Keamanan Big Data didalam Hadoop

Hariyanto

Magister Ilmu Komputer, Universitas Budi Luhur
Jl. Raya Ciledug, Petukangan Utara, Kebayoran Lama, Jakarta Selatan 12260
E- mail: harimeku@gmail.com

Abstrak — Data bisa memberikan manfaat, jika ditangani dengan baik. Big data dalam hal ini – juga harus mendapatkan pengelolaan yang baik, termasuk dalam hal keamanan. Framework Hadoop yang ada pada saat ini; banyak digunakan oleh perusahaan – perusahaan raksasa dunia, dan tak ingin ketinggalan dalam hal keamanan. Serangan dan ancaman seperti Denial-of-Services (DoS) dan Cross-Site scripting (XSS) yang banyak dilakukan oleh hacker dan cracker ke area big data. Dan tool – tool seperti Apache KNOX, Apache Sentry, Apache Ranger, Project Rhino, Kerberos dapat melakukan pencegahan dan perlindungan terhadap big data. Sehingga vulnerability yang ada di framework hadoop dapat terlindungi.

Kata kunci: Hadoop, Big data, Security and Privacy

Abstract — Data can provide benefits, if handled properly. Big data in this case – must also get good management, including in terms of security. The current Hadoop framework; widely used by giant companies in the world, and do not want to be left behind in terms of security. Attacks and threats such as Denial-of-Services (DoS) and Cross-Site scripting (XSS) are mostly carried out by hackers and crackers in the big data area. And tools like Apache KNOX, Apache Sentry, Apache Ranger, Project Rhino, Kerberos can prevent and protect big data. So that the vulnerabilities in the hadoop framework can be protected.

Keyword: Hadoop, Big data, Security and Privacy

I. PENDAHULUAN

Kemajuan teknologi saat ini, mendorong terciptanya data yang begitu besar. Sangking besarnya data yang ada saat ini, di setiap sendi kehidupan dapat dengan mudah kita temukan.

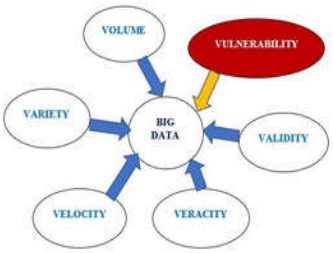
Data yang begitu besarnya menciptakan sebuah istilah yang dinamakan Big Data. Sedangkan Big Data sendiri dapat diartikan jumlah data yang begitu besar yang tidak terstruktur dan semi – struktur dengan lima karakteristik : volume, variety, velocity, veracity dan validity.

Adapun lima karakteristik tersebut dapat kita jabarkan seperti dibawah ini [1] :

- Volume : volume disini mengacu pada tingginya jumlah data yang dihasilkan dari berbagai sumber.
- Variety : Keragaman data disini mengacu pada jenis data yang dikumpulkan dari berbagai sumber seperti media sosial, pemerintah, layanan kesehatan dan lain - lain. Dalam berbagai format yang berbeda, seperti: audio, video, teks, log fle, file dan lain sebagainya.
- Velocity : Kecepatan disini tidak hanya mengacu pada kecepatan pengumpulan data yang besar tetapi juga kecepatan analisis data serta dapat menemukan informasi yang berharga.
- Veracity : Kebenaran yang ada di big data merupakan keabnormalan, noise, dan bias di dalam data tersebut. Untuk menghindari anomali dalam privasi pengguna, maka noise ini harus dikonfigurasi terlebih dahulu.
- Validity : Validasi di big data berarti data harus benar dan akurat bagi pengguna, sehingga data yang valid dapat membantu dalam membuat keputusan yang tepat.

Istilah Big Data diciptakan oleh Charles Tilly dalam kamus Oxford pada tahun 1980. Data yang ada pada hari ini, merupakan hasil dari berbagai sumber – diantaranya : situs media sosial, sensor jarak jauh, sinyal GPS ponsel, catatan transaksi dan log file[14].

Pertumbuhan data yang sangat signifikan menyebabkan ada masalah baru yang tidak hanya terkait dengan volume, variety, velocity, veracity dan validity. Tetapi ada penambahan "V" baru yaitu vulnerability seperti terlihat pada Gambar 1.



Gambar 1. Karakteristik "V" baru yaitu vulnerability

Kerentanan (vulnerability) ini merupakan issue yang cukup menjadi perhatian di dalam big data. Adanya ancaman – ancaman pencurian data menjadi fokus utama dalam big data, yang sebelumnya tidak begitu berpengaruh.

Untuk dapat mengekstrak nilai dari big data, maka diperlukannya kerangka kerja / framework yang dapat menganalisis lebih cepat dari pada tool analitik sebelumnya. Framework Hadoop merupakan alat yang mudah untuk menyimpan sebuah proses data dalam volume yang besar, serta menyediakan akses berkecepatan tinggi di dalam aplikasi. Hadoop digunakan oleh industri besar seperti Google, Yahoo, Facebook, dan lain – lain [15].

Teknologi framework yang telah dikembangkan ini, memastikan adanya privasi dan keamanan data yang berbeda dari teknologi sebelumnya. Karena semakin banyak para pengguna menggunakan teknologi ini untuk menyimpan dan memproses data pribadi mereka, maka bisa menjadi target serangan data yang signifikan [14].

II. METODOLOGI

II.1. Hadoop

Hadoop adalah framework open source berbasis Java dibawah lisensi Apache yang pokok utamanya bekerja pada komputasi terdistribusi dan konsep pemrosesan secara paralel untuk operasi batch. Awal mula munculnya hadoop terinspirasi dari makalah tentang Google MapReduce dan Google File System (GFS) yang ditulis oleh ilmuwan dari Google, Jeffrey Dean dan Sanjay Ghemawat pada tahun 2003. Proses pengembangannya dimulai pada saat proyek Apache Nutch, yang kemudian baru dipindahkan menjadi sub-proyek hadoop pada tahun 2006. Hadoop sendiri dirilis pertama kali pada tahun 2011 untuk mendukung distribusi pada proyek mesin pencari Nutch di Yahoo [14].

Hadoop sendiri terdiri dari kombinasi beberapa komponen diantaranya HDFS untuk penyimpanan / storage , MapReduce untuk memproses data dan YARN untuk memanage resource di dalam cluster [14] seperti yang terlihat pada Gambar 2.

Gambar 2. Ekosistem Hadoop

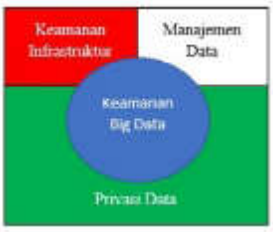
- HDFS
- Hadoop Distributed File System (HDFS) adalah sebuah file sistem berbasis java yang fault – tolerant, terdistribusi dan scalable. Data tersimpan di dalam cluster yang terdiri dari banyak node komputer / server yang masing – masing sudah terinstall hadoop. File – file yang besar dibagi menjadi ke dalam bentuk potongan (chunks) file yang lebih kecil yang berukuran 64kb dan tersimpan di dalam node – node yang tersebar di cluster. Walaupun data tersebut tersebar, pengguna tetap dapat melihat data tersebut.
- MapReduce
- MapReduce merupakan sebuah model programming / algoritma untuk pengelolaan data dalam skala besar dengan komputasi secara terdistribusi dan parallel yang dimana di dalam cluster itu sendiri terdiri dari ribuan komputer. Dalam prosesnya MapReduce terbagi menjadi tiga bagian yaitu Map, Shuffle dan Reduce. Namun dalam prakteknya antara shuffle dan reduce digabung menjadi 1 langkah yaitu Reduce.
- YARN
- YARN merupakan sebuah platform resource – management yang bertanggung jawab untuk mengelola resources dalam clusters dan scheduling. Serta memisahkan Job Tracker / Task Tracker dalam beberapa entitas :
    - Global ResourceManager di node master, berfungsi untuk mengatur semua resource yang digunakan oleh aplikasi di dalam sistem.
    - ApplicationMaster di setiap aplikasi, berfungsi untuk negosiasi resource dengan ResourceManager dan kemudian bekerja sama dengan NodeManager untuk mengeksekusi dan memonitor tasks.

- NodeManager di Agen – Framework setiap node slave, yang bertanggung jawab terhada container, dengan memantau penggunaan resource dari container (cpu, memori, disk, jaringan) dan melaporkannya pada ResourceManager.

II.2. Vulnerability

Vulnerability adalah kelemahan sistem baik secara prosedur, desain, implementasi atau kontrol internal. Kerentanan ini dapat terjadi secara tidak sengaja atau memang disengaja untuk dapat masuk pada celah keamanan yang ada.

Kerentanan ini dapat dibagi menjadi tiga kategori yaitu: keamanan infrastruktur, privasi data, dan manajemen data [14].



Gambar 3. Kategori kerentanan

- Keamanan Infrastruktur

  Jika berbicara tentang keamanan infrastruktur, kita dapat melihat lebih jauh kepada teknologi dan framework yang digunakan, terkait pengamanan dari arsitektur sistem big data itu sendiri. Framework hadoop, merupakan salah satu dari bagian keamanan infrastruktur, yang bisa dikatakan untuk saat ini, hadoop merupakan sinonim dari big data.

- Privasi Data

  Privasi data merupakan hal yang paling utama yang menjadi sorotan orang awam, karena di dalam big data berisi sejumlah informasi pribadi yang digunakan oleh organisasi untuk memperoleh manfaat dari data tersebut. Organisasi seharusnya tidak memiliki kebebasan untuk menggunakan informasi itu tanpa sepengetahuan pengguna.

- Manajemen Data

  Fokus pada bagian ini adalah pada apa yang harus dilakukan setelah data terdapat di dalam lingkungan big data. Tidak hanya bagaimana cara mengamankannya, tetapi juga berbagi data.

III. Hasil dan Pembahasan

III.1. Vulnerability Hadoop

Masalah keamanan infrastruktur, banyak dibahas dalam literatur - dimana menitik beratkan kepada keamanan di hadoop. Karena secara de facto banyak perusahaan yang menggunakannya pada lingkungan big data.

Serangan ancaman yang sering terjadi di hadoop biasanya adalah serangan Denial-of-Service (DoS) dan CROSS-SITE scripting (XSS). Yang mana dapat dilihat pada Tabel 1. Laporan serangan framework hadoop.

Tabel 1. Laporan serangan framework hadoop

Tahun	Total Vulnerability	DoS	XSS
2011	44	15	7
2012	63	19	6
2013	74	25	9
2014	92	23	6
2015	57	19	5
2016	103	15	17
2017	217	29	22
2018	148	15	9
2019	16	1	14

- Denial-of-Service (DoS)

  Serangan DoS terjadi ketika resource tidak tersedia bagi pengguna yang berwenang. Serangan ini dilakukan dengan membanjiri target dengan traffic yang banyak, yang menyebabkan data menjadi crash. Secara umum, ada dua cara dalam serangan DoS : flooding services atau crashing services.

  Serangan flood terjadi ketika sistem mendapatkan traffic yang berlebihan di buffer, yang akhirnya menghentikan traffic itu sendiri. Serangan flood yang sering dilakukan diantaranya : distributed denial of service (DDoS), buffer overflow, SYN flood, Ping to Death, and HTTP flood.

  Di Hadoop sendiri, Name Node dan server otentikasi sangat rentan terhadap serangan DoS. Name Node memiliki master daemon yang bertanggung jawab dalam penjadwalan dan mengkoordinasikan pelaksanaan aplikasi MapReduce pada node data. Serangan DoS ke Name Node dapat menghentikan semua perhitungan MapReduce dan operasi baca – tulis HDFS.

- CROSS-SITE scripting (XSS)

  XSS adalah serangan yang dilakukan dengan cara menyuntikkan kode berbahaya ke dalam aplikasi web yang retan. Berbeda dengan SQL injeksi, SQL injeksi secara langsung menyerang aplikasi web yang digunakan oleh pengguna, sedangkan XSS tidak. Kalau SQL injeksi, si pengguna aplikasi weblah yang berisiko.

  Ada dua jenis serangan XSS : stored / disimpan dan reflected / dipantulkan. Serangan XSS stored atau yang lebih dikenal dengan XSS persistent itu lebih berbahaya dari pada XSS reflected, dimana cara kerjanya dilakukan langsung menyuntikkan kode jahat ke aplikasi web yang rentan. Sedangkan XSS reflected cara kerjanya jika script yang tertananm di dalam sebuah link tertentu, jika terklik maka serangan itu akan aktif. Jadi perbedaan antara XSS persistent dan XSS reflected adalah jika XSS persistent itu dilakukan secara direct, sedangkan XSS reflected terjadi jika ada action dari pengguna itu sendiri.

  Dan di hadoop, UI webnya sangat rentan terhadap serangan ini, dan baru – baru ini, berbagai instalasi basis data telah diserang.

III.2. Keamanan di Hadoop

Ada beberapa tindakan yang dapat dilakukan untuk menutup celah dari kerentanan ini, dan bisa dikatakan bahwa tidak semua sistem yang ada pada saat ini terlepas dari aman, paling tidak ada tindakan untuk meminimalisir dari serangan yang ada.

Untuk menutup celah kerentanan ini, di hadoop terdapat tool yang dapat memberikan perlindungan di dalam cluster hadoop itu sendiri. Adapun tool tersebut yaitu :

- Apache KNOX

  Apache Knox Gateway merupakan titik akses pertama dari beberapa cluster hadoop dengan dasar konsep framework proxy stateless reverse. Knox juga menyediakan autentikasi kepada kelompok pengguna yang sudah terautentikasi, terotorisasi, teraudit dan terpantau oleh sistem. Knox menyembunyikan data serta menyembunyikan perincian instalasi cluster hadoop.

  Menyederhanakan jumlah layanan yang harus dipindahkan oleh pengguna, serta membatasi jumlah titik akses melalui URL dengan satu pintu. Knox memiliki sistem keamanan berparameter berbasis REST API yang dapat mengautentikasi pengguna credentials terhadap AD / LDAP. Hanya pengguna yang berhasil diautentikasi yang boleh diizinkan mengakses cluster hadoop.

- Apache Sentry

  Apache Sentry merupakan granular, otorisasi yang berbasis pada peran dan modul administrasi multitenant untuk hadoop. Sentry dapat memanage akses ke data dan metadata dengan memberikan tingkat hak istimewa dengan sangat akurat kepada pengguna dan aplikasi yang sudah diautentikasikan di dalam kelompok hadoop.

  Ini merupakan standar luar biasa yang mendukung otorisasi untuk beragam model data di hadoop. Keserbagunaan ini dapat mendefinisikan aturan otorisasi guna memvalidasi permintaan akses pengguna atau aplikasi ke resource hadoop.

  Maksud adanya Sentry dapat menjadi mesin otorisasi yand dapat dipasang pada elemen hadoop, seperti Apache Hive, Apache Solr, Impala, dan HDFS.

- Apache Ranger

  Apache Ranger menyediakan framework terpusat yang mengamankan hadoop. Ranger merupakan sistem otorisasi yang dapat menolak akses ke resource cluster hadoop (file HDFS, tabel Hive, dan lain – lain). Ranger juga mendukung pengguna yang sudah diautentikasikan sebelumnya.

  Ketika permintaan dari pengguna datang ke Ranger, maka bisa diasumsikan bahwa permintaan tersebut telah dikonfirmasi. Apache Ranger menggunakan Kerberos guna mengautentikasi dan Apache Knox untuk otorisasi : role-based access control (RBAC). Apache Knox juga mendukung audit HDFS, Hive, dan Hbase, serta Apache Ranger menggunakan wire encryption untuk melindungi data.

- Project Rhino

  Project Rhino memberikan perlindungan data yang bertumpuk di hadoop dengan sistem konsep masuk tunggal / single-sign-on (SSO), otorisasi umum – modul manajemen otentikasi dan dukungan terhadap enkripsi. Rhino meningkatkan enkripsi di cell serta akses kontrol yang halus ke Hbase 0.98 dan enkripsi ke data-at-rest di Apache hadoop. Enkripsi data di hadoop membutuhkan data-at-rest dan data-in-transit; namun, sebagian besar komponen hadoop menyediakan enkripsi untuk transit data saja.

- Kerberos

  Kerberos merupakan protokol autentikasi yang dikembangkan oleh MIT, digunakan untuk memberikan autentikasi kepada pengguna jika ingin mengakses cluster hadoop.

  Protokol Kerberos menggunakan kunci kriptografi rahasia untuk komunikasi yang aman melalui jaringan yang tidak aman. Kerberos merupakan sistem berbasis tiket SSO yang bergantung pada KDC.

  Protokol kerberos dengan menghasilkan tiga jenis tiket autentikasi : (1). Delegasi token merupakan kunci rahasia antara pengguna dan Name Node untuk autentikasi, (2). Token blok akses digunakan untuk mengakses file HDFS yang diautentikasi oleh Name Node dan Data Node yang secara bersamaan mengakses blok data pada Data Node, (3). Token job yang mengenerate JobTracker untuk mengautentikasikan tugas di TaskTrackers. KDE Kerberos terdiri dari tiga komponen : (1). Autentikasi server, (2). Ticket-granting server (TGS), dan (3). Database.

IV. KESIMPULAN

Dalam perkembangannya, data menjadi sesuatu hal yang sangat bernilai atau bisa dikatakan jadi barang komoditi. Keamanan menjadi syarat mutlak yang harus dipersiapkan dalam ruang lingkup big data. Perusahaan – perusahaan yang menggunakan big data, dituntut memberikan sebuah keamanan yang bisa dipertanggung jawabkan, baik itu secara privasi data maupun secara infrastrukturnya.

Hadoop sebagai salah satu framework big data - yang banyak digunakan, memberikan sebuah terobosan dalam hal keamanan. Dengan adanya komponen yang tertanam di dalam hadoop, memberikan kemudahan dalam pengolahan data. Dan dengan tool – tool yang mendukung keamanan di hadoop, membuktikan bahwa keamanan terukur di dalam big data – membuat hadoop patut diperhitungkan.

Dengan tool – tool tersebut dapat mencegah serta melindungi big data dari serangan – serangan hacker maupun cracker. Dan pada akhirnya big data mendapatkan tempat yang aman dalam posisi yang terbaik.

V. DAFTAR PUSTAKA

[1] B. Renu, H. Vaibhav, and J.A. Neelu, “Big Data Security – Chanllenges and Recommendations”, International Journal of Computer Sciences and Engineering, 2016, Vol.04, Issue, 01, E-ISSN: 2347-2693.

[2] B.R. Babak, A. Nafisseh, A. Pouya, and K. Yasaman, “Security and Privacy Challenges in Big Data Era”, International Journal of Control Theory and Applications, International Science Press, 2016, Vol.09, No.43.

[3] S. Shraddha, “BIG DATA SECURITY”, International Journal of Current Research, 2016, Vol.09, Issue, 05, pp.50320 – 50325.

[4] A. Pelin, B. Bhargava, and R. Rohit, “Big Data Analytics for Cyber Security”, Hindawi, 2019, Vol. 2019, Article ID 4109836.

[5] L. Lixiang, O. Kaoru, Z. Zonghua, and L. Yuhong, “Security and Privacy Protection of Social Networks in Big Data Era”, Hindawi, 2018, Vol. 2018, Article ID 6872587.

[6] S. Krzysztof, W. Liqiang, L. Xiangyang, and Y. Dengpan, “Big Data Analytics for Information Security”, Hindawi, 2018, Vol. 2018, Article ID 7657891

[7] A. Sahel, A. Feras, H. Ismail, and G. Gheorghita, “An Effective Classification Approach for Big Data Security Based on GMPLS/MPLSS Networks”, 2018, Vol. 2018, Article ID 8028960.

[8] Z. Dongpo, “Big Data Security and Privacy Protection”, International Conference on Management and Computer Science (ICMCS 2018), 2018, Vol. 77.

[9] Z. Gang, “Big Data and Information Security”, International Journal of Computational Engineering Research (IJCER), 2015, Vol. 05, Issue, 06.

[10] K. Murat, and F. Elena, “Research Challenges at the Intersection of Big Data, Security and Privacy”, frontiers, 2019.

[11] P. Joseph. Carles, I. Carol, S. Mahalakshmi, “Big Data Security an Overview”, International Research Journal of Engineering and Technology (IRJET), 2018, Vol. 05, Issue. 02.

[12] P. Nandhini, “A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce, Int. Journal of Engineering Research and Application, Vol. 08, Issue 04, pp. 65 – 78.

[13] P. Yusuf, “The Hadoop Security in Big Data : A Technological Viewpoint and Analysis”, International Journal of Scientific Research in Computer Science and Engineering, 2019, Vol. 7, Issue 3, pp. 1 – 14.

[14] S.B. Gurjit, S. Amardeep “Big Data : Hadoop framework vulnerabilities, security issues and attacks”, Journal, Elsevier, 2019.

[15] K. Gayatri, A. Agrawal, and K. Ahmad, “Big Data Security challenges : Hadoop Perspective”, International Journal of Pure and Applied Mathematics, 2018, Volume 120, No. 6, 11767 – 11784.

[16] Ninny Bhogal, Shaveta Jain,”A Review on Big Data Security and Handling”, International Research Based Journal, 2017, Vol.6, Issue 1.

[17] Mohammed S.Al-Kahtani, “Security and Privacy in Big Data”, International Journal of Computer Engineering and Information Technology, 2017, Vol. 9, No. 2.

[18] Er. Prachi Jain, and Er. Alisha Gupta, “A brief review on Hadoop architecture and its issues”, International Journal of Engineering Research and General Science, 2017, Vol. 5, Issue 2.

[19] A. Pradeep, S.D. Srikari, and Z. Xiaowen, “Hadoop Eco System for Big Data Security and Privacy”, IEEE LISAT, 2015

[20] Jain. P,”Security issues and their solution in cloud computing”, International Journal of Computing and Business Research, 2012