

Технологични аспекти на модела „3Vs” за представяне на големи обеми от данни

Галина Атанасова, Пламенка Христова, Пано Панов

Abstract: Technological Aspects of Model “3Vs” to Present Big Data: *The paper justifies the necessity to introduce the students from the bachelor degree courses in “Computer Science” how can present and arrange the Big Data. There are explained the essence of the model “3Vs” technology and MapReduce, designed to work with big data. There are presented 10 major issues that are justified by the model. A simple approach for a smooth transition to study these new technologies by students is presented.*

Key words: *Computer Technologies, Big Data, MapReduce, Education.*

ВЪВЕДЕНИЕ

Всеки момент в света се генерират и съхраняват огромно количество данни като тенденцията е за непрекъснато нарастване. Обемът на данните се увеличава с повече от 50% годишно – около 8 пъти по-бързо от пазара на информационни и комуникационни технологии.

Анализът на големите информационни масиви 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 ускорено диагностициране на мозъчни и сърдечни заболявания, анализ и оценка на риска, избягване на задръстванията по пътищата, прогнозиране на добивите, насочване на рекламата и др. [3].

Услугите и технологиите, свързани с обработката на големи информационни масиви, непрекъснато нарастват и създават стотици хиляди нови работни места. Американско проучване показва, че специалистите по големи данни са най-търсените в областта на анализа, статистиката и управлението на информация, като в началото на 2014 год. търсенето на подобни кадри в САЩ е надвишило 500 000 души, а фирмите, които се ръководят при вземането на решения от извлечена от данни информация, имат с 5-6% по-висока производителност от останалите [5].

От друга страна „Големите данни” се явяват предизвикателство както за ИТ инфраструктурата на фирмите, така и за нуждата от добре подготвени специалисти, които да могат целенасочено да селектират и анализират данните, така че да носят нови ползи.

ИЗЛОЖЕНИЕ

Понятието „Големите данни” (*Big Data*) представлява съвкупност от подходи, методи и програмно-технически средства за обработка на огромни обеми структурирани и неструктурирани данни, които се характеризират с голямо многообразие, събират се от различни източници, понякога генерирани в реално време. С помощта на нетрадиционни методи за обработка и анализ на информация се откриват нови хипотези и модели, които се използват ефективно за оптимално управление на човешките дейности в най-различни стопански сфери. Процесът на вземане на решение, който преди се е основавал на предчувствия и интуиция, сега се задвижва от данни и знания.

Моделът „3Vs”, представен на фиг. 1, включва трите основни фактора (обем - *volume*, скорост – *velocity* и многообразие - *variety*), които характеризират твърде точно най-важните свойства на „Големите данни”:

- **Обем**, в смисъл на огромния физически обем на данните;
- **Скорост**, от една страна това е скоростта на прираст на данните, от друга - необходимостта от високоскоростна обработка на данните и получаване на резултатите в реално време;

- **Многообразие**, в смисъл на предоставяне на средства за едновременна обработка на различни типове данни като глас, видео, данни, генерирани от устройства, транзакции, блогове, информация от сайтове и социални мрежи, логове на сървърите на различни системи;



Фигура 1. Елементи на модела „3Vs”

„Големите данни” оперират с набори данни, които са толкова големи и сложни, че е трудно, почти невъзможно, да бъдат съхранявани и обработвани с помощта на традиционните релационни бази данни и SQL приложения за обработка на данни.

Други три „V” фактора определят качеството на входните данни и съществено влияят на избора на технология за обработване на „Големите данни” [2]:

- **Истинност (Veracity)** - Осигуряване на истинността на данните. Предварително филтриране на данни с недопустими отклонения, шум и аномалии. Това е голямо предизвикателство, особено когато обемите са големи и скоростите на обработка са много високи. Във всяко конкретно приложение, трябва да бъде заложена стратегия за осигуряване на верни данни. Да бъде предпазена системата от натрупване на грешни данни;
- **Валидност (Validity)** – Осигуряване на валидността на данните. Валидност означава коректно и точно използване на данните по предназначение. Тя е от ключово значение за вземане на правилни решения;
- **Променливост (Volatility)** – Променливостта на „Големите данни” се отнася до това колко дълго данните са валидни и колко дълго трябва да бъдат съхранявани. В нашия свят на анализ на данните в реално време трябва да може да се определи в кой момент данните вече не са от значение за настоящия анализ.

Десетте проблема, които много успешно се моделират, изследват и решават с помощта на „Големите данни” са [4]:

1. Моделиране на действителния риск (*Modeling true risk*).
2. Изграждане на индикатори и стратегия за запазване на съществуващите клиенти и привличане на нови (*Customer churn analysis*).
3. Създаване на генератор на препоръки/предложения (*Recommendation engine*).
4. Насочване на рекламата към подходяща целева група (*Ad targeting*).
5. Анализ на транзакциите и операциите с постерминали в реално време (*PoS transaction analysis. Real-time Business Activity Management (BAMS)*).

6. Анализ на данни по мрежата за предсказване на колизии (*Analyzing network data to predict failure*).
7. Анализ на заплахите (*Threat analysis*).
8. Наблюдение на търговията (*Trade surveillance*).
9. Търсене на качеството (*Search quality*).
10. Програмни средства за оценка на инвестициите в „Големите данни“ (*Data sandbox*).

„Големите данни“ вече са навлезли в много сектори на живота, както може да се види от фиг. 2:



Фигура 2. Области и дейности, свързани с „Големите данни“

Hadoop и „Големите данни“

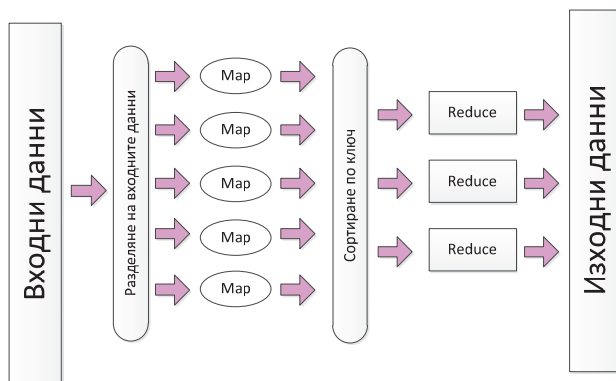
Анализът на „Големите данни“ с помощта на проекта с отворен код Apache Hadoop бързо се утвърждава като предпочитаното решение за справяне с бизнес и технологични тенденции, които разрушават традиционната представа за управление и обработка на данни. Предприятията, които отдавна използват „Големите данни“ получават голямо конкурентно предимство и просперитет.

Hadoop е мащабируема, отказоустойчива разпределена система за съхранение и обработка на данни, много подходяща за работа в облак [1]. Тя е съставена от два основни компонента:

- Hadoop Distributed File System (HDFS) е многократно осигурена, високо надеждна, клъстеризирана файлова система, самолекуваща се от грешки и разрушаване (self-healing), с висока пропускателна способност, пригодена за работа с много големи файлове. Оперира със структурирана и неструктурирана информация. Счита се за една от най-добрите файлови системи за работа с големи файлове в реално време в режим поток

(stream). Файловете в HDFS се записват само веднъж (write once). Отпада функционалността „Архивиране”.

- MapReduce (фиг. 3) е метод за разделяне на дадена задача на части (фаза Map) и разпределянето им между виртуалните машини на кълстера. Всяка виртуална машина решава паралелно с останалите своята част от задачата. Втората фаза (Reduce) събира резултатите и ги обединява в окончателен резултат. MapReduce е мащабируем, посредством увеличаването или намаляването на броя на въвлечените в обработката виртуални машини.



Фигура 3. Принципна схема на метода MapReduce [1]

Обучението на специалисти по „Големи данни” е многостранен и продължителен процес, но изучаването на някои специфични елементи може да бъде включено, макар и в опростен вариант, в съществуващите учебни програми по компютърни науки. Предложеният по-долу пример илюстрира как алгоритъмът на метода MapReduce може да бъде представен в рамките на конвенционален курс по програмиране.

Задача: Да се определи къде и кога е измерена най-ниската и най-високата денонощна температура за даден регион и за определен период от време.

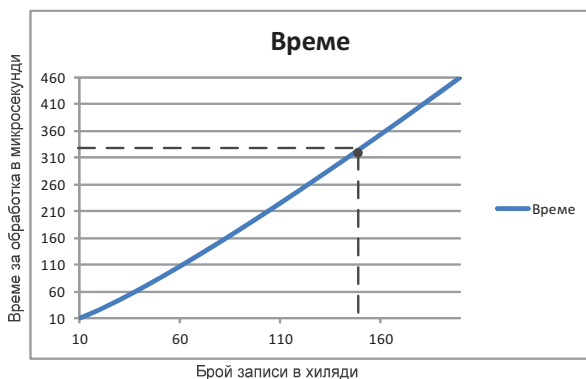
Можем да отнесем тази задача към обработка на „Големи данни”, тъй като ако броят на метеорологичните станции е K , броят на измерванията на температурата за едно денонощие е H , а броят на денонощието е N , то броят на записите V ще е равен на $K \cdot H \cdot N$.

Ако приемливото време за отговор е t сек., а производителността на една машина в брой записи/микросек. е както е показана на Фиг. 4. Да се намери броя на необходимите виртуални машини за извършване на изчисленията за време $\leq t$.

Използване на метода MapReduce за решаване на задачата:

Да приемем, че $K=148$ станции, $H=24$ измервания, $N=36525$ дни и $t \leq 1$ секунда. Общият брой на измерванията (записите) ще бъде равен на 129 736 800.

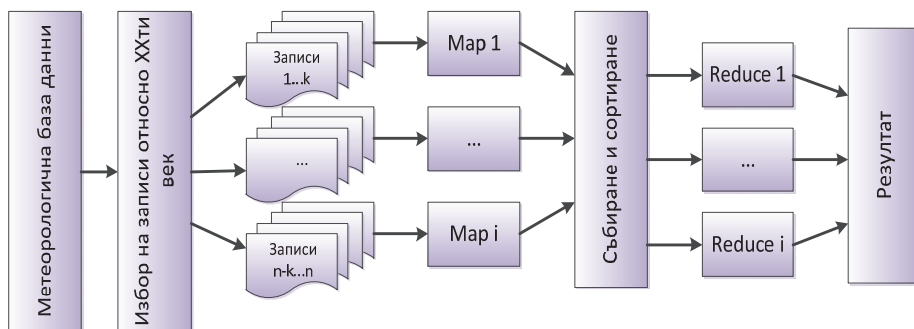
Знаем, че времето за отговор се формира от сумата на времето за вход, времето за Map фазата, времето за същинската обработка, времето за Reduce фазата и извеждане на резултата. Ако (Input Time + Map Time + Reduce Time + Output Time) съставлява 68.7% от общото време, за Processing остават 31.3% или 313 микросек.



Фигура 4. Производителност брой записи/микросек

Графиката на фиг. 4 отчита зависимостта между броя записи на вход (в хиляди) и времето (в микросекунди), необходимо на приложението да ги обработи. Отчита се, че за 313 микросек. приложението може да обработи максимум 145 000 записа. Следователно Map фазата ще стартира паралелно $129\,736\,800 / 145\,000 = 895$ виртуални машини.

В зависимост от обема на входните данни MapReduce моделът изчислява колко виртуални машини е необходимо да бъдат стартирани паралелно, за да бъде спазено времето за отговор. Това свойство на модела се нарича мащабируемост или еластичност. MapReduce моделът използва точно толкова ресурси, колкото изисква конкретната задача. Ако Reduce фазата също е трудоемка, то тя може да се паралелизира с оглед икономия на време (фиг. 5).



Фигура 5. Схема на реализация на задачата с MapReduce

ЗАКЛЮЧЕНИЕ

Търсенето на специалисти в областта на „Големи данни“ е налице и ще нараства в бъдеще. Подготовката засяга широка гама научни области като информатика, социология, политика, статистика, екология, застрахователно дело, медии, WEB и др. Това налага актуализация на учебните програми с оглед включване на базови знания, съответстващи на основните принципи, заложиени в „3Vs“ модела. Това ще създаде предпоставки за плавен преход и предлагане на курсове по Наука за данните (data science, datalogy), Облачни технологии (Cloud

technology) и Големи данни (Big Data). Такива курсове вече се предлагат в много университети по света.

ЛИТЕРАТУРА

- [1] Holmes, A., Hadoop in Practice, Manning, Shelter Island, 2012.
- [2] Leskovec, J., A. Rajaraman, J. D. Ullman. Mining of Massive Datasets, 2014
- [3] North, M., Data Mining for the Masses, Global Text Project Book, Amazon.com, ISBN: 0615684378
- [4] The real-world use of big data, www.ibm.com/services/us/gbs/.../ibv-big-data-at-work.html
- [5] <http://technews.bg/article-69884.html#.VDPdczscTIU>, посетен на 09.04.2014 г.

За контакти:

доц. д-р Пламенка Христова, Катедра "Информатика и информационни технологии", Русенски университет "Ангел Кънчев", тел.: 082-888 326, e-mail: ptx@ami.uni-ruse.bg

гл. ас. д-р Галина Атанасова, Катедра "Информатика и информационни технологии", Русенски университет "Ангел Кънчев", тел.: 082-888 326, e-mail: gea@ami.uni-ruse.bg

хон. преп. Пано Панов, Бакалавърски факултет, НБУ, e-mail: pano.panov@gmail.com