Механика Транспорт Комуникации

ISSN 1312-3823 (print)
ISSN 2367-6620 (online)
том 19, брой 2, 2021 г.
Научно списание **http://www.mtc-aj.com** **статия № 2075**

СИСТЕМА ЗА ОБРАБОТКА НА ГОЛЕМИ ДАННИ В ТРАНСПОРТА

Любен Боянов
lboyanov@unwe.bg

доцент, доктор, УНСС,
Студентски град „Христо Ботев“, 1700 София
РЕПУБЛИКА БЪЛГАРИЯ

Ключови думи: *големи данни, системи за големи данни, транспорт, Internet of Things*

Резюме: *Развитието на информационните технологии (ИТ) през 21 век доведе до големи и революционни промени във всички сектори на живота. ИТ престана да бъде фактор, който влияе само в компютърните системи и мрежи или в мобилните комуникации. Дигиталните технологии навлязоха навсякъде със своите възможности да отчитат и предават данни от всеки обект. Този модел и технология станаха известни като Интернет на обектите (Internet of Things) – свързани хетерогенни обекти в Интернет. Източници на данни от тази парадигма могат да бъдат и транспортните средства и системи, които отчитат и следят дейности в тази област. Погледнато глобално, дигиталните данни създадоха свят, където с голяма скорост се генерират големи количества, много разнообразни по вид и произход данни. Тяхното значение е огромно, защото могат да спомогнат за различни видове анализи и оптимизации на събития и процеси. Данните могат да помагат на бизнеса да повиши ефикасността на процесите си и да открие нови тенденции. Заедно с повсеместното създаване на данни се породи необходимостта от развитие и внедряване на системи за прием, обработка и анализ на тези данни. Настоящата работа представя една такава система, която може да се приложи в много области, включително и в транспорта. Системата е изградена от блокове и модули, които са скалируеми, с отворен код и се интегрират лесно с най-известните програмни продукти от средата Hadoop - системи за големи данни. Описаната система е тествана и представя резултати от пътни събития в района на Лос Анджелис, които са събирани в продължение на три месеца. Системата позволява събиране, анализ и визуализация и на други видове големи данни от областта на транспорта.*

ВЪВЕДЕНИЕ

Краят на 20 век се характеризира с бързото навлизането на персонални компютри в домове и офиси и проникването на Интернет до всяко работно място и място за живот. От интересни и удобни средства за програмиране, въвеждане на текст и игри, компютрите станаха незаменима част от нашето ежедневие, която служи за огромно множество дейности. Заедно с Интернет и последвалото разпространение на мобилните телефони, дигиталните устройства и мрежи започнаха да се прилагат и

използват навсякъде. Това доведе до още по-голямо предлагане и създаване на още повече устройства, като с най-голямо значение за текущата дигитална трансформация на нашата Земя беше създаването и използването на миниатюрни сензори и дигитални идентификатори (RFID). Те са в основата на новото повсеместно свързване - Интернет на Обектите (ИнО), известен с оригиналното си име - Internet of Things (IoT). Технологиите за идентификация и проследяване, сензорни и безжични мрежи, облачните технологии и разпределеното събиране, предаване и обработка на информацията доведе бавно, но сигурно до качествено нова среда за живот и бизнес. ИнО бързо завладя науката, изследователските дейности и внедряванията, като резултатите не закъсняха. Те бяха в различни области на знанието и технологии като най-първи бяха телекомуникации, информатика, електроника и социални науки [1]. ИнО създаде редица „умни“ и „интелигентни“ приложения сред които се появиха редица такива разработки в областта на транспорта [2].

Интернет на обектите започна да генерира големи данни с високи скорости от много места. Появиха се три, четири и пет „V"-та (V3, V4, V5), с които най-често се характеризират големите данни. Това са обем, скорост, разнообразие, достоверност и стойност (volume, velocity, variety, veracity и value). Големите данни породиха нови големи предизвикателства, свързани с тяхното извличане, пренасяне, приемане, съхраняване, обработка и анализ [3]. Големите данни също бързо намериха приложение в различни сектори като логистика [4], финанси [5], транспорт [6] и др. Използването на големи данни за оптимизация и анализ на транспортните системи води до повишаване на тяхната автономност, интелигентност, свързаност и до нови възможности за преодоляване на проблемите по пътища и в градската среда. Интелигентните транспортни системи осигуряват ефективна интеграция и сътрудничество между различните технологии и услуги. Обработката на големи обеми от данни предоставя ефективни, безопасни и удобни решения за множество дейности в транспорта [7].

ПЛАТФОРМИ И ПАКЕТИ ЗА ОБРАБОТКА НА ГОЛЕМИ ДАННИ

С развитието на ИнО и създаването и разпространението на Големи данни се наложи да се създадат и развият нови средства и технологии за тяхното предаване, съхранение и обработка. Принципно данните се приемат и обработват в изчислителни центрове (създадени с тази цел) или ИТ отдели, разположени във фирми или организации. В първите години на новото хилядолетие започна развитието на нов вид услуги за работа с компютърните приложения - облачните услуги (cloud services). Днес те са масово разпространени. Парадигмата на облака стана много популярна по редица причини, най-важните от които са мащабируемост, заплащане за ползване, намаляване на разходите за хардуер и софтуер на мястото на използването и др. Облачните системи предлагат *инфраструктура като услуга* (IaaS) - сървъри, хранилища, мидълуер, виртуализации, мрежи и др. Те също така могат да предоставят и *платформи като услуга* (PaaS) - среди за разработване, тестване, управление на софтуерни приложения и др. Друг клас облачни услуги са *програми като услуги* (SaaS) - софтуерни програми или стек от приложения, управлявани от доставчика на облачни услуги. Облаците могат да предоставят и *контейнери като услуга* (CaaS) - виртуализация, базирана на контейнери (вид софтуерен пакет и работа). Важно предимство на облачната парадигма е, че тя може да реагира на бързи изисквания на клиентите и не изисква от последните умения и ресурси. Сред недостатъците на облачните услуги е, че данните се съхраняват и обработват отдалечено и собственикът на данните няма директен и пряк достъп до тях, което може да доведе до “изтичане“ или „кражба“ на данни от облаци [8].

Друг подход за съхранение и обработка на данни е споменатият по-горе – в центрове и/или във фирма и организация. Днес, когато става въпрос за обработка на големи данни най-често се споменава Hadoop. Софтуер, който е използван от повече от половината от 50-те компании в класацията Fortune [9]. Hadoop е софтуерна библиотека с отворен код (може да се изтегли и види от всеки), която позволява разпределена обработката на големи масиви от данни с прости модели за програмиране [10]. Тази система може да мащабира задания от единични сървъри до хиляди машини, при паралелна обработка и съхранение на данните. Фондацията Apache, която поддържа и развива Hadoop има повече от 300 проекта от този вид с около 2 петабайта (10^{15}) изтеглен изходен код [11]. Един от най-известните и разпространени модули на Hadoop е разпределената файлова система HDFS, работеща на принципа Map-Reduce [12]. Други известни и разпространени пакети на фондацията, които се използват широко от разработчици и потребители на големи данни са NiFi, Apache Kafka, Apache Hive, Apache Impala, Apache Spark, Hue, Apache Hive, Apache Impala и др. [11].

СИСТЕМА ЗА ОБРАБОТКА НА ГОЛЕМИ ДАННИ

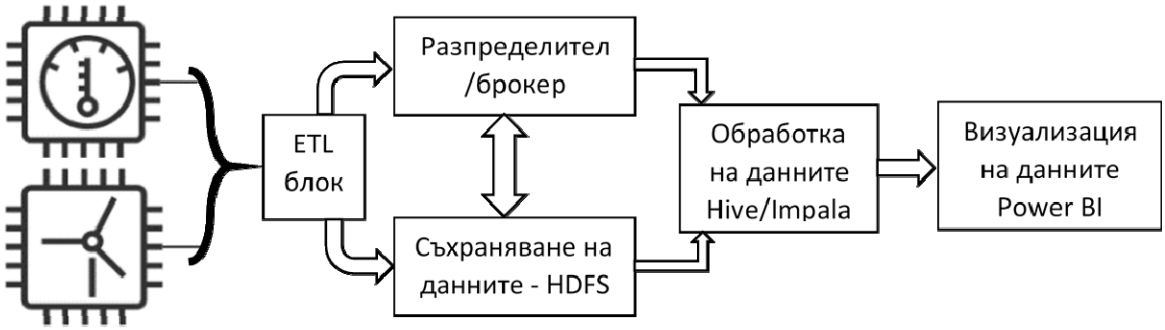
Когато се проектират системи за обработка на данни е важно да се определят минималния брой функционални блокове, около и с които да се изгради дадената система. На първо място трябва да източник на данните. От него ще следват извличане, съхранение, обработка и анализ. Извличането на данните може да става чрез интернет протокол, връзка точка-към-точка (напр. директна сателитна или друга комуникационна връзка) или мобилна мрежа. Източникът на данни може също така да бъде уебсайт. След извличането (или изпращането), данните постъпват в мястото, където е разположена платформата за големи данни. Нейните основни задачи са да трансформира, зарежда, съхранява, обработва и предоставя анализи на данните. Това са основните функционални етапи, необходими за постигане на резултати, които да служат за приложението и крайните потребители.

Първият блок на системата е ETL блока (Extract, Transform and Load). Нашият избор за този модул е Apache NiFi. При него, за улеснение на програмистите е наложена концепцията, базирана на възможност за графична обработка на потоци данни. Всеки поток от данни може да се трансформира и/или пренасочва към друга част на системата.

Вторият блок е брокерът (разпределител на данните по теми). Нашият избор за този модул е Apache Kafka, който е един от най-популярните брокери, използвани в системи за големи данни.

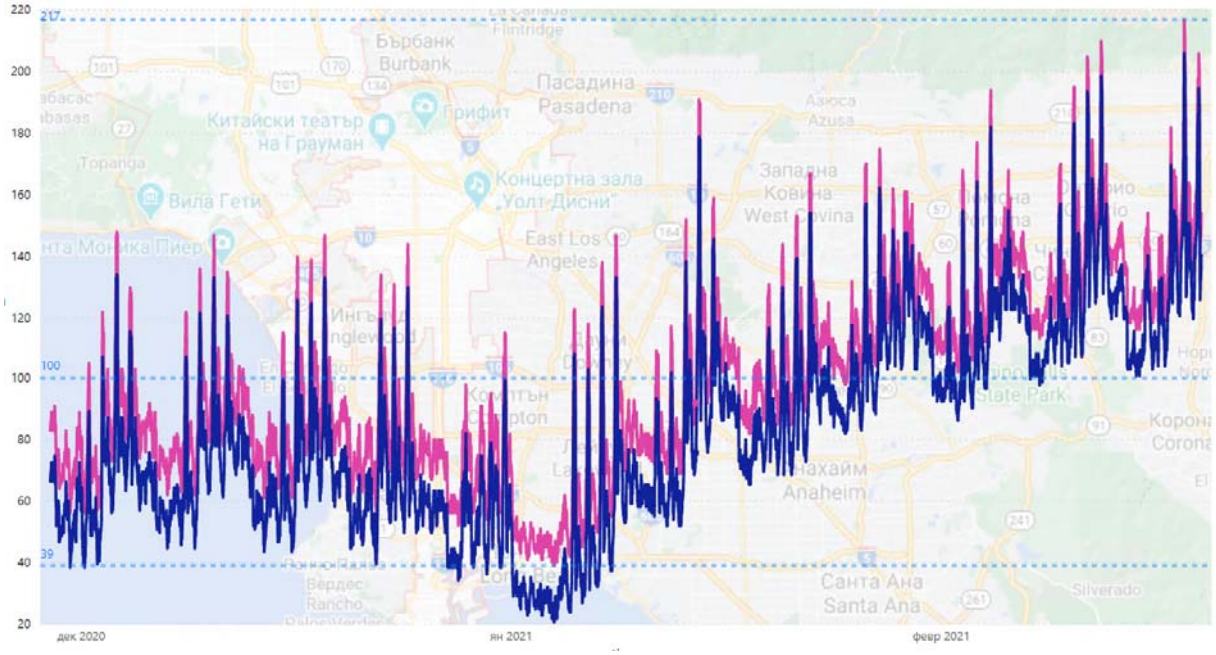
Третият блок е за съхранение на данните – в случая използваме Hadoop HDFS. Следват блокове за обработка на данни – базите данни Hive и Impala, които се използват съответно при по-бавна или по-бърза обработка на данните.

Последният блок е Power BI, с който се извършва визуализацията на резултатите, след тяхната обработка в предния блок. Блоковата структура на тази система е показана на Фигура 1. Системата е реализирана в Центъра за компетентност за обработка на големи данни в УНСС, разработен по проект „BG05M2OP001-1.002-0002 „Дигитализация на икономиката в среда на Големи данни“ (ДИГД)“. Същият разполага с 40 възела, всеки с 128 GB RAM памет и обща памет за постоянно съхранение на данни – 4.5 PBytes.



Фиг. 1. Блокова структура на системата за обработка на данни

За проверка и верификация на работата на системата за обработка на данни са използвани данни от областта на транспорта. Те се предлагат от платформата MapQuest (https://developer.mapquest.com), където в раздел „Трафик“ може да се получат данни в реално време за инциденти по пътя, ремонтните работи и събитията по пътната мрежа в района на Лос Анджелис (ЕлЕй) в САЩ. Транспортните данни, отнасящи се за пътни инцидентни и ремонтни дейности на пътната мрежа на ЕлЕй се зареждат на всеки 10 минути. Това става, като системата за обработка на данни изпраща заявка до платформата на MapQuest. След получаване на данни от заявката, същите се обработват и съхраняват в блока за данни - Apache Hadoop (HDFS). Тези стъпки се повтарят след 10 минути. Данните преминават през различните модули, показани на Фигура 1. Изследванията обхващат период от три месеца – между ноември 2020 и февруари 2021. Бяха създадени и запазени над 12 хиляди записа от транспортни данни от гореспоменатия източник. Фигура 2 представя визуализацията на събраните и отчетени инциденти от пътната мрежа на ЕлЕй за горепосочения период. По абсцисата са посочени дните в трите месеца, а по ординатата – броят пътни инциденти (между 20 и 220).



Фиг. 2. Инциденти по пътната мрежа на Лос Анджелис (м. 10.2020-м. 02.2021)

В червено са дадени инцидентите, обработени с базата данни Hive, а със синьо – с базата данни Impala. Както се вижда, характерът и динамиката на двете графики са еднакви, а разликата между тях се дължи на това, че при Impala зареждането на данни трае около 5 секунди (зареждат се по-ниски стойности), докато при Hive това е около 35 секунди (зареждат се по-точни стойности). От там идва и разликата в стойностите при двете отчитания.

На графиката на Фигура 2 ясно се вижда, че събитията по пътната мрежа имат седмичен характер/цикъл. Пиковете на инцидентите са през уикендите, когато хората предприемат повече пътувания. В периодите след коледните и новогодишните празници има спад на събитията, поради фактът, че голямата част от хората са били в домовете си. Вижда се и плавното покачване на инцидентите, свързано с увеличаване на трафика през първите месеци на годината, свързано с разхлабване на мерките от Ковид-19 кризата. Получените резултати показват, че предложената система е подходяща за обработка и анализ на данни от транспортни събития.

ЗАКЛЮЧЕНИЕ

Съвременните транспортни системи са динамични и сложни, като в същото време са източник на големи данни, идващи от сензори, идентификатори, уеб сайтове и други източници, които показват брой транспортни инциденти, ремонти по пътя и т.н. За обработката на подобни събития е важно да има системи за обработка на големи данни, които са гъвкави и с възможност за разширяемост.

Представената такава система е тествана в Центъра за обработка на големи данни в УНСС с данни от областта на транспорта. Използвани са събития от пътната мрежа на Лос Анджелис в продължение на три месеца, до които има свободен достъп. Показана е пълната работоспособност на системата и възможностите за визуализация и анализ на десетки хиляди транспортни събития. Системата показа стабилна работа за данни, постъпващи в реално време от реални транспортни събития, като същата е в състояние да обработи още по-големи и комплексни данни.

БЛАГОДАРНОСТ

Тази работа е подпомогната от резултати, получени при разработки свързани с проект №BG05M2OP001-1.002-0002 "Дигитализация на икономиката в среда на големи данни", финансиран от Оперативна програма "Наука и образование за интелигентен растеж" 2014-2020 г., процедура BG05M2OP001-1.002 "Изграждане и развитие на центрове за компетентност".

ЛИТЕРАТУРА:

[1] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010, doi: 10.1016/j.comnet.2010.05.010.

[2] V. Chand and J. Karthikeyan, “Survey On The Role Of IoT In Intelligent Transportation System,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, pp. 936–941, Sep. 2018, doi: 10.11591/ijeecs.v11.i3.pp936-941.

[3] С. К. Йорданова С., “Предизвикателствата на големите данни – същност, характеристики и технологии,” vol. Научни трудове на УНСС, том 1/2019, Издателски комплекс УНСС, София, 2019, pp. 17–24.

[4] N. Dragomirov, “Big data in Logistics – definition and sources,” Jan. 2015.

[5] Big Data Finance, “Big Data in Finance: Use Cases, Examples, Challenges, and Getting Started,” 2019. https://bigdatafinance.tw/index.php/finance/1214-big-data-in-finance-use-cases-examples-challenges-and-getting-started (accessed May 03, 2021).

[6] K. Govindan, T. C. E. Cheng, N. Mishra, and N. Shukla, "Big data analytics and application for logistics and supply chain management," *Transportation Research Part E: Logistics and Transportation Review*, vol. 114, May 2018, doi: 10.1016/j.tre.2018.03.011.
[7] B. Belezamo, S. Eken, and C. Avcı, *Special Issue on Big Data in Transportation*. 2020.
[8] G. R. Mettu and D. A. Patil, "Data breaches as top security concern in cloud computing," *International Journal of Pure and Applied Mathematics*, vol. 119, pp. 19–27, Jan. 2018.
[9] CEO, "Top 5 Best Big Data Tools," *The CEO Views*, Aug. 10, 2020. https://theceoviews.com/top-5-best-big-data-tools/ (accessed Feb. 26, 2021).
[10] "Apache Hadoop." https://hadoop.apache.org/ (accessed May 06, 2021).
[11] "Apache Projects List," Feb. 25, 2021. https://projects.apache.org/projects.html?category#big-data (accessed Feb. 25, 2021).
[12] Apache Foundation, "MapReduce Tutorial," 2021. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (accessed Aug. 03, 2021).

SYSTEM FOR BIG DATA PROCESSING IN TRANSPORT

Luben Boyanov
lboyanov@unwe.bg

Associate Professor, PhD, UNWE,
Hristo Botev Students' Town, 1700 Sofia
THE REPUBLIC OF BULGARIA

Key words: *Big data, Systems for Big data processing, transport, Internet of Things*

Abstract: *The advances of Information Technology (IT) in the 21st century have caused major and revolutionary changes in all sectors of human activities. IT has stopped being a factor only in computer systems and networks or in mobile communications. Digital technologies have penetrated everywhere with their capabilities to create and transmit data from any object. This model and technology are known nowadays as the Internet of Things (IoT) - connected heterogeneous objects on the Internet. Data sources from this paradigm can also be vehicles and transportation systems that report and monitor activities in the transport sector. Digital data created a world where big volumes of diverse data are being generated at great velocity. This created another phenomenon – Big data. The importance of Big data was quickly realized as it allows better and innovative type of analysis and optimization. Collected and processed data helps improving business operations and increases the efficiency all kind of practices. Such activities have been closely related to the ubiquitous data creation, which in turn gave rise to systems, capable of extracting, saving, processing, visualizing and analyzing Big data. This work presents one such data processing system that can be applied in many fields, including transportation. The system is constructed from blocks and modules that are scalable, open source, and integrate easily with the most popular and best well-known software products from the Hadoop environment. The presented modular system has been tested and its results demonstrate the applicability for transport data, taken from traffic events in the Los Angeles area. The data has been collected over a period of three months and consists of tens of thousands events. The system also allows for the collection, analysis, and visualization of other types of big data in the transport sector.*