Review of Information Engineering and Applications
2021 Vol. 8, No. 1, pp. 1-9.
ISSN(e): 2409-6539
ISSN(p): 2412-3676
DOI: 10.18488/journal.79.2021.81.1.9

STUDY AND COMPARATIVE ANALYSIS OF PROGRAMMING LANGUAGES USED FOR BIG DATA

Shafagat Mahmudova

SE Department, Institute of Information Technology of ANAS, Baku, Azerbaijan.
Email: shafagat_57@mail.ru Tel: (99412)5688847

Article History
Received: 18 September 2020
Revised: 2 December 2020
Accepted: 21 December 2020
Published: 5 January 2021

Keywords
Big data
IoT
3D printing
Blockchain
Quantum calculations
Technology
Programming languages
MapReduce algorithm
Machine learning.

ABSTRACT

This article provides information about the industrial revolutions and analyzes the Fourth Industrial Revolution. The Fourth Industrial Revolution takes away the potential risks of unsustainable growth and world systems, and its coming is estimated as a challenge. The main technologies of the Fourth Industrial Revolution are identified. Big data technologies enable to handle different types of data at the same time. Big data is very important among technologies. The difference between big data and usual data is that the volume of Big Data starts at terabytes. The types technologies used for big data are explained. The programming languages used for big data are extensively analyzed. Their similarities and differences are compared. Big data will be among the most required information technologies. The article analyzes the literature in this area and data analysis methods for Big data. There are other criteria that emphasize the need for new methods to work with big data.

Contribution/Originality: The study also contributes the exploration of the types of technologies used for big data.

1. INTRODUCTION

The industrial revolution means the mass transition from manual labor to machinery, from manufacturing to factory, etc., which took place in the leading countries of the world in the 18-19th centuries.

The main result of the industrial revolution is the industrialization - the transition from a dominant agrarian economy to industrial production, which resulted in the transformation of agrarian society into an industrial one. The industrial revolution began in England in the late 18th century and became widespread in Europe and America in the first half of the 19th century.

There have been 4 industrial revolutions throughout the history:

1. Discovery of the steam engine.
2. Discovery of electricity.
3. Emergence of computer and automation of production.
4. Emergence of the Internet, etc.

Table 1 provides detailed information on the industrial revolutions 4th (2017).

Table-1. Industrial revolutions, period, innovations and results.

Industrial Revolutions	Period	Innovation	Result
First Industrial Revolution	Late 18th century, early 19th century	Water and steam machines, loom, mechanical devices, transport, metallurgy	Transition from agrarian economy to industrial production, development of transport
Second Industrial Revolution	Second half of the 19th century, early 20th centry	Electricity, oil and chemical industry, telephone, telegraph	Electrification, railways, labor division, group production
Third Industrial Revolution	Late 20th century (1970, etc.)	Digitization, development of electronics, application of information technologies and software in production	Automation and robotics
Fourth Industrial Revolution	Early 21st century (from 2011)	Global networks, Internet of Things, 3D printer, neural networks, biotechnology, artificial intelligence, etc.	Distributed production, energy, use of collaborate network, etc.

The Fourth Industrial Revolution was the mass introduction of cyber-physical systems and services for human needs, including production, living, labor, and leisure [1]. The 4th Industrial Revolution was initiated in 2011 by the businessmen, politicians and scientists, as they identified themselves as a means of increasing the competitiveness of refinery industry of Germany through enhanced integration of "cyber-physical systems" [2].

The Fourth Industrial Revolution (Industry 4.0) is a transition to fully automated digital production, constantly interacting with the external environment in real time by controlled intelligent systems [3].

The Fourth Industrial Revolution, which led to an increase in the economic viability and quality of life, takes away the potential risks of unsustainable growth and world systems, and its arrival is seen as a challenge [4].

One of the most important elements of the fourth industrial revolution is the wireless transmission of data over the network, namely the Internet. The fourth industrial revolution is often described through basic technologies. These technologies include the following collective concepts of future technologies [4] as follows Figure 1.

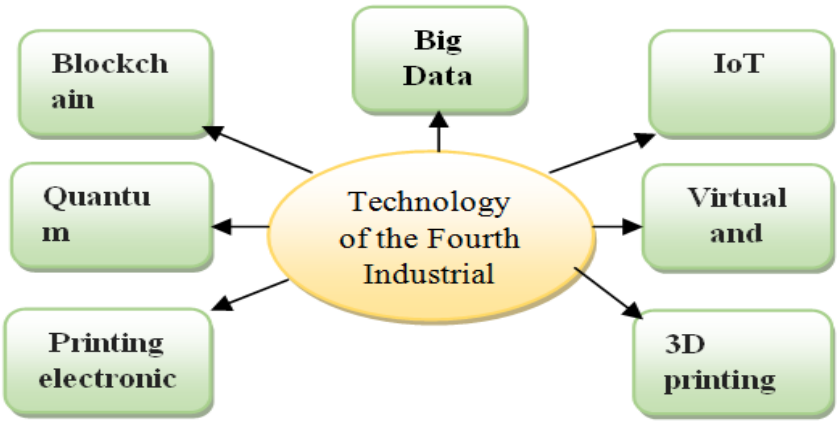

Figure-1. Technology of the fourth industrial revolution.

Summary materials prepared at the 2019 World Economic Forum (Davos) can be used in the field of production

[5].

- Individual robots for production in the field of logistics.
- Drones and sensors for agriculture.
- Digital similarities in business [6].

Industry 4.0 is used as a concept within the framework of the German state's individual automated program to create the individualized products that interact with the external environment. Industry 4.0 is gradually covering the entire world - the United States formed a non-profit consortium of Industrial Internet in 2014 [7] led by industry leaders of General Electric, AT&T, IBM and Intel [8].

2. ABOUT BIG DATA

As mentioned, one of the technologies of the fourth industrial revolution is Big Data.

Big Data analytics is the focus of data science. Many state and private organizations began to collect large amounts of different data which could contain useful information about problems. For example, it became necessary to use Big Data technology in national intelligence, cyber security, marketing and medical informatics [2].

Big Data includes many trends. However in a broad sense, it can be divided into two categories:

1. Big Data Engineering.
2. Big Data Analytics.

These categories are related, but differ from each other.

Big Data engineering deals with carcass processing, data collection and storage, as well as acquiring relevant data for various consumer and internal applications.

Big Data Analytics is an environment developed by Big Data engineering for the use of large amounts of data from external systems. It includes areas of big data analysis, sample analysis, and the development of various classifications and forecasting systems.

Big data analysis includes the analysis of trends, regularities and the development of various systems for classification and forecasting.

Big Data is used to process large and complex data sets. Conventional database control systems are not capable to manage large amounts of unstructured data. Big Data, as a rule, is often defined as a collection of large and complex data sets, and is used when there is a difficulty in managing databases or using traditional software for data processing.

Big Data includes high-speed, high-volume, and there are three types of data:

- Structured data: relational data.
- Semi-structured data: XML data.
- Unstructured data: Word, PDF, etc.

Big Data is truly crucial to our lives, and it is one of the most important technologies in the modern world [9].

In Japan, using Big Data technology, some stores have begun to introduce software to identify individuals on a blacklist of shoppers identified as thieves or "complainants."

In Tampa, Florida, police use Superbowl XXXV, large software that uses Big Data technology to identify criminals based on scanned facial images.

Big Data is used to identify people in the field of biometric technology at universities and airports in many US states.

IBS analysts rated the "global data volume" with the following values Table 2.

Big Data is a combination of technologies designed to perform three operations:

1. Processing large amounts of data compared to "standard" scenarios.
2. Processing large amounts of data, that is, not simply a lot of data, but constantly growing data.
3. Working with parallel and differently structured and complex structured data.

It is believed that these "skills" allow identifying hidden patterns that evade the limited human perception. It provides unprecedented opportunities to optimize many areas of our lives: government, medicine, telecommunications, finance, transportation, manufacturing, etc.

Along with the physical volume of big data, its defining features emphasize the complexity of the task of processing and analyzing this data. In 2001, Meta Group developed a set of VVV attributes (volume, velocity, variety - physical volume, data growth rate and the need for rapid processing, the ability to process different types of data simultaneously) to demonstrate the importance of data management in all three aspects.

The basic principles for working with big data are:

1. Horizontal scalability. This is the basic principle of big data processing. As already mentioned, more and more big information is generated each day. Accordingly, it is necessary to increase the number of computing nodes where this data is distributed and processing should take place without loss of performance.

2. Accident tolerance. This principle follows from the previous rule. A set (sometimes tens of thousands) may include a large number of computing nodes, and their number is likely to increase, and the probability of machine failure will also increase. The methods for working with big data should take into account the possibility of such situations and provide preventive measures.

3. Location of data. As data is distributed over a large number of computing nodes, the cost of data transfer can be unreasonably high if it is physically located on one server and processed on another. Therefore, it is desirable to process the data in the same machine where they are stored.

These principles differ from typical traditional, centralized, and vertical models of well-structured data storage. Accordingly, new approaches and technologies are being developed to work with big data.

Initially, the set of approaches and technologies covered the massive parallel processing of indefinitely structured data, such as NoSQL DBMS, MapReduce algorithms, and Hadoop design tools. In the future, other solutions began to be associated with large information technologies, which provided similar features in the performance of large data series, as well as some devices.

- **MapReduce** is a parallel computing model distributed in computer groups provided by Google. According to this model, the application is divided into a large number of identical elementary tasks performed at the nodes, and then naturally reduced to the final result.
- **NoSQL** - a general term for various unrelated databases and repositories does not include any specific technology or product. Conventional relational databases are well suited to fairly fast and single queries, and with complex and agile queries that are specific to big data, the load exceeds acceptable limits and is inefficient to use.
- **Hadoop** - a framework for the development and implementation of freely distributed utility sets, libraries and advanced programs consisting of hundreds and thousands of nodes. It is considered one of the fundamental technologies of big data.

3. COMPARISON OF PROGRAMMING LANGUAGES USED FOR BIG DATA

Depending on the amount of data being measured R, Python, Scala, and Java programming languages should be used for Big Data [10].

Assume that there is a project with great information. The subject area is understandable, the infrastructure to be used is known, and it is decided to use a certain processing environment, and it is focused on one problem: which programming language to choose?

Below is a brief description of each of these languages to make a decision.

- **R language.** R is a programming language used to process statistics and graphs within the GNU project. This language can be used when an esoteric statistical model is needed for calculations. It is widely used for data analysis and has in fact become the standard for statistical programs.
- **Python.** Python is a high-level programming language aimed at improving the productivity of a developer and a program code. Python's standard library includes many useful functions. Data processing and analysis experts use the Python language. This language has been popular in science for more than a

decade, especially in the areas such as Natural Language Processing (NLP). Python generally supports big data processing. Unlike R, Python is a traditional object-oriented programming language, so it is easy to work with.

- **SQL** Structured Query Language (SQL) is a declarative programming language used to organize, modify, and manage data through a relational database managed by an appropriate database management system. It is one of the main programming languages used for Big Data [11].

C++. C++ is a compiled, static-type general-purpose programming language.

The procedure supports the programming paradigms such as programming, object-oriented programming, and general programming [12]. It has a rich standard library that combines language, algorithms, input/output operators, functions, expressions and other features. C++ combines the features of both high-level and low-level languages. Compared to its predecessor C, it focuses on supporting object-oriented and generalized programming [13].

- **Scala.** Scala is a multi-paradigm programming language used for simple and fast installation of component software, combining the capabilities of functional and object-oriented programming. Scala is one of the four most widely used programming languages; Scala is applied in the financial sector and in companies that process large amounts of data (as Twitter and LinkedIn).
- **Java.** Java is a powerful object-oriented programming language developed by Sun Microsystems (later Oracle). Java is developed by a community organized through the Community Process, and the key technologies that implement it are distributed under the GPL license. The trademark rights belong to Oracle Corporation.

Different programming languages are used for big data projects. Python is the best choice for analysis of large amounts of data with cryptocurrency statistics, and if it is refused to use the R language and a Graphics Processing Unit (GPU) only uses NLP in neural networks. Java or Scala are the most suitable languages to achieve highly effective solutions using all the important processing tools.

Some of the programming languages used for Big Data are shown in Figure 2.

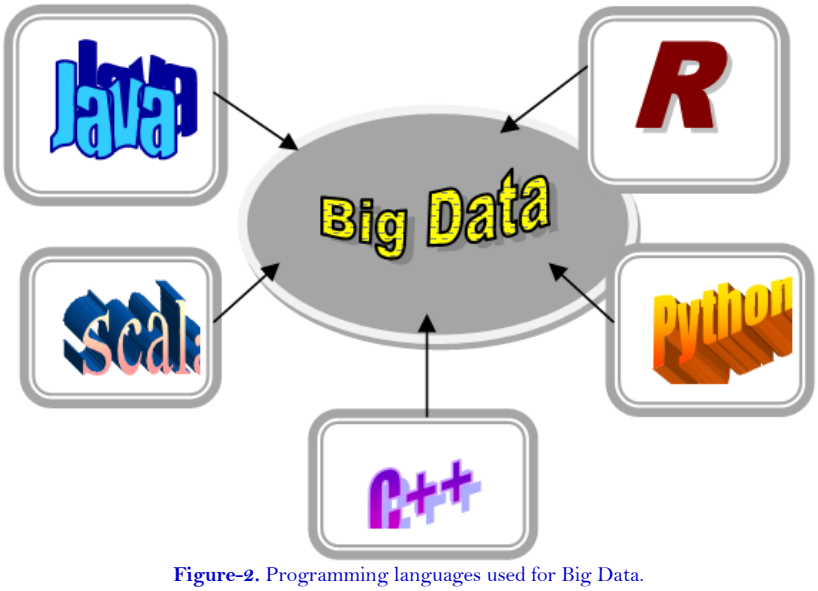

Figure-2. Programming languages used for Big Data.

The TIOBE index survey agency ranks the most commonly used programming languages for Big Data as of March 2020. It is shown in Table 2 and Figure 3.

Table-2. Rankings.

Mar 2020	Programming Language	Ratings
1.	Python	10.11%
2.	C++	6.79%
3.	JavaScript	2.05%
4.	SQL	1.83%
5.	R	

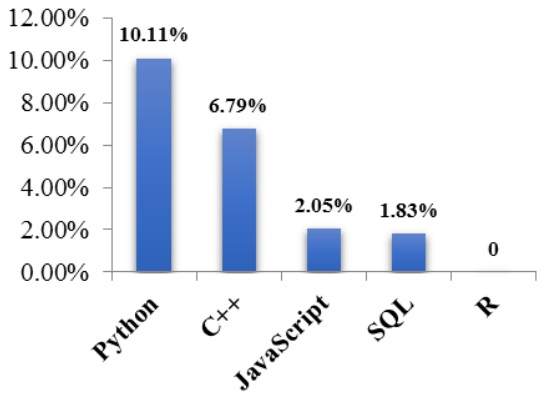

Figure-3. Ranking chart by Programming languages for Big Data.

4. LITERATURE REVIEW

1. Technology Industry 4.0 and Industrial Internet of Things (IIoT) are rapidly evolving for data processing and software that stimulate digitalization in many areas, especially in industrial automation and production systems. A number of advantages offered by these technologies include the use of software to work with Big Data, machine learning (ML) and cloud computing, and the infrastructure used to develop modern platforms for data analysis. Although this area is of great interest, information on the application of data analytics in the context of Industry 4.0 is hardly available in the scientific literature. Therefore, this work provides a platform for data analysis in a process based on the industry 4.0 concept. The platform uses the most advanced IIoT platforms, ML algorithms and software to work with Big Data. The platform emphasizes the use of ML methods to analyze process data when using large data processing tools and taking advantage of existing cloud computing platforms. The simulation results showed that when knowledge of process phenomena is limited, data-based soft sensors are useful tools for analyzing predictive data [14].

2. In the context of a large database, NoSQL is a highly adapted storage model for a column-oriented database system, data storage and transaction analysis. Indeed, the use of NoSQL models simplifies data scaling and is suitable for storing and managing mass data for column storage, especially decision-making queries [15].

3. [16] suggests a collection and input data analysis methodology that describes the use of their open resources by other software applications, following the standard process of CRoss-Industry for data mining. In addition, it allows the study of the use of data as various open sources.

4. In today's digital world, it is difficult to protect confidential information, especially if the user does not have enough information about the importance of data diagnostics before throwing, selling or donating memory devices. Users often use password to protect documents and folders, hard drives, flash drives, and encrypt the SD card in smartphones to prevent unauthorized access to sensitive data. However, almost all of these devices are discarded, sold or donated at the end of their life cycle without proper data cleaning. Deleted data can be easily recovered, so these documents/folders and hard disk space (unused space) should be deleted to prevent unauthorized access to confidential information, as free space can contain pre-deleted confidential information that can be easily recovered. This paper discusses international standards and methods for data deletion programs available in different countries. The results of the experiments show the ability to expand the capabilities of data processing and storage systems for large documents (30% faster) [17].

5. Big Data is a product of the rapid expansion of the Internet data and has stimulated the potential for big data analysis. Many companies need to take advantage of new Big Data technologies to increase competition. [18] mainly discusses the application of Big Data technologies to ensure big data security. Oracle is used jointly with Big Data technology to study and analyze key technologies to develop an application platform for big data security.

5. BIG DATA ANALYSIS METHODS

The term Big Data Analytics refers to large and expanded groups. Existing research organizations have the potential to improve big data for the amount of data collected from these groups [19].

McKinsey, an international consulting firm specializing in strategic management issues, identifies 11 methods and analytical methods applied to big data:

- **Methods of data mining class -** a set of methods for the detection of practically useful knowledge that was not previously known for decision making.
- **Crowdsourcing -** the classification and enrichment of information by the forces of a wide, uncertain circle that carries out this work without entering into an employment relationship.
- **Data merging and integration -** a set of methods that allow for the integration of heterogeneous data from different sources for in-depth analysis (e.g. digital signal processing, natural language processing, including tone analysis, etc.)
- **Machine learning**, including teacher and non-teacher learning - to use models based on statistical analysis or to achieve complex predictions based on basic models.
- **Artificial neural networks**, network analysis, optimization, including genetic algorithms (genetic algorithm - used to solve problems of optimization and modeling by random selection, integration and modification using mechanisms similar to natural selection in nature).
- **Pattern recognition** is a scientific discipline that aims to identify the objects according to several criteria or classes.
- **Predictive analytics -** a class of methods for the analysis of data collected to predict the future behavior of objects and subjects in order to make optimal decisions.
- **Imitation (simulation) -** a method that allows creating the models that describe the processes as they are in reality.
- **Spatial analysis -** a class of methods that uses topological, geometric, and geographic information extracted from data.
- **Statistical analysis -** time series analysis, A / B test (split test - marketing research method); when used, the control group of elements is compared with another group in which one or more indicators are changed.
- **Visualization of analytical data -** a presentation of data in the form of images and diagrams, using interactive features and animations to obtain results and for further analysis.

6. CONCLUSION

Big Data will be among the most required information technologies for a long time. According to the estimations, by 2025, enterprises will generate about 60% of data throughout the world. Almost Information flows are continuously generated by companies in the field of finance, telecommunications, e-commerce, etc. Such business requires technological solutions that will help to efficiently collect, store and handle large amounts of data. The difference between Big Data and regular data is that Big Data is generated at terabytes. Relational systems are not capable to store and process large amounts of data. There are other criteria that emphasize the need for new methods to work with big data.

- **Variety**. Big Data often consists of unstructured data from multiple sources in the form of data in various formats (video and audio files, text, images, etc.). Big Data technologies allow to process different types of data at the same time.
- **Velocity**. Data is generated faster. For example, an online store needs to constantly collect information about the number of customers and purchases. Non-relational databases are more suitable for storing such data, as they can be easily scaled horizontally by adding new servers.
- **Processing rate**. Big Data, despite its significant size, should be handles very fast in real time. For example, the systems recommended in online stores immediately analyze the customer's behavior and provides decisions about which products he/she might like. This high processing rate is achieved through distributed computing [20].

Funding: This study received no specific financial support.
Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.

REFERENCES

[1] K. Il'ya, "Industriya 4.0: Chto takoye chetvertaya promyshlennaya revolyutsiya? Retrieved from: https://hi-news.ru/business-analitics/industriya-4-0-chto-takoe-chetvertaya-promyshlennaya-revolyuciya.html (accessed 15.may 2015)," 2015.

[2] L. Stepan, "The end of the analog world: Industry 4.0, or what the fourth industrial revolution will bring. Retrieved from: https://theoryandpractice.ru/ [Accessed 21.09.2016]," 2016.

[3] Chetvertaya Promyshlennaya Revolyutsiya, *Tselevyye oriyentiry razvitiya promyshlennykh tekhnologiy i innovatsiy. Vsemirnyy ekonomic news forum*: McKinsey & Company, 2019.

[4] M. Sergey, "4-ya promyshlennaya revolyutsiya v Davose. Retrieved from: https://expert.ru/2016/01/21/chetvertaya-promyishlennaya-revolyutsiya/ (accessed 21 yanvar '2016)," 2016.

[5] B. Kateryna, "Challenges and opportunities of industry 4.0 – Spanish Experience," *International Journal of Innovation, Management and Technology*, vol. 9, pp. 1-7, 2018.

[6] V. Pretlík, *Industrial logistics in: Nakladatelství ČVUT v praze*. Prague: Chtu, 2006.

[7] B. Kateryna, "What is in reality industry 4.0?," *Innova Cima*, 2017.

[8] AT&T, "AT&T, "AT&T+Cisco+GE+IBM+Intel = Industrial Internet Consortium |AT&T ,27.03.2014. Retrieved from: https://about.att.com/story/att_cisco_ge_ibm_intel_industrial_internet_consortium.html," 2014.

[9] S. Mahmudova, "Big data challenges in biometric technology," *Education and Management Engineering*, vol. 5, pp. 15-23, 2016.Available at: 10.5815/ijeme.2016.05.02.

[10] P. Ian, "Which freaking big data programming language should I use?," *InfoWorld*, 2016.

[11] R. G. James, N. W. Paul, and J. O. Andrew, *SQL: The complete reference*, 3rd ed. Moscow: Vilyams, 2014.

[12] S. Bjarne, *Programming: Principles and Practice of Use C++, ispravlennoye izdaniye = Programming: Principles and Practice Using*, 248 ed. Moscow: Vilyams, 2011.

[13] S. Bjarne, "The design and evolution of C++'," ed USA: Addison–Wesley, 2007, p. 445.

[14] J. Kabugo, J. S. Jamsa, R. Schiemann, and C. Binder, "Industry 4.0 based process data analytics platform: A waste-to-energy plant case study," *International Journal of Electrical Power & Energy System*, vol. 115, pp. 1-18, 2019.Available at: 10.1016/j.ijepes.2019.105508.

[15] K. Dehdouh, B. O., and B. F., "Big data warehouse: Building columnar NoSQL OLAP cubes," *International Journal of Decision Support System Technology*, vol. 12, pp. 1-24, 2020.Available at: 10.4018/IJDSST.2020010101.

[16] M. Grzenda and J. Legierski, "Towards increased understanding of open data use for software development," *Information Systems Frontiers*, vol. 19, pp. 197–212, 2019.Available at: 10.1007/s10796-019-09954-6.

[17] S. Gnatyuk, V. Kinzeryavyy, T. Sapozhnik, I. Sopilko, N. Seilova, and A. Hrytsak, "Modern method and software tool for guaranteed data deletion in advanced big data systems," in *Proc Appeared in Proceedings of the 2nd International Conference of Artificial Intelligence, Medical Engineering, Education (AIMEE), Moscow, Russia*, 2020, pp. 581-590.

[18] G. Chai, "The research on core development technology of security big data application platform," *Advances in Intelligent Systems and Computing*, vol. 928, pp. 479-486, 2020.Available at: 10.1007/978-3-030-15235-2_70

[19] A. Muruganantham, E. Phong Thanh Nguyen, L. Laxmi, K. Shankar, H. Wahidah, and M. Andino, "Big data analytics and intelligence: A perspective for health care," *International Journal of Engineering and Advanced Technology*, vol. 8, pp. 861-864, 2019.

[20] S. Vladimir, "How can a programmer switch to Big Data, ScenseSoft, 05/14/2019. Retrieved from: Https: //www.sknsoft.by/company," 2019.