DevOps and Big Data: Why should you combine them?

Lluc Feixa Morancho
llucfm@kth.se

May 4, 2023

I/We certify that generative AI, incl. ChatGPT, has not been used to write this essay. Using generative AI without permission is considered academic misconduct.

Contents

1 Introduction

The total amount of data that is created and consumed globally has increased exponentially over the last years and it is expected to grow even more in the coming years, as technology is present in almost every aspect of our lives. That is when terms like Big Data come in to play.

In 2021, around 79 zettabytes of data were generated worldwide, but for 2025, it is expected to double that amount[1], as it can be seen in Figure 1. So, if the data continues to grow, some things have to be done to be able to manage all the information generated in an efficient way and also be able to use greater quality data.

For this reason, combining DevOps and Big Data plays a key role for organizations and companies to build Big Data applications faster, more efficiently and with greater quality, scalability and agility.

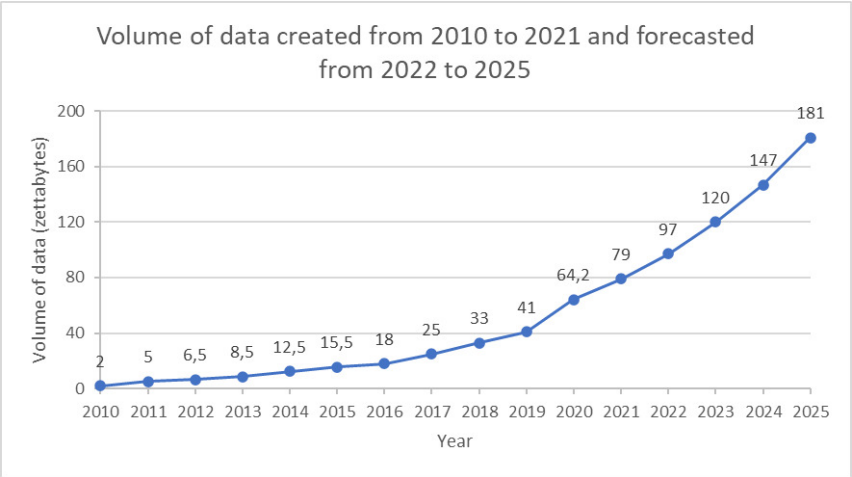

Figure 1: Volume of data created from 2010 to 2021 and forecasted from 2022 to 2025.[2]

2 Big Data

2.1 What is Big Data?

The term Big Data refers to large and complex data sets that can come from different sources. They may arrive in increasing volumes and with more velocity. This complex data comes in such high volumes that the traditional software used for processing it cannot manage it. But then, why should you want to use this large volumes of data? Well, the answer to that is easy, they can be used to solve some business problems that, otherwise, you wouldn't have been able to work out before.

Moreover, when working with Big Data, some other processes are involved in addition to the acquisition of data. The organizations also have to deal with large storage systems to save all the information they receive as well as the infrastructure to share, visualize, transform and test this data.

Apart from what has been mentioned, the most important thing in Big Data is its analysis, because at the end, that is what every organization wants, to examine what they receive as inputs and then know what to do with it. All this process has to be really optimal to be able to work with the maximum efficiency.

2.2 What are the three V's of Big Data?

To help understand more what the concept of Big Data [3] means, there are three properties, called the three V's of Big Data [4] (see Figure 2), that define it. These three V's are Volume, Velocity and Variety. Some sources think that these are the only V's that define Big Data. However, Big Data science experts often cite two or even three more [5], which include Value, Veracity and Variability. The reason why these latter are sometimes not included is the slight ambiguity in their definition.

- **Volume:** This first V refers to the massive amount of data which is growing a lot for each sector. This data can provide better prediction for the future of a company. Nevertheless,

large companies must have the appropriate infrastructure and storage systems to keep all the data that they acquire. An easy example to think about is any social media like Instagram or Facebook that has to store a lot of images and videos every day.

- **Velocity:** Not only it is related to how fast the data is received, but also to how quick it can be analysed in order to make decisions. Some of these decisions have to be made in real time with data arriving in real time, so the velocity of analysis of such amount of data has to be fast. On the other hand, others can collect the data at a lower pace with more time to make conclusions. An example for this V can be the sensors data, as with the increase of the Internet of Things devices more and more sensors will be used, which will create a lot of data at the same time, and some of them need to have a quick response.
- **Variety:** The last of the three main V's refers to the fact that the data can come from many different data sources, that it can be really different in terms of the level of the schema (structured and unstructured data), but also from the type of data received, as it can be text, video, audio and a lot more. Although having different types of inputs is useful, it creates more work to analyse it and make it work. A simple example of it are emails, as every one is different from the others and containing distinct files, so it tends to be unstructured.
- **Value:** This property refers, as its name states, to how valuable or useful the data is to the organization or company that is collecting and using it. It goes without saying that the value from the data strongly depends on the business requirements and business processes of an organization. The other V's mentioned often contribute to the value.
- **Veracity:** It refers to the quality and accuracy of the data and also to the level of trust that an organization or company has in the data that it is acquiring. If the data comes from an unreliable source or has some missing information, veracity can come to risk. As Peter Warden writes in his book [6] "I probably spend more time turning messy source data into something usable than I do on the rest of the data analysis process combined.". As he clearly states, having accurate data is a key point so you don't have to waste more time than the needed on preprocessing the data before analysing it.
- **Variability:** Finally, variability is related to the ways in which the data that is collected can be used, formatted and structured.

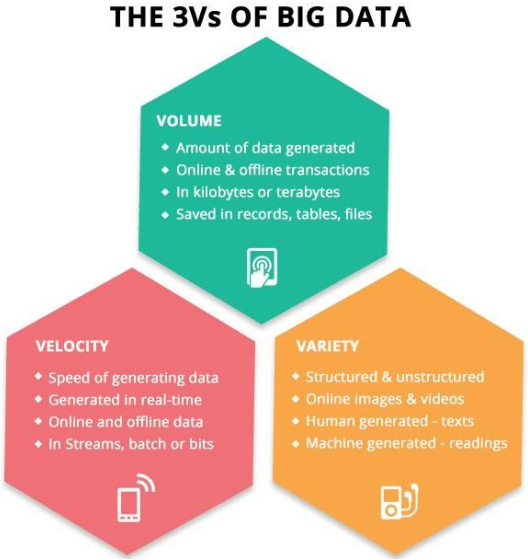

Figure 2: The three V's of Big Data.[7]

2.3 Statistics of Big Data: How is it evolving?

As it has been explained before, the amount of data that is created is increasing, so the Big Data market is also growing [8]. In fact, it is expected to be worth around 100 billion dollars by 2027, so it is a really good choice for organizations to invest in this field. That is why more than 97

percent of them claim to be financing Big Data and AI projects. Nevertheless, the inability to understand and handle all the unstructured data (which represents 80-90 percent of the data that internet users generate) is what holds 95 percent of the companies back.

3 Combining DevOps and Big Data

The first question that may come to your mind when thinking about combining DevOps and Big Data is: Why do we need to combine them? [9] One of the reasons for it is what has been mentioned the most during the essay, which is the huge amount of data that is created and how it has to be managed in an organization or a company. Furthermore, some Big Data projects can be difficult to handle as they have to be delivered fast to keep up with the competition. Another challenge is that the projects have to quickly respond to some changes that may occur.

The traditional approach was to split the organization up in different divisions where the different teams would work in different aspects. However, this lack of collaboration (as everyone is focused only on what they have to do) is not enough and slows down the whole process.

On the other hand, when using DevOps, all participants of the different stages of the process (preprocessing the data, analysing it...) of the software delivery pipeline are brought together. Moreover, if the members of the team are collaborating in this way, they share a more clear final objective, so the efficiency is considerably increased.

Taking all of this into account, it is becoming more and more standardized in Big Data companies to embrace DevOps and include data specialists within the CI/CD process.

3.1 Which are the benefits?

It is needless to say that DevOps does not include analysing data, so for an organization, employing data specialists would be necessary. This can improve the efficiency and power of Big Data operations when combined with DevOps. This combination has some important benefits for organizations, some examples of these can be:

- **Effective Software Updates:** This first benefit refers to the fact that if developers collaborate with data specialists before starting to write the code for the Big Data application they have to work with, they can learn more about which types of data they have to deal with, so the future software updates would be deployed more efficiently.
- **Minimal Error Rates:** Related with the first point, the developers have to know the types of data they are working with, as errors increase when organizations have problems with data managing while writing and testing their software. These errors can be easily reduced (and a lot of time can be saved) if there is a strong bond between DevOps experts and data specialists.
- **Streamlined Processes:** When translating and migrating data, the project that an organization may be working on could be slowed down and a lot of time can be lost. However, combining DevOps and Big Data can streamline the operations and improve the quality of the data a company works with. Moreover, if the workers don't have to worry that much about these processes, they can focus more in other important stuff.
- **Continuous Analytics:** It is similar to the continuous integration (CI) in DevOps. Here, this combination can simplify the analysis of the data and also automate it with algorithms.
- **Consistent Environment:** DevOps states that a development-friendly environment should be as similar as possible to what it would be expected in the real world. This is not really feasible when Big Data takes part. The environment is not really easy to create when a lot of complex data sets and different types of data are involved in the developing of the software. That is when the data specialists can play a key role in providing answers to the developers to help them produce enterprise-level software.
- **Accurate Feedback:** Finally, once the software is deployed, the next step is to collect the feedback to determine which are the strengths and weaknesses of it. This is where DevOps and Big Data can be combined in a very powerful way.

3.2 Which are the challenges?

Although combining DevOps and Big Data offers a lot of benefits to organizations, there are some challenges that may be found during the integration of DevOps and Big Data. These issues have to be solved to make the whole process work and obtain the most of it. Some challenges related to Big Data can be seen in Figure 3, although a few of them are already solved by DevOps, like the scalability problems.

- **Greater Understanding:** It is easy to think that if an organization wants people from different fields working on the same project and strongly collaborating, they will have to learn and have a really good understanding of their colleagues. The DevOps team should learn how to implement analytical models and what can be the Big Data problems. On the other hand, data specialists must be able to know some advanced techniques during the process.
- **More Resources:** An organization that wants to combine DevOps and Big Data will need more resources and cloud computing technology to get the maximum efficiency possible. These services allow IT departments to be more focused on business value than on fixing problems related to the hardware and operating systems part.
- **Testing:** It is really primordial to test the different functionalities of analytic models faster and with more detail.

Figure 3: Big Data challenges.[10]

4 Reflection

After an organization analyses the advantages and disadvantages of implementing this powerful combination, it would seem quite obvious that it would be the right choice to do it. Nevertheless, there are a lot of companies and organizations around the world that don't want to get out of their comfort zone and innovate in their processes. They also might feel that they don't need it as their company is working well. However, there is always room for improvement and this combination could take the organization to the next level and become more efficient and successful. It goes without saying that the amount of data will increase year after year so, why not starting earlier with these in your organization?

5 Conclusions

We live in a strongly dependant-technology society [11] where everybody is generating and consuming tons of data in a daily basis, and the tendency is to increase exponentially throughout

the coming years. Organizations have to adapt to this growth of data, as they have to be able to manage it (saving it in storage systems) and also analysing it and knowing what to do with the results. This large volume, the complexity and the different types of data is something that can not be easily handled with the traditional methods, as they are not efficient enough to keep things working. There is when combinations that at first sight may seem to not be related work together and benefit each other by streamlining the processes. So, the key take-away of this essay is that combining DevOps and Big Data the efficiency of developing software in short and long term is highly incremented and organizations can take a lot of benefits from it.

References

[1] Ogi Djuraskovic. Big data statistics 2023: How much data is in the world? URL: https://firstsiteguide.com/big-data-stats/, 2023. Last accessed 3 May 2023.

[2] Petroc Taylor. Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. URL: https://www.statista.com/statistics/871513/worldwide-data-created/, 2022. Last accessed 3 May 2023.

[3] Oracle. What is big data? URL: https://www.oracle.com/big-data/what-is-big-data/, 2023. Last accessed 3 May 2023.

[4] Nagham Saeed and Laden Husamaldin. Big data characteristics (v's) in industry. *Iraqi Journal of Industrial Research (IJOIR)*, page 2, 2021.

[5] Ben Lutkevich and Ivy Wigmore. 3 v's (volume, velocity and variety). URL: https://www.techtarget.com/whatis/definition/3Vs/, 2023. Last accessed 3 May 2023.

[6] Pete Warden. *Big Data Glossary*. O'Reilly Media, Inc., 2011.

[7] Coforge-Salesforce BU. Understanding the 3 vs of big data - volume, velocity and variety. URL: https://www.coforge.com/blog/understanding-the-3-vs-of-big-data-volume-velocity-and-variety/, 2017. Last accessed 3 May 2023.

[8] Christo Petrov. 25+ impressive big data statistics for 2023. URL: https://techjury.net/blog/big-data-statistics/, 2023. Last accessed 3 May 2023.

[9] Vipul Makwana. Big data and devops – winning combination for global enterprises. URL: https://readwrite.com/big-data-and-devops-winning-combination-for-global-enterprises/, 2022. Last accessed 4 May 2023.

[10] NIX United. 12 big data issues growing companies face. URL: https://nix-united.com/blog/12-big-data-issues-growing-companies-face/, 2022. Last accessed 4 May 2023.

[11] Jenny C. McCune. Technology dependence. *Management Review*, page 10, 1999.