DOI: https://doi.org/10.37075/ISA.2022.1.05

Използване и обработка на големи данни в транспорта – платформи и подходи

Любен Боянов*

Резюме: Настоящата статия сближава темите за извличане, предаване, обработка, използване и анализ на големи данни в областта на транспорта. Тя посочва някои аспекти на приложението на Интернет на обектите и големите данни в сферата на транспортните услуги. Представени са източниците на големи данни в транспортния сектор, като се прави най-обща класификация според мястото на техния произход. Показани са основните протоколи за обмен на данни, както и различни видове облачни платформи, които могат да служат за приемане, съхранение и обработка на данни. Направено е кратко описание на функциите на транспортните контролни центрове, като са посочени някои проблеми, свързани с приемането, съхранението и анализа на големи данни. Представен е подход за разширяване дейността на транспортен контролен център, при което се дава възможност за негова съвместна работа по извличане, обработка и анализ на данни с подходящи средства и адекватна облачна платформа за работа с големи данни.

Ключови думи: транспорт, големи данни, контролни центрове, извличане и обработка на данни.

JEL: C81, O180, R41, R42, O360.

* Любен Боянов е доктор, доцент в катедра „Информационни технологии и комуникации“ на УНСС.

Увод

Създаването на нови, разнообразни, с ниска цена и лесно вграждани сензори, които намират място във всички области на живота (вкл. транспортния сектор), както и разработването и прилагането на дигитални мрежови технологии доведоха до появата и широкото приложение на Интернет на обектите (ИнО, или известно и като Интернет на нещата – Internet of Things – IoT). Множество приложения от ИнО станаха източници на дигитални данни, които започнаха да се генерират и съответно – извличат и използват от хората в техните ежедневни дейности и действащ бизнес. Огромното разнообразие и масовото прилагане на тези технологии доведоха до друг модерен феномен в областта на информационните технологии и дигитализиращия се свят – големите данни. Двете парадигми вече коренно променят екосистемата на Земята, икономиката и други сфери на човешка дейност и тази промяна набира все повече скорост, което пък от своя страна поражда ново по-масово прилагане на ИнО и използване на новогенерирани големи данни (Боянов, 2021).

Областта на транспорта рано възприе и прилага много активно ИнО и големите данни. Организиране и оптимизиране на маршрути, проследяване на пратки, наблюдение на товарите, ефективно използване на складови пространства, предсказваща поддръжка на транспортни средства, отдалечено

управление на транспортни дейности, наред с други, са вече съществуващи и прилагащи се подходи от ИнО и анализа на големи данни (Hussein и др., 2018). Умни транспортни системи, управление на критични ситуации, установяване на аномалии, намаляване на риска от замърсяване на въздуха и по-малко транспортни произшествия предлагат огромни възможности в бъдещето за по-безопасен и устойчив транспорт (Ge et al., 2018; Николова, 2017).

Целта на настоящата статия е да представят различните източници на данни в областта на транспорта, някои от използваните протоколи за извличане и предаване на данни, да се класифицират най-общо видове данни, платформи и среди за тяхната обработка, както и да се посочат някои техни проблеми. В края се представя подход за допълнение към транспортен контролен център, който може да се ползва за извличане, обработка и визуализация на големи данни в този сектор.

Методология на изследването

За разглеждане на източниците и подходите за извличане на данни се обръща внимание на произхода на данните, които в сферата на транспорта могат да бъдат от много различни места. Такива има в самите транспортните средства, в създадената и изграждаща се инфраструктура на транспорта, както и от спомагателни на транспорта средства и приложения.

Източниците на данни (по принцип, не само в транспорта) генерират различни по своята организация данни – структурирани, неструктурирани и полу-структурирани, които се предават с помощта на локални или глобални протоколи за обмен на данни. Направен е кратък преглед на видовете данни и на протоколите за тяхното пренасяне.

След генерирането и предаването данните попадат в облачни и/или сървърни среди или в контролни центрове за обработка и реакция. На тези места или при предаването или извличането се осъществяват действия като агрегиране, почистване, компресиране, комбиниране, филтриране, оценяване, индексиране, отчитане на прагови стойности, търсене, поточна обработка и др. Основните видове обработки в облак стават в обществени, частни или хибридни области, както и в транспортни контролни центрове за управление и споделяне на данни. Отбелязани са някои недостатъци на контролните центрове, които могат да бъдат компенсирани от частен или публичен облак за обработка на големи данни.

В края е представен пример за извличане и обработка на транспортни данни от платформа/облак, която може да се използва в допълнение на контролен център, като примерът е с данни от реални източници на транспортни събития и показва възможностите за ефективно използване и анализ на големи данни в този сектор.

Източници на данни в транспорта

На първо място може да се разгледат данните, които се генерират от сензори и уреди в самото транспортно средство. Примери за такива големи данни са 60 GB данни, генерирани за час от автономна кола на Гугъл, 30 TB данни, генерирани от Боинг 777 по време на трансатлантически полет (International Transport Forum, CPB, 2015), а разположените около 300 сензора във всяка кола от Формула 1 генерират над милион „точки“ телеметрични данни в секунда, предавани от автомобилите към боксовете, което прави стотици терабайти данни за всяко едно състезание (Miller, 2021).

Според някои автори смарт картите и автоматизираните данни са двата големи източника на данни, които най-често се използват за анализи на обществения транспорт. В тези случаи те служат най-често за разбиране на поведението на по-

требителите при пътуване и за оценка на качеството на предлаганите услуги (Welch and Widita, 2019).

Все по-развиващата се транспортна инфраструктура с въвеждането на „умни" транспортни системи, предлагащи решения на редица проблеми в големите градове, е също значим пример за извличане и използване на големи количества данни в транспорта (Serrano, 2019). Умната инфраструктура (както в градовете, така и извън тях) помага да се следи, анализира, осъществяват комуникации и вземат решения на базата на данни от сензори, камери и идентификатори. На този етап това се прави в три направления:

а) събиране и съхраняване на данните, които служат за по-нататъшно разглеждане и анализ;

б) събиране и обработка на данните, които постъпват при оператор, който взема решения (разпространено в наши дни);

в) събиране и автоматична обработка на данните, на базата на които дигитални системи вземат решения и влияят на транспортни събития и сценарии.

Една от добре известните крайни цели в транспортния сектор е пълната автоматизация на транспорта, при което човешкият фактор (с неговите грешки и субективни решения) ще има минимално влияние, което ще осигури ефективност, сигурност, безопасност и устойчивост (Bartuska и Labudzki, 2020) – непрекъснато преследвани цели в областта на транспорта. Добре известно е, че автономните коли имат най-голям брой вградени сензори, радари, сонари, GPS, лидари и камери, като всички те генерират до 4 TB данни всеки ден (Intellias, 2019).

Терминът „големи данни" се отнася за данни, които са в голям обем, пристигат с голяма скорост и имат голямо разнообразие – т.нар. V^3 (volume, velocity, variety). Обемът се счита за голям (макар и това да се мени през годините), когато е GB на минута или час и TB на час или ден, но тези префикси се менят динамично и вероятно след няколко години такива обеми няма да се разглеждат като големи данни. Скоростта е свързана с честотната лента на протоколите или тяхната максимална скорост на обмен на данни, като тя зависи от това дали данните се предават по жичен протокол (тогава са хиляди или милиони мегабайти за секунда) или безжичен (десетки или хиляди мегабайти за секунда). Разнообразието на данните е свързано с различните източници, но най-общо може да се посочат три вида данни – структурирани, неструктурирани и полу-структурирани.

Структурираните данни са тези, които имат ясно изразени връзки и зависимости и се съхраняват в релационна система бази данни. При такава база от данни, връзките между тях трябва да бъдат създадени преди създаването на съдържанието. Както подсказва самото име, неструктурираните данни нямат ясно изразена структура. Те не могат да се съхраняват в редове и колони, както при релационните бази данни. Примери за неструктурирани данни са снимки, видеоматериали и архивирани в папки с файлове документи. При полу-структурираните данни няма отделно описание на вида или структурата на данните и не се изисква предварително дефинирана схема. Такава е възможна, но не е задължителна. Типичен пример за полу-структурирани данни са тези, описани с XML (eXtensible Markup Language) – език за представяне и обмен на данни в WWWeb. Освен споменатите по-горе примери на големи данни от сензори в транспортни средства (те са към структурираните данни, генерирани и предавани с високи скорости), може да се споменат и данни от камери, радари, лидари и сонари (неструктурирани данни – големи обеми).

Големите данни дават възможност за анализи, с помощта на които се предоставят възможности за справяне с много проблеми и предизвикателства (като на-

маляване на себестойността, повишаване на сигурността, предоставяне на нови функционалности, проактивна поддръжка, намаляване на разхода на гориво/енергия, прогнозиране на потреблението и т.н.), пред които са изправени производителите на автомобили и други превозни средства. Анализът на събраните големи данни от превозното средство и пътната инфраструктура, както и от маршрутите на превозното средство може да предложи множество решения, но в същото време изисква знания и опит в сферата на транспорта, към които да се приложат експертизи от сферата на статистиката и информационните технологии. В същото време е важно да се отчете фактът, че големите данни имат голяма стойност и са обект на рискове за киберсигурността, което може да породи заплахи за откуп от недоволни служители или хакери. Големите данни могат да помагат, но могат и да вредят, поради което към тях трябва да се подхожда с комплексни знания и умения.

Протоколи и платформи за споделяне и обмен на данни

Протоколите за предаване на данни в транспорта може да са локални, регионални или глобални. Сред локалните протоколи са тези, които се отнасят за събиране и обмен на данни в рамките на превозното средство или в непосредствена до него близост. Сред тези протоколи (обикновено са с малък обхват) и мрежи за обмен на данни може да отличим CAN, FlexRay, Modbus, NFC, RFID, Bluetooth и др. Регионалните (със среден обхват на обмен на данни) протоколи, системи и архитектури са VANET, MANETs, WiFi, DSRC, LoRA и др. Те най-често служат за обмен на данни между близко движещи се (разположени) превозни средства и/или комуникация с близо намираща се транспортна инфраструктура. От глобалните протоколи и средства за предаване и приемане на данни може да се посочат мобилните или т.нар. клетъчните мрежи. Известни в тази категория са 2G, 3G, 4G и 5G (на която се възлагат големи надежди поради ниското закъснение на сигнала и високата честотна лента), сателитните комуникации (GPS), като съществуват и други, макар не така широко разпространени глобални мрежи за обмен на данни.

Най-популярните платформи за споделяне и обмен на данни са облачните структури (clouds). Облаците (или облачният компютинг) предоставят различни видове дигитални услуги като процесорно време, памет, мрежови услуги, виртуални машини, контейнери, и др. При тях потребителите не знаят къде са разположени съответните използвани от тях ресурси, но те имат достъп от всяко място, свързано с Интернет, сателитна или клетъчна мрежа (а и те почти винаги имат връзка помежду си). При това потребителят заплаща на базата на използваните, заявени от него/тях услуги или ресурси. Облачният компютинг се разглежда в три насоки/вида – публичен, частен и хибриден. Публичните облаци са тези, които са достъпни от всеки регистрирал се в облака потребител и може да използва облачните услуги от всяко място. Думата „публичен“ не означава, че всеки има достъп до данните или ресурсите на друг потребител, а просто, че достъпът може да е за всеки регистрирал се в облака. Частният компютинг/облак е в рамките на частна организация и достъп до него имат само хората от фирмата или от друга организация, на която специално е предоставен достъп до частния облак. Хибридните облаци ползват както публичен, така и частен облак – в първия се разполагат ресурси, които са с по-ниска степен на поверителност, докато във втория – такива, които са с по-висока степен на защита.

Обикновено облачните услуги се прилагат в големи центрове за данни, които имат широколентов (т.е. с висока пропусна

способност и съответно – скорост) достъп до Интернет. Облакът предлага големи изчислителни възможности и умерено или ниско време за реакция, както и огромно (за потребителя – повече от необходимото) количество памет за данни. Това прави облачните услуги мащабируеми (т.е. потребител може да иска да ползва още и още ресурси и те се предоставят веднага), а също така освобождава крайния потребител от обслужване, поддръжка, обновяване на хардуер и софтуер и др.

Друг подход в транспортния сектор е използването на Контролни центрове (КЦ – Control Centers, Traffic Control Centers, Transportation Management Centers, Traffic Operations Centers). Те имат за цел ефективно да разпределят ресурсите и да интегрират управлението на транспортните мрежи, като предоставят динамични и полезни услуги за потребителите. В своята работа КЦ могат да използват дигитални интелигентни приложения и решения за управление на трафика и за реакция при инциденти и извънредни ситуации. Такива приложения и решения са: сензори за откриване на превозни средства; видеонаблюдение; знаци с променящи се съобщения; измервателни уреди на рампа; предупредителни радиосъобщения по магистрали; подсистеми за управление на метеорологичните условия по пътищата (RWMS); различни видове софтуер и телекомуникационни системи, включително комуникации между центрове. Един от проблемите на тези центрове е, че при тяхното създаване те не са били проектирани за събиране и обработка на големи данни, каквито се генерират в днешно време в транспортния сектор. Днес обемът от данни от превозни средства и транспортна инфраструктура расте с всеки изминат ден, а КЦ нямат функционалностите на традиционните облаци, които се мащабират добре, нито пък имат сериозни възможности за аналитична обработка на големи данни. В наши дни е важно подобни контролни центрове да използват и подходите за обработка на данните в края (Edge computing) и в мъглата (Fog computing), с което ще се подобри ефективността и ефикасността на КЦ, но и при този подход остава проблемът със средствата и възможностите за обработка на големи данни.

Един възможен подход е организиране на съвместна работа на КЦ с частен или обществен облак (създавайки в последния случай хибриден облак), при което КЦ ще подава транспортни данни на облака (или пък избрани транспортни данни ще отиват паралелно към центъра и облака), при което облакът ще има възможност да обработва и анализира големите данни, без това да се отразява на основните дейности на УЦ.

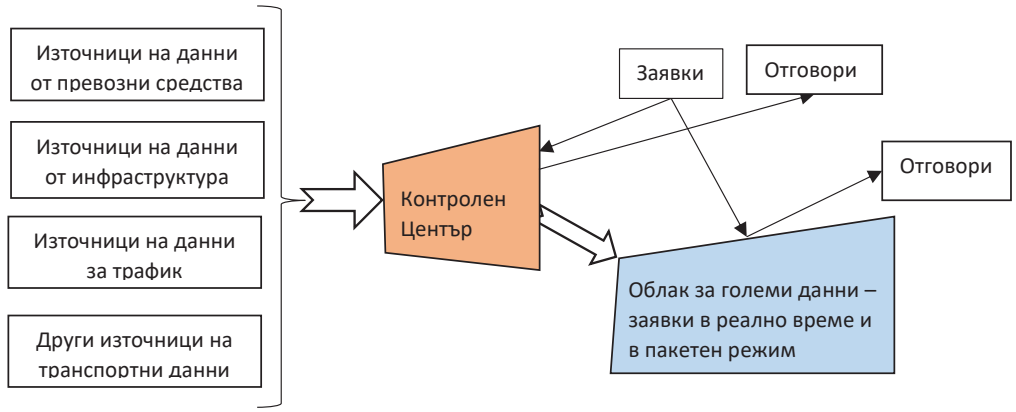
Приложни приноси и насоки за развитие

В УНСС е създадена платформа (частен облак) за работа с големи данни, която се състои от 40 възела, които имат възможност да приемат и обработват големи масиви от данни в реално време или в пакетен режим. Платформата има и големи възможности за съхраняване на данни (над 4 петабайта). Тя е изградена със софтуер с отворен код, създаван или поддържан от фондацията Apache (Apache Foundation, 2021). В центъра на тези софтуерни библиотеки и програми е най-известният продукт за работа с големи данни – Hadoop и неговата файлова система HDFS (“Apache Hadoop”, 2021). Облакът извлича и извършва почистване и първична обработка на данни с помощта на Apache NiFi, извършва разпределение (насочване) на пристигащите потоци по различни „теми“ (задания) от Apache Kafka, създава таблици от данни за обработка в пакетен режим с Apache Hive и таблици за обработка на данни в поточен режим с Apache Impala. Данните се съхраняват в Hadoop HDFS, а визуализацията и анализът се извършват с Power BI. По-подробно

описание на системата може да се намери в (Boyanov, 2021) и (Христов, 2021).

Тази платформа може да бъде приложена към контролен център за транспорт, като общата архитектура на управляващата система с допълващия я частен облак/платформа за обработка на големи данни е показана на фигура 1.

от JSON към TEXT формат. В този процес се извличат определени полета от JSON – броя текущи събития по пътната мрежа. Обработените и почистени данни от NiFi се подават към Apache Kafka, след което отиват към HDFS за съхранение като файлове. Следващата стъпка е от HDFS данните да се заредят/обновят в Hive и

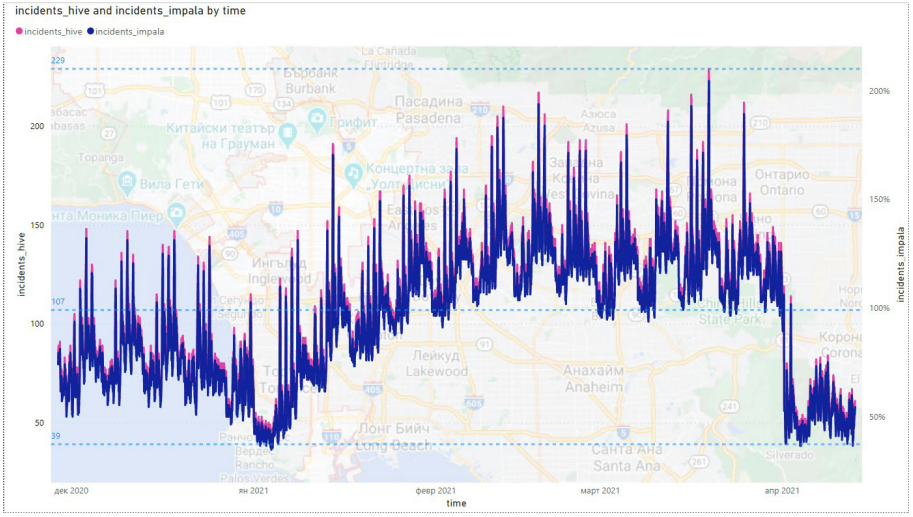


Фигура 1. *Извличане и обработка на големи данни в транспорта*

За тестване на подхода беше използвана американската онлайн услуга MapQuest (MapQuest, 2021), която освен редица други услуги предоставя и данни за трафик в реално време, като в конкретния случай се извличат данни за пътнотранспортни произшествия в даден район. За достъп до данните в MapQuest е необходимо да се създаде безплатен профил за разработчик. В този профил се създават един или повече различни ключове за интерфейс за програмиране на приложения с уникален идентификатор. За целите на тестването е избрана област от пътната мрежа на средната западна област на САЩ с диаметър от около 100 километра. Програмата Apache NiFi, като използва съответния уникален ключ за интерфейса, извършва повикване към MapQuest, след което приема данните и ги обработва. При това действие се извършва почистване на данните и конвертиране

Impala, а накрая се прави визуализация на получените данни в PowerBI. Резултатите от извличането и обработката на данните са показани на фигура 2.

На фигура 2 се вижда графиката на пътнотранспортните инциденти за период от пет и половина месеца, като ясно си личат периодите с по-малко инциденти – около коледните и великденските празници, както и през уикендите. Инцидентите, обработвани в платформата с Apache Hive, са изобразени с червена линия на графиката, а тези с Apache Impala – със синя. Първите се обработват в пакетен режим (поради което са с по-голяма точност), а вторите – в поточен режим (поради което са с по-малка точност, но пък се появяват по-бързо). Извършеният експеримент демонстрира възможността на предложения подход да обработва и големи данни, каквито възможности нямат (в общия случай) контролните центрове за транспорт.



Фигура 2. *Извличане на данни от MapQuest и обработката им в среда за големи данни*

Заключение

Обемът и скоростта, с които днес се генерират, предават и съхраняват данни от всички области на живота, включително и от транспортния сектор, са безпрецедентни. Сензори, идентификатори, карти, изображения от камери, GPS данни, брой и местонахождение на пътнотранспортни произшествия, повреди на инфраструктурата, ремонт на съоръжения и пътища – всичко това създава огромни количества данни, които при подходяща обработка от платформа за такива данни може да предостави нови, качествено по-добри възможности за подобряване на ефективността и безопасността в този сектор. Обединяването или съвместното използване на контролни транспортни центрове с облаци/платформи за големи данни могат да направят процеса на използване и анализ на данни от транспортни събития сравнително лесно реализуем и приложим. Статията демонстрира функционалност на една такава платформа (облак) за транспортни данни, които може да се генерират или препредават от контролен център. Платформата е изградена със софтуер с отворен код, което я прави сравнително лесно реализуема за изследователски или университетски центрове за обработка на данни. Платформата може да се разшири и със средства/софтуер с Изкуствен интелект (такива има в пакетите Apache Spark, Apache Marvin-AI и Apache Spot), с помощта на които да се правят предвиждания или да се предлагат решения, каквито хората трудно могат да предложат поради огромния обем данни, които постъпват в платформата.

Цитирани източници:

Боянов, Л., 2021. Дигиталният свят – промяната, Глобалната дигитална трансформация – обогатяване или обедняване на човечеството. София, изд. „Авангард Прима“.
(Boyanov, L., 2021. Digitalniyat svyat – promyanata, Globalnata digitalna transformatsia – obogatyavane ili obednyavane na chovechestvoto. Sofia, izd. „Avangard Prima“)

Николова, Х., 2017. Приложение на интелигентните транспортни системи за устойчиво развитие на транспорта. *Научни трудове на УНСС,* 1/2017, 77–109.
(Nikolova, H., 2017. Prilozhenie na inteligent-

nite transportni sistemi za ustoychivo razvitie na transporta. *Nauchni trudove na UNSS,* 1/2017, 77–109)

Христов, Я., 2021. Метод за оценка и избор на референтна архитектура за Интернет на Обектите. Дисертационен труд, УНСС. (Hristov, Ya., 2021. Metod za otsenka i izbor na referentna arhitektura za Internet na Obektite. Disertatsionen trud, UNSS)

Apache Foundation, 2021. The Apache Software Foundation [WWW Document]. Welcome to The Apache Software Foundation! URL https://apache.org/ (accessed 5.6.21).

Apache Hadoop [WWW Document], 2021. URL http://hadoop.apache.org/ (accessed 2.15.21).

Bartuska, L., R. Labudzki, 2020. Research of basic issues of autonomous mobility. Transportation Research Procedia, LOGI 2019 - Horizons of Autonomous Mobility in Europe 44, 356–360. https://doi.org/10.1016/j.trpro.2020.02.031

Boyanov, L., 2021. Financial data processing in Big data platforms. UNWE Publishing Complex, *Economic Alternatives* under print.

Ge, M., H. Bangui, B. Buhnova, 2018. Big Data for Internet of Things: A Survey. Future Generation Computer Systems 87, 601–614. https://doi.org/10.1016/j.future.2018.04.053

Hussein, W., L. Kamarudin, H. AL-Hashimi, A. Zakaria, R.B. Ahmad, N.A.H. Binti Zahri, 2018. The Prospect of Internet of Things and Big Data Analytics in Transportation System. *Journal of Physics:* Conference Series 1018, 012013. https://doi.org/10.1088/1742-6596/1018/1/012013

Intellias - Intelligent Software Engineering, 2019. How Big Data in Autonomous Vehicles Defines the Future [WWW Document]. Intellias. URL https://intellias.com/how-big-data-in-autonomous-vehicles-defines-the-future/ (accessed 11.2.21).

International Transport Forum, CPB, 2015. Big Data and Transport, Understanding and assessing options.

MapQuest, 2021. Official MapQuest - Maps, Driving Directions, Live Traffic [WWW Document]. URL https://www.mapquest.com/ (accessed 11.4.21).

Miller, R., 2021. How Cloud Data-Crunching Accelerates the F1 Racing Experience [WWW Document]. Data Center Frontier. URL https://datacenterfrontier.com/how-cloud-data-crunching-power-accelerates-the-f1-racing-experience/ (accessed 11.2.21).

Serrano, W., 2019. Big Data in Smart Infrastructure.

Welch, T.F., A. Widita, 2019. Big data in public transportation: a review of sources and methods. *Transport Reviews* 39, 795–818. https://doi.org/10.1080/01441647.2019.1616849