BAB I
PENDAHULUAN

1.1. Latar Belakang

Munculnya teknologi, perangkat, dan sarana komunikasi baru seperti situs jejaring sosial, menyebabkan jumlah data yang dihasilkan oleh umat manusia berkembang pesat setiap tahun. Jumlah data yang dihasilkan dari awal waktu hingga 2003 adalah 5 miliar *gigabyte*. Jika seseorang menumpuk data dalam bentuk disk, itu mungkin mengisi seluruh lapangan bola. Jumlah yang sama dibuat setiap dua hari pada tahun 2011, dan setiap sepuluh menit pada tahun 2013. Tingkat ini masih tumbuh sangat besar. Meskipun semua informasi yang dihasilkan ini bermakna dan dapat berguna ketika diproses, itu sedang diabaikan (White, 2015).

Gambar 1.1 Generator Big Data di Internet

Karena popularitas internet adalah salah satu alasan utama untuk pertumbuhan komunikasi dan konektivitas yang cepat di dunia, kami melihat munculnya platform Big Data di lingkungan Internet (Phaneendra & Reddy, 2013). Setiap hari, 2,5 miliar *byte* data dibuat dan 90 persen data di dunia saat ini diproduksi dalam dua tahun terakhir. Kemampuan untuk menghasilkan data tidak pernah begitu kuat dan besar sejak penemuan teknologi informasi pada awal abad ke-19. Sebagai contoh lain, pada 4 Oktober 2012, debat presiden pertama antara Presiden Barack Obama dan Gubernur Mitt Romney memicu lebih dari 10 juta

tweet dalam waktu 2 jam. Di antara semua *tweet* ini, momen-momen spesifik yang menghasilkan diskusi paling banyak sebenarnya mengungkapkan kepentingan publik, seperti diskusi tentang *medicare* dan voucher. Diskusi online semacam itu memberikan cara baru untuk merasakan kepentingan publik dan menghasilkan umpan balik secara *realtime*, dan sebagian besar menarik dibandingkan dengan media generik, seperti radio atau siaran TV. Contoh lain adalah Flickr, situs berbagi foto publik, yang menerima 1,8 juta foto per hari, rata-rata, dari Februari hingga Maret 2012. Dengan asumsi ukuran setiap foto adalah 2 *megabyte* (MB), ini membutuhkan 3,6 *terabyte* (TB) penyimpanan setiap satu hari. Memang, sebagai pepatah lama menyatakan: "sebuah gambar bernilai seribu kata," miliaran gambar di Flicker adalah tangki harta karun untuk menjelajahi masyarakat manusia, acara sosial, urusan publik, bencana, dan sebagainya, hanya jika kami memiliki kekuatan untuk memanfaatkan sejumlah besar data (Wu et al., 2014).

Data merupakan komoditas yang penting saat ini terutama bagi para peneliti biologi. Bagaimana tidak, sepuluh tahun yang lalu seorang peneliti memerlukan waktu tiga tahun untuk menemukan gen yang terlibat dalam suatu penyakit. Namun sekarang ini, berkat adanya informasi genom yang tersimpan di database yang besar dimana publik dapat memperoleh data itu, sehingga tugas yang sama mungkin dapat dikerjakan dalam waktu setengah jam saja (Zalzalah, 2015).

Banyak sekali proyek yang akan dilakukan pada 2025 dan dilakukan perbadingan terhadap 3 generator Big Data besar seperti astronomi, YouTube, dan Twitter. Dari perbandingan tersebut diperkirakan bahwa genomik akan menuntut banyak terhadap kebutuhan domain analisis (Stephens et al., 2015).

Ada dua kasus dalam bidang biologi yang membutuhkan penyelesaian secara komputasi *string matching* yaitu kasus *string matching* terhadap sebuah pola atau motif pada DNA tanaman untuk membedakan golongan tanaman berbunga dan golongan tanaman lainnya seperti pakis, moses, jamur, dan alga (Hidayat, Priyandoko, Wardiny, & Islami, 2016). Kemudian kasus pada budidaya melon yang menjadi kendala karena adanya virus dengan gejala *Begomovirous* yang mengakibatkan tanaman melon terjangkit penyakit daun keriting (Wilisiani, Somowiyarjo, & Hartono, 2014). Didapati bahwa enzim restriksi yang

mempengaruhi penyakit daun keriting adalah pola BamHI pada sekuens DNA tanaman. Dua kasus biologi ini kemudian dapat diangkat menjadi studi kasus dalam penelitian yang dilakukan. Pada saat ini, penyelesaian masalah untuk kedua studi kasus ini relatif lama, dimana proses analisa masih menggunakan cara dengan manual yaitu mengamati pada titik-titik tertentu. Hal itu menyebabkan bisa saja terjadi kesalahan analisa dengan *human error* karena analisa dilihat oleh mata manusia. Maka, hal ini yang kemudian menjadi celah atau peluang bagi para peneliti khususnya di bidang ilmu komputer untuk turut terlibat membantu dalam proses analisa sehingga lebih memudahkan dalam melakukan pencarian hasil.

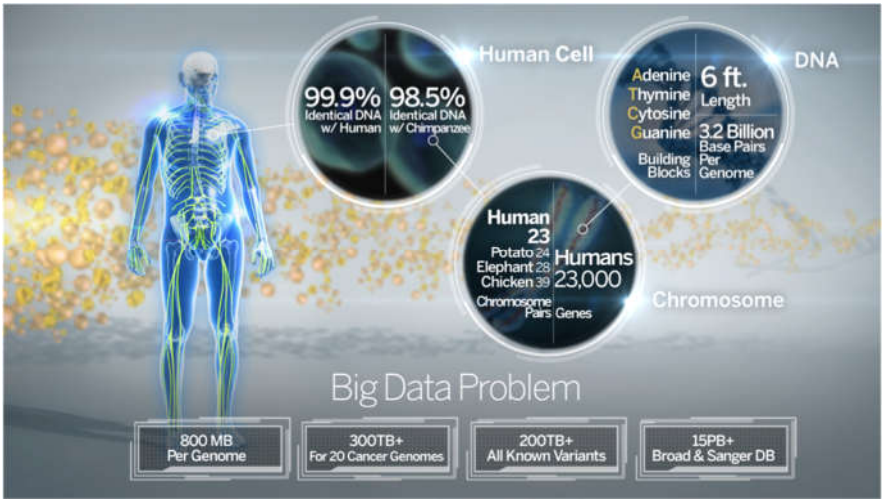
Hal itu lah yang membuat ganomik menjadi hal yand diperhatikan ilmu di bidang Big Data (Stephens et al., 2015). *String matching* adalah tugas utama yang digunakan dalam banyak aplikasi seperti deteksi malware berbasis tanda tangan, biologi komputasi, mesin pencarian pada web, hingga beberapa aplikasi lainnya. Ini merupakan proses menemukan semua *sub-string S* dalam teks sekuens *T*. Algoritma Boyer-Moore adalah salah satu yang paling populer sebab dapat menangani masalah yang semakin kompleks ketika berhadapan dengan data pencarian besar (Zalzalah, 2015).

Pada penelitian ini akan menggunakan variasi dari Algoritma Boyer-Moore, yaitu Algoritma Boyer-Moore Horspool. Sebagaimana penelitian yang dilakukan sebelumnya (Rachman, Riza, & Hidayat, 2017) menjadi rujukan penelitian yang dilakukan dengan melakukan beberapa inovasi terutama pada penggunaan algoritma *string matching* yang setara dan perkembangan terkait teknologi baru yang saat ini masih berkembang yaitu dengan Big Data khususnya Apache Spark.

Tools Big Data saat ini sedang berkembang, dibuktikan dengan beberapa *platform* turut muncul seiring dengan kebutuhan yang saat ini menjadi solusi agar permasalah Big Data dapat diselesaikan. Beberapa *tools* itu adalah Apache Hadoop, Apache Spark, Hive, Kafka, Cassandra dan masih banyak lainnya yang turut berkembang. Bahkan sudah banyak juga beberapa pengembang dalam memupuni Big Data agar dapat lebih mudah digunakan seperti Amazon Web Service, Google Cloud Platform dan *cloud* lainnya dalam mendukung *tools* Big Data ini.

Apache Spark adalah platform komputasi *cluster* yang dirancang agar cepat untuk tujuan umum (Karau, Konwinski, & Wendell, 2015). Di sisi kecepatan, Spark memperluas model MapReduce populer untuk secara efisien mendukung lebih banyak jenis perhitungan, termasuk *query* interaktif dan pemrosesan aliran. Kebanyakan kerangka Big Data saat ini fokus pada analisis data on-disk, yang memungkinkan penyelidikan dataset besar, tetapi sangat menghambat kecepatan. Sebaliknya, Apache Spark berusaha untuk memperbaiki hal ini dengan sangat fokus pada pemrosesan di memori. Karena akses memori secara signifikan lebih cepat daripada akses disk, dan karena kepadatan memori masih berkembang sesuai dengan hukum Moore (tidak seperti kapasitas disk yang telah mengalami hambatan desain), ini sekarang menjadi pendekatan yang layak. Maka dari itu, pada penelitian ini akan digunakan Apache Spark untuk Big Data *platform*-nya.

Beberapa tahun terakhir, dunia ilmiah telah banyak mengembangkan pemetaan terhadap genom manusia bahkan sampai saat ini mulai dengan memetakan genom untuk organisme lainnya. Analisis data sekuens genomik yang muncul dan proyek genom manusia dan organisme lainnya adalah pencapaian penting untuk bioinformatika (Bayat, 2002).



Gambar 1.2 Masalah Big Data pada Biologi dan Kedokteran

Tahun lalu diumumkan bahwa seluruh genom manusia telah dipetakan sebagai hasil dari upaya proyek genom manusia di seluruh dunia dan perusahaan genom pribadi. Namun, dalam beberapa tahun terakhir, dunia ilmiah telah

menyaksikan selesainya seluruh rangkaian genom dari banyak organisme lain. Analisis data sekuens genomik yang muncul dan proyek genom manusia adalah pencapaian penting untuk bioinformatika. Bioinformatika adalah disiplin yang berkembang, dan ahli bioinformatika sekarang menggunakan program perangkat lunak yang kompleks untuk mengambil, memilah, menganalisis, memprediksi, dan menyimpan data urutan DNA dan protein. Pengetahuan yang diperoleh dari data sekuens ini akan memiliki implikasi yang cukup besar tentang pemahaman terhadap biologi dan kedokteran. Pertumbuhan bioinformatika telah menjadi usaha global, menciptakan jaringan komputer yang memungkinkan akses mudah ke data biologis dan memungkinkan pengembangan program perangkat lunak untuk analisis yang mudah. Beberapa proyek internasional yang bertujuan menyediakan basis data gen dan protein tersedia secara bebas untuk seluruh komunitas ilmiah melalui internet (Bayat, 2002).

Maka dengan melihat pentingnya kontribusi penelitian-penelitian baru tekait dengan bidang bioinformatika, pada penelitian ini akan mencoba menggunakan studi kasus bioinformatika sebagai langkah awal. Sebagaimana penelitian yang dilakukan (Rachman et al., 2017) yang telah dipaparkan sebelumnya dan juga penelitian yang dilakukan (Dhiba, Riza, & Setiawan, 2018) yang juga menggunakan studi kasus bioinformatika, yang menjadi latar belakang kasus bioinformatika untuk digunakan.

1.2. Rumusan Masalah

Sesuai latar belakang masalah yang telah diuraikan pada sub bab sebelumnya, maka munculah rumusan masalah sebagai berikut:

1. Bagaimana cara merancang dan mengimplementasikan Apache Spark dengan menggunakan algoritma Boyer-Moore Horspool untuk mengatasi masalah *Internal Transcribed Spacer* dan *Restriction Enzyme*?
2. Bagaimana *output* yang dihasilkan setelah program dirancang?
3. Bagaimana nilai perbandingan waktu antara eksekusi yang menggunakan Apache Spark pada jumlah node 1, 3, 5, 11, dan 16 dengan yang tidak dari

komputasi yang telah dirancang dan diimplementasikan serta nilai akurasinya?

1.3. Tujuan Penelitian

Setelah diketahui rumusan masalahnya, maka tujuan dari penelitian ini adalah:

1. Merancang dan mengimplementasikan Apache Spark dengan menggunakan algoritma Boyer-Moore Horspool untuk menangani masalah *Internal Transcribed Spacer* dan *Restriction Enzyme*.
2. Melakukan uji coba data yang telah dikumpulkan melalui laman FTP NCBI terhadap program yang telah dibuat.
3. Menganalisis terhadap nilai perbandingan waktu dari hasil uji coba dan akurasinya.

1.4. Manfaat Penelitian

Adapun manfaat penelitiannya adalah sebagai berikut:

1. Memberikan pengetahuan tentang Bioinformatika, khususnya tentang merancang dan mengimplementasikan progran *string matching* pada Apache Spark dengan menggunakan algoritma Boyer-Moore Horspool.
2. Melakukan uji coba data yang telah dikumpulkan melalui laman NCBI terhadap program yang telah dibuat.
3. Menganalisis terhadap nilai perbandingan waktu dari hasil uji coba serta nilai akurasinya.

1.5. Batasan Masalah

Adapun batasan masalahanya adalah sebagai berikut:

1. Program ini bekerja untuk data dengan format standar NCBI/Ensembl.
2. Data yang digunakan pada penelitian ini adalah beberapa contoh sekuens DNA tumbuhan seperti Padi pada beberapa *chromosome* di laman NCBI yang dapat diunduh di https://www.ncbi.nlm.nih.gov/nuccore.
3. Jumlah *node* tertinggi yang digunakan pada eksperimen sebanyak 16 *node*.

4. *Core* yang diujikan adalah 4 *core* untuk setiap *node* pada Google Cloud Plaform.
5. Pada komputasi *stand alone* tidak akan dibandingkan secara langsung hasil waktunya dikarenakan perbedaan spesifikasi dan komputasi yang berbeda.
6. Data yang diakses dalam 1 direktori hanya dapat dijadikan satu sekuen panjang sehingga 1 direktori dianggap 1 sekuen spesies.

1.6. Sistematika Penulisan

Pada bagian sistematika penulisan ini akan diuraikan mengenai penjelasan tiap bab.

BAB I PENDAHULUAN

Pada bab ini menjelaskan bagaimana penelitian ini bisa muncul yang isinya menerangkan bagaimana penelitian ini akan dilakukan dan diawali dengan latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini menjelaskan tentang teori pendamping atau pendukung untuk melakukanpenelitian. Teori yang dijelaskan dalam bab ini yaitu mengenai *big data*, Apache Hadoop, Apache Spark, *string matching*, algoritma Boyer-Moore Horspool, Bioinformatics, *Internal Transcribed Specer* dan *Restriction Enzyme* khususnya virus *Bacillus Amyloliquefaciens* (BamHI).

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan langkah-langkah penelitian yang akan dilakukan, dimulai dari desain penelitian, fokus penelitian, alat dan bahan yang digunakan untuk penelitian dan yang terkahir adalah metode penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjabarkan hasil penelitian dan eksperimen yang telah dilakukan. Semua pertanyaan mengenai masalah yang diangkat dalam tema skripsi dibahas pada bab

ini. Beberapa hal di antaranya adalah tentang proses pengumpulan data, pengembangan model, implementasi sistem, studi kasus, desain eksperimen, dan analisa.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran bagi peneliti selanjutnya dari hasil penelitian yang telah dilakukan.