НАУКОМЕТРИЧЕН АНАЛИЗ В ОБЛАСТТА НА BIGDATA[2]

Академик Иван Попчев
докторант Милена Георгиева

SCIENCEMETRIC ANALYSIS IN THE "BIG DATA" AREA

Анотация: *„Големите данни" (Big Data) са един от актуалните термини в компютърната и комуникационна техника. С понижаването на цената на хардуера все повече компании могат да си позволят да съхраняват данни с обем в затабайтове. Цел на това изследване е да проследи актуалността на „Големите данни" през последните десет години. Разработката беше разделена на три отделни етапа. През първият етап бяха дефинирани десет термина, за които се извършиха търсения в четири популярни бази от данни с публикации. На втори етап от изследването бяха включени още три бази от данни, а ключовите думи бяха окрупнени. На трети, последен, етап бяха въведени още четири термини. Целта на това проучване е да изследва развитието на парадигмата за големите данни пред последните десет години.*

Abstract: *The data which is generated and collected nowadays has a volume which requires developing a special paradigms. „Big Data" is the today's buzzword. With the lowering the hardware prices more companies, are able to keep data bigger than Zeta bytes. The aim of this research is to follow how common the „Big data" paradigm is. The work over this material was divided into three parts. During the first part ten different keywords were defined, which were used for a search in four popular scientific data bases. During the second stage of the research three new data bases were used and the keywords were consolidated. During the last stage of the research four new keywords were introduced. The aim of this research is to follow the development of the paradigm of "Big Data" during the last ten years through the methods of the sciencemetric.*

Keywords: *Big data, data science, Data mining, Educational Data mining, Hadoop, MapReduce, Science direct, Google Scholar, Microsoft academic research, In memory computing.*

Увод

Днешната реалност все повече се трансформира в среда ориентирана към данните. Понижаването на цените на хардуера позволява на компании, дори с относително ограничени финансови възможности, да си позволят да съхраняват данни, измерени в Зетабайтове. Наличието на данни с такъв обем, предполага развитие на технологиите за обработката им с цел оптимизиране на процесите генерирали данните. Като следствие от това през последните десет години се наблюдава повишен интерес към така наречените data centric технологии. Целта на това изследване е да проследи развитието на отделни термини, имащи отношение към парадигмата „Големите данни" (Big Data) и „Наука за данните" (Data Science). Беше определена времева рамка от десет години (с начало 2005 год. и край 2015 год.).

Изследването се проведе в три отделни етапа. През първият етап се дефинираха десет термини, за които се извършиха търсения в четири популярни бази от данни за

[2] Работата е по научноизследователски проект на БСУ - Д05.

публикации. На втори етап от изследването се включиха три бази от данни, а ключовите думи бяха окрупнени. На трети, последен, етап се въведоха още три термина.

Първи етап на изследването

Термините, избрани за първия етап, имат за цел да се обхване същността на парадигмата на база трите V-та дефиниращи големите данни. Следните термини бяха избрани за изследване през първия етап на изследването:

- *Big data* – Иновативен подход за събиране, обобщение и обработване на данни, пристигащи от различни източници в почти реално време. Основана характеристика на „Големите данни“ е количеството данни, което по дефиниция трябва да надхвърля количеството данни, което може да се съхрани в традиционните бази от данни [2];
- *Educational data mining* (EDM) – Подход при обработване на данни, генерирани от среди за електронно обучение. Целта, която си поставя подхода е събиране и обработване на тези данни чрез инструменти и алгоритми, предоставени от дейта майнинга [4];
- *Data mining* – Процес на извличане на ценна и неочевидна информация в големи масиви от данни [3, 4];
- *Hadoop* –Архитектура устойчива на отказ, която разпределя и съхранява информацията в повече от един възел (node) и в случай на отказ на някой от тези възли, архитектурата сама открива изпадналия възел и копира информацията на друг. [1,2];
- *MapReduce* – Програмна парадигма, чиято цел е да намали попадналата в системата информация, като преди това я разпределя. [1, 3];
- *E-learning* – електронно обучение;
- *Student performance* – постижения на студенти;
- *Data science* – Относително нов подход за обработване на информация, който съчетава статистически и дейта майнинг методи. [3];
- *Big data analytics* – анализ на големи данни;
- *Learning analytics* – анализ на обучението.

За целите на проучването бяха избрани добре известни портали като Science Direct, Google Scholar, IEEE Xplore, Springer.

Резултати от първи етап на проучването

Данните за проучването бяха изтеглени на дата 03.02.2015 год. Първоначално беше пуснато търсене за термина „Big Data” за съответната година и във всички бази от данни.

Резултати в Science Direct към 03.02.2015											
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Big data	19758	22648	25243	27167	29051	30121	34458	37582	42214	50331	19607
Educational data mining (EDM)	386	439	482	509	623	681	774	923	1111	1278	499
Data mining	6192	7225	7939	8758	9925	9932	12316	13795	15914	19181	7611
Hadoop	-	-	1	1	3	16	30	79	189	241	146
MapReduce	1	2	2	3	7	24	71	109	218	288	167
E-learning	3408	3813	4562	4950	5497	5562	6660	7399	8329	9896	4530
Student performance	1138	1269	1588	1688	1829	1929	2137	2480	2915	3451	1465
Data science	7801	8927	10055	22005	11520	11330	13529	14757	16751	18588	7435
Big data analytics	15	20	42	26	29	51	63	155	352	525	356
Learning analytics	8	22	56	34	46	66	107	233	366	532	365

Таблица 1. Резултати в Science Direct

От таблица 1 става ясно, че броя на публикациите, включващи термина Big Data, бележат чувствително повишение през последните десет години. Трябва да се отбележи, че термините Hadoop и MapReduce бележат възход едва през последните четири години. Именно комбинацията от „Big Data“, „MapReduce“ и „Hadoop“ представляват сериозен научен интерес за следващи разработки. Следващите два термина „E-learning“ и „Student performance“ бяха подбрани заради областта на компетентност, която ще бъде засегната в същинската теза. От получените резултати ясно се вижда, че електронното обучение бележи повишен интерес през последните десет години.

„Data Science“ и „Big Data analytics“ са взаимно заместващи се термини, но от получените резултати може да се направи изводът, че „Data Science“ е предпочитан термин, също така интересът към „Науката за данните“ чувствително се повишава от 2007 година насам.

В таблица 2 са представени резултатите, получени от базата IEEE Xplore. Тук се наблюдава интересен феномен, върнатите резултати за 2014 год. по изследваните ключови думи броя рязко намаляват. Тук за пръв път през 2007 год. се публикува материал имащ отношение към Hadoop. Резултатите отново показват нарастване на публикациите с ключови думи „Big Data”, „Data Mining”, „Data Science” и „Educational Data mining”.

Резултати в IEEE Xplore към 03.02.2015											
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Big data	**7228**	**9079**	**9799**	**12360**	**15184**	**16620**	**15053**	**15053**	**16408**	**7124**	**302**
Educational data mining (EDM)	**777**	**1017**	**1214**	**1798**	**2830**	**2189**	**2118**	**2648**	**2590**	**1231**	**46**
Data mining	**11149**	**13512**	**15577**	**18867**	**22988**	**23206**	**20345**	**20267**	**20765**	**10176**	**507**
Hadoop	**-**	**-**	**6**	**51**	**137**	**368**	**580**	**931**	**1342**	**697**	**21**
MapReduce	**1**	**6**	**24**	**85**	**201**	**463**	**705**	**1051**	**1418**	**753**	**33**
E-learning	**880**	**1125**	**1190**	**1712**	**2131**	**2593**	**2151**	**2334**	**2198**	**856**	**17**
Student performance	**7212**	**8835**	**9844**	**11177**	**13343**	**15379**	**16198**	**17673**	**18283**	**9035**	**568**
Data science	**41650**	**52061**	**57960**	**71203**	**88265**	**100515**	**94018**	**89769**	**86570**	**28227**	**1450**
Big data analytics	**20**	**45**	**67**	**80**	**139**	**179**	**252**	**562**	**1363**	**838**	**47**
Learning analytics	**62**	**127**	**181**	**196**	**318**	**470**	**687**	**1108**	**1771**	**816**	**41**

Таблица 2. Резултати в IEEE Xplore

Следваща посетена база е на издателство Springer. Тук върнатите резултати са доста по-малко на брой, въпреки това релевантността им е изключително висока (табл.3). Публикуваната литература има висока научна стойност, което би могло да обясни малкото на брой публикации.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Big data	6943	54	45	49	59	87	125	375	1,572	3285	990
Educational data mining (EDM)	-	9	11	15	18	56	66	64	86	132	21
Data mining	4570	5283	8873	9251	11666	11714	12235	13240	12793	14101	2462
Hadoop	-	-	3	12	70	138	209	329	447	785	164
MapReduce	4	2	10	25	97	171	268	387	503	818	143
E-learning	966	1321	1416	1749	1742	1820	2006	1870	1866	2303	722
Student performance	372	351	535	468	621	565	817	853	1023	1339	460
Data science	202	130	144	171	193	137	291	238	196	491	100
Big data analytics	-	-	-	-	-	-	-	38	154	295	93
Learning analytics	-	-	-	-	-	-	7	19	88	208	59

Таблица 3. Резултати в Springer

Последната разгледана база през първия етап на проучването е Google Scholar. Съвсем очаквано върнатите резултати са в пъти повече от другите бази. Търсещата машина на Google Scholar притежава един от най-добрите алгоритми за търсене и този факт би обяснил броя на върнатите резултати за „Big Data Analytics" и „Learning Analytics". При резултатите за „Data Mining" се наблюдава ясна тенденция на увеличаване на публикациите по тематиката, но след 2011 година тенденцията се обръща (табл. 4). След 2014 година спадът е ясно различим. Тенденцията отбелязва повишаване броя на публикуваните материали имащи отношение към „Hadoop" и „MapReduce".

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Big data	214 000	224 000	224 000	218 000	207 000	200 000	178 000	169000	143000	131000	44000
Educational data mining (EDM)	14 900	16 500	18 700	20 400	23 500	23 700	25 600	27 900	27 500	25 200	3 700
Data mining	190 000	209 000	206 000	224 000	241 000	260 000	234 000	207 000	142 000	96 900	34 400
Hadoop	180	256	223	1 000	1 310	3 190	4 670	5 840	9 980	11 500	1 020
MapReduce	363	547	531	1 210	1 990	4 000	5 600	7 330	10 800	12 300	1 340
E-learning	25 800	29 000	35 200	40 900	45 500	44 300	45 800	53 700	43 600	39 300	5 600
Student performance	230 000	248 000	276 000	274 000	264 000	264 000	257 000	197 000	161 000	97 100	45 300
Data science	1 080	1 320	1 610	1 720	2 040	3 540	4 480	5 870	8 630	12 400	2 340
Big data analytics	4	6	6	9	14	25	103	584	1 890	3 840	643
Learning analytics	14	15	14	21	19	56	249	734	1 370	2 190	279

Таблица 4. Резултати от Google Scholar

Втори етап на изследването

След обсъждане на междинните резултати, се взе решение да се доразвие изследването, като бъдат включени нови бази от данни и също така се доразвият термините.

Във втората фаза на изследването бяха добавени следните бази от данни academia.edu, Microsoft Academic Research, ACM Digital Library. Използваните ключовите думи са: „Big Data", „Big Data+Analytics+Data Mining" и „Data Science".

Резултати от втори етап на изследването

В сайта www.academia.edu бяха търсени следните термини със съответните резултати (към 14.04.2015 год.): Big data – 1377; Big Data + Analytics + Data Mining – 2638; Data Science - 2341; За съжаление не беше открит способ за разделяне на резултатите по година.

В сайта на Microsoft Academic Research за термина:

- Big data върнати 238 публикации за година 2005;
- Big data върнати 244 публикации за година 2006;
- Big data върнати 320 публикации за година 2007;
- Big data Върнати 339 публикации за година 2008;
- Big data върнати 371 публикации за година 2009;
- Big data върнати 364 публикации за година 2010;
- Big data върнати 326 публикации за година 2011;
- Big data върнати 125 публикации за година 2012;.
- Big data върнати 4 публикации за година 2013;
- Big data върнати 2 публикации за година 2014;

Изводът, който може да се направи е, че въпреки резултатите, получени за периода 2012-2014 година, сочат за спад на публикациите по темата, то в наблюдавания период може да се отчете нарастване на броя публикации в тази база от данни.

При направеното търсене със следните термини Big Data & Analytics & Data Mining, получените резултати имат следния вид:

- 2005 год. – 77 бр.;
- 2006 год. – 80 бр.;
- 2007 год. – 103 бр.;
- 2008 год. – 108 бр.;
- 2009 год. – 109 бр.;
- 2010 год. – 108 бр.;
- 2011 год. – 94 бр.;
- 2012 год. – 41 бр.;
- 2013 год. – 1 бр.;
- 2014 год. – 2 бр.

Въпреки значително по-малкото на брой резултати прави впечатление, че откритите публикации са изключително релевантни към търсенето.

Следва списък на попаденията за комбинация между следните термини Big Data, Analytics и Data Mining:

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	77	80	103	108	109	108	94	94	1	2	1

Таблица 5. Резултати за Big Data, Analytics и Data Mining

Комбинацията между трите ключови думи се оказа с не добра популярност и както се вижда от резултатите в таблица 5 се наблюдава тенденция на спад в броя им.

Това проучване завършва с базата от данни ACM Digital Library, като данните от изследването са представени в таблица 6. Тази база от данни е услуга предоставена от Associated for Computing Machinery. **ACM DL** е пълнотекстова колекция от всички публикации появявали се на конференции, в списания или статии.

Big Data:

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	4687	5178	5943	5804	6207	7964	8354	9008	9193	9047	1565

Таблица 6. Резултати от ACM Digital Library по ключова дума Big Data

В следващата таблица 7 са представени данните за комбинация между ключовите думи Big Data, Analytics и Data mining. Ако сравним получените резултати с предходната база, тук се наблюдава повишаване броя на публикациите имащи отношение към избраните ключови данни.

Big Data + Analytics + Data Mining:

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	9	18	29	51	109	109	147	326	517	636	115

Таблица 7. Резултат за Big Data, Analytics и Data Mining

Интересен резултат се получи при търсене по ключови думи „Data Science“ за 2005 година, а именно 40 219 попадения. Този резултат е в пъти повече от резултатите получени от другите бази от данни (пр. academia.edu – 2341 попадения). Този резултат подсказва, че търсенето не е прецизирано достатъчно и е необходимо ключовата дума да бъде променена. Поради което беше направен опит за навлизане в дълбочина на базата, а именно чрез ключови думи Data Science Bayes Classifier. След промяна на ключовите думи резултатът също се промени, а именно 4486 попадения (табл.8). Беше направен още един опит за подобряване на ключовите думи, а именно добавяне на „missing data”. Резултатите, получени за 2005 година, са 656 попадения.

Година	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	771	894	941	1051	1294	1400	1535	1334	1186	202

Таблица 8. Резултати за Data Base Bayes Classifier

Тази комбинация беше избрана поради актуалността на проблема с възстановяване на липсващи данни, което в един често срещан проблем при обработване на голямо количество данни.

Трети етап на изследването

През третия етап на изследването бяха добавени нови термини на изследване, а именно:

- BigTable – е технология патентована от Google. Представлява система за съхранение на данни с висока производителност. Изградена е на основата на Google File System и още няколко технологии собственост на Google;
- NoSQL – Представлява парадигма за съхранение на данни, концептуално различна от традиционните релационни данни. Името NoSQL (Not only Structured Query Language) има смисъл на нерелационна база от данни. Съществуват няколко подхода за съхранение на информацията: ключ-стойност, графови бази от данни, документ и колона [3];
- In memory computing – Нов подход за обработване на данни. Важното тук е, че цялото количество от данни се зарежда директно в RAM паметта на машината; където се прилагат различни подходи за обработката.

- SAP HANA – Разработка на немската компания SAP. Възприетият подход при тази база от данни е колонна. Това ще рече, че за разлика от подхода при релационните бази от данни, при SAP HANA данните се обработват не по редове (записи), а по колони. Този подход успява, според SAP за една секунда да изчисли всички платени данъци до текущия момент от цялото човечество. [6]

Изследваните бази са: Science direct и Google Scholar. Тези бази бяха избрани поради количеството върнати публикации, а също така и заради възможността да се прави прецизно търсене.

Резултати от трети етап на изследването

Следват резултати от запитвания към двете бази

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	325	359	461	626	796	1210	1590	1850	2120	2010	699

Таблица 9. Резултати по ключова дума BigTable в Google Scholar

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	33	38	48	136	70	440	941	2010	3670	4620	1520

Таблица 10. Резултати по ключова дума NoSQL в Google Scholar

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	**6**	5	10	5	13	20	66	140	251	304	118

Таблица 11. Резултати по ключова дума In memory computing в Google Scholar

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	**65**	86	84	90	120	131	226	418	693	736	383

Таблица 12. Резултати по ключова дума SAP HANA в Google Scholar

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	16598	19169	21667	23030	24624	23813	26479	28003	31791	36146	27200

Таблица 13. Резултати по ключова дума BigTable в Science Direct

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	-	-	-	-	-	1	7	24	87	117	1520

Таблица 14. Резултати по ключова дума NoSQL в Science Direct

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	4170	4719	5206	5491	5817	6126	7100	7978	9199	9872	7917

Таблица 15. Резултати по ключова дума In memory computing в Science Direct

Година	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Върнат резултат	8	12	5	5	4	3	10	3	13	22	20

Таблица 16. Резултати по ключова дума SAP HANA в Science Direct

От направения анализ може да се заключи, че през последните години се наблюдава завишаване на броя публикувани материали, свързани както с технологиите, така и с хардуера, теории, парадигми и софтуер. Интересен факт е, че SAP HANA все още не е често срещана в публикациите.

„In memory computing“ може да бъде определен като относително популярен термин и към момента върху проблематиката се работи усилено. Същевременно, интересна находка са резултатите за термина „NoSQL”. Резултатите от Science Direct най-вероятно се дължат на факта, че работейки върху една база, е необходимо да се споменава конкретния подтип.

Върнатите резултати за „BigTable” от базата Science Direct повече в сравнение с Google Scholar. В една следваща разработка би следвало да се изследват комбинация от ключови думи с цел прецизиране на резултати.

Изводи

Получените резултати в изследването могат да се използват като потвърждение на твърдението, че към текущия момент „Големите данни“ и „Науката за данните“ са „върхът на вълната“ и съчетаването на тези две парадигми с електронното обучение ще доведе до подобряване на качеството както на образованието, така и на образователната система.

Наблюдава се концентрирана работа в областта както на архитектурите, така и на софтуерното обезпечение на парадигмата. Съществуват редица нерешени проблеми, но интересът е с постоянна тенденция на покачване.

Разработват се разнообразни подходи, съществуват голямо количество теории, имащи отношението към „Големите Данни“. Количеството публикации е необхватно, базите данни са също така много на брой.

Изследването ще подпомогне вземането на решение за избора на архитектура, алгоритми и софтуер за обезпечаване на изследвания в областта на електронното обучение и обработката на данни генерирани от системи за управление на електронно образование.

Литература:

1. Plunkett, Tom, et al. Oracle Big Data Handbook, Oracle press, 2014
2. Berman, J. Principles of Big Data, Morgan Kaufmann, 2013
3. EMC Education Service, Data Science & Big Data Analytics,Wiley, 2015
4. Pena-Ayala Al., Educational Data Mining, Springer, Switzerland, 2014
5. Mayer-Schoenberger V., Kenneth Cickier, Big Data die Revolution, die unser Leben veraendern wird, Redline Verlag, Muenchen, 2013
6. SAP, SAP HANA Participant guide, SAP, 2014