INTISARI

ANALISA KINERJA BIG DATA ENVIRONMENT PADA ARSITEKTUR KONTAINER DI DOCKER SWARM DAN KUBERNETES

Wahyuni Puji Lestari
20/466436/PPA/06002

Pertambahan data memunculkan kebutuhan pengolahan big data untuk pengambilan keputuan yang cepat dan akurat. Pemrosesan big data ini memerlukan sumber daya yang besar, sehingga dibutuhkan arsitektur yang dapat menyesuaikan antara sumber daya dengan besarnya pemrosesan data. Lingkungan terdistribusi juga sangat penting untuk meningkatkan analisis Big Data. Karena algoritma yang kompleks, data yang besar, serta komputasi data membutuhkan sumber daya yang sangat tinggi. Evolusi Spark mengungkap beberapa masalah, salah satu aspek penting di antaranya adalah masalah kinerja. Kompleksitas sistem terdistribusi kinerja aplikasi Spark sulit untuk mencapai kecepatan puncak teoritis computer. Kompleksitas kinerja sistem ini juga dipengaruhi oleh arsitektur environment yang digunakan, sehingga pemilihan arsitektur yang tepat dapat mempengaruhi kinerja pemrosesan big data.

Penelitian ini dilakukan analisa perbandingan kinerja pemrosesan data pada penjadwalan standalone di Docker Swarm dan penjadwalan Kubernetes, serta pengaruh penggunaan parameter dan pemodifikasi penjadwalan pekerjaan pada kedua arsitektur. Data yang digunakan untuk melakukan penelitian adalah data text Wikipedia Dump yang akan di komputasi pada pemrosesan data word count, sorting dan powerset.

Hasil dari penelitian didapatkan bahwa penjadwalan Kubernetes lebih unggul daripada penjadwalan standalone di Docker Swarm, dengan peningkatan kecepatan signifikan 25.68% pada word count, 11.94% sorting, dan 14.77% powerset. Hal ini dikarenakan managemen memori pada arsitektur Docker Swarm yang membutukan buffer dan cache yang tinggi sehingga memperlambat waktu pemrosesan data. Penerapan metodologi tuning parameter, tidak menunjukkan peningkatan yang signifikan serta mode penjadwalan pekerjaan FIFO lebih unggul 2.55% dari pada FAIR untuk kecepatan pemrosesan.

Kata Kunci: Spark, Kubernetes, BigData, Docker Swarm, Kontainer

ABSTRACT

BIG DATA ENVIRONMENT PERFORMANCE ANALYSIS ON CONTAINER ARCHITECTURE IN DOCKER SWARM AND KUBERNETES

Wahyuni Puji Lestari
20/466436/PPA/06002

The increase in data raises need for processing big data for fast and accurate decision making. This big data processing requires large resources, so an architecture is needed that can match the resources with the amount of data processing. Distributed environment is also very important to improve Big Data analysis. Because complex algorithms, large data, and data computations require very high resources. The Spark evolution exposed several issues, one important aspect of which was performance issues. The complexity of the distributed system performance of Spark applications makes it difficult to reach the computer's theoretical peak speed. The complexity of the system performance is also influenced by the environment architecture used, so that the selection of the right architecture can affect big data processing performance.

In this study, a comparative analysis of data processing performance was carried out on standalone scheduling in Docker Swarm and Kubernetes scheduling, as well as the effect of using parameters and job scheduling modifiers on both architectures. The data used to conduct the research is Wikipedia Dump text data which will be computed in word count, sorting and powerset data processing.

The results of the study show that Kubernetes scheduling is superior to standalone scheduling in Docker Swarm, with a significant speed increase of 25.68% in word count, 11.94% sorting, and 14.77% powerset. This is due to memory management in the Docker Swarm architecture which requires a high buffer and cache which slows down data processing time. The application of the parameter tuning methodology does not show a significant increase and the FIFO job scheduling mode is 2.55% superior to FAIR for processing speed.

Keyword: Spark, Kubernetes, BigData, Docker Swarm, Kontainer