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 الثاني)

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 II)

د. مؤمن سيد النشرتي

مدرس بقسم المكتبات والوثائق والمعلومات

كلية الآداب - جامعة القاهرة

Email: Navigator001@cu.edu.eg

ORCID: 0000-0003-4503-3378

المستخلص:

تأتي هذه الدراسة باعتبارها 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 لإحدى مصادر المعلومات داخل إحدى المكتبات الرقمية العربية، من خلال بناء ما يعرف بنموذج حقيبة الكلمات Bag of Words، حيث يعد هذا النموذج أحد 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 عمليات التحليل والمعالجة للبيانات. 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 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 الدالة:

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

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.

الجانب التطبيقي للدراســـة: إنشـــاء نموذج حقيبة الكلمات Bag of Words من خلال منصة Hadoop:

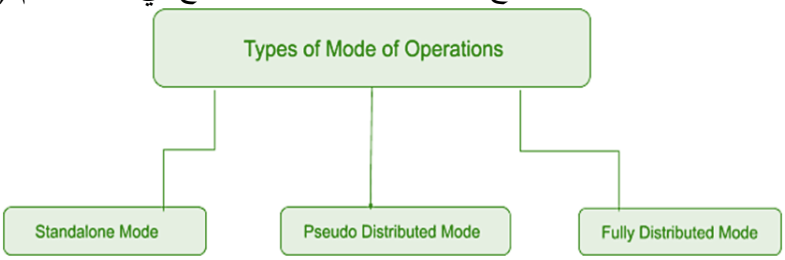
يتمثل الجانب التطبيقي داخل هذه الدراســة في إنشــاء نموذج حقيبة الكلمات Bag of Words، وذلك من خلال إجراء عملية تحليل لإحدى الوثائق الســردية النصــية التي تنتمي لفئة المنفردات من مصادر المعلومات (كتاب)، بحيث تعكس عملية التحليل المستهدفة تمثيل لكل من سـمة الضـخامة Volume للبيانات (حيث سـيُعالج ويُحلَّل ما يقرب من ٣٠٠ ألف كلمة داخل هذا الكتاب)، وسـمة التنوع Variety في ظل معالجة نمط البيانات غير المهيكلة Unstructured Data، التي تتمثل في المحتوى النصي للوثيقة.

تهدف الدراســة التطبيقية إلى التعرف على قدرة منصــة Cloudera Hadoop التقليدية على تحليل مصـــادر المعلومات المتاحة من خلال المكتبات الرقمية العربية، عن طريق

توظيف المنصة في إنشاء نموذج حقيبة الكلمات Bags of Words لإحدى الوثائق المنفردة (أحد الكتب) المتاحة من خلال إحدى المكتبات العربية، معتمدة في ذلك على المكونات الرئيسة الثلاث لمنصة Hadoop (وحدة HDFS، ووحدة MapReduce، ووحدة YARN).

أولًا: فئات أوضاع تنصيب منصة Hadoop Deployment Modes:

تدعم منصة Hadoop ثلاثة أوضاع لتعريفه وتنفيذه عند إجراء عملية تنصيبه على نظم التشغيل المختلفة، ومختلف أنواع أجهزة العمل كما هو موضح في الشكل رقم (٣١):



شكل رقم (٢٣) يوضح الأوضاع الثلاثة لتنصيب منصة Hadoop.

- **الوضع الأول: ويعرف بالوضع المستقل Standalone Mode:**

تعمل جميع الخدمات محليًا على جهاز واحد، وفي هذا الوضع تعمل منصة Hadoopبشكل أسرع عن الوضعين اللاحقين، حيث تكون ملفات HDFS مشابهة لنظام الملفات المتاح لنظام التشغيل (كنظام تشغيل على سبيل المثال MS. Windows)، مثل ملفات NTFS، وملفات FAT32، وعندما تعمل منصة Hadoop في هذا الوضع، فلا داعي لتكوين الملفات الخاصة بمجموعة عمل Cluster الخاصة بالمنصة.

ويستخدم هذا الوضع فقط لأغراض التطوير الصغيرة، كما يُستَخدم أيضًا لأغراض اكتشاف الأخطاء، ونادرًا ما يُستَخدم هذا الوضع في السياق التجاري وإدارة البيانات.

- **الوضع الثاني: وضع التوزيع الزائف Pseudo Distribution Mode:**

في هذا الوضع تعمل كافة الخدمات على الجهاز نفسه، ولكن السمة الرئيسية في هذا الوضع أنه يقوم بإجراء محاكاة كاملة لمجموعة العمل Cluster على جهاز واحد؛ لذا يُعَرف هذا الوضع باسم Single Node، ما يعني أن جميع العمليات داخل مجموعة العمل ستعمل بشكل مستقل مع بعضها بعضًا.

وتُشغَّل كافة الوحدات التكوينية لمنصة Hadoop كوحدة NameNode وDataNode وSecondary Name node وResource Manager و Node Manager بشكل يتسم بالاستقلال.

يعتمد هذا الوضع في ذلك على تنصيب منصة Hadoop على أحد التطبيقات الافتراضية Virtual Machine، كمنصة Java Virtual Machine JVM، والذي يكفل بيئة عمل خاصة بمنصة Hadoop؛ لذا يُعرَف هذا الوضع باسم الوضع المزيف، في ظل كون منصة Hadoop قد نُصِبَت على تطبيق افتراضي وليس نظام التشغيل الرئيس للخادم. أما الأغراض الرئيسية التي يُستخدَم لها هذا الوضع|، فتتمثل في أغراض التعليم والتدريب، والإنشاء داخل المعامل والمختبرات لتدريب الطلاب.

- **الوضع الثالث: الوضع الموزع بالكامل Fully Distributed Mode:**

يعد أكثر الأوضاع أهمية في تنصيب منصة Hadoop، فمن خلال هذا الوضع تُنصَّب المنصة لإجراء عمليات التحليل والإنتاج في السياقات العملية والفعلية؛ أي يستخدم هذا الوضع لأغراض الإنتاج وليس التدريب أو الاختبار.

وينطوي هذا الوضع على إنشاء كافة البنى والوحدات التكوينية لمنصة Hadoop، وليس بصورة افتراضية كالوضع السابق، وتُنشَأ مجموعة العمل الفعلية المكونة من Nodes، وربطها وتوزيع عمليات الحفظ والتحليل عليها للبيانات، حيث تعمل كل خدمة على جهاز منفصل (خادم مخصص)، وتستخدم في إعداد الإنتاج.

فبمجرد تحميل ملف إنشاء منصة Hadoop في صورته المضغوطة، فسيُثَبَّت داخل نظام التشغيل، ثم تُنشَأ مجموعة العمل المكونة من أجهزة الحاسب المتصلة، ثم يُستَخرج ملف المنصة على كافة أجهزة مجموعة العمل، ويُحدَّد أي منها يكون بمثابة DataNode، وأي منها سيكون NameNode، أي تحديد الأجهزة التي ستعمل باعتبارها خوادم رئيسة أو أي منها يعمل باعتباره خادم بيانات لتلك الخوادم الرئيسة في مجموعة العمل (White, 2015).

وفي هذه الدراسة ستُنصَّب منصة Hadoop في وضعيتها الثانية Pseudo Distribution Mode المخصصة لأغراض التدريب والاختبار لعملية الإنشاء للمنصة، معتمدة على نمط توزيع Cloudera QuickStart VM في إصدارته الخامسة CDH 5x (Cloudera, 2023).

ثانيًا: متطلبات تنصيب منصة Hadoop Requirements:

في ظل الاعتماد على الوضع الثاني من فئات تنصيب منصة Hadoop، وهو وضع Pseudo Distribution Mode، فإن ذلك سينطوي على إنشاء بيئة افتراضية Virtual Environment، وذلك اعتمادًا على أحد تطبيقات إنشاء البيئات الافتراضية، الأمر الذي

سيكفل أن تُنصَّب منصة Hadoop على مختلف نظم التشغيل دون قيود تتعلق بتوافقية نظام التشغيل مع المنصة.

وفي هذه الدراسة سيُعتَّمد على برنامج Oracle Virtual Box 7.0 لإنشاء البيئة الافتراضية لمنصة Hadoop، والمرجعية وراء اختيار هذا البرنامج في إجراء الجانب العملي من الدراسة، ما يوفره من إمكانات تتوافق مع عملية التنصيب الكاملة لمنصة Cloudera QuickStart VM، فضلًا عن ترشيح منصة Cloudera لهذا البرنامج في عملية تنصيبه لأغراض التدريب والاختبار لمنصة Hadoop قبيل إجراء عملية التنصيب الكاملة للمنصة.

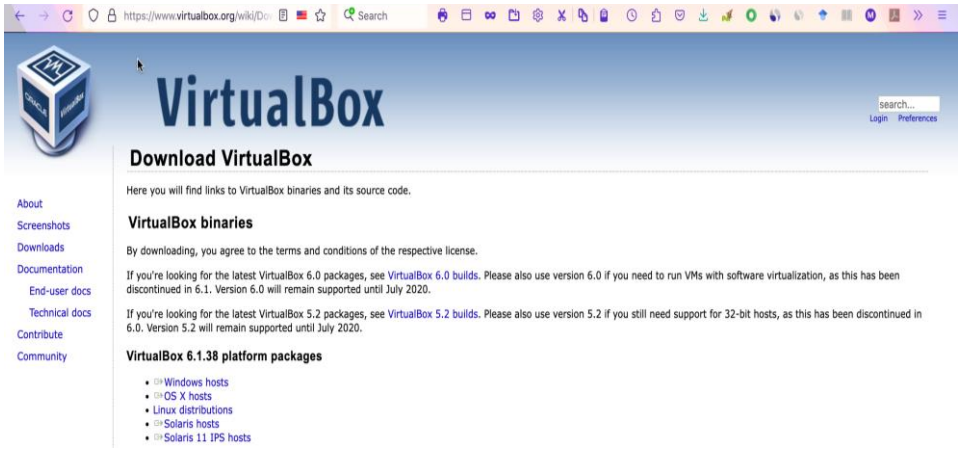
لتكون متطلبات التنصيب للمنصة وتكوين مجموعة العمل على البيئة الافتراضية متمثلة فيما يلي:

- تطبيق Oracle Virtual Box 7.0، لإنشاء البيئة الافتراضية الخاصة بالمنصة، ولبناء مجموعة العمل في صورة فردية Single Node Cluster.
- تطبيق Cloudera Distribution Hadoop QuickStart VM، الذي يكفل كافة مزايا منصة Hadoop لإجراء عمليات التحليل والمعالجة للبيانات الضخمة، ولكن بصورة تتسم بالسهولة والدعم للمستفيد النهائي في التعامل والمعالجة والاسترجاع للنتائج، وسيُعتَمد في هذه الإصدارة على نسخة CDH 5.13.
- متطلبات حاسوبية تتمثل فيما يلي:
  - نظام تشغيل Windows 64 bit، أو Mac OS (الأجهزة ما قبل إصدار معالج M1) أو نظام Unix.
  - ذاكرة RAM لا تقل سعتها عن ٤ جيجابايت.
  - مساحة تخزين فارغة على القرص الصلب لا تقل عن ٢٠ جيجابايت.

ثالثًا: عملية التنصيب لمنصة Cloudera Distribution Hadoop QuickStart VM 5.13:

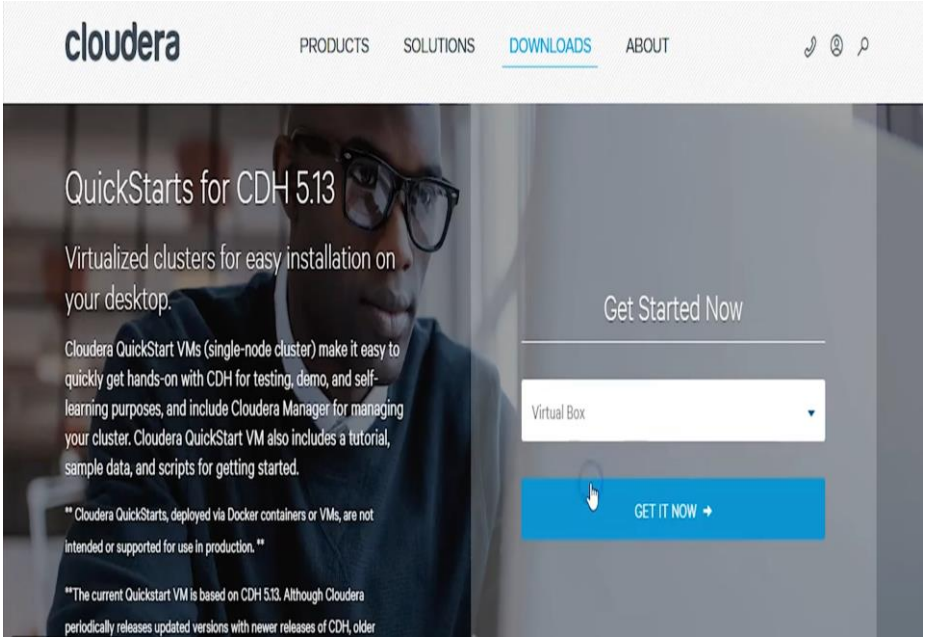
تتم عملية التنصيب تباعًا وفقًا للخطوات التالية:

- يُحمَّل وينصَّب تطبيق Oracle Virtual Box 7.0 في آخر إصدارة له وقت إعداد الدراسة، وهي الإصدارة السابعة 0.7، مع التأكد من أن الإصدارة تتوافق مع نظام التشغيل الخاص بالخادم الذي سيجرى عليه عملية التنصيب من خلال موقع التطبيق نفسه (كما هو موضح في الشكل رقم ٢٤).



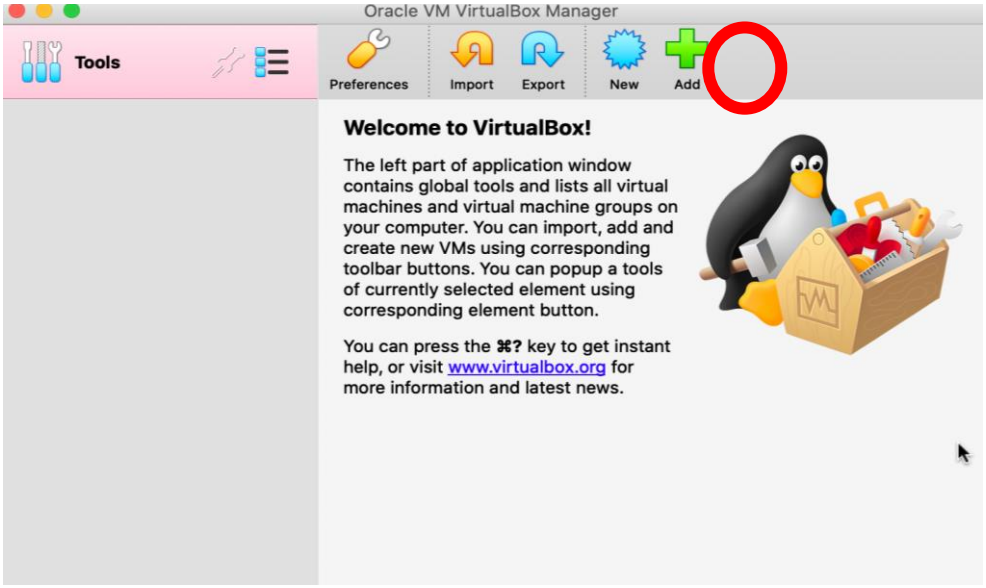
شكل رقم (٢٤) يوضح شاشة التحميل لتطبيق VirtualBox.

- تحميل منصــة Hadoop Cloudera QuickStart VM 5.13 في إصــدارته المتاحة من خلال الموقع المتاح على الرابط الخاص به، كما هو موضح في الشكل رقم (٢٥).



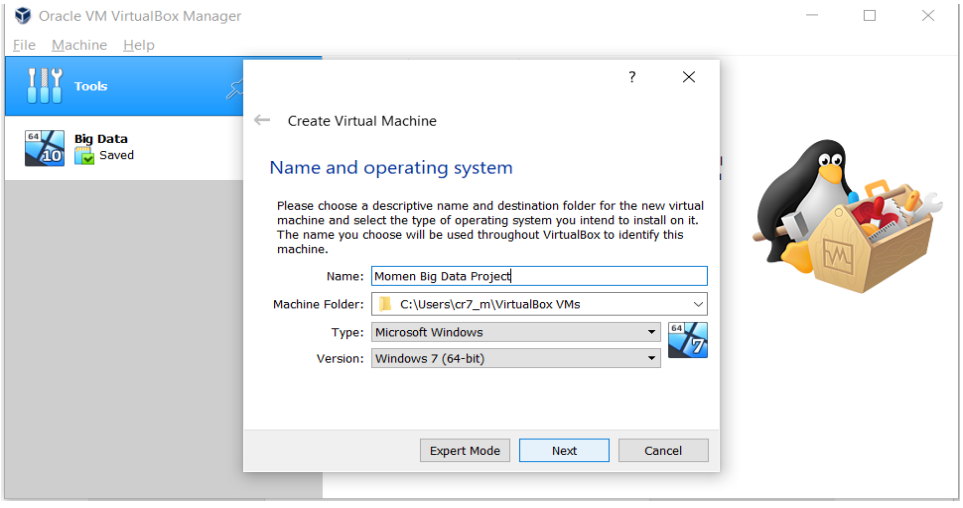
شكل رقم (٢٥) يوضح موقع لتحميل منصة Hadoop Cloudera QuickStart VM.

- تنصيب تطبيق Oracle Virtual Box على الحاسب، وفتح واجهة الاستخدام الخاصة به، لتشغيل منصة Hadoop من خلاله.



شكل رقم (٢٦) يوضح واجهة الاستخدام الرئيسة لتطبيق Oracle VirtualBox.

- يُنقَر على أيقونة New، التي تعني قيام المستخدم بإنشاء نظام أو منصة تشغيلية بصورة افتراضية، لتُفتَح واجهة تعامل لتحديد العناصر التالية:



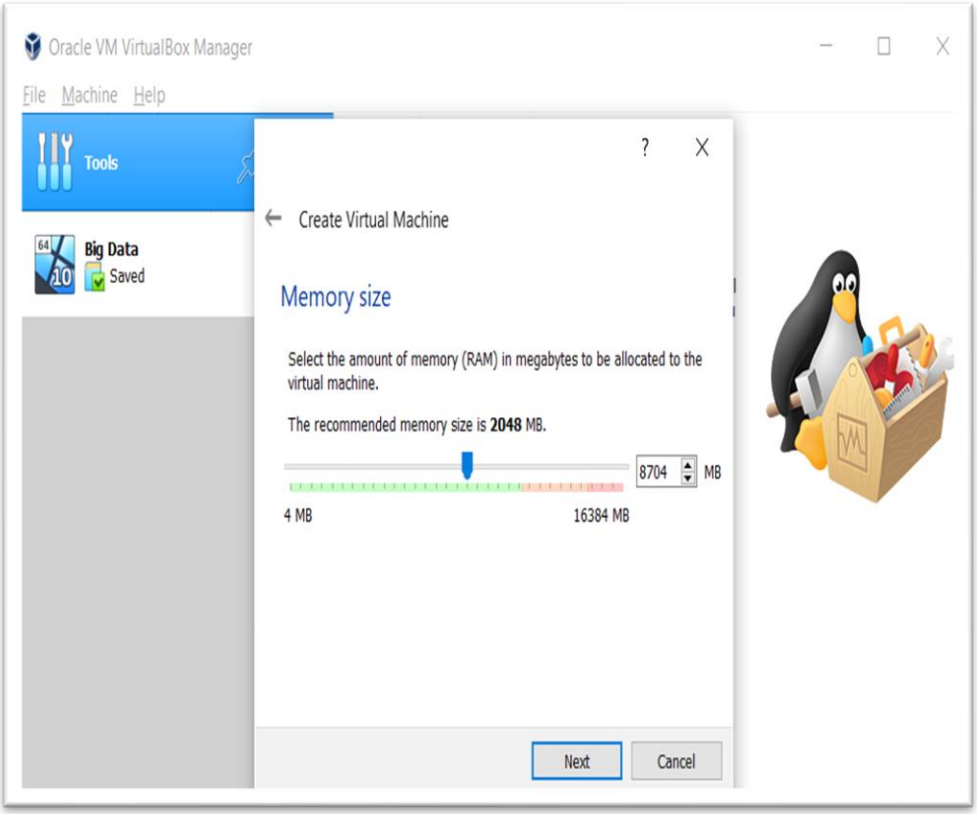
شكل رقم (٢٧) يوضح عملية التنصيب وشاشة إنشاء المنصة.

- اسم المنصة Name: تُختَار التسمية كيفما يشاء المستخدم.
- موقع المنصة Machine Folder: يقوم البرنامج بتحديد موقع المنصة التي ستُنشَأ داخل ملفات الحاسب الخاص بالمستخدم.
- نوع نظام التشغيل المراد تنصيبه Type: يُحدَّد فيه نظام التشغيل الافتراضي

الذي ستُنصَّب عليه منصة Hadoop، وفي هذه الحالة سيكون Microsoft Windows.

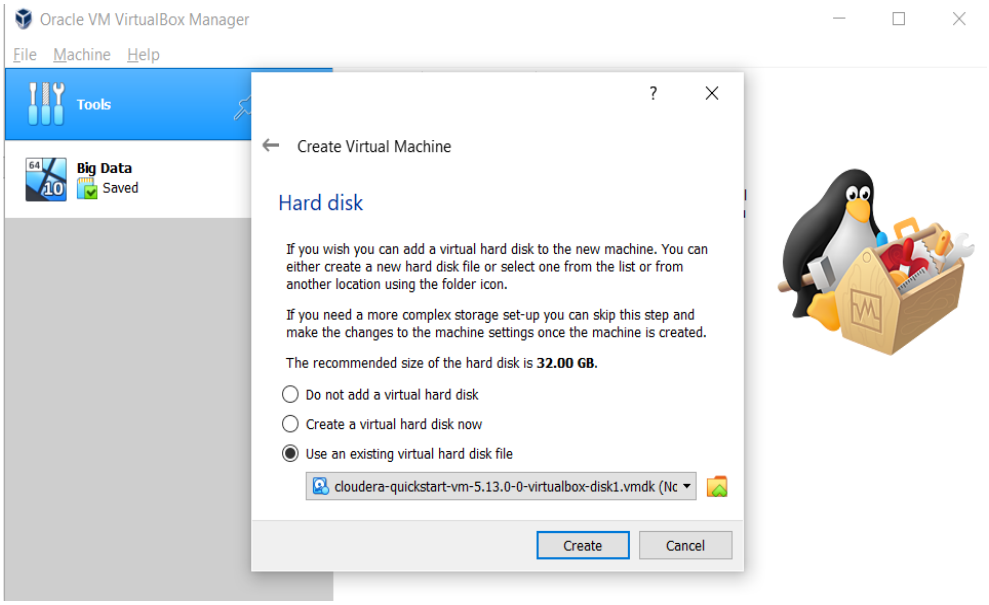
- الإصدارة Version: يُحدَّد فيها نوع إصدارة نظام التشغيل المراد تنصيبه، وتُختار إصدارة Windows 10 (64 bit).

- تُحدَّد مساحة الذاكرة، لتتناسب مع حجم المنصة المراد تنصيبها لتكون المساحة المطلوبة أكبر من سعة ٤ جيجابايت، كما هو موضح في الشكل رقم (٢٨).



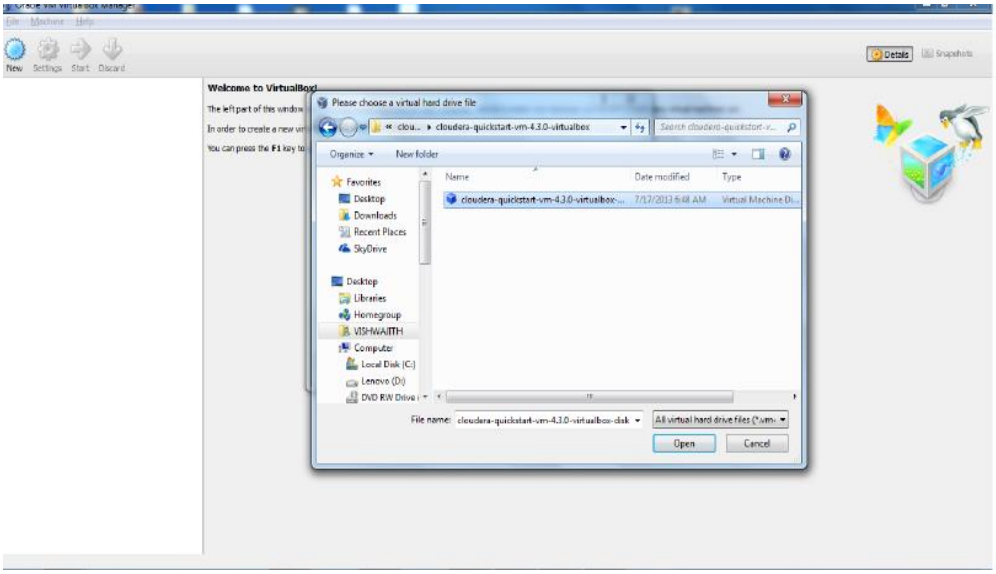
شكل رقم (٢٨) يوضح عملية تحديد مساحة التشغيل للمنصة على تطبيق VirtualBox.

- عقب النقر على علامة Next، تُفتَح شاشة جديدة لاختيار المنصة المراد تشغيلها، وفي هذه الحالة يُختارالخيار المعنون باسم "Use an existing virtual hard Disk file"، كما هو موضح في الشكل رقم (٢٩).



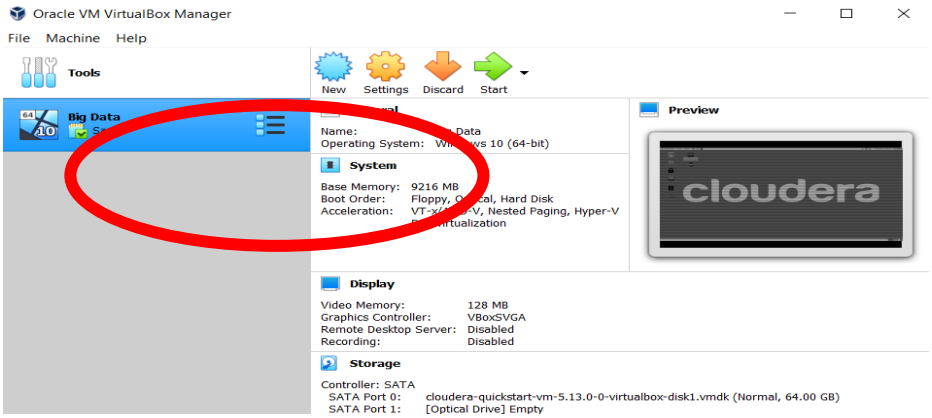
شكل رقم (٢٩) يوضح اختيار منصة Cloudera باعتبارها نظام ملفات

- لفتح شـــاشـــة يُختَار من خلالها المنصـــة المراد تنصـــيبها واختيار ملف Cloudera QuickStart VM 5.13، المنتهي بامتداد vmdk، كما هو موضح في الشكل رقم(٣٠)



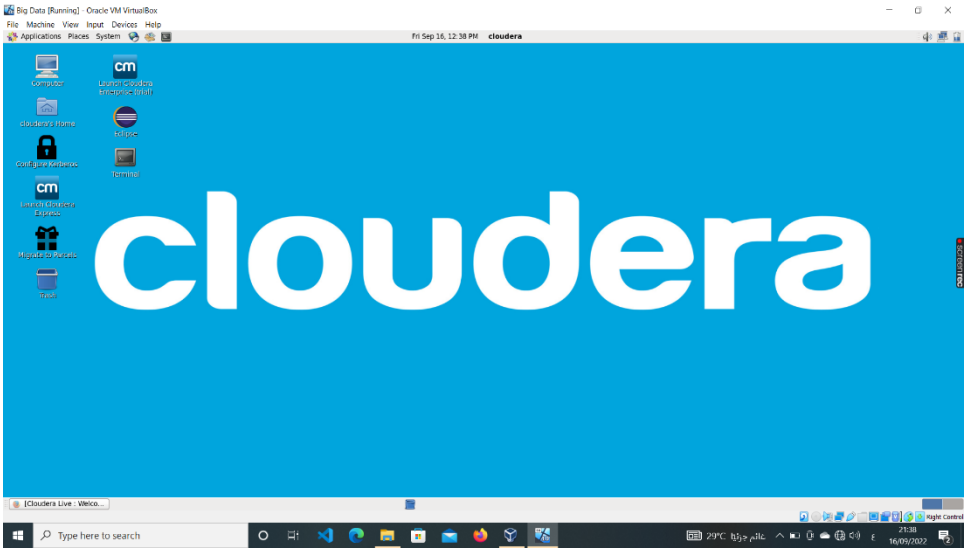
شكل رقم ٣٠ يوضح عملية اختيار المنصة لتفعيلها من خلال تطبيق VM.

- لتحميل المنصـــة بشـــكل كامل على تطبيق VirtualBox، وظهوره في القائمة الجانبية من واجهة تعامل البرنامج، كما هو موضح في الشكل رقم (٣١).



شكل رقم (٣١) يوضح تنصيب منصة Hadoop تحت مسمى Big Data، داخل تطبيق VirtualBox.

- عقب النقر على الأيقونة المعنونة باسم Big Data، تُحمَّل المنصة بالكامل باعتبارها نظام تشغيل، وتُنشأ مجموعة العمل في صورة افتراضية.
- للتعامل لاحقًا مع الأيقونة المعنونة باسم "Terminal"، التي من خلالها ستُكتَب الأوامر لإجراء عملية تحميل البيانات للمنصة وتحليلها من خلال وحدة MapReduce.



شكل رقم (٣٢) يوضح نجاح بناء منصة Cloudera Hadoop وبناء مجموعة العمل luster، وواجهة التعامل للبدء في تحميل البيانات والتعامل معها.

رابعًا: إنشـــاء نموذج حقيبة الكلمات Bag of Words من خلال منصـــة Hadoop Cloudera QuickStart VM 5.13 :

- **الهدف من التطبيق:**

يتمثل الهدف الرئيس من جراء استخدام منصة Hadoop Cloudera QuickStart VM 5.13 في قياس قدرة المنصـــة على إنشـــاء نموذج حقيبة الكلمات Bag of Words، وذلك باســتخدام المنصـــة في إجراء المعالجة لفئة البيانات غير الُمهيكَلَة Unstructured Data، والمتمثلة في النصـــوص والجمل والكلمات الواردة في أحد مصادر المعلومات الثانوية أو المنفردات (كتاب)، لإنشاء نموذج حقيبة الكلمات.

يعد نموذج حقيبة الكلمات Bag of Words إحدى تقنيات معالجة اللغة الطبيعة Natural Languages Processing من خلال الذكاء الاصـــطناعي، فهو يعتبر الحلقة الأولى لتجهيز البيانات النصـــية Data Preparation والخطوة الرئيســـة أيضًـــا قبيل توظيف الخوارزميات المختلفة لمعالجة وإنشـــاء نماذج المعالجة والتحليل والتنبؤ للبيانات النصـــية كخوارزمية TF-IDF، وخوارزمية SVM، وخوارزمية Naïve Bayes |، وغيرها.

يمكن النظر لنموذج حقيبة الكلمات باعتباره عملية هيكلية منطقية لكافة الكلمات الواردة في النصـوص أو الوثائق، حيث تتجلى هيكلية هذا النموذج في حصـر كافة الكلمات الواردة في النص وعدد مرات تكرار ظهورها داخل النص، في صـورة أقرب لجدول يتكون من عمودين، حيث يشـــتمل العمود الأول على الكلمات الواردة في النص، والعمود الثاني على عدد مرات ظهور الكلمات في الوثيقة.

تُعَدّ الغاية من إنشـــاء نموذج حقيبة الكلمات في عملية معالجة اللغة الطبيعية من جانب الحاســـب، تمييز وتحديد كافة الكلمات الواردة للحاســـب تمهيدًا لتشـــغيل الخوارزميات المختلفة على هذه الكلمات، ولإكســـاب الحاســـب والتطبيقات المختلفة القدرة على أن تميز كل كلمة على حدةٍ.

شكل رقم (٣٣) يوضح تصور لنموذج حقيبة الكلمات Bag of Words (Paila, 2020).

- **طبيعة البيانات المستهدفة:**

تستهدف عملية التطبيق والاستخدام للمنصة على إجراء عملية تحليل للوثيقة التي وقع عليه الاختيار، والتي تنتمي بطبيعة الحال لفئة البيانات غير المهيكلة، وقد اعتمدت الدراسة على اختيارها للوثيقة (الملف النصي) المراد تحليلها من خلال المنصة على العديد من العناصر وهي:

- أن تكون الوثيقة عملا منفردا يخضع للتراخيص وإعادة الاستخدام والإتاحة، ولا يخضع للمسؤولية أو قيود المتعلقة بالنشر.
- أن يكون الملف النصي متاح من خلال أحد مشروعات المكتبات الرقمية الرسمية والمجانية ومفتوحة الإتاحة، بما يكفل أن تتسم الوثيقة بالموثوقية والاعتمادية والمصداقية، في منشئها أو في عملية التحويل الرقمي لها.
- أن تتسم الوثيقة بالعمومية دون التخصصية في أحد المجالات لتيسير إدراك المخرجات من جانب القارئ لهذه الدراسة.
- أن تكون الوثيقة معبرا عن الهوية الإسلامية العربية (مع إمكانية صدورها بلغات أخرى غير العربية).
- يفضل أن يكون المصدر المختار قد حول رقميا من صورة مطبوعة (وليس الكتروني المنشأ)، للكشف عما قد يبرز من متغيرات تتعلق بعملية التحليل والتعرف على طبيعة تعامل منصة Hadoop مع المحتوى المرقمن.

ولهذا تم اختيار مشروع المكتبة الرقمية لمركز المعرفة الرقمي Digital Knowledge Hub التابع لمؤسسة محمد بن راشد آل مكتوم للمعرفة[1]، كأحد المكتبات الرقمية الرسمية العربية المعترف بها عالميا والتي لا تنتمي لمكتبات الظل غير القانونية، فضلا عما يتمتع به هذا المشروع من شرعية ورسمية وريادة في النشر والتحويل الرقمي للمنشورات العربية والعالمية، على مستوى الوطن العربي.

يعد المركز أحد أكبر المراكز الرقمية الحاضنة للمحتوى العربي المتاح مجانا، يشتمل المركز على ٣.٥ ملايين مادة رقمية (وفقا لما هو معلن على موقعه) حيث تغطي جميع جوانب المعرفة الموثقة من خلال مصادر المعلومات العربية والأجنبية، وتعد المرجعية وراء اختيارها كمكتبة رقمية يستقى منها الوثيقة المراد إنشاء نموذج الكلمات لها الأسباب التالية:

(1) https://ddl.ae/

١. الإتاحة المجانية لمصادر المعلومات دون التقيد بشروط أو محددات جغرافية أو نوعية، وذلك بصورة قانونية وشرعية تكفل حقوق الملكية الفكرية للأعمال الفنية والفكرية والعلمية.
٢. لا يخضع هذا المشروع لقيود الإتاحة الذي تفرضه بعض المشروعات والمكتبات الرقمية الأخرى، كالقيود الجغرافية في استخدام المكتبة الرقمية، وإتاحة ما تشمله من مصادر داخل حدود الدولة اعتمادًا على معرفات الحاسبات (مثل مشروع بنك المعرفة المصري، أو المكتبة الرقمية السعودية).
٣. كذلك لا يخضع هذا المشروع لقيد مالي، حيث تتاح غالبية المصادر بداخله مجانًا، دون قيود تتعلق بانتماء الباحث لمؤسسة محددة كالمكتبات الرقمية الخاصة بالجامعات.
٤. حجم ما تشمله هذه المكتبة من مصادر معلومات عربية، يفوق ٣.٥ مليون مصدر معلومات متنوع بين المنفردات والدوريات، والمخطوطات، والمصادر غير التقليدية كالخرائط والصور والوسائط المتعددة.
٥. توفير الإتاحة لمصادر المعلومات بأكثر من صيغة للملفات، مما يكفل كفاءة عملية الإتاحة واتساعها.
٦. تركيزه الأساسي على الهوية العربية في سياسة اقتنائه وتجميعه لمصادر المعلومات العربية.
٧. كافة المصادر التي يشتمل عليها تخضع للوصول الحر أو تراخيص المشاع الإبداعي.
٨. تبعيته الإدارية لواحدة من أقوى مؤسسات المعرفة على مستوى العالم، وهي مؤسسة محمد بن راشد آل مكتوم للمعرفة.

وقد وقع اختيار الدراسة لإجراء عملية التحليل على أحد مصادر المعلومات النصية المُرقمَنَة داخل المشروع، وهو كتاب "The Life of Mohammad, the Prophet of Allah" (Dinet, E, 1918)، الذي يعتبر أحد أبرز الأعمال الفكرية الإسلامية في المكتبات العربية.

- **طبيعة عملية التحليل المستهدفة:**

تنتمي عملية التحليل المنفذة في الجانب التطبيقي للدراسة إلى منهجيات التحليل الوصفي Descriptive Analysis، والتي تهدف إلى إيضاح وتقديم إحصائيات

وصفية تتعلق بالبيانات المستهدفة، ونظرًا لطبيعة كون مصادر المعلومات الموجودة بمستودعات المكتبات الرقمية، تنتمي إلى فئة البيانات الكيفية Qualitative Data، فمنهجية التحليل الوصفي هي الأنسب لتنفيذ عملية التحليل.

تعد عملية التحليل الرئيسة المطلوب إجراؤها هي حساب تردد ظهور أو تكرار كل كلمة داخل هذه الوثيقة، 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.
Mohamed	333
Prophet	370
Allah	474
Mecca	150
Muslims	70

جدول رقم (٤) يوضح مخرجات نموذج حقيبة الكلمات Bag of Words.

يعد بناء نموذج حقيبة الكلمات Bag of Words إحدى الحلقات الرئيسة التي تسبق عملية حساب أوزان المصطلحات الخاصة بالوثيقة، والتي تتم من خلال الخوارزميات الخاصة بأوزان المصطلحات مثل خوارزمية TF/IDF، وذلك قبيل إجراء إنشاء قوائم الكلمات المفتاحية Go Lists، والكلمات المستبعدة Stop Lists.

تمثل عملية إحصاء تردد ظهور الكلمات أحد الأساليب والمقاييس الإحصائية، التي تعكس مدى أهمية مصطلح ما بالنسبة لوثيقة معينة أو ضمن مجموعة من الوثائق، وتعتمد خوارزمية TF على نموذج حقيبة الكلمات Bag of Words في ذلك.

تعد خوارزمية TF/IDF إحدى أبرز الخوارزميات الإحصائية الموظفة في العديد من المجالات كمجال البحث والاسترجاع للمعلومات Information Retrieval، وفي مجال ترتيب نتائج محركات البحث الويبية Web Search Engines، والتنقيب في البيانات Data Mining، والنمذجة Modeling (Rajarama,2011).

أوضح (Beel,2016) من خلال دراسة مسحية حول أنظمة الاقتراح في المكتبات الرقمية، أن مقياس TF/IDF يعد أكثر الخوارزميات الإحصائية استخدامًا في تحديد واقتراح مصادر المعلومات للمستفيدين في المكتبات الرقمية بنسبة بلغت ٨٥٪ داخل مكتبات الدراسة.

وفي سياق إجراء الجانب التطبيقي لهذه الدراسة وسمي نموذج الكلمات الخاص بهذه

الدراسة بمسمى The Word Count، بهدف إظهار القيم الرقمية التي تعكس تردد ظهور كل مصطلح داخل مصدر المعلومات بالكامل، تمهيدًا لتوظيف هذا النموذج داخل خوارزمية TF*IDF.

- **استيراد الملف النصي وتضمينه داخل المنصة:**

تنطوي عملية التحليل على أن تُجرَى عملية دفق Copy لمصدر المعلومات داخل المنصة، ثم إجراء عملية حفظ لهذا الوثيقة داخل نظام Cloudera Hadoop، وذلك تمهيدًا لنقل المحتوى النصي الخاص بالوثيقة للوحدة التكوينية الأولى للمنصة، التي تتمثل في وحدة HDFS، والتي ستتولى تباعًا إجراء تقسيم النص إلى كتل من البيانات؛ بحيث تشتمل كل كتلة على جزء من نص الوثيقة بحيث لا يتجاوز هذا الجزء مساحة ١٢٨ ميجابايت، ثم يُرَقَّم ويُعنوَن كل جزء من هذه الأجزاء، وتُعطَى البيانات الوصفية الخاصة بها (الميتاداتا) لتمييزها عن غيرها من الوحدات المناظرة لها.

بناء النظام الفرعي لتحليل البيانات النصية داخل المنصة.

تأتي منصة Cloudera Hadoop بقابلية لبناء العديد من وظائف التحليل والمعالجة، والتي من الممكن أن تكون غير مدرجة في الإصدارة الرئيسة للمنصة، وذلك في ظل كونها منصة مفتوحة المصدر، تكفل أن يقوم المبرمج والمحلل ببناء الوظائف المراد تشغيلها على البيانات.

وفي هذه السياق، يلتزم الباحث بأن يقوم ببناء نموذج حقيبة الكلمات Bag of Words كوظيفة التحليل المراد تشغيلها على الوثيقة، لحساب تكرار ظهور كل كلمة داخل الوثيقة Document Word Count، وفقًا للخطوات التالية.

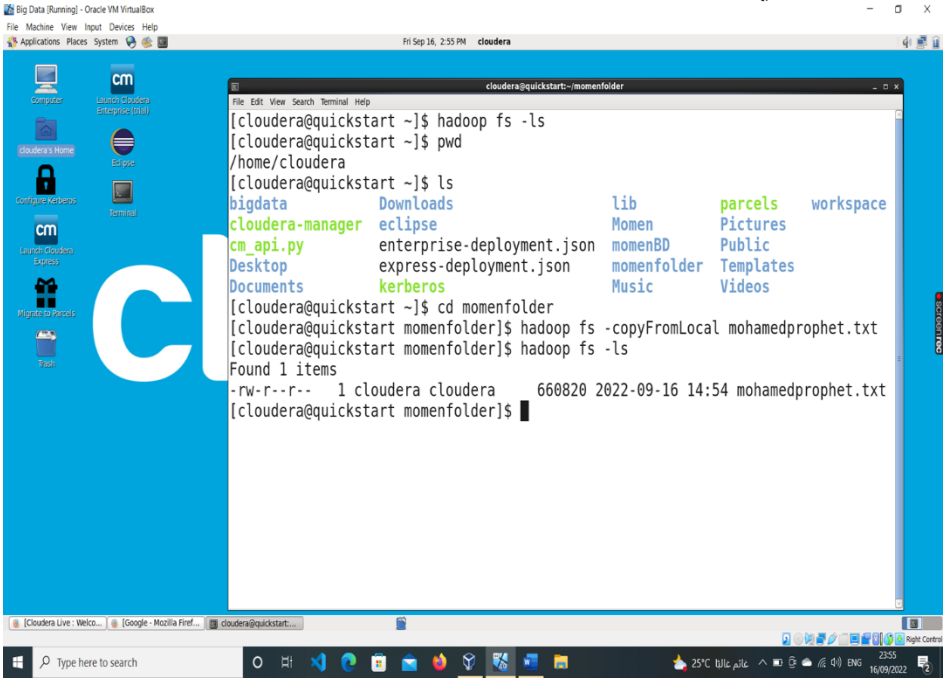
الخطوات:

١. يتم الحصول على الوثيقة المحددة سلفًا وتحويلها إلى صورة نصية مجردة Plain Text دون أي تنسيقات (كتنسيقات برنامج MS. Word) أو ملفات تنسيقية (كملف CSS).

٢. يُنقَر على أيقونة Terminal، التي توجد في واجهة التعامل لمنصة Cloudera Hadoop، والتي تكفل إعطاء الأوامر البرمجية في صورة Commands line بلغة Java للمنصة، ليتم من خلالها استيراد ونقل مصدر المعلومات النصي داخل المنصة، وذلك من خلال كتابة الأمر التالي:

[Cloudera@QuickStart – Momenfolder] $ Hadoop fs – copyFromLocal mohamedprophet.txt

يتولى هذا الأمر البرمجي Command استيراد الوثيقة، التي اختصرت تسميتها إلى Mohamed Prophet، داخل ملفات HDFS داخل منصـــة Hadoop، كما هو موضح في الشكل رقم (٤١).



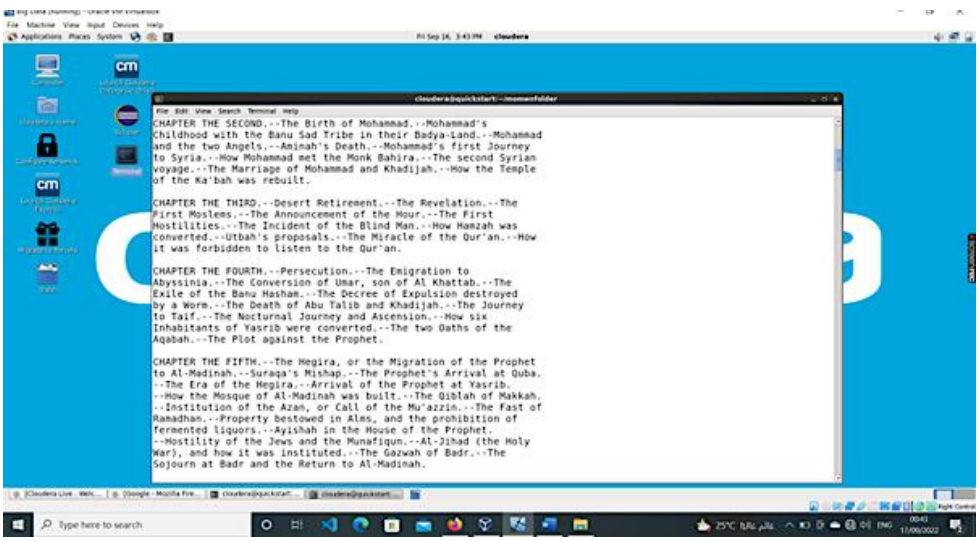
شكل رقم (٣٤) يوضح عملية نقل الوثيقة (مصدر المعلومات) داخل منصة Hadoop.

٣. عقب عملية استيراد الوثيقة داخل منصـة Hadoop، يتم عملية التأكد من طبيعة المحتوى الخاص بالوثيقة كونه نصيًّا لا يشتمل على صور أو إيضاحيات.

٤. ثم تُختَبر قدرة المنصة على قراءة النص الخاص بالوثيقة بصورة كاملة، وذلك من خلال الأمر:

[Cloudera@QuickStart – Momenfolder] $ Hadoop fs – Cat

حيث يُســـتَعرض من خلال هذا الأمر المحتوى بشـــكل كامل للوثيقة، كما هو موضح في الشكل رقم (٣٥).



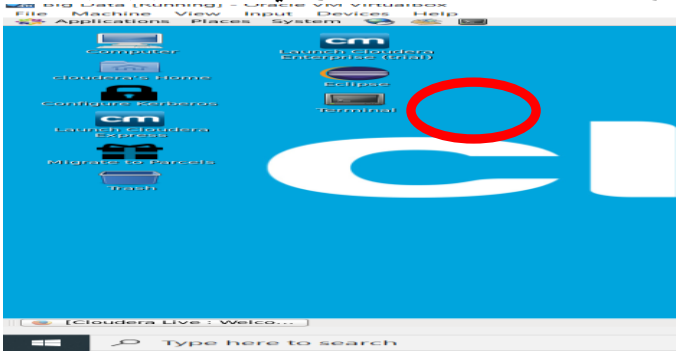
شكل رقم (٣٥) يوضح استعراض المحتوى النصي للوثيقة المراد إنشاء نموذج حقيبة الكلمات الخاص بها.

- **تحديد عملية التحليل المراد تنفيذها من المنصة:**

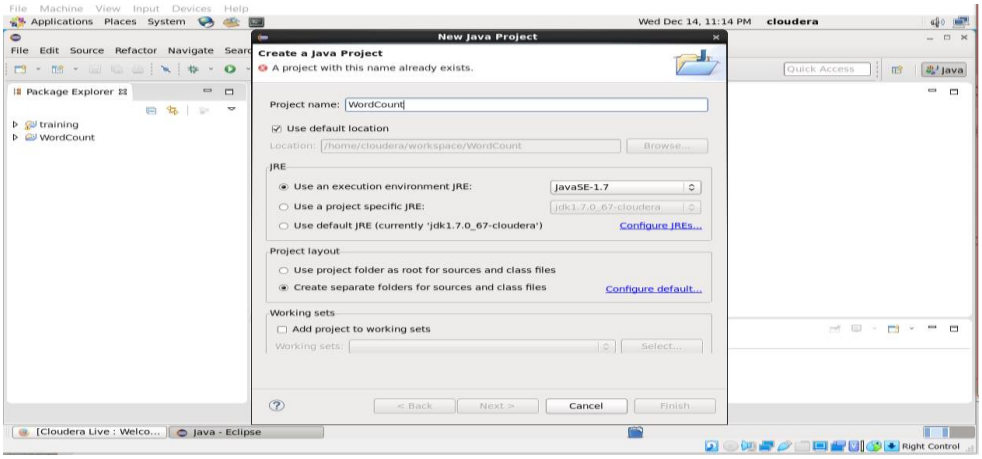
عقب عملية تضمين الوثيقة المراد تحليلها داخل المنصة واستعراض محتواها والتأكد من قدرة المنصة على التعامل مع المحتوى النصي الخاص بها، تأتي مرحلة إنشاء نموذج حقيبة الكلمات المراد تنفيذه على الوثيقة، والذي وسم بمسمى Word Count.

تتم عملية إنشاء نموذج حقيبة الكلمات من خلال الوحدة الفرعية لمنصة Hadoop Cloudera؛ حيث تُبرمَج الوحدة الفرعية المسؤولة عن ذلك، وهي وحدة MapReduce، لتقوم إنشاء هذا النموذج، وتشتمل عمليات بناء الكود البرمجي لوحدة MapReduce على كل من الخطوات التالية:

١. النقر على الأيقونة المعنونة باسم Eclipse، لتفتح الشاشة المسؤولة عن برمجة الوحدة لإنشاء نموذج حقيبة الكلمات كما هو موضح في الشكل رقم (٣٦).



شكل رقم (٣٦) يوضح أيقونة Eclipse التي سيتم من خلالها البرمجة لإجراء عملية التحليل.

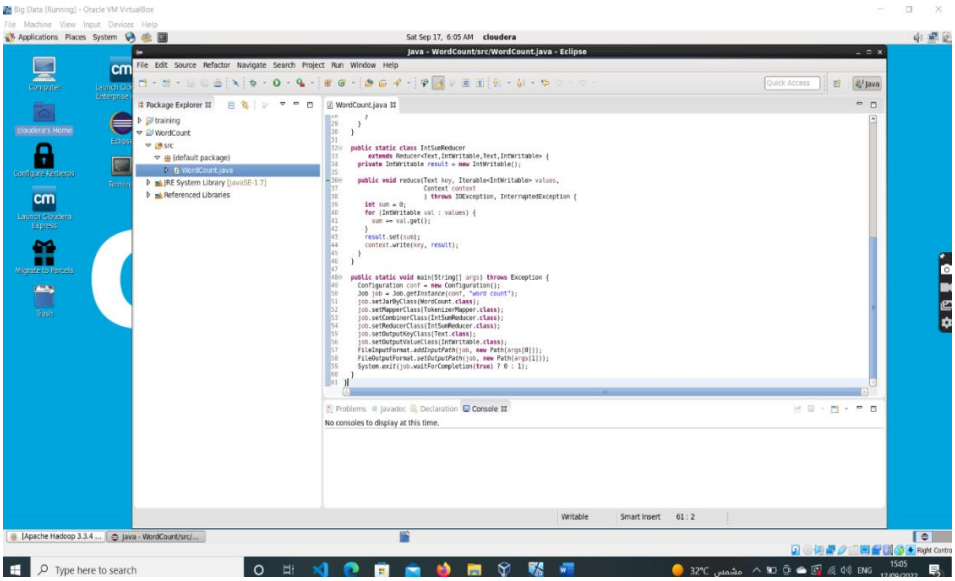


شكل رقم (٣٧) يوضح تسمية نموذج حقيبة الكلمات المسمى بـ WordCount لإحصاء تكرار ظهور الكلمات في الوثيقة.

٢. عقب ذلك تُفتَح شاشة كتابة الكود البرمجي الخاص بإنشاء نموذج حقيبة الكلمات داخل وحدة MapReduce، التي ستجرى من خلالها عملية حصر وإحصاء تكرار ورود المصطلحات الموجودة في الوثيقة، ثم كتابة الكود البرمجي الخاص بإنشاء نموذج حقيبة الكلمات والموسوم باسم نموذج WordCount وفقًا للشكل التالي:

```
١. import java.io.IOException;
٢. import java.util.StringTokenizer;
٣. import org.apache.hadoop.conf.Configuration;
٤. import org.apache.hadoop.fs.Path;
٥. import org.apache.hadoop.io.IntWritable;
٦. import org.apache.hadoop.io.Text;
٧. import org.apache.hadoop.mapreduce.Job;
٨. import org.apache.hadoop.mapreduce.Mapper;
٩. import org.apache.hadoop.mapreduce.Reducer;
١٠. import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
١١. import
org.apache.hadoop.mapreduce.lib.output.FileOutputForm
at;
١٢.
١٣. public class WordCount {
١٤.
١٥. public static class TokenizerMapper
١٦. extends Mapper<Object, Text, Text, IntWritable>{
١٧.
١٨. private final static IntWritable one = new
IntWritable(1);
١٩. private Text word = new Text();
٢٠.
٢١. public void map(Object key, Text value, Context
```

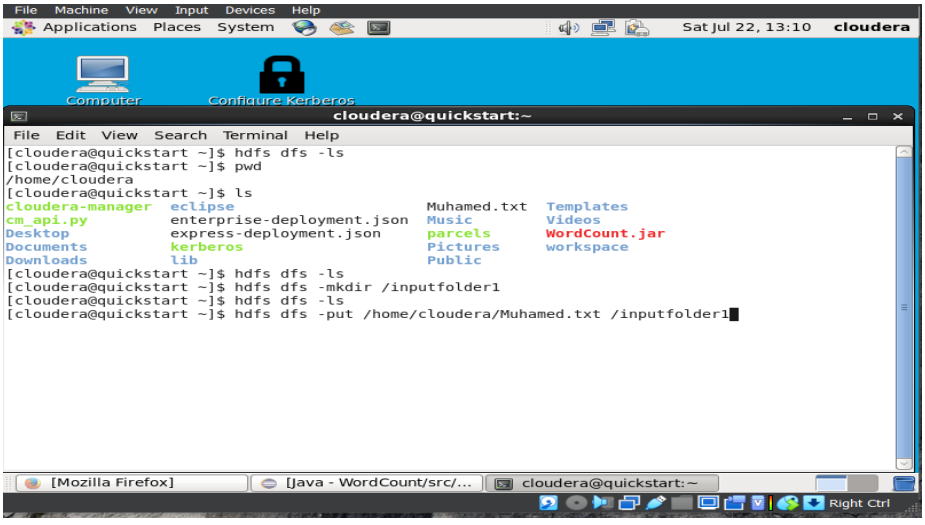
```
                                                context
) throws IOException, InterruptedException {
StringTokenizer itr = new StringTokenizer(value.toString());
while (itr.hasMoreTokens()) {
word.set(itr.nextToken());
context.write(word, one);
}
}
}

public static class IntSumReducer
extends Reducer<Text,IntWritable,Text,IntWritable>
{
private IntWritable result = new IntWritable();

public void reduce(Text key, Iterable<IntWritable>
values,
Context context
) throws IOException, InterruptedException
{
int sum = 0;
for (IntWritable val : values) {
sum += val.get();
}
result.set(sum);
context.write(key, result);
}
}

public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Job job = Job.getInstance(conf, "word count");
job.setJarByClass(WordCount.class);
job.setMapperClass(TokenizerMapper.class);
job.setCombinerClass(IntSumReducer.class);
job.setReducerClass(IntSumReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new
Path(args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
```



شكل رقم (٣٨) يوضح كتابة الكود البرمجي لإنشاء نموذج حقيبة الكلمات داخل منصة Hadoop لإجراء عملية التحليل على الوثيقة المختارة.

٣. ثم تُضـــمَّن الوثيقة داخل منصـــة Hadoop لتطبيق نموذج حقيبة الكلمات عليها (WordCount)، كما هو موضـــح في الشـــكل رقم (٣٨)، وذلك من خلال كتابة الكود التالي:

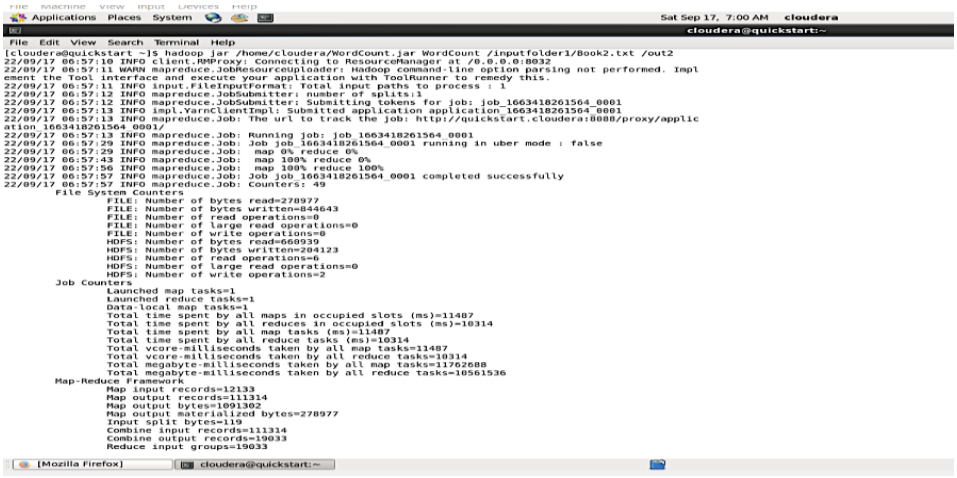
Hdfs dfs Put/Home/Cloudera/Muhamed.txt/ Inputfolder1



شكل رقم (٣٩) يوضح تضمين الوثيقة داخل منصة Hadoop لإجراء عملية التحليل.

عقب ذلك يتم إنشـــاء نموذج حقيبة الكلمات على الوثيقة التي رُفِعَت داخل النظام من خلال كتابة الكود التالي:

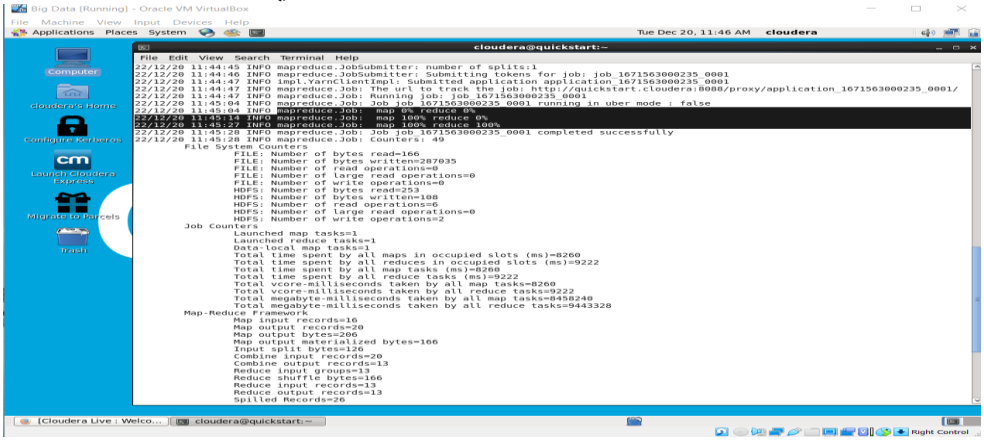
```
Hadoop Jar /Home/Cloudera/WordCount.jar WordCount /MomenFiles/Book1.txt /out1
```



شكل رقم (٤٠) يوضح كتابة كود البدء في عملية التحليل وتنفيذها على الوثيقة.

لتتولى المنصـــة عقب ذلك إجراء عملية تطبيق وتنفيذ نموذج حقيبة الكلمات، الذي بُرمِج على الوثيقة من خلال إجراء مرحلة التحويل Map للمحتوى النصي من صورة غير مهيكلة لصورة مهيكلة.

يلي ذلك إجراء مرحلة التخفيض Reduce للنتائج من خلال تجميع تكرارات المصـــطلح الواحد في صف خاص به، كما هو موضح الشكل رقم (٤١)؛ حيث تم الانتهاء من مرحلة التحويل Mapping بنسبة وصلت لـ١٠٠٪، ثم الانتهاء من عملية Reduce بنسبة ١٠٠٪ في تحليل الوثيقة.

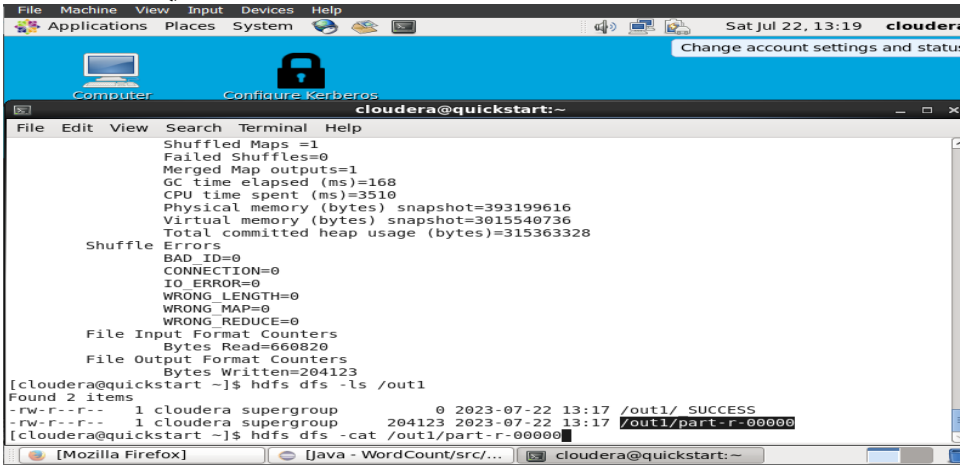


شكل رقم (٤١) يوضح اكتمال عملية التحليل وحصر عدد المصطلحات الواردة في الوثيقة وإنشاء نموذج حقيبة الكلمات الخاص بها.

تأتي المرحلة النهائية في عملية إنشـــاء نموذج حقيبة الكلمات للوثيقة من خلال طلب استعراض نتائج عملية التحليل، التي توضع بشكل أساسي في ملف يعرف باسم Part – r – 00000، من خلال الكود التالي:

Hdfs dfs -cat /out1/part -r -00000

لِيُستَدعَى هذا النموذج ويُستعرض ما يشمله من نتائج من خلال كما هو موضح في الشكل رقم (٤٢).



شكل رقم (٤٢) يوضح استدعاء ملف المخرجات المشتمل على نتائج عملية التحليل.

تتمثل الخطوة الأخيرة في إنشـــاء نموذج حقيبة الكلمات في اســـتعراض تكرار ظهور كل مصطلح داخل الوثيقة، بصورة كمية، كما هو موضح في الشكل رقم (٤٣)، وذلك من خلال كتابة الكود التالي:

Hdfs Dfs -Cat /out1/Part - r - 00000

شكل رقم (٤٣) يوضح قيام منصة Hadoop باستعراض تكرار ظهور كافة الكلمات والمصطلحات الواردة داخل الوثيقة.

وكانت أبرز النتائج التي تُوصِل إليها في عملية تحليل مصدر المعلومات من خلال منصة Hadoop، هي:

- بلغ حجم الملف المراد تحليله نحو ٣ جيجابايت، ليعكس سمة بسيطة من ضخامة البيانات.
- بلغ الوقت المستغرق في عملية التحليل للملف بالكامل نحو ٣-٥ ثانية.
- تمثلت العملية الأولى في قيام المنصة بدفق الملف لنظام HDFS الخاص بتقسيم الوثيقة، وقد قام بتقسيم الملف لعدد ٢٤ كتلة بيانات، حيث بلغ حجمها ٢٤ كتلة بمساحة ١٢٨ للكتلة الواحدة، (٣٠٠٠ ميجابايت /١٢٨ ميجابايت = ٢٣.٢٤ كتلة ببيانات).
- استغرقت عملية الدفق والتقسيم للملف داخل نظام Hdfs أقل من ثانية واحدة.
- بدأ النظام بإجراء عملية التحليل على كتل البيانات، وقد بلغ الوقت المستغرق في إجراء عملية التحويل للبيانات Mapping من وضعها غير المهيكل للهيكلة نحو 0.13 جزء من الثانية الواحدة.
- وبلغت عملية التخفيض Reduce لنتائج عملية التحويل للبيانات نحو 0.12 جزء من الثانية.
- بلغ حجم عمليات التحليل نحو ١١١٤٨٧ مهمة Job، خلال 1.5 ثانية.
- بلغ إجمالي عدد التسجيلات الناتجة من عملية التحليل نحو ١٠٩١٣٠٢ وحدة (كلمة فريدة متبوعة بعدد مرات التكرارات لكل كلمة كما هو في الشكل رقم 50).
- بلغت أكثر الكلمات المفتاحية التي حصلت على أعلى تردد في الكتاب وهي ما يلي:
  - Prophet = 714 مرة.
  - Mohammad= 552 مرة.
  - Allah = 452 مرة.
- بلغت أكثر الكلمات المستبعدة (ذات الدلالة المنخفضة Stop List كحروف الجر والعطف)، التي حصلت على أعلى تردد في الكتاب وهي كلمة (The = 12596 مرة).

١٥- النتائج:

اتضح من استعراض الجانب النظري للدراسة أن قضية معالجة وتحليل البيانات الضخمة تعد من القضايا الحيوية في الوقت الراهن، وتأتي منصة Apache Hadoop باعتبارها

إحدى أبرز المنصات مفتوحة المصدر، التي طورت في هذا السياق بغية المعالجة والتحليل لأنماط البيانات الضخمة، وأن أحد أبرز استثمارها في مجال المكتبات والمعلومات هو توظيفها داخل المستودعات الرقمية لتحليل مصادر المعلومات والخروج بالعديد من المؤشرات، التي لا تقتصر على المؤشرات الوظيفية (كمعدلات الاستخدام لهذه المصادر) بل يتخطى الأمر لإجراء عمليات التحليل والمعالجة على نصوص هذه المصادر واستنتاج العديد من القيمة المضافة، التي تساعد عي إثراء المعرفة لدى المستفيدين.

أوضحت الدراسة أن بنية منصة Hadoop جاءت مشتملة على ٣ عناصر تكوينية رئيسة، هي:

- HDFS لحفظ البيانات ذات الضخامة وتوزيع مهام عملية التحليل عليها.
- MapReduce لإجراء عمليات التحليل على البيانات، وهيكلتها.
- YARN لإدارة موارد الخوادم التي تتكون منها المنصة.

كما اتضح أن منصة Hadoop، لا تعد بمثابة برنامج مفرد يقوم بمعالجة البيانات الضخمة وتحليلها، بل إنها تعد بمثابة نظام تكويني Ecosystem يشتمل على العديد من التطبيقات بجانب البرنامج الرئيس، والتي تهدف إلى تبسيط وإدارة وإجراء عمليات التنسيق، وتحليل كميات ضخمة من البيانات.

اتضح أيضًا من خلال استعراض البنية التكوينية لمنصة Hadoop على سمة الذكاء الاصطناعي، الذي يتمتع بقدرة على التعامل مع الملفات الضخمة بالنمط التوزيعي Distribution Manner من حيث تقسيم وتفتيت الملفات ذات الضخامة في حجمها إلى كتل صغيرة لإجراء التحليل عليها بصورة أسرع، وأكثر دقة، كذلك كفل عملية تكرار ونسخ كتلل. البيانات قدرة النظام على إجراء العديد من عمليات المعالجة في وقت واحد دون الانتظار لتسلسل العمليات.

واتضح كذلك أن ما تتمتع به المنصة من ذكاء اصطناعي كفل لها إجراء عملية تحويل البيانات النصية غير المهيكلة لصورة بيانات مهيكلة، من خلال وحدة MapReduce، كما اتضح جليًّا في الدراسة التطبيقية، قدرة المنصة على إدارة موارد الأجهزة من معالجات ووحدات الذاكرة بصورة فعالة، وهو ظهر من خلال الوقت المستغرق في إجراء عملية التحليل والمعالجة، الذي استغرق معدلات لم تتجاوز الثواني في إجرائها.

كانت إحدى أبرز صعوبات التعامل مع منصة Haddop عدم دعمها لإجراء عمليات المعالجة والتحليل للملفات والوثائق الصادرة باللغة العربية، إذ يقتصر تعاملها مع اللغات الصادرة بالحروف اللاتينية كالإنجليزية، والفرنسية، والإيطالية، وغيرها.

أما على صعيد الجانب التطبيقي، فقد برز نجاح منصة Apache Hadoop في معالجة

مصـادر المعلومات غير المهيكلة، في صـورتها النصـية من خلال القيام ببناء نموذج حقيبة الكلمات Bag of words، بشـــكل كامل لإحدى الوثائق المنفردة، داخل إحدى المكتبات الرقمية العربية، وذلك من خلال الحصر الكامل لتردد كافة الكلمات الواردة داخل النص، كما اتســمت قدرة المنصـــة على التفريق بين مختلف الكلمات الواردة في النص، والتعامل معها باعتبارها وحدات منفصلة.

علي الرغم من قيام المنصــة بتحديد كافة الكلمات، فإن المنصــة اتسـمت في حصـرها لكافة الكلمات الواردة في النص بعدم التفريق بين الكلمات الدالة والكلمات غير ذات الأهمية، ولا عيب في ذلك لأن مهمة حقيبة الكلمات حصـــر التردد لمختلف الكلمات الواردة بالنص دون إعطـاء أوزان لقيمـة كـل كلمـة داخـل النص، نظرًا لوجود الخوارزميـات المتعلقـة بهـذا الأمر كخوارزمية TF/IDF.

كشفت عملية بناء نموذج حقيبة الكلمات، عن صـعوبة تعامل منصـة Hadoop مع الحروف غير اللاتينية، كالحروف العربية في بناء نموذج حقيبة الكلمات للمصـــادر العربية، الأمر الذي يســتدعي الكثير من الجهود في هذا الصـــدد لدعم عمليات التحليل والمعالجة للبيانات الضخمة العربية.

١٦- التوصيات:

توصـــي الدراســـة بشـــكل رئيس بتوجيه المجتمع الأكاديمي العربي المتخصـــص في مجال المكتبات وعلوم المعلومات، نحو استثمار منصة Hadoop في كل مما يلي:

- تضـــمن المنصـــة داخل بنى أنظمة المكتبات الرقمية مفتوحة المصـــدر، أو جعل المنصة تعمل بوصفها منصة فوقية على بنى أنظمة بناء المكتبات الرقمية.
- تطبيق المنصة في تكشيف المحتوى، لما تكفله من قدرات تتعلق بالتكشيف الدلالي، واستخراج الكيانات عوضًا عن التكشيف اللفظي.
- دعم الجانب التقني لمعالجة وتحليل المحتوى العربي للمنصـة في ظل كونها مفتوحة المصدر.
- إجراء عمليات التحليل والمعالجة لمصـــادر المعلومات المتضـــمنة داخل المكتبات الرقمية العربية.
- توظيف تقنيات الحفظ الموزع في إنشـــاء فهارس المكتبات؛ ليكفل لها الســـرعة والفاعلية في استرجاع مصادر المعلومات.
- توظيف منصـــة Hadoop عوضًـــا عن قواعد البيانات العلائقية لما تكفله بنيتها من قدرات في التحليل والمعالجة للبيانات غير المهيكلة.

١٧- المراجع والمصادر العربية:

١- أحمد فايز أحمد سيد. (٢٠١٩). نظم إدارة قواعد البيانات الضخمة: دراسة حالة لنظام أباتشـــي هادوب (Hadoop). أعلم. الاتحاد العربي للمكتبات والمعلومات، يناير ٢٠١٩.

٢- الأكلبي، علي، (٢٠١٨) "أهمية تحليل البيانات الضـــخمة في اتخاذ القرار في جامعة الملك سعود". المؤتمر السنوي الرابع والعشرين لجمعية المكتبات والمتخصصة/ فرع الخليج العربي: مسقط، ٦-٨ مارس ٢٠١٨.

٣- الهادي، محمد (٢٠١٦). ثورة البيانات وتحليلاتها التنموية والتخطيطية. القاهرة. المجلة المصرية للمعلومات، ١٧، ٣٣-٥٥ص.

٤- فرج، أحمد. (٢٠٢٢). استثمار البيانات الضخمة لتطوير آليات البحث والاسترجاع وتخصيص خدمات مؤسسات المعلومات: دراسة استشرافية. المجلة العلمية للمكتبات والوثائق والمعلومات4(11), 7-42. doi: 10.21608/jslmf.2021.53697.1058 .

١٨- المراجع والمصادر الأجنبية:

1. Adams, N. M. (2010). Perspectives on data mining. International Journal of Market Research, 52(1), 11-19. https://doi.org/10.2501/S147078531020103X
2. Aggarwal, Anshul. (2020). “Hadoop: History or Evolution” available at: https://www.geeksforgeeks.org/hadoop-history-or-evolution/
3. Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. Journal of Internet Services and Applications, 6(1), 1-15. https://doi.org/10.1186/s13174-015-0041-5
4. Al-Barashdi, Hafida. Et al. (2018). "Big Data in academic libraries: Literature review and future direction". Journal of Information Studies & Technology (JIS&T), Volume 2018, Issue 2, 2018. https://doi.org/10.5339/jist.2018.13
5. Al-Daihani, S, & Abrahams, A. (2016). “A Text Mining Analysis of Academic Libraries' Tweets”. The Journal of Academic Librarianship, 42:135-143. https://doi.org/10.1016/j.acalib.2015.12.014
6. Al-Mesad, Aseel. (2018). The preparation of big data phenomena in the public sector in Kuwait. SLA/AGC Big Data and its investment prospects, Muscat, 6-8 March

https://search.mandumah.com/Record/870230
7. Ali, A., Qadir, others. (2016). "Big data for development: applications and techniques". Big Data Anal. 1, 2. available at https://doi.org/10.1186/s41044-016-0002-4
8. Amazon Web Services, Inc. (2022). Introduction to Apache Spark. Available at: https://aws.amazon.com/big-data/what-is-spark/
9. Apache Hadoop. Available at: https://hadoop.apache.org/
10. Apache Nutch project. http://nutch.apache.org
11. Apache Spark. (2022). Apache Spark Documentation. Available at: https://spark.apache.org/docs/3.3.0/
12. Azarmi, B., 2016. Scalable Big Data Architecture. Springer. Available at: https://link.springer.com/book/10.1007/978-1-4842-1326-1#toc
13. Bahambhri, Anjul. (2012)."Executive Letter" in "Understanding Big Data". New York. McGraw hill. P 18. https://dl.acm.org/doi/10.5555/2132803
14. Banerjee, A, et al. (2013). “Data analytics: hyped up aspirations or true potential”. Vikalpa. The Journal for Decision Makers, 38(4), 1–11. https://journals.sagepub.com/doi/10.1177/0256090920130401
15. Beel, J., Gipp, B., Langer, S. et al. Research-paper recommender systems: a literature survey. Int J Digit Libr 17, 305–338 (2016). https://doi.org/10.1007/s00799-015-0156-0
16. Bhargava, Sandeep. etc. (2019). Performance Comparison of Big Data Analytics Platforms. International Journal of Engineering, Applied and Management Sciences Paradigms. Volume 54 Issue 2 May 2019. https://doi.org/10.1109/BigData.2017.8258260
17. Boyd, dana; Crawford, Kate (2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. https://dx.doi.org/10.2139/ssrn.1926431
18. Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of ‘big data’? McKinsey Quarterly, (oct 2011). https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/are-you-ready-for-the-era-of-big-data
19. Brueckne, Rich. (2013). “Where Did Big Data Come From?”. Available at: https://insidebigdata.com/2013/02/03/where-did-big-data-

come-from/
20. Cervone, H. (2017). Evaluating social media presence: A practical application of big data and analytics in information organizations. Available at: https://www.emerald.com/insight/content/doi/10.1108/DLP-10-2016-0040/full/html?casa_token=IvNMtTQKI5kAAAAA:BF6uw8YvgkgDiIi0au0FBr1vMRAvdXosMuyjL7sbCJpzuK4FNGfKc9ZQ6rAIAg9ppmb2qtu-7_lSC1K7CDVxc-w5xvG8IK19Lno6TUTqAG_JSDpvdQsawg
21. Chen, M., Mao, S., Zhang, Y., Leung, V.C., (2014). "Big Data: Related Technologies, Challenges and Future Prospects". Springer. available at: https://link.springer.com/book/10.1007/978-3-319-06245-7
22. Chen, M., S. Mao, and Y. Liu. (2014). "Big data: a survey," Mobile Networks and Applications", vol. 19, no. 2. https://doi.org/10.1007/s11036-013-0489-0
23. Chen, P., & Zhang, C. Y. (2014). "Data-intensive applications, challenges, techniques and technologies: a survey on big data". Information Sciences, 275, 31 347. https://doi.org/10.1016/j.ins.2014.01.015
24. Clare, Hopping. (2021). "The best big data technologies". IT Pro. available at: https://www.itpro.co.uk/strategy/28161/the-best-big-data-technologies
25. Cloud bigtable: HBase-compatible, NoSql database. google cloud Google. Available at: https://cloud.google.com/bigtable (Accessed: 25 July 2023).
26. Data Flair Team. (2020). History of Hadoop – The complete evolution of Hadoop Ecosystem. Available at: https://data-flair.training/blogs/hadoop-history
27. Dean, J., and Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA. Retrieved 8 May 2020 from: http://static.googleusercontent.com/media/research.google.com/es/us/archive/mapreduce-osdi04.pdf
28. Delen, D.; Demirkan, H. (2013). "Data, information and analytics as services". Decision Support Syst., 55, p359–363.

https://doi.org/10.1016/j.dss.2012.05.044
29. Dinet, E., & Ben Ibrahim, S. (1918). “The life of Mohammad, the prophet of Allah”. Paris Book Club. https://www.gutenberg.org/ebooks/39523
30. Douglas, K. (2012). “Infographic: big data brings marketing big numbers". Available at: https://www.marketingdive.com/news/infographic-big-data-brings-marketing-big-numbers/28161/
31. Editor, B. (2016, October 28). Why hadoop is important in handling big data?. Big Data Week Blog. https://blog.bigdataweek.com/2016/08/01/hadoop-important-handling-big-data/
32. Emani, C.K., Cullot, N., Nicolle, C., 2015. Understandable big data: a survey. Computer Sci. Rev. 17, 70–81. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1574013715000064
33. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279. https://doi.org/10.1109/TETC.2014.2330519
34. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National science review, 1(2), 293-314. https://doi.org/10.1093/nsr/nwt032
35. Federer, L. (2016). “Research data management in the age of big data: Roles and opportunities for librarians”. Information Services & Use, 36(1-2), 35-43. https://doi.org/10.3233/ISU-160797
36. Gandomi, A., & Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
37. Gantz, John and Reinsel, David. (2011). “Extracting value from chaos”. IDC iView. https://www.sciepub.com/reference/140415
38. Gartner IT Glossary. What is Big Data? URL: http://www.gartner.com/it-glossary/big-data

39. Geeks for Geeks. (2020). "Hadoop: History or Evolution" available at: https://www.geeksforgeeks.org/hadoop-history-or-evolution/.
40. Ghemawat, S., Gobioff, H., and Leung, S.-T. 2003. The Google file system. In 19th Symposium on Operating Systems Principles. Lake George, NY. 29-43. https://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf
41. Gordon-Murnane, L. (2012). "Big Data: A big opportunity for librarians". Online, 36(5), 30-34. https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=26495062
42. Hadoop Tutorial. (2020). Hadoop 1 vs Hadoop 2- The Major Difference You should know. (2020). https://www.hdfstutorial.com/blog/hadoop-1-vs-hadoop-2-differences/
43. Hendrickson, S. (2010). "Getting Started with Hadoop with Amazon's Elastic MapReduce", EMR. https://cloudacademy.com/lab/getting-started-amazon-elastic-mapreduce/
44. Amazon SimpleDB. (2023). https://aws.amazon.com/simpledb/
45. Microsoft, Azure. (2024). https://azure.microsoft.com/en-us/products/azure-sql/database/
46. Softonic International. (2024) https://d3.en.softonic.com/
47.) مركز المعرفة الرقمي 2024.(https://ddl.ae
48. Cloudera. (2024). https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip
49. Apache Drill. (2024). https://drill.apache.org/download/
50. Machine Learning for Data Streams. (2024). https://moa.cms.waikato.ac.nz/
51. SpagoBI. (2024). https://spagobi.software.informer.com/4.1/
52. Apache Spark. (2024). https://spark.apache.org/downloads.html
53. Splice Machine. (2024). https://www.linkedin.com/company/splice-machine/about/
54. Apache Storm. (2024). https://storm.apache.org/downloads.html
55. Cloudera. (2024). https://www.cloudera.com
56. MarkLogic. (2024). https://www.marklogic.com/product/marklogic-database-overview/

57. MongoDB, Inc. (2024). https://www.mongodb.com
58. Sky Tree. (2024). https://www.skytree.net
59. Tableau Software. (2024). https://www.tableau.com
60. Oracle Virtual Box. (2024). https://www.virtualbox.org/wiki/Downloads
61. WibiData. (2024). https://www.wibidata.com
62. IFLA. (2023). “Big Data Special Interest Group” available at: https://www.ifla.org/units/big-data/
63. Internet Live Stats (2016), “Internet live stats”, available at: www.internetlivestats.com/internet-users-by-country/
64. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In 2013 46th Hawaii international conference on system sciences (pp. 995-1004). IEEE. https://doi.org/10.1109/HICSS.2013.645
65. Khan, N., Yaqoob, I., Hashem, I., Inayat, Z., Ali, W., Allam, M., Shiraz, M., and Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. The scientific world journal, https://doi.org/10.1155/2014/712826
66. King, T. (2021, May 19). 80 percent of your data will be unstructured in five years. Best Data Management Software, Vendors and Data Science Platforms. https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/
67. Labrinidis, Alexandros & Jagadish, H.V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment. 5. 2032-2033. http://dx.doi.org/10.14778/2367502.2367572
68. Landset, S., T. M. Khosh goftaar, A. N. Richter, and T. Hasnain. (2015). A survey of open-source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2 (1):1. https://doi.org/10.1186/s40537-015-0032-1
69. Laney, D. (2001). "3-d data management: controlling data volume, velocity and variety". META Group Research. https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1611280
70. Lebdaoui, I, et al. (2014). "An integration adaptation for real-time Data warehousing". International Journal of Software Engineering and its

Applications, 8(11), 115–128.
http://dx.doi.org/10.14257/ijseia.2014.8.11.10

71. Liu, X., N. Iftikhar, and X. Xie.)2014). Survey of real-time processing systems for big data. In Proceedings of the 18th International Database Engineering & Applications Symposium, pages 356–361. ACM. http://dx.doi.org/10.1145/2628194.2628251
72. Lohr, Steve. (2013). "The Origins of 'Big Data': An Etymological Detective Story". New York Times. Retrieved 15 December 2017. https://archive.nytimes.com/bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/
73. Manaseer, Saher. et al. (2018). “Big Data Investment and knowledge Integration using HADOOP Framework in Academic Libraries”. In 24th Annual Conference & Exhibition of the SLA/AGC Big Data and its investment prospects, Muscat, 6-8 March 2018. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1135&context=scholcom
74. Manyika, J., et al & McKinsey Global Institute. (2011). “Big data: The next frontier for innovation, competition, and productivity". Available at: www.mckinsey.com/mgi/publications/
75. Market Watch. (2022). “Hadoop Market Size, Overview with details Analysis, Competitive Landscapes, Forecast to 2022-2030”. Available at: https://www.credenceresearch.com/report/hadoop-market Date: 12/9/2002
76. Marr, Bernard. (2018). “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read”. Forbes, May 21, 2018. Available at: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#114a045560ba
77. Mckinsey Digital. (2022). “The data-driven enterprise of 2025”. available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025#/
78. Micci-Barreca, Daniele. (2021). “Big Data” reaches plateau while interest in Machine Learning grows”. Available at: https://www.linkedin.com/pulse/big-data-reaches-plateau-while-

interest-machine-grows-micci-barreca/
79. Mois, Martin. (2016). “Apache Hadoop Tutorial: The Ultimate Guide. https://www.javacodegeeks.com/2016/02/apache-hadoop-tutorial.html
80. MongoDB: Riding the Data Wave. (2020). Available at https://seekingalpha.com/article/4317681-mongodb-riding-data-wave
81. Mosavi, A. (2018). Industrial Applications of Big Data: State of the Art Survey. In: Luca, D. (eds) Recent Advances in Technology Research and Education. INTER-ACADEMIA 2017. Advances in Intelligent Systems and Computing, vol 660. Springer, Cham. https://doi.org/10.1007/978-3-319-67459-9_29
82. Nada Elgendy and Ahmed Elragal. (2014). Big Data Analytics: A Literature Review Paper. ICDM 2014.Berlin: Springer. Pp. 214–227. http://dx.doi.org/10.1007/978-3-319-08976-8_16
83. Normandeau K (2013), Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity
84. O’Malley, Owen. (2008) “Terabyte Sort on Apache Hadoop,” available at: http://sortbenchmark.org/YahooHadoop.pdf
85. Oracle. (2013). "Big Data Analytics: Advanced Analytics in Oracle Database". Available at: https://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataoracleadvanalytics11gr2-1930894.pdf
86. Oussous. A., F. Z. Benjelloun, A. A. Lahcen and S. Belfkih, "Big data technologies: A survey", J. King Saud Univ. Comput. Inf. Sci., vol. 30, no. 4, pp. 431-448, Oct. 2018. https://doi.org/10.1016/j.jksuci.2017.06.001
87. Oxford College of Marketing. (2023) available at: https://twitter.com/oxcom_marketing/status/1295279209082949634
88. Paila, U. (2020, March 13). Basic feature extraction methods. Practical Machine Learning. https://udibhaskar.github.io/practical-ml/nlp/feature%20extraction/bow/tfidf/hashing%20vectorizer/2020/03/13/Basic-feature-Extraction.html
89. ProQuest. (2013). “Data Mining “Big Data”: A Strategy for Improving Library Discovery”. Available at

https://www.sciencedirect.com/science/article/abs/pii/S009913331500107X

90. Qader, Wisam & M. Ameen, Musa & Ahmed, Bilal. (2019). An Overview of Bag of Words; Importance, Implementation, Applications, and Challenges. 200-204. https://ieeexplore.ieee.org/document/8950616
91. Rainstor.com (2013). available at https://en.wikipedia.org/wiki/RainStor
92. Rajaraman, A., & Ullman, J. (2011). Data Mining. In Mining of Massive Datasets (pp. 1-17). Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/abs/mining-of-massive-datasets/data-mining/E5BFF4C1DD5A1FB946D616D619B373C2
93. Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah, T. Y., & Khan, S. U. (2016). Big data reduction methods: a survey. Data Science and Engineering, 1, 265-284. https://link.springer.com/article/10.1007/s41019-016-0022-0
94. Reinsel, David, John Gantz, and John Rydning. (2017). "Data Age 2025: The Evolution of Data to Life-Critical". Don't Focus on Big Data. An IDC White Paper. https://itupdate.com.au/page/data-age-2025-the-evolution-of-data-to-life-critical-
95. Segaran, Toby; Hammerbacher, Jeff (2009). Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. https://www.amazon.com/Beautiful-Data-Stories-Elegant-Solutions/dp/0596157118
96. Shumway. R, Harrison. M. (2016). "Big Risks Vs. Big Opportunities". Cicero Institute. https://cicerogroup.com/blog/2019/07/17/the-big-data-debate-big-risks-vs-big-opportunities/
97. Singh ،D., and C. K. Reddy. (2014). A survey on platforms for big data analytics. Journal of Big Data, 2 (1):8. https://doi.org/10.1186/s40537-014-0008-6
98. Sivarajah. U, et al. (2017). "Critical analysis of Big Data challenges and analytical methods". Journal of Business Research 70. Available at: http://dx.doi.org/10.1016/j.jbusres.2016.08.001
99. Statista. (2020). "Size of Hadoop and Big Data markets worldwide in 2015 and 2020". Available at:

https://www.statista.com/statistics/587051/worldwide-hadoop-bigdata-market/

100. TechAmerica Foundation's Federal Big Data Commission, Demystifying Big data: A Practical Guide to Transforming the Business of Government, URL: http://www.techamerica.org/Docs/fileManager
101. Ucros, M. A. (2017). "Lady boss, here's why you should study big data". Medium. https://medium.com/@melodyucros/ladyboss-heres-why-you-should-study-big-data-721b04b8a0ca
102. USA, The White House. (2012). "Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments" available at: https://obamawhitehouse.archives.gov/the-press-office/2015/11/19/release-obama-administration-unveils-big-data-initiative-announces-200
103. Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. International journal of production economics, 165, 234-246. https://doi.org/10.1016/j.ijpe.2014.12.031
104. Wang, C., et. (2016). "Exposing library data with big data technology: A review". Computer and Information Science (ICIS), 15th International Conference on. Okayama, Japan: IEEE. http://dx.doi.org/10.1109/ICIS.2016.7550937
105. Wang, Chunning & Xu, Shaochun & Chen, Lichao & Chen, Xuhui. (2016). Exposing library data with big data technology: A review. 1-6. 10.1109/ICIS.2016.7550937, Available at: https://www.researchgate.net/publication/306926515_Exposing_library_data_with_big_data_technology_A_review
106. Watson, H. J. (2014). "Tutorial: big data analytics: Concepts, technologies, and applications". Communications of the Association for Information Systems, 34(1). https://aisel.aisnet.org/cais/vol34/iss1/65/
107. White, T. (2015). Hadoop: The definitive guide (5th ed.). O'Reilly. https://www.oreilly.com/library/view/hadoop-the-definitive/9780596521974/
108. Wilson, Christy. (2017). "Who's Using Hadoop? And What are They

Using It For?". Syncsort available at: https://blog.syncsort.com/2015/06/big-data/whos-using-hadoop-and-what-are-they-using-it-for/

109. Wired. (2013). The Missing Vs in Big Data: Viability and Value. Available at: https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/
110. Woodie, A. (2017) Only a fraction of 160 zettabyte 'Datasphere' to be stored, Datanami. Available at: https://www.datanami.com/2017/04/25/fraction-160-zettabyte-datasphere-stored/ (Accessed: 25 July 2023).
111. Yahoo! Launches World's Largest Hadoop Production Application. (2008). Available at: https://lucene-group-group.iteye.com/group/topic/4244
112. Zikopoulos. Paul. &Others. (2011). "Understanding Big Data: analytics for enterprise class Hadoop and streaming data". New York. McGraw hill ". New York. McGraw hill. P 24. https://www.amazon.com/Understanding-Big-Data-Analytics-Enterprise/dp/0071790535