البيانات الضخمة وتحليلاتها: المفهوم والخصائص والتطبيقات

أحمد خيري عبد الله علي (*)

إن العصر الذي نعيشه الآن ينتج بيانات بمعدل مذهل وغير مسبوق من حيث حجم هذه البيانات وسرعة إنتاجها، مع تنوع وتعدد المصادر التي تصدر عنها هذه البيانات وبالتالي تنوع الصيغ الناتجة عن هذا التعدد، وكل الشواهد والقوانين المطروحة من المتخصصين لا تعطينا علامة تدل على توقف أو تباطؤ هذا الإنتاج الغزيز؛ بل أن الجميع متفق على أن البيانات تتكاثر بسرعه انفجارية، وذلك أجبر المتخصصين في مجال المعلومات والحاسبات والاتصالات على البحث الدائم عن حلول جديدة ومبتكرة للتمكن من تخزين ومعالجة وتحليل وفهم هذه البيانات وبالتالي الاستفادة من كامل القيمة المرجوة – بل والقيمة الكامنة غير الموضوعة في الحسبان – من هذه البيانات، وهذه الحلول الجديدة والتي كان آخرها الحوسبة السحابية والذكاء الصناعي وتعلم الآلة وإنترنت الأشياء وغيرها، كونت ظاهرة لاقت رواجاً وذاع صيتها في السنوات الأخيرة بين الأوساط العلمية والتجارية والاجتماعية والتكنولوجية في العالم كله ، هذه الظاهرة تسمي "البيانات الضخمة Big Data".

من هذا المنطلق فإن الباحث في هذا البحث يحاول إلقاء الضوء على المفاهيم المرتبطة بالبيانات الضخمة ، وذلك من خلال التطرق إلى مجموعة من المفاهيم الأساسية عن البيانات الضخمة وإدارتها بصورة مفصلة، كمحاولة لسد الفجوة المعرفية لدي المتخصصين في مجال المكتبات والمعلومات والإدارة والعلوم الطبية والاقتصادية وغيرها من التخصصات المهتمة بهذا الموضوع كما تستهدف صناع القرار في كل المجالات لتوعيتهم بجوانب هذا الموضوع.

حيث يبدأ بتقديم سرد تاريخي لظهور مصطلح البيانات الضخمة، ثم الوقوف على التعريفات المتعددة لهذا المصطلح من واقع أدبيات الموضوع السابقة للتوقف على تعريف توافقي يجمع السمات المشتركة بما يوافق رأي الباحث، ثم التعرض لخصائص البيانات الضخمة وأنواعها ومصادرها والتقنيات المستخدمة في التعامل معها لإخراج القيمة الكامنة ، يلي ذلك عرض وافي

(*) المدرس المساعد بقسم المكتبات والمعلومات – كلية الآداب – جامعة سوهاج.
هذا البحث من رسالة الدكتوراه الخاصة بالباحث، وهي بعنوان: دور مراكز المعلومات في إدارة البيانات الضخمة: مؤسسات الرعاية الصحية في مصر نموذجًا، تحت إشراف: أ.د. شريف كامل شاهين – كلية الآداب – جامعة القاهرة & د. ناصر أبو زيد الكشكي – كلية الآداب – جامعة سوهاج

لتحليلات البيانات الضخمة big data analytics وأنواعها، وأخيرًا التعرض للمثالب والمخاوف التي تحيط هذه الظاهرة المعلوماتية.

تاريخ البيانات الضخمة:

إن تاريخ البيانات الضخمة "big data" كمصطلح قد يكون وجيزاً، إلا أن الناظر إلى "البيانات الضخمة" بصفتها ظاهرة علمية وثقافية وتكنولوجية واجتماعية يجد أن أصولها تظهر منذ فجر التاريخ البشري قبل أن تظهر الحاسبات الآلية بمفهومها الحديث بآلاف السنوات. ولتتبع أثر البيانات الضخمة يرى الباحث أننا يجب أن نتتبع نمو البيانات وسعات وسائط التخزين وطرق معالجة البيانات التي تحيط بنا والإمساك بها في قنوات تسمح بتحليلها والاستفادة منها بكميات ضخمة – وكلمة ضخمة لها في كل عصر حجم مختلف عبر مرحلتين:

المرحلة الأولي: عصر الثورة المعلوماتية Information revolution:

لا شك أن شرارة البداية لثورة المعلومات هي القدرة على تسجيل المعلومات والتي تعد أحد الخطوط الفاصلة ما بين المجتمعات المتقدمة والبدائية. فقد كان العد البسيط وقياس الطول والوزن من بين أقدم الأدوات المبهرة للحضارات القديمة وكان التدوين ضرورة لثبت هذه المقاييس رغم استخدام العلامات والرموز البسيطة كبداية، وعلى مدار القرون تطورت عملية القياس لتشمل الطول والوزن والمساحة والحجم والوقت، وكان لظهور الأرقام وعلم الرياضيات أثراً واضحاً في التعبير عن الظواهر بصيغة كمية قابلة للقياس والتسجيل.(schonberger & Cukier, 2013, p. 79)

ومع دخول الألفية الثالثة قبل الميلاد ظهرت فكرة المعلومات المسجلة بشكل ملحوظ بعد اختراع الكتابة الذي كان النقطة الفارقة التي بدأ فيها الإنسان تدوين معارفه والتقاط المعلومات لتسجيلها ومن ثم القدرة على استرجاعها مرات ومرات والاستفادة منها على مر السنوات لنفس الأشخاص أو لأجيال متعاقبة، كما سمح ذلك بنمو المعرفة البشرية وتراكم الخبرات وظهور العلوم المختلفة، وبدأ هذا الاختراع كحاجة من الحكومات لتدوين اقتصادها وجمع الضرائب وتمجيد الآلهة والحكام، وأقدم النقوش والرسوم اكتشفت مؤخرا في كهف شوفيه Chauvet بفرنسا وعمرها يزيد عن ٣٠٠٠٠ عام إلا أن اللغة المنطوقة بدأت في الهلال الخصيب عند الفراعنة والسومريون منذ ٨٠٠٠ ق.م (Fang, 1997, p. 2)، زادت بعد ذلك التاريخ الدقة في التقييس فقد استخدموا المقاييس في كل أنشطة حياتهم اليومية، وكان لتطور الكتابة في بلاد ما بين النهرين الفضل في توفير وسيلة دقيقة لمتابعة سير عمليات الإنتاج والصفقات

التجارية، وبهذا مكنت اللغة المكتوبة الحضارات القديمة من قياس الواقع وتسجيله واسترجاعه في وقت لاحق، وبتكاتف عمليتي القياس والتسجيل معاً أصبح من السهل إنتاج البيانات وكانت هذه هي الأساسات الأولي لترميز الواقع[١]"datafication" والذي يعنى تحويل الظواهر إلى الصيغة الكمية القابلة للاستخراج قيمة منها. (schonberger & Cukier, 2013, p.74)

ومع انتشار الكتابة والحساب بدأت شرارة التنوير في الظهور عن طريق ظهور حركة التأليف التي كان أو شرارة واضحة لها مع الحضارة الرومانية التي بدأت في تسيير إطلاق الشرارة الأولي في الثورة المعلوماتية الحديثة حيث أغرقت المدارس الرومانية العالم بالمؤلفات بعدما كانت العلوم تتناقل شفهياً ومن ثم تشكلت المكتبات التي تحفظ فيها هذه المعارف والتي كان أشهرها مكتبة الإسكندرية القديمة التي جمعت الكتابات من كل اللغات ومن كل البقاع القديمة. (Fang, 1997, p. 15)

وكان لانتشار الأرقام العربية التي اخترعت في الهند ثم انتقلت إلى بلاد فارس ثم منها إلى العرب الذين أحدثوا فيها تطورات بالغة الأهمية أثراً واضحا في تطور العمليات الحسابية وظهور عمليات أكثر تعقيداً مقارنة بالأرقام الرومانية التي كانت تحجم وتعيق التطوير في العلوم الرياضية (King, 2004, p. 22)، وانتشار وتطور العلوم الرياضية قدم لنا طريقة رائعة ومنظور جديد للبيانات حيث أصبح بالإمكان تحليلها في صورة كمية وليس فقط تخزينها واسترجاعها، وكان لنظام إقفال الدفاتر المحاسبية وبالأخص نظام القيد المزدوج double-entry البداية الحقيقية لوضع الحقائق الرقمية في جداول والعمل على تحليلها واستخراج الحقائق منها، وهذا النظام المحاسبي ظهر منذ ١٣٤٠م في مدينة جنوا الإيطالية (Brown, 2006, p. 99)

وفي القرن التاسع عشر كان الحماس لفهم الطبيعة من مفهوم كمي كفيلاً بتوضيح العلوم. واخترع الباحثون أدوات ووحدات جديدة للقياس وتسجيل التيار الكهربي والضغط الجوي ودرجة الحرارة وتردد الصوت وغيرها، وبذلك كانت عملية تحويل الظواهر والأحداث للصورة الكمية أكثر سهولة رغم أن التعامل مع الكميات الضخمة من البيانات في العصر التناظري كانت مكلفة ومستهلكة للوقت. وفي كثير من الحالات كانت تتطلب صبراً لا حصر له، أو

[١] صاغ الباحث هذا المقابل "ترميز الواقع" لعدم توافر مقابل للمصطلح "datafication" في اللغة العربية علي حد علم الباحث.

على الأقل تكريس لحياة طويلة، مثلما حدث مع تيكو براهي Tycho Brahe عالم الفلك في ملاحظة النجوم والكواكب في القرن السادس عشر الميلادي ، (Schönberger & Cukier 2013, p. 82)

وكانت الثورة المدوية في عالم المعلومات هي الطباعة التي ابتكرها الكوريون لأول مرة ١٢٣٤م وطورها الألماني يوهان جوتنبرج ١٤٤٧م إلى ما يعرف بالطباعة بالحروف المتحركة ويقال أن الصينيين سبقوه في ذلك إلا أن هذا الاختراع لم يصل إلى أوربا، ورغم ذلك لا ننكر أن ابتكار الورق الذي بدأه الصين تدين له الطباعة بالفضل لأن الطباعة اعتمدت على الحبر والورق الذي كان للصين الفضل الكبير لتطويرهما (S.P.C.K., 1855, p. 15)، ونتج عن سهولة الاستنساخ للمعلومات المكتوبة ورخص ثمنه انتشار المواد المطبوعة وتنامي المجموعات داخل المكتبات بكافة أنواعها وازدهار حركة النشر للمعلومات بكل اللغات، حيث

وكان لظهور وسائل الإعلام الجماهيرية mass media التي بدأت بالراديو عام ١٩٠٦م وتوالت بعد ذلك المخترعات والمكتشفات التي ما لبث أن حققت ثورة اتصالية وشكلت نقلة نوعية كبيرة في وسائل الاتصال الإنساني من خلال ظهور التلغراف والتليفون، ثم التصوير الفوتوغرافي فالفيلم السينمائي ثم التليفزيون وصولا إلى الأقمار الصناعية والفاكس والفيديو والإنترنت والهاتف الخلوي وغير ذلك من وسائل الاتصال والإعلام التقليدية

ظهر الحاسب الآلي في ثلاثينيات القرن العشرين بعد محاولات عديدة ومختلفة الأفكار وكان أشهر هذه الحاسبات الذي اخترعه العالم الألماني كونراد تسوزه Konrad Zuse والذي أسماه Z1 وطور منه أجيال وصلت إلى Z3 الذي كان أول كمبيوتر رقمي تطبيقي عرفه العالم (Zuse, 2013, p. 32)، والجهاز الذي يدعى "هارفرد مارك ١" والذي كان من اختراع عالم الرياضيات الأمريكي "هوارد إيكن"، وكان يتمتع بما يكفي من الدقة في إجراء العمليات الحسابية وصولاً إلى المنزلة ٢٣ بعد الفاصلة العشرية (Norman, 2005, p. 481) وتتالت الأجيال الحاسوبية إلى أن وصلت لما نعاصره من قدرات هائلة في المعالجة والتخزين.

بدأت شبكة المعلومات الدولية الإنترنت في أواخر ستينات القرن العشرين حينما كلفت الحكومة الأمريكية شركة rand بعمل وسيلة تضمن استمرار الاتصالات بين السلطات الأمريكية في حالة حدوث حرب نووية فقامت ببناء أول شبكة مكونة من أربع حاسبات مرتبطة بأربع جامعات تحت إشراف وزارة

الدفاع الأمريكية وسميت هذه الشبكة "إربانت" والتي تحمل اسم وكالة مشاريع الأبحاث المتقدمة بوزارة الدفاع (ARPA) ،بعدها تم ربط ٧٢ جامعة ومركز بحوث عام ١٩٧٢م وكانت جميعها تعمل على أبحاث تخص وزارة الدفاع، ثم أخذت هذه الشبكة في النمو بمعدل حاسب آلي كل ٢٠ يوم، ثم أخذت شبكات أخرى في الظهور مثل ALOHA net في جامعة هاواي وbbn النسخة التجارية لأرباو Teletel في فرنسا NSFNET التي ربطت استراليا وألماني وإسرائيل وإيطاليا واليابان والمكسيك وهولندا kوعمد الباحثين إلى الربط بين الشبكات المختلفة إلى أن تكونت شبكة "الإنترنت" وأصبح لها مرتادين من جميع أنحاء العالم (Norman, 2005, p. 799). وبفضل شبكة الانترنت الذي تطور ليعتمد في شبكاته على الألياف الضوئية وأشعة الليزر والأقمار الصناعية لتنتج نظام الاتصال الرقمي الذي أنجب عصرا ومجتمعا جديداً أطلق علية اسم عصر أو مجتمع المعلومات حيث حدث تقارب بين البشر والأمم إلى حد التفاعل الشديد والسريع الذي أدى إلى اندماج ثقافي وحضاري وبفضل التكنولوجيا تحول العالم شاسع الأركان إلى قرية صغيرة يمكن سماع ومشاهدة أي حادثة به بالصورة والصوت في نفس اللحظة التي تجري بها أو بعدها بثواني قليلة (عامر، ٢٠١١، صفحة ١٢)

كان انتشار الحاسب الآلي وحركة الرقمنة التي حولت الأصول المعلوماتية إلى الصيغة الرقمية ثم إصدار الأصول الجديدة في صورة رقمية هي شرارة البداية لانفجار المعلومات وعدم السيطرة عليها، حيث انتشرت الأصول الرقمية بصورة كبيرة وأصبحت شبكة الإنترنت، فقد توقع فريمونت رايدر أمين المكتبة في جامعة ويسليان في كتابه عن مستقبل المكتبات البحثية أن مكتبات الجامعات الأمريكية سوف تتضاعف في حجمها كل سته عشر عاماً ووفق ذلك توقع أن مكتبة جانعة ييل Yale عام ٢٠٤٠م سوف يصبح بها ٢٠٠ مليون مجلد يشغل أرفف تمتد لـــ ٦٠٠ ميل وهذا يتطلب جيش قوامه أكثر من ٦٠٠٠ مفهرس (Rider, 1944, p. 32)، كما أطلق ديرك رايس في الستينات قانون أسماه "قانون الزيادة الأسية"",law of exponential increase" مفاده أن المجلات الجديدة نمت نموا مطردا وليس خطياً حيث تتضاعف كل نصف قرن، وقد وصل لهذا المعدل اعتمادا على مؤشر النمو في المجلات العلمية والصحف، وربط التقدم في العلم بعدد المواليد في العالم (Price, 1961, p. 69).

وفي أواخر الستينات من القرن العشرين بدأ التحدث عن ضرورة تحسين وسائط التخزين الرقمي وتقليل حجم المساحات لأقل حجم ممكن، حيث نشر

ب.أ.د. دي مان و ب. أ. مارون مقال عن ضغط الملفات الرقمية لملائمة السعات التخزينية المحدودة وقتها وأيضاً ملائمة معدل نقل الملفات البطيء الإمكانات المحدودة لمعالجات لحاسبات الآلية (Marron & de Maine, 1967)

المرحلة الثانية: عصر البيانات الضخمة Big Data:

منذ السبعينات تحولت الكتابات من التحدث عن المعلومات إلى حجم أصغر وهو وحدات البيانات data bits والذي يعتبره الباحث بداية التحول الحقيقي إلى الاهتمام بالبيانات الخام قبل تحولها إلى معلومة واستخراج القيمة الكامنة value في تلك لبيانات، فنجد آرثر ميللر يتحدث عن التعدي على الخصوصيات قائلاً "يبدو أن الكثير ممن يتعاملون مع البيانات يقيسون الأشخاص بما يتوافر عن كل منهم من بيانات داخل ملف حاسب آلي" (Miller, 1972, p. 26)، وفى الثمانينات بدأت وزارة الاتصالات والبريد اليابانية في قياس تفق المعلومات وحجمها في اليابان بدلالة عدد الكلمات وكان من النتائج التي وصلوا لها أن حجم البيانات الجديدة المنتجة يفوق بكثير مدي الاستفادة المأخوذة من هذه المعلومات، وكان من أهم توصياتها تفعيل الطلب على المعلومات التي تعتمد على الاتصال عن بعد وتفعيل التغذية المرتدة التي تحد من تقادم المعلومات وأشارت إلى أن كل شخص يحتاج معلومات تختلف عن الآخرين لذلك يجب الاهتمام بالمعلومات الصغيرة التي تخدم هذه الاحتياجات الفردية ، كما قام مكتب الإحصاء المجري مشروع بحثي لتنمية صناعة المعلومات في المجر ومن ضمن أنشطة هذا المشروع قياس حجم المعلومات بالبت Bit وهذا الإحصاء مستمر إلى يومنا هذا (Hilbert, 2012, p. 1024)

وفي ١٩٨٠م قام "تجومسلاند I.A. Tjomsland" بإعطاء ندوة تحت رعاية IEEE عن أنظمة التخزين الضخمة للبيانات، وقال فيها أن المؤسسات تنشئ كمية كبيرة من البيانات والكثير منها يتقادم ولا تستطيع المؤسسات عزل المتقادم عن المفيد وقال "إن الضرر الواقع من محو البيانات المفيدة أكثر بكثير من تضررنا من ترك البيانات التي لا فائدة منها"(Tjomsland, 1980) وكان ذلك بمثابة إشارة مباشرة إلى قيمة البيانات الضخمة.

في التسعينات نشر بيتر ج. ودينغ مقال بعنوان "saving all bits" يقول فيه أن الاحتفاظ بكل ما لدينا من معلومات شيء لا بديل عنه وهذا يحيلنا إلى

وضع لا نحسد عليه في المستقبل لأن حجم البيانات يفوق حجم الشبكات ووسائط التخزين وقدرة نظم الاسترجاع، فضلا عن مقدرتنا كبشر على فهم هذه البيانات والاستفادة منها، لذا يجب تصميم الآت تستطيع التحكم في تدفق البيانات وتدقيق التسجيلات داخل قواعد البيانات كما يمكن أن تقوم بالتنبؤ وفهم الأنماط التي تشكلها هذه البيانات وبذلك نستطيع تقليل أضرار إهمال المعلومات القيمة التي نملكها واكتشاف معلومات قيمة لم نلقي لها بالاً (denning, 1990, p. 404) وتعد هذه المقالة إشارة صريحة وتنبؤ بصير ببرامج إدارة البيانات الضخمة.

في عام ١٩٩٧م نشر مايكل كوكس وديفيد السوورث بحث في مؤتمر IEEEالثامن عن "التطبيقات المتحكمة في طلب الصور التخيلية" "Application-controlled demand paging for out-of-core visualization" أن هذه البرمجيات تضغط بشكل كبير وتتحدي أنظمة التشغيل وسعات التخزين، وأسميا هذه المشكلة "مشكلة البيانات الضخمة the problem of big data" وكانت هذه هي المرة الأولى الذي يذكر فيه مصطلح البيانات الضخمة في بحث علمي، وعرفا هذه المشكلة: " حينما تكون مجموعات البيانات لا يستطيع المعالج أو وحدات التخزين مواكبتها" وقالا أن الحل الأفضل هو إيجاد موارد بديلة (Cox & Ellsworth, 1997, p. 236).

في ١٩٩٩م نشر ستيف برستون وديفيد نيوارت ومايكل كونكس مقالاً عن تصفح البيانات الصورية المرئية لمجموعات البيانات التي تزيد حجمها عن جيجابايت في الوقت الحقيقي Visually exploring gigabyte data sets in real time وكان أحد أقسام هذا البحث يحمل عنوان "Big Data for Scientific Visualization" واستهل هذا البحث بعبارة "إن القدرات الحاسوبية العالية نعمة كبيرة في العديد من المجالات إلا أنها أيضاً نقمة بلا شك حيث تمطرنا بكمية هائلة من البيانات" وأشار أن استهلاك الفرد الواحد من البيانات يتعدى ٣٠٠ جيجا بايت، والتحدي الأكبر الآن هو فهم هذا الكم الضخم من البيانات وأنماطها وليس فقط تجميع إحصاءات إجمالية عنها (Bryson, Kenwright, Cox, Ellsworth , & Haimes, 1999).

في نوفمبر ٢٠٠٠م قدم فرانسيس اكس ديبولد Francis X. Diebold ورقة بحثية في المؤتمر العالمي الثامن لمجتمع الاقتصاد بعنوان "البيانات الضخمة: نموذج حيوي للقياس والتنبؤ بالاقتصاد الكلي" صرح فيها أن هناك

علوم فيزيائية واجتماعية وبيولوجية جديدة استفادت أو واجهت تحدي وعقبة هي "ظاهرة البيانات الضخمة Big Data" phenomenon" وعرف هذه الظاهرة بأنها " انفجار في كمية (وأحياناً في جودة) البيانات المتاحة والمحتملة ذات الصلة والتي نتجت جراء التطورات غير المسبوقة في تسجيل البيانات وتكنولوجيا التخزين" (Diebold, 2000).

في ٢٠٠١م قام "دوغ لاني" محلل مجموعة (META Group المعروفة الآن باسم Gartner) في تقرير بحثي وعدد من المحاضرات المتعلقة به بتعريف تحديات نمو البيانات وفرصها كعنصر ثلاثي الأبعاد، بمعني زيادة الحجم (كمية البيانات)، السرعة (سرعة البيانات الصادرة والواردة) والتنوع (تنوع أنواع البيانات ومصادرها)، وتقوم Gartner والكثير من الشركات في هذه الصناعة الآن بالاستمرار في استخدام نموذج "the 3Vs" لوصف البيانات الضخمة وأصبح هذا المصطلح هو الأشهر في وصف البيانات الضخمة بالرغم من أن هذا المقال لم يذكر على الإطلاق لفظ البيانات الضخمة big data (Laney, 2001).

بعد كل تلك الدراسات أصبحت البيانات الضخمة كمفهوم ومشكلة معروفة للجميع بينما اقتصرت حلول التحكم في البيانات الضخمة وإدارتها على تعظيم إمكانات المعالجة والتخزين للحاسبات الآلية والمعروف بالترقية Scale up، وهذا ما دفع شركة أوراكل إلى إنتاج الـ Database appliance و أسموه Exadataوهو مجموعة أجهزة في حاوية واحدة بقدرات كبيرة، لكن بقيت البيانات مخزنة في سيرفر واحد و هو جهاز بسعر مرتفع قد يصل ثمنه إلى ٥٠٠ الف دولار وقتها علاوة على اعتمادها على نظم إدارة البيانات التقليدية التي تعتمد علي النظام العلائقي لقواعد البيانات (Trauvitch, n.d)

رغم ذلك كانت الحدود المفروضة على حجم مجموعات البيانات الملائمة للمعالجة في مدة معقولة من الوقت خاضعة لوحدة قياس البيانات اكسابايت exabyte وهي تعادل1000بيتابايت أو 2^{60}بايت، ومع انتشار الهواتف المحمولة الذكية أضافت تطبيقاتها المتنوعة إمكانية التقاط بيانات المستفيدين ومن ثم معرفة الأنشطة اليومية والشبكة الاجتماعية التي يهاتفها ويبادلها الرسائل، بل والأماكن التي يحب أن يتواجد فيها والألعاب التي يفضلها والمواقع التي يتصفحها من هاتفه، كما أمكن عن طريق تطبيقات الهواتف الذكية التعرف على العادات الغذائية والرياضية وأنماط النوم والأمراض التي يتداوى منها

والعقاقير التي يستخدمها والكتب التي يقرأها في أوقات فراغة وانتظاره في المواصلات وغير ذلك الكثير (Adey, 2010, p. 129).

علاوة علي أن العلماء يواجهون عددا من القيود بسبب مجموعات البيانات الضخمة الموجودة في العديد من المجالات غير البيانات الصادرة عن أنشطة البشر، والتي تتضمن الأرصاد الجوية، وعلم الجينات، والمحاكاة الفيزيائية المعقدة والبحوث البيولوجية والبيئية، وتؤثر القيود أيضاً علي بحث الإنترنت، وتقنية الأعمال التجارية والتمويل، وتنمو مجموعات البيانات في الحجم بشكل جزئي، ويرجع ذلك لأنها يتم جمعها بشكل متزايد عن طريق أجهزة استشعار المعلومات المتنقلة، والتقنيات الحسية الجوية (الاستشعار عن بعد)، وسجلات البرامج، والكاميرات، والميكروفونات، وأجهزة تحديد ذبذبات الإرسال وشبكات الاستشعار اللاسلكية، وتضاعفت القدرة التكنولوجية العالمية لتخزين المعلومات للفرد الواحد تقريباً كل ٤٠ شهر من الثمانينات، واعتباراً من عام ٢٠١٢ كان العالم ينشئ ما يقرب من ٢.٥ كوينتيليون بايت (٢.٥ ×18 10) من البيانات يوميا(Laney, 2001) .

في ٢٠٠٣م نشرت جوجل بحثاً بعنوان "Google File System نظام جوجل لتوزيع الملفات" وهو طريقة استحدثتها جوجل لتطوير محرك البحث الخاص بها مكتوبه بلغة الجافا، وتستخدم للتحكم في كيفية تخزين واسترجاع وتنظيم وإدارة ملفات الحاسوب والبيانات التي تحتوي عليها تلك الملفات لتسهيل إيجادها واستخدامها، وأكبر ميزة في هذه الطريقة اعتمادها على الشبكات والأجهزة رخيصة الثمن (Ghemawat, Gobioff, & Leung, 2003)، وفي ٢٠٠٤م نشرت جوجل بحثا بعنوان :"MapReduce Simplified Data Processing on Large Clusters"، ومابريديوس MapReduceهو نموذج برمجة وضعته جوجل على أساس Google File Systemويعمل هذا النموذج البرمجي على تجهيز عدد كبير من مجموعات البيانات، ويستخدم مجموعة واسعة النطاق لأداء المهام والعمليات الموازية تلقائياً (Dean & Ghemawat, 2004).

عام ٢٠٠٥م ولد عملاق البيانات الضخمة هادووب Hadoop، أحد مشاريع مؤسسة أباتشي الرائدة، وهو إطار عمل برمجي مفتوح المصدر مكتوب بلغة جافا لمعالجة للبيانات الموزعة distributed processing dataوالعمل على عدة حواسيب أو عناقيد Clusters في نفس الوقت لمعالجة

البيانات الضخمة وهو مشتق من معمارية MapReduce الخاصة بجوجل وأبحاث نظام ملفات جوجلGFS.وبدء مشروع هادوب من طرف دوج كاتينج Doug Cutting ومايك كافاريلا Mike Cafarella وقتما كانا يعملا في شركة ياهو، وقد اختار دوج اسم "Hadoop" وشعاره ذلك الفيل الظريف من اسم دمية ابنه الصغير التي على شكل فيل! وقد كانت عملية التطوير نابعة في الأساس لدعم مشروع محرك البحث Nutch. (Woodie, 2015)

ومع القدرات العالية على تحليل البيانات ورخص ثمن العتاد اللازم لعمل هذه التجهيزات تلى ظهور تقنيات البيانات الضخمة سرب من التطبيقات والبرمجيات التي تقوم بتحليلات البيانات الضخمة "big data analysis" التي تستفيد من تعلم الآلة والذكاء الاصطناعي والحوسبة السحابية لفهم حزم البيانات في شكلها الخام "raw data" وعدم الحاجة إلى نظم إدارة قواعد البيانات التقليدية التي تتطلب إدخال البيانات وتنظيمها أولاً لتستطيع استرجاع المعلومات منها؛ بل أمكن لبرمجيات إدارة البيانات الضخمة الاستفادة من البيانات الخام وتحليلها في نفس اللحظة التي تنتج فيها في الوقت الحقيقي on time والاستفادة من هذه المعلومات والرؤى الطازجة في اتخاذ القرار السليم في الوقت المناسب بل والتنبؤ من واقع الأنماط المتكررة في البيانات بما سيحدث في المستقبل وبهذا يمكننا اتخاذ قرارات احترازية لتجنب المخاطر وزيادة المنافع (Ohlhor, 2013, p. 4).

ومع انتشار الأقمار الصناعية وأجهزة التصوير الرقمي وأجهزة الاستشعار والمجسات الرقمية في المصانع والمراصد والمتاجر وحتى في الشوارع وإشارات المرور، نتج عن ذلك كمية ضخمة من البيانات الخام عن الظواهر الطبيعية والآلات والإحداثيات وتحركات الأشخاص والمركبات ودرجات الحرارة وعن كل شيء في هذا العالم بداية من حركة المجرات والأجرام السماوية إلى الحمض النووي داخل المجين البشري وتحرك الذرات داخل المواد (Schönberger &Cukier, 2013. p75) ، وأدي ظهور إنترنت الأشياء Internet Of Things إلى ربط هذه المصادر البياناتية data sources عن طريق الشبكات مما جعل الحصول على البيانات الصادرة عنها آني، وبهذا أصبح من السهل التقاط وتخزين معالجة هذه البيانات ومن ثم عمل تحليلات لهذه البيانات الضخمة وسرعة اتخاذ القرار، فمثلا يمكننا معرفة مدي التلف في

محرك أو أحد أجزاء السيارات التي تسير على الطريق وتوقع الأعطال وتصليحها قبل حدوثها.

أما على مستوى الحكومات فقد لاقت البيانات الضخمة إهتماماً كبيراً ففي مارس ٢٠١٢، أعلن البيت الأبيض عن "مبادرة البيانات الضخمة" القومية التي تتألف من ٦ إدارات ووكالات فيدرالية تودع أكثر من ٢٠٠ مليون دولار لمشاريع البيانات الضخمة البحثية. (The White House، ٢٠١٢)

وقد تضمنت المبادرة" National Science Foundation بعثات في الحوسبة" والتي منحت ١٠ مليون دولار علي مدي ٥ سنوات لمعمل AMPLab، كما تلقي AMPLab أيضاً تمويل منDARPA ، وأكثر من اثني عشر راعياً صناعياً ويستخدم البيانات الضخمة لمواجهة مجموعة واسعة من المشاكل بدءاً من الاختناقات المرورية وحتي مكافحة السرطان . (National Science Foundation, n.d)

وشملت مبادرة البيت الأبيض أيضاً التزاماً من وزارة الطاقة لتوفير ٢٥ مليون دولار علي مدار ٥ سنوات لإنشاء معهد إدارة وتحليل وتصور البيانات (SDAV)، والذي يتم قيادته من قبل معمل لورانس بيركلي الوطني التابع لوزارة الطاقة .ويهدف معهد SDAV جمع الخبرات من ٦ مختبرات وطنية و٧ جامعات لتطوير أدوات جديدة لمساعدة العلماء في إدارة وتصور البيانات علي أجهزة الكمبيوتر العملاقة الخاصة بالإدارة.

هذا وقد أعلنت ولاية ماساشوستس الأمريكية عن مبادرة ماساشوستس للبيانات الضخمة في مايو٢٠١٢، والتي توفر التمويل من حكومة الولاية وشركات القطاع الخاص لمجموعة متنوعة من المؤسسات البحثية، وقد استضاف معهد ماساشوستس للتكنولوجيا مركز إنتل للعلوم والتكنولوجيا الخاص بالبيانات الضخمة في مختبر MIT لعلوم الكمبيوتر والذكاء الاصطناعي (csail, 2013) .

وتقوم المفوضية الأوروبية على مدار عامين بتمويل منتدى القطاعين العام والخاص للبيانات الضخمة من خلال برنامجهم السابع لإشراك الشركات والأكاديميات وغيرهم من أصحاب المصلحة في مناقشة قضايا البيانات الضخمة .ويهدف المشروع إلى تحديد استراتيجية خاصة بالبحث والابتكار لتوجيه إجراءات الدعم من المفوضية الأوروبية للتنفيذ الناجح لاقتصاد البيانات

الضخمة. وسوف تستخدم نتائج هذا المشروع كمدخل لمشروعهم التالي Horizon 2020 (European Commission–CORDIS, 2012)

وبذلك أصبحت هذه التقنية الناشئة مسيطرة على عصرنا الحالي لدرجة أن العصر الذي نعاصره الآن أصبح يطلق عليه عصر البيانات الضخمة، ووفقا لجارتنر، وهي شركة رائدة في مجال أبحاث تكنولوجيا المعلومات والاستشارات، فإن البيانات الضخمة تمر الآن بذروة مرحلة التضخم حيث وصلت عوائد مبيعات البيانات الضخمة عام ٢٠١٢م إلى إحدى عشر مليار وستمائة مليون دولار، على أن تقفز إلى خمسون مليون دولار في العام٢٠١٧م، ومن المتوقع نمو سوق تكنولوجيا البيانات الضخمة والخدمات المتعلقة بها بمعدل أسرع بحوالي سبع مرات من سوق تكنولوجيا المعلومات والاتصالات ككل ويتوقع أن تصل إلى القمة الإنتاجية خلال نحو خمس إلى عشر سنوات، أي بعد بضع سنوات من الحوسبة السحابية. (Gartner inc., 2015)

تعريف البيانات الضخمة:

على الرغم من هذا الصيت المدوي للبيانات الضخمة في الأوساط التجارية والأكاديمية والصناعية إلا أن المصطلح لا يزال يحيط به الكثير من الغموض المفاهيمي، ونجد أن استخدامات هذا المصطلح تختلف من باحث لآخر بدرجة أصبحت غير متناسقة مما يزيد الأمر سوءا ويمنع مبدأ تراكم العلم وتطوير الموضوع نفسه، فمن الباحثين يذهب إلى تطور الإمكانات التكنولوجية والحاسوبية، وآخرين ينظرون له بأنه تضخم في حجم البيانات بينما يذهب البعض إلى أنها ظاهرة أثرت على المجتمع والعلم حيث غيرت المفاهيم والثوابت وعدلت في الثقافة والفكر (De Mauro, 2015, p. 97)، لذا كان من الضروري عند تناولنا هذا الموضوع من تحليل التعريفات السابقة والموجودة في أدب الموضوع المنشور عن البيانات الضخمة بغية لوصول إلى تعريف توافقي عن طريق تجميع السمات المشتركة بين هذه التعريفات ووجود مثل هذا المصطلح سيعمل على التكامل بين البحوث ومن ثم تكوين صورة ناضجة ومتماسكة، حيث ترى كل من روندا بوبو وجاراس مارتن أن التوافق في الآراء التي يبديها المجتمع العلمي المنتمي لأحد التخصصات لتعريف المفاهيم يمكن استخدامه كمقياس لتقدم ونضج هذا التخصص (Ronda–Pupo, 2012, p. 188).

وسوف يقوم الباحث بعرض التعريفات في مجموعات متشابهة معاً في وجهة النظر المتبناة في التعريف كالتالي:

- **المجموعة الأولى من التعريفات:**

تركز على السمات الرئيسة المميزة للبيانات الضخمة وهذه التعريفات هي الأكثر شهرة وانتشاراً بين جميع الأوساط، وتستقي هذه السمات من المعوقات التي واجهتها الشركات والمنظمات عند تعاملها مع البيانات خاصة منذ بداية الألفية الثانية للميلاد جراء ظهور التجارة الإلكترونية، ومن أمثلة هذه التعريفات تعريف دوج لاني (Laney, 2001) الذي يعرف البيانات الضخمة عن طريق وصفها بأنها" تلك البيانات التي تمتلك أبعاد ثلاثة هي الحجم Volume الضخم، والسرعة variety في النشأة، والتنوع Velocity في المصادر والصيغ، كما أنه ذكر في التعريف الحاجة الى ممارسات جديدة تعني بالحلول المعمارية والتفضيلات التي تؤثر على اتخاذ القرار" ، وهذا التعريف رغم أنه لم يذكر في عنوانه أو عناوين الأقسام لفظ البيانات الضخمة إلا أنه أفرز التعريف الأكثر شهرة للبيانات الضخمة والمشهور بـــ"the 3 Vs" والذي يستخدمه الجميع حتى هذا الوقت، وشهرة هذا التعريف جعلت بعض التعريفات التي تلته تتبناه مع إضافة سمات أخري للبيانات الضخمة مثل تعريف ديجكس (Dijcks, 2013) الذي أضاف القيمة Value والتي يقصد بها استخدام البيانات لمرات عديدة لاستخراج القيمة الكامنة، وتعريف شرويك وآخرون (Schroeck, Shockley, Smart, Morales, & Tufano, 2012, p. 20) الذي أضاف الصدق Veracity ويقصد به إمكانية استخراج البيانات القيمة من بين الأكوام الخثة والتي يسميها 'dirty data'، وأضافت شركة انتل (Intel, Big Data Analytics:Intel's IT Manager Survey on How Organizations Are Using Big Data, 2012) التعقيد Complex واللا هيكلية unstructured، وأخيراً وليس آخراً نجد اندربال فاندر (Bhandar, 2013) يضيف لتلك الخصائص القابلية للتطاير Volatility وانتهاء الصلاحية Validity ويركز على أهمية الاستغلال الفوري للبيانات.

- **المجموعة الثانية من التعريفات:**

ركزت هذه التعريفات على الاحتياجات التكنولوجية اللازمة لإتمام عملية المعالجة لمجموعة من البيانات الضخمة، فنجد شركة مايكروسوفت (Microsoft، ٢٠١٧) تعرف البيانات الضخمة بأنها "تطبيق قوة تكنولوجية

عظيمة كان آخرها الذكاء الصناعي وتعلم الآلة على مجموعات ضخمة وأحياناً معقدة من المعلومات"، ويقع تعريف المعهد الوطني للمعايير NIST ضمن هذه المجموعة حيث يعرف البيانات الضخمة على أنها "بيانات تتطلب تحويلها إلى بنية قابلة للتخزين وتسمح بإجراء معالجات وتحليلات بارعة وفائقة عليها"، كما تذكر أيضاً تعريف آخر لبنية البيانات الضخمة على أنها "توزيع عمل نظم إدارة البيانات على نقاط مستقلة ومرتبطة معاً لتحقيق التوسع في الإمكانات اللازم لمعالجة هذه الكميات الكبيرة من البيانات بكفاءة" (NIST, 2015, pp. 9–12)، كما تعرفها شركة " (IBM) "تنشأ البيانات الضخمة عن طريق كل شيء من حولنا وفي كل الأوقات كل عملية رقمية وكل تبادل في وسائل التواصل الاجتماعي ينتج لنا البيانات الضخمة، تتناقلها الأنظمة، وأجهزة الاستشعار، والأجهزة النقالة البيانات الضخمة لها مصادر متعددة في السرعة والحجم والتنوع ولكي نستخرج منفعة معنوية من البيانات الضخمة نحتاج إلى معالجة مثالية، وقدرات تحليلية، ومهارات" (IBM, n.d)، وهناك تعريفات أخري في نفس المجموعة تشير لأنها بيانات تخطت الإمكانات التكنولوجية التقليدية وتحتاج لتطوير إمكاناتنا وطرقنا مثل تعريف إد دمبل (Dumbill, 2012, p. 9) الذي يقول" إننا نطلق على البيانات أنها ضخمة حينما تتجاوز قدرة قواعد البيانات التقليدية ونظم إدارتها وبذلك تتطلب نظم بديلة لحل هذه المشكلة"، بينما يشير فيشر (Fisher, 2012, p. 52) بأن الحجم الذي يمكننا القول عنه أنه ضخم يزيد وفق قانون مور وليس حجماً ثابتا مع الزمن ويرتبط هذا الحجم بالقدرات التخزينية التجارية، حيث نحكم على مجموعة بيانات بأنها ضخمة إذا لم يستوعبها القرص الصلب لأفضل جهاز حاسب آلي واحد وبالتالي نضطر إلى تخزينها على عدة أقراص مختلفة.

- **المجموعة الثالثة من التعريفات:**

تركز هذه المجموعة على البيانات الضخمة بصفتها ظاهرة تكنولوجية وثقافية أثرت على المجتمع والطريقة التي نحل بها المشكلات، فتري بويد وكراوفورد أن "السمة الأهم في البيانات الضخمة كونها أكبر من قدرتنا على البحث والتنظيم والإدراك الكلي" وتصف البيانات الضخمة بأنها "ظاهرة ثقافية وتكنولوجية وعلمية تعتمد علي ثلاث ركائز أولها: التكنولوجيا (تعظيم قوة المعالجة ودقة الخوارزميات)، وثانيها التحليلات (تحديد الأنماط والاتجاهات التي توجد في البيانات الضخمة) وأخيراً النظرة البطولية الأسطورية Mythology (يقصد بها الاعتقاد بأن مجموعات البيانات الضخمة تمثل شكل

متقدم جداً من الذكاء والاستخبارات مع وجود لمحة من الصدق والموضوعية والدقة) (boyd & Crawford, 2012, p. 661).

يتشارك في هذه النظرة للبيانات الضخمة تعريف ماير زوكربرج وكوكير اللذان وصفا البيانات الضخمة في كتابهما الذي أسمياه " البيانات الضخمة: الثورة التي ستغير كيفية معيشتنا وعملنا وتفكيرنا" ويحصرا التغيرات التي أثرت بها تحليلات البيانات الضخمة على طريقة تنفيذنا للأمور وعلى المجتمع في ثلاثة تغيرات:

١- المزيد من البيانات more data: حيث أصبح باستطاعتنا جمع كمية هائلة من البيانات عن الظاهرة قد تصل في مجملها الى حد الكمال والتكامل بدلا من طريقة أخذ العينات التي كانت معتمدة في عصر البيانات الصغيرة.

٢- المزيد من الفوضى messiness: وهذا يعني اننا يمكننا التخلي عن القليل من الدقة في المدخلات في مقابل كمية أكبر من البيانات وذلك سيكون أفضل إجمالاً.

٣- الارتباطاتcorrelations : حيث ستكون العلاقات والارتباطات بين مجموعات البيانات واكتشاف الأنماط السائدة أهم من "السببية" التي تفسر لنا كيفية حدوث الظواهر، وسوف نعتمد على هذه الارتباطات في اتخاذ القرار حتى دون معرفة السبب خلف اختيارنا لهذا القرار دون غيره.

(schonberger & Cukier، ٢٠١٣(p. 20 ,

وهناك اتجاهات أخرى في تعريف البيانات الضخمة حيث قام دي مارو وآخرون (De Mauro, 2015, p. 99) بجمع التعريفات التي سبقته للبيانات الضخمة ووجد أن جميع التعريفات تنظر الى البيانات الضخمة من اتجاهات أربعة رئيسة هي:

١- المعلومات Information:

بصفتها الوقود المشغل والمادة الخام الرئيسية المكونة للبيانات الضخمة، خاصة بعد تحويل البيانات من الشكل التناظري الى الشكل الرقمي مثل محاولات رقمنة الكتب المطبوعة الى مجموعات رقمية وأشهر هذه المحاولات مشروع project3 الذي قامت فيه شركة جوجل برقمنة حوالي ١٥ مليون كتاب عام ٢٠٠٤م، وانتشار الأجهزة الذكية وانترنت الأشياء.

٢- التقنيات Technologies:

ويقصد بها المعدات والأدوات التي تتعامل مع البيانات الضخمة لفك تعقيدها والتغلب على سرعتها عبر أداة استعلام query tool متخصصة في التعامل مع البيانات الضخمة والتي أشهرها هو هادووب Hadoop وفق اتجاهات جوجل (google trends, 2017) google trends ، والعنصر التكنولوجي الآخر هو القدرة على التخزين فبرغم الطفرة في وسائط التخزين الا أن الأبحاث تتجه حاليا نحو إيجاد حلول تواكب هذا الفيضان من البيانات (Hilbert, 2012).

٣- الطرق Methods:

ويشير في هذا الى تحويل البيانات الضخمة الى قيمة value عن طرق إدارة البيانات الضخمة وتحليلاتها عن طريق طرق المعالجة المتقدمة التي تتجاوز بكثير الطرق الإحصائية التقليدية، وتتطلب هذه الطرق مهارات خاصة في المتعاملين مع هذه البيانات في الإحصاء والحاسب الآلي والإدارة وغيرها.

٤- الأثر Impact:

ويشير في هذا الى تأثير البيانات الضخمة كظاهرة على المجتمع عن طريق قصص النجاح التي غيرت الانطولوجيات وأساليب التفكير، وغيرت أنماط انتاج البيانات ونشرها والاستفادة منها في كل المجالات أهمها الصناعية والتجارية والعلمية، ويشير أيضا الى الآثار السلبية التي قد تعود على المجتمع وخطورتها على الإرادة البشرية والحريات الشخصية.

ويتبنى الباحث التعريف الذي يقدمه دي مارو (De Mauro, 2015)؛ حيث توصل الى تعريف يراه الباحث جامعاً ينظر فيه الى البيانات الضخمة بصفتها في الأصل "أصول معلوماتية Information assets" حيث يعرف البيانات الضخمة على أنها:

" أصول معلوماتية تتميز بحجم كبير وسرعة وتنوع بحيث تتطلب تقنيات خاصة وطرق تحليلية لتحويلها الى قيمة"

٣/١ خصائص البيانات الضخمة:

بشكل عام، إن خصائص "البيانات الضخمة" تستند إلى التعقيدات والتحديات التي تواجهها المنظمات والأشخاص عند تعاملهم مع هذه البيانات وخصائص البيانات الضخمة تتلخص فيما يلي:

١- الحجم Volume:

في الواقع، إن حجم البيانات الكبير لن يشكل مشكله حقيقيه في التخزين، لكن المشكلة تظهر عندما نريد استرجاع هذه البيانات أو تحليلها .حيث أن سعة التخزين تتزايد بشكل كبير، ولكن العثور على المعلومات المطلوبة داخل تلك الكمية الهائلة من البيانات، وتحليلها هي المشكلة الحقيقية. وذلك لأن تلك البيانات يتم جمعها من مصادر مختلفة (على سبيل المثال، مواقع التواصل الاجتماعي، وصفحات الويب وأجهزة الاستشعار، الخ)، لأغراض محددة النطاق (Dijcks, 2013)

٢- التنوع Velocity:

طبيعة البيانات الخام هي أن تكون متنوعة وغير مصنفه وغير منظمه وأن تأتي بأشكال وصيغ مختلفة، كم أننا يمكن أن نقسم هذه البيانات في صورتها الخام الى ثلاثة أنواع:

- **بيانات مهيكلة dataStructured :**

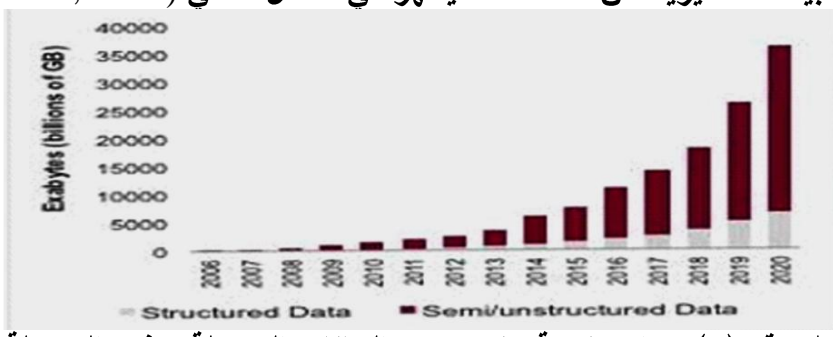
وهي البيانات المنظمة فى صورة جداول او قواعد بيانات تمهيدا لمعالجتها.

- **بيانات غير مهيكلة Unstructured data:**

تشكل النسبة الأكبر من البيانات وهي البيانات التي يولدها الأشخاص يوميا من كتابات نصية وصور وفيديو ورسائل ونقرات على مواقع الانترنت الخ.

- **بيانات شبه مهيكلة dataSemi-structured :**

تعتبر نوعا من البيانات المهيكلة الا ان البيانات لا تصمم في جداول او قواعد بيانات مثل الملفات المكتوبة بصيغة إكس أم أل XML و اتش تي إم ال HTML، وتمثل البيانات غير المهيكلة وشبة المهيكلة الجزء الأكبر من البيانات قد يزيد عن ٨٠% كما يظهر في الشكل التالي (Botteri, 2012):



شكل رقم (١) يوضح نسبة ونمو حجم البيانات المهيكلة وغير المهيكلة وشبه المهيكلة

ونتيجة لهذا التنوع فإن غالبية البيانات المنتجة تكون غير صالحة لاستهلاك المستخدمين بشكل مباشر لذلك، فإنه يتطلب جهدا ضخما لاستخراج

سمات تلك البيانات، لتغييرها إلى شكل موحد منظم قبل أن يصبح من الممكن استخدامها (NIST, 2013, p. 11)

٣- السرعة variety:

ببساطة، السرعة هو مصطلح يطلق على البيانات التي تتغير مع مرور الوقت أو التي يتم توليدها بشكل متكرر، على سبيل المثال، البيانات التي يتم جمعها من أجهزة الاستشعار في السيارات والمصانع والمراصد الفلكية والأقمار الصناعية، وسرعة البيانات في الحقيقة مشكلة يجب حلها، لأن تحليل هذه البيانات عادة يجب أن يكون في الوقت الحقيقي لإعطاء استجابة وردود فعل سريعة لأغراض الرقابة واتخاذ القرار ومنع حدوث الكوارث والخسائر المترتبة على أخذ قرار سريع، وعلاوة على ذلك، فإن هذا النوع من البيانات معرض للضياع إذا كانت قوة المعالجة أو خوارزميات التحليل ليست جيدة بما فيه الكفاية(De Mauro, 2015)

٤- المصداقية Veracity:

ويقصد به إمكانية استخراج البيانات القيمة من بين الأكوام الخثة والتي تسمى "dirty data"، فبعد ظهور الجيل الثاني والثالث من الويب أصبح من الممكن لأي شخص نشر أي نوع من البيانات، لذا أصبح من المهم جدا تحديد مصادر البيانات ومصداقيتها ومستويات الدقة فيها، خاصة أن البيانات الضخمة قد تعتمد الى مدى بعيد على بيانات من صنع البشر كمواقع التواصل الاجتماعي مثل فيس بوك وتويتر ومواقع التسوق وحجز الفنادق وتذاكر الطيران وغيرها. (Schroeck، Shockley، Smart، Morales، و Tufano، ٢٠١٢).

عدم الثبات Volatility:

تعاني البيانات عموما والأحجام الضخمة خاصةً من التناقض وعدم الثبات في صيغها وقيمتها ومكان تواجدها، لأن مكان ومعنى البيانات وما تمثله من قيمة يتغير مع مرور الوقت وفي السياقات مختلفة، وهذا يجعل إدارتها أصعب ونتائج التحليلات التي تجرى عليها تصبح غير مستقرة، خاصة عند التعامل مع البيانات متعددة المصادر والصيغ مثل مواقع التواصل الاجتماعي بصفة خاصة ومواقع الانترنت بوجة عام حيث تتميز بمحتواها القابل للتعديل والمحو من قبل المستخدمين (Dijcks, 2013)

٥- التعقيد Complixity :

نظرا لجمع البيانات من مصادر مختلفة كالمستشعرات ومواقع الانترنت وبيانات الشراء والتحويلات البنكية ونمط النوم والعادات الغذائية والتاريخ المرضي والجنائي وغيرها، لذا تظهر مشكلة جديدة بسبب التركيبة المختلفة للبيانات وتنوع التمثيل للبيانات خاصة مع انتشار انترنت الأشياء Internet of things، الأمر الذي يتطلب تحويل وربط هذه البيانات المختلفة، لإنتاج بيانات مترابطة قابلة لإجراء عمليات التحليل (Intel, 2012)

٤/١ مصادر البيانات الضخمة :

قامت اللجنة الاقتصادية لأوروبا UNECE التابعة للأمم المتحدة بتقديم تقرير بعنوان "ماذا تعنيه البيانات الضخمة للإحصاءات الرسمية" (UNECE, 2013) وقد أوردت فيه تصنيفا لمصادر البيانات الضخمة على النحو التالي :

المصادر الناشئة عن إدارة أحد البرامج:

Administrative (arising from the administration of a program):

سواء كان البرنامج أو الإدارة لها تبعية حكومية أو غير حكومية، مثل السجلات الطبية الإلكترونية وزيارات المستشفيات وسجلات التأمين والسجلات المصرفية وبنوك الطعام وبرامج التعليم والبحث العلمي في الجامعات والمراكز البحثية والتعاملات البنكية للمواطنين وغيرها، حيث تشير التقديرات إلى أنه يتم إنشاء ما يصل إلى ١٥٠ اكسابايت (١٥٠ مليار مليار) من البيانات عالمياً في مجال الرعاية الصحية وحده كل عام. (Solon, 2014)،

المصادر التجارية أو ذات الصلة بالمعاملات، الناشئة عن معاملات بين كيانين:

Commercial or transactional

على سبيل المثال معاملات البطاقات الائتمانية والمعاملات التي تجرى عن طريق الإنترنت بوسائل منها الأجهزة المحمولة، وهذا النوع ازدهر ونما بصورة هائلة مع ظهور التجارة الالكترونية في بداية التسعينات التي أدت الى تضاؤل البعد المكاني وتخفيض تكلفة الدعاية والنقل بالإضافة الى تبادل الآراء وإعطاء التفضيلات وإمكانية التعرف على العميل بنظرة بانورامية نرى بها كل أنشطته وتفضيلاته وتاريخه الشرائي، ولك أن تتخيل أن موقع

Amazon.comيعالج ملايين العمليات الخلفية كل يوم، فضلاً عن استفسارات من أكثر من نصف مليون بائع طرف ثالث، و تملك شركة أمازون أكبر ٣ قواعد بيانات لينوكس في العالم والتي تصل سعتها إلي ٧.٨، 18.5و ٢٤.٧ تيرابايت يعرض للبيع حوالي 27مليون مادة ويقوم ببيع ٤٢٦ سلعة كل ثانية. (فؤاد، ٢٠١٣)

مصادر شبكات أجهزة الاستشعار Sensors:

على سبيل المثال، التصوير بالأقمار الصناعية، وأجهزة استشعار الطرق، وأجهزة استشعار المناخ، ورادارات السرعة على الطرق، فعلى سبيل المثال نجد المشروع الفلكي "مسح سلووان الرقمي للسماء Sloan Digital Sky Surve" وهو مشروع بحثى أمريكي يهدف الى مسح فلكى للسماء باستخدام تليسكوب عملاق متصل بمستشعرات لجمع البيانات عن تكوين النجوم في مجرة درب التبانة عندما تم البدء بجمع البيانات الفلكية في عام ٢٠٠٠، فإنه قد تم جمع بيانات في أسابيعه القليلة الأولي أكثر مما تم جمعه في تاريخ علم الفلك بأكمله، ومع استمراره بمعدل ٢٠٠ جيجا بايت في الليلة، جمع أكثر من ١٤٠ تيرابايت من المعلومات.

مصادر أجهزة التتبع Tracking devices:

على سبيل المثال تتبع البيانات المستمدة من الهواتف المحمولة والنظام العالمي لتحديد المواقع GPS الذي يستخدم في أكثر من جهاز أشهرها الهواتف النقالة والسيارات حيث يرشد الـــ GPS أكثر من 100مليون سائق في كل أنحاء العالم يومياً ويسجل تحركاتهم كما تقوم شركة Windermere Real Estateباستخدام إشارات GPS مجهولة من ما يقرب من ١٠٠ مليون سائق لمساعدة مشتري المنازل الجدد لتحديد أوقات قيادتهم من وإلى العمل خلال الأوقات المختلفة لليوم، ومثال آخر مصادم الهيدرون العظيم يملك ١٥٠ مليون جهاز استشعار تقدم بيانات ٤٠ مليون مرة في الثانية الواحدة. وهناك ما يقرب من ٦٠٠ مليون تصادم في الثانية الواحدة. لكن نتعامل فقط مع أقل من ٠.٠٠١% من بيانات تيار الاستشعار، فإن تدفق البيانات من جميع تجارب المصادم الأربعة يمثل ٢٥ بيتابايت.

مصادر البيانات السلوكية Behavioural:

على سبيل المثال، مرات البحث على الإنترنت عن منتج أو خدمة ما أو أي نوع آخر من المعلومات، ومرات مشاهدة إحدى الصفحات على الإنترنت، فعلى سبيل المثال *متجر وول مارتWal-Mart* : وهى شركة أمريكية للبيع بالتجزئة بعائدات تبلغ ٣٨٧.٦٩ مليار دولار أمريكي وتقوم بمعالجة أكثر من مليون معاملة تجارية كل ساعة، والتي يتم استيرادها إلي قواعد بيانات يقدر أنها تحتوي علي أكثر من ٢.٥ بيتابايت (٢٥٦٠ تيرابايت) من البيانات – وهو ما يوازي ١٦٧ ضعف البيانات الواردة في جميع الكتب الموجودة في مكتبة الكونغرس في الولايات المتحدة (Banjo, 2014)

مصادر البيانات المتعلقة بالآراء Opinion:

على سبيل المثال، التعليقات على وسائط التواصل الاجتماعي، فقد ذكرت البيانات الرسمية الصادرة عن فيسبوك بأن قاعدة مستخدمي الموقع حول العالم توسعت لتسجل مع نهاية الربع الرابع من العام ٢٠١٣م حوالي ١.٨٦ مليار مستخدم نشط للشبكة الاجتماعية الاكثر شعبية حول العالم، منهم حوالي ١.١٩ مليار مستخدم نشط عبر الاجهزة المتنقلة الذكية.ويتم رفع أكثر من ٣٠٠ مليون صورة يوميا على الموقع بينما تخزن داخل قواعد بياناتها حوالى ٥٠ مليار صورة ، واستنادا إلى آخر البيانات المالية الصادرة عن "فيسبوك" العالمية، زادت قاعدة مستخدمي الشبكة بمقدار ١٦٥ مليون مستخدم، وبنسبة تصل إلى% 13 خلال فترة عام، وذلك لدى المقارنة بعدد مستخدمي الشبكة المسجل في نهاية الربع الرابع من العام ٢٠١٣م، والذي بلغ وقتذاك ١.٢٢٨ مليار مستخدم(zeforia, 2017) ، كما تصنف شركة IBM مصادر البيانات الضخمة بصورة أشمل وأبسط بحيث توزعها على ثلاثة مصادر فقط:

(بيانات تنتج من الانسان– بيانات تنتج من التجارة– بيانات تنتج من الآلات)

يرى الباحث أهمية إضافة مصدر مهم جدا للبيانات الضخمة وهو البيانات الناتجة عن الوزارات والدواوين الحكومية وتضم أشكال عديدة مثل التعدادات السكانية والرقم القومي والجوازات وسجلات المرور وسجلات المواليد والوفيات وغيرها، فالحكومات هي المجمع الأصلي الذي يقوم بتجميع البيانات على مستويات شاسعة، ولا تزال هذه الحكومات تنافس القطاع الخاص بقوة عن طريق ذلك الحجم الهائل من البيانات التي يسيطرون عليها ويتحكمون فيها. والفرق الوحيد بينها وبين القطاع الخاص أن الحكومات غالباً ما تجبر

الناس على تزويدهم بالمعلومات بدلا من إقناعهم بذلك أو تقديم شيئا في المقابل مثلما يفعل القطاع الخاص؛ ونتيجة لذلك تظل الحكومات تستمر هذه الكنوز المدفونة والمناجم الهائلة من البيانات حتى دون أن تدري ما تحويه هذه البيانات الضخمة من قيمة ضخمة أيضاً من هنا ظهر مصطلح البيانات الحكومية المفتوحة open data.

البيانات الحكومية المفتوحة (البيانات المفتوحة) "open government data":

ظهرت فكرة لاقت استحسانا كبيرا في الآونة الأخيرة وهي أن أفضل طريقة لاستخراج القيمة الكامنة في هذه البيانات الحكومية هي إعطاء القطاع الخاص والمجتمع حق الوصول لهذه البيانات ليقوموا بهذه المهمة. والمبدأ الذي يستند له هذا الحل أن الحكومة تجمع هذه البيانات نيابة عن المواطنين وبالتالي فإن هؤلاء المواطنون من حقهم الوصول الى معلوماتهم الا في عدد محدود من الحالات مثلا إذا كان الوصول الى هذه البيانات تضر الأمن القومي أو تضر حقوق أو خصوصية الأخرين.

هذه الفكرة نتج عنها عدد لا يحصي من المبادرات التي تدعوا الى جعل البيانات الحكومية مفتوحة "open government data" في جميع أنحاء العالم، وكانت حجة هذه المبادرات أن الحكومات ما هي الا أوصياء على هذه البيانات التي يجمعونها، وأن القطاع الخاص والمجتمع سوف يكونان أكثر ابتكاراً، وعليه قام الدعاة الى هذه المبادرات بمخاطبة الجهات الرسمية لإتاحة حق الوصول الى البيانات للجهات التجارية والمدنية، وقد لاقت فكرة البيانات الحكومية المفتوحة دعما كبيرا حينما أصدر الرئيس باراك أوباما في أول يوم تولى فيه الحكم ٢١يناير ٢٠٠٩م مذكرة رئاسية يأمر فيها رؤساء الوكالات الفيدرالية بتحرير أكبر قدر ممكن من البيانات والاعفاء عنها، حيث كتب في تعليماته "حينما نواجه الشكوك، تفوز الصراحة "In the face of doubt, openness prevails," وكان هذا الإعلان مدويا وملفتا للأنظار خاصة أن موقفه هذا يتعارض تماماً مع الرئيس الذي سبقه والذي أمر الوكالات بعكس ذلك تماماً. (Akin، ٢٠٠٩)

كما أمر أوباما بإنشاء موقع على الانترنت يسمي data.gov يكون مستودعا للمعلومات والوصول اليه مفتوح ومتاح للجميع، بحيث يوضع فيه المعلومات الصادرة عن الحكومة الاتحادية، ونما هذا الموقع بسرعة انفجارية من مجرد نواه بها ٤٧ وحدة بيانات في ٢٠٠٩م الى ما يقارب ٤٥٠٠٠٠

وحدة بيانية تصدر عن ١٧٢ وكالة حكومية بحلول الذكرى الثالثة لتولى أوباما الرئاسة في يوليو من عام ٢٠١٢م، وفي بريطانيا أصدرت الحكومة البريطانية قوانين تشجع البيانات المفتوحة وتدعم إنشاء معهد البيانات المفتوحة Open Data Institute والذي يشارك في إدارته "كيم بيرنرز لي Tim Berners-Lee" مخترع الشبكة العنكبوتية العالمية، كما أعلن الاتحاد الأوربي أيضا عدد من المبادرات للبيانات المفتوحة التي يمكن أن تشمل القارة الأوربية كلها في القريب العاجل، وهناك دول أخرى أعلنت مبادرات تنفيذ استراتيجيات تدعم البيانات المفتوحة مثل استراليا والبرازيل وتشيلي وكينيا، وعلى التوازي من كل ذلك، شكلت مجموعات من مطوري الويب والمفكرين الذين يملكون رؤى لاكتشاف الطرق التي يستفيدون منها أعظم استفادة، هذه المجموعات أمثال: Code في أمريكا وthe Sunlight Foundation في بريطانيا.(Schönberger & Cukier، ٢٠١٣(, p. 116)

٥/١ أدوات تقنيات البيانات الضخمة:

إن الآلية التي تعمل تقنيات البيانات الضخمة بناءً عليها هي التغلب على قيود التخزين والمعالجة للأحجام الضخمة والمختلفة والمتنوعة من البيانات، وبدلا من التوسع الرأسي "vertical scaling" في إمكانات المعالجة والتخزين للحاسبات الآلية وهذه الطريقة مكلفة مادياً، وعوضاً عن ذلك تتوسع أفقياً "horizontal scaling" في عمليات المعالجة بحيث تقسم مهمة المعالجة والتحليل للبيانات على ملايين الحاسبات زهيدة الثمن والتي تكون متصلة عن طريق الشبكات في تكتلات تسمي عناقيد حاسوبية Clusters بدلاً من إسناد المهمة لحاسب آلي واحد ضخم يكلف ملايين الدولارات وقد لا يفي بالغرض، ومن ثم العمل على الاستحواذ علي البيانات بسرعة عالية واكتشافها و/أو تحليلها وذلك عن طريق تخزينها في صيغة "نظام الملفات الموزعة Distributed File System" بحيث يسهل كشف البيانات أينما كانت في كتلة من الحواسيب الخوادم، كما أن أدوات معالجة تلك البيانات موزعة هي أيضاً، وتقع غالباً على نفس الخوادم التي تضم البيانات، هذا ما يفيد في جعل معالجة البيانات أسرع، لكن تنفيذ هذه المهمة يتطلب مجموعة متكاملة من الأدوات التي تعمل فيما بينها على السيطرة والتحكم التام في هذه البيانات. (Warden, 2011, p. 2)

ويري د.م.ويست (West, 2012) أن الأدوات التي تتعامل مع البيانات الضخمة موزعة على فئات ثلاث رئيسة هي:

١- أدوات التنقيب عن البيانات Data mining والتي عادة تتعامل مع بيانات غير مهيكلة (كالنصوص وحركات المستخدمين) والتي تكون موزعة على أجهزة مختلفة عبر الويب.

٢- أدوات التحليل Data Analysis التي تستخدم المقارنة والتصنيف والمقاربة والربط وغيرها من الأدوات التحليلية والتنظيمية للخروج بالنتائج المطلوبة.

٣- أدوات عرض النتائج Dashboard والتي تعرض بشكل مرئي ورسومي النتائج النهائية للتحليل وفقا لما تم تحديده كهدف للتحليل مسبقا.

تحليلات البيانات الضخمة Big Data Analytics:

ان الأدوات والتقنيات سابقة الذكر جميعها تعمل جميعاً لتحقيق هدف رئيس واحد وهو استخراج كامل القيمة الحالية والمستقبلية الموجودة داخل البيانات، وهذه القيمة لا نستطيع التعرف عليها الا بعد استخراج الارتباطات correlations بين البيانات، بالإضافة الى معرفة الانماط التي تسير وفقها تلك البيانات، على سبيل المثال نجد شركة جوجل عرفت الارتباط بين إصابة أي شخص بمرض والبحث عنه في محرك البحث جوجل على شبكة الانترنت ومن هذا المنطلق قامت شركة جوجل عام ٢٠٠٤م بتدشين مشروع google flu trends ويقوم هذا المشروع بالتعرف على المناطق التي تنتشر فيها الإصابة بالأنفلونزا بدون الرجوع الى السجلات والإحصاءات التي تصدرها المستشفيات والمؤسسات الطبية والتي تتقادم قبل صدورها وبذلك ساعدت الولايات المتحدة في مكافحة الانفلونزا بأنواعها (google, n.d).

ان الارتباطات مفيدة في عالم البيانات الصغيرة، لكنها في حالة البيانات الضخمة مبهرة وبراقة، لأننا عن طريقها يمكننا إدراك الواقع بشكل أسهل وأسرع وأكثر وضوحا من ذي قبل، إن الارتباطات في جوهرها هي قياس كمي للعلاقة الإحصائية بين قيمتين من البيانات، ووجود ارتباط قوي يعنى أنه عندما تعمد قيمة أحد هذين القيمتين في التغير فإن القيمة الأخرى يصبح من المرجح بصورة كبيرة أن تتغير قيمتها أيضاً. ولقد رأينا مثل هذا الارتباط في مشروع GOOGLE FLU TRENDS كلما قام الناس في منطقة

جغرافية بعينها بالبحث عن مصطلح معين في محرك البحث جوجل، كان هناك المزيد من البشر مصابين بالأنفلونزا في هذه المنطقة، وعلى العكس، فالارتباط الضعيف يعني أنه عندما يحدث تغير في أحد البيانات يحدث تغير طفيف في المتغير الآخر. على سبيل المثال يمكننا حساب معامل الارتباط بين طول الشعر للفرد ومدي سعادة هذا الشخص، فنجد أن طول الشعر ليس مفيداً في معرفة مدي السعادة، ان الارتباطات تساعدنا على معرفة الحاضر والتنبؤ بالمستقبل: فإذا كانت الظاهرة س تحدث عند حدوث الظاهرة ص، عندها يمكننا انتظار حدوث ص لنتوقع حينها أن الظاهرة س في طريقها للحدوث. (schonberger & Cukier، ٢٠١٣)، p. 30 ,

نجد أيضاً متاجر TARGET أشهر متاجر تجزئة أمريكية تقوم بتحليل أنماط الشراء للزبائن لتحديد التفضيلات والعروض التي ترسلها لكل عميل وأنسب الأماكن التي تضع فيها كل منتج ووصلت دقة هذه التحليلات الى أنها تعرف من أنماط المشتريات أن أحد عملائها السيدات حامل وهل ستلد ولد أم بنت ومتى ستلد وما الماركات التي تفضلها وبناءً على هذه المعلومات تقوم بإرسال العروض الملائمة لكل عميل على حدى، كذلك مجموعة فنادق (CHAIN HOTEL) تستخدم تحليلات البيانات الخاصة بالطقس وإلغاء الرحلات الجوية وبيانات تحديد الموقع GBS للمسافرين الذين تأخرت رحلاتهم لاستقطابهم كنزلاء، وزاد عدد النزلاء أكثر من ٢٠% بعد هذه الطريقة.

ومتاجر بيع المأكولات والبيتزا تقوم بتحليل بيانات الطقس وأماكن انقطاع التيار الكهربي ومواقع الهواتف المحمولة للعملاء فترسل للمناطق التي ينقطع بها التيار الكهربي العروض وقوائم الطعام حيث أن انقطاع التيار الكهربي يعيق ربات المنزل في عملية الطبخ، كما تقوم شركات الإنتاج للأفلام والالبومات الغنائية بتحليل المشاهدات والتفضيلات على مواقع الاستماع ومواقع التواصل الاجتماعي لمعرفة أنواع الأفلام والموسيقي التي يفضلها المشاهدون والمستمعون بل وتحديد الأبطال المفضلين لهم والأغاني التي يفضلونها لتكون هي عنوان الالبوم، وشركات النقل للركاب والبضائع تقوم هي أيضاً بتحليل البيانات التي تصدرها المستشعرات في أساطيل السيارات التي تمتلكها لتقرر أفضل وأقصر الطرق لكي تسير فيها بل وتتنبأ بأي تعطل في أجزاء هذه السيارات قبل حدوثه، علاوة على استخدام التحليلات في كشف الانتحال ومحاولات الاختراق للبنوك وحسابات العملاء، وكشف الحركات الإرهابية

والتشكيلات العصابية، وتحليل إمكانات المنافسين في السوق وغيرها الكثير (Schaeffer, n.d)

ولتستفيد أي مؤسسة الاستفادة القصوى من البيانات التي تمتلكها عن القيام بتحليلات البيانات الضخمة يجب التفريق بين أربعة أنواع من التحليلات نفصلها في التالي:

١- التحليلات الوصفية Descriptive analytics:

هو أبسط أنواع التحليلات ويعمل على تحويل الكميات الضخمة من البيانات المتشابكة والمعقدة الى بيانات سهلة الفهم وذات مغزى، وبهذا يصبح دور هذا النوع من التحليلات هو وصفى الحالة الراهنة استناداً الى البيانات الطازجة الصادرة في الوقت الحقيقي "REAL TIME" ويقوم بتلخيص ما يحدث استناداً الى البيانات الواردة من لوحات التحكم وقوائم البريد الالكتروني وغيرها، ويفيد التمثيل المرئي للبيانات VISUALIZATION في هذا النوع حيث يسهل عملية الوصف، ومن أشهر الأعمال التي تستخدم فيها هذه التحليلات خرائط الظواهر الطبيعية كالزلازل والبراكين والأحوال الجوية من بيانات المستشعرات، والخرائط الملاحية لسير السفن في البحر من المعلومات عن التيارات المائية وحركة الأمواج وارتفاعها، والطرق الجوية من بيانات الطائرات عن المطبات الهوائية والأعاصير وغيرها، كما يمكننا بواسطة هذه التحليلات التعرف على أكثر الطرق ازدحاماً من بيانات تحديد الموقع GIS في الهواتف والسيارات وأكثر الأفلام السينمائية إقبالاً من تحليلات مواقع التواصل الاجتماعي وغيرها الكثير.

٢- التحليلات التشخيصية Diagnostic analytics:

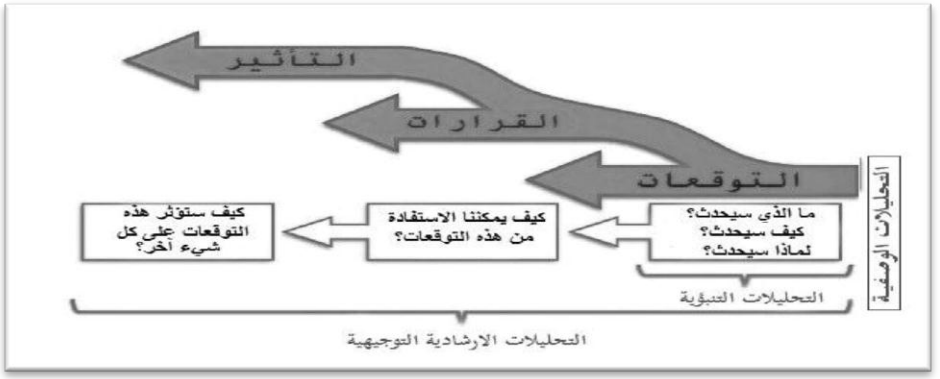
دور ذلك النوع من التحليلات هو النظر في الاحداث الماضية لتحديد ما الذي حدث؟ ولماذا حدث على هذا النحو؟ بمعنى آخر تكشف لنا عن الجزور والأسباب الأساسية التي تسببت في وجود حدث ما؛ على سبيل المثال عند حدوث انخفاض في المبيعات، أو زيادة في عدد المصابين بمرض ما، أو ارتفاع أسهم الشركة بدرجه كبيرة، وغيرها من الأحداث التي تحتاج الى تفسيرات، وغالبا ما نحتاج هذا النوع للأغراض الرقابية لمحاسبة أو مكافئة المتسببين كما يستخدم في اعتماد الاستراتيجيات الثابتة للمنظمة من حيث الميزانية والأرباح وتلافي المخاطر وتغيرات الأداء السلبي وتخفيض الأرباح وغيرها، ويعتمد هذا النوع من التحليلات على التحليلات الوصفية (cyient, n.d)

٣- التحليلات التنبؤية (التوقعية) Predictive analytics :

هو مجال للتحليل الاحصائي للبيانات هدفه استخلاص معلومات حول االتغيرات السلوكية المستقبلية ، بحيث يقوم بتحديد السيناريوهات المستقبلية التي يمكن أن تحدث اعتماداً على ما سبق من تحليلات وصفية وتشخيصية، ويتأسس التحليل التنبؤي على فهم العلاقة بين المتغيرات التي تتسبب في الأحداث و المتغيرات المتغيرة مع الاحداث أو المتوقعة (المظاهر و الأسباب) المنبثقة عن تجارب في الماضي و استعمال هذه العلاقات من أجل توقع المستقبل حيث يقوم بتحديد أنماط البيانات السابقة ويقدم قائمة بالنتائج المحتملة لكل حالة من الحالات، وبهذا فإن نتائج هذا التحليل تكون فرضيات لما سيحدث مستقبلاً، وقد تستخدم في توقع متطلبات العملاء المستقبلية والمخاطر المحتملة والفرص المستقبلية من حيث المبيعات والمكاسب والأصول والقدرة الإنتاجية، ومثال على هذه التحليلات توقع الإصابة بأمراض السرطان والقلب، ومواقع توقع انخفاض أسعار تذاكر السفر مثل farcast.com وتوقع الحالة الجوية وغيرها، ونتائج التحليل التنبؤي تتأثر بشكل كبير بجودة الفرضيات و مستوى تحليل البيانات، أما أشهر استخداماته على الاطلاق هو التقييم الائتماني للأشخاص (CREDIT SCORING)المستخدم في الخدمات المالية. يقوم التقييم الائتماني بدراسة تاريخ العميل الائتماني، إضافة إلى طلبات القروض و معلوماته الأخرى بهدف تقييم العملاء و ترتيبهم على أساس احتمالية دفعهم للدفعات الائتمانية المستقبلية في وقتها المحدد.

٤- التحليلات الارشادية التوجيهية Prescriptive Analytics

دوره الكشف عن الإجراءات التي يجب اتخاذها مستقبلاً وهذا النوع هو الأكثر قيمة لأنه يعطيك القرار وليس المعلومة فقط وهذا هو أقصي طموح وصلت اليها تحليلات البيانات، حيث تستخدم نتائج التحليلات الوصفية والتشخيصية والتنبؤية وتضيف اليها اقتراحات بناءً على القرارات التي تم اتخاذها سابقا في هذه المنظمة أو المنظمات المثيلة



شكل رقم (٢) تكامل أنواع تحليلات البيانات وأثرها على المنظمة

المراجع

١. **أحمد فؤاد.** (٢٤ ١٢، ٢٠١٣). ***مبيعات أمازون.. ٤٢٦ سلعة بالثانية.*** **تاريخ الاسترداد ١ ٦، ٢٠١٧، من سكاي نيوز:** ***http://www.skynewsarabia.com/web/article/508846/%D9%85%D8%***

٢. **تركي العسيري.** (٢٠٠٣). ***برمجة اطار عمل .NET باستخدام Visual .NET Basic.*** **تاريخ الاسترداد ١٦ ٧، ٢٠١٧، من** ***http://www.7ammil.com/index.php/files/guest/vbnetzip?do=download***

٣. **فتحي حسين عامر.** (٢٠١١). ***وسائل الاتصال الحديثة من الجريدة إلى الفيس بوك.*** **القاهرة : العربي للانتاج والتوزيع.**

4. **boyd, d., & Crawford, K. (2012, 6). CRITICAL QUESTIONS FOR BIG DATA. *Information, Communication & Society*, pp. 662-679.**
5. **Bryson, S., Kenwright, D., Cox, M., Ellsworth , D., & Haimes, R. (1999, 8). Visually exploring gigabyte data sets in real time. *Communications of the ACM*, pp. 82-90.**
6. **Dean, J., & Ghemawat, S. (2004, 12). *MapReduce: Simplified Data Processing on Large Clusters*. Retrieved 4 5, 2017, from google Research Publications: *https://research.google.com/archive/mapreduce.html***
7. **Marron, A. B., & de Maine, P. A. (1967 , Nov. 11). Automatic data compression. *Communications of the ACM, 10*.**

8. **Scagliarini, L., & Varone, M. (2016, 8 18). *NLP for Big Data: What everyone should know?* Retrieved 7 16, 2017, from expertsystem.com: *http://www.expertsystem.com/nlp-big-data-everyone-know/***
9. **Adey, P. (2010). *Mobility.* new york: Routledge.**
10. **Agrawal, R. (2016, 10 1). Challenges of big data storage and management. *Global Journal of Information Technology*, pp. 1-10.**
11. **Amar, R., Eagan, J., & Stasko, J. (2005, 10). Low-Level Components of Analytic Activity in Information Visualization. *Information Visualization*, pp. 111-117.**
12. **Aronica, J. (2014, 7 30). *5 Email Marketing Lessons From Amazon*. Retrieved 7 2, 2017, from .klaviyo.com: *https://www.klaviyo.com/blog/5-email-marketing-lessons-from-amazon***
13. **Banjo, S. (2014). *Wal-Mart Notches Web Win Against Rival Amazon.* manhattan: daojones.**
14. **Bhandar, I. (2013). *Big Data Innovation Summit in Boston*. Retrieved 5 3, 2017, from the innovation enterprise: *https://theinnovationenterprise.com/summits/big-data-innovation-boston***
15. **Botteri, P. (2012, 10 24). *Eastern European Champions & the 4 V's of Big Data*. Retrieved 5 9, 2017, from Cracking The Code: http://cracking-the-code.blogspot.com.eg/2012/10/eastern-european-champions-4-vs-of-big.html**
16. **cyient. (n.d). *Diagnostic Analytics*. Retrieved 8 2, 2017, from cyient-insights.com: *http://www.cyient-insights.com***
17. **De Mauro, A. (2015). What is big data? A consensual definition and a review of key research topics. *4th International Conference on Integrated Information* (pp. 97-104). Madrid: AIP Publishing.**
18. **Dean, j., & Ghemawat., S. (2004, 10 3). MapReduce: Simplified Data Processing on Large Clusters.**

***Symposium on Operating Systems Design and Implementation*, pp. 137-150.**

19. **denning, p. j. (1990). Saving All the Bits. *American Scientist*, p. 402.**
20. **Dijcks, J. (2013). *Big Data for the Enterprise.* Retrieved 6 3, 2017, from oracle: *www.oracle.com/us/.../database/big-data-for-enterprise-519135.pdf***
21. **Dumbill, E. (2012). *Planning for Big Data.* California: O'Reilly Media, Inc.**
22. **Eric Brewer .(٢٠١٢) .CAP twelve years later: How the "rules" have changed .*Computer*.٢٩–٢٣ الصفحات ، ٢ ،**
23. **European Commission-CORDIS. (2012, 9 1). *Horizon 2020*. Retrieved 4 5, 2017, from https://ec.europa.eu/eurostat/cros/content/horizon-2020_en**
24. **Fang, E. I. (1997). *A history of information revolutions.* Washington: Butterworth-Heinemann.**
25. **Fisher, D. (2012). Interactions with big data analytics. *Interactions, 19*, pp. 50-59.**
26. **Francis X. Diebold'" .(٢٠٠٠) .Big Data 'Dynamic Factor Models for Macroeconomic Measurement and Forecasting .*the Eighth World Congress of the Econometric Society* .seattle .Retrieved 25/3/2017*http://www.ssc.upenn.edu/~fdiebold/papers/paper40/temp-wc.PDF***
27. **Gartner inc. (2015, 9 6). *Gartner Survey Shows More Than 75 Percent of Companies Are Investing or Planning to Invest in Big Data in the Next Two Years*. Retrieved 5 1, 2017, from gartner.com: *http://www.gartner.com/newsroom/id/3130817***
28. **Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003, 10). *The Google File System*. Retrieved 4 5, 2017, from google Research Publications: https://research.google.com/archive/gfs.html**

29. **google) .n.d .(*Google Flu Trends* .Retrieved 1/8/2017 google.org/flutrends: *https://www.google.org/flutrends/about/***
30. **google trends. (2017, 5 1). *Apache Hadoop*. Retrieved 5 9, 2017, from google trends: https://trends.google.co.uk/trends/explore?q=%2Fm%2F0fdjtq**
31. **Hilbert, M. (2012, 6). How to Measure "How Much Information"? Theoretical, Methodological, and Statistical Challenges for the Social Sciences. *International Journal of Communication*, p. 1042. Retrieved from http://ijoc.org.**
32. **Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big Data For Dummies.* Hoboken: John Wiley & Sons, Inc.**
33. **IBM. (n.d). *What is Big Data?* Retrieved 4 3, 2017, from ibm: *https://www.ibm.com/big-data/us/en/***
34. **Intel. (2012). *Big Data Analytics:Intel's IT Manager Survey on How Organizations Are Using Big Data.* Big Data Analytics.**
35. **Intel. (2012). *Big Data Analytics:Intel's IT Manager Survey on How Organizations Are Using Big Data*. Retrieved 2 25, 2017, from intel.com: *http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer- research-report.pdf***
36. **internet world stats. (2017, 7 1). *Internet Users Statistics for Africa*. Retrieved 10 15, 2017, from internetworldstats: *http://www.internetworldstats.com/stats1.htm***
37. **Kalakota, R. (2013, 5 23). *Data Monetization: Turning Data into $$$*. Retrieved 8 5, 2017, from practicalanalytics.com: https://practicalanalytics.co/2013/05/23/data-monetization-is-the-end-goal/**
38. **Keith D. Foote .(2,6,2016) .*Big Data Processing 101: The What, Why, and How* تاريخ .Retrieved 15/7/2017 ،**

dataversity.net: *http://www.dataversity.net/big-data-processing-101/*

39. King, Z. (2004). *The Story of Our Numbers: The History of Arabic Numerals.* new york: The Rosen Publishing Group.

40. Laney, D. (2001, 3 5). *3D Data Management: Controlling Data Volume, Velocity, and Variety.* Stamford: META group Inc. Retrieved 3 15, 2017, from META group: http://blogs.gartner.com/doug-laney/files/

41. Leek, J. (2015). *The Elements of Data Analytic Style:A guide for people who want to analyze data.* n.p: Leanpub.

42.* Liu, X. (2013, 9 19). *Understanding Big Data Processing and Analytics*. Retrieved 7 15, 2017, from developer.com: *http://www.developer.com/db/understanding-big-data-processing-and-analytics.html

43. Lyko, K., Nitzschke, M., Cyrille, A., & Ngonga , N. (2016). *New Horizons for a Data-Driven Economy:aRoadmap for Usage and Exploitation of Big Data in Europe.* Cham: springer nternational Publishing.

44. M. Schroeck ،R .Shockley ،J .Smart ،D .Romero Morales و ،P .Tufano .(٢٠١٢) .*Analytics: The Real-World Use of Big Data* .IBM report.

45.* Marty, R. (2005, 2). Retrieved 7 18, 2017, from *https://www.slideshare.net/zrlram/big-data-visualization-44258309

46. Michael Cox و ،David Ellsworth .(١٩٩٧) .Application-Controlled Demand Paging for Out-of-Core Visualization .*the 8th IEEE Visualization* p.235 Mississippi State: IEEE.

47. Miller, A. R. (1972). *THE ASSAULT ON PRIVACY - Computers, Data Banks, and Dossiers.* Ann Arbor: University of Michigan Press.

48. **National Science Foundation) .n.d .(*Research Areas* . Retrieved 9/2/2018 nsf.gov: *https://www.nsf.gov/about/research_areas.jsp***

49. NIST. (2013). *Frontiers in Massive Data Analysis.* Washington DC: National Academy of Sciences.

50. NIST. (2015). *Big Data Interoperability Framework: Definitions.* NIST Big Data Public Working Group. Retrieved from http://dx.doi.org/10.6028/NIST.SP.1500-1

51. Norman, J. M. (2005). *From Gutenberg to the Internet: A Sourcebook on the History of Information Technology.* california: Norman Publishing.

52. Ohlhor, F. (2013). *Big Data Analytics: Turning Big Data into Big Money.* Hoboken, New Jersey: John Wiley & Sons, Inc.

53. **Perry, C. (2015, 10 17). *What makes a data visualization memorable?* Retrieved 7 18, 2017, from *https://www.seas.harvard.edu/news/2013/10/what-makes-data-visualization-memorable***

54. Price, D. J. (1961). *Science since Babylon.* Binghamton: Vail-Ballou Press, Inc.

55. Richard Brown .(٢٠٠٦) .*A History of Accounting and Accountants* .new york: Cosimo, Inc.

56. Rider, F. (1944). *The Scholar and the Future of the Research Library. A Problem and Its Solution.* New York: Hadham Press.

57. Ronda-Pupo, G. A. (2012). Dynamics of the evolution of the strategy concept 1962–2008: a co-word analysis. *Strategic Management Journal, 2,* p. 188. doi:10.1002/smj.948

58. S.P.C.K. (1855). *The history of printing.* london: w. clowers.

59. **Schaeffer, C. (n.d). *Big Data in Retail Examples*. Retrieved 7 30, 2017, from .crmsearch.com: *http://www.crmsearch.com/retail-big-data.php***

60. **schonberger, v. m., & Cukier, K. (2013).** ***Big Data: A Revolution that Will Transform how We Live, Work, and Think.*** **new york: houghton miffline harcourt.**
61. **Schönberger, V. r., & Cukier, K. (2013).** ***Big Data: A Revolution that Will Transform how We Live, Work, and Think.*** **new york: Houghton Mifflin Harcourt.**
62. **Skytree, Inc. (2017).** ***Why do Machine Learning on Big Data?*** **Retrieved 7 14, 2017, from skytree.net:** ***http://www.skytree.net/machine-learning/why-do-machine-learning-big-data/***
63. **Solon, O. (2014).** ***A simple guide to Care.data*****. Retrieved 6 6, 2017, from wired.co.uk:** ***http://www.wired.co.uk/article/a-simple-guide-to-care-data***
64. **Techopedia Inc. (n.d).** ***Operational Analytics*****. Retrieved 8 5, 2017, from techopedia.com:** ***https://www.techopedia.com/definition/29495/operational-analytics***
65. **Tjomsland, I. (1980). Digest of Papers: The Gap between MSS Products and User Requirements.** ***Fourth IEEE Symposium on Mass Storage Systems*** **(p. 76). New York: Institute of Electrical and Electronics Engineers.**
66. **Trauvitch, G. (n.d).** ***Oracle's Five Journeys to Cloud Infrastructure*****. Retrieved 4 2, 2017from: https://go.oracle.com/LP=44710?elqCampaignId=80921&sc=ADV_FY17_ME_ENS_E158_E_Search&pcode=EMMK161209P00047&mkwid=s9TpZAvKq|pcrid|176876779646|pkw|exa%20data|pmt|p|pdv|c|sckw=srch:exa%20data**
67. **Tudoran, R. M. (2014).** ***High-Performance Big Data Management Across Cloud Data Centers.*** **Rennes-france: l'unité mixte de recherche-Institut de recherche en informatique et systèmes aléatoires .**