التكامل بين البيانات الكبيرة والحوسبة عالية الاداء: دراسة تشمل وحدات معالجة الرسوم، والتعلم العميق، والذاكرة الداخلية لحوسبة البيانات الكبيرة

محمد عاقب

المستخلص

التقارب بين نماذج البيانات الكبيرة والحوسبة عالية الأداء (HPC) يوفر إمكانات لا محدودة لتطوير نماذج حوسبية جديدة، حل التحديات الكبرى التي طال أمدها، والقيام باكتشافات جديدة.

الهدف من هذه الرسالة هو تحقيق التقارب بين البيانات الكبيرة وتقنيات الحوسبة عالية الأداء باستخدام تطبيقات ذات أهمية عالية، والتي تستهلك في طبيعتها بيانات كثيرة وتحتاج الى معالجة كثيفة. في هذا السياق، تقدم هذه الاطروحة المساهمات التالية. أولاً، تقديم إطار عمل يدعم التقاربين البيانات الكبيرة والحوسبة عالية الأداء باستخدام أربع تقنيات متطورة: البيانات الكبيرة، الحوسبة في الذاكرة، التعلم العميق، ووحدات المعالجة الرسومية (GPUs) . تم تنفيذ إطار العمل باستخدام ار(R)، تنسرفلو (TensorFlow)، كيراس (Keras)، وبايثون (Python). تكمن حداثة المنهجية المتبعة في انها دمجت التقنيات الأربعة المكمّلة لبعضها البعض والتي توفر مجتمعة القدرة على مواجهة تحديات البيانات الكبيرة بطريقة شاملة. اتاحت عملية دمج التقنيات الأربعة القدرة على التحقق من جدوى وفوائد التقارب بين نماذج وتقنيات البيانات الكبيرة والحوسبة عالية الأداء.

ثانياً، قمنا بتطبيق إطار العمل المقترح على أربع تطبيقات عالية التأثير للمدن الذكية باستخدام حالات دراسة مفصلة. تشمل التطبيقات (١) التنبؤ بسرعة، تدفق، ونسبة اشغال حركة المرور على الطرق; (٢) التنبؤ بحوادث السير; (٣) إدارة الكوارث; (٤) واخيراً التبوء الزماني والمكاني لدخول وخروج مترو لندن. الدراسات الأربعة استخدمت بيانات حقيقية مفتوحة المصدر من مصادر مثل نظام قياس الأداء (PeMS) في كاليفورنيا (Caltrans)، النقل من أجل لندن (TfL) المسح المتداول بين الانطلاق والوجهة (RODS)، وإدارة النقل في المملكة المتحدة (DfT). تم التعامل مع حجم البيانات، سرعتها، تنوعها، مصداقيتها، ومشاكل الاندماج الخاص بها من خلال استخدام بيانات حركة المرور لأكثر من احدى عشر عاماً. تم استخدام توليفات مختلفة من مجموعات البيانات الى جانب تكوينات شبكة مختلفة لنماذج التعليم العميق لأغراض التدريب والتنبؤ. أظهرت النتائج الفائدة والاثر العالي لأساليبنا في تطبيقات المدن الذكية الأربعة.

ساهم إطار التقارب المطروح بالإضافة الى حالات الدراسة الأربعة في نماذج تعليم عميق، خوارزميات، منهجيات تحليل وتنفيذ، ومجموعة من التطبيقات البرمجية الخاصة بالمدن الذكية، البيانات الضخمة، الحوسبة عالية الاداء، وتقاربها.

Big Data and HPC Integration: An Investigation

GPUs, Deep Learning and In-Memory Big Data Computing

Muhammad Aqib

Abstract

Convergence of big data and high performance computing (HPC) paradigms provides an unimaginable potential for developing new computing paradigms, solving long-standing grand challenges, and making new explorations and discoveries.

The aim of this thesis is to investigate the convergence of big data and HPC technologies using an important application area that is both data and compute-intensive. In this context, this thesis makes the following contributions. Firstly, it proposes a framework for the convergence of big data and HPC using four different types of cutting-edge technologies: big data, in-memory computing, deep learning and Graphical processing units (GPUs). The framework has been implemented using R, TensorFlow, Keras, and Python. The novelty of our approach lies in the integration of the four technologies that are complementary to each other and collectively provide the potential to address big data challenges in a comprehensive manner. The integration of these four technologies has also allowed investigating the viability and benefits of convergence of big data and HPC technologies and paradigms.

Secondly, we apply the proposed framework to four high-impact smart city applications using detailed case studies. These include (1) road traffic speed, flow, and occupancy prediction; (2) road incident prediction; (3) disaster management; and (4) passengers' entry, exit, and spatio-temporal prediction for London Underground. All four case studies use real open data from sources including California Department of Transportation (Caltrans) Performance Measurement System (PeMS), Transport for London (TfL) Rolling Origin and Destination Survey (RODS), and UK Department for Transport (DfT). Data volume, velocity, variety, veracity, and fusion issues are addressed using over eleven years of traffic data. Different combinations of the datasets along with different network configurations of the deep learning models are investigated for the training and prediction purposes. The results show the usefulness and high-impact of our methods in all four smart city applications.

The convergence framework and the four case studies have contributed multiple novel deep learning models, algorithms, implementation and analytics methodologies, and a range of software products for smart cities, big data, HPC, and their convergence.