ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/upesh-kumar-rapolu/489

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 2 & Iss: 4

Research Article

Designing and Implementation of Azure Databricks in Cloud Environment

Upesh Kumar Rapolu*

Citation: Rapolu UK. Designing and Implementation of Azure Databricks in Cloud Environment. *J Artif Intell Mach Learn & Data Sci 2024*, 2(4), 2236-2238. DOI: doi.org/10.51219/JAIMLD/upesh-kumar-rapolu/489

Received: 03 November, 2024; **Accepted:** 28 November, 2024; **Published:** 30 November, 2024

***Corresponding author:** Upesh Kumar Rapolu, USA, E-mail: Upeshkumar.rapolu@gmail.com

ABSTRACT

The significance of Azure Databricks has been thoroughly analysed within the research paper. The specific design elements of Azure Databricks within the cloud environment have been appropriately discussed. It has also critically evaluated the different ways in which Azure Databricks is used within the domain of cloud computing. In addition, the notable merits and demerits of Azure Databricks have been identified within the research paper in the context of the cloud environment. In the final portion, a number of recommendations have been provided that can be utilised to solve the different kinds of drawbacks that can be related to Azure Databricks.

Keywords: Azure Databricks, Cloud environment, Processing time, Data managament and Cloud computing

1. Introduction

The research paper will critically shed light on the different aspects revolving around the design and implementation of Azure Databricks within the cloud environment. Azure Databricks is an efficient platform that is used for building, managing and sharing various forms of data, AI solutions and analytics. This research paper will also deeply assess the design or architecture of Azure Databricks. It will also evaluate the ways in which Azure Databricks is implemented within the cloud environment. The research paper will highlight the merits and demerits of using Azure Databricks and provide a few recommendations that can help resolve the drawbacks of the incorporation of Azure Databricks in the cloud environment.

2. Designing Azure Databricks in the Cloud Environment

The Azure Databricks software is widely used within the domain of cloud computing. The Microsoft engineers carefully design the software, so that it can perform efficiently in different use cases. This cloud analytic platform mainly uses two distinct architectural components, which are the compute plane and the control plane. The control plane is useful for managing the different kinds of workspace applications, clusters, configurations and notebooks. In the cloud environment, the control plane includes various backend services[1]. Another integral component of Azure Databricks is the compute plane, which processes data efficiently. This component within the Azure Databricks consists of two parts, which are serverless compute and the classic Azure Databricks compute. The different data analysis processes are executive in this component. Additionally, Spark clusters are run which helps to compute the available data within the cloud environment most efficiently. Therefore, the design of Azure Databricks is absolutely valuable for appropriate data management within the domain of the cloud. It needs to be mentioned that the cost estimation formula for Azure Databricks is Total Cost=(Databricks Unit Cost+VM Compute Cost)×Runtime Hours. Additionally, the data processing time estimate can be measured as Processing Time ∝ Data Volume/(Number of Executors×Executor Performance).

Figure 1: Logo of Azure Databricks.

3. Implementing Azure Databricks in the Cloud Environment

Inside a cloud environment, the Azure Databricks are appropriately implemented by the users. The cloud platform allows them to create a dynamic workspace where they are able to appropriately manage the data processing requirements as well as other kinds of machine learning tasks. The services are provided according to the user's Azure subscription. In addition, the users are able to properly control the configuration of the clusters within Azure Databricks. This allows them to manage the different resources and infrastructure of the cloud environment. The proper implementation of Azure Databricks within the domain of cloud computing is absolutely essential since it helps the users experience seamless integration with the other types of Azure services that are available to them on the basis of their subscriptions[2]. Most importantly, the efficient management of the cloud infrastructure enables the user to store information in a reliable manner.

4. Merits of Incorporating Azure Databricks in the Cloud Environment

4.1. Scalability

On the basis of ups and downs in the workload demands, Azure Databricks allows a user to handle large volumes of data without any kind of performance issue.

4.2. Unified platform

It is a singular platform for different operations like data ingestion, analysis and transformation within the cloud environment. The streamlined data management allows the user to work efficiently without any hindrances[3].

4.3. Enhanced security

Security is one of the most important concerns in the domain of cloud computing. This problem is appropriately mitigated by the robust security features that are built within Azure Databricks.

5. Demerits of Integrating Azure Databricks in the Cloud Environment

The disadvantages are stated below:

5.1. High cost

The basic demerits of Azure Databricks are high cost and large-scale workload. In addition to this, the database also depends highly on the learning curve that detects continuous training programs[4]. It is also for the database to have different types of training programs in place for employees. However, it has been noticed that the cost of these training programs is exponentially high. Therefore, despite the advantages of this application, the high cost prohibits implementation of this on multiple scales.

5.2. Limited customization

Azure Databricks does not have customisation provisions. Therefore, it is not possible for the users to determine the outcome according to the exclusive business objectives. Apart from this, it is also important to note that at times of Databricks uses large-scale data, it can lead to cloud bill charges for larger organisations[5]. In addition to this, it also does not provide flexibility to the users.

Figure 2: Data ingestion.

6. Recommendations

The recommendations to mitigate the challenges faced by corporations while implementing Azure Databricks are as follows:

- Continuous training programs need to be implemented. However, it is important to record the expenditure by using relevant cost-saving models[6]. In addition to this providing training in batches will help in reducing the one-time expenditure.
- Using Azure Databricks with some other software to help in customization will be beneficial. For example, CQRS microservices provide options to customise the outcomes. Concerning this it can be stated that software relevant for customisation will be helpful in applying the benefits of Azure Databricks in the workplace.
- Further, the user can also optimise storage by compressing and partitioning the data materials. It will help in cost-saving and beta testing. Another advantage of data optimisation is data segregation in databases[7]. This is beneficial for individuals to access data easily.

6.1. Abbreviations and acronyms

- CQRS- Command Query Responsibility Segregation

6.1.1. Units

- Azure Databricks Units (DBU) are used to process power used in Databricks. It is generally billed on an hourly parameter for the workload.
- Standard DS3 v2 is an interactive type of workload that is measured by 0.75 DBU per hour. The automated workload is measured by 0.55 DBU hourly.
- Standard DS5 v2 is measured by 1.50 DBU of each hour and it is also an interactive type of workload.

6.1.2. Equations

- Azure Databricks is Total Cost= (Databricks Unit Cost+VM Compute Cost) ×Runtime Hours
- Processing Time $\propto$ Data Volume/ (Number of Executors×Executor Performance)
- Equations used for programming languages:

For Python-

```
# Addition, subtraction, multiplication, division
a = 10
```

```
b = 5
sum_ab = a + b # 15
difference = a - b # 5
product = a * b # 50
quotient = a / b # 2.0
```

For SQL-

```
SELECT 10 + 5 AS sum,
 10 - 5 AS difference,
 10 * 5 AS product,
 10 / 5 AS quotient;
```

For Scala-

```
val a = 10
val b = 5
val sum = a + b // 15
val difference = a - b // 5
val product = a * b // 50
val quotient = a / b // 2
```

For R-

```
a <- 10
b <- 5
sum <- a + b # 15
difference <- a - b # 5
product <- a * b # 50
quotient <- a / b # 2
```

7. Conclusion

Azure Databricks is basically a cloud computing service that helps in providing data management and AI solutions. It helps in optimising data with the help of Microsoft. In addition to this which can also be stated that developing the application has provided the contemporary business scenario in creating better databases and accessibility methods.

8. References

1. Shi J, Jin L and Li J. “The Integration of Azure Sphere and Azure Cloud Services for Internet of Things,” applied sciences, 2019;9.
2. Tan J, et al. “Choosing a cloud DBMS,” Proceedings of the VLDB Endowment, 2019;12: 2170-2182.
3. Woo J and Mishra M. “Predicting the ratings of Amazon products using Big Data,” WIREs Data Mining and Knowledge Discovery, 2020;11.
4. Mukhdoomi MA, Oberoi A and Gupta A. “Cloud and Big Data Electronic Age: A Review,” International Journal of Computer Applications, 2020;175.
5. Borra P. “Exploring Microsoft Azure’s Cloud Computing: A ,” SSRN Electronic Journal, 2024;2.
6. Manchana R. “Building a Modern Data Foundation in the Cloud: Data Lakes and Data Lakehouses as Key Enablers,” Journal of Artificial Intelligence, Machine Learning and Data Science, 2023;1.
7. Singu SK. “Designing Scalable Data Engineering Pipelines Using Azure and Databricks,” Journal of Engineering and Technology Advancements, 2021;1.