BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Penelitian ini menggunakan beberapa sumber pustaka yang berhubungan dengan kasus yang akan diteliti diantaranya yaitu:

Oliviandi, S., Osmond, A. B., & Latuconsina, R. (2018) dalam penelitiannya mengimplementasikan Apache Spark dan membandingkannya dengan MapReduce untuk memproses data.

Basuki, K., Novianus Palit, H., & Dewi, L. P. (2015) mengimplementasikan Apache Hadoop untuk mengolah data pinjaman di perpustakaan Universitas Kristen Petra.

Syauqi Ahsan, A., & Asmara, R. (2022) penelitian ini telah menghasilkan dua alternatif desain infrastruktur *big data* serta menunjukkan kelemahan dan kelebihan dari setiap aplikasi yang digunakan, sehingga dapat mengoptimal desain Infrastruktur *big data* sesuai dengan kebutuhan penggunanya.

Karamolegkos, P., Mavrogiorgou, A., Kiourtis, A., & Kyriazis, D. (2023) dalam penelitiannya bertujuan untuk mengembangkan EverAnalyzer, *platform* manajemen *big data* yang secara otomatis mampu melakukan pengumpulan data dan menganalisis data secara *real-time.*

Awaluddin, M., Angelia Mahlil, R., & Ode Muhammad Saidi, L. (2023) penelitian ini bertujuan untuk menerapkan kerangka kerja Hadoop MapReduce untuk memprediksi kelulusan mahasiswa.

Tabel 2.1 Tinjauan Pustaka

No.	Penulis	Objek	Hasil
1	Oliviandi, S., Osmond, A. B., & Latuconsina, R. (2018).	Apache Spark, MapReduce, dan Hadoop Distributed *File* System.	Apache Spark dapat menurunkan waktu pemrosesan data rata-rata 50% sampai 70% dari Hadoop MapReduce.
2	Basuki, K., Novianus Palit, H., & Dewi, L. P. (2015).	Apache Hadoop, MapReduce, dan perpustakaan Universitas Kristen Petra.	Penggunaan Hadoop untuk memproses data yang berukuran besar dapat menghasilkan waktu eksekusi yang rendah walaupun dengan penggunaan rata-rata konsumsi CPU dan *memory usage* meningkat.
3	Syauqi Ahsan, A., & Asmara, R. (2022).	Ekosistem Apache Hadoop.	Penelitian ini membuktikan penerapan analisis kinerja aplikasi *big data* dapat digunakan sebagai sarana optimasi desain infrastruktur *big data*.
4	Karamolegkos, P., Mavrogiorgou, A., Kiourtis, A., & Kyriazis, D. (2023).	EverAnalyzer dan ekosistem Apache Hadoop.	*Platform* manajemen data EverAnalyzer menggunakan MapReduce dan Spark dalam memproses data sehingga *user* mendapatkan wawasan dari data tersebut untuk mengambil keputusan secara akurat.
5	Awaluddin, M., Angelia Mahlil, R., & Ode Muhammad Saidi, L. (2023).	Apache Hadoop MapReduce.	Hadoop MapReduce mampu mengolah data skala besar dengan cepat, implementasi Hadoop MapReduce juga menghasilkan prediksi kelulusan mahasiswa dengan tingkat kesalahan yang rendah.

6	Ian Madiana (2024).	Data Management Platform YAVA247.	Mengembangkan *data management platform* YAVA247 menggunakan ekosistem Apache Hadoop di sistem operasi Ubuntu 22.04.

2.2 Dasar Teori

2.2.1 Big data

Big data adalah istilah yang diberikan pada kumpulan data yang beukuran sangat besar dan kompleks, sehingga tidak memungkinkan untuk diproses menggunakan perangkat pengelola *database* konvensional ataupun aplikasi pemroses data lainnya (Maryanto, 2017).

Big data didasari dari tiga konsep utama yaitu volume yang menggambarkan ukuran data, *velocity* menggambarkan laju pertumbuhan maupun perubahan data, dan *variety* menggambarkan keragaman jenis data.

2.2.2 Ubuntu

Ubuntu merupakan suatu sistem operasi berbasis *open-source* yang dibangun berdasarkan sistem operasi Linux. Ubuntu sendiri merupakan peningkatan dari sistem operasi Debian. Ubuntu memiliki tujuan membangun Distro Linux yang memberikan sistem Linux dalam komputasi *desktop* dan *server* yang selalu baru. Dalam sistem operasi Ubuntu ini, terdapat kelebihan utama yaitu menyediakan peralatan-peralatan yang penting, dan memberikan kebebasan kepada user untuk menginstal atau tidak *source* yang tersedia (Lubis, A. A., dkk 2022).

2.2.3 Apache Maven

Apache Maven merupakan alat untuk membangun dan mengelola proyek yang menggunakan bahasa pemrograman Java. Maven bisa membuat para pengembang lebih mudah, seperti memahami beberapa proyek berbasis Java.

2.2.4 Apache Ambari

Apache Ambari adalah alat untuk menyediakan, mengelola, dan memantau kluster Hadoop. Apache Ambari terdiri dari sekumpulan RESTful API dan antarmuka manajemen berbasis *browser*. Ambari memungkinkan administrator sistem untuk:

1. Menyediakan kluster Hadoop

   Ambari menyediakan panduan langkah demi langkah yang mudah digunakan untuk menginstal layanan Hadoop di sejumlah *host* dan menangani konfigurasi layanan Hadoop untuk kluster.

2. Mengelola kluster Hadoop

   Ambari menyediakan manajemen pusat untuk memulai, menghentikan, dan mengkonfigurasi ulang layanan Hadoop di seluruh kluster.

3. Memantau kluster Hadoop

   Ambari menyediakan *dashboard* untuk memantau kesehatan dan status kluster Hadoop.

2.2.5 Apache Hadoop

Apache Hadoop adalah kumpulan utilitas perangkat lunak *open-source* yang memfasilitasi penggunaan sekelompok komputer yang saling terhubung

dalam jaringan untuk memecahkan masalah terkait sejumlah besar data dan komputasi (Giri, P. R., dkk 2022). Hadoop memilki empat modul utama yaitu:

1. Hadoop Common: Utilitas yang menyediakan *library* bagi modul Hadoop lain.
2. Hadoop Distributed *File* System (HDFS): Sistem *file* terdistribusi yang menyediakan akses berkecepatan tinggi ke data aplikasi.
3. Yet Another Resource Negotiator (YARN): Kerangka kerja yang mengatur penjadwalan dan manajemen sumber daya kluster.
4. MapReduce: Sistem yang memproses data yang sangat besar secara paralel.

2.2.6 Apache Zookeeper

Zookeeper adalah layanan terpusat untuk memelihara dan koordinasi untuk aplikasi terdistribusi. Dalam sistem terdistribusi, ZooKeeper berfungsi untuk menyediakan konsensus di antara beberapa *node*. Fungsi ini sangat penting untuk memastikan konsistensi dan koordinasi yang diperlukan dalam operasi terdistribusi yang dapat diandalkan.

2.2.7 Apache Phoenix

Apache Phoenix adalah lapisan di atas HBase yang memungkinkan akses SQL dan menyediakan titik masuk yang mudah bagi pengguna dan aplikasi lain. Meskipun HBase adalah penyimpanan tanpa skema, Apache Phoenix memerlukan skema dan pengetikan data untuk menyediakan fungsionalitas SQL-nya (Cherepanova, E., dkk 2021).

2.2.8 Apache Spark

Apache Spark merupakan teknologi komputasi kluster yang cepat, yang dirancang untuk perhitungan cepat. Hal ini didasarkan pada Hadoop MapReduce dan memperluas model dari MapReduce untuk efisiensi lebih banyak jenis perhitungan, yang mencakup *query* interaktif dan *stream processing*. Fitur utama Apache Spark adalah komputasi kluster di *memory* yang meningkatkan kecepatan pemrosesan aplikasi. Apache Spark dirancang untuk mencakup berbagai macam beban kerja seperti *batch aplication, iterative algorithms, interactive queries* dan *streaming* (Oliviandi, S., dkk 2018).

2.2.9 Apache Livy

Apache Livy adalah layanan yang memudahkan interaksi dengan kluster Spark melalui antarmuka REST. Livy memungkinkan pengiriman *spark jobs* atau kode *snippet*, pengambilan hasil sinkron atau asinkron, serta manajemen *Spark Context* melalui antarmuka REST sederhana atau pustaka klien RPC. Apache Livy juga menyederhanakan interaksi antara Spark dan *server* aplikasi, sehingga memungkinkan penggunaan Spark untuk aplikasi *web* atau seluler secara interaktif.

2.2.10 Apache Zeppelin Notebook

Apache Zeppelin merupakan kode editor berbasis *web* yang juga dikenal sebagai *notebook*. Fungsinya adalah untuk menyederhanakan proses analisis data secara interaktif, eksplorasi data, visualisasi data dan berkolaborasi ke Hadoop dan Spark. Apache Zeppelin mendukung sejumlah bahasa pemrograman dan skrip, termasuk SQL, Python, Scala, dan berbagai bahasa pemrograman lainnya