مركز أوال للدراسات والتوثيق
AWAL CENTRE FOR STUDIES AND DOCUMENTATION

أرشيفو
ARCHIVO
العدد 6 - حزيران/ يونيو 2017

ثقافة أرشيفية

البيانات الضّخمةإعصارٌ يحتوي العالم

زينب البزال

مع الازدياد الهائل في كميّات البيانات المُنبثقة من مصادر رقميّة متعدّدة، أصبحت الحاجة إلى تحليلها ملحّة، ما يجعل هذه البيانات الَّتي كانت تُعتَبَر مُظلمة وغير مجدية طيلة سنوات، ذات أهميّة كبرى. وبما أنَّ الشّركات أصبَحت مُلزَمَة بتقديم نتائج تحليل مباشرة، فإنَّ عملية تحليل بياناتها الضّخمة أصبحت ضرورة مهمّة.

تُعتَبر شركات غوغل وأمازون وفايسبوك وغيرها رائدة في مجال تحليل البيانات الضّخمة، فغوغل يجري عمليّات بحثيّة في ملايين المواقع لإيجاد نتائج جيّدة ومطابقة لطلبات البحث، وأمازون يعرض لمستخدميه مقترحات شخصيّة ذكيّة للشراء، وفيسبوك يعرض لمالك الحساب مقطعًا قصيرًا من الفيديو يشمل مختصرًا لأهمّ الأحداث من ضمن المنشورات النصيّة والصور ومقاطع الفيديو و«الإعجابات» والتعليقات الموجودة لديه منذ إنشاء حسابه.

هذه الشّركات تستخدم المعلومات المستخرجة من البيانات الضّخمة للحصول على قيمة تنافسيّة. فلو فكّرنا في نظام أمازون الذي يقدّم اقتراحات للمستخدم، فإنّ الشركة هنا تستخدم تاريخ المشتريات لدى المستخدم، إضافةً إلى ما تعلمه عنه من أنماط الشراء وأنماط شراء المستخدمين الَّذين يشبهونه، من أجل تقديم بعض المقترحات الجيّدة والمفيدة له. إنّها آلة تسويقيّة. كما أنّ إمكانيّات التحليل الضّخمة لدى هذه الشّركات جعلتها ناجحة بشكل كبير.

ما هي البيانات الضّخمة؟

«البيانات الضّخمة» هي مصطلحٌ «ساخن» في عالم المعلومات والمعلوماتيّة. قبل الدّخول في تعريفها وتقديم معلومات تفصيلية عنها، نعرّف في المقابل مصطلح «البيانات الصّغيرة»، فما يُسمّى عادةً بالبيانات هو في الواقع «بيانات صغيرة».

تُوصف البيانات الصّغيرة بأنّها ذات حجم وشكل يمكّنها من أن تكون متاحة ومُخْبِرَة وفعّالة، فهي عادةً تُعطي معلومات من شأنها الإجابة عن أسئلة محدّدة أو معالجة مشكلة محدّدة. من الأمثلة على ذلك: نتائج مباريات، تقارير مبيعات محدّدة، أسعار منتجات محدّدة وأنماط تغيّرها... تتميّز هذه البيانات بأنّها محدّدة من حيث الشّكل وطريقة الحفظ، ممّا يتيح استخراجها وتحليلها بطريقة واضحة.

في المقابل، فإنّ مصطلح «البيانات الضّخمة» يصف أيّ حجم من البيانات التي قد تكون مهيكلة بالكامل

أو جزئيًا أو غير مهيكلة نهائيًا، والتي تمكّننا من استخراج المعلومات.

تتميّز البيانات الضَّخمة بثلاثة أبعاد:

1. ضخامة الحجم.
2. التنوّع الكبير في الشَّكل وطريقة الحفظ.
3. السّرعة المطلوبة للمعالجة والتّحليل.

ورغم أنّ مصطلح البيانات الضّخمة لا يعبّر عن قيمة محدّدة لحجم البيانات، فإنّه غالبًا ما يستخدم للتعبير عن «تيرابايتس»، «بيتابايتس»، وحتّى «هيغزابايتس» من البيانات المحصّلة عبر الزّمن.

من أين تأتي البيانات الضّخمة؟

يتمّ جمع البيانات الضّخمة من عددٍ لا يُحصى من المصادر، من مثل سجلّات المبيعات التجاريّة، سجلّات الاختبارات العلميّة، وعمليّات البحث عبر محرّكات البحث... قد تكون هذه البيانات خامًا (أوليّة) أو خاضعة لمعالجة مسبقة باستخدام أدوات برامج الحاسوب قبل أن يتمّ تحليلها. أمّا بالنسبة إلى حفظ البيانات، فقد تكون محفوظة بطريقة مهيكلة، من مثل (SQL Database)، أو غير مهيكلة، أو متدفّقة من أجهزة الاستشعار.

يقوم كلّ مشروع تحليليّ للبيانات الضَّخمة بعمليّات البحث والرّبط وتحليل مصادر البيانات. بعد ذلك، يعطي جوابًا أو نتيجةً، بناءً على طلبٍ محدّد. ولهذه الغاية، فقد توسّعت عمليّات تحليل البيانات لتشمل مجالات التعلّم الآلي والذّكاء الاصطناعي.

نظرة تاريخيّة إلى البيانات الضّخمة

أتاح ظهور الإنترنت في العام 1991 لكلّ شخص بأن يكون متّصلًا، بحيث يستطيع تحميل معلوماته وتحليل المعلومات المحمّلة من قبل الآخرين. وفي العام 1997، ظهر محرّك البحث «غوغل» الَّذي سرعان ما أصبح الأكثر شهرةً واستخدامًا في العالم. وفي العام 1999، استُخدم مصطلح «البيانات الضّخمة» لأوّل مرّة، وذلك في رسالة أكاديميّة باللّغة الإنكليزيّة بعنوان «Visually Exploring Gigabyte Datasets in Real Time». وفي العام نفسه، استُخدم مصطلح «إنترنت الأشياء» في خلال تقديم عرض عمل من قِبَل كيفن آشتون.

في العام 2000، حاول بيتر لايمان وهال فاريان (اقتصادي لدى غوغل حاليًّا)، أن يحدّدا حجم البيانات الرقميّة السّنوية في العالم ونموّها، فتوصّلا إلى أنّ إنتاج العالم السّنوي من المطبوعات

والأفلام والصّور وغيرها يحتاج إلى 1.5 جيغابايت من حجم التخزين، أي ما كان يعادل 250 ميغابايتًا لكلّ شخص. بعد ذلك، حدّد دوغ لاني في العام 2001، الأبعاد الثلاثة لإدارة البيانات، وهي: حجم البيانات، تنوّع شكلها، وسرعة استخراجها.

دخل الويب مرحلة 2.0 (Web 2.0) في العام 2005، ممّا دفع إلى ازدياد حجم البيانات بشكل كبير جدًّا، إذ أصبحت تُنتَج مباشرةً من المستخدم، بدلًا من مقدّم الخدمة على الويب. في هذه المرحلة، ظهر الفايسبوك، ما أتاح لملايين المستخدمين تحميل البيانات ومشاركتها مع أصدقائهم. وفي هذه المرحلة أيضًا، تمّ إنشاء هادوب (Hadoop)، وهو نظام مفتوح المصدر (open source)، صُنع خصّيصًا لحفظ مجموعات البيانات الضخمة المتنوّعة (صوتيّة، فيديو، نصّية) وتحليلها.

في العام 2010، قال إيريك شميديت، المدير التنفيذي لدى غوغل، في مؤتمر، إنّ كميّة البيانات التي تصدر خلال يومين، توازي البيانات الَّتي صدرت منذ بداية الحضارة الإنسانيّة وحتّى العام 2003. ساهمت أجهزة الهواتف المحمولة بشكل كبير في ذلك، فقد أصبح الكثير من الأشخاص يستخدمون هواتفهم الجوّالة للوصول إلى البيانات الرقميّة. وبعد أربع سنوات، بدأ «الموبايل إنترنت» يحلّ محلّ الديسكتوب. وهنا، أصبحت الحاجة إلى تحليل البيانات الضّخمة ملحّة.

خفض الإنفاق: تحقّق تكنولوجيا البيانات الضّخمة، مثل هادوب، والتّحليلات المبنيّة على التقنيّات السحابيّة (Cloud-based Computing)، مزايا كبيرة لجهة التكلفة، عندما يتعلّق الأمر بتخزين كميّات ضخمة من البيانات.

السرعة والفعاليّة في اتّخاذ القرار: يتميّز نظام هادوب بسرعة تحليليّة عالية، إضافةً إلى إمكانيّة تحليل بيانات من مصادر جديدة، ما يسمح بالقيام بالتّحليل واتّخاذ القرار بشكل مباشر.

إيجاد منتجات وخدمات جديدة: لقد أصبح بالإمكان استحداث منتجات وخدمات جديدة ترضي المستخدم بشكل أكبر.

التكنولوجيا الرئيسيّة المستخدمة في التَّحليل

إدارة البيانات: يجب أن تكون البيانات ذات جودة عالية ومحكمة جيدًا قبل القيام بتحليلها. فمع تدفّق البيانات باستمرار داخل المنظّمة وخارجها، كان من الضروري إنشاء عمليّات متكرّرة لبناء معايير جودتها والحفاظ عليها. وبمجرّد أن تصبح البيانات موثوقة، يتوجّب على المؤسّسات إنشاء برامج رئيسيّة لإدارتها.

التنقيب في البيانات (Datamining): هي التكنولوجيا الّتي تساعد على فحص كميّات ضخمة من البيانات، من أجل اكتشاف الأنماط فيها. هذه الأنماط من شأنها المساهمة في تحليلات لإعطاء إجابة عن مسألة عمليّة.

هادوب (Hadoop): هو برنامج مفتوح المصدر لإطار عمل (open source software framework)، يمكّن من تخزين كميّات ضخمة من البيانات. لقد أصبح هادوب التكنولوجيا الرئيسية لممارسة الأعمال التجارية، بسبب الزيادة المستمرة في حجم البيانات وتنوّع أشكالها، كما أنّ نموذج الحوسبة الموزّعة لديه، يمكّن من تحليل البيانات بسرعة كبيرة.

التحليل في الذاكرة (In-memory analytics): من خلال تحليل البيانات من ذاكرة النظام (بدلًا من محرك القرص الثابت)، يمكن استخلاص رؤى فورية من البيانات والتصرّف بسرعة بناءً عليها.

التحليلات التنبؤية (Predictive analytics): تستخدم تقنيَّة التحليلات التنبؤية الخوارزميات الإحصائيَّة وتقنيات التعلّم الآلي لتحديد احتمالات النتائج المستقبليّة، استنادًا إلى البيانات التاريخيَّة.

تنقيب النّصوص (text mining): تمكّن هذه التكنولوجيا من تحليل البيانات النصّية من شبكة الإنترنت، وتعليقات المستخدمين، والكتب والمصادر النّصيّة الأخرى، للكشف عن رؤى لم تتم ملاحظتها من قبل.

لقد أصبح معلومًا لدى معظم المؤسّسات الّتي تمتلك البيانات، بأنّ عمليّة اقتناء جميع البيانات الّتي تتدفّق إليها وحفظها، تمكّنها من القيام بعمليّات تحليليّة تؤدي إلى فائدة كبيرة. فمنذ عشرات السنين، وقبل بروز مصطلح «البيانات الضّخمة» بعقود، كانت الكثير من الشركات تستخدم أنظمة تحليليّة بدائيّة للقيام بعمليات تحليليّة، لكنَّ التميّز الّذي تقدمه اليوم الطرق الجديدة لتحليل البيانات الضَّخمة، هو عامل السّرعة والفعاليّة. ففي السّابق، كانت البيانات تُجمع أولًا، ومن ثمّ يجري القيام بتحليلها لاتخاذ قرارات مستقبليّة. أمّا اليوم، فقد أصبحت عمليّة التحليل تتمّ بسرعة عالية جدًا وبشكل آنيٍّ، ممّا يمكّن المؤسّسات من امتلاك قيمة تنافسيّة كانت تفتقر إليها من قبل، فالقيام بهذا التّحليل يؤدّي إلى جني أرباح أكبر، ويجعل المستخدم أكثر رضًى.

زينب البزال: اختصاصيَّة في مجال البرمجة، وهي مبرمجة محلّلة في إدارة الإحصاء المركزي.
للتواصل عبر الإيميل: albazzalzainab@gmail.com