Spectrum Journal of Innovation, Reforms and Development
Volume 29, July - 2024 | ISSN (E): 2751-1731
Website: www.sjird.journalspark.org

BIG DATA, DATA SCIENCE AND MACHINE LEARNING AS THE LATEST MODERN TRENDS

Каримова Лайло Рахимбергановна,
студентка бакалавриата 2 курс, Ташкентский Государственный Технический Университет, Узбекистан, г. Ташкент

Abstract
This scientific article examines such important areas of IT – Information technologies as Big Data, Data Science and Machine Learning. Also, everything is supported by examples in various IT companies and not only. The results are summarized and the possibilities of further application in the near future are described.

Keywords: Big Data, Data Science, Data Scientist, Machine Learning, Artificial Intelligence, IT, Neural Networks, Big Data, Data Science, Machine Learning, Artificial Intelligence.

Introduction

Today, when humanity has moved into the era of the fourth industrial revolution, it is impossible to imagine a business without qualified IT specialists, especially when it is necessary to work with huge amounts of information. Our modern society, in the era of information and global leap, is rapidly developing, and the volume and flow of data are constantly growing, which leads us to new discoveries. As a result, new meanings, new terms appear, which in the scientific and practical world have been called Big Data, Data Science and Machine Learning, or translating these terms into Russian - "big data", "data science" and "machine learning".

Big Data

Big Data is a combination of many different technologies and methods for collecting, processing, and analyzing unstructured and structured data in large volumes. A few years ago, big data was an innovation trend that was only used in the high-tech sector. At the moment, big data is present in all spheres and sectors of human life and occupies a huge place in our daily life, and among other things, it can be found, used and applied for both commercial and non-commercial purposes and environments.

Big Data technologies allow you to process large amounts of data, systematize them, analyze and identify patterns where the human brain would never notice them. This opens up completely new possibilities for using data. Big Data doesn't just mean large packets of data, it's huge stored and processed arrays of hundreds of gigabytes, or even petabytes of data. In short, we can define Big Data as technologies for processing the total amount of information to obtain certain information.

With the development of BigData, the technologies of global companies have also evolved. At the moment, BigData is the lot of not only the giants of the IT world. This direction, thanks to a set of cloud services from IBM, Amazon, Google, becomes available to almost any company working in

the IT field. And such solutions as Clickhouse, Cassandra, InfluxDB allow even individual developers who want to create their own business projects to enter the field of working with Big Data.

Competent use of BigData today is a prerequisite for the development of large IT companies. Without analyzing the behavior of its users, without the ability to predict, guided only by experience and intuition, it is currently extremely difficult to remain competitive with such large companies as Amazon and Google. A configured and working BigData system allows you to provide valuable information in seconds, obtained and compiled from the analysis of billions of actions of the company's customers.

In business today, the concept of Data Driven Management has already emerged, which means managing a company based solely on information obtained from data analysis. And such methods of managing companies show brilliant results. Facebook, Google, Mail.ru, and Yandex have long used analytics to make decisions. Also, today traditional businesses are also interested in BigData, whose representatives need new tools to improve efficiency.

Basic principles of working with Big Data.

1. Horizontal scalability.

Since there can be a large number of data when working with it, the system in which the data is stored must be able to expand. If the volume of data has doubled, then the number of clusters should increase by 2 times, and by analogy, when increasing not twice, but by another certain figure.

2. Fault tolerance

Horizontal scalability means that there are a huge number of machines working with data in the cluster. And, accordingly, it cannot be ruled out that these machines will fail for one reason or another. Yahoo's Hadoop cluster, for example, has more than 42,000 machines. Methods of working with BigData should take this possibility into account and continue to work without visible losses if a certain number of machines fail.

3. Locality of data

In large systems, data is distributed across a large number of machines. If the data is on one machine and processed on another, the cost of transferring that data can even exceed the cost of processing. Therefore, an important issue in the design of Big Data is the principle of data locality, or in other words, the processing of information in the same place where it is originally stored.

The global use of Big Data has caused the emergence of new trends, one of which can be called Data Science. Most of the largest companies today use Data Science to provide their customers with personalized offers. A prime example of this is Google, AdSense, which collects information about users and shows contextual ads.

Data Science

Data Science is a branch of computer science that studies the problems of analyzing, processing, and presenting data in digital form. It combines methods for processing data in conditions of large volumes and a high level of parallelism, static methods, data mining methods and artificial Intelligence applications for working with data, as well as methods for designing and developing databases.

The term Data Science was first introduced and characterized in a book by the Danish scientist Peter Naur in 1974, although it is believed that Naur used the term as early as the 1960s. However, the term Data Science gained its popularity only in the first decade of the 21st century, largely due to the popularization of the concept of Big Data.

As a consequence of the emergence of this science, Data Science generates a demand for new professions in this field, one of which is Data Scientist.

A Data Scientist is a data scientist or engineer with high skills in mathematics, programming and analytics. This field and profession can be described as the one that is at the top and will occupy leading positions for a long time, since Data Science specialists with high mathematical and analytical qualities are in great demand in the labor market now and, according to analysts' forecasts, there will be a very high demand for them for a long time.

If we consider this field and profession in more detail, it can be noted that a Data Scientist is a specialist who works very closely in the mathematical field, delving into more complex categories and subcategories of mathematics, such as mathematical statistics, probability theory and linear algebra, and also knows how to apply mathematical knowledge in practical terms, using various software tools.

All of the above is the main difference between a Data Scientist and an ordinary mathematician. This profession requires deep theoretical and real practical knowledge of methods of statistical data analysis, skills in building mathematical models (for example, neural networks), working with large data sets and a unique ability to find patterns.

Summarizing all of the above, it should be noted that a Data Scientist is a specialist who understands many areas and areas in the field of IT (information technology), such as analytics, business intelligence, artificial intelligence, machine learning, deep learning, and much more.

In the process of studying the features of Big Data concepts and the prospects for the development of Data Science, it is impossible not to touch upon such an important area in IT, which has already been mentioned above, as machine learning.

Machine Learning.

There is no exact universally accepted definition of Machine Learning, so the interpretations of machine learning from various major representatives of the IT industry and research companies will be presented below.

- "Practical Use of Algorithms to Analyze Data, Study It, and Then Predict Something" (NVIDIA).
- "The Science of How to Teach Computers to Function Without Explicit Programming" (Stanford University).
- "Technology based on algorithms that can learn from embedded data without the aid of programming tools" (McKinsey & Co).
- "Algorithms that are able to independently choose a method for solving important problems by generalizing the examples embedded in the system" (University of Washington).
- "A field whose function is to find ways to create computer systems that can self-learn and improve themselves as experience is accumulated, as well as to search for the fundamental laws by which all learning processes work" (Carnegie Mellon University).
- Having familiarized ourselves with all the above definitions, below we will put forward our generalized interpretation of Machine Learning.

- Machine Learning is a subfield of Artificial Intelligence and Data Science that specializes in the use of data and algorithms to mimic human learning, or, in other words, build trainable models for various purposes: for example, process automation, automatic text translation, image recognition. It is such a direction as machine learning that helps to rank content in various social networks and create voice or text assistants that communicate in natural language, creating the illusion of a real interlocutor, for example, Siri from Apple or Alice from Yandex.

Types of machine learning.
Machine learning can be divided into two types:
1. Deductive learning (expert systems).
In this case, there is formulated and formalized knowledge. For example, it can be a database that indicates that if the temperature exceeds 30 degrees, then you need to turn on the air conditioner, and if it is raining outside, then you need to close the windows. It is necessary to derive from them a new rule that can be applied to a specific specific case. Expert systems are more often referred to as a branch of cybernetics, the science of managing information in complex systems, than to machine learning.

2.Inductive Learning, which in turn is divided into:
• Study with a teacher.
An example of possible tasks: based on the previous exchange rate, you need to predict the exchange rate for tomorrow; distinguish cats from dogs by images (in this case, there should initially be information on which picture and where the cat and dog are depicted).

• Unsupervised learning.
An example of a possible task is to divide a group of site users based on their interests or demographics. Typically, you want to know how many groups are already in your data.

• Reinforcement learning.
An example of a possible challenge is a series of Super Mario games in which a computer (agent) interacts with the environment (game level) and receives either positive or negative points.

• Active learning.
An example of a possible task: hinting words on the keyboard layout of a smartphone.
Many methods of inductive learning are not so much about learning as they are about extracting information.
If we take a closer look at the capabilities of supervised machine learning algorithms, then it is worth considering, for example, a few tasks or problems that can be solved by machine learning algorithms:
• Determination of the postal code by handwritten numbers on the envelope;
• Finding a benign tumor based on medical images;
• Detection of fraudulent activity in credit card transactions;
• Predicting failures of high-tech and complex industrial equipment;
• Identification and recognition of images captured by unmanned aerial vehicles.

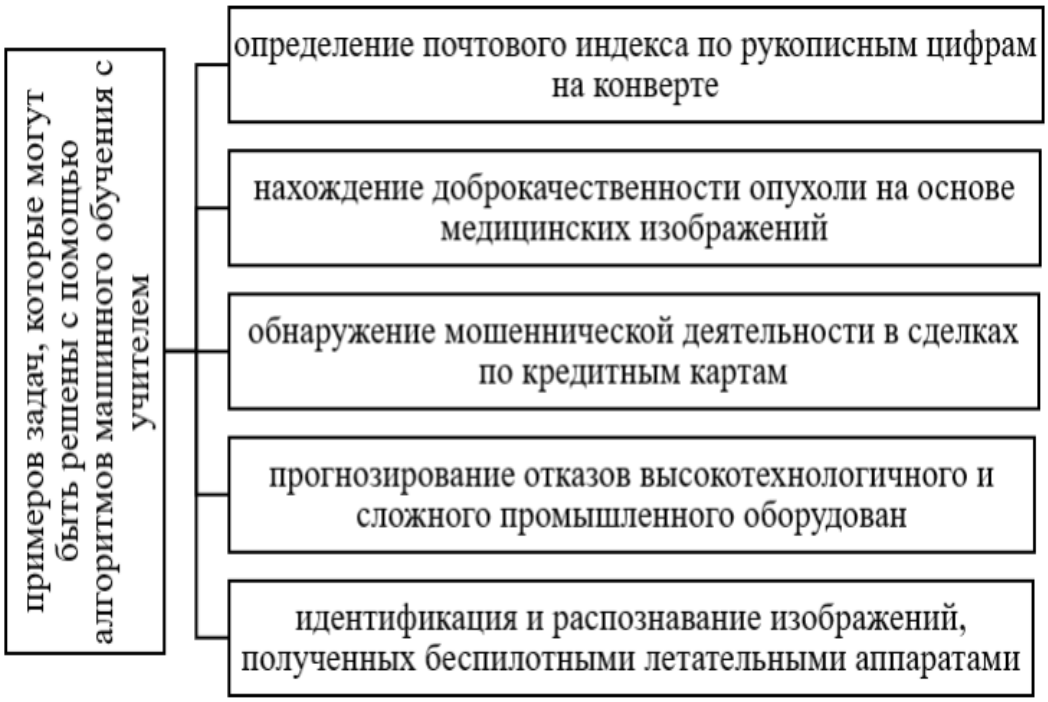

Figure 1. Examples of tasks solved by ML

Summing up the results of the above study, the following conclusions can be drawn. In the era of the development of information technology and the transition of mankind to the fourth industrial revolution, a huge amount of data in digital format or big data – Big Data – has appeared. Many different technologies for storing, computing, mathematical tools for analyzing and processing data have appeared. All this leads to the emergence of new business processes, scientific fields and professions. One of these areas is Data Science and Machine Learning. Today, the market needs highly qualified specialists who understand these areas, this is more relevant than ever and will be relevant for a very long time. The use of data and data science at the moment is not limited to just one field of IT, every business needs specialists who understand large amounts of information and are able to competently analyze it and work with the data obtained. It is these facts that make Big Data, Data Science, and Machine Learning very important and relevant for the modern world.

References

1. Veretennikov A.V. Big Data: Big Data Analysis Today – 2017. — № 32 (166). — P. 9-12.
2. Lee R. Big Data, Cloud Computing, and Data Science Engineering. — Cham: Springer. — 2020. — 214 p.
3. Silen D., Meisman A., Ali M. Fundamentals of Data Science and Big Data. Python and Data Science. St. Petersburg: Piter. — 2017. 336 p.
4. A great guide to Data Science for beginners: terms, application, education, and entry into the profession. [Electronic resource]. Available at: https://netology.ru/blog/01-2020-gid-po-data-science (accessed: 4.01.2022).
5. Müller A., Guido S. Introduction to Machine Learning with Python. A guide for data scientists. Moscow: Williams. — 2017. - 393 p.
6. Overview of the Data Scientist profession. [Electronic resource]. Available at: https://habr.com/ru/company/netologyru/blog/329068/ (accessed: 5.01.2022).

7. Cheng Q., Li H., Wu Q., Ngan K. Hybrid-Loss Supervision for Deep Neural Network. — Neurocomputing. — 2020. — Vol. 388. — P. 78–89.
8. Big Data from A to Z. Part 1: Principles of working with big data, the MapReduce paradigm // Habrahabr. [Electronic resource]. Available at: https://habrahabr.ru/company/dca/blog/267361/ (accessed: 4.01.2022).