DETEKSI GENOMIC REPEATS MENGGUNAKAN ALGORITMA BOYER-MOORE DENGAN APACHE SPARK STREAMING

SKRIPSI

Diajukan untuk Memenuhi Sebagian dari
Syarat Memperoleh Gelar Sarjana Komputer
Program Studi Ilmu Komputer

Oleh
Farhan Dhiyaa Pratama
1503677

PROGRAM STUDI ILMU KOMPUTER
DEPARTEMEN PENDIDIKAN ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
BANDUNG
2019

DETEKSI GENOMIC REPEATS MENGGUNAKAN ALGORITMA BOYER-MOORE DENGAN APACHE SPARK STREAMING

Oleh

Farhan Dhiyaa Pratama

NIM 1503677

Sebuah Skripsi yang Diajukan untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Komputer di Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

Universitas Pendidikan Indonesia

Agustus 2019

FARHAN DHIYAA PRATAMA
1503677

DETEKSI GENOMIC REPEATS MENGGUNAKAN ALGORITMA BOYER-MOORE DENGAN APACHE SPARK STREAMING

DISETUJUI DAN DISAHKAN OLEH PEMBIMBING:

Pembimbing I,

Lala Septem Riza, M.T., Ph.D.

NIP. 197809262008121001

Pembimbing II,

Erna Piantari, M.T.

NIP. 920171219890224201

Mengetahui,
Ketua Departemen Pendidikan Ilmu Komputer

Lala Septem Riza, M.T., Ph.D.

NIP. 197809262008121001

PERNYATAAN

Dengan ini penulis menyatakan bahwa skripsi dengan judul "Deteksi Genomic Repeat menggunakan Algoritma Boyer-Moore dengan Apache Spark Streaming" ini beserta seluruh isinya adalah benar-benar karya penulis sendiri. Penulis tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, penulis siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya penulis ini.

Bandung, Agustus 2019

Yang Membuat Pernyataan,

Farhan Dhiyaa Pratama

NIM 1503677

DETEKSI GENOMIC REPEATS MENGGUNAKAN ALGORITMA BOYER-MOORE DENGAN APACHE SPARK STREAMING

Oleh

Farhan Dhiyaa Pratama — farhansmg@gmail.com

1503677

ABSTRAK

Dalam satu dekade terakhir para ilmuwan harus melakukan penelitian laboratorium selama 3 tahun untuk menganalisa DNA. Salah satu kasus dari analisa DNA yang membutuhkan waktu dan tenaga dalam skala besar tersebut adalah untuk menganalisa penyakit yang disebabkan oleh pola genom yang berulang atau disebut dengan *genomic repeats*. Dalam menganalisa masalah *genomic repeats* dilakukan analisa *string matching* atau *pattern matching* dimana akan mencari sebuah pola dalam sebuah teks yang berukuran besar. Algoritma Boyer-Moore memproses pola dan membuat dua tabel, yang dikenal sebagai tabel Boyer-Moore *Bad Character* (bmBc) dan tabel Boyer-Moore *good-suffix* (bmGs). Untuk setiap karakter dalam set alfabet, tabel *bad character* menyimpan nilai pergeseran berdasarkan kemunculan karakter dalam pola. Algoritma ini membentuk dasar untuk beberapa algoritma pencocokan pola. Untuk itu, penelitian ini membuat sebuah model komputasi untuk mendapatkan pola genom yang berulang atau *genomic repeats* secara cepat dan efektif dengan memodifikasi dan mengimplementasikan algoritma Boyer-Moore pada *Big Data Platform* yaitu Apache Spark Streaming. Hasil penelitian ini menunjukkan adanya percepatan antara penggunaan *Big Data platform* dengan perancangan 2 skenario. Skenario pertama yaitu penggunaan cluster dengan 4 *cores* dan beberapa *worker node* dan skenario kedua yaitu penggunaan cluster dengan 2 *worker node* dan beberapa jumlah *core*. Penelitian ini juga membuktikan bahwa model komputasi yang dibangun menunjukkan adanya percepatan terhadap penelitian terdahulu dengan menggunakan *stand alone*.

Kata Kunci: *Big Data, Algoritma Boyer-Moore, Apache Spark Streaming, genomic repeats*

GENOMIC REPEATS DETECTION USING BOYER-MOORE ALGORITHM WITH APACHE SPARK STREAMING

Arranged by

Farhan Dhiyaa Pratama —farhansmg@gmail.com

1503677

ABSTRACT

In the past decade scientists have been doing laboratory research for 3 years to analyze DNA. One of the cases of DNA analysis that requires time and effort on a large scale is to analyze diseases caused by repetitive genomic patterns or called genomic repeats. In analyzing the problem of genomic repeats an analysis of string matching or pattern matching is carried out which will look for a pattern in a large text. The Boyer-Moore algorithm processes patterns and creates two tables, known as the Boyer-Moore Bad Character (bmBc) table and the Boyer-Moore good-suffix (bmGs) table. For each character in the alphabet set, bad character tables store shift values based on the appearance of characters in the pattern. This algorithm forms the basis for several pattern matching algorithms. For this reason, this research creates a computational model to get repetitive genomic patterns or genomic repeats quickly and effectively by modifying and implementing the Boyer-Moore algorithm on the Big Data Platform, namely Apache Spark Streaming. The results of this study indicate an acceleration between the use of Big Data platforms with the design of 2 scenarios. The first scenario is the use of clusters with 4 cores and several worker nodes and the second scenario is the use of clusters with 2 worker nodes and a number of cores. This study also proves that the computational model that was built shows the acceleration of previous research using stand alone.

Keywords: Big Data, Boyer-Moore Algorithm, Apache Spark Streaming, genomic repeats.

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah swt. karena hanya dengan kehendak, berkat, serta karunia-Nya lah penulis dapat menyelesaikan skripsi yang berjudul "Deteksi Genomic Repeats menggunakan Algoritma Boyer-Moore dengan Apache Spark Streaming" ini dapat terselesaikan.

Penyusunan skripsi ini ditunjukan untuk memenuhi dan melengkapi salah satu syarat untuk penyusunan skripsi yang merupakan syarat untuk mendapatkan gelar sarjana komputer atas jenjang studi S1 pada Program Studi Ilmu Komputer Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam Universitas Pendidikan Indonesia.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih terdapat banyak kekurangan dan keterbatasan yang perlu disempurnakan. Oleh karena itu, penulis sangat mengharapkan saran maupun kritik yang membangun agar tidak terjadi kesalahan yang sama dikemudian hari dan dapat meningkatkan kualitas ke tahap lebih baik.

Bandung, Agustus 2019

Penyusun

UCAPAN TERIMAKASIH

Alhamdulillahhirabilalamin, puji dan syukur kehadirat Allah SWT Yang telah melimpahkan rahmat dan hidayah-Nya sehingga penulis diberikan kelancaran dalam menyelesaikan penulisan skripsi ini. Dalam proses menyelesaikan penelitian dan penyusunan skripsi ini, peneliti banyak mendapat bimbingan, dorongan, serta bantuan dari berbagai pihak. Oleh karena itu, pada kesempatan ini peneliti mengucapkan terimakasih serta penghargaan yang setinggi-tingginya, kepada:

1. Kedua orang tua yaitu Doddy Hadi Rukmana dan Evi Mustika Sari yang selalu memberikan doa dan dukungan moral dan materil, serta selalu menjadi penyemangat utama dalam menempuh pendidikan tinggi sehingga penulis dapat menyelesaikan skripsi ini.
2. Bapak Lala Septem Riza, M.T., Ph.D. selaku pembimbing I atas segala waktu yang dicurahkan untuk membimbing penulis demi terselesaikannya skripsi ini.
3. Ibu Erna Piantari, M.T., selaku pembimbing II yang telah memberikan saran kepada penulis selama proses penyelesaian penelitian dan penulisan skripsi.
4. Bapak Prof. Dr. H. Munir, M.IT., selaku Kepala Departemen Pendidikan Ilmu Komputer FPMIPA Universitas Pendidikan Indonesia.
5. Bapak Eddy Prasetyo Nugroho, M.T., selaku Ketua Program Studi Ilmu Komputer.
6. Bapak dan Ibu Dosen Prodi Pendidikan Ilmu Komputer dan Ilmu Komputer yang telah berbagi ilmu yang sangat bermanfaat kepada penulis.
7. Aulia Fauziah Nasuha selaku orang terdekat penulis yang selalu memberikan segala dukungan untuk penulis dalam menjalani hidup.
8. Sahabat itok meti, Ammar, Dimas, Yogi, Arga, Adie, Rahman, Hafidz, Fiko, Trisna, Adit, Fakhri yang senantiasa memberikan dukungan, semangat, canda dan tawa kepada penulis baik selama proses perkuliahan maupun selama proses pengerjaan skripsi ini.
9. Naufal dan teh Fidel selaku grup bimbingan dan topik *Big Data* pertama dengan pak Lala yang telah melalui pahit manis penelitian ini bersama, juga selalu memberi dukungan, semangat, dan melengkapi satu sama lain.

10. Pengurus Inti BEM KEMAKOM 2017/2018 yang telah berjuang bersama sama dalam menyebar kebaikan untuk KEMAKOM.
11. DEPKOMINFO BEM KEMAKOM 2016/2017, yang telah membentuk keluarga baru di KEMAKOM pada tahun pertama penulis berorganisasi, serta memberikan kedekatan luar biasa hingga saat ini.
12. DEPKOMINFO BEM KEMAKOM 2017/2018, yang telah membentuk keluarga baru di KEMAKOM pada tahun kedua penulis berorganisasi, serta memberikan kedekatan luar biasa hingga saat ini
13. Kelas C 2015, yang sama-sama berjuang dari awal perkuliahan dari awal hingga ke titik akhir perkuliahan.
14. Semua pihak yang telah membantu peneliti dalam menyelesaikan skripsi ini yang tidak dapat peneliti sebutkan satu persatu.

Semoga semua amal baik yang telah diberikan kepada penulis mendapatkan balasan yang berlipat dari Allah SWT. Aamiin.

Bandung, Agustus 2019

Farhan Dhiyaa Pratama

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR PUSTAKA

A.P.Czernilofsky, W.DeLorbe, R.Swanstrom, H.E.Varmus, J.M.Bishop, E. T. and R. G. (1980). volume 8 Number 131980 Nucleic A c i d s Research, *8*(13), 2967–2984. https://doi.org/10.1093/nar/8.13.2967.

A Vouk, M. (2008). Cloud computing–issues, research and implementations. *Journal of Computing and Information Technology*, *16*(4), 235–246. https://doi.org/10.1109/ITI.2008.4588381

Aggarwal, C. C. (2010). *Social Network Data Analytics*. https://doi.org/10.1007/978-1-4419-8462-3_4

Al Kindhi, B., & Sardjono, T. A. (2015). Pattern Matching Performance Comparisons as Big Data Analysis Recommendations for Hepatitis C Virus (HCV) Sequence DNA. *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, 99–104. https://doi.org/10.1109/AIMS.2015.27

Ashley, C. T., & Warren, S. T. (1995). Trinucleotide repeat expansion and human disease. *Electrophoresis*, *16*(1), 1698–1704. https://doi.org/10.1002/elps.11501601282

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79–80*, 3–15. https://doi.org/10.1016/j.jpdc.2014.08.003

Barkham, R., & Saiz, A. (2018). Urban Big Data : City Management and Real Estate Markets Urban Big Data : City Management and Real Estate Markets. *MIT Center for Real Estate and DUSP*.

Batarfi, O., El, R., Ayman, S., Reza, G. F., Ahmed, S. B., & Sherif, B. (2015). Large scale graph processing systems : survey and an experimental evaluation. *Cluster Computing*, *18*(3), 1189–1213. https://doi.org/10.1007/s10586-015-0472-6

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Begenau, J., Gsb, S., Maryam, N., Princeton, F., & Veldkamp, L. (2017). Big Data in Finance and the Growth of Large Firms.

Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, *11*, 21.

Boyer, R. S., & Moore, J. S. (1977a). A fast string searching algorithm. *Communications of the ACM*, *20*(10), 762–772. https://doi.org/10.1145/359842.359859

Boyer, R. S., & Moore, J. S. (1977b). A fast string searching algorithm. *Communications of the ACM*, *20*(10), 762–772. https://doi.org/10.1145/359842.359859

Buard, J., & Jeffreys, A. J. (1997). Big, bad minisatellites. *Nature Genetics*, *15*(4), 327–328. https://doi.org/10.1038/ng0497-327

Bughin, J., Chui, M., & Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. *McKinsey Quarterly*, 75–86. https://doi.org/10.1109/MC.2012.358

Calladine, C. R., Drew, H., Luisi, B., & Travers, A. (2004). *Understanding DNA: The Molecule and How it Works*. Elsevier.

Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*. https://doi.org/10.1038/371215a0

Chowdhury, N. M. M. K., & Boutaba, R. (2010). A survey of network virtualization. *Computer Networks*, *54*(5), 862–876. https://doi.org/10.1016/j.comnet.2009.10.017

Chung, W. (2014). International Journal of Information Management BizPro : Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, *34*(2), 272–284. https://doi.org/10.1016/j.ijinfomgt.2014.01.001

Cockcroft, S., & Russell, M. (2018). Big Data Opportunities for Accounting and Finance Practice and Research. *Australian Accounting Review*, (October), 1–11. https://doi.org/10.1111/auar.12218

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large

Clusters. *Communications of the ACM*, *51*(1), 107. https://doi.org/10.1145/1327452.1327492

Donald E. Knuth, James H. Morris, J. and V. R. P. (1977). FAST PATTERN MATCHING IN STRINGS. *SIAM Journal on Computing*, *6*(2), 323–350.

Douglas, C., Lowe, J., Malley, O. O., & Reed, B. (n.d.). Apache Hadoop YARN : Yet Another Resource Negotiator.

Edgar, R. C., & Myers, E. W. (2005). PILER: Identification and classification of genomic repeats. *Bioinformatics*, *21*(SUPPL. 1), 152–158. https://doi.org/10.1093/bioinformatics/bti1003

Endo, P. T., Gonçalves, G. E., Kelner, J., & Sadok, D. (2009). A Survey on Open-source Cloud Computing Solutions. *VIII Workshop Em Clouds, Grids e Aplicações*, *508*, 3–16. Retrieved from http://sbrc2010.inf.ufrgs.br/anais/data/pdf/wcga/st01_01_wcga.pdf

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Fau, A., Mesran, & Ginting, G. L. (2017). Analisa Perbandingan Boyer Moore Dan Knuth Morris Pratt Dalam Pencarian Judul Buku Menerapkan Metode Perbandingan Eksponensial (Studi Kasus : Perpustakaan STMIK Budi Darma). *Jurnal Times (Technology Informatics & Computer System)*, *6*(1), 12–22.

Gandomi, A., & Haider, M. (2015). International Journal of Information Management Beyond the hype : Big data concepts , methods , and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Guo, R., Zhao, Y., Zou, Q., Fang, X., & Peng, S. (2018). Bioinformatics applications on Apache Spark, (August). https://doi.org/10.1093/gigascience/giy098/5067872

Hahn, U. (2000). of Automatic Researchers are investigating summarization tools and methods that, (November), 29–36.

Han, Z., & Sql, A. S. (2016). Spark : A Big Data Processing Platform Based On Memory Computing. *Seventh International Symposium on Parallel Architectures, Algorithms and Programming*, 172–176. https://doi.org/10.1109/PAAP.2015.41

Haponiuk, M., Pawe[a]kowicz, M., Przybecki, Z., & Nowak, R. M. (2017). CuGene as a tool to view and explore genomic data Keywords :, *10445*, 1–8. https://doi.org/10.1117/12.2280533

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115. https://doi.org/10.1016/j.is.2014.07.006

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hirschberg, J., Hjalmarsson, A., & Elhadad, N. (2010). " You ' re as Sick as You Sound ": Using Computational Approaches for Modeling Speaker State to Gauge Illness and Recovery, 305–322. https://doi.org/10.1007/978-1-4419-5951-5

Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). *Learning spark: lightning-fast big data analysis*. O'Reilly Media, Inc.

Kayyali, B., Knott, D., & Kuiken, S. Van. (2013). The big-data revolution in US health care : Accelerating value and innovation. *McKinsey & Company*, (April), 1–6. https://doi.org/10.1145/2537052.2537073

Lee, K.-H., Lee, Y.-J., Choi, H., Chung, Y. D., & Moon, B. (2012). Parallel data processing with MapReduce. *ACM SIGMOD Record*, *40*(4), 11. https://doi.org/10.1145/2094114.2094118

Liao, X., Gao, Z., Ji, W., & Wang, Y. (2015). An Enforcement of Real Time Scheduling in Spark Streaming.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). Review What is bioinformatics ? An. *Gene Expression*, *40*(4), 83–100.

https://doi.org/10.1053/j.ro.2009.03.010

McAfee, A., & Brynjolfsson, E. (2012). Big data: The Management Revolution. *Harvard Business Review*, (October), 1–9. https://doi.org/00475394

Mishra, A. K., Hellerstein, J. L., Cirne, W., & Das, C. R. (2010). Towards characterizing cloud backend workloads: insights from Google compute clusters. *ACM SIGMETRICS Performance Evaluation Review*, *37*(4), 34–41. https://doi.org/10.1145/1773394.1773400

Mohammadi, M., & Al-Fuqaha, A. (2018). Enabling Cognitive Smart Cities Using Big Data and Machine Learning: Approaches and Challenges. *IEEE Communications Magazine*, *56*(2), 94–101. https://doi.org/10.1109/MCOM.2018.1700298

Murdoch, T. B. T. B., & Detsky, A. S. A. S. (2013). The inevitable application of big data to health care. *Jama*, *309*(13), 1351–1352. https://doi.org/10.1001/jama.2013.393

Nair, L. R., Shetty, S. D., & Shetty, S. D. (2018). Applying spark based machine learning model on streaming big data for health status prediction. *Computers and Electrical Engineering*, *65*, 393–399. https://doi.org/10.1016/j.compeleceng.2017.03.009

Nekrutenko, A., Eberhard, C., Houwaart, T., Coraor, N., Rebolledo-Jaramillo, B., Chilton, J., … Backofen, R. (2017). Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers. *PLOS Computational Biology*, *13*(5), e1005425. https://doi.org/10.1371/journal.pcbi.1005425

Neumeyer, L., Robbins, B., Nair, A., & Kesari, A. (2010). S4: Distributed stream computing platform. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 170–177. https://doi.org/10.1109/ICDMW.2010.172

Nguyen, H., Case, D. A., & Rose, A. S. (2018). NGLview-interactive molecular graphics for Jupyter notebooks. *Bioinformatics*, *34*(7), 1241–1242. https://doi.org/10.1093/bioinformatics/btx789

Oppitz, M., & Tomsu, P. (2017). Inventing the cloud century: How cloudiness keeps changing our life, economy and technology. *Inventing the Cloud Century: How*

Cloudiness Keeps Changing Our Life, *Economy and Technology*, 1–609. https://doi.org/10.1007/978-3-319-61161-7

Orr, H. T., & Zoghbi, H. Y. (2014). Trinucleotide Repeat Disorders. *Encyclopedia of the Neurological Sciences*, *30*, 525–533. https://doi.org/10.1016/B978-0-12-385157-4.00649-7

Pahadia, M., Srivastava, A., Srivastava, D., & Patil, N. (2015). Genome Data Analysis Using MapReduce Paradigm. *2015 Second International Conference on Advances in Computing and Communication Engineering*, 556–559. https://doi.org/10.1109/ICACCE.2015.68

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *Proceedings of the LREC*, 1320–1326.

Patil, H. A. (2010). " Cry Baby ": Using Spectrographic Analysis to Assess Neonatal Health Status from an Infant ' s Cry. https://doi.org/10.1007/978-1-4419-5951-5

Perez, F., & Granger, B. E. (2015a). Project Jupyter : Computational Narratives as the Engine of Collaborative Data Science. *Retrieved September*, (April), 1–24. Retrieved from http://archive.ipython.org/JupyterGrantNarrative-2015.pdf

Perez, F., & Granger, B. E. (2015b). Project Jupyter : Computational Narratives as the Engine of Collaborative Data Science. *Retrieved September*, (April), 1–24.

Rachman, A. B. (2018). *DETEKSI GENOMIC REPEATS MENGGUNAKAN ALGORITMA KNUTH-MORRIS-PRATT PADA R HIGH-PERFORMANCE COMPUTING PACKAGE*. Bandung: Universitas Pendidikan Indonesia.

Ramírez-Gallego, S., Fernández, A., García, S., Chen, M., & Herrera, F. (2018). Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. *Information Fusion*, *42*, 51–61. https://doi.org/10.1016/j.inffus.2017.10.001

Richard Cole. (1994). TIGHT BOUNDS ON THE COMPLEXITY OF THE BOYER-MOORE STRING MATCHING ALGORITHM. *SIAM Journal on Computing*,

23(5), 1075–1091.

Richard, G.-F., Kerrest, A., & Dujon, B. (2008). Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. *Microbiology and Molecular Biology Reviews*, *72*(4), 686–727. https://doi.org/10.1128/MMBR.00011-08

Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online*, *24*, 4–15. https://doi.org/10.7152/acro.v24i1.14671

Ryza, S., Laserson, U., Owen, S., Wills, J. (2015). *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. O'Reilly Media.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

Salloum, S., Dautov, R., Chen, X., Xiaogang, P., & Huang, J. Z. (2016). Big data analytics on Apache Spark. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-016-0027-9

Shanahan, J. G., Street, H., Street, H., & Francisco, S. (2015). Large Scale Distributed Data Science using Apache Spark. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2323–2324.

Sheik, S. S., Aggarwal, S. K., Poddar, A., Balakrishnan, N., & Sekar, K. (2004). A fast pattern matching algorithm. *Journal of Chemical Information and Computer Sciences*, *44*(4), 1251–1256. https://doi.org/10.1021/ci030463z

Shibata, Y., Matsumoto, T., & Takeda, M. (2000). A Boyer – Moore Type Algorithm for, 181–194.

Sommerville, I. (2011). *Software engineering 9th Edition*.

Sukamto, R. A., & Shalahuddin, M. (2011). *Modul Pembelajaran Rekayasa Perangkat Lunak (Terstrktur dan Beroientasi Objek)*. Bandung: Modula.

Taylor, R. C. (2010). An overview of the Hadoop/MapReduce/HBase framework and its

current applications in bioinformatics. *BMC Bioinformatics*, *11*(SUPPL. 12), S1. https://doi.org/10.1186/1471-2105-11-S12-S1

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Wyman, A. R., & White, R. (1980). A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *77*(11), 6754–6758. https://doi.org/10.1073/pnas.77.11.6754

Zaharia, B. Y. M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., … Gonzalez, J. (2016). Apache Spark : A Unified Engine for Big Data Processing. *Communications of the ACM*, *59*(11), 56–65.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark : Cluster Computing with Working Sets. *HotCloud 2010*.

Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., & Stoica, I. (2013). Discretized Streams : Fault-Tolerant Streaming Computation at Scale. *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, (1), 423–438.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026