IONOS

Big data in the age of commercial forking

Solutions to the licensing dilemma

IONOS

Contents

1 Introduction

Open source is often used in big data. But the fact that software is licensed as open source does not necessarily mean that it is free of charge and can be used without licence terms. Companies, organisations and even government agencies need to be aware of which terms and conditions must be fulfilled to use a piece of software, and how much it actually costs.

Operation in proprietary data centres – on premises – is often expensive, complicated and difficult to manage. In this case, it may make sense to use cloud-based, open source solutions. If so, data security, data privacy and data sovereignty are key considerations. This white paper will illustrate how you can bring these three things under one roof, and make big data analysis significantly easier.

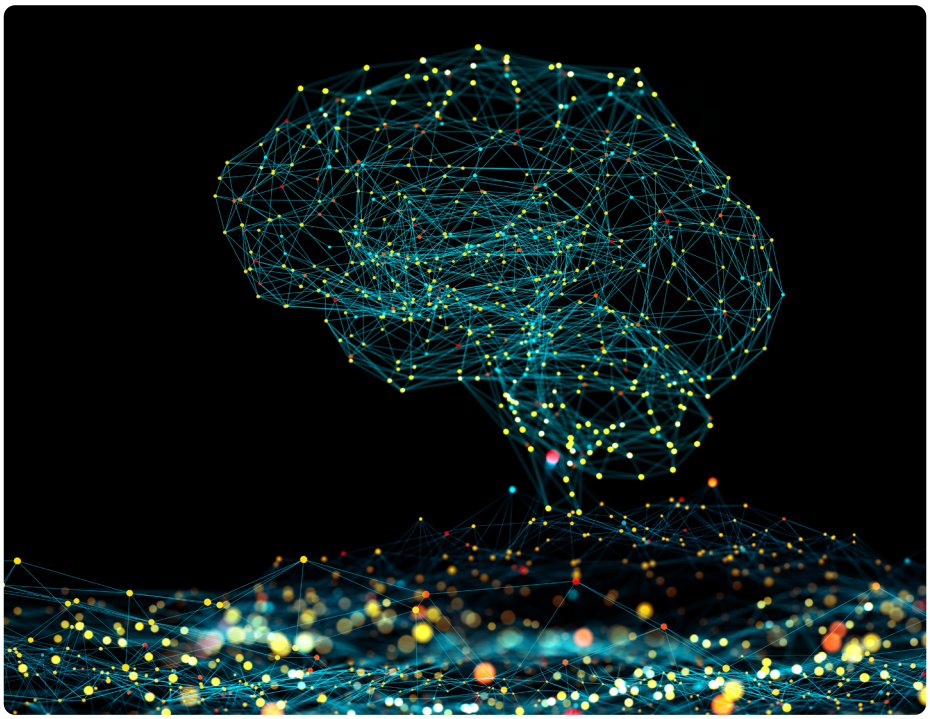
IONOS

2 The status quo: Big data, data science and AI

An increasing number of artificial intelligence (AI), machine learning (ML), and deep learning (DL) technologies are being used in the field of big data and data science projects. Major quantities of data can also be processed far more efficiently using AI technologies than though traditional means. This trend often relies on several solutions that achieve their full potential only when they are used together. Connecting these services is not only challenging from a technical standpoint, but also in terms of licensing rights.

Open source, AI and big data

Big data projects specialise in storing and analysing large quantities of data, whereas AI solutions help to analyse and process the data. Open source applications are used in numerous instances in all the areas involved – i.e. big data, data science and AI. However, there are a few things that managers in companies, organisations and government agencies need to be aware of.



The use of AI technologies in data analysis is a major trend that's continuing at a rapid pace in 2022. Data science machine learning platforms (DSML) are becoming ever more widespread and are mostly based on several solutions, tools and frameworks. Dataiku, Databricks and Microsoft are some well-known solution providers.

Generally speaking, cloud platforms are ideal in this context because both licence terms and scalability can be implemented far more easily. AI and big data require powerful hardware and a lot of storage space that must be effectively available in the network.

Cloud solutions provide the necessary scalability, and open source solutions can play a big role if used optimally.

> Companies, organisations and government agencies want to process their data in increasingly large quantities, at increasingly fast speeds, and with greater effectiveness. At the same time, the data must be securely stored, data privacy must be preserved and even data sovereignty is tremendously important. In essence, this is only possible with data science coupled with AI technologies from European data centres. AI-specific infrastructure stacks (MLOps and AIOps), infrastructure-as-a-service (IaaS) and platform-as-a-service (PaaS) are also among the major trends and ensure that companies, government agencies and organisations can quickly and securely access the necessary hardware and software.

Big data is already operating with real-time analyses in many fields, which naturally puts high demands on a company's infrastructure. Added to this are the constant changes in the infrastructure. Besides on-premises usage – i.e., in the local data centre – an increasing number of companies are using the cloud or hybrid networks, and the use of multi-cloud infrastructures is becoming increasingly widespread too.

Modern data stacks and cloud warehouses

Modern data stacks offer various tools and technologies that can be used to analyse and process data. These are often different, interconnected applications, which can be run in the local data centre (on-premises) and in the cloud alike. However, this can result in a combination of numerous tools and solutions that make little sense to operate locally. Often, there isn't sufficient computing power, storage capacity and staffing to carry out the projects.

The complexity of big data and data science

For several years, there has been a rapidly growing trend of using big data and data science solutions in the cloud, partly driven by cloud warehouses. With this technology, the system obtains its data from many different sources, extracts it and processes it. This also includes comprehensive analyses using AI technologies as well as visualising the information, sometimes in real time as well. Examples include Amazon Redshift, Snowflake, Google BigQuery, and Microsoft Synapse. However, there are numerous other related solutions that quickly extract, load and transform data and largely use open source, even for software-as-a-service (SaaS) solutions. In this field, open source is an important foundation that plays just as big of a role in local, on-premises installations as in SaaS and PaaS.

New tools and solutions

Another example is the DBT (Data Build Tool). This open source tool helps to transform data and is increasingly used by data scientists to transform data from sources using SQL queries. Data engineering also plays a pivotal role here, where tools such as Spark are used. Using this kind of solution requires knowledge about licensing but also specialised proficiency in operating the solution itself. Conversely, a solution is far easier to use when deployed in the cloud, particularly in terms of both the licences and the actual usage, and even with regard to data privacy, data security and – of course – data sovereignty as well.

Due to new trends, solutions and tools, technologies are blending, and licences play just as much of a role as the actual operation of the solutions. For instance, data lakes and data warehouses coalesce into data lakehouses or unified analytics warehouses. Furthermore, AI/ML technologies are also used here, and they help to transform, process and subsequently analyse data that is more or less structured. Technologies are fusing together, which plays just as much of a role in the context of licensing as it does in the actual deployment and implementation of the solutions in the network or in the cloud.

New trends, solutions and tools bring about a coalescence of technologies.

In many cases, it is far more appropriate to use cloud solutions for big data than to use locally installed open source solutions. Licence terms tend to be far clearer in the cloud; furthermore, cloud solutions can also be scaled and used more easily, operated far more securely and be deployed to users faster and more effectively.

IONOS

3 The challenges

3.1 Risks when using open source for big data

As an approach to distributing software, open source has significant advantages. In most cases, the programs are available at least partly free of charge, and since the source code is public, security loopholes are quickly identified and the community helps to close them. However, it also poses some challenges.

- **Complex licensing can make using open source problematic**
  In the majority of cases, software solutions in big data are built with far greater complexity. Although open source is often used, in many cases it is used in combination with commercial products. This makes the licensing more complex because conventional licences and the often complex licence terms of the open source components become intertwined. This is something that needs to be taken into account to ensure proper licensing, and it also plays a significant role when planning costs.
- **Software forks make licensing even more difficult and often more expensive as well**
  Applications often have what are known as software forks. A software fork is when new projects split from established applications and continue to be developed separately, which naturally makes usage far more complex. If afterwards new licensing is used or if a former open source project switches to a commercial licence, it becomes very difficult to use open source software properly, legally and with the right licensing.

  Software forks are used in open source software far more often than you might think, and they already have a long-standing tradition. LibreOffice was forked from OpenOffice.org, and the web content management system Joomla! was created from Mambo. In 2002, VNC became paid software with version 3.3, which resulted in additional implementations under GPL that use the same protocol.

What exactly is GPL?

The GNU General Project Licence was created from the GNU project in 1983. Its objective is for the general public to receive free access to software while simultaneously preserving the right to make modifications. The GPL was the first licence that also gave users access to a software's source code. The condition is that modified versions of the software must also be freely available.

This last example shows how quickly a formerly no-cost, open source solution can become a commercial product. Using software forking to include open source technologies in commercial software that – in many cases – is also intended to be sold at a high price undermines the principle of open source and exploits it for private profit. Users do not benefit in any way from the fact that parts of the commercial software are based on open source.

- **There is often a lack of support when using open source software**
  The use of open source software poses some risks that we will illustrate in detail in this white paper. Besides the problems with licensing changes discussed in other sections, there are additional challenges that should be taken into account. Among them, for instance, is the frequent lack of support available when using open source solutions.

  Open source software often comes with no guarantee, and there is also often no guarantee that maintenance will be taken care of. Furthermore, open source software initiatives do not always receive adequate financing. This can result in the project either being cancelled or continued by a different manager under a different licence.

  Vulnerabilities in open source software are accessible to the public, which means that – in principle – they can be quickly rectified by the community. However, hackers also have access to the vulnerabilities and can exploit them. New functions and error corrections may not be added to open source solutions quickly enough. Naturally, this depends on the supporting community and on the developers who work on the source code.

- **Open source licence terms can change and thus pose problems to users**
  There are also more recent examples of changes due to commercial forking. Using the search engine Elasticsearch as an example, we can see how – even in big data – open source solutions can very quickly lead to licensing problems, and the use of open source can quickly become expensive and/or pose legal problems.

  If companies operate these kinds of solutions themselves, any challenges related to operating a local big data platform will also involve the issues of licensing, legally compliant operation and perhaps exchanging individual components. Furthermore, very few companies, organisations or government agencies can master these tasks themselves as well.

- **Popular open source software can quickly be turned into commercial software**
  Although the search engine Elasticsearch is open source, the company Elastic – which is where the software is developed in-house – has changed the software's licence terms.[1] Anyone who uses the open source solution in their environment must grapple with the legal and practical entanglements. Elasticsearch has been published as open source under the Apache 2 licence since 2010. At the same time, Elasticsearch is based on the Lucene library, which is also available as open source.

1. Banon (2021): Revisiting the issue of 'open', part II – Continuation of our blog 'Doubling down on open' from three years ago, available online (in German) at: https://www.elastic.co/de/blog/licensing-change.

In 2012, developer Shay Banon founded the company Elastic, which ultimately became publicly listed in 2018. In addition to the open source Elasticsearch, Elastic has published enhancements that are available under their proprietary licensing terms called the Elastic licence. This shows how complex it is to properly licence Elastic for use.

- **Open source licence terms are not set in stone**
  An increasing number of Elasticsearch's functions are falling under this new Elastic licence and as a result, they are by definition no longer open source. One of the causes is that companies such as Amazon Web Services (AWS) have begun to offer Elasticsearch as cloud software-as-a-service (SaaS). In response to this, Elastic placed components of Elasticsearch under the Elastic licence, making it virtually impossible for the software to be reliably usable as open source.

  AWS also responded to this and forked Elasticsearch for its platform, and the new solution is available as open source. However, Elastic has sued AWS as well as software service providers that offer solutions for the new software called Elasticsearch OpenDistro on AWS.[2]

- **New licences limit the freedom of open source**
  Furthermore, Elastic wants to publish new versions of Elasticsearch under a special licence that makes it virtually impossible to connect and use legally hosted versions of Elastic. Another prominent example of this kind of licence policy is MongoDB. This is another case where the licence was altered to make it practically impossible for third-party providers to host it. To accomplish this, they introduced the new Server Side Public License (SPPL) – which is now used for Elasticsearch as well – for the software in 2018. Anyone who places products under this licence, such as MongoDB or Elasticsearch, must fully publish the code of modified versions.

Server Side Public License

Server-side public licence (SPPL) is a software licence that was introduced by MongoDB in 2018. The licence is not recognised as an open source licence. If developers use SPPL-licensed software, then they also need to publish the source code of their own software it if is used in a public cloud or web service. This also applies to any other source code associated with the project. For instance, these also include software for system management, user interfaces, storage backend or backups. Many critics see its main purpose as preventing SPPL-licensed software from being used in other services.

2. Förster (2021): Amazon is to blame for ending open source, available online (in German) at: https://www.heise.de/news/Elastic-Amazon-ist-schuld-am-Open-Source-Ende-5030541.html.

Therefore, Elasticsearch is actually no longer open source and poses a risk – at least from a legal standpoint. The Open Source Initiative (OSI) did not accept the new licence terms from Elasticsearch since they severely limit freedoms for developers and for users as well. If one company largely controls parts of a piece of software and its licences, it can very quickly change the rules of the game. This will turn users and service providers into pawns for licensing; they will hardly be able to use the product legally with proper licensing, and they definitely cannot offer it to customers. This change in Elasticsearch's licences has resulted in a new fork called OpenSearch, which is available on GitHub. The developers use the Apache 2 licence here, while Elasticsearch now uses SPPL.

Elastic announced at the end of 2020 that they intended for Elasticsearch to be permanently licensed under the Apache 2 licence. This shows how quickly the situation can change and how reliable the licence statements are. Elasticsearch is no longer open source under the new SPPL licence. Anyone who uses services based on Elasticsearch in future should have its use reviewed by a lawyer since it can pose legal and, ultimately, massive business risks.

- **New software forks may be lower quality than the source**
  When open source software becomes commercial software as is effectively the case with Elasticsearch, it often results in forks that use the open source code. However, this in turn poses the risk of the forks not being authorised to use certain parts of the code and the need to develop new, fully independent code.

  Naturally, this takes a toll on the product's quality and compatibility in the long run since it is a completely new piece of software. Furthermore, taking Elasticsearch as an example, you can expect many developers to retire from the project and stop contributing to the open source components. Elastic no longer calls the code in Elasticsearch 'open source', but rather 'open code'.

3.2 Decline in the use of flexible open source licences

There has been a general decline in the use of open source licences for new software. As early as 2017, Stephen O'Grady from Redmonk observed that the GNU General Public License (GPL) was been used half as often from 2010 to 2017, and the trend is continuing well into the 2020s.[3]

Open source doesn't necessarily mean free of charge
Although an increasing number of software products are using open source Apache and MIT licences, only the GPL offers the maximum open use of source code. For instance, the Linux distributions Debian, Ubuntu or Fedora also fall under the GPL licence.

If a project uses code from another GPL project, then that project must also be licensed under GPL. GPL products may be sold commercially, but the GPL is very strict about this. This is one of the reasons why many developers are moving away from GPL and are using other licences. Anyone who runs open source software on their local network have to deal with it whether they like it or not, but those who use open source software in the cloud delegate these tasks to the PaaS or IaaS service provider.

An increasing number of companies are openly considering making money with open source. Apart from pure services and distributions, companies wish to use licences that enable them to respond more flexibly when the software is distributed accordingly. The compromise between providing no-cost open source software and the possibility of making money is difficult for many companies.

GPL versus Apache 2, MPL and MIT

The GPL effectively requires comprehensive and no-cost use of open source, which is the basic understanding of open source among large segments of the user base as well as among decision-makers in companies. However, developers and companies that provide software must be able to monetise their efforts. Apache 2 and MIT licences are significantly more flexible for this purpose, and so is the MLP licence. The Mozilla Public License (MPL) is also used often in big data. It permits the use of other types of licences, even when using other MPL projects in your own project. Anyone who develops a new big data solution and uses MPL projects in the process can publish their project under the Apache 2 or MIT licence without any problems. As a result, these licences are far more flexible, but the MPL-licensed code must be clearly separated from the rest of the code.

The attitudes of many developers have changed over the past few years. There should continue to be open source solutions, but there also needs to be a way to monetise them. Therefore, it is expected that more and more solutions will be published under the Apache 2 and MIT licences in the future. One prominent example of the Apache licence is the big data framework Apache Hadoop.

The Apache 2 licence is published by the Apache Software Foundation (ASF). When using this licence, developers may use any other licences, but they need to mention the licences of the products that are used and document any changes. The code may be used in closed source code scenarios and utilised for commercial purposes.

3. O'Grady (2017): The State of Open Source Licensing, available online at: https://redmonk.com/sogrady/2017/01/13/the-state-of-open-source-licensing/.

The MIT licence from the Massachusetts Institute of Technology has been in existence since the 1980s. It is one of the simplest licences on the market and one of the most flexible as well, which is why it is also used in many big data and data science projects. This licence has virtually no restrictions as far as using open source components is concerned. You only need to include the original copyright and the licence terms of the software that is used. It fully exonerates authors from liability.

In future, GPL will probably be used primarily for free software. Hence, companies, organisations and government agencies that use open source software should plan in good time about how to offset a change in licences or modifications to licensing conditions. The principle that open source means forever free of charge is long gone.

Summary

Licence changes can pose a massive business risk if companies, organisations or government agencies use a piece of software for essential services because it is very possible that changes can prevent continued legal use entirely, requiring comprehensive changes in the local infrastructure.

3.3 Flexibility and security in big data and data science

Open source software is frequently used in big data and data science. Among other things, this is due to the fact that major companies such as Netflix, Twitter, Facebook, Microsoft, and Google develop solutions that they first use in-house to process their enormous data volumes. As soon as the tool is stable enough, it is provided to the public as open source software. The advantage for companies is that the solution undergoes continuous development and improvement by a major community. The challenge for companies who wish to use this solution consists of finding the right tools and properly licensing them while simultaneously ensuring that the components work together in harmony.

Access to open source tools using Cloudera and Hortonworks as an example

Since at least 2019, the licensing problem mentioned above has reached a new peak in the big data market as well. When Cloudera and Hortonworks were merged, nearly all of Cloudera's solutions became paid services, and since then, anyone accessing the service provider's solutions needs to pay a subscription to use the software.

This was followed by price increases for products that were already fee-based and extensive changes in the use of cloud solutions. Since the spring of 2021, all of Cloudera's customers have become painfully aware of these changes to the licensing policy.[4] These events make it abundantly clear that companies, organisations and government agencies must always consider alternatives so that they are prepared for any changes if necessary.

4. Cloudera (2021): Cloudera pricing & licensing updates, available online at: https://www.cloudera.com/products/pricing/pricing-update.html.

That is why Cloudera and Hortonworks offer distributions or even collections of various different tools that are available as an all-in-one solution. Although the individual tools continue to be open source, the complete, all-in-one solution is fee-based, and the full distributions are commercially sold.

This white paper has already demonstrated the tremendous dynamics that the open source market is subject to as far as big data and data science are concerned. There are already numerous technologies, applications, solutions and tools, and new approaches and applications are constantly being added to the mix; many of these solutions work closely together, merge with or complement each other. This is why it is important for companies, organisations and government agencies to avoid subjecting themselves to a rigid infrastructure, but rather have the ability to respond flexibly and quickly. In addition, there is also the need to always be up to date, have the ability to quickly update software, and continuously optimise the environment as well.

Flexibility and security play a fundamental role in big data and data science.

Many companies are also taking the infrastructure as code approach in this regard, which achieves maximum flexibility for big data environments. With infrastructure as code, all solutions can be shown in the big data infrastructure as code that can be expanded and quickly adjusted at any time thanks to versioning and the CI/CD approach. However, such approaches can rarely be used in proprietary data centres since time and the required expertise are often lacking.

Finally, price also plays a role, and as the Cloudera/Hortonworks example shows, licensing models are an important factor in planning a data science environment, and modularity is just as important. Functions in the infrastructure must be quickly replaceable while simultaneously remaining scalable as well. The relationship between scalability and price development plays an important role here too.

Security is absolutely critical for big data and big data science. Securely handling data is tremendously important, and not just because of the GDPR. In its Report on the State of IT Security in Germany 2020, the German Federal Office for Information Security (BSI) warned of a sharp increase in the number of attacks on the IT infrastructure of companies and organisations. Just the number of new malware variants alone increased by roughly 117.4 million in 2020, which amounts to around 320,000 new malware programs per day. This increase once again amounts to over a third compared to the previous year.

According to a representative study conducted by Bitcom in 2021[5], criminal cyber attacks caused 220 billion euros' worth of damage to the German economy across all industries in 2020, which is double the damage of the previous year.

The attacks resulted in numerous security incidents linked to the successful encryption of data by ransomware. The companies affected were blackmailed or experienced accompanying computer sabotage that included rendering IT systems useless. These incidents alone quadrupled in 2020 compared to 2018/2019. Almost 10 percent of the surveyed companies regard it as an existential threat. DDoS attacks have been known about for years and they are increasing every year; they increased by 10 percent from 2019 to 2020. In August 2021, Microsoft fended off its largest DDoS attack to date on 70,000 computers on their Azure Cloud.

An increase in the complexity of the environments also increases the complexity of the security infrastructures that are needed to safeguard the data. In this case, companies should recruit a partner who is able to provide powerful, flexible and yet secure infrastructure.

4 The solution

4.1 Big data managed services

With optimal planning, open source components for building and operating scalable data and streaming infrastructures can be the right approach for running big data in companies. These mainly include the following components:

- Modern data warehouses and data lakehouses
- Event streaming
- Machine learning and artificial intelligence

It is important to keep an eye on modularity while also ensuring alternatives. Generally speaking, it is never a good idea to focus on one single software component. The examples in this white paper show that the consequences can be expensive, risky and ultimately dangerous to the company's economic viability as well.

5. bitkom (2021): The German economy in the crosshairs; more than 220 billion euros worth of damage per year, available online (in German) at: https://www.bitkom.org/Presse/Presseinformation/Angriffsziel-deutsche-Wirtschaft-mehr-als-220-Milliarden-Euro-Schaden-pro-Jahr.

This approach enables companies to achieve maximum flexibility and simultaneously the greatest data privacy while also significantly reducing the complexity of configuration and operation. Of course, it also has benefits for maintenance and updating the complete environment, which will continue to play a crucial role in the coming years. In this scenario, companies do not need to worry about security updates. The tools work together optimally, and the service provider also ensures that it stays that way when performing updates.

4.1.1 Infrastructure as code

Big data managed services enable big data infrastructures to be operated with a focus on the user and labour-saving provision. This is where the previously mentioned infrastructure as code approach has a positive impact. It offers several benefits when implementing new functions, updates, and even for replacing components in case licensing terms or when other things change. The infrastructure as code approach affects the following aspects:

- Automatic provisioning
- Configuration
- Monitoring
- Updates
- Maintenance

4.1.2 Open source with infrastructure as code

Open source with infrastructure as code uses conventional solutions for big data and data science that are based on open source, and it can also incorporate common software solutions for big data in the company's IT. At first, it does not matter if the solutions here are used in the local data centre (on-premises) or in the cloud, because hybrid networks and even multi-cloud infrastructures are also feasible without any issues. In this case, it would be ideal to use preconfigured distributions from the following fields:

- Big data
- Stream processing
- Business intelligence
- Machine learning
- Artificial intelligence

As we have already seen, combinations of these technologies in particular can be meaningfully used in this context as well, and this is also aided by the modular approach that allows you to include additional modules or replace them at any time. Such a managed service solution is complemented with extra security thanks to professional protection of the systems behind big data.

4.1.3 Practical example: MARISPACE-X in Gaia-X

Data sovereignty naturally plays a pivotal role in data processing for big data and data science. Together with its partner Stackable, IONOS Cloud is contributing a cloud infrastructure and services for aggregating and analysing data to the European Gaia-X project. Gaia-X is a project that is also being advanced by the governments of Germany, France and other EU member states. It is a multi-cloud architecture that is independent from other regions and offers European customers comprehensive data sovereignty via harmonised data exchange spaces.

Data sovereignty through a proprietary cloud infrastructure in Europe with Gaia-X

Gaia-X aims to build a data infrastructure for Europe that is independent from other regions of the world. The project was presented by the German Federal Ministry of Economics at the 2019 Digital Summit. The model for Gaia-X is the European Airbus consortium, and companies such as Deutsche Telekom, Bosch, the Fraunhofer Institute, SAP, and Siemens, among others, are contributing to the project. A majority of the software that is used for Gaia-X is based on open source, and AI is playing an important role for an increasing number of European companies. This demonstrates that European companies, government agencies and organisations do not necessarily want to entrust American corporations with sensitive data in this field, and this is precisely where Gaia-X comes into play.

MARISPACE-X offers a treasure trove of data that can save lives

MARISPACE-X from IONOS Cloud focuses on gathering maritime data. Many European companies and organisations gather vast quantities of geoinformation from the world's oceans and process them for their own specific purposes. However, the data gathered by many different companies has not traditionally been shared to enable efficient usage beyond the scope of individual areas. Drones, satellites, sensors, and a multitude of measurement information run in parallel to each other, collect data, and process it as well. If this data was to coalesce in a secure data room, it would create one of the most comprehensive sources of data for the maritime economy, and MARISPACE-X is following this exact approach.

This treasure trove of data is of extreme interest to European companies for numerous use and business cases. One example of this is the search for and retrieval of ammunition from the world's oceans. Ammunition poses a major hazard for shipping since grenades explode on a regular basis. In addition, the ammunition leaks poisons that can also endanger maritime life. 1.6 million tonnes of ammunition have sunk in the Baltic Sea alone, and it is estimated that an additional 1.3 million tonnes are in the North Sea.

Big data and AI help the search for ammunition in the world's oceans

north.io GmBH from Kiel in northern Germany collaborates extensively on MARISPACE-X and provides AI-assisted software that researches historical documents for ammunition sites. It also calculates the ammunition risks for various regions and performs a big data analysis on clearing ammunition. These kinds of applications require huge volumes of data and enormous computing power.

The challenges go far beyond what companies or organisations can affordably provide in their own data centres, which is also the reason why north.io is using the IONOS Cloud. For analysis, the company makes parallel use of mass data from other partner companies and the analysis of this data by other partners in the MARISPACE-X group. One example of this is Stackable and their free and – unlike the software components described earlier in this white paper – truly open distribution of optimally integrated open source projects for modern data platforms.

A mutual storage platform creates new business models

As Rainer Sträter, Head of Global Platform Hosting at IONOS, said "With MARISPACE-X, all partners are seeking to create a mutual, virtual pool of all available maritime data and thereby create new business models as well." To make this possible, IONOS – the consortium leader – is providing the cloud infrastructure for MARISPACE-X. In addition, the data is merged and analysed on a mutual storage platform that uses open source components.

To ensure effective data analysis, the required computing performance is also part of the project. Communicating with the computing power, the mutual storage platform and the huge volume of data, analysis applications can determine precisely where the hazardous substances are and the efforts that are required to retrieve them.

Climate protection, offshore wind energy, and the Internet of Underwater Things with big data

Beyond finding ammunition, additional applications of MARISPACE-X include calculating the locations of wind farms and the optimal installation of undersea cables. The project has also been focusing on offshore wind energy, the Internet of Underwater Things (IoUT) and biological climate protection since mid-2021.

Satellite images and underwater sensors help to search for algae surfaces and seagrass beds that play a crucial role in storing carbon dioxide. MARISPACE-X helps to analyse existing areas and can simultaneously use the analysis to find new growth areas. Soil condition and ocean currents are also part of its wide-ranging calculations.

Mutual data exchange with international data spaces

MARISPACE-X focuses on mutual data exchange among the participating partners. In addition, there are also connectors available that connect data sources and the processing systems to each other. The standard that is used here – called International Data Space Association (IDSA) – is fast, secure and subject to European guidelines and legislation on data privacy and data security. Only European companies comprehensively adhere to such guidelines since the data collected is stored only in EU data centres.

IONOS Cloud provides the computing power and storage while also delivering the necessary compression of the data. Alongside north.io and IONOS, numerous other European companies that provide data and perform collective calculations and analyses with MARISPACE-X are also part of the consortium. Its members include the companies TrueOcean GmbH, Stackable GmbH, MacArtney Germany GmbH, Siemens Gamesa Renewable Energy, Quality Positioning Services B.V., WINDEA, Offshore GmbH & Co. KG, OffCon24 and Wallaby Boats.

Even scientific organisations, public institutions and associations contribute to the project – these include the Fraunhofer Institute for Computer Graphics Research, the GEOMAR Helmholtz Centre for Ocean Research Kiel – AG Deep Sea Monitoring, the University of Rostock, the University of Kiel, the Maritime Cluster Northern Germany, the German Association for Marine Technology, TransMarTech Schleswig-Holstein, the Schleswig-Holstein Chamber of Commerce and Industry, Labs Network Industrie 4.0 e.V., as well as Schleswig-Holstein's Ministry of Energy, Agriculture, Environment, Nature, and Digitalisation and the Ocean Data Alliance.

4.2 Big data frameworks and platforms as a managed service in the European cloud

The possibilities mentioned above are just a few examples of the functions that big data offers in conjunction with AI. The analysis largely uses open source solutions that are merged, managed, updated and maintained in a managed cloud.

Apache Kafka – managed big data from the European cloud
Apache Kafka is one of the most well-known open source solutions in big data. Originally developed by LinkedIn, this solution is able to import and process large volumes of data from many different sources. Apache Kafka's main benefit is its very high data throughput; it can use real-time events, logs, and other data, and the software can do event streaming very well. These events can be saved in files for additional processing. In contrast, Hadoop is excellent at processing huge volumes of data in files, which can in turn be created by Kafka.

Kafka combines the RAM, cache, storage systems and storage management of the local operating system, which enables efficient distribution of computing and storage tasks. If the underlying infrastructure is able to quickly provide this data like in the IONOS Cloud, for example, then Kafka can work wonders.

Apache Zookeeper, which is also open source, is subsequently used so that the components can be integrated with each other and be available at peak performance. Zookeeper is a central service for maintaining configuration information, naming objects, and the distributed synchronisation of group services. The solution is mainly intended to prevent uncontrolled proliferation in the infrastructure and to help achieve a consistent configuration, which is ideal for a stable and effective cloud infrastructure.

Kafka can also be used to process major volumes of data and send them directly to the Hadoop system. Apache Kafka also works together with Apache Strom, which can read data from other streams – including from Apache Kafka – before they are consistently stored and finalised.

Big data in new dimensions with Apache Spark and NiFi
In major Hadoop or big data environments, standard options and queries are often not sufficient to efficiently analyse data. The Apache project Spark has tackled this problem and offers efficient, real-time analysis of data in Hadoop clusters.

The Apache project Spark offers efficient, real-time analysis of data in Hadoop clusters.

Spark achieved a new world record in the 100 terabyte class in the Daytona Gray Sort Benchmark. The previous world record was 72 minutes and was achieved by a Hadoop MapReduce cluster. Spark beat the old record at 23 minutes and using one-tenth of the computing power at that. Clearly, Spark can make inroads into areas of big data processing that Hadoop is unable to handle.

Apache Spark enhances the abilities of Hadoop clusters with real-time queries. To accomplish this, the framework provides in-memory technologies, meaning it can save queries and data directly to the cluster nodes' RAM. Since the queries can also be distributed among several nodes in parallel, performance is significantly increased. To use it, you need powerful hardware that can be provided as needed via the IONOS Cloud.

Apache Spark is intended to replace MapReduce in Hadoop and offers much faster data queries. According to the developers, the speed is a hundred times faster. The framework is already being used by major companies that need to process large volumes of data. Prominent examples include NASA, Intel and IBM. Online music service Spotify also optimises its playlists with Spark, and Spark is likewise used in the IONOS Cloud, i.e. in the MARISPACE-X project.

Apache NiFi also plays a significant role in this context. It is an additional open source solution that is based on the Apache licence. NiFi's task is to automate the flow of data between systems. When working in harmony, Spark, NiFi, and Kafka can be used to process, transform, and even analyse enormous volumes of data. And when this takes place on a shared platform that is optimised for collaboration and has the corresponding computing and storage power, it results in tremendous benefits as far as the speed of the analyses are concerned.

Apache Druid: Low-latency data storage

Apache Druid is open source data storage for analyses, which enables low-latency business intelligence queries of event data. Real-time access is just as possible as fast data aggregation, and this open source solution can be used as an alternative to conventional data warehouses. Druid focuses on event data and time series, although it can also analyse other data.

In many environments, Druid is also used in parallel with data warehouses. It can also make sense to obtain this data from Apache Druid wherever real-time analysis, interactive interfaces with updated data or ultra-modern query apps are used. In such a scenario, Druid can in turn obtain its data from a data warehouse, process it, and make it available. Afterwards, the data warehouse can deliver reports and archive data in parallel.

Running Druid can make sense if high performance is needed during analysis.

Druid supports numerous sources of data, including in the cloud. Examples include Amazon S3, Apache Kafka, Azure Event Hub, Amazon Kinesis, Google Cloud Storage, Azure Data Lake, local data and, of course, data from databases as well as locally stored data. After connecting the source, Apache Druid reads it and indexes the data, and Druid also supports Hadoop.

Apache Druid can be integrated with Apache Hive and Apache Ambari, and you can even create OLAP cubes with SQL or invoke existing Druid cubes as part of a real-time analysis.

When optimally configured, open source can be a huge help with the analysis

The examples in this section of the white paper show that there are numerous open source solutions that can process and analyse vast volumes of data. When tools such as Kafka, Spark, NiFi, Druid, and Hadoop work together, companies, organisations and even government agencies can quickly and securely analyse nearly any amount of data. However, this requires enormous computing power, storage capacities and tremendous amounts of specialised knowledge. Stackable, a startup from Germany with a data platform that only uses software with non-restrictive open source licensing, sheds light on the licensing madness and enables trustworthy big data in a thoroughly transparent environment.

In order that the solution can be worked on jointly – and only then will it result in an optimum benefit – hardware and software must act in concert and be optimally configured, and this also entails regular updates and maintenance. Essentially, an environment like this can be created only in the cloud, ideally by a reliable partner who operates in accordance with European law.

When big data analyses are executed in European clouds such as the IONOS Cloud, a lot of attention is paid to preserving data sovereignty. At the same time, the service provider ensures that the solutions always work together in harmony, are up to date and that the hardware used works perfectly with the software. Furthermore, the modular structure ensures that individual components can be replaced while preserving the entirety of the solution.

IONOS Cloud's data centre is located in Europe and is subject to the GDPR.

4.3 IONOS and the Stackable cloud data warehouse

The Stackable cloud data warehouse provides a modern BI solution consisting of powerful data processing and distribution pipelines. With S3 object storage, even the data storage is highly scalable.

Fast and comprehensive SQL enquiries and OLAP-based business analyses are used to fill charts and dashboards with relevant data, which are created based on many different open-source solutions. HBase and the Hadoop platform are mainly used for this, along with Kafka, Spark, Superset, Atlas, and many other open source tools that connect Stackable to an ecosystem for big data managed services. Developers consistently enhance, update and improve the open source tools in collaboration with the open source community.

The data centre for the cloud-based solution is located in Europe and is subject to the GDPR. It is also important that the underlying cloud provider be a European company and therefore provide maximum protection from the US CLOUD Act. ISO 27001-certification is also important. Local customer support for the solution is available, along with local consulting on building a cloud data warehouse. As a result, customers benefit from fast turnaround times and get quick assistance with building, operation or expansion.

Companies and organisations can easily create their virtual data centre via the intuitive Data Centre Designer (DCD) user interface. Stackable big data solutions are based on the container orchestration system Kubernetes. Furthermore, Stackable also provides pre-built configurations. The setup process is significantly easier than with comparable solutions from US service providers; furthermore, the cloud data warehouse differentiates itself from similar offerings via exceptionally qualified support in the local language at no extra cost.

By using vanilla Kubernetes, customers also benefit from maximum flexibility, since it allows for the entire environment to be moved to another data centre as well. Kubernetes has developed into a market standard. Therefore, there is no danger of vendor lock-in at any time when using the Stackable cloud data warehouse. The solution is structured in a heavily modular manner and is based on well-known and popular open source tools. The cloud platform can be changed at any time, and a move to a proprietary data centre is always possible too. The Stackable cloud data warehouse offers high transparency and is available as an open source solution. The company uses the original form of the tools used, whose source code can be viewed at any time, making it transparent.

This gives companies and organisations the ability to build a huge cloud data warehouse without having to provide in-house knowledge for building a cluster. Currently there are very few experts who are able to build such a comprehensive ecosystem, and this is unlikely to change in the foreseeable future. This is where companies and organisations stand to benefit from the know-how of the experts at Stackable. Stackable bundles its expertise with that of the IONOS Cloud and b.intelligent, a proven specialist in big data consulting. However, anyone in a company who has adequate knowledge can connect their proprietary services to the Stackable cloud data warehouse at any time using software containers. Lastly, the system is fully transparent thanks to the open source components.

Since the system uses Kubernetes, additional containers and tools can be added at any time, and hybrid clouds and multi-cloud environments can also be used as well. Whoever uses common cloud environments can also implement authentications based on Azure Active Directory and Active Directory, and use numerous other systems. Here, too, the system is based on open source, which makes it possible to expand quickly. The experts at Stackable also provide assistance with the implementation, and Stackable also ensures that commercial forking does not cause any licensing problems when using open source software components.

5 Summary

In a managed cloud, all these challenges are simply the only way to comply with the strategies of comprehensive data analyses in the coming years. Companies, organisations and government agencies of all sizes can benefit from this since they do not have to provide their own hardware and software and do not need to operate any clusters. Booking the right cloud infrastructure from European data centres also preserves data privacy, data security and data sovereignty.

In addition, managed services ensure that the services and thus all the data is stored and processed with the utmost security. In addition, setup is far easier than when operating a standalone cluster in a local data centre. And lastly, such solutions are easier to use, simpler to manage, can be scaled far better, and are also far more affordable than using software in a proprietary data centre.

Furthermore, managed services make it easier for users to get started – particularly with newer technologies such as big data and AI – since they require far less IT admin. If truly transparent open source software modules are bundled in a distribution and provided in a transparent way, the user benefits far more from the agility and security advantages of community-driven software.

IONOS

About IONOS

With more than eight million customer contracts, IONOS is a leading European provider of cloud infrastructure, cloud services and hosting services. IONOS offers everything from domains, classic websites and do-it-yourself solutions, to online marketing tools, and even full-fledged servers and IaaS. This offering is tailored to freelancers, business owners and consumers as well as corporate customers with complex IT requirements.

The IONOS Cloud is the European cloud alternative. The product portfolio comprises the Compute Engine – an IaaS compute engine with a proprietary code stack for virtualisation, managed Kubernetes for container applications, a private cloud powered by VMware and S3 object storage. Established mid-sized and large companies, regulated industries, the digital economy and the public sector can access all the services that they need for success in the cloud.

IONOS was created in 2018 from the merger of 1&1 Internet and the Berlin-based IaaS service provider ProfitBricks. IONOS is part of publicly listed United Internet AG (ISIN DE0005089031). STRATO, Arsys, Fasthosts, home.pl, InterNetX, SEDO, United Domains and World4You are part of the IONOS brand family.

Get more information at https://cloud.ionos.co.uk

IONOS

Contact

IONOS Cloud Ltd
Discovery House
154 Southgate Street
Gloucester
GL1 2EX
United Kingdom

Phone +44 333 336 2984
Email product@ionos.co.uk
Website https://cloud.ionos.co.uk

Copyright