی همه‌ی رکوردهاش از یه ساختار مشخص پیروی میکردند مثل دیتابیس‌های relational که همه‌ی ردیف‌های یه جدول، ستون‌های یکسانی داشتند و هیچ ردیفی نبود که یک ستون کمتر و یا یک ستون بیشتر داشته باشه. تراکنش‌های بانکی structured هستند.

توی سال‌های اخیر نسبت داده‌های structured به unstructured کم شده‌است. داده‌های غیرساخت‌یافته مثل یک متن ، ویدئو، عکس و وویس. (مثلا ساختار دو داکیومنت یکسان نیست ، یه داکیومنت ممکنه پایان نامه باشه و یه داکیومت دیگه ممکنه یه فایلِ اکسل از حقوق کارکنانِ یه شرکت باشه). البته یه نوع سوم از دیتا هم وجود داره تحت عنوان semi structured یعنی اینکه توش ساختار وجود داره ولی این ساختار مثل

جدولای relational خیلی فیکس و ثابت نیست به عنوان مثال داده‌های xml ، json که ساختار دارند یعنی مثلا توی xml تگ‌هایی هست که باز و بسته می‌شوند ولی دو تا سند xml الزاما tagهای یکسانی ندارند.

توی bigData تمایل به سمت داده‌های semi structured و unstructured است البته نه اینکه توی بیگ‌دیتا با داده‌ی structured سر و کار نداریم بلکه عمدتا سیستم‌های بیگ جنسشون غیرساخت‌یافته است.

ویژگی veracity

قبلاًها دیتایی که داشتیم، دیتای فوق‌العاده clean بود یعنی مثلا توی یه تراکنش بانکی مشخصه که چه کسی برای چه کسی چه مبلغی در چه تاریخی واریز کرده. این دیتا هیچ عدم قطعیت و نویزی نداره اما توی بیگ‌دیتا منابع مختلف داده داریم که هر کدومشون ممکنه نویز داشته باشند مثلا فرض کنید یه سری سنسور قرار دادیم تا اطلاعات دما، رطوبت و ... رو دریافت کنیم. دیتای دما ممکنه با توجه به کیفیت سنسور نویز داشته باشه یا مثلا توی شبکه‌ی اجتماعی میخواهیم کامنت‌های یک پست رو تحلیل احساسات کنیم، ورودی یه سری متن است که میخواهیم تحلیل کنیم. تحلیل روی این متون 100 درصد درست نیست پس تصمیم‌گیری‌هایی که در نهایت از بیگ دیتا میشه اغلبشون دارای عدم قطعیت هستند.

نتیجه‌گیری

آیا الزاما همه‌ی ویژگی‌ها یعنی همه‌ی Vها باید وجود داشته باشه که بگیم bigdata ؟ خیر اگه حداقل یکی از این ویژگی‌ها وجود داشته باشه میگیم bigData .

آیا دیتای دیجی‌کالا بیگ هست؟

با توجه به اینکه اغلب افراد برای خرید کالاها و مقایسه‌ی قیمت‌ها یه سری به دیجی کالا میزنند و با توجه به اعتمادی که مردم به این شرکت دارند، تعداد کوئری به دیجی‌کالا بالا است (البته بالا بودن کوئری به دیجی کالا در حد youtube و google نیست اما در حد خودش بالا است) ، یعنی مهمترین ویژگی موجود در دیجی کالا velocity هست. از طرفی حجم دیتایی که دیجی کالا نگهداری میکنه، بالا هست سابقا که به ازای هر محصول یه سری عکس و متن صرفا ذخیره میکرد ، آن چنان حجمش زیاد نبود ، از چند وقت گذشته یه سری قابلیت جدید معرفی کرده مثل اینکه کاربران عکس و فیلم از کالایی که خریدند میذارن و این باعث میشه حجم دیتا بیشتر از قبل بشه (volume بالا) پس به طور کلی میشه دیجی کالا رو bigData حساب آورد.