جامعة القاهرة

كلية الحاسبات والمعلومات

قسم علوم الحاسب

الفهرسة لتحسين تحليل البيانات الضخمة

إعداد

حسين شحاتة عبد العزيز

رسالة مقدمة إلى كلية الحاسبات والمعلومات

جامعة القاهرة

كجزء من متطلبات الحصول على درجة الماجستير

في علوم الحاسب

تحت إشراف

أ.د/ فاطمة عبدالستار حسن عمارة

قسم علوم الحاسب

كلية الحاسبات والمعلومات

جامعة القاهرة

د/ محمد حلمى خفاجى

قسم علوم الحاسب

كلية الحاسبات والمعلومات

جامعة الفيوم

كلية الحاسبات والمعلومات

جامعة القاهرة

جمهورية مصر العربية

يناير ٢٠١٦

الملخص العربى

تعد معالجة البيانات الضخمة واحدة من أعقد العمليات الحسابية حاليا نظرا للنمو الضخم فى حجم البيانات وابعادها وكذلك انماط البيانات الغير مهيكله. الدراسات الحديثة تتوقع ازدياد مضطرد فى حجم البيانات يصل الى zettabytes . هذه البيانات يتم انتاجها عن طريق اجهزة الاستشعار , شبكات التواصل الاجتماعى وكذلك سجلات البيانات الناتجة عن مكينات المصانع.

تعد قواعد البيانات العلائقية غير مناسبة لتحليل تلك البيانات الضخمة وكذلك تطبيق الفهرسة او الارتباطات العلائقية عليها نظرا لانها تؤثر بالسلب على الوقت اللازم لتحميل ومعالجة تلك البيانات وكذلك استرجاعها.

اصبحت MapReduce من اهم الانظمة البرمجية الفعالة فى معالجة وتحليل البيانات الضخمة وعلى الجانب الاخر Hadoop يعد من اهم البرمجيات المبنية على طريقة Map/Reduce لتحليل ومعالجة البيانات الضخمة. وتعد HIVE من البرمجيات التى تعمل كمحرك قاعدة بيانات والتى تعمل بدون روابط علائقية او فهرسة للبيانات والتى بنيت كمترجم يحول جملة SQL الى مهمة MapReduce يمكنها ان تنفذ بواسطة Hadoop .لكن مراحل تنفيذ استعلامات الربط (الشبكية) من اعقد المراحل التى يقوم بها HIVE حيث انه يستهلك الوقت وكذلك مساحات تخزين مؤقته وكذلك ذاكرة عشوائية كبيرة خاصة ب JAVA Heap والتى قد تؤدى الى حدوث تدفق زائد فى الذاكرة المحجوزة اثنا التشغيل مما يؤدى الى اعادة تشغيل المهمة مرة اخرى.

تعد قواعد البيانات النجمية من اكثر انماط قواعد البيانات التى تحتاج الى استعلامات ربط من اجل جمع المعلومات واستخراج التقارير لصانعى القرار مما يجعلها بطئية نسبيا فى حالة استخدام HIVE كمحرك قاعدة بيانات.

يوجد العديد من الطرق التى حاولت اضافة الفهرسة الى نظام الملفات HDFS عن طريق اضافة الفهرس بطريقة ثابته او متغيرة او اثناء عملية التشغيل او قبلها حتى تتمكن من تحسين الوقت اللازم لجمع المعلومات المطلوبة من استعلامات الربط ولكنها لا تزال تعانى من البطئ خاصة فى

حالة استعلامات الربط لانها لا تزال تعتمد على مراحل تنفيذ HIVE لاستعلامات الربط والتى مازالت معقدة.

فى هذه الرسالة سنقوم بعرض طريقتين مختلفتين كتعديل وتحسين لمراحل تنفيذ استعلامات الربط فى HIVE من أجل تحسين كفاءة HIVE وكذلك تقليل الوقت المطلوب من اجل تنفيذ هذه الاستعلامات .وقد بنيت هاتين الطريقتين باسلوب الفهرسة الثابت والمعد سلفا والذى نقوم ببناءه فى مرحلة تحويل البيانات من نمط قاعدة البيانات النجمية الى نمط الجدول الكبير الذى يحوى كل بيانات قاعدة البيانات فى جدول واحد لتصبح فى حالة ربط مبدئ وكأننا قمنا بتنفيذ استعلام ربط لجمع كل البيانات فى جدول واحد كبير واضافة حقل جديد لها الجدول يمكننا من ربط الصفوف ببعضها ويمثل هذا الحقل (المؤشر او الفهرس) الذى يربط البيانات بعضها ببعض.

واحدة (Join Once) ويمكن استخدام هذه البيانات المجمعه اكثر من مرة بعد ذلك (Use Many) والتى قمنا بأختصارها الى (JOUM: Join Once Use Many).

يوفر دمج البيانات فى جدول واحد الوقت اللازم لعمليات الاستعلام وكذلك الذاكرة العشوائية المطلوبه والمساحة التخزينية المؤقته التى يحتاجها HIVE اثنا تنفيذ استعلامات الربط ، ومن اهم مميزات الطريقتين التى نقوم بعرضهم انهما لا يقوم بتعديل فى الكود المصدرى ل HIVE ولكنهما تضيف طبقات من الاكواد بعضها يسبق مراحل HIVE والاخرى تليها مما يجعلها ان تعمل مع الانظمة القديمة التى بنيت باستخدام HIVE دون اى تعديل مطلوب لهذه الانظمة مما يجعل هاتين الطريقتين تمثل اسلوب الفائز-الفائز لكل الاطراف.

تسمى الطريقة الاولى JOUM والتى تعتمد على جدول كبير واحد لكل البيانات والثانية Keys/Facts والتى تعتمد على تحويل الجدول الواحد الى جدولين احدهم يحمل المفاتيح الخاصة بالبيانات والاخرى يحمل الحقائق الخاصة بالبيانات المكررة على المفاتيح التى اضيفت من قبل فى الجدول الاول وتساعد عملية التقسيم هذه فى تحسين الوقت وتقليل المساحة التخزينية الثابتة للبيانات.

قمنا بتقييم الطريقتين السابقتين بإستخدام معيار البيانات TPC-H والتى بنيت بنمط قواعد البيانات النجمية والتى تعد من أكثر معايير البيانات المستخدمه ; وتظهر نتائج التقييم ان الطريقتين تعديا اداء

HIVE بنسبة (53.1%,49.8%) من نسبة الوقت اللازم ل HIVE وكذلك (27%,28%) من نسبة المساحة التخزينية المؤقتة اللازمة ل HIVE وكذلك (26%,19%) من نسبة المساحة التخزينية الثابتة للبيانات كحمل زائد وكذلك عدد اقل من عمليات مشاكل الذاكرة.

فى المجمل تعد الطريقيتين مناسبتين لتحليل البيانات الضخمة على الرغم من وجود زيادة فى حجم البيانات عند تحويلها من نمط قاعدة البيانات النجمية الى الجدول الواحد الكبير وذلك بسب حقل الفهرسة الذى تم اضافته الى البيانات وتعد هذه الزيادة مقبولة جدا اذا ما قورنت بنسبة التوفير فى المساحة المؤقته والتى نحتاجها اثنا كل عملية التشغيل لاستعلام من استعلامات الربط.