توظيف تقنيات معالجة البيانات الضخمة في بناء نموذج حقيبة الكلمات Bag of Words لتحليل مصادر المعلومات بالمكتبات الرقمية : دراسة تطبيقية باستخدام منصة Apache Hadoop (الجزء الأول)(*)

Using Big Data Platforms to Build a Bag of Words to Analyze Information Resources in Digital Libraries : An Applied Study Based on Apache Hadoop (Part I)

د. مؤمن سيد النشرتي

مدرس بقسم المكتبات والوثائق والمعلومات

كلية الآداب - جامعة القاهرة

Email: Navigator001@cu.edu.eg

ORCID: 0000-0003-4503-3378

١ - المستخلص:

تأتي هذه الدراسة بوصفها أولى الدراسات التطبيقية العربية المتخصصة في مجال المكتبات وعلوم المعلومات، والتي تركز على معالجة وتحليل البيانات الضخمة من خلال استخدام منصة Apache Hadoop، حيث هدفت الدراسة لإجراء عملية تحليل لأحد مصادر المعلومات داخل إحدى المكتبات الرقمية العربية، من خلال بناء ما يعرف بنموذج حقيبة الكلمات Bag of Words، حيث يعد هذا النموذج إحدى المراحل الأساسية في معالجة وتكشيف الوثائق من خلال تقنيات الذكاء الاصطناعي. كما تكشف الدراسة، من خلال عملية بناء نموذج حقيبة الكلمات BoW، مدى قدرة منصة Hadoop على معالجة البيانات النصية غير المهيكلة Unstructured Data، معتمدة في ذلك على المنهج الوصفي في رصد الدوافع لتطوير منصات البيانات الضخمة، وإيضاح مفهوم البيانات الضخمة The Big Data في إطارها العلمي المجرد عن السياقات التخصصية، ثم التطرق للجانب التكويني لمنصة Hadoop، والتطبيقات المساندة للمنصة ودورها في دعم تحليل ومعالجة البيانات. أما المنهج التجريبي، فقد اعتمدت الدراسة عليه في بناء نموذج حقيبة الكلمات من خلال منصة Hadoop، وتعد أبرز النتائج التي توصلت لها الدراسة هي قدرة منصة Hadoop على معالجة البيانات غير المهيكلة النصية بصورة كاملة، والتي تعكس غالبية مصادر المعلومات

(*) يُنْشَر القسم الثاني من الدراسة في العدد رقم ٣٥ (سبتمبر ٢٠٢٥).

المقتناة في المكتبات الرقمية، ونجاح المنصة في بناء نموذج حقيبة الكلمات بصورة مكتملة، ولكن تبرز صعوبة تتعلق بتعامل منصة Hadoop مع اللغة العربية في التحليل والمعالجة.

الكلمات المتاحية:

تحليل البيانات الضخمة– تقنيات البيانات الضخمة – Hadoop – نموذج حقيبة الكلمات Bag of words – المكتبات الرقمية – MapReduce– HDFS

The Abstract:

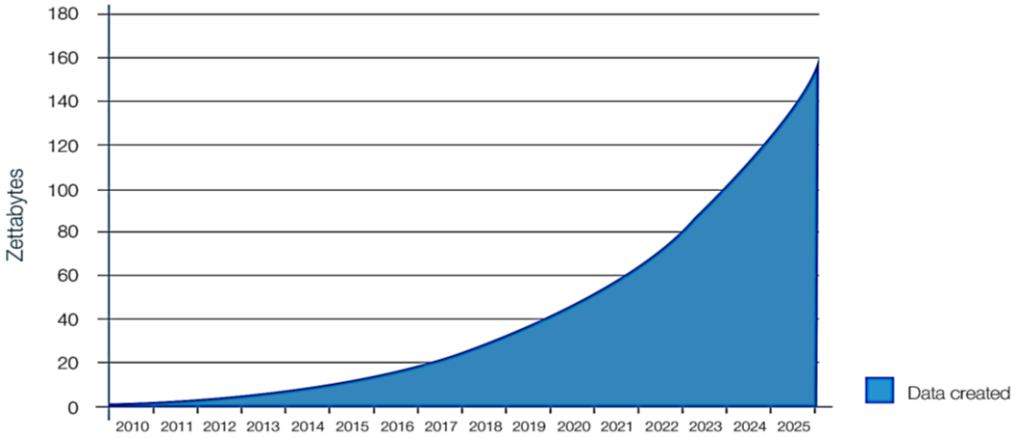
This study is considering as the first Arabic applied studies in the library and information science, where it focuses on processing and analyzing unstructured bigdata by Using the Apache Hadoop platform. The study aimed to build the Bag of Words model (the bag of Words Model is considered as the first and basic stage in AI processing and indexing documents.) to one of the information resources that is being at one of Arab digital libraries by analysis the full text of that resource. The study dependent on the Descriptive approach to discover the motivations that leaded to developing big data platforms, then studying the architecture and main components of the Hadoop platform, at the same time the study has dependent to the experimental approach, to Build the model of bag of words for the information Resource. The most prominent findings of the study are the ability of the Hadoop platform to fully process unstructured textual data, and the success of the platform in building the bag of words model in a complete manner. The study difficulty back to the analysis of Arabic content by Hadoop platform.

Keywords: The Big Data –The Big Data Techniques – The Bag of Words Model – Hadoop – The Library and Information Science – Digital Libraries.

٢- المقدمة:

أشار التقرير السنوي الصادر من وكالة البيانات الدولية في عام ٢٠١١ International Data Corporation، إلى أن حجم البيانات التي أنشئت في العالم من قبل الأفراد والهيئات قد بلغ نحو ١٠٢١ تريليون إكسابايت من البيانات، وأن حجم هذه البيانات قد تضاعف بنحو ٩ أضعاف هذا الرقم خلال الخمس سنوات المنصرفة لهذا التاريخ، لتشكل هذه الأرقام ناقوس خطر ينبئ بظهور ظاهرة البيانات الضخمة (Gantz, 2011).

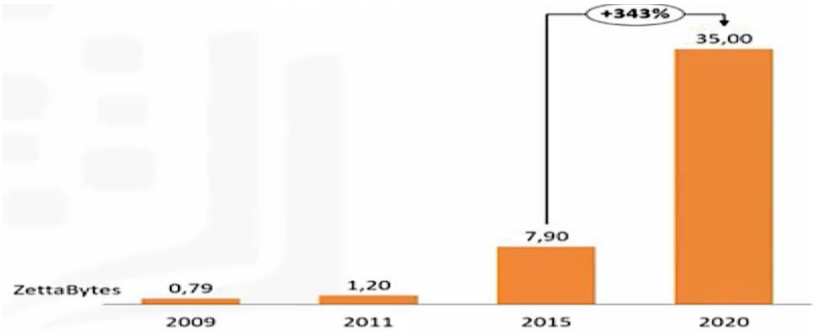
لتردف الوكالة هذا التقرير بنظير له عام ٢٠١٧ ليرصد فيه توقعًا بحجم ما يُولَّد من البيانات ومعدلات تطورها وتصاعد معدلاتها ليصل توقعاتها إلى أن بحلول عام ٢٠٢٥ سيصل حجم ما يُولَّد من بيانات لنحو ١٦٠ زيتابايت (= 160,000,000,000,000جيجا بايت) (Reinsel, 2017).



شكل رقم (١) يوضح المعدلات المتوقعة لنمو وتوليد البيانات حتى عام ٢٠٢٥.

ليستكمل هذا الأمر بما رصده وأعلنه أحد التقارير الصادرة في عام ٢٠١٨ من مجلة Forbes الأمريكية، بأن متوسط معدلات ما ينتجه العالم من بيانات في اليوم الواحد من قبل الأفراد والهيئات قد تجاوز نحو ٢,٥ كوينتيليون بايت ذات التنوع والتباين في فئاتها وأشكالها (أي ما يعادل ١٥ ضعف ما تمتلكه جوجل من البيانات) (Marr, 2018).

ويوضح تقرير (Mckinsey, 2022) المراقب لتطور حجم البيانات، أن في كل ثانية يتولد نحو ١,٧ ميجابايت من البيانات، وأن حجم البيانات منذ عام ٢٠١٥ لعام ٢٠٢٠ قد تضاعف بنسبة ٣٤٣٪ شكل رقم (٢).



شكل رقم (٢) يوضح تضاخم حجم البيانات (Mckinsey, 2022).

ليس هذا فحسب، بل إن ضخامة البيانات أصبحت لا تقتصر على حجمها، بل على تنوع مصادر توليدها وإنشائها، فبجانب المصادر التقليدية لإنشاء البيانات، أمسى هناك مصادر غير تقليدية تضخ أنماطًا جديدة من البيانات، ولعل من أبرز هذه المصادر الأجهزة المحمولة، وأجهزة الاستشعار عن بعد، والهوائيات والكاميرات، وشبكات الاستشعار

اللاسلكية، وهوائيات وتقنيات RFID وغيرها، وهذا ما أكده (Segaran, 2009).

أما عن سرعة تدفق البيانات، فقد قام موقع (Internet Live Stats, 2016) برصد معدلات التدفق والبث للبيانات على صعيد العالم في الثانية الواحدة علي منصات الويب المختلفة (بدءًا من شبكات التواصل الاجتماعي، ومرورًا بحجم المستخدمين، وحجم ما يُرسَل من بريد إلكتروني علي صعيد العالم... وغير ذلك) من خلال بناء عدادات لحساب حجم ما يُبث ويتدفق من هذه البيانات، وقد أشار إلى أن حجم ما يُتدَفق من بيانات على هذه المنصات يتجاوز الملايين في الثانية الواحدة.

شكل رقم (٣)

يوضح عدادات حجم ما يُولَّد من بيانات خلال الثانية الواحدة (Internet Live Stats, 2022)

ولا يقتصر الأمر على هذا الحد، بل يبرز تحد آخر اتسمت به البيانات في وقتنا الحاضر، مفاده أن نحو ٩٠٪ من حجم ما يتولد من هذه البيانات يتسم بكونه غير مهيكل "Unstructured"؛ أي لم تصمم هذه البيانات لتتوافق مع أنماط قواعد البيانات الراهنة أو أن تُضَمن بداخلها.

وعليه أمست الحاجة إلى التعبير عن هذه الظاهرة – والتي لا تتعلق بتضخم حجم البيانات المنشأة فحسب، بل بتنوعها بين المهيكل منها والمسرد، وبتنوع أشكالها بين النصي والصوري وذي الوسيط المتعدد، وبتنوع مصادرها بين الموثوق والمزيف – بصك اصطلاح لها عرف بالبيانات الضخمة "The Big Data"، ليصبغ هذا المصطلح العصر الحالي ليسمى بحقبة البيانات الضخمة Big Data Era (Marr, 2019).

٣- قائمة المصطلحات:

- **البيانات الضخمة:** "هي أنماط البيانات التي تتخطى قدرات أنظمة ومستودعات وقواعد البيانات التقليدية الراهنة على تخزينها وإدارتها وإجراء مختلف العمليات عليها بفاعلية وكفاءة" (Manyika, 2011)، (Oracle,2013).
- **تحليل البيانات الضخمة:** "هي العمليات أو الإجراءات التي تنطوي على تطبيق مجموعة من التقنيات غير التقليدية لمعالجة مجموعات البيانات الضخمة، لاستخراج مجموعة من المؤشرات مما هو متوافر من كيانات وعلاقات؛ للإفادة منها في اتخاذ القرار" (Adams,2010).
- **منهجيات تحليل البيانات الضخمة:** هي الطرق التي تكفل عمليات استخراج المؤشرات والدلائل من البيانات الضخمة (Al Nuaimi, 2015).
- **تقنيات تحليل البيانات الضخمة:** تعرف بأنها الخوارزميات التي طورت في العديد من المجالات الموضوعية المتخصصة (كمجال هندسة البرمجيات – والرياضيات التطبيقية – ومجال الاقتصاد – ومجال الإحصاء) بهدف توفير نمط من الذكاء في عمليات المعالجة والتحقق واستخراج المؤشرات من مجموعات البيانات (Sivarajah, 2017)، (Lebdaoui, 2014).
- **نموذج حقيبة الكلمات The Bag of Words:** أحد نماذج معالجة اللغة الطبيعية، والتي يُعتَمد عليها بصورة رئيسة في تجهيز النصوص Pre-Processing قبيل معالجتها، ويعتمد هذا النموذج على تحويل النصوص لعدد من المتجهات الرقمية Numerical Vectors، ويستخدم بشكل رئيس في تصنيف النصوص، واسترجاع المعلومات (Qader, 2019).

٤- أهمية الدراسة:

ليس هناك أدل على كون ظاهرة البيانات الضخمة أمست تشكل أحد أبرز القضايا الرئيسة الحديثة لمجتمع المكتبات والمعلومات، من قيام الاتحاد الدولي لجمعيات المكتبات ومؤسسات المعلومات IFLA بإنشاء مجموعة اهتمام موضوعية، خاصة بظاهرة البيانات الضخمة تحت مسمى IFLA Big Data Special Interest Group، والتي هدفت من خلالها إلى رصد تحديات البيانات الضخمة في المجتمع البليوجرافي، وأوجه استثمار تقنياتها في الإفادة مما تملكه المكتبات من رصيد للبيانات، وقد قامت مجموعة الاهتمام في هذا الصدد بتحديد عدد من القضايا التي أوصت بضرورة أخذها بعين الاعتبار من جانب

مجتمع المكتبات في تعاملها مع قضايا البيانات الضخمة، والتي يعد أبرزها:

- **قضايا تقنيات البيانات الضخمة في المكتبات**، والتي تشتمل على عدد من العناصر، أبرزها:
  - متطلبات البنية التحتية وأطر العمل، والتطبيقات المطلوبة لإدارة وتحليل البيانات الضخمة.
  - توفير تقنيات التنقيب عن البيانات، ونمذجتها، وعرضها المرئي Visualization.
  - إدراك منهجيات تحليل البيانات، وعمليات اكتشافها. (IFLA, 2023).

وعلى هذا تأتي هذه الدراسة لتبرز وتعالج إحدى القضايا التي أقرها الاتحاد الدولي للمكتبات ومؤسسات المعلومات (IFLA)، وهي **قضية تقنيات البيانات الضخمة في المكتبات**، في محاولة لإلقاء الضوء عليها للقارئ العربي المتخصص.

أما على الصعيد العام، فليس هناك ما يستدل به على أهمية قضية البيانات الضخمة مما قامت به إدارة الرئاسة الأمريكية في ٢٠١٢، بإطلاق مبادرة عرفت باسم "مبادرة البحث والتطوير في البيانات الضخمة Big Data Research and Development Initiative" مع رصد ميزانية أولية بلغت نحو ٢٠٠ مليون دولار (USA, The White House, 2012).

٥- مشكلة الدراسة:

تكمن مشكلة الدراسة بشكل رئيس في التعامل مع سمة الضخامة The Big، التي وسمت بها البيانات منذ مطلع العقد الأول من القرن الحالي، والتي عكست قضية رئيسة طرأت على عمليات معالجة البيانات، وهي قضية **التعامل مع أنماط غير تقليدية من البيانات، والتي عجزت أمامها التقنيات والأنظمة والبرمجيات الراهنة عن اكتشافها واستيعابها وحفظها ومعالجتها وتحليلها واسترجاعها والإفادة منها.**

وتتضح مشكلة الدراسة بصورة جلية فيما ساقه (Wang, 2016) في مراجعته العلمية، بأن ما تشمله المكتبات بمختلف فئاتها وأشكالها، من العديد من مصادر المعلومات ومصادر البيانات كالكتب والمقالات والتقارير في أشكالها التقليدية والإلكترونية، ذات التنوع في أنساقها، يضع المكتبات من المؤسسات التي تتعامل مع البيانات الضخمة تحت الضوء في الكشف عن قدرتها في استغلال هذه التقنيات في تقديم المعرفة والمعلومات باعتبارها قيمًا مضافة من هذه المصادر المتنوعة وبصورة تتسم بالفاعلية والكفاءة. وعلى الرغم من ذلك، فإن **غالبية المكتبات على صعيد العالم بمنأى عن توظيف تقنيات البيانات الضخمة**

بالصورة التي تمكنها من توليد المعرفة من هذه المصادر أو استخراج المؤشرات، حيث يظل ما تملكه المكتبات من تسجيلات ببليوجرافية أو مصادر معلومات معزولًا عن أن يعالج بالصورة التحليلية التي تكفل فرص تقديم خدمات أفضل، الأمر الذي يتطلب إجراء العديد من الدراسات البحثية للكشف عن فرص توظيف تقنيات تحليل البيانات الضخمة بالصورة التي تتسم بالفاعلية والكفاءة.

ومن ثم يمكن تلخيص مشكلة الدراسة، في **الكشف عن فرص توظيف تقنيات البيانات الضخمة في المكتبات لاستخراج المعرفة والقيمة المضافة لما تملكه من الرصيد المعلوماتي والمعرفي الرقمي.**

٦- أهداف الدراسة:

أوضح تقرير مؤسسة IT Pro الصادر عام ٢٠٢١ لتقييم نظم وتقنيات معالجة وتحليل البيانات الضخمة أن منصة Apache Hadoop ما زالت – منذ إطلاقها – تحتل عرش تطبيقات معالجة البيانات الضخمة، وتتصدر المركز الأول على صعيد التوظيف والاستخدام، لما تكفله هذه المنصة من بيئة عمل تعمل على احتواء مختلف التطبيقات الأخرى المناظرة لها بداخلها (Clare, 2021).

وعلى هذا يتمثل **<u>الهدف الرئيس لهذه الورقة البحثية</u>** في **دراسة ورصد منصة** Apache Hadoop Ecosystem **بالتحليل والدراسة** باعتبارها إحدى أبرز منصات معالجة وتحليل وتخفيض وهيكلة البيانات الضخمة في الوقت الراهن.

وينبثق من هذا الهدف الرئيس عدد من الأهداف الفرعية، والتي يمكن إجمالها على النحو التالي:

١. دراسة **البنية التكوينية لبرنامج** Hadoop **وأوجه القوة والضعف في بنيته.**
٢. رصد **ما قد طور في سياق هذا البرنامج من برمجيات مصاحبة له.**
٣. **إجراء دراسة تطبيقية من خلال تحليل أحد أشكال مصادر المعلومات النصية (البيانات غير المهيكلة) بإنشاء نموذج حقيبة الكلمات** Bag of Words **لهذا المصدر باستخدام منصة** Apache Hadoop.

٧- تساؤلات الدراسة:

يمكن تحديد **<u>تساؤلات الدراسة</u>** في ضوء ما سبق من أهداف على النحو التالي:

- ما طبيعة منهجيات التحليل التي يمكن أن تعتمد عليها تقنيات معالجة وتحليل البيانات

الضخمة؟

- كيف يمكن لنظام Hadoop Ecosystem من خلال بنيته التكوينية البرمجية أن يقوم بمعالجة وتحليل مجموعات البيانات الضخمة وغير المهيكلة؟
- ما جوانب النقص في نظام Hadoop، التي أدت إلى تطوير مجموعة من البرمجيات الملحقة به لتشكل ما يعرف بنظام Hadoop Ecosystem؟
- كيف يمكن إجراء تحليل للبيانات النصية لمصادر المعلومات من خلال منصة Hadoop؟

٨- مجال الدراسة وحدودها:

- **الحدود الموضوعية:** على الصعيد النظري تهتم الدراسة بالتأصيل لمفهوم ظاهرة البيانات الضخمة في مجال المكتبات والمعلومات. أما على الصعيد التطبيقي، فتقوم باختيار منصة Hadoop لتحليل البيانات الضخمة وإجراء عملية تحليل لمجموعة من البيانات الضخمة غير المهيكلة من خلالها.
- **الحدود النوعية:** تتمثل في إبراز الدور الذي تلعبه منصة Hadoop في معالجة تحليل البيانات غير المهيكلة، من خلال بناء نموذج حقيبة الكلمات Bag of Words لأحد مصادر المعلومات النصية بداخل إحدى المكتبات الرقمية العربية.
- **الحدود اللغوية:** تتمثل في التعامل مع الإنتاج الفكري الصادر باللغة العربية والإنجليزية حول ظاهرة الدراسة، وما صدر بخلاف هاتين اللغتين تقوم الدراسة بالتعامل معه على صعيد مستخلصه العلمي.
- **الحدود الزمنية:** تتمثل في رصد محاولات التأصيل النظري لظاهرة البيانات الضخمة، منذ صدور أول عمل علمي محكم تناول هذه الظاهرة عام ١٩٩٦، حتى وقت الانتهاء من الدراسة.

٩- منهج الدراسة وأدواتها:

تعتمد الدراسة على استخدام منهجين أساسيين لتحقيق أهدافها، هما:

١. **المنهج الوصفي التحليلي**، وذلك للأغراض التالية:

- استعراض الأسانيد الأدبية والنظرية للتأصيل لظاهرة البيانات الضخمة في مجال المكتبات وعلوم المعلومات.
- تحليل البنية المعمارية والتكوينية لتطبيق Apache Hadoop، ورصد أبرز مظاهر القوة والضعف فيه.

٢. **المنهج التجريبي**، القائم على إنشاء نموذج حقيبة الكلمات Bag of Words لأحد مصادر المعلومات؛ للكشف عن قدرة منصة Hadoop في تحليل البيانات الضخمة النصية غير المهيكلة.

أما فيما يتعلق بأدوات الدراسة، فسيُعتَمد إن شاء الله تعالى بشكل رئيس لتحقيق أهداف الدراسة على الأدوات الآتية:

- مراجعة الإنتاج الفكري الراجع والجاري لحصر أبعاد ظاهرة البيانات الضخمة.
- مجموعة من الجلسات العلمية المباشرة لاستخدام منصة Hadoop في تحليل البيانات غير المهيكلة، واختبار المنصة وتقييمها، وإنشاء نموذج حقيبة الكلمات Bag of Words لأحد مصادر المعلومات.

١٠- ظاهرة البيانات الضخمة: مراجعة علمية:

نأت هذه المراجعة العلمية بنفسها عن ذاتية اختيار مفردات الإنتاج الفكري من دراسات وأبحاث لتقوم باستعراضها بوصفها دراسات سابقة عليها، (ذاك الاختيار الذي قد يتسم بصورة أو بأخرى بالعديد من مظاهر التحيز وعدم الموضوعية من جانب الباحث في اختيار أبحاث دون غيرها لاستعراضها باعتبارها حلقات في مسيرة تطور الموضوع)، ولذلك فقد عمدت الدراسة للتوجه نحو المراجعات العلمية والمحكمة التي قد صدرت حول مختلف جوانب ظاهرة البيانات الضخمة، عوضًا عن الدراسات والأبحاث في صورتها المنفردة والموزعة حول هذا الموضوع، فضلًا عن ما تستند إليه هذه المراجعات العلمية المحكمة من معايير اختيار وتقييم لما تقوم باستعراضه من أبحاث، جاءت العديد من المراجعات العلمية التي رصدت ظاهرة البيانات الضخمة سواء على صعيد التقنيات أو أنماط التحليل، وقد اتسمت هذه المراجعات بالتجريد في تناولها لقضايا البيانات الضخمة (Ali, 2016)، (Chen, 2014)، (Emani, 2015)، (Fan, 2014) Oussous, 2018))، (Wamba, 2015)، (Kaisler, 2013)، (Fahad, 2014)، حيث بلغ الحد الأدنى لحجم الاستشهادات المرجعية لهذه المراجعات نحو ١٠٠٠ عمل علمي.

ومن ثم فقد اشتملت هذه المراجعة العلمية، على استعراض أبرز إسهامات الدراسات الواردة في المراجعات العلمية (لتكون بمثابة مراجعة علمية للمراجعات العلمية الواردة حول هذا الموضوع)، الأمر الذي كفل العديد من المزايا (التي قد لا يكفلها استعراض الأبحاث في صورة منفردة)، والتي تتمثل في:

١. تقديم صورة تتسم بالشمول عن ظاهرة الدراسة (البيانات الضخمة).

٢. تتبع التطور الزمني لظاهرة البيانات الضخمة.
٣. الإحاطة بالقطاعات الموضوعية والتوجهات البحثية لهذه الظاهرة.
٤. تحديد الفجوات البحثية النوعية التي لم تُغطَ في الأبحاث العلمية السابقة لهذه الدراسة، وإبراز الإسهام العلمي الذي ستقدمه هذه الدراسة.
٥. توفير مبررات إعداد هذه الدراسة وفقًا لما ارتضته من أهداف وتساؤلات وحدود تتميز بها عن نظائرها من الأبحاث.

ومن ثم فإن استعراض الدراسات السابقة في هذه الدراسة سيتأتي إن شاء الله تعالى من خلال تحديد القطاعات والأقسام التكوينية، التي اشتملت عليها العديد من المراجعات العلمية التي تناولت قضية البيانات الضخمة، والتي ستعكس الدراسات السابقة لهذه الدراسة على النحو التالي:

١. طبيعة البيانات في صورتها المجردة.
٢. إرهاصات البزوغ لظاهرة البيانات الضخمة.
٣. منهجيات تحليل البيانات الضخمة.
٤. تقنيات تحليل البيانات الضخمة.
٥. تطبيقات تحليل البيانات الضخمة.
٦. البيانات الضخمة في مجال المكتبات والمعلومات.
٧. الإسهام العلمي لهذه الدراسة.

١- طبيعة البيانات في صورتها المجردة:

قبيل العقد الأول من القرن الحالي، كانت المصادر الرئيسة لتوليد وإنتاج البيانات تتمثل فيما يقوم البشر بإنتاجه، وذلك من خلال أنشطة التأليف، أو البحث، أو الإبداع، أو التوثيق، أو التطوير، أو الكشف، أو ما يقوم به البشر من إجراءات فنية ووظيفية وعملية (كإنتاج البيانات الببليوجرافية من واقع إجراءات ونشاط الفهرسة)، وذلك خلال سياقات إنتاج البيانات المتنوعة والمتباينة بين السياق الأكاديمي، أو التجاري، أو الاقتصادي، أو الترفيهي.

وفي سياق هذه الأنماط من البيانات المولدة من جانب البشر، طورت العديد من البرمجيات، والنظم، والتقنيات، لأغراض اكتشافها ومعالجتها وحفظها ثم استرجاعها، وذلك في ظل قدرة هذه الأنظمة على استيعاب ما ينتج إليها من بيانات، ولتتخذ البيانات في هذا السياق نمطين رئيسين، هما:

- نمط البيانات المهيكلة Structured Data.

- نمط البيانات شبه مهيكلة Semi Structured.

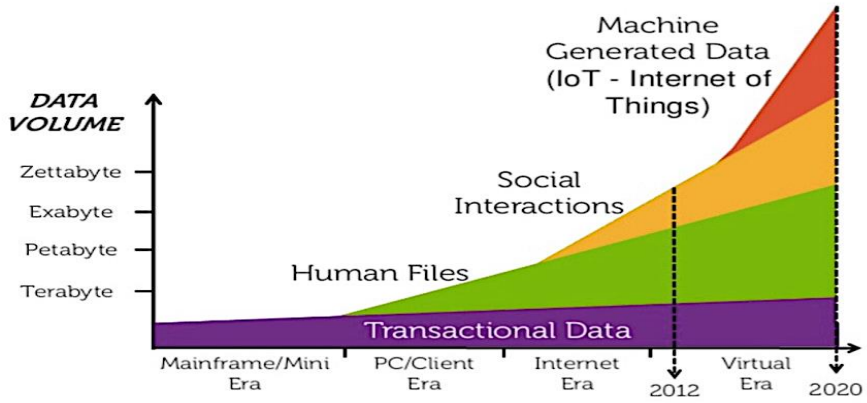
يقصد بالبيانات المهيكلة، بالبيانات التي يمكن أن تخضع للمعالجة والتحليل وفقًا لنماذج بيانات مسبقة، كإخضاع البيانات النصية داخل قواعد البيانات وفقًا لنموذج البيانات العلائقي Relational database model.

حيث يكفل نموذج البيانات القدرة على أن تكون البيانات قابلة للإدراك من جانب الآلة والبشر على السواء في حفظها ومعالجتها واسترجاعها، (كنموذج MARC ونموذج BIBFRAME للبيانات الببليوجرافية، والذي يكفل للحاسب القدرة على فهم البيانات واسترجاعها).

أما البيانات شبه المهيكلة فهي البيانات التي تُنشَأ وفقًا لأحد أنماط التمثيل المخصصة كنمط لغة XML، الذي يعمل على هيكلة البيانات النصية من خلال التعريف بأنواع البيانات وطرق تمثيلها، وليتخذ مصممو التقنيات والنظم والبرمجيات من هذين النمطين محورًا لهم في تصميماتهم وإنشائهم وتطويرهم لهذه البرمجيات، حتى تتمكن من استيعاب ما يرد لها من هذين النمطين.

ولكن مع بزوغ القرن الواحد والعشرين، وبالتحديد في العقد الأول منه، أصبحت عملية إنتاج البيانات لا تقتصر فقط على البشر، بل شاركتهم في ذلك الآلات ، لتصبح بذلك عملية إنتاج البيانات معتمدة على مصدرين، هما:

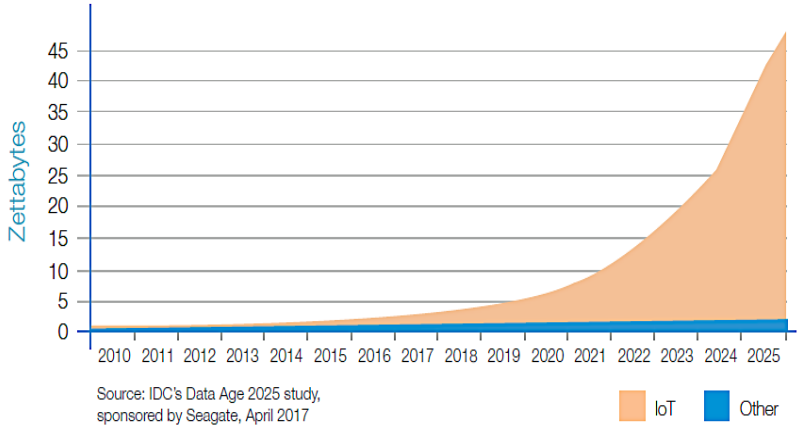
- المصدر التقليدي؛ ويتمثل فيما ينتجه البشر من بيانات.
- المصدر غير التقليدي؛ ويتمثل في الآلة وما تنتجه من بيانات.



شكل رقم (٤) يوضح تفوق معدلات نمو البيانات المولدة من جانب الآلة عن معدلات البيانات المولدة من جانب البشر (Ucros, 2017).

اتسمت عملية إنتاج الآلة للبيانات (بوصفها مصدرًا غير تقليدي) بالغزارة الشديدة،

وبمعدلات إنتاج تفوق ما قام البشر بإنتاجه منذ أن اخترعوا الكتابة، ليس هذا فحسب، بل إن عملية إنتاج البيانات من قبل الآلة أسفر عن وجود نمط جديد من أنماط البيانات، وهو النمط غير المهيكل للبيانات The Unstructured Data.



شكل رقم (٥) يوضح ارتفاع معدلات حجم البيانات المولدة من جانب تقنيات إنترنت الأشياء (الآلة) في مقابل مصادر البيانات الأخرى (Woodie, 2017).

يقصد بنمط البيانات غير المهيكل The Unstructured Data بأنه النمط الذي لا يستند على نموذج محدد للبيانات Data Model، في إنشائه أو معالجته أو آلية حفظه واسترجاعه، كالصور، وملفات الفيديو، والبريد الإلكتروني، وغيرها.

ومن ثم فإن افتقار البيانات لنموذج البيانات Data Model، قد فرض على عمليات معالجة البيانات، العديد من المتطلبات والموارد المسبقة بغية أن تسهم هذه المتطلبات في تجهيزها للمعالجة.

وعلى هذا، فقد أسهمت كل من الأحجام المطردة الناتجة من عمليات توليد وإنتاج الآلة للبيانات، والأنماط غير المهيكلة من البيانات، على بزوغ ما يعرف بظاهرة البيانات الضخمة، ولتقف مختلف التقنيات والبرمجيات والنظم الراهنة عاجزة بشكل كامل عن معالجة هذا النمط أو الحجم من البيانات، وليشكل هذا الأمر بداية حقبة البيانات الضخمة، أي عصر البحث والتطوير والاستبدال لما طُوِّر من تقنيات ونظم لديها القدرة على التعامل مع البيانات الناتجة من المصادر التقليدية، بنظم وبرمجيات وتقنيات بأخرى لديها القدرة على أن تتعامل مع ما تنتجه الآلة – بوصفها مصدرًا غير تقليدي – من بيانات غير مهيكلة.

٢- إرهاصات البزوغ لظاهرة البيانات الضخمة:

اتضح مما عُرِضَ أن ما يُضَخ من بيانات من خلال المصادر التقليدية وغير التقليدية لتوليدها، خلال الدقيقة الواحدة يتخطى قدرات نظم وتقنيات وبرمجيات ومستودعات التحليل التقليدية الراهنة في التعامل معه، لتشكل عمليات الضخ هذه، عددًا من التحديات، والتي عرفت لاحقًا بتحديات البيانات الضخمة، والتي كانت بمثابة الدافع الرئيس للعديد من الجهود نحو تطوير تقنيات لمعالجة وتحليل هذه الأنماط غير التقليدية من البيانات الضخمة.

يرصد (Bahambhri, 2012) تطور ظاهرة البيانات الضخمة من خلال التتبع التاريخي لأثر التقنيات في معالجة البيانات، فيرصد خمس محطات رئيسة لهذا التطور، والتي تمثلت فيما يلي:

- **المحطة الأولى** تمثلت في تطوير أول نموذج أولي Prototype لهيكلة البيانات في صورة علائقية في معامل مركز أبحاث Almaden في مدينة San Jose بالولايات المتحدة، والذي عرف فيما بعد بقواعد البيانات العلائقية Relational Database Management Systems في حقبة السبعينيات من القرن المنصرف، وقد أعقب هذا النموذج الأولي بروز العديد من البرمجيات التي هدفت إلى إنشاء قواعد البيانات العلائقية، والتي يعد أبرزها كل من DB2, Oracle, SQL/DS.
- أما **المحطة الثانية**، فقد جاءت نتيجة طبيعية لزيادة أعداد قواعد البيانات وتنوعها على صعيد بنيتها وعلى صعيد محتواها، فجاء ما يعرف بمستودعات قواعد البيانات Data Warehousing، والتي هدف منها إلى تسهيل وإدارة وتحليل البيانات القادمة من قواعد بيانات متنوعة استجابة لحاجة أسواق العمل ومتخذي القرار، وكان ذلك في حقبة التسعينيات من القرن الماضي.
- أما **المحطة الثالثة**، فقد رُصِدَت في مطلع القرن الحالي، وعرفت بحقبة لغات التكويد للبيانات المهيكلة على الويب، وبرزت من خلال استخدام لغة النص الفائق الموسع Extensible Markup Language XML، تلك اللغة التي عملت على أن تنشأ البيانات النصية المجردة Plain Text في صورة شبه مهيكلة، في صورة أقرب لنماذج Models قواعد البيانات، ودون الحاجة لأن تسكن في قواعد بيانات، وليجرى عليها عمليات التخزين والمعالجة والهيكلة والإدارة بشكل مستقل، وليصاحب هذا النمط من البيانات فئة جديدة من البرامج والتطبيقات، هدفت لإنشاء هذا النمط من البيانات، وعرفت هذه التطبيقات باسم نظم إدارة المحتوى Content

Management System، ويجب الإشارة إلى أن هذا التطور لم يقض على المرحلة السابقة من الاعتماد على قواعد البيانات، بل تماشى كلاهما معًا معتمدين على منصة عمل كفلت الحرية في الاختيار والإنشاء وهي الويب.

- أما **المحطة الرابعة**، فتمثلت في ظهور شبكات التواصل الاجتماعي، والتي أفضت هي الأخرى إلى محتوى يتسم بسمتين رئيستين:

  - آنية الوجود للمعلومات Real Time: فالمعلومات وفقًا لهذه السمة لا تخضع لمراحل النشر التقليدية من فرز وتقييم ومراجعة، بل أصبح التداخل في حلقات النشر والإتاحة للمعلومات سمة هذه المنصات، فالمستفيد القارئ هو في الوقت نفسه باحث وناشر، فضلًا عن أن هذه المنصات كفلت فرص تقديم التغذية المرتدة من خلال التفاعل والمشاركة، لتكفل قيمًا مضافة للمحتوى نفسه.
  - تقنيات الدفع للمعلومات Push Techniques: لم يعد المستفيد والباحث بحاجة إلى إجراء عمليات بحث لاحقة وراجعة للإنتاج الفكري، بل تمثلت الحاجة في أن يقوم بالوجود المعنوي على منصات الدفع ليصل إليه كل طرح جديد يقع في دائرة اهتمامه.

- أما **المحطة الخامسة**، فتمثلت في بروز ظاهرة البيانات الضخمة، وقد جاء الإسهام هذه المرة من جانب الآلة والتطبيقات التي أمست مصدرًا من مصادر توليد البيانات، بصورة تتضاعف وتتنوع في عمليات إنتاجها بشكل أسي عما كان عليه الوضع آنفًا، ليستكمل الأمر ببروز تقنيات إنترنت الأشياء IOT، لتدرك حينها كيانات الأعمال الكبرى Business Enterprises ضرورة الحاجة للبحث عن طرق ومنهجيات غير تقليدية لإدارة هذه البيانات الضخمة ذات التنوع والضخامة والتسارع في النمو، والتي تنتجها أو تملكها أو تطلبها، بصورة تكفل فعالية القدرة على اتخاذ القرار، فأدركت أن الحل الأمثل يكمن في إجراء التحليلات على هذا الكم المتنوع والمتسارع من البيانات، ليبرز العديد من المجالات المعرفية الجديدة كالتنقيب عن البيانات Data Mining، وخوارزميات تعليم الآلة Machine Learning Algorithms، ولتنتهي هذه المحطة فيما يعرف بمجال تحليل البيانات الضخمة Big Data Analysis، ليكون هدفه الرئيس لا يتجلى فحسب في إدارة هذا الكم من البيانات، بل في العمل أيضًا على استقرائها لاستنباط قيم مضافة تكفل لتلك الكيانات القدرة على البقاء والريادة والتنافس في أسواقها.

أما عن صك مصطلح البيانات الضخمة، فقد سبقه العديد من الإرهاصات، والتي جاءت في العديد من الكتابات العلمية والتقنية والاقتصادية، والتي كان من أبرزها كتابات Erik Larson لمجلة The Washington Post عام ١٩٨٩م حينما أورد تساؤل في طيات مقاله الموسوم بعنوان "The Devil in the White City" عن البريد غير المرغوب فيه The Spam Mails، الذي يصله على حساب بريده الإلكتروني، متسائلًا عن كيفية وإمكانية الإفادة من هذا الحجم من البيانات الضخمة، التي ترد من الشركات وإليها لاستخدام هذه البيانات في أغراض ذات جدوى (Brueckne, 2014).

أما الفضل في صك هذا المصطلح بوصفه ظاهرة علمية، فيعزى إلى John Mashey في منتصف التسعينيات من القرن الماضي، والذي استخدمه في سياقات علمية وندوات تعريفية، في إشارة إلى البيانات التي تنتجها الشركات العملاقة لمعالجة البيانات واستثمارها لها في الصناعات الثقافية والإبداعية، ليبرز في النهاية هذا المصطلح باعتباره عنوانًا لعمل علمي أكاديمي محكم بعنوان "Big Data and the Next Wave of Infrastress " عام ١٩٩٨(Lohr, 2013).

٣- منهجيات تحليل البيانات الضخمة:

يشير كل من (Jagadish & Labrinidis, 2012) إلى أن أحد أبرز تحديات توظيف منهجيات التحليل في معالجة البيانات، **الفجوة** بين **طبيعة البيانات** وبين **طبيعة منهجيات التحليل** المستخدمة، فملامح هذه الفجوة تتضح في كون البيانات تمثل كيانات تتسم طبيعتها بالديناميكية، بينما تتصف هذه المنهجيات بأنها قد صُمِمَت للعمل على أنماط تتسم بالثبات، ليس هذا فحسب، بل إن معالجة البيانات الضخمة لا يتعلق بالعمل على معالجة خصائصها من حيث الحجم والتنوع والسرعة، بل يجب العمل على تحليل شبكة العلاقات الكامنة بين أنماط البيانات، تلك الشبكة التي تتسم بدرجة من التعقيد بالصورة التي تجعل منها تحديًا آخر في معالجة وتحليل البيانات الضخمة.

أما فيما يتعلق بأنواع منهجيات تحليل البيانات الضخمة، فيوضح (Rehman, 2016) أنها تتنوع بين:

- **منهجيات التحليل الوصفية** Descriptive
- **منهجيات التحليل التشخيصية** Diagnostic
- **منهجيات التحليل التنبؤية** Predictive
- **منهجيات التحليل والوظيفية** Perspective ..

وتمثل **منهجية التحليل الوصفي** أولى أنماط المنهجيات التي طورت، وتعد هذه المنهجية أبسط منهجيات تحليل البيانات الضخمة، حيث تنطوي على عرض وتقديم المعلومات والبيانات التي تعكس الصورة التفصيلية للوضع القائم للعمليات والإجراءات التي تقوم بها المؤسسة، من خلال أساليب إحصائية بسيطة تعمل على تقديم مختلف المتغيرات عن حدث معين من خلال تقارير أداء معيارية بسيطة.

وتعد أكثر الأنماط استخدامًا في مختلف تطبيقات تحليل البيانات حيث توظف هذه التطبيقات هذه المنهجية لعرض نتائج ومخرجات عمليات التحليل الراجعة للبيانات (أي البيانات التي سُجلَت وحُصِل عليها وتعكس الأنشطة التي قامت بها المؤسسة)، معتمدة في ذلك على العديد من الأدوات، كتقارير الأداء Reports، ولوحات قراءة البيانات Dashboards، وبطاقات الهدف Scorecards، والنمذجة المرئية Data Visualization، والتي تعمل على تقديم صورة وصفية لما يُحَلّل من بيانات (Watson, 2014).

تأتي منهجية التحليل التشخيصية Diagnostic باعتبارها منهجية لاحقة لمنهجية التحليل الوصفية، وتعد التحليلات التشخيصية أحد الأنواع الأكثر تقدمًا لتحليلات البيانات الضخمة، التي يمكنك استخدامها للتأكد من البيانات والمحتوى والتحقق منه.

تستهدف هذه المنهجية الإجابة عن التساؤلات السببية مثل "لماذا حدث ذلك؟"، الأمر الذي يكفل إمكانية فهم أسباب بعض السلوكيات والأحداث المتعلقة بالمؤسسة أو الفرد.

ويشتمل هذا النمط من منهجيات التحليل على العديد من الأدوات والتقنيات المستخدمة كتقنيات البحث عن الأنماط Patterns في مجموعات البيانات، وتصفية البيانات، واستخدام نظرية الاحتمالات، وتحليل الانحداري Regression Analysis.

أما النوع الثالث من منهجيات التحليل الرئيسة للبيانات، فيتمثل في **منهجية التحليل التنبؤي** Predictive، والذي يُعرَّف بأنه "نمط التحليل القائم على توظيف المفاهيم والأساليب الرياضية والإحصائية؛ لإيضاح واستكشاف أنماط العلاقات بين البيانات بغرض التنبؤ والتوقع بالنتائج المستقبلية من واقع تحليل أنماط العلاقات بين مجموعات البيانات ودرجات الاعتمادية والاستقلالية، ومن واقع درجات التغيير التي تحدث في مجموعات البيانات" (Delen, 2013).

تعمل هذه المنهجية على الإجابة عن سؤالين يتعلقان بعملية تحليل البيانات، وهما: ما الذي سيحدث؟ ولماذا سيحدث؟

ويشير كل من (Gandomi & Haider, 2015) إلى أن أبرز التقنيات التي تعتمد

عليها هذه المنهجية هي تقنيات التعلم الآلي Machine Learning Techniques، والشبكات العصبية Neural Networks، فكلاهما يستخدمان في تطوير ما يعرف بمنصات اكتشاف المعرفة Knowledge Discovery Frameworks، وخلاصة الأمر أن منهجيات التحليل التنبؤية تعمل على التوقع من خلال التحليل لسجلات العمليات وأنماط التعامل خلال فترة زمنية.

أما فيما يتعلق **بمنهجية التحليل الوظيفية** Perspective، فيعد الهدف الرئيس من تطوير هذه الفئة من منهجيات التحليل، قياس الأثر المترتب على العلاقة بين نتائج التحليل الواردة من منهجيات التحليل الوصفي والتنبؤي، وما يجب على المؤسسة أن تقوم به من تحسين لسياسات إجراء العمليات.

تعمل هذه المنهجية على الإجابة عن سؤالين يتعلقان بعملية تحليل البيانات، وهما: ماذا سأفعل تجاه الحدث؟ ولماذا أفعل ذلك؟

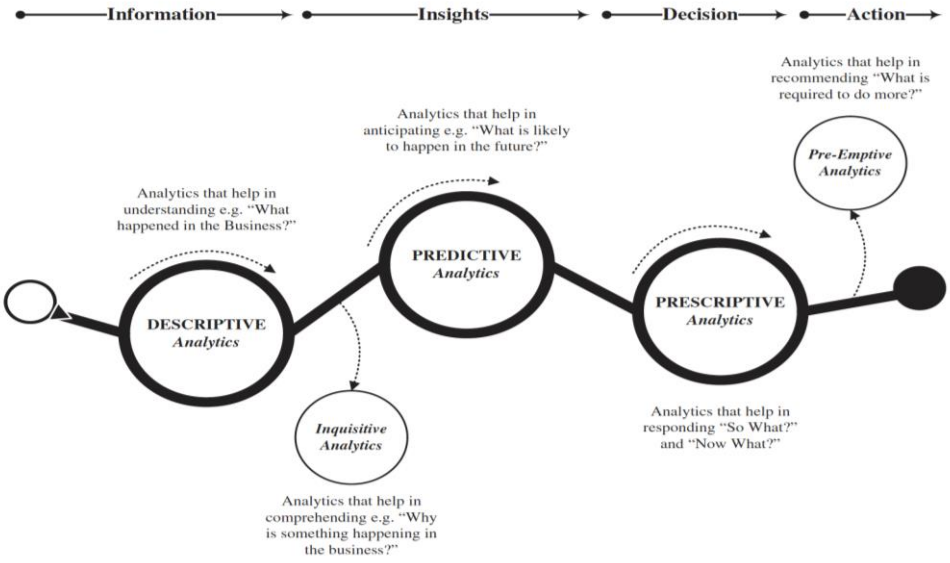
فيشير (Banerjee, 2013) وآخرون إلى أن الجانب التطبيقي لتوظيفها في عمليات تحليل البيانات يتسم بالندرة، والمرجعية في ذلك تعود إلى أن غالبية مستودعات وقواعد البيانات الراهنة تتسم بالتقيد بعدد من الأبعاد التحليلية، التي لا تسمح بأن يتولد أو يطبق من خلالها هذا النمط من المنهجيات.

يجب الإشارة إلى أن هذه المنهجيات التحليلية السابقة كافة، تستلزم أن يتوافر مجموعة من المحليين ليس لديهم فقط مهارات قراءة البيانات، بل القدرة على استنباط الحقائق من الأرقام والإحصائيات، وربط هذه الحقائق بعمليات اتخاذ القرار (Rehman, 2016).

Types of Big Data Analytics				
	Descriptive	**Diagnostic**	**Predictive**	**Prescriptive**
Answers the question...	What happened?	Why did it happen?	What will happen next?	What should I do?
Level of advancement	Low	Medium	High	Very high
Incorporates AI and machine learning?	Not usually	Sometimes	Usually	Always
Level of popularity	Used by almost all organizations	Used by many organizations	Used by a smaller but growing group of organizations	Not yet widespread

شكل رقم (٦)

يوضح أنواع منهجيات تحليل البيانات الضخمة وخصائص كل منهجية.



شكل رقم (٧)

يوضح القطاعات الرئيسة والفرعية لمنهجيات التحليل الوظيفية للبيانات (Sivarajah, 2017)

٤- تقنيات تحليل البيانات الضخمة The Big Data Techniques:

ارتكزت غالبية الجهود في مضمار معالجة تحديات البيانات الضخمة على أن مواجهة هذه التحديات لا تتأتى إلا من خلال تطوير آليات وتقنيات غير تقليدية ومبتكرة، حيث تتسم بقدرتها على إدارة هذه الأنماط غير التقليدية من البيانات، وذلك على صعيد معالجتها بالصورة التي تمكن من يتعامل معها من الخروج بمؤشرات وقيم مضافة غير مسبوقة.

وقد أسفرت هذه الجهود عن تطوير العديد من التقنيات، وقد وُظِفَت هذه التقنيات في العديد من البرمجيات Software، جاء البعض منها تجاريًّا والبعض الآخر مجانيًّا والبعض الآخر مفتوح المصدر، ويعد من أبرز هذه البرمجيات:

(Apache Hadoop [1] - Google BigTable [2] - Amazon SimpleDB [3] - WibiData [4] - MarkLogic [5] - Apache Spark [6] - Splice Machine [7] - Tableau [8] -

1 https://hadoop.apache.org
2 https://cloud.google.com/bigtable
3 https://aws.amazon.com/simpledb/
4 https://www.wibidata.com
5 https://www.marklogic.com/product/marklogic-database-overview/
6 https://spark.apache.org
7 https://splicemachine.com
8 https://www.tableau.com

MongoDB [9] Cloudera [10] - SkyTree [11] - Microsoft Azure [12]).

وفي هذا الصدد قامت العديد من الدراسات العلمية بحصر هذه البرمجيات والنظم، وتحديد تقنياتها بهدف عقد المقارنات بين مختلف هذه البرمجيات أو منصات تحليل معالجة البيانات الضخمة (Singh, 2014)، (Liu, 2014)، (Landset, 2015).

ويعد أبرز ما أفضت إليه هذه الدراسات المقارنة ما أشار إليه (Bhargava, 2019) أن أكثر التقنيات ضرورة في معالجة البيانات الضخمة، من حيث الاستخدام والكفاءة والفعالية والقدرة على التكيف مع فئات وأشكال البيانات المختلفة هي منصة Apache Hadoop، لما تكفله من سمة عظيمة تتجلى في قدرتها علي أن تعمل مع مختلف التطبيقات والأنظمة الأخرى، من خلال بناء ما يعرف بنظام Hadoop Ecosystem والذي يكفل أن يتم تضمين التقنيات الأخرى بداخله في صورة أشبه بالبنية الطبقية Layers، لتعمل معًا بصورة متجانسة لمعالجة البيانات الضخمة وفقًا للسياق المراد معالجته، كما سيرد لاحقًا.

كما أشار (Chen et all, 2014)، في دراسة مسحية، إلى أن منصة Apache Hadoop تأتي على رأس هذه التطبيقات المستخدمة في مجال معالجة البيانات الضخمة. وأشار أيضًا تقرير معهد McKinsey الصادر عام ٢٠١١ إلى أن التنوع وراء تعدد التقنيات المستخدمة في تحليل البيانات الضخمة، يعود إلى تعدد المجالات المعرفية التي قامت بتطوير هذه التقنيات وفقًا لأغراض محددة لها، ولعل أبرز هذه التخصصات التي كان لها السبق في تطوير هذه التقنيات، هي: مجال هندسة البرمجيات Software Engineering – والرياضيات التطبيقية Applied Mathematics – ومجال الاقتصاد Economics– ومجال الإحصاء Statistics وبحوث العمليات Operation Research (Manyika, 2011).

أشار (A. Oussous, 2018) إلى أن العديد من نماذج البيانات وأطر العمل والتقنيات الجديدة التي قدمت في صدد معالجة البيانات الضخمة، جاءت نتيجة للعديد من المشاريع المختلفة من جميع أنحاء العالم.

ويعد أبرز التقنيات والخوارزميات الموظفة في مجال تحليل البيانات الضخمة ما يلي:

- تقنية A/B testing: خوارزمية تعمل على إجراء مجموعة من المقارنات لتحديد أنماط التغيير في البيانات والمعطيات.

9 https://www.mongodb.com
10 https://www.cloudera.com
11 https://www.skytree.net
12 https://azure.microsoft.com/en-us/products/azure-sql/database/

- تقنية Association rule learning: خوارزميات تعمل على اكتشاف أنماط العلاقات الموجودة بين البيانات في قواعد البيانات.
- تقنية Cluster analysis: خوارزمية إحصائية لتصنيف الكيانات داخل عدد من القطاعات بشكل آلي، بناءً على إدراك أوجه الاتفاق والاختلاف بين هذه الكيانات، ويعد أبرز تطبيقاتها تصنيفات فئات المستفيدين بناءً على اتجاهاتهم البحثية.
- تقنية Crowdsourcing: تقنية تعمل على التقاط عدد ضخم من البيانات من مصادر متعددة، وإجراء عمليات التحليل عليها، وتوظف هذه التقنية بشكل كبير في شبكات التواصل الاجتماعي.
- تقنيات Data Mining: مجموعة من التقنيات التي تعمل على استخراج مجموعة من العوامل Patterns من مجموعات البيانات، من خلال عمليات الدمج بين العديد من خوارزميات التحليل والترتيب والتصنيف، لتتولد في ضوئها النتائج والمؤشرات.
- تقنية Machine Learning: إحدى تقنيات الذكاء الاصطناعي التي تعتمد على إكساب الحاسب القدرة على إجراء عمليات المعالجة بصورة قريبة من العقل البشري، معتمدًا في ذلك على العديد من البيانات التجريبية والواقعية التي يُدرَب الحاسب من خلالها.
- تقنية Natural Language Processing (NLP): تعمل على إكساب الحاسب القدرة على فهم ومعالجة وتحليل اللغات البشرية.
- تقنية Neural Networks& Deep Learning: خوارزمية تعمل على أن تكون شبيهه بشبكة الأعصاب التي توجد في عقل الإنسان؛ لإكساب الحاسب والتطبيقات القدرة على الإدراك والتعرف على الكيانات.
- تقنية Pattern Recognition: وهي خوارزمية تعمل على إجراء عمليات التعرف والإدراك من خلال إجراء عمليات التحليل لعدد من المعطيات النصية والصوتية والمرئية.
- تقنية Regression: خوارزمية إحصائية تعمل على دراسة كيفية أن التعديل في أنماط البيانات المستقلة يؤثر في تغيير القيم التابعة لها، وتستخدم هذه التقنية بشكل كبير في أغراض التنبؤ والتوقع.
- تقنية Signal Processing: خوارزميات حسابية طورت لتحليل الإشارات الصادرة من المجسات، وأجهزة البث الترددية.
- تقنية Sentiment Analysis: خوارزميات تعمل على تحليل المعلومات الموضوعية من

المواد النصية، وتحديد إذا ما كانت تتسم بالإيجابية أو الحيادية أو السلبية تجاه موضوع أو قضية ما.

- تقنية Simulation: مجموعة من الخوارزميات التي تستخدم في النمذجة، والتنبؤ، ووضع مخططات وسيناريوهات مستقبلية.
- تقنية Time Series Models: مجموعة من الخوارزميات التي تعمل على إجراء عمليات تحليل البيانات وفقًا لمتغيرات زمنية، ويعد أبرز الأمثلة التطبيقية لهذه الخوارزميات تحليل المتغير الزمني لقيم الأسهم المالية (Manyika, 2011).

٥- برمجيات تحليل البيانات الضخمة The Big Data Software:

ذهب كل من (Chen & Zhang, 2014) إلى أن تطبيقات تحليل البيانات الضخمة يمكن أن تسكن داخل ٣ قطاعات رئيسية، هي:

- القطاع الأول: تطبيقات تحليل البيانات في صورة دفعات Batch Analysis.
- القطاع الثاني: تطبيقات تحليل البيانات المتدفقة Stream Processing.
- القطاع الثالث: تطبيقات التحليل التفاعلي للبيانات Interactive Analysis.

يشير **القطاع الأول** من هذه التطبيقات إلى عمليات التحليل، التي تعتمد على أساس أن تُحفَظ البيانات أولًا بداخل ملفات الحفظ، ثم إجراء عمليات التحليل والمعالجة من خلال المعالجات، ومن أشهر أنماط هذه الملفات (HDFS, Amazon S3, GPFS)، وتكفل تطبيقات هذا القطاع العديد من عمليات التحليل مثل (تقديم التوصيات بوصفها نتائج، وإدارة مصادر البيانات، ومراقبة أداء العمليات والإجراءات، والنمذجة المرئية، والجدولة لتنفيذ المهام، وحساب القيم من مجموعات البيانات) وتمثل تطبيقات هذا القطاع النمط التقليدي من تقنيات تحليل البيانات الضخمة، ويعد أبرز تطبيقات هذا القطاع (تطبيق Hadoop، وتطبيق Microsoft Dryad، وتطبيق Apache Mahout).

أما **القطاع الثاني**، فيشير إلى عمليات تحليل البيانات المتدفقة Stream Data كبيانات السندات المالية، والمؤشرات الطبية (كتحليل معدلات نبضات القلب بشكل آني)، حيث إن عملية التحليل تتم على البيانات دون أن تُحفَظ داخل ملفات حفظ البيانات، وبجانب استخدامات تطبيقات هذا القطاع في إجراء عمليات التحليل الآني للبيانات، فإنها تستخدم لأغراض تتعلق بالذكاء الاصطناعي، وإجراء عمليات المعالجة المستمرة لأنماط البيانات، والتنقيب عن البيانات، والتصنيف، معتمدة في ذلك على نماذج التحليل الآنية Real-Time

Data Analytical Models، ويعد أبرز التطبيقات لهذا القطاع برنامج Storm[١٣] المطور من شركة Twitter، وتطبيق Spark[١٤]، وبرنامج MOA[١٥] مفتوح المصدر.

بينما يشير **القطاع الثالث** إلى التطبيقات التي تعرف باسم تطبيقات التحليل التفاعلية، والتي تكفل السهولة والبساطة في إجراء عمليات التحليل من خلال أن توفر للمستفيد التعبير المرئي عن استفساراته Queries Visualizing القائم على استخدام الصور والرسوم والأمثلة، وذلك من خلال محركات بحث تعرف باسم Query by Example Engine (QbE)، ويعد الهدف الرئيس من هذه الفئة من التطبيقات توفير القدرة على تشغيل الآلاف من خوادم البيانات في وقت واحد، وإجراء عمليات التحليل لتريليونات التسجيلات في ثواني، كما تكفل أيضًا إجراء نمط التحليل الدلالي لمجموعات البيانات، ويعد من أبرز هذه التطبيقات، تطبيق Apache Drill[١٦]، وتطبيق SpagoBI[١٧]، وتطبيق D3[١٨].

كما قدم معهد McKinsey تقريرًا قام فيه بحصر تطبيقات تحليل البيانات الضخمة، حيث عكس هذا الحصر مختلف التطبيقات دون التقيد بمجال موضوعي أو تطبيقي أو قيود محددة في تغطيته لها، وتسوق هذه الدراسة بعض من هذه التطبيقات على النحو التالي:

- تطبيق Cassandra: تطبيق مفتوح المصدر يعتمد على تحليل البيانات وفق منهجية نظم الملفات الموزعة، وقد طور من جانب شركة Facebook.
- تطبيق Hadoop: تطبيق مجاني ومفتوح المصدر، يسمح بإجراء العديد من أنماط تحليل البيانات.

تطبيق Dynamo: نظام لتحليل البيانات يعتمد على تقنية النظم الموزعة، وقد طور من جانب شركة Amazon.

- تطبيق Big Table: تطبيق طور من جانب شركة Google لإدارة البيانات من خلال مجموعة من الملفات الموزعة.
- تطبيق Google File System: أحد أنظمة الملفات الموزعة التي تقوم بحفظ وتحليل البيانات.
- تطبيق Mashup: تطبيق يعمل على استخدام ودمج البيانات من مصدرين أو أكثر لإنشاء خدمة جديدة (Brown, 2011).

13 https://storm.apache.org/downloads.html
14 https://spark.apache.org/downloads.html
15 https://moa.cms.waikato.ac.nz/
16 https://drill.apache.org/download/
17 https://spagobi.software.informer.com/4.1/
18 https://d3.en.softonic.com/

٦- البيانات الضخمة في مجال المكتبات والمعلومات:

قبيل استعراض الدراسات التي تناولت تطبيق نظام Hadoop في مجال المكتبات والمعلومات، يجدر الإشارة إلى أن تطوير نظام Hadoop قد استند على وثيقتين مصدريتين، حيث عكست الوثيقة الأولى الجانب المفاهيمي والنظري (Ghemawat, 2003)، والتي قدمها مهندسو شركة google باعتبارها إطارًا نظريًّا للحاجة إلى تطوير نظام ملفات لحفظ البيانات، يتسم بكونه لا مركزيًّا وكونه موزعًا على عدد من الأجهزة، وقد صدرت هذه الوثيقة تحت عنوان The Google File System.

أما الوثيقة الثانية، فعكست الجانب العملي والتطبيقي في تطبيق نظام Hadoop والقيام بعمليات تحليل البيانات بشكل عام (O'Malley, 2008)، وقد قدمت من جانب O'Malley، وهو مهندس البرمجيات في شركة Yahoo، الذي أسند إليه مهمة توظيف تقنيات Google File system، وتقنية MapReduce داخل نظام Hadoop، لمعالجة وتحليل مجموعة بيانات Data Set تتسم بضخامتها وتنوعها (كما سبق الإشارة من قبل إلى أن شركة Google قد قامت بطرح كلتا التقنيتين باعتبارهما إطارين نظريين دون القيام بتطبيقهما عمليًّا).

نجح (O'Malley, 2008) في توظيف كلتا التقنيتين داخل منصة Hadoop، حيث قام بتشكيل أول مجموعة عمل Cluster مكونة من ٩١٠ عقد Node (حاسب)، مشتملة على ١٠ مليار تسجيلة من البيانات، ليجرى ١٨٠٠ مهمة Task تعكس عمليات التحليل والمعالجة والفرز لهذه التسجيلات في الوقت نفسه، لتسفر عملية التشغيل عن نجاح منصة Hadoop في إجراء المهام التحليل بشكل كامل (١٨٠٠ مهمة) لكافة تسجيلات البيانات في وقت بلغ ٢٠٩ ثوانٍ أي نحو **٣ دقائق و٤٨ ثانية**، ليسجل رقمًا قياسيًّا في كونه أول نظام لمعالجة وتحليل للبيانات يعالج هذا الحجم من البيانات مستغرقا هذا التوقيت.

تأتي على رأس أبرز الدراسات التطبيقية التي عكست توظيف واستثمار تقنيات البيانات الضخمة في مجال **المكتبات**، ما قامت به شركة (ProQuest, 2013) من إعداد دراسة تطبيقية لتحليل أنماط سلوك المستفيدين في عمليات البحث على أداة البحث Summon (وهي أداة اكتشاف لمصادر المعلومات تُضَمّن داخل بنية فهارس المكتبات لإجراء عمليات دمج بين نتائج البحث الواردة من قواعد البيانات المشتركة بها المكتبة، وبين نتائج بحث الفهرس في واجهة تعامل واحدة)، وقد قامت فيها بعملية تحليل لمئات الملايين من سجلات الحفظ المشتملة على عمليات البحث، التي قام المستفيدون بها على هذه الأداة عبر مئات

الفهارس، لتكشف عن عدد من المؤشرات، والتي كان أبرزها أن سلوك المستفيد في البحث يعتمد على استخدام ما بين ٢- ٤ مصطلحات في الاستفسار الواحد، فضلًا عن العديد من القيم المضافة التي كان لها بالغ الأثر في أن تقوم الشركة بإضافة عدد من الخصائص الداعمة لعمليات البحث والاسترجاع بالصورة التي تكفل رضاء المستفيد والباحث في فهرس المكتبة.

كذلك أوضح (Wang, 2016) وآخرون أن مجتمع المكتبات يجب أن يذهب إلى أبعد من عمليات التنظيم والاسترجاع لمصادر المعلومات، حيث يجب أن يعيد النظر في دوره داخل مجتمعات المعلومات، وذلك في ظل ما تملكه المكتبات من أحجام معلومات ضخمة ومتنوعة، بالصورة التي تستلزم أن تدرك المكتبات أبعاد عمليات تحليل Analysis وتحويل Transformation وإنشاء Creating البيانات.

وهو ما أكده (Federer, 2016) في تناوله للدور الذي يجب أن يلعبه اختصاصي المعلومات في سياق ظاهرة البيانات الضخمة، هذا الدور الذي لن يقتصر على ما كان يقوم به آنفًا من عمليات تجميع ووصف لمصادر المعلومات في شكليها التقليدي والرقمي، وتقديم الخدمات من خلالهما، بل سيتخطى ذلك في توظيف تقنيات وتطبيقات تحليل البيانات الضخمة لتقديم قيم مضافة للمستفيد من هذه المصادر وعنها، وليس هذا فحسب بل إن دور اختصاصي المعلومات في إدارة البيانات، سوف يتعاظم من خلال تعاونهم مع الباحثين في ظل ما فرضته سياسات العديد من الدوريات العلمية على الباحثين من ضرورة وضع خطط لإدارة البيانات Data Management Plans، تعكس ما قاموا به من عمليات إدارة لبيانات أبحاثهم وفرص استثمار ما قد وصلوا له من نتائج على الصعيد البحثي، الأمر الذي يلزم الباحثين بمراجعة اختصاصيي المعلومات لإدراك هذه الجوانب من إدارة البيانات.

هذا الأمر الذي أسفر عن توجه عدد من المكتبات الأكاديمية كمكتبة New York University (NYU)، ومكتبة معهد MIT لتقديم ما يعرف بخدمات البيانات Data Services، تلك الخدمات ستفرض على أمناء المكتبات تأهيلهم بعدد من المهارات تشمل التنظيم والبحث والإتاحة لمجموعات البيانات، والوعي بمصادر البيانات غير التقليدية، وإدراك قضايا الملكية الفكرية (Gordon, 2012).

وفي سياق إفادة المكتبات من مؤشرات شبكات التواصل الاجتماعي، كشفت دراسة (Cervone, 2017) عن جدوى إفادة المكتبات من مؤشرات عمليات الاستخدام لصفحات المكتبات، حيث تكفل هذه المؤشرات التعرف على النشاطات المبذولة وعمليات التفاعل التي

تتم على صفحات المكتبات في هذه الشبكات، وكان من أهم نتائج الدراسة أن الكثير من المكتبات تعمل على تفعيل وسائل التواصل الاجتماعي من أجل التسويق لخدماتها المعلوماتية من خلالها.

أما عن الإسهام العربي المتخصص في ظاهرة البيانات الضخمة، فتجدر الإشارة إلى أن هذا القسم لا يستطرق إلى الدراسات النظرية للظاهرة، والتي زخرت بالكثير من الدراسات النظرية والتوظيفية والتحليلية دون التطرق للجانب التطبيقي بإجراء عملية تنصيب لأحد منصات تحليل البيانات الضخمة، وإجراء عملية تحليل لأحد مصادر البيانات الضخمة من خلال منصات التحليل المتعارف عليها، وعلى هذا يركز هذا القسم على استعراض الدراسات التطبيقية العربية لظاهرة البيانات الضخمة، دون النظرية أو الاستكشافية او التحليلية.

على الرغم من أن دراسة (الهادي، ٢٠١٦) تعد من الدراسات النظرية في مجال تحليل البيانات الضخمة، فإنها تعد من أولى الدراسات العربية المتخصصة في مجال البيانات الضخمة، ومع ذلك فقد اقتصرت على الجانب الوصفي والمسحي في التعامل مع قضية تحليل البيانات الضخمة، ويعد أبرز ما أشارت إليه هذه الدراسة أن مجال البيانات الضخمة يحمل تأثيرًا قويًّا على رسم السياسات واتخاذ القرارات التخطيطية والتنموية، وذلك على صعيد القطاع العام والخاص، هذا التأثير يتمثل فيما تفرضه هذه الظاهرة من تحديات على الصعيد السياسي والتنموي.

كذلك جاءت دراسة كل من (Al-Daihani & Abrahams, 2016) باعتبارها دراسة تطبيقية لتحليل مجموعة من التغريدات Tweets الخاصة ببعض المكتبات الأكاديمية، وقدد أوضحت الدراسة أن أبرز تقنيات تحليل البيانات التي يمكن أن توظف في سياق المكتبات الأكاديمية هي تقنية تحليل النصوص Text Mining، والتي يمكن توظيفها في العديد من أنشطة المكتبات كصفحات التواصل الاجتماعي الخاصة بالمكتبة، وأن هذا النمط من عمليات التحليل سوف يسفر عن العديد من المؤشرات، يكون لها بالغ الأثر في عمليات اتخاذ القرار.

وفي دراسة أخرى، رصد (الأكلبي، ٢٠١٨) أحد الأنظمة المحلية التي قامت جامعة الملك سعود بتطويرها لإدارة مستودعات بياناتها، وقدرتها على التعامل مع البيانات الضخمة، ارتكزت الدراسة بصورة أساسية على توضيح البنية التكوينية للنظام، وأن النظام يتسم بالفاعلية في تحقيق أهدافه من استخراج للمؤشرات.

وعلى الرغم من عدم استعراض *(Al-Barashdi, 2018)* أحد الجوانب التطبيقية

لتوظيف ظاهرة البيانات الضخمة في المكتبات، فإن ما قدمته في مراجعتها للإنتاج الفكري الصادر حول توظيف تقنيات البيانات الضخمة في المكتبات الأكاديمية، يعد إسهامًا بارزًا في هذا المجال، فقد قدمت في دراستها حصر بأبرز التقنيات التي يمكن للمكتبات الأكاديمية أن تقوم بتوظيفها في تحليل بياناتها، والتي كان أبرزها التقنيات الإحصائية، وتقنيات التعلم الآلي، وتقنيات التنقيب عن البيانات، كذلك قامت بإجراء حصر عن لأنسب البرمجيات المناسبة لتوظف في المكتبات الأكاديمية، والتي تمثلت في كل من برنامج *Hadoop*، وبرنامج *Apache Mahout*، وبرنامج *Microsoft Dryad*.

تعد أقرب الدراسات لهذه الدراسة هي دراسة (فرج، ٢٠٢٢)، والتي أوضح فيها الجانب الهيكلي لمنصة Hadoop، وخصائصها، وبيان مراحل إدارة البيانات الضخمة في مؤسسات المعلومات بمراحل تجميع وتدفق وتخزين وعرض وتحليل البيانات، وذلك لأغراض تجهيزها وتوظيفها لتعزيز البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات، لكنها لم تتطرق لتوظيف المنصة في إجراء عملية تحليل أو معالجة لمجموعات البيانات الضخمة.

٧- الإسهام العلمي لهذه الدراسة:

اتضح من خلال ما استُعرض من الدراسات العربية، التي عكست الجانب التطبيقي في تناول ظاهرة البيانات الضخمة في مجال المكتبات والمعلومات، أن التركيز الرئيس فيها تمحور حول قياس الإفادة من هذه الظاهرة في هذا المجال، وأثر ما تلقيه هذه الظاهرة من تحديات في التعامل مع البيانات في مجال المكتبات والمعلومات، وإن كانت دراسة كل من *Al-Daihani & Abrahams* قد تناولت توظيف لأحد تقنيات تحليل النصوص لمحتوى صفحات المكتبات الأكاديمية على موقع *Twitter*، لكنها لم تتطرق لتوظيف تطبيق أو منصة خاصة لتحليل البيانات الضخمة، وفي الوقت نفسه افتقر جُل الدراسات العربية للجانب التجريبي والتطبيقي ودراسات الحالة لكل من منهجيات وخوارزميات وتطبيقات تحليل البيانات الضخمة في مجال المكتبات والمعلومات. وفي ضوء ذلك يتمثل الإسهام العلمي الرئيس لهذه الدراسة فيما ارتضته من أهداف لنفسها، والتي يمكن إجمالها في المحاور التالية:

- تحليل البنية التكوينية لنظام *Hadoop*.
- اختبار نظام *Hadoop* في بناء نموذج حقيبة الكلمات *Bag of words* لتحليل أحد مصادر المعلومات غير المهيكلة، ورصد فاعلية المعالجة والتحليل له.

الجانب النظري للدراسة:

١- المفهوم والتعريف بظاهرة البيانات الضخمة:

جاء مصطلح البيانات الضخمة The Big Data بوصفه مصطلحًا جديدًا نسبيًّا، يحمل في دلالته ومفهومه العديد من القضايا التي تتعلق بالبيانات، ويعمل على إيجاد الحلول التي تكفل فرص الاستثمار الأمثل للبيانات، وتوفيرها لمن يحتاجها في الوقت المناسب، وبأقصر السبل، كما يعمل هذا المصطلح على اكتشاف قيم مضافة جديدة من البيانات اعتمادًا على منهجيات الاستدلال، ومما تم استقراءه من بيانات متاحة، والاستنباط مما قد يتولد من قيم مضافة أو قرارات، كذلك يبرز من بين ثنايا هذا المصطلح العديد من التحديات التي تتعلق بتجميع هذه البيانات، وتنظيمها، وإدارتها على نحو فعال، فضلًا عن توفير آليات الاسترجاع التكاملية ليبرز من خلالها ما يعرف بالمعرفة ذات الحكمة ""Wisdom Knowledge، والتي تحتل ذروة سنام مجتمعات المعرفة.

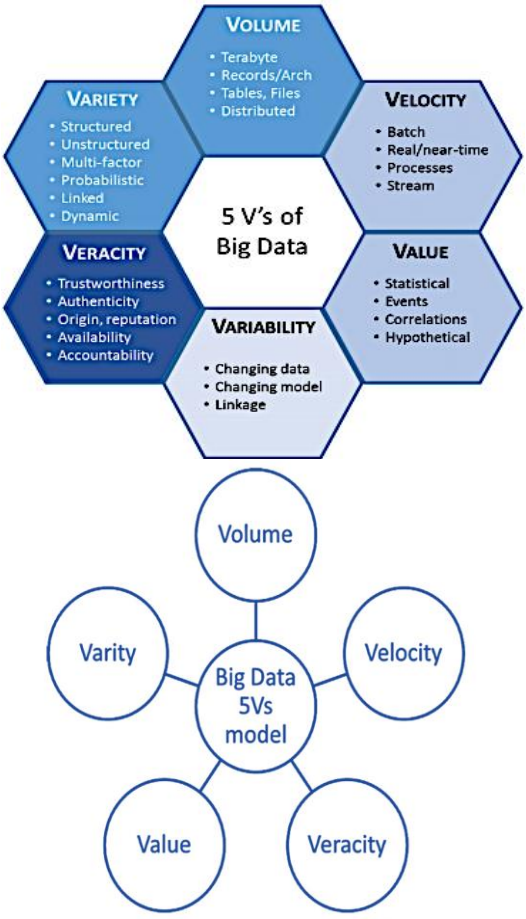
كما أن هذا المصطلح قد أصبح في الآونة الأخيرة يشير إلى القدرة على استخدام التحليلات التنبؤية لاستنباط التوجهات البحثية والاقتصادية، فضلا عن الاعتماد على تحليل أنماط سلوك المستفيد لرصد توجهاته على مختلف الأصعدة الحياتية له (Elgendy, 2014)، ولعل أبرز مثال على هذا الأمر ما يظهر في الكثير من المواقع في خدمات وسمت بعنوان "هذا الكتاب موصى لك به أو مرشح لك "Recommend it for You" وهي خدمة ناتجة عن تحليل بيانات استخدام المستفيد للمعلومات وأنماط إفادته منها.

ذهب كل من (Boyd & Crawford, 2011) إلى أن اقتصار استخدام مصطلح البيانات الضخمة للدلالة على الظاهرة من منظور حجم البيانات فقط (كما هو الحال في الخطاب العلمي والاقتصادي والاجتماعي والسياسي في الإشارة لها على أنها مجموعات البيانات الكبيرة التي تتطلب حاسبات وتطبيقات فائقة لمعالجتها)، هو استخدام يعلوه الضعف في الدلالة، فضخامة البيانات في الوقت الراهن يمكن أن تعالج من جانب العديد من التطبيقات القياسية المتاحة على الأجهزة المحمولة، ومن ثم فإن دلالة هذا المصطلح يجب أن تذهب إلى أبعد من وصف الضخامة للبيانات، لتشير إلى أنماط العلاقات التي تربط بين هذه البيانات بعضها بعضًا.

قام (Laney, 2001) بنمذجة هذا المفهوم في نموذج ثلاثي عكس أبرز خصائص البيانات الضخمة، والتي تمثلت في الحجم Volume، والسرعة Velocity، والتنوع Varity.

ليتبع ذلك قيام شركة IBM بتوسعة هذا النموذج ليشمل خاصيتين أخريين هما

المصداقية Veracity، والقيمة Value، ليظهر النموذج في صورة خماسية الأبعاد كما هو موضح في الشكل رقم (١٦).



شكل رقم (٨)

يوضح نماذج تحديد خصائص البيانات الضخمة

ويستتبع (Normandeau, 2013) أنه بجانب كل من (ضخامة حجم البيانات، وارتفاع معدلات تدفقها، وتنوعها، وعدم القدرة على التحقق من مصداقيتها، وصعوبة توليد القيم المضافة منها)، كخصائص للبيانات الضخمة، يجب أن يؤخذ في الحسبان خاصيتين أخريين، هما خاصية الصلاحية Validity، والتي تشير إلى تحدي التحقق من صحة البيانات ودقتها، وخاصية قابلية التغيير Volatility، والتي تشير إلى تحدي تحديد الفترة التي ستظل فيها البيانات صالحة للاستخدام، وتحديد الفترة التي يجب على المؤسسات استبقائها وحفظها أو التخلص منها.

أما عن التعريف بالظاهرة، فقد وردت العديد من التعريفات الرامية إلى تحديد هوية ظاهرة تضخم البيانات، ومن أبرزها ما يلي:

- "هي البيانات التي تقع خارج قدرة التقنيات الحالية لتخزينها وإدارتها وإجراء مختلف العمليات عليها بفاعلية وكفاءة" (Manyika, 2011).
- "هي أنماط من البيانات تتسم بالضخامة على صعيد الحجم وتزايد سرعة إنشائها مما يتطلب أشكالًا مبتكرة وفعالة من حيث التكلفة لمعالجة البيانات لتعزز القدرة على صنع القرار، وتوفير الإحصائيات المناسبة" (Gartner, 2023).
- "مصطلح يهدف إلى تحديد نمط من المعالجة لبيانات تتسم بالضخامة على صعيد حجمها وسرعتها والتعقيد في التقاط المعلومات منها وتوزيعها وإدارتها وتحليلها" (TechAmerica, 2023).

في هذا السياق أوضح كل من (Boyd & Crawford, 2011) أن مفهوم البيانات الضخمة يشير إلى ٣ عناصر أساسية تتمثل في:

- الإشارة إلى العلاقات التي تربط بين البيانات بعضها بعضًا.
- الإشارة إلى التحديات التي تواجه إدارة ومعالجة البيانات الضخمة داخل نظم ومستودعات البيانات.
- الإشارة إلى منهجيات وآليات التحليل التي في ضوئها تُستخَرج المؤشرات والقيم المضافة.

٢- تقنيات معالجة البيانات الضخمة: منصة Apache Hadoop:

١- نشأة منصة Hadoop وطبيعتها:

منصة Haddop [١٩]، إحدى أبرز البرمجيات المجانية ومنصات العمل مفتوحة المصدر التي طورت في سياق مجال تحليل البيانات الضخمة.

استهلت الجهود في مجال معالجة البيانات الضخمة على صعيد تطوير التقنيات في عام ٢٠٠٢، بشروع كل من Doug Cutting & Mike Cafarella في وضع مسودة لتطوير محرك بحث لديه القدرة على تكشيف مليار صفحة مجمعة ومتاحة على الويب، على أن يتسم هذا المحرك ببنية مفتوحة المصدر، وقد أطلقا على ذلك المشروع مسمى Nutch.

وعقب دراسة الأمر على صعيد التنفيذ، اتضح أن تكلفة هذا المشروع على صعيد

19 Http://hadoop.apache.org

العتاد المادي Hardware تبلغ نحو نصف مليون دولار، وتكلفة عملية التشغيل له تصل لنحو ٢٠ ألف دولار شهريًّا، وذلك في ظل الاعتماد على التقنيات التقليدية، لذا تعثر إنشاء المشروع في ظل البحث عن آليات وتقنيات تعمل على تقليل تكلفة التخزين، ويكفل فعالية معالجة البيانات منها (Aggarwal, 2020).

وفي الوقت نفسه بالتحديد قامت شركة جوجل في عام ٢٠٠٣ بتطوير تقنية جديدة لحفظ ملفات البيانات، عرفت باسم نظام جوجل The Google File System، والتي صدرت تفاصيلها في ورقة بحثية وسمت بذات المسمى.

وتعتمد هذه التقنية على إنشاء ملفات حفظ لديها القدرة على حفظ وتخزين البيانات وفقًا لنمط موزع Distributed Manner، من خلال تقسيم ملفات البيانات ذات المساحات الكبيرة إلى كتل ذات مساحات ثابتة، ثم حفظها بشكل موزع على عدد من الحاسبات المتصلة فيما بينها، بصورة تتسم بالتكاملية، مع القيام بتكرار هذه الكتل على عدد من الحاسبات، وذلك بهدف القيام بإجراء أكثر من عملية معالجة أو تحليل على ذاتها نسخ الكتلة الواحدة دون الحاجة للانتظار من فراغ المهام بشكل متعاقب، ومن ثم يكفل هذا الأمر تحقيق فعالية السرعة في المعالجة.

عرفت هذه التقنية بالملفات الموزعة Distributed Files، وتوضح Google أن ابتكار هذه الملفات جاء في ظل الحاجة لمنصة مبتكرة لديها القدرة على توليد وحفظ ومعالجة وتحليل ما تملكه وتتعامل معه الشركة من بيانات تتسم بضخامتها وتنوع مصادرها، وتعتمد هذه التقنية على أن يقوم مدير النظام بإنشاء ما يعرف بمجموعة العمل Clusters، يقوم من خلالها بحفظ ملفات البيانات على الآلاف من الأجهزة (الخوادم Servers)، لتُكتَب وتُقرأ البيانات بتنوع أشكالها على هذه الملفات الموزعة بشكل متزامن (Ghemawat, 2003).

جاء هذا الأمر ليشكل حلًا مناسبًا لمشروع Nutch لإجراء عمليات حفظ وتخزين البيانات ذات الأحجام الضخمة والتنوع في فئاتها، ولكن تبقي قضية المعالجة والتكشيف والتحليل لهذه البيانات المحفوظة بداخل نمط موزع من الملفات قضية وعائقًا لم يحل من خلال ما طرحته شركة Google في الورقة السابقة.

ليقوم مطورو شركة Google مرة أخرى بتقديم ورقة بحثية في عام ٢٠٠٤، يُطرح من خلالها آلية لمعالج البيانات الضخمة داخل هذه الملفات الموزعة، وقد وسمت هذه الآلية بمسمى تقنية MapReduce.والتي هدفت من خلالها إلى توفير إطار برمجي يكفل إجراء عمليات المعالجة والتحليل والاسترجاع المتعدد على العديد من البيانات المحفوظة داخل

الملفات الموزعة في الوقت نفسه وبشكل تكاملي، وذلك بشكل متوازي في وقت واحد، دون التقيد بالنمط التقليدي بأن تتم المعالجة في صورة جدولة للمهام الأولى ثم الثانية (Dean, 2004).

جدير بالذكر أن شركة Google قد قامت بطرح كل من الورقتين البحثيتين، دون الإجراء في تنفيذهما في ذلك الوقت، ليقوم كل من Doug Cutting & Mike Cafarella بتنفيذيهما في مشروعهما Apache Nutch وجعله منصة مفتوحة المصدر.

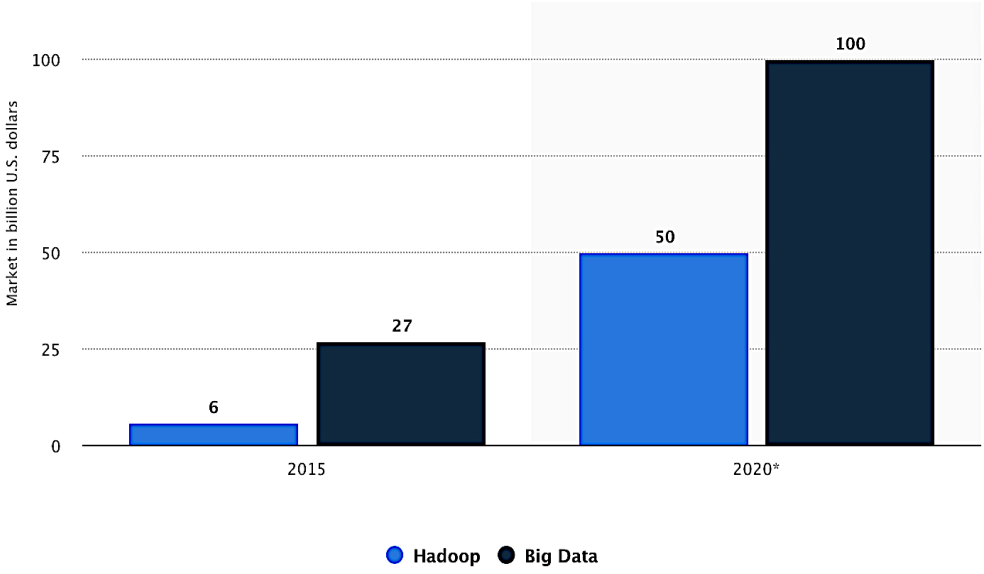
عقب تصميمهما للمشروع، اتضح لكل من Cutting & Cafarella في عام ٢٠٠٥ أن مشروعهما يتسم بالمحدودية على صعيد ما يتاح له من بيانات، إذ بلغ أقصى حجم تتكون منه مجموعة العمل الخاصة بالمشروع نحو ما يقرب ما بين ٤٠-٥٠ وحدة عمل Node (حاسب)، وهو الأمر الذي لا يعكس استثمارًا لمزايا وفعالية ما يُوظَف من تقنيات في معالجة البيانات الضخمة.لذا عملا كل منهما على البحث عن بيانات ضخمة، لتطبيق كلتا التقنيتين عليها، فعمدا على التوجه نحو شركة Yahoo، في ظل ما تملكه هذه الشركة من أحجام بيانات تتسم بالضخامة في ذلك الوقت (اعتمادًا على محرك بحثها ودليلها الموضوعي)، ذاك الأمر الذي كفل لهما القدرة على التحقق من فعالية كل من التقنيتين في عمليات الحفظ والمعالج، لتقوم شركة Yahoo، بتطبيق مشروعه على مجموعة عمل Cluster كونتها من ألف وحدة حاسب، ليجرى عليها عمليات حفظ واسترجاع من خلال هاتين التقنيتين، لتنجح الشركة في عملية معالجة البيانات الضخمة غطت من خلالها مليارات عمليات البحث والتكشيف والتحليل لملايين الصفحات المتاحة على الويب. عقب نجاح اختبار التقنيتين قامت شركة Yahoo بإتاحة مشروع منصة Hadoop في عام ٢٠١١ باعتباره نظامًا مفتوح المصدر لمعالجة البيانات الضخمة تحت مسمى الإصدار الأول Hadoop 1.0 (*Yahoo,2008*).

توالت عقب ذلك إصدارات نظام Hadoop ليصدر الإصدار الثالث منه عام ٢٠١٨، وتتوالى عليه العديد من الإضافات والبرمجيات من جانب العديد من المطورين والشركات، لتشكل منصة Hadoop ما يعرف بـ Hadoop Ecosystem، ويمكن رصد أبرز المحطات الرئيسة في تاريخ تطور منصة Hadoop كما هو موضح في الشكل رقم (١٧) (Aggarwal, 2020).

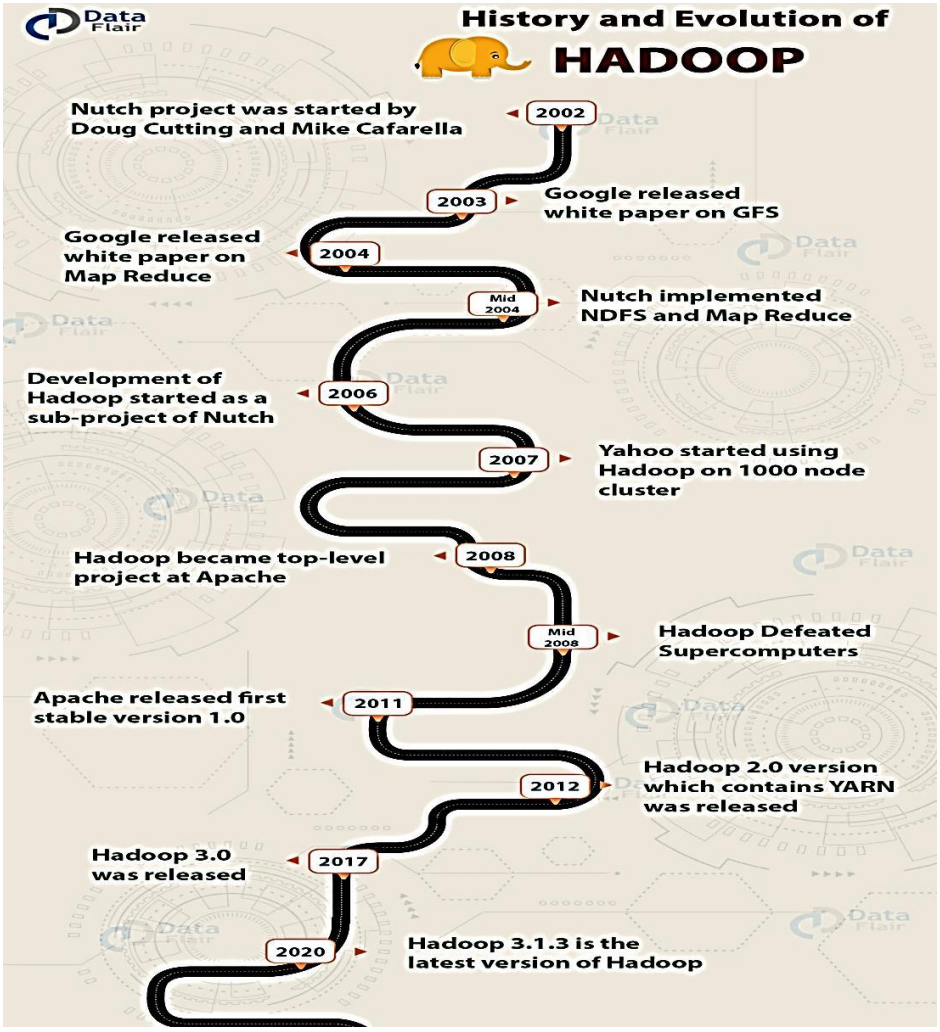
في عام ٢٠١٧ تتبعت وكالة البيانات الدولية IDC، معدلات استخدام وتبني نظام Hadoop في مجال تحليل البيانات في مختلف القطاعات المهنية، وأفضى تتبعها إلى أن

نظام Hadoop يستحوذ على أكثر من ٣٥٪ من إجمالي سوق تبني برمجيات ونظم البيانات الضخمة في المؤسسات على صعيد العالم (*Wilson, 2017*).

على جانب آخر قام موقع Statista (المعني بتقديم الإحصائيات في مجال تقنيات المعلومات على صعيد العالم) بإجراء مسح إحصائي رصد فيه معدلات نمو استثمار نظام Hadoop من خلال تبنيه من جانب الشركات والمؤسسات لتحليل ما تملكه من بيانات، واتضح أن نظام Hadoop قد نمت معدلات الاستثمار فيه من ٦ مليار دولار في عام ٢٠١٥ إلى ٥٠ مليار دولار في عام ٢٠٢٠، كما أن نظام Hadoop بهذه القيمة يستحوذ على نحو نصف إجمالي الاستثمارات التي تستثمر في مجال تحليل البيانات الضخمة (*Statista, 2017*).



شكل رقم (٩) يوضح استحواذ نظام Hadoop على نحو ٥٠٪
من إجمالي الاستثمارات في مجال تحليل البيانات الضخمة في عام ٢٠٢٠.



شكل رقم (١٠)

يوضح الخط الزمني لنشأة وتطور نظام Hadoop (*Data Flair Team ,2020*).

بلغت معدلات استحواذ نظام Hadoop على عمليات تحليل البيانات الضخمة داخل المؤسسات على صعيد العالم نحو ٥٠٪، وهذا ما أشار إليه تقرير (Syncsort, 2017) الصادر عام ٢٠١٧، وإن نحو ٣٥٪ من المؤسسات العاملة في حقل البيانات تقوم بالاعتماد على هذا البرنامج في تحليل ما لديها من بيانات، وذلك على صعيد العالم، كما أعلنت شركة Yahoo أنها قامت في يونيو من عام ٢٠١٢ باستخدام هذا البرنامج على نحو ٤٢ ألف خادم للبيانات، في ٤ مراكز للبيانات، لمعالجة ما تملكه من بيانات، كما قامت شبكة التواصل الاجتماعي Facebook بإعلان أنها تقوم بمعالجة نحو ١٠٠ مليار اكسابايت من

البيانات يوميًّا من خلال هذا البرنامج، وأن ٩٤٪ من مستخدمي البرنامج يقومون باستخدامه لأغراض تحليل البيانات، بجانب عدد من الاستخدامات الأخرى (*Khan,2015*).

٢- البنية التكوينية لبرنامج Hadoop:

ينمتي نظام Haddop إلى فئة من أنواع البرامج على صعيد البنية المعمارية، تعرف ببرمجيات المستوى المرتفع لخوادم الأباتشي Top Level Apache، (أي أن هذه النظم تبنى فوق خوادم الأباتشي في صورة أشبه بالطبقات Layers)، وفي الوقت ذاته تُبنَى العديد من البرمجيات (كبرنامج - Zookeeper – وبرنامج AVRO) وغيرها من البرامج على برنامج Hadoop نفسه، والتي تكون بالنسبة له برمجيات ذات مستوى مرتفع له، لتعرف معمارية برنامج Hadoop وفقًا لهذا الوضع بالنظام التكويني لبرنامج Hadoop Ecosystem.

إن المنهجية العلمية التي بُنيت على أساسها ملفات برنامج Haddop، تتمثل في أن نموذج عملها Model ينمتي إلى نماذج (الوظيفة – للبيانات Function –to- Data)، وليس لنماذج (البيانات – للوظيفة Data- To – Function).

ويمكن توضيح أوجه الفرق بين كلا النموذجين في أن النمط الأول يكون التركيز فيه على القيام بالوظائف (وهي عمليات التحليل)، دون الاهتمام بطبيعة البيانات أو فئاتها، (نظرًا لأن أنواع وأشكال البيانات كافة تحفظ في هذه الملفات دون التفريق أو التمييز)، أي أن الهدف الرئيس من إنشاء التطبيق أو البرنامج يتمثل في التركيز على القيام بوظائف دون النظر إلى طبيعة البيانات التي سيتم تضمينها داخل النظام.

أما النموذج الثاني، فيشير إلى التركيز على البيانات، (أي أن الملفات تحدد طبيعة ما يجب أن تشتمل عليه من بيانات وأحجامها كما هو الحال في نظم إدارة قواعد البيانات التقليدية فيما يعرف بمرحلة تصميم جداول البيانات)، أي أن الهدف الرئيس من هذا النموذج يتمثل في إنشاء وتطوير برامج يكون هدفها حفظ ومعالجة البيانات.

كما ينمتي هذا البرنامج على صعيد بنيته التكوينية إلى نظم الملفات الموزعة المجمعة Distributed Clustered File System (أي يعتمد على توزيع ما يشتمل عليه من ملفات البيانات على العديد من الخوادم).

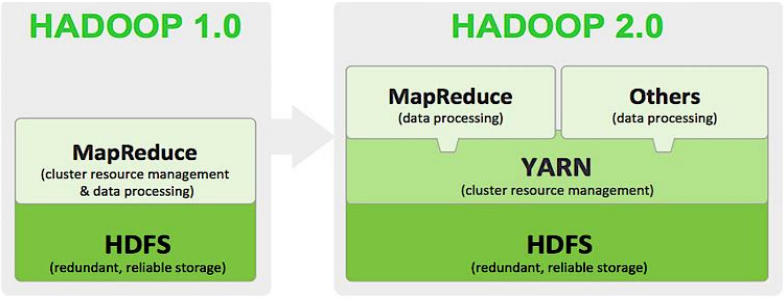
استنبط تطوير منصة Hadoop من ٣ تقنيات:

- التقنية الأولى: هي تقنية نظم ملفات جوجل الموزعة Google Distributed Files System (GFS)، والتي طور في سياقها نمط ملفات يعرف باسم Hadoop

.Distributed File System (HDFS)

- أما التقنية الثانية، فهي تقنية MapReduce، وتمثل المعالجات التي تقوم بعمليات تحليل البيانات التي توجد داخل هذه الملفات بصورة متوازية.
- وفي عام ٢٠١٢، نظرًا لعدد من التحديات والصعوبات وأوجه القصور التي اعتلت تقنية MapReduce، طُورَت تقنية ثالثة عُرِفَت باسم YARN (Yet Another Resource Negotiator)، والهدف منها أن تكون بمثابة تقنية لديها القدرة على إدارة موارد الحاسبات المكونة لمجموعة العمل (ويقصد بالموارد في هذا السياق الوحدات التكوينية المادية للحاسبات مثل المعالج CPU، ووحدة الذاكرة العشوائية RAM، وغيرها من تقنيات العتاد المادي) (*Mois, 2016*).

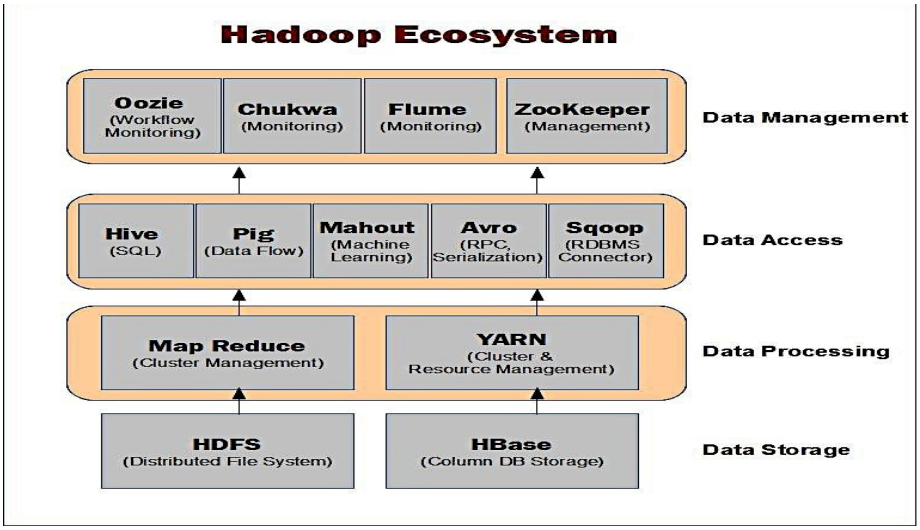
ليظهر نظام Hadoop مشتملًا على ٣ تقنيات حتى الوقت الراهن، وليبلغ عدد إصداراته ٣ إصدارات، وتسمى الثالثة بمسمى Hadoop 3.1.3، والتي صدرت عام ٢٠١٨ (*Khan,2015*).



شكل رقم (١١) يوضح فروق البنية التكوينية بين الإصدار الأول الثاني، لنظام Hadoop((Hadoop Tutorial, 2020

تشتمل البنية التكوينية لبرنامج Haddop على ٤ نظم فرعية، تجتمع معًا لتشكل ما يعرف بمنصة Hadoop Ecosystem، وهذه النظم تتمثل في:

- نظام Hadoop للملفات الموزعة Hadoop Distributed File System (HDF).
- إطار العمل الخاص بعمليات التحليل والمعالجة MapReduce.
- تقنية إدارة ومراقبة العتاد المادي والبرمجي للخوادم، وتعرف باسم YARN.
- التطبيقات التشغيلية والوظيفية المساندة لبرنامج Hadoop.



شكل رقم (١٢) يوضح البنية التكوينية لنظام Hadoop

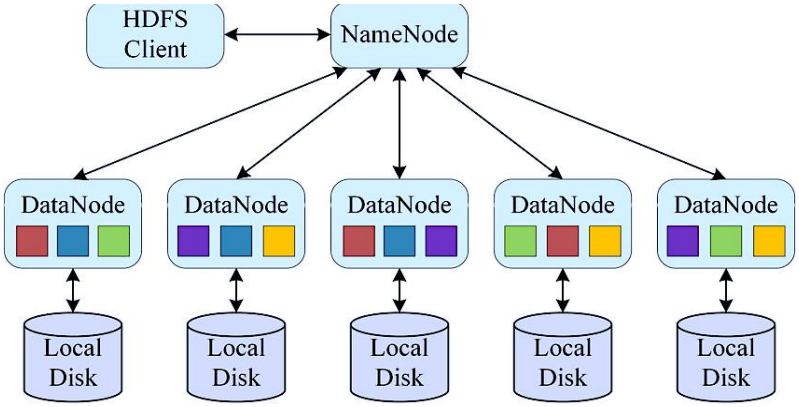
لتشكيل ما يعرف بنظام Hadoop Ecosystem (Daniele, 2021).

١- **نظام** Hadoop **للملفات الموزعة** Hadoop Distributed File System (HDFS):

يأتي نظام الملفات الموزعة في برنامج Haddop، بمثابة المكون الرئيس الأول لهذا النظام، ويعد بمثابة المستودع الرئيسي لحفظ البيانات داخل هذا النظام، ويعد الهدف الرئيس من تطوير هذا النمط من الملفات هو التعامل مع البيانات ذات الأحجام والتنوع والتدفق والخصائص الضخمة التي لا يمكن معالجتها على آلة أو خادم مفرد.

تعتمد البنية التكوينية لبرنامج Hadoop على أن يقوم بإنشاء مجموعة من ملفات التخزين لمختلف أنماط وأشكال وأنواع البيانات، هذه الملفات تحفظ على مجموعة من الخوادم، لتشكل هذه الخوادم في صورتها المجمعة ما يعرف بمجموعة العمل Cluster، وبداخل مجموعة العمل الواحدة تُقسَم الخوادم إلى فئتين:

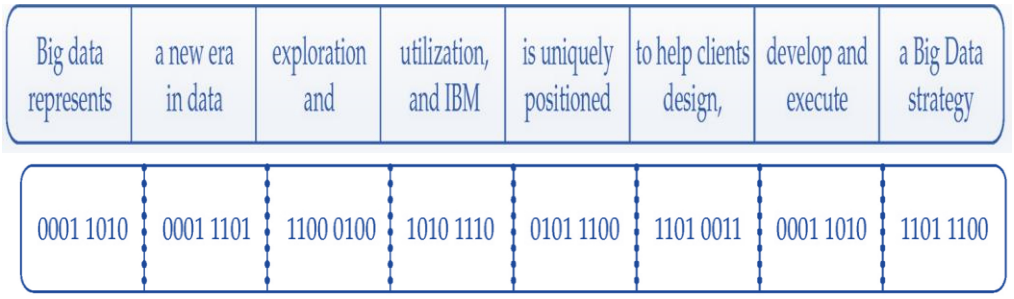
- الفئة الأولى تعرف باسم Data Node: وتشير إلى الخوادم التي تشتمل على البيانات التي سيتم عليها عمليات المعالجة.
- أما الفئة الثانية من خوادم مجموعة العمل، فتعرف باسم Name Node: وتشير إلى كونها الخوادم الرئيسة في مجموعة العمل.



شكل رقم (١٣)

يوضح أنواع خوادم منصة Hadoop.

بداخل خوادم البيانات Data node تُقسَم البيانات المراد معالجتها إلى أقسام صغيرة، تعرف باسم كتل أو وحدات البيانات Data Blocks، كما هو موضح في الشكل رقم (٢١).



شكل رقم (١٤) يوضح عملية تقسيم برنامج Hadoop،

لملف نصي على كتل البيانات وفقًا للمساحة التخزينية الخاصة بهذه الكتل.

ويعد الهدف الرئيس من عمليات التقسيم والإنشاء لكتل البيانات هو السماح للبرنامج بالقيام بالعديد من عمليات التحليل المختلفة بشكل متوازي ومتعدد Multi-Functions على البيانات في وقت واحد، أي على سبيل المثال أن تُجرَى عملية تحليل البيانات نصية في الوقت نفسه الذي تُجرَى عمليات تحليل على بيانات صورية.

يعمل برنامج Hadoop على توليد خادم رئيس لمجموعة الخوادم التي تشكل مجموعة العمل Cluster، هذا الخادم يعرف باسم الخادم الرئيس NameNode.

يتولى هذا الخادم الرئيس إدارة خوادم البيانات التي تشتمل عليها مجموعة العمل Cluster، وذلك من خلال العديد من المهام، والتي يعد أبرزها:

- مهمة إنشاء الميتاداتا الخاصة لمختلف كتل البيانات التي توجد داخل خوادم البيانات،

حيث يقوم الخادم الرئيس بتسمية كتلة البيانات، وتحديد اسم الخادم الذي توجد بداخله، وتحديد طبيعة البيانات التي توجد داخل كل كتلة.

- يقوم هذا الخادم مقام الدليل لمختلف كتل البيانات التي تشتمل عليها مجموعة العمل، فبجانب حفظه لبيانات الميتاداتا الخاصة بكل خادم والميتاداتا لكتلة البيانات، يقوم بتحديد وتوزيع مهام عمليات التحليل المطلوب إجراؤها على البيانات وفقًا للكتل.
- كذلك تظهر أهمية الخادم الرئيس NameNode عند قيام المستخدم بإنشاء كتلة بيانات جديدة داخل مجموعة العمل، فإن هذه الوحدة تقوم بتحديد الخادم الذي سوف يتم إنشاء وتسكين هذا الملف بداخله.
- كذلك تتولى هذه الوحدة القيام بجدولة المهام The Job (عمليات التحليل والمعالجة)، المراد تنفيذها على البيانات، وبتحديد الكتل التي تشتمل على هذه البيانات المراد تحليلها ومعالجتها والمنتشرة على خوادم مجموعة العمل، هذا الأمر الذي يكفل تقليل حجم عمليات المراسلات والاتصالات بين الخوادم لتحديد أماكن حفظ البيانات، ومن ثم فإن تقليل هذه المراسلات والاتصالات سيكفل تعزيز وتحسين البرنامج بالقيام بما أسند إليه من مهام تحليلية للبيانات.

جدير بالذكر في هذا المقام أن وحدة العمل الرئيسة أو الخادم الرئيس The NameNode لا يعد من مهامها أن تتبع ما يُعالج أو يُحلَّل من بيانات، أي بعبارة أخرى تتولى هذه الوحدة القيام بتحديد أماكن تخزين البيانات، وبتوصيل المهمة أو الإجراء المراد تنفيذه إلى كتلة البيانات المشتملة على البيانات المراد تحليلها، فقط دون أن يتطرق إلى مخرجات عمليات التحليل أو ما نُفِذَ من إجراءات على هذه البيانات بالاسترجاع.

تقوم كافة وحدات DataNode بإرسال نبضات تعرف باسم Heartbeats لوحدة NameNode كل ثلاث ثوان، وفي حالة عدم تلقي وحدة NameNode هذه النبضات من DataNode، فسيتم اعتباره بمثابة وحدة معطلة، وفي حالة عدم قيام وحدة DataNode بإرسال نبضات لفترة تصل إلى ١٠ دقائق، فستقوم وحدة NameNode بإعادة نسخ أي كتل كانت مخزنة داخل هذه الوحدة، وعلى هذا يعد من الضروري أن تتسم وحدة NameNode بمواصفات خاصة عن مختلف خوادم البيانات التي توجد داخل مجموعة العمل، بجانب ضرورة توفير خادم احتياطي منه the Secondary NameNode بحيث يشتمل على مختلف ما تشتمل عليه هذه الوحدة الرئيسة من بيانات تتعلق بكتل بيانات مجموعة العمل.

يتمثل الهدف الرئيس لتقنية HDFS من إنشائها الملفات في صورة موزعة، الاستغلال الأمثل لمختلف الخوادم المكونة والموجودة داخل مجموعة العمل Cluster الواحدة، حيث أن

كل خادم يشتمل على مجموعة من وحدات التخزين Hard Disk Drivers، والتي تقوم تقنية HDFS باستغلالها بالشكل الأمثل أيضًا، من خلال أن يقوم بتعيين مسبق لأماكن حفظ البيانات داخل كل وحدة تخزين على حدةٍ، ثم القيام بتعيين المهام وعمليات التحليل والمعالجة المراد تنفيذها، وفقًا لأماكن البيانات المراد معالجتها، هذا المبدئ الذي يعرف باسم التعيين المكاني لكتل البيانات Data Locality.

تبرز العديد من مظاهر القوة التي تتمتع بها تقنية HDFS في حفظها للبيانات داخل هذه الخوادم، والتي يعد أبرزها **مبدأ التكرار** Redundancy القائم على أن يتم توزيع وتكرار حفظ كتل ووحدات البيانات الواحدة على أكثر من خادم، داخل مجموعة العمل الواحدة، ويعد الغرض الرئيس من ذلك، تجنب ما قد يحدث من مشكلات قد تؤدي إلى فشل وحذف هذه الكتلة أو الوحدة من أحد الخوادم.

ومن ثم فإن تكرار نسخها على عدد من الخوادم يعد إجراءً احترازيًّا لعدم فقد البيانات من الملفات، هذا من جانب، كذلك يعد من ضمن الأغراض الجوهرية من هذا التكرار لنسخ البيانات، القيام بعمليات التشغيل والتحليل على البيانات بالتوازي، أي إجراء العديد من العمليات والتحليلات على البيانات في وقت واحد، ومن ثم يقوم البرنامج بإجراء أكثر من عملية تشغيل أو تحليل على كتلة البيانات المستهدفة في وقت واحد.

أي بعبارة أخرى إذا حُدِدَ عدد ٣ عمليات تحليل أو تشغيل مطلوب إجراؤها على كتلة رقم ١ من البيانات، فإنه لن يُجرَى التحليل بشكل تسلسلي على هذه الكتلة، بحيث تبدأ عملية التحليل الثانية بمجرد الفراغ من الأولى، الأمر الذي يتسم باستغراق الكثير من الوقت في انتظار الانتهاء من عمليات التحليل.

كما أن هذا التسلسل في إجراء التحليل على كتلة واحدة بشكل متتالي يسفر عنه تحميل زائد واستنفاد لطاقة كتلة البيانات، أو قد يسفر عنه حدوث العديد من المشكلات التشغيلية لها، وعلى هذا اعتمدت تقنية HDFS على أن تقوم بتكرار نسخ كتل البيانات داخل العديد من الخوادم.

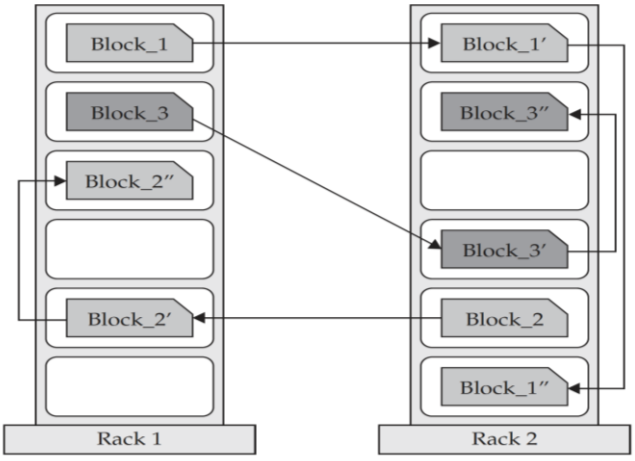
يتولى الخادم الرئيس Name Node في مجموعة العمل مهمة إنشاء النسخ المكررة من كتل البيانات في خوادم مجموعة العمل، حيث يقوم الخادم الرئيس بإنشاء نسختين مكررتين لكل كتلة أو وحدة بشكل رئيس، أي أن كل كتلة أو وحدة بيانات تتكون من نسخة رئيسة ونسختين إضافيتين، كذلك يوفر الخادم الرئيس إمكانية وقابلية زيادة عدد النسخ المكررة للوحدة وفقًا لرؤية مدير النظام.

أما عن منطقية توزيع النسخ المكررة على خوادم البيانات، فتتمثل في أن توجد نسخة احتياطية بجانب الكتلة الرئيسة من البيانات على الخادم الذي توجد فيه كتلة البيانات، وتوجد

نسخة إضافية على خادم آخر غير الخادم الذي توجد فيه كل من النسخة الرئيسة والاحتياطية، كما هو موضح في الشكل رقم (٥)، حيث يشتمل هذا الشكل على خادمين يشتملان على ٣ كتل من البيانات تم تسميتهم باسم Block 1,2,3، حيث نُسِخَت كل واحدة من هذه الكتل أو الملفات نسختين، وعنونت النسخة الأولى لكل كتلة بالملحقة (′)، أما النسخة الثانية فقد عنونت بالملحقة (″)، للإشارة إلى النسخة الثانية من كتلة البيانات، وتشير كلمة Rack إلى الخادم الذي يشتمل على كتل البيانات.

تعتمد منهجية توزيع كتل البيانات وما أنشئ من نسخ لها على أن تنشأ وفقًا لمقاصة محددة، تتمثل في أن كلًا من الملفين الأول والثالث يتم إنشاؤهما على خادم واحد Rack 1، أما الملف الثاني فينشأ على خادم آخر Rack 2، أما النسخ الخاصة لكل من الملفين الأول والثالث فيتم إنشاؤهما على الخادم الذي يوجد عليه الملف الثاني، في حين أن نسخ الملف الثاني تنشأ على خادم الملفين الأول والثالث.

تعد الحكمة الرئيسة لهذه المنهجية في التوزيع أن لا توجد مختلف النسخ الرئيسة من الملفات على خادم واحد، بينما توجد النسخ المكررة على خادم آخر، بل كل خادم يشتمل على كتل رئيسة ويشتمل أيضًا على نسخ لكتل أخرى غير الكتل الرئيسة التي توجد لديه، هذا الأمر الذي يكفل لمجموعة العمل أن تقوم بعمليات الإصلاح لمختلف المشكلات أو حالات الأعطال التي قد تحدث لإحدى كتل البيانات بصورة تلقائية.

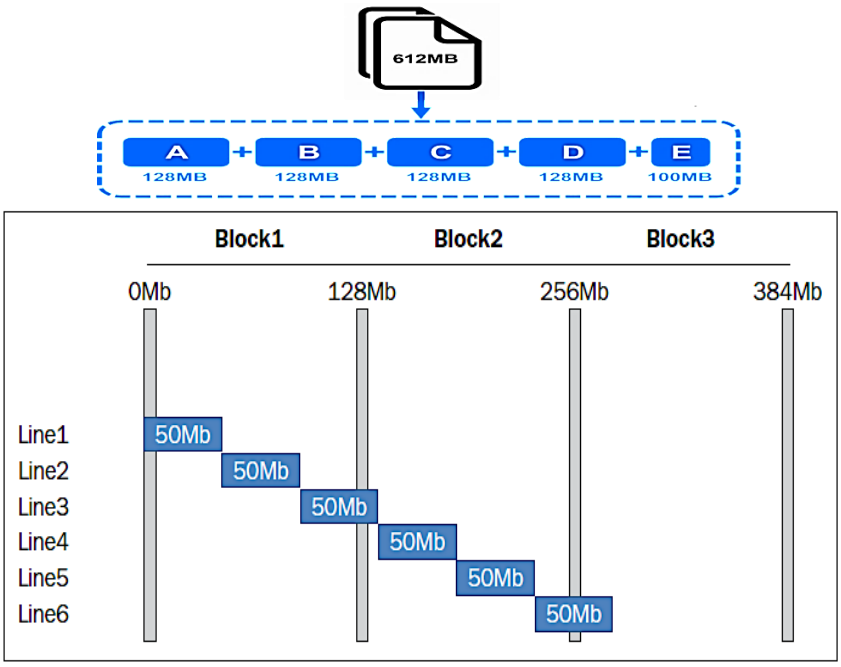


شكل رقم (١٥)

يوضح تكرار عملية النسخ لكتل البيانات على خوادم مجموعة العمل داخل ملفات HDFS.

تلعب المساحة التخزينية لكتل البيانات دورًا مهمًا في عمليات المعالجة والتحليل لها، فالمساحة التخزينية لكتلة أو وحدة البيانات الواحدة داخل برنامج Hadoop تأتي بصورة

معيارية في حجم ١٢٨ ميجابايت، كما هو موضح في الشكل رقم (٢٤).



شكل رقم (١٦) يوضح كيفية ملء كتل البيانات داخل منصة Hadoop وفقًا للمساحة الثابتة، وهي مساحة ١٢٨ ميجابايت.

أما عن كيفية التعامل مع هذه الملفات من جانب المستفيد، فقد كفلت منصة Hadoop واجهة تعامل خاصة به غير رسومية، تعتمد على كتابة الأوامر بصورة خاصة بها، وجدير بالذكر في هذا المقام الإشارة إلى أن على الرغم من كون واجهة تعامل نظام Hadoop تعتمد على كتابة الأوامر السطرية Command Shell في التعامل مع ما يشمله من بيانات، فإن هذه الواجهة لا تنتمي إلى فئة واجهات التعامل التي تعرف باسم (واجهات التعامل القائمة على أوامر نظم التشغيل Portable Operating System Interface For Unix POSIX).

ويمكن توضيح هذا المفهوم من خلال إدراك أن أي برنامج يُنَصَّب على نظم التشغيل كتنصيب برنامج MS. Word على نظام تشغيل Windows، فإن واجهة التعامل الخاصة بهذا البرنامج أو التطبيق تعتمد على الأوامر التي يشتمل عليها نظام التشغيل المنصب عليه، أي أن تنفيذ أمر النسخ Copy على برنامج word. MS يكون من خلال الأمر Ctrl+C، كما هو الحال في نسخ الملفات على نظام التشغيل Windows، ولكن يختلف الأمر في واجهة تعامل نظام Hadoop، فعلى الرغم من كونه نظامًا يعمل على نظم تشغيل

سطرية كنظام Unix، فإن له أوامره الخاصة في إدارة الملفات، أي أن الأوامر التقليدية التي يشتمل عليها نظام التشغيل Unix لا تطبق في واجهة تعامل هذا البرنامج، وبعبارة أخرى فإن واجهة تعامل نظام Hadoop لها أوامرها الخاصة التي لا تشابه أوامر نظم التشغيل التي يُنَصّب عليها، ويمكن استعراض بعض من أوامر واجهة تعامل برنامج Hadoop في إدارة الملفات كما يلي:

- cat: الأمر الخاص بنسخ الملفات.
- chmod: الأمر الخاص بتغيير صلاحيات القراءة والكتابة لمجموعة من الملفات.
- chown: قم بتغيير مالك أو صاحب هذه الملفات.
- copyFromLocal: قم بنسخ الملفات من تطبيق ما، وإدخالها داخل ملفات برنامج Hadoop.
- copyToLocal: قم بنسخ الملفات من ملفات برنامج Hadoop لملفات تطبيق ما.
- expunge: قم بتفريغ الملفات التي توجد في سلة المهملات كافة.
- ls: قم بعرض قائمة الملفات التي توجد في دليل ما.
- mkdir: قم بإنشاء دليل لملفات نظام Hadoop.
- mv: قم بنقل الملفات من دليل لآخر.
- rm: قم بحذف الملف وإرساله إلى سلة المهملات.
- rm-skiptrash: قم بحذف الملفات مباشرة دون إرسالها لسلة المهملات.

تعتبر إحدى أبرز المشكلات التي يتسم بها نظم الملفات HDFS، أن الملفات بداخله أو كتل البيانات، تحفظ لأغراض إجراء المعالجة والتحليل عليها فقط، أي لا تتمتع بإمكانية أن يُعَدَّل ما تشتمل عليه من محتوى، وهو الأمر الذي يعرف بمنطق Write Once, Read Many، أي أن بمجرد أن يتم دفق الملفات داخل نظم HDFS، لا يمكن أن يُعَدَّل عليها في ظل وجودها على هذا النظام، أي توجد فقط لأغراض المعالجة لما تشمله هذه الكتل من بيانات أو محتوى دون التعديل على محتوى هذه الكتل.

الأمر الذي أسفر عن أن عمليات التعديل على محتوى الملفات، يتم خارج نظم HDFS، على نظم التشغيل Local System، ثم دفقها داخل نظم HDFS.

٢- نموذج أو إطار عمل MapReduce لمعالجة وتحليل البيانات في برنامج Hadoop:

ينظر إلى نموذج MapReduce، على كونه إطار العمل أو التطبيق البرمجي الرئيس Programming Paradigm لمنصة Hadoop، والذي تُنَفذ من خلاله عمليات التشغيل

والتحليل المراد تنفيذها على كتل البيانات التي توجد في مجموعات العمل التي تشتمل عليها خوادم الملفات والبيانات لنظام Hadoop، أي بعبارة مختصرة، فإن كافة عمليات التحليل والمعالجة والتشغيل والمهام المراد تنفيذها على البيانات يقوم بتنفيذها هذا التطبيق.

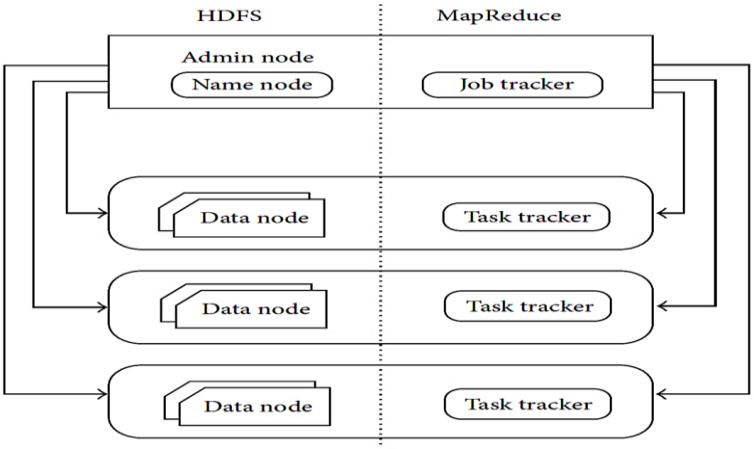
يشتمل هذا التطبيق في تنفيذه لعمليات تحليل البيانات، على العديد من المفاهيم، حيث تعرف عمليات التحليل المراد تنفيذها على كتل البيانات بمسمى العملية Job، حيث يقوم برنامج MapReduce باستقبال طلب عملية تحليل ما (العملية job) من المستفيد، والقيام بإرسال هذه العملية داخل بنية ملفات النظام من خلال وحدة فرعية تعرف باسم متتبع العملية JobTracker.

يقوم متتبع العملية JobTracker بالاتصال بخادم الملفات الرئيس NameNode لإعلامه بوجود عملية تحليل ما، ليقوم خادم الملفات الرئيس بإيجاد وتحديد أماكن كتل البيانات الموزعة على مختلف خوادم مجموعة العمل، والمراد إجراء عمليات التحليل عليها، وعند تحديد كتل البيانات المراد معالجتها، تُقَسَّم العملية Job، إلى مجموعة من المهام Task، بحيث يُحَدّد لكل كتلة المهمة الخاصة بها، ثم جدولتها لكل وحدة، ومن ثم تُجرَى عمليات التحليل كما هو موضح في الشكل رقم (٢٥).

عقب تحديد المهام لكل كتلة، يقوم تطبيق MapReduce بتوليد ما يعرف بمتتبع المهام TaskTracker، والذي يتولى عملية متابعة إجراء تنفيذ المهمة على كتلة البيانات المحددة لها، ويقوم بتوضيح حالة تنفيذ المهمة، وفي حالة وجود خطأ أو مشكلة في تنفيذ المهمة داخل كتلة البيانات، يقوم هذا المتتبع بإرسال تقرير لمتتبع العمل JobTracker، والذي يقوم بسرعة بإعادة جدولة المهمة التي لم تنفذ لكتلة بيانات أخرى (وذلك في ظل عمليات النسخ التي تمت لهذه الكتل داخل مجموعة العمل).

تعتمد منهجية عمل تقنية MapReduce داخل منصة Hadoop على وحدتين فرعتين، ولكل منهما وظيفته الخاصة به والتي تتعلق بإجراء عمليات تحليل البيانات:

- الوحدة الأولى: هي وحدة التحويل للبيانات، وتعرف باسم Map.
- الوحدة الثانية: هي وحدة (تخفيض) النتائج، وتعرف باسم Reduce.



شكل رقم (١٧) يوضح قيام JobTracker بتقسيم ما يشمله من عمليات تحليل إلى عدد من المهام Tasks، وفي المقابل يقوم الخادم الرئيس NameNode بتحديد كتل البيانات المقابلة لكل مهمة يراد تنفيذها عليها.

يمكن توضيح مهام كل من وحدة تحويل البيانات Map ووحدة النتائج Reduce، من خلال المثال التالي:

بافتراض وجود مكتبة رقمية تشتمل على العديد من الوثائق والكتب في صورتها الرقمية، بحيث يعكس كل كتاب (أو منفردة) كتلة من البيانات النصية غير المهيكلة Unstructured Data، حيث تتكون كل كتلة من النسق التقليدي للكتاب من حيث اشتماله على العديد من الفصول والفقرات والجمل والكلمات، حيث ترد النصوص مسردة بشكل تتابعي دون أن تتسم بهيكلة أو مخطط قد يشتمل على صورة الحقول والقيم.

وتعد عملية التحليل الرئيسة job المطلوب إجراؤها اختيار إحدى هذه الكتل من البيانات (أي اختيار كتاب ما)، وحساب تردد ظهور أو تكرار كل كلمة داخل هذا الكتاب، بصورة إحصائية، وبصورة مهيكلة، أي بعبارة أخرى، تقتضي عملية التحليل هذه بشكل ضمني تحويل البيانات من نسقها غير المهيكل إلى نسق مهيكل، بحيث يرد أمام كل كلمة عدد مرات ظهورها أو تكرارها داخل الكتاب، على النحو التالي:

Word	Freq.
River	20
Deer	37
Bear	47
Car	15

شكل رقم (١٨)

يوضح صورة مخرجات عملية التحليل لأحد الكتب من خلال تقنية Map Reduce

تتولى منصة Hadoop إجراء عملية التحليل الإحصائية لهذا الكتاب من خلال تشغيل الوحدة الفرعية الأولى لوحدة MapReduce، وهي وحدة Map والتي تتولى إجراء مسح للملف النصي بالكامل وتحديد الكلمات الواردة في بنية الكتاب وتمييزها كافة.

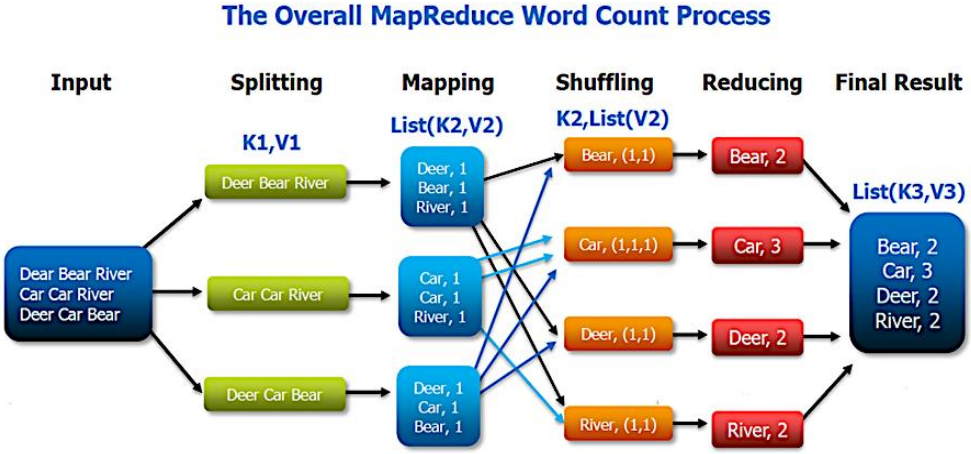
ليعقب ذلك أن تقوم هذه الوحدة بإجراء تقسيم Splitting لهذا النص لعدد من كتل البيانات Blocks بحيث تتكون كل كتلة من جزء من الكتاب مشتملة على كلماته في مساحة محددة لا تتجاوز ١٢٨ ميجابايت.

ثم يتولى إجراء عملية التحويل Mapping لطبيعة البيانات من كونها بيانات غير مهيكلة (نصية) إلى بيانات مهيكلة في صورة (حقل Field or Keys، وقيمة Values)، وهنا تكمن قيمة منصة Hadoop في تحويل البيانات غير المهيكلة لبيانات مهيكلة، بحيث يُنشأ عمودان داخل كل كتلة بحيث يعكس العمود الأول كافة الكلمات الواردة داخل كتلة البيانات في النص، ويطلق على العمود في هذه الحالة (المفتاح Keys)، يجدر الإشارة إلى أن في هذا العمود ستُسجل كل كلمة دون الأخذ في الاعتبار تكرارها، أي أن الكلمة الواحدة ستظهر أكثر من مرة داخل هذا العمود. أما العمود الثاني، فيطلق عليه مسمى القيمة Values، ويشير إلى القيمة الخاصة بكل كلمة تم تسجيلها العمود الأول Keys، أي أن الكلمة سيُسَجل ظهورها، حتي وإن كانت مكررة داخل الكتلة.

ليعقب ذلك أن تقوم هذه الوحدة الفرعية أيضًا بإجراء عملية التجميع Shuffling لمرات تكرار كل مصطلح من مختلف كتل البيانات التي وزع عليها النص، بحيث تظهر كل كلمة وأمامها كل عدد مرات ظهورها من كل كتلة (أي على سبيل المثال وردت كلمة Mohamed في ٥ كتل فيُعَبر عنها داخل هذه العملية بهذه الصورة Mohmed, 1,1,1,1,1)، وبهذه العملية ينتهي دور الوحد الفرعية الأولى وهي وحدة Map.

لتبدأ عمل الوحدة الفرعية الثانية وهي وحدة تخفيض النتائج Reduce، والتي يعكس مسماها طبيعة ما تقوم به، وهو تخفيض حجم النتائج التي تُوُصِل إليها، (وهنا تكمن أيضًا قيمة منصة Hadoop، من خلال قدرتها على تخفيض حجم البيانات الضخمة لصورة تسهل من إدراك مخرجات عمليات التحليل)، تتمثل مدخلات هذه الوحدة فيما أُخرِج من نتائج من الوحدة الفرعية Map، بحيث تتولى عملية مسح لمختلف النتائج التي توُصِل إليها، من خلال الوحدة الفرعية الأولى Map، وتحليلها لاستخراج نتيجة واحدة تعكس الإجابة أو النتيجة المطلوب تنفيذها لأمر التشغيل أو التحليل وهو إيجاد قيمة واحدة تعكس مرات ظهور كل كلمة داخل النص بالكامل، (أي على سبيل المثال تقوم بتجميع قيم مرات ظهور كلمة

Mohamed من صورته السابقة 1,1,1,1,1 Mohmed, لتكون في هذه الصورة (Mohmed, 5)، كما هو موضح في الشكل رقم (٢٧).



شكل رقم (١٩) يعكس وجود كتل البيانات، والتي تقوم وحدة التحويل Map باستخراج البيانات وتسكينها وفقًا لقطاعات ثنائية (المفتاح والقيمة)، لتقوم وحدة Shuffling بإرسال هذه القطاعات إلى وحدة Reduce.

في ضوء المثال السابق يتضح أن الوظيفة الرئيسة للوحدة الأولى في تطبيق MapReduce، هي:

إجراء عملية مسح لكتل البيانات التي توجد داخل خوادم البيانات، وذلك عقب قيام JobTracker بعملية الاتصال بالخادم الرئيس لتحديد كتل البيانات المراد إجراء عمليات التحليل أو التشغيل عليها، وعقب جدولة المهام لكل وحدة، لتبدأ بعد ذلك وحدة Map بهيكلة البيانات التي توجد في كتل البيانات وفقًا للمهمة المراد تنفيذها، هذه الهيكلة تسفر عن تكوين مجموعة من القطاعات تعتمد على عمودين (المفتاح، والقيمة)، وتُسَكّن البيانات المراد تحليلها واستخراجها وفقًا لهذه البنية، ثم تقوم بتقديم المخرجات لوحدة فرعية أخرى تتمثل في وحدة Shuffling والتي تتولى إرسال مخرجات وحدة تحويل البيانات Map، إلى وحدة تخفيض النتائج Reduce؛ أي أن مخرجات وحدة Map، هي مدخلات وحدة Reduce.

عقب استلام وحدة تخفيض النتائج للبيانات التي أرسلت من جانب وحدة تحويل البيانات Map، من خلال وحدة Shuffling، تقوم وحدة Reduce بتخفيض مختلف النتائج التي توصل إليها لتقديم نتيجة نهائية تضاهي المهمة Task المطلوب تنفيذها على كتل البيانات، ثم القيام بحفظ هذه النتائج داخل ملفات مجموعة العمل، وتقديمها لمن يطلبها.

ويمكن إجمال مهمة تطبيق MapReduce في عمليات معالجة البيانات وتحليلها على النحو التالي (الجدول رقم ٣):

جدول رقم (٣)

يوضح المهام التي تقوم بها وحدة MapReduce في معالجة وتحليل البيانات.

الوحدة الفرعية	المهمة الخاصة بالوحدة الفرعية
المهمة أو العمل Job	-تتمثل في عمليات التحليل المراد إجراؤها على كتل البيانات.
متتبع العمل JobTracker	-يقوم بإجراء عملية الاتصال بالخادم الرئيس لمجموعة العمل NameNode ليحدد كتل البيانات التي سوف تتم عليها عمليات التحليل. -يقوم بتقسيم عمليات التحليل لمجموعة من المهام، والتي تُقابل بكتل البيانات وتُنَفذ عمليات التحليل عليها. -يقوم بجدولة المهام لكل كتلة داخل مجموعة العمل.
وحدة تحويل البيانات Map	-يقوم بإجراء عملية إعادة الهيكلة للبيانات وفقًا للمهمة المراد تنفيذها. -تتمثل هذه الهيكلة في صورة عمود يتكون من (مفتاح، وقيمة).
وحدة إرسال البيانات Shuffling	-تقوم هذه الوحدة باستلام البيانات المهيكلة وفقًا للعملية المراد تنفيذها، وإرسالها إلى وحدة تخفيض البيانات. -تقوم بعملية فرز وترتيب النتائج قبل إرسالها لوحدة تخفيض النتائج Reduce.
وحدة تخفيض النتائج Reduce.	-تقوم بدمج النتائج التي تُرسَل من قبل وحدة Map، من خلال وحدة Shuffling، واستخلاص النتيجة النهائية التي تلبي الغرض الرئيس من المهمة المراد تنفيذها على كتلة البيانات.
النتائج Result	-تُخزَن النتائج التي تُوصِّل إليها داخل ملفات مجموعة العمل. -وتقديمها للمستفيدين من خلال هذه الملفات.

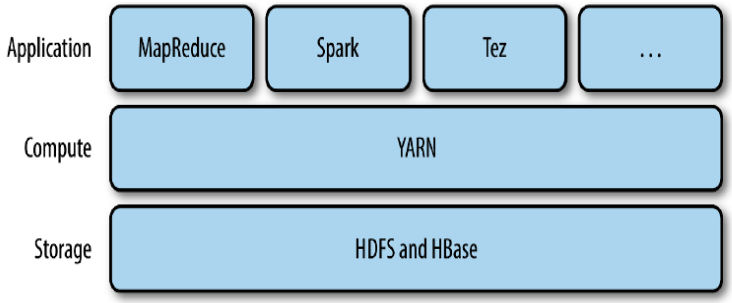
٣- وحدة YARN لإدارة موارد خوادم مجموعة عمل Hadoop:

تعرف هذه الوحدة الفرعية باسم Apache YARN (Yet Another Resource Negotiator)، طُوِرَت هذه الوحدة في الإصدار الثاني من منصة Hadoop 2.0، لإزالة المشكلات التي تواجه وحدة Job Tracker في الإصدار الأول من منصة Hadoop 1.0، وذلك بغية تحسين عمليات إدارة الموارد الخاصة بمجموعة العمل Cluster، (يقصد بالموارد

في هذا السياق هو الوحدات المادية للتخزين والمعالجة كRAM ، وCPU).

حيث أشار الموقع الرسمي لتطوير منصة Hadoop، إلى أن وحدة MapReduce داخل إصدار الجيل الأول من منصة Hadoop 1.0، تتسم بالبطء وعدم الفاعلية في عملية التحليل عند تجاوز عدد الحاسبات والخوادم المكونة لمجموعة العمل لأكثر من ٤٠٠٠ خادم، ومن ثم فقد تتسبب قابلية التوسع بعد ذلك في تدهور الأداء وزيادة فشل المهام.مما استدعى الحاجة للتفكير في إنشاء وحدة خاصة بإدارة موارد مجموعة العمل Cluster، للحفاظ على كفاءة وحدة MapReduce، بحيث تتولى تحديد المصادر المناسبة لكي تنهض هذه الوحدة MapReduce بمهام التحليل والمعالجة، وقد تمثل هذا المدير التنفيذي في الوحدة الفرعية التي عرفت باسم YARN، لتكون بمثابة مدير للموارد Resource Manager لتتولى تحديد حجم CPU التي تتطلبها عملية التحليل، وحجم الذاكرة المطلوبة كذلك لحفظه وتخزينه، وبالفعل طُوِرَت وحدة YARN.

تتمثل منهجية عمل الوحدة الفرعية YARN في توفير واجهات عمل برمجية APIs، والتي تكفل للمستفيد أن يقوم باستقبال أوامره المتعلقة بالتحكم والإدارة لمختلف المصادر التي تتكون منها مجموعة العمل Cluster داخل منصة Hadoop، تعمل هذه الواجهات عوضًا عن أن يقوم المستفيد بكتابة الأوامر أو الطلبات بشكل مباشر للوحدة الفرعية YARN، بل يتم ذلك الأمر من خلال العديد من وحدات وسيطة قد تكون وحدة MapReduce إحداها، كما هو موضح في الشكل رقم (١٦).



شكل رقم (٢٠)

يوضح منهجية التعامل غير المباشر مع وحدة YARN داخل منصة Hadoop.

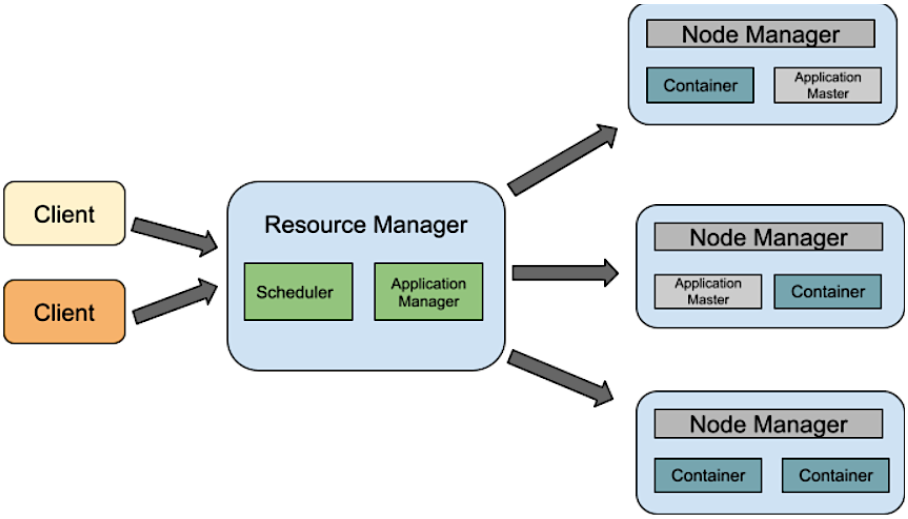
تعمل الوحدة الفرعية YARN من خلال وحدتين أساسيتين تشتمل عليهما، هما:

- وحدة Resource Manager: حيث تنشأ لكل مجموعة عمل وحدة خاصة بها تتولى عملية

إدارة مصادر المعالجة والتحليل (التي يقصد بها العتاد المكون لمجموعة العمل)، والتي تتولى عملية تحديد وتعيين العتاد الخاص بإجراء كل عملية تحليل تُطلَب من خلال المستفيد.

تنطوي مسؤولية وحدة Resource Manager على أن تقوم بجدولة الأوامر التي تتلقاها من المستفيد من خلال الوحدة الفرعية Scheduler، والتي تتولى عملية الجدول لعميات التحليل المطلوبة، وفي هذا المقام لا تتولى هذه الوحدة القيام بأية إجراءات تتعلق بالمراقبة أو الإدارة، أو التتبع لهذه المهام. لتتولى وحدة فرعية أخرى داخل وحدة Resource Manager، والتي تعرف باسم Application Manager، مهام عملية الإدارة والمراقبة في تنفيذ عمليات التحليل.

- أما الوحدة الفرعية الأخرى، فتتمثل فيما يعرف باسم Node Manager، والتي تتولى عملية إدارة مختلف وحدات البيانات أو ما عرف سابقًا باسم خوادم البيانات التي تشتمل على كتل البيانات المراد تحليلها أو معالجتها، وذلك من خلال تجهيز المصادر (العتاد مثل CPU - RAM) لإجراء وتنفيذ المهام أو العمل Job (عملية. التحليل)، وذلك من خلال ما يعرف بوحدة الحاوية Container، والتي يجهز فيها ما حُدِّد من مصادر لمعالجة وتنفيذ المهام المطلوبة من خادم البيانات Data Node.



شكل رقم (٢١)

يوضح البنية التكوينية لوحدة Yarn وكيفية عملها داخل منصة Hadoop.

٣- التطبيقات والبرمجيات التشغيلية والتنفيذية المساندة لمنصة Hadoop

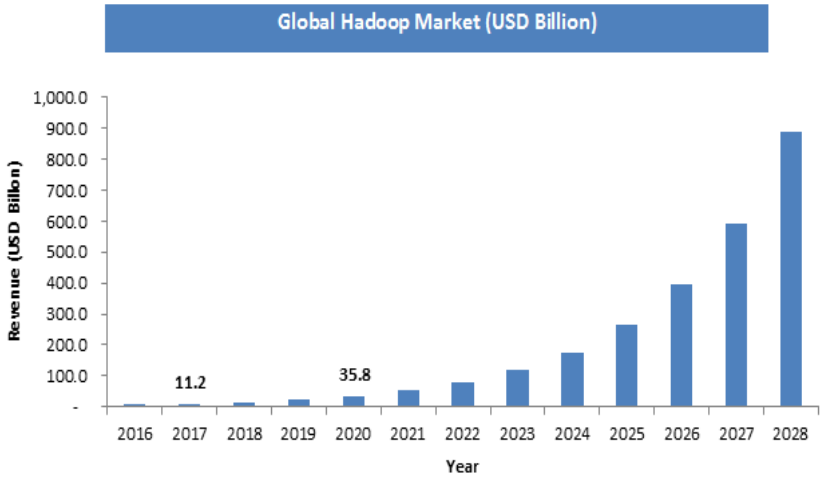
تنطوي عملية التشغيل والتنصيب لمنصة Hadoop على أن يدرك القائم بالتشغيل أو المطور طبيعة احتياجه من المنصة وطبيعة المهام المنشودة، والتي سيستخدم لتحقيقها المنصة، حيث سيتشكل في ضوء طبيعة الاستخدام والاحتياجات التقنيات التي سيقوم المطور باختيارها، وذلك الأمر الذي سيشكل ما يعرف بمنصة Hadoop Ecosystem.

وتنطوي عملية الاختيار للتقنيات لتنصيب منصة Hadoop، أن يدرك المطور خياريين مهمين، هما:

- أن يعمد المطور لأحد المنصات والتوزيعات الأكثر شيوعًا واستخدامًا، والتي تمتاز بجمعها وإنشائها للتقنيات المطلوبة لأغراض وأهداف مشروع المطور، ومن أشهر هذه التوزيعات Cloudera، ومنصة Hortonworks.
- أن يقوم بتجميل منصة Hadoop في وضعها المجرد (أي بدون اشتمالها على الوحدات الإضافية)، ويقوم بشكل منفصل بتجميع وإنشاء التقنيات المطلوبة، وهذا اشبه بالتنصيب اليدوي Manual.

خلال السنوات السابقة، تطورت منصة Hadoop لتكون بمثابة أقوى المنصات والأنظمة لتخزين البيانات الضخمة وتحليلها، لتستخدم المنصة على نطاق واسع في شتى المؤسسات في مختلف المجالات في الوقت الراهن، فوفقًا لإحدى الدراسات الحديثة الصادرة من موقع Marketwatch حول حجم السوق ومعدلات استخدام منصة Hadoop، أوضحت أنه من المتوقع أن يتجاوز حجم سوق Hadoop أكثر من ٣٠٠ مليار دولار بحلول عام ٢٠٢٥، موضحة أن سوق Hadoop العالمي في وضع يسمح لها بالنمو المذهل في السنوات القادمة كما هو موضح في الشكل رقم (Market Watch, 2022).

وعلى هذا، جذب سوق التعامل مع منصة Hadoop، العديد من الشركات والمؤسسات العاملة في مجال البيانات وتكنولوجيا المعلومات كشركة IBM وشركة Microsoft وشركة Oracle، وكذلك برزت العديد من الشركات المتخصصة في منصة Hadoop مثل شركة Cloudera وشركة Hortonworks وشركة MapR، لتكون هذه الشركات بمثابة مؤسسات داعمة لمنصة Hadoop بشكل غير مباشر، في ظل اعتماد تطبيقات هذه الشركات ما طورته من برمجيات على منصة Hadoop (White, 2015).



شكل رقم (٢٢)

يوضح اتساع قاعدة استخدام منصة Hadoop، وارتفاع حجم أرباح العائدات من استخدامها.

كما سبق الإشارة إلى أن التعامل مع ملفات ونظام Hadoop يتم من خلال واجهة تعامل خاصة بها، تعتمد على مجموعة من أوامر التشغيل Commands Shell الخاصة بها، أي أن واجهة التعامل الخاصة بالبرنامج في إدارة الملفات لا تعتمد على الأوامر التقليدية التي توجد داخل نظم التشغيل التقليدية كنظام Unix أو Linux، وذلك على صعيد التعامل مع ملفات النظام HDFS التي توجد داخل خوادم مجموعة العمل.

على جانب آخر نجد أن تطبيق MapReduce الخاص ببرنامج Hadoop، تعتمد واجهة التعامل الخاصة به API على أن تُكتَب أوامرها من خلال لغة الجافا Java، مما يجعل التعامل مع هذا التطبيق مقصورًا على مبرمجين هذه اللغة فقط أو من لديه القدرة على التعامل بها.

وعلى هذا جاءت العديد من التطبيقات الفوقية لتبني فوق نظام Hadoop، (وذلك في ظل كون هذا البرنامج مفتوح المصدر)، والتي هدفت من تبسيط درجة التعقيد في النموذج البرمجي للنظام وتبسيط التعامل معه، حيث اتسمت هذه التطبيقات بالبساطة في نمط تكويدها، والعمل على أن تكون واجهة تعامل تبني فوق هذا البرنامج لتسهل من عمليات استخدامه والتعامل معه.

تنوعت هذه التطبيقات في لغات البرمجة الخاصة ببنائها والأغراض التي طورت من أجلها، ومنهجية تعاملها مع ملفات النظام HDFS وطبيعة استخدامها داخل برنامج Hadoop، وطبيعة من يقوم بالتعامل معها، لتشكل هذه التطبيقات وملفات النظام ما يعرف

بالنظام التكويني للبرنامج Hadoop Ecosystem، وقد جاءت على النحو التالي:

- **تطبيق Apache Spark:**

أحد أبرز التطبيقات المستخدمة في الوقت الراهن لمعالجة البيانات الضخمة بجانب منصة Hadoop، يعرف هذا التطبيق على صفحته بأنه "نظام مفتوح المصدر لمعالجة البيانات بشكل موزع على مجموعة من الحاسبات وخاصة لمعالجة البيانات الضخمة المولدة بشكل آني Real Time Data.

كانت بداية تطبيق Apache Spark في عام ٢٠٠٩ بوصفه مشروعًا بحثيًّا في جامعة كاليفورنيا في بيركلي داخل معمل AMPLab، يضم كل من الطلاب والباحثين وأعضاء هيئة التدريس في الجامعة، ويركز على مجالات التطبيق كثيفة التوليد للبيانات ككاميرات المراقبة وتطبيقات إنترنت الأشياء IOT، والتي ترتفع من خلالها معدلات توليد البيانات الآنية.

كان الهدف من Spark، إنشاء إطار عمل جديد ومُحسّن للمعالجة التكرارية السريعة مثل التعلم الآلي، وتحليل البيانات التفاعل، مع الاحتفاظ بقابلية التوسع، والتسامح مع أخطاء Hadoop MapReduce.

نُشرت الورقة الأولى التي تعلن عن ميلاد هذا التطبيق تحت بعنوان " Spark: Cluster Computing with Working Sets" في يونيو ٢٠١٠، وفي يونيو ٢٠١٣، دخلت Spark تحت مظلة تطبيقات Apache Software Foundation،

أما عن عملية تشغيل تطبيق Spark، فيمكن أن يعمل بشكل مستقل، ولكن في أغلب الأحيان يستخدم بشكل كبير على منصة Apache Hadoop.

أحد أبرز المزايا التي كفلها تطبيق Spark، وإن كان يعد السبب لتطويره قدرته على التعامل ومعالجة وتحليل البيانات الآنية Real Time Data، دون الحاجة لإجراء عمليات الحفظ المسبق لإجراء التحليل عليها تباعًا، وذلك اعتمادًا على عمليات الحفظ المؤقت، تحسين عمليات الاستفسار المقدمة على البيانات، مهما اتسم حجم البيانات بالضخامة، وتباينت أنواعها.

أما على صعيد البنية التكوينية، فيتكون تطبيق Spark من مكونين رئيسين، هما:

- المكون الأول: وحدة التشغيل The Driver:

وهي الوحدة المسؤولة عن استقبال أكواد المستخدمين المطلوبة لإجراء عمليات تحليل البيانات، ثم تحويلها إلى مهام متعددة تُوزَّع على مختلف حاسبات مجموعة العمل Cluster.

- المكون الثاني: وحدة التنفيذ The Executors:

وهي الوحدة التي تتولى المهام التي أرسِلت من جانب وحدة التشغيل السابقة، وتتولى وحدة التنفيذ إجراد هذه المهام على مختلف Nodes، والتي حُدِدَت لتنفيذ المهام المعينة لها.

كذلك يكفل هذا التطبيق واجهة تعامل API تدعم كل من لغة Java, Python, Scala, R، كما تدعم إعادة استخدام الكود عبر العديد من البيانات، ويدعم التعامل مع البيانات في صورة دفعات، وفي صورة منفردة، ويدعم الاستعلامات التفاعلية، والتحليلات الآنية، ومعالجة الرسوم البيانية وتحليلها (White, 2015) (Amazon, 2022) (Apache Spark, 2022).

- تطبيق Pig & PigLatin:

طُوِّر هذا التطبيق من جانب شركة Yahoo، وقد تعددت الأغراض الرئيسة التي طُوِّر من أجلها هذا التطبيق، فمنها العمل على تبسيط استخدام وكتابة أوامر تطبيق MapReduce، لتحليل البيانات، وتقليل الوقت المستغرق في كتابة هذه الأوامر من جانب مَنْ يستخدمه، كذلك من أجل معالجة كافة أنواع البيانات (ولعل هذا هو السبب الرئيسي في تسمية هذا البرنامج بهذا الاسم؛ للدلالة على قدرته في معالجة وتحليل مختلف أنواع البيانات).

يعتمد هذا البرنامج على مكونين رئيسين في تحقيق أغراضه، الأول لغة كتابة الأوامر، والتي سميت باسم PigLatin. أما المكون الثاني، فيتمثل في بيئة تشغيل أو منصة عمل Runtime Environment تتسم في بكونها آنية التنفيذ لما يكتب لها من أوامر من جانب PigLatin.

تتسم لغة PigLatin بكونها لغة بسيطة للغاية لكتابة كل من أوامر تشغيل ومعالجة تحليل البيانات MapReduce، حيث يبدأ هذا التطبيق بتنفيذ عمليات التحليل من خلال الأمر Load، والذي يقوم بتحميل البيانات المراد معالجتها من ملفات النظام HDFS، ثم تُجرَى عمليات التحليل من خلال العديد من أوامر التشغيل والتحويل للبيانات، والتي تعمل بشكل كامل تحت كل من الوحدة الفرعية الأولى والثانية لتطبيق MapReduce، وفي النهاية يستخدم الأمر Dump، الذي يقوم بعرض نتائج عمليات التحليل على الشاشة أو تخزين ما تُوصِل إليه من نتائج داخل ملف ما.

تعتمد آليات تشغيل تطبيق Pig داخل برنامج Hadoop، على ثلاث آليات:

- الآلية الأولى: هي أن يُضَمَّن التطبيق داخل منصة البرنامج لكتابة الأوامر Hadoop Script.
- الآلية الثانية: هي أن يُضَمَّن باعتباره تطبيقًا Java Program داخل برنامج Hadoop.

- الآلية الثالثة: أن تُكتَب أوامره في صورة أوامر سطرية، هذه الأوامر التي تعرف باسم Pig Grunt.

- تطبيق HIVE:

يعد هذا التطبيق أحد أبرز التطبيقات الفوقية التي تبني على نظام Hadoop، وقد طُوِّر هذا التطبيق من جانب شركة Facebook، ويعد الغرض الرئيس من تطوير هذا التطبيق توفير لغة استعلام مهيكلة تتسم بالبساطة، ويمكن أن يتم تطبيقها على ملفات البيانات التي يشتمل عليها برنامج Hadoop، يعود السبب الرئيس وراء بساطة هذه اللغة في كونها بنيت على غرار لغة الاستعلام المهيكلة التقليدية SQL، وعلى هذا فإن المطورين المتآلفين مع لغة SQL يمكن لهم من خلال هذا التطبيق كتابة مجموعة الأوامر للتعامل مع برنامج Hadoop.

أما عن آلية التشغيل لاستعلامات هذا التطبيق داخل برنامج Hadoop، فيتم ذلك من خلال العديد من الطرق مثل:

- أن تُشَغَّل من خلال واجهة تعامل أمرية Command Line Interface.
- أو من خلال قاعدة بيانات مبنية بلغة الجافا متصلة بهذا التطبيق Java Database Connectivity.
- أو من خلال قاعدة بيانات مفتوحة مرتبطة بالتطبيق Open Database Connectivity (ODBC).

- تطبيق Flume:

يعرف Flume، بأنه أحد أبرز التطبيقات التي يقوم برنامج Hadoop، باستخدامها بهدف القيام بتجميع وتحويل ومعالجة كميات ضخمة من البيانات، وقد طُوِّر هذا التطبيق ليسمح لمستخدمي البرنامج بدفق البيانات من أحد مصادر البيانات File Source إلى ملفات نظام Hadoop HDFS.

ووفقًا للبنية المعمارية لهذا التطبيق فإن الكيانات التي يتعامل معها المبرمج أو المستخدم في إدخال البيانات من خلاله لبرنامج Hadoop تتمثل في ثلاثة كيانات أساسية، هي: المصادر Sources، وكيان المعالجة Decorators، وكيان الهدف Sink.

حيث يشير كيان المصادر إلى كافة مصادر البيانات التي تُدفَق داخل برنامج Hadoop لإجراء المعالجة عليها وتحليلها.

أما كيان الهدف Sink، فيشير إلى الهدف من عملية تحليل محددة. أما كيان

المعالجة Decorator، فيشير إلى الأوامر التي يتم من خلالها القيام بعمليات تحليل البيانات من صورة لأخرى.

- **تطبيق Zookeeper:**

يعد هذا التطبيق أحد التطبيقات مفتوحة المصدر التي بنيت وفقًا لتقنية Apache، والتي يعتمد عليها برنامج Hadoop في إدارته للخوادم التي تتكون منها مجموعة العمل Cluster، سواء كانت هذه الخوادم خوادم رئيسة NameNode، أو خوادم بيانات DataNode.

يتولى هذا التطبيق عمليات التسمية للخوادم، والهيكلة، وإنشاء المجموعات من الخوادم، وخدمات إدارة التعريفات للخوادم Configuration Management، وخدمات المزامنة بين الخوادم في تنفيذ مهام التطبيقات التي تتعلق بإجراء عمليات التحليل والاستعلام على البيانات التي تشملها هذه الخوادم.

يجدر الإشارة أولًا إلى أن عمليات التعامل والتفاعل مع هذا التطبيق تتم من خلال واجهة تعامل تكتب فيها الأوامر الخاصة به من جانب المستفيد بلغة Java أو بلغة C.

- **تطبيق HBase:**

يعد هذا التطبيق أحد أبرز التطبيقات التي يعتمد عليها برنامج Hadoop، في إدارة ما يشمله من ملفات بيانات HDFS، وينتمي هذا التطبيق إلى فئة التطبيقات مفتوحة المصدر الفوقية، لإدارة وهيكلة البيانات في صورة قواعد بيانات، وقد طور هذا التطبيق استنادًا على تطبيق BigTable الذي قامت شركة Google، بإنشائه لحفظ وإدارة الكميات الهائلة من البيانات في صورة جداول، تتمثل الوظيفة الرئيسة لهذا التطبيق في القيام بإدارة البيانات التي توجد في صورة أعمدة داخل جداول البيانات.

أما عن وجه الاختلاف الرئيس لهذا التطبيق عن نظم إدارة قواعد البيانات التقليدية فتتضح في منهجية عمله، والتي تتمثل في قيامه بإنشاء ما يعرف بعائلة الأعمدة Column Families، حيث إن هذا التطبيق يقوم بإنشاء مجموعة من الجداول تهيكل من خلالها البيانات في صورة مجموعة من الصفوف التي تعكس تسجيل الكيانات أو التسجيلات، وفي صورة مجموعة من الأعمدة التي تعكس خصائص هذه الكيانات أو التسجيلات (الحقول)، وكل جدول من هذه الجداول يقوم بتوفير بيان أو تحديد خاصية أساسية من ضمن خصائص هذه الكيانات تعرف باسم Primary Key، هذه الخاصية التي تعد بمثابة المفتاح الرئيس للتعامل مع ما يشمله هذا الجدول من كيانات، وتسجيلات.

أما جوهر تعامله ومعالجته للبيانات والكيانات التي تشتمل عليها جداول هذا التطبيق، فتتمثل في التعامل مع أعمدة هذه الجداول أي خصائصها، وليس الكيان نفسه (أي التعامل مع أصغر وحدة للكيان وهي خاصيته)، ليقوم من خلال هذا التعامل والمعالجة لهذه الخصائص بإنشاء ما يعرف بعائلة الأعمدة Column Families، والتي تعكس مجموعة من الحقول والخصائص المجمعة الخاصة بالتسجيلات والكيانات التي تشتمل عليها هذه الجداول.

- تطبيق Lucene:

يعد أحد أبرز التطبيقات مفتوحة المصدر لمعالجة البيانات، وقد طور من خلال تقنية Apache، ويعد من أكثر التطبيقات انتشارا، واستخدامًا في مجال عمليات البحث داخل النصوص أو البحث النصي داخل مجموعات البيانات والوثائق، بل أن العديد من محركات البحث الشهيرة والمشروعات التطبيقية تعتمد عليه بشكل أساسي في إجراء عمليات التكشيف للنصوص والبحث بداخلها واسترجاعها، بل يمكن القول إن العديد من مستخدمي الويب يقومون بالبحث من خلاله دون إدراك أنهم يعتمدون عليه في إجراء عمليات البحث عن النصوص.

يعمل هذا التطبيق على توفير آلية للتكشيف والبحث داخل النصوص الكاملة معتمدًا على لغة Java، ويجدر الإشارة في هذا الصدد إلى أن هذا التطبيق يكفل عملية تصدير النتائج للعديد من لغات البرمجة كلغة C++, Python, Perl، وغيرها من اللغات.

أما منهجية وآلية عمل هذا التطبيق، فتتسم بالبساطة الشديدة، حيث يعتمد على أن يعمد إلى مجموعة من النصوص، أو الوثائق ليقوم هذا التطبيق بهيكلتها من خلال تقسيم هذه البنية أو هذه الوثائق داخل مجموعة من الحقول ثم يقوم ببناء كشاف على هذه الحقول، هذا الكشاف الذي يعد بمثابة حجر الأساس لهذا التطبيق، والذي يشكل الأساس في إجراء عمليات البحث داخل هذه النصوص.

٥- المنصات التجارية لنظام Hadoop:

تعد أحد أبرز الجوانب السلبية التي اعتلت منصة Hadoop، أن المستخدمين لها قد ينتهي بهم الأمر بمنصة تتكون من إصدارات مختلفة من الوحدات، نظرًا لأن كل وحدة تكوينية داخل منصة Hadoop لها منحنى التطوير الخاص بها، ومن ثم بزوغ مشكلة عدم توافق الإصدارات للبني التكوينية داخل المنصة الواحدة Hadoop. على جانب آخر تبرز مشكلة أخرى، تتمثل في أن عملية الدمج بين الوحدات التكوينية غير المتوافقة داخل المنصة الواحدة قد تؤدي إلى العديد من المشكلات الأمنية للمنصة وما تشمله من بيانات.

ولمواجهة هذه المشكلات، قامت العديد من الشركات العاملة في مجال تقنيات البيانات

الضخمة، كشركة IBM، Cloudera، MapR، Hortonworks، بتطوير العديد من الإصدارات الخاصة بمنصة Hadoop وإتاحتها (في صورة جعلها منصة مفتوحة) في صورة تسهل على المستفيد النهائي التعامل مع منصة Hadoop، وبحيث تضمن التوافق وقضايا الحماية والأمن، والأداء لجميع الوحدات التكوينية المدمجة بشكل سليم، ويعد أبرز هذه المنصات التوزيعية القائمة على منصة Hadoop هي منصة Cloudera Hadoop (Oussous, 2018).

- منصة Cloudera CDH:

تعد إحدى أكثر المنصات التجارية استخدامًا، لما تكفله من قدرات تتعلق بعملية التنصيب والإدارة للبيانات بصورة مدعومة بشكل كامل من منصة Hadoop الرئيسة، وتوفر هذه المنصة العديد من المزايا التقنية كالإدارة المركزية لمجموعة العمل ودعم لغة الاستعلام المهيكلة، فضلًا عن التحكم الكامل في منصة Hadoop.

طور تطبيق Cloudera Distribution Hadoop (CDH) من جانب شركة Cloudera الأمريكية، ليصبح منصة عمل مفتوحة المصدر، تعمل على تسهيل التعامل مع منصة Apache Hadoop، وقد قدم هذا البرنامج في عام ٢٠٠٨، وقد صمم خصيصًا لتلبية متطلبات الشركات والمؤسسات العاملة في مجال تحليل البيانات، وفي مارس ٢٠٠٩، أعلنت شركة Cloudera عن توافر أول توزيع لها لمنصة Hadoop كحزمة واحدة، مشتملة في ذلك على أضافة العديد من الميزات المتعلقة بواجهة المستخدم والجوانب الأمنية والمراقبة وما إلى ذلك.

الأمر الذي كفل للمستخدم العادي، (غير المطور) القيام بتنصيب منصة Hadoop، والتعامل معه من خلال عدد من الأكواد البرمجية تتسم بالبساطة والسهولة، من خلال إصدار الشركة والذي عرف لاحقًا باسم Cloudera QuickStart VM.

تشتمل منصة Cloudera على مجموعة من المكونات الإضافية الداخلية، والتي تُضَاف للمكونات الأساسية لمنصة، Hadoop حيث تكفل هذه المكونات إدارة أفضل للمجموعة العمل وتنفيذ المهام.

ويعد أبرز المكونات الإضافية التي تشتمل عليها منصة Cloudera ما يلي:

- تطبيق Impala: ويتمثل في كونه محرك بحث يتمتع بنمط الاسترجاع في الوقت الحقيقي Real Time، ويعمل بشكل متوازٍ، ويعتمد في الأساس على لغة SQL للبحث عن البيانات في ملفات HDFS، ويعد IMPALA أسرع محرك للاستعلام في منصات توزيع Hadoop الأخرى، وهي منافس مباشر لشركة Spark.

- تطبيق Cloudera Manager: ويمثل هذا التطبيق وحدة التحكم في منصة Cloudera لإدارة وتنصيب مكونات Hadoop داخل مجموعة العمل Cluster.
- تطبيق Hue: ويمثل هذا التطبيق وحدة التحكم التي تتيح للمستخدم التفاعل مع البيانات وتشغيل البرامج النصية لمكونات Hadoop المختلفة الموجودة في مجموعة العمل Cluster (Azarmi, 2016).

وسيُعتَمد على تطبيق Cloudera Distribution Hadoop في القسم العملي من الدراسة.