دار المنظومة
DAR ALMANDUMAH
الرواد في قواعد المعلومات العربية

العنوان:	نظم إدارة قواعد البيانات الضخمة: دراسة حالة لنظام أباتشي هادوب Hadoop Apache
المصدر:	اعلم
الناشر:	الاتحاد العربى للمكتبات والمعلومات
المؤلف الرئيسي:	سيد، أحمد فايز أحمد
المجلد/العدد:	ع23
محكمة:	نعم
التاريخ الميلادي:	2019
الشهر:	يناير
الصفحات:	121 - 184
رقم MD:	970704
نوع المحتوى:	بحوث ومقالات
اللغة:	Arabic
قواعد المعلومات:	HumanIndex
مواضيع:	البيانات الضخمة، أباتشي هادوب، انترنت الأشياء، الحوسبة السحابية، نظم إدارة قواعد البيانات العلائقية، نظم إدارة قواعد البيانات الضخمة
رابط:	http://search.mandumah.com/Record/970704

© 2021 دار المنظومة. جميع الحقوق محفوظة.
هذه المادة متاحة بناء على الإتفاق الموقع مع أصحاب حقوق النشر، علما أن جميع حقوق النشر محفوظة. يمكنك تحميل أو طباعة هذه المادة للاستخدام الشخصي فقط، ويمنع النسخ أو التحويل أو النشر عبر أي وسيلة (مثل مواقع الانترنت أو البريد الالكتروني) دون تصريح خطي من أصحاب حقوق النشر أو دار المنظومة.

نظم إدارة قواعد البيانات الضخمة:
دراسة حالة لنظام أباتشي هادوب Apache Hadoop

د. أحمد فايز أحمد سيد
مدرس تكنولوجيا المعلومات بقسم المكتبات و المعلومات
بكلية الآداب و العلوم الإنسانية
جامعة قناة السويس، مصر

مستخلص:
لقد نتجت البيانات الضخمة عن الاستخدام المتزايد لشبكة الانترنت وشبكات التواصل الاجتماعي والأجهزة المحمولة وإجراء الأبحاث العلمية المختلفة والحوسبة السحابية وغيرها، واصبح مع هذه الزيادة الهائلة في البيانات هناك صعوبة في تحليل ومعالجة هذه البيانات بالنظم التقليدية المعروفة، مما أدى إلى استحداث طرق جديدة وفعالة للتنقيب عن البيانات واسترجاعها وتحليلها من هذا الكم الضخم من البيانات سواء كانت هذه البيانات مهيكلة أو غير مهيكلة. وتهدف هذه الدراسة إلى تحليل ووصف نظام إدارة قواعد البيانات الضخمة (هادوب) ومقارنته بنظم إدارة قواعد البيانات التقليدية. ولقد تم اتباع منهج دراسة الحالة لدراسة ماهية نظام إدارة قواعد البيانات الضخمة (هادوب) ثم تطوره ومكوناته، ثم المنهج المقارن للمقارنة بين نظم إدارة قواعد البيانات التقليدية ونظم إدارة قواعد البيانات الضخمة، واستخدم الباحث الإنتاج الفكري لمراجعة كل ما تم دراسته حول البيانات الضخمة في مجال المكتبات وعلم المعلومات وكذا نظم إدارة قواعد البيانات الضخمة، واستخدم الباحث الإنتاج الفكري لمراجعة كل ما تم دراسته حول البيانات الضخمة في مجال المكتبات وعلم المعلومات وكذا نظم إدارة قواعد البيانات الضخمة، ومدى استخدامها في المكتبات المختلفة وكيفية معالجتها.

الكلمات المفتاحية: البيانات الضخمة، أباتشي هادوب، انترنت الأشياء، الحوسبة السحابية، نظم إدارة قواعد البيانات العلائقية، نظم إدارة قواعد البيانات الضخمة.

أولا: الإطار المنهجي للدراسة

1/1 مقدمة

تشكل البيانات الضخمة تحديا كبيرا للمختصين لجمعها وتحليلها وقراءتها واستخراج النتائج منها، حيث تتميز هذه البيانات الضخمة بتعدد مستويات إنتاجها وتداولها وخلال فترات وجيزة وسريعة، كما تأتي من مصادر وأشكال متنوعة، هذا بالإضافة إلى اختلاف درجة مصداقية هذه البيانات، مما يتطلب لتحليل هذه البيانات الضخمة، التحكم في كيفية استخدام الطرق الرئيسية لاستخراجها وتحليلها وتطبيقها، فضلا عن نقاط القوة والضعف للتطبيقات والبرمجيات المختلفة التي تساعد على تحليل البيانات الكبيرة. ويوجد العديد من الأدوات والتقنيات التي تستخدم لتحليل البيانات الكبيرة، مثل: Hadoop ،MapReduce، GridGain ، HPCC، Storm، Cassandr

2/1 مشكلة الدراسة

تنبع مشكلة الدراسة من ظهور مصطلح البيانات الضخمة والذي نشأ عن تطور البيانات بشكل كبير على شبكة الإنترنت ووسائل التواصل الاجتماعي، هذا الكم من البيانات يحوي الكثير من المعلومات المفيدة جدًا ليس في مجال العلوم فقط ولكن أيضًا مفيدة لقطاع المال والأعمال، والقصور في التعامل معها يشكل عائقًا أمام محركات البحث تمامًا كما يعيق التقدم في كل تلك المجالات الأخرى. كل هذه البيانات من الصعب التعامل معها في زمن معقول نسبيًا باستخدام البرمجيات الحسابية العادية للتحليل والعرض ونظم إدارة البيانات على الحاسبات الشخصية. وتحتاج لأنواع جديدة من البرامج وأساليب التحليل والحوسبة السحابية لكشف المعلومات القيمة التي تحتويها وتمكننا من فهم البشر والكوكب كله، وتوقع تطورات وعواقب ما يحدث، لذا تحاول هذه الدراسة رصد بدايات البيانات الضخمة وتعريفها وملامحها، ودراسة تعريف الهادوب وتطوره كنظام لإدارة قواعد البيانات الضخمة ثم مقارنة نظم إدارة قواعد البيانات التقليدية بنظم إدارة قواعد البيانات الضخمة.

3/1 أهمية الدراسة

للبيانات الضخمة أهمية كبرى فهي تقدم ميزة تنافسية عالية للمكتبات ومراكز المعلومات اذا استطاعت الاستفادة منها ومعالجتها لأنها تقدم فهما أعمق للمستفيدين منها ومتطلباتهم، وتساعد في اتخاذ القرارات المناسبة والملائمة داخل المكتبات بطريقة أكثر فعالية وذلك بناء على

المعلومات المستخرجة من قواعد بيانات المكتبات وبالتالي زيادة الكفاءة وتقديم خدمات أفضل. وتتمثل أهمية الدراسة في:

أ) تعد هذه الدراسة من أولى الدراسات التي تتناول موضوع البيانات الضخمة باللغة العربية بشكل عام، وبنظم إدارة قواعد البيانات الضخمة هادوب بشكل خاص.

ب) مصطلح البيانات الضخمة في بداياته وخاصة بالعالم العربي، ويحتاج للعديد من الدراسات التي تتناول المصطلح وتطبيقاته في مجال المكتبات والمعلومات.

ج) أن توضح الدراسة تعريفات البيانات الضخمة وتطورها وملامحها التي يمكن الاستفادة منها في التخصصات المختلفة.

د) أن تبين الدراسة نظام هادوب والفروق بين نظم إدارة قواعد البيانات التقليدية ونظم إدارة قواعد البيانات الضخمة.

4/1 أهداف الدراسة

تهدف الدراسة إلى تحليل ووصف نظام إدارة قواعد البيانات الضخمة (هادوب)، وتتمثل الأهداف الفرعية في:

1) توضيح ماهية البيانات الضخمة من خلال تعريفها وبداياتها وتطورها.
2) شرح الملامح المميزة للبيانات الضخمة، والمصادر الرئيسية لها.
3) دراسة تعريف نظام إدارة قواعد البيانات الضخمة هادوب وتطوره.
4) تحليل مكونات نظام إدارة قواعد البيانات الضخمة هادوب.
5) عقد مقارنة بين نظم إدارة قواعد البيانات التقليدية والضخمة (هادوب).

5/1 تساؤلات الدراسة

هناك العديد من التساؤلات المتعلقة بالدراسة والتي تتمثل فيما يلي:

1) ما تعريف البيانات الضخمة؟ وما نظام إدارة قواعد البيانات الضخمة هادوب؟
2) متى ظهر مصطلح البيانات الضخمة ونظام هادوب؟
3) لماذا أطلق على البيانات الضخمة هذا الاسم؟
4) من ينتج ويستخدم البيانات الضخمة؟
5) كيف تستخدم نظم إدارة قواعد البيانات الضخمة هادوب؟
6) ما الفرق بين نظم إدارة قواعد البيانات التقليدية والضخمة (هادوب)؟

6/1 حدود الدراسة وأبعادها

الحدود الموضوعية: نظام إدارة قواعد البيانات الضخمة هادوب.

الحدود الزمنية: منذ بدايات ظهور مصطلح البيانات الضخمة عام 1998م، ونظام هادوب عام 1999م.

7/1 منهج الدراسة وأدواتها

فرضت طبيعة الدراسة أن يكون منهج دراسة الحالة هو المنهج الذي تتم به دراسة نظام إدارة قواعد البيانات الضخمة (هادوب)، ثم المنهج المقارن للمقارنة بين نظم إدارة قواعد البيانات التقليدية ونظم إدارة قواعد البيانات الضخمة، واستخدم الباحث الإنتاج الفكري لمراجعة كل ما تم دراسته حول البيانات الضخمة في مجال المكتبات وعلم المعلومات وكذا نظم إدارة قواعد البيانات الضخمة، والمعايشة الإلكترونية لمواقع الهادوب وطبيعه استخدامه.

8/1 الدراسات السابقة

عند البحث عن المصطلح ببوابة بنك المعرفة المصري، تم التوصل إلى (50.233) نتيجة باللغة الانجليزية، وعند تحديد المصدر بتخصص المكتبات وتكنولوجيا المعلومات، وصلت عدد نتائج البحث (2.682) مقالة علمية وفصول من كتب، وتبين أن معظم هذه النتائج تركز على تطبيق البيانات الضخمة في مجال الطب والمحاسبة والرياضيات والتعليم والكيمياء، وبعد إجراء فلترة لحصر النتائج التي تتعلق بالعلوم الاجتماعية والإنسانية وتخصص المكتبات والمعلومات، تمثلت في قواعد البيانات التالية، والتي بلغ عدد نتائج البحث فيها 452 تبدأ من عام 2013م وحتى 2018م

EBSCOhost: Academic Search Complete	14
EBSCOhost: Library, Information Science & Technology Abstracts (LISTA)	30
Elsevier: ScienceDirect Journals	185
Emerald Insight	32
IET: Digital Library	9
SAGE: Journals	54
SAGE: Knowledge	10
Wiley Online Library: Journals	118

ومن أكثر الدراسات المتعلقة بموضوع البيانات الضخمة في تخصص المكتبات والمعلومات ما يلي:

1) دراسة (Bhat, W., 2018) حول الحفظ طويل المدى للبيانات الضخمة: آفاق تكنولوجيات التخزين الحالية في المكتبات الرقمية، وتهدف هذه الدراسة إلى التحقيق من آفاق تقنيات التخزين الحالية للحفظ طويل المدى للبيانات الضخمة في المكتبات الرقمية، لذا تم إجراء مراجعة منهجية ونقدية للأدبيات ذات الصلة لاستكشاف آفاق تكنولوجيات التخزين الحالية للحفظ طويل المدى للبيانات الضخمة في المكتبات الرقمية. وقد تم البحث في قواعد البيانات الحاسوبية على شبكة الإنترنت لتحديد الأدبيات ذات الصلة المنشورة بين عامي 2000م و 2016م. وتم وضع معيار محدد لعمليات الإضافة والاستبعاد وتطبيقه في مرحلتين متميزتين لتحديد أهم الأوراق ذات الصلة بالموضوع. ولقد خلصت الدراسة إلى أن تقنيات التخزين الحالية لا تصلح للحفظ طويل المدى للبيانات الضخمة في المكتبات الرقمية. ولا يمكنها الوفاء بجميع متطلبات التخزين، ولا تخفيف النفقات المالية للمكتبات الرقمية. وتشير الدراسة أيضا إلى أن الانتقال إلى تكنولوجيا التخزين الناشئة في المكتبات الرقمية هي حل قابل للتطبيق على المدى الطويل. وتوصي الدراسة بضرورة الاستعانة بالجهود البحثية الابتكارية المستمرة في مجال تقنيات التخزين الحالية للحد من تأثير نقص التخزين على المكتبات الرقمية، وإتاحة إمكانية الاستعانة بتكنولوجيا التخزين الناشئة وتطبيقها. وفي الوقت نفسه، يتطلب بذل جهود أكثر جدية من الأكاديمين ورجال الصناعة نحو إجراء العديد من الأبحاث والدراسات التطويرية من أجل تعزيز تكنولوجيا التخزين الناشئة حتى يمكن للمكتبات الرقمية الاستعانة بها وتطبيقها في أسرع وقت ممكن. تكشف الدراسة أن المكتبات الرقمية، بالإضافة إلى تكبد نفقات مالية كبيرة، ستعاني من فقدان محتمل للمعلومات بسبب النقص في التخزين من أجل الحفاظ على البيانات الضخمة على المدى الطويل، مع استخدام تكنولوجيا التخزين الحالية. ولذلك، ينبغي على صانعي السياسات والممارسين أن يختاروا بدقة تكنولوجيا التخزين من أجل الحفاظ على البيانات الضخمة في المكتبات الرقمية على المدى الطويل.

هذا النوع من الدراسات الشاملة يتيح مساهمة جديدة في البحث العلمي، حيث تحقق في إمكانيات تكنولوجيا محرك الأقراص المغناطيسية، وتقنية محرك الأقراص الصلبة، وتقنيات الحد من البيانات للحفاظ على المدى الطويل من البيانات الضخمة في المكتبات الرقمية. وأوضحت الدراسة أن الفئات الأربعة يمكنهم التغلب على المشكلة وهم: الأكاديميين والممارسين وصناع القرار، ورجال الصناعة بعد الفهم العميق لها، وأن يكون لديهم تفاصيل تقنية لاختيار تقنيات التخزين بدقة، ومزيد من التبصر لوضع سياسات مستدامة، والفرص لمعالجة مختلف المشاكل البحثية.

2) دراسة (Zheng, L., QuHui, Y., Zhang,H.,Shi, H. and Wang, X., 2017) تناولت أمن المعلومات للمكتبات الجامعية في عصر البيانات الضخمة، ففي عصر البيانات الضخمة تنمو المعرفة والوصول إليها وتخزينها بسرعة فائقة، وحيث يتوافر في الجامعات مراكز معلومات الشبكة، فالمكتبات هي دعم مهم للتعليم والبحث. ولقد وضعت هذه الدراسة ثلاثة طرق لتوحيد الوضع الحالي للمكتبات الجامعية وقضايا أمن المعلومات، تتمثل في: تقنيات موثوقة، ومعدات متطورة وإدارات فعالة. وسيتم تدريس المعلمين والطلاب كيفية استخدام وعدم الانخداع بالمعلومات الكاذبة أيضا.

3) دراسة (Xiaodan, D. and Wei, W., 10-12 March 2017) حول مناقشة نمط خدمة المكتبة الجامعية في عصر البيانات الضخمة، حيث تنتمي موارد معلومات المكتبة تنتمي إلى نظام البيانات الضخمة، ويعد تطبيق تكنولوجيا البيانات الضخمة في خدمة المكتبة إصلاحا لإدارة المكتبة. وتدعم الأمة بقوة تطوير صناعة خدمات المعلومات. وتواجه المكتبة أيضا مع حالة خدمة المجتمع والمشاريع لتعظيم الاستفادة من البيانات. مع التطور السريع للإنترنت، أصبحنا نعيش في عصر البيانات الضخمة. وفي ظل البيئة الجديدة، تعددت مطالب المستفيدين من المعلومات المختلفة مما يتوجب متطلبات جديدة لخدمة المعلومات من المكتبة. وتشير هذه الدراسة إلى أن الابتكار في المعلومات قد يواجه تحديات نمط الخدمة التقليدية للمكتبة، ويجب اقتراح التغييرات في نمط خدمة المكتبة مع العصور.

4) دراسة (Xie, Z. and Fox, E., 2017) حول تطوير البنية التحتية السيبرانية للمكتبة لمشاركة البيانات الضخمة وإعادة استخدامها. يتيح العلم المكثف للبيانات فرصا جديدة، فضلا عن التحديات التي تواجه المكتبات البحثية. إن التحدي للبنية التحتية السيبرانية، على الرغم من أنها تكنولوجية في المقام الأول، إلا أنها تنطوي أيضا على عوامل اجتماعية واقتصادية وبشرية، وبالتالي يتطلب فهم عميق للأدوار التي ينبغي أن تلعبها المكتبات البحثية في دورة حياة البحث. تناقش هذه الدراسة الأساس المنطقي والدوافع وراء المشروع البحثي للتحقق من استراتيجيات البنية التحتية للبيانات الضخمة للمكتبة.

5) دراسة (Prindle, S. and Loos, A., 2017) حول أخلاقيات المعلومات والمكتبات الأكاديمية: خصوصية البيانات في عصر البيانات الضخمة. إن تطور ممارسات جمع البيانات في التعليم له تأثير قوي وتحويلي على ممارسات التقييم التربوي. وقد أدى الضغط المتزايد لجمع وتحليل معلومات الطالب لأغراض التقييم إلى خلق حالة فريدة للمكتبات الأكاديمية التي تعد تقليديا خصوصية بيانات المستفيد أمرا بالغ الأهمية. تسلط هذه الدراسة الضوء على التحديات المتميزة التي تواجهها مهنة المكتبات وهي تتحرك نحو الواقع الجديد

للبيانات الضخمة وتدعو بشدة إلى إجراء حوار حول هذه القضايا في الأدب المهني ومهنة المكتبات.

6) دراسة (Andresic, D. and Saloun, P., 9- 10 July 2017) حول كفاءة تحليل البيانات الضخمة على جهاز واحد باستخدام أباتشي سبارك ومكتبات خريطة التنظيم الذاتي. يستخدم أباتشي سبارك عادة كمنصة تحليلية للبيانات الضخمة على عناقيد من الحاسوبات القوية، كما أنها تستخدم في المقام الأول ذاكرة الحاسوب الرئيسية للتقييم. وتضيف محاولتنا مكتبات برامج لخرائط ذاتية التنظيم على حزمة واحدة لتحليل البيانات الضخمة، ذات كفاءة وسرعة بما فيه الكفاية حتى على جهاز حاسوب واحد قياسي. هذا النهج المبتكر يتيح تحليل البيانات الضخمة للباحثين مع الموارد المحدودة. لذا تم تأكيد الفكرة التجريبية في الدراسة وتم وصفها كدراسة حالة لاستخدام البيانات المتاحة وتحليل المشاعر من تويت المقابلة والارتباط مع بيانات البورصة.

7) دراسة (Li, J., Lu, M., Dou, G. and Wang, S., 2017) حول إطار تطبيق البيانات الضخمة وتحليل جدوى استخدامها في المكتبة، وتهدف هذه الدراسة إلى إدخال مفهوم البيانات الضخمة، وتقديم نظرة شاملة للقراء لفهم إطار تطبيق البيانات الضخمة في المكتبات، ولقد استخدم الباحثون منهج تحليل النص وطريقة التحليل الاستقرائي لفهم مفهوم البيانات الضخمة، وتلخيص التحديات والفرص المتاحة لتطبيق البيانات الضخمة في المكتبات، واقتراح إطار تطبيق البيانات الضخمة في المكتبات، ثم استخدموا طريقة المسح الاستبانة لجمع البيانات من أخصائي المكتبات لتقييم جدوى تطبيق تطبيقات البيانات الضخمة في المكتبات. وتوصلت الدراسة للعديد من النتائج منها: تشتمل التحديات الأساسية في تطبيق البيانات الضخمة في المكتبات: دقة البيانات، وتقليل حجم البيانات وضغطها، وسرية البيانات وأمنها، ونظام معالجة البيانات الضخمة والتكنولوجيا المتعلقة بها، وتتمثل الفرص المتاحة في تطبيق البيانات الضخمة في المكتبات إثراء قاعدة بيانات المكتبة، وتعزيز مهارات أمناء المكتبات، وتعزيز خدمة تبادل الإعارة بين المكتبات، وتوفير خدمات المعرفة الشخصية. ويمكن النظر في إطار تطبيق البيانات الضخمة في المكتبات من خمسة أبعاد: الموارد البشرية، وموارد الأدب، ودعم التكنولوجيا، وابتكار الخدمات، وبناء البنية التحتية. وتعتقد معظم المكتبات أنه يمكن تطبيق إطار البيانات الضخمة، وتميل إلى تطبيقها بالفعل، والعقبات الرئيسية التي تحول دون تطبيق إطار البيانات الضخمة هي الموارد البشرية ومستوى تكنولوجيا المعلومات. وتتمثل قيمة هذه الدراسة في توافر العديد من التطبيقات والحلول العملية للمكتبات لتطبيق إطار البيانات الضخمة.

8) دراسة (Hao, W, 2017) حول نظام خدمة المعلومات الشخصية للمكتبة في ظل بيئة البيانات الضخمة، والتي جاءت نتيجة تطور معلومات الشبكة، وتنوع موارد معلومات المكتبة ونموها على نطاق واسع. حيث تم في البداية تقديم تعريف لبيانات المكتبة، ثم تمت مناقشة تنظيم موارد بيانات المكتبة. وأخيرا، تم عرض عملية تصميم نظام إدارة المنظمة والخاص بنظام خدمة المعلومات الشخصية في بيئة بيانات المكتبة الضخمة. وتوصلت نتائج الدراسة إلى أن تخزين موارد البيانات في المكتبة يمكن أن يحل مشكلة توفير الموارد التي يمتلكها مصادر متعددة.

9) دراسة (Ahmed, W. and Ameen, K., 2017) حول تعريف البيانات الضخمة وقياس الاتجاهات المرتبطة بها في مجال إدارة المكتبات والمعلومات. وتهدف هذه الدراسة إلى تعريف البيانات الضخمة واستخلاص مفاهيمها. واستكشاف اتجاهات بحوث البيانات الضخمة في مجال إدارة المكتبات والمعلومات، هذا بالإضافة إلى استكشاف الاتجاهات البحثية، وتحليل الأبحاث المكشفة في شبكة المعرفة لتومسون رويترز. وتعتمد هذه الدراسة على الإنتاج الفكري والدراسة السنتوميرتك، حيث تم وضع تعريف رسمي من خلال مراجعة الإنتاج الفكري. وعلاوة على ذلك، تم إجراء تحليل سسينتوميتريك للأوراق المكشفة في شبكة المعرفة لمعهد المعلومات العلمية من تومسون رويترز لاستكشاف الاتجاهات البحثية المرتبطة بالبيانات الضخمة في مجال علم المكتبات والمعلومات، وذلك باستخدام برمجيات فوسفيور .Vosviewer. وتشير نتائج الدراسة إلى إعادة تشكيل تعريف البيانات الضخمة، وكذا اتجاهات البحث الرئيسية المرتبطة بالبيانات الضخمة. وأشار التحليل إلى "المخاطر" و"الصناعة" و"السوق" و"الجدارة الائتمانية" و"تحليلات البيانات الضخمة"، وهي أكثر الاتجاهات البحثية المتكررة المرتبطة بالبيانات الضخمة. وتلخص الدراسة ما يجب على مدير البيانات الناجح تعلمه، أولا: تحدد الدراسة تعريف البيانات الضخمة. ثانيا، تصف الدراسة اتجاهات البحث الحالية المرتبطة بالبيانات الضخمة. ثالثا، بناء على الاتجاهات المستكشفة، يسترشد مديرو البيانات والمكتبة ومتخصصي إدارة المعلومات بالمطلوب تعلمه ليكونوا مديرين ناجحين للبيانات. وحيث يتوقع في المستقبل الآلاف من مديري معرفة البيانات، لذا تعد هذه الدراسة دليلا للاتجاهات المرتبطة بالبيانات الضخمة. وتعود أصالة هذه الدراسة إلى أنها الأولى من نوعها التي توفر إعادة تشكيل لتعريف البيانات الضخمة، وهي تصور الاتجاهات البحثية المرتبطة بها في مجال إدارة المكتبات والمعلومات.

10) دراسة (Johnson, V, 2017) حول الاستفادة من خبرة المكتبات التقنية في إدارة البيانات الضخمة، ولقد كان أمناء مكتبة شل أستراليا Shell Australia هم مجتمع الدراسة في مشروع تحسين قواعد البيانات والعمليات المستخدمة في إدارة البيانات الجيوفيزيائية للشركة. تعد البيانات الجيوفيزيائية تحدي للإدارة؛ لأنها بيانات ضخمة معقدة، ولا يمكن فهرستها بالطريقة التقليدية، وأثناء المشروع يتعاون أخصائي المكتبات مع الجيوفيزيائيين ومحللي البيانات الجيوفيزيائية، وأخصائي دعم تكنولوجيا المعلومات، ومهندسي قواعد البيانات، وشملت إسهاماتهم كمهنيين في مجال علم المكتبات والمعلومات تقديم المشورة بشأن حقول واصفات البيانات المطلوبة، وتطوير المفردات المحكمة، وصياغة المسميات، وتحديد مصطلحات ومعايير البحث المطلوبة، وتحديد الفرص المتاحة للوظائف الإضافية، واختبار قواعد البيانات، وتوفير التغذية المرتدة، وتعبئة قاعدة البيانات، ووضع الإجراءات. وقد جلب المكتبيون منظورا طويل الأجل لاستعراض إدارة البيانات الجيوفيزيائية، وهو أمر ضروري لحفظ البيانات إلى ما بعد دورة الأعمال. وقد أتاح التعاون لأمناء المكتبات فرصة لمعرفة المزيد عن البيانات الجيوفيزيائية ودورة حياتها، وأتاح فرصة لإظهار قيمة مهارات علم المكتبات والمعلومات في إدارة البيانات الضخمة.

11) دراسة (Kim, Y. and Cooke, L., 2016) حول تحليل البيانات الضخمة لعمليات المكتبة العامة وخدماتها باستخدام طريقة تشيرنوف فيس Chernoff face. تهدف هذه الدراسة إلى إجراء تحليل البيانات الضخمة لعمليات المكتبات العامة وخدماتها في مدينتين من دولتين مختلفتين باستخدام طريقة تشيرنوف فيس. ولقد تم تصميم الدراسة لتقييم خدمات المكتبة من خلال تحليل تشيرنوف فيس، وتم جمع بيانات ضخمة عن المكتبات العامة في لندن وسول London and Seoul، على التوالي، ومن معهد تشارترد للمالية العامة والمحاسبة Chartered Institute of Public Finance and Accountancy وموقع الحكومة الكورية Korean لرسم تشيرنوف فيس. وحددت الدراسة المسحية العلاقة بين المتغيرات وملامح الوجه البشري. على الرغم من محدودية قدرتها على التعامل مع عدد كبير من المتغيرات (تم تحليل ثمانية في هذه الدراسة) إلا أن طريقة تشيرنوف فيس تتيح بسهولة المقارنة بين عدد كبير من حالات التحليل، ولقد استخلصت الدراسة حوالي (58) رسم تشيرنوف فيس من البيانات المنسقة باستخدام لغة البرمجة R. وكشفت الدراسة أن أداء معظم الحكومات المحلية في لندن أفضل من تلك التي في سول. وتعزى هذه النتيجة إلى أن الحكومات المحلية في لندن تقوم بتشغيل المزيد من المكتبات، وتستثمر ميزانيات أعلى، وتخصيص عدد أكبر من الموظفين، وتقتني كم أكبر من المجموعات عن الحكومات المحلية

في سول. وأسفرت هذه الإدارة عن زيادة استخدام المكتبات في لندن عن سول. وتحقق الدراسة من صحة فائدة استخدام طريقة تشيرنوف فيس لتحليل البيانات الضخمة لخدمات المكتبة. وتوفر طريقة تشيرنوف فيس لتحليل البيانات الضخمة تقنية جديدة لتقييم خدمات المكتبة، كما تقدم رؤى قد لا تكون واضحة ويمكن تمييزها باستخدام أساليب تحليلية تقليدية. وتتمثل أصالة هذه الدراسة في أنها الدراسة الأولى التي تستخدم طريقة تشيرنوف فيس لتحليل البيانات الضخمة لخدمات المكتبة.

12) دراسة (Mutula, S., 2016) حول صناعة البيانات الضخمة: التأثير على علوم المكتبات والمعلومات. تناقش الدراسة تأثير صناعة البيانات الضخمة في المكتبات وعلوم المعلومات. وتشمل الموضوعات فوائد استخدام تحليلات البيانات الضخمة في جميع المجالات الأكاديمية، ويعود النمو المتزايد للبيانات الضخمة والاهتمام بها إلى انخفاض تكلفة تخزين البيانات، والحاجة إلى أخصائي مكتبات لفهم البيانات الضخمة واستخدامها لتسهيل العمليات البحثية الأساسية. ولقد اقرت العديد من مدارس المكتبات وعلوم المعلومات في الولايات المتحدة البيانات الضخمة كمقرر لمناهجها الدراسية بما في ذلك جامعة بيتسبرج University of Pittsburgh.

13) دراسة (Chen, H., Doty, P., Mollman, C., Niu, X., Yu, J.and Zhang, T., 2015) حول تقييم المكتبة وتحليل البيانات في عصر البيانات الضخمة: الممارسة والسياسات. لقد وفرت التكنولوجيا الناشئة للمكتبات وأخصائي المكتبات طرقا وأساليب جديدة لجمع البيانات وتحليلها في عصر المساءلة لتبرير قيمتها ومساهماتها. على سبيل المثال، قام جالاجير وباور ودولار (2005) Gallagher, Bauer and Dollar بتحليل الدراسة واستخدام المجلات الإلكترونية من جميع مصادر البيانات المحتملة، واكتشفا أن المستفيدين في مكتبة ييل الطبية Yale Medical Library يفضلون الشكل الإلكتروني للمقالات عن النسخة المطبوعة، وبعد ذلك تمكنا من اتخاذ الخطوات اللازمة لضبط اشتراكات المجلة. ويدعو العديد من أخصائي المكتبة إدارة المكتبة القائمة على البيانات لتعزيز وتحديد ميزانية المكتبة المقترحة.

هذا ويتضح من الدراسات السابقة أنها تناقش الحفظ طويل المدى للبيانات الضخمة بالمكتبات الرقمية، وأمن المعلومات للمكتبات الجامعية في عصر البيانات الضخمة، وخدمات المكتبة الجامعية في عصر البيانات الضخمة، وتطوير البنية التحتية السيبرانية للمكتبة لمشاركة البيانات الضخمة وإعادة استخدامها، أخلاقيات المعلومات والمكتبات الأكاديمية في عصر البيانات الضخمة، وكفاءة تحليل البيانات الضخمة على جهاز واحد باستخدام أباتشي سبارك ومكتبات خريطة التنظيم الذاتي، وإطار تطبيق البيانات الضخمة وتحليل جدوى استخدامها في المكتبة، ونظام خدمة المعلومات الشخصية للمكتبة في ظل بيئة البيانات الضخمة، وتعريف البيانات الضخمة وقياس الاتجاهات المرتبطة بها في مجال إدارة المكتبات والمعلومات، والاستفادة من خبرة المكتبات التقنية في إدارة البيانات الضخمة، وتحليل البيانات الضخمة لعمليات المكتبة العامة وخدماتها باستخدام طريقة تشيرنوف فيس Chernoff face، وصناعة البيانات الضخمة وتأثيرها على علوم المكتبات والمعلومات، وتقييم المكتبة وتحليل البيانات في عصر البيانات الضخمة: الممارسة والسياسات، وتتشابه كل الدراسات السابقة في تعريف البيانات الضخمة ومحاولة تطبيقها في المكتبات النوعية، واختلف البعض في التركيز على الأمن والحماية، والبعض الأخر ركز على إدارة المكتبات وخدماتها في ظل بيئة البيانات الضخمة، وتختلف الدراسة الحالية عن الدراسات السابقة في تناول البيانات الضخمة بالتفصيل من حيث تعريفها وتطورها وملامحها الخمسة ومصادرها، وكذا نظام الهادوب كأحد أنظمة إدارة قواعد البيانات الضخمة مفتوح المصدر، ثم المقارنة بينه وبين نظم إدارة قواعد البيانات التقليدية.

9/1 مصطلحات الدراسة

- البيانات الضخمة: مجموعة من حزم البيانات الضخمة والمعقدة، والتي يصعب التعامل معها ومعالجتها باستخدام نظم إدارة قواعد البيانات التقليدية من حيث الحفظ والبحث والتحليل واستخراج النتائج والمشاركة والنقل في غضون فترة زمنية مقبولة لتلك البيانات؛ فهي بيانات متعددة الأنواع والمصادر بحجم يصل إلى المئات من التيرابايت أو البيتابايت أو أكثر للحزمة الواحدة من البيانات.
- ®Apache Hadoop: إن أباتشي هادوب جزءا من مشروع أباتشي برعاية مؤسسة برمجيات أباتشي Apache Software Foundation فهو منصة برمجية مفتوحة المصدر مكتوبة لغة الجافا، يدعم معالجة البيانات وتخزينها، مصمم لمعالجة مجموعات ضخمة من البيانات عبر مئات إلى الآلاف من العقد الحوسبية التي تعمل بالتوازي، فهو يوفر حلا فعالا لتخزين

مجلدات هائلة من البيانات الضخمة الموزعة ولا يتطلب صيغة محددة. ويشير هذا المصطلح لمهمتين منفصلتين ومتمايزتين يقوم بها هادوب، أولها توفير عرض النطاق الترددي العالي وإمكانية التخزين على العناقيد، والمسئول عنها نظام ملف توزيع هادوب، والثانية توزيع مجموعة ضخمة من البيانات عبر عدة خوادم والمسئول عنها MapReduce.

- نظم إدارة قواعد البيانات: عبارة عن مجموعة من البرامج الحاسوبية التي تتحكم في تنظيم وتخزين وإدارة البيانات المتاحة بقواعد البيانات، ويدير النظام العديد من قواعد البيانات كما يمكن للعديد من المستفيدين الوصول لهذه القواعد في نفس الوقت.

ثانيا: الإطار الفكري للدراسة

2/1 تمهيد

لقد كان لشبكة الإنترنت وشبكات التواصل الاجتماعي دورا بارزا في زيادة معدلات استخدام البيانات ومن ثم زيادة حجمها بل وتضخمها بشكل يصعب تحليلها واستخدامها كما كان عليه من قبل. ولقد أدى هذا إلى استحداث طرق جديدة وفعالة للتنقيب عن البيانات واسترجاعها وتحليلها من هذا الكم الضخم من البيانات سواء كانت هذه البيانات منظمة (مثل: البيانات التي تم تنظيمها وتحليلها وإداراتها بقواعد البيانات المختلفة) أو غير منظمة (مثل: رسائل البريد الإلكتروني، والفيديو، والصور، والمحتوى الخاص بوسائل التواصل الاجتماعي، وغيرها من البيانات التي ينتجها المستفيدون)، ومن ثم يمكن إتخاذ القرارات المناسبة، ولا نغفل أن هذه الطرق تحتاج إلى كوادر بشرية متخصصة يطلق عليهم علماء البيانات (data scientist) لديهم مهارات خاصة في تصنيف البيانات وتنظيمها، وتحليلها وبرمجتها واستخراج النتائج منها ثم اكتشاف حلول لها. ونظرا للوقت والجهد والتكلفة الكبيرة التي تحتاجها البيانات الضخمة لتحليلها ومعالجتها، يعتمد التقنيون على أنظمة الذكاء الاصطناعي Artificial Intelligence التي لديها القدرة على التعلم والاستنتاج ورد الفعل على أوضاع لم تبرمج في الآلة باستخدام خوارزميات معقدة للعمل عليها، بالإضافة إلى تقنيات الحوسبة السحابية لإتمام عملها.

2/2 تعريف البيانات الضخمة

على الرغم من أن مصطلح البيانات الضخمة أصبح متداولا في العديد من التخصصات منها تكنولوجيا المعلومات وأمن المعلومات وتحليل البيانات والرياضيات والصحة والتعليم والأسواق المالية والطاقة والإعلام وغيرها، إلا أنه لا يمكن أن نحدد تعريف واحد يتناسب وكل التخصصات، وعند تحليل المصطلح فنجده عبارة عن جزأين الأول هو البيانات، والبيانات: هي

الصورة الخام للمعلومات قبل عمليات الفرز والترتيب والمعالجة ولا يمكن الاستفادة منها بصورتها الأولية قبل المعالجة. ويمكن ان تقسم البيانات الخام الى ثلاثة انواع:

1) بيانات مهيكلة: هى البيانات المنظمة في جداول أو قواعد بيانات.
2) بيانات غير مهيكلة: تشكل النسبة الأكبر من البيانات، وهى البيانات التي يولدها الأشخاص يوميا من كتابات نصية وصور وفيديو ورسائل ونقرات على مواقع الانترنت... الخ.
3) بيانات شبة مهيكلة: تعتبر نوعا من البيانات المهيكلة، إلا أن البيانات لا تكون في صورة جداول أو قواعد بيانات.

والجزء الثاني هو الضخامة وتشير إلى كبر حجم البيانات، وتعدد مصادرها، وتغيرها بشكل سريع، وتنوعها، مما يصعب تخزينها ومعالجتها باستخدام قواعد البيانات التقليدية، ومن ثم فهي بحاجة إلى أدوات وبرامج معقدة وغير تقليدية لتحليل هذه البيانات.

لقد عرفت موسوعة ويبوبيديا Webopedia (Beal, V., 2017) البيانات الضخمة بأنها عبارة تستخدم للدلالة على حجم هائل من البيانات المنظمة وغير المنظمة، لدرجة من الصعب معالجتها باستخدام قواعد البيانات وتقنيات البرمجيات التقليدية. وفي معظم سيناريوهات المشاريع نجد حجم البيانات ضخم جدا أو يتحرك بسرعة كبيرة أو يتجاوز سعة المعالجة الحالية. ولقد نشأ المصطلح مع شركات بحث الويب، ويستخدم للاستعلام عن البيانات المنظمة الموزعة بشكل هائل، مثل: تتكون البيانات الضخمة والمتمثلة بالبيتابايت petabytes والتي تساوي (1.024) تيرابايت terabytes أو بالإكسابايت exabytes والتي تساوي (1.024) بيتابايت، من بليون أو ترليون من التسجيلات لملايين الأشخاص من مصادر مختلفة (مثل: الويب، المبيعات، مركز تواصل الزبائن، التواصل الاجتماعي، بيانات الهواتف المحمولة وغيرها)، فهي البيانات المنظمة الفضفاضة الغير مكتملة والتي لا يمكن الوصول إليها.

وتعرفها موسوعة انفستوبيا (Investopedia, LLC., 2017) بأنها عبارة عن نمو في حجم البيانات المنظمة وغير المنظمة، وسرعة في إنشائها وتجميعها، ونطاق تغطية كم من نقاط البيانات، وتأتي البيانات الضخمة من مصادر متعددة، وتصل في صيغ مختلفة.

أما موسوعة تكوبيديا (Techopedia Inc., 2017) فتشير إلى أن البيانات الضخمة عبارة عن عملية تستخدم عندما لا تتمكن تقنية التنقيب عن البيانات وتقنيات المعالجة التقليدية في كشف رؤى ومعنى البيانات الأساسية، فهي البيانات غير المنظمة أو الحساسة للوقت أو ببساطة البيانات

الضخمة جدا التي لا يمكن معالجتها بمحركات قواعد البيانات العلائقية، ويتطلب هذا النوع من البيانات مسار مختلف للمعالجة يطلق عليه البيانات الضخمة التي تستخدم المتوازيات الهائلة على الأجهزة المتاحة والمعدة لذلك.

في حين يعرف قاموس أكسفورد (Oxford University Press, 2017) البيانات الضخمة بأنها عبارة عن مجموعات من البيانات الهائلة، والتي يمكن تحليلها بواسطة الحاسب الآلي للكشف عن الأنماط والاتجاهات والمؤسسات وخاصة فيما يتعلق بسلوك الإنسان وتفاعله، والكثير من شركات تكنولوجيا المعلومات في إتجاهها نحو إدارة وحفظ البيانات الضخمة.

يعرف قاموس ميرام وبستر (Merriam-Webster, 2017) البيانات الضخمة بأنها عبارة عن تراكم من البيانات الضخمة للغاية والمعقدة لدرجة لا يمكن معالجتها بأدوات إدارة قواعد البيانات التقليدية، كما أوضح عدم إمكانية تحديد مدى حداثة المصطلح، ففي عام 1980م تقدم تشارلز تيلي Charles Tilly بورقة بحثية استخدم فيها البيانات الضخمة، ولم يكن يستخدمها بنفس الطريقة المستخدمة في الوقت الحالي، وإنما كان يطلقها على البيانات الضخمة للناس والتي يشير فيها إلى المؤرخين العاملين في مجالات البيانات الغنية، في حين يمكن أن تشير البيانات الضخمة في الوقت الحالي إلى مجموعات من البيانات الضخمة، أو إلى النظم والحلول المتطورة لإدارة هذا التراكم الضخم من البيانات.

ولقد أُجريت العديد من الدراسات حول أصول وبدايات مصطلح البيانات الضخمة من وجهة نظر الأكاديمين ورجال الصناعة وعلم الحاسب والاحصاء والاقتصاد، ومنها جون ماشي (Mashey, J., 1998) وآخرون بشركة سيليكون جرافيك Silicon Graphics أنتج عمل جماعي غير منشور وغير أكاديمي ذات علاقة وثيقة بالبيانات الضخمة في منتصف التسعينيات، وعادة ما تتضمن البيانات الضخمة مجموعات بيانات ذات أحجام تتخطى قدرة أدوات البرامج المعروفة والمستخدمة لاختيار البيانات وتنظيمها ومعالجتها خلال فترة زمنية مقبولة (Snijders, C., Matzat, U. and Reips, U, 2012).

ولقد ظهرت أول المراجع الأكاديمية البارزة عام 1998م بواسطة وايس واندوركيا Weiss and Indurkhya في علوم الحاسب الآلي، ودي بولد Diebold عام 2000م في الإحصاء والاقتصاد. كما انتج دوجلاس لاني بشركة جارتنر عمل بارز غير منشور وغير أكاديمي، وحاليا ظهر وانتشر مصطلح البيانات الضخمة في العديد من التخصصات (Diebold, F., 2012).

وفي تقرير بحثي ومحاضرات ذات العلاقة بالبيانات الضخمة عام 2001م (Laney, D, 2001)، وكان جو هيلرشتاين (Hellerstein, J., 2017) وهو عالِم متخصص في الحاسب الآلي يعمل في جامعة كاليفورنيا، بيركلي، الولايات المتحدة الأمريكية– من أوائل من أشاروا في نوفمبر 2008م إلى دنو 'ثورة صناعية في البيانات. في حين تحدثت مجلة الإيكونوميست (The Economist Newspaper Limited, 2010) عن 'طوفان بيانات' في أوائل عام 2010م.

عرفت شركة ميتا (جارتنر حاليا) تحديات وفرص نمو البيانات كعناصر ثلاثية الأبعاد، مثل: زيادة الحجم (كم البيانات)، والسرعة (سرعة البيانات الصادرة والواردة)، والتنوع (تنوع أنواع البيانات ومصادرها). وتستمر شركة جارتنر والعديد من الشركات الأخرى في استخدام النموذج الثلاثي "3Vs" لوصف البيانات الضخمة(Beyer, M, 2011) .

وفي عام 2011م أوضح تقرير معهد ماكينزي الدولي (Manyika, J. et.al., 2011) المكونات الرئيسية والفرعية للبيانات الضخمة، كالتالي: هي تقنيات لتحيل البيانات، مثل: تعلم الآلة، ومعالجة اللغة الطبيعية؛ وتكنولوجيا البيانات الضخمة مثل الذكاء الاصطناعي والحوسبة السحابية وقواعد البيانات؛ وإمكانية العرض مثل الرسومات والخرائط والصور وغيرها من الوسائل العارضة للبيانات.

وقامت شركة جارتنر (Gartner IT Glossary) المتخصصة في أبحاث واستشارات تقنية المعلومات عام 2012م بتحديث تعريف البيانات الضخمة ليصبح: بأنها "الأصول المعلوماتية كبيرة الأحجام وسريعة التدفق وكثيرة التنوع، والتي تتطلب طرق معالجة مجدية اقتصادياً ومبتكرة من أجل تطوير المعالجة الآلية وطرق اتخاذ القرارات ".

كما تعرفها شركة آي بي إم (IBM) يشير الذكاء الاصطناعي والموبايل ووسائل التواصل الاجتماعي وإنترنت الأشياء إلى تعقد البيانات وأشكال ومصادر جديدة للبيانات. إن تحليلات البيانات الضخمة عبارة عن استخدام التقنيات التحليلية المتقدمة لمجموعات ضخمة ومتنوعة من البيانات متضمنة البيانات المنظمة وشبة المنظمة وغير المنظمة، من مصادر متعددة وبأحجام مختلفة تتراوح ما بين التيرابايت إلى الزيتابايت. إن البيانات الضخمة مصطلح يشيرإلى مجموعات البيانات يفوق حجمها أو نوعها قدرة قواعد البيانات التقليدية العلائقية على التقاط البيانات وإدارتها ومعالجتها، وتتميز بأحد أو كل الملامح التالية: الضخامة في الحجم، والسرعة، والتنوع. وتتنوع مصادر البيانات الضخمة من أجهزة الاستشعار والأجهزة والفيديو/

الصوت، والشبكات وملفات الدخول وتطبيقات المعاملات والويب ووسائل التواصل الاجتماعي، ويولد الكثير من هذه البيانات في الوقت المناسب وعلى أوسع نطاق.

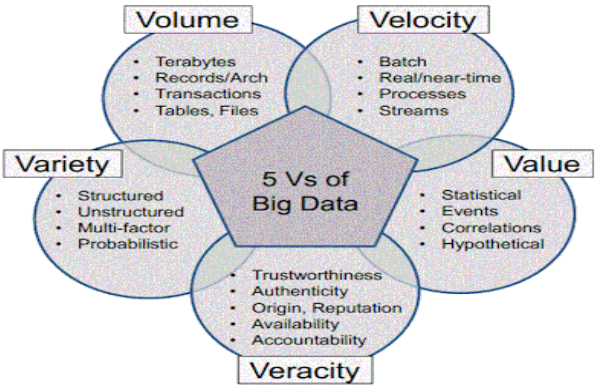
وتكمن فلسفة البيانات الضخمة في بيانات مهيكلة وشبة مهيكلة وغير مهيكلة، وتعتمد بشكل كبير على البيانات غير المهيكلة. (Dedić, N., Stanier, C., 2017)، أما بالنسبة لحجم البيانات الضخمة فهي هدف متغير باستمرار، حيث تراوحت البيانات الضخمة عام 2012م ما بين بضع عشرات من التيرابايت إلى العديد من البيتابايت من البيانات. (Everts, Sarah, 2016) وتتطلب البيانات الضخمة مجموعة من التقنيات والتكنولوجيا ذات أشكال جديدة من التكامل تكشف عن رؤى من مجموعات بيانات مختلفة ومعقدة وعلى نطاق واسع (Hashem,I., Yaqooba,A., Anuar,N., Mokhtar,S., Gania,A. and Khanb,S., 2015).

هذا ومن خلال التعريفات السابقة يمكن وضع تعريف إجرائي للبيانات الضخمة بأنها مجموعة من حزم البيانات الضخمة والمعقدة، والتي يصعب التعامل معها ومعالجتها باستخدام نظم إدارة قواعد البيانات التقليدية من حيث الحفظ والبحث والتحليل واستخراج النتائج والمشاركة والنقل في غضون فترة زمنية مقبولة لتلك البيانات؛ فهي بيانات متعددة الأنواع والمصادر بحجم يصل إلى المئات من التيرابايت أو البيتابايت أو أكثر للحزمة الواحدة من البيانات.

2/ 3 ملامح البيانات الضخمة:

يعد عصر البيانات الضخمة هو العصر التالي لعصر المعلومات وتكنولوجيا الاتصالات، حيث بدأ عصر المعلومات عام 1971م، حيث أصبح بالإمكان الوصول إلى كميات هائلة من البيانات عبر شبكة الإنترنت والحاسبات الآلية والهواتف المحمولة، ومع تزايد حجم البيانات المستمر ظهر عصر البيانات الضخمة. ولقد حددت الحكومة البريطانية عام 2012م البيانات الضخمة بأنها واحدة من ثمان تقنيات مستقبلية عظيمة (Budd, C., 2016).

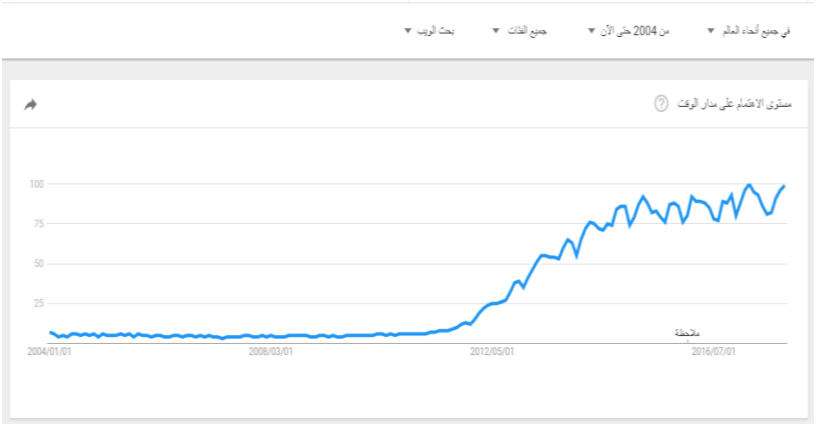
ولايزال تعريف جارتنر للبيانات الضخمة ووصفها 3Vs يستخدم على نطاق واسع، والذي ينص على أن البيانات الضخمة تمثل أصول معلومات تتميز بكبر حجمها وسرعتها العالية وتنوعها الهائل، والتي تتطلب تكنولوجيا محددة وطرق تحليلية لتحويل البيانات إلى قيمة (Mauro,A., Greco,M., and Grimaldi,M, 2016). كما أضافت بعض الشركات الكبرى مثل آي بي إم بعدا رابعا وهو مدى موثوقية البيانات لأصحاب القرار في المؤسسات (Wired, 2013) وربما بعدا خامسا وهو القيمة (GilPress, 2013)



شكل رقم (1) ملامح البيانات الضخمة 5V (Demchenko,Y., Ngo,C., Laat,C., Membrey,P. and Gordijenko,D., 2015)

1/3/2 الحجم Volume:

وهي حجم البيانات المستخرجة من مصدر ما، وهو ما يحدد قيمة وامكانات البيانات لكي تصنف من ضمن البيانات الضخمة؛ وقد يكون الخاصية الأكثر أهمية في تحليل البيانات الضخمة كما أن وصفها بالضخمة لا يحدد كمية معينة؛ فكما ذكرنا آنفا بأن الحجم يقاس عادة بالبيتابايت او بالإكسابايت. وهناك أدلة على أن البيانات الضخمة تزداد يوميًا بشكل لا يمكن أن تستطيع قواعد البيانات التقليدية وأساليب تحليل البيانات التقليدية أن تتعامل معه. من أبرز الأمثلة التي نتجت عن تأثير الإنترنت في تجميع المعلومات هو الموسوعة البريطانية تأسست عام 1768 (أي أكثر من 350 عاماً وباللغة الإنجليزية فقط)، استطاعت موسوعة ويكيبيديا التي تم تأسيسها عام 2001 وتضم 280 لغة من أن تتفوق عليها في حجم المعلومات التي وصلت إلى 30 مليون مقالة بكل اللغات (تبلغ اللغة الإنجليزية وحدها نحو 5 ملايين مقالة). (شركة أرقام بزنس إنفو، 2016) وكذا عمليات البحث على جوجل أحد المقاييس التي تُظهِر هذا: حيث لم يتزايد عدد عمليات البحث التي تشمل المصطلح حتى المدة من 2011 إلى 2012م، حيث تم نشر أربعة تقارير كبرى خلال هذه الفترة من: مبادرة الأمم المتحدة للإشارة العالمية (UN Global Pulse, 2012)، والمنتدى الاقتصادي العالمي (World Economic Forum, 2012).



شكل رقم (2) اتجاهات بحث جوجل حول البيانات الضخمة (Google Trends, 2016)

وطبقًا لشركة آي بي إم يتم إنتاج أكثر من 2.5 "كونتيليون" بايت Quintillion bytes من البيانات يوميا، أي 90% من حجم البيانات حول العالم قد تم إنتاجها في العامين الآخيرين فقط بسبب زيادة الأجهزة الرقمية والتقنيات التي تنتج هذه البيانات(Bisk Education, 2017) .

كما أشارت شركة إنتل Intel إلى أن حجم البيانات التي أنتجها البشر منذ أن بدأ يعرف الكتابة حتى عام 2003 تقدر بنحو 5 اكسابايت (Exabyte) أي 5 ملايين تيرابايت)، حيث أنه يمكن للاكسابايت الواحد تسجيل 36000 سنة من الفيديو عالي الجودة HD، لكن هذا الرقم تضاعف 500 مرة حتى عام 2012 ليصل إلى 2.7 زيتابايت Zettabyte، حيث يساوي الزيتابايت 1000 اكسابايت) (UN Global Pulse, 2012).

للبيانات الضخمة مصدرٌ آخر وهو الهواتف المحمولة والذكية، فاليوم يتجاوز عدد من الهواتف المحمولة في العالم عدد سكانه، كما ما توفره من مواد يتم مشاركتها أو كتابتها من صور ونصوص وفيديو، وهو رقم يتزايد كل ثانية نظرا لحجم المستخدمين الضخم الموجود حاليا على مواقع التواصل الاجتماعي؛ حيث طبقا لاحصائية منتصف عام 2017م وصل عدد مستخدمين الفيس بوك (2.01) بليون مستخدم نشط شهريا، و(800) مليون مستخدم لانستجرام، و(500) مليون مستخدم للينكد إن، و(1.5) بليون مشاهد لقناة اليوتيوب (Smith, C., 2017). كما يستخدم تطبيق "WhatsApp" الواتساب نحو 600 مليون مستخدم منهم 70% مستخدمين نشطين. كما يتم إرسال أكثر من 700 مليون صورة يوميًا بين مستخدمي التطبيق وأكثر من 17 مليار رسالة ترسل وتستقبل يوميًا بين المستخدمين، ولا تقف عند هذا الحد، بل

هناك احتمال وجود 2500000000000000000 محادثة في الوقت نفسه. (الجندي، محمد ، 2014)

هذه البيانات الضخمة باختصار يتم تجميعها في قواعد للبيانات والقيام بعمليات البحث والمشاركة والتحليل والمقارنة واستخلاص النتائج، ثم يتم تخزين كافة المعلومات مهما كانت صغيرة أو تافهة، مئات ومئات العمليات التي تجرى عليها والتي تحتاج بالطبع لعدد كبير من الأجهزة ومن الأشخاص للقيام بها.

2/3/2 التنوع Variety:

ويقصد بها تنوع البيانات المستخرجة، والتي تساعد المستخدمين سواء كانوا باحثين أو محللين على اختيار البيانات المناسبة لمجال بحثهم وتتضمن بيانات منظمة أو مهيكلة في قواعد بيانات وبيانات غير منظمة أو غير مهيكلة تأتي من طابعها غير الممنهج، مثل: الصور ومقاطع وتسجيلات الصوت وأشرطة الفيديو والرسائل القصيرة وسجلات المكالمات وبيانات الخرائط ...(GPS) وغيرها الكثير؛ وتتطلب وقتاً وجهداً لتهيئتها في شكل مناسب للتجهيز والتحليل (Gartner, Inc., 2011).

ويمكن توضيح البيانات المنظمة أو المهيكلة بأنها عبارة عن أجزاء صغيرة من بيانات 'صماء' مثل: الأرقام أو الحقائق، والتي وصفها أليكس 'ساندي' بنتلاند (Pentland, 2012) بأنها "فتات الخبز الرقمية. ويطلق على هذه البيانات مهيكلة أو منظمة لأنها تشكل مجموعات من المتغيرات التي يمكن توصيفها، وتصنيفها، وتنظيمها بسهولة (في أعمدة وصفوف) بغرض التحليل المنهجي. ومن الأمثلة على ذلك سجلات تفاصيل المكالمات الهاتفية التي تجمعها شركات المحمول. وتعد سجلات تفاصيل المكالمات الهاتفية واصفات للبيانات تسجل استخدام المشتركين لهواتفهم المحمولة – بما في ذلك رمز تحقيق الهوية، وموقع برج الهاتف الذي حدد مسار المكالمة لكل من المتصل والمتلقي، كحد أدنى– ووقت المكالمة ومدتها. وتحفظ شركات الاتصالات الكبرى سجلات تفاصيل المكالمات الهاتفية يوميًا(Bouillet, E. et.al., July 16 - 20, 2012) .

وهناك نوع ثاني من البيانات الضخمة يشمل: ملفات الفيديو، والمستندات، والتدوينات، وغيرها من محتوى وسائل التواصل الاجتماعي. معظم هذه البيانات 'غير مهيكلة'؛ ولذلك يصبح تحليلها أصعب. وهي تختلف عن البيانات 'المهيكلة' من حيث إنها تخضع لآراء مؤلفيها، وقد ترسم صورة خادعة لأنها غير موضوعية. على سبيل المثال، قد تكتب أنت تدوينة تذكر فيها مقاطعتك منتجًا معينًا، ولكن بيان بطاقة الائتمان الخاصة بك – استنادًا إلى المشتريات

الفعليـــــة – قـــــد يكشـــــف عـــــن وجـــــود اختيـــــار مختلـــــف . أما النوع الثالث من البيانات الضخمة فهو ما يتم عن بُعد بواسطة أجهزة استشعار رقمية، ويعكس تصرفات البشر. وقد تكون هذه الأجهزة 'عدادات ذكية' مثبّتة في المنازل لتسجيل استهلاك الكهرباء، أو صور الأقمار الاصطناعية التي يمكنها التقاط معلومات مادية –مثل الغطاء النباتي– كمؤشر على إزالة الغابات(Prashad, Lela, 2011) .

وتشتمل البيانات الضخمة بالإضافة إلى ما سبق السجلات الإدارية، وبيانات عن الأسعار أو الطقس، أو الكتب التي تم تحويلها إلى الصيغة الرقمية ، والتي يمكن أن تشكل نوعا رابعا للبيانات الضخمة .

ولقد قامت اللجنة الاقتصادية لأوروبا، بتقديم تقرير بعنوان" ماذا تعنيه البيانات الضخمة للإحصاءات الرسمية وقد أوردت فيه تصنيفا لمصادر البيانات الضخمة على النحو التالي (Shalah, R, 2017):

- المصادر الناشئة عن إدارة أحد البرامج، سواء كان برنامجا حكوميا أو غير حكومي، كالسجلات الطبية الإلكترونية وزيارات المستشفيات وسجلات التأمين والسجلات المصرفية وبنوك الطعام.
- المصادر التجارية أو ذات الصلة بالمعاملات، الناشئة عن معاملات بين كيانين، على سبيل المثال معاملات البطاقات الائتمانية والمعاملات التي تجرى عن طريق الإنترنت بوسائل منها الأجهزة المحمولة.
- مصادر شبكات أجهزة الاستشعار، على سبيل المثال، التصوير بالأقمار الصناعية، وأجهزة استشعار الطرق، وأجهزة استشعار المناخ.
- مصادر أجهزة التتبع، على سبيل المثال تتبع البيانات المستمدة من الهواتف المحمولة والنظام العالمي لتحديد المواقع.
- مصادر البيانات السلوكية، على سبيل المثال، مرات البحث على الإنترنت عن منتج أو خدمة ما أو أي نوع آخر من المعلومات، ومرات مشاهدة إحدى الصفحات على الإنترنت.
- مصادر البيانات المتعلقة بالآراء، على سبيل المثال، التعليقات على وسائط التواصل الاجتماعي.

3/3/2 السرعة Velocity:

ويقصد بها سرعة إنتاج واستخراج البيانات لتغطية الطلب عليها؛ حيث تعتبر السرعة عنصراً حاسماً في اتخاذ القرار بناء على هذه البيانات، وهو الوقت الذي نستغرقه من لحظة وصول هذه البيانات إلى لحظة الخروج بالقرار بناء عليها. وستقدم الخطة المستقبلية المتعلقة بوجود

شبكة الجيل الخامس G 5 قدرة وصول إلى البيانات بمعدل 1 جيجابايت في الثانية الواحدة بالنسبة لعشرات العاملين في الوقت نفسه وفي نفس المكتب. هذه المعلومات يتم تجميعها من خلال أنشطة المستخدمين عبر مواقع الإنترنت، ولعل هذا ما يفسر سر تقديم هذه الخدمات بالمجان، فكل نشاط بسيط يتم عبر الهواتف المحمولة أو الحاسبات الآلية، يتم تجميعها لدى الشركة، على سبيل المثال: الأشياء التي يتم البحث عنها في الإنترنت، ولقطات وأفلام الفيديو التي يتم مشاهدتها، وسماع الموسيقى أو الملفات الصوتية من المواقع الخاصة بها، والمنشورات على الفيس بوك والتغريدات على توتير أو كافة الأنشطة على المواقع الاجتماعية، عدد الساعات التي يقضيها المستخدم، وعدد الصفحات التي يطالعها، كل معلومة مهما بلغ حجمها يتم تجميعها في قاعدة بيانات ضخمة.

ولقد كانت الشركات تستخدم لمعالجة مجموعة صغيرة من البيانات المخزنة في صورة بيانات مهيكلة في قواعد بيانات عملية تسمى بال"Batch Process " حيث كان يتم تحليل كل مجموعة بيانات واحدة تلو الأخرى في انتظار وصول النتائج. مع الازدياد الضخم في حجم البيانات وسرعة تواترها أصبحت الحاجة أكثر إلحاحا الى نظام يضمن سرعة فائقة في تحليل البيانات الضخمة في الوقت اللحظي "Real Time" أو سرعة تقارب الوقت اللحظي. أدت تلك الحاجة الى ابتكار تقنيات وحلول مثل Apache و SAP HANA وHadoop وغيرها الكثير. (صلاح، رامي، 2017) إلا أن هادوب Hadoop يعد من أشهر هذه الأدوات، و"هادوب" هو برنامج أو منصة برمجية مفتوحة المصدر مكتوبة بلغة الجافا لتخزين ومعالجة البيانات الضخمة بشكل موزع مثل تخزين بيانات ضخمة على عدة أجهزة ومن ثم توزيع عملية المعالجة على هذه الأجهزة لتسريع نتيجة المعالجة. (الجعيد، عبدالرحمن عويض، 2017)

4/3/2 صحة البيانات الضخمة Veracity:

تشير إلى فوضى أو عدم الثقة في البيانات، حيث نجد مع تعدد أشكال البيانات الضخمة واختلاف جودتها وقلة التحكم في جودتها (مثل: مشاركات توتير مع وضع علامات والاختصارات والكلمات العامية وعدم الثقة في دقة المحتوى والموثوقية)، إلا أن البيانات الضخمة والتكنولوجيا التحليلية تيسر حاليا إمكانية التعامل مع هذه الأنواع من البيانات، وكثيرا ما تفتقر الأحجام الضخمة إلى الجودة أو الدقة(Marr, B., 2014) . فيمثل التمتع بالقدرة على تقييمها عنصراً جوهرياً في وضع الأساس لاتخاذ القرارات المهمة بناءً عليها. ويجب الأخذ بالاعتبار تباين هذه البيانات وعدم التأكد من صحتها في عملية اتخاذ القرار (Wired, 2013).

5/3/2 القيمة Value:

أخيرا وليس آخرا، القيمة والتي يصفها البعض بأعلى هرم البيانات الضخمة. (Banco Bilbao Vizcaya Argentaria, 2017) حيث يجب أن تكون البيانات الضخمة ذات قيمة. وهذا يعني أنك إذا كنت تريد الاستثمار في البنية التحتية اللازمة لجمع البيانات وتفسيرها على نطاق النظام، فمن المهم التأكد من أن الإحصاءات الناتجة تعتمد على بيانات دقيقة وتؤدي إلى تحسينات قياسية.(Jain, A., 2016)

ومن الأمثلة التي تساعد البيانات الضخمة المؤسسات في عمليات اتخاذ القرار بشكل أفضل (KUWAIT10'S WORLD، 2012):

1. تحويل 12 تيرابايت من التغريدات يوميا إلى تحليل لمشاعر الناس حول منتج ما.
2. تحويل 350 مليار متر من القراءات السنوية إلى تحليل لتوقعات استهلاك الطاقة.
3. تحليل تفاصيل سجلات 500 مليون مكالمة يومية في الوقت الحقيقي للتنبؤ بتوجه العملاء بشكل أسرع.
4. التدقيق على 5 ملايين معاملة تجارية يوميا لكشف أي عملية احتيال محتملة.
5. مراقبة البث المباشر لمائة كاميرا مراقبة لتحديد النقاط التي تجذب الاهتمام.

وتظهر الحسّاسات التي يُمكنها تقديم مراقبة مستمرة لحالتنا الصحية - كإحدى التقنيات التي تتقدم بسرعة في هذا المجال. وستدعم شبكة الجيل الخامس بضع مئات الآلاف من عمليات الاتصال المتزامنة لعمليات نشر ضخمة لتلك الحساسات. وفي الواقع، فالمستقبل ليس بعيدا، وقريباً ستتصل أجهزتنا مع بعضها البعض بوجود القليل من التدخّل البشري، أو حتى بدونه، فعلى سبيل المثال ستتحدث آلة الطبخ مع آلة الجلي والتنظيف، ومع السوبرماركت أيضا في كل لحظة يتم فيها تحضير وجبة، وهذا ما يُعرف بإنترنت الأشياء (Internet of things) .

وتأتي كمية معتبرة من البيانات المهمة بالنسبة للعلوم الاجتماعية من الطريقة التي نستخدمها للتعامل مع أجهزتنا، والمعلومات التي تُعطيها تلك الأجهزة عن نمط حياتنا. ففي كل مرة نشتري فيها شيئاً ما من أمازون(Amazon) ، أو نستخدم حسابنا البنكي، أو نُشغّل جهازا كهربائيا، أو هاتفنا، أو نكتب بريدا إلكترونيا، فإننا نُنشئ بيانات ستحتوي معلومات يُمكن من حيث المبدأ تحليلها. وعلى سبيل المثال يُمكن تحديد عادات الشراء، أو المواقع المتتبعة وتسجيلها. ويُمكن استخدام الرياضيات في كل مراحل هذه العملية، لكن يجب ألا نفقد أبدا البعد الأخلاقي عند القيام بذلك (Budd, C., 2016) .

4/2 مصادر البيانات الضخمة

تتعدد مصادر البيانات الضخمة عالميا بشكل كبير، كما ينتج عن هذه المصادر كما هائلا من البيانات يصعب مع مرور الوقت معالجتها وتخزينها، حيث تشير بعض التوقعات مع حلول عام 2020م، سيتم إنشاء حوالي 1.7 ميجابايت من المعلومات الجديدة في كل ثانية لكل إنسان على هذا الكوكب. كما سيزداد الكون الرقمي المتراكم من البيانات من 4.4 زيتابايت إلى حوالي 44 زيتابايت، أو 44 تريليون جيجابايت. أي كل ثانية تظهر بيانات جديدة. حيث يتم إجراء على سبيل المثال (40.000) عملية بحث في الثانية الواحدة على جوجل فقط، أي ما يساوي (3.5) مليون عملية بحث يوميا و(1.2) تريليون عملية بحث سنويا. (Kumar,V., 2017)كما سيكون لدينا أكثر من (6.1) مليار مستخدم للهواتف الذكية في جميع أنحاء العالم (تجاوز الاشتراكات الأساسية للهاتف الثابت). وسوف يمر ثلث جميع البيانات على الأقل عبر السحابة (شبكة من الخوادم الموصولة عبر الإنترنت). ومن المتوقع أن ينمو سوق Hadoop (برمجيات المصدر المفتوح للحوسبة الموزعة) بمعدل نمو سنوي مركب يبلغ 58 في المائة، وهو ما يتجاوز بليون دولار بحلول عام 2020، حيث يمكن (94٪) من مستخدمين هادوب إجراء تحليلات على كميات ضخمة من البيانات لم يكن بالإمكان تنفيذها من قبل؛ ويمكن (88٪) تحليل البيانات بطريقة أكثر تفصيلا، في حين يمكن (82٪) الاحتفاظ بمزيد من بياناتهم.(Mulcahy, M., 2017)

وفي مجال الأعمال نجد أن (6) مليون مطور على مستوى العالم يعمل على البيانات الضخمة وإجراء التحليل المتقدم عليها، ولقد وصلت نفقات تكنولوجيا البيانات الضخمة ($57) بليون دولار مع أواخر عام 2017م، كما وصل تحليل سوق الأعمال الذكي عالميا ($18.3) بليون دولار في نهاية عام 2017م، وسينمو عدد علماء البيانات خمسة أضعاف علماء البيانات التقليدين مع حلول عام 2020م، وسيقوم كل شخص بانتاج (1.7) ميجابايت من البيانات كل ثانية، لذا تحاول (85%) من الشركات أن تكون قائدة للبيانات، ولم ينجح في ذلك إلا نسبة (37%) منهم. (Hoppe, G., 2017)وسنحاول فيما يلي حصر بعض المجالات المساهمة بشكل كبير في إنتاج البيانات الضخمة، والتي تعد من الأسباب الرئيسة في تضخمها:

1/4/2 خدمات الاتصالات:

يعد تطور خدمات الاتصالات من العناصر الأساسية المساهمة في البيانات الضخمة، حيث يرتبط كم البيانات التي يتم مشاركتها وإنتاجها بالمستفيدين الذين تتوافر لديهم خدمات اتصال عالية الأداء. كما يعد استخدام الهواتف المحمولة عاملا مهما في إنتاج البيانات ومعالجة الطلبات. ويرتبط ذلك ارتباطا مباشرا بالزيادة الكبيرة في اشتراكات خدمات الاتصالات المتنقلة

التي شهدتها السنوات الأخيرة. وأدت هذه الخدمات، التي تعززها انتشار الشبكة وتوافرها، إلى أن تكون أسرع الطرق الهائلة لانتاج البيانات في أي قطاع من قطاعات تكنولوجيا المعلومات.

2/4/2 التواصل الاجتماعي:

تعد وسائل التواصل الاجتماعي المتطورة من أقوى العوامل المحفزة للبيانات غير المهيكلة، التي تنتجها الأجهزة المحمولة. كما يعزى نمو الإمكانات الهائلة للشبكات الاجتماعية إلى استخدامها على نطاق واسع ليس فقط في مجال المستهلكين ولكن أيضا في سيناريوهات الشركات، حتى أصبحت جاهزة للتسويق والاتصالات الاجتماعية. وفي الواقع إن مستقبل التعاون التكنولوجي الموحد سيكون مثل التكنولوجيا الاجتماعية، ومن الأمثلة على كمية البيانات المتدفقة من وسائل التواصل الاجتماعي، نجد أكثر من بليون مستخدم لليوتيوب، والذين تتراوح أعمارهم بين 18- 34 سنة، وهم من (88) دولة على مستوى العالم، و(76) لغة، كما يتم مشاهدة بليون ساعة يوميا.(YouTube, 2017)

3/4/2 العلم والبحث:

تعد العلوم والبحوث مجالات تسهم بشكل كبير في البيانات الضخمة، قامت مراكز البحوث خلال السنوات السابقة باستثمارات كبيرة في التقنيات من أجل إجراء التجارب ودراسة أو محاكاة الظواهر الفيزيائية. كما يتم من خلال العديد من الأنشطة العلمية بجمع وتحليل كميات هائلة من البيانات غير المهيكلة. ومثال على ذلك، يمكننا أن نأخذ في الاعتبار تجارب المسارع النووي الضخم (Large Hadron Collider (LHC، (CERN Accelerating science, 2017) والتي تنطوي على نحو 150 مليون جهاز استشعار يرسل البيانات بمعدل 40 مليون عينة / ثانية. ويبلغ مجموع المعلومات التي تنتجها التجارب الأربعة الرئيسية حوالي 700 ميجا بايت في الثانية قبل تكرارها على مواقع معالجة الأماكن الجغرافية، وقد تصل لحوالي 200 بيتا بايت بعد تكرارها. وعلى الرغم من أن المسارع النووي الضخم يعد أحد التجارب المنتجة للبيانات الضخمة، إلا ان هناك الكثير من المبادرات التي تعد نماذج للبيانات الضخمة، وتكويد الجينوم البشري، أو محاكاة المناخ أو مسح السماء الرقمي سلوان Sloan Digital Sky Survey وغيرها.

4/4/2 الحكومة والقطاع العام:

تقوم جميع الدول تقريبا في العالم بتنفيذ البرامج والتي تركز على جمع البيانات الضخمة واستخدامها. وفي الواقع، يقوم القطاع العام والجهات الحكومية بجمع كميات هائلة من البيانات لأغراض مختلفة. ولقد كانت هذه المنظمات من أكثر الجهات المنتجة للوثائق لرقمنتها وتخزينها بدعم من تكنولوجيا المعلومات، وذلك للحد من استهلاك الورق بشكل أساسي. وبناء على ذلك، فهناك كميات هائلة متراكمة من البيانات الرقمية، وهي الآن تدرس طرقا فعالة

لتحليلها وربطها، ولتقدير هذه الظاهرة بوضوح، يمكن التفكير في البيانات الرقمية الناجمة عن الرعاية الصحية، والعمليات الإدارية، والإشراف على المناطق، وهلم جرا.

في 29 مارس 2012م قدمت الإدارة الأمريكية "مبادرة بحث وتطوير البيانات الضخمة Big Data Research and Development Initiative"(Wactlar, H., 2012) : برنامج استثماري ضخم للبحث والتطوير في مجال البيانات الضخمة، يتعلق بالدفاع والصحة والطاقة والمسح الجيولوجي.

والفكرة الأساسية هي إتاحة البيانات مفتوحة المصدر التي تجمعها الحكومة للشركات والأفراد والقطاع غير الربحي، من أجل توفير "تطبيقات" مفيدة وخدمات، وتعزيز الديمقراطية والمشاركة والشفافية. (Welcome to Open Government Data) كما تركز المبادرات الأخرى التي تتخذها الحكومات المختلفة حول العالم على استخدام البيانات الكبيرة والمتنامية باستمرار التي ينتجها القطاع العام لتقديم فوائد جديدة في المجتمع. ومثال على ذلك، وفقا لتقرير ماكينزي، هناك ثلاث مجالات استراتيجية في القطاع العام، حيث يمكن للبيانات الكبيرة إنتاج قيمة مضافة لا تقل عن 150 بليون يورو في الاتحاد الأوروبي. (Manyika,J., Chui,M., Brown,B., Bughin,J., Dobbs,R., Roxburgh,C., and Byers,A., 2011)

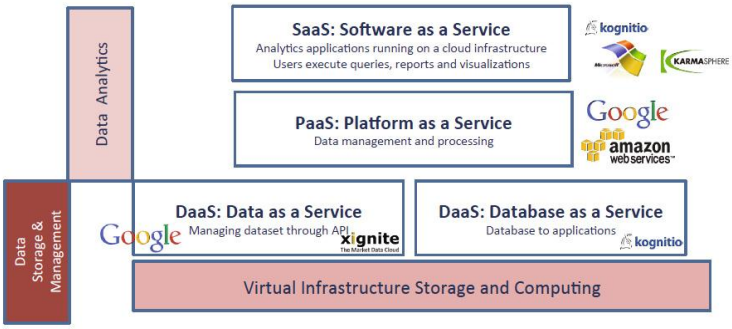
5/4/2 إنترنت الأشياء:

يعد الاتصال بين الماكينة والآلة اتجاها آخر شائعا بين المصادر، وهو ما يمثل، مرة أخرى، عنصرا أساسيا تمكينيا للاتصال اللاسلكي المتنقل. وقد تم تجهيز الآلات الحديثة مثل السيارات والقطارات ومحطات الطاقة والطائرات، وغيرها بأعداد متزايدة من أجهزة الاستشعار لجمع البيانات الضخمة. كما تغطي أجهزة الاستشعار مناطق ضخمة جدا مماثلة لأغراض الرصد / المراقبة، فضلا عن العديد من أشياء العالم الحقيقي (ويعرف أيضا باسم الأشياء) والمجهزة بأجهزة تحديد الهوية بموجات الراديو RFID والتي تقدم الكثير من المعلومات الرقمية المفيدة لمستخدميها (على سبيل المثال، الكتب في مكتبة أو منتجات فى متجر). ومن الشائع أن يكون هناك الآلاف أو حتى مئات الآلاف من أجهزة الاستشعار التي توفر معلومات عن أداء وأنشطة آلة أو موقع معين وجعلها متاحة من خلال شبكة الإنترنت. وعلى نحو مماثل، يمكن أن تتاح بيانات ملايين الأجهزة المستقلة الخاصة بتحديد الهوية بموجات الراديو وخصائصها عبر الشبكة. كما تعد العلاقة بين الأشياء وارتباطها، وعلى الرغم من أنها معقدة إلا أنها احد التكنولوجيا الناشئة التي لديها إمكانات هائلة لتغيير حياتنا. وطبقا للوضع الحالي فهناك أحد ترليون من الأجهزة المرتبطة ببعضها والتي تنتج (100) مليون من الاتصالات عبر القطاعات من آلة إلى آلة. ويعني المزيد من الاتصالات المزيد من الأتمتة والسيطرة، ثم المزيد من عمليات

التحسين والأمن. وتستثمر الشركات بقوة في هذا القطاع مما يؤدي إلى نمو كبير في البيانات المنتجة. (Manyika, J., Chui, M., Bughin,J., Dobbs,R., Bisson,P. and Marrs,A., 2013)

6/4/2 السحب:

تشارك السحب أيضا بقوة في إنتاج البيانات الضخمة، لأن الكثير من الخدمات التي تعتمد على البيانات غير المهيكلة يتم إرسالها كنمط من أنماط السحابة. فعلى سبيل المثال، تعتمد رسائل البريد الإلكتروني أو الرسائل الفورية أو العديد من الشبكات الاجتماعية على البنى التحتية السحابية لأنها توفر الطريقة الوحيدة التي تتيح لعدد كبير من المستفيدين الوصول إلى هذه الخدمات من خلال شبكة الإنترنت. وعلى الجانب الآخر، هناك تطبيقات وخدمات للشركات، حيث يعتبر نموذج السحابة الخاصة من النماذج المتميزة السريعة التي تجعل من السهل وبأسعار معقولة نشر معماريات جديدة لتخزين البيانات على نطاق واسع، ودعم إنتاج البيانات ومعالجتها، وتكرارها أو تخزينها بتكلفة منخفضة. وتغطي نماذج الخدمات السحابية (البرامج كخدمة، والمنصات كخدمة، والبيانات كخدمة) ,The SaaS (Software as a Service) PaaS (Platform as a Service) and DaaS (Data as a service) معظم الاحتياجات المترابطة بمعالجة البيانات الضخمة وتتيح إمكانية تحمل فرص البيانات الضخمة.



شكل رقم (3) البيانات الضخمة والخدمات السحابية

كما يمكن النظر إلى نظام البيانات كخدمة بشكل خاص على أنه شكل من أشكال الخدمة المنظمة على غرار البرامج كخدمة أو البنية التحتية كخدمة والذين يوفروا تسهيلات لتحليل البيانات التي يوفرها مقدم خارجي لمساعدة المنظمات على فهم واستخدام الأفكار المكتسبة من مجموعات البيانات الضخمة بهدف الحصول على ميزة تنافسية. وكثيرا ما تعتمد البيانات الضخمة كخدمة على التخزين السحابي لتوفير الوصول الفعال والمرن للبيانات للمنظمة التي تمتلك المعلومات وكذلك إلى مقدم الخدمة الذي يعمل معها. هذا بالإضافة إلى أن يمكن أن

تكون خدمات البيانات السحابية أرخص طريقة لتخزين البيانات. هذا ويتأثر إنتاج البيانات الكبيرة باستخدام الأجهزة المحمولة، والشبكات الاجتماعية، وظهور خدمات سحابة مركزية البيانات. وتتأثر كل هذه العوامل بدورها بتوافر البيانات الضخمة لتطورها ونموها الديناميكي. (Scarfo,A. and Palmieri,F., 2015)

ثالثا: الدراسة التحليلية لنظم إدارة قواعد البيانات الضخمة (هادوب)

تتمثل أهمية تكنولوجيا البيانات الضخمة في توفير تحليلات أكثر دقة، والتي تؤدي بدورها إلى اتخاذ قرارات حاسمة تؤثر في زيادة الكفاءة التشغيلية، وتخفيض التكاليف، وتقليل المخاطر بالنسبة للأعمال التجارية. لذا يتطلب للاستفادة من قوة البيانات الضخمة بنية تحتية قادرة على إدارة ومعالجة كميات ضخمة من البيانات المنظمة وغير المهيكلة في الوقت المناسب، مع إمكانية حماية خصوصية البيانات وأمنها. وهناك تقنيات مختلفة في الأسواق التجارية، مثل: الأمازون، آي بي إم، ومايكروسوفت، وما إلى ذلك، تتعامل مع البيانات الضخمة. والجدير بالذكر أن هناك فئتين من تكنولوجيا معالجة البيانات الضخمة، والتي تتمثل في:

أولا: البيانات التشغيلية الكبرى: وهي النظم التي توفر قدرات تشغيلية للأعمال الزائدة التفاعلية في وقت مناسب لها، حيث يتم التقاط البيانات وتخزينها بشكل أولي، ومثال على هذه النظم MongoDB. ولقد تم تصميم نظم البيانات الضخمة NoSQL للاستفادة من أبنية الحوسبة السحابية الجديدة التي ظهرت على مدى العقد الماضي لإتاحة إمكانية تشغيل الحسابات الضخة بشكل رخيص وفعال، مما يسهل إدارة كم أعمال البيانات الضخمة التشغيليةوأرخص وأسرع في التنفيذ. ويمكن لبعض نظم NoSQL توفير نظرة عميقة على الأنماط والاتجاهات اعتمادا على البيانات الواقعية مع حد أدنى من التكويد وبدون الحاجة لعلماء البيانات والبنية التحتية الإضافية.

ثانيا: البيانات الضخمة التحليلية: وتشمل هذه النظم نظم قواعد بيانات المعالجة الضخمة المتوازية (MPP) Massively Parallel Processing ونظام MapReduce الذي يوفر قدرات تحليلية هائلة للتحليل السابق والمعقد المتعلق بمعظم أو كل البيانات، حيث يوفر MapReduce طريقة جديدة لتحليل البيانات المكملة للامكانيات التي توفرها SQL، ونظام يعتمد على MapReduce يمكنه توسيع نطاقه من خادمات فردية إلى الآلاف من الماكينات المرتفعة والمنخفضة.

1/3 مفهوم نظام إدارة قواعد البيانات الضخمة (هادوب) ومكوناته

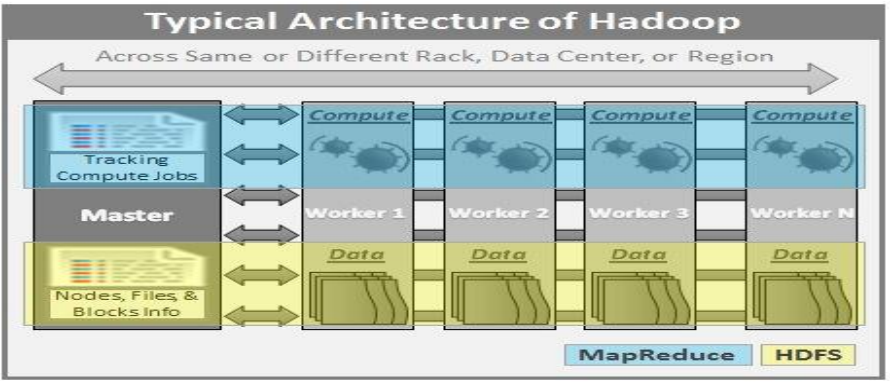
ظهر أباتشي هادوب ™ نتيجة الحاجة لمعالجة كم لا حصر له من البيانات الضخمة، فلقد كانت شبكة الإنترنت سببا رئيسيا في توليد المزيد من المعلومات يوميا، وأصبح من الصعوبة بما كان فهرسة وتكشيف أكثر من مليار صفحة من المحتوى يوميا. وللتغلب على هذه المشكلة حاولت شركة جوجل اختراع نمط جديد من معالجة البيانات والذي عُرف MapReduce. وبعد مرور عام من نشر جوجل الورقة البيضاء التي تصف MapReduce، استوحى دوج كتينج ومايك كافاريلا Cutting and Mike Cafarella من هذه الورقة هادوب لتطبيق هذه المفاهيم في إطار برمجي مفتوح المصدر لدعم توزيع مشروع محرك البحث نوتش Nutch، ولقد تم تصميم هادوب ببنية تحتية قابلة للكتابة المبسطة مرة واحدة فقط. (Rouse, Margaret , 2018)

ولم يقتصر هادوب على فهرسة محتوى الويب، وإنما أصبح يستخدم في العديد من الصناعات لمجموعة كبيرة من المهام التي تتقاسم جميعا موضوع مشترك لكثير من البيانات المتنوعة والمختلفة الأحجام والسرعة- على حد سواء منظمة أو غير منظمة. ويستخدم هادوب الآن على نطاق واسع في جميع القطاعات، بما في ذلك التمويل والإعلام والترفية والحكومة والرعاية الصحية وخدمات المعلومات وتجارة التجزئة وغيرها من الصناعات مع متطلبات البيانات الضخمة، لكن لا تزال هناك قيود على البنية التحتية للتخزين.

ولقد تم تصميم هادوب لمعالجة كميات ضخمة من البيانات والتي تبدأ من تيرابايت إلى بيتابايت وأكثر. ومع هذا الكم من البيانات الضخمة، فمن غير المنطقي أن يتم حفظ هذا الكم على القرص الصلب لحاسب آلي. وما يميز هادوب أنه مصمم لمعالجة كميات ضخمة من البيانات بكفاءة من خلال ربط العديد من أجهزة الحاسوب للعمل بشكل متوازي. وباستخدام نموذج MapReduce، يمكن لهادوب التقاط الاستعلام عن مجموعة من البيانات، ثم توزيعها، وتشغيلها على عدة عقد متوزاية، هذا التوزيع على عدة حاسبات يحل مشكلة هذا الكم من البيانات الضخمة الذي لا يتناسب على جهاز حاسب آلي واحد.

هذا وتوفر مجموعة برمجيات هادوب أساليب اقتصادية متطورة لتخزين البيانات ومعالجتها على نطاق واسع، فهي تتيح للمنظمات مرونة فائقة في الاستفادة من البيانات بجميع أشكالها وأحجامها، كما يمكن للمستفيدين تشغيل حزم من البرامج والأجهزة نظام التشغيل وبرمجيات هادوب من خلال كل العناقيد وإدارتها بالكامل من خلال واجهة واحدة.

يتضمن أباتشي هادوب نظام الملفات الموزعة Distributed File System (HDFS)، الذي يقوم بتقسيم بيانات المدخلات ويخزنها على العقد الحاسوبية، وهذا يتيح إمكانية معالجة البيانات بشكل متوازي باستخدام كافة الأجهزة في العنقود، والجديد بالذكر أن نظام الملفات الموزعة لهادوب مكتوب بلغة الجافة ويعمل على العديد من أنظمة التشغيل المختلفة.



شكل رقم (4) البنية الأساسية لهادوب (Sindol, D., 2018)

ولقد تم تصميم هادوب من البداية لاستيعاب تطبيقات لعدة نظم للملفات وإتاحتها، حيث يستخدم نظام الملفات الموزعة HDFS ونظام ملفات S3 بشكل كبير، ويتاح أيضا العدد من النظم الأخرى منها نظام ملف ماب ار MapR. (MapR Technologies, Inc., 2017)

إن أباتشي هادوب ®Apache Hadoop جزءا من مشروع أباتشي برعاية مؤسسة برمجيات أباتشي Apache Software Foundation فهو منصة برمجية مفتوحة المصدر مكتوبة لغة الجافا، يدعم معالجة البيانات وتخزينها (Rouse, Margaret , 2018)، مصمم لمعالجة مجموعات ضخمة من البيانات عبر مئات إلى الآلاف من العقد الحوسبية التي تعمل بالتوازي، فهو يوفر حلا فعالا لتخزين مجلدات هائلة من البيانات الضخمة الموزعة ولا يتطلب صيغة محددة. ويشير هذا المصطلح لمهمتين منفصلتين ومتمايزتين يقوم بها هادوب (IBM, 2018) ، أولها توفير عرض النطاق الترددي العالي وإمكانية التخزين على العناقيد، والمسئول عنها نظام ملف توزيع هادوب، والثانية توزيع مجموعة ضخمة من البيانات عبر عدة خوادم والمسئول عنها MapReduce. (AlTaie, M., 2015)

ويمكن توضيح مكونات هادوب فيما يلي:

نماذج أساسية لاطار عمل هادوب أباتشي:

1) هادوب الأساسي Hadoop Common: والمعروف kernel النواة هي مركز أساسي لنظام تشغيل الحاسوب، وتوفر الخدمات الأساسية لجميع أجزاء نظام التشغيل الأخرى، وقد يتعارض كيرنيل مع الهيكل shell (الجزء الخارجي من نظام التشغيل والذي يتفاعل مع أوامر المستفيدين)، ويستخدم المصطلحان مع نظام تشغيل يونيكس أكثر من الحاسوب المركزي لشركة آي بي إم أو نظم ميكروسوفت ويندوز. (What is Operating System, Kernel and Types) ويحتوي هادوب الأساسي على مكتبات ومتطلبات أساسية تحتاجها وحدات الهادوب الأخرى.(Hadoop Common)

2) نظام الملفات الموزعة هادوب (HDFS): هو نظام لتوزيع الملفات التي تقوم بتخزين البيانات على أجهزة سلعية تقوم بتوفير نطاق ترددي عالي وشامل لكل العناقيد . (Hanson, J. , 2011)

3) مفوض هادوب لمصادر أخرى Hadoop Yet Another Resource Negotiator (YARN): هي منصة مسئولة عن إدارة الموارد الحاسوبية في عناقيد واستخدامها في جدولة تطبيقات المستفيدين. (Vavilapalli, V., Murthy,A., Douglas,C., Agarwal,S., Konar,M., Evans,R., Graves,T., Lowe, J., Shah,H., Seth,S., Saha,B., Curino,C., O'Malley,O., Radia,S., Reed,B. and Baldeschwieler,E., 2013)

4) هادوب MapReduce : هو تطبيق لنموذج برمجة MapReduce لمعالجة البيانات على نطاق واسع. (MapReduce Tutorial, 2017)

حزم برمجيات تكميلية: تدعم هادوب مجموعة من المشروعات ذات العلاقة والتي تكمل الإمكانيات الأساسية لهادوب، وتتمثل هذه البرمجيات التكميلية في:

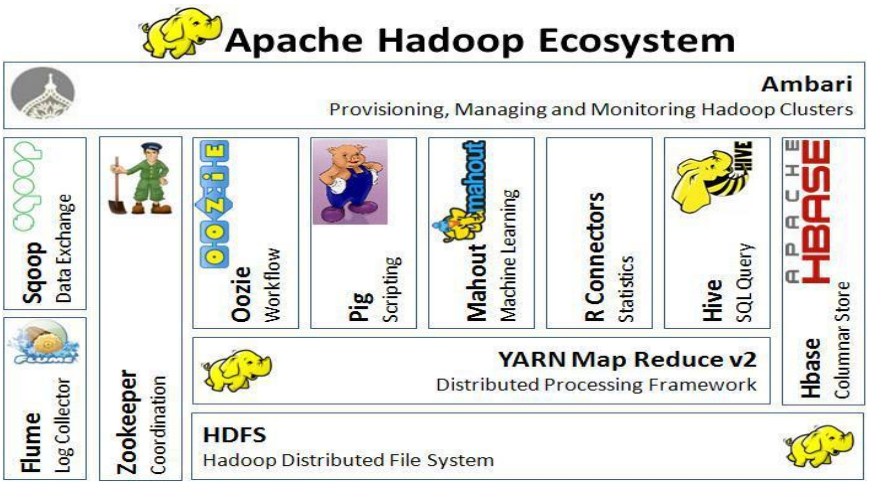
1) أباتشي فلوم Apache Flume: أداة تستخدم لجمع، ونقل كميات ضخمة من البيانات المتدفقة إلى نظام الملفات الموزعة. (The Apache Software Foundation, 2017)

2) أباتشي قاعدة اتش Apache HBase: قاعدة بيانات موزعة مفتوحة المصدر وغير علائقية.(The Apache Software Foundation, 2018)

3) أباتشي Apache Hive: مستودع بيانات يوفر تلخيص للبيانات والاستعلام والتحليل . (The Apache Software Foundation , 2014)

4) كلوديرا امبالا Cloudera Impala: قاعدة بيانات تقوم بعالجة متوازية هائلة لهادوب، وقامت بإنشائها شركة برمجيات كلوديرا، ويتم إصدارها حاليا كبرمجيات مفتوحة المصدر. (Cloudera, Inc., 2018)

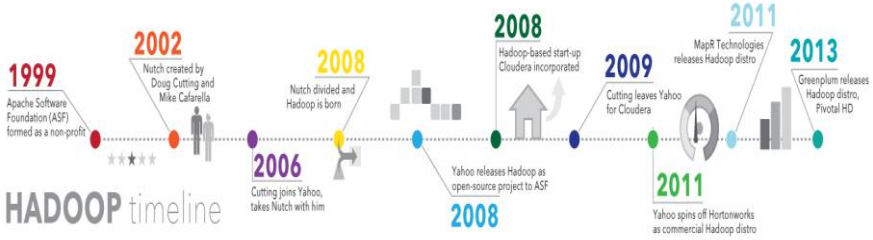
5) أباتشي أوزي Apache Oozie: نظام جدولة لسير العمل على الخادم ويقوم بإدارة وظائف هادوب. (The Apache Software Foundation, 2018)
6) أباتشي فونيكس Apache Phoenix: محرك بحث علائقي لهادوب يعتمد على قاعدة بيانات HBase وهو مفتوح المصدر، يمكنه القيام بمعالجة متوازية على نطاق واسع . (Apache Software Foundation., 2018)
7) أباتشي بيج Apache Pig: منصة رفيعة المستوى لإنشاء البرامج التي تعمل على هادوب. (The Apache Software Foundation., 2017)
8) أباتشي سكوب Apache Sqoop: أداة لنقل البيانات المجمعة بين هادوب ومستودعات البيانات المهيكلة، مثل قواعد البيانات العلائقية. (The Apache Software Foundation., 2018)
9) أباتشي سبارك Apache Spark: محرك سريع لمعالجة البيانات الضخمة قادر على تدفق ودعم قواعد بيانات SQL، وتعلم الآلة ومعالجة الرسم البياني. (The Apache Software Foundation, 2018)
10) أباتشي ستورم Apache Storm: نظام لمعالجة البيانات مفتوح المصدر. (Apache Software Foundation, 2015)
11) أباتشي زوكيبر Apache ZooKeeper: مصدر مفتوح يقوم بإنشاء خدمة التسجيل وتزامنها وتسميتها للنظم الموزعة الضخمة. (The Apache Software Foundation, 2017)



شكل رقم (5) مكونات نظام هادوب (AlTaie, M., 2015)

2/3 بدايات نظم إدارة قواعد البيانات الضخمة (هادوب) وتطوره:

لقد ظهرت شبكة الويب العالمية في أواخر التسعينيات وبدايات الألفية الثانية، ومن ثم تم إنشاء محركات البحث والفهارس للمساعدة في إيجاد المعلومات ذات الصلة وسط هذا المحتوى من النصوص، وفي السنوات الأولى، كان البشر هم من يقوموا باسترجاع النتائج، ومع نمو الويب من عشرات إلى ملايين الصفحات، كان هناك حاجة لعملية أتمتة لتنظيم واسترجاع النتائج من هذا الكم، لذا ظهرت زواحف الويب، ومعظمها كانت عبارة عن مشروعات بحثية بالجامعات، وأطلقت حينها محركات البحث الأولى (ياهو، ألتافيستا، الخ). وكان أحد هذه المشاريع محرك بحث مفتوح المصدر يسمى نوتش Nutch - من بنات أفكار دوج كتينج ومايك كافاريلا Doug Cutting and Mike Cafarella. أرادوا أن يسترجعوا نتائج البحث على شبكة الإنترنت بشكل أسرع من خلال توزيع البيانات والحسابات عبر أجهزة حاسوبية مختلفة بحيث يمكن إنجاز مهام متعددة في وقت واحد. وخلال هذه الفترة، كان جوجل مشروع محرك بحث في مرحلة التطور، وكان يعتمد على نفس المفهوم وهو تخزين ومعالجة البيانات بطريقة موزعة ومميكنة بحيث يتم استرجاع نتائج البحث بشكل أسرع. وفي عام 2006، انضم كتينج لشركة ياهو وأخذ معه مشروع نوتش، فضلا عن أفكاره التي تعتمد على عمل جوجل المبكر في أتمتة عمليات تخزين البيانات الموزعة ومعالجتها، ولقد تم تقسيم مشروع نوتش Nutch مع الإبقاء على جزء الزاحف كما هو كنوتش، وأصبح الجزء الخاص بالحوسبة الموزعة ومعالجتها هو هادوب (تم تسميته باسم ابن الفيل اللعبة الخاصة بكيتنج). وفي عام 2008م، أصدرت ياهو هادوب كمشروع مفتوح المصدر، ويديره حاليا إتحاد برمجيات أباتشي غير الربحي وهو مجتمع عالمي لمطوري البرمجيات والمساهمين. (Hadoop)



شكل رقم (6) بدايات هادوب وتطوره (Hadoop)

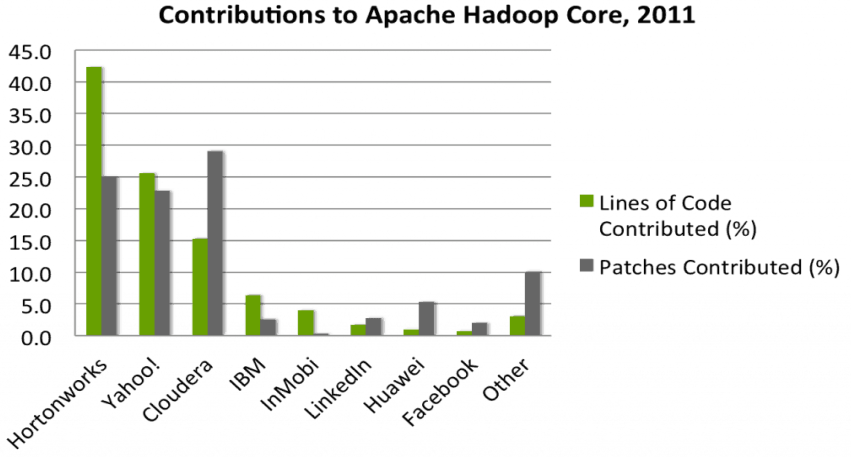
ويمكن سرد تفاصيل تطور هادوب فيما يلي:

1) عام 2003م تم إطلاق ورقة نظام ملف جوجل. (Ghemawat,S., Gobioff,H. and Leung,S., 2003)
2) عام 2004م تم ظهور معالجة البيانات المبسط على عناقيد ضخمة MapReduce. (Dean, J., and Ghemawat, S., 2004)
3) عام 2006م: خلال شهر يناير ظهرت المشروعات الفرعية لهادوب مع القوائم البريدية وويكي، وظهر هادوب من محرك نوتش، وخلال شهر فبراير من نفس العام كما انفصل MapReduce +NDFS عن أباتشي نوتش ليتم إنشاء هادوب، وتم تسمية هادوب باسم الدمية الخاصة كيتينج، وعليه أطلقت إول إصدارة من هادوب Hadoop 0.1.0 في شهر أبريل، وتم ترتيب هادوب على 188 عقدة خلال 47.9 ساعة، وفي شهر مايو قامت شركة ياهو بتشغيل 300 عنقود لأجهزة هادوب، ووصلت 600 جهاز أو آلية في شهر أكتوبر. (Woodie, A., 2015)
4) عام 2007م: في شهر ابريل قامت شركة بتشغيل عنقودين يشتملان على (1000) ماكينة، ولم يكن على صفحة دعم هادوب إلا ثلاث شركات فقط في شهر يونيو، وتم إطلاق أول هادوب يتضمن قاعدة إتش HBase في شهر أكتوبر، وقامت ياهو في نفس الشهر أيضا بإنشاء معامل بيج وتبرعت به لاتحاد برمجيات أباتشي.(White, T., 2009)
5) عام 2008م: خلال شهر يناير ظهر مفوض هادوب للمصادر الأخرى JIRA YARN وتواجدت (20) شركة على صفحة دعم هاوب، وفي شهر فبراير تحول كشاف ويب ياهو إلى هادوب، واصبح إنتاجه 10.000 عنقود لهادوب بشكل أساسي، أما في شهر مارس كانت أول قمة لهادوب، وسجل هادوب الرقم القياسي العالمي كأسرع نظام لفرز تيرابايت من البيانات الذي يعمل على 910 عقدة في العنقود، وسجل هادوب تيرابايت في 209 ثانية وذلك في شهر أبريل، وفي شهر مايو فاز هادوب بأفضل ترتيب لتيرابايت (نظام ترتيب السجلات العالمي)، فأصبح معيار لترتيب التيرابايت في شهر يوليو، وفي شهر أكتوبر أمكان تحميل 10 تيرابايت يوميا في عناقيد ياهو، وتم إنشاء موزع هادوب كلوديرا، وفي شهر نوفمبر تطبيق Google MapReduce قام بترتيب واحد تيرابايت في 68 ثانية (White, T., 2012).
6) 2009م: في شهر مارس أجرى ياهو 17 عنقود ب24.000 ماكينة، وقام هادوب بترتيب بيتابايت في شهر ابريل، لذا استخدم ياهو هادوب لترتيب تيرابايت في 62 ثانية في شهر مايو، وتم إطلاق ثاني قمة لهادوب في شهر يونيو، أما شهر يوليو فتميز بظهور عدة تطورات لهادوب وهي تغيير هادوب المركزي Core إلى هادوب الأساسي Common ، وتم

إنشاء موزع هادوب MapR، وأصبح كل من نظام توزيع ملفات هادوب و MapReduceنظم فرعية منفصلة. (The Apache Software Foundation., 2017)

7) 2010م: في شهر يناير تم إضافة دعم كيربيروس Kerberos لهادوب، وتطورت قاعدة بيانات اتش HBase أباتشي في شهر مايو، أما شهر يونيو شهد إطلاق ثالث قمة لهادوب و4000 عقدة لياهو لكل 70 بيتابايت، و2300 عنقود فيس بوك لكل 40 بيتابايت، كما شهد سبتمبر تطوير كل من Apache Hive وApache Pig . (Thinkbiganalytics.com, 2010)

8) 2011م: تطور أباتشي زوكيبر Apache Zookeeper، وساهمت شركات فيس بوك ولينكد إن وإي باي وآي بي ام IBM and eBay ,LinkedIn ,Facebook مجتمعين 200.000 خط من الأكواد وذلك خلال شهر يناير. وفي شهر مارس حاز أباتشي هادوب على أعلى جائزة من جوائز الجارديان الابتكارية الإعلامية Media Guardian Innovation Awards، وأصبح لدى ياهو 42 كيلو عقد هاوب وتخزين مئات من البيتابايت في نفس الشهر (يونيو) الذي كانت فيه قمة هادوب السنوية الثالثة بحضور 1700 شخص، أما شهر أكتوبر شهد النقاش حول أي من الشركات المساهمة أكثر في هادوب. (Harris, D., 2013)



شكل رقم (7) المساهمات في أباتشي هادوب الأساسي عام 2011م (Baldeschwieler, E., 2011)

9) 2012م: في شهر يناير انتقل مجتمع هادوب وانفصل عن MapReduce واستبدله YARN وبعد مرور ستة أشهر وفي شهر يونيو عقدت قمة هادور سان جوس San Jose بحضور 2100 شخص، وتم إتاحة نسخة أباتشي هادوب 1.0 في شهر نوفمبر. (Woodie, A., 2015)

10) 2013م: في شهر مارس عقدت قمة هادوب في أمستردام بحضور 500 شخص، كما دخل يارن YARN في نفس الشهر في خطة إنتاج ياهو، ثم عقد قمة هادوب سان جوس San Jose بحضور 2700 شخص في شهر يونيو، وفي شهر أكتوبر تم إتاحة إصدارة أباتشي هادوب 2.2. (The Apache Software Foundation., 2017)

11) 2014م: شهد شهر فبراير إتاحة إصدارة أباتشي هادوب 2.3 وأصبح مشروع أباتشي سبارك في المستوى الأعلى، أما شهر ابريل تم عقد قمة هادوب في أمستردام بحضور (750) شخص، وتم إتاحة إصدارة أباتشي هادوب 2.4 وعقد قمة هادوب بسان دوس بحضور 3200 شخص وذلك في شهر يونيو، أما شهر أغسطس شهد إتاحة إصدارة أباتشي هادوب 2.5، وأتيحت الاصدارة الأحدث في الشهر التالي لها مباشرة نوفمبر وهي إصدارة أباتشي هادوب 2.6.(DALEY, R., 2014)

12) 2015م: تمت قمة هادوب الأوربية في شهر ابريل، وبعدها بشهرين وبالتحديد في شهر يونيو تم إتاحة إصدارة أباتشي هادوب 2.7.

13) 2017م: تم إتاحة ثلاث إصدارات لاباتشي هادوب 2.8 خلال شهر مارس و 2.9 خلال شهر نوفمبر و 3.0 خلال شهر ديسمبر. (Apache Hadoop Releases, 2017)

3/3 مقارنة بين نظم إدارة قواعد البيانات التقليدية ونظم إدارة قواعد البيانات الضخمة (هادوب)

إذا ما تم المقارنة بين قواعد البيانات التقليدية وقواعد البيانات الضخمة، سنجد أن المؤسسة لديها جهاز حاسب آلي لتخزين ومعالجة البيانات. ويستعين المبرمجون ببائعي قواعد البيانات مثل أوراكل وآي بي ام وغيرهم لتخزين البيانات المتزايدة، ويتفاعل المستفيد في هذا النظام مع التطبيق، والذي بدوره يعالج جزء من تخزين البيانات وتحليلها. ويعمل هذا النظام بشكل جيد مع التطبيقات التي تعالج البيانات الأقل حجما، والتي يمكن استيعابها من قبل خوادم قاعدة البيانات القياسية، أو قدرة المعالج على معالجة البيانات. ولكن عندما يتعلق الأمر بالتعامل مع كميات ضخمة من البيانات القابلة للتطوير والزيادة، فمن الصعوبة بما كان معالجة تلك البيانات من خلال قاعدة بيانات واحدة.

لذا حاولت شركة جوجل إيجاد حلا لهذه المشكلة باستخدام خوارزمية تسمى MapReduce. تقسم هذه الخوارزمية المهمة إلى أجزاء صغيرة لاستخدامها على عدد من أجهزة الحاسبات الآلية، وتقوم بجمع النتائج منها وعند تكامل النتائج تشكل مجموعة بيانات من النتائج. وباستخدام الحل الذي قدمته شركة جوجل، قام دوج كتينج Doug Cutting وفريقه بتطوير مشروع مفتوح المصدر أطلقوا عليه هادوب، ويدير هادوب تطبيقات باستخدم خوارزمية

MapReduce، حيث تتم معالجة البيانات بشكل متوازي. باختصار، يستخدم هادوب في تطوير التطبيقات التي يمكنها إتمام التحليل الإحصائي على كميات هائلة من البيانات (Tutorials Point, 2014) .

إن أباتشي هادوب جزءا من فهو منصة برمجية مفتوحة المصدر مكتوبة لغة الجافا، يدعم معالجة البيانات وتخزينها (Rouse, Margaret , 2018)، مصمم لمعالجة مجموعات ضخمة من البيانات عبر مئات إلى الآلاف من العقد الحوسبية التي تعمل بالتوازي، فهو يوفر حلا فعالا لتخزين مجلدات هائلة من البيانات الضخمة الموزعة ولا يتطلب صيغة محددة. ويشير هذا المصطلح لمهمتين منفصلتين ومتمايزتين يقوم بها هادوب (IBM, 2018) ، أولها توفير عرض النطاق الترددي العالي وإمكانية التخزين على العناقيد، والمسئول عنها نظام ملف توزيع هادوب، والثانية توزيع مجموعة ضخمة من البيانات عبر عدة خوادم والمسئول عنها MapReduce. (AlTaie, M., 2015) ، ويوضح الباحث في الجدول التالي الفروق بين نظم إدارة قواعد البيانات العلائقية والضخمة (هادوب)

جدول رقم (1) مقارنة بين نظم إدارة قواعد البيانات العلائقية والضخمة (هادوب)

	نظم إدارة قواعد البيانات العلائقية	إدارة قواعد البيانات الضخمة (هادوب)
المصطلح	DBMS Database Management System	Apache Hadoop®
تعريفها	قواعد البيانات هي مساحات تخزين مرتبة منطقيًا لكل الأنواع المختلفة من البيانات. لكل نوع من قواعد البيانات –باستثناء غير المهيكلة– نموذج يقدم هيكلة للبيانات التي يتم التعامل معها.أنظمة إدارة قواعد البيانات هي تطبيقات (أو مكتبات) تدير قواعد البيانات بأشكالها وأحجامها وأنواعها المختلفة. (فاخوري، عبدالرحيم، 2015)	هو إطار عمل برمجي مفتوح المصدر والذي يدعم التطبيقات الموزعة للبيانات الضخمة، ومرخص وفقاً لرخصة (Apache Hadoop) كما أنه يدعم تشغيل التطبيقات على مجموعات كبيرة من الأجهزة السلعية (GFS) (Apache v2) الخاصة بجوجل وأبحاث نظام ملفات جوجل (MapReduce)
البيانات	تتوافر على الحاسبات الصغيرة والشخصية والكبيرة، وتتصف بالتكاملية وعدم التكرارية وإمكانية المشاركة عليها	القدرة على تخزين ومعالجة كميات هائلة من أي نوع من البيانات، بسرعة. مع تزايد حجم البيانات والأصناف باستمرار، وخاصة من وسائل التواصل الاجتماعي وإنترنت الأشياء.من أفضل الأمثلة على تلك البيانات أيضا التي يتم إنتاجها من الماكينة وبيانات الدخول، والمكتوبة بصيغ تخزين منها JSON, Avro and ORC

	نظم إدارة قواعد البيانات العلائقية	إدارة قواعد البيانات الضخمة (هادوب)
حجم البيانات	الجيجابايت= 1024 ميجابايت	البيتابايت= 1024 تيرابايت
شكل البيانات	عادة ما تكون البيانات مرتبة في صفوف وأعمدة	تقبل هادوب عمليا أي نوع من البيانات، فتقوم بتخزين البيانات في أشكال متنوعة
معالجة البيانات	إن قواعد البيانات العلائقية جيدة في تخزين ومعالجة مجموعات البيانات محددة نماذج البيانات مسبقا والثابتة، وبالنسبة للبيانات غير المهيكلة فتفتقر قواعد البيانات العلائقية إلى خفة الحركة وقابلية التوسع.	يجعل أباتشي هادوب من الممكن تحليل كميات ضخمة من البيانات المهيكلة وغير المهيكلة ومعالجتها بتكلفة زهيدة، ومعالجة البيانات بدون تحديد نظام مسبق لها.
المعدات	المعدات: أدوات التخزين، مثل: الأقراص المغناطيسية، والضوئية، والقرص الصلب.	القدرة الحاسوبية. هادوب نموذج للحوسبة الموزعة يعالج البيانات الضخمة بسرعة. ويستخدم العقد المتوازية والعناقيد من خلال أجهزة سلعية أساسية منخفضة التكلفة.
الإدارة البرمجية	البرامج: برنامج مدير قاعدة البيانات DBMS هو نظام يدير كافة العمليات التي تجرى على قاعدة البيانات مثل عملية انشاء الجداول او التعديل عليها اوحتى حذفها من قاعدة البيانات، وايضا عملية بناء العلاقات بين الجداول وهو عبارة عن برنامج بالغ التعقيد باهظ الثمن ويحتاج من مختص قاعدة البيانات قدراً علياً و معرفياً كبيراً للتعامل إذ أنه يتولى السيطرة على العناصر الآلية والبرمجية للقاعدة بالتعاون مع نظام التشغيل.	هادوب من السهل إدارته حيث تتيح نظم الحوسبة البديلة عالية الأداء high performance computing (HPC) بتشغيل البرامج على مجموعات كبيرة من الحواسيب، ولكنها تتطلب عادة تكوينا ثابتا للبرنامج، وتخزين للبيانات على نظام شبكة تخزين منفصلة storage area network (SAN). وتتطلب الجدولة على نظام الحوسبة البديلة عالية الأداء إدارة دقيقة، وحيث أن تنفيذ البرنامج حساس لفشل العقدة، لذا فإن إدارة عناقيد هادوب أسهل. كما يعالج هادوب بشكل غير مرئي مشاكل التحكم في الوظائف مثل فشل العقدة، فاذا كان هناك فشل في العقدة، يتأكد هادوب من تشغيل الحاسبات على عقد أخرى وتم استرداد البيانات المخزنة من هذه العقد.
المستخدم	▪ مخططو البرامج الذين يكتب برامجهم ويستخدمون إمكانيات قاعدة البيانات. ▪ مختصو قواعد البيانات، وهم المسئولون عن صيانة وتشغيل قاعدة البيانات. ▪ المستخدمون لقواعد البيانات الذين يتعاملون مع قاعدة البيانات عبر النهايات الطرفية.	اشهر مستخدمي الهادوب: Amazon, Akamai, Apple, AVG, eBay, Electronic Arts, Facebook, Google IBM, ImageShack, LinkedIn, Microsoft, The New York Times, Twitter, Yahoo
أمثلة لها	نظام إدارة قواعد بيانات علائقيّة Oracle, MySQL, Microsoft SQL	هناك العديد من الأدوات والتقنيات التي تستخدم لتحليل البيانات الكبيرة مثل Hadoop :

	نظم إدارة قواعد البيانات العلائقية	إدارة قواعد البيانات الضخمة (هادوب)
	(RDBMS Server, PostgreSQL and IBM DB2 Definition, 2017)	MapReduce , GridGain , HPCC , Storm , Cassandra
البرمجة	أداء العمليات الحسابية السابقة على كميات ضخمة من البيانات، عادة ما تكون موزعة لكن كتابة البرامج للنظم الموزعة هو أمر صعب للغاية.	يتيح نموذج البرمجة المبسط هادوب للمستخدمين إمكانية الكتابة بسرعة واختبار البرمجيات في النظم الموزعة، ويجعل من السهل كتابة البرامج الموزعة.
لغات البرمجة	SQL ، OQL ، XQuery ، SQL/XML	معظم أعمال إعداد البيانات في هادوب يتم حاليا عن طريق كتابة التعليمات البرمجية بلغات البرمجة مثل Hive, Pig or Python
قابلية التطوير	تتمثل المشكلة المتعلقة بنظم إدارة قواعد البيانات العلائقية التقليدية عدم إمكانية توسيع نطاقها لمعالجة كميات ضخمة من البيانات.	هادوب هو منصة تخزين قابلة للتطوير بدرجة كبيرة، لأنه يمكن تخزين وتوزيع مجموعات بيانات ضخمة على نطاق البيتابايت عبر مجموعات من مئات من خوادم منخفضة التكلفة تعمل بشكل متوازي.
التكلفة	تتحمل الشركات تكاليف باهظة لتوليد قيمة من البيانات حتى يتم تخزينها ومعالجتها في قواعد البيانات	هادوب هو مفتوح المصدر ويعمل على الأجهزة السلع منخفضة التكلفة، يمكن استخدام جميع أنواع البيانات، على حد سواء منظم وغير منظم، لاستخراج قيم أكثر وضوحا من كم ضخم من البيانات.
المرونة	لابد من إنشاء مخططات منظمة قبل تخزين البيانات، أي لابد من معالجتها قبل تخزينها	لم يكن لديك لإنشاء مخططات منظمة قبل تخزين البيانات. يمكنك تخزين البيانات بأي تنسيق، بما في ذلك تنسيقات شبه منظمة أو غير منظمة، ثم تحليل وتطبيق المخطط على البيانات عند القراءة ويتضمن ذلك بيانات غير منظمة مثل النصوص والصور ومقاطع الفيديو
الموثوقية	عند فشل حقل من حقول القاعدة يؤثر على قاعدة البيانات بأكملها	مجموعات الحوسبة الكبيرة عرضة لفشل العقد الفردية في الكتلة. هادوب هو في الأساس مرونة - عند فشل عقدة يعاد توجيهها إلى العقد المتبقية في الكتلة ويتم إعادة نسخ البيانات تلقائيا في التحضير لفشل عقدة في المستقبل
إمكانية الوصول	يمكن الوصول لها بشكل تفاعلي أو دفعة دفعة	يتم الوصول إليها على شكل دفعات أو مجموعات
التحديث	يمكن القراءة والكتابة عدة مرات	يتم الكتابة مرة واحدة والقراءة عدة مرات
الهيكل	يتم تصميمه على الكتابة	يتم تصميمه على القراءة

ويتضح من الجدول السابق أن نظام إدارة قواعد البيانات العلائقية يدير كافة العمليات التي تجرى على قاعدة البيانات مثل عملية انشاء الجداول او التعديل عليها او حتى حذفها من قاعدة البيانات، وايضا عملية بناء العلاقات بين الجداول وهو عبارة عن برنامج بالغ التعقيد باهظ الثمن ويحتاج من مختص قاعدة البيانات قدراً علمياً ومعرفياً كبيراً للتعامل إذ أنه يتولى السيطرة على العناصر الآلية والبرمجية للقاعدة بالتعاون مع نظام التشغيل، ويتعامل مع البيانات المهيكلة فقط والتي لا يتعدى حجمها الجيجابايت، وتعمل على أجهزة الحاسبات الشخصية ولا تتطلب مساحات تخزينية كبيرة.

أما نظام إدارة قواعد البيانات الضخمة فهو إطار عمل برمجي مفتوح المصدر والذي يدعم التطبيقات الموزعة للبيانات الضخمة، ومرخص وفقاً لرخصة (Apache Hadoop) كما أنه يدعم تشغيل التطبيقات على مجموعات كبيرة من الأجهزة السلعية (Apache GFS) (v2الخاصة بجوجل وأبحاث نظام ملفات جوجل (MapReduce) ، ولديه القدرة على تخزين ومعالجة كميات ضخمة من أي نوع من البيانات المهيكلة وغير المهيكلة، بسرعة تصل إلى البيتابايت وبتكلفة زهيدة، ومعالجة البيانات بدون تحديد نظام مسبق لها.

رابعا: النتائج والتوصيات

1/4 النتائج

توصلت الدراسة إلى العديد من النتائج التي يمكن حصرها فيما يلي:

1) هناك العديد من الدراسات الأجنبية التي تناولت البيانات الضخمة في مجال المكتبات والمعلومات إلا أنها لازالت في مهدها، حيث تناقش هذه الدراسات الحفظ طويل المدى للبيانات الضخمة بالمكتبات الرقمية، وأمن المعلومات للمكتبات الجامعية في عصر البيانات الضخمة، وخدمات المكتبة الجامعية في عصر البيانات الضخمة، وتطوير البنية التحتية السيبرانية للمكتبة لمشاركة البيانات الضخمة وإعادة استخدامها، أخلاقيات المعلومات والمكتبات الأكاديمية في عصر البيانات الضخمة، وكفاءة تحليل البيانات الضخمة على جهاز واحد باستخدام أباتشي سبارك ومكتبات خريطة التنظيم الذاتي، وإطار تطبيق البيانات الضخمة وتحليل جدوى استخدامها في المكتبة، ونظام خدمة المعلومات الشخصية للمكتبة في ظل بيئة البيانات الضخمة، وتعريف البيانات الضخمة وقياس الاتجاهات المرتبطة بها في مجال إدارة المكتبات والمعلومات، والاستفادة من خبرة المكتبات التقنية في إدارة البيانات الضخمة، وتحليل البيانات الضخمة لعمليات المكتبة العامة وخدماتها باستخدام طريقة تشيرنوف فيس Chernoff face، وصناعة البيانات الضخمة وتأثيرها على علوم المكتبات

والمعلومات، وتقييم المكتبة وتحليل البيانات في عصر البيانات الضخمة: الممارسة والسياسات.

2) التعريف الإجرائي للبيانات الضخمة بأنها مجموعة من حزم البيانات الضخمة والمعقدة، والتي يصعب التعامل معها ومعالجتها باستخدام نظم إدارة قواعد البيانات التقليدية من حيث الحفظ والبحث والتحليل واستخراج النتائج والمشاركة والنقل في غضون فترة زمنية مقبولة لتلك البيانات؛ فهي بيانات متعددة الأنواع والمصادر بحجم يصل إلى المئات من التيرابايت أو البيتابايت أو أكثر للحزمة الواحدة من البيانات.

3) تتسم البيانات الضخمة بخمس سمات وهي: حجم البيانات المستخرجة من مصدر ما والذي يقاس عادة بالبيتابايت او بالإكسابايت، وتنوع البيانات المستخرجة سواء كانت بيانات مهيكلة في قواعد بيانات وبيانات غير مهيكلة تأتي من طابعها غير الممنهج، وسرعة إنتاج واستخراج البيانات والقدرة على اتخاذ القرار بناء على هذه البيانات التي تقدر سرعتها 1 جيجابايت في الثانية الواحدة بالنسبة لعشرات العاملين في الوقت نفسه وفي نفس المكتب، وعدم الثقة في البيانات، حيث نجد مع تعدد أشكال البيانات الضخمة واختلاف جودتها وقلة التحكم في جودتها، وأخيرا القيمة أي إذا كنت تريد الاستثمار في البنية التحتية اللازمة لجمع البيانات وتفسيرها على نطاق النظام، فمن المهم التأكد من أن الإحصاءات الناتجة تعتمد على بيانات دقيقة وتؤدي إلى تحسينات قياسية.

4) تتمثل مصادر البيانات الضخمة في ستة مصادر أساسية، وهي التي تساعد في إنتاج بيانات هائلة وكذا استخدامها، ومن هذه المصادر: خدمات الاتصالات عالية الآداء والهواتف المحمولة المتاحة لدى المستفيدين، وشبكات التواصل الاجتماعي، والمؤسسات العلمية ومراكز البحوث المختلفة، والمؤسسات والمنظمات الحكومية والقطاع العام، وإنترنت الأشياء الذي ييسر الاتصال اللاسلكي المتنقل بين الآلات والأجهزة وبعضها باستخدام المجسات، والحوسبة السحابية.

5) إن أباتشي هادوب Apache Hadoop® جزءا من مشروع أباتشي برعاية مؤسسة برمجيات أباتشي Apache Software Foundation فهو منصة برمجية مفتوحة المصدر مكتوبة لغة الجافا، يدعم معالجة البيانات وتخزينها، مصمم لمعالجة مجموعات ضخمة من البيانات عبر مئات إلى الآلاف من العقد الحوسبية التي تعمل بالتوازي، فهو يوفر حلا فعالا لتخزين مجلدات هائلة من البيانات الضخمة الموزعة ولا يتطلب صيغة محددة. ويشير هذا المصطلح لمهمتين منفصلتين ومتمايزتين يقوم بها هادوب، أولها توفير عرض النطاق الترددي العالي وإمكانية التخزين على العناقيد، والمسئول عنها نظام ملف توزيع هادوب، والثانية توزيع مجموعة ضخمة من البيانات عبر عدة خوادم والمسئول عنها MapReduce.

6) يمكن تقسيم مكونات نظام إدارة قواعد البيانات الضخمة هادوب إلى نماذج أساسية تقوم بتشغيل البيانات الضخمة على الأجهزة وتوزيعها على العناقيد وإدارتها ومعالجتها وتتمثل في أربعة نماذج: هادوب الأساسي، ونظام الملفات الموزعة هادوب، ومفوض هادوب لمصادر أخرى، وهادوب MapReduce؛ وحزم برمجيات تكميلية: تكمل الإمكانيات الأساسية لهادوب، وتتمثل هذه البرمجيات التكميلية في إحدى عشر برنامجا وهم: أباتشي فلوم، وأباتشي قاعدة اتش، وأباتشي Apache Hive، وكلوديرا امبالا Cloudera Impala، وأباتشي أوزي Apache Oozie، وأباتشي فونيكس Apache Phoenix، وأباتشي بيج Apache Pig، وأباتشي سكوب Apache Sqoop، وأباتشي سبارك Apache Spark، و أباتشي ستورم Apache Storm، وأباتشي زوكيبر Apache ZooKeeper.

7) لقد اطلقت ورقة نظام ملف جوجل عام 2003م، ثم ظهرت معالجة البيانات المبسطة على عناقيد ضخمة عام 2004م، وخلال عام عام 2006م: خلال شهر يناير ظهرت المشروعات الفرعية لهادوب مع القوائم البريدية وويكي، وظهر هادوب من محرك نوتش، وخلال شهر فبراير من نفس العام كما انفصل MapReduce +NDFS عن أباتشي نوتش ليتم إنشاء هادوب، وتم تسمية هادوب باسم الدمية الخاصة كيتينج، وعليه أطلقت إول إصدارة من هادوب Hadoop 0.1.0 في شهر أبريل، وتم ترتيب هادوب على 188 عقدة خلال 47.9 ساعة، وفي شهر مايو قامت شركة ياهو بتشغيل 300 عنقود لأجهزة هادوب، ووصلت 600 جهاز أو آلية في شهر أكتوبر، وظل يتطور حتى تم إتاحة ثلاث إصدارات لاباتشي هادوب 2.8 خلال شهر مارس و 2.9 خلال شهر نوفمبر و 3.0 خلال شهر ديسمبر من عام 2017م.

8) نظام إدارة قواعد البيانات العلائقية يدير كافة العمليات التي تجرى على قاعدة البيانات مثل عملية انشاء الجداول او التعديل عليها او حتى حذفها من قاعدة البيانات، وايضا عملية بناء العلاقات بين الجداول وهو عبارة عن برنامج بالغ التعقيد باهظ الثمن ويحتاج من مختص قاعدة البيانات قدراً علمياً ومعرفياً كبيراً للتعامل إذ أنه يتولى السيطرة على العناصر الآلية والبرمجية للقاعدة بالتعاون مع نظام التشغيل، ويتعامل مع البيانات المهيكلة فقط والتي لا يتعدى حجمها الجيجابايت، وتعمل على أجهزة الحاسبات الشخصية ولا تتطلب مساحات تخزينية كبيرة.

9) نظام إدارة قواعد البيانات الضخمة فهو إطار عمل برمجي مفتوح المصدر والذي يدعم التطبيقات الموزعة للبيانات الضخمة، ومرخص وفقاً لرخصة (Apache Hadoop) كما أنه يدعم تشغيل التطبيقات على مجموعات كبيرة من الأجهزة السلعية (Apache GFS) (v2الخاصة بجوجل وأبحاث نظام ملفات جوجل (MapReduce) ، ولديه القدرة على

تخزين ومعالجة كميات ضخمة من أي نوع من البيانات المهيكلة وغير المهيكلة، بسرعة تصل إلى البيتابايت وبتكلفة زهيدة، ومعالجة البيانات بدون تحديد نظام مسبق لها.

2/4 التوصيات

ولقد توصلت الدراسة إلى العديد من الدراسات، من أهمها:

1. إجراء العديد من الدراسات الأكاديمية حول البيانات الضخمة ومدى استخدامها في المكتبات المختلفة وكيفية معالجتها.
2. تقرير البيانات الضخمة كمقرر تدريسي في أقسام المكتبات والمعلومات.
3. عقد المزيد من الندوات التعريفية وورش العمل والمؤتمرات حول البيانات الضخمة وتطبيقاتها في المجالات المتنوعة.
4. تدريب أخصائي المكتبات على نظم إدارة قواعد البيانات الضخمة.
5. التعاون بين المكتبات فيما يتعلق بالبيانات الضخمة وتحليلها وإدارتها.

دراسة أمن وحماية قواعد البيانات الضخمة.

Bibliography

- Apache Software Foundation. (2015). Apache Storm. Retrieved 01 24, 2018, from http://storm.apache.org/
- Beal, V. (2017). Big Data. Retrieved 11 19, 2017, from Webopedia: https://www.webopedia.com/TERM/B/big_data.html
- DALEY, R. (2014, 06 06). Hadoop Summit 2014 – Big Data Keeps Getting Bigger. Retrieved 01 25, 2018, from Hitachi Vantara : http://www.pentaho.com/blog/2014/06/06/hadoop-summit-2014-big-data-keeps-getting-bigger
- Harris, D. (2013, 03 04). The history of Hadoop: From 4 nodes to the future of data. Retrieved 01 25, 2018, from Gigaom: https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/
- Sindol, D. (2018). Getting Started with HDInsight - Part 1 - Introduction to HDInsight. Retrieved 01 24, 2018, from SQL Server Indexing Tips and Tricks --- Free MSSQLTips Webcast!: https://www.mssqltips.com/sqlservertip/3413/getting-started-with-hdinsight--part-1--introduction-to-hdinsight/
- The Apache Software Foundation. (2018). Apache Oozie Workflow Scheduler for Hadoop. Retrieved 01 24, 2018, from Oozie: http://oozie.apache.org/
- The Apache Software Foundation. (2018). Apache Spark™. Retrieved 01 24, 2018, from Apache Spark™ : https://spark.apache.org/
- Ahmed, W. and Ameen, K. (2017). Defining big data and measuring its associated trends in the field of information and library management. Library Hi Tech News, 34(9), 21-24. Retrieved 01 31, 2018, from http://mplb1ci.ekb.eg/MuseProxyID=1103/MuseSessionID=0015z4e/MuseProtocol=http/MuseHost=www.emeraldinsight.com/MusePath/doi/full/10.1108/LHTN-05-2017-0035
- AlTaie, M. (2015, 09 07). Hadoop as the Backbone of Big Data Technologies. Retrieved 01 24, 2018, from Big Data: http://blog.agroknow.com/?p=3779
- Andresic, D. and Saloun, P. (9- 10 July 2017). Efficient big data analysis on a single machine using apache spark and self-organizing map libraries. In S. M. Bielikova M. (Ed.), 12th International Workshop on Semantic and Social Media Adaptation and Personalization,

SMAP 2017 (pp. 1-5). Bratislava; Slovakia: Institute of Electrical and Electronics Engineers Inc. doi:10.1109/SMAP.2017.8022657

- Apache Hadoop Releases. (2017, 12 18). Retrieved 01 25, 2018, from The Apache Software Foundation: http://hadoop.apache.org/releases.pdf
- Apache Software Foundation. (2018). Apache Phoenix: OLTP and operational analytics for Apache Hadoop. Retrieved 01 24, 2018, from https://phoenix.apache.org/
- Baldeschwieler, E. (2011, 10 07). REALITY CHECK: CONTRIBUTIONS TO APACHE HADOOP. Retrieved 01 25, 2018, from Hortonworks Inc: https://hortonworks.com/blog/reality-check-contributions-to-apache-hadoop/
- Banco Bilbao Vizcaya Argentaria. (2017). The five V's of big data. Retrieved 01 11, 2018, from https://www.bbva.com/en/five-vs-big-data/
- Beyer, M. (2011, 06 27). Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Retrieved 11 19, 2017, from https://www.gartner.com/newsroom/id/1731916
- Bhat, W. (2018, 01 29). Long-term preservation of big data: Prospects of current storage technologies in digital libraries. Library Hi Tech. Retrieved 01 31, 2018, from http://0015z4c.1103.y.https.www.scopus.com.mplb1ci.ekb.eg/record/display.uri?eid=2-s2.0-85040860429&origin=inward&txGid=4e3326ba8a1c7d411a46fa7928edfb04
- Bisk Education. (2017). What is Big Data? Retrieved 11 20, 2017, from https://www.villanovau.com/resources/bi/what-is-big-data/
- Bouillet, E. et.al. (July 16 - 20, 2012). Processing 6 billion CDRs/day: from research to production (experience report). Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems, (pp. 264-267). Berlin, Germany. Retrieved 11 20, 2017, from https://dl.acm.org/citation.cfm?doid=2335484.2335513
- Budd, C. (2016, 05 16). Big Data. Plus Magazine. Retrieved 05 16, 2017, from https://plus.maths.org/content/big-data
- Budd, C. (2016, 05 16). Eight great reasons to do maths. Retrieved 11 20, 2017, from https://plus.maths.org/content/great-eight
- CERN Accelerating science. (2017). The Large Hadron Collider. Retrieved 01 11, 2018, from https://home.cern/topics/large-hadron-collider

- Chen, H., Doty, P., Mollman, C., Niu, X., Yu, J.and Zhang, T. (2015). Library assessment and data analytics in the big data era: Practice and policies. Proceedings of the Association for Information Science and Technology , 52(1), 1–4. doi:10.1002/pra2.2015.14505201002 V
- Cloudera, Inc. (2018). Apache Impala (incubating). Retrieved 01 24, 2018, from https://www.cloudera.com/products/open-source/apache-hadoop/impala.html
- Dean, J., and Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation. an Francisco, CA,. Retrieved 01 25, 2018, from https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
- Dedić, N., Stanier, C. (2017). Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. Innovations in Enterprise Information Systems Management and Engineering. ERP Future 2016. Lecture Notes in Business Information Processing. 285. Springer, Cham. Retrieved 11 18, 2017, from https://link.springer.com/chapter/10.1007/978-3-319-58801-8_10
- Demchenko,Y., Ngo,C., Laat,C., Membrey,P. and Gordijenko,D. (2015, 03 24). Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. Retrieved 01 11, 2018, from https://www.researchgate.net/publication/273945634_Big_Security_for_Big_Data_Addressing_Security_Challenges_for_the_Big_Data_Infrastructure
- Diebold, F. (2012). On the Origin(s) and Development of the Term "Big Data". Penn Institute for Economic Research, Department of Economics . Philadelphia: University of Pennsylvania . Retrieved 11 19, 2017, from https://poseidon01.ssrn.com/delivery.php?ID=199020086001026016121011113124125099104015006077091033071075020106070094067003066103097114000125006036111067121065071127066124062015046052031092099125019014068096048075035098094090121068082069013101085108074006O
- Everts, Sarah. (2016). Information Overload. Distillations, 2(2), 26–33. Retrieved 11 18, 2017, from https://www.chemheritage.org/distillations/magazine/information-overload

- Gartner IT Glossary. (n.d.). What is Big Data? Retrieved 11 18, 2017, from https://research.gartner.com/definition-whatis-big-data
- Gartner, Inc. (2011). Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Press Release. Retrieved 11 20, 2017, from https://www.gartner.com/newsroom/id/1731916
- Ghemawat,S., Gobioff,H. and Leung,S. (2003). The Google File System. SOSP'03. New York, USA: Bolton Landing. Retrieved 01 25, 2018, from https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
- GilPress. (2013, 07 25). Big Data 3 V's: Volume, Variety, Velocity (Infographic). Retrieved 11 20, 2017, from https://whatsthebigdata.com/2013/07/25/big-data-3-vs-volume-variety-velocity-infographic/
- Google Trends. (2016). Big Datat. Retrieved 11 20, 2017, from Google Trends: https://trends.google.com/trends/explore?date=all&q=big%20data
- Hadoop. (n.d.). Retrieved 01 24, 2018, from SAS | The Power to Know: https://www.sas.com/en_us/insights/big-data/hadoop.html
- Hadoop Common. (n.d.). Retrieved 01 24, 2018, from Techopedia: https://www.techopedia.com/definition/30427/hadoop-common
- Hanson, J. . (2011, 11 01). An introduction to the Hadoop Distributed File System. Retrieved 01 24, 2018, from IBM Developers Work: https://www.ibm.com/developerworks/library/wa-introhdfs/index.html
- Hao, W. (2017, 02). Personalized information service system of a library under the big data environment. AGRO FOOD INDUSTRY HI-TECH, 28(1), 1701-1704. Retrieved 01 25, 2018, from http://mplb1ci.ekb.eg/MuseProxyID=1106/MuseSessionID=0015ebw/MuseProtocol=http/MuseHost=apps.webofknowledge.com/MusePath/full_record.do?product=WOS&search_mode=GeneralSearch&qid=1&SID=D4QFZVcZLBnCboz7GKl&page=1&doc=3
- Hashem,I., Yaqooba,A., Anuar,N., Mokhtar,S., Gania,A. and Khanb,S. (2015, 01). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98-115. Retrieved 11 18, 2017, from http://www.sciencedirect.com/science/article/pii/S0306437914001288?via%3Dihub

- Hellerstein, J. (2017). The Commoditization of Massive Data Analysis. Retrieved 11 20, 2017, from O'Reilly Media, Inc.: http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html
- Hoppe, G. (2017, 05 09). 10 Surprising Big Data Statistics. Retrieved 01 12, 2018, from Business Intelligence: https://blog.capterra.com/10-surprising-big-data-statistics/
- IBM. (2018). Apache Hadoop: built for big data, insights, and innovation. Retrieved 01 12, 2018, from IBM Analytics: https://www.ibm.com/analytics/hadoop
- IBM. (n.d.). What is Big Data Analytics? Retrieved 11 18, 2017, from IBM Analytics: https://www.ibm.com/analytics/hadoop/big-data-analytics
- Investopedia, LLC. (2017). Big Data. Retrieved 11 19, 2017, from Investopedia: https://www.investopedia.com/terms/b/big-data.asp
- Jain, A. (2016, 09 17). The 5 Vs of Big Data. Retrieved 01 11, 2018, from https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/
- Johnson, V. (2017). Leveraging Technical Library Expertise for Big Data Management. JOURNAL OF THE AUSTRALIAN LIBRARY AND INFORMATION ASSOCIATION, 66(3), 271-286. doi:10.1080/24750158.2017.1356982
- Kim, Y. and Cooke, L. (2016, 11 30). Big data analysis of public library operations and services by using the Chernoff face method. Journal of Documentation, 73(3), 466-480. doi:https://doi.org/10.1108/JD-08-2016-0098
- Kumar,V. (2017, 03 26). Big Data Facts. Retrieved 01 12, 2018, from AnalyticsWeek: https://analyticsweek.com/content/big-data-facts/
- KUWAIT10'S WORLD. (2012, 08 26). ما هو Big Data (بيئات البيانات الضخمة)؟. Retrieved 11 20, 2017, from https://kuwait10.net/2012/08/26/big-data/
- Laney, D. (2001, 02 16). 3D Data Management: Controlling Data Volume, Velocity and Variety. Retrieved 11 19, 2017, from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
- Li, J., Lu, M., Dou, G. and Wang, S. (2017, 06 05). Big data application framework and its feasibility analysis in library. Information Discovery and Delivery, 45(4), 161-168. Retrieved 01 25, 2018, from

http://mplb1ci.ekb.eg/MuseProxyID=1103/MuseSessionID=0015yy3/MuseProtocol=http/MuseHost=www.emeraldinsight.com/MusePath/doi/full/10.1108/IDD-03-2017-0024

- Li, Y. and Diane, H. . (2014, 11). A multi-dimensional LIS: A report from the 12th conference on library and information science across the Taiwan straits. Bulletin of the Association for Information Science and Technology, 41(1), pp. 45–48. doi:10.1002/bult.2014.1720410116

Manyika, J. et.al. (2011). Big data: The next frontier for innovation, competition, and productivity . McKinsey Global Institute. Retrieved 11 20, 2017, from

- https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
- Manyika, J., Chui, M., Bughin,J., Dobbs,R., Bisson,P. and Marrs,A. (2013). Disruptive technologies: Advances that will transform life, business, and the global economy. Retrieved 01 11, 2018, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/disruptive-technologies
- Manyika,J., Chui,M., Brown,B., Bughin,J., Dobbs,R., Roxburgh,C., and Byers,A. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved 01 11, 2018, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
- MapR Technologies, Inc. (2017). Hadoop & Big Data. Retrieved 01 24, 2018, from MapR: Data Technologies: https://mapr.com/products/apache-hadoop/
- MapReduce Tutorial. (2017). Retrieved 01 24, 2018, from Apache Software Foundation: https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- Marr, B. (2014, 03 16). Big Data: The 5 Vs Everyone Must Know. Retrieved 01 11, 2018, from Big Data: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know
- Mashey, J. (1998, 04 25). Big Data ...and the Next Wave of InfraStress. Retrieved 11 17, 2017, from http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf
- Mauro,A., Greco,M., and Grimaldi,M. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122-135. Retrieved 11 20, 2017, from http://www.emeraldinsight.com/doi/full/10.1108/LR-06-2015-0061

- Merriam-Webster. (2017). big data. Retrieved 11 19, 2017, from https://www.merriam-webster.com/dictionary/big%20data
- Miaoab, J., Guanac, Q. and Huac, S. (2017, 07 20). pRPL + pGTIOL: The marriage of a parallel processing library and a parallel I/O library for big raster data. Environmental Modelling & Software, 96, 347-360. Retrieved 01 26, 2018, from https://www.sciencedirect.com/science/article/pii/S1364815216304893
- Mulcahy, M. (2017, 02 22). Big Data – Are You In Control? Retrieved 01 12, 2018, from Big Data Statistics & Facts for 2017 : https://www.waterfordtechnologies.com/big-data-interesting-facts/
- Mutula, S. (2016, 10). Big Data Industry: Implication for the Library and Information Sciences. African Journal of Library, Archives & Information Science. , 26(2), 93-96. Retrieved 01 31, 2018, from http://mplb1ci.ekb.eg/MuseProxyID=1103/MuseSessionID=0015z4z/MuseProtocol=http/MuseHost=web.a.ebscohost.com/MusePath/ehost/detail/detail?vid=0&sid=49ab23ff-dad6-4df5-966f-2f686cf7e8b4%40sessionmgr4010&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=a9h&AN=120156796
- Oxford University Press. (2017). big data. Retrieved 11 19, 2017, from English Oxford Living Dictionaries: https://en.oxforddictionaries.com/definition/big_data
- Pentland, A. ". (2012, 08 30). REINVENTING SOCIETY IN THE WAKE OF BIG DATA. Retrieved 11 20, 2017, from https://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data
- Prashad, Lela. (2011, 11 28). SOCIAL IMPACT THROUGH SATELLITE REMOTE SENSING - VISUALIZING ACUTE AND CHRONIC CRISES BEYOND THE VISIBLE SPECTRUM. UNITED NATIONS GLOBAL PULSE. Retrieved 11 20, 2017, from https://www.unglobalpulse.org/node/14543
- Prindle, S. and Loos, A. (2017, 09 01). Information ethics and academic libraries: Data privacy in the era of big data. Journal of Information Ethics, 26(2), 22-33. Retrieved 01 31, 2018, from

http://0015z4i.1103.y.https.www.scopus.com.mplb1ci.ekb.eg/record/display.uri?eid=2-s2.0-85039442280&origin=inward&txGid=f5cdeb8c8e8eee025bf9d4f6d1ef947c

- RDBMS Definition. (2017, 12 16). Retrieved 01 25, 2018, from Software Terms: https://techterms.com/definition/rdbms
- Rouse, Margaret . (2018). Hadoop. Retrieved 01 22, 2018, from Using big data and Hadoop 2: New version enables new applications: http://searchcloudcomputing.techtarget.com/definition/Hadoop
- Scarfo,A. and Palmieri,F. (2015). How the Big Data Is Leading the Evolution of ICT Technologies and Processes. In Modeling and Processing for Next-Generation Big-Data Technologies (Vol. 4, pp. 283-317). Retrieved 01 11, 2018, from https://link.springer.com/chapter/10.1007/978-3-319-09177-8_12
- Shalah, R. (2017, 02 28). نبذة عنBig Data. عالم البرمجة. Retrieved 11 20, 2017, from http://www.3alampro.com/articles/general-articles/about-big-data/
- Smith, C. (2017, 10). How Many People Use Facebook, Youtube and other Social Media. Retrieved 11 20, 2017, from DMR: https://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/
- Snijders, C., Matzat, U. and Reips, U. (2012). "Big Data": Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science, 7(1), 1-5. Retrieved 11 18, 2017, from http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf
- Techopedia Inc. (2017). Big Data. Retrieved 11 19, 2017, from Techopedia: https://www.techopedia.com/definition/27745/big-data
- The Apache Software Foundation . (2014). APACHE HIVE TM. Retrieved 01 24, 2018, from https://hive.apache.org/
- The Apache Software Foundation. (2017). Welcome to Apache Flume. Retrieved 01 24, 2018, from Apache Flume: https://flume.apache.org/
- The Apache Software Foundation. (2017). Welcome to Apache ZooKeeper™. Retrieved 01 24, 2018, from Apache ZooKeeper™: https://zookeeper.apache.org/
- The Apache Software Foundation. (2018). Welcome to Apache HBase™. Retrieved 01 24, 2018, from Apache HBase: https://hbase.apache.org/

- The Apache Software Foundation. (2017, 06 20). Welcome to Apache Pig! Retrieved 01 24, 2018, from https://pig.apache.org/
- The Apache Software Foundation. (2017, 12 18). Welcome to Apache™ Hadoop®! Retrieved 01 25, 2018, from Apache™ Hadoop: http://hadoop.apache.org/
- The Apache Software Foundation. (2018). Apache Sqoop. Retrieved 01 24, 2018, from http://sqoop.apache.org/
- The Economist Newspaper Limited. (2010, 02 25). Data, data everywhere. The Economist Newspaper. Retrieved 11 20, 2017, from http://www.economist.com/node/15557443

- Thinkbiganalytics.com. (2010, 07 12). Yahoo! Updates from Hadoop Summit 2010. Retrieved 01 25, 2018, from Thinkbiganalytics.com: https://www.thinkbiganalytics.com/2010/07/12/yahoo-updates-from-hadoop-summit-2010/
- Tutorials Point. (2014). Hadoop Tutorials Point. Retrieved 01 11, 2018, from https://www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf
- UN Global Pulse. (2012). Big Data for Development: Challenges and Opportunities. Retrieved 11 20, 2017, from http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseMay2012.pdf
- Vavilapalli, V., Murthy,A., Douglas,C., Agarwal,S., Konar,M., Evans,R., Graves,T., Lowe, J., Shah,H., Seth,S., Saha,B., Curino,C., O'Malley,O., Radia,S., Reed,B. and Baldeschwieler,E. (2013). Apache Hadoop YARN: yet another resource negotiator. 4th annual Symposium on Cloud Computing (SOCC '13. New York, NY, USA: ACM. Retrieved 01 24, 2018, from https://54e57bc8-a-62cb3a1a-s-sites.googlegroups.com/site/2013socc/home/program/a5-vavilapalli.pdf?attachauth=ANoY7crK0tJxTUnMEOblTvzjze5tGS1Ipi1-5qDOLVNHMfo6A2V2DYZCENxqmmJlRuLd5kUyd98tvhw43cDdbRky-XpaR1fiKnRkwJDF8pPEvCwb0FEyKAX0sxEwg-frSPJD43WKoGD6XQlc_
- Wactlar, H. (2012, 06). Big Data R&D Initiative. Retrieved 01 11, 2018, from https://www.nist.gov/sites/default/files/documents/itl/ssd/is/NIST-BD-Platforms-05-Big-Data-Wactlar-slides.pdf

- Welcome to Open Government Data. (n.d.). Retrieved 01 11, 2018, from Open Government Data: https://opengovernmentdata.org/
- What is Operating System, Kernel and Types. (n.d.). Retrieved 01 24, 2018, from http://www.idc-online.com/technical_references/pdfs/data_communications/What_is_Operating_System_Kernel_and_Types_of_kernels.pdf

White, T. (2009). Hadoop: The Definitive Guide (1st ed.). O'Reilly Media, Inc. Retrieved 01 25, 2018, from

- http://barbie.uta.edu/~jli/Resources/MapReduce&Hadoop/Hadoop%20The%20Definitive%20Guide.pdf
- White, T. (2012). Hadoop: The Definitive Guide (3rd ed.). O'Reilly. Retrieved 01 25, 2018, from https://www.isical.ac.in/~acmsc/WBDA2015/slides/hg/Oreilly.Hadoop.The.Definitive.Guide.3rd.Edition.Jan.2012.pdf
- Wired. (2013, 05). The Missing V's in Big Data: Viability and Value. Retrieved 11 20, 2017, from https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/
- Woodie, A. (2015, 04 15). From Spiders to Elephants: The History of Hadoop. Retrieved 01 25, 2018, from Datanami: https://www.datanami.com/2015/04/15/from-spiders-to-elephants-the-history-of-hadoop/
- World Economic Forum. (2012). Big Data, Big Impact: New Possibilities for International Development. Retrieved 11 20, 2017, from http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf
- Xiaodan, D. and Wei, W. (10-12 March 2017). Discussion on university library service pattern in big data Era. 2nd IEEE International Conference on Big Data Analysis, ICBDA 2017 (pp. 597-600). Beijing; China: Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ICBDA.2017.8078705
- Xie, Z. and Fox, E. (2017, 10 17). Advancing library cyberinfrastructure for big data sharing and reuse. Information Services and Use, 37(3), 319-323. Retrieved 01 25, 2018, from

http://0015ebu.1106.y.https.www.scopus.com.mplb1ci.ekb.eg/record/display.uri?eid=2-s2.0-85032589123&origin=inward&txGid=fbc75d2676fbb308e8cd13ad40ace890

- Xie, Z. and Fox, E. (2017, 10 17). Advancing library cyberinfrastructure for big data sharing and reuse. (L. B. Armbruster C., Ed.) Information Services and Use, 37(3), 319-323. doi:10.3233/ISU-170853
- YouTube. (2017). YouTube: Statistics . Retrieved 01 11, 2018, from YouTube: https://www.youtube.com/yt/about/press/

Zheng, L., QuHui, Y., Zhang,H.,Shi, H. and Wang, X. (2017). A Research on Information Security of University Libraries in the Era of Big Data. In Advances in Intelligent Systems and

- Apache Software Foundation. (2015). Apache Storm. Retrieved 01 24, 2018, from http://storm.apache.org/
- Beal, V. (2017). Big Data. Retrieved 11 19, 2017, from Webopedia: https://www.webopedia.com/TERM/B/big_data.html
- DALEY, R. (2014, 06 06). Hadoop Summit 2014 — Big Data Keeps Getting Bigger. Retrieved 01 25, 2018, from Hitachi Vantara : http://www.pentaho.com/blog/2014/06/06/hadoop-summit-2014-big-data-keeps-getting-bigger
- Harris, D. (2013, 03 04). The history of Hadoop: From 4 nodes to the future of data. Retrieved 01 25, 2018, from Gigaom: https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/
- Sindol, D. (2018). Getting Started with HDInsight - Part 1 - Introduction to HDInsight. Retrieved 01 24, 2018, from SQL Server Indexing Tips and Tricks --- Free MSSQLTips Webcast!: https://www.mssqltips.com/sqlservertip/3413/getting-started-with-hdinsight--part-1--introduction-to-hdinsight/
- The Apache Software Foundation. (2018). Apache Oozie Workflow Scheduler for Hadoop. Retrieved 01 24, 2018, from Oozie: http://oozie.apache.org/
- The Apache Software Foundation. (2018). Apache Spark™. Retrieved 01 24, 2018, from Apache Spark™ : https://spark.apache.org/
- Ahmed, W. and Ameen, K. (2017). Defining big data and measuring its associated trends in the field of information and library management. Library Hi Tech News, 34(9), 21-24. Retrieved 01 31, 2018, from

http://mplb1ci.ekb.eg/MuseProxyID=1103/MuseSessionID=0015z4e/MuseProtocol=http/MuseHost=www.emeraldinsight.com/MusePath/doi/full/10.1108/LHTN-05-2017-0035

- AlTaie, M. (2015, 09 07). Hadoop as the Backbone of Big Data Technologies. Retrieved 01 24, 2018, from Big Data: http://blog.agroknow.com/?p=3779
- Andresic, D. and Saloun, P. (9- 10 July 2017). Efficient big data analysis on a single machine using apache spark and self-organizing map libraries. In S. M. Bielikova M. (Ed.), 12th International Workshop on Semantic and Social Media Adaptation and Personalization, SMAP 2017 (pp. 1-5). Bratislava; Slovakia: Institute of Electrical and Electronics Engineers Inc. doi:10.1109/SMAP.2017.8022657
- Apache Hadoop Releases. (2017, 12 18). Retrieved 01 25, 2018, from The Apache Software Foundation: http://hadoop.apache.org/releases.pdf
- Apache Software Foundation. (2018). Apache Phoenix: OLTP and operational analytics for Apache Hadoop. Retrieved 01 24, 2018, from https://phoenix.apache.org/
- Baldeschwieler, E. (2011, 10 07). REALITY CHECK: CONTRIBUTIONS TO APACHE HADOOP. Retrieved 01 25, 2018, from Hortonworks Inc: https://hortonworks.com/blog/reality-check-contributions-to-apache-hadoop/
- Banco Bilbao Vizcaya Argentaria. (2017). The five V's of big data. Retrieved 01 11, 2018, from https://www.bbva.com/en/five-vs-big-data/
- Beyer, M. (2011, 06 27). Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Retrieved 11 19, 2017, from https://www.gartner.com/newsroom/id/1731916
- Bhat, W. (2018, 01 29). Long-term preservation of big data: Prospects of current storage technologies in digital libraries. Library Hi Tech. Retrieved 01 31, 2018, from http://0015z4c.1103.y.https.www.scopus.com.mplb1ci.ekb.eg/record/display.uri?eid=2-s2.0-85040860429&origin=inward&txGid=4e3326ba8a1c7d411a46fa7928edfb04
- Bisk Education. (2017). What is Big Data? Retrieved 11 20, 2017, from https://www.villanovau.com/resources/bi/what-is-big-data/
- Bouillet, E. et.al. (July 16 - 20, 2012). Processing 6 billion CDRs/day: from research to production (experience report). Proceedings of the 6th ACM International Conference on

Distributed Event-Based Systems, (pp. 264-267). Berlin, Germany. Retrieved 11 20, 2017, from https://dl.acm.org/citation.cfm?doid=2335484.2335513

- Budd, C. (2016, 05 16). Big Data. Plus Magazine. Retrieved 05 16, 2017, from https://plus.maths.org/content/big-data
- Budd, C. (2016, 05 16). Eight great reasons to do maths. Retrieved 11 20, 2017, from https://plus.maths.org/content/great-eight
- CERN Accelerating science. (2017). The Large Hadron Collider. Retrieved 01 11, 2018, from https://home.cern/topics/large-hadron-collider
- Chen, H., Doty, P., Mollman, C., Niu, X., Yu, J.and Zhang, T. (2015). Library assessment and data analytics in the big data era: Practice and policies. Proceedings of the Association for Information Science and Technology , 52(1), 1–4. doi:10.1002/pra2.2015.14505201002 V
- Cloudera, Inc. (2018). Apache Impala (incubating). Retrieved 01 24, 2018, from https://www.cloudera.com/products/open-source/apache-hadoop/impala.html
- Dean, J., and Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation. an Francisco, CA,. Retrieved 01 25, 2018, from https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
- Dedić, N., Stanier, C. (2017). Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. Innovations in Enterprise Information Systems Management and Engineering. ERP Future 2016. Lecture Notes in Business Information Processing. 285. Springer, Cham. Retrieved 11 18, 2017, from https://link.springer.com/chapter/10.1007/978-3-319-58801-8_10
- Demchenko,Y., Ngo,C., Laat,C., Membrey,P. and Gordijenko,D. (2015, 03 24). Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. Retrieved 01 11, 2018, from https://www.researchgate.net/publication/273945634_Big_Security_for_Big_Data_Addressing_Security_Challenges_for_the_Big_Data_Infrastructure

- Diebold, F. (2012). On the Origin(s) and Development of the Term "Big Data". Penn Institute for Economic Research, Department of Economics . Philadelphia: University of Pennsylvania . Retrieved 11 19, 2017, from https://poseidon01.ssrn.com/delivery.php?ID=199020086001026016121011113124125099104015006077091033071075020106070094067003066103097114000125006036111067121065071127066124062015046052031092099125019014068096048075035098094090121068082069013101085108074006O
- Everts, Sarah. (2016). Information Overload. Distillations, 2(2), 26–33. Retrieved 11 18, 2017, from https://www.chemheritage.org/distillations/magazine/information-overload
- Gartner IT Glossary. (n.d.). What is Big Data? Retrieved 11 18, 2017, from https://research.gartner.com/definition-whatis-big-data
- Gartner, Inc. (2011). Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. Press Release. Retrieved 11 20, 2017, from https://www.gartner.com/newsroom/id/1731916
- Ghemawat,S., Gobioff,H. and Leung,S. (2003). The Google File System. SOSP'03. New York, USA: Bolton Landing. Retrieved 01 25, 2018, from https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
- GilPress. (2013, 07 25). Big Data 3 V's: Volume, Variety, Velocity (Infographic). Retrieved 11 20, 2017, from https://whatsthebigdata.com/2013/07/25/big-data-3-vs-volume-variety-velocity-infographic/
- Google Trends. (2016). Big Datat. Retrieved 11 20, 2017, from Google Trends: https://trends.google.com/trends/explore?date=all&q=big%20data
- Hadoop. (n.d.). Retrieved 01 24, 2018, from SAS | The Power to Know: https://www.sas.com/en_us/insights/big-data/hadoop.html
- Hadoop Common. (n.d.). Retrieved 01 24, 2018, from Techopedia: https://www.techopedia.com/definition/30427/hadoop-common
- Hanson, J. . (2011, 11 01). An introduction to the Hadoop Distributed File System. Retrieved 01 24, 2018, from IBM Developers Work: https://www.ibm.com/developerworks/library/wa-introhdfs/index.html

- Hao, W. (2017, 02). Personalized information service system of a library under the big data environment. AGRO FOOD INDUSTRY HI-TECH, 28(1), 1701-1704. Retrieved 01 25, 2018, from http://mplb1ci.ekb.eg/MuseProxyID=1106/MuseSessionID=0015ebw/MuseProtocol=http/MuseHost=apps.webofknowledge.com/MusePath/full_record.do?product=WOS&search_mode=GeneralSearch&qid=1&SID=D4QFZVcZLBnCboz7GKl&page=1&doc=3
- Hashem,I., Yaqooba,A., Anuar,N., Mokhtar,S., Gania,A. and Khanb,S. (2015, 01). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98-115. Retrieved 11 18, 2017, from http://www.sciencedirect.com/science/article/pii/S0306437914001288?via%3Dihub
- Hellerstein, J. (2017). The Commoditization of Massive Data Analysis. Retrieved 11 20, 2017, from O'Reilly Media, Inc.: http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html
- Hoppe, G. (2017, 05 09). 10 Surprising Big Data Statistics. Retrieved 01 12, 2018, from Business Intelligence: https://blog.capterra.com/10-surprising-big-data-statistics/
- IBM. (2018). Apache Hadoop: built for big data, insights, and innovation. Retrieved 01 12, 2018, from IBM Analytics: https://www.ibm.com/analytics/hadoop
- IBM. (n.d.). What is Big Data Analytics? Retrieved 11 18, 2017, from IBM Analytics: https://www.ibm.com/analytics/hadoop/big-data-analytics
- Investopedia, LLC. (2017). Big Data. Retrieved 11 19, 2017, from Investopedia: https://www.investopedia.com/terms/b/big-data.asp
- Jain, A. (2016, 09 17). The 5 Vs of Big Data. Retrieved 01 11, 2018, from https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/
- Johnson, V. (2017). Leveraging Technical Library Expertise for Big Data Management. JOURNAL OF THE AUSTRALIAN LIBRARY AND INFORMATION ASSOCIATION, 66(3), 271-286. doi:10.1080/24750158.2017.1356982
- Kim, Y. and Cooke, L. (2016, 11 30). Big data analysis of public library operations and services by using the Chernoff face method. Journal of Documentation, 73(3), 466-480. doi:https://doi.org/10.1108/JD-08-2016-0098
- Kumar,V. (2017, 03 26). Big Data Facts. Retrieved 01 12, 2018, from AnalyticsWeek: https://analyticsweek.com/content/big-data-facts/

- KUWAIT10'S WORLD. (2012, 08 26). ما هو Big Data (بيئات البيانات الضخمة)؟. Retrieved 11 20, 2017, from https://kuwait10.net/2012/08/26/big-data/
- Laney, D. (2001, 02 16). 3D Data Management: Controlling Data Volume, Velocity and Variety. Retrieved 11 19, 2017, from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
- Li, J., Lu, M., Dou, G. and Wang, S. (2017, 06 05). Big data application framework and its feasibility analysis in library. Information Discovery and Delivery, 45(4), 161-168. Retrieved 01 25, 2018, from http://mplb1ci.ekb.eg/MuseProxyID=1103/MuseSessionID=0015yy3/MuseProtocol=http/MuseHost=www.emeraldinsight.com/MusePath/doi/full/10.1108/IDD-03-2017-0024
- Li, Y. and Diane, H. . (2014, 11). A multi-dimensional LIS: A report from the 12th conference on library and information science across the Taiwan straits. Bulletin of the Association for Information Science and Technology, 41(1), pp. 45–48. doi:10.1002/bult.2014.1720410116
- Manyika, J. et.al. (2011). Big data: The next frontier for innovation, competition, and productivity . McKinsey Global Institute. Retrieved 11 20, 2017, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
- Manyika, J., Chui, M., Bughin,J., Dobbs,R., Bisson,P. and Marrs,A. (2013). Disruptive technologies: Advances that will transform life, business, and the global economy. Retrieved 01 11, 2018, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/disruptive-technologies
- Manyika,J., Chui,M., Brown,B., Bughin,J., Dobbs,R., Roxburgh,C., and Byers,A. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved 01 11, 2018, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
- MapR Technologies, Inc. (2017). Hadoop & Big Data. Retrieved 01 24, 2018, from MapR: Data Technologies: https://mapr.com/products/apache-hadoop/

- MapReduce Tutorial. (2017). Retrieved 01 24, 2018, from Apache Software Foundation: https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
- Marr, B. (2014, 03 16). Big Data: The 5 Vs Everyone Must Know. Retrieved 01 11, 2018, from Big Data: https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know
- Mashey, J. (1998, 04 25). Big Data ...and the Next Wave of InfraStress. Retrieved 11 17, 2017, from http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf
- Mauro,A., Greco,M., and Grimaldi,M. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122-135. Retrieved 11 20, 2017, from http://www.emeraldinsight.com/doi/full/10.1108/LR-06-2015-0061
- Merriam-Webster. (2017). big data. Retrieved 11 19, 2017, from https://www.merriam-webster.com/dictionary/big%20data
- Miaoab, J., Guanac, Q. and Huac, S. (2017, 07 20). pRPL + pGTIOL: The marriage of a parallel processing library and a parallel I/O library for big raster data. Environmental Modelling & Software, 96, 347-360. Retrieved 01 26, 2018, from https://www.sciencedirect.com/science/article/pii/S1364815216304893
- Mulcahy, M. (2017, 02 22). Big Data – Are You In Control? Retrieved 01 12, 2018, from Big Data Statistics & Facts for 2017 : https://www.waterfordtechnologies.com/big-data-interesting-facts/
- Mutula, S. (2016, 10). Big Data Industry: Implication for the Library and Information Sciences. African Journal of Library, Archives & Information Science. , 26(2), 93-96. Retrieved 01 31, 2018, from http://mplb1ci.ekb.eg/MuseProxyID=1103/MuseSessionID=0015z4z/MuseProtocol=http/MuseHost=web.a.ebscohost.com/MusePath/ehost/detail/detail?vid=0&sid=49ab23ff-dad6-4df5-966f-2f686cf7e8b4%40sessionmgr4010&bdata=JnNpdGU9ZWhvc3QtbGl2ZQ%3d%3d#db=a9h&AN=120156796
- Oxford University Press. (2017). big data. Retrieved 11 19, 2017, from English Oxford Living Dictionaries: https://en.oxforddictionaries.com/definition/big_data

- Pentland, A. ". (2012, 08 30). REINVENTING SOCIETY IN THE WAKE OF BIG DATA. Retrieved 11 20, 2017, from https://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data
- Prashad, Lela. (2011, 11 28). SOCIAL IMPACT THROUGH SATELLITE REMOTE SENSING - VISUALIZING ACUTE AND CHRONIC CRISES BEYOND THE VISIBLE SPECTRUM. UNITED NATIONS GLOBAL PULSE. Retrieved 11 20, 2017, from https://www.unglobalpulse.org/node/14543
- Prindle, S. and Loos, A. (2017, 09 01). Information ethics and academic libraries: Data privacy in the era of big data. Journal of Information Ethics, 26(2), 22-33. Retrieved 01 31, 2018, from http://0015z4i.1103.y.https.www.scopus.com.mplb1ci.ekb.eg/record/display.uri?eid=2-s2.0-85039442280&origin=inward&txGid=f5cdeb8c8e8eee025bf9d4f6d1ef947c
- RDBMS Definition. (2017, 12 16). Retrieved 01 25, 2018, from Software Terms: https://techterms.com/definition/rdbms
- Rouse, Margaret . (2018). Hadoop. Retrieved 01 22, 2018, from Using big data and Hadoop 2: New version enables new applications: http://searchcloudcomputing.techtarget.com/definition/Hadoop
- Scarfo,A. and Palmieri,F. (2015). How the Big Data Is Leading the Evolution of ICT Technologies and Processes. In Modeling and Processing for Next-Generation Big-Data Technologies (Vol. 4, pp. 283-317). Retrieved 01 11, 2018, from https://link.springer.com/chapter/10.1007/978-3-319-09177-8_12
- Shalah, R. (2017, 02 28). نبذة عنBig Data. عالم البرمجة. Retrieved 11 20, 2017, from http://www.3alampro.com/articles/general-articles/about-big-data/
- Smith, C. (2017, 10). How Many People Use Facebook, Youtube and other Social Media. Retrieved 11 20, 2017, from DMR: https://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/
- Snijders, C., Matzat, U. and Reips, U. (2012). "Big Data": Big Gaps of Knowledge in the Field of Internet Science. International Journal of Internet Science, 7(1), 1-5. Retrieved 11 18, 2017, from http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf
- Techopedia Inc. (2017). Big Data. Retrieved 11 19, 2017, from Techopedia: https://www.techopedia.com/definition/27745/big-data

- The Apache Software Foundation . (2014). APACHE HIVE TM. Retrieved 01 24, 2018, from https://hive.apache.org/
- The Apache Software Foundation. (2017). Welcome to Apache Flume. Retrieved 01 24, 2018, from Apache Flume: https://flume.apache.org/
- The Apache Software Foundation. (2017). Welcome to Apache ZooKeeper™. Retrieved 01 24, 2018, from Apache ZooKeeper™: https://zookeeper.apache.org/
- The Apache Software Foundation. (2018). Welcome to Apache HBase™. Retrieved 01 24, 2018, from Apache HBase: https://hbase.apache.org/
- The Apache Software Foundation. (2017, 06 20). Welcome to Apache Pig! Retrieved 01 24, 2018, from https://pig.apache.org/
- The Apache Software Foundation. (2017, 12 18). Welcome to Apache™ Hadoop®! Retrieved 01 25, 2018, from Apache™ Hadoop: http://hadoop.apache.org/
- The Apache Software Foundation. (2018). Apache Sqoop. Retrieved 01 24, 2018, from http://sqoop.apache.org/
- The Economist Newspaper Limited. (2010, 02 25). Data, data everywhere. The Economist Newspaper. Retrieved 11 20, 2017, from http://www.economist.com/node/15557443
- Thinkbiganalytics.com. (2010, 07 12). Yahoo! Updates from Hadoop Summit 2010. Retrieved 01 25, 2018, from Thinkbiganalytics.com: https://www.thinkbiganalytics.com/2010/07/12/yahoo-updates-from-hadoop-summit-2010/
- Tutorials Point. (2014). Hadoop Tutorials Point. Retrieved 01 11, 2018, from https://www.tutorialspoint.com/hadoop/hadoop_tutorial.pdf
- UN Global Pulse. (2012). Big Data for Development: Challenges and Opportunities. Retrieved 11 20, 2017, from http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseMay2012.pdf
- Vavilapalli, V., Murthy,A., Douglas,C., Agarwal,S., Konar,M., Evans,R., Graves,T., Lowe, J., Shah,H., Seth,S., Saha,B., Curino,C., O'Malley,O., Radia,S., Reed,B. and Baldeschwieler,E. (2013). Apache Hadoop YARN: yet another resource negotiator. 4th annual Symposium on Cloud Computing (SOCC '13. New York, NY, USA: ACM. Retrieved 01 24, 2018, from

https://54e57bc8-a-62cb3a1a-s-sites.googlegroups.com/site/2013socc/home/program/a5-vavilapalli.pdf?attachauth=ANoY7crK0tJxTUnMEOblTvzjze5tGS1Ipi1-5qDOLVNHMfo6A2V2DYZCENxqmmJlRuLd5kUyd98tvhw43cDdbRky-XpaR1fiKnRkwJDF8pPEvCwb0FEyKAX0sxEwg-frSPJD43WKoGD6XQlc_

- Wactlar, H. (2012, 06). Big Data R&D Initiative. Retrieved 01 11, 2018, from https://www.nist.gov/sites/default/files/documents/itl/ssd/is/NIST-BD-Platforms-05-Big-Data-Wactlar-slides.pdf
- Welcome to Open Government Data. (n.d.). Retrieved 01 11, 2018, from Open Government Data: https://opengovernmentdata.org/
- What is Operating System, Kernel and Types. (n.d.). Retrieved 01 24, 2018, from http://www.idc-online.com/technical_references/pdfs/data_communications/What_is_Operating_System_Kernel_and_Types_of_kernels.pdf
- White, T. (2009). Hadoop: The Definitive Guide (1st ed.). O'Reilly Media, Inc. Retrieved 01 25, 2018, from http://barbie.uta.edu/~jli/Resources/MapReduce&Hadoop/Hadoop%20The%20Definitive%20Guide.pdf
- White, T. (2012). Hadoop: The Definitive Guide (3rd ed.). O'Reilly. Retrieved 01 25, 2018, from https://www.isical.ac.in/~acmsc/WBDA2015/slides/hg/Oreilly.Hadoop.The.Definitive.Guide.3rd.Edition.Jan.2012.pdf
- Wired. (2013, 05). The Missing V's in Big Data: Viability and Value. Retrieved 11 20, 2017, from https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/
- Woodie, A. (2015, 04 15). From Spiders to Elephants: The History of Hadoop. Retrieved 01 25, 2018, from Datanami: https://www.datanami.com/2015/04/15/from-spiders-to-elephants-the-history-of-hadoop/
- World Economic Forum. (2012). Big Data, Big Impact: New Possibilities for International Development. Retrieved 11 20, 2017, from http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf

- Xiaodan, D. and Wei, W. (10-12 March 2017). Discussion on university library service pattern in big data Era. 2nd IEEE International Conference on Big Data Analysis, ICBDA 2017 (pp. 597-600). Beijing; China: Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ICBDA.2017.8078705
- Xie, Z. and Fox, E. (2017, 10 17). Advancing library cyberinfrastructure for big data sharing and reuse. Information Services and Use, 37(3), 319-323. Retrieved 01 25, 2018, from http://0015ebu.1106.y.https.www.scopus.com.mplb1ci.ekb.eg/record/display.uri?eid=2-s2.0-85032589123&origin=inward&txGid=fbc75d2676fbb308e8cd13ad40ace890
- Xie, Z. and Fox, E. (2017, 10 17). Advancing library cyberinfrastructure for big data sharing and reuse. (L. B. Armbruster C., Ed.) Information Services and Use, 37(3), 319-323. doi:10.3233/ISU-170853
- YouTube. (2017). YouTube: Statistics . Retrieved 01 11, 2018, from YouTube: https://www.youtube.com/yt/about/press/
- Zheng, L., QuHui, Y., Zhang,H.,Shi, H. and Wang, X. (2017). A Research on Information Security of University Libraries in the Era of Big Data. In Advances in Intelligent Systems and Computing (Vol. 690, pp. 112-116). Retrieved 01 25, 2018, from http://0015ebs.1106.y.https.link.springer.com.mplb1ci.ekb.eg/chapter/10.1007/978-3-319-65978-7_17

- الجعيد، عبدالرحمن عويض .(11 08, 2017) .تحليل البيانات الكبيرة Big Dataوتحسين التعليم . مدونة تعليم جديدhttps://www.new . Retrieved 01 11, 2018, from
-educ.com/%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA
%D8%A7%D9%84%D9%83%D8%A8%D9%8A%D8%B1%D8%A9-
%D9%88%D8%AA%D8%AD%D8%B3%D9%8A%D9%86-
%D8%A7%D9%84%D8%AA%D8%B9%D9%84%D9%8A%D9%85
- الجندي، محمد .(05 12, 2014) .ماهي البيانات الضخمة؟ .مجلة لغة العصر ,11 01 Retrieved . 2018, from http://aitmag.ahram.org.eg/News/2891.aspx
- شركة أرقام بزنس إنفو .(24 02, 2016) .ما هي البيانات الضخمة Big data؟ وما أهميتها وخطورتها على مستقبل العالم؟ .أرقام ديجيتال . Retrieved 01 11, 2018, from : http://digital.argaam.com/article/detail/116012

- صلاح، رامي .(28 02, 2017) .نبذة عن .Big Dataعالم البرمجة Retrieved 01 11, 2018, from .
http://www.3alampro.com/articles/general-articles/about-big-data/
- فاخوري، عبدالرحيم .(26 08, 2015) .مقارنة بين أنظمة إدارة قواعد البيانات العلاقية SQLite :مع MySQLمع PostgreSQL. Retrieved 01 25, 2018, from أكاديمية حاسوب
:https://academy.hsoub.com/devops/servers/databases/%D9%85%D9%82%D8%A7%D8%B
1%D9%86%D8%A9-%D8%A8%D9%8A%D9%86-
%D8%A3%D9%86%D8%B8%D9%85%D8%A9-
%D8%A5%D8%AF%D8%A7%D8%B1%D8%A9-
%D9%82%D9%88%D8%A7%D8%B9%D8%AF-
%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA-%D8%A7

Big Database Management Systems: A Case Study for Apache Hadoop

Dr. Ahmed Faye z Ahmed Sayed

Abstract:

Big data has resulted from the growing use of the Internet, social networks, mobile devices, various scientific research, cloud computing, etc. With this massive increase in data, it has been difficult to analyze and process this data with known traditional systems. This massive data has been leading to the development of new and effective methods of data mining Retrieved and analyzed from such a large amount of data, whether structured or unstructured. The objective of this study is to analyze and describe the huge database management system (HADOP) and compare it with traditional database management systems. The case study was followed to study what a big database management system is, its development and components, and then a comparative approach to compare traditional database management systems with big database management systems. The researcher used Litreature to review all papers about big data and its relation to Libraries and information science, as well as big database management systems, and examine sites of Hadoop and its uses. The study has achieved several results, the most important of which is the excellence of the big database management system. It is an open source software framework that supports large distributed data applications, is licensed under the Apache Hadoop license. It also supports the operation of applications on large groups of commodity devices (Apache v2) (MapReduce). It has the ability to store and process huge amounts of structured and unstructured data, and to process data without having to specify a predefined system. The study recommends that we need to do many academic studies on the big data and the extent of use in different libraries and how to process it.

Keyword: Big Data, Apache Hadoop, Internet of Things, Cloud Computing, Relational Database Management Systems, Massive Database Management Systems.