أطار لأدراة وتحليل قواعد البيانات العملاقة

ابراهيم محمد الحداد

الملخص

إن رقمنة البيانات ،الأفراد و الشركات يولدون كمية هائلة من البيانات بصورة يومية و في جميع أنحاء العالم. رقمنة البيانات ، تحويل البيانات التناظرية إلى البيانات الرقمية ، سهل إطلاق العديد من مشروعات الرقمنة مثل مشروع مكتبة كتب Google التي تم فيها مسح ملايين الكتب ضوئيًا وتخزينها كمكتبة إلكترونية. إن المليارات من الهواتف المحمولة والأجهزة اللوحية وأجهزة الكمبيوتر المحمولة المزودة بأجهزة استشعار مثل الكاميرات والتي تشغل تطبيقات التواصل الاجتماعي مع اتصالها بالإنترنت يؤدي لتوليد كمية هائلة من البيانات. علاوة على ذلك ، تولد الشركات والمؤسسات كمية هائلة من بيانات المعاملات وتجمع ملايين الميغابايت من البيانات عن عملائها ومورديها ومنتجاتها. كل مصادر البيانات هذه والعديد من المصدر الآخرى ساهمت في بناء ظاهرة البيانات العملاقة. تتميز البيانات الضخمة بالحجم الهائل وبنية البيانات المتنوعة والتغيير السريع ؛ أنظمة إدارة البيانات الحالية ، مثل قواعد البيانات المتوازية ، تفشل في التعامل مع هذه الخصائص الفريدة. البيانات العملاقة، بهذه الخصائص، تواجة متخصصي الحاسوب و المعلومات من اكاديميين و من قطاع الأعمال بالعديد من التحديات التي تدفعهم إلى تطوير تقنيات جديدة لمعالجة هذه التحديات والتغلب عليها. تستفيد الشركات والمؤسسات من المعلومات التي تم جمعها من خلال تطوير الخوارزميات والنظم لتحسين تحليل البيانات واستكشافها. مع حجم البيانات وتنوع مصادرها ، ستكون المعلومات التي يتم الحصول عليها نتيجة لتحليلها أكثر أهمية وفائدة. في هذا العمل ، نقترح إطارًا جديدًا لإدارة البيانات الكبيرة وتحليلها. الإطار الجديد المقترح في أول تنصيب يرسل مستخلص البيانات الوصفية إلى جميع نقاط البيانات. تم تصميم هذه المستخلصات لتناسب بنية البيانات المخزنة في كل نقطة بيانات. ثم يتم

استخدام البيانات الوصفية المستخرجة لتصنيف كل مجموعة من مجموعات البيانات باستخدام خوارزميات نماذج الموضوعات. بعد ذلك يتم تنظيم جميع الموضوعات في مجموعة البيانات كشجرة من أجل تسهيل ربط البيانات المشتركه من جميع المصادر المختلفة. عند تلقي أي مهمة تحليل ، يتم استخدام شجرة البحث لتحديد موقع البيانات المتعلقة بالمهمة ، ثم يتم إرسال نسخة من مهمة التحليل إلى نقاط البيانات التي تحتوي على هذه البيانات. لتقييم أداء الإطار المقترح ، قمنا بإجراء عدد من التجارب حيث قمنا بتنفيذ العديد من مهام تحليل البيانات لاستخدام النموذج المقترح الجديد. في كل تجربة ، تم استخدام ثلاثة معايير لقياس أداء النموذج الجديد ، وهي وقت المعالجة والبيانات المتوسطة ووقت إعداد البيانات. كما أجرينا نفس التجارب ولكن باستخدام MapReduce لأداء نفس مهمة التحليل باستخدام نفس البيئة. أظهرت التجارب تحسنا مهما في أداء النظام المقترح.

Chapter I. BIG DATA MANAGEMENT AND ANALYSIS FRAMEWORK

Chapter II. EBRAHEEM M.ALHADDAD

ABSTRACT

Data digitization, individuals and businesses are generating tremendous amount of data on daily basis all around the world. Digitization, the conversion of analogue data to digital data, facilitates launching many digitization projects such as Google Books Library Project in which millions of books were scanned and stored as an electronic library. Billions of mobile phones, tablets and laptops equipped with sensors such as cameras and running social media applications while connected to the Internet generates a huge amount of data. Moreover, businesses and organizations generate huge amount of transactional data and collect millions of megabytes data about their customers, suppliers, and products. All these data sources and many others build up the big data phenomena. Big data is characterized by the huge volume, diverse data structures and rapid change; existing data management systems, such as parallel databases, fail to cope with such unique data properties. Big data, with this characteristics, confront computer and information technology specialists from both academic and business with a lot of challenges forcing them to develop new technologies to address and overcome these challenges. Companies and institutions benefit from the collected data through the algorithms and systems development for better data analysis and exploration. With the size of the data and the diversity of its sources, the information obtained as a result of its analysis will be more important and useful. In this work we propose a novel framework for big data management and analysis. The new proposed framework at the first insulation send metadata extractors to all data nodes. These extractors are designed to adequate the structure of data stored at each data node. The extracted metadata is then used to classify each data set instance using topic modeling algorithms. Then all topics in the data set are organized as a tree in order to facilitate mapping the related data from all different sources. When any analysis job is received, the mapping tree is used to locate the relevant data, then a copy of the analysis task is sent to data

nodes which contains this data. To evaluate the performance of the proposed framework, we carried out a number of experiments where we executed several data analysis tasks to using the new proposed model. In each experiment, three

criteria were used to measure the performance of the new model, namely pro- cessing time, intermediate data and data preparation time. We also performed the same experiments but using MapReduce to perform the same analysis task using the same environment. Experiments have shown an improvement in the performance of the proposed system.