الإنترنت وعصر «البيانات الضخمة BIG DATA»

هل فكرت يومًا ما هي كميّة المعلومات التي يقدّر أنها تتوافر للبشر الآن؟ إذا وُضعَت هذه المعلومات على أسطوانات رقميّة مدمَجَة «CD»، ثم جُعلَت في صفوف متراصة، لصنعت الأسطوانات المدمَجَة 5 طرق بين الأرض والقمر، مع ملاحظة أن المسافة بين الأرض والقمر هي قرابة 384.4 ألف كيلومتر، ما يعني أن معلومات البشر تساوي قرابة 1.9 مليون كيلومتر من الأسطوانات الرقمية المدمَجَة.

كم يبلغ متوسط حصّة الفرد من هذه المعلومات؟ عندما كانت «مكتبة الإسكندرية» هي مخزن معلومات البشر، بل المكان الذي كان أسطورة في ضخامة ما يحتويه من المعلومات المؤرشفة في وثائق مكتوبة، لم يكن ما تحتويه يزيد عن 1200 «إكسا بايت» Exabyte، مع ملاحظة أن كل «إكسابايت» تساوي بليون غيغابايت. لنعد إلى الفرد العادي في القرن الـ21. ما هي حصّته من المعلومات؟ إنها 320 «مكتبة إسكندرية» في عزّ ازدهارها وذروة تألّقها أيام الرومان. لو وُزّعت المعلومات على البشر، لنال كل شخص يعيش الآن أكثر مما احتوته «مكتبة الإسكندريّة» عندما كانت معقلاً ومخزنًا للفكر البشري، بما يزيد عن 320 ضعفًا.

ماهي البيانات (Data): هي الصورة الخام للمعلومات قبل عمليات الفرز والترتيب والمعالجة ولا يمكن الاستفادة منها بصورتها الأولية قبل المعالجة.

المعلومات (Information): هي البيانات التي خضعت للمعالجة والتحليل والتفسير والتي يمكن الاستفادة منها في استنباط العلاقات المختلفة بين الظواهر واتخاذ القرارات.

تُقسم البيانات الخام إلى ثلاثة أنواع:

بيانات مهيكلة: وهي البيانات المنظمة في صورة جداول أو قواعد بيانات تمهيدًا لمعالجتها.

بيانات غير مهيكلة: تشكل النسبة الأكبر من البيانات وهي البيانات التي يولدها الأشخاص يوميًّا من كتابات نصية وصور وفيديو ورسائل ونقرات على مواقع الإنترنت إلخ.

بيانات شبه مهيكلة: تعد نوعًا من البيانات المهيكلة إلا أن البيانات لا تصمم في جداول أو قواعد بيانات.

لنتابع مع لعبة الأرقام قليلاً. من المثير أيضًا أن كثيرًا من هذه المعلومات باتت مكتوبة بـ«لغة» واحدة: لغة الكومبيوتر. وتتوفر قسم كبير من هذه المعلومات على هيئة ملفات إلكترونيّة رقميّة مكتوبة على الكومبيوتر، تشمل الكتب والنصوص والصحف والمجلات والمنشورات والدراسات والصور والخرائط والرسومات والأفلام والأغاني والموسيقى والرسائل وما يكتب يوميًّا على مواقع شبكات التواصل الاجتماعي ورسائل الخليوي. بات كل شيء يكتب ويوضع على الكومبيوتر، فيصبح جزءًا من ملفات رقميّة مكتوبة بلغة الإلكترونات التي تتألف من عددين هما صفر وواحد! لنتأمّل ثانية. صارت كل المعلومات قابلة لأن توصف وتخزّن وتُصنّف وتبحث وترصف، عبر أداة واحدة هي الإلكترونات وملفاتها الرقميّة. يطلَق على هذا الأمر تسمية «الرقمَنَة» Digitization، وهي من الملامح الأبرز لزمن البيانات الضخمة. لم تعش البشرية شيئًا كهذا من قبل، وهو أمر يجدر تذكّره كثيرًا عند التفكير في البُعد المختلف لمسألة البيانات الكبيرة في عصر المعلوماتية.

ما هي البيانات الضخمة؟

البيانات الضخمة عبارة عن مجموعة من البيانات ذات أحجام تتخطي قدرة البرامج التي يشيع استخدامها لالتقاط وإدارة ومعالجة وتخزين وتحليل تلك البيانات في غضون مدة زمنية مقبولة.

وتتوافر البيانات الضخمة حولنا ففي كل دقيقه يولّد العالم من حولنا ما يقرب من 1.7 مليون بليون بايت من البيانات من مواقع التواصل الاجتماعي والبريد الإلكتروني لمواقع الأعمال الخاصة كأمازون واي- باي وأعمال البيع بالتجزئة للمشاريع العلمية والقومية العملاقة.

منذ عام 2012 كانت الحدود المفروضة على حجم مجموعات البيانات الملائمة للمعالجة في مدة معقولة من الوقت خاضعة لوحدة قياس البيانات إكسابايت. وتضاعفت القدرة التكنولوجية العالمية لتخزين المعلومات للفرد الواحد تقريبًا كل 40 شهر بدءًا من العام 1980 حيث في عام 2012 تم تقدير البيانات المُنتجه ب 2.5 كوينتيليون بايت (2.5 × 1018) من البيانات يوميًّا 80% منها غير مهيكلة مقابل 20% فقط من البيانات المهيكلة.

وبحسب شركة إنتل أن حجم البيانات التي ولدها البشر منذ بداية التاريخ وحتى عام 2003 ما قدره 5 إكسابايت، لكن هذا الرقم تضاعف 500 مرة خلال عام 2012 ليصل إلى 2.7 زيتابايت، ويتوقع أن يتضاعف هذا الرقم ثلاث مرات حتى عام 2015.

ويقوم كل فرد منا بإطلاق 2.2 مليون تيرا بايت من البيانات يوميًّا، و هناك 12 تيرابايت من التغريدات يوميًّا مع 25 تيرا بايت من سجلات الدخول على فيس بوك يوميًا وعلى تويتر أكثر من 200 مليون مستخدم نشط يكتبون أكثر من 230 مليون تغريدة يوميًّا.

حجم البيانات كان في 2009 حوالي 1 زيتا بايت (تريليون غيغابايت) وفي 2011 ارتفع إلى 1.8 زيتا بايت

تقول IBM إننا ننتج 2.5 كوينتيليون بايت من البيانات كل يوم (الكوينتيليون هو الرقم واحد متبوعًا بـ18 صفرًا). هذه البيانات تنبع من كل مكان، مثل المعلومات حول المناخ والتعليقات المنشورة على مواقع التواصل الاجتماعي والصور الرقمية والفيديوهات ومعاملات البيع والشراء.

تعد البيانات الضخمة الجيل القادم من الحوسبة والتي تعمل على خلق القيمة من خلال مسح وتحليل البيانات.

ومع مرور الزمن أصبحت البيانات التي ينتجها المستخدمين تنمو بشكل متسارع لعدة أسباب، منها بيانات المشتريات في محلات السوبر ماركت والأسواق التجارية وفواتير الشحن والمصارف والصحة وشبكات التواصل الاجتماعية.

ومع تطوير تقنيات التعرف على الوجه والأشخاص، فإنها ستتمكن من العثور على المزيد من التفاصيل والمعلومات عن أي شخص، ومع تزايد عدد الأجهزة المتصلة بالإنترنت، الأجهزة التي لم نعتد عليها أن تتصل بالشبكة العالمية مثل السيارات والبرادات والغسالات فإنها كلها تساهم في زيادة حجم البيانات المنتجة.

وتعد اليوم Hadoop من أفضل تقنيات التعامل مع البيانات الضخمة، وهي مكتبة مفتوحة المصدر مناسبة للتعامل مع البيانات الضخمة المتنوعة والسريعة، وتستخدم شركات كبرى خدمة Hadoop، مثلاً هناك لينكد إن شبكة التواصل الاجتماعية المتخصصة بالوظائف والعمل تستخدم الخدمة من أجل توليد أكثر من 100 مليار مقترح على المستخدمين أسبوعيًّا.

لكن ما الفائدة من البيانات الضخمة؟

تقول IBM إن البيانات الضخمة تعطيك فرصة اكتشاف رؤى مهمة في البيانات، وتقول أوراكل أن البيانات الضخمة تتيح للشركات أن تفهم بعمق أكثر زبائنها.

قدرت شركة سيسكو أنه وبحلول عام 2015 فإن حركة الزيارات على الإنترنت بالشكل الإجمالي ستتجاوز 4.8 زيتابايت (أي 4.8 مليار تيرا بايت) سنويًّا.

ومن الأمثلة العلمية لفائدة البيانات الضخمة:

– مصادم الهيدرون يملك 150 مليون جهاز استشعار تقدم بيانات 40 مليون مرة في الثانية الواحدة. وهناك ما يقرب من 600 مليون تصادم في الثانية الواحدة. لكن نتعامل فقط مع أقل من 0.001 % من بيانات تيار الاستشعار، فإن تدفق البيانات من جميع تجارب المصادم الأربعة يمثل 25 بيتابايت.

– موقع Amazon.com يعالج ملايين العمليات الخلفية كل يوم، فضلاً عن استفسارات من أكثر من نصف مليون بائع طرف ثالث. وتعتمد أمازون على نظام اللينوكس بشكل أساسي ليتمكن من التعامل مع هذا الكم الهائل من البيانات، و تملك أمازون أكبر 3 قواعد بيانات لينوكس في العالم التي تصل سعتها إلي 7.8، 18.5 و24.7 تيرابايت.

– سلسلة المتاجر Walmart تعالج أكثر من مليون معاملة تجارية كل ساعة، التي يتم استيرادها إلى قواعد بيانات يُقدر أنها تحتوي علي أكثر من 2.5 بيتابايت (2560 تيرابايت) من البيانات – وهو ما يوازي 167 ضعف البيانات الواردة في جميع الكتب الموجودة في مكتبة الكونغرس في الولايات المتحدة.

– يعالج فيس بوك 50 مليار صورة من قاعدة مستخدميه. ويقوم نظام حماية بطاقات الائتمان من الاحتيال " FICO Falcon Credit Card Fraud Detection System" بحماية 2.1 مليار حساب نشط في جميع أنحاء العالم.

– تقوم شركة Windermere Real Estate باستخدام إشارات GPS مجهولة من ما يقرب من 100 مليون سائق لمساعدة مشتري المنازل الجدد لتحديد أوقات قيادتهم من وإلى العمل خلال الأوقات المختلفة لليوم.

ينصحنا علماء عالم المعلوماتية وتقدّمها، على غرار ما ورد في كتاب صدر أخيرًا في أمريكا بعنوان «البيانات الكبيرة: ثورة تغيّر طريقة حياتنا وعملنا وتفكيرنا» Big Data. A revolution that would Change the Way we Live، Work & Think للبروفسور كينيث كروكر؟ ما هو الشيء الذي يرون أن الفكر البشري الآن يكسبه من الانتقال إلى «عصر البيانات الكبيرة»، بمعنى ما هو الأثر المستفاد من هذا التدفّق الهائل في المعلومات؟ لا تتسرع في الإجابة، لأنها ربما تحمل كثيرًا من الصدمة. ثمة 3 مسارات أساسية (مهارات، طرق في التفكير، أساليب في التدبير: سمّها ما شئت) هي:

1 – ضرورة هجران الدقّة لمصلحة التشديد على أهمية الخطأ ووجوده Get Rid of Accuracy. لا نكران للخطأ، بل الأهم وضعه دائمًا في الاعتبار والتفكير بأنه موجود فعليًّا.

2 – وضع الجهد باتجاه تجميع المعلومات والمزيد منها، ووضع المعلومات المتراكمة موضع الاستخدام اليومي والمستمر، Collect & Use a Lot of Data، خصوصًا مع تفتّح آفاق واسعة في الحصول على المعلومات وتنويعها وتجميعها وتخزينها، وبكلفة تتدنى باستمرار.

3 – التركيز على فهم العلاقات والروابط والميول بدل التشديد على البحث عن الأسباب والعلل Give Up On Causes- Accept Correlations! ربما لا يعجب كثيرًا من العقول في العالم العربي هذا الكلام في هذا الكتاب الذي تجدر ترجمته إلى اللغة العربيّة، لأنه ينطق عن لحظة الحياة الحاضرة مع تلمّس جريء للمستقبل.