The Benefits of Enterprise Data Warehouse Implementation in Cloud vs. On-premises

Darko Golec
IBM Slovenia, Slovenia
Ivan Strugar
Faculty of Economics & Business, Zagreb, Croatia
Drago Belak
IBM Slovenia, Slovenia

Abstract

When we think about running enterprise applications on-premises, enterprises do two things for their servers, databases, and storage. Enterprises provision for peaks and put a lot of infrastructures to handle peak demand, although a lot of this capacity is not used at normal times. The other thing is a few instances that each application needs to have, typically between five and six. Multiplying this number by many times due to various applications causes a lot of costs and creates capacity that is not used. For such reasons, the enterprise applications in the cloud seem reasonable. In the cloud, two things are possible again. Instead of overprovisioning for peaks, enterprises can scale the capacity on on-demand and spin up instances on demand. This means a certain amount of cost-saving by running at a normal level instead of overprovisioning. In this paper, various factors will be considered, and the benefits for enterprise data warehouse implementation in the cloud vs. on-premises will be stated.

Introduction

In the era where data is a valuable resource that business has, data warehouse solutions are more important than ever (Vaisman et al., 2014). By definition, a data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data supporting management's decision-making process (Chen, 2004). Data warehouse solutions form a core component of the analytical infrastructure and include a range of useful features for data management and consolidation (Davenport et al., 2013). Some data warehouse solutions have built-in functionalities like machine learning algorithms or artificial intelligence (Nemati et al., 2002).

The availability of cloud computing is essential. More than one-third (36%) of EU enterprises will use cloud computing in 2020, mostly for e-mail and file storage (Eurostat, 2021). In both Slovenia and Croatia, the percent of enterprises using cloud computing was 39 % in 2020. Cloud computing is a key objective to increase European data sovereignty, as outlined in the European Commission's Data Strategy, Digital Strategy, Industrial Strategy, and the EU recovery plan (European Commission, 2021). A strategy is intended for the investments to build a world-class digital infrastructure to securely process data in Europe and to build a trusted, open environment for storing, sharing, and re-using data and results.

With more and more organizations becoming data-driven, the need for data warehouses that could handle the high number of users and high volumes of data has become a must. A dilemma is which data warehouse model is better – cloud or on-premises. In this paper, we compare on-premises and cloud applications, discuss the benefits of choosing cloud vs. on-premises enterprise data warehouse, describe one possible approach of choosing between on-premises and cloud applications, and bring four emerging issues (big data, Covid-19, data migration, and repatriation) (Bailey et al., 2020).

Comparing On-premises and Cloud Applications

The major difference between on-premises and cloud models is that on-premises hardware, software, and applications are on-site, which means an organization manages its own data center without a third party (EM360 Tech, 2020). In the case of the cloud, it is all done off-side, and others are responsible for monitoring and maintaining a data center. Warehouse solutions can be more flexible when they get delivered through the cloud. Like other as-a-service solutions, business leaders can add and remove features to suit the changing needs of their organization (Kaur et al., 2012).

Different parameters can be compared for a cloud vs. on-premises data warehouse implementation, such as deployment, cost, security, maintenance, and flexibility (Kurunji et al., 2012). In the case of on-premises, enterprises themselves are responsible for maintaining and handling deployment processes. In some cases, organizations authorize third parties to maintain and handle the data. In the case of the cloud, resources are provided at the service provider's end and accessed by the public. Generally, the cost of operations for the on-premises model is higher than compared to the cloud model. Every hardware incurs the cost, server, storage device, and software, compared to the cloud model, where renting is used, and only used resources are billed. The cloud model provides more flexibility in terms of payment options.

Organizations are still skeptical about the security in the cloud. Healthcare organizations, government bodies, or financial institutions all have sensitive information they need to maintain (Shaikh et al., 2021; Alouffi et al., 2021). A broad set of policies is also applied in the cloud (Vaishnav et al., 2021). A dedicated cloud

helps to convince more and more organizations to move to the cloud for storing and maintaining the data where only specific organizations can access it. This makes the experience more secure. European Commission facilitated a platform to develop Codes of Conduct for data protection in the cloud, which resulted in two Codes of Conduct. A piece of information about the processing of personal and non-personal data in the cloud can be found in the Guidance on mixed datasets (European Commission, 2021).

In terms of maintenance, an organization is responsible for maintaining their servers, software, data backup, storage devices, or disaster recovery, following an on-premises model. Compared to a cloud model, no need to worry about regular maintenances. Automatic maintenance times can be set up, and then solution provider takes care of version upgrade on their own, and they do not charge apart from hardware resources being used (Ying et al., 2009). Besides, the cloud model is more scalable and flexible in comparison to the on-premises model. Changing the infrastructure or upscaling the server takes more time than the cloud, where scaling can be flexible as per requirements (Agrawal et al., 2011).
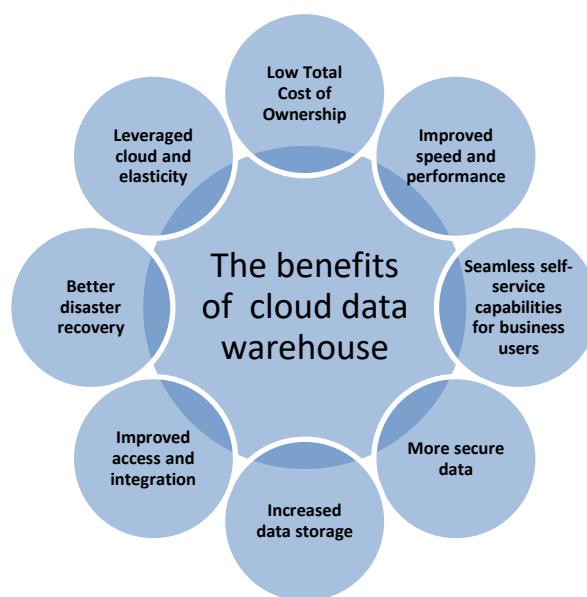
The Benefits of Cloud Data Warehouse

Nowadays, the cloud has been challenging the status quo in data warehousing, which was based on an on-premises model. Cloud data warehouse is getting more interesting for many enterprises as it allows an enterprise of any size to manage their large data without risking the purchase of a large server (Brumec et al., 2013). Distributed processing became a necessity for processing a large amount of data organizations have stockpiled (Dageville et al., 2016). Massively Parallel Processing (MPP) is one of the benefits of taking a lot of data and processing it over numerous different computation nodes. Important characteristics are distributed computation, scalability, and elasticity (SeattleDataGuy, 2020). The reasons why provisioning cloud data warehouses are speed of development, industry mandates, big data enablement, and accelerated analytics (Oracle, 2021).

Some other reasons why migrating the data warehouse to the cloud are slow-performing queries, which cannot meet business requirements, heavy maintenance activities and performance activities required on the on-premises infrastructure, maximum capacity achieved on the existing infrastructure, which cannot be scaled up/out or only at a very high cost, difficulties with a structured and unstructured data handling in a real-time to support the business operations, and overspending on the hardware or software license fees. The advantages of the cloud approach of data processing are related to the standards with high requirements for reliability and resiliency of all the network and engineering infrastructure (Shakhovska et al., 2018).

Figure 1 outlines the benefits of the cloud data warehouse, while Table 1 presents the detailed elaboration (Talend, 2021).

Figure 1
The benefits of the cloud data warehouse

Source: Authors' work according to Talend (2021)

Table 1
The Benefits of Cloud Data Warehouse

Benefit	Description
Low Total Cost of Ownership	Expensive hardware, time-wasting upgrades, maintenance, or outage management are not required.
Improved speed and performance	It is essential to use the cloud data warehouse to handle the growth of data amount. Different business solutions must be linked to support data-driven decisions. A speed and performance are important.
Seamless self-service capabilities for business users	Data marts can be spin-up seamlessly without impacting the ongoing activities. Queries can be performed on the same tables while new data is ingested.
More secure data	Cloud solutions have become very secure. They support services like multi-factor authentication, which makes communication very secure. Typically, virtual private networks are provided.
Increased data storage	Data storage can be flexible. The "Pay-as-you-go" model is provided, and the data storage subscription can be increased or decreased over time.
Improved access and integration	Structured and unstructured data are supported, and different querying languages can be used.
Better disaster recovery	A high number of nodes is used to store the data. It asynchronously duplicates the data and makes it constantly obtainable, which does not interrupt the current work.

Disaster recovery is less risky and less costly in case an issue occurs. The needs for the conventional secondary data center are smaller compared to on-premises applications. |

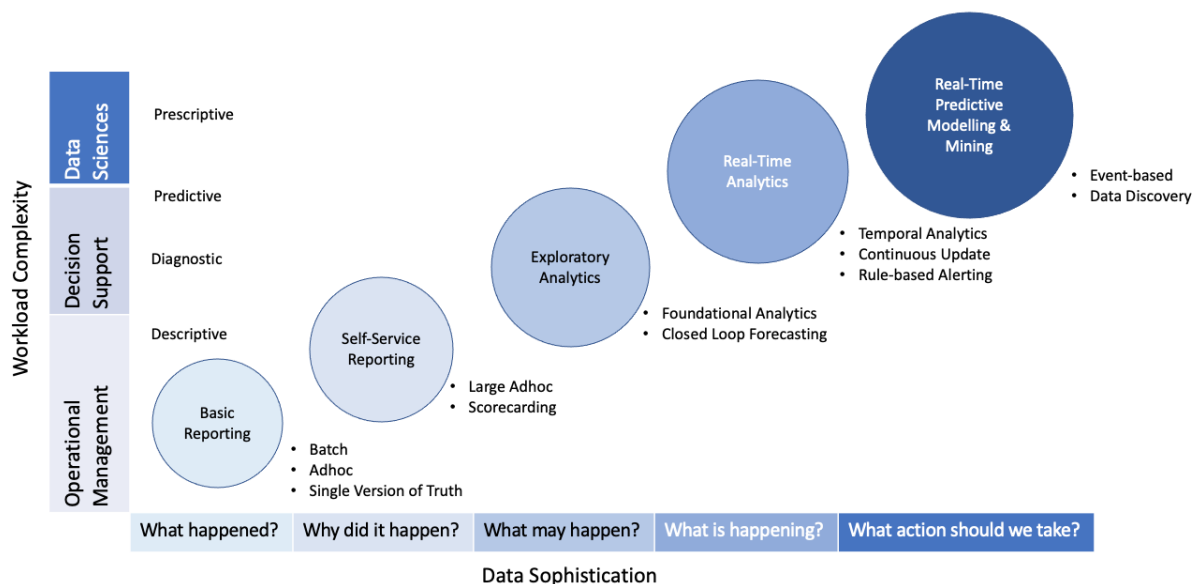|	
Leveraged cloud and elasticity	Elasticity is an important benefit. In case of high demand, more capacity can be added, and in case of low demand, storage or compute power can be shrunk to reduce the costs. Especially good candidates for the cloud are applications that might not have predictable traffic. Such applications can leverage the benefits of scaling or load balancing.

Source: Talend (2021)

Choosing between On-premises and Cloud Applications

When choosing a strategy about having a data warehouse on-premises or in the cloud, the organization needs to find out where the organization is at the analytical maturity curve (Figure 2).

The analytical maturity curve states the relationship between workload complexity and data sophistication. Handling more sophisticated data requires more complex workloads. The third bubble, called Explanatory Analytics, is the data warehouse bubble. The required sources are brought together at this stage, and benefits are gained from a 360-degree view of the organization. The third bubble is also the most complex because a data warehouse is planned in this bubble, and strategic decision on a cloud vs. on-premises is made at this stage. What is the data growth rate, what is the data quality like, how much hardware is required to implement an on-premises model or the total costs? Are some typical questions to be answered? After that, it is easier to go for the fourth or fifth bubble, which is about predictive and prescriptive analytics.

Figure 2
Analytical Maturity Curve

Source: Author's illustration

Cloud Strategy

A cloud strategy covers the transformation from on-premises to a full cloud implementation (Pragmatic Works, 2019). When choosing the strategy, it is important to map out and understand the situation and find out the best possible path. Some of the factors to look at before choosing the cloud strategy are the price of the hardware or the price of the software.

An on-premises environment comes with a dedicated cost – buying the machines, licenses, implementing the security, practices, and others. Most of it is on business to purchase, implement, and be in charge of. The first step would be getting rid of hardware while continuing with software purchasing as we move up. At that point, virtual machines can be rented through an Infrastructure as a Service provider (IaaS). IaaS works similarly to traditional computer hardware, but it operates in a virtual capacity. Platform as a Service (PaaS) refers to cloud services that provide a framework that can be quickly and easily built. This model allows focusing on the application software without managing operating systems, software updates, and other infrastructure matters. Software as a Service (SaaS) is the most prevalent cloud service and provides a whole set of software applications. SaaS applications are often accessible through a web browser, removing the need to install applications on individual workstations.

Staying on the cloud perspective, infrastructure, platform, and software can be rented, and organizations can transform to a cloud strategy without software and hardware purchasing. Operating systems, data, servers, applications, and storage are all managed by the cloud provider. Depending on the needs of the business, enterprises can pursue one or all three types of cloud services. Many large businesses retain a combination of traditional hardware and cloud computing solutions.

According to EM360 Tech (2020), the top solutions for a cloud data warehouse are Amazon Redshift, Snowflake, Google BigQuery, IBM Db2 Warehouse, Microsoft Synapse, and Oracle Autonomous Data Warehouse.

Emerging Issues related to the Cloud Data Warehouse

Cloud and Big Data

Big data expands a traditional data warehouse by offering new data types and new use cases. An emergence of traditional data warehouses, Big Data concepts like data streaming and multi-structured data, and Cloud concepts like elasticity, on-demand, and cost-effectiveness, expands the possibilities for building a modern data warehouse. In a data warehouse 4.0, those capabilities are natively provided, and a combination of infrastructure and architecture is relevant to the digital age. The core capabilities of a data warehouse 4.0 in a digital era are elastic scalability, cost-effectiveness, reliability, simplified operations, flexibility, separation of computing and storage, and insights - a rich variety of use cases (Ramakrishna, 2021). A data warehouse 4.0 can help in delivering better customer experiences and new capabilities. It must support connecting and query the data transparently across cloud, multi-cloud, and on-premises.

Cloud and Opportunities emerging from COVID-19

The COVID-19 pandemic has compelled people to work from their homes. During the lockdown, the increased number of employees working from home increased collaboration or conferencing services. It stresses back-end support services and network traffic. The cloud computing components and applications are vital for overcoming the ongoing COVID-19 situation (Singh et al., 2021).

The increased demand poses threats to the cloud. To manage the increased load or deliver an uninterrupted customer experience, robust and abundant cloud architecture is needed. Gartner (2021) stated that the "cloud computing model is inherently designed to handle fluctuating demand and, if implemented correctly, should be able to support increased requests such as those created by the COVID-19 crisis". New technologies, such as virtual reality or new partnerships with telecom providers present opportunities to ramp up the cloud offerings. By taking advantage of these opportunities, the cloud providers can better understand how much of the cloud-based work becomes the norm rather than the exception.

The organizations where digital transformation had been delayed or had not been considered have realized that moving their data analytics to the cloud, such as data warehouse, is the key to coping with and surviving the COVID-19 disruption.

Data Migration

Migrating the existing data from on-premises to the cloud environment is challenging, risky, and time-consuming. Often the code needs to be rewritten for the ETL processes, data to be reformatted to fit the target, and data to be optimized for the new platform.

Cloud Repatriation

The cloud is not the one right and only way despite the industry hype. Cloud organizations are nowadays finding some analytics workloads better on-premises, so they are pulling those workloads back to on-premises. The reversal of the cloud trend has been dubbed Cloud Repatriation (Roberts, 2021). It is explained why any smart organization would decide to do the exact opposite of what all the industry advice says is the right thing to do. The cloud issues can relate to regulatory compliance like GDPR, hidden costs of the cloud, cloud lock-in services, and a network latency due to larger and less predictable internet workload, which greatly affects how usable communication or enjoyable experience can be.

Conclusion

In this paper, the benefits of enterprise data warehouse implementation in the cloud vs. on-premises were described. A general trend in data warehousing is toward the cloud, which can provide a significant competitive advantage if done well.

Both models have some benefits, and organization strategy based on the analytical maturity curve can indicate what data warehouse model from the two available models is more appropriate for enterprise implementation. Instead of overprovisioning for peaks, enterprises can scale the capacity and spin up data warehouse instances on demand. Cloud emerging issues are also important, convincing an enterprise to a combined type of enterprise data warehouse - a hybrid blend approach.

References

1. Agrawal, D., El Abbadi, A., Das, S., Elmore, A. J. (2011), Database scalability, elasticity, and autonomy in the cloud, in International Conference on Database Systems for Advanced Applications, Springer, Berlin, Heidelberg.
2. Alouffi, B., Hasnain, M., Alharbi, A., Alosaimi, W., Alyami, H., Ayaz, M. (2021), "A Systematic Literature Review on Cloud Computing Security: Threats and Mitigation Strategies", IEEE Access, Vol. 9, pp. 57792-57807.
3. Bailey, M., Kanthan, C. (2020), "Cloud Growth, Migration, and Repatriation Continue to Gain Momentum. IDC Study", available at: https://www.idc.com/getdoc.jsp?containerId=US46119020 (1 May 2021)
4. Brumec, S., Vrček, N. (2013), "Cost effectiveness of commercial computing clouds", Information Systems, Vol. 38 No. 4, pp. 495-508.
5. Chen, X. (2004), "E-Business Data Warehouse Design and Implementation", available at: http://iceb.johogo.com/proceedings/2004/Paper/CN073-paper.pdf (1 May 2021)
6. Dageville, B., Cruanes, T., Zukowski, M., Antonov, V., Avanes, A., Bock, J., Unterbrunner, P. (2016), "The snowflake elastic data warehouse", in Proceedings of the 2016 International Conference on Management of Data, pp. 215-226.
7. Davenport, T. H., Dyché, J. (2013), "Big data in big companies", International Institute for Analytics, Vol. 3, pp. 1-31.
8. EM360 Tech (2020), "Top 10 Cloud Data Warehouse Solution Providers", available at: https://em360tech.com/data_management/tech-features-featuredtech-news/top-10-cloud-data-warehouse-solution-providers (7 April 2021)
9. European Commission (2021), "Shaping Europe's digital future", available at: https://digital-strategy.ec.europa.eu/en/policies/cloud-computing (8 July 2021)
10. Eurostat (2021), "Cloud computing - statistics on the use by enterprises", available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cloud_computing_-_statistics_on_the_use_by_enterprises (8 July 2021)
11. Gartner (2021), "When Cloud Meets COVID-19: Threats and Opportunities", available at: https://www.gartner.com/en/conferences/hub/cloud-conferences/insights/when-cloud-meets-covid-19-threats-opportunities (6 July 2021)
12. Kaur, H., Agrawal, P., Dhiman, A. (2012), "Visualizing clouds on different stages of DWH- an introduction to data warehouse as a service", in 2012 International Conference on Computing Sciences, IEEE, Phagwara, pp. 356-359.
13. Kurunji, S., Ge, T., Liu, B., Chen, C. X. (2012), "Communication cost optimization for cloud Data Warehouse queries", in 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings, IEEE, Taipei, pp. 512-519.
14. Nemati, H. R., Steiger, D. M., Iyer, L. S., Herschel, R. T. (2002), "Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing", Decision Support Systems, Vol. 33 No. 2, pp. 143-161.
15. Oracle (2021), "Data Warehouse In The Cloud, Bringing Decades of Data Management Innovations to the Cloud", available at: https://www.oracle.com/a/ocom/docs/data-warehouse-in-the-cloud-brief-3157435.pdf (1 June 2021)
16. Pragmatic Works (2019), "On Premise vs Cloud Modern Data Warehouse", available at: https://www.youtube.com/watch?v=i4t8c3K58Fo (31 May 2021)
17. Ramakrishna, R. (2021), "Data Warehouse 4.0 - Reimagining Data Warehouse for Cloud & Big Data", available at: https://canvas.lntinfotech.com/images/PolarSled/WhitePapers/Data-Warehouse-4.0_Canvas-PolarSled-WP.pdf (7 July 2021)
18. Roberts, P. (2021), "Why is Cloud Repatriation Happening? Real-Time Insights", available at: https://www.rtinsights.com/why-is-cloud-repatriation-happening (17 July 2021).
19. SeattleDataGuy (2020), "What Are the Benefits of Cloud Data Warehousing?", available at: https://medium.com/smb-lite/what-are-the-benefits-of-cloud-data-warehousing-a7322947a479 (1 June 2021)

20. Shaikh, A. H., Meshram, B. B. (2021), "Security issues in cloud computing", in Intelligent Computing and Networking, Springer, Singapore, pp. 63-77.
21. Shakhovska, N., Boyko, N., Pukach, P. (2018), "The Information Model of Cloud Data Warehouses", in Advances in Intelligent Systems and Computing III, pp. 182-191.
22. Singh, R. P., Haleem, A., Javaid, M., Kataria, R., Singhal, S. (2021), "Cloud Computing in Solving Problems of COVID-19 Pandemic", Journal of Industrial Integration and Management, Vol. 6 No. 2, pp. 209-219.
23. Talend (2021), "8 Benefits of a Cloud Data Warehouse", available at: https://www.talend.com/resources/cloud-data-warehouse-benefits (10 May 2021)
24. Vaishnav, J., Prasad, N. H. (2021), "Security Aspects in Cloud Tools and Its Analysis - A Study", in Inventive Systems and Control, Springer, Singapore, pp. 927-937.
25. Vaisman, A., Zimányi, E. (2014), "Data warehouse systems", Springer.
26. Ying, Z., Yong, S. (2009), "Cloud storage management technology", in 2009 Second International Conference on Information and Computing Science, IEEE Manchester, pp. 309-311.

About the authors

Darko Golec, Ph.D., works as a Project Manager at IBM Slovenia. His expertise is in the Analytics field, including Data Warehousing and Business Intelligence. Darko Golec is also a lecturer at a few different faculties in Slovenia. His main research interests are data modeling and digital management. Darko Golec published several papers in international journals. The author can be contacted at **darko.golec@gmail.com**

Ivan Strugar is a Professor at the Department of Informatics at the Faculty of Economics & Business. He graduated from the Faculty of Economics & Business – Zagreb, where he received his Ph.D. degree in Social Sciences, Humanities, and Theology in Economics in 1997. He is a lecturer at the Faculty of Economics, teaching information technology courses at undergraduate, graduate, postgraduate and doctoral levels at the various Universities in Croatia, Slovenia, Bosnia, and Herzegovina. His interests include business information systems application in accounting, banking and tourism, cloud computing, collaboration, document management systems, Web, intranet, and e-learning development and application. The author can be contacted at **istrugar@efzg.hr**

Drago Belak works as a Software Technical Architect at IBM Slovenia. He is an Enterprise Architect with over 25 years of banking industry experience and over 15 years in retail, health, pharma, energy, and other industries with a development and system engineering background. He leads digital transformation implementations of IBM solutions at several enterprise clients working with DevOps teams. The author can be contacted at **drago.belak@gmail.com**