

ملخص البحث رقم (4)

عنوان البحث باللغة العربية

iHOME: تحسين استعلام JOIN المستند إلى الفهرس لتخزين البيانات الضخمة المحدود

اسماء المؤلفين:

RadhyaSahal ▪ MarwahNihad ▪ Mohamed H. Khafagy ▪ Fatma A. Omara

مكان النشر وتاريخه:

Journal of Grid Computing Volume 16, Issue 2, Pages 345-380, 2018

ملخص البحث باللغة العربية :

أصبح تحسين الاستعلام في البيانات الضخمة اتجاهًا بحثيًا واعدًا نظرًا لتشعب أنظمة تحليل البيانات الضخمة مثل نظام Hadoop. حيث يزداد تحسين الاستعلام صعوبة في تنفيذ استعلامات JOIN من خلال لغة استعلام Hadoop ، Hive ، عبر مخازن بيانات كبيرة محدودة. وفقًا لعملنا السابق ، تم تقديم تحسين HiveQL لاستعلام JOIN عبر نظام بيئة متعددة الجلسات (HOME) عبر نظام Hadoop لتحسين أدائه عن طريق تخزين النتائج الوسيطة لتجنب الحسابات المتكررة. تعتبر زيادة الوقت ومخازن البيانات الكبيرة العيب الرئيسي لنظام HOME ، خاصة في حالة استخدام مخازن فعلية إضافية أو استئجار مخازن افتراضية إضافية. في هذا البحث ، يُقترح نظام قائم على الفهرسة لإعادة استخدام البيانات يسمى فهرسة HiveQL Optimization لـ JOIN من خلال بيئة بيانات كبيرة متعددة الجلسات (iHOME) للتغلب على النفقات العامة HOME من خلال تخزين فهارس الصفوف المرتبطة فقط بدلاً من تخزين النتائج الوسيطة الكاملة مباشرة. علاوة على ذلك ، يعالج نظام iHOME المقترح ثماني حالات من استعلامات JOIN التي تم تصنيفها إلى ثلاث مجموعات ؛ مشابه لـ iHOME و Compute-on-iHOME و Filter-of-iHOME. وفقًا للنتائج التجريبية لنظام iHOME باستخدام معيار TPC-H ، فقد وجد أن وقت تنفيذ ثمانية استعلامات JOIN باستخدام iHOME على Hive قد تم تقليله. أيضًا ، يتم تقليل حجم البيانات المخزنة في نظام iHOME بالنسبة إلى نظام HOME ، بالإضافة إلى تخزين البيانات الكبيرة. لذلك ،

من خلال زيادة حجم البيانات المخزنة ، يضمن نظام iHOME قابلية التوسع في المساحة ويتغلب على قيود التخزين.

ملخص البحث باللغة الانجليزية :

Query optimization in Big Data becomes a promising research direction due to the popularity of massive data analytical systems such as Hadoop system. The query optimization is getting hard to efficiently execute JOIN queries on top of Hadoop query language, Hive, over limited Big Data storages. According to our previous work, HiveQL Optimization for JOIN query over Multi-session Environment (HOME) system has been introduced over Hadoop system to improve its performance by storing the intermediate results to avoid repeated computations. Time overheads and Big Data storages limitation are considered the main drawback of the HOME system, especially in the case of using additional physical storages or renting extra virtualized storages. In this paper, an index-based system for reusing data called indexing HiveQL Optimization for JOIN over Multi-session Big Data Environment (iHOME) is proposed to overcome HOME overheads by storing only the indexes of the joined rows instead of storing the full intermediate results directly. Moreover, the proposed iHOME system addresses eight cases of JOIN queries which classified into three groups; Similar-to-iHOME, Compute-on-iHOME, and Filter-of-iHOME. According to the experimental results of the iHOME system using TPC-H benchmark, it is found that the execution time of eight JOIN queries using iHOME on Hive has been reduced. Also, the stored data size in the iHOME system is reduced relative to the HOME system, as well as, the Big Data storage is saved. So, by increasing stored data size, the iHOME system guarantees the space scalability and overcomes the storage limitation

البحث مشتق من رسالة علمية

يقع البحث ضمن مجالات البحث بالقسم العلمي

عميد الكلية

أ.د / نبيلة محمد حسن