



Dr. Wahyudi, S.T., M.T.

# SAINS DATA DAN ANALISIS BIG DATA

Menemukan, Menganalisis,  
Memvisualisasikan dan Menyajikan Data



0858 5343 1992  
eurekamediaaksara@gmail.com  
Jl. Banjaran RT.20 RW.10  
Bojongsari - Purbalingga 53362



**SAINS DATA DAN ANALISIS BIG DATA:  
Menemukan, Menganalisis, Memvisualisasikan  
dan Menyajikan Data**

**Dr. Wahyudi, S.T., M.T.**



**PENERBIT CV. EUREKA MEDIA AKSARA**

**SAINS DATA DAN ANALISIS BIG DATA:  
Menemukan, Menganalisis, Memvisualisasikan  
dan Menyajikan Data**

**Penulis** : Dr. Wahyudi, S.T., M.T.

**Desain Sampul** : Ardyan Arya Hayuwaskita

**Tata Letak** : Uli Mas'uliyah Indarwati

**ISBN** : 978-623-151-730-2

Diterbitkan oleh : **EUREKA MEDIA AKSARA, OKTOBER 2023**  
**ANGGOTA IKAPI JAWA TENGAH**  
**NO. 225/JTE/2021**

**Redaksi:**

Jalan Banjaran, Desa Banjaran RT 20 RW 10 Kecamatan Bojongsari  
Kabupaten Purbalingga Telp. 0858-5343-1992

Surel : eurekamediaaksara@gmail.com

Cetakan Pertama : 2023

**All right reserved**

Hak Cipta dilindungi undang-undang

Dilarang memperbanyak atau memindahkan sebagian atau seluruh isi buku ini dalam bentuk apapun dan dengan cara apapun, termasuk memfotokopi, merekam, atau dengan teknik perekaman lainnya tanpa seizin tertulis dari penerbit.



## KATA PENGANTAR

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadiran Allah SWT berkat rahmat, karunia dan pertolonganNya, penulis dapat menyelesaikan buku berjudul **"Sains Data Dan Analisis Big Data: Menemukan, Menganalisis, Memvisualisasikan dan Menyajikan Data"**. Shalawat serta salam semoga senantiasa tercurah atas Nabi Muhammad SAW, para kerabat, serta pengikutnya hingga hari kiamat nanti.

Buku ini hadir untuk menambah literasi tentang teknologi informasi. Buku ini merupakan seri ketiga dari beberapa buku sains data. Buku ini menjelaskan perspektif lain bagaimana cara Menemukan, Menganalisis, Memvisualisasikan dan Menyajikan Data, dan Siklus Analisis Data.

Penulis menyadari bahwa dalam penulisan buku ini masih banyak terdapat kekurangan, untuk itu penulis mengharapkan kritik dan sarannya guna penyempurnaan buku ini di masa mendatang

Padang, September 2023

Penulis

## DAFTAR ISI

<b>KATA PENGANTAR.....</b>	<b>iii</b>
<b>DAFTAR ISI.....</b>	<b>iv</b>
<b>DAFTAR TABEL .....</b>	<b>vi</b>
<b>DAFTAR GAMBAR.....</b>	<b>vii</b>
<b>BAB 1 PENGANTAR ANALISIS BIG DATA .....</b>	<b>1</b>
A. Gambaran Umum Big Data .....	1
1. Data Struktur .....	7
2. Perspektif Analisis tentang Repositori Data.....	15
B. Keadaan Praktik dalam Analisis .....	20
1. BI Versus Ilmu Pengetahuan Data.....	22
2. Current Analytical Architecture .....	24
3. Faktor Pendorong Big Data .....	29
4. Ekosistem Big Data yang Sedang Berkembang dan Pendekatan Baru untuk Analisis.....	31
C. Peran Kunci untuk Ekosistem Big Data Baru .....	37
D. Contoh Analisis Big Data.....	43
<b>BAB 2 SIKLUS ANALISIS DATA.....</b>	<b>47</b>
A. Gambaran Umum Siklus Hidup Analisis Data .....	48
1. Peran Kunci untuk Proyek Analisis yang Sukses ...	49
2. Latar Belakang dan Gambaran Umum Siklus Hidup Analisis Data .....	52
B. Fase 1: Penemuan .....	57
1. Mempelajari Domain Bisnis .....	57
2. Sumber daya.....	59
3. Membingkai Masalah .....	61
4. Mengidentifikasi Pemangku Kepentingan Utama..	63
5. Mewawancarai Sponsor Analisis .....	64
6. Mengembangkan Hipotesis Awal .....	67
7. Mengidentifikasi Sumber Data Potensial.....	68
C. Fase 2: Persiapan Data .....	71
D. Fase 3: Perencanaan Model.....	86

E. Fase 4: Membangun Model.....	95
F. Fase 5: Mengkomunikasikan Hasil.....	100
G. Fase 6: Mengoperasionalkan.....	104
H. Studi Kasus: Jaringan Inovasi Global dan Analisis (GINA) .....	110
<b>DAFTAR PUSTAKA.....</b>	<b>125</b>

## DAFTAR TABEL

<b>Tabel 1</b> Jenis Repositori Data, dari Perspektif Analisis .....	19
<b>Tabel 2</b> Pendorong Bisnis untuk Analisis Tingkat Lanjut .....	20
<b>Tabel 3</b> Contoh Inventaris Dataset .....	80
<b>Tabel 4</b> Penelitian tentang Perencanaan Model dalam Industri Vertikal .....	89
<b>Tabel 5</b> Rencana Analitik dari Proyek EMC GINA.....	122



## DAFTAR GAMBAR

<b>Gambar 1.1</b>	Apa yang Mendorong Banjir Data .....	4
<b>Gambar 1.2</b>	Contoh-contoh yang Dapat Dipelajari Melalui Genotipe, dari 23andme.com.....	6
<b>Gambar 1.3</b>	Pertumbuhan Big Data Meningkatkan Terhadap Data Tidak Terstruktur .....	8
<b>Gambar 1.4</b>	Contoh Data Terstruktur .....	10
<b>Gambar 1.5</b>	Contoh Data Semi-Terstruktur.....	11
<b>Gambar 1.6</b>	Contoh Hasil Pencarian EMC Data Science .....	13
<b>Gambar 1.7</b>	Contoh Data Tidak Terstruktur: Video Tentang Ekspedisi Antartika.....	14
<b>Gambar 1.8</b>	Membandingkan BI dengan Ilmu Data .....	24
<b>Gambar 1.9</b>	Arsitektur Analitik yang Umum.....	25
<b>Gambar 1.10</b>	Evolusi Data dan Munculnya Sumber-Sumber Big Data .....	31
<b>Gambar 1.11</b>	Ekosistem Big Data yang Sedang Berkembang .....	36
<b>Gambar 1.12</b>	Peran Kunci dari Ekosistem Big Data yang Baru .....	37
<b>Gambar 1.13</b>	Profil Seorang Ilmuwan Data.....	42
<b>Gambar 1.14</b>	Visualisasi Data Jaringan Sosial Pengguna Menggunakan InMaps.....	46
<b>Gambar 2.1</b>	Peran kunci untuk proyek analitik yang sukses .....	50
<b>Gambar 2.2</b>	Gambaran Umum Siklus Hidup Analisis Data .....	54
<b>Gambar 2.3</b>	Fase Penemuan.....	58
<b>Gambar 2.4</b>	Tahap Persiapan Data.....	73
<b>Gambar 2.5</b>	Fase Perencanaan Model .....	88
<b>Gambar 2.6</b>	Fase Pembangunan Model .....	97
<b>Gambar 2.7</b>	Fase Mengkomunikasikan Hasil.....	101
<b>Gambar 2.8</b>	Fase Operasionalisasi Model.....	105
<b>Gambar 2.9</b>	Hasil Utama Dari Proyek Analisis yang Sukses.....	109
<b>Gambar 2.10</b>	Visualisasi Sosial Graf Pengirim Ide dan Finalis.....	117
<b>Gambar 2.11</b>	Visualisasi Sosial Graf Dari Para Pemberi Pengaruh Inovasi .....	117



**SAINS DATA DAN ANALISIS BIG DATA:  
Menemukan, Menganalisis,  
Memvisualisasikan  
dan Menyajikan Data**



# BAB

# 1

# PENGANTAR ANALISIS BIG DATA

Banyak yang telah ditulis tentang Big Data dan kebutuhan analitik tingkat lanjut dalam industri, akademisi, dan pemerintah. Ketersediaan sumber data baru dan munculnya peluang analisis yang lebih kompleks telah menciptakan kebutuhan untuk memikirkan kembali arsitektur data yang ada untuk memungkinkan analisis yang memanfaatkan Big Data. Selain itu, terdapat perdebatan yang signifikan dengan sebaik-baiknya. Bab ini menjelaskan beberapa konsep utama untuk memperjelas apa yang dimaksud dengan Big Data, mengapa analitik tingkat lanjut diperlukan, bagaimana Data Science berbeda dari Business Intelligence (BI), dan peran baru apa yang baru apa yang dibutuhkan untuk ekosistem Big Data yang baru.

## **A. Gambaran Umum Big Data**

Data diciptakan secara konstan, dan dengan kecepatan yang terus meningkat. Ponsel, media sosial, teknologi pencitraan untuk menentukan diagnosis medis. Banyak ciptakan data baru, dan harus disimpan di suatu tempat untuk beberapa tujuan. Perangkat dan sensor

# BAB

# 2

# SIKLUS ANALISIS DATA

Proyek sains data berbeda dari sebagian besar proyek Intelijen Bisnis tradisional dan banyak proyek analisis data karena proyek sains data lebih bersifat eksploratif. Untuk alasan ini, sangat penting untuk mengaturnya dan memastikan para peserta teliti dan ketat dalam pendekatan, agar tidak menghambat eksplorasi.

Banyak masalah yang tampak besar dan menakutkan pada awalnya dapat dipecah menjadi bagian-bagian yang lebih kecil atau fase-fase yang dapat ditindak lanjuti dengan mudah. Proses yang baik memastikan metode yang komprehensif dan komprehensif dan diulang untuk melakukan analisis. Selain itu, membantu memfokuskan waktu dan energi di awal proses untuk mendapatkan pemahaman tentang masalah bisnis yang harus dipecahkan.

Kesalahan umum yang sering terjadi dalam proyek data science adalah terburu-buru dalam pengumpulan dan analisis data, yang perlu meluangkan waktu yang cukup untuk merencanakan dan menentukan jumlah pekerjaan yang dilakukan, memahami kebutuhan, atau bahkan mbingkai masalah bisnis dengan benar. Akibatnya, para

## DAFTAR PUSTAKA

- C. B. B. D. Manyika, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.
- D. R. John Gantz, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC, 2013.
- <http://www.willisresilience.com/emc-datalab> [Online].
- C. Duhigg, *The Power of Habit: Why We Do What We Do in Life and Business*, New York: Random House, 2012.
- K. Hill, "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did," *Forbes*, February 2012. <http://hadoop.apache.org> [Online].
- T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, October 2012.
- J. Manyika, M. Chiu, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.
- "Scientific Method" [Online]. Available: [http://en.wikipedia.org/wiki/Scientific\\_method](http://en.wikipedia.org/wiki/Scientific_method).
- "CRISP-DM" [Online]. Available: [http://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining).
- T. H. Davenport, J. G. Harris, and R. Morison, *Analytics at Work: Smarter Decisions, Better Results*, 2010, Harvard Business Review Press.

D. W. Hubbard, How to Measure Anything: Finding the Value of Intangibles in Business, 2010, Hoboken, NJ: John Wiley & Sons.

J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton, MAD Skills: New Analysis Practices for Big Data, Watertown, MA 2009.

“List of APIs” [Online]. Available:  
<http://www.programmableweb.com/apis>.

B. Shneiderman [Online]. Available:  
<http://www.ifp.illinois.edu/nabhcs/abstracts/shneiderman.html>.

“Hadoop” [Online]. Available: <http://hadoop.apache.org>.

“Alpine Miner” [Online]. Available: <http://alpinenow.com>.

“OpenRefine” [Online]. Available: <http://openrefine.org>.

“Data Wrangler” [Online]. Available:  
<http://vis.stanford.edu/wrangler/>.

“CRAN” [Online]. Available: <http://cran.us.r-project.org>.

“SQL” [Online]. Available:  
<http://en.wikipedia.org/wiki/SQL>.

“SAS/ACCESS” [Online]. Available:  
[http://www.sas.com/en\\_us/software/data-management/access.htm](http://www.sas.com/en_us/software/data-management/access.htm).

“SAS Enterprise Miner” [Online]. Available:  
[http://www.sas.com/en\\_us/software/analytics/enterprise-miner.html](http://www.sas.com/en_us/software/analytics/enterprise-miner.html).

“SPSS Modeler” [Online]. Available: <http://www-03.ibm.com/software/products/en/category/business-analytics>.

“Matlab” [Online]. Available:  
<http://www.mathworks.com/products/matlab/>.

“Statistica” [Online]. Available: <https://www.statsoft.com>.

“Mathematica” [Online]. Available:  
<http://www.wolfram.com/mathematica/>.

“Octave” [Online]. Available:  
<https://www.gnu.org/software/octave/>.

“WEKA” [Online]. Available:  
<http://www.cs.waikato.ac.nz/ml/weka/>.

“MADlib” [Online]. Available: <http://madlib.net>.

K. L. Higbee, *Your Memory – How It Works and How to Improve It*, New York: Marlowe & Company, 1996.

S. Todd, “Data Science and Big Data Curriculum” [Online].  
Available: [http://stevetodd.typepad.com/my\\_weblog/data-science-and-big-data-curriculum/](http://stevetodd.typepad.com/my_weblog/data-science-and-big-data-curriculum/).

T. H Davenport and D. J. Patil, “Data Scientist: The Sexiest Job of the 21st Century,” *Harvard Business Review*, October 2012.