Prosiding Seminar Implementasi Teknologi Informasi dan Komunikasi
Vol. 2, No. 2, Agustus 2023
ISSN: 2809-9834, DOI: 10.31284/p.semtik.2023-2.4656

Perancangan Data Infrastruktur dengan Menerapkan Teknik *Data Wrangling* Studi Kasus: Data *Users* di Narasio Data

Wahyu Widyanto, Victoria Lucky Mahendra, Fa'iz Abiyyu Rizqullah Saputra, Rani Rotul Muhima*

Institut Teknologi Adhi Tama Surabaya

*Penulis korespondensi. E-mail: rani.muhima@itats.ac.id

ABSTRACT

In the digital age, the amount of complex and diverse data continues to increase from various sources. However, differences in data formats, schemas, and quality can make it difficult to merge, transform, and analyze data. One of the major challenges is data quality issues that can result in errors in decision-making. Therefore, it is important to clean the data properly so that data fusion produces accurate results. The data wrangling technique with column mapping and merging method is applied to solve the problem of merging data from two different sources. The research stages include defining the problem and case study provided by the product team of PT Berfikir Revolusioner Indonesia (Narasio Data), collecting excel and csv data files derived from training and event registration activities. The excel files used 38 files while there were 10 csv files. The results showed that the data wrangling technique with the column mapping and merging method can clean and combine columns and contents from training and event registration data into a new data set called the users data set. Therefore, information obtained from various origins can be repurposed and enhance future data analysis.

Keywords

cleaning data,
column mapping,
column merging,
data wrangling

ABSTRAK

Dalam era digital, jumlah data yang kompleks dan beragam terus meningkat dari berbagai sumber. Namun, perbedaan dalam format, skema, dan kualitas data dapat menyulitkan penggabungan, transformasi, dan analisis data. Salah satu tantangan utama adalah masalah kualitas data yang dapat mengakibatkan kesalahan dalam pengambilan keputusan. Oleh karena itu, penting untuk membersihkan data dengan baik agar penggabungan data menghasilkan hasil yang akurat. Teknik *data wrangling* dengan metode *column mapping and merging* diterapkan untuk mengatasi masalah penggabungan data dari dua sumber yang berbeda. Tahapan penelitian meliputi pendefinisian masalah dan studi kasus yang diberikan oleh tim produk PT. Berpikir Revolusioner Indonesia (Narasio Data), pengumpulan data file excel dan csv yang berasal dari kegiatan *training* dan *event registration*. File excel yang diolah sebanyak 38 file sedangkan file csv ada sebanyak 10 file. Hasil penelitian menunjukkan bahwa teknik *data wrangling* dengan metode *column mapping and merging* dapat membersihkan dan menggabungkan kolom serta isi dari data *training* dan *event registration* menjadi satu set data baru yang disebut sebagai set data *users*. Dengan demikian, informasi yang diperoleh dari berbagai sumber dapat digunakan kembali dan meningkatkan analisis data di masa mendatang.

PENDAHULUAN

Perkembangan data dalam era digital telah mengalami peningkatan yang mencolok, memberikan pengaruh yang signifikan pada berbagai aspek kehidupan manusia. Era digital telah menciptakan sekumpulan data yang besar dan kompleks, termasuk informasi dari berbagai sumber seperti media sosial, transaksi bisnis, dan sumber data lainnya [1]. Namun, data yang kompleks dapat menimbulkan dampak negatif terhadap efisiensi dan efektivitas organisasi. Salah satu tantangan dalam penggunaan data secara efektif adalah masalah kualitas data, di mana data yang tidak terjaga dengan baik dapat menyebabkan kesalahan dalam pengambilan keputusan [2].

Perbedaan format, skema, dan kualitas data dapat menyulitkan upaya penggabungan, transformasi, dan analisis data karena terdapat beragam jenis data yang perlu dikelola, termasuk data yang terstruktur, tidak terstruktur, data *real-time*, dan lain sebagainya sehingga menciptakan kesulitan dalam mengidentifikasi pola maupun maksud penting dari sebuah data [3]. Ketika terdapat data dari dua atau lebih sumber yang berbeda digunakan dalam proses *data fusion* atau proses penggabungan data, maka data dengan kualitas buruk dapat menyebabkan masalah setelahnya. Oleh karena itu, pembersihan data yang memadai dapat berdampak signifikan pada hasil penggabungan data nantinya [4] [5].

Data wrangling adalah salah satu metode dalam pembersihan data yang tidak struktur maupun tidak konsisten dengan cara mendeteksi anomali kemudian memperbaiki atau bahkan menghapus data sesuai kebutuhan [6]. Pemrosesan dengan metode tradisional atau disebut dengan *data munging* sudah tidak cocok diterapkan kedalam kumpulan data yang kuantitasnya bersifat besar, karena memakan banyak waktu dan rentan terjadinya kesalahan [7] [8].

Penelitian tentang manfaat dari *data wrangling* pernah ditulis oleh *Florian Endel*, yang dimana pada penelitian tersebut dijelaskan bahwa proses *data wrangling* dapat membantu dalam pengembangan suatu data pada proyek yang sebelumnya dilakukan dengan cara tradisional dan merupakan pekerjaan yang membosankan [8], tidak hanya itu penerapan *data wrangling* memungkinkan pemrosesan permintaan bisnis dapat berjalan lebih cepat dengan solusi yang tepat untuk analis [9]. Selain itu terdapat juga penelitian yang menerapkan data wrangling menggunakan *tool pandas* pada Bahasa pemrograman python, pada penelitian tersebut pandas digunakan dalam mencari sebuah anomali data dan memperbaiki data-data yang tidak konsisten [10].

Pada penelitian ini, teknik *data wrangling* digunakan untuk menghindari kesalahan dalam proses membersihkan data, metode yang dipakai pada penelitian ini adalah *column mapping and merging* dan standarisasi data untuk menggabungkan dua data dari sumber berbeda sehingga menciptakan sebuah satu infrastruktur set data baru. Data yang diolah dalam penelitian ini berasal dari PT. Berpikir Revolusioner Indonesia (Narasio Data) yang berbentuk file csv dan excel.

TINJAUAN PUSTAKA

Data Infrastruktur

Data infrastruktur merupakan dasar atau struktur yang dibangun untuk mengumpulkan, menyimpan, mengelola, dan memproses data dalam suatu organisasi atau sistem. Ini melibatkan penggunaan teknologi, perangkat keras, perangkat lunak, jaringan, dan prosedur yang digunakan untuk memanajemen semua tahapan data, mulai dari pengumpulan hingga analisis dan pelaporan [11] [12].

Manajemen data melibatkan serangkaian langkah dalam mengatur, mengelola, dan menjaga data secara efektif dan efisien. Ini meliputi berbagai strategi untuk mengumpulkan, menyimpan, melakukan pemrosesan data, serta menjaga kebersihan dan keamanan data agar dapat diakses dan dimanfaatkan dengan optimal [13]. Pengolahan data melibatkan tahapan pengubahsuaian dan analisis data dengan tujuan untuk menghasilkan pengetahuan yang bernilai. Ini mencakup berbagai teknik seperti pemrosesan secara paralel, analisis statistik, pembelajaran mesin, dan eksplorasi data [14].

Data Wrangling

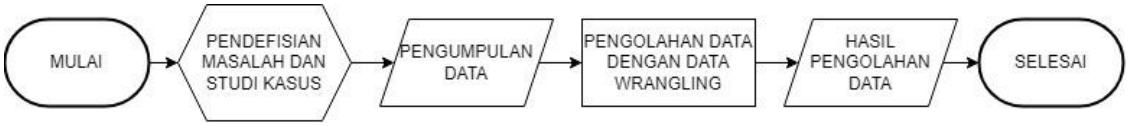
Teknik *data wrangling* disebut juga *data* cleaning merupakan sekumpulan proses data yang kompleks dan berulang dengan tujuan analisis dan visualisasi data. Proses ini melibatkan langkah-langkah kompleks dalam membersihkan kualitas data. Jika terjadi kesalahan pada data, proses *data wrangling* dapat berulang secara iteratif. Tujuannya adalah mencapai hasil yang diinginkan, seperti catatan yang akurat, bebas bug, dan dapat diproses atau diimpor tanpa kesalahan [15].

Masalah kualitas data yang biasanya dipecahkan dalam database maupun sistem informasi menggunakan *data wrangling* antara lain [15]:

- Atribut yang tidak dipertahankan
- Penyalahgunaan atribut untuk mendapatkan informasi tambahan
- Kesalahan data yang disebabkan oleh input yang salah, seperti kesalahan membaca, dan sebagainya
- Kesalahan pengetikan
- Ketidakakuratan data
- Data yang hilang (*missing value*)
- Data yang berlebihan dan tidak konsisten
- Format data yang salah.
- Duplikasi data.
- Informasi yang sudah usang.

METODE

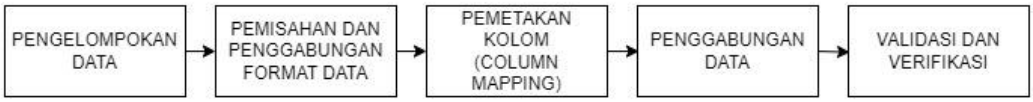
Teknik yang diterapkan dalam penelitian ini adalah penerapan *data wrangling* dengan metode *column mapping and merging* dan standarisasi data, untuk mempermudah peneliti melakukan proses *data wrangling*, maka disini peneliti menggunakan *python jupyter notebook* dan *tools data processing* seperti *numpy*, *pandas* dan *pandasql*. Adapun skema dari penelitian ini seperti yang ditunjukkan pada Gambar 1.



Gambar 1. Skema Penelitian

Tahap pertama yang dilakukan adalah pendefinisian masalah dan studi kasus. Permasalahan dan studi kasus ini langsung diberikan secara langsung oleh tim produk dari PT. Berpikir Revolusioner Indonesia (Narasio Data), yang dimana permasalahan yang sedang dialami perusahaan tersebut adalah bagaimana membuat sebuah infrastruktur data *Users* melalui penggabungan beberapa set data dari dua jenis sumber kegiatan yang berbeda yaitu kegiatan *Training* dan *Event Registration*.

Tahap selanjutnya adalah pengumpulan data, pada tahap ini, peneliti melakukan pengumpulan data berupa file excel dan csv yang masing-masing didapatkan dari kegiatan *training* atau *event registration* yang diselenggarakan oleh perusahaan. File excel yang didapatkan dalam pengumpulan data ini ada sebanyak 38 file sedangkan file csv ada sebanyak 10 file. Kemudian pada tahap *data wrangling* dibagi menjadi beberapa tahap lagi seperti pada Gambar 2.



Gambar 2. Tahapan *data wrangling*

Setelah tahapan terakhir dari *data wrangling* atau tahap validasi dan verifikasi yang dimana pada tahapan tersebut adalah presentasi akhir ke tim produk dari perusahaan telah selesai dilakukan maka tahap terakhir adalah rekap hasil pengolahan data yang sudah menjadi satu infrastruktur data *users* yang seutuhnya dan bebas dari inkonsistensi data.

HASIL DAN PEMBAHASAN

Hasil pertama dalam melakukan tahapan pengolahan data dengan *data wrangling* adalah mengelompokkan kategori data berdasarkan struktur dari nama file data yang ada, kategori yang dikelompokkan disini berupa *Attendance, Feedback,* Peserta, dan Registrasi serta terdapat juga pengelompokkan berdasarkan jenis data nya, 1 adalah data *Training* dan 2 adalah data *Event Registration.* Table 1 adalah potongan hasil dari pengelompokkan data.

Tabel 1. Pengelompokkan data

Id	nama file	data_id	kategori
1	List of participants Advanced Machine Learning Feb 2021	1	Peserta
2	Attendance Basic Analytics Batch 1-not	1	Attendance
3	List of participants Basic Analytics Feb 2021-true	1	Peserta
45	Mar 2022 - Registration - Python Hacks for Daily Use	2	Registrasi
46	July 2022 - Feedback Form - Why Does Data Analytics Matter for Accountants	2	Feedback
47	July 2022 - Registration Form - Why Does Data Analytics Matter for Accountants	2	Registrasi

Selanjutnya adalah tahap pemisahan dan penggabungan format data, pada tahap ini untuk setiap kolom pada set data yang sudah dikelompokkan berdasarkan jenis maupun kategori datanya dilakukan penyesuaian kolom-kolom sehingga set data satu dengan set yang lain dapat memiliki struktur data yang sama. Kolom-kolom yang diambil dari kumpulan set data ini berupa data diri, data asal institusi pendidikan, dan data pekerjaan. Gambar 3 menampilkan hasil dari pemisahan dan penggabungan format data yang telah dilakukan.

#	Column	Non-Null Count	Dtype
0	Student ID	18 non-null	float64
1	Nama Lengkap	269 non-null	object
2	Alamat Email	269 non-null	object
3	No Telepon	18 non-null	float64
4	STUDENT ID	30 non-null	float64
5	NAMA	30 non-null	object
6	E-MAIL	30 non-null	object
7	NO.TELP	30 non-null	float64
8	Nomor Telepon	251 non-null	object
9	Submitted At	35 non-null	object
10	Nama lengkap	100 non-null	object
11	Alamat email	100 non-null	object
12	Nomor telepon	59 non-null	float64
13	Nomor telepon (WA)	41 non-null	float64
14	Asal Kota	31 non-null	object
15	Jurusan	19 non-null	object
16	Jabatan	22 non-null	object
17	Asal Sekolah	16 non-null	object
18	Nama Perusahaan	11 non-null	object

(a)

#	Column	Non-Null Count	Dtype
0	Nama lengkap	5398 non-null	object
1	E-mail	444 non-null	object
2	Nomor Telepon/HP	543 non-null	float64
3	Nama Universitas/Sekolah	207 non-null	object
4	Timestamp	543 non-null	object
5	Institusi Asal	252 non-null	object
6	Nama Universitas	84 non-null	object
7	Jabatan	1065 non-null	object
8	Alamat email	4854 non-null	object
9	Nomor telepon	4855 non-null	float64
10	Asal Sekolah/Universitas	2468 non-null	object
11	Jurusan/Program Studi	523 non-null	object
12	Nama Perusahaan	1611 non-null	object
13	Jabatan dalam perusahaan	730 non-null	object
14	Submitted At	7007 non-null	object
15	Asal kota	1657 non-null	object
16	Institusi asal	1278 non-null	object
17	Nama Lengkap	2152 non-null	object
18	Alamat Email	2152 non-null	object
19	Nomor Telepon	1901 non-null	object
20	Asal Instansi	1120 non-null	object
21	No Telepon (WA)	251 non-null	float64
22	Asal Sekolah	190 non-null	object
23	Jurusan	190 non-null	object
24	Nama Sekolah/Universitas	7 non-null	object
25	Full Name	23 non-null	object
26	Telp No	23 non-null	float64
27	E-mail	99 non-null	object

(b)

Gambar 3. Hasil dari pemisahan dan penggabungan format data a) Pada data *Training*, b) Pada data *Event Registration*

Setelah set data dari jenis data *Training* maupun data *Event Registration* telah diformat kolomnya sesuai dengan kebutuhan, maka selanjutnya adalah melakukan penyesuaian pada kolom-kolom yang memiliki makna yang sama pada kedua set data dengan menerapkan metode *Column Mapping* sehingga kedua set data tersebut memiliki struktur yang sama satu sama lain, misalnya seperti pada kolom Nama Lengkap, Nama lengkap, Full Name, NAMA dan sebagainya menjadi satu kolom yaitu kolom *Name*. Hasil dari *Column Mapping* dari set data *training* dan set data *event registration* disajikan pada Gambar 4.

#	Column	Non-Null Count	Dtype
0	User_id	48 non-null	object
1	Name	399 non-null	object
2	Email	399 non-null	object
3	Phone_number	399 non-null	object
4	Regency	31 non-null	object
5	Province	0 non-null	float64
6	Born_year	0 non-null	float64
7	Gender	0 non-null	float64
8	Background	19 non-null	object
9	Actor_role	22 non-null	object
10	Institution_name	26 non-null	object
11	Is_client	0 non-null	float64
12	Created_at	35 non-null	object
13	Updated_at	0 non-null	float64

(a)

#	Column	Non-Null Count	Dtype
0	User_id	0 non-null	float64
1	Name	7573 non-null	object
2	Email	7549 non-null	object
3	Phone_number	7573 non-null	object
4	Province	0 non-null	float64
5	Born_year	0 non-null	float64
6	Background	713 non-null	object
7	Actor_role	1795 non-null	object
8	Institution_name	7573 non-null	object
9	Is_client	0 non-null	float64
10	Created_at	7007 non-null	object
11	Updated_at	0 non-null	float64

(b)

Gambar 4. Hasil *Column Mapping* a) Pada data *Training*, b) Pada data *Event Registration*

Sekarang kedua set data telah memiliki struktur kolom yang mirip satu sama lain, maka tahap selanjutnya adalah melakukan penggabungan set data *Training* dengan set data *Event Registration* menjadi set data yang bernama *Users*. Dan tahapan terakhir adalah melakukan Validasi dan Verifikasi kepada mitra pemilik data atas kesesuaian hasil pengolahan data yang telah dilakukan. Gambar 5 adalah hasil dari pengolahan data menggunakan Teknik wrangling tepatnya menggunakan metode *Column Mapping* dan *Column Merging*. Pada Gambar 5 data yang disajikan hanya 5 data awal dan 5 data terakhir selain itu data tersebut merupakan data *dummy* atau data palsu demi menjaga kerahasiaan data perusahaan. Nilai *NaN* pada data yang disajikan tetap dibiarkan (tidak ditangani) karena permintaan dari mitra.

User_id	Name	Email	Phone_number	Regency	Province	Born_year	Gender	Background	Actor_role	Institution_name	Is_client	Created_at	Updated_at
1	Kajen Narpati	xmaryati@sihom	+62223313138	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Gada Narpati	bharyanto@gma	+6274705065657	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Dimaz Lazuardi	dian.padmasari@	+6291753731869	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Eva Utami	puti.wahyuni@g	+6235620851483	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	Tania Wulandari	aslijan34@yaho	+629635904569	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	Hesti Farida	lanjar.pudjiastuti	+6251632972961	NaN	NaN	NaN	NaN	NaN	Product Web	Perum Wasita Nurd	NaN	NaN	NaN
NaN	Olivia Wulandari	taufik.situmoran	+62256696160	NaN	NaN	NaN	NaN	NaN	Central Ident	UD Sihotang Anggr	NaN	NaN	NaN
NaN	Eja Latupono	marsudi.nurainia	+628083896713	NaN	NaN	NaN	NaN	NaN	Lead Ideatio	Perum Natsir (Perse	NaN	NaN	NaN
NaN	Unggul Winarno	ffirmansyah@jar	+627103681557	NaN	NaN	NaN	NaN	NaN	Relational Da	Perum Habibi	NaN	NaN	NaN
NaN	Caket Nashirudc	tomi61@gmail.c	+6254488356333	NaN	NaN	NaN	NaN	NaN	Human Team	Perum Dongoran Su	NaN	NaN	NaN

Gambar 5. Hasil pengolahan data

KESIMPULAN

Data wrangling atau pembersihan data merupakan proses yang harus diterapkan dalam mendeteksi, memperbaiki, hingga menghapus sebuah *record* data baik dalam bentuk dataset, database, maupun file excel dan csv yang memiliki anomali data didalamnya. Teknik yang dipakai adalah Teknik *data wrangling* dengan fokus pada metode *Column Mapping* dan *Column Merging*. Dimana kedua metode ini digunakan untuk membersihkan dan menggabungkan kolom dan isi dari set data *training* dan *event registration* menjadi set data baru yang Bernama *users*. Pada tahapan melakukan *column mapping*, data *training* yang sebelumnya memiliki 18 kolom sekarang tersisa menjadi 13 kolom karena terdapat proses penggabungan kolom yang memiliki makna sama, sedangkan pada data *event registration* yang sebelumnya memiliki 27 kolom sekarang hanya tersisa 11 kolom saja. Dari 13 dan 11 kolom pada set data *training* dan *event registration* dilakukan penggabungan (*Column Merging*) yang dimana banyak kolom yang dihasilkan adalah mengikuti set data dengan kolom terbanyak yaitu set data *training* yang memiliki 13 kolom.

Hasil proses *wrangling* didapatkan bahwa dua data atau lebih yang berbeda sumber dapat digabungkan menjadi satu dengan syarat data-data tersebut masing memiliki makna yang mirip satu sama lain, dengan begini Informasi dari kumpulan set data yang ada dapat dengan mudah dilakukan analisis kedepannya.

DAFTAR PUSTAKA

[1] Binus University, “Sejarah dan Evolusi Big Data – Himpunan Mahasiswa Sistem Informasi,” *Binus University*, May 29, 2023. https://student-activity.binus.ac.id/himsisfo/2023/05/ sejarah-dan-evolusi-big-data/ (accessed Jun. 12, 2023).

[2] N. Putu, A. Widiari, M. Agus, D. Suarjaya, and D. Putra Githa, “Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali,” *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, pp. 137–145, Jul. 2020, doi: 10.24843/JIM.2020.V08.I02.P07.

[3] W. Swapnil and Y. Anil, “Big Data: Characteristics, Challenges and Data Mining,” *Int J Comput Appl*, pp. 975–8887.

[4] H. Mueller and J. Freytag, “Problems , Methods , and Challenges in Comprehensive Data Cleansing,” 2005.

[5] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis for feature level fusion with application to multimodal biometrics," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 1866–1870, May 2016, doi: 10.1109/ICASSP.2016.7472000.

[6] I. Setiawan, A. Mutia Dawis, and P. Studi Sistem dan Teknologi Informasi Fakultas Sains Dan Teknologi, "DATA SCIENCE: PENDEKATAN DAN LANGKAH PRAKTIS DENGAN EXCEL," *Journal of Innovation And Future Technology (IFTECH)*, vol. 5, no. 1, pp. 11–22, Feb. 2023, doi: 10.47080/IFTECH.V5I1.2457.

[7] F. Ridzuan and W. M. N. Wan Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Comput Sci*, vol. 161, pp. 731–738, Jan. 2019, doi: 10.1016/J.PROCS.2019.11.177.

[8] F. Endel and H. Piringer, "Data Wrangling: Making data useful again," *IFAC-PapersOnLine*, vol. 48, no. 1, pp. 111–112, Jan. 2015, doi: 10.1016/J.IFACOL.2015.05.197.

[9] M. M. Patil and B. N. Hiremath, "A Systematic Study of Data Wrangling," *International Journal of Information Technology and Computer Science*, vol. 10, no. 1, pp. 32–39, Jan. 2018, doi: 10.5815/IJITCS.2018.01.04.

[10] S. Ghosh, K. Neha, and Y. Praveen Kumar, "Data wrangling using python," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 11, pp. 3491–3495, Sep. 2019, doi: 10.35940/IJRTE.B1427.0982S1119.

[11] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures Their Consequences*. SAGE Publications Ltd, 2014. doi: 10.4135/9781473909472.

[12] J. Gray, C. Gerlitz, and L. Bounegru, "Data infrastructure literacy," *Original Research Article*, pp. 1–13, 2018, doi: 10.1177/2053951718786316.

[13] T. C. Redman, "Data driven : profiting from your most important business asset," p. 257, Accessed: Jun. 13, 2023. [Online]. Available: https://www.perlego.com/book/836939/data-driven-profiting-from-your-most-important-business-asset-pdf

[14] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. Elsevier, 2011. doi: 10.1016/C2009-0-61819-5.

[15] O. Azeroual, "Data Wrangling in Database Systems: Purging of Dirty Data," *Data 2020, Vol. 5, Page 50*, vol. 5, no. 2, p. 50, Jun. 2020, doi: 10.3390/DATA5020050.