INTISARI

Mahadata menawarkan potensi besar dalam memaksimalkan pencarian wawasan dan pengambilan keputusan berkat perkembangan teknologi seperti komputasi awan dan IoT. Di Indonesia, data lokasi dari penggunaan ponsel pintar—dikenal sebagai Mobile Positioning Data (MPD)—merupakan sumber mahadata yang berharga terutama dengan masifnya penggunaan ponsel pintar pada masyarakat. Analisis MPD sangat bermanfaat untuk mengetahui perilaku dan preferensi masyarakat, tetapi proses pengolahannya menantang karena ukuran datanya yang besar dan kecepatannya yang cepat.

Meskipun olah data pada lingkungan penelitian yang dimiliki dengan Python dasar efektif dan efisien untuk data kecil, pendekatan ini menjadi tidak memadai untuk mengolah mahadata berukuran besar dan dinamis. Hal ini dikarenakan oleh adanya *Global Interpreter Lock* pada Python yang membatasi proses komputasi Python biasa. Maka dari itu, diperlukan solusi analisis mahadata untuk mendukung pengolahan data MPD berukuran besar pada lingkungan penelitian yang dimiliki yang dapat memanfaatkan sumber daya dari kluster komputer pada lingkungan penelitian secara efektif dan efisien.

Solusi yang dikembangkan mengintegrasikan Apache Spark dan Hadoop dengan orkestrasi kontainer Docker pada Docker Swarm di kluster komputer penelitian yang terdiri dari tiga mesin. Setelah solusi selesai dikembangkan, solusi tersebut kemudian diujicobakan terhadap lingkungan python standar pada data MPD dengan ukuran yang bervariasi dan diukur waktu eksekusi serta penggunaan sumber dayanya. Ujicoba dilakukan dengan mengeksekusi fungsi analisis MPD yang terdiri dari tahap *load* data, pembersihan data, operasi olah data geolokasi (*Reverse Geocoding*), penyaringan data, serta dua tahap pemrosesan untuk uji kasus analisis waktu tinggal dan aggregat perpindahan pada level kelurahan.

Hasil dari pengujian mengindikasikan bahwa teknologi mahadata terdistribusi dapat mengakomodasi pengolahan data MPD hingga ukuran 26GB, yang mana Python dasar tidak mampu melakukannya. Selain itu, solusi terdistribusi ini menunjukkan waktu eksekusi yang lebih cepat pada 4 dari 6 tahapan yang diujicobakan. Dari segi sumber daya komputasi, solusi terdistribusi ini juga mencatatkan penggunaan CPU, memori, dan pertukaran data yang lebih efisien.

Dengan demikian, solusi ini menawarkan sebuah alternatif yang kuat dan efisien untuk analisis mahadata, khususnya dalam konteks MPD. Solusi ini juga telah berhasil mendemonstrasikan keunggulannya dalam hal efisiensi waktu dan sumber daya dibandingkan dengan metode tradisional.

Kata kunci : Komputasi Paralel, Komputasi Terdistribusi,Mahadata, Infrastruktur Mahadata, *Mobile Positioning* Data

ABSTRACT

The growing potential of big data is enhancing insights and decision-making processes, especially with the advent of technologies like cloud computing and the Internet of Things (IoT). In Indonesia, Mobile Positioning Data (MPD) obtained from smartphones is a valuable big data resource due to the widespread use of these devices. Although analyzing MPD offers crucial insights into public behavior and preferences, handling it is quite challenging due to its large volume and high velocity.

While data processing within the existing research environment using basic Python is effective and efficient for small-scale data, this approach falls short when dealing with large and dynamic Big Data. This limitation arises from the presence of the Global Interpreter Lock in Python, which constrains the processing capabilities of regular Python computing. Hence, there is a need for a Big Data analysis solution that supports the processing of large-scale MPD data within the research environment, effectively harnessing the computing resources of a research cluster efficiently.

The developed solution integrates Apache Spark and Hadoop with container orchestration using Docker Swarm on a research cluster consisting of three machines. After the solution's development, it is tested and compared to execution in standard Python environment using MPD data of varying sizes, measuring execution time and resource utilization. Testing involves the execution of an MPD analysis function comprising data loading, data cleansing, geolocation data processing (Reverse Geocoding), data filtering, and two stages of processing for testing stay time analysis and aggregated displacement analysis at the neighborhood (kelurahan) level.

The test results indicate that distributed Big Data technology can accommodate the processing of MPD data up to a size of 26GB, a task beyond the capabilities of basic Python. Furthermore, this distributed solution demonstrates faster execution times in 4 out of the 6 tested stages. In terms of computational resources, this distributed solution also records more efficient CPU usage, memory utilization, and data exchange.

Consequently, this solution offers a robust and efficient alternative for Big Data analysis, particularly in the context of MPD. Moreover, it successfully demonstrates its advantages in terms of time and resource efficiency when compared to traditional methods.

Keywords : *Parallel Computing, Distributed Computing, Big Data, Big Data Infrastructure, Mobile Positioning Data*