Received: February 2023
Accepted: September 2023
DOI: 10.7862/rz.2023.mmr.14

Szymon DZIUBAK[1]

REVIEW OF CLOUD DATABASE BENEFITS AND CHALLENGES

The volume of data is increasing rapidly, that is why using cloud computing to store and process data may be inevitable. Providers offer many database services in public cloud that include many types of traditional relational and non-relational databases, as well as special purpose databases. Organizations can then migrate their data to cloud databases, however, decisions makers need to be aware of cloud benefits and challenges. Data in cloud is globally distributed, computing resources can be scaled up or down according to demand, cloud providers guarantee high level of service availability, many manual database administration tasks are automated. Data partitioning, replication and scaling ensure high performance. Certain applications of databases are cheaper than in private environments, however, sometimes using cloud database may be more expensive. Security is considered as a one of the main cloud database concerns due to storing data in external infrastructure. Due to regulations, organizations have to consider data privacy issues. Shared infrastructure offered in cloud is beneficial, however, sometimes isolated environments are better for cloud databases.

Keywords: cloud computing, databases, cloud database benefits, cloud database challenges.

1. INTRODUCTION

According to NIST definition, cloud computing is a model that enables convenient access to shared pool of IT resources (Mell, Grance, 2011). It provides measured service that is available on-demand through broad network access. Resources are available from shared pool and can be rapidly provisioned and released. From the perspective of this paper, it is important to characterize 3 cloud service models:

- Infrastructure as a Service (IaaS). It is a set of infrastructure resources that provides the most manageability and control. Simultaneously it requires the most clients work to release solution to production.
- Platform as a Service (PaaS). It provides a platform to create application, while cloud provider carries about underlying infrastructure.

[1] Szymon Dziubak, SGH Warsaw School of Economics, Poland; e-mail: sdziuba@sgh.waw.pl. ORCID: 0000-0001-5475-6899.

- Software as a Service (SaaS). In this model client receives ready solution. It requires the lowest configuration and management effort however it provides the lowest customization.

Cloud computing is gaining popularity. Amazon Web Services (AWS), Microsoft Azure and Google Cloud are the top cloud computing providers (Bajpai, 2023). In Q1 2023 AWS revenue increased by 20% year to year to 21.4B $, Intelligent Cloud of Azure by 18% Y/Y to 21.5B $ and Google Cloud by 32% to 7.3B $. This is significant increase taking into account worldwide IT spendings increase by 2.4%.

Databases are systems that are storing, maintaining and sharing data (Atzeni, De Antonellis, 1993). These are simply a collection of data that can be viewed as a related information (Halpin, Morgan, 2010). Databases can be manual, however, nowadays those are mainly automated, based on computer program. Databases are usually built and maintained using database management systems, which are IT systems. Databases were created in 1960s, while the most popular type of databases – relational databases – were invented in the beginning of 1970s. NoSQL databases are databases that are object--oriented, store semi-structured or non-structured data like XML or are databases for special purposes like stream processing (Strauch et al., 2011). NoSQL database is a great choice for applications to overcome relational database limitations. The main purpose of NoSQL databases is to store and process huge amount of data. Lakshman et al. (2009) compares performance of Cassandra (NoSQL) and MySQL (relational database). Cassandra is able to write 50GB of data 2500 times faster than MySQL. Catania et al. (2014) say that we observer data explosion due to new devices (for example sensors) and social networks. It is important to store and analyse the data that are produced there. Parallel Relational Data Warehouse may be the answer to this challenge. Parallelization and data partitioning, which are core concepts of new Relational Data Warehouses, can be achieved using cloud computing capabilities. NoSQL databases can be divided into following categories (Nayak et al., 2013):

- Key-Value, it stores key-value pairs, where key represents identifier and value is actual data,
- Column-Oriented, this type has column values in the place of rows,
- Document, it stores data as a document in various formats, mainly JSON, however, among others, XML and PDF are also applicable,
- Graph, data is stored as a graph there, with nodes, edges and relationships between them.

Cloud databases are gaining on popularity due to overcoming the obstacles of traditional databases stored on-premise (Bhatti, Rad, 2017). These are principally used in applications that require huge amount of data from different areas. Databases are usually used in cloud as a Platform as a Service model, however can be also other service model.

Objectives of this paper is to explain:

O1. What cloud databases are available in the market?

O2. What are the benefits of using cloud database?

O3. What are the cloud database challenges?

There are published papers on benefits and challenges of cloud services. This paper is focusing strictly on databases comparing their benefits and challenges taking into account up-to-date offerings of cloud providers. It makes the profile of this paper more practical.

2. DATABASES IN CLOUD

According to Yoon (2011), database deployed in virtualized cloud infrastructure is called cloud database. Cloud enables virtual model supporting 'everything-as -a-service', that is why databases can be also included in cloud products and services.

Relational databases can be shifted to cloud – the new model is called Relational Cloud (Curino et al., 2011). It enables users to move majority of responsibilities to service operator, for example scaling, backup, privacy, access control. Cloud providers can offer ready-made database service or DYI (do it yourself) service (Arora, Gupta, 2012). Ready--made are database services where cloud provider installed, configured and maintained a cloud server. DYI service means that majority of responsibilities are on cloud user. DYI service can be useful while migrating database to cloud with little or no change (lift-and--shift).

Zhao et al. (2014) considers NoSQL as a technology that may challenge the dominance of relational databases. NoSQL databases can be scaled horizontally, have more flexibility and schema-less models. These features can be achieved by migrating or building database in cloud due to cloud features like scalability.

As described in Introduction, 3 main cloud computing providers are Amazon Web Services, Microsoft Azure and Google Cloud Platform. Each of the mentioned cloud computing providers share databases services. Table 1 presents cloud providers' offerings of database products.

Each cloud provider offers relational databases. It can be fully own by provider (like Azure SQL that is Microsoft technology) or it can be based on open-source or other commercial solution (like databases for PostgreSQL or MySQL). It is important to point out on lift-and-shift approach to migrate databases. It is moving application or database to cloud without changing its architecture (Malhotra, 2022). Microsoft Azure and Google Cloud offers services to migrate databases without significant changes – SQL Server on Azure Virtual Machines offers SQL Server on Virtual Machine to be migrated from on-premise environment and Bare Metal Solution for Oracle offers moving Oracle databases to Google Cloud. Both, SQL Server and Oracle database are relational databases. Each cloud provider offers cloud data warehouse.

Different types of NoSQL databases are offered by selected cloud providers – AWS and Azure offers NoSQL databases of each type (document, column-oriented, key-value and graph) while Google Cloud offers native solutions only for document and column--oriented databases.

Serverless computing is describing the situation when code is executed without control of underlying architecture (Baldini et al., 2017). Operational issues (provisioning, scalability, monitoring etc.) are managed by cloud provider without client's control. Serverless is the term placed in between PaaS and SaaS cloud models. Cloud providers offers databases in serverless model where all operational and infrastructure concerns are managed by service provider and client's work is to database functionalities. These databases can be relational (see Amazon Aurora Serverless) including relational data warehouse (see Serverless SQL pool in Azure Synapse Analytics) or NoSQL (see Amazon DynamoDB).

Table 1. Cloud database services

Provider	Type	Name
AWS	relational database	Amazon Aurora
AWS	relational database	Amazon Aurora Serverless
AWS	NoSQL (document)	Amazon DocumentDB
AWS	NoSQL (key-value)	Amazon DynamoDB
AWS	NoSQL (in-memory caching)	Amazon ElastiCache
AWS	NoSQL (column-oriented)	Amazon Keyspaces
AWS	NoSQL (in-memory caching)	Amazon MemoryDB for Redis
AWS	NoSQL (graph)	Amazon Neptune
AWS	NoSQL (ledger database)	Amazon Quantum Ledger Database (QLDB)
AWS	relational database	Amazon RDS
AWS	relational database (data warehouse)	Amazon Redshift
AWS	NoSQL (time-series database)	Amazon Timestream
Azure	relational database	Azure SQL
Azure	NoSQL (multiple models)	Azure CosmosDB
Azure	relational database	Azure SQL Database
Azure	relational database	Azure Database for PostgreSQL
Azure	relational database	Azure Database for MySQL
Azure	relational database	Azure Database for MariaDB
Azure	relational database (lift-and-shift)	Azure SQL Managed Instance
Azure	NoSQL (in-memory caching)	Azure Cache for Redis
Azure	NoSQL (column-oriented)	Azure Managed Instance for Apache Cassandra
Azure	relational database (data warehouse)	Azure Synapse
Google Cloud	relational database	AlloyDB for PostgreSQL
Google Cloud	NoSQL (column-oriented)	Cloud Bigtable
Google Cloud	NoSQL (document)	Firestore
Google Cloud	NoSQL (in-memory caching)	Memorystore
Google Cloud	relational database	Cloud Spanner
Google Cloud	relational database	Cloud SQL
Google Cloud	relational database (lift-and-shift)	Bare Metal Solution for Oracle
Google Cloud	NoSQL (document)	Firebase Realtime Database
Google Cloud	relational database (data warehouse)	BigQuery

Source: Autor's own study based on information shared by cloud providers.

Cloud computing providers offer many types of databases in many models. That is why migrating database from other environment or building database from scratch in public cloud shall satisfy client's requirements. By numerous services, functionalities beyond these in local environment can be achieved. In next section cloud database benefits and challenges will be compared in order to cloud database usage decision support.

3. CLOUD DATABASE BENEFITS

Global distribution

Distribution refers to the situation where database or other application is located in multiple regions and can be stored in any available location (Guay Paz, 2018). Cloud computing is characterized by automation, so distribution configuration is handled by cloud providers and it is usually done automatically. Iosup (2014) claims that many cloud providers offers native distribution for serverless applications, less frequently for IaaS solutions – for lift-and-shift approach that require IaaS service model it can be valuable remark that distribution is not as automated and may require clients work. Azure CosmosDB is a service that can be distributed globally to any of available Azure region. It is replicated across all regions associated with Azure CosmosDB account and it is a notable example of global distribution of databases. Relational Cloud SQL from Google enables to create replicas in a different location from the primary database (called cross-regions replicas).

Elasticity

Elasticity or scalability is the ability to quickly add or remove computing resources according to demand. Cloud databases are highly scalable, usually automatically according to variable number of requests but it can be also configurated manually (Guay Paz, 2018). Cloud storages usually don't have limitations of maximum computing power that it uses. Amazon Redshift offers quick scaling possibilities through API or console. Scaling is possible from little to huge amount of data.

High availability

Most cloud services have Service Level Agreement that guarantee the level of service availability, level of consistency and throughput. Azure Cosmos DB have SLA equal to 99.99%, so the service should not be inaccessible more than 52.5 minutes per year. Amazon DynamoDB is offering SLA at the level of 99.999% that is the guarantee that service will not be available maximum 5.25 minute per year.

Performance

Performance refers to ability to serve all requests with low latency. Data partitioning refers to data division into partitions – parts of the data that can be accessed and managed separately (Tamer Özsu, Valduriez, 2020). Data replication means storing the consistent data in more than one node (Kemme et al., 2010). Data partitioning and data replication are ways to achieve high performance improvement while working with data (Zhao et al., 2014b). It refers also to high availability because data is stored in many locations. NoSQL databases hosted on cloud are also characterized by minimalizing response time and maximizing throughput (Sakr et al., 2014). Cloud providers are offering high performance; however, it is related with higher costs. Azure Database for MySQL offers Intelligent performance recommendations tool that works as advisor for database optimization. Azure CosmosDB offers many partitioning strategies.

Costs

Migrating data to cloud databases can be especially valuable when data needs to be analysed quickly with low costs. Cloud databases offers pay-as-you-go cost model so that databases can be provisioned and deployed quickly, used for any purpose and then deleted if no longer needed (Alade, 2017). Example of this can be lambda architecture where data

is analysed in batch layer (that can be on-premise) and streaming layer (it can be built in cloud) (Alexandre da Silva et al., 2016). Data from streaming layer can be deleted after analysing it so costs are lowered. On the other hand the project of migrating database and data solutions can induce high cost due to architecture change and IT specialists and consulting services (Baig et al., 2019). Costs depend on the type of service, configuration and usage, which is why it is difficult to evaluate what is more beneficial in terms of costs – public cloud or other environments.

Automation

Database automation refers to automation of database administrative tasks for example creating backups, improving security, provisioning and configuring. Automation is included in some cloud databases mainly in serverless model (Hilprecht et al., 2020). Amazon DynamoDB offers automated backup and restoring and securing data with encryption at rest. Google's Cloud SQL offers automated data provisioning, capacity management, while Azure SQL Database offers automated updates, provisioning and backups. These are only examples of cloud database automation that point out that using databases hosted in public cloud computing can be time-saving due to automation of manual and repetitive tasks.

4. CLOUD DATABASE CHALLENGES

Security

Security refers to protecting infrastructure, software and data from malicious use. Some security concerns were addressed by cloud computing provider (Lehner & Sattler, 2013). They secure cloud infrastructure however it is not possible to fully secure services they offer. Different virtualization techniques are not fully able to separate databases from other software stack. It is highlighted that cloud provider had to be trusted entity because data needs to be secured also from the provider – this is the biggest issue from security view. Ramachandran et al. (2017) is describing data security issues in cloud. Data breach and data loss are two main cloud threads. Data breach refers to situation when data is stolen by attackers. Costs of data breaches are growing as the size of stored data is increasing. Data can be encrypted to minimize data breach costs. Data loss refers to situation when the data is lost as a consequence of malicious attacks or accident. As highlighted by Ramachandran et al. cloud provider can accidentally erasure or loss data.

Data privacy

Personal data is at the centre of governments interests. European Union and other countries and organizations are restricting data migration between countries borders due to data privacy concerns (Lynn, Fox, 2020). Many cloud providers are abiding regional regulations; however, any client's needs to check if certain cloud provider is certified to applicable law. Moreover, some operations in cloud are prohibited, for example personal data can be moved within EU borders but it is not allowed to move it outside European Union states (according to GDPR).

Shared infrastructure

Computing resources are provided from shared pool by public cloud computing providers. Underlying architecture such as CPU or GPU can be not fully isolated among clients solutions – it is multi-tenant environment (Alani, 2017). In on-premise or private

cloud environments entire infrastructure is intended for one organization when in public cloud underlying infrastructure is shared. This may cause issues with noisy neighbour, the co-tenant that is using all or nearly all computing resources, so that client's database applications can't work properly. To avoid noisy neighbour issue, clients can use dedicated hosts if these are available. The availability depends on type of cloud service and it is usually offered as a database in virtual machines not as a serverless solution.

5. DISCUSSION

Migrating database to cloud should be an individual decision each time that depends on type of the data and the purpose of the solution. Decision makers should be aware of cloud limitations and challenges; however, they should not forget about benefits that cloud database is bringing.

This paper has 3 objectives.

O1. What cloud databases are available in the market?

Cloud providers offer many cloud databases. These databases include relational and non-relational databases for each of main cloud provider. Relational databases are traditional relational databases, data warehouses and special virtual machines to move the data in lift-and-shift approach. Non-relational databases are document, column-oriented, key-value and graph databases as well special purpose databases (in-memory caching and time-series databases). Cloud providers are offering then many database products that meet majority organization's needs.

O2. What are benefits of using cloud database?

Through cloud database organizations can store their data automatically replicated in many geographical locations. Computing resources are scaled up or down automatically according to demand to keep database available and minimize resources usage. Cloud providers ensure high availability of the service described in Service Level Agreement. Through data replication, partitioning and scalability cloud databases can achieve high performance with low costs in some applications. However, costs depend on how database will be used – traditional approach to infrastructure (on-premise, private cloud) may appear to be more economic. Most cloud databases activities are automated, so these require less manual work.

O3. What are cloud database challenges?

Cloud databases can be beneficial, however decision makers should be aware of cloud limitations and challenges. Security is considered as a one of the biggest cloud database challenges – cloud databases need to have underlying infrastructure and data inside secured. Cloud providers are spending a lot on cloud security, however, storing confidential data in private environments needs to be also considered. Many regulations regarding data privacy may restrict data migration to cloud or migration between cloud locations. Public cloud is shared infrastructure, that is an advantage in terms of resources utilization and costs, however certain applications may require isolated environments.

Further research questions may include qualitative analysis of real cloud databases implementations cases, encountered obstacles and benefits that migration redounded. Technology is changing over time, that is why preparing similar cloud database benefits and challenges comparison may be crucial in the future.

6. CONCLUSIONS

Cloud computing providers offers many database services that are available publicly for organizations. These databases are divided into SQL (relational) and NoSQL databases (Sharma, Dave, 2012).

Relational databases are created to store structured data in tabular format. Examples of structured data are financial transactions, customers' data or inventory data. Non-Relational databases are storing data in non-tabular form. These are used to store higher amount of data than relational ones. Moreover, non-relational databases can store different type of data like documents. If anyone fills up and sends a file in JSON format it can be stored in No-SQL document database without changing structure and format of this file. Choosing between relational and non-relational database depends on clients requirements.

Cloud computing is a technology that provides waste amount of computing resources to clients. It is shared by external companies (mainly AWS, Azure and Google Cloud) through internet. Services are available on-demand and are preconfigured by provider to minimize client's work.

This paper presents theoretical and practical approach to cloud-based databases. It presents benefits and challenges of this technology taking into account what cloud providers offers and what is available in the market. It can be useful for further theoretical discussion as well as for practical implications. Conclusions of this paper may help IT decision makers decide whether organization should use cloud-based databases or keep using databases in on-premise model.

REFERENCES

Alade, A.T. (2017). *Business Intelligence Tools for Informed Decision-Making: An Overview* [In:] Hosseinian-Far, A., Ramachandran, M., Sarwar, D. (Eds.), *Strategic Engineering for Cloud Computing and Big Data Analytics*. Springer International Publishing. DOI: 10.1007/978-3-319-52491-7_12.

Alani, M.M. (2017). C*loud Security: A Security Management Perspective* [In:] Hosseinian-Far, A., Ramachandran, M., Sarwar, D. (Eds.), *Strategic Engineering for Cloud Computing and Big Data Analytics*. Springer International Publishing. DOI: 10.1007/978-3-319-52491-7_10.

Alexandre da Silva, V., Julio C.S. dos, A., Edison Pignaton, de F., Thomas J.L., Claudio F.G. (2016). *Strategies for Big Data Analytics through Lambda Architectures in Volatile Environments. "IFAC-PapersOnLine"*, 49(30). DOI: 10.1016/j.ifacol.2016.11.138

Arora, I., Gupta, A. (2012). *Cloud databases: A paradigm shift in databases. "International Journal of Computer Science Issues (IJCSI)"*, 9(4).

Atzeni, P., De Antonellis, V. (1993). *Relational database theory*. Benjamin-Cummings Publishing Co., Inc.

Baig, M.I., Shuib, L., Yadegaridehkordi, E. (2019). *Big data adoption: State of the art and research challenges. "Information Processing & Management"*, 56(6). DOI: 10.1016/ j.ipm.2019.102095.

Bajpai, P. (2023). *The Top Companies in the Cloud Computing Space*. Access on the Internet: https://www.nasdaq.com/articles/the-top-companies-in-the-cloud-computing-space.

Baldini, I. et al. (2017). *Serverless Computing: Current Trends and Open Problems* [In:] Chaudhary, S., Somani, G., Buyya, R. (Eds.), *Research Advances in Cloud Computing. "Springer Singapore"*. DOI: 10.1007/978-981-10-5026-8-1.

Benkrid, S., Bellatreche, L., Cuzzocrea, A. (2014). *Designing Parallel Relational Data Warehouses: A Global, Comprehensive Approach* [In:] Catania, B. et al. (Eds.), *New Trends in Databases and Information Systems (Vol. 241). "Springer International Publishing"*. DOI: 10.1007/978-3-319-01863-8-16.

Bhatti, H.J., Rad, B.B. (2017). *Databases in Cloud Computing: A literature review. "International Journal of Information Technology and Computer Science"*, 9(4).

Curino, C., Jones, E., Popa, R. (2011). *Relational Cloud: A Database-as-a-Service for the Cloud.* MIT DSpace. Access on the internet: https://dspace.mit.edu/handle/1721.1/62241.

Guay Paz, J.R. (2018). *Introduction to Azure Cosmos DB* [In:] Guay Paz, J.R., *Microsoft Azure Cosmos DB Revealed. "Apress".* DOI: 10.1007/978-1-4842-3351-1_1.

Halpin, T., Morgan, T. (2010). *Information modeling and relational databases*. Morgan Kaufmann.

Hilprecht, B., Binnig, C., Röhm, U. (2020). *Learning a Partitioning Advisor for Cloud Databases*. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. DOI: 10.1145/3318464.3389704.

Iosup, A., Prodan, R., Epema, D. (2014). IaaS Cloud Benchmarking: Approaches, Challenges, and Experience [In:] Li, X., Qiu, J. (Eds.), *Cloud Computing for Data-Intensive Applications.* Springer New York. DOI: 10.1007/978-1-4939-1905-5_4.

Kemme, B., Jiménez-Peris, R., Patiño-Martínez, M. (2010). *Database replication. "Synthesis Lectures on Data Management"*, 5(1).

Lakshman, A., Malik, P. (2009). *Cassandra – Structured Storage System over a P2P Network*. Access on the internet: http://static.last.fm/johan/nosql-20090611/cassandra_nosql.pdf.

Lehner, W., Sattler, K.-U. (2013). *Data Cloudification* [In:] Lehner, W., Sattler, K.-U., *Web-Scale Data Management for the Cloud.* Springer New York. DOI: 10.1007/978--1-4614-6856-1_1.

Lynn, T., Fox, G. (2020). *Data Privacy and Trust in Cloud Computing: Building Trust in the Cloud Through Assurance and Accountability*. Springer International Publishing AG.

Malhotra, Y. (2022). *How You Can Implement Well-Architected 'Zero Trust' Hybrid--Cloud Computing Beyond 'Lift and Shift': Cloud-Enabled Digital Innovation at Scale with Infrastructure as Code (IaC), DevSecOps and MLops. "SSRN Electronic Journal"*. DOI: 10.2139/ssrn.4131044.

Mell, P., Grance, T. (2011). *The NIST Definition of Cloud Computing*. National Institute of Standards and Technology. Access on the internet: http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf.

Nayak, A., Poriya, A., Poojary, D. (2013). *Type of NOSQL databases and its comparison with relational databases. "International Journal of Applied Information Systems"*, 5(4).

Sakr, S., Sakr, S., Liu, A., Bouguettaya, A. (2014). D*atabase Replication of NoSQL Database-as-a-Service* [In:] Zhao, L., Liu, A., Bouguettaya, A., Zhao, L., *Cloud Data Management*. Springer International Publishing. DOI: 10.1007/978-3-319-04765-2_5.

Sharma, V., Dave, M. (2012). *Sql and nosql databases. "International Journal of Advanced Research in Computer Science and Software Engineering"*, 2(8).

Strauch, C., Sites, U.-L. S., Kriha, W. (2011). *NoSQL databases*. Lecture Notes, Stuttgart Media University, 20(24).

Tamer Özsu, M., Valduriez, P. (2020). *Correction to: Principles of Distributed Database Systems* [In:] Özsu, M.T., Valduriez, P., *Principles of Distributed Database Systems* (p. C1–C2). Springer International Publishing. DOI: 10.1007/978-3-030-26253-2_13.

Yoon, J.P. (2011). *Access Control and Trustiness for Resource Management in Cloud Database*s [In:] Fiore, S., Aloisio, G. (Eds.), *Grid and Cloud Database Management*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-20045-8_6.

Zhao, L., Sakr, S., Liu, A., Bouguettaya, A. (2014). Cloud-Hosted Data Storage Systems [In:] L. Zhao, L., Sakr, S., Liu, A., Bouguettaya, A., *Cloud Data Management*. Springer International Publishing. DOI: 10.1007/978-3-319-04765-2_3.