BAB I

PENDAHULUAN

1.1. Latar Belakang

Dalam memproses data yang sangat besar atau yang dikenal dengan istilah *Big Data* kita pasti memerlukan sebuah perangkat lunak yang bisa memproses dan mengolah data seoptimal mungkin. Banyak sekali perangkat lunak di internet yang disediakan untuk hal tersebut. Tetapi pada penelitian kali ini hanya akan menganalisis perbandingan antara dua perangkat lunak yaitu *Apache Spark* dan *Hadoop Mapreduce*.

Analisis perbandingan merupakan teknik analisis suatu aplikasi atau *software* yang dilakukan dengan cara memasukkan proses terhadap suatu aplikasi dan membandingkan antara satu dengan yang lain, dengan menunjukkan informasi atau data lain berupa kecepatan memproses data atau besar memori yang dibutuhkan untuk memproses sebuah data pada suatu aplikasi.

Big Data merupakan istilah yang berlaku untuk informasi yang tidak dapat diproses atau dianalisis menggunakan alat tradisional.(Zikopoulos et al., 2012) Lalu, (Dumbill, 2012) menerangkan bahwa *Big Data* adalah data yang melebihi proses kapasitas dari konvensi sistem database yang ada. Data terlalu besar dan terlalu cepat atau tidak sesuai dengan struktur arsitektur database yang ada. Dalam mendapatkan nilai dari data, maka harus memilih jalan alternatif untuk memprosesnya.

Big Data dapat dianalisis untuk menjadi informasi atau pengetahuan yang berharga. Ada beberapa teknik dalam proses analisis *Big Data* antara lain, *Association Rule Learning, Classification tree analysis, Generic algorithms, Machine Learning, Regression analysis, Sentimental Analysis,* dan *Social network analysis* (K. D. Cahyo, 2018).

Penelitian dalam lingkungan *Big Data* telah banyak dilakukan. Salah satu penelitian yang dilakukan oleh (Oliviandi et al., 2018) yang melakukan pengujian analisis *Big Data* dengan *Apache Spark* pada *Big Data* berbasis HDFS, bahwa penggunaan *Apache Spark* sangat tepat karena dapat menurunkan *response time* rata-rata 50% sampai 70% dari *Hadoop Mapreduce*. Namun, kelemahannya dalam optimalisasi kemampuan, membutuhkan spesifikasi *hardware* yang mumpuni.

Sementara itu penelitian yang dilakukan (Demidova et al., 2016) melakukan pengujian analisis *Big Data* dengan mengimplementasikan *Support Vector Machine* dengan modifikasi *Particle Swarm Optimization*. Hasilnya menunjukkan Selama percobaan ditemukan bahwa *SVM esembles* menunjukkan akurasi mulai dari 85,75% hingga 91,5%. Sementara keakuratan *SVM two level classifier* menunjukkan persentase sebesar 97,26%. Dengan demikian, *SVM two level classifier* meningkatkan akurasi klasifikasi hampir 3% dibandingkan untuk akurasi dari salah satu pengklasifikasi *SVM esembles*.

Perbedaan hasil penelitian ini terjadi karena masing-masing peneliti memilih representasi data yang berbeda, sedangkan performa suatu algoritma sangat bergantung pada representasi dari data yang digunakan (Goodfellow et al., 2016).

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di Annual Workshop on Computational Learning Theory. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti margin hyperplane (Duda & Hart (1973), Cover (1965), Vapnik (1964), dan sebagainya.), kernel (Aronszajn, 1950) dan konsep-konsep pendukung yang lain. Belum pernah ada upaya merangkaikan komponen-komponen tersebut hingga tahun 1992 (L. Cahyo, 2018).

Berdasarkan seluruh uraian tersebut, terdapat peluang pengembangan penelitian tentang proses perbandingan antara *Apache Spark* dengan *Hadoop Mapreduce* dengan menggunakan algoritma SVM. Maka dari itu pada penelitian ini akan membahas tentang "Analisis Perbandingan Performa *Apache Spark* Dan *Hadoop Mapreduce* Pada *Mapreduce Framework* Menggunakan Algoritma *Support Vector Machine*".

1.2. Rumusan Masalah

Berdasarkan uraian pada latar belakang, maka rumusan masalah dari penelitian ini sebagai berikut:

1. Bagaimana menganalisis perbandingan performa *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework*?
2. Bagaimana implementasi algoritma *SVM* pada *Apache Spark* dan *Hadoop Mapreduce*?

3. Bagaimana tingkat akurasi algoritma *Support Vector Machine* pada proses perbandingan *Apache Spark* dan *Hadoop Mapreduce*?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini adalah :

1 Menganalisis perbandingan performa *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework.*

2 Mengimplementasikan algoritma *SVM* pada *Apache Spark* dan *Hadoop Mapreduce*.

3 Membuat perbandingan akurasi dengan algoritma *Support Vector Machine* pada *Apache Spark* dan *Hadoop Mapreduce* sehingga menghasilkan teknologi mana yang paling bagus.

1.4. Tujuan Penelitian

Tujuan penelitian tugas akhir ini adalah sebagai berikut:

1. Menganalisis data berskala besar dengan menggunakan *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework.*
2. Memodelkan algoritma *SVM* pada *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework.*
3. Mengukur tingkat akurasi dengan algoritma *Support Vector Machine* pada proses perbandingan *Apache Spark* dan *Hadoop Mapreduce*.

1.5. Manfaat Penelitian

Manfaat penelitian dari tugas akhir ini adalah sebagai berikut:

1. Menghasilkan hasil analisis dari sebuah data berskala besar dengan menggunakan *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework.*
2. Menghasilkan sebuah grafik dan persebaran dari pengimplementasi *Machine Learning SVM* pada *Apache Spark* dan *Hadoop Mapreduce* sehingga memudahkan pembacaan data.
3. Menghasilkan tingkat akurasi dari algoritma *Machine Learning SVM* dalam proses perbandingan pada teknologi *Apache Spark* dan *Hadoop Mapreduce.*

1.6. Metodologi Penelitian

1. Studi Literatur

   Studi pustaka dan literatur yang dilakukan, dipelajari, dan dikaji dalam menyelesaikan penelitian ini diambil dari buku, jurnal, internet, serta dokumen-dokumen yang berkaitan dengan penelitian.

2. Pengumpulan Data

   Data yang dibutuhkan dalam penelitian ini berupa dataset yang telah diperoleh dari berbagai repositori yang berisi berbagai macam data untuk membantu penelitian yang akan dilakukan.

3. Perancangan Program

   Perancangan program dibuat sesuai dengan analisis kebutuhan. Rancangan yang dibuat memuat proses memasukkan data dan inisialisasi data, proses pelatihan model, proses visualisasi dan analisis data, pengujian

algoritma SVM pada *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework.*

4. Implementasi Program

Implementasi dibuat untuk rancangan program yang telah disusun. Pada implementasi program, harus memuat proses memasukkan data dan inisialisasi data, proses pelatihan model, proses visualisasi dan analisis data, pengujian algoritma SVM pada *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework* seperti yang telah dijabarkan pada rancangan program.

5. Pengujian

Pengujian dilakukan untuk mengetahui kemampuan dari algoritma SVM serta model yang telah dibuat dengan beberapa parameter yang diuji. Hasil dari penelitian yang dilaksanakan berupa nilai akurasi dari SVM model. Kemudian, ditarik kesimpulan mana teknologi terbaik antara *Apache Spark* dan *Hadoop Mapreduce* pada *Mapreduce Framework*.

1.7. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam laporan tugas akhir ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini memaparkan mengenai gambaran umum yang meliputi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini memuat uraian penelitian-penelitian terkait yang sudah pernah dilakukan dalam penelitian lain dan hubungannya dengan masalah penelitian yang sedang dilakukan. Selain itu, memuat dasar-dasar teoritis maupun penjelasan umum mengenai *Big Data*, *machine learning*, algoritma SVM, *Apache Spark, Hadoop Mapreduce*, pendekatan analisis data, dan penjelasan umum lainnya yang berhubungan dengan penelitian.

BAB III METODOLOGI

Bab ini menguraikan mengenai alur penyelesaian masalah terhadap penelitian yang dilakukan. Selain itu juga menjelaskan analisis kebutuhan dan perancangan sistem yang akan digunakan. Pada bab ini dibahas mengenai alur penyelesaian masalah mulai dari studi literatur, pengumpulan data, perancangan program, implementasi program, sampai pengujian program.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memuat uraian mengenai hasil dan proses pencapaian dalam menyelesaikan penelitian ini. Dimulai dari hasil studi literatur, diskusi dan konsultasi,

pengumpulan data, perancangan program, implementasi program, sampai pengujian program.

BAB V KESIMPULAN DAN SARAN

Bab ini merupakan bab akhir dari penulisan laporan yang berisi mengenai simpulan-simpulan yang merupakan hasil analisis pada bagian sebelumnya serta saran yang perlu diperhatikan berdasarkan keterbatasan yang ditemukan.