BIG DATA ANALITIK

DENGAN APACHE SPARK

UNDANG-UNDANG REPUBLIK INDONESIA
NOMOR 28 TAHUN 2014
TENTANG HAK CIPTA

PASAL 113
KETENTUAN PIDANA
SANKSI PELANGGARAN

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp100.000.000 (seratus juta rupiah).

2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

3. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/atau pidana denda paling banyak Rp1.000.000.000,00 (satu miliar rupiah).

4. Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rp4.000.000.000,00 (empat miliar rupiah).

BIG DATA ANALITIK

DENGAN APACHE SPARK

I Made Suartana

Global Aksara Pers

BIG DATA ANALITIK DENGAN APACHE SPARK

Diterbitkan pertama kali dalam bahasa Indonesia
oleh Penerbit Global Aksara Pers

ISBN:978-623-462-882-1
xii + 90 hal.; Ukuran Unesco (15,8 x 23 cm)
Cetakan Pertama, Juli 2025

Penulis : I Made Suartana
Penyunting : Alaika M. Bagus Kurnia PS.
Desain kover : Tito Ramadhan
Layouter : Hamim Thohari Mahfudhillah

Diterbitkan oleh:

CV. Global Aksara Pers
Anggota IKAPI, Jawa Timur, 2021,
No. 282/JTI/2021
Jl. Wonocolo Utara V/18 Surabaya
+628977416123/+628573269334
globalaksarapers.com

KATA PENGANTAR

Di era digital saat ini, ledakan volume data yang dikenal sebagai Big Data telah menjadi sebuah tantangan sekaligus peluang besar bagi berbagai organisasi. Kemampuan untuk mengelola dan menganalisis data dalam skala besar dan beragam format menjadi kunci untuk mendapatkan wawasan berharga yang sebelumnya tidak mungkin diperoleh. Sejak akhir 1990-an, analitik Big Data telah berkembang pesat dan kini menjadi alat yang sangat penting untuk memahami pelanggan, proses bisnis, dan tren global. Seiring dengan perkembangan teknologi ini, kita dapat menantikan aplikasi yang lebih transformatif di masa yang akan datang.

Buku ini hadir untuk membahas secara komprehensif bagaimana kerangka kerja (framework) Apache Spark dapat dimanfaatkan secara efektif untuk pemrosesan dan analitik Big Data. Spark, yang dirancang untuk kecepatan dan kemudahan penggunaan, menawarkan kemampuan analitik canggih yang dapat berjalan hingga 100 kali lebih cepat di memori dibandingkan solusi lain seperti Hadoop MapReduce. Fleksibilitasnya memungkinkan integrasi dengan berbagai teknologi lain, seperti Kafka untuk streaming data dan Cassandra untuk penyimpanan, serta mendukung berbagai bahasa pemrograman termasuk Python, Java, dan Scala.

Melalui studi kasus praktis dalam analisis data trafik jaringan komputer, buku ini akan memandu pembaca dalam mengelola kumpulan data besar dan tidak terstruktur untuk mengidentifikasi pola-pola penting. Dengan menggunakan dataset KDD CUP 99 sebagai contoh, pembaca akan diajak untuk mempraktikkan berbagai tahapan dalam siklus hidup analitik Big Data, mulai dari pengumpulan dan pembersihan data, analisis eksplorasi, hingga penerapan model machine learning untuk klasifikasi.

Pembahasan akan mencakup konsep-konsep fundamental seperti 7V Big Data (Volume, Variety, Velocity, Variability, Veracity, Visualization, dan Value), jenis-jenis analitik (Deskriptif, Diagnostik, Prediktif, dan Preskriptif), serta komponen inti dan operasi utama dalam Apache Spark. Lebih lanjut, buku ini akan mengupas penggunaan Spark SQL untuk kueri data terstruktur, dan pustaka MLlib untuk membangun model klasifikasi seperti Regresi Logistik dan Pohon Keputusan (Decision Tree).

Meskipun Spark menawarkan banyak keunggulan, perlu diakui bahwa platform ini masih memerlukan pengembangan lebih lanjut di beberapa area, seperti keamanan dan integrasi dengan perangkat Business Intelligence. Namun demikian, dengan kemampuan yang dimilikinya saat ini, Apache Spark tetap menjadi salah satu teknologi terdepan dan paling menjanjikan dalam ekosistem Big Data.

Buku ini diharapkan dapat menjadi panduan yang solid bagi para praktisi data, mahasiswa, dan siapa pun yang tertarik untuk menguasai analitik Big Data menggunakan Apache Spark.

Penulis

DAFTAR ISI

BAB 1

PENDAHULUAN

Dalam buku ini, akan dibahas bagaimana kerangka kerja Apache Spark membantu pemrosesan Big Data dan analitik dengan standar API dari Apache Spark. Spark didasarkan pada sistem penyimpanan file HDFS yang sama dengan Hadoop, sehingga Apache Spark dan MapReduce dapat digunakan bersama-sama jika sebelumnya sudah mengimplementasikan sistem penyimpanan Big Data dengan Hadoop.

Pemrosesan Spark juga bisa digabungkan dengan Spark SQL, Machine Learning, dan Spark Streaming. Dengan beberapa integrasi dan adaptor di Spark, Teknologi ini juga dapat digunakan dengan teknologi lain. Contohnya adalah penggunaan Spark, Kafka, dan Apache Cassandra secara bersamaan dimana Kafka dapat digunakan untuk streaming data masukkan, Spark untuk melakukan komputasi, dan terakhir database Cassandra NoSQL untuk menyimpan data hasil komputasi. Namun perlu diingat, Spark masih membutuhkan peningkatan lebih lanjut di area seperti keamanan dan integrasi dengan perangkat Business Inteligent.

Sejarah analitik Big Data dapat ditelusuri kembali ke masa awal komputasi, ketika organisasi pertama kali mulai menggunakan komputer untuk menyimpan dan menganalisis

data dalam jumlah besar. Namun, baru pada akhir 1990-an dan awal 2000-an analitik Big Data benar-benar mulai berkembang pesat, karena organisasi semakin banyak beralih ke komputer untuk membantu mereka memahami volume data yang berkembang pesat yang dihasilkan oleh bisnis mereka.

Sekarang ini, analitik Big Data telah menjadi alat yang sangat penting. Dengan memanfaatkan teknologi Big Data, organisasi dapat memperoleh wawasan tentang pelanggan mereka, bisnis mereka, dan dunia di sekitar mereka yang sebelumnya tidak mungkin dilakukan. Seiring bidang analitik Big Data terus berkembang, kita dapat berharap untuk melihat aplikasi yang lebih menakjubkan dan transformatif dari teknologi ini di tahun-tahun mendatang.

Dalam buku ini dibahas penggunaan pendekatan Big Data Analitik dengan menggunakan teknologi Apache Spark dalam mengelola data trafik jaringan. Trafik jaringan merupakan kumpulan data besar yang tidak terstruktur. Kemampuan analitik big data digunakan untuk mengelola data trafik jaringan dan melihat pola dari data. Dataset KDD CUP 99 digunakan sebagai contoh data untuk uji coba.

BAB 2

BIG DATA

A. Big Data

Istilah Big Data sendiri menunjukkan volume yang sangat besar, kecepatan tinggi dalam terciptanya atau kemunculan data, dan keragaman struktur dan format data. Big data yang dihasilkan dapat berupa data terstruktur, data semi-terstruktur atau data tidak terstruktur. Sistem basis data yang ada kesulitan untuk memproses, menganalisis, menyimpan, dan mengelola data semacam itu. Tantangan Big Data adalah perlindungan data, kurasi, pengambilan, analisis, pencarian, visualisasi, penyimpanan, transfer, dan penyebarluasan data.

B. 7 V Big data

Big Data lebih dari sekadar memiliki banyak data atau memilikinya dalam jumlah besar. Big Data adalah salah satu jenis data yang berasal dari berbagai sumber dan terdiri dari berbagai jenis data yang berbeda dalam berbagai format yang berbeda. Dalam konteks Big Data, mengacu pada kumpulan data dengan ukuran yang sangat besar sehingga sistem basis data standar tidak mampu memproses informasi secara tepat waktu dan efisien. Namun, Big Data memiliki lebih dari sekadar

ukurannya, dan di sinilah hal-hal menjadi menarik. Awalnya Big Data digambarkan terdiri dari tiga dimensi (3V): volume tinggi, kecepatan tinggi, dan variasi tinggi. Namun, ada versi "V" lain yang dapat digunakan untuk lebih memahami sifat dan konsekuensi Big Data yang sebenarnya.

7 V ke Big Data:

1. Volume
2. Variety
3. Velocity
4. Variability
5. Veracity
6. Visualization
7. Value

1. Volume

Sejumlah besar data adalah fitur Big Data yang paling membedakan. Big Data didefinisikan sebagai "BESAR" dalam konteks ini dengan volume frase. Dengan banyaknya data yang dihasilkan setiap hari, kita tahu bahwa gigabyte tidak cukup untuk menyimpan data dalam jumlah besar. Akibatnya, data sekarang disimpan dalam bentuk Zettabytes, Exabytes, dan Yottabytes, bukan megabyte. Jumlah kumpulan data yang harus dievaluasi dan diproses disebut sebagai volume data. Kumpulan data ini sekarang secara teratur berukuran lebih besar dari terabyte dan petabyte dan harus diproses secara *real-time*. Volume data yang sangat besar memerlukan pengembangan teknologi pemrosesan baru dan berbeda yang berbeda dari penyimpanan standar dan kemampuan pemrosesan.

2. Variety

Variasi mengacu pada banyak jenis sumber data, yang menghasilkan peningkatan keragaman data, yang mencakup segala hal mulai dari data tersimpan dan terstruktur yang disimpan dalam database hingga data tidak terstruktur, data semi-terstruktur, dan data dalam berbagai bentuk. Big Data dapat diklasifikasikan menjadi tiga jenis: data terstruktur, data semi-terstruktur, dan data tidak terstruktur. Sekarang ini, data yang dihasilkan dalam jumlah besar hanyalah data yang tidak terstruktur seperti file audio, file video, gambar, file teks, dan sebagainya. Jenis data ini sulit untuk dipetakan karena sifatnya tidak mengikuti aturan apa pun, sehingga sulit untuk memisahkannya dari informasi penting.

Salah satu masalah tersulit dari Big Data ditandai oleh keragamannya. Keberagaman bisa dari data tidak terstruktur, dan bisa berisi berbagai jenis data, mulai dari XML hingga video hingga pesan teks.

3. Velocity

Data dapat diproses dan diakses disebut sebagai kecepatan. Tingkat di mana data dipindahkan, diproses, dan didapat di dalam dan di luar organisasi telah meningkat secara substansial dalam beberapa tahun terakhir. Secara tradisional, model bisnis intelijen membutuhkan waktu berhari-hari untuk diproses, tetapi persyaratan analitik saat ini menuntut agar data direkam dan diproses dalam waktu *real-time*, yang dimungkinkan oleh aliran informasi berkecepatan tinggi yang tersedia saat ini.

Dalam kasus posting media sosial, video YouTube, file audio, dan foto, yang diunggah dalam jumlah ratusan setiap detik, konten tersebut harus tersedia sesegera mungkin.

Ketersediaan dan di mana-mana perangkat yang terhubung ke internet, baik secara nirkabel maupun melalui koneksi kabel, memungkinkan data ditransmisikan hampir secara *real-time*. Saat ini, informasi dipertukarkan dengan kecepatan sangat tinggi.

4. Variability

Variabilitas berbeda dari keragaman karena tidak dapat diprediksi. Dalam statistik, variabilitas mengacu pada fakta bahwa data selalu berubah. Konsep variabilitas terutama berkaitan dengan pemahaman dan interpretasi interpretasi yang benar dari data mentah. Algoritma harus dapat memahami konteks di mana mereka beroperasi dan menguraikan makna yang tepat dari setiap kata di lingkungan khusus mereka.

5. Veracity

Kebenaran didefinisikan sebagai kemampuan untuk mengambil informasi yang dapat dipercaya. Karena kualitas data yang akurat tinggi, dimungkinkan untuk memanfaatkan data tersebut. Ini sangat penting untuk bisnis yang bisnis intinya didasarkan pada penyebaran informasi. Sangat penting untuk memastikan bahwa informasi yang Anda kumpulkan akurat, serta mencegah informasi yang salah dari sistem. Ini juga mengacu pada keandalan atau kualitas data yang diterima dan diproses oleh perusahaan untuk mendapatkan wawasan yang relevan dari data. Namun, beberapa orang berpendapat bahwa, mengingat banyaknya informasi yang sudah tersedia, kejujuran benar-benar merupakan karakteristik sekunder dari Big Data.

6. Visualization

Kemampuan untuk menyampaikan data kepada manajemen untuk tujuan pengambilan keputusan disebut sebagai visualisasi. Ini mengacu pada membuat data yang telah dikumpulkan dan dievaluasi dapat dipahami dan mudah dipahami. Tidak mungkin untuk memanfaatkan data mentah kecuali disajikan dengan cara yang tepat. Data dapat ditampilkan dalam berbagai format, antara lain file excel, dokumen word, grafik, dan sebagainya. Yang paling penting adalah informasinya mudah dibaca, dipahami, dan diperoleh terlepas dari formatnya; untuk alasan ini, visualisasi data sangat penting.

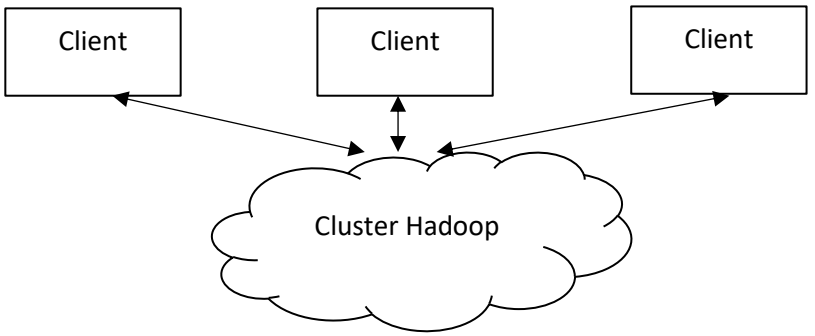
7. Value

Dalam Big Data, data berharga memiliki nilai yang diketahui, yang tercermin dalam pengembalian investasi dari pengelolaan data. Penting bagi setiap pengguna untuk menyadari bahwa perusahaan membutuhkan semacam nilai setelah upaya dan sumber daya dihabiskan untuk V yang disebutkan di atas. Big Data dapat membantu pengguna dalam memberikan nilai jika dikumpulkan dan ditangani secara efektif.

C. Hadoop

Hadoop adalah sumber terbuka (*open source*), penyimpanan Big Data, dan kerangka kerja perangkat lunak pemrosesan data berkecepatan tinggi. Seperti yang ditunjukkan pada Gambar, hadoop menggunakan kumpulan perangkat keras untuk menyimpan dan memproses data besar secara terdistribusi. Memiliki kemampuan penyimpanan data yang luar biasa, pemrosesan data dengan kecepatan tinggi membuat

Hadoop lebih cocok untuk pemrosesan data dalam jumlah besar. Hadoop cluster adalah sekumpulan mesin yang melibatkan kemampuan penyimpanan yang sangat besar, terhubung bersama dalam satu lokasi yaitu cloud. Mesin cloud ini kemudian digunakan untuk penyimpanan dan pemrosesan data. Dari pengguna individu dapat mengirimkan pekerjaan mereka ke cluster. Pengguna mungkin ada di beberapa lokasi dari cluster Hadoop. Sistem file terdistribusi, pemrosesan lebih cepat, transfer data lebih cepat, toleransi kesalahan yang baik menjadikan Hadoop sangat efisien dan andal.



Untuk memberikan ketersediaan data yang lebih baik dan toleransi kesalahan, replikasi data dilakukan. Pengguna tidak perlu khawatir tentang mempartisi data, dan penetapan tugas ke *node*, komunikasi antar *node*. Karena Hadoop menangani semuanya, pengguna dapat berkonsentrasi pada data dan operasi pada data tersebut.

D. MapReduce

MapReduce adalah teknik pemrosesan dan model program untuk komputasi terdistribus. Algoritma MapReduce berisi dua tugas penting, yaitu *Map* dan *Reduce*. *Map* mengambil

satu set data dan mengubahnya menjadi set data lain, di mana masing-masing elemen dipecah menjadi tupel (pasangan kunci/nilai). Kedua, *reduce* tugas, yang mengambil luaran dari *map* sebagai nilai masukkan dan menggabungkan tupel data tersebut menjadi satu set tupel yang lebih kecil. Sesuai urutan nama MapReduce, tugas *reduce* selalu dilakukan setelah pekerjaan *map*.

Keuntungan utama MapReduce adalah kemudahan untuk menskalakan pemrosesan data melalui beberapa *node* komputasi. Di bawah model MapReduce, primitif pemrosesan data disebut *mapper* dan *reduction*. Memecah aplikasi pemrosesan data menjadi *mapper* dan *reduction* terkadang tidak sepele. Namun, begitu kita menulis aplikasi dalam bentuk MapReduce, menskalakan aplikasi untuk menjalankan ratusan, ribuan, atau bahkan puluhan ribu mesin dalam sebuah cluster hanyalah perubahan konfigurasi. Skalabilitas sederhana inilah yang menarik banyak programer untuk menggunakan model MapReduce.

BAB 3

BIG DATA ANALITIK

Volume data yang harus ditangani telah mencapai ke tingkat yang tidak terbayangkan dalam dekade terakhir, dan pada saat yang sama, harga penyimpanan data telah berkurang secara sistematis. Tantangan di era Big Data adalah memahami kumpulan besar data ini. Di sinilah muncul konsep Big Data analitik, untuk menyediakan kerangka kerja untuk mengatur pekerjaan yang dibutuhkan oleh suatu organisasi dan memberikan wawasan yang jelas dari Big Data. Big Data Analitik sebagian besar melibatkan pengumpulan data dari berbagai sumber, menggabungkannya sedemikian rupa sehingga tersedia untuk diproses oleh analis dan akhirnya memberikan informasi yang berguna. Proses mengkonversi sejumlah besar data mentah tidak terstruktur, yang diambil dari berbagai sumber menjadi produk data yang berguna untuk organisasi merupakan inti dari Big Data Analitik.

A. Big Data Analitik

Analitik Big Data adalah proses menemukan pola, tren, dan hubungan dalam kumpulan data masif yang tidak dapat ditemukan dengan teknik dan alat manajemen data tradisional.

Cara terbaik untuk memahami ide di balik analitik Big Data adalah dengan membandingkannya dengan analitik data biasa.

1. Pendekatan tradisional. Analitik biasanya terjadi setelah periode waktu atau peristiwa tertentu. Jika Anda adalah pemilik toko online, Anda dapat melihat data yang terakumulasi selama seminggu dan kemudian menganalisisnya. Misalnya, Anda menghitung pelanggan mana yang menggunakan voucer diskon yang dikirimkan kepada mereka melalui email.
2. Big Data. Analitik biasanya terjadi secara *real-time* saat data dihasilkan dan penemuan disajikan hampir secara instan. Katakanlah, Anda mengoperasikan armada pengiriman barang dan Anda perlu mengetahui lokasi persis masing-masing armada serta penundaan rute secara *real-time*.

Data yang dihasilkan dari berbagai sumber termasuk sensor, file log, dan media sosial, dapat digunakan baik secara mandiri maupun sebagai pelengkap data transaksional yang sudah dimiliki banyak organisasi. Selain itu, bukan hanya pengguna bisnis dan analis yang dapat menggunakan data ini untuk analitik lanjutan, tetapi juga peneliti data yang dapat menerapkan Big Data untuk membangun proyek pembelajaran mesin prediktif.

B. Jenis Big Data Analitik

Berikut adalah empat jenis analitik Big Data:

1. Analisis Deskriptif

   Ini meringkas data masa lalu ke dalam bentuk yang mudah dibaca orang. Ini membantu dalam membuat laporan, seperti pendapatan perusahaan, laba, penjualan, dan

sebagainya. Juga, ini membantu dalam tabulasi metrik media sosial.

2. Analisis Diagnostik
   Ini dilakukan untuk memahami apa yang menyebabkan masalah pada awalnya. Teknik seperti penelusuran, penambangan data, dan pemulihan data adalah contohnya. Organisasi menggunakan analitik diagnostik karena memberikan wawasan mendalam tentang masalah tertentu.
3. Analisis Prediktif
   Jenis analitik ini melihat data historis dan saat ini untuk membuat prediksi masa depan. Analitik prediktif menggunakan penambangan data , AI, dan pembelajaran mesin untuk menganalisis data saat ini dan membuat prediksi tentang masa depan. Ini berfungsi untuk memprediksi tren pelanggan, tren pasar, dan sebagainya.

4. Analisis Preskriptif
   Jenis analitik ini menentukan solusi untuk masalah tertentu. Analitik perspektif bekerja dengan analitik deskriptif dan prediktif. Sebagian besar waktu, itu bergantung pada AI dan pembelajaran mesin.

C. Proses Utama Big Data Analitik

Analitik Big Data mencakup proses pengumpulan, pemrosesan, pemfilteran/pembersihan, dan analisis kumpulan data yang luas sehingga organisasi dapat menggunakannya untuk mengembangkan, menumbuhkan, dan menghasilkan produk yang lebih baik. Mari kita lihat lebih dekat prosedur ini.

1. Integrasi Data

Proses mengidentifikasi sumber dan kemudian mendapatkan Big Data. Perlu dicatat bahwa pengumpulan data biasanya terjadi secara *real-time* atau mendekati *real-time* untuk memastikan pemrosesan segera dilakukan. Teknologi modern memungkinkan pengumpulan data terstruktur (kebanyakan data dalam format tabular) dan tidak terstruktur (semua jenis format data) dari berbagai sumber termasuk situs web, aplikasi seluler, database, file datar, sistem manajemen hubungan pelanggan (CRM), IoT sensor, dan sebagainya.

Data mentah harus menjalani proses ekstraksi, transformasi, dan loading, sehingga aliran data ETL atau ELT dibuat untuk mengirimkan data dari sumber ke repositori terpusat untuk penyimpanan dan pemrosesan lebih lanjut. Dengan pendekatan ETL, transformasi data terjadi sebelum mencapai repositori target seperti datawarehouse, sedangkan ELT memungkinkan untuk mengubah data setelah dimuat ke sistem target.

2. Penyimpanan dan Pemrosesan Data

Berdasarkan kompleksitas data, data dapat dipindahkan ke penyimpanan seperti cloud data warehouse awan tempat bisnis intelijen dapat mengaksesnya menggunakan alat saat dibutuhkan. Ada beberapa solusi berbasis cloud modern yang biasanya menyertakan penyimpanan, komputasi, dan komponen infrastruktur klien. Lapisan penyimpanan memungkinkan data yang berasal dari sumber yang berbeda diatur dalam partisi untuk pengoptimalan dan kompresi lebih lanjut. Lapisan komputasi adalah kumpulan mesin pemrosesan yang digunakan untuk melakukan tugas komputasi apa pun

pada data. Ada juga lapisan klien tempat semua aktivitas manajemen data terjadi.

Saat data tersedia, data harus diubah menjadi bentuk yang paling mudah dicerna untuk mendapatkan hasil yang dapat ditindaklanjuti pada kueri analitik. Untuk tujuan itu, ada opsi pemrosesan data yang berbeda. Pilihan pendekatan yang tepat mungkin bergantung pada tugas komputasi dan analitis perusahaan serta sumber daya yang tersedia.

3. Pembersihan Data

Sebelum dianalisis secara menyeluruh, data baik yang kecil maupun yang besar perlu dibersihkan dengan benar untuk memastikan kualitas terbaik dan memberikan hasil yang akurat. Singkatnya, proses pembersihan data melibatkan pembersihan untuk setiap kesalahan, duplikasi, ketidakkonsistenan, redudansi, format yang salah, dll., dan dengan demikian memastikan kegunaan dan relevansi data untuk analitik. Setiap data yang tidak relevan atau cacat harus dihapus atau diperbaiki. Beberapa alat kualitas data dapat mendeteksi kekurangan dalam kumpulan data dan membersihkannya.

4. Analisis data

Analisis data adalah tahap dimana Big Data berubah menjadi pengetahuan yang dapat diaplikasikan, antara lain, mendorong perkembangan dan daya saing perusahaan. Untuk memahami sejumlah besar data, ada beberapa teknik dan praktik analisis daya. berikut beberapa contoh analisis data:

- *Natural language processing* adalah teknologi yang digunakan untuk membuat komputer memahami dan merespons bahasa manusia, apakah itu teks atau kata-kata yang diucapkan.

- *Text mining* adalah pendekatan analitik lanjutan yang digunakan untuk memahami Big Data yang datang dalam bentuk tekstual seperti email, tweet, dan posting blog.
- *Sensor data analysis* adalah pemeriksaan data yang terus menerus dihasilkan oleh berbagai sensor yang dipasang pada objek fisik. Jika dilakukan tepat waktu dan benar, ini tidak hanya membantu memberikan gambaran lengkap tentang kondisi peralatan, tetapi juga mendeteksi perilaku yang salah dan memprediksi kegagalan.
- *Outlier analysis* atau deteksi anomali adalah teknik yang digunakan untuk mengidentifikasi titik data dan peristiwa yang menyimpang dari data lainnya. Ini banyak diterapkan dalam kegiatan deteksi penipuan.

BAB 4

APACHE SPARK

Apache Spark adalah kerangka kerja (*framework*) yang bersifat *open source* untuk pemrosesan Big Data. Apache Spark dikembangkan dengan tujuan kecepatan dan kemudahan dalam penggunaannya. Apache Spark juga dilengkapi dengan kemampuan analitik yang canggih. Apache Spark awalnya dikembangkan pada tahun 2009 di AMPLab UC Berkeley, dan versi *open-source* Spark dikembangkan pada tahun 2010 sebagai bagian dari proyek Apache.

Apache Spark memiliki beberapa keunggulan dibandingkan dengan teknologi big data dan MapReduce lainnya seperti Hadoop dan Storm. Spark memberi kita kerangka kerja terintegrasi yang komprehensif untuk mengelola persyaratan pemrosesan big data dengan berbagai kumpulan data (*dataset*) yang sifatnya beragam seperti data teks, data grafik, dan jenis data lainnya. Spark juga memungkinkan kita untuk mengolah data berbagai sumber data seperti *batch, real-time*, dan streaming data. Spark memungkinkan aplikasi di *cluster* Hadoop untuk berjalan hingga 100 kali lebih cepat di memori dan 10 kali lebih cepat bahkan saat berjalan di media

penyimpanan (*Hard disk*). Spark memungkinkan kita dapat dengan cepat menulis aplikasi dengan beberapa Bahasa pemrograman seperti Java, Scala, atau Python. Spark juga dilengkapi satu set *built-in* dengan lebih dari 80 operator bahasa tingkat tinggi yang dapat digunakan secara interaktif untuk mengolah data di dalam perintah *Shell*.

Selain operasi MapReduce[1], Spark mendukung kueri SQL[2], data streaming, pembelajaran mesin, dan pemrosesan data grafik. Pengembang dapat menggunakan f*ramework* Spark secara terpisah (berdiri sendiri) atau menggunakannnya dengan komponen atau teknologi pengolahan data yang lain.

A. Konsep Spark

"Fitur utama Spark adalah menyimpan dataset yang berfungsi pada memori cache cluster, untuk memungkinkan komputasi yang lebih cepat."

Spark memanfaatkan paralelisme tugas pada banyak pekerja, seperti MapReduce. Spark bekerja dengan cara yang sama:

- Pada satu mesin tunggal untuk pengujian dan sampel data dengan jumlah kecil, tanpa manajer cluster
- Pada cluster untuk volume data yang besar, dengan cluster manajer.

Spark bukanlah versi modifikasi dari Hadoop , tetapi Hadoop adalah cara untuk mengimplementasikan Spark. Spark dapat dibangun dengan komponen Hadoop dengan cara berikut:

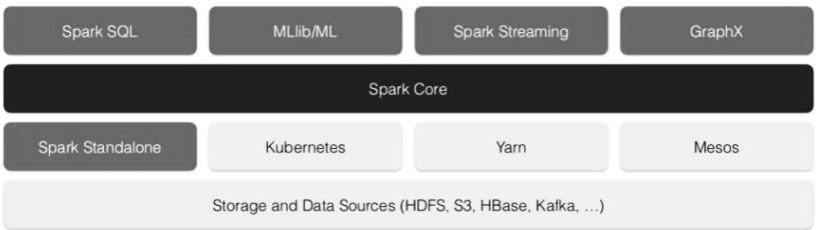
- Spark di atas HDFS

[1] *Pembaca diasumsikan sudah paham konsep Map Reduce*
[2] *SQL disini mengacu pada structure query language*

- Spark di atas HDFS dan Yarn
- Spark di atas HDFS dan di MapReduce

B. Komponen Utama Spark

Diagram berikut menggambarkan komponen utama dari Spark.



- Spark Core, *Engine* umum yang mendasari untuk platform Spark. Fungsionalitas lain Dibangun di atas Spark Core, yang menyediakan komputasi dalam memori dan referensi kumpulan data dalam sistem penyimpanan eksternal.
- Spark SQL: Data-frame SparkSQL memungkinkan penggunaan RDD (*Resilient Distributed Datasets*), memberikan dukungan untuk data terstruktur dan semi-terstruktur.
- Spark Streaming: Spark Streaming memungkinkan *pseudo-streaming*, yaitu melakukan *micro-batch* (misalnya Setiap 50ms) dan melakukan transformasi RDD pada *micro-batch* tersebut.
- MLib: SparkML adalah kerangka kerja pembelajaran mesin terdistribusi di atas Spark karena arsitektur Spark berbasis memori terdistribusi. MLib mencakup seluruh kerangka kerja ML: seperti pra-pemrosesan, validasi silang, algoritma dalam sistem terdistribusi.