# ПРЕДИЗВИКАТЕЛСТВАТА НА ГОЛЕМИТЕ ДАННИ – СЪЩНОСТ, ХАРАКТЕРИСТИКИ И ТЕХНОЛОГИИ

Станимира Йорданова[*], Камелия Стефанова[**]

## Увод

Съвременният свят става все по-динамичен и това налага фирмите и организациите да отговарят адекватно на бързо променящите се бизнес условия. Големи данни (Big Data) е нов термин, който се отнася за изключително големи, сложни и бързо променящи се набори от данни. Големите данни са подход за откриване на важни модели на поведение в набора от данни, тяхното анализиране и използването им в последващи обработки.

Терминът „Големи данни“ описва големи набори данни, които са толкова обемни и комплексни, че не могат да бъдат събирани, селектирани, обработвани или управлявани посредством широко използваните и прилагани досега софтуерни инструменти в рамките на приемливи периоди от време. Генерирането, обработката, съхранението, споделянето, преносът, анализът и визуализирането на големи данни създават нови предизвикателства за развитието на технологиите – нарастване на компютърната мощност, по-големи ИТ устройства за съхранение и високоскоростни канали за данни. Събирането на данни в реално време поставя нови възможности и предизвикателства за разбиране на данните, напр. комбиниране на административни данни с големи данни, идващи от различни източници – като търговски данни, мобилни устройства и сензори, социални медии и други обществено достъпни данни. Това, от една страна, изисква нови средства и методи, а от друга – нови умения и визия за работа с данни.

Основната задача на големите данни е събраното количество данни да се обработи по такъв начин, че да предостави смислена информация на потребителя. Тенденциите, които се налагат с големите данни, са свързани с допълнителна обработка и анализи за откриване на нови модели и по-задълбочено разбиране на процесите и явленията в различни области. Според прогноза на International Data Corporation (IDC), пазарът на технологии и услуги в областта на “големите данни“ ще расте средно с 26,4% годишно,

[*] Станимира Йорданова, асистент, доктор, катедра „Информационни технологии и комуникации“, УНСС, email: syordanova@unwe.bg

[**] Камелия Стефанова, професор, доктор, катедра „Информационни технологии и комуникации“, УНСС, email: kstefanova@unwe.bg

като през 2018 г. се предвижда да достигне 41,5 млрд. долара. Така сегментът Големи данни ще расте 6 пъти по-бързо от IT пазара като цяло. Голямото предизвикателство пред бизнеса, икономиката и обществото е да прилагат предимствата на новите технологии и да извличат стойност от огромните потоци информация, с които разполага всяка организация. Цифровата революция поставя нови предизвикателства за организационни промени и въвеждане на нов подход в разбирането на данни и информация.

Настоящата статия има за цел да направи характеристика на големите данни и обзор на свързаните с тях технологии за съхранение и управление на големите данни чрез анализ и обобщаване на информацията от различни публикации в областта на големите данни. Резултатите от изследването могат да бъдат използвани от експерти по ИТ и бизнес анализатори при проучване и избор на подходящи технологии и архитектури за анализ на големи данни.

Статията се състои от четири части. Част първа представя причините за нарастването на обема на данните. Развитието на характеристиките на големите данни през годините и основните десет характеристики, които очертават основните предизвикателства, свързани с извличане, съхраняване, обработване и анализ на данните са представени в част втора. В част трета е направено разграничение между типовете големи данни и са анализирани и представени основни източници на големи данни от гледна точка на типа и начина на създаване на данните. В последната част са описани особеностите на езерото за данни като хранилище за големи данни и са представени популярни технологии за разработване на езеро за данни.

## Защо данните стават големи?

Логичният въпрос е "Колко големи?" и всъщност обемът ли е единствената характеристика, която се има предвид за „големи данни". Границата, след която данните стават "големи" е условна и тенденцията е тя да се отмества към все по-големи обеми, тъй като компютърната техника постоянно се усъвършенства и става все по-достъпна. Названието предполага, че става въпрос за анализ на големи обеми от данни. Според доклад на Маккинзи институт (McKinsey Institute), озаглавен "Големите данни: новата граница за иновации, конкуренция и производителност" (Big data: The next frontier for innovation, competition and productivity), терминът големи данни означава набор от данни, размерът на които надхвърля възможностите на типичните бази данни за съхранение, управление и анализ на информация. Според прогнозата в доклада "Изследване на цифровата вселена" (Digital Universe Study, IDC), данните ще нараснат с около 50 пъти до 2020г. вследствие на повишаващия се брой на вградени системи като сензори, информационни

платформи, медицински устройства и т.н., а неструктурираната информация като файлове, имейли и видео ще съставлява 90% от всички данни.

Една важна особеност на големите данни е стремежът за обработка на огромния информационен поток като цяло, с цел получаване на по-точни резултати от провежданите анализи. "Големите данни" предполагат нещо повече от анализ на огромни обеми информация. Проблемът не е в това, че организациите създават огромни обеми от данни, а в това, че голяма част от тях е във формат, който не съответства на традиционния структуриран формат на базите данни – неструктурирани данни, като видео записи, текстови документи, машинен код, геопространствени данни и т.н. Цялата тази информация се пази в разнообразни хранилища, много често извън пределите на организацията. В резултат, макар и да имат достъп до огромен обем от данни, организациите нямат необходимите инструменти, за да намират зависимости между тези данни и да правят значими изводи. Сред многобройните източници на големи данни са измерващи устройства, събития от радиочестотни идентификатори, потоци съобщения от социалните мрежи, метеорологични данни, потоци от данни за местоположението на абонати на клетъчни мрежи, устройства за аудио и видеорегистрация и т.н. Масовото разпространение на изброените технологии и модели за използване на различни типове устройства и интернет услуги са отправната точка, от която започва проникването на големите данни във всички сфери на човешката дейност.

Днес данните се обновяват все по-често. Стига се до ситуация, в която традиционните методи за анализ на информацията не могат да се справят с огромните обеми постоянно обновявани данни, което води до необходимостта от въвеждането на нови технологии за големи данни. Всеки ден се създават огромни количества от данни и информация във всички дейности и точки по света. Тези данни трябва да се анализират, за да се открие ценната, полезна информация, която да обслужва бизнеса във вземането на оптимални решения.

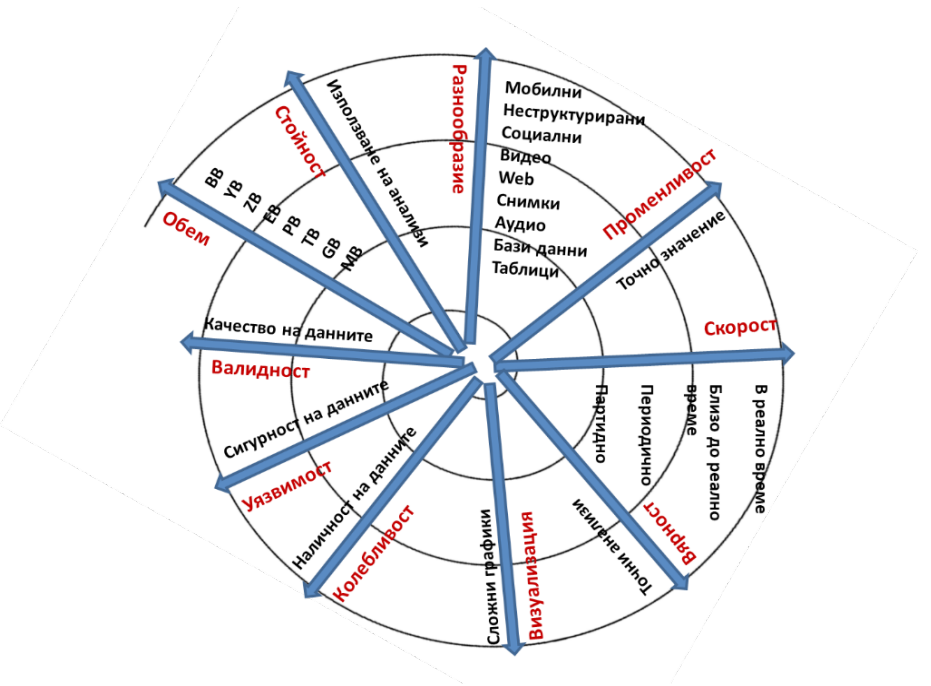
## Десетте V-та на големите данни

През 2001 г. консултантска компания Гарднър (Laney, Gartner Group), която прогнозира тенденциите в развитието на информационната индустрия, обособява три предизвикателства пред управлението на данни и ги събира под определенията обем, скорост и разнообразие на данните (Data Volume, Data Velocity и Variety Data). Те се превръщат в 3-те V-та на големите данни. През годините V-тата продължават да се развиват и обогатяват и около десетилетие по-късно стават 4V's, след това 7V's и 10V's. Тоест вече има 10 предизвикателства и особености на големите данни, започващи с буквата

„V" на английски език, които постоянно се обогатяват и се добавят нови. През 2017г. Том Шафър (изследовател в областта на анализа на данни) от компанията Елдър Рисърч представя актуализиран списък с 42 V's на големите данни.

Главната цел на V-базираната характеристика е да се подчертаят и обобщят някои от най-сериозните предизвикателства, свързани с внедряването на големи данни. Организациите от различните сектори на икономиката изучават възможностите, които предлагат анализите на големи данни (за маркетингови иновации, ангажираност на клиентите, постигане на крайната цел и мн. други бизнес цели), но и доста от тях не съумяват успешно да приложат тези технологии и да постигнат желаната възвръщаемост на инвестициите от тях.

На Фиг. 1 са представени само първите 10 V's от характеристиките и свойствата на големите данни, които да очертаят основните предизвикателства и предимства на прилагането на големи данни (Firican, 2017).




**Фиг. 1.** 10 V's на големите данни

а) Volume (обем) – Обемът може би е най-известната характеристика на „големите данни", тъй като обемът на данните се увеличава много бързо – 90% от всички съвременни данни са създадени през последните няколко години според доклада "Изследване на цифровата вселена" (Digital Universe Study). Ето няколко примера: 300 часа видеоклипове се качват в YouTube

всяка минута (Transforming Data with Intelligence, TDWI). През 2017 г. са били направени 1,5 трилиона снимки. Тъй като една и съща снимка обикновено се съхранява на няколко места, общият брой на съхранените снимки също е 4,7 трилиона през 2017 г. Компанията TDWI изчислява, че през 2017 г. глобалният мобилен трафик възлиза на 6,9 екзабайта на месец (6,9 милиарда гигабайта).

b) Velocity (скорост) – Тази характеристика на големите данни се отнася до скоростта, с която се генерират, произвеждат, създават или обновяват данните. Информационният склад на Facebook съхранява повече от 300 питабайта данни, а същевременно, Facebook заявява, че 600 терабайта се въвеждат всеки ден. Google обработва средно повече от 40 000 заявки за търсене на всяка секунда, което е повече от 3,5 милиарда търсения на ден (Transforming Data with Intelligence, TDWI).

c) Variety (разнообразие) – Когато става дума за големи данни, трябва да се има предвид, че компаниите трябва да се справят не само със структурирани данни, но и с полуструктурирани и предимно с неструктурирани данни.

d) Variability (променливост) – Променливостта в контекста на големите данни се отнася до няколко различни аспекта. Едното е броят на несъответствията в данните. Те трябва да бъдат открити чрез методи за откриване на аномалии и отклонения, за да се получат смислени анализи. Големите данни също се променят поради множеството измерения на данните, които са резултат от множество различни видове данни и източници. Променливостта може да се отнася и до непоследователната скорост, при която големи данни се зареждат в базата данни.

a) Veracity (истинност) – Това е една от трудните характеристики на „големи данни“. Докато всички изброени по-горе свойства се увеличават, истинността (доверието в данните) спада. Истинността на данните се отнася по-скоро до източника или достоверността на източника на данни, контекста и колко е смислено да се използва за анализа. Познаването на истинността на данните от своя страна, помага да се разберат по-добре рисковете, свързани с анализите и бизнес решенията, базирани на конкретен набор от данни.

b) Validity (валидност) – Подобно на истинността, валидността се отнася до точността и коректността на данните. Според проучване, направено от CrowdFlower през 2016 г., около 60% от времето на анализаторите на данни се изразходва за изчистване на данните преди да може да се пристъпи към анализ. Ползите от анализи на големи данни зависят именно от валидността на данните, които се анализират. Това е причината фирмите да въвеждат нови практики в управлението на данните, за да си осигурят качествени данни.

c) Vulnerability (уязвимост) – Големите данни носят много рискове за сигурност. Нарушение или пробив в такива данни водят до огромни

негативни последствия, особено лични данни (пр. Cambridge Analitica и Александър Когън, милионите откраднати данни от лични профили във Facebook през 2018).

d) Volatility (колебливост) – Големите данни са динамични, развиващи се, пространствено-времеви данни, динамични редове, сезонни с нестатично поведение. Поради скоростта и обема на големи данни тяхната колебливост трябва да бъде внимателно обмислена. Трябва да се установят правила за наличност на данни, както и да се осигури бързо извличане на информация, когато е необходимо. Когато се борави с „големи данни“, разходите и сложността на процеса на съхранение и извличане се увеличават.

e) Visualization (визуализация) – Не може да се разчита на традиционни графики, когато трябва да представят милиарди данни – затова са необходими нови, различни начини за интегриране, обработка и показване на резултатите. Комбинирайки това с разнообразието и скоростта на големите данни, както и сложните взаимоотношения между тях, може да се заключи, че развиването на смислена визуализация не е никак лесна задача. Голямо предизвикателство тук се явява и бизнес разбирането на данните и полезната форма, която да удовлетворява оптималната интерпретация.

f) Value (стойност) – Може би една от най-важните специфики е ценността. Другите характеристики на данните стават безсмислени, ако бизнесът не може да извлече полза от разбирането на данните и анализа – напр. по-добро познаване на поведението на клиентите, оптимизиране на процесите и подобряване на бизнес постиженията, и др.

g) Много важен въпрос за големите данни е: колко стари могат да бъдат данните, за да се считат за ненужни и нерелевантни. Преди големите данни, организациите са съхранявали информацията за неопределено време – няколко терабайта данни може да не създават големи разходи за съхранение и да не причиняват проблеми с производителността. В момента ситуацията е различна. Данните идват бързо, но и бързо остаряват, променят се и не носят стойност. Освен това компаниите са задължени да се съобразяват с определени правила и закони за съхранение на данни и за заличаването им. Също така, когато става въпрос за лични данни, към настоящия момент, актуалният Закон за защита на личните данни предвижда възможност за всяко физическо лице да поиска заличаването на личните си данни, когато обработването им не отговаря на изискванията на този закон. Следователно, някои от другите важни характеристики на големите данни са: местоположение (venue), речник (vocabulary), неопределеност (vagueness) и др. Необходимо е да се отбележи, че редът на подредба на характеристиките на големите данни не е редът им на значимост. В различните приложения приоритетите се определят индивидуално.

## Източници на структурирани и неструктурирани данни

Големите данни са два основни типа – структурирани и неструктурирани/ полуструктурирани. Структурирани са данните, които имат определен формат и структура, и обикновено се съхраняват в бази данни. Традиционните източници на структурирани данни са системите за управление на връзките с клиентите (Customer Relationship Management, CRM), за планиране на ресурсите на предприятието (Enterprise Resource Planing, ERP) и финансови данни. Често тези данни са интегрирани в складове за данни (Data Warehouses), подготвени за последващ анализ. Организациите фокусират инвестиции си в системи, работещи със структурирани данни, тъй като прилагането им би довело до подобряване на представянето им. Неструктурираните данни нямат специфичен формат и структура като техният обем в организациите бързо нараства. Използвайки технологии за анализ на текст, текстовите данни могат да се обработят, анализират, структурират и използват по различни начини. Например, неструктурирани данни се създават при клиентски разговори в социалните медии, които се използват в анализите на социалните медии. Освен това неструктурирани данни са бележките в кол центровете, електронните писма, писмените коментари в различните проучвания и други документи, които се анализират с цел разбиране поведението и опита на клиентите. След като неструктурираните данни се структурират, може да се комбинират с други структурирани в склада за данни, след което да се приложат бизнес интелигентни средства или методи за извличане на знания от данни (data mining) с цел получаване на нова значима за организацията информация. Големите структурирани и неструктурирани данни могат да бъдат два типа в зависимост от създаването им. В таблица 1 са представени примери за големи структурирани и неструктурирани данни и техните източници.

**Таблица 1**: Примери за големи структурирани и неструктурирани данни

| Големи данни | Данни, създадени от компютър или машина без човешка намеса | Данни, създадени от човека при взаимодействие с компютри |
|---|---|---|
| 1 | 2 | 3 |
| Структурирани | Сензорни данни като радиочестотните данни, данни от GPS системи (Global Positioning System), от смартфоните и др. Компаниите се интересуват от тези данни за проследяване на обекти от разстояние, както и за управление на веригата за доставки и инвентарен контрол, както и за анализ и разбиране на поведението на клиентите. | Входни данни в системи, които се въвеждат от хората. |
| | Данни от уеб логове, които се създават при работата на сървърите, приложенията, мрежите и могат да се използват за предсказване на пробиви в сигурността. | Данни, които се генерират при всяко кликване върху връзка в уебсайт данни от игрите, където всеки ход, който се прави в една игра, може да бъде записан. Тези данни могат да бъдат анализирани и да подпомогнат разбирането на поведението на клиентите. |
| | Данни, създадени в момента на продажбите, когато касиерът използва баркода на закупения продукт .<br><br>Финансови данни | Данни от игрите, където всеки ход, който се прави в една игра, може да бъде записан. Тези данни могат да бъдат анализирани и да подпомогнат разбирането на поведението на клиентите. |
| Неструктурирани | Сателитните снимки, които предоставят данни например за времето. | Документи като отчети, дневници, проучвания, електронна поща. Текстът в тези документи е неструктурирани данни. |
| | Научни данни от сеизмични изображения, атмосферни и физични данни. | Социални медии и големите социални медийни платформи като YouTube, Facebook, Twitter, LinkedIn и Flickr, чиито потребители създават текстови данни. |
| | Снимки и видео от наблюдения, трафик и сигурност. | Мобилни телефони, които предоставят неструктурирани данни като текстови съобщения и данни за местоположение на потребителите. |

Технологиите за съхранение на данни се развиват главоломно от години насам, като следващата огромна крачка напред най-вероятно се прави в момента. Натрупването на данни е важна задача на модерния дигитален свят, тъй като всяко едно съприкосновение със софтуер или уебсайт се запазва, записва или изисква все повече информация, за да бъде осъществено. Тези процеси водят до необходимостта от създаване и разширяване на дигитални хранилища за данни, които достигат до огромни размери.

**Технологии за съхранение и управление на големи данни**

През 2010 г. Джеймс Диксън от компанията Пентахо, лидер сред доставчиците на бизнес интелигентни средства, дефинира понятието езеро за данни (data lake) като данни, които се излъчват от един източник и се съхраняват в естественото им състояние.

Езеро за данни е хранилище за много различни видове и източници на данни, структурирани или неструктурирани, вътрешни или външни с цел улесняване чрез различни начини достъпа и анализа на данните. Три са изискванията към изграждането на езеро от данни (Subramanyam, 2018):

- Да събира и съхранява данни от един или повече източници в оригинална сурова форма и евентуално в различните си обработени форми
- Да позволява гъвкав достъп до данните от различни приложения – достъп до структурирани таблици и колони, както и достъп до неструктурирани данни
- Транзакционните данни трябва строго да се управляват, за да се предотврати превръщането на езерото в блато.

Езерото за данни е своеобразно хранилище за структурирани данни от транзакционни системи и неструктурирани данни от уеб логове, социални медии, устройства за интернет на нещата (Internet of Things) и други източници на големи данни като прави възможно комбинирането на данните с цел по-точни анализи чрез използването на уникални тагове на метаданни. Без тези тагове, езерото за данни бързо може да се превърне в блато. В съвременната архитектура на данните, която трябва да позволява да бъдат съхранявани данни от различни източници с възможност потребителите да интегрират, обогатяват и анализират данните без ограниченията, възниква въпроса как може да се интегрира съществуващ склад за данни с езеро на данни.

Складът за данни (Data Warehouse) е хранилище за данни, оптимизирано за анализ на релационни данни, идващи от транзакционните системи и оперативни бизнес приложения. Структурата на данните и схемата се дефинират предварително, за да се оптимизират за бързи SQL заявки, където резултатите обикновено се използват за оперативни справки и анализи.

Данните се изчистват, обогатяват и трансформират, за да могат да действат като „единствен източник на истината“, на който потребителите могат да се доверят. Езерото за данни е различно хранилище, което съхранява релационни данни от транзакционните системи и оперативни бизнес приложения и нерелационни данни от мобилни приложения, сензорни устройства и социални медии. Структура на данните не се дефинира. Това означава, че може да се съхраняват всички данни в оригинален формат без да се проектира определена структура с необходимост от предварително определяне на въпроси, които се нуждаят от отговор в бъдеще. Различията между склада за данни и езерото за данни са представени в обобщен вид на таблица 2.

**Таблица 2:** Различия между склада за данни и езерото за данни

| | Склад за данни | Езеро за данни |
|---|---|---|
| **Данни** | Структурирани данни | Структурирани, неструктурирани, полуструктурирани данни |
| **Обработка на данните** | Дефинирана структура на данните преди зареждане им<br>Извличане, трансформиране и зареждане на данните в склада | Липсва дефинирана структура на данните<br>Извличане и зареждане на данните в средата на езерото, трансформиране при необходимост |
| **Разходи** | Високи разходи | Ниски разходи (от гледна точка на софтуер и хардуер) |
| **Зрялост на технологиите** | Висока зрялост на технологиите | Технологии в процес на усъвършенстване |
| **Потребители** | Бизнес потребители | Изследователи на данни<br>Анализатори на данни |

- Съхранявани данни**:** В склада за данни се съхраняват структурирани данни, които са обработени и трансформирани за нуждите на последващи справки и анализи от бизнес потребители. В езерото за данни могат да се съхраняват структурирани, неструктурирани/полуструктурирани данни в суров вид.
- Обработка на данните**:** В традиционния склад за данни в процесите на извличане, трансформиране и зареждане на данните, трансформирането на данните се извършва преди зареждането в склада. Структурата (схемата) на данните трябва да бъде дефинирана преди зареждането им в склада „schema on write”. Правилното трансформиране и

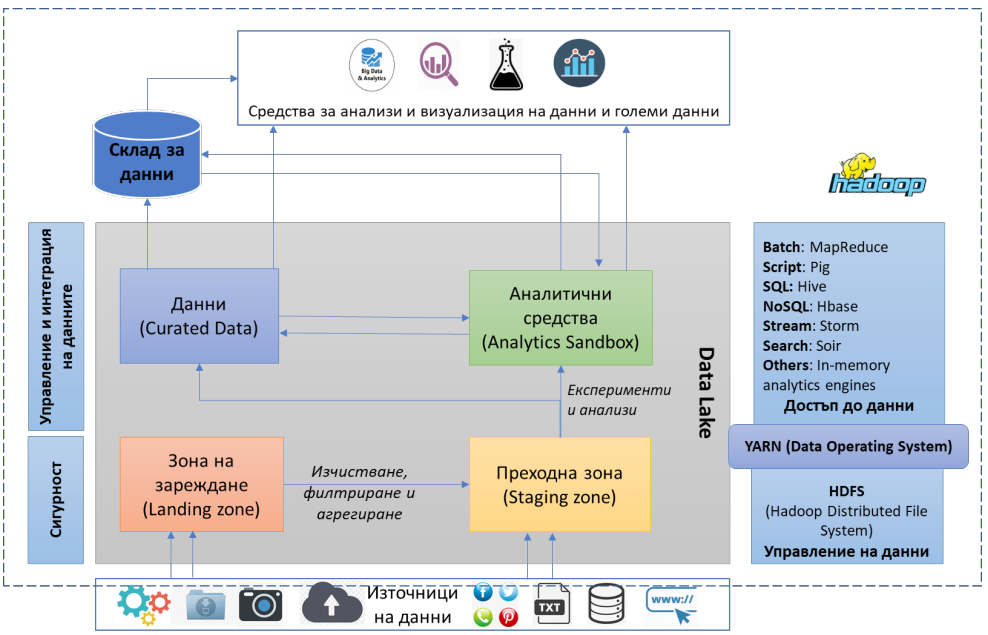
зареждане отнема време и усилия, защото определянето на структурата на данните зависи от предварителното дефиниране на случаите на използване на данните, които се развиват и променят във времето. Ако се появят нови бизнес изисквания, които променят фундаментално първоначалната структура на данните, тогава склада за данни трябва да се преструктурира. Затова е необходима по-гъвкава структура на данните. За разлика от склада за данни, езерото за данни използва част от процесите на извличане, трансформиране и зареждане на данните. Данните се извличат и зареждат в езерото в оригиналния си формат, след което могат да се трансформират, ако се появи необходимост. Техниката се нарича „schema on read“ и се използва при управлението на големи данни като съхраняването на данни става сравнително лесно с по-малко инвестиции (Coates, 2016). Езерото за данни дава възможност на разработчиците и изследователите на данни лесно да конфигурират своите модели, заявки и приложения в движение.

- Разходи**:** Една от основните характеристики на технологиите за големи данни е, че разходите за съхраняване на данните са сравнително ниски в сравнение с тези при складовете за данни. Причините за това са две: повечето технологии са с отворен код и лицензирането и подкрепата от общността от разработчици са безплатни, както и са проектирани за инсталация на хардуер с ниски разходи.
- Зрялост на технологиите**:** Технологиите за складове за данни, които се развиват от десетилетия, са много по-зрели от тези за езера за данни. Трябва да се отбележи обаче, че в областта на големите данни в момента се полагат значителни усилия за развиването на технологиите за управление на данни извън складовете за данни.
- Потребители**:** Основни потребители на склада за данни и аналитичните средства са бизнес потребители, докато езерото за данни в този етап на зрялост е най-подходящ за използване от изследователи на данни или от анализатори на данни със значителен опит в намирането на полезна информация от сурови данни.

Съвременните архитектури за управление на разнообразни типове данни от различни източници се проектират с изискването за подпомагане на бизнес стратегията на организацията и за достатъчна гъвкавост във възможността за съхраняване на нови типове данни в бъдеще. Скот Гидли (Gidley, 2017) изтъква важността на метаданните в този процес на успешно управление на данните в езерото. Той посочва, че е необходима рамка, която да обхваща техническите, оперативните и бизнес метаданни, по начин, по който прави възможно откриване и използване на определени данни в зависимост от случая на употреба. Автоматизираното хващане на метаданните при тяхното пристигане в средата на езерото и обвързването им със специ-

фични дефиниции, например бизнес речник на предприятието, гарантира, че всички потребители ще тълкуват данните по един и същ начин, използвайки този набор от правила и концепции. Неправилната архитектура на метаданните може да попречи на потребителите да използват пълния набор от данни за експерименти или анализи в аналитичната и експериментална зона на езерото.

Архитектурата на езерото за данни може да обхваща няколко зони, представени на фиг. 2 (Makaranka, 2018).




**Фиг. 2.** Езеро за данни и Hadoop

Зоната на зареждане на данните (landing zone) съдържа суровите данни от източниците без обработка и трансформация. Тази зона се управлява от ИТ експертите, които автоматизират процеса по зареждане на данните в езерото. Суровите данни подлежат на изчистване, филтриране и/или агрегиране при прехода към следващата зона. Данните могат да постъпят в преходната зона (staging zone) по два начина: от зоната за зареждане, ако са данни от сензори или директно от източника на данни без предварителна обработка, ако например са коментари на клиенти в социалните мрежи. В преходната зона данните изчакват използване от приложения. Експерименталната и аналитична зона (analytics sandbox) е отделна среда в архитектурата на езерото за данни, предназначена да бъде използвана от множество потребители

и поддържана с помощта на ИТ. Тази среда се контролира от анализаторите и позволява да се инсталират и използват инструменти за анализ на данни, както и да управляват планирането и обработката на данните. От всички споменати зони задължителна е преходната зона, докато всички останали са незадължителни. В средата анализаторите изследват и експериментират с данни от различни източници като прилагат модели или алгоритми към необработените данни с цел изследване и разбиране на данните, създаване на прототипи, изграждане на нови хипотези и случаи на използване на данните. Включването на тази зона в архитектурата на езерото осигурява среда, в която данните са винаги налични при поискване и позволява бърз достъп до тях с цел обработка и анализ, без да е нужно да реализират големи бизнес интелигентни проекти. Друго съществено предимство за бизнеса и ИТ екипа е, че тази среда дава на бизнес потребителите възможност да експериментират като изграждат прототипи на решения, свързани с данните. Благодарение на тези прототипи, те могат да формулират своите изисквания към ИТ експертите, което спестява много време и усилия.

В последната зона (curated zone) се съхраняват организирани и подредени големи данни, готови за анализ. Съществуват различни мнения дали тази зона с данни трябва да се счита за част от езерото за данни. Някои автори считат, че тази зона е част от езерото, защото съхранява, както традиционни данни, така и големи данни, докато складът за данни само традиционни данни. Различни са начините за използване на езерото за данни. Например езерото е много добро решение за съхраняване на данни от сензори и интелигентни устройства (интернет на нещата), които поради обема си са потрудни за съхранение, и може да поддържа анализ на тези данни в реално време. Експериментално може да се анализират данните в езерото с цел доказване на тяхната стойност преди да бъдат трансформирани за последващи анализи, което е възможно поради специфичния подход в процесите по извличане, зареждане и трансформиране на данните. Понякога някои данни се използват рядко, но трябва да са налични за анализ, затова една от стратегиите за използване на езерото е да съхранява и архивира исторически данни. Други стратегии за използване на езерото е като подготвителен етап (staging area) преди склада за данни или като част от склада за данни, където да се съхраняват данни, които не се използват често и са достъпни чрез заявки. Предлагат се и решения, базирани на Ламбда архитектури (Coates, 2017) или като възможност за разширяване на склада за данни с възможности за разпределена обработка.

Разпределената файлова система (Distributed File System) на Hadoop, софтуерна рамка, управлявана от фондацията Apache, е най-популярната сред множеството възможни технологии за изграждане на езеро за данни, защото дава възможност за съхраняване на разнообразни типове данни – сен-

зорни данни, видео и аудио файлове и текстови записи. Файловата система поддържа широка гама от техники за обработка и заедно със системата MapReduce е основен компонент на технологията Hadoop, която осигурява паралелна обработка на данни между изчислителните възли с цел увеличаване на скоростта на изчисленията и намаляване на латентността.

Разпределената файлова система е универсален и устойчив подход към управлението на файлове в среда за големи данни, защото данните се записват веднъж и след това се четат многократно за разлика от файловите системите, при които данните постоянно се четат и записват. Системата работи като разделя големи файлове на по-малки парчета, наречени блокове. Блоковете се съхраняват във възли за обработка на данни (data nodes) и отговорност на контролния възел (Namenode) е да знае кои блокове в кои възли съставят пълния файл. Контролният възел също управлява целия достъп до файловете, включително четене, писане, създаване, изтриване и репликиране на блокове с данни върху възлите за обработка на данни. Разпределената файлова система е устойчива срещу повреди в сървърите, защото блокове се репликират навсякъде в клъстера.

Системата MapReduce предоставя високоефективно паралелно/разпределено обработване на данни от алгоритъма MapReduce. Тя осигурява всички възможности, необходими за разбиване на големи данни в управляеми парчета и паралелното им обработване на разпределен клъстер. MapReduce е начин за изпълнение на набор от функции върху голямо количество данни в пакетен режим. Функцията „map" разпределя програмния проблем или задачи в голям брой системи и обработва разположението на задачите по начин, който балансира натоварването и управлява възстановяването от откази. След като разпределеното изчисление приключи, друга функция „reduce" обединява всички елементи обратно, за да осигури целения резултат. Предимствата на MapReduce се реализират само когато се прилагат към разпределен клъстер от сървъри като създава среда, в която същата функция може да се приложи едновременно на много машини.

Екосистемата Hadoop обединява непрекъснато разширяваща се колекция от инструменти и технологии, специално създадени, за да осигуряват разработването, внедряването и поддръжката на решения за големи данни. В Hadoop може да се инсталират допълнителни софтуерни пакети като YARN, Pig, Hive, HBase, Phoenix, Spark, ZooKeeper, Flume, Apache Sqoop, Oozie, Storm. Съществуват и други технологии за съхраняване и обработка на големи данни в езерото (като Apache Cassandra или MongoDB). Изборът на подходяща технология зависи от правилната преценка на предимствата и ограниченията на всяка от тях и конкретните изисквания в управлението на данните.

## Заключение

Генерирането и съхраняването на големи данни е непрекъснат и постоянно развиващ се процес. Необходимостта от използване на големите данни води до развиване и разширяване на съвременната архитектура на данните, която трябва да позволява да бъдат съхранявани данни от различни източници с възможност потребителите да интегрират, обогатяват и анализират данните без ограниченията. С навлизането и развитието на технологиите с отворен код се появи нова вълна в управлението на данните – използване на всички налични данни чрез интегрирана система. Езерото за данни е добър подход за съхраняване на големи данни, който отговаря на изискванията и предизвикателствата на големите данни. Но неправилното и необмислено проектиране и използване на езеро за данни носи значителни рискове, свързани с качеството на данните, сигурността и контрола на достъп и използването им. Технологии като MapReduce, Hadoop и други позволяват да се обработва огромно количество данни ефективно и навреме. Технологията Hadoop, обаче, като разпределен подход за управление и анализ на файлове, където данните се разделят, съхраняват и обработват на разпределена платформа, не може да замести подходите при управление на традиционната система за съхранение на данни. Hadoop не поддържа актуализации или промени в съществуващия запис, което е важен процес в системата за съхранение на данни. Затова се търсят варианти за интегриране и/или разширяване на традиционния склад за данни, за да поддържа управлението и анализа на неструктурирани данни и данни в реално време.

Големите данни се явяват предизвикателство както за ИТ инфраструктурата на бизнес организациите, така и за нуждата от добре подготвени специалисти, които да могат целенасочено да селектират и анализират данните за целите на бизнеса. Анализът на големите данни и използването на методи за извличане на бизнес правила и генериране на модели, допринасят за вземане на правилни и навременни решения в жизнено важни области като анализ и оценка на риска, прогнозиране на печалбата, оптимизиране на решенията, насочване на рекламата и др.

*Референции:*

Ashworth L. and Gillespie T. (2018), Who is Dr Aleksandr Kogan, the Cambridge academic accused of misusing Facebook data? https://www.varsity.co.uk/news/15192 (accessed 20 October 2018)

Cisco Connected World Technology Report (2017), The Potential and Challenge of Data, достъпно на: https://www.cisco.com/c/dam/en/us/

solutions/enterprise/connected-world-technology-report/Global-Data-CCWTR-Chapter3-Media-Briefing-Slides.pdf (accessed 10 February 2019)
Coates M. (2016) Data Lake Use Cases and Planning Considerations, достъпно на: https://www.sqlchick.com/entries/2016/7/31/data-lake-use-cases-and-planning (accessed 5 January 2019)
Coates M. (2017), Zones in a Data Lake, December 30, 2017, достъпно на: https://www.sqlchick.com/entries/2017/12/30/zones-in-a-data-lake (accessed 5 January 2019)
Data Science report (2016), CrowdFlower, https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf (accessed 10 February 2019)
EMC Digital Universe with Research & Analysis by IDC (2014), The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, April 2014, достъпно на: https://www.emc.com/leadership/digital-universe/ (accessed 10 February 2019)
Firican G. (2017), The 10 Vs of Big Data, TDWI-Transforming Data with Intelligence, достъпно на: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx (accessed 5 January 2019)
Gidley S. (2017), Tips for managing metadata in a data lake: Metadata is central to a modern data architecture, May 4, 2017, O'Reilly Media, достъпно на: https://www.oreilly.com/ideas/tips-for-managing-metadata-in-a-data-lake (accessed 5 January 2019)
Hurwitz J., Nugent A., Halper F., Kaufman M. (2013), Big Data For Dummies, publisher: For Dummies, Release Date: April 2013, ISBN: 9781118644171
James D. (2010), Union of the State – A Data Lake Use Case, достъпно на: https://jamesdixon.wordpress.com/?s=data+lake (accessed 10 February 2019)
Laney D. (2001), 3D Data Management: Controlling Data Volume, Velocity and Variety, Gartner, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 20 October 2018)
Makaranka I. (2018), Alternative Approaches to Implementing Your Data Lake, ScienceSoft, May 21, 2018, достъпно на: https://www.scnsoft.com/blog/data-lake-implementation-approaches (accessed 10 February 2019)
McKinsey Institute (2011), Big data: The next frontier for innovation, competition and productivity, May 2011, достъпно на: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed 10 February 2019)
Rozados V, Ivan & Tjahjono, B. (2014), Big Data Analytics in Supply Chain Management: Trends and Related Research. 10.13140/RG.2.1.4935.2563.

Schmarzo B. (2017), Data Warehouse and Data Lake Analytics Collaboration, October 19, 2017, достъпно на: https://infocus.dellemc.com/william_schmarzo/data-warehouse-data-lake-analytics-collaboration/ (accessed 5 January 2019)
Shafer T. (2017), The 42 V's of Big Data and Data Science, достъпно на: https://www.elderresearch.com/blog/42-v-of-big-data, (accessed 20 October 2018)
Subramanyam S. (2018), 3 Requirements for an Enterprise Data Lake, достъпно на: https://orzota.com/2018/02/16/enterprise-data-lake/ (accessed 20 October 2018)
Thilina G. (2015), Hadoop and MapReduce, достъпно на: https://hub.packtpub.com/hadoop-and-mapreduce/ (accessed 20 October 2018)

# ПРЕДИЗВИКАТЕЛСТВАТА НА ГОЛЕМИТЕ ДАННИ – СЪЩНОСТ, ХАРАКТЕРИСТИКИ И ТЕХНОЛОГИИ

**Резюме**

В съвременния свят обемът, сложността, разнообразието, темповете на растеж и скоростта на генерираните данни достигат невероятни нива като данните не могат да бъдат събирани, селектирани, обработвани или управлявани посредством широко използваните и прилагани досега софтуерни инструменти. Необходимостта от извличане на смислена и значима информация за потребителя от големи данни води до развитието на нови подходи и технологии, които да обхванат процесите на съхраняване, обработка и анализ на големите данни от различни източници. Настоящата статия има за цел да направи характеристика на големите данни и обзор на свързаните с тях технологии за съхранение и управление на големите данни, които изискват въвеждане на нов подход в разбирането на данните и информацията.

**Ключови думи:** големи данни, езеро за данни, Hadoop

**JEL:** C88

# BIG DATA CHALLENGES – DEFINITION, CHARACTERISTICS AND TECHNOLOGIES

Stanimira Yordanova[*], Kamelia Stefanova[**]

**Abstract**

Volume, complexity, variety and velocity of generated data today reach incredible levels and data cannot be collected, selected, processed or managed through widely used and applied software tools. The need to extract meaningful information from big data evokes the development of new approaches and technologies for storing, processing and analyzing big data coming from different sources. This paper aims to present the definition, characteristics and technologies for storing and managing big data that require a new approach to understanding data and information.

**Keywords**: Big Data, Data Lake, Hadoop

**JEL**: C88

[*] Stanimira Yordanova, Assistant Professor, PhD, Department of Information and Communication Technologies, UNWE, email: syordanova@unwe.bg

[**] Kamelia Stefanova, Professor, PhD, Department of Information and Communication Technologies, UNWE, email: kstefanova@unwe.bg