نمذجة وتحليل البيانات الطبية الضخمة في بيئة الحوسبة السحابية

الباحث المهندس: شادي علي بليدي
كلية الهندسة المعلوماتية – جامعة دمشق

ملخص:

إن الانخفاض الكبير في تكلفة خدمات الرعاية الصحية، واستخدام الموارد، وقابلية الصيانة واعتماد التقنيات الجديدة، هي بعض الفوائد التي يمكن لمراكز الرعاية الصحية والمستشفيات الحصول عليها من نظام المعلومات الطبية القائم على السحابة.

كما أن الطبيعة المعقدة والموزعة والمتعددة التخصصات للبيانات الطبية فرضت وجود قيود على إمكانيات تحليل البيانات بالطرق التقليدية الخاصة بالوصول إلى البيانات وتخزينها ومعالجتها وتحليلها وتوزيعها ومشاركتها. وبالتالي أصبحت التقنيات الجديدة والفعالة ضرورية للحصول على ثروة المعلومات والمعرفة التي تقوم عليها البيانات الطبية الضخمة. تعد تحليلات البيانات الضخمة مجال نمو له القدرة على توفير رؤية مفيدة في نظام المعلومات الصحية. يمكن للبيانات الضخمة توحيد جميع البيانات المتعلقة بالمريض للحصول على المزيد من الخيارات لعرض سجلات المرضى لتحليل وتوقع اكتشاف المرض المبكر. البيانات الضخمة تدعم وتحسن الممارسات السريرية، وتطوير الأدوية الجديدة وعملية تمويل الرعاية الصحية.

يقترح هذا البحث ويناقش نموذجاً لتحليل البيانات الطبية الضخمة في بيئة الحوسبة السحابية، من خلال دراسة وظائف Hadoop المختلفة، وإجراء تحليلات البيانات على مجموعة بيانات الرعاية الصحية باستخدام تقنياته المختلفة.

كلمات مفتاحية: الحوسبة السحابية ، البيانات الطبية الضخمة ، Hadoop ، MapReduce ، Hive.

Modeling and analyzing big medical data in a cloud computing environment

Eng: Shadi A. Blidi

Abstract

The drastic reduction in the cost of healthcare services, utilization of resources, maintainability and the adoption of new technologies are some of the benefits that healthcare centers and hospitals can get from cloud-based medical information system.

The complex, distributed, and highly interdisciplinary nature of medical data has underscored the limitations of traditional data analysis capabilities of data accessing, storage, processing, analyzing, distributing, and sharing. New and efficient technologies are becoming necessary to obtain the wealth of information and knowledge underlying medical Big Data. Big data analytics is a growth area with the potential to provide useful insight in health information system. Big Data can unify all patient related data to get more option to view patient records to analyze and predict early disease detection. Big data supports and improve clinical practices, new drug development and health care financing process.

This research proposes and discusses a model of analyzing medical Big Data in a cloud computing environment, by studying different Hadoop functionalities in details and perform data analytics on a health care data set using Hadoop.

Keywords: cloud computing, Medical big data, Hadoop, MapReduce, Hive.

1- مقدمة:

التحدي الحاسم الذي تواجهه مؤسسات الرعاية الصحية هو تحليل البيانات الطبية المتضخمة على نطاق واسع. حيث أنه مع النمو السريع لتطبيقات الرعاية الصحية المختلفة، تولد الأجهزة المختلفة المستخدمة في الرعاية الصحية أنواعًا مختلفة من البيانات. وبالتالي يجب معالجة البيانات وتحليلها بفعالية من أجل اتخاذ قرارات أفضل.

يحتاج مزودو الرعاية الصحية في جميع أنحاء العالم إلى الحصول على معلومات في الوقت الحقيقي لتوفير رعاية صحية جيدة. لذلك يسعون لاستخدام تطبيقات البرمجيات كخدمة تقدمها معظم مزودي الخدمات السحابية. ونظرًا لأن معلومات الرعاية الصحية سرية وتوفر رعاية صحية ذات جودة أفضل، يجب على مختلف أصحاب المصلحة تبادل معلومات المرضى والمعلومات السريرية بطريقة آمنة.

تتوفر بيانات الرعاية الصحية حتى الآن في شكل سجلات طبية إلكترونية (EMR) وسجلات صحية إلكترونية (EHR) وسجلات المرضى الطبية (PMR). وبالتالي فإن جمع السجلات الطبية الرقمية على مدى فترة من الزمن سيؤدي إنشاء مجموعات البيانات الضخمة. البيانات الطبية الضخمة الناتجة عن مجموعات الرعاية الصحية تحتاج إلى التحليل الفعال ومحاولة حل المشكلات المختلفة التي تواجهها [1].

إن الرعاية الصحية الجيدة، وتخفيض التكلفة الطبية، وكفاءة اتخاذ القرارات لتوفير الرعاية الصحية المناسبة، وإيجاد أنماط لعمليات إعادة الإدخال غير الضرورية للمستشفيات هي بعض المشكلات التي يمكن حلها بواسطة تقنية البيانات الضخمة والحوسبة السحابية. حيث يمكن تحليل قيمة جديدة من مجموعات البيانات الضخمة من خلال بناء أدوات تحليلية فعالة تساعد المرضى والأطباء وأصحاب المصلحة المختلفين في مجال الرعاية الصحية.

بما أن بيانات الرعاية الصحية متوفرة في مجموعات بيانات ضخمة وبأشكال مختلفة، فإن البيئة السحابية هي الطريقة الفعالة لتخزين ومعالجة هذه البيانات. حيث أنه يتم اليوم نشر معظم تطبيقات البرمجيات في مراكز بيانات (Data Centers). ومن أجل إجراء العمليات الحسابية المعقدة، تعد الحوسبة السحابية بمثابة بنية مهيمنة يمكنها أداء حساب البيانات على نطاق واسع بكفاءة من خلال توفير الموارد القابلة للتطوير [2].

حلت الحوسبة السحابية معظم المشكلات المتعلقة ببيانات الرعاية الصحية مثل توحيد تبادل سجلات الرعاية الصحية والخصوصية وأمان الشبكة. الأمن هو القضية الرئيسية أثناء مشاركة بيانات الرعاية الصحية في السحب الالكترونية [2].

يعمل إطار Hadoop على حل معظم المشكلات المتعلقة بمعالجة البيانات الضخمة. MapReduce هو نموذج الحوسبة المستخدمة في Hadoop لأنه يوفر نظام الملفات الموزعة (Hadoop HDFS) الذي يخزن البيانات على العقد. تحتاج مجموعات بيانات الرعاية الصحية إلى التحليل على مجموعات (Hadoop Clusters) في بيئة الحوسبة السحابية لحل مختلف القضايا في صناعة الرعاية الصحية [3].

2- أهداف البحث:

يهدف هذا البحث إلى الاستفادة من تقنيات تكنولوجيا المعلومات الحديثة في تطوير نموذج سحابي لإدارة وتحليل المعلومات الطبية الضخمة، بما يساهم في تقديم الخدمات الطبية والتحليلات المتعلقة بها بالسرعة المناسبة، وبما يضمن مساعدة المجتمع الطبي للوصول إلى الحلول المثلى ما أمكن في التشخيص والعلاج.

يمكن تصنيف أهداف البحث في العديد من الجوانب:

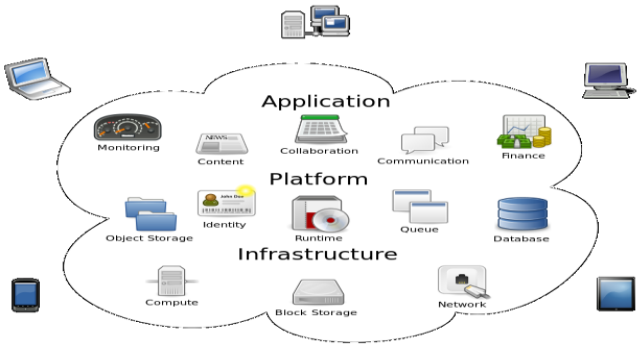
- استعراض ومراجعة مفاهيم ومنهجيات العمل الحالي في استخدام تقنيات البيانات الضخمة والحوسبة السحابية في مجال الرعاية الصحية بشكل عام، واقتراح مزيد من التطوير في هذا المجال.
- استعراض التحديات المرتبطة باعتماد تقنيات البيانات الضخمة والحوسبة السحابية في إدارة الرعاية والبحوث الصحية واستكشاف الحلول المناسبة.
- بالإضافة إلى تطوير وتقديم نموذج بنية نظام معلومات طبية معتمد على تقنيات البيانات الضخمة وتكنولوجيا الحوسبة السحابية للوصول إلى نظام معلومات طبي سحابي شامل يساهم في تطوير تحليل البيانات الطبية الضخمة وضمان جودة نتائجها.

3- مواد وطرق البحث:

3-1- الحوسبة السحابية (Cloud Computing)

الحوسبة السحابية هي مصطلح يشير إلى المصادر والأنظمة الحاسوبية المتوافرة تحت الطلب عبر شبكة الويب، والتي تستطيع توفير عدد من الخدمات الحاسوبية المتكاملة دون التقيد بالموارد المحلية بهدف التيسير على المستخدم. وتشمل تلك الموارد مساحة لتخزين البيانات والنسخ الاحتياطي والمزامنة الذاتية، كما تشمل قدرات معالجة برمجية وجدولة للمهام والبريد الالكتروني والطباعة عن بعد. ويستطيع المستخدم عند اتصاله بالشبكة التحكم في هذه الموارد عن طريق واجهة برمجية بسيطة تبسط العمل وتتجاهل الكثير من التفاصيل والعمليات الداخلية [4].

و قد عرف المركز القومي للمعايير والتكنولوجيا "السحابة" على أنها: نموذج لتوفير وصول مناسب ودائم في أي وقت إلى الشبكة، لمشاركة مجموعة كبيرة من المصادر الحاسوبية والتي يمكن نشرها وتوفيرها بأدنى مجهود أو تفاعل مع موفر الخدمة [4].



الشكل (1) البنية العامة للحوسبة السحابية [5]

طبقات الحوسبة السحابية

يمكن استعراض الحوسبة السحابية على أنها مجموعة من الخدمات التي يمكن تقديمها كبنية طبقات متصلة على السحابة الالكترونية كما يلي [5]:

- **البنية التحتية كخدمة (Infrastructure as a Service – IaaS)**

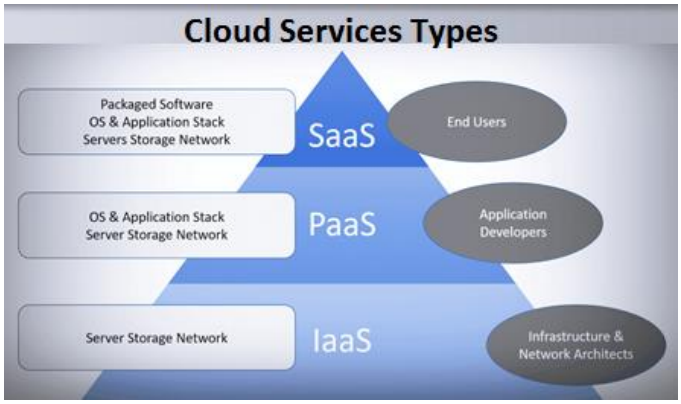
ويعتمد هذا النموذج في أبسط مفهوم له على مخدم افتراضي قائم على السُحب يوفر خدمات الشبكات والتخزين وخدمات البنية التحتية الأخرى، ولا يقوم العميل بإدارة مركز البيانات. أو التحكم فيه، ولكن يمكنه التحكم في البيانات وأنظمة التشغيل.

- **المنصة كخدمة (Platform as a Service –PaaS)**

ومن خلال هذا النموذج من الحوسبة السحابية يستطيع العملاء استعمال تطبيقاتهم على البنية التحتية لمقدم خدمات الحوسبة السحابية، وايضاً يستطيع العميل التحكم في البيانات وفي جزء من البيئة المضيفة.

- **التطبيقات البرمجية كخدمة (Software as a Service – SaaS)**

وفيه يستطيع العملاء النفاذ إلى تطبيقات مُقدِّم خدمات الحوسبة السحابية من خلال شبكة الإنترنت، وهو الشكل الأكثر شيوعاً لخدمات الحوسبة السحابية، وتستخدمه أغلب شبكات التواصل الاجتماعي ومقدمي خدمات البريد الالكتروني.



الشكل (2) طبقات الحوسبة السحابية [5]

وتستخدم الحوسبة السحابية تكنولوجيا الحوسبة الافتراضية (Virtualization Technology) في نموذج "البنية التحتية كخدمة" الخاص بها، حيث أنه يساعد على توفير الطاقة والتكلفة والمساحة في مراكز البيانات. فالحوسبة الافتراضية تعد حجر الأساس في بنيان السحابة.

- أنماط الحوسبة السحابية

هناك أربعة أنواع من الحوسبة السحابية [5]:

1- السحابة العامة (Public Cloud).

2- السحابة الخاصة (Private Cloud).

3- السحابة الهجينة (Hybrid Cloud).

4- السحابة المجتمعية (Community Cloud).

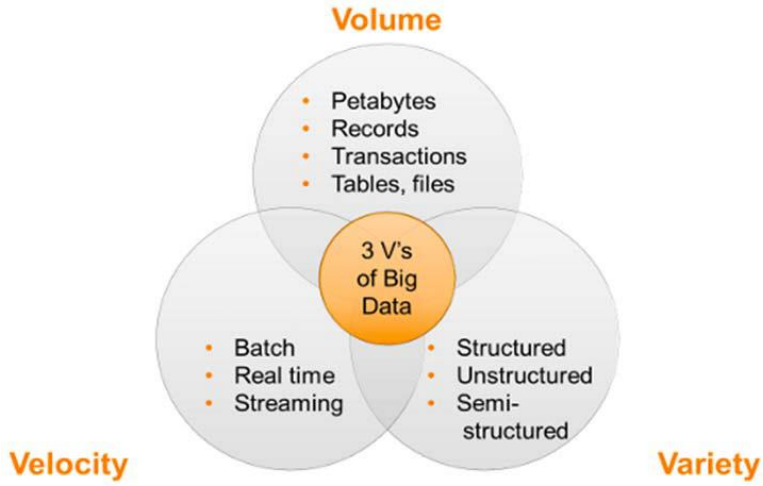
3-2- البيانات الضخمة & Hadoop

كمية البيانات التي يتم إنشاؤها كل يوم في العالم تتزايد بشكل هائل. حيث أن زيادة حجم الوسائط الرقمية والاجتماعية وانترنيت الأشياء تغذيها أكثر. إن معدل نمو البيانات ينمو بشكل مذهل وتأتي هذه البيانات بسرعة، بشكل مجموعات متنوعة (غير منظمة بالضرورة) وتحتوي على ثروة من المعلومات التي يمكن أن تكون مفتاحًا لاكتساب ميزة في الشركات المتنافسة. إن القدرة على تحليل هذا الكم الهائل من البيانات أدت إلى حقبة جديدة من نمو الإنتاجية والابتكار [6].

البيانات الضخمة هي المصطلح الخاص بمجموعة من مجموعات البيانات الضخمة والمعقدة بحيث يصبح من الصعب معالجتها باستخدام أدوات إدارة قواعد البيانات التقليدية أو تطبيقات معالجة البيانات. تشمل التحديات مجالات التقاط هذه البيانات وتنظيمها وتخزينها والبحث عنها ومشاركتها ونقلها وتحليلها وتصورها.

سمات البيانات الضخمة

تُستخدم الثلاثية – الحجم والسرعة والتنوع – عادةً لوصف الجوانب المختلفة للبيانات الضخمة. تسهل هذه السمات الثلاث تحديد طبيعة البيانات ومنصات البرامج المتاحة للتحليل [7].



الشكل (3) سمات البيانات الضخمة [7]

- **الحجم (Volume)**

يعد الحجم هو الجانب الأكثر تحديا للبيانات الضخمة، لأنه يفرض الحاجة إلى تخزين قابل للتطوير ونهج موزع للاستعلام. الشركات الكبيرة لديها بالفعل كمية كبيرة من البيانات التي تم تجميعها وحفظها على مر السنين. يمكن أن تكون في شكل نظام سجلات وحفظ السجلات... الخ. يصل مقدار هذه البيانات بسهولة إلى النقطة التي قد لا تتمكن فيها أنظمة إدارة قواعد البيانات التقليدية من التعامل معها.

- **السرعة (Velocity)**

حالياً تتدفق البيانات إلى المنظمات بسرعة كبيرة. حيث أتاحت تقنيات الويب والجوال توليد تدفق البيانات إلى مقدمي الخدمات بشكل كبير. كما أحدث التسوق عبر الإنترنت ثورة في تفاعلات المستهلك ومزود الخدمة.

- **التنوع (Variety)**

إن البيانات الناتجة عن الوسائط الاجتماعية والرقمية نادراً ما تكون بيانات منظمة. تدعم قواعد البيانات التقليدية "الكائنات الكبيرة" (LOB)، ولكن لها حدودها إن لم يتم توزيعها. من الصعب احتواء هذه البيانات في هياكل إدارة قواعد البيانات العلائقية التقليدية والأنيقة وليست بيانات صديقة للتكامل وتحتاج إلى الكثير من عمليات التعديل قبل أن تتمكن التطبيقات من إدارتها. وهذا يؤدي إلى فقدان المعلومات. إذا فقدت البيانات، فهذه خسارة لا يمكن استردادها.

مفهوم Hadoop

بشكل عام Hadoop مرتبط ارتباط وثيق مع البيانات الضخمة . وهو منصة برمجية مفتوحة المصدر تديرها مؤسسة (Apache Software Foundation). إنه النظام الأساسي الأكثر شهرة لتخزين وإدارة كمية هائلة من البيانات بكفاءة وفعالية من حيث التكلفة.

التعريف الرسمي لـ Hadoop من قبل Apache: هو مكتبة برامج و إطار يتيح المعالجة الموزعة لمجموعات البيانات الضخمة عبر مجموعات من أجهزة الكمبيوتر باستخدام نماذج برمجة بسيطة.

Hadoop هو إطار مفتوح المصدر من قبل Apache، وقد اخترع طريقة جديدة لتخزين ومعالجة البيانات .لا تعتمد على أجهزة عالية التكلفة وعالية الكفاءة. بدلاً من ذلك، تستفيد من فوائد المعالجة المتوازية الموزعة لكميات هائلة من البيانات عبر خوادم منخفضة التكلفة. تقوم هذه البنية الأساسية بتخزين البيانات وكذلك معالجتها، ويمكن بسهولة توسيع نطاقها حسب الاحتياجات المتغيرة. [8].

تم تصميم Hadoop للعمل على الأجهزة الموزعة ويمكن أن يرتفع أو ينخفض الأداء دون انهيار النظام. وهو يتألف من ثلاث وظائف رئيسية هي: التخزين والمعالجة وإدارة الموارد. يتم استخدامه حاليًا من قبل الشركات الكبرى مثل (Yahoo، eBay ، LinkedIn، Facebook).

خصائص Hadoop

يمتلك Hadoop العديد من الميزات والخصائص التي تجعله من أهم أدوات التعامل مع البيانات الضخمة، من هذه الخصائص [9]:

- التسامح مع الخطأ: التسامح مع الخطأ هو قدرة النظام على البقاء وظيفياً دون انقطاع ودون فقدان البيانات حتى لو فشل أي من مكونات النظام. أحد الأهداف الرئيسية لـ Hadoop هو أن تكون متسامحة مع الخطأ. نظرًا لأن عنقود Hadoop يمكن أن تستخدم آلاف العقد التي تعمل على الأجهزة المختلفة، فإنها تصبح عرضة بدرجة كبيرة للفشل. Hadoop .يحقق التسامح مع الخطأ عن طريق البيانات. ويوفر أيضا القدرة على مراقبة المهام الجارية وإعادة تشغيل المهمة تلقائيًا إذا فشلت.

- البنية التكرارية: Hadoop يكرر أساسا البيانات في كتل عبر عقد البيانات. ولضمان كل كتلة هناك كتلة احتياطية من نفس البيانات الموجودة في مكان ما عبر عقد البيانات. العقدة الرئيسية تتبع هذه العقد مع أسلوب تعيين البيانات. وفي حالة فشل أي من العقدة، تتولى العقدة الأخرى التي توجد بها كتلة البيانات الاحتياطية، جعل البنية الأساسية آمنة. لدى RDBMS التقليدية نفس المخاوف ويستخدم مصطلحات مثل: المثابرة والنسخ الاحتياطي والاسترداد. هذه المخاوف تتصاعد مع البيانات الضخمة.
- Scale القياس التلقائي لأعلى / لأسفل: تعتمد Hadoop اعتمادًا كبيرًا على نظام الملفات الموزعة، وبالتالي فهي تأتي مع إمكانية إضافة أو حذف عدد العقد المطلوبة في الكتلة بسهولة.

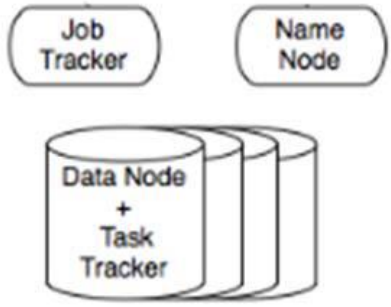
- نقل الحساب إلى البيانات: يتم تنفيذ أي استفسارات حسابية حيث توجد البيانات. هذا تجنب الحمل المطلوبة لإحضار البيانات إلى البيئة الحسابية. يتم حساب الاستعلامات بشكل متوازٍ ومحلياً، ويتم دمجها لاستكمال مجموعة النتائج.

مكونات Hadoop

فيما يلي سيتم شرح أهم المكونات الأساسية في إطار Hadoop [10]:

- **نظام الملفات الموزعة في Hadoop – HDFS**

HDFS هو نظام ملفات موزع مصمم للعمل على الأجهزة الاستهلاكية الرخيصة. HDFS لديه هندسة (Master/Salve). وهو أسلوب للكتابة مرة وقراءة عدة مرات.

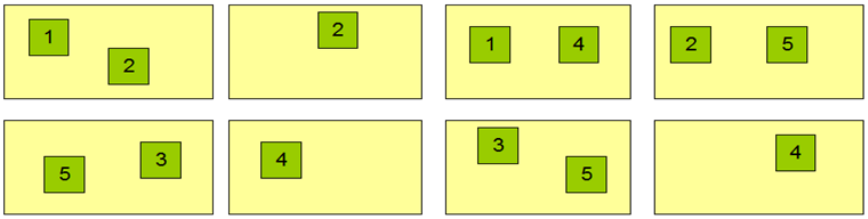


الشكل (4) تصور مبسط لعنقود Hadoop [10]

يتكون عنقود HDFS من NameNode واحد، وهو جهاز خادم رئيسي يدير نظام الملفات وينظم وصول العملاء إلى نظام الملفات. بالإضافة إلى العديد من عقد البيانات لكل عنقود. يتم تقسيم البيانات إلى كتل وتخزينها على عقد البيانات هذه. يحتفظ NameNode بخريطة توزيع البيانات. تعد نقاط البيانات مسؤولة عن عمليات قراءة وكتابة البيانات أثناء تنفيذ تحليل البيانات.

Hadoop Administrator يمكنه تحديد أجزاء البيانات التي سيتم حفظها على أي رفوف. هذا لمنع فقدان جميع البيانات في حالة فشل الحامل بأكمله وكذلك لتحسين أداء الشبكة من خلال تجنب الاضطرار إلى نقل أجزاء كبيرة

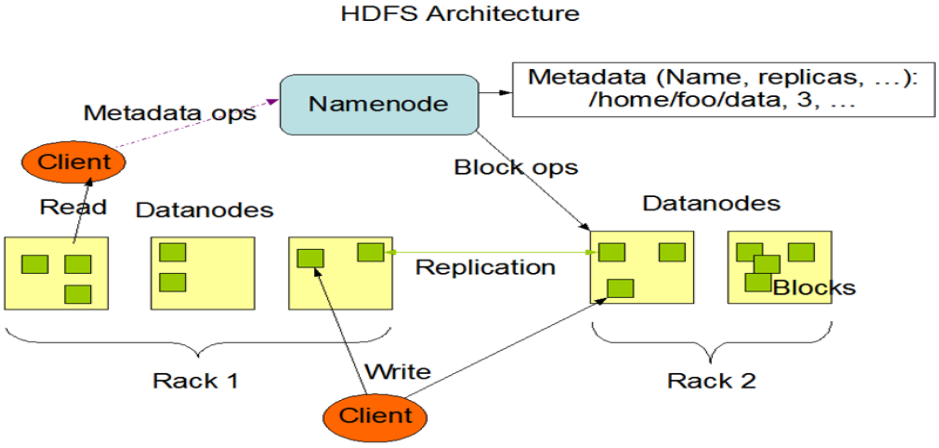
من البيانات الضخمة عبر الرفوف. يمكن تحقيق ذلك عن طريق نشر كتل البيانات المنسوخة على الأجهزة على رفوف مختلفة.



الشكل (5) النسخ المتماثل للبيانات على عقد Hadoop [10]

تعد NameNode و DataNode خوادم سلعية، وعادة ما تكون أجهزة Linux. تدير Hadoop برامج مختلفة على هذه الأجهزة لجعلها NameNodeأو DataNode . تم تصميمHDFS باستخدام لغة Java. يمكن تحويل أي جهاز يمكن تشغيل Java عليه ليكون بمثابةNameNode أو DataNode .

يحتوي العنقود النموذجي على جهاز مخصص يقوم بتشغيل برنامج NameNode فقط. يقوم كل جهاز من الأجهزة الأخرى الموجودة في نظام المجموعة بتشغيل مثيل واحد من برنامج DataNode. يقوم NameNode بإدارة جميع بيانات تعريف HDFS.



الشكل (6) البنية التفصيلية في Hadoop [10]

- **مفهوم MapReduce**

MapReduce عبارة عن إطار عمل برمجي مقدم من قِبل Google لإجراء معالجة متوازية على مجموعات البيانات الضخمة. على افتراض أن سعة تخزين البيانات الضخمة موزعة على عدد كبير من الأجهزة، يحسب كل جهاز البيانات المخزنة محليًا، مما يساهم بدوره في المعالجة الموزعة والمتوازية. هناك جزءان لمثل هذا الحساب – الخريطة والتقليل (Map & Reduce).

DataNodes المعيّنة إلى مرحلة الخريطة، تأخذ بيانات الإدخال الخام وتستند إلى نوع الحساب المطلوب ثم تنتج بيانات وسيطة يتم تخزينها محليًا. عقد التقليل (Reduce Nondes) تأخذ هذه المخرجات الوسيطة وتجمعها لاشتقاق المخرجات النهائية التي يتم تخزينها بعد ذلك في HDFS.

Hadoop يحاول جمع البيانات والحساب. يحاول NameNode بمعرفته بكيفية توزيع البيانات، تعيين المهمة إلى العقدة التي توجد بها البيانات محليًا. يمكن للمبرمجين كتابة خريطة مخصصة وتقليل الوظائف، وتعتني وظيفة MapReduce تلقائيًا بتوزيع المهام وموازنتها عبر مجموعة من آلات السلع في المجموعة الموجودة أسفلها. كما أنها تدير الاتصالات بين الماكينات تاركة المبرمجين للتركيز على وظائف الحد من الخريطة الفعلية.

يستخدم Hadoop إطار الحوسبة المتسامح والموثوق والموزع والمتعامل مع الأعطال لتحليل مجموعات البيانات الضخمة الموزعة على HDFS.

- **مفهوم Hive**

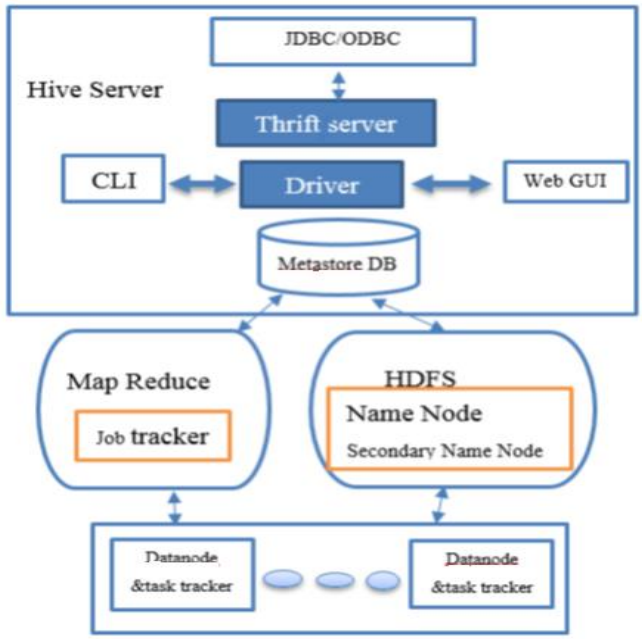
Hive هي بنية تحتية لمستودع البيانات مبنية على أعلى طبقة في نظام Hadoop الموزع، وتوفر أدوات لتمكين ETL بسهولة من ضم مجموعات البيانات المختلفة وتجميعها وتصفيتها. كما يسمح للمبرمجين ببناء وظائف MapReduce مخصصة.

يوفر Hive واجهة استعلام مثل SQL تسمى HiveQL والتي تقوم داخليًا بعمليات (MapReduce).

Hive مفيدة للغاية عند معالجة كميات كبيرة من البيانات. ويعد سهل في الاستخدام لأنه يبسط تعقيد Hadoop. تدعم الكثير من الشركات تقنية Hive، وذلك لسبب بسيط يتمثل في دعم استعلامات SQL القائمة على Hadoop.

- **هيكلية Hive**

الشكل التالي يوضح الهيكلية التفصيلية لـ Hive وعلاقاته مع باقي تقنيات Hadoop



الشكل (7) هيكلية Hive [10]

من الشكل نجد مكونات Hive

- Thrift server: هذا المكون اختياري. يسمح ذلك للعميل البعيد بإرسال طلبات إلى خلية لاسترداد النتائج.
- Driver: يعد برنامج التشغيل مكونًا مهمًا جدًا يأخذ جميع الطلبات من CLI (واجهة سطر الأوامر) أو واجهة ويب أو خادم Thrift، ويقوم بجمع البيانات وتحسينها وتنفيذها.

- Meta Store: يخزن هذا المكون جميع معلومات البنية الخاصة بالجداول والأقسام المختلفة في المستودع بما في ذلك معلومات حول نوع العمود وخصائصه، أيضاً المتسلسلات ومزيلات التسلسل الضرورية لقراءة وكتابة البيانات وملفات HDFS المقابلة حيث يتم تخزين البيانات.

3-3- الحوسبة السحابية والبيانات الطبية الضخمة (دراسة مرجعية):

تشارك الحوسبة السحابية من خلال منصة الإنترنت على نطاق واسع في مهام تجميع وتخزين وإدارة ومعالجة البيانات الضخمة. وبسبب العمليات المتكررة للمعالجة الضخمة للبيانات، يعمل العديد من الباحثين لدعم نموذج برمجة معالجة البيانات الجماعية [11].

نموذج برمجة معالجة البيانات الشامل الأكثر شعبية في العالم هو MapReduce الذي صممه جوجل. يقسم نموذج البرمجة MapReduce المهمة إلى العديد من المهام الفرعية، ويمكن لهذه المهام الفرعية أن تجدول ما بين عقد المعالجة المتاحة، مما يجعل معالجة العقد لهذه المهمة أسرع.

تتمثل الموارد الأساسية للحوسبة السحابية في الخدمات، ويتم إصدار وظائف البرنامج في شكل خدمات، وغالبًا ما تكون هناك حاجة لإيصال رسالة التعاون بين الخدمات المختلفة. لذلك، تعد البنية التحتية للاتصالات الموثوقة والآمنة وعالية الأداء ضرورية لنجاح الحوسبة السحابية.

يسمح نظام الملفات الموزعة للمستخدم بالوصول إلى ملف الخادم البعيد، حيث يشبه زيارة نظام الملفات المحلي، ويمكن للمستخدمين أخذ البيانات المخزنة في خوادم بعيدة متعددة. في الغالب، يحتوي نظام الملفات الموزع على آلية احتياطية، وآلية تتحمل الأخطاء لضمان صحة قراءة البيانات وكتابتها. بناءً على نظام الملفات الموزع ووفقًا لخصائص التخزين السحابي، فإن خدمة التخزين السحابي تجعل منها البنية المناسبة لتكوين وتحسين الخدمات الصحية [11].

يناقش المؤلفون في [12] التحديات والفرص في تصميم أنظمة لاسلكية قابلة للتطوير لاحتضان حقبة "البيانات الضخمة". حيث قوم بمراجعة بنيات الشبكات الحديثة وتقنيات معالجة الإشارات القابلة للتكيف لإدارة حركة البيانات الضخمة في الشبكات اللاسلكية.

وبدلاً من عرض البيانات الضخمة على الأجهزة المحمولة كعبء غير مرغوب فيه، فإنهم يقدمون طرقًا للاستفادة من حركة البيانات الضخمة، من أجل بناء شبكة لاسلكية كبيرة تدرك البيانات مع جودة خدمة لاسلكية أفضل وتطبيقات هواتف محمولة جديدة. تتناول هذه المقالة التحديات والفرص التي نواجهها في عصر البيانات الضخمة اللاسلكية. وقد حددوا العقبات الرئيسية التي تعترض معالجة إشارات البيانات الضخمة وتصميم الشبكات فيما يتعلق بحجم حجم المشكلة وهياكل المشكلات المعقدة. ومع ذلك، فإن الأبحاث المتعلقة بالبيانات الضخمة الخاصة بالاتصالات اللاسلكية والشبكات ليست واعدة فحسب، بل لا مفر منها أيضًا في ضوء الانفجار المستمر لحجم البيانات.

تقترح الورقة [13] حلاً جديدًا للحوسبة السحابية للمستشفيات الحكومية في دول العالم الثالث للوصول بشكل أفضل إلى المعلومات الطبية للمريض.

تقدم الدراسة [14] نموذجاً للحوسبة السحابية لدمج أنظمة معلومات المستشفيات التي تستند إلى بنية موجهة نحو الخدمة. يتيح النموذج المقترح وصولاً ملائمًا لجميع مكونات النظام بما في ذلك مزود الخدمة والمكونات التنظيمية.

تقترح الورقة [15] وتنفذ نظام السجلات الطبية الإلكترونية (CloudeMR) القائم على السحابة لتحسين تقديم نظام الرعاية الصحية في المجتمعات الريفية. وتقدم بنية النظام الشاملة جنبًا إلى جنب مع المكونات الوظيفية.

يقترح البحث [16] تطوير السجلات الصحية الإلكترونية (EHR) للتكامل مع مقدمي الرعاية الصحية في جميع أنحاء الهند ولتطبيقه مع البنية التحتية السحابية.

أجرت الدراسة [17] بحثًا عن تنفيذ الحوسبة السحابية في مؤسسات الرعاية الصحية، حيث ركزت على النموذج المتكامل لنظام معلومات السجلات الطبية للمرضى باستخدام التنسيق القياسي للمعلومات الصحية لتبادل البيانات.

تبحث الورقة [18] في تأثير الحوسبة السحابية على تحسين خدمات الرعاية الصحية. حيث تعرض تفاصيل التصميم المعماري لنظام سجلات الصحة الشخصية يسمى "MedCloud" الذي يستخدم ويدمج الخدمات من النظام البيئي Hadoop بالتزامن مع قواعد الخصوصية والأمن.

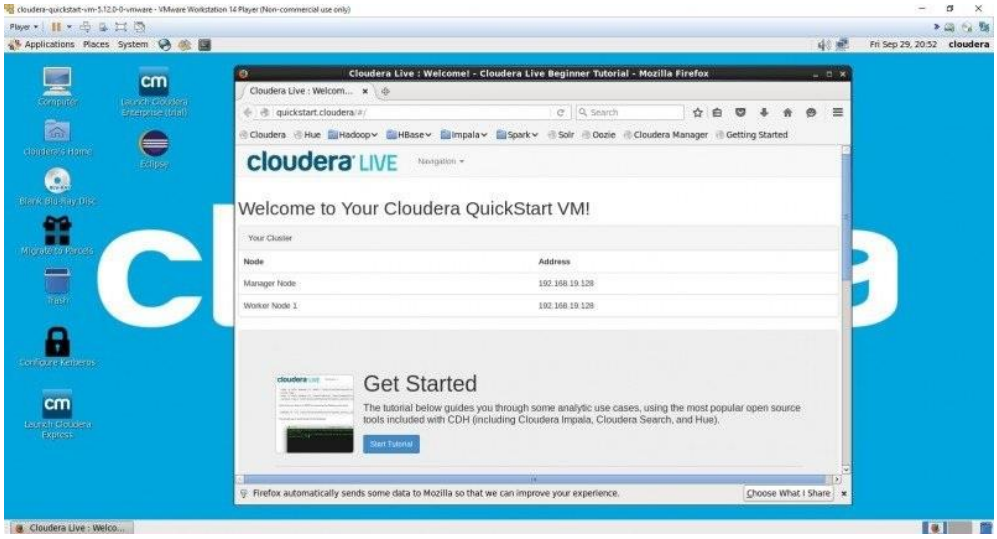
تقدم الورقة [19] هيكلية قائمة على السحابة والتي تربط القطاعات الرئيسية لأي إطار رعاية صحية من المريض والطبيب والأعراض والمرض. تركز الورقة بشكل أساسي على كيفية ترابط هذه الأجزاء وكيفية استنباط البيانات المناسبة منها. كتطبيق، فإنه يظهر واجهة محلل الرعاية الصحية الأساسية التي تأخذ البيانات كمدخلات ويتم التنقيب عن البيانات باستخدام بعض تقنيات استخراج البيانات.

3-4- Cloudera QuickStart VM 5.13.x

CDH (Cloudera Distribution Hadoop):

هو توزيع Apache Hadoop مفتوح المصدر مقدم من Cloudera Inc وهي شركة برمجيات أمريكية.

يحتوي Cloudera quickstart VM على عينة من منصة Cloudera لـ " Big Data". ويتوفر VM من Cloudera بتوزيعات VMware و VirtualBox و KVM ، وكلها تتطلب نظام تشغيل مضيف 64 بت. يعمل جهاز VM هذا على نظام CentOS 6.2 ويشمل CDH4.3 و Cloudera Manager 4.6 و Cloudera Impala 1.0.1 و Cloudera Search .9 Beta.



الشكل (8) بيئة عمل Cloudera

4- التصميم والتنفيذ:

الإطار المفاهيمي للنموذج المقترح

إن الحصول على البيانات الطبية الضخمة، كخطوة أساسية لعمليات معالجة البيانات، يهدف إلى جمع كمية كبيرة ومتنوعة من البيانات الطبية من حيث الحجم والنوع بطرق متنوعة.

بالتالي للتأكد من دقة هذه البيانات وموثوقيتها، يجب تطبيق تقنيات جمع أو استخراج البيانات الموزعة عالية السرعة والموثوقة على أساس النظام الأساسي لتحقيق تكنولوجيا تكامل البيانات عالية السرعة لتحليل البيانات الطبية وتحويلها وتحميلها، بما يراعي ضمان اتساق البيانات وأمانها.

إن تقنية تخزين البيانات الطبية الضخمة وإدارتها تحتاج إلى حل المشكلات على مختلف المستويات، مادياً ومنطقياً.

على المستوى المادي، من الضروري إنشاء نظام ملفات موثوق وموزع، مثل HDFS، لتوفير تقنية تخزين البيانات الطبية الضخمة متاحة دائماً، وقادرة على التسامح مع الأخطاء، وفعالة ومنخفضة التكلفة، إضافة إلى توفير إطار سحابي يوفر السهولة للوصول إلى البيانات زمانياً ومكانياً.

على المستوى المنطقي، من الضروري تطوير تقنية نمذجة البيانات الضخمة لتوفير إدارة ومعالجة البيانات الموزعة غير العلائقية وأيضاً توفير القدرة لتكامل البيانات غير المتجانسة، إضافة إلى القدرة التنظيمية لمثل هذه الأنواع من البيانات.

مما سبق يمكن تقسيم النظام السحابي للبيانات الطبية الضخمة إلى عدة مراحل:

- مرحلة جمع البيانات: مهمتها تنسيق تخزين البيانات الطبية الضخمة والمتنوعة، من مختلف المصادر الطبية، مثل المستشفيات ومراكز الرعاية الصحي.
- مرحلة تخزين البيانات: يتم فيها تخزين جميع البيانات التي تم جمعها من المرحلة السابقة في إطار نظام البيانات الضخمة.
- مرحلة استخراج البيانات واستكشاف المعرفة: وهي أهم مرحلة في هذه المنظومة. حيث تهدف إلى الاستعلام عن البيانات الطبية وتحليلها وتقديم الحلول والمعلومات التي تساعد في اتخاذ القرارات الطبية المناسبة.
- مرحلة التطبيق: وهي صلة الوصل بين المستخدمين والمنظومة، يستطيع من خلالها المستخدمون إدخال أو استعراض البيانات المتاحة حسب طبيعة وسماحيات الاستخدام.

الإطار التحليلي للنموذج المقترح

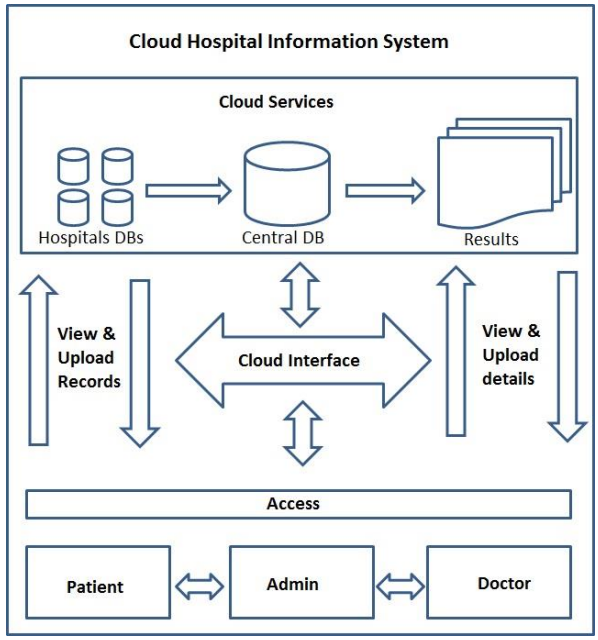
يمكن تقسيم العمل ضمن النموذج المقترح إلى عدة نماذج داخلية توضح تدفق البيانات والعمليات الداخلية ضمن المنظومة بشكل عام.

حيث يمكن القول أن عمل المنظومة يتحدد بشكل رئيسي عبر ثلاثة حلول:

- نموذج الحل السحابي للمنظومة الصحية بإطارها العام.
- نموذج تراسل البيانات ضمن السحابة متضمناً العمل مع البيانات الضخمة.
- نموذج تحليل البيانات الطبية الضخمة.

أولاً: نموذج الحل السحابي للمنظومة الصحية

يوضح الشكل (9) النموذج العام للحل السحابي للمنظومة الصحية، حيث يتضمن العلاقات بين كيانات المنظومة الصحية المقترحة.

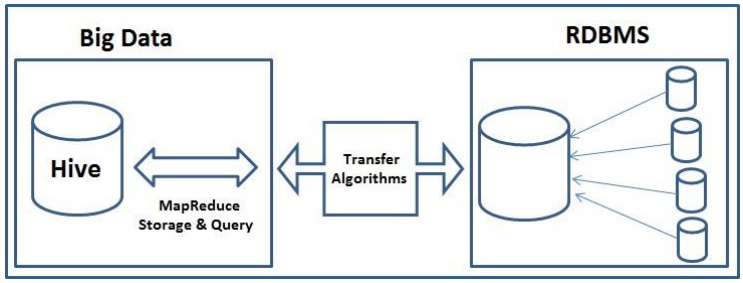


الشكل (9) نموذج الحل السحابي للمنظومة الصحية

ثانياً: نموذج تراسل البيانات الطبية ضمن السحابة

في النموذج المقترح يتم تجميع بيانات محددة من قواعد بيانات المستشفيات والمراكز الصحية، وتخزينها في قاعدة بيانات مركزية. هذه البيانات المحددة هي التي سيتم الاستفادة منها مستقبلاً في منصة بوابة الصحة الالكترونية أولاً، وثانياً في نقلها وتخزينها في إطار عمل البيانات الضخمة من أجل عمليات التحليل الطبي والحصول على المعرفة.

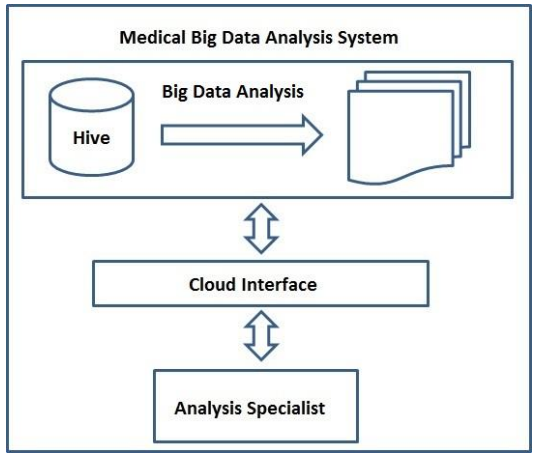
إن نظام التراسل بين قاعدة البيانات العلائقية وبيئة البيانات الضخمة هو ثنائي الاتجاه. وذلك ليتم الاستفادة من خصائص التخزين في بيئة البيانات الضخمة في حال حدوث انهيار أو فقدان في نظم البيانات العلائقية. حيث يتم استعادة البيانات بشكل سريع إلى قاعدة البيانات العلائقية المركزية. كما هو موضح في الشكل (10).



الشكل (10) نموذج تراسل البيانات الطبية ضمن السحابة

ثالثاً: نموذج تحليل البيانات الطبية الضخمة

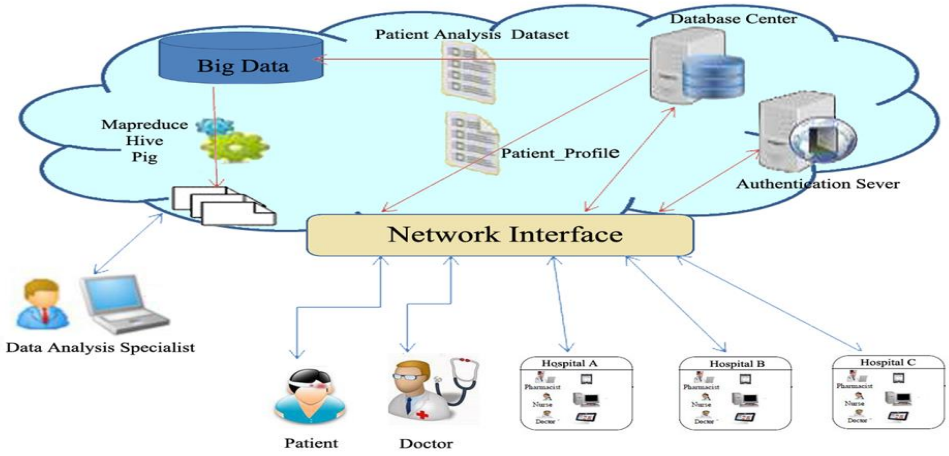
يوضح الشكل (11) نموذج منظومة تحليل البيانات الطبية اعتماداً على تقنيات البيانات الضخمة.



الشكل (11) نموذج تحليل البيانات الطبية الضخمة

هيكلية النموذج المقترح:

يوضح الشكل التالي (12) الهيكلية المقترحة لنموذج النظام السحابي للبيانات الطبية الضخمة. حيث يوضح هذا النموذج مختلف مراحل معالجة البيانات ضمن المنظومة الطبية، بدءاً من إدخال البيانات، مروراً بجمعها من مختلف المصادر، ثم نقلها إلى بيئة نظم البيانات الضخمة وتخزينها، وأخيراً تحليل هذه البيانات للحصول على المعلومات المطلوبة في استكشاف المعرفة الطبية.



الشكل (12) هيكلية نموذج الرعاية الصحية المقترح

تحليل وتطبيق النموذج المقترح

ينقسم تنفيذ النموذج إلى ثلاثة مراحل رئيسية:

1- الحل السحابي للمنظومة الصحية.

2- تراسل البيانات بين قواعد البيانات العلائقية و Hadoop.

3- تحليل البيانات الطبية باستخدام تقنيات البيانات الضخمة.

المرحلة الأولى: الحل السحابي للمنظومة الصحية

يتم تصميم واجهة مستخدم موحدة يمكن تطبيقها من قبل كل المستشفيات والمراكز الصحية، وكل مستشفى يمتلك قاعدة بياناته الخاصة حفاظاً على أمان المعلومات. يتم ربط قواعد بيانات هذه المستشفيات بقواعد بيانات خاصة موحدة تستخدم لهدفين: قاعدة بيانات خاصة بالمرضى يمكنهم الاستعلام منها عن الوضع الصحي وإجراءاته في جميع المستشفيات، وقاعدة بيانات خاصة لتحليل البيانات الطبية المستخلصة من جميع المشافي والتي تساعد في نظم اتخاذ القرار.

تدفق البيانات بين كل قواعد البيانات تلك يخضع لمجموعة من الشروط والخوارزميات: يأخذ المريض موعد دخول للمشفى بعد تحديد بعض المعلومات مثل القسم والطبيب.... إلخ، عند دخول المريض للمشفى سواء دخل بموعد أو لا يتم تسجيل كل المعلومات المتعلقة به في قاعدة البيانات الخاصة بهذه المشفى، وعند الانتهاء من المعاينة

والفحوص وإجراء العلاج يتم تخزين هذه البيانات بالتوازي في قاعدة بيانات المشفى وفي قاعدتي البيانات الخاصتين بالاستعلام والتحليل مع مراعاة السرية والأمان. حيث تخزن هذه المعلومات اعتمادا على الرمز الطبي للمريض دون التطرق إلى معلوماته الشخصية.

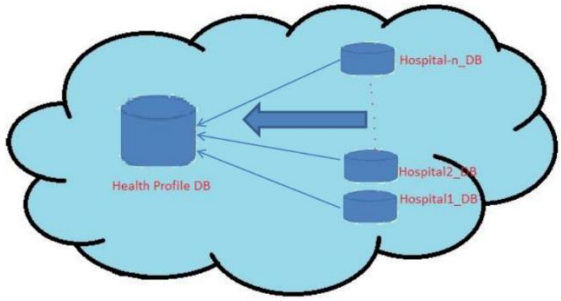
الجدول التالي يوضح قاعدة البيانات المستخرجة من قاعدة البيانات المركزية (من جميع قواعد بيانات المستشفيات) والتي ستستخدم في الاستعلام عن الملف الطبي:

Field	Description of the field and its scope
id	Record id
Medical_id	Patient's medical identity
First Name	First name of the patient
Last Name	Last name of the patient
Age	Age of the patient
Gender	Gender of the patient
City	Patient's address (City)
Country	Patient's Country
Hospital_Name	Name of hospital
Date_of_Hospital_visit	Date on which patient visited the hospital
Temperature	Temperature of the patient during the hospital visit
Blood_group	، B−، B+، A−، A+، O−،Blood group. Scope:O+ AB −،AB+
Blood_ Needed	Whether patient needed blood transfusion (Yes،No)
Disease	Patient's disease
Treatment	Medicine used by the patient
Died	(Yes،No)

الجدول (1) قاعدة البيانات العامة لاستعلام المريض

قاعدة البيانات السابقة (استعلام المرضى) تعتبر قاعدة بيانات عامة يتم ملؤها بشكل أوتوماتيكي باستخدام قوادح خاصة تستخرج البيانات من قواعد بيانات المشافي. الشكل (13).

بحيث أنه بعد أن يتم إنهاء جميع الإجراءات الطبية الممكنة للمريض، وتخزين البيانات للإجراءات المتخذة، يتم أخذ نسخة مختصرة لهذه الإجراءات وتخزينها في قاعدة البيانات الطبية العامة، وذلك للاستفادة من هذه البيانات في العلاجات المستقبلية للمرضى في مستشفيات أو مراكز أخرى. أي أنه يمكن الحصول على الملف الطبي الكامل للمريض من مختلف مراكز الرعاية الصحية من أي مكان وفي أي وقت.



الشكل (13) تراسل البيانات في منظومة الصحة السحابية

المرحلة الثانية: تراسل البيانات بين قواعد البيانات العلائقية و Hadoop

في هذه المرحلة سيتم دراسة نقل سجلات البيانات الطبية من قاعدة البيانات الصحية المركزية إلى قواعد بيانات خاصة بالبيانات الضخمة لتخزينها وإجراء مختلف التحليلات عليها، مع مراعاة الخصوصية للمرضى، أي أنه سيتم فقط نقل البيانات العامة دون الشخصية، والتي يمكن أن تفيد في عمليات التحليل مثل العمر، المدينة، الجنس....

الجدول التالي يوضح قاعدة البيانات المستخرجة من قاعدة البيانات المركزية إلى (Hadoop) والتي سيتم نقلها إلى بيئة عمل البيانات الضخمة لإجراء مختلف التحليلات عليها:

Patient Analysis Dataset:

Field	Description of the field and its scope
id	Record id
Age	Age of the patient
Gender	Gender of the patient
City	Patient's address (City)
Country	Patient's Country
Date_of_Hospital_visit	Date on which patient visited the hospital
Temperature	Temperature of the patient during the hospital visit
Blood_group	Blood group. Scope:O+، O−، A+، A−، B+، B−، AB+، AB −
Blood_ Needed	Whether patient needed blood transfusion
Disease	Patient's disease
Treatment	Medicine used by the patient
Died	(Yes،No)

الجدول (2) قاعدة بيانات المرضى الخاصة بالتحليل

أغلب الدراسات التي تتعلق بتحليل البيانات الطبية الضخمة اعتمدت على قاعدة بيانات جاهزة تم تحميلها بشكل يدوي إلى بيئة عمل البيانات الضخمة ثم تطبيق التقنيات عليها. وفي بعض الحالات يتم استخدام تطبيقات جاهزة للنقل المستمر والمتزامن.

في حالة النموذج المقترح تم الحصول أيضاً على قاعدة بيانات طبية جاهزة اخترنا منها مجموعة محددة من الأعمدة التي تلائم قاعدة البيانات الخاصة بالتحليل مع إجراء بعض

التعديلات مثل إضافة عمود الرمز الطبي وتغيير بعض البيانات مثل المدن والعناوين والتي توافق الجدول السابق...

ثم تخزين هذه السجلات في قاعدة البيانات المركزية (تم تخزين حوالي 20000 سجل) لتطبيق عمليات النقل الأولي وخوارزميات النقل المتزامن عليها بشكل مبدئي. كما تم تحميل قواعد بيانات طبية ضخمة أخرى لإجراء مختلف التحليلات في بيئة البيانات الضخمة لضمان مقاربة صحة النتائج.

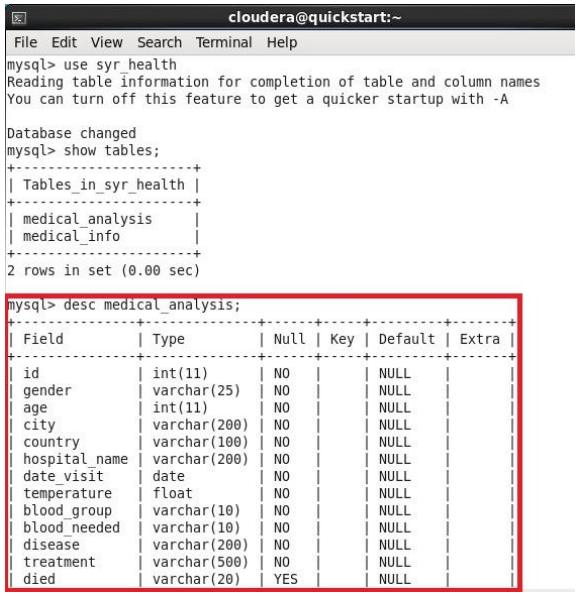
تم اتباع عدد من الخطوات من أجل تجهيز قاعدة البيانات الطبية على محرك قواعد البيانات العلائقية RDBMS، قبل العمل على نقلها إلى Hive في Hadoop.

وهذه الخطوات كالتالي:

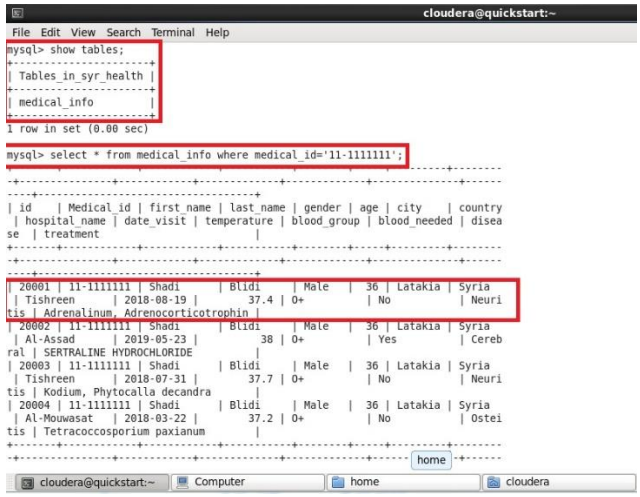
بعد إنشاء قاعدة البيانات الخاصة بالسجلات الصحية (syr_health)، تم تحميل قاعدة البيانات التي تم تجهيزها مسبقاً والتي تضم حوالي (20000) سجل صحي. وتم توزيعها إلى جدولين. الجدول الأول خاص بالاستعلام من بوابة الصحة الالكترونية وهو خاص بالمرضى والمراكز الصحية (medical_info).

أما الجدول الثاني فهو خاص بالتحليل، حيث أنه لا يضم معلومات شخصية، فقط يضم السجلات التي تهم في مجال تحليل البيانات الطبية بشكل عام (medical_analysis). وهذا الجدول هو الذي سيتم التعامل معه في نقل البيانات إلى Hadoop. حيث سيتم هناك جميع عمليات التحليل والتقييم.

الشكل (14) تحميل البيانات إلى قاعدة البيانات العلائقية الطبية



الشكل (15) بنية جدول السجلات الطبية الخاص بالتحليل



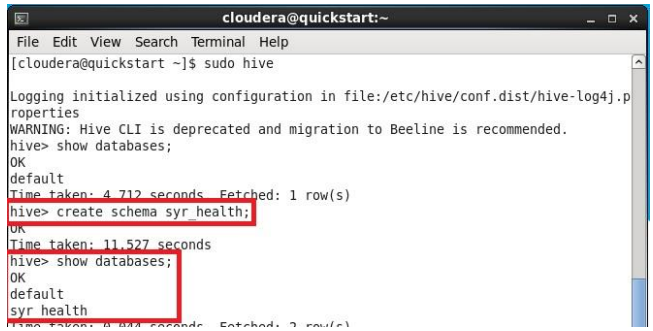
الشكل (16) الاستعلام عن البيانات الطبية

الآن بعد تحميل قاعدة بيانات السجلات الطبية، هنالك عدة خطوات لنقلها إلى Hadoop.

Hive عبارة عن إطار عمل (جدول افتراضي) لتخزين البيانات مبني على الجزء العلوي من Hadoop. وستتم إدارة هذه الجداول بالكامل من قبل محرك قاعدة بيانات Hive. يتم تخزين جميع الجداول في الدليل الافتراضي التالي: (/user/hive/warehouse).

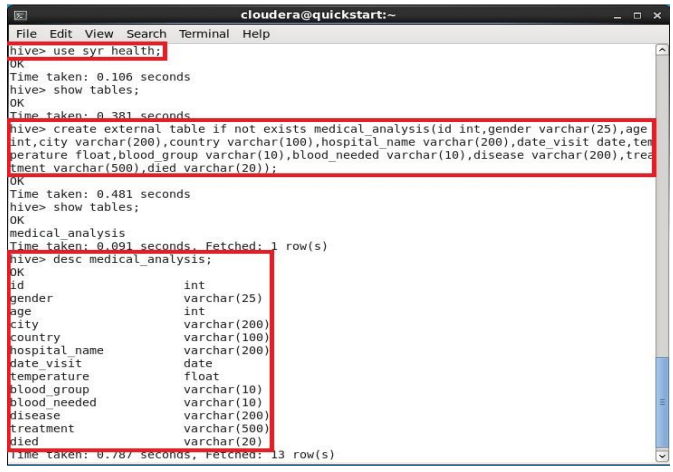
سيتم إنشاء جدول لبيانات السجلات الطبية في Hive ، لاستقبال السجلات القادمة من قاعدة البيانات العلائقية وتخزينها لإجراء التحليلات عليها لاحقاً. كما يلي:

إنشاء قاعدة البيانات الطبية في Hive:



الشكل (17) إنشاء قاعدة البيانات الطبية في Hive

إنشاء جدول السجلات الطبية في Hive بهيكلية متوافقة مع جدول السجلات الطبية في قاعدة البيانات العلائقية:

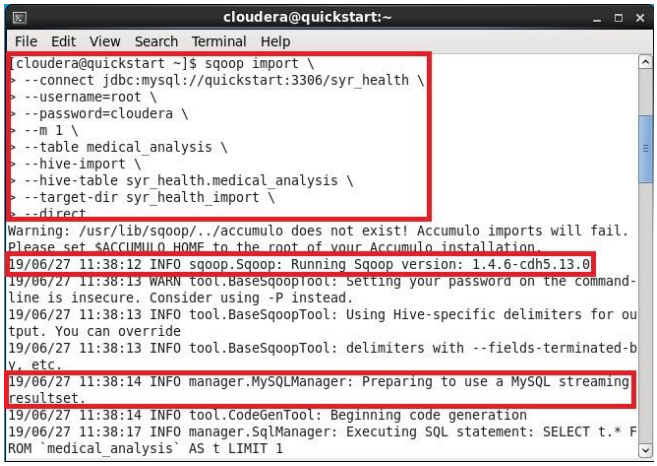


الشكل (18) إنشاء جدول السجلات الطبية في Hive

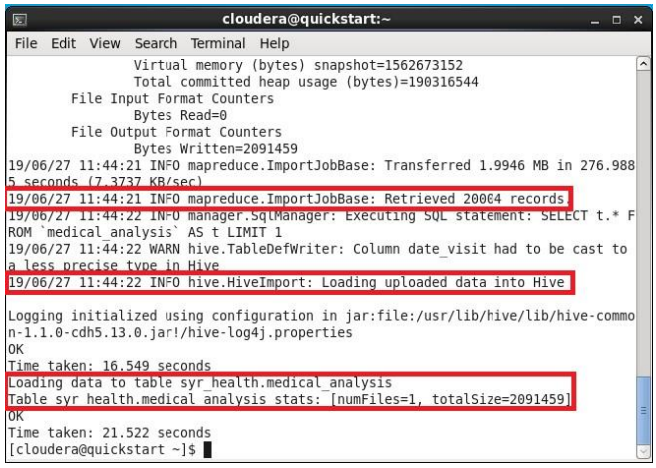
والآن نستخدم الأداة Sqoop لاستيراد البيانات من MySql إلى Hive، وفق الخوارزمية التالية:

```
sqoop import \
--connect jdbc:mysql://quickstart:3306/syr_health \
--username=root \
--password=cloudera \
--m 1 \
--table medical_analysis \
--hive-import \
--hive-table syr_health.medical_analysis \
--target-dir /syr_health_import \
--direct
```

حيث تقوم هذه الخوارزمية بنقل السجلات الطبية من الجدول medical_analysis في MySql إلى الجدول medical_analysis في Hive.

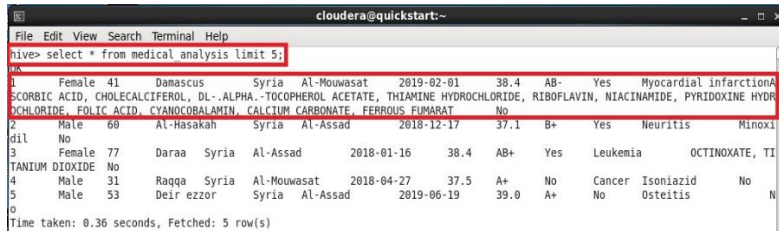


الشكل (19) خوارزمية استيراد البيانات من MySql إلى Hive



الشكل (20) عمليات Sqoop في استيراد البيانات

بعد الانتهاء من عملية الاستيراد يتم اختبار وصول البيانات بالشكل السليم.



الشكل (21) اختبار الاستعلام عن البيانات في Hive

بالتالي أصبحت قاعدة بيانات السجلات الطبية في Hadoop جاهزة لإجراء عمليات التحليل المختلفة.

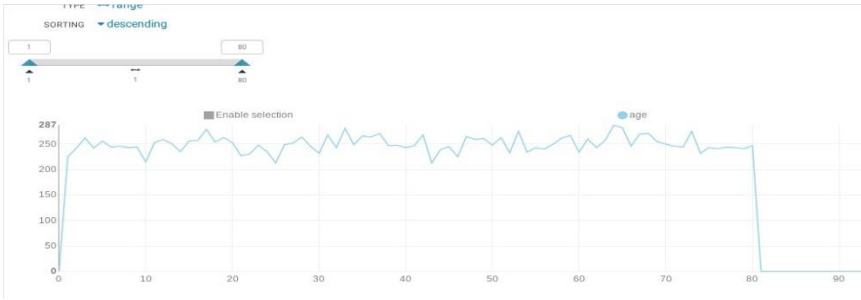
المرحلة الثالثة: تحليل البيانات الطبية الضخمة

بعد أن تمت معالجة نقل البيانات الطبية إلى نظام تخزين البيانات الضخمة في Hadoop تأتي مرحلة تحليل هذه البيانات وإظهار النتائج.

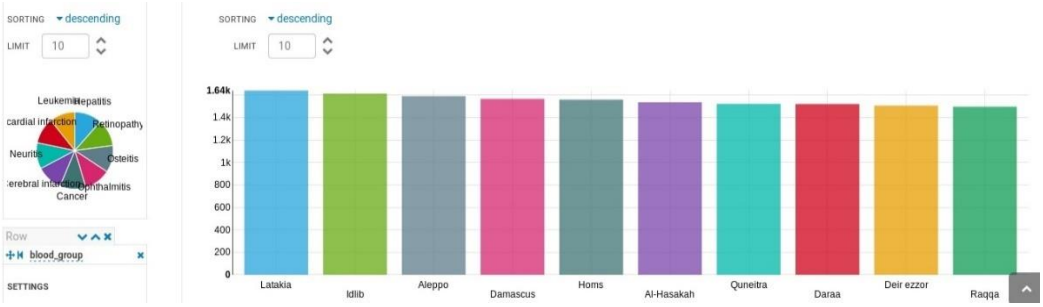
فيما يلي سيتم استخدام أدوات بيانية لتحليل وتمثيل البيانات تفيد في مجال المقارنات والأبحاث المتقدمة.

i- باستخدام الأداة (Hue Dashboard)

Hue Dashboard هي أداة تحليل بيانات مدمجة مع منظومة (Hue UI)، تقدم العديد من الإمكانيات المتقدمة في مجال الاستعلام التحليلي الدقيق والواسع النطاق. قمنا باستخدام هذه الأداة لدراسة بعض الحالات المتعلقة بقاعدة البيانات الطبية المخزنة كما يلي:



الشكل (22) تمثيل بياني لمعدل أعداد المرضى حسب العمر



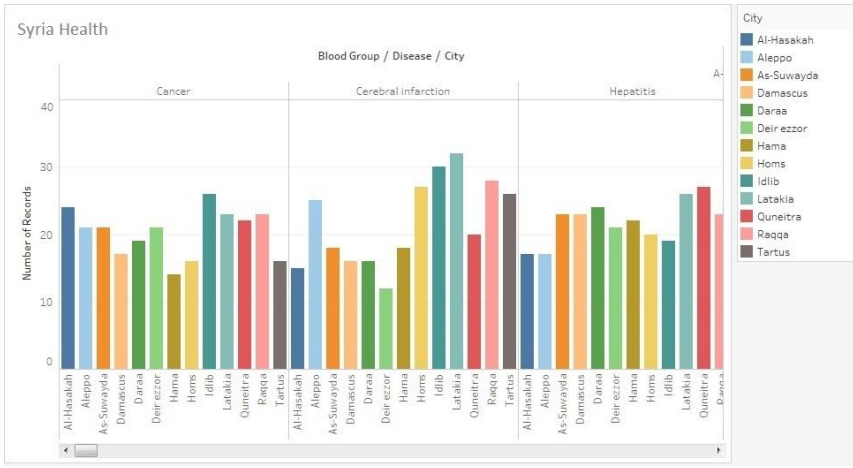
الشكل (23) تمثيل بياني لأعداد المرضى في كل محافظة، مع نسبة كل مرض

ii- باستخدام برنامج (Tableau)

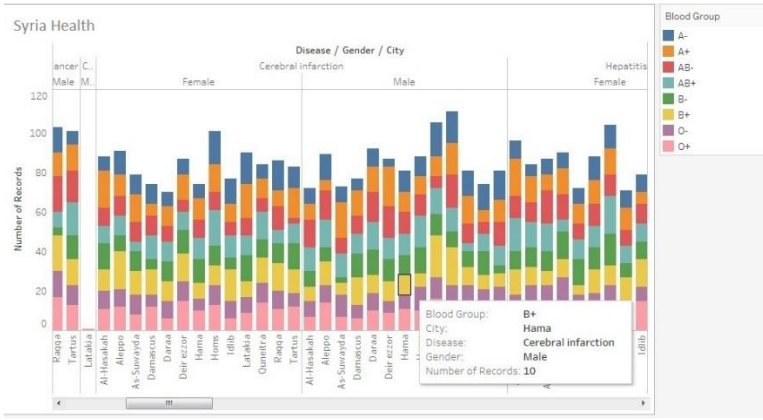
Tableauهي أداة قوية لتمثيل البيانات وتطوير ذكاء الأعمال، يأتي مع واجهة سهلة الاستخدام لإجراء تحليل عميق. معTableau ، يمكن للمستخدم الحصول بسرعة على تحليلات قيمة من فضاء بيانات Hadoop الواسع. ويتيح Tableau أيضًا معرفة عميقة بلغات الاستعلام المتقدمة ويجعل البيانات الضخمة أكثر سهولة من خلال التحليل البصري.

Hadoop هو أحد مصادر البيانات لـTableau ، وباستخدام روابط أصلية بسيطة، يمكن توصيل Tableau بسهولة بـHadoop، Hive، Impala .

فيما يلي مجموعة من التحليلات التي قمنا بها على قاعدة البيانات الطبية باستخدام الأداة (Tableau).



الشكل (24) توزع المرض في كل محافظة حسب الزمرة الدموية

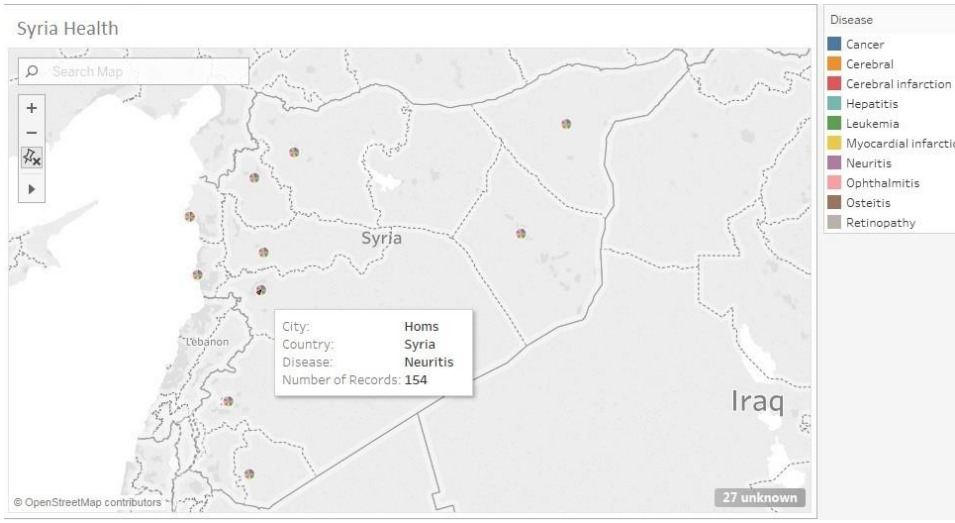


الشكل (25) توزع المرض في كل محافظة حسب الجنس والزمرة الدموية

الشكل (26) نسبة الوفيات السنوية حسب كل مرض في محافظة دمشق

الشكل (27) نسبة الوفيات السنوية في كل محافظة حسب المرض

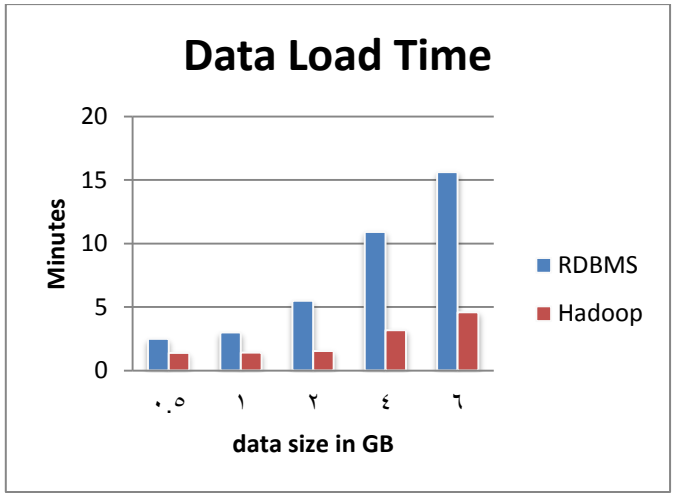
- خريطة نسب توزع الأمراض في كل محافظة



الشكل (28) خريطة نسب توزع الأمراض في كل محافظة

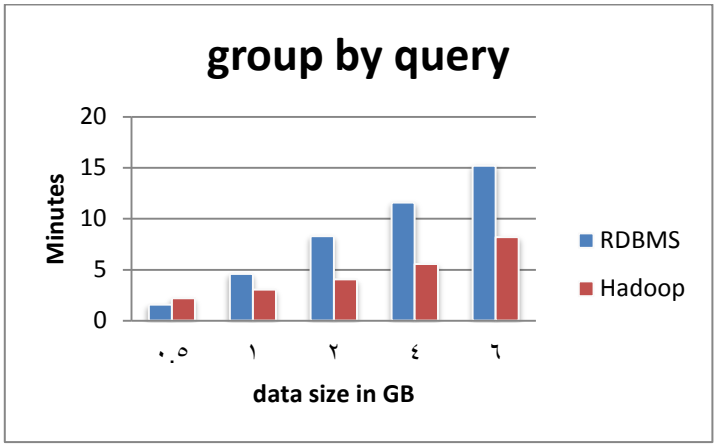
5- مقارنات واستنتاجات وتوصيات:

مقارنة أداء تحميل البيانات الطبية بين RDBMS و (Hive في Hadoop)

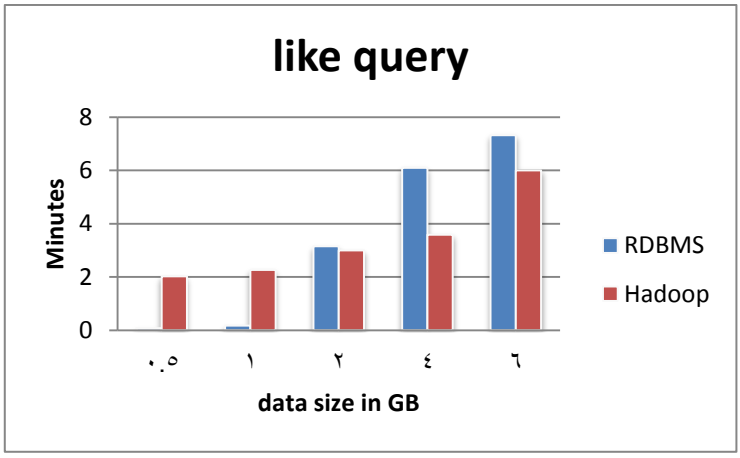


الشكل (29) مقارنة وقت تحميل البيانات بين RDBMS و Hadoop

مقارنة أداء تنفيذ الاستعلامات الطبية بين RDBMS و Hadoop



الشكل (30) مقارنة تنفيذ استعلام group by بين RDBMS و Hadoop



الشكل (31) مقارنة تنفيذ استعلام Like بين RDBMS و Hadoop

تُظهر التحليلات السابقة زيادة كبيرة في الأداء مع Hadoop بسبب بنيتها الموزعة مقابل RDBMS ذات الجزء الواحد. بالتالي وجد أن الفرق في الوقت اللازم لتحميل البيانات على RDBMS مقابل Hadoop ينمو بشكل كبير مع حجم البيانات.

كما نجد أن فاعلية Hadoop في تنفيذ الاستعلامات على البيانات الضخمة تزيد كلما تزايدت هذه البيانات، على عكس RDBMS. وهذا من شأنه زيادة الأفضلية لنظم استعلامات البيانات الضخمة في الاستعلام عن البيانات الطبية الضخمة والمعقدة الهيكلية، مما يساعد في تسريع الوصول إلى التحليلات المناسبة وبالتالي الوصول الأسرع إلى اتخاذ القرارات الطبية المتصلة.

بالتالي نجد بأن تخزين البيانات الطبية ومعالجتها في بيئة Hadoop أثبت أفضلية مطلقة في التعامل مع هذه البيانات من حيث التخزين والاستعلام والتحليل.

التوصيات:

في ظل توفر وسائل الاتصال والسرعات العالية للإنترنت، لابد من ربط المراكز الصحية والمستشفيات الحكومية منها والأهلية بمركز رئيسي واحد للبيانات وذلك لمتابعة حالة المرضى والاطلاع على تاريخهم الصحي في حال قام مريض بزيارة أكثر من طبيب، أو أكثر من مشفى. ولتقديم هذه الخدمة يُفضل توظيف التقنيات الحديثة التي تؤمن عمل مثل هذه المنظومات، كتقنيات البيانات الضخمة والحوسبة السحابية.

بعد ذلك، سيتكون لدينا كم هائل من البيانات والإحصاءات الصحية للمرضى، بالتالي يستطيع متخذو القرار الاستفادة منها لإصدار الأحكام والقرارات الملامسة للواقع. وذلك بعدة طرق كإخراج التقارير الدورية لحالة المرضى بشكل عام أو محدد، وكذلك معرفة أعداد المرضى لحالات خاصة في فترات زمنية معينة لأخذ التدابير الوقائية والاستعدادية، ويمكن أيضاً تطبيق وسائل تحليل واستخراج البيانات الطبية للاستفادة من هذا الكم الهائل من البيانات.

وتتلخص وسائل تحليل واستخراج البيانات الطبية في عملية استخراج البيانات والإحصاءات المترابطة والتي تشترك بخصائص وصفات متشابهة من بيانات تاريخية

متوفرة لتشكل المعلومات، وتقوم بمعالجتها وتحليلها والربط بين العوامل المشتركة بينها ومن ثم اكتشاف المعرفة الطبية. فعند توفر كل هذا الكم الكبير من البيانات الصحية، يمكن اكتشاف العوامل المشتركة المسببة لأمراض محددة، فيتم تصنيف الأمراض المنتشرة ومدى خطورتها عند تناول غذاء محدد وفي وقت معين لأشخاص في مرحلة عمرية متشابهة، وكذلك يمكن إيجاد أعراض الأمراض المنتشرة ومن ثم تجميع وعنقدة المناطق والمدن حسب وجود الأعراض فيها ليتم معالجتها حسب الأولوية وهكذا.

إن من الجوانب المهمة والمساهمات الأساسية لهذا البحث هو أنه يفتح الآفاق لدراسات مستقبلية يمكن العمل عليها من أجل تحسين النتائج، وفتح باب اعتماد النموذج المقترح في الدراسة وتطويره ليلائم بيئات العمل المختلفة، وإمكانية تطبيقه في قطاعات عمل أخرى مع إجراء بعض التعديلات.

المراجع العلمية:

[1] BOLLINENI, P. K., & NEUPANE K, 2011- Implications for adopting cloud computing in e-Health.

[2] MEMON A. , NAEEM M. R., TAHIR M., AAMIR M., & WAGAN A., 2014- A New Cloud Computing Solution for Government Hospitals to Better Access Patients' Medical Information, American Journal of Systems and Software, 2(3), 56-59.

[3] SHAH T., RABHI F., & RAY P., 2015- Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions, Cluster Computing, 18(1), 351-367.

[4] MICHAEL A., ARMANDO F., REAN G., ANTHONY D. J., RANDY K., ANDY K., ... & MATEI Z, 2010- A view of cloud computing, Communications of the ACM, *53*(4), 50-58.

[5] AL-REFAI A., & PANDIRI S., 2011- Cloud Computing: Trends and Performance Issues.

[6] WIKIPEDIA. Big data, 2014- http://en.wikipedia.org/wiki/Big_data، accessed April 2014.

[7] VITRIA. The Operational Intelligence Company, 2014- http://blog.vitria.com، accessed

April 2014.

[8] APACHE HADOOP. What Is Apache Hadoop?, 2014- http://hadoop.apache.org/، accessed

April 2014.

[9] WIKIPEDIA. Apache Hadoop, 2014- http://en.wikipedia.org/wiki/Apache_Hadoop، accessed April 2014.

[10] WHITE T., 2012- Hadoop: The definitive guide. " O'Reilly Media, Inc.".

[11] CATLEY C., SMITH K., MCGREGOR C., & TRACY M.,2009- Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data Streams: A neonatal intensive care unit case study, In 2009 22nd IEEE International Symposium on Computer-Based Medical Systems, (pp. 1-5). IEEE.

[12] JAVIER A, CARMEN C. Y. POON, ROBERT D. MERRIFIELD, STEPHEN T. C. WONG, GUANG-ZHONG YANG 2015- "Big Data for Health", IEEE Journal of biomedical and health informatics, Vol.19 No.4.

[13] DHIRAJ D. J., KOMAL S. B., TRUPTI V. P., SOHAIL SH., 2018-'Medical Data Mining for General Hospital', International Research Journal of Engineering and Technology (IRJET),Volume: 05 Issue: 05.

[14] SETAREH S., REZAEE A., FARAHMANDIAN V., HAJINAZARI P., & ASOSHEH A., 2014 - A cloud-based model for hospital information systems integration. IEEE, In 7'th International Symposium on Telecommunications (IST'2014) (pp. 695-700).

[15] BOYINBODE O., & TORIOLA G., 2015-CloudeMR: A Cloud Based Electronic Medical Record System. International Journal of Hybrid Information Technology, 8(4), 201-212.

[16] KAVITHA R., KANNAN E., & KOTTESWARAN S., 2016- Implementation of cloud based electronic health record (EHR) for Indian healthcare needs. Indian Journal of Science and Technology, 9(3), 1-5.

[17] PARDAMEAN B., & RUMANDA R. ,2011- Integrated model of cloud-based E-medical record for health care organizations. In 10th WSEAS international conference on e-activities (pp. 157-162).

[18] SOBHY D., EL-SONBATY Y., & ELNASR M. A., 2012- MedCloud: healthcare cloud computing system. IEEE, In 2012 International Conference for Internet Technology and Secured Transactions (pp. 161-166).

[19] PAREKH M., & SALEENA B., 2015- Designing a cloud based framework for healthcare system and applying clustering techniques for region wise diagnosis. Procedia Computer Science, 50, 537-542.