

An Overview of ETL Cloud Services: An Empirical Study Based on User's Experience

Andreea VINES*

Bucharest University of Economic Studies, Bucharest, Romania

**Corresponding author, andreea.vines@csie.ase.ro*

Laura TANASESCU

Bucharest University of Economic Studies, Bucharest, Romania

laura.tanasescu@csie.ase.ro

Abstract. *Cloud technologies are being used more and more by companies in solution development and are able to replace traditional on-premises tools. The purpose of this paper is to provide the approach of IT employees to using different cloud technologies for developing data solutions. The objectives presented are based on two research questions – to determine the main challenges faced in data development and to see which are the most used tools and technologies by employees. The case study created to answer the questions mentioned above, consists of a survey addressed to employees from different companies that work on data development, including data engineers and data architects, and see their preferences for using services from key players in the cloud computing industry. After all responses were analyzed, it was identified that companies prefer to use cloud services instead of on-premises tools. Regarding which cloud provider they want to use, it is subjective as all cloud providers offer different advantages or disadvantages for the data integration solutions they are offering. However, even with all the cloud options and resources available, there are still challenges for big data development around the infrastructure available, data security and governance.*

Keywords: Cloud Technologies, Data Solutions, Big Data challenges, Employees Survey, Developing Opportunities.

Introduction

In an ever-changing business world, organizations tend to produce and store more and more data. Along with this increased volume of data, there is also a need for better analytical solutions, capable of analyzing the volume and variety of data, which contains structured and unstructured information. All these analytical solutions are important within the organization, as they help the company to make different business decisions. Some examples of business decisions might be understanding the needs of customers, future demand for products, or designing marketing strategies. (Assunção et al., 2015)

As the volume of data increases considerably, we are moving towards a big data area where traditional data storage solutions on a single machine (also named on-premises) are no longer sufficient, hence it appears a correlation between cloud computing and big data to overcome these shortcomings. This integration brings different benefits such as elasticity, cost-effectiveness, resource pooling, and on-demand self-service. Of course, these advantages also come with a series of challenges, the most well-known being data security and various cyber-attacks that can affect cloud infrastructure. Despite these challenges, big data and cloud computing will shape the business world, to allow organizations to make knowledge-driven decisions and maximize their goals. (Balachandran & Prasad, 2017)

Big data is being used in several domains such as business, engineering, scientific computing and health care. It overcame the weakness of data warehouse systems as allows integrating of multiple sources of data. (Chen & Zhang, 2014) Several studies were developed to cover this area; Philip et al. presented a survey on big data problems, including opportunities and challenges, current techniques, and technologies, observing that around 50% of enterprises think that big data will help increase operational efficiency. (Kune et al., 2015). Buyya et al. (2009) also presented a survey on big data computing using cloud resources, concluding that this complex process demands people with expertise in cleaning and understanding data, and selecting proper methods to analyze the results, as all this wide range of resources and possibilities might overwhelm non-experienced users. (Assunção et al., 2015)

The focus of the current paper is to perform a comprehensive analysis of the cloud technologies that are being used in the development of data solutions nowadays. Through this paper, it can be observed the trend regarding the choice of the appropriate cloud provider for solution development – either for data storage or for data processing or transformation. Hence, this study will provide an analysis of the challenges encountered in the process of developing big data solutions, considering their main characteristics. Thus, it will try to provide an answer to the following question: what are the main challenges faced by employees in the development of data solutions and how would they be rated?

Moreover, cloud providers that offer storage or processing, or transforming solutions will also be analyzed and ranked in a comparison based on different criteria such as price, scalability, documentation accessibility, and many others that will be presented after. Therefore, the second question of the study that we will try to answer is: what are the tools used for data processing, data storing, and data ingesting and how would they be rated?

To fulfill the purpose of the work, a questionnaire was created and sent to a wide number of employees who work in different IT positions (data engineer, data architect, data scientist, etc.) in developing data solutions. Using this questionnaire, their opinion regarding today's trends in this domain will be observed and it can be analyzed after. The survey contains multiple types of questions, starting with some control questions to identify university studies, experience several data projects they were involved in. It also contains questions that will be used to answer the purpose of the work, these being both multiple-choice and grid questions with several options. The form also contains some open questions added to see if there are also some other providers that are being used, but not mentioned in the survey or to see employees' experience around data development.

Therefore, the present work is structured as follows: introduction, literature review that focuses on the characteristics of cloud computing and big data, a methodology that contains the research questions and the dataset description, results obtained after the questionnaire's answers are interpreted and analyzed, and conclusions.

Literature review

Big Data and Cloud Computing

Business Intelligence solutions aim to analyze data that come from different sources to provide different insights that will help the business to make decisions. As the volume increases, it becomes harder for traditional on-premises tools to respond to all the requirements.

Big Data is defined by some characteristics, also defined as the “4 Vs”. Its main characteristic is the volume, due to the largeness of the data; the data size of a big data ecosystem can range from dozens of terabytes up to a few zettabytes and it is still massively growing.

According to a study by International Data Corporation, it is expected that the volume will exceed 175 zettabytes by 2025. (Zhou et al., 2020)

Another characteristic is the variety which indicates the heterogeneity of data that is being analyzed, including different formats such as unstructured, semi-structured, and structured data. The third “V” is considered velocity which determines the speed of processing data, including different methods such as batch or real-time processing. Veracity is also a challenge to ensure data accuracy and consistency to make sure that the values that are provided are correct. While these 4 Vs have been constantly used to define big data ecosystems, more Vs have been introduced, such as vulnerability, volatility, validity, variability, visualization, and not the last value (Anuradha, 2015; Avram, 2014).

All these challenges of big data were divided into three main categories. The first category contains data challenges that are related to the characteristics of the data itself referring to the 10 Vs mentioned previously. Another category should be for process challenges that are related to the technologies, tools, and techniques required to process, transform, and store data and to model it by choosing the right model for the analysis to provide the expected results. The last category is also important, containing management challenges that refer to organizational aspects such as data governance, privacy, and security (Sivarajah et al., 2017).

Cloud computing has several capabilities to support the challenges mentioned above, offering a diverse suite of resources capable of processing, and storing data, hence on-premises tools are being replaced with cloud resources in the product development of data solutions.

One of the definitions of cloud computing provided by Marks and Lozano is that is on-demand access to IT resources that are outside the organization, where the resources are shared with others, have a subscription payment method (usually called pay-as-you-go) and the resources are accessed through the Internet (Marks & Lozano, 2010).

Another definition of Kim presents cloud computing as a convergence of two major trends that can be observed nowadays – IT efficiency determined by how the power of modern computers is utilized more efficiently through scalable resources and by business agility that transforms IT into a competitive tool designed for rapid deployment, parallel batch processing and near real-time analysis (Bahrami & Singhal, 2015).

Data processing tools from on-premises to cloud

There are several on-premises tools for ETL (extract-transform-load) that are created to transform and store data. They were being used to simplify the data management process by reducing the development effort. They are designed to save time and money as most of them are code-free meaning that there is no need for developers to write code, since most of the transformations can happen using drag-and-drop activities (Barateiro, J., & Galhardas, 2005). Therefore, they have different capabilities and can be integrated with multiple connections to extract data from heterogeneous data sources using structured or unstructured data. The processing part involves different transformations such as filtering, cleaning, and aggregations to remove dirty data and create a model that can be further used in the analysis and to make sure that it won't affect the quality of the information or will lead to erroneous decisions (Kherdekar & Metkewar, 2016).

According to different surveys and research papers, it seems to be visible that all technology providers offer different ETL tool, but the most used per a Gartner analysis are: (Barateiro, J., & Galhardas, 2005) (Kherdekar & Metkewar, 2016) (Amanpartap & Jaiteg, 2018) (Mukherjee & Kar, 2017)

- Informatica Power Center that was launched in 1996 offers a lot of in built transformations and functionalities.

- Microsoft SQL Server Integrated Services that was launched in 1997 as a component of SQL Server database that performs data transformations being less expensive than other tools.
- Oracle Data Integrator appeared in 1999 offering different data transformation and merging high volumes of data. One important feature is the repository which represents a collection of all metadata.
- SAS Data Integration Studio was created in 1996 by SAS Enterprise company combining both data warehousing and intelligence applications designed for traditional business processing using a drag-and-drop interface.

In the research papers mentioned, they are being analyzed observing that all of them have different advantages and disadvantages. However, considering the launched dates for all of them, one disadvantage can be that all of them seem to be quite old and need to be installed using an on-premises machine – for example, Microsoft SQL Server Integrated Services needs to be launched on a Windows machine. It is true that there can be integrated with cloud computing and use, for example, a virtual machine on the cloud to avoid the dependency on a local machine, but all top cloud provider competitors offer their own resources capable of designed ETL or ELT flows directly in the cloud, without the necessity of having a local computer.

Microsoft offers Azure Data Factory which is an ETL tool for data integration, working across both on-premises and cloud, designed to move, transform, and process raw data from disparate systems with a multitude of available connectors. It offers the possibility to be integrated with different services such as Databricks and Hadoop and run scripts via it using drag-and-drop activities that can be orchestrated and triggered at specific times. (Klein, 2017) (Diaz & Freato, 2018) The pricing is based on the number of orchestration activities and the volume of data that is being processed. There are also different additional charges for premium features.

AWS Glue is a fully managed cloud ETL tool designed by Amazon Web Services that is used to read and load data from different sources having heterogeneous formats by designed pipelines with automatic executions via triggers. (Stefanov, 2019) The charging is based on the number of ETL jobs and development endpoints that are being run and the amount of data loaded. It allows using lambda functions to build a serverless full-fledged ETL pipeline. (Sreemathy et al., 2021)

GCP Dataflow is a fully managed cloud-native data processing service for batch and streaming data, allowing developers to build data processing pipelines to ingest, transform and analyze data in batches or real-time. It is designed to be highly scalable and easy to use, requiring no infrastructure setup or maintenance. Data flow can be used to run serverless ETL pipelines and enforce different data privacy rules. (Sreemathy et al., 2021)

Data processing tools play a crucial role in the context of big data, helping to manage, analyze and extract value from large and complex data sets. As part of analyzing the correlation between big data and data processing tools, the current paper focuses trying to respond to the following research questions:

Research question 1: *What are the main challenges faced in data development and how would you rate them?* This question will provide more insights into the challenges that might appear during the life-cycle development of data solutions, using answers from employees that have experience in working on different projects designing data products. This was a question of a survey where all the activities were presented and the respondents rated them from one to five, considering that one means that it is the easiest activity and five for the most complex one.

Research question 2: *What are the tools used for data processing, data storing and data ingesting and how are they being rated?* This question will provide more insights into the tools data are being used in different IT companies to develop data solutions, observing that more and

more employees tend to use cloud resources instead of traditional on-premises tools. The survey focused on the most used cloud processing tools and asked the respondents to rate them from one to five stars depending on different criteria such as pricing, user experience, documentation, number of connectors available, customer support, service availability, and many others that will be described in the Results section. There were also some open questions for the respondents to complete with some other tools that are not mentioned but are being used by them.

Methodology

The current paper uses a survey as a comprehensive research method used to collect information to produce statistics; all this information is collected by asking questions and using the answers provided as data for the analysis. All this information is collected from only a fraction of the population representing a sample that corresponds to the survey targets and meets all the criteria and eligibility to respond to the questions. (Kitchenham & Pfleeger, 2008)

The questionnaire was sent via email, messages, and LinkedIn posts to employees that work in various companies and have experience working with data projects. All the responses are anonymous. It used an expert sampling method which involved collecting responses from experts in the area, to the study to have a better view of the results and to confirm the validity of the research. (Etikan & Bala, 2017) Hence, the survey was shared directly to employees with experience in the data domain that have been working on different data integration solutions and have various jobs as data engineers, data scientists or architects. Sample volume measures the subset of the population, and the current paper aims to use a sample of 104 responses. It was also aimed to obtain a confidence level of 95%, but it's important to note that it doesn't represent a guarantee of accuracy, being only used to measure the level of confidence of the sample that reflects the population.

The survey contains 20 questions divided into two sections. The first section is dedicated to personal information, and it contains six control questions to analyze the target, composed of multiple-choice questions. They were added to gain the following information from the respondents – education level, overall work experience, work experience on data projects, number of data projects he/she was involved in, current role, and the business sector. All questions are mandatory and have the possibility to add a new response if it's not included in the predefined list.

The second section contains the questions needed to respond to the research questions mentioned previously. For most of the questions it was used a like-art scale with several options where respondents can rate from one to five. For example, the first question from this section refers to big data challenges that are being rated depending on different criteria that define big data characteristics (variety, volume, velocity, visualization) and infrastructure, governance, and compliance. The next question was added to respond to RQ1 where all the activities from the life cycle of a project were added to be rated from one to five. The next questions refer to cloud providers and their experience with them. There were added the most four ETL tools (Azure Data Factory, AWS Glue, GCP Dataflow, and Informatica Cloud Data Integration) and respondents rated them depending on different criteria such as pricing, user experience, documentation, service availability, scalability and performance, pre-build connectors, customer support, and security/compliance. Similar to the ETL tools questions, there were also added some questions to see which is their opinion of some well-known data storage tools (Azure Data Lake, AWS S3, Google File System, and Snowflake for Data Lake Analytics), considering different criteria such as pricing, service availability, security and compliance, business continuity and disaster recovery. Moreover, there were added three open questions to see which are other tools that are being used

but not mentioned. The last open question refers to the experience with data projects, where respondents had the possibility to share their experience with data projects using cloud services. The open questions were not mandatory, and the technology questions were not mandatory either as not all respondents have worked with all the technologies mentioned, so they can only rate those that they have tested before.

Results and discussions

This section reports the results of the survey based on the collected data, containing 104 answers that were collected in January 2023 through the survey described in the previous version. It was created in Google Forms and shared via email, messages and LinkedIn posts to employees that work in various data related jobs. Considering all channels where the questionnaire was shared, it used a simple random sampling methodology where the respondents were from Romania and the United Kingdom, working in various multinational or national companies across different domains. All responses were collected into a single spreadsheet for further analysis, including descriptive statistics.

Based on that number of respondents, they all work in Information Technology (IT) jobs such as data engineering (63.5% which means 66 respondents), data/cloud architecture (6.7% which means 7 persons), or data science (7.7% which means 8 persons). The rest of the respondents (22.1%) work on various jobs such as data visualization, project delivery, or data testing, as it is shown in Figure 2.

What is your role on the current project/job?

104 răspunsuri

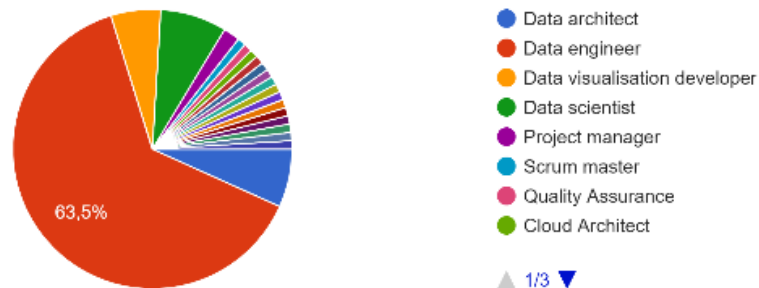


Figure 1. Distribution of responses based on the current job

Source: Authors' own research.

Considering the education level of the respondents, it can be observed in Figure 2 that most of them have a bachelor's (40.4%) or a master's degree (55.8%). There were also some respondents that did not have a bachelor or a master's degree and they gained the knowledge from IT training.

What is your education level?
104 răspunsuri

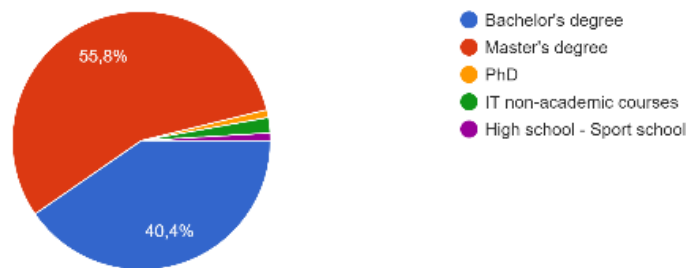


Figure 2. Distribution of responses based on the education level

Source: Authors' own research.

Regarding the level of experience, they were classified into three categories with 1 to 3 years of experience representing 22.1% from the total data, middle level with 4 to 7 years representing 42.3% and the rest of 35.6% is the third category with more than 7 years of experience. However, most of them worked with data for 4 to 7 years (representing 39.4 % of the total of 104 respondents), being involved in at least 4 data projects. In the Figure 3, it can be noticed the employees' responses similar for the questions regarding the number of years of overall experience and experience with data, meaning that they were focusing on data jobs from the beginning – for example, from the respondents with 4 to 7 years of experience, 32 of 44 of them have been worked with data for 4 to 7 years and only 12 of them (27.27%) have been only worked for maximum 3 years, meaning that previously they were focusing on different jobs.

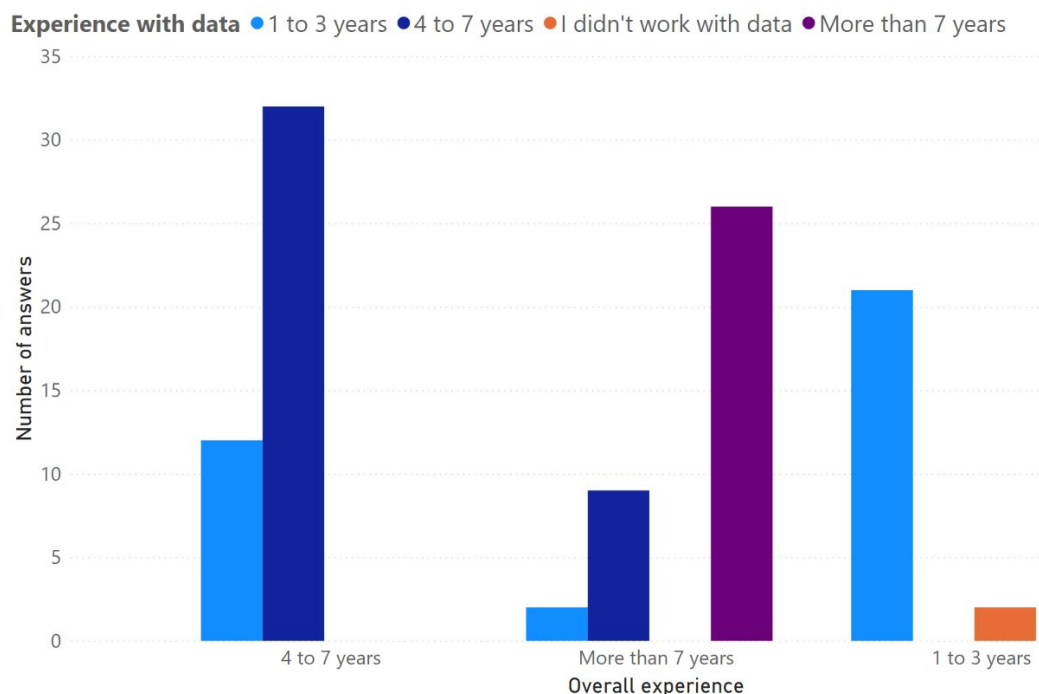


Figure 3. Distribution of responses based on the work experience and experience with data projects

Source: Authors' own research.

Another question was related to the business sector to understand the business environment that is being used the most for developing data projects. It can be observed that the respondents worked on different projects in various business sectors, but most of them were in financial services where more than half of the respondents have been worked on (62 of 104). The second most used business sector is commerce including retailing and wholesaling projects which 51.9% of the employees have worked on. The other frequently encountered sectors from the answers are health care (selected by 36.5 % of the respondents) and transport with a percentage of 22.1%. As can be observed in the figure, for the rest of the business sectors, the percentages are less than 5%, meaning that less than 6 employees have been working on data projects there.

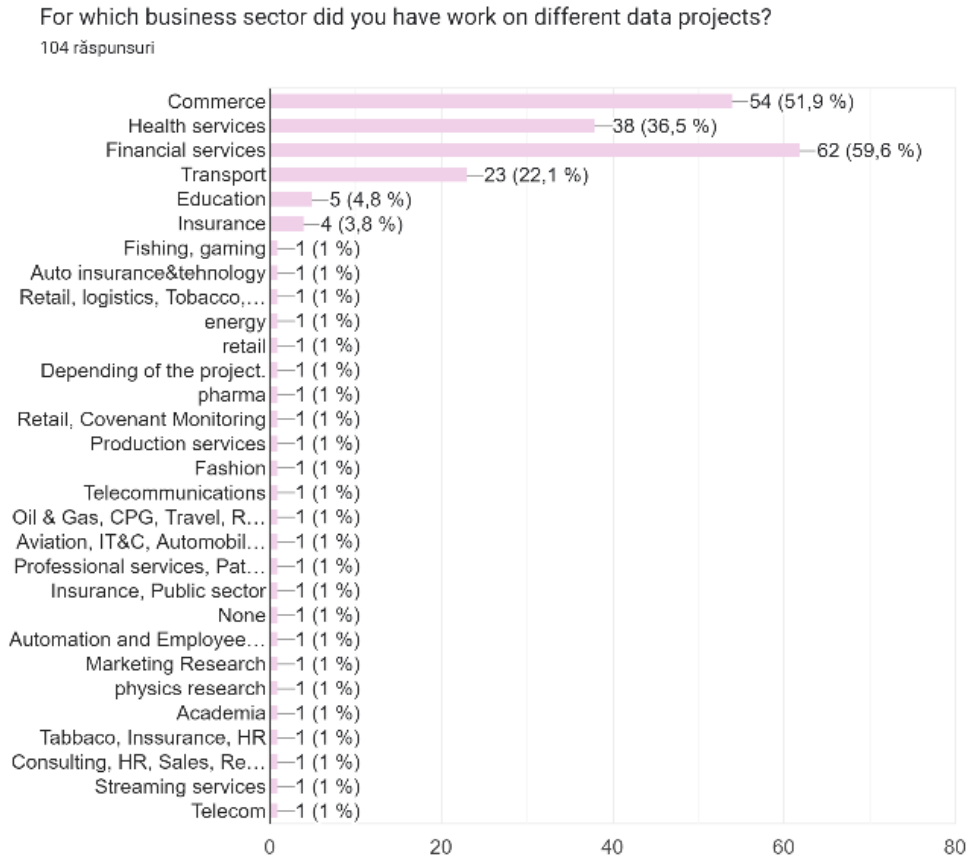


Figure 4. Distribution of responses based on the business sector

Source: Authors' own research.

One of the research questions was *What are the main challenges faced in data development and how would you rate them?* With the responses for this question, it was aimed to determine the primary challenges encountered in data development and to gauge their severity. Participants in the survey were asked to assign a numerical rating between one to five to each stage of the project development, and the results are depicted in Figure 5.

How would you define the level of complexity for the following activities that describe project life cycle ?

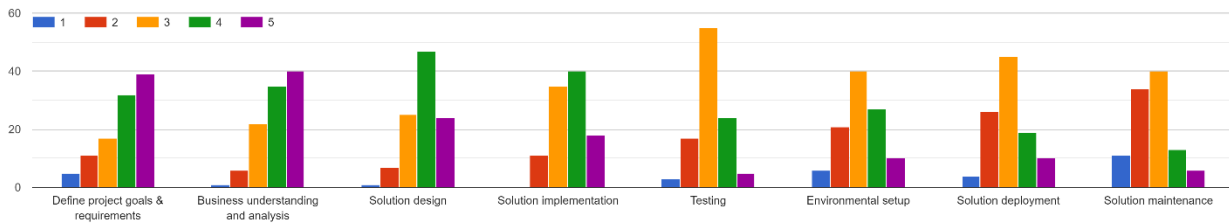


Figure 5. Distribution of responses for the project life cycle activities

Source: Authors' own research.

It can be noticed in Figure 5 that almost half of them consider that the most challenging part of developing a data solution is understanding the business requirement and the analysis and defining the project goals and requirements. Solution design is also quite challenging being scored with four points by 45% of the employees. Most of the employees that fill out the survey considered that testing is at the middle point between challenging and less challenging, being rated with three points by most of the respondents 55 of 104. It can be also observed a trend for the option with two points which is solution maintenance that is being selected by 34 of the respondents. Analyzing the responses by summing all the points assigned to each activity by all the employees, it can be confirmed that the most challenging activity is understanding the business requirements (with 419 points out of a total of 520), closely followed by defining the project objectives and requirements with 401 points as it can be observed in Table 1. Considering that for each from two to five starts, the respondents are similar for a project life cycle activity, it can be summarized that along most of the projects, the challenging part is understanding the goals and requirements and the less challenging is the solution maintenance which happens after the development. So, the projects start with challenging activities, but their complexity decreases over time. This behavior is expected as the project is clearer after the testing period and near the deployment period.

Table 1. Project life-cycle activities rated based on the survey results

Challenge	Number of points	Percentage of total
Business understanding and analysis	419	80.58%
Define project goals & requirements	401	77.12%
Solution design	398	76.54%
Solution implementation	377	72.50%
Environmental setup	326	62.69%
Testing	323	62.12%
Solution deployment	317	60.96%
Solution maintenance	281	54.04%

Source: Authors' own research.

Regarding the next questions, all employees that took part in the survey have rated some big data challenges from one to five points, observing that 30 of them consider that at this moment data infrastructure solutions and data governance are the most challenging, being rated with five points. By summing all the points assigned to each category, it can be observed in Table 2, that the most challenging is data infrastructure (having 397 points out of a possible total of 520), followed by data governance and policy with 385 points. On the other hand, most of the respondents consider that data visualization is not that challenging, having 285 points.

Table 2. Big Data challenges rated based on the survey results

Challenge	Number of points	Percentage of total
Data infrastructure	397	76.35%
Data governance and policy	385	74.04%
Data integration	381	73.27%
Data growth	366	70.38%
Data velocity	361	69.42%
Data compliance and regulation	355	68.27%
Data variety	348	66.92%
Data visualization	281	54.04%

Source: Authors' own research.

The second research question was: *What are the tools used for data processing, data storing and data ingesting and how are they being rated?* It aimed to investigate the tools utilized for data processing, storage, and ingestion and to assess their performance. The objective was to identify the leading players in the industry and the tools that are commonly used by the survey participants in the development of data solutions.

It can be clearly observed in Figure 6 that most of the respondents are working with Microsoft Azure resources representing 68.3%, followed by Amazon Web Services at 29.8% and Google Cloud Platform at 15.4%. It can be also noticed that there are also different cloud providers and from the total of 104 respondents only 7 of them are not working with cloud tools and use on-premises technologies, representing only 6.7%. From these statistics can be also flagged that most of the companies switched from on-premises to cloud resources for the project development of data solutions.

Which is the main cloud provider do you use for data projects?

104 răspunsuri

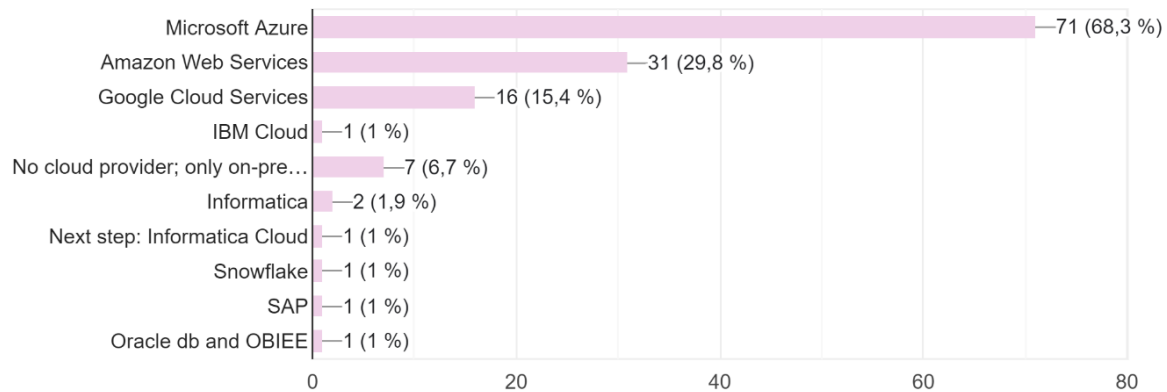


Figure 6. Distribution of responses for the cloud providers that are being used for data solutions

Source: Authors' own research.

Hence it can be observed in the Figure 6 that there are three top providers that are being used by the respondents, next questions of the survey will focus on the data integration solutions offered by the cloud providers and how they are being rated by employees taking into account the following criteria – pricing, user experience, official documentation, service availability, scalability and performance, pre-build connectors available, customer support and the latest one security and compliance. In the table below (Table 3), it can be noticed that the most used technology for data integration is Azure Data Factory, rated by 58 of the respondents. The fourth column displays the total points possible for each cloud resource, by multiplying the number of answers with the maximum of five stars that can be assigned to the criteria.

Table 3. Total number of answers and points assigned to each cloud provider

Cloud resource	Cloud Provider	Number of answers	Total points possible
Azure Data Factory	Microsoft	58	290
AWS Glue	Amazon	23	115
GCP Data Flow	Google	19	95

Source: Authors' own research.

In Table 4, it can be observed that most of the respondents consider that the service availability of Azure Data Factory is very good, offering five stars and a total of 274 which represents 94.48%. Another feature of the resource that is being appreciated by the respondents is the scalability and performance, having to compute power that performs operations quickly. However, they consider that the tool can be quite expensive, depending on the features and integrators that are being used.

On the other side, the percentages offered by 23 respondents for AWS Glue service are less compared with Azure Data Factory. The biggest percentage is also for scalability and performance

of the resource and most of the respondents consider that the customer support is not as satisfied by comparison with the one offered by Microsoft.

GCP Data Flow is being used by only 19 respondents which consider offering fewer points based on criteria. It can also be observed that the resource is more expensive than the other tools mentioned and doesn't offer the same wide range of pre-built connectors for integration.

Table 4. Comparison between data integration tools of the key cloud providers

Criteria	<i>Azure Data Factory</i>	<i>AWS Glue</i>	<i>GCP Data Flow</i>
Pricing	74.14%	70.43%	68.42%
User Experience	88.28%	76.52%	74.74%
Documentation	89.31%	75.65%	71.58%
Service Availability	94.48%	79.13%	83.16%
Scalability and Performance	94.48%	79.13%	83.16%
Pre-build Connectors and Integrations	87.93%	75.65%	74.74%
Customer Support	77.24%	66.96%	71.58%
Security and Compliance	93.10%	79.13%	83.16%

Source: Authors' own research.

It can be noted that Azure Data Factory received the highest ratings across all criteria, particularly in the areas of security and compliance, where its percentage score was significantly higher compared to the other data processing tools offered by Amazon and Google.

The last question of the survey contained an open question for the respondents to share their experience using a cloud provider. From the answers obtained, most of them consider that Microsoft Azure offers more benefits of scalability and pricing, with the consumption of resources and services being a great feature that resources can be paused whenever they are not used. They also consider that for the data integration part, GUI offered by Azure Data Factory is useful.

Although some feedback indicated that AWS offers greater computing options and flexibility. Some respondents consider that AWS Glue offers a superior user interface and a set of features for data ingestion, but its code management resources are quite robust. Another respondent shared that AWS infrastructure was preferred for the project he was involved in due to its superior pricing, performance, and an existing internal partnership.

It is important to keep in mind that all these opinions and experiences are subjective, and they may vary based on project requirements, types of resources and computing power needed, existing partnerships with different cloud providers and the familiarity with some cloud resources.

Conclusion

This research paper presented an empirical study of cloud technologies that are being used by employees in developing data solutions. Data was collected through a questionnaire and all the answers received provided clues for understanding how different cloud tools for data ingestion are

being evaluated by the respondents based on their work experience. The focus of the paper was on three key players from the market, Microsoft Azure, AWS, and GCP, by comparing all three resources that they offer for data processing or ETL/ELT, which are Azure Data Factory, AWS Glue, and GCP Data Flow. Based on the responses received, it can be mentioned that Azure Data Factory is the one that satisfies the most needs of large organizations, offering most of the advantages compared with the other two. Microsoft Azure Data Factory is used by most employees, offering better pricing options, service availability, and security and compliance. From another survey realized for on-premises ETL tools, it was also observed that the on-premises ETL solution offered by Microsoft, SQL Server Integration Services (also known as SSIS) is considered superior compared with other on-premises providers, provided by IBM or Oracle. However, from the answers received, it can be also noticed that even with all those cloud providers and options available, data infrastructure required to process all data and data governance are still some concerns on big data technology evolution.

From the answers provided in the survey, it can be concluded that on-premises tools are being replaced with cloud technologies that offer a vast area of resources that can be used for different purposes, such as data integration, data storage, compute power, or code management. On-premises vs cloud tools can also be distinguished from different perspectives such as deployment, cost control, security, and compliance, hence cloud computing offers different solutions to the most current enterprise IT problems nowadays, being part of the 4.0 Industry.

The current paper has also some limitations, including the sample size of 104 respondents, that might not offer the same results for a larger population if there is a high degree of sampling error. Another limitation can also be the bias as the sample was selected using employees from mainly multinational companies that are tented to work with cloud resources. Hence the paper aims to analyze cloud providers, we consider that the results should not be different to change the conclusions. The study was focused only on analyzing top cloud vendors and the resources offered by them. However, there were also some open questions where the respondents were able to add other different cloud providers they are using, but it was not observed some major data tools other than those presented in the current paper. One opportunity that can be considered as future work is to use a mixed-methods approach, that involves combining the survey with other data sources as qualitative interviews or observational data, that can help to mitigate these limitations and to provide a more compressive understanding of the results.

As future work, this survey can be extended to a wider group to cover a larger number of respondents, not only one hundred, to see if the conclusions will be different compared with the sample provided in the current paper. The survey can also be upgraded by adding more questions to analyze the key players in the cloud computing industry, not only through data integration resources they offer, but also to include different other resources such as data storage or compute power. It can also gain more insights from the respondents on different other topics related, such as data governance and policy, which are the most challenging as per the survey results.

References

- Amanpartap, S., & Jaiteg, S. (2018). A Comparative Review of Extraction, Transformation and Loading Tools. *Database Systems Journal*, 42-51.
- Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, 3-15.

- Avram, M. G. (2014). Advantages and challenges of adopting cloud computing from an enterprise perspective. *Procedia Technology*, 12, 529-534.
- Kitchenham, B.A., Pfleeger, S.L. (2008). Personal Opinion Surveys. *Guide to Advanced Empirical Software Engineering*, 63-92.
- Bahrami, M., & Singhal, M. (2015). The role of cloud computing architecture in big data. *Information granularity, big data, and computational intelligence*, 275-295.
- Balachandran, B. M., & Prasad, S. (2017). Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence. *Procedia Computer Science*, 112, 1112-1122.
- Barateiro, J., & Galhardas, H. (2005). A Survey of Data Quality Tools. *Datenbank Spectrum*, 15-21.
- Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, 25(6), 599-616.
- Chen, C. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 314-347.
- Diaz, F., Freato, R., Diaz, F., & Freato, R. (2018). Orchestrate Data with Azure Data Factory. *Cloud Data Design, Orchestration, and Management Using Microsoft Azure: Master and Design a Solution Leveraging the Azure Data Platform*, 263-325.
- Etikan, I., & Bala, K. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, 5(6), 00149.
- Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. *Procedia Computer Science*, 319-324.
- Klein, S. (2017). *IoT Solutions in Microsoft's Azure IoT Suite* (pp. 273-289). Berkeley, CA: Apress.
- Kune, R., Konugurthi, P. K., Agarwal, A., Chillarige, R. R., & Buyya, R. (2016). The anatomy of big data computing. *Software: Practice and Experience*, 46(1), 79-105.
- Marks, E. A., & Lozano, B. (2010). *Executive's Guide to Cloud Computing*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Mukherjee, R., & Kar, P. (2017, January). A comparative review of data warehousing ETL tools with new trends and industry insight. In *2017 IEEE 7th International Advance Computing Conference (IACC)* (pp. 943-948). IEEE
- Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 263-286.
- Sreemathy, J., Brindha, R., Nagalakshmi, M. S., Suvekha, N., Ragul, N. K., & Praveennandha, M. (2021, March). Overview of ETL tools and talend-data integration. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1650-1654). IEEE.
- Stefanov, G. (2019). Analysis of Cloud based ETL in the Era of IoT and Big Data. In *9th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education (ICAICTSEE – 2019)*. SOFIA.
- Kherdekar, V. A., & Metkewar, P. S. (2016). A technical comprehensive survey of ETL tools. *International Journal of Applied Engineering Research*, 11(4), 2557-2559.
- Zhou, C., Su, F., Pei, T., Zhang, A., Du, Y., Luo, B., . . . Zhu, Y. (2020). COVID-19: challenges to GIS with big data. *Geography and Sustainability*, 77-87.