The big data Ecosystem and its Environs

Babak Bashari Rad and Pouya Ataei
Asia Pacific University of Technology and Innovation Technology Park Malaysia, Kuala Lumpur, Malaysia

Summary
By the virtue of advanced devices, sensors, and social networks, big data arose to confront practitioners with a complete shift in the way they operationalize data. This has changed the context for many industries, and challenged leaders to adopt to big data ecosystem. This paper aims to explore big data ecosystem with attention to its architecture, key role players, and involving factors. In this study, firstly, the paradigm shift is discussed, and secondly the new ecosystem has been elaborated. This ecosystem is then dissected with attention to key role players, big data computation architecture, and skills required.

Key words:
big data, big data ecosystem, big data role players, big data traits and qualities.

1. Introduction

In recent times, through a shift from good-dominance to a service-oriented engineering, "Big Data" has emerged to challenge the industry. This concept has invited an extensive competition to practitioners looking for competitive advantage, escalated productivity, and added-on opportunity. Scholarly researches predicted that the big data is the "game-changing innovation" [1] is the "dawn of new industrial revolution" [2], is the "new category of economic asset" [3], and the "fourth paradigm of science" [4].With prodigious advancements in data generation by prevalence of smart phones, social networks, sensors, and large-scale service-oriented systems, today, big data is a far-reaching trend of information technology. Big data arose from the analysis through Information Systems (IS), Software Engineering (SE), and Business Analytics (BA) theories and practices [1]. Under the trajectory of big data, business intelligence and analytics transmuted with cutting-edge techniques and technologies so as to keep in line with market demand [5]. As a result, business processes faced major differentiators in organizations from various industries, and became increasingly analytic oriented to wring every last drop of value from these processes [6]. Withal, the hype is often vague, and so does little to address the far-reaching bottle-necks. This led to confusions among managers and executive and left individuals to make their own under informed and highly subjective understanding toward the big data, and what it can do and what it can't [7]. According to Gartner survey, big data is ascending toward the peak of inflated expectations [8]. With all the hype, however, big data is still in infant stage and requires further research and practices to lead organizations out of the dark.

Manuscript received March 5, 2017
Manuscript revised March 20, 2017

2. The New Wave

An unprecedented magnitude of data, allegedly, the 'data tsunami' has come to change the world [9]. In recent years, by virtue of advancements in internet speed, smartphones, and IOT, there has been mass production of data worldwide. A study by IDC's Digital Universe, states that, the amount of data is doubling every two years and is predicted to reach 40ZB by 2020 [9]. To support this fact, the online microblogging services Twitter, gleans and crunches approximately 12 TB of data per day, whereas Facebook take in more than five hundred million likes per day [10]. On the other hand, Cisco Internet Solutions Group (IBSG) anticipate that there will be 50 billion devices connected to the internet by 2020 [11]. Furthermore, according to CGOC (Compliance, Governance and Oversight Council), data volumes doubles every 18-24 months for majority of organizations and 90% of the data in the world has been created in the last two years [12]. According to the same study, in March 2012, the Obama government declared big data research and advancement initiative, which were utilized to address important problems facing the government. IT giants such as SAG, Oracle, IBM, Microsoft, SAP, HP, and Google, have spent a whopping $15 billion on data management and analytics [13]. A quick view at Google Trends, reveals the fact that big data queries have grown excessively, getting 75% increase from Jan 2012 to June 2016 [14]. Figure 3-1 provides an overview of these query trends.

3. Big Data Ecosystem

Organizations, starting to realize the state of context and the content where the interplay between the user and the provider is in well simpatico using big data, have their archaic ideas crushed on what is true understanding toward the end user and the market. As big data begin to rise, state of practice in analytics evolve [15]. Today, the industry sees an advanced and intricate implementations of big data, such as Mechanical Turk that uses crowdsourcing, and Hortonworks that brings value-added tools to the market. As the new ecosystem begins to shape, there are four main role players within this nexus [16]. These four main roles players are as following;

- Data Generators: this group belongs to data devices that generate new data about data. So for each megabyte of new data created and additional gigabyte is generated.
  - For instance, loyalty card that started to get popular lately were built with the idea of collecting information about spending habits, most visited stores, best hours of the day, and bestselling products. Having these data gleaned and analyzed, the business executive can have actionable knowledge toward the sale patterns and make better future decisions [17].
  - Blizzard Entertainment, one of the leading video game developer and publisher based in Irvine uses massive big data platforms for all of its games, tools, and operations being offered. The company utilizes robust pipelines to collect global information that power analytics, operations, machine learning, and discovery [18].
- Data Collectors: as the name is descriptive enough, this group collect data about users and devices with attention to their attributes and attitudes.
  - Nielsen, being the long dominant player in the collection of data in the television industry, tracks activity on mobile devices, internet, and cable television in order to gain insight on consumer sentiments, reputation of the brand, and consumer reaction to public relations events [19].
  - Furthermore, retails stores, using the RFID chip, track the path a customer takes through their store in order to gain insights on products having most foot traffic [20].
- Data Aggregators: this group belongs that collects data and draw patterns based on it. These organizations gleans data from various sources such as retail stores, sensors, websites, and smartphones, analyze it, generate insights, and then sell it as a product to other organizations that are in need [15].
  - Axciom is a company that concentrates on business and residential listings in U.S. and Canada. Company gleans data about business names, phone numbers, addresses, classification codes, and coordination. In addition, Axciom provides developers with more than 100 million indexed keywords and phrases [21]. These information will be utilized to generate patterns of behaviors.
- Data Consumers: this group benefit from the data gleaned and crunched by others within the data value chain.
  - There are numerous scenarios in which the data is being used to benefit the business. For instance, Donald Trump's presidential campaign in 2016, used data analysis based on a hyper-targeted psychological approach that deemed to be successful [22].
  - Moreover, Google Chrome web browser consumes data based on search indexes, clicks on search results, Email, YouTube, and user's browsing sessions in order to gain actionable knowledge and competitive advantage [23].

The loop among these 4 groups gives life to the big data ecosystem. Data generators using variety of different devices, produce deluge of data that will be gleaned by data collectors. Data collectors then pass these colossal data-to-data aggregators. This is where the complex analytics will be applied and patterns will be unearthed. From there on, insights and patterns will be sold for competitive advantage and business intelligence to demanding organizations. However, in the case of giant companies like Google, the whole process of collection, aggregation and consumption happens within the company, which eliminates dependency on outside organizations. This works efficiently for major role players with huge capitals, nevertheless, small to medium sized companies may require to have some phases of this loop outsourced. Open source technologies have contributed positively, helping organizations figure out their data needs, howbeit, there may be shortage of skills necessary to elicit actionable knowledge and business insights.

4. Big Data Architecture

With the advent of big data and colossal volumes of data standing in need of analysis and value generation, a paradigm shift in computing materialized. The practices of scaling computation resources both in hardware and software layer revolutionized. Today, most big data technologies are based on distributed architecture of computation, which is a horizontal scaling to handle large datasets. Whereas, vertical scaling implies adding more peripherals such as processors and RAMs to increase the performance of the system, horizontal scaling means scaling by adding more machines into the pool of resources. U.S. Navy rear admiral Grace Hooper (1906-1992), who was the inventor of the first compiler for a computer programming language, explained the need of using a distributed architecture eloquently. She stated that in the preindustrial time, oxen were common for heavy pulling, when log was overloaded and the ox could not budge, people did not try to increase the large of ox, instead they added one more oxen [15]. The point is that, there will always be a demand for a greater computation; hence, instead of producing bigger and more powerful computers, a better approach is to build a distributed system comprised of different computers that can efficiently share workloads.

4.1 MapReduce

MapReduce is a computation architecture for parallel processing of workloads on distributed computing system

[24]. Jeffrey Dean and Sanjay Ghemwat first described the algorithm in the paper called "MapReduce Simplified Data Processing on Large Cluseters". The algorithm works its way by enabling split of a single computation task to multiple nodes for distributed processing [24]. By the reason that tasks are usually broken down into multiple nodes, the processing power of the system is determined by the number of nodes. Today, numerous open-source and commercial technologies are implemented based on MapReduce Architecture. Apache Hadoop is a prominent implementation of MapReduce, which is utilized for data processing underlying distributed computing architecture.

4.2 MapReduce Functionality

The 'Map' part of the algorithm takes care of the splitting tasks in computing nodes, while the 'Reduce' is responsible for collecting results in an individual computation and combine them to get the final result [25]. In This algorithm, the mapping functions reads the input data and creates a set of intermediate records for the computation [24]. These produced intermediate records by map function, take the form of a (key, value) pair. As a part of this function, these records are assigned to different computing nodes with usage of a hashing function. Computing nodes then perform the received computing operation and return the result to the reduce function. Reduce function collects computing nodes results of the computation to generate a final output. For instance, if we are about to calculate the number of occurrences of each word in a text file, the 'Map' function takes the file, detach words, and sends them to different nods. The function then splits the file into words and assigns a digit "1" to them to structure a key-value pair for further computation. Once done, the output from the sort operation is fed to 'Reduce' function, which collects output from different nodes, aggregate them, and generate the final output consisting of the frequency of words [24]. Figure 1 portrays MapReduce functionality.

5. Human Factor

The industrial shape-shift and the rise of new big data ecosystem has underpinned the foundation of new roles, platforms and analytical methods [15]. The point of concentration in this section is the emerging roles and their categories.

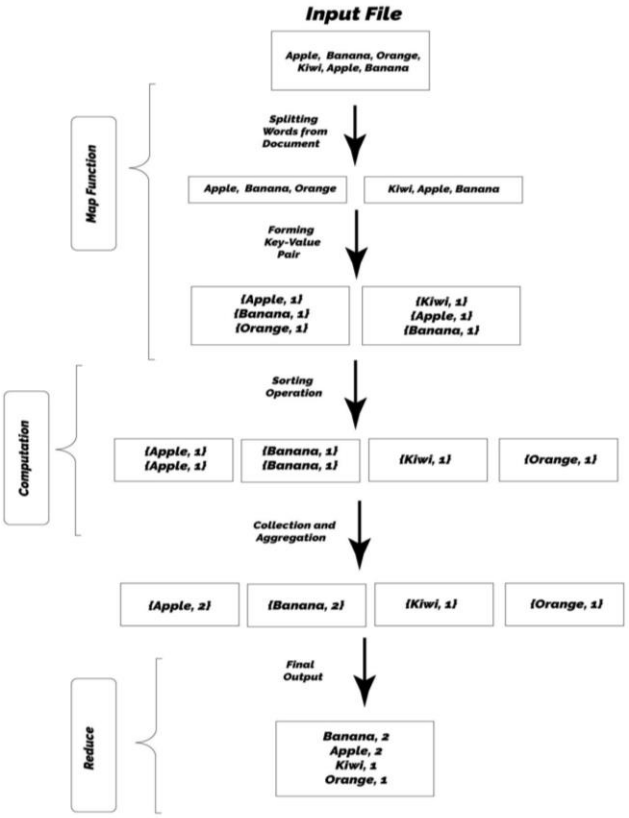


Fig. 1: MapReduce Functionality

These categories are as followings:

- **Deep Analytical Talents:** This category belongs to technical savvies, with high technical skills. Members of this category are those who are proficient in applying complex analytical algorithms to structured and unstructured data in variant scales. Besides, they possess good knowledge of mathematics, statistics, and machine learning. They are usually provided with rigorous workspaces where they have infrastructure and technologies available to perform large-scale analytical data experiments [10]. Some instances of current industrial professions that fit into this category includes economists, statisticians, mathematicians, and indeed data scientists. According to Mckinsey [10], the United States will be confronting a talent gap of 140,000- 190,000 people with deep analytical talent by 2018.
- **Data Savvy Professionals:** This category compare to the previous, possesses lower technical skills but has knowledge of basic statistics and machine learning [10]. Some instance of this category are financial analysts, life scientists, operation managers, business managers, and market researchers. According to the same report by [10], talent gap for this category in U.S will be 1.5 million by the year 2018.

- **Technology and Data Enables:** This group is constituent of people that have technical expertise to support analytical projects, such as managing large-scale data architectures, and administration of analytical sandboxes [10]. Members of this category need to carry skills correspondent to software engineering, database administration, and programming. Figure 2 presents an overview of these categories.

These three categories form a basis of collaboration that addresses complex big data challenges. Most of today's organizations are known of two latter categories mentioned, but the first category tends to be unique. To clear thoughts, data scientist will be elaborated as commonplace role in big data ecosystem. Activities essential to data scientist role with given detail of the skill required are as followings:

- **Bridge business challenges to analytics challenges:** This is a skill to identify business problems, consider the contributing factors, and choosing suitable analytical method for rectification.

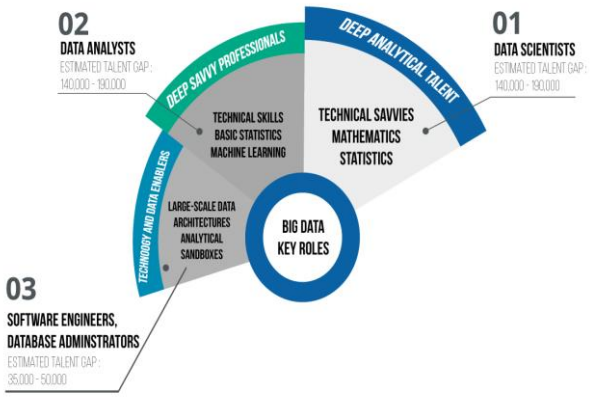


Fig. 2: Key Role Players

- **Design, implement and deploy statistical models:** These are series of operations that constitutes the focal point of data scientist role. As its descriptive enough, this skill entails applying complex analytical methods to variety of business problems [10].
- **Develop insights and actionable knowledge:** It is important to realize the fact that only applying advanced methods to large-scale data sets doesn't drive new business intelligence. Instead, it is important to learn how to unearth new patterns and draw insights out of data and communicate them effectively [15]. That's where the pedal meets the metal.

According to Mckinsey [10], data scientists generally possess five main traits and qualities, which are listed as following:

- **Quantitative Skills:** mathematics, statistics
- **Technical Skills:** software engineering, programming, machine learning, deep learning
- **Critical Thinking:** an objective analysis and evaluation of the issue
- **Skepticism:** multidimensional testing and conceptualization toward problems
- **Creative:** creatively approaching problems and drawing solutions
- **Curious:** self-motivating engrossed individuality
- **Communication Skills:** eloquent expression of business value and collaboration with other stakeholders

Figure 3 illustrates these traits and qualities.

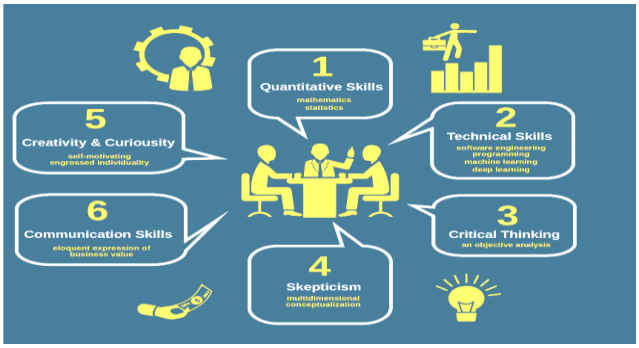


Fig. 3: Data Scientists Profile

6 Conclusion

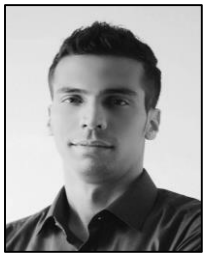
Colossal volumes of data standing in need of analysis, laid the foundation of big data. Big data came as a paradigm shift in the industry that confronted practitioners with diverse challenges. Organizations started to realize that, the give and take between the user and the business in an effective manner, can transform their traditional ideas on what's correct understanding toward the client and the market. This brought along alterations in both technological and conceptual paradigms, such as promotion of horizontal scaling, intelligent distributed computing algorithms like MapReduce, and various big data methodologies. Throughout this study, this ecosystem has been elaborated, with outlining of major game-changing factors such as key role players, required skills, algorithms, and technologies. Howbeit, this is just a substantial fraction of big data ecosystem, and further researches can be exerted to elaborate various technology and analysis related concepts such as big data on the cloud, predictive analysis, big data models, big data challenges, privacy, and security concerns.

References

[1] Chen, H.-M., et al. Big Data as a Service: A Neo-Metropolis Model Approach for Innovation. in 2016 49th Hawaii International Conference on System Sciences (HICSS). 2016. IEEE.

[2] Huberty, M., Awaiting the second big data revolution: from digital noise to value creation. Journal of Industry, Competition and Trade, 2015. 15(1): p. 35-47.
[3] Xinhua, E., et al. Big Data-as-a-Service: Definition and architecture. in Communication Technology (ICCT), 2013 15th IEEE International Conference on. 2013. IEEE.
[4] Strawn, G.O., Scientific Research: How Many Paradigms? Educause Review, 2012. 47(3): p. 26.
[5] Chen, H., R.H. Chiang, and V.C. Storey, Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly, 2012. 36(4): p. 1165-1188.
[6] Abbasi, A., S. Sarker, and R. Chiang, Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 2016. 17(2): p. 3.
[7] Fox, S., Getting real about innovations: formulating innovation descriptions that can reduce ontological uncertainty. International Journal of Managing Projects in Business, 2012. 5(1): p. 86-104.
[8] Fox, S. and T. Do, Getting real about Big Data: applying critical realism to analyse Big Data hype. International Journal of Managing Projects in Business, 2013. 6(4): p. 739-760.
[9] Barrachina, A.D. and A. O'Driscoll, A big data methodology for categorising technical support requests using Hadoop and Mahout. Journal of Big Data, 2014. 1(1): p. 1.
[10] Mckinsey, B.D., The Next Frontier for Innovation. Competition, and Productivity, 2011.
[11] Qin, E., et al. Cloud computing and the internet of things: Technology innovation in automobile service. in International Conference on Human Interface and the Management of Information. 2013. Springer.
[12] Austin, D., eDiscovery Trends: CGOCs Information Lifecycle Governance Leader Reference Guide. 2012.
[13] Zheng, Z., J. Zhu, and M.R. Lyu. Service-generated big data and big data-as-a-service: an overview. in 2013 IEEE international congress on Big Data. 2013. IEEE.
[14] Bughin, J., Big data, Big bang? Journal of Big Data, 2016. 3(1): p. 1.
[15] Dietrich, D., B. Heller, and B. Yang, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. 2015: John Wiley & Sons.
[16] Davenport, T.H. and J. Dyché, Big data in big companies. International Institute for Analytics, 2013.
[17] Mayer-Schönberger, V. and K. Cukier, Big data: A revolution that will transform how we live, work, and think. 2013: Houghton Mifflin Harcourt.
[18] El-Nasr, M.S., A. Drachen, and A. Canossa, Game analytics: Maximizing the value of player data. 2013: Springer Science & Business Media.
[19] E. Prescott, M., Big data and competitive advantage at Nielsen. Management Decision, 2014. 52(3): p. 573-601.
[20] Zaslavsky, A., C. Perera, and D. Georgakopoulos, Sensing as a service and big data. arXiv preprint arXiv:1301.0159, 2013.
[21] Bakalash, R., G. Shaked, and J. Caspi, Method of and apparatus for data aggregation utilizing a multidimensdional database and multi-stage data aggregation operations. 2001, Google Patents.
[22] Mezzofiore, G. How a little-known data firm helped Trump become president. 2016; Available from: http://mashable.com/2016/11/10/donald-trump-polling-data/#pwnOPux3FSqp.
[23] O'Driscoll, A., J. Daugelaite, and R.D. Sleator, 'Big data', Hadoop and cloud computing in genomics. Journal of biomedical informatics, 2013. 46(5): p. 774-781.
[24] Bhagattjee, B., Emergence and taxonomy of big data as a service. 2014, Massachusetts Institute of Technology.
[25] Dean, J. and S. Ghemawat, MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008. 51(1): p. 107-113.

Dr Babak Bashari Rad received his B.Sc. of Computer Engineering (Software) in 1996 and M.Sc. of Computer Engineering (Artificial Intelligence and Robotics) in 2001 from University of Shiraz; and Ph.D. of Computer Science (Information Security) in 2013 from University Technology of Malaysia. Currently, he is the programme leader of postgraduate studies and senior lecturer in the School of Computing, Asia Pacific University of Technology and Innovation (APU), Kuala Lumpur Malaysia. His main research interest covers a broad range of various areas in computer science and information technology including Information Security, Malware Detection, Machine Learning, Artificial Intelligence, Image Processing, Robotics, Cloud Computing, Big Data, and other related fields.



Pouya Ataei received the B.S. dual degrees in Software Engineering from Asia Pacific University and Staffordshire University in 2015. During 2015-16, he was an active researcher in the industry. His current research interests include Big Data, IOT, Cloud Computing, Software Engineering, and Security. He is currently pursuing a M.S. in Software Engineering from Staffordshire University.