ISSN (Online) 2747-0563
Volume 1 Tahun 2020

Exploratory Data Analysis pada Kasus COVID-19 di Indonesia Menggunakan HiveQL dan Hadoop Environment

Tresna Maulana Fahrudin[1*],Prismahardi Aji Riyantoko[2],Kartika Maulida Hindrayani[3],I Gede Susrama Mas Diyasa[4]

[1234] Program Studi Sains Data, Fakultas Ilmu Komputer
Universitas Pembangunan Nasional "Veteran" Jawa Timur
[2]prismahardi.aji.ds@upnjatim.ac.id, [3]kartika.maulida.ds@upnjatim.ac.id,
[4]igsusrama.if@upnjatim.ac.id

**Corresponding author email: [1]tresna.maulana.ds@upnjatim.ac.id*

***Abstrak*— Kasus COVID-19 menjadi pandemi dan hampir menghentikan aktifitas dan rutinitas seluruh warga negara di dunia, termasuk Indonesia. Pandemi ini telah berlangsung di Indonesia di mulai awal bulan Maret hingga saat ini. Pemerintah terus berupaya untuk mensosialisaiskan dan menginformasikan jumlah kasus terkonfirmasi positif, kasus pasien sembuh, pasien meninggal dan pasien dalam perawatan tiap harinya. Namun, dengan seiring bertambahnya kasus tiap hari, perlu upaya para akademisi dan praktisi untuk turut membantu pemerintah untuk menemukan solusi bersama. Exploratory Data Analysis, salah satu strategi yang tepat untuk melakukan analisis data dengan didukung teknologi Big Data, semakin mempermudah untuk melakukan eksplorasi secara mendalam. Berguna untuk menemukan pola-pola tersembunyi, mengungkap informasi yang belum diketahui sebelumnya, dan memberikan dampak untuk tindak lanjut ke depannya. Untuk merealisasikan hal tersebut, pertama, mengumpulkan data kasus COVID-19 dari Badan Nasional Penanggulangan Bencana, Republik Indonesia, lalu data di-load ke dalam Hadoop Environment dan diakses menggunakan HiveQL. Hasil eksperimen menunjukkan bahwa analisis statistik deskriptif menemukan total jumlah kasus COVID-19 tiap kategori paling sedikit terjadi pada bulan Maret sedangkan paling besar terjadi di bulan September 2020. Rata-rata harian kasus COVID-19 tiap bulannya selalu bertambah, dari 50 kasus per hari di bulan Maret hingga melonjak menjadi 3.740 per hari di bulan September. Analisis PDF-CDF menunjukkan laju pertumbuhan kasus terkonfirmasi positif COVID-19 semakin meningkat, belum ditemukan tanda-tanda melandai, dan analisis korelasi menunjukkan terdapat pengaruh kuat antara pertambahan jumlah kasus terkonfirmasi positif dengan jumlah kasus pasien sembuh dan kasus pasien meninggal sebesar 0.94 dan 0.9 masing-masing.**

***Kata Kunci*— COVID-19 Indonesia, Exploratory Data Analysis, HiveQL, Hadoop Environment, Big Data**

I. PENDAHULUAN

Wolrd Health Organization (WHO) menetapkan *Novel Coronavirus Disease* 2019 atau yang lebih dikenal dengan COVID-19 sebagai pandemi global yang menyerang seluruh dunia pada tanggal 11 Maret 2020 [1] . Skala dampak dari pandemi ini belum pernah terjadi sebelumnya, dan beberapa penelitian menyebutkan bahwa membutuhkan beberapa dekade untuk memulihkan kondisi ini baik dalam bidang sosial dan ekonomi [2]. Hal ini juga akan mengganggu *Sustainable Development Agenda* (SDGs) 2030 dimana program ini telah disepakati oleh seluruh pemimpin dunia [3], termasuk Indonesia. Namun, kekurangsiapsiagaan untuk menangani pandemi ini berdampak pada penyebaran yang progresif dan cepat, sehingga membuat banyak pemerintah di seluruh dunia tidak siap untuk mengatasi hal ini.

Indonesia negara terpadat keempat di dunia, sehingga diperkirakan akan mengalami pandemi yang lebih lama dibandingkan negara yang berpenduduk sedikit lainnya [4]. Ketika virus corona baru SARS-CoV2 menginfeksi penduduk Cina pada bulan Desember 2019 hingga Februari 2020, Indonesia melaporkan tidak ada kasus infeksi sama sekali. Namun pada tanggal 2 Maret 2020, Presiden Republik Indonesia, Joko Widodo melaporkan bahwa terdapat 2 kasus infeksi COVID-19 pertama terkonfirmasi positif di Indonesia. Pada bulan berikutnya, hingga 2 April 2020 terdapat 1.790 kasus terkonfirmasi positif, 113 kasus baru, 170 pasien meninggal, dan 112 pasien sembuh. Hingga bulan September 2020 jumlah kasus terkonfirmasi positif di Indonesia sebesar 287.008, 4.284 kasus baru, 10.740 pasien meninggal, dan 214.947 pasien sembuh [5].

Semakin cepatnya laju pertumbuhan data menimbulkan banyak tantangan baru [6], berdampak pada terbentuknya data yang semakin bervariasi, pola data yang semakin rumit dan padat dari waktu ke waktu. Data kasus COVID-19 di Indonesia dan bahkan berbagai kasus di negara lain yang bertambah tiap hari menjadi *new oil* bagi sejumlah *data analyst* untuk melakukan analisis dan eksplorasi data. Hal ini dapat memberikan manfaat dengan ditemukannya informasi baru, fenomena baru, dan suatu *hidden knowledge* yang belum terpikirkan sebelumnya [7]. *Exploratory Data Analysis* atau yang dikenal dengan EDA merupakan salah satu strategi untuk melakukan analisis data, erat kaitannya dengan era *Big Data* saat ini.

Untuk melakukan analisis dan eksplorasi data COVID-19 di Indonesia, maka dibutuhkan beberapa tahapan mulai tahap persiapan hingga implementasi. Tahap pertama, *data preparation,* data kasus COVID-19 dikumpulkan dari sumber terpercaya yang setidaknya mewakili data kasus pasien terkonfirmasi positif, sembuh, meninggal dan yang sedang dirawat. Tahap kedua, yakni menyiapkan Hadoop Environment sebagai pendukung Big Data. Hadoop sangat tepat sebagai basis *framework* untuk melakukan analisis data dalam jumlah besar, berbeda halnya dengan basis data relasional. Tahap ketiga, HiveQL atau Hive Query Language, sebagai *tools* untuk melakukan query data dalam jumlah besar

ISSN (Online) 2747-0563
Volume 1 Tahun 2020

yang tersambung ke Hadoop. Tahap keempat, yakni visualisasi, terdiri dari *descriptive statistics, histogram, cumulative distribution function (CDF)-probability density function (PDF),* dan *correlation.*

II. PENELITIAN TERKAIT

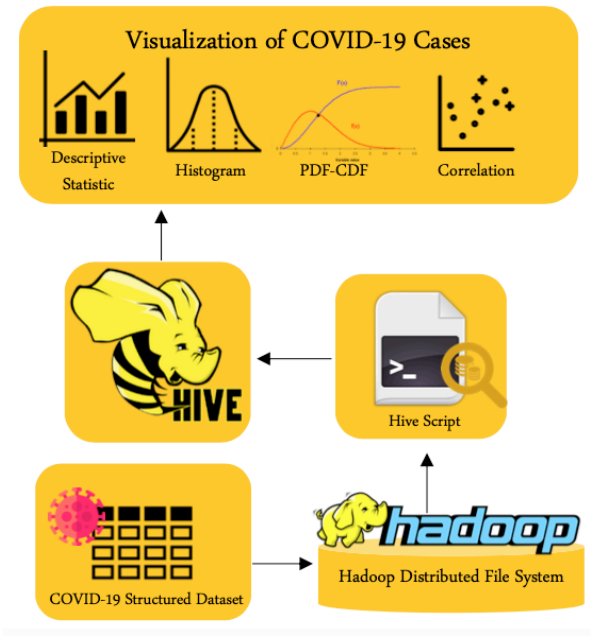
Beberapa penelitian yang berhubungan dengan domain penelitian yang diangkat antara lain penelitian dari Paolo, dkk [8] mengangkat topik tentang COVID-19 di Italia menggunakan metode Extreme Data Mining. Pada penelitiannya mengusulkan algoritma Topological Weighted Centroid (TWC) untuk menyelesaikan permasalahan epidemi COVID-19 di Italia dengan data yang sangat terbatas, namun memperoleh informasi dan *knowledge* baru yang relevan. Menurutnya, Italia memiliki peran sentral dalam epidemi ini karena tingginya jumlah penduduk yang terinfeksi COVID-19. Melalui algoritma kecerdasan buatan yang diusulkan, mereka mencoba untuk menganalisis evolusi fenomena dan memprediksi langkah berikutnya di masa mendatang menggunakan kumpulan data yang berisi koordinat geospasial (bujur dan lintang) dari kasus pertama yang tercatat. Setelah diketahui koordinat dimana terdapat kasus penularan, dilanjutkan dengan beberapa analisis antara lain menentukan titik wabah (*outbreak point*) dan ”peta panas” (*heat map*) (TWC alpha), distribusi probabilitas penularan (TWC beta), kemungkinan adanya penyebaran penyakit dalam waktu dekat dan yang akan datang (TWC gamma dan TWC theta). Output dari penelitiannya adalah memodelkan bagaimana situasi yang mungkin terjadi menjelang akhir epidemi dalam kaitannya dengan tingkat penularan di suatu daerah ke dalam *heat map.*

Petar, dkk [9] mengangkat topik tentang implementasi Data Mining dan analisis rekaman data literatur penelitian ilmiah yang berkaitan dengan mortalitas, imunitas, dan pengembangan vaksin COVID-19 pada gelombang pertama pandemi. Dalam penelitiannya, mereka menyelidiki respon penelitian ilmiah dari awal terjadinya pandemi, dan meninjau temuan utama tentang bagaimana *early warning system* yang dikembangkan pada epidemi sebelumnya. Sumber dokumen literatur ilmiah didapatkan dari Web of Science Core Collection lalu digali melalui serangkaian analisis seperti peta struktur konseptual, analisis korespondensi ganda, dan mengidentifikasi beberapa hubungan antara kata kunci, sinonim dan konsep, terkait dengan mortalitas, imunitas, dan pengembangan vaksin COVID-19. Hasil penelitiannya menemukan bahwa Universitas di China sangat kuat mendominasi penelitian tentang topik literatur penelitian tentang COVID-19 selama masa pandemi, bahkan berkolaborasi dengan Amerika Serikat. Selain itu, Amerika Serikat sendiri mendominasi volume literatur ilmiah di bidang penemuan vaksin COVID-19. Penelitiannya menyimpulkan bahwa didapatkan pengetahuan yang terintegrasi dan berkorelasi dari 276 dokumen tentang COVID-19 dan mortalitas, 71 dokumen tentang COVID-19 dan imunitas, dan 189 dokumen tentang vaksin COVID-19.

Mathieu [10] mengangkat topik tentang COVID-19 dan set data media melalui pendekatan Textual Data Mining berbasis periode dan lokasi. Menurutnya, kosakata yang digunakan dalam berita tentang penyakit seperti COVID-19 berubah dalam beberapa periode. Aspek ini dibahas berdasarkan set data media yang bersumber dari MEDISYS melalui dua studi. Yang pertama berfokus pada ekstraksi terminologi dan yang kedua pada prediksi periode sesuai konten tekstual menggunakan pendekatan pembelajaran mesin (*machine learning*). Pada eksperimen yang telah dilakukan, waktu pengambilan data dibagi ke dalam 3 bagian yakni bulan Maret 2020, May 2020 dan Juli 2020, lalu dihubungkan terhadap lokasi yang berbeda antara lain UK, Spanyol dan Perancis. Beberapa *terms* yang didapatkan dari penelitiannya menggunakan kata kunci “mask” yakni periode 1 {*‘face mask’, ‘gas mask’, ’protective mask’, ’mask’,* dan *‘masks’*}, periode 2 {*‘coronavirus mask’, ’masks’, ’mask mess’, ’surgical mask’,* dan *‘fase masks’*}, dan periode 3 {*‘masks’, ’mask’, ’mandatory mask’, ’mandatory mask-wearing’, ’mandatory masks’*}. Untuk mengetahui kesesuaian *terms* yang diekstrak terhadap kategori tiap periode, penelitiannya melibatkan algoritma *machine learning* seperti Naïve Bayes, Support Vector Machine dan Random Forest untuk mengetahui kepresisiannya.

III. DESAIN SISTEM

Desain sistem pada penelitian ini ditunjukkan pada Gbr. 1, dimana terdapat aliran data dimulai dari input data kasus COVID-19 berupa data terstruktur, lalu data tersebut disimpan ke dalam sebuah basis data terdistribusi yang bernama HDFS (Hadoop Distributed File System). Selanjutnya data yang sudah tersimpan dapat dilakukan query menggunakan HiveQL (Hive Query Language), lalu query yang dibuat akan dikirimkan oleh Hive ke Hadoop Environment yang telah terkonfigurasi sebelumnya. Terakhir, query data yang diminta akan ditampilkan ke dalam bentuk tabel atau berupa visualisasi seperti diagram statistik deskriptif, histogram, PDF-CDF dan analisis korelasi.



Gbr. 1 Desain sistem yang diusulkan pada penelitian

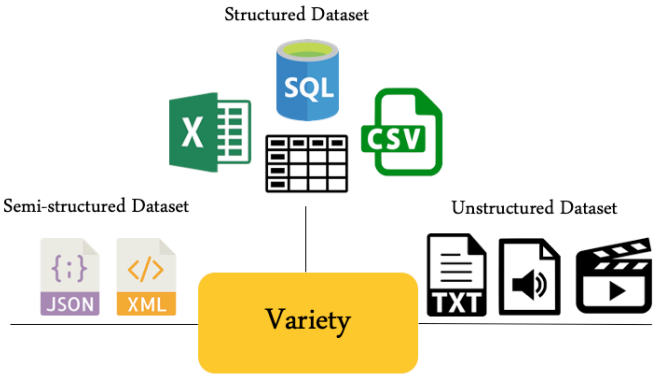
ISSN (Online) 2747-0563
Volume 1 Tahun 2020

Berikut akan dijelaskan lebih detail desain sistem ke dalam beberapa sub bab terkait *data preparation,* arsitektur Hadoop, MapReduce dan HDFS, Hive Query Language dan *Graphical Exploratory Data Analysis*.

A. *Data Preparation*

Data preparation atau persiapan data merupakan tahap awal untuk mendapatkan berbagai sumber data serta teknik untuk menjadikan set data agar mudah terbaca oleh sistem. Bentuk dan jenis set data cukup beragam, namun jika dikaitkan dengan konsep 3V Big Data dari sisi *variety* dapat dibagi menjadi 3 jenis yaitu data terstruktur, semi-terstruktur dan tidak terstuktur [11]. Seperti yang ditunjukkan pada Gbr. 2, jenis set data beserta contohnya antara lain:

- *Structured dataset*: Tabular, CSV (Comma Separated Values), SQL (Structured Query) Language dan Microsoft Excel Open XML Spreadsheet
- *Semi-structured dataset*: JSON (JavaScript Object Notation), XML (eXtensible Markup Language)
- *Unstructured dataset*: TXT (Text File), Audio, dan Image dan Video

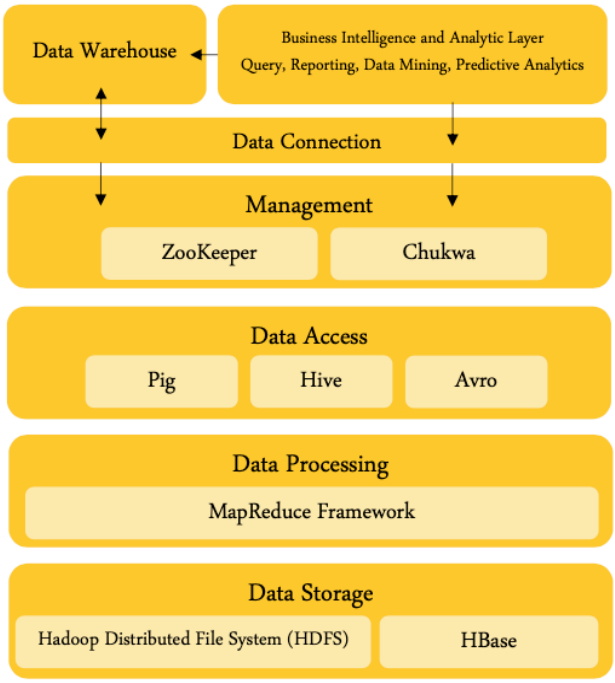


Gbr. 2 Structured, semi-structured, dan unstructured dataset

B. *Hadoop Architecture, MapReduce dan HDFS*

Apache Hadoop adalah teknologi Big Data terkenal yang memiliki banyak komunitas pendukung. Teknologi Big Data ini telah dirancang untuk mempunyai performa kinerja yang tinggi dan kompleksitas yang dihadapi saat memproses dan menganalisis Big Data dibanding menggunakan teknologi tradisional. Salah satu kelebihan utama pada Hadoop adalah kemampuannya untuk memproses kumpulan data besar dengan cepat, komputasi klaster paralel dan sistem file terdistribusi. Faktanya, tidak seperti teknologi tradisional, Hadoop tidak menyalin di memori seluruh data yang ada untuk melakukan komputasi. Sebagai gantinya, Hadoop mengeksekusi *task* dimana lokasi data tersimpan [11].

Hadoop mengurangi beban jaringan dan server dari komunikasi yang cukup besar, misalnya, Hadoop hanya membutuhkan beberapa detik saja untuk melakukan query data dalam terabyte, tidak menghabiskan waktu 20 menit atau lebih seperti halnya teknologi penyimpanan data tradisional [12]. Keuntungan lain dari Hadoop adalah kemampuannya untuk menjalankan program dan memastikan toleransi kesalahan (*fault-tolerance*), biasanya ditemui di lingkungan terdistribusi. Untuk menjamin itu, Hadoop mencegah kehilangan data dengan mereplikasi data di server.



Gbr. 3 Arsitektur Big Data secara umum

Kekuatan platform Hadoop didasarkan pada dua sub komponen utama: Hadoop Distributed File System (HDFS) dan MapReduce *framework.* Selain itu, pengguna dapat menambahkan modul di atas Hadoop sesuai kebutuhan menyesuaikan *requirement* aplikasi masing-masing. Gbr. 3 menunjukkan fleksibilitas arsitektur Big Data untuk menambahkan modul-modul pada tiap *layer.* Komunitas Hadoop telah berkontribusi untuk memperkaya ekosistemnya dengan menyediakan beberapa modul *open source*.

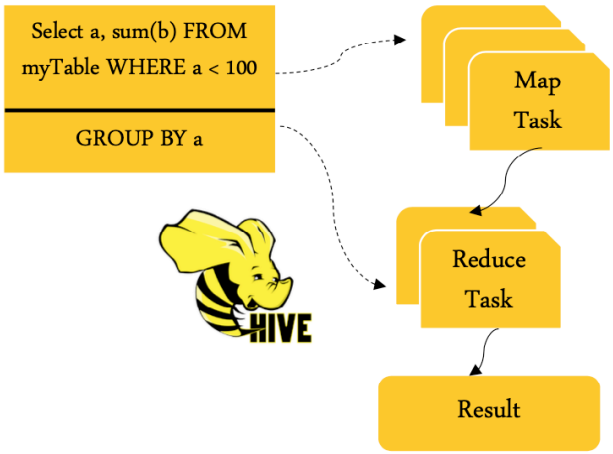
C. *Hive Query Language*

Seperti yang ditunjukkan pada Gbr. 3, Hive berada pada *layer* pemrosesan data. Hive memungkinkan untuk merepresentasikan data ke dalam basis data terstruktur yang lebih mudah dipahami oleh pengguna yakni berbentuk tabel. Tabel tersebut mewakili direktori HDFS dan dibagi menjadi beberapa partisi. Setiap partisi kemudian dibagi menjadi beberapa keranjang (*bucket*). Kelebihan lain, Hive juga menyediakan bahasa yang mirip SQL yang disebut HiveQL [13], memungkinkan pengguna untuk mengakses dan memanipulasi data berbasis Hadoop yang disimpan dalam HDFS atau HBase.

Di sisi lain, Hive tidak cocok untuk transaksi *real-time* [14], ini didasarkan pada operasi latensi rendah. Seperti Hadoop, Hive dirancang untuk pemrosesan skala besar sehingga *job* kecil pun dapat memakan waktu beberapa menit. Secara teknis, HiveQL mengubah query seperti operasi *'join', 'summarization', 'group by'* menjadi MapReduce *jobs* yang diproses sebagai *batch tasks*. Pada Gbr. 4 ditunjukkan bagaimana HiveQL membagi tugas bersama *Map task* dan *Reduce task.* Operasi *'select'* dan *'where'* dibebankan kepada

ISSN (Online) 2747-0563
Volume 1 Tahun 2020

Map task, sedangkan operasi *'group by'* dibebankan kepada *Reduce task.*



Gbr. 4 Script HiveQL bekerjasama dengan MapReduce Task

Karena skema yang digunakan pada Hive adalah *schema-on-read* [15], maka data yang diunggah ke HDFS menggunakan Hive tidak melalui proses validasi yang memenuhi kriteria skema yang diinginkan. Akibatnya, proses *load* akan lebih cepat, namun dari sisi query relatif lebih lambat. Oleh karena itu, Hive akan bekerja lebih baik jika data yang diproses berukuran besar, sebaliknya, jika data yang diproses hanya berukuran kecil, maka Hive membutuhkan waktu untuk mengaktifkan *Map task dan Reduce task*. Hive juga tidak memiliki dukungan SQL penuh seperti melakukan *insert, update* dan *delete* baris data, karena Hive sendiri didesain untuk melakukan *load data* dan bekerja secara *batch processing.*

D. *Graphichal Exploratory Data Analysis*

Jika mengikuti teori analisis data, *Exploratory Data Analysis* (EDA) dapat dicirikan sebagai berikut [16]:

- Penekanan pada pemahaman substantif data yang menjawab pertanyaan luas tentang "apa yang sedang terjadi?"
- Penekanan pada representasi grafik dari data
- Fokus pada pembuatan model eksperimen dan pembuatan hipotesis secara iteratif dari spesifikasi model, analisis residual, dan re-spesifikasi model
- Pengukuran yang tepat, proses pencarian pengetahuan yang berulang dan analisis subset
- Bersikap skeptisisme (kecurigaan dan keingintahuan) serta fleksibilitas terkait metode mana yang akan diterapkan

Bentuk EDA yang disajikan dalam grafik memiliki tujuan yang sama yaitu menyediakan ringkasan statistik dari data mentah (*raw* data) [17] yang hanya berupa representasi angka saja menjadi sebuah informasi yang berguna dengan pemahaman secara singkat. Berikut akan dijelaskan lebih detail terkait bentuk grafik EDA yang dapat disajikan dalam bentuk *descriptive statistics, histogram, PDF-CDF,* dan *correlation.*

1) Descriptive Statistics

Statistik deskriptif merupakan teknik analisis data untuk menggambarkan kondisi keseluruhan data dalam sebuah metrik atau suatu standar pengukuran data [17]. Pilihan metrik untuk statistik deskriptif cukup banyak, antara lain:

- *Count*: jumlah banyaknya data dari suatu set data
- *Mean*: rata-rata data atau ukuran kecenderungan terpusat dari kumpulan suatu set data. Rumus *mean* ditunjukkan pada persamaan (1), dimana $\bar{x}$ adalah *mean,* x_i adalah individu data ke-*i* dari suatu set data, dan *n* adalah jumlah banyaknya data.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

- *Standard Deviation:* ukuran jumlah variasi atau sebaran sejumlah nilai data. Rumus *standard deviation* ditunjukkan pada persamaan (2), dimana σ adalah *standard deviation,* $\bar{x}$ adalah *mean, x* adalah individu data dari suatu set data, dan *n* adalah jumlah banyaknya data

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \quad (2)$$

- *Min*: nilai terkecil dari suatu set data
- *Max*: nilai terbesar dari suatu set data
- *InterQuartile Range* (25%, 50%, dan 75%): ukuran variabilitas dari suatu kumpulan data yang dibagi menjadi beberapa kelompok data yang disebut kuartil. Rumus *InterQuartile* ditunjukkan pada persamaan (3), namun sebelumnya data harus diurutkan terlebih dahulu secara *ascending*. Q1 adalah kuartil pertama atau terletak 25% di bagian kiri dari keseluruhan data. Q2 adalah kuartil kedua atau terletak 50% di bagian tengah dari keseluruhan data. Q3 adalah kuartil ketiga atau terletak 75% di bagian kanan dari keseluruhan data.

$$Q1 = (N+1)\frac{1}{4}, Q2 = (N+1)\frac{2}{4}, Q3 = (N+1)\frac{3}{4} \quad (3)$$

2) Histogram

Histogram merepresentasikan distribusi data numerik yang menghubungkan sebaran nilai suatu variabel dengan frekuensinya [17]. Histogram lebih tepat diterapkan untuk sebaran nilai data bertipe kontinu, dimana nilai-nilai tersebut akan dikelompokkan ke dalam sebuah interval kelas. Jika direpresentasikan ke dalam grafik 2-dimensi, maka axis *x* menunjukkan interval kelas, sedangkan axis *y* menunjukkan frekuensi. Rumus histogram ditunjukkan pada pada persamaan (4), dimana *w* adalah histogram, *R* adalah rentang yang didapatkan dari selisih nilai maksimum dan minimum, dan *n* adalah jumlah interval.

$$w = \frac{R}{n}, where\ R = X_{max} - X_{min} \quad (4)$$

ISSN (Online) 2747-0563
Volume 1 Tahun 2020

3) PDF dan CDF

Probability Density Function (PDF) dan *Cumulative Distribution Function* (CDF) merupakan suatu fungsi yang digunakan untuk mencari nilai probabilitas dari suatu kejadian pada waktu tertentu maupun sejumlah kejadian tertentu [18]. PDF dari suatu variabel acak merupakan plot antara variabel acak dan frekuensinya, sehingga akan menghasilkan distribusi probabilitas dari variabel acak. Rumus PDF ditunjukkan pada persamaan (5), dimana $p(x_i)$ adalah kepadatan probabilitas, $H(x)$ menyatakan jumlah munculnya suatu kejadian x.

$$p(x_i) = \frac{H(x_i)}{\sum_{j=1}^{n} H(x_j)} \quad (5)$$

CDF adalah fungsi yang menjumlahkan nilai probabilitas sampai suatu kejadian tertentu. CDF menggunakan nilai dari PDF, dimulai dari menjumlahkan suatu probabilitas dengan probabilitas selanjutnya, hingga terbentuk nilai kepadatan kumulatif sebesar 1. Rumus CDF ditunjukkan pada persamaan (6), dimana $p(X \leq x_k)$ adalah suatu nilai probabilitas sebanyak k. Misal $p(x_1)$ *CDF* adalah $p(x_1)$ PDF, $p(x_2)$ *CDF* adalah $p(x_1)$ PDF + $p(x_2)$ PDF, dan $p(x_n)$ *CDF* adalah $p(x_1)$ PDF + $p(x_2)$ PDF + $p(x_n)$ PDF.

$$p(X \leq x_k) = \sum_{i=1}^{k} p(x_i) \quad (6)$$

4) Correlation Function

Untuk mengetahui kekuatan hubungan antara suatu variabel (random) dengan variabel yang lain, salah satunya dapat menggunakan analisis korelasi. Analisis korelasi ini akan menghasilkan suatu koefisien korelasi yang menyatakan apakah suatu variabel (random) dipengaruhi oleh variabel lain. Koefisien korelasi ini dinyatakan dalam persentase (%). Rumus untuk mendapatkan koefisien korelasi dinyatakan dalam r mengikuti persamaan (7), dimana X diasumsikan sebagai variabel prediktor, dan Y diasumsikan sebagai variabel respon.

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2\right]}\sqrt{\left[n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2\right]}} \quad (7)$$

IV. HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan hasil eksperimen beserta pembahasannya terkait dengan data terstruktur kasus COVID-19 di Indonesia, konfigurasi Hadoop Environment, dan HiveQL serta *Graphical Exploratory Data Analysis.*

A. *Data Terstruktur Kasus COVID-19 Dataset di Indonesia*

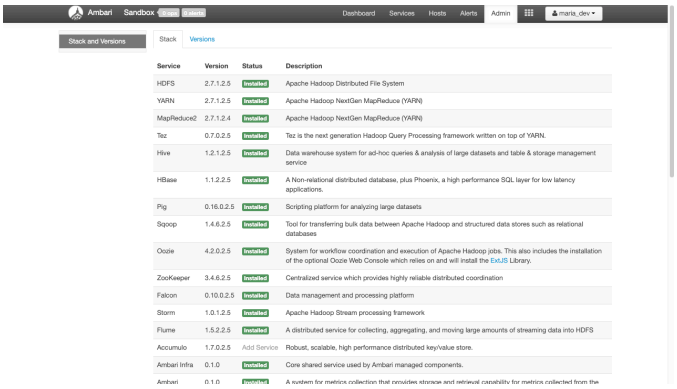
Data kasus COVID-19 dalam penelitian ini didapatkan dari website resmi Badan Nasional Penanggulangan Bencana, Republik Indonesia atau dapat diakses di https://bnpb-inacovid19.hub.arcgis.com/datasets. Pada Tabel 1 menunjukkan struktur tabel data kasus COVID-19 di Indonesia yang terdiri dari kolom jumlah kasus terkonfirmasi positif, kasus pasien sembuh, kasus pasien meninggal dan kasus pasien dalam perawatan, serta indeks hari atau bulan. Satu baris data merepresentasikan jumlah frekuensi kasus pada masing-masing kategori, dan total terdapat 213 baris data. Salah satu hal yang menarik adalah kasus pasien dalam perawatan per hari dapat dilaporkan minus (-), artinya pada hari itu tidak ada pelaporan kasus pasien dalam perawatan baru, namun memungkinkan pasien dilaporkan sembuh atau meninggal pada hari berikutnya.

TABEL I
DATA KASUS COVID-19 DI INDONESIA BULAN MARET-SEPTEMBER 2020

Hari ke-	Kategori			
	Kasus Terkonfirmasi per hari	Kasus Sembuh per hari	Kasus Meninggal per hari	Kasus Dalam Perawatan per hari
1	2	0	0	2
2	0	0	0	0
3	0	0	0	0
. . .	. . .	. . .	. . .	. . .
211	3509	3856	87	-434
212	4002	3567	128	307
213	4284	4510	139	-365

Pada penelitian ini data yang dilibatkan tidak menggunakan data kasus COVID-19 kumulatif, tetapi menggunakan kasus per hari. Jika menggunakan data kasus COVID-19 kumulatif, analisisnya akan mengarah ke laju pertumbuhan kasus. Selain itu, jika menggunakan data kasus kumulatif, tentunya jumlah kasus hari ini akan dipengaruhi oleh jumlah kasus sebelumnya, dan jumlah kasus kedepan akan dipengaruhi jumlah kasus hari ini dan sebelumnya. Padahal, penelitian ini bertujuan untuk menganalisis dan mengeksplorasi pola-pola kasus per hari atau per bulan agar dapat diketahui probabilitas kejadian dan korelasinya pada kategori kasus masing-masing.

B. *Konfigurasi Hadoop Environment*



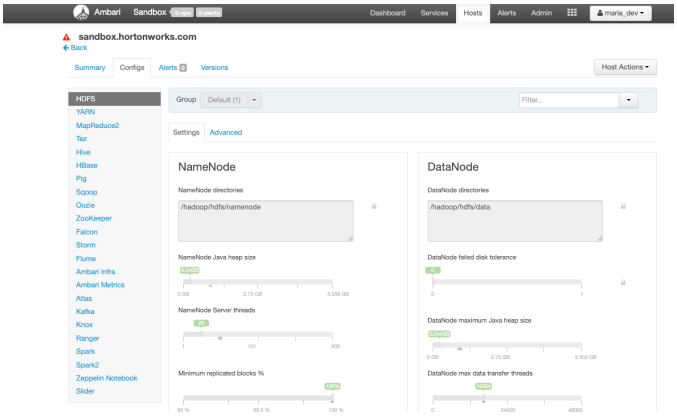
Gbr. 5 Modul-modul Arsitektur Big Data pada Hortonworks Sandbox

Untuk mengimplementasikan EDA pada penelitian ini ke dalam arsitektur Big Data, software yang cukup direkomendasikan salah satunya adalah Hortonworks (Docker)

ISSN (Online) 2747-0563
Volume 1 Tahun 2020

Sandbox. Pada Hortonworks Sandbox ini, Hadoop menjadi fondasi atau dalam arsitektur Big Data sebagai *layer* penyimpanan data dalam jumlah besar dan data tersebut telah dibentuk menjadi file terdistribusi yang disebut HDFS.

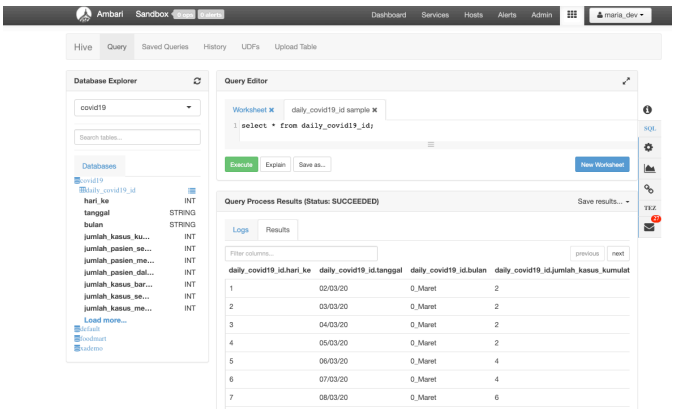
Pada Gbr. 5 menunjukkan modul-modul yang sudah terinstal pada Hortonworks Sandbox antara lain Hadoop, YARN, MapReduce2, Hive, Pig, Zookeper dan modul-modul lainnya, sedangkan Pada Gbr. 6 menunjukkan modul-modul yang dapat dikonfigurasi ulang menyesuaikan kebutuhan pengguna.



Gbr. 6 Arsitektur Big Data pada Hortonworks Sandbox

C. *HiveQL dan Graphical Exploratory Data Analysis*

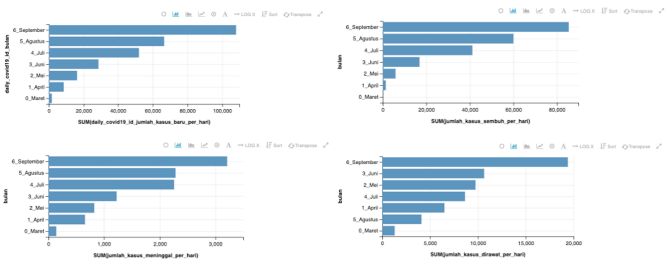
Pada bagian ini akan dibahas lebih mendalam hasil eksperimen yang sudah dilakukan untuk mengeksplorasi data kasus COVID-19 di Indonesia menggunakan HiveQL dan Hadoop environment pada berbagai teknik *Exploratory Data Analysis*. Gbr. 7 menunjukkan Hive Query Editor, dimana user dapat melakukan berbagai query. Pada ruang kerja tersebut, pada bagian sisi kiri terdapat menu *database explorer* untuk melihat struktur basis data, tabel dan kolom serta atributnya. Pada bagian tengah terdapat *worksheet* yang dapat ditambah sesuai kebutuhan, tempat untuk mengeksekusi query dan sebagai tempat output query akan ditampilkan.



Gbr. 7 Hive Query Editor pada Hortonworks Sandbox

1) *Descriptive Statistics pada Kasus COVID-19 di Indonesia*

Penggunaan statistik deskriptif pada penelitian ini bertujuan untuk mengetahui jumlah dan rata-rata tiap kategori kasus COVID-19 yang terjadi di Indonesia per bulan. Apakah terdapat kecenderungan pada suatu bulan yang mengakibatkan lonjakan pada kategori tertentu. Pada Gbr. 8 menunjukkan bahwa total kasus pada tiap kategori cenderung terjadi penambahan kasus tiap bulan, diagram *bar* berwarna biru menunjukkan kasus yang paling sedikit terjadi di bulan Maret dan yang paling banyak terjadi di bulan September. Tentu yang diharapkan adalah kasus terkonfirmasi, kasus meninggal, dan kasus dalam perawatan semakin menurun, sedangkan kasus pasien sembuh semakin meningkat.



Gbr. 8 Diagram *Horizontal Bar:* Total Jumlah Kasus COVID-19 di Indonesia pada Tiap Kategori di Bulan Maret-September 2020

Tabel II menunjukkan informasi jumlah kasus tiap kategori dan dilengkapi dengan indeks bulan, dimana pada bulan Maret jumlah kasus terkonfirmasi positif sebanyak 1.528 dan *trend* kasus semakin melonjak pada bulan September terdapat 112.212 kasus. Hal yang menarik adalah adanya lonjakan kasus terkonfirmasi positif di bulan September, diiringi dengan jumlah kasus pasien yang sembuh di bulan September sebanyak 88.988. Hal ini yang patut menjadi perhatian agar jumlah kasus pasien sembuh semakin bertambah mengungguli jumlah kasus pasien terkonfirmasi positif, namun juga tetap menekan jumlah pasien yang meninggal agar semakin menurun tiap bulannya.

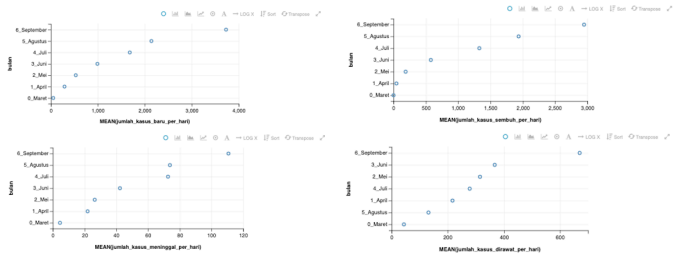
TABEL III
TOTAL JUMLAH KASUS COVID-19 DI INDONESIA PADA TIAP KATEGORI BULAN MARET-SEPTEMBER 2020

Bulan	Kategori			
	Kasus Terkonfir-masi Baru per hari (SUM)	Kasus Sembuh per hari (SUM)	Kasus Meninggal per hari (SUM)	Kasus Dalam Perawatan per hari (SUM)
Maret	1528	81	136	1311
April	8590	1441	656	6493
Mei	16355	5786	821	9748
Juni	29912	17498	1263	11151
Juli	51991	41101	2255	8635
Agustus	66420	60052	2286	4082
September	112212	88988	3323	19901

Pada Gbr. 9 menunjukkan diagram *scatter* yang menghubungkan antara axis *x* yakni jumlah rata-rata kasus COVID-19 tiap kategori dengan axis *y* yakni bulan. Terlihat bahwa titik berwarna biru di bulan September mengungguli

ISSN (Online) 2747-0563
Volume 1 Tahun 2020

seluruh bulan lainnya dari berbagai kategori kasus yang ada, sedangkan bulan Maret menjadi rata-rata titik terendah terjadinya kasus COVID-19 pada masing-masing kategori. Hal ini akan menjadi pertanyaan analisis baru, apakah rata-rata kasus COVID-19 per hari di bulan Oktober juga akan mengungguli rata-rata kasus di bulan sebelumnya.



Gbr. 9 Diagram *Scatter:* Rata-rata Jumlah Kasus COVID-19 di Indonesia Tiap Kategori di Bulan Maret-September 2020

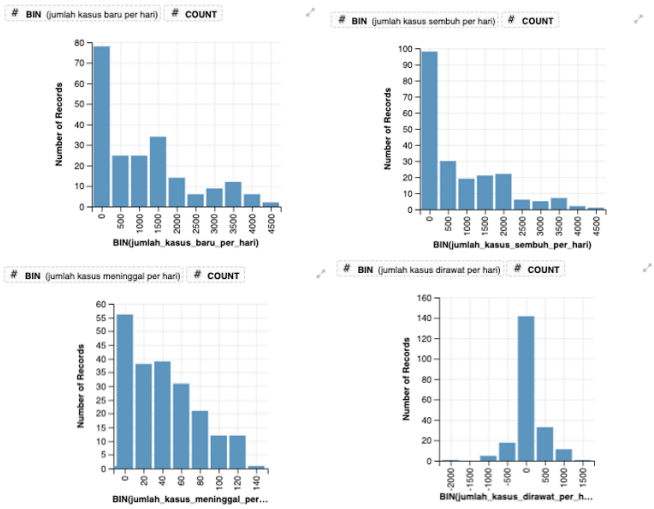
Pada Tabel III semakin terlihat bahwa rata-rata perubahan *trend* jumlah kasus terkonfirmasi positif tiap bulan terjadi penambahan yang cukup signifikan. Pada bulan Maret, rata-rata jumlah kasus terkonfirmasi positif sebanyak 51 kasus per harinya, namun di bulan April *trend* jumlah kasusnya berubah menjadi 286 kasus per hari. Di bulan Mei tepat di bulan Ramadan, rata-rata jumlah kasus berubah menjadi 527 kasus per hari, sedangkan di akhir Mei dan awal-akhir bulan Juni yakni bertepatan pada Hari Raya Idul Fitri, rata-rata kasus terkonfirmasi positif melonjak menjadi 997 kasus per hari. Hingga bulan September, rata-rata kasus terkonfirmasi positif menjadi 3.740 per hari, tentu banyak hal yang perlu dievaluasi untuk menekan angka kasus COVID-19 ini.

TABEL IIIII
RATA-RATA JUMLAH KASUS COVID-19 DI INDONESIA TIAP KATEGORI BULAN MARET-SEPTEMBER 2020

Bulan	Kategori			
	Kasus Terkonfir-masi Baru per hari (AVG)	Kasus Sembuh per hari (AVG)	Kasus Meninggal per hari (AVG)	Kasus Dalam Perawatan per hari (AVG)
Maret	50,933	2,7	4,533	43,7
April	286,333	48,0333	21,867	216,433
Mei	527,580	186,645	26,484	314,451
Juni	997,067	583,267	42,1	371,7
Juli	1677,129	1325,84	72,742	278,548
Agustus	2142,58	1937,16	73,742	131,677
September	3740,4	2966,27	110,77	663,37

2) Histogram pada Kasus COVID-19 di Indonesia

Histogram digunakan pada penelitian ini bertujuan untuk mengetahui sebaran nilai pada tiap kategori kasus COVID-19 jatuh pada interval mana saja. Apakah sebaran nilai tersebut jatuh pada grafik yang mencondong ke kiri (*skewed left histogram*), mencondong ke kanan (*skewed right histogram*) atau justru berbentuk diagram lonceng (*normal histogram*). Tentu bentuk grafik yang diharapkan menyesuaikan dengan kebutuhan analisis dan asumsi-asumsi yang digunakan dalam proses analisis. Pada Gbr. 10, terlihat bahwa jumlah kasus terkonfirmasi positif (kanan-atas) menunjukkan diagram *bar* yang mencondong ke kiri, terdapat angka kasus 0 yang mendominasi dibandingkan *bin* lainnya. Namun, jumlah kasus terkonfirmasi positif lainnya mulai dari *bin* 500-4.500 kasus, tetap menjadi perhatian bersama untuk bagaimana menindaklanjutinya.



Gbr. 10 Diagram *Bar:* Histogram Tiap Kategori Kasus COVID-19 di Indonesia Bulan Maret-September 2020

3) PDF dan CDF pada Kasus COVID-19 di Indonesia

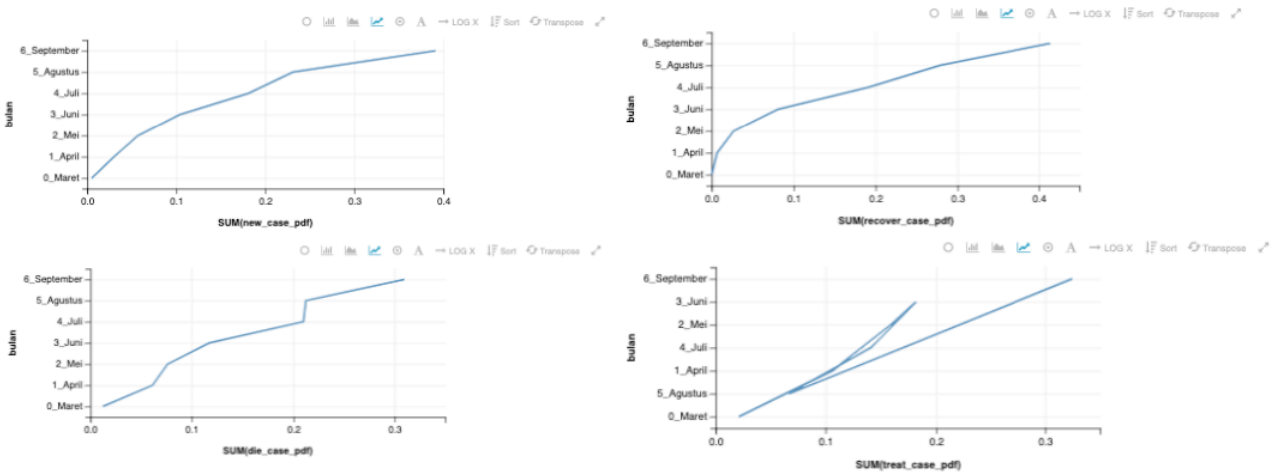
PDF dan CDF cukup membantu untuk mengetahui probabilitas suatu kasus dan laju pertumbuhan suatu kasus tiap harinya. PDF pada penelitian ini digunakan untuk mengetahui distirbusi probabilitas terjadinya kasus COVID-19 tiap kategori. Untuk mengimplementasikan analisis PDF pada HiveQL, gunakan *syntax* query berikut:

Query PDF: SELECT [COLUMN], [my_case / sum(my_case)] OVER() AS my_case_pdf FROM [TABLE];

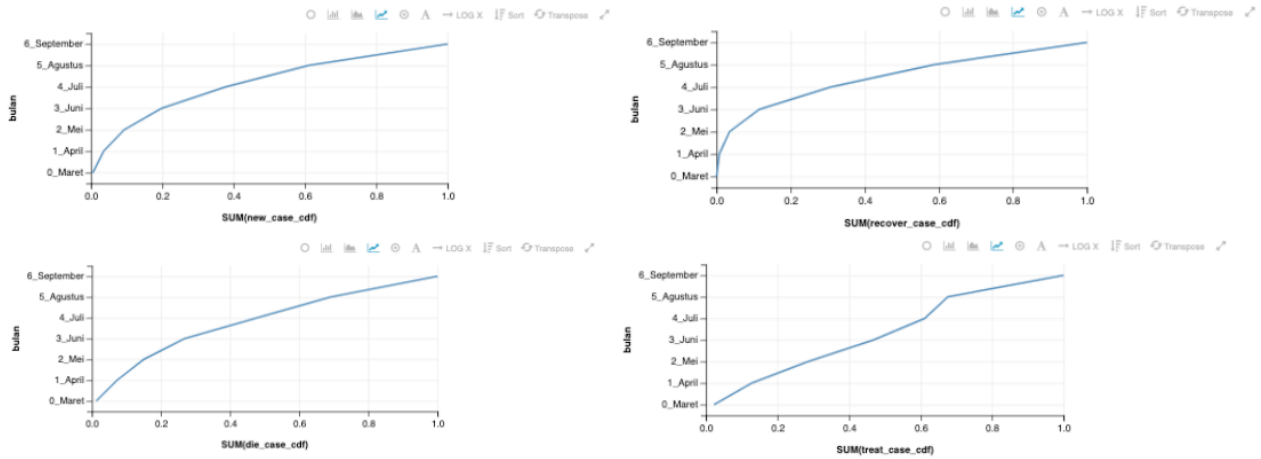
Pada Gbr. 11 menunjukkan diagram *line,* dimana terlihat bahwa kasus terkonfirmasi positif, kasus sembuh dan kasus meninggal cukup signifikan kenaikannya tiap bulan. Hal ini semakin memperkuat analisis statistik deskriptif sebelumnya, baik dari total jumlah kasus maupun rata-rata. Probabilitas kasus paling besar terjadi di bulan September yakni mendekati angka 0.3 hingga 0.4. Terkait dengan jumlah pasien yang dirawat pada bulan Agustus sempat terjadi penurunan, namun naik Kembali pada bulan-bulan berikutnya.

Berbeda dengan analisis PDF, analisis CDF digunakan untuk melihat laju pertumbuhan kasus tiap periode, misal pada periode bulan. CDF ini juga menjadi teknik yang digunakan oleh pemerintah untuk melihat pertumbuhan kasus COVID-19 secara kumulatif, artinya kumulatif kasus hari ke-*i* merupakan penjumlahan dari probabilitas kasus hari ke *i-1* dan ke-*i*, sedangkan kumulatif kasus hari ke *i+1* merupakan penjumlahan dari probabilitas kasus hari ke-*i, i-1* dan *i+1*.

ISSN (Online) 2747-0563
Volume 1 Tahun 2020



Gbr. 11 Diagram *Line: Probability Density Function* Tiap Kategori Kasus COVID-19 di Indonesia Bulan Maret-September 2020



Gbr. 12 Diagram *Line: Cumulative Distribution Function* Tiap Kategori Kasus COVID-19 di Indonesia Bulan Maret-September 2020

Untuk mengimplementasikan analisis CDF pada HiveQL, gunakan *syntax* query berikut:

Query CDF: SELECT [COLUMN], SUM(my_case) OVER(ORDER BY my_index) as my_case_cdf FROM [TABLE];

Pada Gbr. 12 menunjukkan laju pertumbuhan kasus terkonfirmasi positif yang terus meningkat, belum ada tanda-tanda adanya grafik yang melandai. Pada analisis CDF, titik maksimum nilai grafik selalu mendekati 1, karena merupakan hasil penjumlahan dari tiap pergerakan data sebelumnya.

4) Correlation Function pada Kasus COVID-19 di Indonesia

Analisis korelasi berguna untuk mengetahui hubungan suatu variabel dengan variabel lainnya, apakah ada pengaruh antara kedua variabel tersebut. Pada data COVID-19 ini terdapat 4 variabel, dimana 6 pasang hasil analisis korelasi ditemukan dalam penelitian ini. Untuk mengimplementasikan analisis korelasi pada HiveQL, gunakan *syntax* query berikut:

Query: CORR (my_column1, my_column2) FROM [TABLE];

TABEL IVV
PERBANDINGAN NILAI KOEFISIEN KORELASI ANTAR KATEGORI KASUS COVID-19 DI INDONESIA

Korelasi Kasus	Nilai Korelasi
Kasus Terkonfirmasi-Kasus Sembuh	0.94
Kasus Terkonfirmasi-Kasus Meninggal	0.9
Kasus Terkonfirmasi-Kasus Dalam Perawatan	0.38
Kasus Sembuh-Kasus Meninggal	0.89
Kasus Sembuh-Kasus Dalam Perawatan	0.049
Kasus Meninggal-Kasus Dalam Perawatan	0.211

ISSN (Online) 2747-0563
Volume 1 Tahun 2020

Pada Tabel IV terlihat bahwa koreasi antara kasus terkonfirmasi positif dengan kasus sembuh dan kasus meninggal cukup signifikan yakni sebesar 0.94 dan 0.9 masing-masing, artinya terdapat pengaruh kuat di antara keduanya. Bertambahnya kasus orang yang meninggal tiap harinya 90% dipengaruhi oleh bertambahnya kasus terkonfirmasi positif tiap harinya, begitu pula kasus pasien sembuh. Namun, terdapat hal lain dimana nilai korelasi antar dua variabel justru kecil, misal korelasi antara kasus dalam perawatan dengan kasus terkonfirmasi positif, kasus meninggal dan kasus sembuh, mungkin terdapat faktor lain yang mempengaruhi korelasi tersebut.

V. KESIMPULAN

Kasus COVID-19 menjadi suatu hal menarik untuk dilakukan analisis dan eksplorasi secara mendalam, guna menemukan pola-pola tersembunyi, mengungkap informasi yang belum diketahui sebelumnya, dan tentunya memberikan dampak untuk tindak lanjut ke depannya. Hasil eksperimen menunjukkan bahwa melalui teknik *Exploratory Data Analysis* dan didukung dengan lingkungan Big Data, cukup membantu untuk dilakukan eksplorasi terhadap data COVID-19. Hasil analisis menunjukkan bahwa teknik statistik deskriptif menemukan total jumlah kasus COVID-19 tiap kategori paling sedikit terjadi pada bulan Maret sedangkan paling besar terjadi di bulan September 2020. Rata-rata harian kasus COVID-19 tiap bulannya selalu bertambah, dari 50 kasus per hari di bulan Maret hingga melonjak menjadi 3.740 per hari di bulan September. Selain itu, analisis PDF-CDF menunjukkan laju pertumbuhan kasus terkonfirmasi positif COVID-19 semakin meningkat, belum ditemukan tanda-tanda melandai. Terakhir, analisis nilai korelasi menunjukkan terdapat pengaruh kuat antara pertambahan jumlah kasus terkonfirmasi positif dengan jumlah kasus pasien sembuh dan kasus pasien meninggal sebesar 0.94 dan 0.9 masing-masing. Hasil analisis korelasi yang lain ditemukan nilai korelasi yang kecil antara kasus pasien dalam perawatan terhadap kasus terkonfirmasi positif, kasus pasien meninggal, dan kasus pasien sembuh. Untuk keberlanjutan penelitian ini, perlu diimplementasikan analisis regresi karena bentuk data terstruktur kasus COVID-19 ini seperti *time series*. Dengan harapan mendapatkan model prediktif yang *fit* terhadap kasus COVID-19 di Indonesia dan memanfaatkan model sebagai penunjang prediksi berakhirnya COVID-19 ke depan.

REFERENSI

[1] World Health Organization, "Critical Preparedness, Readiness and Response Actions for COVID-19," 2020.

[2] United Nations, "Launch of Global Humanitarian Response Plan for COVID-19," 2020.

[3] R. Djalante, J. Lassa, D. Setiamarga, A. Sudjatma, M. Indrawan, B. Haryanto, C. Mahfud, M. S. Sinapoy, S. Djalante, I. Rafliana, L. A. Gunawan, G. A. K. Surtiari and H. Warsilah, "Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020," *Progress in Disaster Science,* vol. 6, no. 100091, pp. 1-9, 2020.

[4] Asian Development Bank, "ADB Approves $3 Million Grant to Support Indonesia's Fight Against COVID-19," 2020.

[5] Satuan Tugas Penanganan COVID-19, "Peta Sebaran COVID-19 di Indonesia," 2020. [Online]. Available: https://covid19.go.id/peta-sebaran, tanggal akses 11 Oktober 2020.

[6] B. Sharma, "Processing of Data and Analysis," *Biostatistics and Epidemiology International Journal,* vol. 1, no. 1, pp. 3-5, 2018.

[7] P. M. Buscema, F. D. Torre, M. Breda, G. Massini and E. Grossi, "COVID-19 in Italy and Extreme Data Mining," *Physica A,* vol. 557, no. 124991, pp. 1-7, 2020.

[8] P. Radanliev, D. D. Roure and R. Walton, "Data Mining and Analysis of Scientific Research Data Records on Covid- 19 Mortality, Immunity, and Vaccine Development - In The First Wave of The Covid-19 Pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 14, no. 5, pp. 1121-1132, 2020.

[9] M. Roche, "COVID-19 and Media Datasets: Period- and Location-Specific Textual Data Mining," *Data in Brief,* vol. 33, no. 106356, pp. 1-3, 2020.

[10] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar and A. V. Vasilakos, "Big Data: From Beginning to Future," *International Journal of Information Management,* vol. 36, no. 6, pp. 1231-1247, 2016.

[11] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen and S. Belfkih, "Big Data Technologies: A Survey," *Journal of King Saud University – Computer and Information Sciences,* vol. 33, no. 4, pp. 431-448, 2018.

[12] D. Usha and A. J. A.P.S, "A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce," *International Journal of Current Engineering and Technology,* vol. 4, no. 2, pp. 602-606, 2014.

[13] S. Sakr, Big Data 2.0 Processing Systems: A Survey, Switzerland: Springer, 2016.

[14] H. Bansal, S. Chauhan and S. Mehrotra, Apache Hive Cookbook, Birmingham: Pack Publishing Ltd, 2016.

[15] A. Loganathan, A. Sinha, M. V. and S. Natarajan, "A Systematic Approach to Big Data Exploration of the Hadoop Framework," *International Journal of Information & Computation Technology,* vol. 4, no. 9, pp. 869-878, 2014.

[16] J. T. Behrens, "Principles and Procedures of Exploration Data Analysis," *Psychological Methods,* vol. 2, no. 2, pp. 131-160, 1997.

[17] K. Sahoo, A. K. Samal, J. Pramanik and S. K. Pani, "Exploratory Data Analysis using Python," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 8, no. 12, p. 2019, 2019.

[18] C. Chesneau, T. Hussain and H. S. Bakouch, "A New Cumulative Distribution Function Based on m Existing Ones," *UPB Scientific Bulletin, Series A: Applied Mathematics and Physics,* vol. 80, no. 3, pp. 75-82, 2018.