ГОЛЕМИ ДАННИ. ВЪВЕДЕНИЕ

Големите данни се появяват в различни приложения и от различни източници. Те са записани във файл или поредица от файлове, като данните в тях не са структурирани. Manayika (Manayika at al.:2011) дефинира големите данни като „количество данни, надхвърлящо възможностите на днешните технологии да ги обработят, съхранят и изчислят ефективно“. Наред с това, в Zikopoulos (Zikopoulos at al.:2012) и в (Berman at al.:2013) големите данни се определят посредством три характеристики: обем, многообразие и честота. Тези три характеристики за първи път са представени от Gartner (Gartner:2012) и с помощта им се описват предизвикателствата в областта на големите данни. Терминът *големи данни* може да се опише освен с трите характеристики на Gartner (https://www.gartner.com/en): обем, многообразие и скорост (volume, variety, velocity), също така и с четвърта характеристика – стойност (value). Така нареченото определение 4V (Четирите V) е най-широко прието, тъй като пояснява значението на големите данни и необходимостта от тях (Gantz at al.:2011).

Кратки характеристики [Big Data Approach and its applications in Various Fields: Review, 2019, https://doi.org/10.1016/j.procs.2019.08.084]

‘Big Data’ can be defined by the huge volume of data generated, in real time, from various digital sources like sensors, smartphones, social media and others. These data can have many different types as videos, audio, images, text and so on [Big Data Analytics and Its Applications, 2017, https://arxiv.org/abs/1710.04135].
Several researchers have defined ‘Big Data’ by its Volume, Velocity and Variety (3Vs definition). That is means respectively, according to authors in [4], data size is large, the data will be created rapidly, and the data will be collected in multiple types and captured from different sources:

- Volume: Denotes the large amount of data which is generated every second from different sources.
- Velocity: Measures the speed at which data can be collected, analyzed and exploited.
- Variety: Means that the incoming data can have different types as: Structured, Semi Structured, Unstructured and Multi Structured.

Lakshen and al. talked about 5Vs definition of big data instead of 3Vs. They added in [5] Value and Veracity:

- Value: The useful data among this large Volume of data
- Veracity: Denotes the correctness and accuracy of the data.

Also, according to research in [6], four other characteristics of Big data are broached to build 9Vs:

- Variability: It refers to data whose meaning is constantly changing.
- Validity: Means the data is correct and accurate for the intended use.
- Volatility: Means how long does company need to store data.
- Visualization: Presentation of data in readable and accessible manner for better decision making

To summarize, Big data are massive and rapidly-expanding, but it's also noisy, messy, constantly-changing, in hundreds of formats and virtually worthless without analysis and visualization.

[4] J. Campos, P. Sharma, U. G. Gabiria, E. Jantunen, and D. Baglee, "A Big Data Analytical Architecture for Asset Management," Procedia CIRP, vol. 64, pp. 369–374, 2017. https://doi.org/10.1016/j.procir.2017.03.019

[5] G. A. Lakshen, S. Vranes, and V. Janev, "Big data and quality: A literature review," 2016 24th Telecommun. Forum, pp. 1–4, 2016. https://ieeexplore.ieee.org/document/7818902

[6] S. S. Owais and N. S. Hussein, "Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 3, pp. 254–258, 2016. https://pdfs.semanticscholar.org/d204/960c9ee540630b444afcdfe5c0509baa9e4c.pdf

Първата цел на специалистите, анализиращи големи данни, е да анализират данните и в резултат на това да се получат структурирани данни, независимо от тяхната дължина. Може да си представим структурираните данни, като таблица, на която колоните са характеристиките на данните, а редовете са самостоятелни наблюдения и всяко наблюдение притежава в някаква степен съответната характеристика. Ако за дадено наблюдение няма информация доколко притежава някоя характеристика, то съответното поле остава празно и се интерпретира, че има липсваща стойност (missing value) или стойност nan. Следващата стъпка при анализа на структурираните данни е да се разделят наблюденията на класове, т.е. изпълнява се процес на клъстъризация на данните. Тази съвкупност от структурирани данни (независимо дали са разделени на класове или не) се нарича множество от данни, което може да се запише в различни видове файлове, като най-често използваните формати са текстовите файлове и csv

файловете, който формат се поддържа от MS Excel. Съществуват много интернет сайтове, които съхраняват различни множества от данни, заедно с техните описания. След като едно множество от данни е разделено на класове то може да се анализира в посока на създаване успешен прогнозен модел за наблюденията в множеството. Вместо термина множество от данни може да се използва множество от наблюдения.

Големите данни следват жизнения цикъл, показан на фигурата по-долу [фиг. 1] от генериране на данни до визуализация на данни чрез събиране на данни, съхранение на данни и анализ на данни

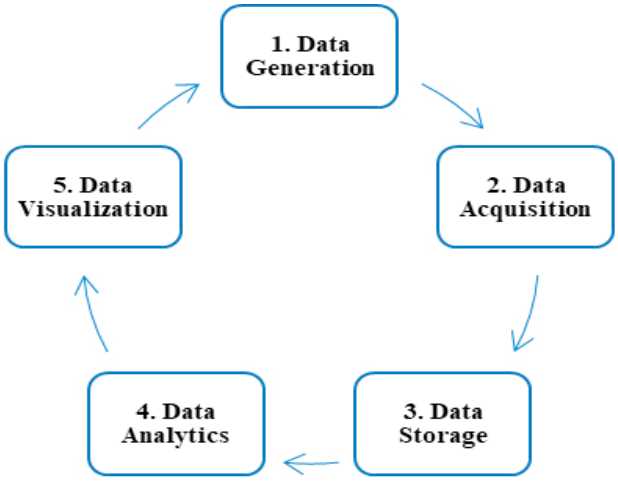

Fig. 1: Big Data life cycle

- Data Generation: Collect of data from various sources (sensors, video, click streams …).
- Data Acquisition: The process of obtaining information from data by:

(a) Selection data: Select pertinent data which is useful for the analysis

(b) Pre-processing data: detect, clean, and filter the unnecessary and inconsistent data.

- Data Storage: persistently storing.
- **Data Analytics**: Analytics refers to the process of deriving actionable insights using qualitative and quantitative techniques [C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," J. Big Data, vol. 2, no. 1, pp. 1–32, 2015. https://doi.org/10.1186/s40537-015-0030-3] [Fig. 2].

(a) Data Transformation: After Gathering, Selection and pre-processing data, transforming preprocessed data into data-mining-capable format is required.

(b) Data analysis: After transforming data, analysis can be done using various statistical methods and data mining algorithms such as regression, classification, clustering [M. Dave and H. Gianey, "Different clustering algorithms for Big Data

analytics: A review," Proc. 5th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2016, pp. 328–333, 2017, https://ieeexplore.ieee.org/document/7894544].

- Data Visualization: Representation of data insights in an interactive way going through:

(a) Evaluation: Measure the results of data analysis;

(b) Interpretation: displaying the output of data analysis by an interactive way.

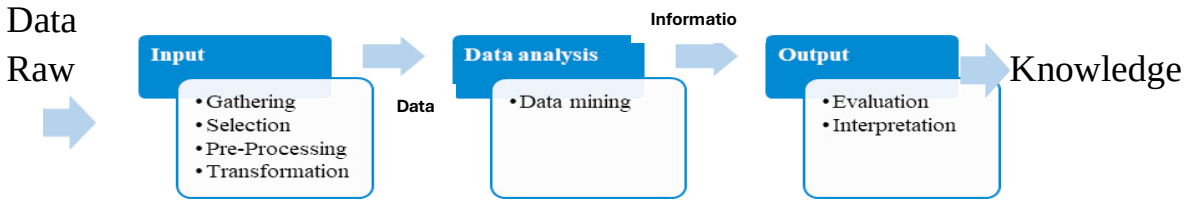

Fig 2 Data analytics process

Big Data Analytics [Big Data Approach and its applications in Various Fields: Review, 2019, https://doi.org/10.1016/j.procs.2019.08.084]

Nowadays, the data that need to be analyzed are big, contain heterogeneous data types, and even include streaming data which may change the statistical and data analysis approaches. Therefore, Traditional tools cannot be able to analyze this category of data. [10]. So, New approach of big data called 'Big Data Analytics' was born.

"Big data analytics" refers to advanced technologies designed to work with large volumes of heterogeneous data to improve the traditional Data Analytics Process mentioned in [Fig.2].

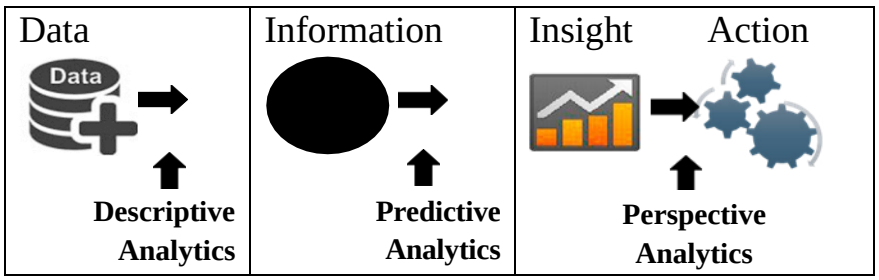

Fig 3. Types of Analytics Techniques

A lot of researchers talk about sophisticated types of analytics techniques as shown [Fig.3]: Descriptive Analytics, Predictive Analytics and Prescriptive Analytics:

- Descriptive Analytics: Gives information about What happened. In this technique, based on historical data,

new insights are developed using statistical descriptions (such as Statistic Summary, Correlations and

Sampling…) and Clustering (such as K-means…) [11].

- Predictive Analytics: Predicts the future outcomes using new statistical methods and predictive algorithms such as 'Decision Tree'. It provides information on what likely happens in the future and what actions can be taken [12].

- Prescriptive Analytics: It is a type of predictive analytics. It helps to derive a best possible outcome by

analyzing the possible outcomes by responding to the question So what? Now What?

[10] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics," IEEE Trans. Ind. Informatics, vol. 13, no. 4, pp. 1891–1899, 2017.

[11] G. Park, L. Chung, L. Khan, and S. Park, "A Modeling Framework for Business Process Reengineering Using Big Data Analytics and A Goal-Orientation," 2017.

[12] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015.

ПРИЛОЖЕНИЯ !

Big Data and Business Process Management (BPM) Synergy

Business Process (BP) is a succession of activities designed by humans and systems that intends to achieve business goals [14]. Business Process Management (BPM) is a way to collect and treat data outcoming from processes in real time to support decision making. The lifecycle of BPM Project contains successive steps as it's shows in [Fig. 4].

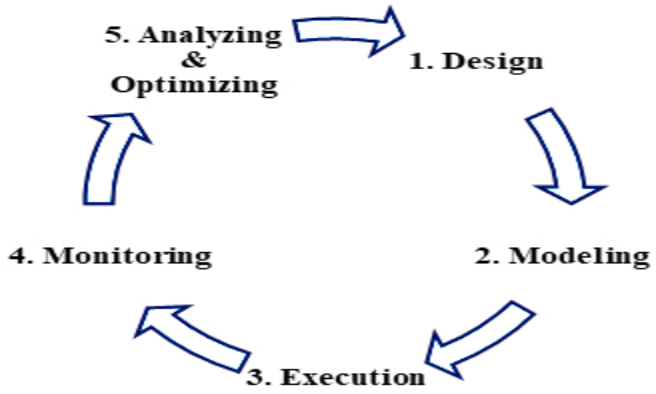

Fig. 4: BPM lifecycle

- Design: Modeling the design of what is currently being used and what will be used.
- Modeling: Analyzing process and performing "what if" analysis and then comparing the various process options to determine optimal improvements.
- Execution: Once the processes have been designed and simulated, they will be integrated into the information system for execution.
- Monitoring: Managing and supervising the processes.
- Analysis and optimization: Iterate for continuous improvement.

By the appearance of Big Data, organizations will need to integrate mobile data, social networks, digital video, and sensor data into their Business Process. So, they must be aware of Big Data challenges to make an efficient and intelligent BP that aims to bring huge value for process decision makers and process actors. Some decisions are based on subjective judgment however, increasingly important decisions need to be based on hard data based on big data analytics.

This is why big data analytics play a critical role in BPM by [A. Hassani and S. Ayachi, "ScienceDirect ScienceDirect A framework for Business Process Data Management based on Big A frameworkfor Business Data Process Data Management based on Big Approach Data Approach," Procedia Comput. Sci., vol. 121, pp. 740–747, 2017, https://doi.org/10.1016/j.procs.2017.11.096] [Fig. 5].

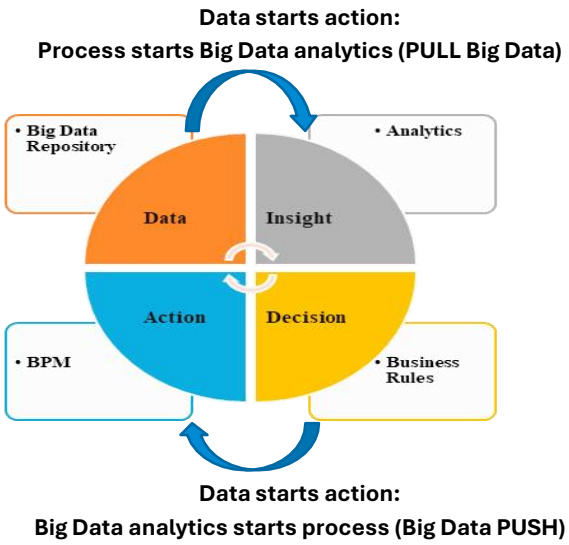

Fig. 5: Big data-BPM lifecycle

The big data opens up BPM to new data sources and opportunities to analyze processes more deeply and simulate potential improvements:

- Big Data Repository: refers to the large volume of data from different sources.
- Analytics: means a collection of reports, graphs and other analytical capabilities (Computation, Visualization...)
- Business Rules: After analytics, Managers can create "business rules" to make decisions that can be exercised

through human judgment or fired automatically.

- BPM: Once all the decisions are made, the design and execution of the BP, in the opportune moment, can be triggered.
- Learning from experience: The massive amount of data transformed into actionable knowledge needs to be stored and analyzed in order to learn from experience and to continually change and improve the algorithms used to support decision making based on real metrics.

To conclude, Big Data Analytics / BPM synergy appears clearly in [14]:

- More Insight: Generally, more analysis should lead to more actionable insight;

- Process Improvement: Some processes can explicitly benefit from the results of Big Data analytics. There is no place for instinctive feeling decision while now decision can be supported with large-scale analysis;
- Predictive Analytics: To predict which inputs contribute to successful outputs or the opposite;
- Learning from experience: Business Process work may have already identified key metrics and these data can be used as input for Big Data implementation instead of reworking on global big data;
- Process starts action (PULL Big Data): Process starts Big Data analytics;
- Data starts action (Big Data PUSH): Big Data analytics starts process.

Big Data and Human Resources Management (HRM) Synergy

Big Data in Human Resources Management refers to the large volume of employees, customers, and transactional data in organizations. By analyzing these Big data by specific tools, HR processes (Recruitment Process, Training Process, Employees Career Management Process…) become more relevant which help companies to make decisions and create bigger benefits.

- Recruitment Process: Sometimes traditional recruitment process leads to bad hires Because most of the time, the interviewer may not have correct information about candidates that's leads to false results. Combining big data collected and traditional recruitment process may help recruiter to search for potential talents and everything he would possibly want to know about them is on their profile (personal picture, living conditions, social relationships…etc.). HR manager can match the candidate's skills and personal beliefs and the company's needs. Hence, company can avoid invest in bad hires.

- Training Process: As it's known, talent training can lead to an increase in employees' level of knowledge and skill, it can also enhance their work performance. By traditional talent training organized by the company, professional trainers can be hired to ensure training which is spend a lot of material, human and financial resources. Usually, such training takes the traditional form of classroom instruction which cannot meet all the needs of employees. Using Big data context, any employee can easily search and access the information that he needs to know on the Internet at anytime and anywhere.

- Employees Career Management Process: By analyzing all the gathered information of employees such as: interest on job, professional experience, performance …, HR Manager could find new ways to motivate employees and make them more engaged. Companies can combine traditional career management and career management of Big Data to make planning of new more effective talent-retention programs and to avoid employee turnover.

Big Data Approach for Telecommunication sectors

A. *Big Data opportunities in telecom sectors*

- Boost customer experience: To increase customer loyalty, the operator needs to have an overview of its traceability: its web user account, its behavior towards his phone (web browsing, downloads...), his missed calls and its shares on social networks. Today, thanks to Big Data Approach, Telecom operators can collect this information from customers' mobiles. Operators can manage large subscriber databases that get richer each time a customer calls, creates a text message or uses the Internet. By using these databases, operators can:

a) Predict churn: Traditionally, Telecom operators could detect dissatisfaction of one of its subscribers if he has complained or if he has abandoned the service or thanks to purchase history. Today, with Big Data Analytics Techniques, operators can reveal new indicators of customers unsubscribing. They can identify those who are looking for new services on competitors' websites or posting negative contributions on a forum or on social networks. So, the operator can predict customers' churn by targeting "at risk" customers.

b) Detect upselling opportunities: Today, using Big Data Analytics Techniques, Telecom operators can get a global idea about purchase of their customers, they can propose to them a product or a service slightly higher and more expensive than their interests and make them aware of the ability to get even more of what he is looking for.

c) Identify cross-selling opportunities: Through e-commerce websites, Telecom operators can collect 'Big' Data on customers purchase. They can then make aware them of ways to accessorize their purchase.

- Network Optimization: Today, with big data analytics techniques, operators seek to collect, store and analyze data generated by user devices and

network devices like routers, switches, base stations, and so on. All these data and others Key performance Indicators (KPIs) can help operators to well monitor network performance to ensure proper operation without problem.

- Data Monetization: Some telecommunication companies see these massive volumes of data as a marketable resource to generate new revenue by aggregating and selling these data, they search to turn that data into money. Many telecom operators also seek to commercially exploit this customer information, i.e. to generate new revenue by aggregating and selling this data. Some of them consider this an excellent financial opportunity.

To sum up, promoting loyalty, anticipating and reducing churn, offering upselling and cross-selling, optimizing network and personalization services are key areas where telecom operators can take advantage of Big Data Analytics.

B. *Big Data challenge in telecom sectors*

Certainly, Big Data has opportunities in the telecommunications sectors. However, it also has divers' challenges such as: Big Data process, Real time Analytic, security and Revenue losses challenges.

- Big Data process and Analytics challenge: The complexity of the data to be managed today is a real

challenge. Traditional tools of treatment of data seem simple compared to the contextual big data collected.

- Security and data governance challenge: The volume and diversity characteristics of Big Data lead to a new level of complexity in data security when telecom operators integrate new sources of information.
- Revenue losses challenge: Because of the competitive environment, gradual and continuous loss of landlines and increase of social networks, telecom operators have lost a large amount of revenue.

In fact, Big Data is still in development and its related techniques and tools are far from mature. So, Human Resources Management, Telecom Sectors and others face also challenges in the use of Big Data in terms of Storage, Analytics and Management.

В зависимост от предметната област, от която произтичат данните се построяват *регресионни* модели, които да моделират зависимости между отделни характеристики на данните и да прогнозират бъдещи стойности на зависимата характеристика при промяна на стойности в независимата характеристика или при появило се ново наблюдение в съответната предметна област. Построеният модел трябва да прогнозира каква стойност да очакваме за зависимата характеристика (променлива) на новото наблюдение.

За други множества от данни е важно да се построи прогнозиращи модели, които да класифицират наблюденията от множеството и да формират класове от наблюдения в самото множество. Това са модели за *класифициране* на данните.

Областта която обработва големите данни се нарича machine learning. Под този термин се разбират алгоритми за анализ на (големи) данни. За последните 10 години алгоритмите за анализ на данни са придобили широко разпространение и дълбочина при анализа на данните. В последната година възникна необходимостта от „посредничество“ потребителя и алгоритмите за анализ на данни. С усилени темпове се развива направлението Generative AI (Artificial Intelligence) . Чрез тази технология алгоритмите за анализ на данни създават съдържание в различини форми като текст, програмен код, аудио, видео, кратки клипове и други (https://shorturl.at/Fiv6H).

Проф. Иван Иванов

Октомври 2024