Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Developing A Big Data Model for Gathering and Structuring Container Data

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Master in
Computer Science
By**

Arkan Abdulilah Khader

Supervised by

Assistant Professor Dr. Rawaa Putros Polos

1443 A.H. **2022 A.D**

Abstract

Docker Containers are the technology for the encapsulation of an application with its different components and dependencies into one image package. The containers can run in any environment that has a Docker engine without any problems or additional components.

Docker Swarm mode is a cluster of nodes provided by the Docker engine. Docker swarm mode groups Docker Engines together into one "virtual engine" that pools resources and communicates with one or more Swarm managers to execute instructions.

Performance metrics gathering and analysis of the Docker Swarm cluster contains hundreds of nodes that run thousands of containers and services that produce large amounts of performance, statistical, and log data that are essential not only for identifying problems but also for improving the performance of the images of their containers.

The goal of this thesis is to provide a big data model with a new approach for gathering and per-process wide range of performance, statistical, and status metrics about Docker swarm containers, which will provide a comprehensive view of what happens inside the Docker cluster. Then store, analysis and visualise the gathered metrics using a big data model. The architecture of the big data model consists of a group of techniques (HDFS, Apache Hive, Apache Ambari). the HDFs used as distributed file storage, when the analytics phase is done using Apache Hive and the Hadoop Distributed File System (HDFS). Hive is used to simplify the work with parallel processing model in Hadoop ecosystem, and it is used to perform pre-programmed queries that work on the data inside the HDFS for the purpose of extracting useful conclusions from the gathered data.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

تطوير نموذج بيانات ضخمة لغرض تجميع وهيكلة بيانات الحاويات

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة الماجستير في
علوم الحاسوب

من قبل
اركان عبدالاله خضر الحبو

بإشراف
أ.م.د. رواء بطرس بولص

١٤٤٣هـ ٢٠٢٢م

الخلاصة

تعتبر تقنية الحاويات "Docker containers" تقنية جديدة تقوم على مبدء عمل حزمة للبرامج مع كافة محتوياتها سواء كانت مكتبات او اعتماديات بحيث تضمن تشغيل البرامج في اي بيئة حاسوبية مدعومة بدون الحاجة للاعتماد على مكونات من خارج الحزمة المخصصة.

توفر منصة "Docker" نمط عناقيد يدعى "Docker Swarm cluster" وهو نمط يمكن المطورين من تكوين عناقيد من الحاويات ولاغراض متعددة بسهولة وسرعة حيث يقوم هذا النمط بتنظيم وتسهيل عملية التواصل وادارة توزيع الموارد وتنفيذ الخدمات بين اجهزة العنقود الواحد.

ان عملية مراقبة وتحليل اداء مجموعة كبيرة من الحاويات قد تصل للمئات او الالف ضمن العناقيد والتي بدورها تنتج كميات ضخمة من بيانات الاداء والبينات الاحصائية ليست عملية سهلة بالنسبة لمنصات المراقبة وتحليل الاداء التقليدية خصوصاً في حال رغبة المطورين بالحصول والوصل الى معلومات دقيقة عن حالة الحاويات وتتبع المشاكل التي قد تحصل للحاويات اثناء عملها.

هدف هذه الرسالة , هو تصميم نموذج يعتمد على طرق جديدة لجمع طيف واسع من بيانات الاداء والحالة والبيانات الاحصائية من الحاويات ضمن العناقيد ومن ثم تنظيمها وهيكلتها من اجل خزنها بأستخدم تقنية الخزن الموزعة الخاصة بالبيانات الكبيرة "HDFS" وتحليلها باستخدم تقنية "Apache Hive " و "Apache Ambari " و التي بدورها تسهل عملية التحليل واستخلاص من المعلومات من خلال توفيرها لمجموعة من الاستعلامات والايعازات التي تم برمجتها خصيصاً لمراقبة وتحليل البينات التي جمعها من الحاويات.