"Verification and Validation of Big Data Using Mapreduce Based PK-Means Algorithm"

Prepared By

Natheir abdulqader Asaad Milhem

supervisor

Dr.Ashraf Mousa Saleh

Co-supervisor

Dr. Ahmad Mohammad Khasawneh

Abstract

Internet of Things (IoT) constitutes a valuable source of big data. Big data is usually generated from the sensors that are connected to electronic devices. The sourcing capacity depends on the ability of the sensors to provide accurate real-time information. IoT has a significant source of big data from multi-purpose devices, sensors, and appliances. The size of big data received from the Internet of Things requires a large space for cloud storage, and this affects the speed of its performance, especially if the data is redundant and not classified according to its quality, and this is one of the most important challenges in the world of big data. This research proposes a verification and validation (V&V) framework to verify the transmission of a huge volume of generated data and vast distribution of IoT devices which requires huge cloud computing storage. To reduce data size and to ensure (V &V) of big data transmission, different techniques and various algorithms have been applied. The risks of big data can be reduced by applying a systematic (V &V). Consequently,

this research proposes a framework for the (V & V) technique in big data using MapReduce programming model with PK–Means algorithm in the Apache Hadoop environment. The processing of big data requires a set of applications that guarantee the data transmission from the IoT appliances to its designated storage and analysis site. To evaluate the performance and reliability of the proposed framework, three test phases will be applied (the input stage, the processing stage, and the output stage). Moreover, the validation of the first stage includes the generated log files from IoT appliances and HDFS input. The second stage consists of the Map-Reduce model with PK-means. The last stage constitutes verification of HDFS output using K-nearest neighbor(KNN). With the proposed method, a hybrid technique between the two machine learning techniques (supervised and unsupervised). The proposed method with using seven different IBM datasets, achieved high accuracy, up to 87%, by transforming the unstructured data to structured data and enhancements in validation and verification.

Keywords: Internet of things (IoT), Big Data(BD), Hadoop, MapReduce, PK-Means Algorithms, Machine Leaning, K-nearest neighbor.

"التحقق والمصادقة من البيانات الضخمة باستخدام المعالجة المتوازية استناداً على خوازمية التجميع"

اعداد

نذير عبدالقادر اسعد ملحم

أشرف

د. اشرف موسى صالح

مشرف مشارك

د. احمد محمد خصاونة

الملخص

إنترنت الأشياء هو المصدر القيم للبيانات الضخمة. عادة ما يتم إنشاء البيانات الضخمة من أجهزة الاستشعار المتصلة بالأجهزة الإلكترونية. حجم البيانات الضخمة الواردة من إنترنت الأشياء يتطلب الكثير من مساحة التخزين وهذا يؤثر على سرعة معالجتها ، خاصة إذا كانت البيانات زائدة عن الحاجة وغير مصنفة حسب جودتها ، وهو ما يعد من أهم التحديات في عالم البيانات الضخمة. يقترح هذا البحث إطار عمل للتحقق والمصادقة (V&V) من نقل كميات هائلة من البيانات التي تم إنشاؤها والتوزيعات الكبيرة لأجهزة إنترنت الأشياء التي تتطلب مساحة تخزين ضخمة في الحوسبة السحابية. لتقليل حجم البيانات ولضمان نقل البيانات الضخمة (V&V) ، تم تطبيق تقنيات مختلفة وخوارزميات مختلفة. يمكن تقليل مخاطر البيانات الضخمة من خلال تطبيق منهجي (V&V). وفقًا لذلك ، يقترح هذا البحث إطارًا لتقنية (V&V) في البيانات الضخمة باستخدام نموذج البرمجة MapReduce مع خوارزمية PK – Means في بيئة Apache Hadoop. تتطلب معالجة البيانات الضخمة مجموعة من التطبيقات التي

تضمن نقل البيانات من أجهزة إنترنت الأشياء إلى موقع محدد للتخزين والتحليل. لتقييم أداء وموثوقية الإطار المقترح ، سيتم تطبيق ثلاث مراحل اختبار (مرحلة الإدخال ومرحلة المعالجة ومرحلة الإخراج). علاوة على ذلك ، يتضمن التحقق من صحة المرحلة الأولى ملفات السجل التي تم إنشاؤها من جهاز إنترنت الأشياء ومدخلات HDFS. تتكون المرحلة الثانية من نموذج Map-Reduce بوسائل PK. المرحلة الأخيرة هي التحقق من إخراج البيانات المخزنة ب HDFS بواسطة K-nearest neighbor. الطريقة المقترحة هي تقنية هجينة بين جهازي تعلم آلي (خاضع للإشراف وغير خاضع للرقابة). حققت الطريقة المقترحة باستخدام سيعة مجموعات بيانات مختلفة من IBM دقة عالية تصل إلى 87٪ مع نقل البيانات الأولية إلى بيانات منظمة وتحسينات في التحقق المصادقة.