International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 10 | Oct - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Automation For Databricks and Metastore Catalog Creations in Cloud Environment

Upesh Kumar Rapolu
Deepthi Rapolu
Naresh Kumar Rapolu
Houston
Upeshkumar.rapolu@gmail.com

***Abstract*- In the domain of cloud environment, this specific paper has thoroughly examined the automatic and metadata catalogue creation. It has highlighted the way manual catalogue management creates risk of inefficiency, non-compliance and errors while automation improves scalability, security and governance. In addition, the integration of Databricks across AWS, Azure and GCP has been elaborated here for evaluating the necessity of building metastore catalogues for a better role-based access and metadata programming. The implementation strategy has been discussed in the study, such as Infrastructure as Code, APIs, CI/CD pipelines, and terraform. The multifaceted advantages and drawbacks of the automation process have also been articulated here.**

I. INTRODUCTION

In recent cloud data platforms, automation has emerged as a major phenomenon. With organisations dealing with huge amounts of data every day, manual processes are slow, prone to errors and difficult to scale. In terms of reducing the risk parameters and building a reliable database, automation has helped to establish a more efficient system. Databricks has become a central platform in cloud ecosystems such as GCP, AWS and Azure. It provides a unified environment for data engineering, machine learning and data science. In this ecosystem, Databricks integrates with compute, storage and governance tools that enable enterprises to handle data at scale effectively and efficiently. There also exists various issues regarding manually handling the metastore catalogues. Manual work often leads to inconsistencies, delays and compliance risks. In this context, the major aim of this paper revolves around devising automation strategies in the cloud environment with the purpose of generating a better functional catalogue for a better governance.

II. IMPLEMENTATION STRATEGIES

Defining target environment and resource requirements: The first step is to identify the target cloud environment, such as AWS, GCP and Azure and list all resources required for automation. In the initial step, identification of target cloud environment becomes very essential along with listing the required resources. This includes storage accounts, Databricks, cluster identity management system and networking configuration. The mentioned resources needs to be ensured at the beginning stages so that a level of consistency can be maintenance proud the implementation process.

Automated catalogue provisioning workflow: Different tools such as Terraform, AWS CloudFormation and Azure ARM templates can be used after establishing the environment with the purpose of catalogue provisioning, utilising IaC. These roles allow administrators to script the setup of metastore catalogues, permissions, and schemas[1]. The process ensures repeatability and avoids human errors. In addition, Databricks tools including CLI and APIs can be leveraged for the dynamic creation of catalogues and updating them during the introduction of new data sets and projects.

Security, governance, and role-based access control: Automation must also cover governance and security, where Role-based access control (RBAC) can be implemented to define which users or teams can view, manage and edit specific data assets. Security and governance also need to be ensured during the implementation of automation process with the incorporation of RBAC for defining the teams and users with particular assets of data with the power of managing and editing them. Inclusion of GDPR and HIPAA also ensures to protect the sensitive data without the requirement of checking the database manually.

Monitoring, logging, and error handling: With the purpose of tracking the access level and any sort of

failure in creating automated catalogue, monitoring and login are found to be of paramount importance. A better visibility into the system behaviour can be achieved through logs and on the other hand, alerts can ensure of proper configuration for detecting security breaches. Error-handling strategies like rollbacks and retirements help maintain system stability. For example, using Terraform with Databricks providers, administrators can define catalogue schemes and access policies in code[2]. Therefore, when executed, Terraform automatically provisions these resources in the target cloud environment. This approach not only speeds up deployment but also ensures version control across projects and auditability.

Figure 1: Implementation strategies

III. OVERVIEW OF DATABRICKS AND METASTORE CATALOG

Role of Metastore catalog and databricks

Databricks AI platform and cloud native analytics are designed in terms of processing the large volume of data while enabling collaborative data science and machine learning. On Azure, with the integration of Databricks with Azure Data Lake Storage, Active Directory and Synapse, it is becoming easier in terms of focusing on the enterprises for running the compliant and securing the analytics workflows. On AWS, this particular understanding is leveraging the S3 for scalable storage. Moreover, AWS Glue is utilising for the metadata management and for fine grained access of IAM while focusing on the control and making it one of the most important and effective platforms for the AI workloads and big data. In GCP, the Databricks commonly pair with the cloud storage, big query and the roles of iam in terms of simplifying the large-scale data engineering and advanced analytics.

Significance of Metastore for Governance/Scheme

The catalogue of metastore is responsible for serving as the foundation for governing and organising data assets in Databricks. It is providing a structured environment in terms of focusing on the defined schemas, managing the tables and databases while enforcing the consistency of the naming conventions. Beyond the structure, it is also focusing on the role based on access control ensuring that the data that are sensitive and only accessible for the authorised users. Metastore governance is also supporting the compliance that are important for the GDPR and HIPAA while offering the tracking of lineage and auditing trails.

IV. AUTOMATION APPROACHES IN CLOUD ENVIRONMENT

IaC: Infrastructure as Code (IaC), are allowing the teams in terms of defining and deploying the resources of infrastructure programmatically. Tools such as the Terraform are autotomating the creation of clusters, what spaces and metastore catalog. On Azure, the templates of ARM are supporting the consistency of provisions related to the resources of Databricks while focusing on the CloudFormation performances of AWS. Similarly, it is also focusing on the functions of AWS hosted deployments. Moreover, IAC is also responsible for ensuring the consistency, repeatability and the fostering of faster provision compared to that of the manual setup.

API Automation: Exposure of Databricks REST APIs and CLI tools that are allowing the administrators for scriptic and automating the creation of catalogue while focusing on the management of schemes and permission settings. These approaches are ideal for the environment dynamically within which the catalogs are demanding for the creation of demanded understandings of new projects or the data sources[3]. The API automation is also providing the granular control and our integrity directly within the workflows of enterprise while reducing the manual overhead and minimising the challenges of configuration.

CI/CD Pipelines: The automations can be further extended by the integration of catalogue creation within the pipelines of CI/CD. Utilisation of tools are also allowing the organisations and terms of defining the pipelines that can be deployed with the resources of Databricks while focusing on the transformative reinforcing of governance with the help of automated validation and testing steps. Together the approaches are also delivering a complete framework of automation for the deployment of Databricks.

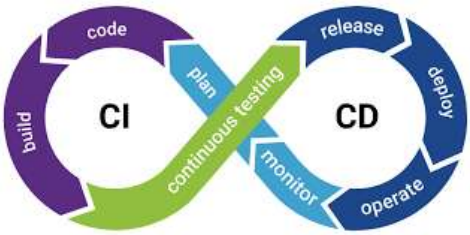

Figure 2: CI/CD Pipelines

V. KEY BENEFITS AND CHALLENGES

Benefits

Automation of Databricks and metastore catalogues offers several important benefits. It ensures scalability as catalogues can be proposed for multiple projects and teams without manual work. Automation also brings consistency and speed, which helps to reduce downtime and human errors[4]. Therefore, by embedding governance rules in a script, organisations must comply with industry standards. Hence, it can be stated that automation supports faster innovation while keeping data secure, which can help the organisations to keep consistency in their data and with high security.

Challenges

There are some challenges which come along with automation, such as integrating across multiple cloud platforms, which can be complex. This is because each provider has its own tools and policies. There are also security risks if automation scripts are not properly tested, since misconfiguration can expose sensitive data[5]. Version management of automation scripts requires careful planning to avoid conflicts. Lastly, cost optimisation is a challenge as automatic provisions can lead to unused and over-provisioned resources if not monitored.

VI. CONCLUSION

As conclusion, across all the platforms, it can be stated that Databricks is responsible for providing automation, scalability and unified management for the pipelines of data which is important for modern enterprises. Furthermore, without a central catalogue, the management of the metadata across the multiple cloud deployments are becoming fragmented which is also leading towards the challenges of non-compliance and inconsistencies. Automatic catalogue and also creating identification of the challenges by ensuring the security of data, discovery understanding and governing at the scales.

Abbreviations and Acronyms

- **DBX** – Databricks
- **IAM** – Identity and Access Management
- **IaC** – Infrastructure as Code
- **CI/CD** – Continuous Integration / Continuous Deployment
- **API** – Application Programming Interface
- **RBAC** – Role-Based Access Control
- **ADLS** – Azure Data Lake Storage
- **S3** – Simple Storage Service (AWS)
- **GCP** – Google Cloud Platform
- **VM** – Virtual Machine

Units

- Time measured in minutes/seconds,
- Data size represented in TB (terabytes) / GB (gigabytes).
- Compute usage measured in vCPUs and RAM (GB).
- Network throughput in Mbps/Gbps.
- Storage latency in milliseconds (ms).

Equations

Provisioning Time Reduction = (ManualTime–AutomatedTime)/ManualTime × 100%

ACKNOWLEDGEMENT

I sincerely thank my supervisors, peers, and contributors for their continuous guidance, valuable insights, and support throughout this work.

REFERENCES

[1] F. M. S. Morton, G. S. Crawford, J. Crémer, D. Dinielli, A. Fletcher, P. Heidhues, and M. Schnitzer, "Equitable interoperability: the 'supertool' of digital platform governance," *Yale Journal on Regulation*, vol. 40, p. 1013, 2023. https://openyls.law.yale.edu/bitstream/handle/20.500.13051/18329/Equitable%20Interoperability%20The%20Supertool%20of%20Digital%20Platform%20Governance.pdf?sequence=1&isAllowed=y

[2] L. Madabathula, "Autonomous Data Ecosystem: Self-Healing Architecture with Azure Event Hub and Databricks," *Journal of Computer Science and*

Technology Studies, vol. 7, no. 8, pp. 866–873, 2025. https://al-kindipublishers.org/index.php/jcsts/article/download/10649/9397

[3] M. Lamothe, Y. G. Guéhéneuc, and W. Shang, "A systematic review of API evolution literature," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–36, 2021. https://www.ptidej.net/publications/documents/CSUR22.doc.pdf

[4] S. Rivera, A. Prokaieva, and A. Baker, *Databricks ML in Action*, 2024. https://sciendo.com/2/v2/download/chapter/9781800564008/10.0000/9781800564008-001.pdf?Token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJ1c2VycyI6W3sic3ViIjoyNTY3ODUxNywicHVicmVmIjoiNzY0NDg4IiwibmFtZSI6Ikdvb2dsZSBHb29nbGVib3QgLSBXZWIgQ3Jhd2xlciBTRU8iLCJ0eXBlIjoiaW5zdGl0dXRpb24iLCJsb2dvdXRfbGluayI6Imh0dHBzOi8vY29ubmVjdC5saWJseW54LmNvbS9sb2dvdXQvNjdmOWM5MjBkMDBmOTRlNjViMzcyOTU2YjRiM2ZhNzYiLCJhdXRoX21ldGhvZCI6ImlwIiwiaXAiOiI2Ni4yNDkuNjkuMTY2In1dLCJpYXQiOjE3NDQ0MjUyMDYsImV4cCI6MTc0NTYzNDgwNn0.5VWFzNLKtkuOQckEcBkKVQMRvTd5reTSpyFs4ibG_0I

[5] V. R. Gudelli, "CloudFormation and Terraform: Advancing Multi-Cloud Automation Strategies," *International Journal of Innovative Research in Management, Pharmacy and Sciences (IJIRMPS)*, vol. 11, no. 2, 2023. https://www.researchgate.net/profile/Venkata-Gudelli/publication/389590452_Cloud_Formation_and_Terraform_Advancing_Multi-Cloud_Automation_Strategies/links/67c886fbcc055043ce6defd5/Cloud-Formation-and-Terraform-Advancing-Multi-Cloud-Automation-Strategies.pdf