بحوث ودراسات

استثمار البيانات الضخمة لتطوير آليات البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات: دراسة استشرافية

د. أحمد فرج أحمد

أستاذ إدارة المعلومات المساعد
كلية الآداب – جامعة أسيوط
ahmed.farag@aun.edu.eg

تاريخ الاستلام: 15 ديسمبر 2020 **تاريخ القبول: 20 يناير 2021**

المستخلص:

تتسم مجتمعات المعرفة باستمرارية نمو حجم بياناتها بشكل كبير مما أدى لظهور مفاهيم "البيانات الضخمة"، وقد دفع ذلك مؤسسات المعلومات إلى التركيز على تطوير تقنيات مبتكرة توفر القدرة على التقاط وتخزين وتحليل بياناتها، ومن ثم الخروج بنتائج لتطوير آليات البحث والاسترجاع وتعزيز المحتوى وتخصيص خدمات المعلومات.

ويعزز تطبيق البيانات الضخمة في مؤسسات المعلومات تحولها نحو عهد جديد يتسم بتقديم خدمات معلومات ذكية وذات طابع ابتكاري، وتوسيع دائرة المستفيدين المستهدفين من خدماتها.

ومن هذا المنطلق هدفت الدراسة التعرف على دوافع اعتمادها لتحليل وإدارة أنشطة مؤسسات المعلومات، والكشف عن إشكاليات توظيف البيانات الضخمة في مؤسسات المعلومات وسبل تخطيها، وآليات إدارة هذه البيانات باستخدام منصة هادوب "Hadoop"، مع التعريف بأهمية بهذه المنصة وخصائصها وبيان بنيتها الهيكلية، وتعمل الدراسة على استشراف مراحل إدارة البيانات الضخمة لأغراض تطوير البحث والاسترجاع من خلال ربط المستودعات الرقمية، وتخصيص المحتوى وخدمات المعلومات، وتعزيز تجربة المستخدم.

وكشفت الدراسة عن مرور إدارة البيانات الضخمة في مؤسسات المعلومات بمراحل تجميع وتدفق وتخزين وعرض وتحليل البيانات، وذلك لأغراض تجهيزها وتوظيفها لتعزيز البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات، وكذلك إمكانية

ممارسة أنظمة إدارة البيانات الضخمة لدور فعال في تخطي إشكاليات إدارة ومعالجة المحتوى النابع من عدة مستودعات رقمية، وتزويد المستخدمين بتجربة بحث ثرية، وتوفير خدمات بحثية تقود لنتائج ترتبط بتلبية الاحتياجات المعلوماتية للمستخدمين، وتوصي الدراسة بضرورة قيام مؤسسات المعلومات العربية بالعمل على تبني آليات متطورة لتحليل البيانات التي تمتلكها، واستثمارها لتعزيز تحقيق رؤيتها ورسالتها وأهدافها الإستراتيجية.

وقد فرضت طبيعة الموضوع استخدام المنهج الوصفي مع التركيز على أداة تحليل المحتوى في دراسة تجارب مؤسسات المعلومات العالمية التي طبقت أساليب تحليل البيانات الضخمة والاستفادة منها في سياق خدماتها، وقد تم إجراء مسح لأدبيات الإنتاج الفكري بمختلف أشكاله في ضوء أحدث ما وصلت إليه أدبيات الموضوع.

الكلمات المفتاحية: البيانات الضخمة – اقتصاديات المعرفة- مؤسسات المعلومات – إنترنت الأشياء- المستودعات الرقمية- خدمات المعلومات – البحث والاسترجاع- تحسين تجربة المستخدم.

أولاً: الإطار المنهجي:

1/1. المقدمة:

تتسم مؤسسات المعلومات باستمرارية النمو المضطرد ليس فقط في حجم مجموعاتها من مصادر المعلومات، ولكن أيضًا في حجم البيانات التي تنتجها وتتعامل معها، وتتمثل أبرز مصادر حصولها على البيانات الضخمة في النشاطات والأحداث والفعاليات والخدمات التي تقدمها، وكذلك استخدامها لخدمات الشبكة العالمية ومواقعها، وصفحاتها على شبكات التواصل الاجتماعي، والمستشعرات أو المجسات "Sensors" التي يتم توظيفها في أنظمة التعريف بترددات الراديو "RFID" و "I beacon" وغيرها من تقنيات إنترنت الأشياء، وكذلك تطبيقات الأجهزة الرقمية مثل: الهواتف الذكية "Smart phones"، والمساعدات الرقمية الصوتية "Digital Voice Assistants"، بالإضافة إلى خدمات تحديد المواقع "GPS"، وأنظمة الحوسبة السحابية، إلى غير ذلك من مصادر البيانات الضخمة في مؤسسات المعلومات.

وجدير بالذكر قد يؤثر ذلك بشكل سلبي على قدرة مؤسسات المعلومات في إدارة خدماتها ومصادرها نتيجة لعجز الأنظمة الحالية ([1]) عن إدارة البيانات والمصادر بحجمها الضخم، مما دفع المؤسسات المتطورة إلى التركيز على تبني تقنيات مبتكرة لتجميع والتقاط وتحليل وتخزين البيانات، ومن ثم الخروج بنتائج ومؤشرات لها قيمة مضافة في اتخاذ القرارات الداعمة للنشاطات الإدارية والفنية والخدمية والمساهمة في تعزيز اقتصادها المعرفي.

وتختص البيانات الضخمة بكونها غير مهيكلة "Unstructured data" أي: إنها غير منظمة وغير مرتبة وفق قوالب متوافقة مع متطلبات أنظمة إدارة قواعد البيانات العلائقية Relational Database Management System "RDBMS" المعروفة، وتواجه مؤسسات المعلومات العديد من التحديات المتصلة بإدارة وتنظيم البيانات غير المهيكلة والتي من أبرزها: القلق من سرعة نمو المصادر والمقتنيات بشكل يؤثر على فعالية الخدمات، وعدم القدرة على استيعاب التدفق المستمر للبيانات الضخمة والسيطرة عليها، ومعوقات

(1) تعتمد مؤسسات المعلومات على الأنظمة الآلية المتكاملة في المكتبات LIS Libraries Integrated Systems وكذلك على أنظمة إدارة قواعد البيانات العلائقية RDMS.

تحليل وتنظيم وإدارة كم ضخم من المحتوى، وتحديات النسخ الاحتياطي "Data Backup" وما يرتبط بها من نقل كميات كبيرة من البيانات تحتاج إلى إعدادات وتجهيزات خاصة، وصعوبات في توفير المساحة التخزينية، وبالإضافة لتكرار البيانات Data redundancy، وكذلك ندرة التقنيات "Resources limitations" القادرة على معالجة كميات هائلة من البيانات مما قد يقود إلى صعوبة الاستفادة منها، وتواجه مؤسسات المعلومات أيضًا تحديات تتعلق بقدرة أنظمة معلوماتها على استخراج وتحليل البيانات، وقضايا البحث والاسترجاع، ومشكلة تنوع أشكال البيانات التي تتعامل معها من النصوص، والصور، والرسومات، والوسائط المتعددة، والمشاركات، والملفات وغيرها، وكذلك قضايا الخصوصية في البيانات الضخمة، وندرة العنصر البشري المتخصص والقادر على التعامل مع متطلبات ومنصات إدارة البيانات الضخمة.

ومن هذا المنطلق تعمل الدراسة الحالية على اقتراح آليات عمل لمؤسسات المعلومات تعتمد على منصات إدارة البيانات الضخمة لتخطي التحديات والإشكاليات الآنفة الذكر والعمل على تطوير وتخصيص خدماتها، وذلك من خلال استعراض مفاهيم البيانات الضخمة وأهميتها ومجالاتها وخصائصها وقيمتها المضافة، وتحديد دوافع استخدامها في مؤسسات المعلومات لتحليل وإدارة أنشطتها، واستعراض خصائص وبنية أحد أهم منصات إدارة البيانات الضخمة وهي هادوب "Hadoop"، وتتناول كذلك مراحل سير العمل لتوظيف البيانات الضخمة لتطوير البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات، وآليات استثمارها في ربط المستودعات الرقمية، وتخصيص وتعزيز المحتوى الموجه إلى المستفيدين، وتحسين تجربة المستخدم.

2/1. إشكالية الدراسة وتساؤلاتها:

اعتمدت مؤسسات المعلومات في إدارة مصادرها وخدماتها على أنظمة معلومات متكاملة تتوافق مع منظومة إدارة قواعد البيانات العلائقية، ولكن مع التوجه نحو تعزيز توظيف البيانات الضخمة، شرعت هذه المؤسسات في التحول نحو عهد جديد يتسم بتقديم خدمات معلومات ذكية وذات طابع ابتكاري، وتبرز إشكالية ذلك في **عدم قدرة أنظمة إدارة قواعد البيانات على التعامل مع البيانات الضخمة** لأنها صممت لحل مشاكل محددة تستهدف البيانات المنظمة أو المهيكلة، وصعوبة تعاملها مع حجم ضخم من البيانات، ووجود أنواع وأشكال من البيانات لا تتوافق معها، وكذلك تكلفة تخزينها للبيانات تكون مرتفعة، وبالإضافة إلى البطء في معالجة البيانات الكبيرة الحجم، وذلك مقارنة بتقنيات إدارة البيانات الضخمة.

وانطلاقًا من هذه الإشكاليات برزت تساؤلات الدراسة في الآتي:

- كيف يمكن لمؤسسات المعلومات تخطي تحديات إدارة البيانات الضخمة؟
- ما دور منصة هادوب Hadoop في إدارة البيانات الضخمة؟
- ما آليات استثمار أنظمة البيانات الضخمة لتطوير البحث والاسترجاع في مؤسسات المعلومات؟
- كيفية توظيف تقنيات البيانات الضخمة لتعزيز وتخصيص محتوى وخدمات مؤسسات المعلومات؟

3/1. أهمية الدراسة ومبررات اختيارها:

تكمن الأهمية الأولى للدراسة في كونها تعالج موضوع يحظى باهتمام كافة القطاعات العاملة في إدارة المعلومات وتقنياتها وأنظمتها، وذلك لما له من دور في فتح آفاق الاستفادة من محتوى ضخم غير منظم تعجز أنظمة إدارة البيانات العلائقية عن تحليله وتنظيمه، وبالتركيز على مؤسسات المعلومات على اختلاف أنواعها، ترتكز أهمية هذه الدراسة في تناولها أطر التعريف بأحد أهم المنصات

المستخدمة في إدارة البيانات الضخمة، وبيان بنيتها الهيكلية وكيفية استخدامها، واستشراف دوافع الاستثمار فيها، وإلقاء الضوء على مراحل معالجة وتوظيف البيانات الضخمة لتطوير البحث والاسترجاع، وذلك من خلال ربط المستودعات الرقمية لتحقيق أداء بحث أفضل لمحركات وأدوات البحث، واستشراف آليات تخصيص خدمات مؤسسات المعلومات وتعزيز المحتوى والتعرف على رغبات المستفيدين المعلوماتية، وبالتالي بناء خدمات مخصصة وفق الاحتياجات، وكذلك العمل على تحسين تجربة المستخدم.

4/1. أهداف الدراسة:

تتمثل أبرز الأهداف التي تعمل الدراسة على تحقيقها في:

- التعرف على إشكاليات وتحديات البيانات الضخمة في مؤسسات المعلومات وسبل تخطيها.
- بيان كيفية إدارة البيانات الضخمة باستخدام منصة هادوب.
- تحليل مراحل توظيف البيانات الضخمة لتطوير البحث والاسترجاع في مؤسسات المعلومات.
- استشراف طرائق استثمار البيانات الضخمة لتخصيص محتوى وخدمات مؤسسات المعلومات.

5/1. حدود الدراسة:

- **الحدود الموضوعية**: تتناول الدراسة موضوع إدارة البيانات الضخمة وآليات تحليل دورها في تطوير واستحداث خدمات معلومات ذكية تتوافق مع بيئة العمل في مؤسسات المعلومات المتطورة.
- **الحدود النوعية**: تركز الدراسة على مؤسسات المعلومات بكافة أنواعها سواء أكانت وطنية أو أكاديمية أو متخصصة وغيرها.
- **الحدود اللغوية**: تم الاعتماد على تطبيقات وأنظمة إدارة البيانات الضخمة الداعمة للغات العربية والإنجليزية.

6/1. منهج الدراسة وأدوات جمع البيانات:

نظرًا لطبيعة موضوع الدراسة، وبعد الاطلاع على أبرز أدبيات الإنتاج الفكري المتاح باللغات العربية والإنجليزية والفرنسية والمرتبط بموضوع الدراسة، تم الاعتماد على منهج البحث الوصفي التحليلي وذلك لتحقيق أهداف الدراسة، ومعالجة إشكاليتها والإجابة على استفساراتها، وقد تمثلت آليات جمع المادة العلمية التي تم استخدامها في أسلوب تحليل المحتوى يمثل أسلوب بحثي يقوم على وصف المحتوى بشكل موضوعي سواء أكان كميا أو نوعيا، ويهدف في الأساس إلى جمع المعلومات عن طريق الرجوع للمصادر البحثية، وقد مر بمجموعة من المراحل والتي بدأت باختيار عينة المصادر وذلك عن طريق تحديد مجتمع البحث، وثانيًا تحديد الفترة الزمنية التي سيتم إجراء البحث فيها، وثالثًا اختيار العينة وهي خطوة بحثية رئيسية تضمن تحديد التوجهات والخطوط العريضة للبحث، وتضم الخطوة الرابعة تحقيق الأهداف وذلك عن طريق تحديد وسائل جمع البيانات والمعلومات، وخامسًا تصنيف المعلومات والمواد التي تم جمعها من العينة، وأخيرًا خطوة تحليل البيانات.

6/1. الدراسات السابقة:

أفرز بحث أدبيات الإنتاج الفكري الذي تم القيام به في قواعد ومصادر معلومات الناشرين والمتاحة من خلال بنك المعرفة المصري، الوصول إلى العديد من الدراسات الأكاديمية والتي ركزت على البيانات الضخمة، ولعل منها دراسة (Ruan & Wang,2016)

والتي ألقت الضوء على تأثير البيانات الضخمة تجاه حقبة جديدة من تحول المكتبات، وقد خلصت بأن الخدمات الذكية للمكتبات المبنية بواسطة البيانات الضخمة هي التوجه الابتكاري لنموذج خدمات المعلومات التي تستهدفها المكتبات، وخلصت أيضًا إلى أهمية أن تعمل المكتبات على صيانة وتطوير أنظمتها بحيث يمكن تطبيق البيانات الضخمة وتعزيز خدماتها الذكية.

وفي دراسة (الأكلبي، 2017) والتي توصل من خلالها بأن البيانات الضخمة تزداد ضخامة بسرعة هائلة وتحتاج إلى خطط معالجة على المستوى الوطني، و الفائدة من البيانات الضخمة ما زالت محدودة مقارنة بالفرص والقيم غير المستغلة، والحاجة إلى توظيف متخصصين في مجالات تحليل ومعالجة البيانات وأمن المعلومات، وأوصت دراسته بضرورة سن الأنظمة والتشريعات المنظمة وذلك على المستويات الوطنية والإقليمية والدولية بشكل واضح لموضوع الملكية الفكرية وخصوصية المعلومات.

كما ناقش كل من (Golub & Hansson, 2017) قضايا تمتع المكتبات بتاريخ طويل من التركيز على جمع وتنظيم وتخزين وتوفير الوصول إلى مصادر المعلومات لفئات متباينة من المستخدمين، وتمت الإشارة إلى تقديم البيانات الضخمة توسعًا كميًا في الاتصالات العلمية وتنظيم ومشاركة البيانات، والتي تمثل ثلاث مجالات مرتبطة بالبيانات في علم المكتبات والمعلومات، وتم مناقشتها في هذه الورقة في ضوء التطورات الحالية، وكذلك من منظور تحقيق أهمية مجال البحث وكيفية استمرار البيانات الضخمة والتقنيات الجديدة وبيئات البحث الشبكية في الزيادة من حيث العدد والحجم، وخلصت الدراسة إلى تميز علوم المكتبات والمعلومات بالتطور السريع لأدوات تستهدف تلبية الفرص الناشئة، من خلال المبادرات التعليمية وتطوير مجالات بحث جديدة مثل معالجة وتحليلات البيانات.

وهدفت دراسة (Zhan & Widén, 2017) إلى تكوين فهم شامل للبيانات الضخمة، وقد اقتصر مجال الدراسة على المكتبات بسبب موقعها الفريد في إدارة واستخدام البيانات الضخمة،وبالتالي، التركيز على فهم البيانات الضخمة في المكتبات وفقًا لكيفية تعريفها في تلك المهنة والتي تساعد في توضيح فهم المكتبات الحالي للبيانات الضخمة، وتمت مراجعة المقالات التي تحتوي على تعريفات للبيانات الضخمة وتم جمع (35) تعريفًا، ونظرًا لأن عدد التعريفات التي قامت هذه الدراسة بتحليلها يعتبر قليلا نسبيًا، تم إجراء كل من تحليل المحتوى والوصف الإحصائي على هذه التعريفات.

وقد أشار كل من (Al-Barashdi & Al-Karoui, 2018) إلى مواجهة المكتبات الأكاديمية حال تنفيذ تحليلات البيانات الضخمة لتحديدين أساسيين: يتمثل الأول في: الحجم الهائل والسرعة وتنوع البيانات، ويرتكن الثاني على: تعقيد تقنياتها وخوارزمياتها، وبالتالي هدفت هذه الدراسة إلى استكشاف التقنيات والأدوات التي يمكن تطبيقها في المكتبات الأكاديمية من أجل تحليل البيانات الضخمة، ومن ثم تحديد كيفية استثمارها لتعزيز اقتصاديات المكتبات الأكاديمية، كما حاولت هذه الدراسة الإجابة على كيفية مشاركة المكتبيين في البيانات الضخمة، ومستقبل التطورات البحثية، والثغرات في دراسات البيانات الضخمة المتعلقة بالمكتبات الأكاديمية، وقام الباحثان بإجراء مراجعة شاملة لأدبيات تحليل البيانات الضخمة للمكتبات الأكاديمية، وقد أسفرت النتائج عن (37) ورقة تتعلق بالبيانات الضخمة في المكتبات الأكاديمية، وأشارت النتائج إلى أنه على الرغم من الكم الكبير من الأبحاث التي أجريت حول هذا الموضوع، إلا أن عددًا قليلًا من الدراسات ناقش تأثير البيانات الضخمة على المكتبات الأكاديمية، بما في ذلك أدوات وتقنيات التحليل.

وعلى نفس السياق هدفت دراسة (العميري، 2018) إلى التعرف على واقع البيانات الضخمة في المكتبات الأكاديمية بسلطنة عمان، وقد اعتمدت على المنهج الوصفي الكمي، وقد بلغت عينة الدراسة (106) من أخصائيي المكتبات من (28) مكتبة أكاديمية، وكشفت نتائج الدراسة أن عمليات البحث عن المصادر الإلكترونية حققت أعلى نسبة بلغت (69 ٪) من بين أكثر مصادر جمع البيانات الضخمة التي تم الاستفادة من بياناتها في المكتبات الأكاديمية، بينما اتضح أن التنبؤ بالاحتياجات المستقبلية للمكتبات من أبرز مجالات

الاستفادة من البيانات الضخمة، فقد بلغت (75.5 ٪)، وتبين أن قلة وعي أمناء المكتبات بمدى أهمية البيانات الضخمة وتحليلها جاءت بأعلى نسبة والتي وصلت (71.4 ٪)، كما اتضح أن أكثر التحديات التي تواجه البيانات الضخمة هي التكلفة الباهظة وتوفير الإمكانيات التقنية بنسبة (68.9 ٪).

وقد أكدت دراسة (الحاتمية وآخرون، 2018) على أهمية البيانات الإحصائية باعتبارها ركيزة أساسية في صنع القرار والتخطيط الإستراتيجي، وتشكل كذلك جزء كبيرا من البيانات الضخمة، حيث تترابط مع بيانات أخرى لتعطيها قيمة مضافة، وقد تبقى هذه البيانات في أحيان كثيرة عديمة الفائدة إذا لم تحظ بالتحليل والإدارة الجيدة لعمليات حفظها واسترجاعها وإتاحتها، وأفادت الدراسة بأنه غالبًا ما يعزى إلى المراكز الإحصائية في الدول بمهمة توفير وانتاج واتاحة البيانات الإحصائية، حيث تلعب دورا مهمًا وحيويًا في دعم مختلف القطاعات والمؤسسات من خلال توفير البيانات الإحصائية، ويمثل المركز الوطني للإحصاء والمعلومات بسلطنة عمان مصدرا مهما لتوفير البيانات الإحصائية الرسمية والمستخدمة في تلبية متطلبات كافة القطاعات، وإيمانًا بأهمية البيانات الإحصائية هدفت هذه الدراسة إلى استكشاف واقع إنتاج البيانات الإحصائية، والتعرف على طبيعة إتاحتها والوسائل المستخدمة في نشرها، وطرق استثمارها، وكذلك التعرف على العوامل المؤثرة في إنتاج البيانات الإحصائية في المركز الوطني للإحصاء والمعلومات بسلطنة عمان.

ومن خلال عرض النماذج السابقة من أدبيات الإنتاج الفكري يلاحظ أنها قد اتسمت بالتركيز على البيانات الضخمة من منظور المفاهيم والأهمية وكيفية استخدامها لتوفير نموذج خدمات المعلومات التي تستهدفها المكتبات وبخاصة كيفية مشاركة المكتبات الأكاديمية في البيانات الضخمة، بينما تعمل الدراسة الحالية على معالجة مراحل سير العمل في إدارة البيانات الضخمة لتطوير البحث والاسترجاع وتخصيص خدمات المعلومات وكيفية توظيفها لربط مستودعات متعددة لتوفير مقومات بحث معلوماتي أفضل وتخصيص المحتوى وتحسين تجربة المستخدم.

ثانياً: الإطار النظري:

يستعرض الإطار النظري للدراسة مفاهيم وأهمية البيانات الضخمة والخصائص التي تتمتع بها، ويلقي الضوء على نماذج من الإشكاليات التي تواجه البيانات الضخمة مع محاولة بيان سبل تخطيها، ويعالج أيضًا دور **منصة هادوب في إدارة البيانات الضخمة** من خلال التعريف بها وبأهميتها وخصائصها، وبيان طرائق توظفيها واستخدامها، إلى جانب تحليل بنيتها الهيكلية.

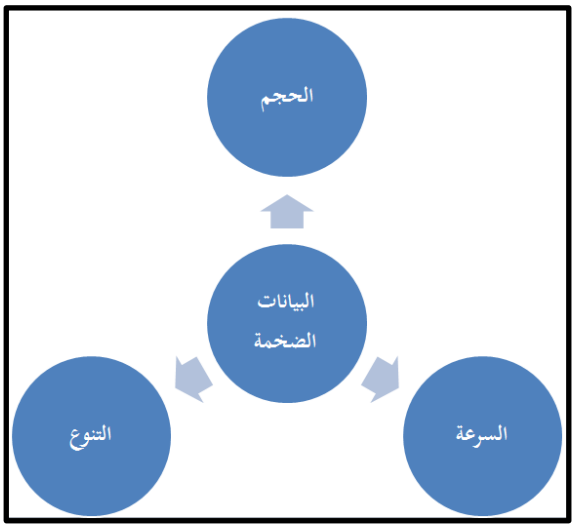
1/2. البيانات الضخمة: المفاهيم والأهمية والخصائص:

تُقسم البيانات وفقًا لآلية المعالجة والتنظيم إلى ثلاثة أنواع رئيسية: يتمثل النوع الأول في: البيانات المهيكلة "Structured data" وهي بيانات منظمة ومرتبة وقابلة للبحث والاسترجاع بسهولة من خلال أنظمة إدارة قواعد البيانات، ولغة الاستفسار الهيكلية أو البنيوية "Structured Query Language" SQL، ويطلق على النوع الثاني: البيانات شبه المهيكلة "Semi structured data" والتي من أبرز أمثلتها: ملفات لغة التكويد القابلة للامتداد والتوسعة "XML files"، وأخيرًا هناك البيانات غير المهيكلة "Unstructured data" أو غير المنظمة والتي تنطوي على مشاكل في التنظيم، وتحتاج إلى وقت وجهد للبحث والوصول للمعلومات ومن نماذجها ملفات الفيديو، والتقارير، ورسائل البريد الإلكتروني، والصور، ومحتوى منصات شبكات التواصل الاجتماعي وغيرها.

وقد أبرز بحث أدبيات الإنتاج الفكري العديد من المفاهيم التي تتعلق بالبيانات الضخمة، ولعل من أكثرها تداولًا ما أشار إليه كل من (Al-Barashdi & Al-Karoui, 2018) والذي ينظر إليها كونها بيانات ضخمة للغاية من حيث الحجم، والسرعة في التوالد

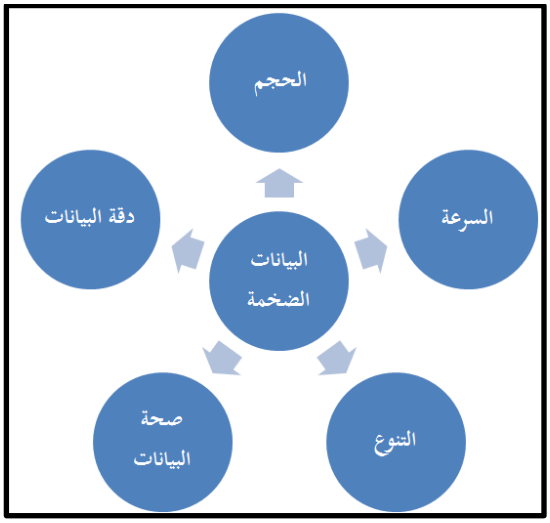
والانتشار، وبالإضافة إلى مواجهتها لصعوبات في المعالجة والإدارة بحيث يتعذر- على الأنظمة التقليدية المستخدمة في إدارة البيانات المهيكلة- معالجتها.

وقد قام (Laney, 2001) بتصور خصائص البيانات الضخمة باستخدام نموذج (3Vs)، والذي يتضمن كما هو مبين بالشكل رقم (1) الحجم "Volume" والذي يعكس حجم البيانات، وسرعة البيانات "Velocity" سواء من حيث التوالد و/ أو الانتشار، وأخيرًا التنوع "Variety" والمقصود به تنوع أشكال وأنواع البيانات التي يتم التعامل معها من الصور والرسومات والملفات والمشاركات وغيرها.



شكل (1) نموذج (3Vs) للبيانات الضخمة

وقد شرع كل من (Lomotey & Deters, 2014) بتطوير نموذج (3Vs) والخروج بنموذج آخر أطلق عليه (5Vs)، والذي ضم إليه -كما هو موضح بالشكل رقم (2) – صحة البيانات "Veracity" والقيمة "Value"، وتعكس "صحة البيانات" ضرورة التأكد من تخزين البيانات الضخمة بصورة سليمة وتفادي وجود مشاكل أثناء إجراءات النقل والإدارة والمعالجة، وبينما ترتكن "القيمة" إلى استثمار البيانات الضخمة لتحقيق الأرباح وتعزيز الاقتصاد المعرفي وضمان استمرارية تطوير الخدمات والمنتجات.



شكل (2) نموذج (5Vs) للبيانات الضخمة

وعلى نفس المنوال قام Firican, (2017) باستعراض نموذج (10Vs) فبالإضافة إلى ما سبق الإشارة إليه في نموذج (5Vs)، وجد في نموذج (10Vs)، -كما هو مبين بالشكل رقم (3) - خاصية المتغيرات "Variability" والتي تؤكد على أسس تحقيق التوافق مع التغييرات المحتملة التي قد تطرأ على النظام، ففي بعض الأنظمة قد تحدث مشكلات يتطلب تخطيها مجهودات برمجية تستلزم وقتا وجهدا وتكلفة، ثم تأتي خاصية الصلاحية "Validity" ويقصد بها دقة التحليل والاستفادة من تحليلات البيانات الضخمة في ضوء الحاجة إلى اعتماد ممارسات فعالة لضمان الجودة والاتساق، وتشير خاصية التأثير العالي "Vulnerability" إلى القابلية واحتماليات الخرق، وبخاصة مع إثارة البيانات الضخمة مخاوف أمنية، وظهور العديد من نماذج التعدي على خصوصية البيانات الشخصية لمستخدمي العديد من التطبيقات، وتركز خاصية التقلب "Volatility" على التوقيتات الزمنية التي تمكن من النظر إلى البيانات باعتبارها غير ذات صلة أو أصبحت متقادمة وغير مفيدة، وتحديد الفترة الزمنية الواجب فيها الاحتفاظ بالبيانات، وكانت تميل المؤسسات قبل استخدام البيانات الضخمة إلى تخزين بياناتها إلى أجل غير مسمى بصرف النظر عن حجمها، ويجب مراعاة خاصية التقلب نظرًا لسرعة نمو البيانات الضخمة، والحاجة إلى وضع قواعد حاكمة لتحديث البيانات ومدى توفرها، وكذلك ضمان الاسترجاع السريع للمعلومات عند الحاجة، وأخيرًا تعكس خاصية العرض المرئي للبيانات وتصورها "Visualization" الاستفادة من أنماط عرض البيانات وفق أشكال متنوعة في شكل تقرير أو لوحة مخصصة للقيادة والتحكم "Dashboard"، وكذلك وفق متغيرات متعددة مثل: السن والمكان والجغرافيا إلى غير ذلك.

وتتوافر نماذج أخرى من مفاهيم البيانات الضخمة تشير إلى استخدام المصطلح للإشارة لمجموعات البيانات "datasets" ذات الهياكل الكبيرة والمعقدة، ووفق هذا المنظور قام كل من (Gantz & Reinsel, 2011) بالنظر إلى البيانات الضخمة بأنها جيل جديد من التقنيات والمعماريات المصممة لاستخراج القيمة المضافة اقتصاديًا من كميات ضخمة ومتنوعة من البيانات، وتوظف البيانات الضخمة أيضًا سلسلة من التقنيات لتخزين وتحليل مجموعات البيانات الكبيرة والمعقدة، بما في ذلك "NoSQL" و"MapReduce" وتعلم الآلة "Machine learning" والتي تعني التنبؤ بالمستقبل.

شكل (3) نموذج (10Vs) للبيانات الضخمة

وقد لاحظ (Akoka & Comyn-Wattiau & Laoufi, 2017) نموًا كبيرًا في حجم الإنتاج الفكري والدراسات العلمية المرتبطة بتحليل البيانات الضخمة على مدى السنوات الخمس الماضية، وأشاروا إلى تنوع الاهتمام بين الباحثين في دراسة قضايا تحديد الأهداف وآثارها، وتطبيق المعايير القياسية، وكذلك مجموعة واسعة من الاستخدامات والتطبيقات، وقد خلصوا لأن تحليل البيانات الضخمة يركز على تقنيات التجميع "Clustering" والتصنيف والتنظيم "Classification" وكذلك على نماذج التنبؤ "Prediction models".

وانطلاقًا من المفاهيم السابقة تخلص هذه الدراسة إلى إمكانية النظر إلى البيانات الضخمة باعتبارها كميات كبيرة من البيانات، والتي غالبًا ما تنشأ مصادر متعددة وتظهر بتنسيقات وأشكال مختلفة، وتفتقر إلى التنظيم والهيكلة، وتتطلب أنماطًا جديدة من المعالجة، وفي المقابل تتسم بالعديد من الخصائص والمميزات والتي دفعت مؤسسات ومراكز المعلومات المتطورة إلى أن تتعامل معها بشكل متزايد، والتوجه نحو توظيفها من أجل تطوير خدماتها لدعم مستخدميها، ولتمكين تحسين دعم اتخاذ وصنع القرارات الإدارية والفنية والتقنية.

2/2. إشكاليات توظيف البيانات الضخمة في مؤسسات المعلومات وسبل تخطيها :

استعرض (الأكلبي، 2017) العديد من التحديات التي تواجه مؤسسات المعلومات، ولعل من أبرزها الخوف من عدم قدرتها على استيعاب تدفق البيانات الضخمة المستمر والسيطرة عليها، وكذلك تحديات ترتبط بأنظمتها التقليدية المستخدمة في استخراج وتحليل البيانات، في ظل سرعة نمو البيانات وما ترتبط بها من مشاكل المعالجة والمساحة التخزينية، وقضايا البحث والاسترجاع، بالإضافة إلى تهديدات تواجه خصوصية البيانات الضخمة، إلى غير ذلك.

وتمثل أنظمة التخزين السحابية أحد الحلول المقترحة للتغلب على قضايا تخزين البيانات ومعالجتها، ومع ذلك لا يلقى هذا الحل قبول بعض المؤسسات وبخاصة في حالة وجود بيانات حساسة تخضع لمقتضيات المحافظة على السرية، والتخوف من اختراقها لاسيما مع بروز مشكلات ذات علاقة بأمن المعلومات، **وتقترح هذه الدراسة لتخطي مثل هذه الإشكالية** تخصيص مستودع بيانات سحابية خاص بمؤسسة المعلومات يتم تشفيره وتخزين البيانات فيه، وتبني إجراءات صارمة للمحافظة على الخصوصية وتأمين البيانات والمعلومات.

ويمكن أن تلجأ مؤسسات المعلومات إلى توفير تقنيات متخصصة في إدارة البيانات بكفاءة، والتي من نماذجها استخدام تقنيات معالجة البيانات الديناميكية المتقدمة وحساب كثافة البيانات وإدارتها.

وينبغي على مؤسسات المعلومات لكي تتمكن من تخطي معوقات البحث والاسترجاع، تبني آليات تقود إلى تهيئة البنية التحتية لمستودع البيانات بقدرة استيعابية كافية، وتحليل البيانات وتوصيفها وتنظيمها، وكذلك توظيف قدرات وإمكانيات محركات البحث الدلالية لتعزيز قدرات البحث والاسترجاع، والتي تمثل إحدى متطلبات إدارة البيانات الضخمة، وقد حصر كل من (عبد الله والهنائي، 2018) إمكانات ومتطلبات استخدام البيانات الضخمة والإفادة منها في استعدادات البنية التحتية والإدارية والتنظيمية، ورصد البيانات الضخمة واقتنائها، وخبرات التعامل مع البيانات الضخمة، والمعرفة بأدوات تنظيم البيانات الضخمة ومهارات استخدامها.

وتحظى قضايا الخصوصية والاستخدام الآمن للمعلومات والحد من الاختراقات المحتملة لها باهتمام كبير من جانب المتخصصين في إدارة البيانات الضخمة، وبخاصة أنها تركز على التنقيب عن البيانات واكتشاف المعرفة، ومن ناحية أخرى، قد تنشأ مخاطر جديدة لاختراق النظام بسبب إمكانية الوصول إلى كمية كبيرة من البيانات، وقد أشار(Wang et al, 2016) بأن أبحاث ودراسات توظيف البيانات الضخمة في المكتبات لم تأخذ قضايا أمن البيانات في الاعتبار ولم توليها الاهتمام الكافي.

ويجب أن تنوه **الدراسة الحالية** بأن هناك قطاعا يرى أن المشاركات ونشر المستخدمين للبيانات الشخصية من خلال شبكات التواصل الاجتماعي قد لا يندرج ضمن المفهوم العام للخصوصية، ويستند هذا القطاع في رؤيته بأن منصات البيانات الضخمة تستهدف من وراء تحليل وإدارة وتخزين هذه البيانات، زيادة قيمتها وتقديم خدمات مخصصة وفعالة في متناول المستخدمين، وبالتالي ووفق هذا السياق لا تخترق البيانات الضخمة الخصوصية.

وبعد هذا العرض المركز حول القضايا والقيود التي تواجه توظيف البيانات الضخمة في مؤسسات المعلومات ومعالجة آليات تخطيها، تعمل الفقرات التالية على فحص دور منصة هادوب في إدارة البيانات الضخمة.

3/2. إدارة البيانات الضخمة باستخدام منصة هادوب:

1/3/2. هادوب: النشأة والتطور:

تمثل منصة هادوب نظاما متكاملا مفتوح المصدر يستهدف معالجة وإدارة وتخزين أنواع متباينة من البيانات، وترجع الجذور الأولى لهذه المنصة إلى عام (2005م) من خلال تصور قام به كل من "Doug Cutting" و "Mike Cafarella" لمشروع محرك بحث مفتوح المصدر على الويب أطلق عليه "Nutch"، وكانت الفكرة الرئيسية تكمن في عرض نتائج البحث على الشبكة العنكبوتية العالمية بشكل أسرع عبر تبني آلية عمل توزيع البيانات "Distribute data" ومعالجتها وتخزينها عبر العديد من أجهزة الحاسبات الآلية لأغراض إنجاز المهام بشكل متزامن، وتعظيم فرص تخزين كم ضخم من البيانات وتحقيق قدرات معالجة عالية (Sultana , 2015)

وبالتوازي -في نفس الفترة- قامت مؤسسة "Google" بعمل مشروع "MapReduce" لتصنيف ومعالجة البيانات، واعتمدت فكرته على نفس مفهوم مشروع "Nutch" والتي تتمثل في تخزين ومعالجة البيانات بطريقة موزعة ومؤتمتة بحيث يمكن استرجاع نتائج البحث على الويب بشكل أسرع، من خلال تقسيم أو توزيع البيانات ووضعها على أجهزة خوادم متصلة مع بعضها البعض، (Taylor, 2010)

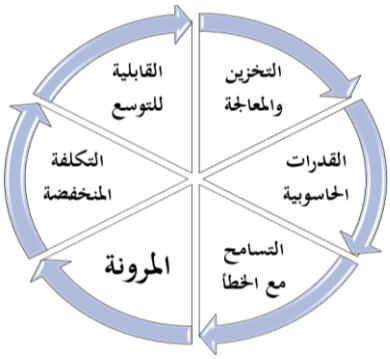
وانضم "Cutting" في عام (2006م)، إلى مؤسسة "Yahoo" وأخذ معه مشروع "Nutch" بالإضافة إلى أفكار تستند إلى عمل مشروع "Google" لأتمتة تخزين البيانات الموزعة ومعالجتها، وتم تقسيم مشروع "Nutch" إلى قسمين: مثل الأول: زاحف الويب واحتفظ بالمسمى "Nutch"، والقسم الآخر: أصبح جزء الحوسبة والمعالجة الموزعة اعتمد على تطبيق مفهوم "MapReduce" وأطلق عليه تسمية هادوب "Hadoop".

وفي عام (2008م) أطلقت ياهو منصة هادوب كمشروع مفتوح المصدر، والذي تتم في الوقت الراهن إدارته وصيانته من خلال مؤسسة برمجيات أباتشي Apache Software Foundation (ASF) غير الربحية ([2]).

وبعد استعراض النشأة والبدايات الأولي لمنصة هادوب، **ولأغراض التمهيد لمعالجة هدف الدراسة الثاني والذي يتعلق بكيفية إدارة البيانات الضخمة باستخدام منصة هادوب**، تتناول الجزئية التالية معالجة أهمية هذه المنصة وبيان أبرز الخصائص والملامح المميزة لها، وكذلك استشراف أوجه وكيفية الاستخدام والتوظيف لها.

Hadoop.2/3/2 : الأهمية والخصائص والاستخدام:

تركز الفقرات التالية على تحليل آليات حل مشكلات إدارة البيانات الضخمة باستخدام منصة هادوب، وذلك من خلال استعراض الأهمية وأبرز الخصائص التي يتمتع بها، بالإضافة إلى إلقاء الضوء على كيفية استخدامه، ويمكن من خلال فحص أدبيات الإنتاج الفكري حصر أهمية منصة هادوب وقيمتها المضافة في إدارة البيانات الضخمة في النقاط الرئيسية التالية كما هو مبين في الشكل رقم (4):



شكل (4) خصائص منصة هادوب

- **تخزين ومعالجة كميات ضخمة من البيانات**: ويعد هذا أحد الاعتبارات الرئيسية وبخاصة مع تزايد أحجام البيانات وتنوعها باستمرار، وتعدد مصادرها لاسيما من وسائل وشبكات التواصل الاجتماعي وإنترنت الأشياء (IoT)، وتمتاز منصة هادوب بالسرعة الفائقة في معالجة الكم الضخم والمتنوع من البيانات، كما تتمتع بقدرات تخزينية عالية لاعتمادها على توظيف عدد كبير من الحاسبات الشخصية ومحطات العمل.

(2) لمزيد من المعلومات حول Apache Software Foundation ASF يمكن الاطلاع على الرابط التالي https://www.apache.org

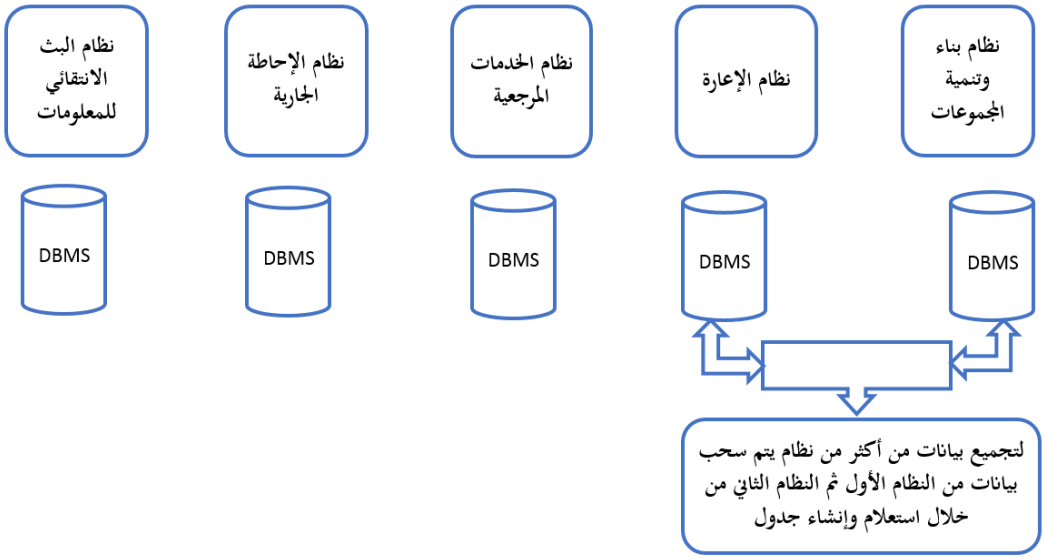
- **القدرة الحاسوبية وقوة المعالجة**: تمثل منصة هادوب نموذجا للحوسبة الموزعة لمعالجة البيانات الضخمة، وكلما زادت عدد الحاسبات التي يتم استخدامها بالمنصة، كلما زادت قوة المعالجة وسرعة الحصول على النتائج، و هادوب يتوافق مع العمل على الحاسبات الشخصية ومحطات العمل، وبالتالي ليست هناك حاجة إلى أجهزة ذات قدرات ومواصفات فنية عالية في المعالجة.
- **التسامح مع الخطأ**: المقصود وفق هذا الإطار حماية معالجة البيانات والتطبيقات في حالة فشل أحد الأجهزة أو تعطلها، ولا توجد احتمالية لفقدان البيانات لأن النظام يتبني إجراءات حماية صارمة، تتمثل في النسخ الإجباري لكل كتلة من البيانات "Block" وبشكل تلقائي في ثلاثة أماكن مختلفة، وفي حال فقدان النظام لبيانات يقوم باستخراج نسخة أخرى للمستخدم، وبالتالي تكون البيانات والتطبيقات محمية من أي فشل محتمل في أداء الأجهزة، وبعبارة أخرى في حالة حدوث مشكلة في عقدة "Node" (جهاز) معينة، تتم إعادة توجيه المهام تلقائيًا إلى العقد الأخرى للتأكد من عدم فشل الحوسبة الموزعة.
- **المرونة**: لا يتطلب نظام هادوب ضرورة المعالجة المسبقة للبيانات قبل تخزينها، ويقوم بتخزين البيانات بشكلها الأساسي كما هو دون الالتزام بشكل معين، وذلك بعكس أنظمة إدارة قواعد البيانات العلائقية، وبالتالي يمكن تخزين أي حجم من البيانات بأية صيغة وتحديد واتخاذ قرار حول شكل المعالجة والاستخدام في وقت لاحق عندما تدعو الحاجة لذلك.
- **التكلفة المنخفضة**: هادوب نظام مفتوح المصدر يكون استخدامه مجانًا ودون أية تكلفة مادية مقابل الحصول على تراخيص الاستخدام، وكما –سبقت الإشارة– أنه يتوافق مع الحاسبات الشخصية، وبالتالي لا توجد اشتراطات خاصة حول مواصفات الأجهزة لتخزين كميات كبيرة من البيانات.
- **القابلية للتوسع**: يمكن ببساطة تطوير منصة هادوب لتوسيع قدرات المعالجة والتعامل مع المزيد من البيانات الإضافية عن طريقين: الأول: من خلال التوسع الرأسي "Horizontal scaling" والذي يتضمن إضافة أجهزة جديدة إلى مجموعة الأجهزة المتصلة مع النظام، وأما الطريق الثاني فيكون عبر التوسع الأفقي "Vertical scaling" والذي يرتكز على تطوير إمكانات ومواصفات الحاسبات الموجودة بالفعل مثل: تعزيز قدرات الذاكرة والقرص الصلب ووحدة المعالجة المركزية وغيرها.

ومن أبرز الخصائص الرئيسية التي تتمتع بها منصة هادوب ما ذكره (Rodrguez-Mazahua et al, 2016) في عدم التعامل مع البيانات ككتلة واحدة؛ بل يتم توزيعها إلى عدد من الكتل "Blocks" وكل كتلة تخزن في خادم، ويطلق على كل مجموعة من الكتل مسمى "Cluster"، وهناك ما يسمى بالعقدة الرئيسية أو المركزية "Central node" أو "Main node" والتي تتحكم في كل بيانات الأجهزة "العقد" الأخرى، ولا تحتوي على أية بيانات، وتجدر الإشارة بأن "MapReduce" يتولى مسئولية اتصال الأجهزة معًا وتوزيع المهام عليها وذلك في ضوء التنسيق مع الجهاز أو العقدة الرئيسة "Main Nodes" للحصول على المعلومات والبيانات منها، وبالتالي توفير الكثير من الوقت لأداء المهمة المطلوبة.

ومع تجاوز الأهداف الأصلية التي أنشئت منصة هادوب لتحقيقها والتي تتمثل في البحث عن الملايين من صفحات الويب والحصول على نتائج ذات صلة، تتطلع العديد من مؤسسات المعلومات إلى اعتبار هذه المنصة بمثابة النظام الأساسي لإدارة التعامل

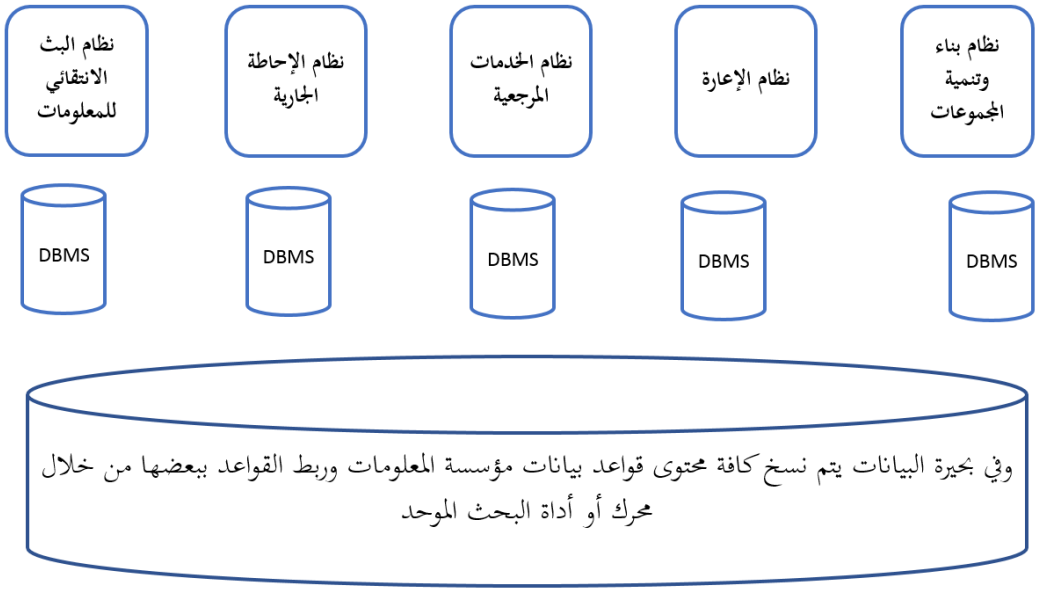
ومعالجة البيانات الضخمة، وتشمل أبرز الاستخدامات الشائعة التي يمكن تطبيقها في إطار مؤسسات المعلومات لتطوير مهامها وخدماتها في الآتي:

- **التخزين والأرشفة**: قادت التكلفة المتواضعة للأجهزة الشخصية والتي يعمل من خلالها نظام هادوب إلى توفير فرص تخزين ودمج أنواع عديدة من البيانات الناتجة عن العديد من المصادر منها: المعاملات "Transactional" والوسائط الاجتماعية "Social media" وأجهزة الاستشعار "Sensors" والآلات "Machines" وغيرها، وتتيح التكلفة المنخفضة فرصة الاحتفاظ بالمعلومات التي لا تعتبر مهمة بالنسبة لمؤسسة المعلومات في الوقت الراهن، ولكن قد تكون هناك رغبة لتحليلها وإدارتها لاحقًا.
- **الاستكشاف والتحليل**: نظرًا لتصميم هادوب لغرض التعامل مع أحجام ضخمة من البيانات وفق مجموعات متنوعة من الأشكال والنماذج، فإنه يمكن تشغيل خوارزميات تحليلية، ويمكن أيضًا أن يساعد تحليل البيانات الضخمة مؤسسة المعلومات على العمل بكفاءة أكبر، والكشف عن إمكانيات جديدة تستمد مزايا تنافسية وتوفير فرص للابتكار بأقل استثمار ممكن.
- **بحيرة البيانات** Data lake: تُمثل مستودعًا باستطاعته استيعاب قدر هائل من البيانات الخام في شكلها الأولي ودمجها معًا، وتكمن فكرتها في تخزين البيانات بتنسيقها الأصلي أي: كما هي بكل خصائصها وأنواعها في مكان واحد، وتجميع كافة بيانات أنظمة إدارة قواعد بيانات مؤسسة المعلومات في مكان واحد، مثل: نظام بناء وتنمية المجموعات، ونظام الإعارة، ونظام الخدمة المرجعية، ونظام البث الانتقائي للمعلومات والإحاطة الجارية، ونظام الفواتير، ونظام خدمة العملاء، ونظام الموارد البشرية وذلك في مكان واحد، والمتعارف عليه لكل نظام قاعدة البيانات الخاصة به والتي تكون عادة منفصلة ومستقلة عن الأنظمة الأخرى، وفي حالة الربط بين نظامين أو أكثر أو قاعدتين أو أكثر من قواعد البيانات كما هو مبين في الشكل رقم (5)، تتمثل منهجية التنفيذ في سحب البيانات المطلوبة من النظام (أ) وبالتحديد من نظام إدارة قاعدة البيانات الخاصة به ووفق الشكل رقم (5) يتمثل في نظام بناء وتنمية المجموعات، وسحب البيانات من النظام (ب) وهو نظام الإعارة، ثم القيام باستخراج جدول جديد يضم النتائج المطلوبة، ويطلق على هذه الآلية "ETL Extract Transaction Loading"، وهذه الطريقة هي المتبعة في إجراء الاستعلام في مستودعات البيانات "Data Warehouse" بمؤسسات المعلومات بالرغم من تعاملها فقط مع البيانات المهيكلة.



شكل (5) آلية استخراج البيانات من أكثر من نظام

ويستخلص من ذلك أنه تكمن فكرة بحيرة البيانات في وضع كل البيانات الخاصة بجميع الأنظمة وقواعد بياناتها في مكان واحد كما هو موضح في الشكل رقم (6)، وفي بحيرة البيانات يتم نسخ كافة محتويات قواعد بيانات مؤسسة المعلومات في مكان مركزي ضخم وربط القواعد ببعضها من خلال محرك أو أداة البحث الموحد "Federated Search".



شكل (6) نسخ قواعد بيانات الأنظمة في قاعدة بيانات ضخمة مركزية وفق منظومة بحيرة البيانات

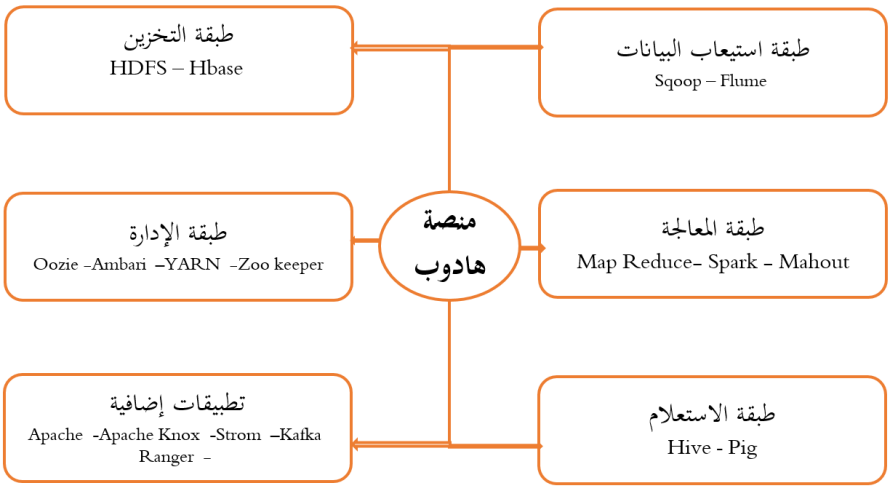
- **التكامل مع مستودع البيانات**: يمثل مستودع البيانات المكان الذي تختزن فيه مؤسسة المعلومات بياناتها لفترات طويلة من الزمن، وتختص حصريًا بحفظ البيانات المهيكلة -المنظمة- والمتوافقة مع أنظمة إدارة قواعد البيانات، ويمكن

لكل من مستودعات البيانات ومنصة هادوب التعايش معًا والعمل سويًا داخل مؤسسة المعلومات، حيث إن الهدف النهائي لكل مؤسسة هو الحصول على نظام أساسي مناسب لتخزين ومعالجة البيانات وإدارتها.

وبعد استعراض أهمية منصة هادوب وأبرز الخصائص التي تتمتع بها، وكذلك إلقاء الضوء على ملامح الاستخدام المحتملة، يستهدف الجزء التالي من الدراسة إلقاء الضوء على المكونات الرئيسية والتي تعكس بنية منصة هادوب من الأدوات والتقنيات.

3/3/2. البنية الهيكلية لنظام هادوب:

تتألف البنية الهيكلية العامة لمنصة هادوب من مجموعة من المكونات الرئيسية وتضم كل منها بطبيعة الحال -كما هو مبين في الشكل رقم (7) – مجموعة من التقنيات والأدوات، والتي تشكل معًا ما يطلق عليه نظام هادوب البيئي "Hadoop Ecosystem".



شكل (7) مكونات وبنية منصة هادوب

تتألف منصة هادوب من عدد من الطبقات وتؤدي كل منها مهام معينة ويتمثل أبرزها في الآتي:

- **طبقة استيعاب البيانات Data ingestion layer**: وتضم مجموعة الأدوات المسئولة عن إدخال البيانات إلى نظام هادوب ومن أمثلتها Flume وSqoop
- **طبقة التخزين Data storage**: والمسئولة عن تخزين البيانات ومن نماذج تقنياتها HDFS – Hbase
- **طبقة المعالجة Data Processing**: ويقع على عاتقها مهمة معالجة البيانات والتي من أمثلتها تقنيات – Spark MapReduce - Mahout
- **طبقة الإدارة Cluster Management**: وتتضمن تطبيقات مسئولة عن إدارة النظام ومن نماذجها Zoo keeper– Oozie – Ambari – YARN
- **طبقة الاستعلام Querying layer**: والتي تركز على إدارة إجراءات البحث والاسترجاع ومن نماذج أدواتها - Hive Pig- Hcatalog
- **تطبيقات إضافية مساندة** ومن أبرزها Apache Ranger – Apache Knox – Strom – Kafka

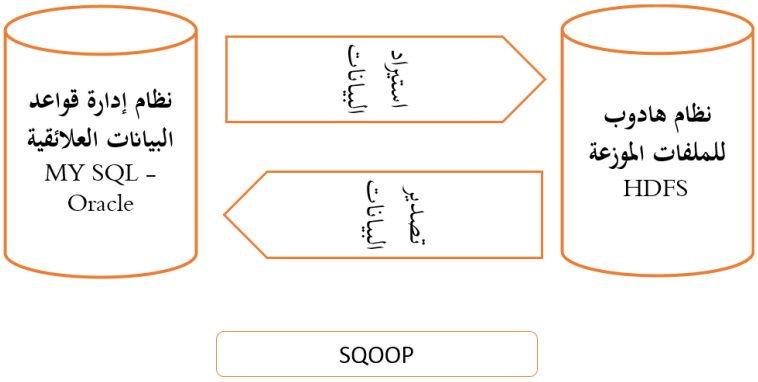
وتستعرض الفقرات التالية نماذج من أنظمة كل طبقة من طبقات منصة هادوب، مع التركيز على ملامح كل منها وخصائصها المميزة إلى جانب إلقاء الضوء على آلية العمل.

أولًا: طبقة استيعاب البيانات: Data ingestion layer

كما سبقت الإشارة هي الطبقة المسئولة عن إدخال وتجميع البيانات إلى منصة هادوب، وتضم مجموعة الأدوات لعل من أهمها Scoop وFlume.

1. أداة نقل البيانات: Apache Sqoop ([3])

تعتبر أداة اتصال ونقل لكونها مصممة لنقل كميات كبيرة من البيانات بشكل فعال من منصة هادوب إلى قواعد البيانات العلائقية والعكس، وبالتالي تستخدم هذه الأداة بشكل أساسي لاستيراد وتصدير البيانات -كما هو مبين في الشكل رقم (8) -، حيث تستورد البيانات من مصادر خارجية وتتمثل تلك المصادر عادة في أنظمة إدارة قواعد البيانات العلائقية وتنقلها إلى مكونات ذات الصلة بمنصة هادوب وبخاصة نظامها للملفات الموزعة "HDFS" أو "HBase" أو "Hive". وكما – سبقت الإشارة- تقوم أيضًا بتصدير البيانات من هادوب إلى تلك المصادر الخارجية، ومن نماذج قواعد البيانات العلائقية التي تعمل وتتوافق مع أداة "Sqoop" تأتي أنظمة "Teradata" و"Netezza" و"Oracle" و "MySQL".



شكل (8) آلية عمل Apache Sqoop

2. أداة إدارة السجلات: Flume ([4])

تتبنى فكرة تقنية "Sqoop" نفسها، ولكنها تعمل على مستوى السجلات "Logs" وتتولى مسئولية جمع وتوليف كميات كبيرة من البيانات، والتي يمكن الحصول عليها من خوادم متعددة، وإدارة نقلها بكفاءة من مصادرها وإرسالها مرة أخرى إلى "نظام هادوب للملفات الموزعة" "Hadoop Distributed File System" "HDFS"، وبناء عليه تمثل "Flume" خدمة موزعة وتحظى بالموثوقية لتوليد

(3) لمزيد من المعلومات حول Sqoop يمكن الاطلاع على الرابط https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial

(4) لمزيد من المعلومات حول Flume يمكن الاطلاع على الرابط https://flume.apache.org

السجلات بفعالية مستهدفة من وراء ذلك تجميع ونقل كمية كبيرة من بيانات السجل، وتتسم بنيتها بالبساطة والمرونة والقابلية للتوسع، واعتمادها على تدفق البيانات من المصدر إلى بيئة هادوب، وكذلك تسامحها مع الخطأ المحتمل في عمل الحاسبات الآلية.

ثانياً: طبقة التخزين: Data storage

تعد الطبقة المسئولة عن تخزين البيانات ولعل من أبرز الأدوات العاملة في نطاقها "HDFS" و" Hbase"

1. نظام هادوب للملفات الموزعة: HDFS [5]

يمثل "نظام هادوب للملفات الموزعة" "Hadoop Distributed File System" أداة التخزين الأساسية في منصة هادوب، ويقوم بعرض طرق لتخزين كميات ضخمة من الملفات الصغيرة بأشكال متنوعة، وذلك في العديد من الأجهزة، وبشكل موثوق حتى في حالة فشل بعض الأجهزة عن العمل.

ويتألف نظام هادوب للملفات الموزعة من مكونين أساسيين: يطلق على الأول: "NameNode" ويعمل بصفته جهاز "عقدة" رئيس "Master" وذلك في مجموعة الكتل في هادوب "Hadoop clusters"، ويقوم هذا المكون بتخزين البيانات الوصفية، مثل: البيانات حول عدد الكتل والنسخ المتماثلة وغيرها من البيانات والتفاصيل الأخرى، ويتولى أيضًا تعيين المهام التي ينبغي أن يقوم بها الجهاز "العقدة" التابع "Slave node"، والتي ينبغي نشرها على أجهزة موثوقة كونها تمثل حجر الزاوية في "نظام هادوب للملفات الموزعة"، ويتمثل المكون الثاني في: "DataNode" ويكون بمثابة "تابع" في إطار مجموعة الكتل في هادوب، وتكمن مسئوليته في تخزين البيانات في نظام هادوب للملفات الموزعة، وإجراء عملية القراءة والكتابة حسب طلب المستخدمين.

و الملف يتم تقسيمه إلى عدة أجزاء تسمى: كتل "Bulks"، ويتم نسخ كل جزء بشكل افتراضي وتلقائي على ثلاثة أماكن يأخذ كل منها القيمة (128) ميجا بايت، على سبيل المثال: إذا كانت مساحة الملف (300) ميجابايت، فيحتل القسم الأول مساحة (128) ميجا بايت، والثاني يأخذ نفس القيمة، وأما القسم الثالث فيشغل المساحة التخزينية المتبقية وهي (44) ميجا بايت فقط.

2. أداة قواعد البيانات: HBase [6]

ينظر إلى "Apache HBase" باعتبارها قاعدة بيانات لا تتشابه ولا تتوافق مع مبادئ لغة الاستعلام الهيكلية أو البنيوية "NoSQL"، وتختزن البيانات المهيكلة في جداول يمكن أن تحتوي على ملايين الصفوف والأعمدة، وتوفر هذه الأداة إمكانية الوصول المتزامن أو اللحظي لقراءة البيانات و/ أو كتابتها في نظام هادوب للملفات الموزعة.

ويضم "HBase" مكونين أساسيين: الأول: "HBase Master" والمسئول عن إجراءات إدارة واجهة إنشاء الجداول وتحديثها وحذفها إلى غير ذلك، بينما يتألف المكون الثاني من: "Region Server" والمسئول عن التعامل مع طلبات القراءة والكتابة والتحديثات والحذف من جانب المستخدمين.

---

(5) لمزيد من المعلومات حول HDFS يمكن الاطلاع على الرابط https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

(6) لمزيد من المعلومات حول Apache Hbase يمكن الاطلاع على الرابط https://hbase.apache.org

ثالثاً: طبقة المعالجة: Data Processing layer

وهي الطبقة المسئولة عن معالجة البيانات والتي من أمثلتها تقنيات MapReduce- Spark – Mahout

1. أداة المعالجة: MapReduce ([7])

يطلق عليها طبقة معالجة البيانات الضخمة سواء أكانت مهيكلة أو غير المهيكلة والمخزنة في نظام هادوب للملفات الموزعة، وتمثل المحرك الذي يعمل في الخلفية لتنفيذ المهام، وتكمن آلية العمل في تقسيم المهمة أو الوظيفة إلى مجموعة من المهام المستقلة والفرعية، وكذلك إلى عدد من المراحل وتوزيعها على عدد من الأجهزة لتنفيذها بسرعة، وتنقسم المعالجة التي تقوم بها تقنية "MapReduce" إلى مرحلتين رئيستين وهما: "Map" و "Reduce"، وتمثل "Map" المرحلة الأولى من المعالجة ويحدد فيها ما يسمى بالكود المنطقي المعقد "Complex logic code"، بينما يطلق على المرحلة الثانية "Reduce" ويحدد فيها المعالجة الخفيفة الوزن مثل التجميع/ الجمع Aggregation/Summation.

2. تقنية: Mahout ([8])

ينظر إلى هذه التقنية كونها إطار عمل مفتوح المصدر، يُستخدم لإنشاء خوارزمية تعلم آلي "Machine learning" قابلة للتطوير، وبمجرد تخزين البيانات في نظام هادوب للملفات الموزعة، توفر هذه التقنية أدوات علوم البيانات للعثور تلقائيًا على أنماط ذات معنى في مجموعات البيانات الضخمة.

3. تقنية: Spark([9])

تركز على تحليل البيانات لحظيًا "أي: في الوقت نفسه" "Real time data analysis"، وبالتالي تمثل نظام معالجة موزع مفتوح المصدر يستخدم للتعامل مع الأحمال المحتملة للبيانات الضخمة، ويستخدم التخزين المؤقت في الذاكرة وتحسين تنفيذ الاستعلام على البيانات بأي حجم، ويعد Spark محركًا سريعًا وعامًا لمعالجة البيانات على نطاق واسع، ويكمن سبب السرعة في أنه يعمل على الذاكرة العشوائية "RAM"، وهذا يجعل المعالجة أسرع بكثير مقارنة بمحركات الأقراص، ويعني الجزء العام "General" أنه يمكن استخدامه لأشياء متعددة مثل: تشغيل لغة البرمجة الهيكلية "SQL" الموزعة، وإنشاء البيانات واستيعابها في قاعدة البيانات، وتشغيل خوارزميات التعلم الآلي، والعمل مع الرسوم البيانية، ودعم تدفقات البيانات، إلى غير ذلك.

رابعاً: طبقة الإدارة: Management

تشتمل هذه الطبقة على مجموعة التطبيقات المسئولة عن إدارة النظام أو كما يطلق عليها إدارة المصادر Resource Management، ومن تطبيقاتها Oozie – Ambari –Zoo keeper -YARN

(7) لمزيد من المعلومات حول MapReduce يمكن الاطلاع على الرابط https://www.edureka.co/blog/mapreduce-tutorial

(8) لمزيد من المعلومات حول Mahout يمكن الاطلاع على الرابط http://mahout.apache.org

(9) لمزيد من المعلومات حول Spark يمكن الاطلاع على الرابط https://spark.apache.org

1. أداة إدارة البيانات: YARN ([10])

تعمل أداة "Yet Another Resource Negotiator" "YARN" على إدارة المصادر في منصة هادوب، ومن مسئوليتها إدارة ومراقبة أعباء العمل وتنفيذ الضوابط الأمنية التي يتم تحديدها، وتعتبر منصة مركزية لتقديم أدوات إدارة البيانات عبر مجموعات الكتل داخل هادوب، ويمكن أيضًا لمنصة هادوب بفضل تقنية "YARN" دعم بنية بحيرة البيانات والسابق ذكرها في جزئية خصائص هادوب.

وتتميز هذه التقنية بدعمها لأدوات معالجة البيانات وإدارة التدفق الآني أو اللحظي للبيانات "Real-time streaming". وتتألف "YARN" من مكونين: يطلق على الأول إدارة المصادر أو الموارد "Resource Manager" ويعمل على الجهاز الرئيسي " Master machine"، ومن ثم فإنه يدير الموارد والمصادر من خلال مدير التطبيقات "Application Manager" ويعمل على جدولة التطبيقات التي تعمل في "YARN" عن طريق المجدول "Scheduler"، ويضم المكون الثاني: مدير العقدة Node Manager والذي يعمل على الجهاز التابع "Slave machine"، ويتواصل باستمرار مع مدير أو إدارة المصدر ليبقى محدثًا.

2. أداة التنسيق: Zookeeper ([11])

تعتبر بمثابة خدمة مركزية تحافظ على معلومات إعدادات التكوين أو الضبط والتسمية وتوفير المزامنة الموزعة، كما تقوم بإدارة التنسيق بين أعباء ومهام مجموعة كبيرة من الأجهزة العاملة ضمن منظومة منصة هادوب، ولعل من أبرز فوائد Zookeeper تأتي السرعة مع تحمل أعباء العمل، وكذلك قدرته على الاحتفاظ بجميع المعاملات وترتيبها.

3. أداة المراقبة والتأمين: Ambari ([12])

تمثل منصة مفتوحة المصدر لإدارة ومراقبة وتأمين هادوب، وبالتالي أصبحت إدارة منصة هادوب تتسم بالبساطة لكون Ambari توفر نظامًا أساسيًا متسقًا وآمنًا للتحكم التشغيلي.

ومن فوائد هذه الأداة تأتي بساطة عمليات التثبيت والتكوين والإدارة، وإنشاء وإدارة مجموعات الكتل "Clusters" على نطاق واسع بسهولة وكفاءة، وكذلك ضبط إعدادات الأمان المركزي حيث تقوم بتكوين مواصفات ومتطلبات أمان المجموعة عبر النظام الأساسي، وتقليل تعقيدات الإدارة، وهناك القابلية للتوسيع وإدخال الخدمات المخصصة.

4. أداة تدفق البيانات: Oozie ([13])

تتولى هذه الأداة جدولة سير عمل وتدفق البيانات، وبالتالي تم دمج إطار عملها بالكامل مع إدارة وظائف منصة هادوب، كما تدعم وظائف كل من: "MapReduce" و "Pig" و "Hive" و "Sqoop" و "YARN"، وتتسم هذه الأداة بقابليتها للتطوير والمرونة، وكذلك إمكانية بدء المهام والوظائف وإيقافها وتعليقها وإعادة تشغيلها بسهولة، وإعادة تشغيل مهام سير العمل الفاشلة، وهناك نوعان

(10) لمزيد من المعلومات حول YARN يمكن الاطلاع على الرابط https://intellipaat.com/blog/apache-hadoop-yarn

(11) لمزيد من المعلومات حول Zookeeper يمكن الاطلاع على الرابط https://zookeeper.apache.org

(12) لمزيد من المعلومات حول Ambari يمكن الاطلاع على الرابط https://ambari.apache.org

(13) لمزيد من المعلومات حول Oozie يمكن الاطلاع على الرابط https://oozie.apache.org

أساسيان من وظائف: "Oozie" يضم الأول: تخزين وتشغيل مهام سير العمل المكونة من وظائف هادوب مثل: "MapReduce"، "Pig"، "Hive"، ويتمثل النوع الثاني في: التنسيق والذي يدير وظائف سير العمل بناءً على جداول زمنية محددة مسبقًا.

خامساً: طبقة الاستعلام: Querying layer

تشكل الطبقة المسئولة عن إدارة الاستعلامات والاستفسارات التي تجرى على البيانات في إطار منصة هادوب، ومن نماذج التقنيات التي توجد في هذه الطبقة Pig – Hive - Hcatalog

1. أداة الاستعلام والتحليل: Hive [14]

تعد مستودع بيانات مفتوح المصدر تستخدم للاستعلام، وكذلك تحليل ومعالجة مجموعات البيانات الضخمة سواء أكانت منظمة أو غير المنظمة، والمخزنة في "نظام هادوب للملفات الموزعة". وتتوافق هذه الأداة مع لغة الاستعلام المهيكلة "البنيوية" "SQL"، ويستخدم "Hive" اللغة المسماة HQL (HiveQL)، والتي تشبه لغة الاستعلام المهيكلة، وتقوم "HiveQL" تلقائيًا بترجمة الاستفسارات المشابهة لـلغة الاستعلام المهيكلة إلى وظائف "MapReduce"، وتعطى دقة في عملية استخلاص البيانات وذلك بطريقة أكثر ذكاء.

2. أداة الاستعلام والتحليل: Pig [15]

تمثل منصة تم تطويرها لتنفيذ الاستعلامات على مجموعات البيانات الضخمة المخزنة في نظام هادوب للملفات الموزعة، وتتشابه مع لغة الاستعلام المهيكلة، ويقوم "Pig" بتحميل البيانات وتطبيق المرشحات (الفرز) المطلوبة وتفريغ البيانات بالتنسيق والشكل المطلوب من جانب المستخدم.

ومن أبرز خصائص تلك الأداة تأتي القابلية للتوسيع حيث يمكن لمستخدميها إنشاء وظائف مخصصة لتلبية متطلبات المعالجة الخاصة بهم، وكذلك خاصية التحسين الذاتي والتلقائي للنظام "Self-optimizing"، ولذلك يمكن للمستخدم التركيز على الدلالات والتعامل مع جميع أنواع البيانات، حيث تقوم الأداة بتحليل كل من البيانات المهيكلة وغير المهيكلة.

3. أداة: Hcatalog [16]

تعد مكونًا رئيسًيا في "Hive" ويمكن النظر إليها كأداة تمكن المستخدم من تخزين البيانات بأي تنسيق "شكل" وبنية، كما تدعم مكونات هادوب لقراءة البيانات وكتابتها بسهولة، ومن مزاياها توفير تقنيات ومتطلبات فلترة البيانات وأرشفتها، ومساعدة المستخدم على تخطي أعباء تخزين البيانات، وتمكين إخطارات توافر البيانات.

(14) لمزيد من المعلومات حول Hive يمكن الاطلاع على الرابط https://hive.apache.org

(15) لمزيد من المعلومات حول Pig يمكن الاطلاع على الرابط https://pig.apache.org

(16) لمزيد من المعلومات حول Hcatalog يمكن الاطلاع على الرابط https://www.tutorialspoint.com/hcatalog/index.htm

سادساً: التطبيقات الإضافية:

1. أداة: Kafka [17]

تكون بمثابة وسيط بين المستودع المختزن به البيانات والمستخدم وذلك بدلًا من نسخ البيانات بين المرسل والمستخدم، وتعمل على معالجة تدفق البيانات "Streaming" في الوقت الفعلي وبشكل مستمر ومتزامن، وتعتبر مكتبة لبناء تطبيقات التدفق، وتحديدًا التطبيقات التي تحول موضوعات المدخلات إلى مخرجات أو استدعاء الخدمات الخارجية، وتحديث قواعد البيانات إلى غير ذلك.

2. أداة تدفق وبث البيانات: Strom [18]

تقوم بما يقوم به "Flume" ولكنه يزيد، بمعنى: في "Flume" يتم نقل بيانات من مكان (أ) إلى مكان (ب) فإنها تنتقل كما هي بدون أية تعديلات، ولكن مع "Storm" يمكن إجراء بعض التعديلات عليها، وذلك بالإضافة إلى إمكانية نقلها.

ويمثل بالتالي نظام لمعالجة تدفق البيانات بشكل موثوق وبثها بشكل مجاني ومفتوح المصدر وموزع، ويمكن استخدامه مع أية لغة برمجة، ويتسم بسهولة الاستخدام، ويحتوي على العديد من حالات الاستخدام من أمثلتها التحليلات في الوقت الفعلي (اللحظية أو الآنية)، والتعلم الآلي عبر الشبكة العالمية، وغيرها، ويتميز بالسرعة والقابلية للتطوير، وتحمل الأخطاء، وضمان معالجة البيانات، وسهولة إعداده وتشغيله.

3. أداة التأمين: Knox [19]

يعد إنشاء هوية المستخدم بمصادقة قوية أساس الوصول الآمن في هادوب، ويحتاج المستخدمون إلى تعريف أنفسهم بشكل موثوق، ثم نشر هذه الهوية عبر منصة هادوب، وتستهدف هذه الأداة تأمين الوصول، وتمثل نظام وصول المستخدمين لهادوب دون تقليل معدلات الأمان، وتعمل Knox أيضًا على تبسيط أمان المستخدمين الذين يصلون إلى بيانات في هادوب ويقومون بتنفيذ المهام.

4. أداة المراقبة: Ranger [20]

تمثل إطار عمل لتمكين مراقبة توفير أمان شامل للبيانات عبر بيئة نظام منصة هادوب، ويحتاج تأمين البيانات داخل هادوب إلى التطوير لدعم حالات الاستخدام المتعددة للوصول إلى البيانات، مع توفير إطار عمل للإدارة المركزية لسياسات الأمان والمراقبة والتحكم في وصول المستخدم.

ثالثاً: التحليل والمناقشة:

قامت الدراسة بحصر- كما هو مبين في الجدول رقم (1)- نماذج من المبادرات التي قادتها بعض مؤسسات المعلومات المتطورة لتوظيف البيانات الضخمة لتعزيز خدماتها وأدائها، وذلك بهدف الاسترشاد بها في تحقيق أهداف الدراسة والإجابة على تساؤلاتها.

---

(17) لمزيد من المعلومات حول Kafka يمكن الاطلاع على الرابط https://kafka.apache.org

(18) لمزيد من المعلومات حول Storm يمكن الاطلاع على الرابط https://storm.apache.org

(19) لمزيد من المعلومات حول Apache Knox يمكن الاطلاع على الرابط https://knox.apache.org

(20) لمزيد من المعلومات حول Apache Ranger يمكن الاطلاع على الرابط https://ranger.apache.org

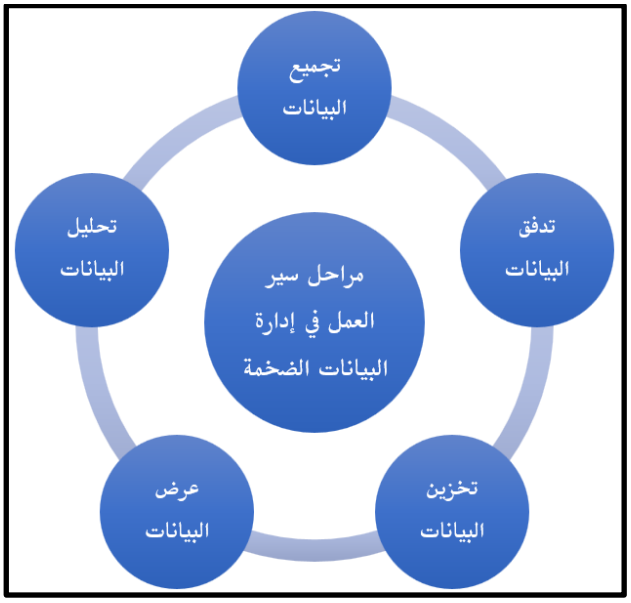
جدول (1) مبادرات البيانات الضخمة في مؤسسات المعلومات

المبادرة	العنوان على الشبكة العنكبوتية العالمية
KNEXT	**https://knext.ischool.umd.edu/**
Canada's Libraries and Archives	**https://cwrc.ca/rsc-src/**
National Science Digital Library NSDL	**https://nsdl.oercommons.org/**
UCL Research Data Repository	**https://www.ucl.ac.uk/library/research-support/research-data-management/ucl-research-data-repository**
Twitter Archive at the Library of Congress	**https://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/**
Harvard University Library: Big Data Applications for Books	**https://library.harvard.edu/services-tools/harvard-library-apis-datasets**
Harvard Library. "Harvard Library Open Metadata	**https://emeritus.library.harvard.edu/open-metadata#Harvard-Library-Bibliographic-Dataset**
Creation of a Metadatabase for Geophysical Data in Australia	**http://www.ga.gov.au/about/facilities/geophysical-network/metadata-and-data-archival**
Jisc & HESA Library Data Labs Project	**https://libraryservices.jiscinvolve.org/wp/2016/10/library-data-labs-project/**
Brooklyn Public Library (BPL): Big Data for the Visualization of User Data	**https://diginomica.com/how-the-brooklyn-public-library-data-visualization-a-better-library-with-tableau**
University of Michigan Library Open Access Bibliographic Records	**https://www.lib.umich.edu/open-access-bibliographic-records**

وتركز هذه الجزئية من الدراسة على مناقشة وتحليل المراحل التي يتم من خلالها توظيف البيانات الضخمة لأغراض تطوير إجراءات البحث والاسترجاع، وتخصيص الخدمات التي تقدمها مؤسسات المعلومات المتطورة، وذلك في ضوء معالجة أسس ربط المستودعات الرقمية لتعزيز البحث المعلوماتي، وآليات تخصيص المحتوى والخدمات في بيئة البيانات الضخمة، ودراسة الأطر التي تمكن من تحسين تجربة المستخدم.

1/3. مراحل توظيف البيانات الضخمة لتطوير خدمات مؤسسات المعلومات:

ينبغي أن تمر إدارة البيانات الضخمة في مؤسسات المعلومات بمراحل سير عمل تتسم بكونها متداخلة وليست متتالية، وذلك لأغراض تجهيزها وتوظيفها لتعزيز البحث والاسترجاع وتخصيص خدمات المكتبات وغيرها من مؤسسات المعلومات، وتتمثل هذه المراحل كما هو مبين بالشكل رقم (9) في تجميع وتدفيق وتخزين وعرض وتحليل البيانات.



شكل (9) مراحل سير العمل في إدارة البيانات الضخمة

أولًا: مرحلة تجميع البيانات: Data collection

يقصد بهذه المرحلة جمع البيانات وتحديد مصادرها المختلفة وأنواعها، وكذلك آليات نقلها وتخزينها، وهناك مصادر متنوعة داخل وخارج مؤسسات المعلومات تعتبر روافد للتجميع، ومن ثم الحصول على البيانات والمعلومات ومن أمثلتها التطبيقات الحاسوبية، ومواقع وبوابات الشبكة العالمية، وشبكات التواصل الاجتماعي، والأجهزة الذكية، والخدمات الإلكترونية، والحوسبة السحابية، والأجهزة والمعدات، وما إلى ذلك ([21])، وتمثلت من هذا المنطلق أهم مصادر تجميع البيانات الضخمة لمؤسسات المعلومات في:

- التحليل الموضوعي ويضم التصنيف والتكشيف الآلي، والتوسيم الاجتماعي"Social tagging"، الفوكسونومى"Folksonomy".
- التحليل الببليوجرافي والذي يشتمل على توصيف البيانات والمحتوى، ومخططات هيكلة البيانات الوصفية.
- صفحات مؤسسات المعلومات المتاحة على شبكات ومنصات التواصل الاجتماعي والتي من نماذجها – Facebook Twitter – LinkedIn – Academic Social Networks، إلى غير ذلك.
- البيانات والمعلومات الناتجة عن استخدام المساعدات الرقمية الصوتية "DVA" وغيرها من تقنيات الثورة الصناعية الرابعة.
- نتائج دراسات سلوكيات وممارسات المستخدمين المعلوماتية من خلال تطبيقات الأجهزة وبخاصة الهواتف الذكية لتعظيم الاستفادة من خدمات مؤسسات المعلومات.
- البيانات والمعلومات الناتجة عن خدمات المعلومات الرقمية التي توفرها مؤسسات المعلومات في متناول روادها.

(21) تم معالجة هذه الجزئية بالتفصيل في الإطار النظري لهذه الدراسة تحت عنوان البيانات الضخمة: المفاهيم والأهمية والخصائص.

- إلى غير ذلك.

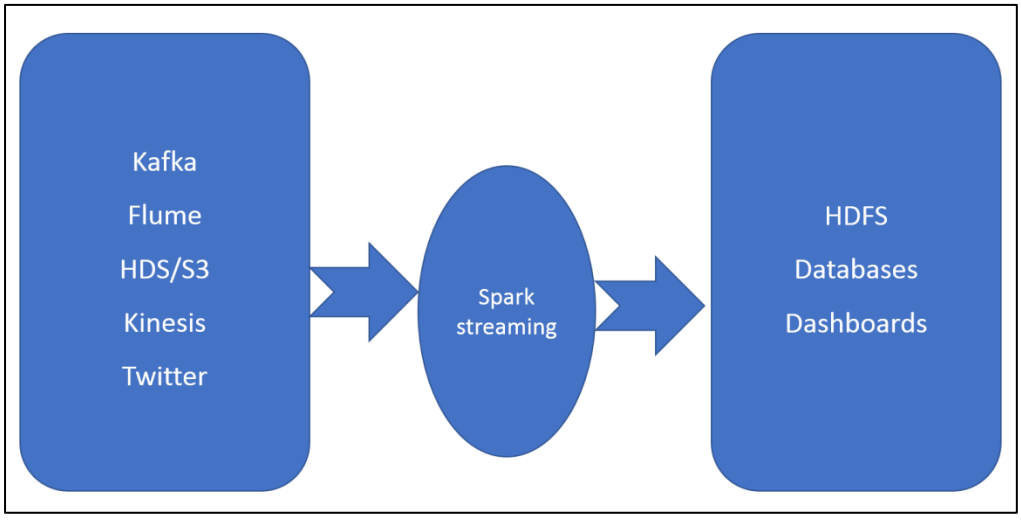
وينبغي على مؤسسات المعلومات الأخذ بعدة اعتبارات في تجميع البيانات واستيعابها مثل: توافر متطلبات الاتصال الجيد بالشبكة العالمية، وبروتوكولات اتصالات أسرع وأسهل ومخصصة لإنترنت الأشياء، والحوسبة والبنية السحابية مثل: خدمات Amazon – eBay - Azure وغيرها، ومع ضرورة مراعاة التقنيات المتباينة الموجهة لإدارة البيانات الضخمة، وظهور قواعد بيانات موزعة، والتدفق المستمر للبيانات، وتعدد مصادرها وتنوعها، وانتشار ظاهرة البيانات الموزعة.

وقد قادت التطورات التقنية المتلاحقة وبخاصة المنبثقة من الثورة الصناعية الرابعة والتي تعتمد على تطبيقات الذكاء الاصطناعي "AI"، إلى تعظيم توظيف هذه التقنيات وخدماتها في مؤسسات المعلومات المتقدمة، كما ساعدت جائحة كورونا (COVID-19) التي يمر بها العالم إلى توجه تلك المؤسسات إلى العالم الرقمي، والتوجه لتوفير خدمات المستفيدين منها عن بُعد.

ومن خلال تحليل مكونات منصة هادوب، يمكن أن تمارس تقنيات إدخال وتجميع البيانات الضخمة مثل: "Flume" و "Sqoop" دور مهم في هذه المرحلة من مراحل سير العمل.

ثانياً: مرحلة تدفق البيانات: Data Streaming

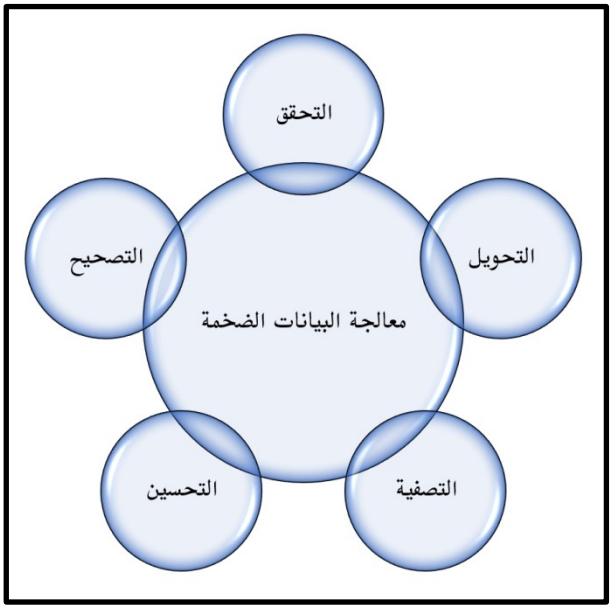
يتسم تدفق البيانات في مؤسسات المعلومات – نتيجة تعدد وتنوع مصادر الحصول عليها– بالعديد من الخصائص لعل من أبرزها الاستمرارية والسرعة والآنية والكم الهائل، وبالتالي لابد من توافر أدوات وتطبيقات تتحكم في إدارة تدفق البيانات ومن أمثلتها: "Spark streaming" وكذلك "Kafka" كما هو مبين بالشكل رقم (10).



شكل (10) آلي4ة التحكم في تدفق البيانات من خلال Spark Streaming

وتهدف معالجة وتحليل البيانات في مؤسسات المعلومات إلى العمل على تجهيزها لأغراض التحليل لتوظيفها في العمليات الفنية، وكذلك إتاحة خدمات معلومات مخصصة ومطورة في متناول المستخدمين، وتنطوي معالجة البيانات على تصحيح أية أخطاء و/ أو أوجه قصور بالبيانات وتحديد إجراءات تصويبها، لتكون في شكل معياري قابل للتحليل، وتمثل هذه الخطوة إجراء جوهريا وبخاصة في حالات البيانات غير المهيكلة (غير المنظمة)، والتي تقتضي وضع معايير واضحة للتصنيف وتحليل محتوى مصادر المعلومات وتحديد

الحقول والمتغيرات، وتتضمن معالجة البيانات -كما هو مبين في الشكل رقم (11)- العديد من الإجراءات ولعل من أبرزها: التحقق والتحويل والتصفية والتحسين والتصحيح (Golub & Hansson, 2017) .



شكل (11) إجراءات معالجة البيانات

أ. **التحقق** Validation: ويمثل إحدى إجراءات معالجة البيانات الضخمة ويهدف إلى التأكد من البيانات المدخلة آمنة ولا تهدف إلى تخريب النظام المتكامل لمؤسسة المعلومات أو أي شيء من هذا القبيل، وتلجأ عملية التحقق من البيانات إلى استخدام بعض القواعد للتأكد من البيانات التي يتم إدخالها للنظام صحيحة ومفيدة وآمنة، ويكمن غرض التحقق من صحة البيانات في توفير ضمانات تحري الكفاءة والدقة والاتساق لأي نوع من أنواع المدخلات المختلفة للمستخدم في تطبيق أو نظام يتوافق مع إدارة البيانات الضخمة.

ب. **التحويل** Transformation: تعتبر من المراحل المهمة في تجهيز البيانات للتحليل والتنقيب عنها، وذلك لكونها تستهدف المساعدة في الحصول على أفضل النتائج والممارسات في عمليات تحليل محتوى مصادر المعلومات أثناء الإجراءات الفنية بمؤسسات المعلومات ورفع الكفاءة، وتعمل على تبسيط فهم الأنماط والارتباطات التي يتم استكشافها.

ج. **التصفية** Filtering: تشير الفلترة إلى مجموعة الإستراتيجيات أو الحلول لتحسين البيانات والمعلومات، ويعني هذا أن البيانات يتم تنقيحها والعمل على تخصيصها لتلائم احتياجات المستخدم، مع الأخذ بعين الاعتبار تفادي إدراج أية بيانات أخرى يمكن أن تكون متكررة أو غير ذات صلة أو حتى حساسة، ويمكن استخدام أنواع عديدة من عوامل تصفية البيانات لتخصيص نتائج البحث واسترجاع المعلومات، ولعل من أبرز الأمثلة على ذلك في أنظمة مؤسسات المعلومات استخدام خيارات البحث المتقدم.

د. **التحسين** Enrichment: تستهدف عملية تحسين المعلومات إثراء البيانات أو زيادتها من خلال استكمال البيانات المفقودة أو غير المكتملة وإضافة بيانات جديدة، وعادًة ما يتم إثراء البيانات باستخدام مصادر بيانات خارجية مثل أنظمة إدارة قواعد البيانات، وتكون مجمعة وفق سياق ذي صلة، وتتضمن دمج التحديثات والمعلومات الجديدة في قاعدة بيانات مؤسسات المعلومات لتحسين الدقة

والمساعدة في اتخاذ القرارات، ويمثل هذا الإجراء دورا مهما في رفع القيمة المضافة للإجراءات الفنية بمؤسسات المعلومات، وبالتالي رفع فعالية وكفاءة الخدمات وتعظيم ارتباطها باحتياجات المستفيدين.

ه. **التصحيح** Cleansing: تمثل إجراءات تنظيف البيانات وتصحيحها واكتشاف وتصحيح أو إزالة بيانات تالفة أو غير دقيقة، وتشير أيضًا إلى تحديد أجزاء غير كاملة أو غير صحيحة أو غير دقيقة أو غير ذات صلة بالبيانات، ثم العمل على تبني آليات تستهدف استبدال أو تعديل أو حذف البيانات غير المرغوب فيها، ويفيد مثل هذا الإجراء مؤسسات المعلومات في التعامل مع ظاهرة تقادم المعلومات وكيفية إدارة البيانات والمعلومات المتقادمة والتي تحتاج إلى تحديث سواء بالحذف أو بالتعديل أو بالإضافة.

ثالثاً: مرحلة تخزين البيانات: Data storage

اعتمدت مؤسسات المعلومات لفترات طويلة على الطرق التقليدية في تخزين البيانات بواسطة أنظمة إدارة قواعد البيانات العلائقية، ولغة الاستعلام الهيكلية "البنيوية"، والتي لا تتناسب ولا تتوافق مع طبيعة وسمات البيانات الضخمة، ونتيجة لذلك ظهرت تقنيات يمكن أن تعمل مؤسسات المعلومات على توظيفها كونها تمكن من تخزين البيانات الضخمة بشكل أفضل، ومنها على سبيل المثال قواعد البيانات غير مهيكلة "No SQL"، وقواعد البيانات الموزعة، وتتسم هذه التقنيات بكونها منخفضة التكلفة إلى حد كبير مقارنة بأنظمة قواعد البيانات العلائقية، وتخزن فيها البيانات بشكل مستقل، ومن أمثلتها نظام هادوب للملفات الموزعة وأنظمة " Mango DB"(22) و "Amazon DynamoDB"(23) و "Redis"(24) و Neo 4J(25) و JASON JavaScript Object Notation(26) و Hbase.

وتؤكد هذه الدراسة على تمثيل تقنيات التخزين والحوسبة السحابية لمؤسسات المعلومات أحد الحلول المقترحة لتجاوز معوقات المساحة التخزينية للبيانات ومعالجتها، وبخاصة في ضوء تخصيصها لمستودع بيانات سحابية خاص بها ويوظف فيه أنظمة التشفير المتقدمة، وكذلك تطبيق إجراءات صارمة للمحافظة على الخصوصية وأمن البيانات والمعلومات.

وتشير الدراسة الحالية إلى أهمية توجه مؤسسات المعلومات المتطورة نحو تبني أدوات تخزين البيانات الضخمة والتي من نماذجها نظام هادوب للملفات الموزعة، كونه يتيح أساليب لتخزين كميات ضخمة من الملفات الصغيرة وذلك وفق أشكال متنوعة وبشكل آمن في العديد من الأجهزة، ويمكن أيضًا توظيف أداة قواعد البيانات "Hbase" والتي تختزن البيانات المهيكلة في جداول يمكن أن تحتوي على ملايين الصفوف والأعمدة، والتي توفر إمكانية الوصول المتزامن أو اللحظي لقراءة البيانات أو كتابتها في نظام هادوب للملفات الموزعة.

(22) لمزيد من المعلومات حول MangoDB يمكن الاطلاع على الرابط https://www.mongodb.com

(23) لمزيد من المعلومات حول Amazon DynamoDB يمكن الاطلاع على الرابط https://aws.amazon.com/dynamodb

(24) لمزيد من المعلومات حول Redis يمكن الاطلاع على الرابط https://redis.io

(25) لمزيد من المعلومات حول Neo 4J يمكن الاطلاع على الرابط https://neo4j.com

(26) لمزيد من المعلومات حول JASON يمكن الاطلاع على الرابط https://www.json.org/json-en.html

رابعاً: مرحلة عرض البيانات: Data visualization

تصور وعرض البيانات هو التمثيل الرسومي للبيانات والمعلومات، وذلك باستخدام العناصر المرئية مثل: المخططات والرسوم البيانية والخرائط، وتوفر أدوات تصور البيانات طريقة تسهل رؤية وفهم الاتجاهات والقيم والأنماط في البيانات، وتعد أدوات وتقنيات تصور وعرض البيانات ضرورية في منظومة البيانات الضخمة نظرًا لتحليل كميات هائلة من المعلومات واتخاذ قرارات تعتمد على البيانات، ومن أهم الأنواع العامة الشائعة تأتي الرسوم البيانية، والجداول، والخرائط، وهناك أمثلة أكثر تحديدًا لطرق تصور البيانات منها: المخطط المساحي، والمخطط الشريطي، والرسم البياني النقطي، وخرائط التوزيع النقطي، والمخططات الزمنية مثل: جانت "Gantt"، والمصفوفات، والشبكة، والمخططات ثنائية وثلاثية الأبعاد.

وتتنوع أهداف إدارات وأقسام ووحدات مؤسسات المعلومات في الوصول للبيانات والمعلومات من خلال آليات تخصيص عرض وتصور البيانات، فمنها: الاختصار "Summarization"، والبحث "Querying"، والترتيب "Sorting"، والتجميع "Aggregation"، والتقارير "Reporting".

خامساً: مرحلة تحليل البيانات: Data analysis

تأتي هذه المرحلة بعد جمع البيانات والمعلومات وتنظيمها وترتيبها لتسهيل تحليلها، وذلك من أجل إخراجها وإبرازها على شكل معلومات يتم استخدامها بهدف الإجابة على أسئلة معينة، وتكمن آلية عمل تحليل البيانات في استخدام مؤسسات المعلومات لمنصات تخزين البيانات الضخمة، وبمجرد أن تصبح البيانات كاملة وجاهزة، يتم تحليلها بواسطة برامج تحليل عالية الجودة ومن نماذجها: لغة بايثون "Python"، أو لغة "R"، ولغة "SQL"، ولغة "Scala" وغيرها من لغات البحث والاستعلام، ويتم دعم هذه اللغات من خلال تقنيات "SQL-on-Hadoop"، حيث تمتلك هذه البرامج أدوات خاصة للقيام بعملية التحليل، وأبرزها:

- أدوات التنقيب عن البيانات: والتي تقوم بتنقيح البيانات، بالإضافة إلى البحث عن جميع أنماط البيانات.
- أدوات التحليل التنبؤي: والتي تعتمد على أنظمة الذكاء الاصطناعي في بناء نماذج الاحتمال والتخطيط والتنبؤ بالتطورات المستقبليّة، وبسلوكيات ومتطلبات المستفيدين من خدمات مؤسسات المعلومات.
- أدوات التعلم الآلي: والتي تعمل على تحليل كميات ضخمة من البيانات بالاعتماد على نماذج من الخوارزميات المتقدمة.

ويفيد التعامل مع حجم ضخم للغاية من البيانات في الاستفادة منه في دراسة الماضي والتعرف على الحاضر والتنبؤ والتخطيط للمستقبل، ووفقًا لما ذكره (سرحان، 2016) هناك أساليب مختلفة لتحليل البيانات يمكن أن تتبع مؤسسات المعلومات أحداها أو أكثر، وتتمثل في التسلسل الزمني "Time series"، والترابط "Correlation"، والتقارب "Clustering"، والتصنيف "Classification"، والتقدير "Estimation"، والترابطية "Connectionism"، والعلاقات (بين البيانات) "Association"، والتشابه "Similarity"، والتحليل النصي "Text analysis"، والارتباط "Correlation"، وهناك أدوات وتقنيات تقوم بهذا العمل ويقتصر دور العنصر البشري على إدخال البيانات للحصول على أنماط تحليل مناسبة لها، ومثلت "Python" أحد أهم التطبيقات التي استخدمتها مؤسسات المعلومات المتطورة في تحليل بياناتها.

وبعد هذا العرض لمراحل توظيف البيانات الضخمة في مؤسسات المعلومات والتي تمثلت في تجميع وتدفق وتخزين وعرض وتحليل البيانات، تركز الفقرات التالية على استشراف دورها في تطوير البحث والاسترجاع، وتخصيص خدمات مؤسسات المعلومات، والكشف عن آليات تحسين تجربة المستخدم.

2/3. البيانات الضخمة وتطوير البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات:

مع ظهور تطبيقات وأنظمة متطورة قادرة على معالجة وتحليل البيانات الضخمة واستثمار مخرجاتها في دعم أسس ومقومات اقتصاد المعرفة، أيقنت العديد من مؤسسات المعلومات –وبخاصة موردي وناشري قواعد المعلومات العالمية– أن نماذج أعمالها التقليدية "Business Model" لم تعد متوافقة مع التقدم التقني المتسارع سواء على مستوى بناء وتنمية المجموعات وأسس تطويرها، والعمليات الفنية التنظيمية، وكذلك خدماتها المعلوماتية وربطها بتلبية احتياجات المستخدمين المتنامية والمعقدة.

ومن هذا المنطلق تتناول الجزئية التالية أبرز آليات استثمار البيانات الضخمة في مؤسسات ومراكز المعلومات عبر الإجابة على الاستفسار الرئيسي المرتبط بالدراسة والذي يدور حول كيفية استخدام مؤسسات المعلومات لإطار البيانات الضخمة لتحسين البحث والاسترجاع وتخصيص المحتوى والخدمات، وتتركز الإجابة حول سبل تنفيذ البيانات الضخمة لربط العديد من المستودعات الرقمية لرفع مستوى أداء محركات وتقنيات البحث والاسترجاع وتوفير نتائج أكثر ارتباطًا بالاستفسارات المطروحة، وكذلك كيفية استثمار هذه التقنيات لتخصيص المحتوى وتعزيز الخدمات، وتحسين وإثراء تجربة المستخدم ودعم الاقتصاد المعرفي لمؤسسات المعلومات.

1/2/3. تطوير البحث والاسترجاع عبر ربط المستودعات الرقمية:

استخدمت المستودعات الرقمية من قبل الجامعات والمكتبات وغيرها من مؤسسات المعلومات لتخزين بياناتها الببليوجرافية ومصادر معلوماتها العلمية و/ أو المؤسسية، ومن ثم إتاحة البيانات الوصفية على الشبكة العنكبوتية العالمية (الويب) وذلك من خلال بروتوكول "OAI-PMH"(27). ومع ظهور تقنيات الويب الدلالي برز اهتمام مزودي المستودعات الرقمية بنشر وإثراء المحتوى الخاص بهم باستخدام تقنيات البيانات المرتبطة "Linked Data".

وتتضمن عادة المستودعات الرقمية أنواع متعددة من مصادر المعلومات منها: الكتب الإلكترونية ومقالات الدوريات والمجلات وأخبار ومقالات الصحف وأعمال المؤتمرات والمستخلصات والكشافات، بالإضافة إلى المصادر المتاحة للوصول الحر المجاني والتي تكون عادة ذات قيمة مضافة، وتعمل مؤسسات المعلومات خاصة مع زيادة حجم المحتوى وتنوعه على جذب مجموعات متباينة ومتنوعة من المستخدمين للاشتراك في خدماتها.

ومن الثابت أن تجميع وتنظيم وتكشيف المحتوى من مستودعات رقمية متعددة يجعل البحث سريعًا وبسيطًا للمستخدمين، ويمكن النظر إلى الربط بين المستودعات الرقمية للناشرين على سبيل المثال باعتبارها خدمات تهدف إلى تسهيل وصول مؤسسات المعلومات لمصادر الناشرين والموردين –سواء التي يتم الحصول عليها عن طريق الشراء كما هو الحال في الكتب الرقمية أو الاشتراكات مثلما هو مطبق في الدوريات الإلكترونية وقواعد المعلومات العلمية أو المصادر المتاحة للوصول المجاني أو الحر–. وتجدر الإشارة بأن عملية وصول مؤسسات المعلومات والاطلاع على النص الكامل للمصادر تتم في ضوء تراخيص واتفاقيات الاستخدام الموقعة بينها

(27) لمزيد من المعلومات حول برتوكول OAI-PMH يمكن الاطلاع على الرابط https://www.openarchives.org/pmh

من جهة والناشرين وموردي المحتوى من جهة أخرى (فرج، 2013)، وبالتالي تتولى أنظمة البيانات الضخمة على عاتقها إدارة مثل هذه الصلاحيات والتي من نماذجها "Apache Knox" – "Apache Ranger".

ويمكن لأنظمة إدارة البيانات الضخمة ممارسة دور فعال في تخطي إشكاليات إدارة ومعالجة المحتوى ذي القيمة العالية والنابع من عدة مستودعات رقمية، وتزويد المستخدمين بتجربة بحث ثرية وتوفير خدمات بحثية تقود إلى نتائج أكثر ارتباطًا بتلبية الاحتياجات المعلوماتية للمستخدمين، لأنه إذا لم يتمكن المستخدمون من الوصول للمحتوى، فبالتالي لا يمكنهم الاشتراك فيه، ومن هنا يمارس البحث الجيد بالنسبة لمؤسسات المعلومات دور فعال في المحافظة على المستخدمين، ويبدأ البحث الفعال عن المحتوى والخدمات بضرورة معالجة البيانات وإثرائها وتكشيفها بشكل مناسب، وكل هذه الإجراءات يمكن القيام بها بشكل منهجي من خلال بنية البيانات الضخمة (Teets & Goldner, 2013)

وتحاول الدراسة الحالية حصر أهم خصائص ومميزات توظيف أنظمة البيانات الضخمة لربط المستودعات الرقمية في النقاط التالية:

- نشر مجموعات ضخمة من البيانات وإتاحة في متناول الباحثين الوصول إلى مستودعات بيانات رقمية موثوقة وعالية الأداء.
- قابلية توصيف البيانات للتخصيص، حيث يمكن للمستخدمين والمؤسسات المسئولة عن إدارة المحتوى تحديد واستخدام مخططات البيانات الوصفية الخاصة بها لوصف مجموعات المصادر.
- التحكم المرن في الوصول، بمعنى: مجموعات المصادر المنشورة قد تكون مقيدة الوصول والإتاحة أو قد تكون مشاركتها حصريًا مع مجموعة معينة من المستخدمين أو مشاركتها بشكل عام، وبالتالي من الضروري أن يكون هناك مرونة في أساليب التحكم في الوصول إلى مصادر المستودعات الرقمية.
- المرونة في تدفقات البيانات الموجهة للمستخدم وذلك من خلال إتاحة وصول المستخدمين إلى مستودع تخزين بيانات موثوق وعالي الأداء.
- اختيار المعرف الثابت "Digital Object Identifier" DOI (28)، حيث ترتبط البيانات المختزنة في المستودعات الرقمية بمعرف ثابت يمكن الإشارة إليه بشكل فريد في مصادر المعلومات.
- إمكانات بحث فعالة لأنه يمكن وصف البيانات المكشفة بمجموعة متنوعة من البيانات الوصفية القياسية والمخصصة والمحددة المجال والتي تسهل وترفع من أداء إجراءات البحث والاسترجاع.

وفي ضوء ذلك يمكن لمؤسسات المعلومات تصور آلية الربط مع المستودعات الرقمية في السماح بإجراء البحث بشكل تزامني في قاعدة بيانات كل مستودع رقمي، بهدف إتاحة نتائج صادرة من عدة مستودعات في نفس الوقت للمستخدمين، وذلك وفق برتوكولات وتراخيص استخدام مع محركات بحث الناشرين وموردي خدمات مصادر المعلومات، وتجميع نتائج الاستعلامات البحثية، وتنسيق عرضها من خلال الفرز وحذف المكررات إلى غير ذلك.

(28) لمزيد من المعلومات حول DOI يمكن الاطلاع على الرابط https://www.doi.org

2/2/3. البيانات الضخمة ودورها في تخصيص المحتوى والخدمات:

يمكن أن تمارس تقنيات البيانات الضخمة دورا مهما في مساعدة مؤسسات المعلومات لتخصيص وتعزيز المحتوى الموجه إلى المستفيدين، ومن الثابت أن يتطلب التخصيص في المقام الأول الحصول على البيانات الشخصية للمستخدمين وتجميع أكبر قدر ممكن منها، وتحليل سجلات البحث "Search Logs" وسلوكيات وتوجهات المستخدمين بشكل منهجي وتطبيق آليات التعلم الآلي، وذلك للتعرف على مفضلات المستخدمين واستخلاص المحتوى الذي يساهم في تخصيص الخدمات بشكل فعال.

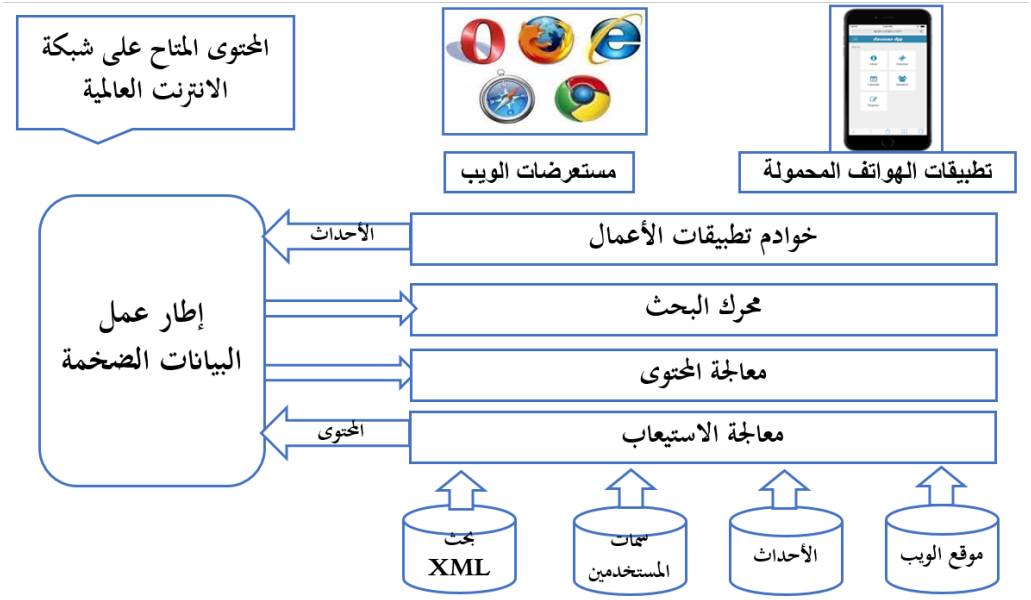
وتتسم مؤسسات ومراكز المعلومات والمستودعات الرقمية الأكاديمية للجامعات، وكذلك الجهات التي تتعامل مع المحتوى من منظور تجاري ويمثلها ناشرو وموردو خدمات قواعد المعلومات ومنهم على سبيل المثال: Proquest -Ebsco -Elsevier، بعملها المستمر على زيادة عدد زوارها والمشتركين في مصادرها وخدماتها، وقد ألقى (Nelson, 2020) الضوء على الإشكالية الرئيسية التي فرضت نفسها وتكمن في كيفية قيام هذه الجهات بالاستفادة من الكم الهائل من المصادر لتحسين البحث والتخصيص.

وتمثل أنظمة البيانات الضخمة أداة رئيسية يمكن أن تعتمد عليها مؤسسات المعلومات لتعظيم فرص الاستفادة والاستثمار في مصادرها وتعزيز خدماتها، حيث يمكن أن تحقق هذه المؤسسات نجاحات في تخصيص المحتوى والخدمات باستخدام البيانات الضخمة وذلك من خلال:

- تجميع البيانات الأولية حول المستخدم والمحتوى ويمكن أن يكون مصدر هذه البيانات الملفات الشخصية للمستفيدين، وطلبات البحث، والملفات التي يتم تنزيلها، والكلمات الدالة والمفتاحية المستخدمة من قبل المستفيدين، والبيانات الوصفية وإعداد التصنيف Taxonomy وغيرها.
- معالجة وتحليل بيانات سجل المستخدم والمحتوى داخل نظام هادوب.
- تغذية محركات البحث بالمحتوى وتوصيف البيانات وغيرها من الآليات التي تمكنه من تقديم النتائج ذات الصلة والتوصيات الفريدة عبر واجهة المتصفح سواء أكان موقع ويب أو تطبيق هاتف ذكي وإلى غير ذلك.
- من خلال ميكنة التعلم الآلي "Automated Machine Learning" يمكن إجراء تحليلات متقاطعة قوية Cross analytics بين عادات تصفح المستخدمين (البيانات غير المهيكلة) والمحتوى (البيانات المهيكلة) لتقديم تجربة بحث وتصفح مخصصة.
- البيانات الشخصية مثل: الموقع الحالي والمؤهل الدراسي والتخصص العام والدقيق والعمر والجنس وتاريخ الاتصال الأولي وما إلى ذلك من كافة البيانات المتاحة والتي يسمح بجمعها، وإذا كان المستخدم يعمل في مؤسسة المعلومات، فيمكن أن يتضمن ذلك أيضًا بيانات الموظفين مثل: سنوات الخبرة، والإدارة والوحدة والقسم الإداري التابع له، والمسمى الوظيفي، وما إلى ذلك.
- تسجيل النشاطات التي يقوم بها المستخدم على بوابة مؤسسة المعلومات مثل: الصفحات التي تم عرضها، والأقسام التي تم زيارتها واستخدام المحتوى والخدمات بها.
- النشاطات على مواقع وصفحات الويب الأخرى مثل: النقرات التي يقوم بها المستخدم على الإعلانات وروابط الإحالة وما إلى ذلك.
- معرفة عضوية المستخدم في مجموعات الاهتمامات الموضوعية إن وجدت، والتعرف على اهتمامات ونشاطات هذه المجموعات على سبيل المثال: إذا كان المستخدم معروفًا بأنه عضو في مؤسسة أو جمعية أو مجموعة مكتبية.

- النشاطات التي يقوم بها المستخدم في مصادر خارجية مثل: شبكات التواصل الاجتماعي وغيرها.
- نشاطات تتم من خلال أنظمة وتطبيقات أخرى.
- نشاطات عمليات الشراء.

وتقترح الدراسة من خلال الشكل رقم (12) البنية الهيكلية العامة التي يمكن من خلالها توظيف البيانات الضخمة لأغراض تخصيص محتوى وخدمات مؤسسات المعلومات، وتتألف هذه البنية من بوابة مؤسسة المعلومات على الشبكة العالمية، وقاعدة بيانات موزعة تتضمن الأحداث والفعاليات، وقاعدة بيانات موزعة بسمات واهتمامات المستخدمين والمستفيدين من مصادر وخدمات مؤسسة المعلومات، وقاعدة بيانات البحث والاسترجاع والتي تتوافق مع معيار "XML".



شكل (12) بنية البيانات الضخمة لتخصيص المحتوى والخدمات

وترتبط المكونات السابقة بمجموعة من الطبقات والتي تتمثل في طبقة استيعاب وإدخال البيانات وتجميعها وتعمل على معالجة هذه المدخلات "Ingestion processing"، ثم تأتي طبقة معالجة المحتوى "Content processing"، يليها الطبقة المسئولة عن إجراء البحث ويمثلها محرك أو أداة البحث "Search engine"، وأخيرًا الطبقة الخاصة بخوادم تطبيقات الأعمال " Business application servers"، وهي الطبقة التي تتعامل بشكل مباشر مع طلبات المستخدمين سواء من خلال تطبيقات الهواتف المحمولة و/ أو مستعرضات الويب المتعارف عليها للاستفادة من المحتوى والخدمات بشكل مخصص.

وتتواصل هذه الطبقات بشكل مباشر مع إطار عمل البيانات الضخمة والتي تستقبل المدخلات المتمثلة في البيانات غير المهيكلة والمتعلقة بالأحداث والمحتوى، وبعد إدارة ومعالجة مثل هذه البيانات يتم إرسالها إلى كل من طبقة معالجة المحتوى ومحرك البحث، وجدير بالذكر إطار البيانات الضخمة يكون على اتصال بالمحتوى والمصادر التي يمكن الحصول عليها عبر الشبكة العالمية.

3/2/3. البيانات الضخمة وتحسين تجربة المستخدم :

مع انتشار استخدام تقنيات الكتب الرقمية وأجهزة القراءة الإلكترونية وتطبيقات الهواتف الذكية، عملت العديد من مؤسسات المعلومات على تبني آليات لدعم تجارب المستخدمين "UE User Experiences" منها بشكل أفضل، ومن الآليات التي تم اعتمادها: توفير أنماط الوصول السهل لتصفح المصادر والاستفادة من الخدمات وذلك بهدف جذب أكبر عدد ممكن من المستخدمين والمشتركين، والمحافظة على قاعدة بياناتها من المستخدمين والمشتركين بخدماتها، لأن ذلك يمثل معيارا مهما لتعزيز إيراداتها ودعم اقتصاداتها والمحافظة على استمرارية تقديم الخدمات وتطويرها.

ومن الآليات التي يمكن أن تعتمد عليها مؤسسات المعلومات في إثراء تجربة المستخدم في البحث والاستعلام والحصول على رضا المستفيدين، تأتي آلية العمل على توفير مقاييس صارمة "Metrics" لتقييم العلاقة بين أداء البحث والعائد كلما تم جمع المزيد من سجلات المستخدمين "User Logs" والمحتوى.

ويمثل تسجيل نتائج البحث عملية تقنية وقابلة للقياس لتحسين الأداء لأداة البحث المستخدمة باستمرار، وذلك على الرغم من تطلب ذلك الكثير من التعديلات المتكررة والجهود البرمجية المضنية، ويمكن كذلك القيام بكافة إجراءات الاختبار في وضع عدم الاتصال "Testing offline"، وبالتالي ضمان بأن كافة الأمور تسير بشكل صحيح قبل الدخول في النقل والإتاحة المباشرة، وقد قلل ذلك بشكل كبير من اضطرابات الأعمال والتخطيط المجهد للفرق الفنية كما زاد من رضا المستخدمين النهائيين من أداء تقنيات البحث المستخدمة في مؤسسات المعلومات.

وقد أكد (Taylor, 2010) قدرة البيانات الضخمة على استخراج معلومات حول مجموعات مصادر المكتبات وغيرها من مؤسسات المعلومات، ومن ناحية أخرى إمكانية تتبع وتسجيل نشاطات مستخدمي المصادر والخدمات وتخزين تلك البيانات في مستودع البيانات على نطاق واسع ثم إجراء تحليل البيانات، ويمكن بعد ذلك استخدام نتائج التحليل لأغراض تحسين تجربة المستخدم وتحقيق الرضا من خدمات المعلومات المقدمة، ويعمل تحسين وتطوير تجربة المستخدم Optimize User Experience على تعزيز العائد المادي من الخدمات المقدمة.

رابعاً: النتائج والتوصيات:

توصلت الدراسة إلى مجموعة من النتائج والتوصيات ويمكن حصر أبرز هذه النتائج في النقاط الرئيسية التالية:

- اتسمت البيانات الضخمة بأنها تنشىء مصادر متعددة وتظهر بتنسيقات وأشكال متباينة، وتفتقر إلى التنظيم والهيكلة، وتتطلب أنماطًا جديدة من المعالجة، وتتمتع بالعديد من الخصائص التي دفعت مؤسسات ومراكز المعلومات المتطورة إلى التوجه نحو توظيفها لتطوير خدماتها لدعم مستخدميها، ولتحسين دعم اتخاذ وصنع القرارات.
- برزت تحديات تواجه مؤسسات المعلومات منها ضعف القدرة على استيعاب تدفق البيانات الضخمة المستمر والسيطرة عليها، ومشاكل معالجة البيانات، وقضايا البحث والاسترجاع، بالإضافة إلى تهديدات تواجه الخصوصية والاستخدام الآمن للمعلومات والحد من الاختراقات المحتملة.
- قادت خصائص منصة هادوب وطبيعة استخداماتها وبنيتها الهيكلية ومجموعة من التقنيات والأدوات التي تشكل مكوناتها إلى كونها المنصة الرئيسية في إدارة البيانات الضخمة.
- مرت إدارة البيانات الضخمة في مؤسسات المعلومات بمراحل تجميع وتدفق وتخزين وعرض وتحليل البيانات، وذلك لأغراض تجهيزها وتوظيفها لتعزيز البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات.
- يمكن لأنظمة إدارة البيانات الضخمة ممارسة دور فعال في تخطي إشكاليات إدارة ومعالجة المحتوى النابع من عدة مستودعات رقمية، وتزويد المستخدمين بتجربة بحث ثرية، وتوفير خدمات بحثية تقود لنتائج ترتبط بتلبية الاحتياجات المعلوماتية للمستخدمين.
- يمكن أن تمارس تقنيات البيانات الضخمة دورا مهما في مساعدة مؤسسات المعلومات في تخصيص وتعزيز المحتوى الموجه إلى المستفيدين.
- اقترحت الدراسة نموذجا يمكن من خلاله توظيف البيانات الضخمة لأغراض تخصيص محتوى وخدمات مؤسسات المعلومات، ويتألف من بوابة مؤسسة المعلومات على الشبكة العالمية، وقواعد بيانات موزعة تتضمن الأحداث والفعاليات، وسمات واهتمامات المستخدمين والمستفيدين من مصادر وخدمات مؤسسة المعلومات، والبحث والاسترجاع.
- تتوافر للبيانات الضخمة مقومات استخراج معلومات حول مجموعات مصادر مؤسسات المعلومات، وتتبع وتسجيل نشاطات مستخدمي المصادر والخدمات، وتخزين تلك البيانات في مستودع البيانات على نطاق واسع، ثم إجراء تحليل البيانات، واستخدام نتائج التحليل لأغراض تحسين تجربة المستخدم وتحقيق الرضا من خدمات المعلومات المقدمة.

ومن هذا المنطلق توصي هذه الدراسة بالآتي:

- قيام مؤسسات المعلومات بتقييم بنيتها التحتية التقنية والتأكد من توافقها مع متطلبات توظيف البيانات الضخمة.
- قيام مؤسسات المعلومات بالعمل على تبني آليات متطورة لتحليل البيانات التي تمتلكها، واستثمارها لتعزيز تحقيق رؤيتها ورسالتها وأهدافها الإستراتيجية.
- تعظيم مؤسسات المعلومات لمصادر وأدوات الحصول على البيانات الضخمة ومنها: صفحاتها على شبكات التواصل الاجتماعي.
- تبني آليات تستهدف تشجيع وتوجيه الطلاب والباحثين نحو إجراء مزيد من الدراسات حول البيانات الضخمة وتقنياتها وأنظمتها وسبل توظيفها في مؤسسات المعلومات وغيرها من القطاعات ذات الصلة.
- تنفيذ ورش العمل والبرامج التدريبية من جانب مؤسسات المعلومات لتعريف منسوبيها بأهمية وخصائص البيانات الضخمة ومحاور توظفيها لتحقيق الأهداف الإستراتيجية لهذه المؤسسات.
- تخصيص مستودع بيانات سحابية خاص بمؤسسة المعلومات يتم تشفيره وتخزين البيانات فيه، وتبني إجراءات صارمة للمحافظة على الخصوصية وأمن البيانات والمعلومات.

خامساً: المراجع:

الأكلبي، على بن ذيب، (2017)، تحويل البيانات الضخمة إلى قيمة مضافة. مجلة مكتبة الملك فهد الوطنية. 23 (2).

الحاتمية، أسماء بنت سعيد بن راشد& المعمرية، وهالة بنت خميس بن حمود& الحراصي، نبهان حارث، (2018)، البيانات الإحصائية والبيانات الضخمة: واقع إنتاج واستثمار البيانات الإحصائية في المركز الوطني للإحصاء والمعلومات بسلطنة عمان، في المؤتمر الرابع والعشرون لجمعية المكتبات المتخصصة فرع الخليج العربي: البيانات الضخمة وآفاق استثمارها: الطريق نحو التكامل المعرفي، سلطنة عمان.

سرحان، عماد عمر، (2016)، البيانات الضخمة وإنترنت الأشياء، سلسلة المشورة الرقمية تعلم، استرجعت من **https://www.slideshare.net/EmadOmarSarhan/ss-88037740**

عبدالله، خالد عتيق سعيد& الهنائي، عبدالله بن سالم، (2018)، البيانات الضخمة في مكتبات جامعة السلطان قابوس: واقعها ومستوى الاستفادة منها من وجهة نظر موظفيها، في المؤتمر الرابع والعشرون لجمعية المكتبات المتخصصة فرع الخليج العربي: البيانات الضخمة وآفاق استثمارها: الطريق نحو التكامل المعرفي (ص ص.1-29)، سلطنة عمان.

العميري، منال حمدان سعيد، (2018)، البيانات الضخمة في المكتبات الأكاديمية في سلطنة عمان: الواقع والتحديات، في المؤتمر الرابع والعشرون لجمعية المكتبات المتخصصة فرع الخليج العربي: البيانات الضخمة وآفاق استثمارها: الطريق نحو التكامل المعرفي، سلطنة عمان.

فرج، أحمد فرج، (2013)، أنظمة البحث التجميعي: المفاهيم والبناء الهيكلي وآليات التقويم في مؤسسات المعلومات الأكاديمية، مجلة اعلم، 9-41.

Akoka, J.& Comyn-Wattiau, I.& Laoufi, N. (2017). Research on Big Data - A systematic mapping study. *Comput. Stand. Interfaces, 54*, 105-115.

Al-Barashdi H.& Al-Karousi R. (2018). Big Data in academic libraries: literature review and future research directions. *Journal of Information Studies and Technology*, 2(13), Retrieved October 01, 2020, from https://doi.org/10.5339/jist.2018.13

Andrea, D.M.& Marco, G.& Michele, G. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65 (3), 122-135.

Gantz, J.& Reinsel, D. (2011). *Extracting Value from Chaos. IDC's Digital Universe Study*, Retrieved September 27, 2020, from http://www.sciepub.com/reference/140415

Golub, K.& Hansson, J. (2017). Big Data in Library and Information Science: A Brief Overview of Some Important Problem Areas. *Journal of universal computer science*, 23(11), 1098-1108.

Firican, G. (2017). *The* 10 *Vs of Big Data*. Retrieved July 02, 2020, from https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group Research* Note, 6 (70).

Lomotey, R. K.& Deters, R. (2014). *Towards knowledge discovery in Big Data. In the* 8*th international symposium on service-oriented system engineering.* (pp. 181–191). Retrieved from IEEE Computer Society.

Nelson, P. (2020). *How Big Data Helps Online Publishers Boost Revenue and Retention*. Retrieved from https://www.searchtechnologies.com/blog/big-data-online-publishing

Rodrguez-Mazahua, L.& Rodrguez-Enrquez, C.A.& Snchez-Cervantes, J.L.& Cervantes, J.& Garca-Alcaraz, J.L.& Alor-Hernndez, G. (2016). A general perspective of Big Data: Applications, tools, challenges and trends. *The Journal of Supercomputing*, 72 (8), 3073–3113.

RUAN, J.& WANG, S. (2016). Study on Innovation of Smart Library Service Model in the Era of Big Data. Advances in Computer Science Research, volume 50. 4*th International Conference on Electrical & Electronics Engineering and Computer Science* (*ICEEECS* 2016)

Sultana, A. (2015). *Using Hadoop to Support Big Data Analysis*: *Design and Performance Characteristics. Culminating Projects in Information Assurance*, 27, Retrieved October 01, 2020, from https://repository.stcloudstate.edu/msia_etds/27

Taylor, R.C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. In: *proceedings of the* 11*th Annual Bioinformatics Open Source Conference* (*BOSC*). Retrieved from https://doi.org/10.1186/1471-2105-11-S12-S1

Teets, M.& Goldner, M. (2013). Libraries' Role in Curating and Exposing Big Data. *Future Internet*, 5 (3), 429-438.

Wang, C.& Xu, S.& Chen, L.& Chen, X. (2016). *Exposing library data with big data technology*: *A review. In IEEE/ACIS* 15*th International Conference on Computer and Information Science* (*ICIS*). (pp. 1-6). Okayama. Retrieved from doi: 10.1109/ICIS.2016.7550937

Zhan, M.& Widén, G. (2017). Understanding big data in librarianship. *Journal of Librarianship and Information Science*, 51 (2), 561-576.

●●●

Big Data Investment to Develop Search

and Retrieval Mechanisms and Customize information Institutions Services: an Exploratory Study

Dr. Ahmed Farag Ahmed
Assistant Professor -Department of Library, Archives and Information Technology - Faculty of Arts - Asut University

ahmed.farag@aun.edu.eg

The study aimed to identify the motives for big data adoption for the analysis and management of the information institutions activities. And disclosure of the problems of employing big data in information institutions and ways to overcome them, and the mechanisms of managing this data using the "Hadoop" platform, with the definition of the importance of this platform, its characteristics and its structural structure.

The study revealed that the management of big data in information institutions has passed through the stages of collecting, storing, displaying and analyzing data, for the purpose of processing and employing it to enhance search, retrieval and customize services of information institutions. As well as the ability of big data management systems to play an effective role in overcoming the problems of managing and processing content stemming from several digital repositories, providing users with a rich research experience, and providing research services that lead to results related to meeting the information needs of users. The study recommends the need for Arab information institutions to work on adopting advanced mechanisms to analyze the data they own, and to invest them to enhance the achievement of their vision, mission and strategic goals.

The descriptive approach was used with a focus on the content analysis tool in studying the experiences of information institutions that applied methods of analyzing big data and use it in the context of their services. A survey of the literature review was conducted considering the latest findings on the subject.

Key words*: *Big data - Knowledge economies - information institutions - Internet of things - digital repositories - information services - search and retrieval - User experience