BAB II

LANDASAN TEORI

2.1 Big Data

2.1.1. Data, Informasi, dan Pengetahuan

Data adalah suatu bahan mentah yang kelak dapat diolah lebih lanjut menjadi sesuatu yang lebih bermakna. Data inilah yang nantinya akan disimpan dalam database (Irsyad, 2017).

Data adalah elemen yang paling dasar, bersifat diskrit, dan belum diproses, sehingga belum memiliki makna. Contoh: angka, kata, kode, tabel, dan basis data (Halim, 2018).

Berdasarkan pengertian di atas,data merupakan fakta-fakta yang masih mentah dan masih harus diolah supaya dapat dipahami dan digunakan orang lain sehingga dapat membentuk sebuah informasi.

Informasi bermuara pada data yang telah diproses sehingga memberi nilai tambah atau pengetahuan bagi yang menggunakannya. Suatu data yang telah diolah atau diproses menjadi suatu informasi dapat digunakan untuk menghasilkan keputusan baik keputusan jangka pendek maupun keputusan jangka panjang (Irsyad, 2017).

Berdasarkan pengertian di atas, informasi adalah data-data yang telah diolah sedemikian rupa sehingga dapat berguna untuk menambah pengetahuan bagi penerimanya.

Pengetahuan adalah milik dari isi pikiran. Jadi pengetahuan merupakan hasil proses dari usaha manusia untuk tahu. Pengetahuan yang diperoleh merupakan informasi yang ditangkap oleh panca indra manusia (Darmawan, 2016).

Menurut (Jr & Prince, 2011) Pengetahuan (*knowledge*) terdiri dari data atau informasi yang telah terorganisasi dan proses untuk memberikan pemahaman, pengalaman, dan pembelajaran, serta keahlian terhadap problema bisnis yang sedang dihadapi.

Berdasarkan pengertian di atas, pengetahuan (*knowledge*) adalah suatu rangkaian informasi yang diterima dan digunakan untuk mengidentifikasi, menciptakan atau menjelaskan sesuatu dan pemahaman akan fakta, kebenaran atau informasi yang diperoleh melalui pengalaman atau pembelajaran.

2.1.2. Big Data

Big Data merupakan kumpulan data yang volume datanya super besar, memiliki keragaman sumber data yang tinggi, sehingga perlu dikelola dengan metode dan perangkat bantu yang kinerjanya sesuai (Maryanto, 2017).

Big data digambarkan sebagai kumpulan data yang besar dengan penurunan biaya penyimpanan data dan peningkatan daya komputasi (Tattersall & Grant, 2016).

Sedangkan (Dumbill, 2013) berpendapat 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.

Menurut pengertian para ahli di atas, dapat disimpulkan bahwa *Big Data* adalah kumpulan data yang sangat besar yang tidak bisa diproses jika hanya menggunakan aplikasi pemrosesan data tradisional.

2.1.3. Dimensi-Dimensi *Big Data*

Ada 5 dimensi awal dalam Big Data yaitu 3V: Volume, Variety dan Velocity.

Gambar 2.1 Dimensi *Big Data* (Kannadasan et al., 2015)

a. Volume

Berbagai perusahaan selalu tertimbun dengan data yang terus tumbuh dari semua jenis sektor, maka dari itu perusahan dapat dengan mudahnya mengumpulkan terabyte bahkan petabyte-informasi.

1. Mengubah 12 terabyte Tweet dibuat setiap hari ke dalam peningkatan sentimen analisis produk.

2. Mengkonvert 350 miliar pembacaan tahunan untuk lebih baik dalam memprediksi kemampuan beli pasar.

Mungkin karakteristik di atas paling mudah dimengerti karena besarnya data. Volume juga dapat mengacu pada jumlah massa data, bahwa suatu organisasi berusaha untuk memanfaatkan data untuk meningkatkan pengambilan keputusan yang banyak diterapkan di berbagai perusahaan di banyak negara. Volume data juga terus meningkat dan belum pernah terjadi sampai setinggi ini sehingga tidak dapat diprediksi jumlah pasti dan juga ukuran dari data sekitar lebih kecil dari petabyte sampai zetabyte. Dataset *Big Data* sekitar 1 terabyte sampai 1 petabyte perperusahaan jadi jika *Big Data* digabungkan dalam sebuah organisasi/group perusahaan ukurannya mungkin bisa sampai zetabyte dan jika hari ini jumlah data sampai 1000 zetabyte, besok pasti akan lebih tinggi dari 1000 zetabyte.

b. Variety

Volume data yang sekian banyak tersebut bertambah dengan kecepatan yang begitu cepat sehingga sulit bagi kita untuk mengelola hal tersebut. Terkadang hanya dengan kurun waktu 2 menit sudah menjadi terlambat. Proses dalam waktu sensitif seperti penangkapan penipuan, data yang besar harus digunakan sebagai aliran ke dalam perusahaan Anda untuk memaksimalkan nilainya.

1. Meneliti 5 juta transaksi yang dibuat setiap hari untuk mengidentifikasi potensi penipuan

2. Menganalisis 500 juta detail catatan panggilan setiap hari secara *real-time* untuk memprediksi melonjak pelanggan lebih cepat.

Berbagai jenis data dan sumber data. Variasi adalah tentang cara mengelola kompleksitas dari beberapa jenis data, termasuk *structured* data, *unstructured* data dan *semi-structured* data. Organisasi perlu mengintegrasikan dan menganalisis data dari *array* yang kompleks dari kedua sumber informasi *Traditional* dan *non traditional* informasi, dari dalam dan luar perusahaan. Dengan begitu banyaknya sensor, perangkat pintar (*smart device*) dan teknologi kolaborasi sosial, data yang dihasilkan dalam bentuk yang tak terhitung jumlahnya, termasuk data text, web data, tweet, sensor data, audio, *video*, *click stream*, *log file* dan banyak lagi.

c. Velocity

Big Data adalah setiap jenis data-data baik yang terstruktur maupun tidak terstruktur seperti *teks*, data sensor, audio, video, *click stream*, *file log* dan banyak lagi. Wawasan baru ditemukan ketika menganalisis kedua jenis data ini bersama-sama.

1. Memantau 100 video masukan langsung dari kamera pengintai untuk menargetkan tempat tujuan.
2. Mengeksploitasi 80% perkembangan data dalam gambar, video, dan dokumen untuk meningkatkan kepuasan pelanggan.

Kecepatan dimana data dibuat, diolah dan dianalisis terus menerus. Berkontribusi untuk kecepatan yang lebih tinggi adalah sifat penciptaan

data secara *real-time*, serta kebutuhan untuk memasukkan *streaming* data ke dalam proses bisnis dan dalam pengambilan keputusan. Dampak *Velocity latency*, jeda waktu antara saat data dibuat atau data yang ditangkap, dan ketika itu juga dapat diakses. Hari ini, data terus-menerus dihasilkan pada kecepatan yang mustahil untuk sistem tradisional untuk menangkap, menyimpan dan menganalisisnya. Jenis tertentu dari data harus dianalisis secara *real time* untuk menjadi nilai bagi bisnis.

d. Veracity

Konsep ini berkaitan dengan kualitas dan akurasi data yang dikumpulkan. Terkadang, data yang dikumpulkan mungkin memiliki bagian yang hilang, tidak akurat, atau bahkan tidak memberikan wawasan yang berguna. Veracity merujuk pada tingkat kepercayaan yang ada dalam data yang dikumpulkan secara keseluruhan.

Data yang tidak akurat atau tidak lengkap dapat menyebabkan kebingungan dan bahkan risiko kesalahan yang serius, terutama dalam bidang medis. Jika informasi tentang obat yang dikonsumsi oleh pasien tidak lengkap atau tidak benar, nyawa pasien dapat terancam.

Oleh karena itu, nilai dan kebenaran data sangat penting untuk menentukan kualitas dan wawasan yang dihasilkan dari analisis data.

e. Value

Hal yang paling penting dalam konteks bisnis adalah Nilai. Jika sistem Big Data tidak dapat menghasilkan nilai dari keseluruhan proses

dalam waktu yang wajar, maka proses tersebut tidak akan bermanfaat untuk terlibat dalam bisnis.

Teoritisnya, Big Data seharusnya memberikan nilai bagi bisnis. Besar atau kecilnya nilai tersebut tergantung pada tim analisis dan peneliti untuk memikirkan, merancang, membangun, dan menyampaikannya.

Nilai adalah salah satu karakteristik utama yang dibahas dalam bisnis, dan tingkat nilai tertentu harus diproyeksikan di awal proyek Big Data.

2.1.4. Arsitektur *Big Data*

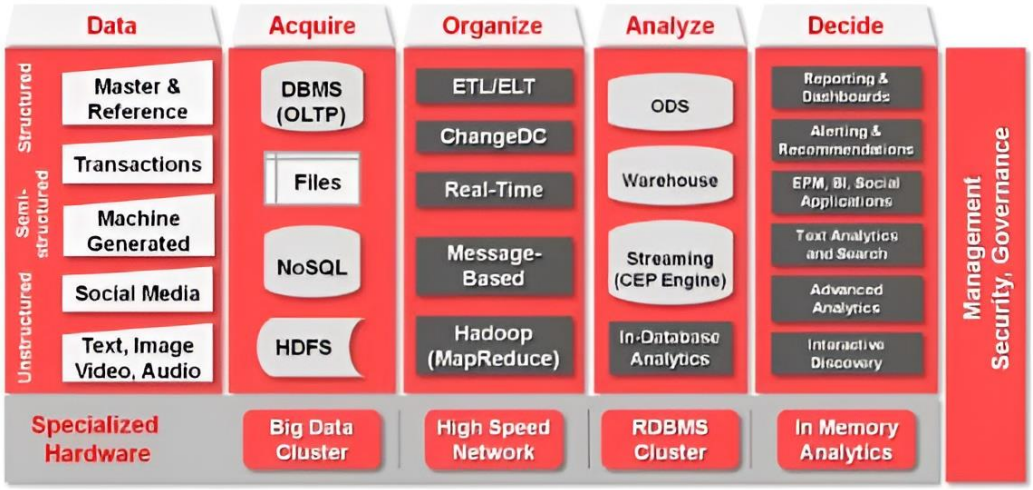
Memahami level aspek arsitektur yang tinggi dari *Big Data*, sebelumnya harus memahami arsitektur informasi logis untuk data yang terstruktur. Pada gambar di bawah ini menunjukkan dua sumber data yang menggunakan teknik integrasi (*ETL*/*Change Data Capture*) untuk mentransfer data ke dalam DBMS *data warehouse* atau operational *data store*, lalu menyediakan bermacam-macam variasi dari kemampuan analisis untuk menampilkan data. Beberapa kemampuan analisis ini termasuk ; *dashboards*, laporan, *EPM/BI Applications*, ringkasan dan *query statistic*, interpretasi *semantic* untuk data tekstual, dan alat visualisasi untuk data yang padat. Informasi utama dalam prinsip arsitektur ini termasuk cara memperlakukan data sebagai asset melalui nilai, biaya, resiko, waktu, kualitas dan akurasi data.

Gambar 2.2 *Traditional Information Architecture Capabilities* (Heller & Piziak, 2015)

Mendefinisikan kemampuan memproses untuk *big data architecture*, diperlukan beberapa hal yang perlu dilengkapi; volume, percepatan, variasi, dan nilai yang menjadi tuntutan. Ada strategi teknologi yang berbeda untuk *real-time* dan keperluan *batch processing*. Untuk *real-time*, menyimpan data nilai kunci, seperti NoSQL, memungkinkan untuk performa tinggi, dan pengambilan data berdasarkan indeks. *Batch processing*, digunakan teknik yang dikenal sebagai *Mapreduce*, memfilter data berdasarkan pada data yang spesifik pada strategi penemuan. Setelah data yang difilter ditemukan, maka akan dianalisis secara langsung, dimasukkan ke dalam *unstructured database* yang lain, dikirimkan ke dalam perangkat *mobile* atau digabungkan ke dalam lingkungan *data warehouse* tradisional dan berkorelasi pada data terstruktur.

Gambar 2.3 *Big Data Information Architecture Capabilities* (Heller & Piziak, 2015)

Kekuatan informasi ada dalam kemampuan untuk asosiasi dan korelasi. Maka yang dibutuhkan adalah kemampuan untuk membawa sumber data yang berbeda-beda, memproses kebutuhan bersama – sama secara tepat waktu dan analisis yang berharga.



Gambar 2.4 *Oracle Integrated Information Architecture Capabilities* (Heller & Piziak, 2015)

Ketika bermacam – macam data telah didapatkan, data tersebut dapat disimpan dan diproses ke dalam DBMS tradisional, *simple files*, atau sistem cluster terdistribusi seperti NoSQL dan *Hadoop Distributed File System* (HDFS).

Secara arsitektur, komponen kritikal yang memecah bagian tersebut adalah layer integrasi yang ada di tengah. Layer integrasi ini perlu untuk diperluas ke seluruh tipe data dan domain, dan menjadi jembatan antara data penerimaan yang baru dan tradisional, dan pengolahan kerangka. Kapabilitas integrasi data perlu untuk menutupi keseluruhan spektrum dari kecepatan dan frekuensi. Hal tersebut diperlukan untuk menangani kebutuhan ekstrim dan

volume yang terus bertambah banyak. Oleh karena itu diperlukan teknologi yang memungkinkan untuk mengintegrasikan Hadoop / Mapreduce dengan *data warehouse* dan data transaksi.

Layer berikutnya digunakan untuk *Load* hasil reduksi dari *big data* ke dalam *data warehouse* untuk analisis lebih lanjut. Diperlukan juga kemampuan untuk mengakses data terstruktur seperti informasi profil pelanggan ketika memproses dalam *big data* untuk mendapatkan pola seperti mendeteksi aktivitas yang mencurigakan.

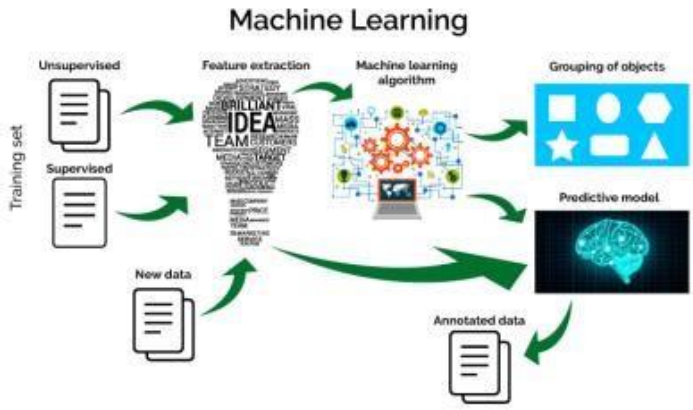
Hasil pemrosesan data akan dimasukkan ke dalam ODS tradisional, *data warehouse,* dan data marts untuk analisis lebih lanjut seperti data transaksi. Komponen tambahan dalam layer ini adalah *Complex Event Processing* untuk menganalisis arus data secara *real-time*. *Layer business intelligence* akan dilengkapi dengan analisis lanjutan, dalam analisis *database* statistik, dan visualisasi lanjutan, diterapkan dalam komponen tradisional seperti laporan, *dashboards*, dan *query*. Pemerintahan, keamanan, dan pengelolaan operasional juga mencakup seluruh spektrum data dan lanskap informasi pada tingkat enterprise.

Dengan arsitektur ini, pengguna bisnis tidak melihat suatu pemisah, bahkan tidak sadar akan perbedaan antara data transaksi tradisional dan big data. Data dan arus analisis akan terasa mulus tanpa halangan ketika dihadapkan pada bermacam – macam data dan set informasi, hipotesis, pola analisis, dan membuat keputusan.

2.2. *Machine Learning*

Machine Learning adalah aplikasi kecerdasan buatan (AI) yang menyediakan sistem kemampuan untuk secara otomatis belajar dan meningkatkan pengalaman tanpa diprogram secara eksplisit (Varone et al., 2020).

Machine learning merupakan serangkaian teknik yang dapat membantu dalam menangani dan memprediksi data yang sangat besar dengan cara mempresentasikan data-data tersebut dengan algoritma pembelajaran (Danakusumo, 2017).



Gambar 2.5 *Machine Learning* (Pantech, 2018)

Istilah *machine learning* pertama kali didefinisikan oleh Arthur Samuel di tahun 1959. *Machine learning* adalah salah satu bidang ilmu komputer yang memberikan kemampuan pembelajaran kepada komputer untuk mengetahui sesuatu tanpa pemrogram yang jelas (Samuel, 1959).

Machine Learning (ML) atau pembelajaran mesin merupakan pendekatan dalam AI yang banyak digunakan untuk menggantikan atau menirukan perilaku manusia untuk menyelesaikan masalah atau melakukan otomatisasi. Sesuai namanya, ML mencoba menirukan bagaimana proses manusia atau makhluk cerdas belajar dan menggeneralisasi (Hania, 2017).

Dalam pembelajaran *machine learning*, terdapat beberapa skenario-skenario. seperti:

1. *Supervised Learning*

   Penggunaan skenario *supervised learning*, pembelajaran menggunakan masukan data pembelajaran yang telah diberi label. Setelah itu membuat prediksi dari data yang telah diberi label.

2. *Unsupervised Learning*

   Penggunaan skenario *Unsupervised Learning*, pembelajaran menggunakan masukan data pembelajaran yang tidak diberi label. Setelah itu mencoba untuk mengelompokan data berdasarkan karakteristik-karakteristik yang ditemui.

3. *Reinforcement Learning*

   Pada skenario *reinforcement learning* fase pembelajaran dan tes saling dicampur. Dalam mengumpulkan informasi pembelajar secara aktif dengan berinteraksi ke lingkungan sehingga untuk mendapatkan balasan untuk setiap aksi dari pembelajar. Saat ini telah banyak pendekatan machine learning yang digunakan untuk deteksi spam, *Optical character recognition (OCR)*, pengenalan wajah, deteksi penipuan online, *NER (Named Entity Recognition)*, dan *Part-of-Speech Tagger*.

2.3. *Support Vector Machine*

Support Vector Machine (SVM) adalah sistem pembelajaran yang menggunakan ruang hipotesis berupa fungsi-fungsi linier dalam sebuah ruang fitur (*feature space*) berdimensi tinggi, dilatih dengan algoritma pembelajaran yang

didasarkan pada teori optimasi dengan mengimplementasikan learning bias yang berasal dari teori pembelajaran statistik (Munawarah et al., 2016).

Menurut (Nugroho et al., 2003) SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*.

Sementara itu, (Islami, 2019) SVM adalah metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space*. Tulisan ini membahas teori dasar SVM dan aplikasinya dalam bioinformatika, khususnya pada analisis ekspresi gen yang diperoleh dari analisis *microarray.*

Support Vector Machine (SVM) merupakan salah satu metode dalam supervised learning yang biasanya digunakan untuk klasifikasi (seperti *Support Vector Classification*) dan regresi (*Support Vector Regression*). Dalam pemodelan klasifikasi, SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya. SVM juga dapat mengatasi masalah klasifikasi dan regresi dengan linier maupun non linier (Samsudiney, 2019).

2.4. *Apache Spark*

Apache Spark adalah mesin analitik terpadu untuk pemrosesan data skala besar. Ini awalnya dikembangkan pada 2009 di AMPLab UC Berkeley, dan bersumber terbuka pada 2010 sebagai proyek Apache (Apache Software Foundation, 2018).

Sedangkan menurut (Wijaya, 2018) *Apache Spark* adalah *engine* (perangkat lunak) analisis terpadu super cepat untuk memproses data dalam skala besar; meliputi *Big Data* dan *machine learning*. Secara lebih detailnya, *Apache Spark* dapat didefinisikan sebagai *engine* (perangkat lunak) untuk memproses data dalam skala besar secara *in-memory*, dilengkapi dengan API pengembangan yang elegan dan ekspresif guna memudahkan para pekerja data dalam mengeksekusi pekerjaan-pekerjaan yang membutuhkan perulangan akses yang cepat terhadap data yang diproses, seperti halnya *streaming*, *machine learning*, maupun SQL, secara efisien.

Sebagai perangkat lunak untuk memproses data dalam skala besar, *Apache Spark* memiliki sejumlah keunggulan, diantaranya:

1. Kecepatan. *Apache Spark* mampu bekerja 100 kali lebih cepat dibanding Hadoop. Berkat penggunaan *state-of-the-art DAG scheduler, query optimizer*, dan *physical execution engine*, *Apache Spark* dapat mencapai performa tinggi baik dalam hal pemrosesan data yang sifatnya batch maupun streaming.
2. Mudah Digunakan. Dapat menggunakan bahasa program Java, Scala, Python, R, dan SQL untuk mengembangkan aplikasi yang menggunakan *Apache Spark*. Spark menyediakan lebih dari 80 operator level tinggi yang dapat memudahkan pengembang untuk membangun aplikasi secara paralel. *Apache Spark* juga dapat digunakan secara interaktif dari shell Scala, Python, R, dan SQL.
3. Memiliki cakupan yang luas. *Apache Spark* menggabungkan SQL, *streaming*, dan analitik yang kompleks menyediakan setumpuk library perangkat lunak meliputi SQL dan DataFrames, MLlib untuk *machine learning*, GraphX, dan

Spark Streaming. Pengembang aplikasi dapat menggabungkan semua library ini dengan mulus dalam satu aplikasi yang sama.

4. Dapat dijalankan Dimana-mana. *Apache Spark* dapat dijalankan di Hadoop YARN, Apache Mesos, Kubernetes, dengan mode *standalone* maupun *cluster*, atau di platform cloud seperti EC2. Pada dasarnya, Spark dapat mengakses berbagai tipe sumber data seperti halnya HDFS, Apache Cassandra, Apache HBase, Apache Hive, dan ratusan sumber data lainnya.

2.5. *Hadoop Mapreduce*

Menurut (RYANTO, 2017) Hadoop adalah sebuah *Framework* untuk melakukan proses komputasi terdistribusi data yang berukuran besar pada sekumpulan komputer yang saling terhubung (*cluster computing*) dengan menggunakan model pemrograman yang sederhana. Hadoop merupakan *Framework* yang sedang dikembangkan oleh *Apache Software Foundation* yang bersifat *open source* untuk komputasi yang handal, *scalable*, dan terdistribusi.

Hadoop *Framework* merupakan *software* yang bersifat *open source* yang diciptakan untuk mengolah data dan menyimpan data dengan kapasitas yang sangat besar hingga petabyte dimana pengolahan data-data tersebut dilakukan dengan cara mendistribusikan data-data tersebut kedalam beberapa komputer yang telah di cluster (Hurwitz et al., 2013).

2.6. Penelitian Terkait

Penelitian terdahulu sebagai kajian penelitian yang akan dilakukan sangat penting untuk mengetahui hubungan antara penelitian yang dilakukan sebelumnya dengan penelitian yang dilakukan saat ini serta dapat menghindari adanya

duplikasi. Hal ini bermanfaat untuk menunjukkan bahwa penelitian yang dilakukan. mempunyai arti penting sehingga dapat diketahui kontribusi penelitian terhadap ilmu pengetahuan.

Tabel 2.1 Penelitian Terkait

No	Peneliti/Tahun	Judul	Problem	Metode / Algoritma / Teknik / Model / Sensor / Platform	State Of The Art / Keterbaruan
1.	(Gopalani & Arora, 2015)	*Comparing Apache Spark and Mapreduce with Performance Analysis using K-Means*	Analisis perbandingan *Apache Spark* dan Mapreduce menggunakan algoritma K-Means	Algoritma : *K-Means*	Hasil penelitian menghasilkan Tools yang paling akurat & cepat antara *Apache Spark* dan Mapreduce

Lanjutan Tabel 2.1 Penelitian Terkait

2.	(RYANTO, 2017)	Analisis Kinerja *Framework* Big Data Pada Cluster Tervirtualisasi: *Hadoop Mapreduce* Dan *Apache Spark*	Perbandingan kinerja antara *Hadoop Mapreduce* dan *Apache Spark*	Teknik : *big data processing*	Hasil penelitian menghasilkan Tools yang paling baik antara *Hadoop Mapreduce* Dan *Apache Spark*
3.	(L. B. D. Cahyo, 2018)	Implementasi Metode *Support Vector Machine* Untuk Melakukan	Penderita *medulloblastoma* antara 18%-20% dari semua tumor otak pada anak-	Algoritma : *Support Vector Machine* Teknik : *Microarray*	Berdasarkan hasil analisis SVM mampu memprediksi kelas

Lanjutan Tabel 2.1 Penelitian Terkait

		Klasifikasi Pada Data Bioinformatika	anak serta 70% dari penderita *medulloblastoma* terdeteksi pada usia 10 tahun ke bawah.		Penderita dengan akurasi 95% dengan nilai AUC 98%
4.	(Oliviandi et al., 2018)	Implementasi *Apache Spark* Pada Big Data Berbasis Hadoop *Distributed File System*	Big data merupakan kumpulan data dalam skala besar, yang mempunyai karakteristik data yang variatif, sangat cepat pertumbuhannya dan kompleks datanya.	Algoritma : *Mapreduce*	Skenario yang digunakan adalah memproses *wordcount* suatu data dengan besar data yang berbeda yang bertujuan

Lanjutan Tabel 2.1 Penelitian Terkait

					untuk menganalisis *response time* dan penggunaan hardware dari kedua platform tersebut.

Lanjutan Tabel 2.1 Penelitian Terkait

			dengan teknologi Apache Hadoop dan *Apache Spark*.		mengerjakan kueri dibandingkan Hive.
6.	(K. D. Cahyo, 2018)	Studi dan Implementasi *Apache Spark* MLLIB Untuk Analisis Big Data	Dibutuhkan komputer dengan kekuatan komputasi yang sangat tinggi untuk menganalisis data dengan ukuran yang sangat besar.	Metode : *Machine Learning*	Kinerja dari fungsi-fungsi MLlib sangat baik untuk komputasi pada ukuran data yang besar.

2.6 Penelitian Terdekat

Berikut adalah beberapa penelitian terdekat dengan penelitian ini :

Tabel 2.2 Penelitian terdekat

No	Peneliti/Tahun	Judul	Problem	Metode / Algoritma / Teknik / Model / Sensor / Platform	State Of The Art / Keterbaruan
1.	(Gopalani & Arora, 2015)	*Comparing Apache Spark and Mapreduce with Performance Analysis using K-Means*	Analisis perbandingan *Apache Spark* dan Mapreduce menggunakan algoritma K-Means	Algoritma : *K-Means*	Hasil penelitian menghasilkan Tools yang paling akurat & cepat antara *Apache Spark* dan Mapreduce

Lanjutan tabel 2.2 Penelitian terdekat

2.	(RYANTO, 2017)	Analisis Kinerja *Framework* Big Data Pada Cluster Tervirtualisasi: *Hadoop Mapreduce* Dan *Apache Spark*	Perbandingan kinerja antara *Hadoop Mapreduce* dan *Apache Spark*	Teknik : *big data processing*	Hasil penelitian menghasilkan Tools yang paling baik antara *Hadoop Mapreduce* Dan *Apache Spark*

Tabel 2.2 menjelaskan beberapa penelitian terkait yang dijadikan acuan untuk penelitian dalam bidang *Big Data Analysis*. Penelitian berjudul "*Comparing Apache Spark and Mapreduce with Performance Analysis using K-Means*" yang dilakukan (Gopalani & Arora, 2015) dengan menggunakan algoritma *K-Means.*

Pada penelitian yang berjudul "*Analisis Kinerja Framework Big Data Pada Cluster Tervirtualisasi: Hadoop Mapreduce Dan Apache Spark*" (RYANTO, 2017), telah dijelaskan untuk melakukan pemrosesan dan analisis data yang sangat besar, kedua *Framework* direkomendasikan untuk berjalan pada server fisik, tetapi dalam membangun cluster dengan server fisik membutuhkan biaya yang tidak sedikit. Cluster fisik membutuhkan energi yang tinggi dan kaku dalam pengelolaannya. Maka, teknologi virtualisasi menjadi solusi dalam membangun cluster yang fleksibel dan rendah biaya. Dalam penelitian ini, akan dibuat cluster *Hadoop Mapreduce* dan *Apache Spark* yang tervirtualisasi memanfaatkan *hypervisor Proxmox Virtual Environment* lalu akan dilakukan analisis pada kinerja komputasi dan kinerja I/O cluster dari kedua *Framework*.

Hasil penelitian menunjukkan pada pengujian kinerja komputasi menunjukkan *Apache Spark* lebih cepat 3-5 kali lipat pada single node cluster tervirtualisasi dan lebih cepat 1-4 kali lipat pada multi node cluster tervirtualisasi dibandingkan dengan kinerja komputasi *Hadoop Mapreduce*. Serta pada pengujian kinerja I/O cluster, *throughput* yang diberikan lebih besar ketika digunakan bersama dengan *Apache Spark*.

Berdasarkan dua penelitian yang telah disebutkan, maka diusulkan penelitian yang berjudul

"Analisis Perbandingan Performa Apache Spark Dan Hadoop Mapreduce Pada Mapreduce Framework Menggunakan Algoritma Support Vector Machine".