تحليلات البيانات الكبيرة وأهمية تطبيقها في المنشاءات التجارية والصناعية

أشرف عبد العزيز القماش[١]– هدى قرشى محمد[٢]– حسن محمد شحاتة[٢]
السيد محمد حلمى خاطر[٣]
١) باحث بمعهد الدراسات والبحوث البيئية، جامعة عين شمس ٢) كلية الهندسة، جامعة عين شمس ٣) المركز القومى للبحوث

المستخلص

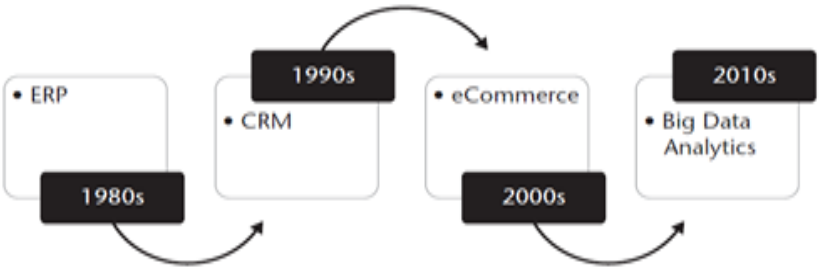
مع الاعتمادية الكاملة على نظم المعلومات الحديثة والتقنيات الرقمية فانه يتم استخراج كم هائل من البيانات كل يوم، يقاس حجمها ب إكسا بايت ($١٠^{١٨}$ بايت)، وذلك ناتج من استخدام تطبيقات عملية في الحياة اليومية مثل إنترنت الأشياء والحوسبة السحابية. يتطلب تحليل هذا الكم الضخم المتراكم من البيانات الكبيرة الكثير من الجهود على مستويات متعددة لاستخراج المعرفة اللازمة لاتخاذ القرار الصحيح في الوقت الملائم. لذلك فان التصدي بالتحليل لهذه البيانات الكبيرة هو مجال غير مسبوق للبحث والتطوير الحالي. ويعتبر الهدف الأساسي من هذا البحث هو استكشاف التأثير المحتمل لتحديات البيانات الكبيرة، ودراسة قضايا البحوث المفتوحة، والأدوات المختلفة المرتبطة بها. ويمكن اعتبار هذا البحث يوفر نافذة صغيرة لاستكشاف عالم البيانات الكبيرة الآخذ في النمو بتسارع رهيب في مراحله المختلفة والمتعددة. بالإضافة إلى ذلك فانه يفتح أفاقًا جديده للباحثين لتطوير الحلول، استنادا إلى التحديات والقضايا البحثية المفتوحة. كما يساهم في نشر التبادل المعرفي للنهمين لمواكبة التطور في هذا المجال من مع غير الباحثين المتخصصين.

المقدمة

خلال انعقاد الدورة الحادية عشرة للندوة العالمية لمؤشرات الاتصالات/تكنولوجيا المعلومات والاتصال (WTIS)، بالمكسيك خلال الفترة من ٤ إلى ٦ ديسمبر ٢٠١٣م، تم

الترقيم الدولي ISSN 1110-0826

اعتماد تقنية البيانات الكبيرة باعتبارها تمثِّل طاقة هائلة لتعزيز التنمية عن طريق تهيئة معلومات حالية، بتكلفة منخفضة بالقياس إلى البيانات المتوفرة من مصادر أخرى (Kakhani, *et al.*, 2015). تمتلك تقنية البيانات الكبيرة إمكانية تحليل بيانات الحساسات/أجهزة الاستشعار، ومواقع الانترنت، وبيانات شبكات التواصل الاجتماعي بأنواعها المختلفة، حيث أن تحليل هذه البيانات يتيح وجود علاقات بين مجموعات من البيانات المستقلة، تتيح امكانية استكشاف نواحي جديدة مختلفة؛ ومنها على سبيل المثال، التنبؤ بالاتجاهات التجارية للشركات، وربط الاستشارات القانونية، ومكافحة الجريمة، وتحديد ظروف حركة تدفق البيانات، وغيرها. وتوفر هذه التنبؤات لصانع القرار أدوات مبتكرة لفهم أفضل للعملاء والأسواق، وإدارة المخاطر على نحو فعال. ومن اللافت للنظر عند دراسة المسار الزمنى لتطور تقنية البيانات الكبيرة، بأن البيانات أصبحت عاملاً رئيسياً للإنتاج. ربما أكثر أهمية من الأرض والعمل نفسه ورأس المال، وفي المقابل فإن كم المعلومات الناتج سوف تدفع تحول العمليات والنماذج التجارية ، نحو تحقيق مستويات أعلى من الجودة والكفاءة والفاعلية. في العالم الرقمي، يتم إنشاء البيانات من مصادر مختلفة والانتقال السريع من التقنيات الرقمية أدى إلى نمو البيانات الكبيرة. ويوفر اختراقات تطورية في العديد من الحقول مع مجموعة من مجموعات البيانات الكبيرة. بشكل عام، هو يشير إلى مجموعة من مجموعات البيانات الكبيرة والمعقدة التي يصعب معالجتها باستخدام إدارة قواعد البيانات التقليدية أدوات أو تطبيقات معالجة البيانات، والشكل (١) التالي يشير إلى المسار الزمنى لتطور استخدام تقنية البيانات الكبيرة.



شكل (١): المسار الزمنى لتطور تقنية البيانات الكبيرة

ومع التوجه العالمي نحو التحول الرقمي، لوحظ ان الانتقال السريع لاستخدام التقنيات الرقمية أدى إلى نمو سريع في استخدامات البيانات الكبيرة التي يتم إنشاء البيانات بها من مصادر مختلفة ووسائط متعددة. مما اتاح طفرات تطويرية في العديد من المجالات بصورة غير مسبوقة، وتوفير مجموعات غير نمطية من البيانات الكبيرة. بشكل عام، فان التحول الرقمي يشير إلى مجموعة من مجموعات البيانات الكبيرة والمعقدة التي يصعب معالجتها باستخدام إدارة قواعد البيانات التقليدية بنفس أدوات أو تطبيقات معالجة هذه البيانات الحالية. وتوافر المعلومات المتاحة حول تصنيف هذه البيانات من حيث هيكلتها (منظم / غير منظم / شبه هيكلي). ليصبح الهدف الرئيسي من تكوين البيانات الكبيرة هو تحليلها بعد معالجة هذا الكم الهائل من البيانات مع مراعاة السرعة والتنوع في مصادرها وصحتها ودقتها باستخدام مختلف التقنيات التقليدية والحسابية الذكية. وتعتبر محاولة وضع التعريف الدقيق للبيانات الكبيرة، سوف تساعد في الحصول على دعم صنع / اتخاذ القرار، واكتشاف و تحسين الأداء باستخدام ادوات وطرق مبتكرة وفعالة من حيث التكلفة أيضا. من المتوقع أن يتم نمو البيانات الكبيرة بشكل متسارع بحلول عام ٢٠١٥ [Lynch, 2008]. ويتجلى ذلك في تطبيقات تكنولوجيا المعلومات والاتصالات، والبيانات الكبيرة مرشحة لتكون قاطرة

التقدم في الجيل القادم من صناعات تطوير تكنولوجيا المعلومات (Jin, *et al.*, 2015)، التي بنيت على نطاق واسع على تقنيات 3G / 4G / 5 G، معتمدة على المتوفر حاليا من البيانات الكبيرة، والحوسبة السحابية، وإنترنت الاشياء، والأعمال الاجتماعية. مع الاخذ في الاعتبار ان استخراج المعرفة الدقيقة من البيانات الكبيرة المتاحة ستصبح هي الهدف النهائي. ولحل معضلة ان معظم طرق البحث التقليدية عن البيانات لم تعد قادرة على التعامل مع مجموعات البيانات الكبيرة بنجاح. وتعتبر المشكلة في تحليل البيانات الكبيرة الناتجة من عدم التنسيق الكامل بين أنظمة قواعد البيانات وكذلك مع أدوات التحليل مثل استخراج البيانات والتحليل الإحصائي، من التحديات التي تنشأ عندما نرغب في اكتشاف المعرفة والتمثيل لتطبيقاتها العملية. تكمن المشكلة الأساسية في كيفية الوصف الكمي و خصائص البيانات الكبيرة. هناك حاجة لدراسة المعرفة الناتجة من ثورة البيانات الكبيرة (Kitchin, 2014). بالإضافة إلى ذلك، فان التحليلات المبنية على دراسة درجة التعقيد في انشاء البيانات الكبيرة سوف تساعدنا في فهم خصائص وتشكيل أنماط معقدة من البيانات الكبيرة، وتبسيط تمثيلها، لنحصل على المعرفة مجردة، وتوجيه تصميم نماذج الحوسبة والخوارزميات المبنية على البيانات الكبيرة. وقد أجريت الكثير من البحوث من قبل مختلف الباحثون على البيانات الكبيرة واتجاهاتها (Rio, *et al.*, 2014). وتجدر الإشارة إلى أن معظم البيانات المتاحة في شكل البيانات الضخمة حاليا غير مفيدة للتحليل أو دعم اتخاذ القرارات، وبالرغم من ذلك فان الكيانات الصناعية الكبيرة والأوساط الأكاديمية مهتمة بنشر نتائج تحليلات البيانات الكبيرة.

هدف البحث

وفى هذا البحث سيتم التركيز على التحديات التي تواجهنا في تصميم وانشاء البيانات الكبيرة والتقنيات المتاحة. بالإضافة إلى محاولة مناقشة قضايا البحوث المفتوحة في هذا الصدد.

وقد تم تقسيم هذا البحث إلى الأقسام التالية:

- تعريف البيانات الكبيرة، بداية نشأتها و خصائصها المحددة له.
- طرح التحديات التي تنشأ خلال تحليلات البيانات الكبيرة.
- مناقشة القضايا البحثية التي يمكن أن تساعدنا لمعالجة البيانات الكبيرة واستخراج المعرفة.
- دراسة بعض ادوات التحليل والتقنيات المتاحة لمعالجة البيانات الكبيرة.
- ملاحظات ختامية.
- النتائج.

تعريف البيانات الكبيرة

بدأ مصطلح البيانات الكبيرة في الظهور في أوائل التسعينيات، وزاد انتشاره وأهميته بشكل متنامي بمرور الزمن. غالباً ما يُنظر إلى البيانات الكبيرة على أنها جزء لا يتجزأ من استراتيجية بيانات الشركة، حيث انه هناك اعتقاد أنه عند ظهور مشكلة محددة فهي سوف تساعدنا في الحصول على دعم صنع القرار، و التحسين في الأداء بصورة مبتكرة وفعالة من حيث انخفاض التكلفة أيضا.

الخصائص العشرة للبيانات الكبيرة (10 V's)

[https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx]

للبيانات الضخمة خصائص محددة يمكن أن تساعدك على فهم كل من التحديات والمزايا التي تشملها هذه التقنية وهى:

الحجم Volume: الحجم هو على الأرجح أفضل خاصية معروفة للبيانات الكبيرة؛ هذه ليست مفاجأة، بالنظر إلى أن ٩٠ % من جميع بيانات اليوم تم إنشاؤها في العامين الماضيين. وعلى سبيل المثال فإنه:

- يتم تحميل ٣٠٠ ساعة من الفيديو على YouTube كل دقيقة.
- تم التقاط ما يقدر بنحو ١,١ تريليون (10^{12}) صورة في عام ٢٠١٦، ومن المتوقع أن يرتفع هذا العدد بنسبة ٩ % في عام ٢٠١٧. وبما أن الصورة نفسها عادةً ما تحتوي على مثيلات متعددة مخزنة عبر أجهزة مختلفة، وخدمات مشاركة الصور أو المستندات، وكذلك خدمات الوسائط الاجتماعية، من المتوقع أيضاً أن يزيد إجمالي عدد الصور المخزنة من ٣,٩ تريليون (10^{12}) صورة في عام ٢٠١٦ إلى ٤,٧ تريليون (10^{12}) صورة في عام ٢٠١٧.
- في عام ٢٠١٦، بلغ معدل الحركة المتنقلة العالمية ٦,٢ *إكسا بايت* (10^{18} بايت) في الشهر.

السرعة Velocity: تشير السرعة إلى السرعة التي يتم بها إنشاء البيانات أو إنتاجها أو إنشاؤها أو تحديثها. بالتأكيد، يبدو من المثير للإعجاب أن مستودع بيانات فيسبوك يخزن ما يصل إلى ٣٠٠ بيتا بايت (10^{15} بايت) من البيانات، لكن السرعة التي يتم بها إنشاء بيانات جديدة يجب أن تؤخذ في الاعتبار. فيس بوك يستقبل ٦٠٠ *تيرا بايت* (10^{12} بايت) من البيانات الواردة يوميا. تقوم Google وحدها بمعالجة ما يزيد عن "٤٠،٠٠٠ استفسار بحث كل ثانية" في المتوسط، والتي تُترجم تقريباً إلى أكثر من ٣,٥ جيجا (10^{9}) عملية بحث يومياً.

التنوع Variety: عندما يتعلق الأمر بالبيانات الكبيرة، لا يتعين علينا التعامل مع البيانات المهيكلة فحسب، بل أيضاً البيانات شبه المهيكلة وغير المهيكلة أيضاً. كما يمكنك الاستنتاج من الأمثلة المذكورة أعلاه، أن معظم البيانات الكبيرة غير منظمة، ولكن إلى

الترقيم الدولي ISSN 1110-0826

جانب الصوت والصورة وملفات الفيديو وتحديثات الوسائط الاجتماعية وتنسيقات نصية أخرى، هناك أيضاً ملفات السجل، وبيانات النقر، وبيانات الجهاز والحساسات المتصلة بالمعدات والاجهزة، بيانات تتبع الأماكن وخلافه.

التغير Variability: يشير التباين في سياق البيانات الكبيرة إلى بعض الأشياء المختلفة، والمتغيرة أيضاً بسبب تعدد أبعاد البيانات الناتجة عن أنواع ومصادر بيانات متباينة متعددة. يمكن أن تشير المتغيرات أيضاً إلى السرعة غير المتسقة التي يتم بها تحميل البيانات الكبيرة في قاعدة البيانات الخاصة بك.

الصدق Veracity: هذه هي واحدة من الخصائص المؤسفة للبيانات الكبيرة. مع زيادة أي من الخصائص المذكورة أعلاه أو كلها، تنخفض الصدقية (الثقة في البيانات). تشير زيادة الصدقية إلى موثوقية مصدر البيانات، وسياقه، ومدى جدوى التحليل المستند إليه. على سبيل المثال، ضع في اعتبارك مجموعة بيانات من الإحصاءات حول ما يشتريه الأشخاص في المطاعم وأسعار هذه العناصر على مدار السنوات الخمس الماضية. قد تسأل: من الذي أنشأ المصدر؟ ما هي المنهجية التي اتبعوها في جمع البيانات؟ هل تم تضمين بعض المأكولات فقط أو أنواع معينة من المطاعم؟ هل قام منشئو البيانات بتلخيص المعلومات؟ هل تم تعديل المعلومات؟ الإجابات على هذه الأسئلة ضرورية لتحديد صحة هذه المعلومات. وتساعدنا معرفة صحة البيانات على فهم المخاطر المرتبطة بالتحليل وقرارات العمل التي يتم اتخاذها بشكل أفضل بناءً على هذه المجموعة من البيانات.

الصلاحية Validity: على غرار الدقة، تشير الصلاحية إلى مدى دقة وتصحيح البيانات للاستخدام المقصود منها. وفقًا لمجلة فوربس ، يستهلك ٦٠ % من وقت علماء البيانات في تنظيفها قبل التمكن من إجراء أي عمليات تحليل. إن الاستفادة من تحليلات

الترقيم الدولي ISSN 1110-0826

البيانات الكبيرة ليست بنفس جودة البيانات الأساسية، لذلك تحتاج إلى اتباع اساليب واطر حاكمة لضمان جودة البيانات المتسقة، والتعاريف الشائعة، والبيانات الوصفية.

الضعف Vulnerability: البيانات الكبيرة ستجلب معها مخاوف أمنية جديدة. حيث أن اختراق البيانات مع البيانات الكبيرة سيسبب العديد من الكوارث الكبرى. ولسوء الحظ، كان هناك العديد من الاختراقات للبيانات الكبيرة. مثل تسريبات Wikileaks، وأيضا في مايو ٢٠١٦ قام أحد المتطفلين باسم Peace بنشر بيانات على شبكة الإنترنت المظلمة لبيعها، والتي زُعم أنها تضمنت معلومات عن ١٦٧ مليون (٦١٠) حساب على linkedIn ، بالإضافة الي ٣٦٠ مليون (٦١٠) بريد إلكتروني وكلمة مرور مستخدم.

التقلب Volatility: إلى أي مدى يجب أن تكون بياناتك قديمة قبل أن تعتبر غير ذات صلة أو تاريخية أو غير مفيدة بعد الآن؟ كم من الوقت تحتاج البيانات إلى الاحتفاظ بها؟ قبل بروز تقنية البيانات الكبيرة، كانت تميل المؤسسات إلى تخزين البيانات إلى أجل غير مسمى، قد لا ينتج عن حفظ وتخزين المئات من تيرا بايت من البيانات نفقات تخزين عالية؛ ويمكن الاحتفاظ بها حتى في قاعدة البيانات الاصلية دون التسبب في مشاكل في الأداء. وكذلك في إعداد البيانات التقليدية، قد لا توجد سياسات حاكمة لعمليات أرشفة البيانات، لكن بسبب سرعة وحجم البيانات الكبيرة، فإن ذلك يحتاج إلى دراسة متأنية. هنا سنحتاج إلى وضع قواعد لعملية اعداد البيانات واتاحتها عند طلبها وكذلك ضمان الاسترجاع السريع للمعلومات عند الحاجة. مع التأكد من ارتباطها الواضح باحتياجاتك وعملياتك ومعاملاتك التجارية. لأنه عند التعامل مع البيانات الكبيرة، سترتفع بصورة كبيرة التكاليف وتزداد تعقيدا عملية التخزين والاسترجاع.

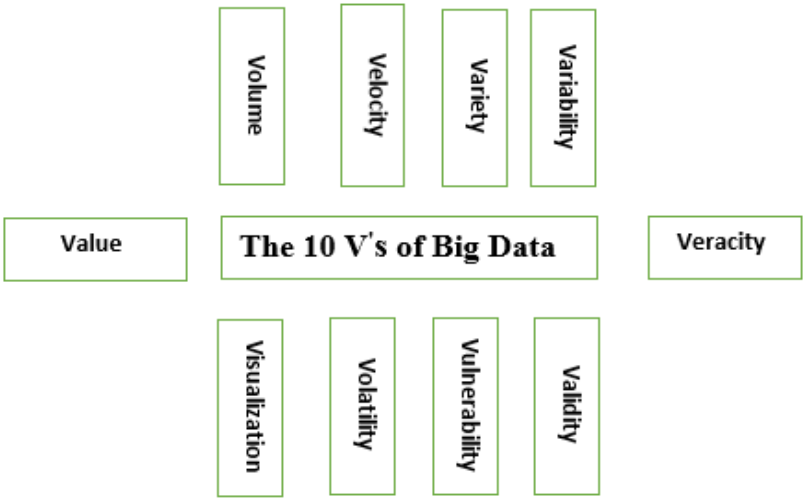
التصور Visualization: من الخصائص الأخرى للبيانات الضخمة مدى صعوبة التصور. تواجه أدوات التصور الكبيرة للبيانات الحالية تحديات تقنية بسبب القيود المفروضة على التكنولوجيا في الذاكرة وضعف قابلية التوسع والوظائف ووقت الاستجابة.

الترقيم الدولي ISSN 1110-0826

لا يمكنك الاعتماد على الرسوم البيانية التقليدية عند محاولة رسم مليار نقطة بيانات، لذلك تحتاج إلى طرق مختلفة لتمثيل البيانات مثل تجميع البيانات أو استخدام خرائط الأشجار، أو انفجار الشمس، أو الإحداثيات الموازية، أو مخططات الشبكة الدائرية، أو أشجار المخروط. ادمج هذا مع العديد من المتغيرات الناتجة عن تنوع البيانات الكبيرة وسرعتها والعلاقات المعقدة بينهما، ويمكنك أن ترى أن تطوير تصور مفيد ليس بالأمر السهل.

القيمة Value: أخيراً، ولكن الأهم من ذلك كله هو القيمة. الخصائص الأخرى للبيانات الكبيرة لا معنى لها إذا لم تستمد قيمة الأعمال من البيانات. يمكن العثور على قيمة كبيرة في البيانات الكبيرة، بما في ذلك فهم عملائك بشكل أفضل، واستهدافهم وفقًا لذلك، وتحسين العمليات، وتحسين أداء الجهاز أو العمل. تحتاج إلى فهم الإمكانات، إلى جانب الخصائص الأكثر تحديا، قبل الشروع في استراتيجية البيانات الكبيرة.و الشكل (٢) التالي يوضح الخصائص العشرة للبيانات الكبيرة (10 V's).

الترقيم الدولي ISSN 1110-0826



شكل (٢): الخصائص العشرة للبيانات الكبيرة (10 V's)

الترقيم الدولي ISSN 1110-0826

التحديات في تحليلات البيانات الكبيرة

في السنوات الأخيرة، تراكمت البيانات الكبيرة في عدة مجالات مثل الرعاية الصحية، الإدارة العامة، البيع بالتجزئة، المعاملات البنكية، الكيمياء الحيوية، وغيرها من البحوث العلمية متعددة التخصصات. كما تواجه التطبيقات المعتمدة على الويب تقنية البيانات الكبيرة بشكل متكرر، كامثلة على ذلك تطبيقات الحوسبة الاجتماعية، وعرض النصوص والوثائق عبر الإنترنت، وفهرسة البحث على الإنترنت. الحوسبة الاجتماعية تشمل تحليل الشبكة الاجتماعية، والتجمعات عبر الإنترنت، ونظم التوصية، والأنظمة المسموعة، وأسواق التنبؤ حيث الإنترنت تتضمن فهرسة البحث ISI وIEEE Xplore وScopus و Thomson Reuters وما إلى ذلك بالنظر إلى مزايا البيانات الكبيرة التي توفرها فرص جديدة في مهام معالجة المعرفة للباحثين الحاليين والمستقبليين. لكن الفرص تستتبعها دائما بعض التحديات. للتعامل مع هذه التحديات نحتاج إلى معرفة مختلف التعقيدات الحسابية، وتأمين المعلومات، وطريقة الحساب، لتحليل البيانات الكبيرة. على سبيل المثال، فإن العديد من الأساليب الإحصائية التي تعمل بشكل جيد لحجم البيانات الصغيرة لا يتم القياس عليها عند التعامل مع البيانات الكبيرة. كمثال التحديات المختلفة فى القطاع الصحي تم البحث عن أساليب مختلفة من قبل الكثير من الباحثين [Nambiar, *et al.*, 2013]. هنا يتم تصنيف تحديات تحليلات البيانات الكبيرة إلى أربعة فئات رئيسية هي تخزين البيانات وتحليلها، معرفه الاكتشافات والتعقيدات الحسابية، التدرجية وتصور البيانات، وأمن المعلومات. نناقش هذه القضايا لفترة وجيزة في الأقسام الفرعية التالية. تخزين البيانات وتحليلها في السنوات الأخيرة تضخم حجم البيانات بشكل كبير على وسائط مختلفة مثل الأجهزة المحمولة، التقنيات الحسية الجوية، والاستشعار عن بعد، وتحديد قراءات ترددات الراديو وخلافه. وعند تخزين هذه البيانات بإنفاق عالي التكلفة يتم تجاهلها أو حذفها لعدم توافر مساحة كافية لتخزينها. لذلك، فإن التحدي الأول لتحليل البيانات الكبيرة هي وسائط التخزين وارتفاع سرعة المدخلات /

الترقيم الدولي ISSN 1110-0826

المخرجات. في مثل هذه الحالات، يجب أن تكون إمكانية الوصول إلى البيانات هي الأولوية القصوى لاكتشاف المعرفة، حيث يجب الوصول إليها بسهولة وبسرعة عند الحاجة لمزيد من التحليل. في الماضي، تم استخدم وحدات الأقراص الصلبة لتخزين البيانات ولكن، فإنه أبطأ مع العشوائية في عمليات الإدخال / الإخراج. ومع ذلك فان المتاح من تقنيات التخزين لا يكفى للحصول على الأداء المطلوب لمعالجة البيانات الكبيرة. والتحدي الآخر مع تحليل البيانات الكبيرة الناتج عن تنوع البيانات. مع التزايد المستمر من مجموعات البيانات، زادت مهام التصنيف واختيار البيانات بشكل كبير. كما تعتبر أتمتة معالجة وتطوير خوارزميات جديدة للتعلم الآلي تحد اخر كبير في السنوات الأخيرة (Huang, 1997). ويعتبر تطوير تقنيات مثل hadoop وmap Reduce تجعل من الممكن لجمع كمية كبيرة من البيانات شبه منظم هيكليا او غير منظم في فترة زمنية معقولة. وتعتبر كيفية تحليل هذه البيانات بشكل فعال للحصول منها على معرفة أفضل هو التحدي الرئيسي (Das, *et al.*, 2013). في هذه الحالة يتوجب علينا توجيه المزيد من الاهتمام لتصميم أنظمة تخزين ورفع كفاءة أدوات تحليل البيانات التي توفر مخرجات تأتي من مصادر مختلفة. بالضافة لتصميم خوارزميات لتطوير أجيال من الاجهزة الذكية باستخدام تطبيقات تقنية تعلم الأجهزة لتحليل البيانات، وذلك لتحسين الكفاءة والتدرج باكتشاف المعرفة لتشمل عددا من التطبيقات الفرعية مثل المصادقة، الأرشفة، الإدارة، الحفظ، استرجاع المعلومات والتصور والتمثيل للبيانات بطرق متعددة. وكمثال على ذلك، فان مواقع التسوق عبر الإنترنت مثل سوق كوم، الأمازون، الخليج الإلكتروني لديها الملايين من المستخدمين والمليارات من البضائع لبيع كل منها شهر. هذا يولد الكثير من البيانات. تحقيقا لهذا، فان بعض الشركات تستخدم أداة – Tableau – لتصور البيانات الكبيرة. لديها القدرة لتحويل البيانات الكبيرة والمعقدة إلى صور بديهية. وذلك لمساعدة موظفي الشركة على تصور مدى ملاءمة البحث، مراقبة تحديثات العملاء، وتحليل بياناتهم ومعاملاتهم. ومع ذلك فإن أمن المعلومات في تحليل

البيانات الكبيرة التي ترتبط بكمية هائلة من البيانات، وتحليلها، واستنباط أنماط ذات مغزى منها. يستلزم من المنظمات ان تكون لديها سياسات مختلفة لحماية المعلومات الحساسة الخاصة بهم. حيث يمثل الحفاظ على المعلومات الحساسة مشكلة كبيرة في تحليل البيانات الكبيرة، وذلك لوجود مخاطر أمنية كبيرة حقيقية مرتبطة بعملية تامين البيانات الكبيرة (Zhu, *et al.*, 2015)، وهى تتطلب الكثير من التحسين المستمر، ويعتبر التحدي هو تطوير أمن متعدد المستويات مع الحفاظ على الخصوصية.

قضايا البحث المفتوح في تحليلات البيانات الكبيرة: أصبحت عملية تحليلات البيانات الكبيرة وعلم البيانات من أهم مراكز التنسيق البحثي في مختلف الصناعات والأوساط الأكاديمية. حيث ان علم البيانات يهدف إلى البحث فى البيانات الكبيرة واستخراج المعرفة منها. تشمل تطبيقات البيانات الكبيرة ومنها نمذجة عدم اليقين، تحليل البيانات غير المؤكد ،التعلم الآلي، التعلم الإحصائي، التعرف على الأنماط، تخزين البيانات، ومعالجة الإشارات. وفى هذا القسم سيتم مناقشة لقضايا البحث المتعلقة بتحليل البيانات الكبيرة مصنفة إلى ثلاث مجموعات وهي إنترنت الأشياء (IOT)، الحوسبة السحابية، الحوسبة الكمية.

إنترنت الأشياء: أعاد الإنترنت هيكلة العلاقات العالمية، حاليا، الآلات هي المسيطرة على عدد لا يحصى من الأدوات الذاتية عبر الإنترنت وإنشاء إنترنت الأشياء (IOT). وبالتالي، الأجهزة أصبحت مستخدم للإنترنت، مثل البشر مع متصفحات الويب. إنترنت الأشياء تجذب انتباه العديد من الباحثين لفرصها الواعدة وكذلك التحديات المرتبطة. وطبعا سيكون لها تأثيرات اقتصادية واجتماعية لبناء المستقبل من حيث توافر المعلومات والشبكات وتكنولوجيا الاتصالات. لقد أصبح مفهوم إنترنت الأشياء أكثر أهمية في العالم الواقعي بسبب التطوير المستمر في الأجهزة المحمولة، وتقنيات الاتصالات المضمنة والشاملة، الحوسبة السحابية، وتحليلات البيانات. وكذلك يواجه التحديات في التعامل مع الحجم الكبير

الترقيم الدولي ISSN 1110-0826

من البيانات والسرعة والتنوع. حيث ان توفر خدمة الإنترنت، وتقنية إنترنت الأشياء، تمكن الأجهزة من التواجد في عدد لا يحصى من الأماكن ويسهل تطبيقاتها. ويمثل اكتساب المعرفة من بيانات إنترنت الأشياء هو التحدي الأكبر الذى تواجه تلك التقنية. مما يستلزم التطوير المستمر للبنية التحتية لتحليل بيانات إنترنت الأشياء. ويعتبر حاليا تعلم الآلة وتقنيات الذكاء الحسابي هي الحل الوحيد للتعامل مع البيانات الكبيرة من إنترنت الأشياء (Chen, *et al.*, 2012). الشكل (٣) يصور لمحة عامة عن إنترنت الأشياء وعملية اكتشاف المعرفة.

شكل (٣): إنترنت الأشياء

الحوسبة السحابية: جعلت الحوسبة السحابية تطوير تقنيات المحاكاة الافتراضية لعمليات الحوسبة الفائقة السرعة أكثر يسرا وبأسعار معقولة. الحوسبة السحابية هي بنية تحتية مخفية في برامج المحاكاة الافتراضية. وهى أنظمة تتصرف مثل الكمبيوتر الحقيقي، ولكن مع

الترقيم الدولي ISSN 1110-0826

المرونة في تفاصيل المواصفات مثل عدد المعالجات، عدد الاقراص الصلبة والمساحات التخزينية والذاكرة ونظام التشغيل. يعتبر استخدام هذه التقنية واحد من أقوى تقنيات تهيئة بيئة عمل لتطبيقات وأدوات تحليل البيانات الكبيرة. وتتمثل تحدياتها في إدارة مجموعات البيانات، وتنوعها، والسرعة، تخزين البيانات، معالجة البيانات، وإدارة الموارد (Assuno, *et al.*, 2015). وهذه التقنية تساعد في تطوير نماذج الأعمال لجميع أنواع التطبيقات التي تتعامل مع البنية التحتية وأدوات التحليل، مما يدعم تطبيقات البيانات الكبيرة واستخدام تحليل البيانات في عمليات التنمية. تشكل البيانات الكبيرة إطاراً لمناقشة الحوسبة السحابية كخيارات متعددة، اعتمادا على الاحتياجات الخاصة، حيث يمكن للمستخدم الذهاب إلى السوق وشراء خدمات البنية التحتية من الخدمة السحابية من مزودين مثلِ Amazon and Google IBM, والبرمجيات كخدمة (ادارة العلاقات) من العديد من الشركات مثلNetSuite, Cloud9 and Job science. ميزة أخرى هي التخزين في السحابة الذي يوفر وسيلة ممكنة لتخزين كم كبير من البيانات. تحد اخر واضح هو الوقت والتكلفة اللازمة لتحميل وتنزيل البيانات الكبيرة في البيئة السحابية. ولكن تبقى الخصوصية والمخاوف المتعلقة باستضافة البيانات على الخوادم العامة، وتخزين البيانات من الدراسات المختلفة. كل هذه القضايا سوف تأخذ البيانات الكبيرة والحوسبة السحابية إلى مستوى عال من تطوير. الشكل (٤) يصور الحوسبة السحابية.

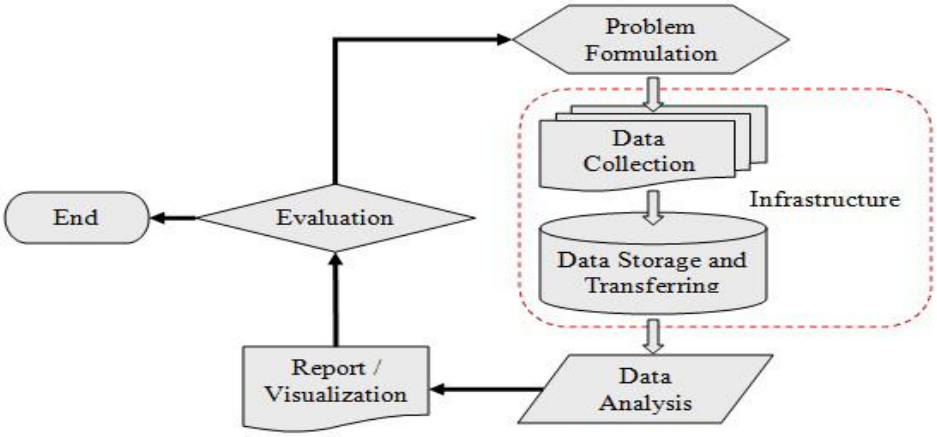
الترقيم الدولي ISSN 1110-0826

شكل (٤): الحوسبة السحابية

الحوسبة الكمية: كمبيوتر الكم لديه ذاكرة أضعافا مضاعفة أكبر من حجمها الفعلي ويمكن التعامل الأسي مع مجموعة من المدخلات في وقت واحد (Nielsen, *et al.*, 2000). ويمكن أن يكون هذا هو الحل للمشاكل الصعبة على أجهزة الكمبيوتر الحالية، وبطبيعة الحال تطبيقات تقنية البيانات الكبيرة اليوم. وتعتبر التقنية الرئيسية في كمبيوتر الكم هي وسيلة لدمج ميكانيكا الكم وتكنولوجيا معالجة المعلومات. في الكمبيوتر التقليدي، فإن المعلومات يتم تقديمها بواسطة سلاسل طويلة من bits التي تشفر ٠،١. من ناحية أخرى يستخدم كمبيوتر الكم qubit (sometimes qbit) الذي يشفر ٠،١ إلى اثنين من دوال الكم المميزة. وبالتالي، ويمكن الاستفادة من ظاهرة التراكب والتشابك. على سبيل المثال، تتطلب ١٠٠ وحدة بت في الأنظمة الكمية ٢١٠٠ من القيم المعقدة ليتم تخزينها في نظام الكمبيوتر العادي. وهذا يعني أنه يمكن حل العديد من مشاكل البيانات الكبيرة بشكل أسرع من خلال الحجم الأكبر لتوسيع نطاق أجهزة الكمبيوتر الكمي مقارنة مع أجهزة الكمبيوتر العادية.

الترقيم الدولي ISSN 1110-0826

أدوات معالجة البيانات الكبيرة: تتوفر أعداد كبيرة من الأدوات لمعالجة البيانات الكبيرة. في هذا القسم، نناقش بعض التقنيات الحالية لتحليل البيانات الكبيرة مع التركيز على ثلاث أدوات ناشئة مهمة وهي Map Reduce، Apache Spark، Storm. أكثر من تركز الأدوات المتاحة على معالجة الدُفعات ومعالجة التدفق والتحليل التفاعلي. معظم أدوات معالجة الدُفعات تعتمد على بنية Apache Hadoop مثل Apache Mahout وDryad. تستخدم تطبيقات التدفق غالبا لتحقيق واقعية الوقت التحليلي. بعض الأمثلة على منصات التدفق واسعة النطاق هي Apache Spark، Storm. وحيث ان عملية التحليل التفاعلي تسمح للمستخدمين بالتفاعل مباشرة في الوقت الحقيقي لتحقيق تحليلاتهم الخاصة. وكذلك تعتبر Apache Drill منصة للبيانات الكبيرة التي تدعم التحليل التفاعلي (Ingersoll, 2009). هذه الأدوات تساعدنا في تطوير مشاريع البيانات الكبيرة (Li *et al.*, 2012)، وتم شرح تدفق اعمال مشروع للبيانات الكبيرة في الشكل (٥).



شكل (٥): تدفق اعمال مشروع للبيانات الكبيرة

الترقيم الدولي ISSN 1110-0826

وفيما يلى شرح لبعض الأدوات المستخدمة لمعالجة البيانات الكبيرة، مع مراعاة عدم ترجمة أسماء الأدوات لتحقيق فائدة اكبر للباحثين:

- Apache Hadoop احد اكثر منصات البرمجيات استخداما لتحليل البيانات الكبيرة. وهو يتألف من نواة Hadoop، ونظام الملفات الموزعة (HDFS) . وتستخدم نموذج برمجة لمعالجة مجموعات البيانات الكبيرة يعتمد على طريقة Map Reduce. يتم تطبيق الطريقة في خطوتين هما (Map step ، Reduce Step). يعمل Hadoop على نوعين من العقد مثل العقدة الرئيسية والعقدة المنفذة. تقسم العقدة الرئيسية عملية الإدخال إلى مشاكل فرعية أصغر ثم تقوم بتوزيعها على العقد المنفذة في Map step. بعد ذلك تجمع العقدة الرئيسية النواتج لجميع المشاكل الفرعية في Reduce Step.
- Apache Mahout يهدف Apache Mahout إلى توفير تقنيات تعلم الآلة القابلة للتطوير وتطبيق تحليل البيانات الكبيرة على نطاق واسع وذكي. تعمل الخوارزميات الأساسية فى Apache Mahout، بما في ذلك التجميع، التصنيف، استخراج الأنماط، الخوارزميات التطويرية، والترشيح التعاوني القائم على الدُفعات على قمة منصة Hadoop من خلال إطار عمل Map Reduce. الهدف منApache Mahout هو بناء مجتمع ديناميكى ومتجاوب ومتنوع لتسهيل المناقشات حول المشروع وحالات الاستخدام المحتملة. الهدف الأساسي من Apache mahout هو توفير أداة لمواجهة التحديات الكبيرة. الشركات المختلفة التي طبقت خوارزميات تعلم الآلة القابلة للتطوير هي Google وIBM وAmazon وYahoo وTwitter وFace Book.
- Apache Spark يعتبر تطبيق Apache Spark إطاراً مفتوح المصدر لمعالجة البيانات الكبيرة مصمماً لمعالجة السرعة والتحليلات المتطورة. إنه سهل الاستخدام وقد تم تطويره في الأصل عام ٢٠٠٩ في UC Berkeley AMP Lab. واصبح مفتوح المصدر منذ عام ٢٠١٠ كمشروع Apache. يتيح لك Spark كتابة التطبيقات بسرعة

الترقيم الدولي ISSN 1110-0826

في java أو Scala أو python. بالإضافة إلى مخطط Map Reduce، فإنه يدعم استعلامات SQL وبيانات التدفق والتعلم الآلي ومعالجة بيانات الرسم البياني. يعمل Spark أعلى البنية الأساسية لنظام الملفات الموزعة (HDFS) الموجودة في Hadoop لتوفير وظائف محسنة وإضافية. يتكون Spark من مكونات أساسية وهي برنامج التشغيل ومدير نظام المجموعة والعقد المنفذة. يخدم برنامج التشغيل كنقطة انطلاق لتنفيذ تطبيق على نظام Spark. يخصص مدير الكتلة الموارد والعُقد المنفذة للقيام بمعالجة البيانات في شكل مهام. سيكون لكل تطبيق مجموعة من العمليات تسمى التنفيذيين المسؤولين عن تنفيذ المهام. الميزة الرئيسية هي أنه يوفر الدعم لنشر تطبيقات الشرارة في مجموعات Hadoop الحالية. أهم المميزات لـــ Apache Spark: التركيز الأساسي على مجموعات البيانات الموزعة والمرنة (RDD) ، التي تخزن البيانات في الذاكرة وتوفر التسامح مع الخطأ دون النسخ المتماثل. وهو يدعم الحساب التكراري ، ويحسن السرعة واستخدام الموارد. كما أنه بالإضافة إلى Map Reduce، فإنه يدعم أيضًا تدفق البيانات، وتعلم الآلة، وخوارزميات الرسم البياني. ميزة أخرى هي أنه يمكن للمستخدم تشغيل برنامج التطبيق بلغات مختلفة مثل Java أو R أو Python أو Scala.

- Dryad : نموذج برمجة شائع آخر لتنفيذ برامج متوازية وموزعة للتعامل مع قواعد السياق الكبيرة على الرسم البياني لتدفق البيانات. يتكون من مجموعة من عقد الحوسبة، ويستخدم المستخدم موارد كتلة الكمبيوتر لتشغيل البرنامج الخاص به بطريقة موزعة. في الواقع ، يستخدم مستخدم Dryad عدة آلاف من الأجهزة متعددة المعالجات. الميزة الرئيسية هي أن المستخدمين لا يحتاجون إلى معرفة أي شيء عن البرمجة المتزامنة. يعمل على رسم بياني موجه حسابي يتكون من رؤوس حسابية وقنوات اتصال. لذلك، يوفر Dryad عددًا كبيرًا من الوظائف، بما في ذلك إنشاء رسم بياني للوظائف، وجدولة

الترقيم الدولي ISSN 1110-0826

الآلات للعمليات المتاحة، ومعالجة إخفاقات الانتقال في المجموعة، وجمع مقاييس الأداء، وتصور المهمة، واستدعاء السياسات المعرفة من قبل المستخدم، وتحديث الرسم البياني الوظيفي ديناميكيًا استجابة لهذه القرارات الأساسية.

- Storm: هو نظام حساب الوقت الحقيقي الموزعة والتسامح مع الخطأ لمعالجة البيانات ذات الحجم الكبير. تم تصميمه خصيصاً للمعالجة في الوقت الفعلي، بالخلاف مع Hadoop والمصم هو لمعالجة الدفعات. بالإضافة إلى ذلك، من السهل أيضاً الإعداد والتشغيل، والقابلة للتوسع، والتسامح مع الأخطاء لتوفير أداء تنافسي.حيث يقوم المستخدمون بتشغيل different topologies لتنفيذ مهام Storm، بينما تطبق منصة Hadoop مخطط Map Reduce للتطبيقات المقابلة. حيث يتمثل الاختلاف الأساسي في أن Map Reduce تقلل من انتهاء المهمة في نهاية المطاف، في حين أن topologies تعالج الرسائل طوال الوقت، أو حتى ينهيها المستخدم. تتكون Storm من نوعين من العقد مثل العقدة الرئيسية والعقدة المنفذة. تقوم العقدة الرئيسية والعقدة المنفذة بتنفيذ نوعين من الأدوار مثل السحابة والمشرف على التوالي. السحابة هو المسؤول عن توزيع التعليمات البرمجية عبر Storm، وجدولة المهام وتعيينها إلى العقد العاملة، ومراقبة النظام بأكمله. يلتزم المشرف بالمهام الموكلة إليهم بواسطة السحابة. بالإضافة إلى ذلك، يبدأ وينهي العملية حسب الضرورة بناءً على تعليمات السحابة. يتم تقسيم التكنولوجيا الحاسوبية بأكملها وتوزيعها على عدد من العمليات المنفذة وتنفذ كل عملية عامل جزءاً من الهيكل.
- Apache Drill: هو نظام موزع آخر للتحليل التفاعلي للبيانات الكبيرة. يتمتع بمزيد من المرونة لدعم العديد من أنواع لغات الاستعلام وتنسيقات البيانات ومصادر البيانات. كما أنه مصمم خصيصاً لاستغلال البيانات المتداخلة. كما أنه يهدف إلى زيادة عدد الخوادم إلى عشرة آلاف خادم أو أكثر، والوصول إلى القدرة على معالجة اكسا بايتا من

الترقيم الدولي ISSN 1110-0826

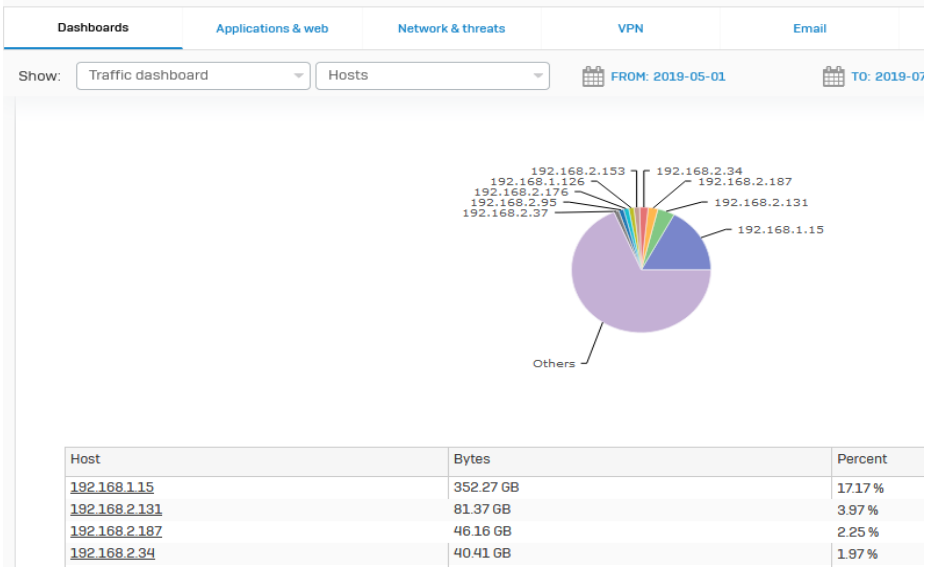
البيانات وتريليونات السجلات في ثوانٍ. ويستخدم HDFS للتخزين، و Map Reduce لإجراء تحليل الدُفعات.

- Jaspersoft حزمة Jaspersoft عبارة عن برنامج مفتوح المصدر ينتج تقارير من أعمدة قاعدة البيانات. إنها منصة تحليلية للبيانات الكبيرة قابلة للتطوير ولديها القدرة على عرض البيانات بسرعة على منصات التخزين الشائعة ، بما في ذلكMango DB وCassandra وRedis وما إلى ذلك. ومن الخصائص المهمة لـــ Jaspersoft إمكانية استكشاف البيانات الكبيرة بسرعة دون الاستخراج والتحويل والتحميل (ETL). بالإضافة إلى ذلك ، تتمتع أيضاً بقدرة على إنشاء تقارير قوية بلغة ترميز النص التشعبي (HTML) ولوحات المعلومات بشكل تفاعلي ومباشر من مخزن البيانات الضخم دون متطلبات ETL. يمكن مشاركة هذه التقارير التي تم إنشاؤها مع أي شخص داخل أو خارج مؤسسة المستخدم.
- Splunk : عبارة عن منصة ذكية في الوقت الفعلي تم تطويرها لاستغلال البيانات الكبيرة التي تم إنشاؤها بواسطة تعلم الآلة. فهو يجمع بين التقنيات السحابية الحديثة والبيانات الكبيرة. كما يساعد المستخدم في البحث عن البيانات التي أنشأها الجهاز ومراقبتها وتحليلها من خلال واجهة الويب. يتم عرض النتائج بطريقة سهلة الاستخدام مثل الرسوم البيانية والتقارير والتنبيهات. Splunk يختلف عن أدوات معالجة التدفق البيانى الأخرى، حيث تشمل خصائصه فهرسة البيانات المنظمة وغير المنظمة التي يتم إنشاؤها بواسطة تعلم الآلة، والبحث في الوقت الفعلي، واستنباط النتائج التحليلية، Dash Board. ويعتبر هدف Splunk الأهم هو توفير مقاييس معيارية للعديد من التطبيقات، وتشخيص المشكلات للبنية التحتية للنظام وتكنولوجيا المعلومات ، والدعم الذكي للعمليات التجارية.

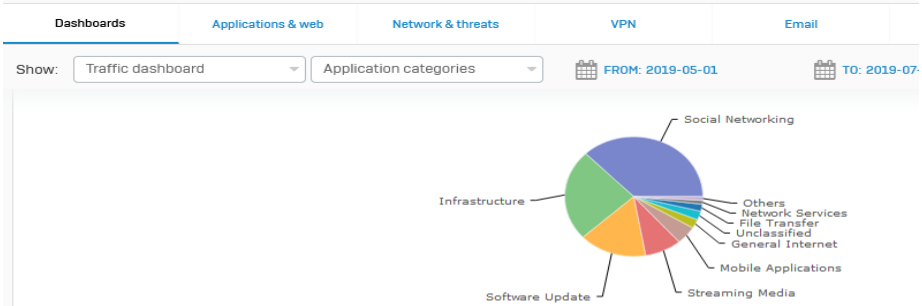
الترقيم الدولي ISSN 1110-0826

النتائج

أصبحت تكنولوجيا المعلومات توظف أدواتها لتحليل ومعالجة البيانات الكبيرة والتي تهدف الى استرجاعها من الأنظمة المختلفة في العديد من المجالات ومنها المجال الصناعي والمجال الطبي والقطاع الحكومي واقتصاد المؤسسات والمجال التعليمي. حيث تتم عملية استرجاع البيانات ومعالجتها واستخدامها بغرض تطوير المنتجات او استحداث منتج جديد، وتوفير المعلومات اللازمة لمساعدة متخذي القرار للمنتج في مراحل الإنتاج وتطوير عملية التسويق الإلكتروني وتطوير الأدوية والمساعدة في اكتشاف الأمراض وتحقيق الأمن الوطني والمساهمة في تحسين وتحقيق اهداف التنمية للدول. ومرفق فيما يلى بعض النتائج الفعلية لاستخدامات شركة تجارية بتطبيق ادوات تحليل البيانات الكبيرة لاستخداماتها المختلفة عن الفترة من ١-٥-٢٠١٩ حتى ٣١-٧-٢٠١٩ في شكل ٦،٧،٨.

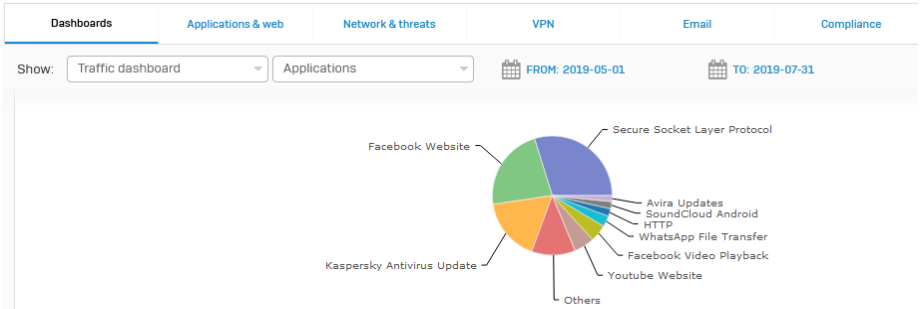


شكل (٦): تحليل للبيانات الكبيرة حسب المستخدم



Category	Bytes	Percent
Social Networking	936.89 GB	37.26 %
Infrastructure	618.48 GB	24.6 %
Software Update	400.39 GB	15.92 %
Streaming Media	216.38 GB	8.6 %
Mobile Applications	122.9 GB	4.89 %
General Internet	60.78 GB	2.42 %
Unclassified	60.08 GB	2.39 %

شكل (٧): تحليل للبيانات الكبيرة حسب نوع التطبيق المستخدم



Application/proto:port	Category	Risk	Bytes	Percent
Secure Socket Layer Protocol	Infrastructure	1	611.69 GB	29.82 %
Facebook Website	Social Networking	3	462.52 GB	22.55 %
Kaspersky Antivirus Update	Software Update	2	349.88 GB	17.06 %
Others	Unclassified	?	242.16 GB	11.35 %
Youtube Website	Streaming Media	3	113.53 GB	5.53 %
Facebook Video Playback	Streaming Media	2	95.19 GB	4.64 %
WhatsApp File Transfer	Mobile Applications	3	60.86 GB	2.97 %

شكل (٨): تحليل للبيانات الكبيرة حسب التطبيق المستخدم

الترقيم الدولي ISSN 1110-0826

مناقشة النتائج: من القراءة التحليلية لنتائج هذا البحث مصنفة حسب المستخدم، نوع التطبيق المستخدم، التطبيق المستخدم، يتبين لنا الاستخدام المتنامى لمواقع التواصل الاجتماعى في التطبيقات التجارية واعتمادية هذه المؤسسة التجارية عليها في اعمال التسويق والبيع الالكترونى الخاصة بعمليات التوسع المطرد المخطط لها من قبل هذه المؤسسة.

ملاحظات ختامية: مع تولد قناعة لدى المتخصصين ان كمية البيانات التي يتم جمعها من مختلف التطبيقات في جميع أنحاء العالم عبر مجموعة واسعة من المجالات اليوم سيتضاعف تقريبا كل عامين. فإنه لا يوجد سبب محدد لديه يمنعنا من تحليل هذا الكم المتجمع من البيانات للحصول على معلومات مفيدة تنمى المعرفة. وهذا بالطبع يستلزم تطوير التقنيات التي سيتم استخدامها لتسهيل تحليل البيانات الكبيرة.

ويعتبر التطوير المطرد في أجيال أجهزة الكمبيوتر القوية هو اكبر نعمة لتنفيذ هذه التقنيات التي تؤدي إلى تحويل البيانات إلى أدوات معرفية. وهذا لا يعني انها مهمة سهلة للحصول على بيانات عالية الأداء معالجة على نطاق واسع، بما في ذلك استغلال التوازي الحالي وإعادة هيكلة التصميم المستقبلي للكمبيوترات الفائقة القدرة للمساعدة في استخراج البيانات. علاوة على ذلك، هذه البيانات قد تنطوي على عدم اليقين في العديد من الأشكال المختلفة. وبالتالي هذا يطرح العديد من القضايا البحثية في صناعة البحوث والمجتمع في أشكال الالتقاط و الوصول إلى البيانات بشكل فعال. بالإضافة إلى ذلك، معالجة سريعة في حينها لتحقيق الأداء العالي والإنتاجية العالية، مع الوضع في الاعتبار ان تخزين ذلك كله بكفاءة للاستخدام الامثل في المستقبل قضية أخرى. كذلك، فان عمليات البرمجة اللازمة لتحليل البيانات الكبيرة هو تحد هام اخر يتوجب مجابهته [Acharjya, *et al.*, 2015]. كذلك تركيز البحث في مجال تعلم الآلة عند معالجة البيانات، تطوير خوارزميات التنفيذ،

والتحسين المستمر للعديد من هذه الأدوات التي ستحتاج إلى تغيير جذري للاعتماد عليها مستقبلا.

REFERENCES

Lynch C. (2015): Big data: How do your data grow? Nature, 455 pp.28-29.

Acharjya D. P.; Dehuri S. and Sanyal S. (2008): Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4.

Ingersoll G. (2009): Introducing apache mahout:Scalable,commercial friendly machine learning for building intelligent applications,White Paper, IBM Developer Works,1-18.

Li H.; Fox G. and Qiu J. (2012): Performance model for parallel matrix multi-plication with dryad: Data flow graph run time, Second International Con-ference on Cloud and Green Computing, 675-683.

Zhu H.; Xu Z. and Huang Y. (2015): Research on the security technology of big data information, International Conference on Information Technology and Management Innovation, pp.1041-1044.

Nielsen M. and Chuang I. (2000): Quantum Computation and Quantum Information, Cambridge University Press, New York, USA.

Assuno M.; Calheiros R.; Bianchi S.; Netto, M. and Buyya, R. (2015): Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79,3-15.

الترقيم الدولي ISSN 1110-0826

Kakhani M.; Kakhani S. and Biradar R. (2015): Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8),228-232.

Kitchin R. (2014): Big Data,new epistemologies and paradigm shifts, Big Data Society, 1(1), 1-12.

Nambiar R.; Sethi A.; Bhardwaj R. and Vargheese R. (2013): A look at challenges and opportunities of big data analytics in healthcare,IEEE International Conference on Big Data,17-22.

Rio S.; Lopez V.; Bentez J. and Herrera F. (2014):On the use of map reduce for imbalanced big data using random forest, Information Sciences, 112-137.

Das T. and Kumar P. (2013): Big data analytics: A framework for unstructured data analysis, International Journal of Engineering and Technology, 153-156.

Jin X.; Wah B.; Cheng X. and Wang Y. (2015): Significance and challenges of big data research, Big Data research, 2(2), 59-64.

Chen X. and Jin Z. (2012): Research on key technology and applications for internet of things, Physics Procedia, 561-566.

Huang Z. (1997): A fast clustering algorithm to cluster very large categorical data sets in data mining, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery.

BIG DATA ANALYTICS AND IMPORTANCE OF THIER APPLICATION IN COMMERCIAL AND INDUSTRIAL FOUNDATIONS

Ashraf A. El-Kamash[(1)]; Hassan M. Shehata[(2)]
Hoda K. Mohamed[(2)] and El Sayed M. Khater[(3)]
1) Post-Graduate Student at Institute of Environmental Studies and Research, Ain Shams University 2) Faculty of Engineering, Ain Shams University 3) National Research Center

ABSTRACT

With full reliability on modern information systems and digital echnologies, a huge amount of data is extracted every day (measured in exa byte (10^{18}) byte), resulting from the use of practical applications in everyday life such as Internet Objects and Cloud Computing. Analysis of this large amount of data (Big Data) requires a lot of effort at multiple levels to extract the knowledge needed to make the right decision at the right time. Therefore, addressing the analysis of these Big Data is an unprecedented field of research and development. The objective of this research is to explore the potential impact of Big Data challenges, to examine open research issues, and the various tools associated with them such as Hadoop, MapReduce. This research can be seen as providing a small window to explore the world of Big Data growing with tremendous acceleration in its various stages. In addition, it opens up new perspectives for researchers to develop solutions, based on open research challenges and issues. It also contributes to the dissemination of knowledge exchange for the two men to keep pace with the development in this field with non-specialized researchers.

الترقيم الدولي ISSN 1110-0826

Keywords: Big Data, Hadoop, Map Reduce, Internet Objects, Cloud computing.