

**MODUL CLOUD COMPUTING DAN HL7 DALAM PELAYANAN
KESEHATAN
BIG DATA
PERTEMUAN 10 (ONLINE)**



Disusun Oleh
Syefira Salsabila

Menurut Gartner, “Big data adalah data yang merupakan aset informasi dimana variasinya banyak, kecepatannya tinggi, dan ukurannya besar yang secara biaya (bisa) lebih murah dan dalam bentuk-bentuk (teknologi) inovatif untuk proses automasi, pengambilan keputusan dan memungkinkan untuk mendapatkan informasi mendalam.” *Big data* merupakan istilah untuk data sangat besar dan kompleks yang tidak dapat dikelola (*capture, store, manage, analyze*) dengan *software* dan *tool* pemrograman database biasa/konvensional.

Dalam kesehatan masyarakat kita terutama tertarik terhadap perilaku dan perubahan perilaku dari masyarakat, institusi kesehatan, tenaga kesehatan dan politisi. Ketersediaan data yang besar dan cepat dari segala sumber merupakan alat yang bermanfaat untuk kesehatan masyarakat. Tetapi, ketersediaan *big data* juga memberikan tantangan tersendiri dalam pemanfaatannya. Tantangan pemanfaatan *big data* dalam Kesehatan Masyarakat ini antara lain ketersediaan dan akses data, struktur dan klasifikasi data, privasi dan kerahasiaan data perorangan, serta bias dalam analisis faktor risiko.

Data yang telah dibuat dan disimpan pada tingkat global hari ini hampir tak terbayangkan jumlahnya. Data tersebut terus tumbuh tanpa henti. Sayangnya sampai saat ini, baru sebagian kecil data yang telah dianalisis. Pentingnya *big data*, tidak hanya seputar pada jumlah data yang di miliki, tetapi hal yang penting adalah bagaimana mengolah data-data tersebut untuk pengambilan keputusan. Pada *Big Data*, data terlalu besar dan terlalu cepat atau tidak sesuai dengan struktur arsitektur database konvensional. Untuk mendapatkan nilai dari data, harus digunakan teknologi untuk mengekstrak dan memperoleh informasi yang lebih spesifik.

Banyak pihak yang mencoba memberikan definisi terhadap *Big Data* semuanya mengacu pada 3V: *Volume, Variety, Velocity*, dan ada yang menambahkan unsur *Veracity* dan *Value*. *Volume* (kapasitas data) berkaitan dengan ukuran media penyimpanan data yang sangat besar atau mungkin tak terbatas hingga satuan petabytes atau zettabytes; *Variety* (keragaman data) terkait tipe atau jenis data yang dapat diolah mulai dari data terstruktur hingga data tidak terstruktur; sedangkan *Velocity* (kecepatan) terkait dengan kecepatan memproses data yang dihasilkan dari berbagai sumber, mulai dari data *batch* hingga *real time*, sementara karakteristik *Veracity* (kebenaran) dan *Value* (nilai) terkait dengan ketidakpastian data dan nilai manfaat dari informasi yang dihasilkan.

Teknologi *Big Data* dapat dimanfaatkan oleh banyak pihak, baik Perusahaan/corporate besar, Usaha Kecil dan Menengah (UKM), maupun pemerintah. Meski pemanfaatan ‘Big Data’ terbilang rumit dan mahal, namun UKM bermodal kecil pun bisa memanfaatkannya asal tahu persis tujuan bisnisnya sehingga memudahkan proses identifikasi data yang dibutuhkan. Manfaat *Big Data* sudah dirasakan khususnya sektor private/perusahaan, diantaranya untuk:

- a. Mengetahui respons masyarakat terhadap produk-produk yang dikeluarkan melalui analisis sentimen di media sosial.
- b. Membantu perusahaan mengambil keputusan secara lebih tepat dan akurat berdasarkan data.
- c. Membantu meningkatkan citra perusahaan di mata pelanggan. Perencanaan usaha, dengan mengetahui perilaku pelanggan seperti pada perusahaan telekomunikasi dan perbankan.
- d. Mengetahui tren pasar dan keinginan konsumen.

Selain dimanfaatkan untuk analisis bisnis, muncul harapan besar bahwa teknologi *Big Data* juga akan dimanfaatkan secara luas di pemerintahan. Beberapa peluang pemanfaatan *Big Data* di sektor publik antara lain:

- a. Mendapatkan *feedback* dan respon masyarakat sebagai dasar penyusunan kebijakan dan perbaikan pelayanan publik. *Feedback* tersebut dapat diperoleh dari sistem informasi layanan pemerintah maupun dari media sosial.
- b. Membuat layanan terpadu dengan segmen khusus sehingga layanan bisa lebih efektif dan efisien.
- c. Menemukan solusi atas permasalahan yang ada, berdasarkan data. Contohnya, dengan menganalisa informasi cuaca dan informasi pertanian terkait data tingkat kesuburan tanah, pemerintah dapat menetapkan atau menghimbau jenis varietas tanaman yang ditanam oleh petani pada daerah dan waktu tertentu.

One of the main purposes of Big Data focuses on healthcare. The aim is to use Big Data and the increasingly powerful computational systems to shift from generic to personalised medicine. The intention is to decode each patient's genetic sequences in order to personalise and know for certain which treatment would be most efficient for each case.

Another aim of Big Data is control of epidemics. A georeferenced factor is added to traditional databases and, together with the information obtained from official as well as informal means such as blogs, searches, social media sites, etc., an estimation of the progression of the illness can be made.

An enormous amount of information is collected in healthcare and, for this reason, computer analysis techniques are not sufficient, requiring that new algorithms and methods be found to carry out these analyses more efficiently. Tools such as Mahout's Apache Hadoop, Google's MapReduce and the NoSQL databases have been developed more as format than a tool. Relational databases (such as SQL) often cannot be used with Big Data. One solution is to use non-relational databases (NoSQL). Examples of these include MongoDB, CouchDB or SimpleDB.

Hospitals are another area where Big Data analysis of patients' tests may give better early diagnostic results. In this manner, patients' drug dosage could be lowered. Detection of infectious foci in hospitals or the evolution of possible viruses that could become an epidemic, such as the H1N1 virus, are also key issues. When said flu virus broke out, some companies were already using Big Data to analyse its progression. Google, for example, analysed via geolocation and the number of hits matching the search where the virus was spreading with the purpose of finding a model to predict its progression and warn the Centers for Disease Control and Prevention (CDC).

Finally, although Big Data use is on the rise in every field of healthcare, it is important to mention diabetes. The disease is affecting an increasingly larger number of persons across the world. It is the cause of a considerable number of deaths in the noncommunicable disease category. The onset of diabetes can appear at any time, even at birth, and requires lifelong treatment. For this reason, Big Data is being used to search for a cure as there is a large amount of data and information available.

The use of Big Data is having a positive impact on the field of health. Massive amounts of information are now being generated, with approximately 1.8 ZB in 2011 and rapid growth continuing since that time. All of this information cannot be analysed with traditional methods, meaning that alternatives must be found. Big Data is not the solution as it only makes reference to the information. However, it is the first step towards methods and tools that can analyse it. Going beyond the above, the aim is not only to analyse the information but to ensure that it makes sense as different interpretations can be made from the same results.

In the hospital industry, Big Data and its analysis are helping to cut costs when handling clinical analyses from electronic medical records. This will make updated patient

information available to doctors in real time, thus giving them a more solid base to make a final diagnosis.

The ultimate objective of Big Data is cutting research costs and improving the quality of healthcare. The intention is to find faster better analysis of the medical tests conducted on patients, faster analysis of the information obtained from experiments and help medical staff to have greater knowledge of medicine. Obtaining results in real time makes working with different groups such as chemists, doctors or social workers more interactive and provides better outcomes.

Determining risk factors is one of the most important possible final aims for which Big Data can be used. This consists of using the information collected from clinical experiments on the use of medication to more easily and reliably determine the side effects of said drugs. There are a large number of subjects in clinical trials of this type and they can even be conducted at several centres simultaneously to determine additional factors. Big Data can link all this information and identify the risk factors being searched for. In conclusion, there is still a great deal to be explored in the field of Big Data in healthcare.

Big Data in Health: a Literature Review from the Year 2005. Isabel de la Torre Díez¹

The volume of data is growing exponentially, and it is expected that by 2020 there will be more than 16 zettabytes (16 Trillion GB) of useful data. We are on the verge of an era where every device is online, where sensors are ubiquitous in our world generating continuous streams of data, where the sheer volume of data offered and consumed on the Internet will increase by orders of magnitude, where the Internet of Things will produce a digital fingerprint of our world.

Big data is the emerging field where innovative technology offers new ways of extracting value from the tsunami of new information. The ability to effectively manage information and extract knowledge is now seen as a key competitive advantage. Many organizations are building their core business on their ability to collect and analyse information to extract business knowledge and insight. Big data technology adoption within industrial sectors is not a luxury but an imperative need for most organizations to survive and gain competitive advantage.

The potential for big data is expected to impact all sectors, from healthcare to media, from energy to retail. The positive transformational potential has already been identified in a number of key sectors. Big data brings together a set of data management challenges for working with data under new scales of size and complexity. Many of these challenges are not new. What is new however are the challenges raised by the specific characteristics of big data related to the 3 Vs:

- Volume (amount of data): dealing with large scales of data within data processing (e.g. Global Supply Chains, Global Financial Analysis, Large Hadron Collider).
- Velocity (speed of data): dealing with streams of high frequency of incoming real-time data (e.g. Sensors, Pervasive Environments, Electronic Trading, Internet of Things).

- Variety (range of data types/sources): dealing with data using differing syntactic formats (e.g. Spreadsheets, XML, DBMS), schemas, and meanings (e.g. Enterprise Data Integration).

Table 3.1 Definitions of big data

| Big data definition | Source |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|
| “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” | Laney (2001), Manyika et al. (2011) |
| “When the size of the data itself becomes part of the problem and traditional techniques for working with data run out of steam” | Loukides (2010) |
| Big Data is “data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time” | Jacobs (2009) |
| “Big Data technologies [are] a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis” | IDC (2011) |
| “The term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications” | Wikipedia (2014) |
| “A collection of large and complex data sets which can be processed only with difficulty by using on-hand database management tools” | Mike 2.0 (2014) |
| “Big Data is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies.” By extension, the platform, tools and software used for this purpose are collectively called “Big Data technologies” | NESSI (2012) |
| “Big data can mean big volume, big velocity, or big variety” | Stonebraker (2012) |

Within the field of Business Management, Value Chains have been used as a decision support tool to model the chain of activities that an organisation performs in order to deliver a valuable product or service to the market. In a Data Value Chain, information flow is described as a series of steps needed to generate value and useful insights from data.

The Big Data Value Chain, can be used to model the high-level activities that comprise an information system. The Big Data Value Chain identifies the following key high-level activities:

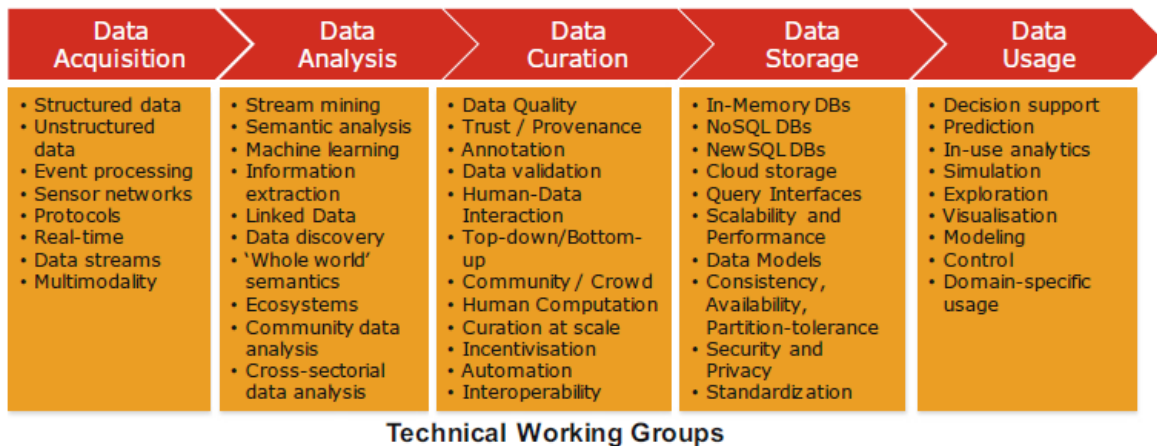


Fig. 3.1 The Big Data Value Chain as described within (Curry et al. 2014)

Data Acquisition is the process of gathering, filtering, and cleaning data before it is put in a data warehouse or any other storage solution on which data analysis can be carried out. Data acquisition is one of the major big data challenges in terms of infrastructure requirements. The infrastructure required to support the acquisition of big data must deliver low, predictable latency in both capturing data and in executing queries; be able to handle very high transaction volumes, often in a distributed environment; and support flexible and dynamic data structures. Data acquisition is further detailed in this chapter.

Data Analysis is concerned with making the raw data acquired amenable to use in decision-making as well as domain-specific usage. Data analysis involves exploring, transforming, and modelling data with the goal of highlighting relevant data, synthesising and extracting useful hidden information with high potential from a business point of view. Related areas include data mining, business intelligence, and machine learning.

Data Curation is the active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage. Data curation processes can be categorised into different activities such as content creation, selection, classification, transformation, validation, and preservation. Data curation is performed by expert curators that are responsible for improving the accessibility and quality of data. Data curators (also known as scientific curators, or data annotators) hold the responsibility of ensuring that data are trustworthy, discoverable, accessible, reusable, and fit their purpose. A key trend for the curation of big data utilises community and crowd sourcing approaches.

Data Storage is the persistence and management of data in a scalable way that satisfies the needs of applications that require fast access to the data. Relational Database Management Systems (RDBMS) have been the main, and almost unique, solution to the storage paradigm for nearly 40 years. However, the ACID (Atomicity, Consistency, Isolation, and Durability) properties that guarantee database transactions lack flexibility with regard to schema changes and the performance and fault tolerance

when data volumes and complexity grow, making them unsuitable for big data scenarios. NoSQL technologies have been designed with the scalability goal in mind and present a wide range of solutions based on alternative data models.

Data Usage covers the data-driven business activities that need access to data, its analysis, and the tools needed to integrate the data analysis within the business activity. Data usage in business decision-making can enhance competitiveness through reduction of costs, increased added value, or any other parameter that can be measured against existing performance criteria.

Within the health sector big data technology aims to establish a holistic approach whereby clinical, financial, and administrative data as well as patient behavioural data, population data, medical device data, and any other related health data are combined and used for retrospective, real-time, and predictive analysis. In order to establish a basis for the successful implementation of big data health applications, the challenge of data digitalization and acquisition (i.e. putting health data in a form suitable as input for analytic solutions) needs to be addressed. As of today, large amounts of health data are stored in data silos and data exchange is only possible via Scan, Fax, or email. Due to inflexible interfaces and missing standards, the aggregation of health data relies on individualized solutions with high costs.

In hospitals patient data is stored in CIS (clinical information system) or HER (electronic health record) systems. However, different clinical departments might use different systems, such as RIS (radiology information system), LIS (laboratory information system), or PACS (picture archiving and communication system) to store their data. There is no standard data model or EHR system. Existing mechanisms for data integration are either adaptations of standard data warehouse solutions from horizontal IT providers like Oracle Healthcare Data Model, Teradata's Healthcare Logical Data Model, IBM Healthcare Provider Data Model, or new solutions like the i2b2 platform. While the first three are mainly used to generate benchmarks regarding the performance of the overall hospital organization, the i2b2 platform establishes a data warehouse that allows the integration of data from different clinical departments in order to support the task of identifying patient cohorts. In doing so, structured data such as diagnoses and lab values are mapped to standardized coding systems. However, unstructured data is not further labelled with semantic information. Besides its main functionality of patient cohorts identification, the i2b2 hive offers several additional modules. Besides specific modules for data import, export, and visualization tasks, modules to create and use additional semantics are available. For example, the natural language processing (NLP) tool offers a means to extract concepts out of specific terms and connect them with structured knowledge.

Today, data can be exchanged by using exchange formats such as HL7. However, due to non-technical reasons such as privacy, health data is commonly not shared across organizations (phenomena of organizational silos). Information about diagnoses, procedures, lab values, demographics, medication, provider, etc., is in general provided in a structured format, but not automatically collected in a standardized manner. For

example, lab departments use their own coding system for lab values without an explicit mapping to the LOINC (Logical Observation Identifiers Names and Codes) standard. Also, different clinical departments often use different but customized report templates without specifying the common semantics. Both scenarios lead to difficulties in data acquisition and consequent integration.

Regarding unstructured data like texts and images, standards for describing high-level meta-information are only partially collected. In the imaging domain, the DICOM (Digital Imaging and Communications in Medicine) standard for specifying image metadata is available. However, for describing meta-information of clinical reports or clinical studies a common (agreed) standard is missing. To the best of our knowledge, for the representation of the content information of unstructured data like images, texts, or genomics data, no standard is available. Initial efforts to change this situation are initiatives such as the structured reporting initiative by RSNA or semantic annotations using standardized vocabularies. For example, the Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus of the US National Library of Medicine to capture topics of texts in the medical and biological domain. There also exist several translations to other languages. Since each EHR vendor provides their own data model, there is no standard data model for the usage of coding systems to represent the content of clinical reports. In terms of the underlying means for data representation, existing EHR systems rely on a case-centric rather than on a patient-centric representation of health data. This hinders longitudinal health data acquisition and integration.

Easy to use structured reporting tools are required which do not create extra work for clinicians, i.e. these systems need to be seamlessly integrated into the clinical workflow. In addition, available context information should be used to assist the clinicians. Given that structured reporting tools are implemented as easy-to-use tools, they can gain acceptance by clinicians such that most of the clinical documentation is carried out in a semi-structured form and the quality and quantity of semantic annotations increases.

From an organizational point of view, the storage, processing, access, and protection of big data has to be regulated on several different levels: institutional, regional, national, and international level. There is a need to define who authorizes which processes, who changes processes, and who implements process changes. Therefore, a proper and consistent legal framework or guidelines [e.g. ISO/IEC 27000] for all four levels are required.

IHE (integrating the healthcare enterprise) enables plug-and-play and secure access to health information whenever and wherever it is needed. It provides different specifications, tools, and services. IHE also promotes the use of well established and internationally accepted standards (e.g. Digital Imaging and Communications in Medicine, Health Level 7). Pharmaceutical and R&D data that encompass clinical trials, clinical studies, population and disease data, etc. is typically owned by the pharmaceutical companies, research labs/academia, or the government. As of today, a lot of manual effort is taken to collect all the datasets for conducting clinical studies and related analysis. The manual effort for collecting the data is quite high.

1. Teknologi Hadoop untuk Pemantauan Kondisi Vital Pasien

Beberapa rumah sakit di seluruh dunia telah menggunakan Hadoop untuk membantu stafnya bekerja secara efisien dengan Big Data. Tanpa Hadoop, sebagian besar sistem layanan kesehatan hampir tidak mungkin menganalisis data yang tidak terstruktur.



Children's Healthcare of Atlanta merawat lebih dari 6.200 anak di unit ICU mereka. Rata-rata durasi tinggal di ICU Pediatrik bervariasi dari satu bulan sampai satu tahun. Children's Healthcare of Atlanta menggunakan sensor di samping tempat tidur yang membantu mereka terus melacak kondisi vital pasien seperti tekanan darah, detak jantung dan pernafasan. Sensor ini menghasilkan data yang sangat besar, dan sistem yang lama tidak mampu untuk menyimpan data tersebut lebih dari 3 hari karena terkendala biaya storage. Padahal rumah sakit ini perlu menyimpan tanda-tanda vital tersebut untuk dianalisa. Jika ada perubahan pola, maka perlu ada alert untuk tim dokter dan asisten lain.

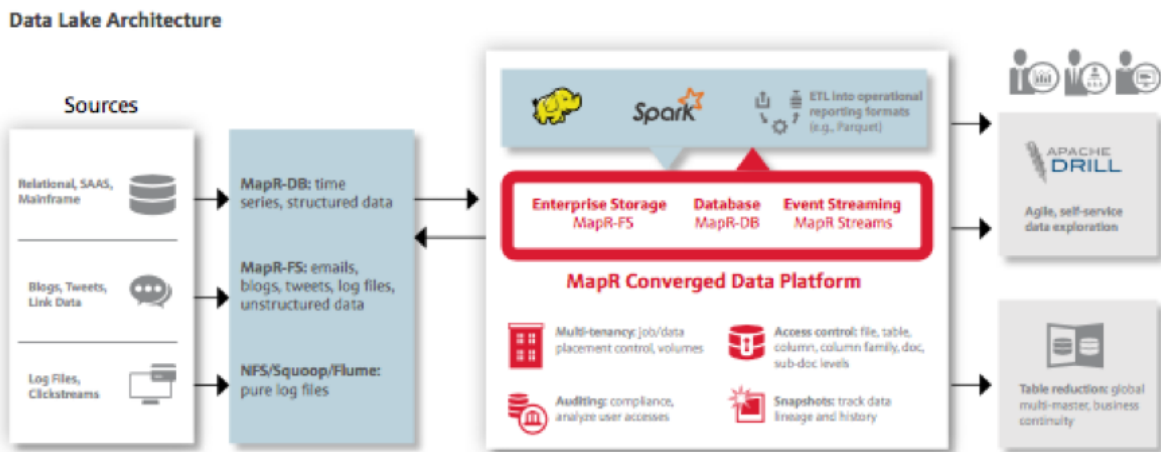
Sistem tersebut berhasil diimplementasikan dengan menggunakan komponen ekosistem Hadoop : Hive, Flume, Sqoop, Spark, dan Impala.

Setelah keberhasilan project tersebut, project berbasis Hadoop selanjutnya yang mereka lakukan adalah riset mengenai asma dengan menggunakan data kualitas

udara selama 20 tahun dari EPA (Environment Protection Agency). Tujuannya: mengurangi kunjungan IGD dan rawat inap untuk kejadian terkait asma pada anak-anak.

2. Valence health : peningkatan kualitas layanan dan reimbursements

Valence health menggunakan Hadoop untuk membangun data lake yang merupakan penyimpanan utama data perusahaan. Valence memproses 3000 inbound data feed dengan 45 jenis data setiap harinya. Data kritikal ini meliputi hasil tes lab, data rekam medis, resep dokter, imunisasi, obat, klaim dan pembayaran, serta klaim dari dokter dan rumah sakit, yang digunakan untuk menginformasikan keputusan dalam peningkatan baik itu pendapatan ataupun reimbursement. Pertumbuhan jumlah klien yang pesat dan peningkatan volume data terkait semakin membebani infrastruktur yang ada.



Sebelum menggunakan big data, mereka membutuhkan waktu hingga 22 jam untuk memproses 20 juta records data hasil laboratorium. Penggunaan big data memangkas waktu siklus dari 22 jam menjadi 20 menit, dengan menggunakan hardware yang jauh lebih sedikit. Valence Health juga mampu menangani permintaan pelanggan yang sebelumnya sulit untuk diselesaikan. Misalnya jika seorang klien menelpon dan mengatakan bahwa ia telah mengirimkan file yang salah 3 bulan yang lalu, dan perlu untuk menghapus data tersebut. Sebelumnya dengan solusi database tradisional, mereka memerlukan 3 sampai 4 minggu. Dengan memanfaatkan MapR snapshot yang menyediakan point-in-time recovery, Valence dapat melakukan roll-back dan menghapus file tersebut dalam hitungan menit.

3. Hadoop dalam Pengobatan Kanker dan Genomics

Salah satu alasan terbesar mengapa kanker belum dapat dibasmi sampai sekarang adalah karena kanker bermutasi dalam pola yang berbeda dan bereaksi dengan cara yang berbeda berdasarkan susunan genetik seseorang. Oleh karena

itu, para peneliti di bidang onkologi menyatakan bahwa untuk menyembuhkan kanker, pasien perlu diberi perawatan yang disesuaikan dengan jenis kanker berdasarkan genetika masing-masing pasien.

Ada sekitar 3 miliar pasangan nukleotida yang membentuk DNA manusia, dan diperlukan sejumlah besar data untuk diorganisir secara efektif jika kita ingin melakukan analisis. Teknologi big data, khususnya Hadoop dan ekosistemnya memberikan dukungan yang besar untuk paralelisasi dan proses pemetaan DNA.

David Cameron, Perdana Menteri Inggris telah mengumumkan dana pemerintah sebesar £ 300 juta pada bulan Agustus, 2014 untuk proyek 4 tahun dengan target memetakan 100.000 genom manusia pada akhir tahun 2017 bekerja sama dengan perusahaan Bioteknologi Amerika Illumina dan Genomics Inggris. Tujuan utama dari proyek ini adalah memanfaatkan big data dalam dunia kesehatan untuk mengembangkan personalized medicine bagi pasien kanker.

Arizona State University mengadakan sebuah proyek penelitian yang meneliti jutaan titik di DNA manusia untuk menemukan variasi penyebab kanker sedang berlangsung. Proyek ini merupakan bagian dari Complex Adaptive Systems Initiative (CASI), yang mendorong penggunaan teknologi untuk menciptakan solusi bagi permasalahan dunia yang kompleks.

Dengan menggunakan Apache Hadoop, tim peneliti universitas dapat memeriksa variasi dalam jutaan lokasi DNA untuk mengidentifikasi mekanisme kanker dan bagaimana jaringan berbagai gen mendorong kecenderungan dan efek kanker pada individu.

"Proyek kami memfasilitasi penggunaan data genomik berskala besar, sebuah tantangan bagi semua institusi penelitian yang menangani precision medicine," kata Jay Etchings, direktur komputasi riset ASU. Ekosistem Hadoop dan struktur data lake terkait menghindarkan setiap peneliti dan pengguna klinis untuk mengelola sendiri jejak data genomik yang besar dan kompleks.

4. UnitedHealthcare: Fraud, Waste, and Abuse

Saat ini setidaknya 10% dari pembayaran asuransi Kesehatan terkait dengan klaim palsu. Di seluruh dunia kasus ini diperkirakan mencapai nilai miliaran dolar. Klaim palsu bukanlah masalah baru, namun kompleksitas kecurangan asuransi tampaknya meningkat secara eksponensial sehingga menyulitkan perusahaan asuransi kesehatan untuk menghadapinya.

UnitedHealthCare adalah sebuah perusahaan asuransi yang memberikan manfaat dan layanan kesehatan kepada hampir 51 juta orang. Perusahaan ini menjalin kerja sama dengan lebih dari 850.000 orang tenaga kesehatan dan sekitar 6.100 rumah sakit di seluruh negeri. Payment Integrity group/divisi

integritas pembayaran mereka memiliki tugas untuk memastikan bahwa klaim dibayar dengan benar dan tepat waktu. Sebelumnya pendekatan mereka untuk mengelola lebih dari satu juta klaim per hari (sekitar 10 TB data tiap harinya) bersifat ad hoc, sangat terikat oleh aturan, serta terhambat oleh data yang terpisah-pisah. Solusi yang diambil oleh UnitedHealthCare adalah pendekatan dual mode, yang berfokus pada alokasi tabungan sekaligus menerapkan inovasi untuk terus memanfaatkan teknologi terbaru.

Dalam hal pengelolaan tabungan, divisi tersebut membuat “pabrik” analisis prediktif di mana mereka mengidentifikasi klaim yang tidak akurat secara sistematis dan tepat. Saat ini Hadoop merupakan data framework berplatform tunggal yang dilengkapi dengan tools untuk menganalisa informasi dari klaim, resep, plan peserta, penyedia layanan kesehatan yang dikontrak, dan hasil review klaim terkait.

Mereka mengintegrasikan semua data dari beberapa silo di seluruh bisnis, termasuk lebih dari 36 aset data. Saat ini mereka memiliki banyak model prediktif (PCR, True Fraud, Ayasdi, dll.) yang menyediakan peringkat provider yang berpotensi melakukan kecurangan, sehingga mereka dapat mengambil tindakan yang lebih terarah dan sistematis.

5. Liaison Technologies: Streaming System of Record for Healthcare

Liaison Technologies menyediakan solusi berbasis cloud untuk membantu organisasi dalam mengintegrasikan, mengelola, dan mengamankan data di seluruh perusahaan. Salah satu solusi vertikal yang mereka berikan adalah untuk industri kesehatan dan life science, yang harus menjawab dua tantangan : memenuhi persyaratan HIPAA dan mengatasi pertumbuhan format dan representasi data.

Dengan MapR Stream, permasalahan data lineage dapat terpecahkan karena stream menjadi sebuah SOR (System of Record) dengan berfungsi sebagai log yang infinite dan immutable dari setiap perubahan data. Tantangan kedua, yaitu format dan representasi data, bisa digambarkan dengan contoh berikut: rekam medis pasien dapat dilihat dengan beberapa cara yang berbeda (dokumen, grafik, atau pencarian) oleh pengguna yang berbeda, seperti perusahaan farmasi, rumah sakit, klinik, atau dokter.

Dengan melakukan streaming terhadap perubahan data secara real-time ke basis data, grafik, dan basis data MapR-DB, HBase, MapR-DB JSON, pengguna akan selalu mendapatkan data paling mutakhir dalam format yang paling sesuai.

Meningkatkan hasil pelayanan pada pasien dengan biaya yang sama atau bahkan lebih murah adalah tantangan bagi penyedia layanan kesehatan manapun, mengingat di Amerika keseluruhan biaya perawatan kesehatan meningkat sampai sekitar 15%. Transformasi digital adalah kunci untuk membuat mewujudkan tujuan ini. Digitalisasi, peningkatan komunikasi, dan big data analytics menjadi landasan untuk mendukung usaha transformasi tersebut.

Berbagai macam use-case untuk big data analytics sangat terkait dengan kemampuan Hadoop dan ekosistemnya dalam mengolah dan menyimpan data yang beraneka ragam, serta menyajikannya untuk analisa yang mendalam. Dalam memilih platform big data khususnya distribusi Hadoop, salah satu hal yang penting untuk dipertimbangkan adalah kemampuan untuk menangani berbagai macam tipe data dari sumber data yang terpisah-pisah : data klinis dalam satu silo, data farmasi di silo yang lain, data logistik, dan lain sebagainya. Platform yang dipilih hendaknya cukup fleksibel sehingga tenaga kesehatan dapat menggunakan data yang kompleks seperti catatan dokter, hasil scan, dan lain sebagainya untuk mendukung analisis terhadap pasien, bukan hanya untuk pengarsipan saja.

Beberapa *tools* yang tersedia untuk penerapan *Big Data* dapat dikategorikan menjadi 2, berbayar (*personalized software*) dan tidak berbayar (berbasis *open source software*). Tahapan aktivitas dan dukungan teknologi pada *Big Data* ditunjukkan pada bagan dibawah ini.

- a. *Acquired* berhubungan sumber dan cara mendapatkan data.
- b. *Accessed* berhubungan dengan daya akses data; data yang sudah dikumpulkan memerlukan tata kelola, integrasi, *storage* dan *computing* agar dapat dikelola untuk tahap berikutnya. Beberapa *processing/computing tools*: Hadoop, Nvidia CUDA, Twitter storm, GraphLab. Beberapa *tools* terkait *storage*: neo4J, Titan, HDFS.
- c. *Analytic* berhubungan dengan informasi yang akan didapatkan, hasil pengelolaan data yang telah diproses. Analitik yang dilakukan dapat berupa *descriptive* (penggambaran data), *diagnostic* (mencari sebab akibat berdasar data), *predictive* (memprediksi kejadian dimasa depan) maupun *prescriptive analytics* (merekomendasikan pilihan dan implikasi dari setiap opsi). Beberapa *analytics tools*: MLPACK, Mahout, R.
- d. *Application* terkait visualisasi dan reporting hasil dari analitik. Contoh *application tool*: RStudio.

Perangkat Lunak Tidak Berbayar (*Open Source Software*)

Salah satu aplikasi berbasis *open-source* yang populer untuk *Big Data* yaitu Hadoop. Hadoop merupakan framework untuk menjalankan aplikasi pada komoditas perangkat keras yang memiliki cluster besar. Hadoop menerapkan prinsip komputasi yang disebut *Map/Reduce*, yaitu model pemrograman untuk memproses data bervolume besar secara paralel dan independen dengan cara membagi tugas (*task*) kedalam beberapa fragmen kerja. Selain itu, Hadoop menyediakan sistem file terdistribusi (*HDFS*)

yang akan menyimpan data dalam titik-titik (*nodes*) komputasi. Dengan kedua teknik tersebut, kegagalan di satu titik diatasi secara otomatis oleh sistem. Komponen-komponen dalam Hadoop yaitu: Nama *node*, Data *node*, *Secondary-name node*, *Job Tracker*, *Task Tracker*.

Saat ini sudah tersedia *tools* yang mampu menghasilkan koding *map reduce* dari Hadoop. *Tools* tersebut memiliki *user interface* yang memungkinkan pengguna untuk meng-klik dan *drag* untuk membuat aliran pengolahan data dan secara otomatis menghasilkan dan mengeksekusi kode pada prosesor yang didistribusikan secara paralel. Dengan *tools* tersebut pengolahan data dapat dilakukan tanpa/sedikit koding. Selain Hadoop, *tools* yang populer digunakan serta memiliki fungsi dalam analitik dan penyajian data yang beragam adalah R. R merupakan bahasa pemrograman dan *software framework* untuk analisis statistik dan grafik. Keunggulan dari software R terletak pada pendekatan analisis data dengan menggunakan visualisasi untuk menemukan karakteristik yang penting dari suatu dataset.



Perangkat Lunak Berbayar (*Proprietary Software*)

Big Data analytics menjadi model bisnis baru bagi beberapa vendor/perusahaan. Seperti yang dilakukan oleh PT. Telkom Indonesia Tbk. sejak tahun 2014, dengan menawarkan *Big Data* untuk menunjang pertumbuhan di industri lain. Telkom telah menjadi penyedia *data analytics* bagi instansi/lembaga pemerintah pusat dan daerah, industri airlines, dan health care. Telkom juga memfasilitasi pertukaran data miliknya dengan sumber data lainnya, sebab dalam dunia operator telekomunikasi data yang dikelola sangat banyak seperti data transaksi, data lokasi serta data kunjungan ke website. Sebagai contoh, dengan jumlah pelanggan 156 juta orang, diketahui dalam satu menit terjadi kicauan di Twitter sebanyak 98 ribu, update status Facebook 695 ribu, serta pembicaraan di instant messaging 11 juta, data ini tentu dapat bercerita banyak jika diolah.

Selain itu, ada pula vendor/perusahaan yang secara spesifik bergerak di bidang *IT solution* yang menyediakan solusi berbasis teknologi *Big Data* sesuai dengan kebutuhan klien (*personalized*). Ada vendor yang membuat *tool analytic* secara parsial, misalnya *tool* untuk integrasi data saja, penyimpanan data, analisis data, dan presentasi

data hasil analisis. Tetapi ada pula vendor yang menyediakan *tool*/ secara lengkap, mulai dari integrasi data sampai analitis dan presentasi data, seperti Oracle, SAP, IBM, Microsoft, Teradata.

DAFTAR PUSTAKA

- Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard Business Review*, 84, 98–107.
- Choo, C. W. (1996). The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. *International Journal of Information Management*, 16, 329–340. doi:[10.1016/0268-4012\(96\)00020-5](https://doi.org/10.1016/0268-4012(96)00020-5).
- Curry, E., Ngonga, A., Domingue, J., Freitas, A., Strohbach, M., Becker, T., et al. (2014). D2.2.2. Final version of the technical white paper. Public deliverable of the EU-Project BIG (318062; ICT-2011.4.4).
- Curry, E., Freitas, A., & O’Ria’in, S. (2010). The role of community-driven data curation for enterprises. In D. Wood (Ed.), *Linking enterprise data* (pp. 25–47). Boston, MA: Springer US.
- DG Connect. (2013). A European strategy on the data value chain. European Commission. (2014). *Towards a thriving data-driven economy, Communication from the commission to the European Parliament, the council, the European economic and social Committee and the committee of the regions*, Brussels.
- Gossain, S., & Kandiah, G. (1998). Reinventing value: The new business ecosystem. *Strategy and Leadership*, 26, 28–33.
- IDC. (2011). *IDC’s worldwide big data taxonomy*.
- Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, 52, 36–44. doi:[10.1145/1536616.1536632](https://doi.org/10.1145/1536616.1536632).
- Kim, H., Lee, J.-N., & Han, J. (2010). The role of IT in business ecosystems. *Communications of the ACM*, 53, 151. doi:[10.1145/1735223.1735260](https://doi.org/10.1145/1735223.1735260).
- Koenig, G. (2012). Business ecosystems revisited. *Management*, 15, 208–224.
- Laney, D. (2001). *3D data management: Controlling data volume, velocity, and variety*. Technical report, META Group.
- Loukides, M. (2010). What is data science? O’Reily Radar.

- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, p. 156.
- Mike 2.0. (2014). *Big data definition – Mike 2.0*.
- Moore, J. F. (1993). Predators and prey: A new ecology of competition. *Harvard Business Review*, 71, 75–86.
- Moore, J. F. (1996). *The death of competition: Leadership and strategy in the age of business ecosystems*. New York: HarperCollins.
- Moore, J. F. (2006). Business ecosystems and the view from the firm. *Antitrust Bulletin*, 51, 31–75.
- NESSI. (2012). *Big data: A new world of opportunities*. NESSI White Paper.
- O'Riaín, S., Curry, E., & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems*, 13, 141–162. doi:[10.1016/j.accinf.2012.02.002](https://doi.org/10.1016/j.accinf.2012.02.002).
- Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library and Archives Journal*, 1, 1–3.
- Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: Free Press. doi:[10.1182/blood-2005-11-4354](https://doi.org/10.1182/blood-2005-11-4354).
- Rayport, J. F., & Sviokla, J. J. (1995). Exploiting the virtual value chain. *Harvard Business Review*, 73, 75–85. doi:[10.1016/S0267-3649\(00\)88914-1](https://doi.org/10.1016/S0267-3649(00)88914-1).
- Stonebraker, M. (2012). What does 'big data' mean. *Communications of the ACM, BLOG@ ACM*.
- Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. *Ecology*, 16, 284–307.
- Wikipedia. (2014) Big data. Wikipedia article. http://en.wikipedia.org/wiki/Big_data