КАКВО ПРЕДСТАВЛЯВАТ ГОЛЕМИТЕ ДАННИ?

Николай Бизев[1]

Резюме

Според световния компютърен гигант IBM през 2012 г. човечеството е произвело 2.8 зета байта дигитална информация. За сравнение, един зета байт е равен на 1000 байта, повдигнати на седма степен. Според експертите тези данни могат да запълнят 57.5 милиарда устройства iPad с памет от 32 гига байта (GB). Запълнените устройства могат да послужат за изграждането на две Китайски стени[2].

Натрупването на дигитална или дигитализирана информация през последното десетилетие е най-скоростното в историята на човечеството досега. Бурното развитието на Интернет от края на 20-ти век и появата на „ултра“ популярните социални мрежи като Facebook, Tweeter, Google + и др. дадоха допълнителен тласък на създаването, натрупването и складирането на огромни масиви от данни. Освен в компютърните системи, произвеждаме данни от сензори, датчици, камери, мобилни телефони, фотоапарати, сканиращи устройства, медицинска апаратура, спътници и др. Всички тези видове данни попадат под един знаменател - т.нар. **големи данни.** Така, например, за извършването на един полет от Ню Йорк до Лос Анжелис самолет Боинг 737 създава данни в размер на 240 теребайта[3]. Автори като Виктор Шюнбергер и Кенет Кукиер[4] наричат явлението по създаване, събиране и анализиране на „големи данни“ „революция, която ще промени начина ни на живот, начина ни на мислене и работа“.

Все повече внимание на „големите данни“ обръщат институциите на международни организации и отделните държави. Европейската Комисия прие през 2014 г. стратегия за дигитализирането на Европа, като на „големите данни“ е отредено приоритетно място. В САЩ, президентът Обама вижда развитието на сектора като един от инструментите за изход от финансовата и икономическата криза от последните 5 години. Целта на настоящата статия е да предостави информация за дефиницията на понятието „големи данни“, неговите характеристики и ползи. За да може да отговори на тези очаквания, статията ще изследва въпросите: Какво са „големите данни“? Какъв е интересът за институционализиране на „големите данни“? Кои са успешните примери за употреба на „големи данни“?

WHAT IS BIG DATA?

Nikolay Bizev,
Senior expert, Institute of Public administration

Abstract

According to IBM, world leading innovation company, in 2012 the humanity has created 2.8 zeta bytes of digital information. For comparison, 1 zeta byte is equal to 1000 bytes raised to the power of seven (7). IBM experts are

---

[1] **Николай Бизев** е магистър по Европейски публични отношения от университета в Маастрихт и бакалавър специалност „Публична администрация“ от катедра „Публична администрация”, Философски факултет на Софийски университет „Св. Климент Охридски”. Основни области на научен интерес – прозрачност в публичното управление, отворени данни, големи данни, административна реформа. По-важни проекти, в които Николай Бизев е участвал са: Въвеждане на комплексно административно обслужване за предоставяне на качествени услуги на гражданите и бизнеса“, финансиран по Оперативна програма „Административен капацитет“, Администрация на Министерския съвет, 2013; Изграждане на капацитет на ИПА за изследвания, обучение и приложение на иновативни европейски практики в доброто управление, финансиран по Оперативна програма „Административен капацитет“, Институт по публична администрация, 2015 г.

[2] http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html

[3] https://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/

[4] Big Data: A Revolution that Will Transform how We Live, Work, and Think, 2013, Viktor Mayer-Schönberger Kenneth Cuki

convinced that these information can fulfill 57.2 billion of Ipad devices wit 32 GB memory. These devices will be enough to build the Chinese wall twice.

Ovr the last decade (2000-2010) the accumulation of digital or digitalized information is the fastest in the human history until now. The tremendous development of Internet in the late 20th century and the occurrence of "super" popular social networks like Facebook, Twitter, Google + and many others added extra boost to creation, accumulation and storage of massive amounts of data. Data can be produced not only by computer systems. We produce data though sensors, mobile devices, cameras, smart phones, scanning devices, medical equipment, satellites and etc. All this information can be characterized as "Big data". For example, a single flight from New York to Los Angeles operated by Boing 737 creates 240 terabytes of data. Viktor Mayer-Schönberger Kenneth Cuki, the authors of the bestseller book "Big Data: A Revolution that Will Transform how We Live, Work, and Think" share the opinion that the "big data" phenomenon a "revolution that will change the way we live, the way we think and work".

Big data causes interest in many international organisations and institutions. In 2014 the European Commission launched a strategy called "Digital Europe", including big data as an instrument for economic development. The president of US, Obama shared a communication with information that in the next 5 years US will invest heavily in state agencies and work on big data.

The purpose of this article is to examine issues related to the big data definition, its characteristics and дефиницията на понятието „големи данни", неговите характеристики и benefits. In order to find an explanation the article will answer the questios: What is big data? What is the institutional interest in the field? Are there a success stories in working with big data?

Какво представляват големите данни?

Според световния компютърен гигант IBM през 2012 г. човечеството е произвело 2.8 зета байта дигитална информация. За сравнение, един зета байт е равен на 1000 байта, повдигнати на седма степен. Според експертите тези данни могат да запълнят 57.5 милиарда устройства iPad с памет от 32 гига байта (GB). Запълнените устройства могат да послужат за изграждането на две Китайски стени[5].

Натрупването на дигитална или дигитализирана информация през последното десетилетие е най-скоростното в историята на човечеството досега. Бурното развитието на Интернет от края на 20-ти век и появата на „ултра" популярните социални мрежи като Facebook, Tweeter, Google + и др. дадоха допълнителен тласък на създаването, натрупването и складирането на огромни

[5] http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html

масиви от данни. Освен в компютърните системи, произвеждаме данни от сензори, датчици, камери, мобилни телефони, фотоапарати, сканиращи устройства, медицинска апаратура, спътници и др. Всички тези видове данни попадат под един знаменател - т.нар. **големи данни.** Така, например, за извършването на един полет от Ню Йорк до Лос Анжелис самолет Боинг 737 създава данни в размер на 240 теребайта[6]. Автори като Виктор Шюнбергер и Кенет Кукиер[7] наричат явлението по създаване, събиране и анализиране на „големи данни“ „революция, която ще промени начина ни на живот, начина ни на мислене и работа“.

Все повече внимание на „големите данни“ обръщат институциите на международни организации и отделните държави. Европейската Комисия прие през 2014 г. стратегия за дигитализирането на Европа, като на „големите данни“ е отредено приоритетно място. В САЩ, президентът Обама вижда развитието на сектора като един от инструментите за изход от финансовата и икономическата криза от последните 5 години. Целта на настоящата статия е да предостави информация за дефиницията на понятието „големи данни“, неговите характеристики и ползи. За да може да отговори на тези очаквания, статията ще изследва въпросите: Какво са „големите данни“? Какъв е интересът за институционализиране на „големите данни“? Кои са успешните примери за употреба на „големи данни“?

Явлението „големи данни“? Концепцията 4Vs + 1 extra

„Големите данни“, като явление все още нямат точна и безспорна дефиниция, която да обясни пълното им значение. Много учени и институции работят над темата и търсят свои определения. Трудностите за намиране на точна дефиция на „големите данни“ идват от постоянната промяна на явлението. Институтът Макинзи (McKinsey Global Institute) в своето изследване от 2011 г. „Големите данни: следващата граница за иновациите, конкуренцията и производителността“ дава следното определение:

„Масиви от данни, чиито размер е над възможностите на типичните софтуерни инструменти на база данни за събиране, съхраняване, управление и анализиране“ .

Данни:
1. Скоростта на натрупване
2. Обем
3. Разнообразие
4. Автентичност
5. Стойност

Експертите на института Макинзи предполагат, че с напредването на технологиите във времето, размерът на масиви от данни, които се определят като „големи данни“ ще се увеличи. Те отчитат и развитието на техниката, която би трябвало да увеличи възможността си за архивиране и анализиране на тези данни. Също така определението може да варира от сектора, в зависимост от това, кои са общодостъпни видове софтуерни инструменти и какви размери на масиви от данни са достъпни за определен отрасъл. Според експертите, „големите данни“ в

[6] https://gigaom.com/2010/09/13/sensor-networks-top-social-networks-for-big-data-2/
[7] Big Data: A Revolution that Will Transform how We Live, Work, and Think, 2013, Viktor Mayer-Schönberger Kenneth Cuki

много сектори варират от няколко десетки терабайта до няколко петабайта (в хиляди терабайта). Друга популярна дефиниция за „големите данни“ е така наречената:

„4 Vs (data velocity, data volume, data variety, data veracity)“

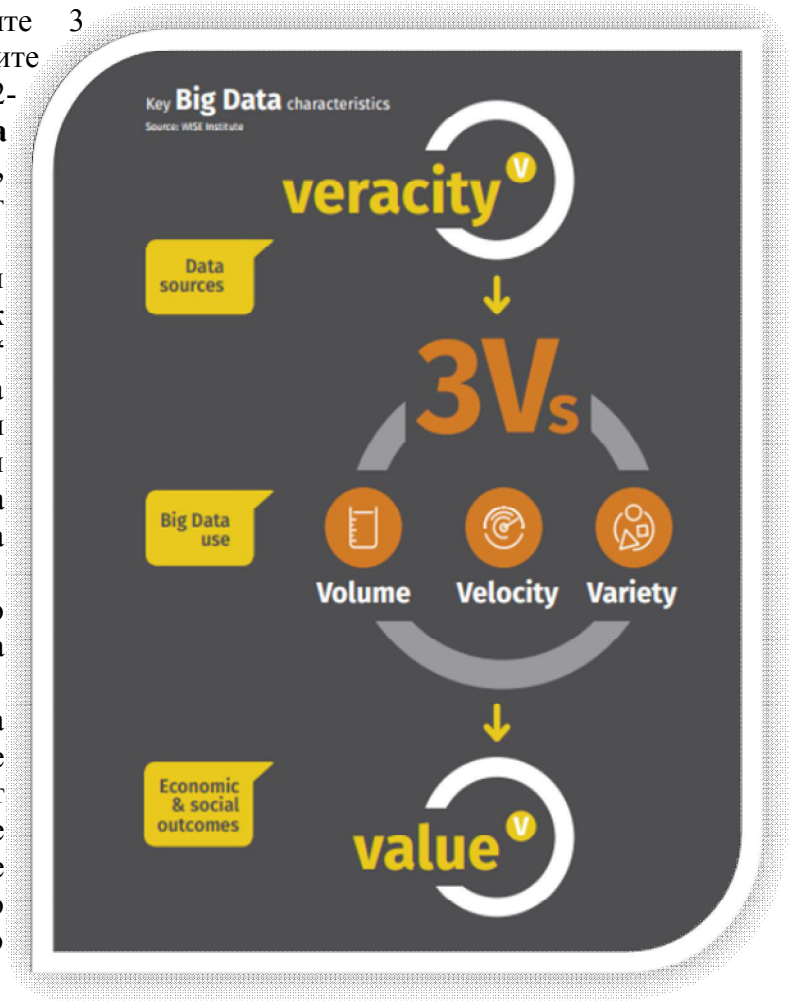
Към тази дефиниция се придържат организации като Организацията за икономическо сътрудничество и развитие (ОИСР), компанията IBM и др. Характеристиката 4Vs се отнася към скоростта на натрупване на данните, обема на данните, разнообразието и автентичността на данните. 4Vs по-точно се отнасят към:

Скоростта на натрупване (Velocity): „Големите данни“ се характеризират с висока скорост на натрупване, тъй като често се създават в реално време. Натрупване на данни, например, се предизвиква от данни, изпратени от електронно устройство (умен телефон) към сървър при употреба на мобилно приложение.

Обем на данните (Volume): Размерът на наличните данни, произведени от устройства, компютърни системи и други. Смята се, че в световен мащаб се произвеждат около 2.5 милиарда мегабайта на ден.

Разнообразие на данните (Variety): „Големите данни“ много често са неструктурирани. Те могат да бъдат под формата на снимки, музикални файлове, съдържание в интернет, видео файлове, здравна информация и др.

В допълнение на изброените 3 основни характеристики на „големите данни“ някои експерти добавят още 2- Автентичност **(Veracity), източника на данните** и Стойност **(Values), крайният продукт, резултат от данните.**



В изследването си „Големи и отворени данни: Двигател за растеж или пропуснати възможности“ Варшавският институт за икономически изследвания добавя тези две нови характеристики, за да изчисли потребността от натрупването на „големите данни“ в икономиката на ЕС.

Европейската Комисия също работи със своя дефиниция за явлението:

Терминът "големи данни" се отнася за големи количества от различни видове данни, произведени с висока скорост от голям брой различни видове източници. Обработването на данни е силно променливо, тъй като натрупването им се случва в реално

време. Това изисква нови средства и методи, като например мощни процесори, софтуер и алгоритми за управление и анализ[8].

Различните дефиниции все пак отчитат един общ елемент: именно големината на данните. Най-просто казано, „големите данни“ са всякакъв вид дигитална информация, която подлежи на компютърна обработка.

Какъв е интересът за институционализиране на „големите данни“?

Според Варшавския институт за икономически изследвания[9], „големите данни“ могат да донесат над 206 милиарда евро до 2020 г. за Европа като допълнителен брутен вътрешен продукт. Според изследователската институция този БВП е равен на увеличение от 1.9 % в реалната икономика на Европа.

Съгласно стратегическия документ ЕВРОПА 2020 за постигане на устойчив растеж в ЕС е необходимо да се направи преход към дигитализиране на споделеното пазарно пространство между държавите членки. Като инструмент за приближаване към устойчивия растеж Европейската комисия залага на „големите данни“ и ги припознава като приоритетна област. Макар че в сравнение със САЩ, ЕС изостава, институционализирането на явлението „големи данни“ вече е факт. В комюнике от 2 юни 2014 г. на Европейската Комисия до Съвета на ЕС, Парламента на ЕС и др. институции ЕК за първи път обръща внимание на „големите данни“ като инструмент за постигане на заложените в стратегията ЕВРОПА 2020 цели. Според изследване на ЕК, отделни държави членки като Обединеното кралство са увеличили работата си по „големите данни“ с 240 % спрямо 2013 г. Активността на ЕК по темата доведе до свикването на Съвет по икономическите въпроси на ниво министри в Брюксел през март 2015. Заключенията от Съвета затвърдиха позициите на „големите данни“ в дневния ред на европейските институции. Основно заключение от Съвета бе нуждата от провеждане на проучване по темата, за да се оценят реалните ползи от въвеждането на инициатива за „големите данни“. Изследвания като това на Варшавския икономически институт показват значителен ръст на БВП при въвеждането на употреба на „големите данни“ в редица сектори. Според ЕК „големите данни“ имат приложение в сектори като здравеопазване, храни, сигурност, изменението на климата и ефективното използване на ресурсите в енергетиката, интелигентни транспортни системи и интелигентните градове.

[8] Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions: Towards a thriving data-driven economy, {swd(2014) 214 final}
[9] http://www.bigopendata.eu/wp-content/uploads/2014/01/bod_europe_2020_full_report_singlepage.pdf

Администрацията на президента на Съединените щати обяви още през март 2012 г[10]., че ще финансира проекти на 6 държавни агенции в областта на „големите данни“ с бюджет от 200 милиона долара. Според президента „големите данни“ могат да спомогнат за „справяне с най-сложните предизвикателства, които стоят пред нацията“. Администрацията на президента на САЩ ще финансира програма за проучване и разработване на проекти, свързани с „големи данни“ в областта на медицината, отбраната, енергетиката, секторите на научни иновации и развитие на човешки ресурс, които да работят с „големи данни“. Инициативата „Големи данни“ в САЩ се ръководи от междуведомствена комисия, която управлява и следи изпълнението на проектите, които се финансират по програмата.

Кои са успешните примери за употреба на „големи данни“? Примерът на Google

През 2009 г. екип на компанията Google публикува статия „Откриване на грипни епидемии, използвайки данни на търсачката“[11] в световноизвестното изследователско списание Nature. В нея екипът описва механизъм за предсказване на епидемии от различни щамове на грипни вируси чрез анализ на данни от интернет търсачка на компанията. Заявките за търсене са обработени чрез статистически методи, в комбинация със сложни математически алгоритми за изчисление на зависимости: най-общо казано, анализ на „големи данни“.

Интернет търсачката Google получава над 3 милиарда единични търсения на ден. Всички те се архивират и запазват. Това позволява на компанията да създаде огромен архив от данни, който може да обработва. През 2009 г. Google пуска своя публикация в научното издание Nature, в която твърди, че чрез анализ на данни от търсенията на своите клиенти компанията може да предскаже къде и кога ще настъпи епидемия от зимен грип в САЩ. Механизмът, който Google използва на пръв поглед изглежда прост: анализаторите на

[10] Комюнике на президента на САЩ, Барак Обама от 29 март 2012 г във връзка с инициативата големи данни.
[11] Откриване на грипни епидемии, използвайки данни на търсачката за заявки, брой 457, Февруари 2009 г., Списание Nature

компанията правят проучване на търсенията в търсачката, свързани с вируса на грипа, симптомите и свързани думи. Те правят проучване на складираните данни в периода между 2003 г. и 2008 г. В допълнение към събирането на данни те се опитват да създадат връзка между търсенията в търсачката и данните за разпространение на вируси, събирани от центрове за контрол и превенция на заболявания в Съединените щати. Освен набора от думи, които са най-често търсени от клиентите на Google в периода на зимните грипове, екипът използва данни за разпространението на вирусите по територия и времетраенето на епидемиите. В резултат от работата на експертите в Google са създадени 45 ключови думи, проверени в 50 милиона търсения. Математически алгоритми изчисляват възможността грипа да се появи на точно определено място, в точно определено време на база на търсенията в интернет търсачката. През 2009 г. интернет компанията представя продукт, който може да предскаже къде и кога ще избухне грипна епидемия въз основа на анализ на „големи данни“. Продуктът на интернет компанията Google представлява приложение към интернет браузъра Chrom.

Приложението следи в реално време разпространението на грипни вируси и е в състояние да предвиди избухване на грипна епидемия. В резултат от това Националните центрове за превенция на заболяванията в САЩ проявяват интерес към разработката и сътрудничат при набирането на информация. По този начин те предоставят допълнителен ресурс на Google, за да допълни своя съществуващ. Към момента Центърът за превенция на заболявания в САЩ работи заедно с екип на интернет компанията Google за предсказване на грипни епидемии. Резултатите от взаимната работа на компанията и администрацията е достъпно интернет приложение, което по всяко време може да предскаже разпространение на вирус в САЩ и да представи информация на повече от 30 езика.

Примерът на Google от 2009 г. доказва, че големите данни имат огромно практическо приложение. Изборът на сфера като здравната доказва, че „големите данни“ реално могат да се използват в много различни сектори, стига да има достатъчно голям набор дигитална или дигитализирана информация за изготвяне на анализ. „Големите данни“ всъщност поставят огромно предизвикателство пред модерните системи за мониторинг и анализ. Те дават възможност за предсказване, до голяма степен предричане на евентуални бъдещи състояния чрез анализ на огромни потоци от информация. За частния сектор и бизнеса това предсказване на бъдещо състояние със сигурност е положително.

„Големите данни“ могат да бъдат ползвани в системи за планиране. Както в много случаи бизнесът вероятно по-бързо ще усвои практиката от държавата. Тук идва въпросът какво може да прави държавата с „големите данни“? Къде е ролята на държавата в този процес на развитие? Дали трябва да се намеси като регулатор и да определи докъде да стига инициативата за проучване и използване на данни или да стане проводник на информация като отвори достъп до своите масиви, за да стимулира икономическото развитие. Вече е ясно, че САЩ инвестират усилено в сектора. ЕС до някаква степен е осъзнал нуждата като от 2014 г. постави политически приоритет на това явление. Дебатът по темата остава отворен. Предстои да видим как на практика може да бъде измерен ефекта от „големите данни“.