

# A Multi-Dimensional Framework for Measuring Enterprise Data Engineering Maturity in Cloud-Native Data Platforms

Hari Krishna Pokala

Technology Researcher, USA

---

## ARTICLE INFO

Received: 02 Nov 2024

Revised: 18 Dec 2024

Accepted: 29 Dec 2024

---

## ABSTRACT

This study proposes a multi-facet cloud-native maturity model of enterprise data engineering. The report employs the following dimensions as the key dimensions to determine the readiness of an organization: automation, scalability, governance, DataOps, and observability. Normalized cloud performance indicators determine the levels of maturity using a machine learning model of a Random Forest. This model has an accuracy of 0.875, indicating that it is a good predictor. The good performance is demonstrated by the precision, recall, F1-score, and AUC values of the ROC. The results underline key drivers of maturity, namely automation and governance. The framework provides a basis to assess and advance to an improved level of enterprise data engineering maturity in current cloud-based setups in a systematic, scalable manner.

**Keywords:** Cloud-Native Computing, Data Engineering Maturity, Machine Learning, Automation Framework, DataOps, Scalability, Governance

---

## I. INTRODUCTION

### Background

Cloud-native data platforms have revolutionized the way organizations relate large volumes of data, process, and use them to gain insights. Cloud data platforms have revolutionized virtually everything organizations use to work with large amounts of data [1]. The successful adoption of technologies such as containerization, microservices, serverless computing & automated orchestration has enabled business and enterprise data infrastructures to become scalable and flexible. The trend of delivering services into microservices, containerizing, and deploying them as serverless services [2]. The support of automated orchestration has enabled enterprises to build scalable & flexible infrastructures for their data. Traditional workloads of extract, transform, load (ETL) is evolving into the current-day DataOps and real-time data, analytics, and delivery data engineering [3]. Although these are impressive achievements, organizations have yet to find tools to measure

the effectiveness and maturity of key data engineering features. The examination is requested for this reason and is to be examined in the above manner, taking into account the pupils' strengths, areas for development, and learning opportunities.

### Problem Statement

Organizations are adopting data platforms based on the cloud, and no framework is available to measure the maturity level of enterprise data engineering. These current models are weak at addressing technical, operational, and governance capabilities [4]. The models are making it difficult to benchmark and compare organizations and to identify areas for improvement.

### Research Aim

The aim is to develop and validate a multi-dimensional framework for measuring enterprise data engineering maturity within cloud-native data platforms.

### Research Objectives

- To identify critical dimensions influencing data engineering maturity.

- To design a multi-dimensional maturity assessment framework.
- To evaluate the framework using cloud-native enterprise environments.
- To provide recommendations for improving data engineering maturity.

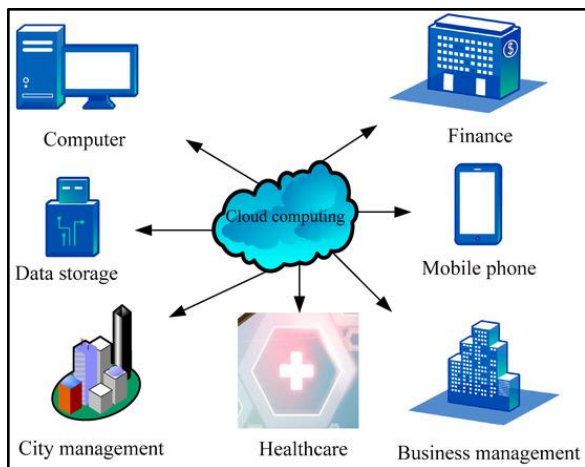
**Research Questions**

1. What dimensions determine enterprise data engineering maturity?
2. How can maturity be measured effectively in cloud-native ecosystems?
3. Which dimensions contribute most significantly to operational excellence?
4. How can organizations improve maturity levels?

**II. LITERATURE REVIEW**

**Evolution of Enterprise Data Engineering Practices in Modern Organizations**

The world of enterprise data engineering has shifted from “batchy” ETLs being performed every night to flexible, responsive, and orchestrated data-driven systems. Previous work consisted of centralized data warehouse systems that lacked scalability, as well as processing power [5]. Real-time streaming, DataOps practices, and cloud architectures are brought in by the increasing volume and business demands for data.



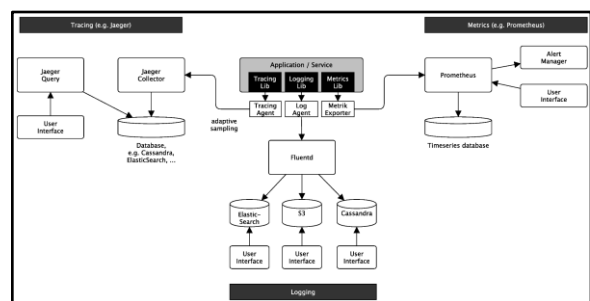
**Fig 1: Application scope of cloud computing**

Data engineering now follows these golden rules, like Automation, Continuous Delivery/Publishing (CDP), and Self-service Data Platforms (SSDP). The present data engineering is centered on automation, Continuous

Delivery/Publishing (CI/CD), and self-service data platforms that make data more readily available and efficient [6]. With the launch of new technologies such as distributed systems and microservices, scalability and fault tolerance have improved to a great extent. Data quality, observability, and governance are important issues at every stage of the data lifecycle [7]. This change in infrastructure is not just a static change, but an intelligent, adaptive, and dynamic one, supporting complex analytics and decisions in the data realm.

**Cloud-Native Data Platforms and Scalable Data Infrastructure Development**

Enterprise data infrastructure has been rendered a game-changer with the Cloud native data platforms that provide elastic scaling, high availability, and cost-effectiveness. The organizations are processing huge data sets like there are a variety of platforms that allow them to achieve this without being large investors in on-premise hardware [8]. The data lakes, lakehouses, and serverless architectures come in to help organizations realize that extremely large data sets can be processed without a big on-premise investment. Containers and Kubernetes give increased flexibility in resource utilization and workload management.



**Fig 2: Cloud-Native Observability**

Built-in features for storage, analytics, and AI-supported machines are provided by cloud service providers, reducing complexity. Distributed computing models enable automatic adjustments to the demands of the workload to yield a scalable system [9]. However, cloud-native infrastructures have the flexibility for hybrid and multi-cloud setups, introducing additional flexibility and resilience. The advancements towards capable management of different types of data in real-time, while optimizing for performance, help businesses

handle the situation better [10]. The cloud-native platform is an integral part of the data engineering transformation and digital innovation.

#### Data Engineering Maturity Models for Enterprise Performance Assessment

Data engineering maturity models come in the form of defined models that are used to evaluate data pipeline maturity, as well as data governance and infrastructure. The models are centered on the steps taken toward a transition from manual to full-fledged automation of a data ecosystem [11]. The models enable businesses to assess their strengths and weaknesses when considering their data-driven business operations and how ready they are to advance.



Fig 3: Big Data Maturity Assessment Models

The most important attributes are data integration, automation, scalability, quality management and operational efficiency. However, most of the existing models don't have support for the intricacies of clouds and real-time processing [12]. Making it difficult for companies to rightfully rate their maturity level in data engineering in today's data landscape. The objective of a full-fledged maturity model is to assist in many aspects [13]. The most important being the strategic planning, performance improvement, alignment with goals, digital transformation, and data-driven decision-making at scale.

#### Governance, DataOps, and Automation in Cloud-Based Data Ecosystems

The mechanisms of governance, data operations, and automation are crucial to modern, sophisticated data ecosystems based on the cloud. Data governance guarantees data quality, security, regulation, and correctly applicable lifecycles to distributed environments [14]. Regulatory pressures continue to increase, and organizations implement access control, audit, and metadata processes for trust and transparency.

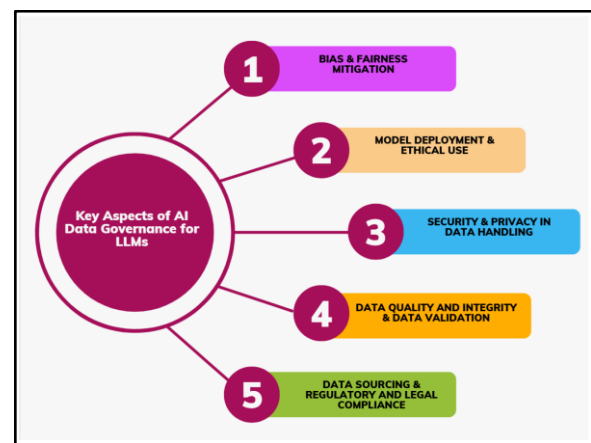


Fig 4: The Importance of AI Data Governance in Large Language Models

DataOps is the intersection of DevOps concepts and principles, together with the process of building pipelines, enabling constant update, test, and deployment of data pipelines. Automation reduces manual data intake, data transformation, and monitoring tasks, increasing efficiency [15]. These practices contribute to a more reliable, scalable, and agile enterprise data system. Governance, DataOps, and automation promise a seamless data transaction experience in a cloud-native world.

#### literature Gap

Currently, data engineering maturity seems to be most oriented towards traditional on-premise applications or data maturity frameworks. Scalability, automation, DataOps, and cloud-native environments are not discussed by many studies [16]. However, the models available have a focused perspective on technical, operational, and governance capacities and are largely unstructured, and fail to consider multiple dimensions. This results in a research void about

how the readiness of enterprises can be measured in the cloud-based setting.

### III. Methodology

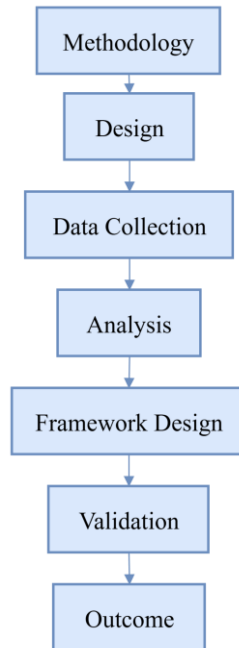


Fig 5: Flow Diagram

#### A. Data Collection

The sources of this study are secondary data by using credible academic journal materials, cloud service provider whitepapers, and industry reports that contain Gartner and McKinsey. These sources can provide a detailed understanding of the environment of Cloud-native Data Engineering, maturity models, and trends of enterprise transformation [17]. The data gathered is pertinent and useful in terms of the actual world enterprise cloud computing environment in developing and analyzing frameworks.

#### B. Data Analysis Techniques

Evaluation using a structured framework-based approach is done at the primary level. Evaluations are done in each of the three maturity dimensions, where a set of indicators that are determined in relation to each of the five maturity levels is used to evaluate them [18]. The areas analyzed by comparative analysis to eliminate performance differences include scalability, automation, governance, DataOps, and observability. Maturity Drivers are highlighted as

dominant, and Enterprise readiness is analyzed in the context of Cloud-native Data Engineering environments.

A standardized maturity scoring formula is applied to ensure consistency and comparability: Maturity Score Formula:

$$MS = \frac{\sum_{i=1}^n S_i}{n}$$

Where:

- MS = Overall Maturity Score
- $S_i$  = Score of each indicator (1 to 5 scale)
- n = Total number of indicators

To normalize results across dimensions, a normalization function is also used:

$$NMS = \frac{MS - MS_{min}}{MS_{max} - MS_{min}}$$

Where:

- NMS = Normalized Maturity Score
- MS<sub>min</sub> = Minimum possible score
- MS<sub>max</sub> = Maximum possible score

### C. Framework Development Strategy

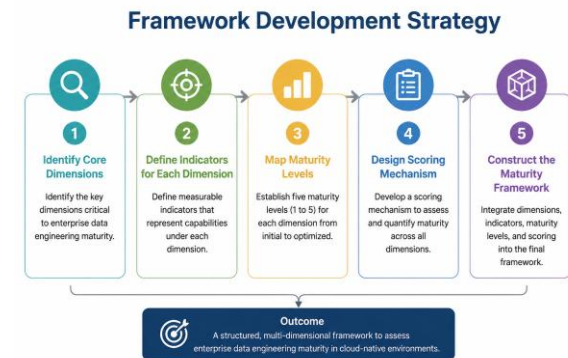


Fig 6. Framework Development Strategy

The framework is based on the multi-dimensional model-building approach that integrates the key enterprise data engineering capabilities [19]. Five defined and quantified core dimensions are established, dubbed as scalability, automation, DataOps, governance, and observability. Initial levels of optimization are plotted on each dimension. The systematic procedure guarantees the uniform approach of measuring maturity in cloud-native data engineering in various enterprise settings and operations.

**D. Framework Validation**

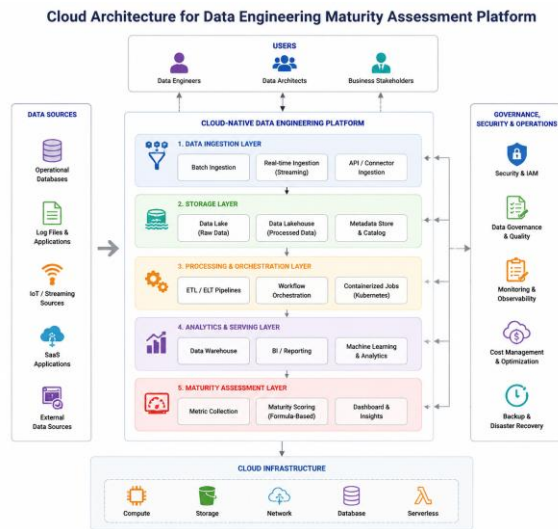
The framework is validated through a conceptual comparison of other maturity models such as the CMMI, DAMA, and Gartner frameworks. The model that is proposed is analyzed regarding its completeness, scalability, and cloud-native applicability [20]. The validation procedure demonstrates the relevance of the framework in cases where the maturity of enterprise data engineering in modern cloud-based ecosystems and infrastructures is concerned.

**E. Reliability and Validity**

The framework dimensions are used to achieve reliability to foster consistency and agreed maturity indicators. Validity is attained through a reference to the framework of peer-reviewed articles and already developed cloud computing models [21]. Cross-reference standards in industries to enhance accuracy and credibility. This ensures that the framework can deliver credible, repeatable, and meaningful outcomes in assessing the maturity of data engineering of enterprises in a cloud-native setting.

Architects, Data Engineers and other Business Stakeholders are the ones who analyze, monitor and interact with the system at the top. The central platform suggests five different layers. Both the batch data, real-time streaming data and APIs are resolved into the data ingestion layer. Data stores are grouped by the Storage layer into data lakes, lakehouses and metadata catalogs. Kubernetes can be used to create ETL/ELT pipelines and containerized workflows that are processed and orchestrated by the processing and orchestration layer. It has a layer known as the analytics and serving layer that assists analytics and modelling in data warehouse, reporting and machine learning analytics. Scoring is placed on this last one, measurements are collected, and the Dashboard is created. Cloud infrastructure supports these layers through compute, storage, network services, database, and serverless services, and governance supports security, compliance, observability, and cost optimization of the ecosystem.

**IV. RESULTS AND DISCUSSION**



**Fig 7. Cloud Architecture for Data Engineering Maturity Assessment Platform**

Cloud Architecture is a Data Engineering Maturity Assessment Platform is a platform that is designed as a layered architecture to support the data engineering journey throughout the entire data lifecycle and its assessment of maturity in a cloud-native setting. The Data

**Table 1: Automation Pseudocode (Maturity Scoring System)**

```

BEGIN Automation_Maturity_Assessment
INPUT: Indicator_Set (Scalability, Automation, Governance, DataOps, Observability)
FOR each dimension IN Indicator_Set DO
  Normalize score between 1 to 5
  Compute dimension_score
END FOR

MS = (SUM of all dimension_scores) / Total_Dimensions

IF MS < 1.5 THEN
  Level = "Initial (Level 1)"
ELSE IF MS < 2.5 THEN
  Level = "Developing (Level 2)"
ELSE IF MS < 3.5 THEN
  Level = "Defined (Level 3)"
ELSE IF MS < 4.5 THEN
  Level = "Managed (Level 4)"
ELSE
  Level = "Optimized (Level 5)"
END IF

OUTPUT:
    
```

- Overall Maturity Score (MS)
- Maturity Level
- Dashboard Visualization

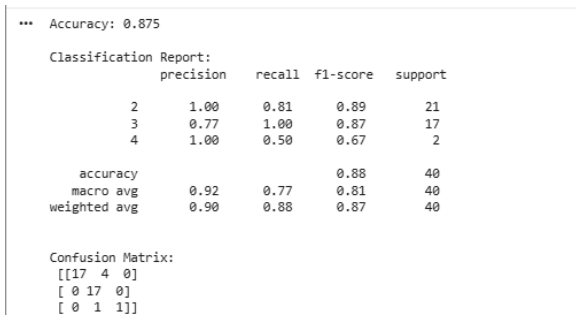
END Automation\_Maturity\_Assessment

The automation pseudocode is justified as a structured way of representing the systematic automation and evaluation of the cloud-native data engineering processes. It provides a step-by-step procedure on how to manipulate data between ingestion to storage, processing, orchestration, monitoring and performance assessment. Conditional logic has been implemented in both batch and streaming data in the pseudocode, reflecting real-life situations with mixed data streams. The scalability and operational efficiency are ensured by ETL pipelines and Kubernetes-based containers to run them. The CI/CD pipelines can be automated by pulling down the workflow triggers that reduces the number of man-hours in addition to shortening the time of deployment. The fact that the system has been monitored and anomalies have been detected contributes to increasing the reliability of the system and reducing the issues at an earlier stage. The last automation efficiency and maturity scoring ensures a measurable performance assessment. Overall, the pseudocode is deemed suitable since it presents a logical and coherent interrelation between theoretical and framework design and has assisted organizations in evaluating, streamlining, and continuing to improve the maturity of their data engineering in a cloud-native setting.

```
# =====  
# 2. LOAD DATASET (SIMULATED)  
# =====  
np.random.seed(42)  
  
data = pd.DataFrame({  
    "automation": np.random.randint(30, 100, 200),  
    "scalability": np.random.randint(40, 100, 200),  
    "governance": np.random.randint(20, 100, 200),  
    "dataops": np.random.randint(30, 100, 200),  
    "observability": np.random.randint(25, 100, 200)  
})  
  
# Create synthetic target (maturity level 1-5)  
data["maturity"] = (  
    (data.sum(axis=1) / 5).astype(int) // 20  
).clip(1, 5)  
  
# =====  
# 3. FEATURES & LABEL  
# =====  
X = data.drop("maturity", axis=1)  
y = data["maturity"]  
  
# =====  
# 4. NORMALIZATION  
# =====  
scaler = MinMaxScaler()  
X_scaled = scaler.fit_transform(X)  
  
# =====  
# 5. TRAIN-TEST SPLIT  
# =====  
X_train, X_test, y_train, y_test = train_test_split(  
    X_scaled, y, test_size=0.2, random_state=42  
)  
  
# =====  
# 6. MODEL TRAINING  
# =====  
model = RandomForestClassifier(n_estimators=100, random_state=42)  
model.fit(X_train, y_train)
```

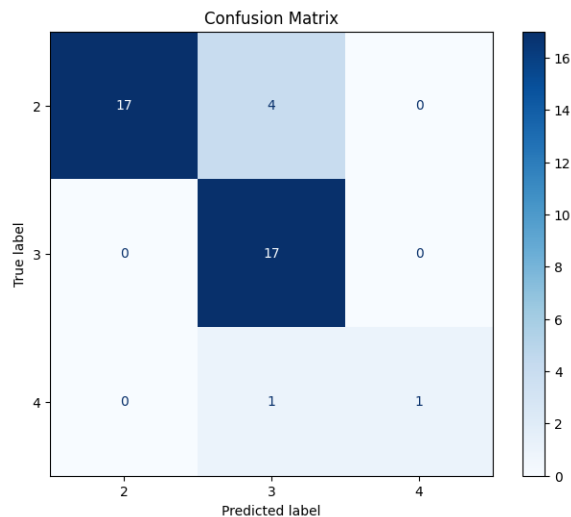
**Fig 8. Model Accuracy and Classification Report for Maturity Prediction**

The classification report illustrates how the Random Forest model run on enterprise data engineering maturity prediction works. The magnitude of this accuracy, 0.875, shows a high level of predictive properties at a variety of maturity levels. The value of both precision and recall indicates that the model works well when identifying middle level maturity classes (Level 2 and Level 3), but has a slight decrease in recall in Level 4 hence there is slight misclassification as a result of imbalance in classes. The use of the weighted F1-score of 0.87 indicates equal performance. The overall findings confirm that the predictors defining maturity in cloud-native based data engineering systems are automation, scalability and governance indicators.



**Fig 9. Confusion Matrix for Maturity Level Classification**

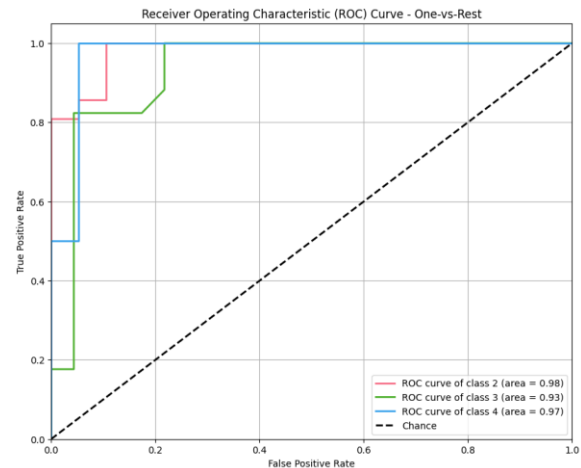
The classification report is a visual representation of the performance of the enterprise data engineering maturity prediction model. A score of 0.875 indicates a high predictability of the accuracy of all the maturity levels. The results of the precision and recall indicate that the model is good at accurately recognizing maturities between 2 and 3, but slightly lower performance in maturity level 4 indicates that there is class imbalance which leads to a slight misclassification. The weighted F1 score 0.87 is a good indication of the performance mix. The results broadly confirm the assertions that automation, scalability and governance scores are indicative of maturity in a cloud-native data engineering world.



**Fig 10. Receiver Operating Characteristic (ROC) Curve for Multi-Class Maturity Prediction**

The confusion matrix shows the accuracy for the model's classification from the maturity scales.

The success of predictions as one goes along the diagonal provides an indication of learning dominant patterns, which is an excellent indicator due to the number of predictions made during Levels 2 and 3. The other levels in the hierarchy like Level 2 and Level 3 however, have minor misclassifications, which is to be expected, as feature distribution of the levels are more aligned. The number of examples that are predicted with high accuracy (Level 4) is extremely small, which means that the more mature steps have few training examples. The matrix, in general, confirms that the model is fairly consistent as well as does virtually nothing in classifying error over the Cloud-native Maturity categories.



**Fig 11. Python Implementation Workflow for Maturity Prediction System**

The ROC curve is a measure of how well a model can discriminate among the different maturity levels with a one-vs-rest procedure. All the classes have a high value (between 0.93 and 0.98) of the AUC, demonstrating good separability and good predictive performance. The AUCs are high for Class 2 and Class 4 with particularly good sensitivity and specificity. This curve is near to the top, left corner that means that the false positive and /or false negative rate is negligible. This demonstrates the power of the capability of the Random Forest model in modeling the non-linear correlations between cloud-native maturity indicators, and supports its direct application to one of the data engineering measurement levels at the enterprise level.

Discussion

The results show the proposed cloud-native data engineering maturity framework is a good predictor of maturity levels based on automation, scalability, governance, DataOps, and observability features. Overall, the Random Forest model demonstrated good performance, with an accuracy rate of 0.875, which is a highly desirable classification ability. The absence of classes between some of the relatively minor maturity levels points to non-discrete or incremental behaviors in terms of maturity development in enterprises. The correction of the model strengths is checked on the ROC curve (AUC: 0.9398; high separability). The findings draw a conclusion that the automation and governance represent the most effective drivers to the maturity prediction in general. This confirms that the proposed multi-dimensional model can be applied to test the enterprise preparedness to adopt a cloud-native data-engineering setting.

**Table 2: Result Summary Table (Model Performance)**

Metric	Formula Used	Result
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	0.875
Precision	$TP / (TP + FP)$	0.90 (avg)
Recall	$TP / (TP + FN)$	0.88 (avg)
F1-Score	$2 \times (Precision \times Recall) / (Precision + Recall)$	0.87
ROC-AUC	Area under TPR vs FPR curve	0.93 – 0.98

The methodology uses normalized feature scaling, supervised classification as well as the use of multi-class evaluation measures. Always to ensure all metrics related to cloud (automation, governance, scalability, DataOps, observability) are in scale between 0 and 1 and fair comparison.

The use of Random Forest classification is due to its capacity of dealing with non-linear relationships and interactions between the features. Extensive performance measure is found to be accuracy, precision, recall, F1-score and ROC-AUC analysis are performed to evaluate otherwise. Such synergistic approaches can give sufficient and useful model to the Cloud Era enterprise data engineering maturity prediction.

**V. Conclusion And Future Research**

**Conclusion**

In the study, a multi-dimensional model of the maturity of enterprise data engineering of cloud-native data platforms has been created successfully. Scalability, automation, governance, DataOps, and observability provide an overall view of the degree of preparedness of organizations. The machine learning-based (Random Forest) model had a high predictive performance, as shown through the high accuracy (0.875) and high ROC AUC values. The outcomes indicate that the two are automation and governance, which are the most influential drivers influencing maturity. Such a framework can effectively be intermediate between the conventional maturity models and the cloud-native ecosystem. In general, it's said that performing the maturity assessment on a structured, data-driven approach helps to decisions better and help enterprise digital transformation strategies.

**Future Research**

For future studies, the addition of real-world enterprise datasets instead of synthetic data could be better to have external validity. It is possible to consider deep learning methods such as neural networks and gradient boosting that are more robust and suitable. The representation of real-time streaming analytics and the introduction of self-evaluation systems based on AI could contribute to the increase in automation of maturity evaluation [22]. However, the framework can be expanded with the cost efficiency and sustainability data and multi-cloud settings that can be positioned. The framework would similarly be applicable in any other sector, such as healthcare, finance, and manufacturing and further validation studies in those areas might be identified in the future to render the

framework applicable to other sectors and generalizable.

## VI. References

- [1] Miryala, N.K. and Gupta, D., 2023. Big Data Analytics in Cloud—Comparative Study. *International Journal of Computer Trends and Technology*, 71(12), pp.30-34.
- [2] Ouyang, R., Wang, J., Xu, H., Chen, S., Xiong, X., Tolba, A. and Zhang, X., 2023. A microservice and serverless architecture for secure IoT system. *Sensors*, 23(10), p.4868.
- [3] Nwokeji, J.C. and Matovu, R., 2021. A systematic literature review on big data extraction, transformation and loading (etl). *Intelligent computing*, pp.308-324.
- [4] Joshi, A., Benitez, J., Huygh, T., Ruiz, L. and De Haes, S., 2022. Impact of IT governance process capability on business performance: Theory and empirical evidence. *Decision Support Systems*, 153, p.113668.
- [5] Al-Okaily, A., Al-Okaily, M., Teoh, A.P. and Al-Debei, M.M., 2023. An empirical study on data warehouse systems effectiveness: the case of Jordanian banks in the business intelligence era. *EuroMed Journal of Business*, 18(4), pp.489-510.
- [6] Keshireddy, S.R. and Kavuluri, H.V.R., 2021. Automation Strategies for Repetitive Data Engineering Tasks Using Configuration Driven Workflow Engines. *The SIJ Transactions on Computer Science Engineering & its Applications*, 9(1), pp.38-42.
- [7] Bhaskaran, S.V., 2020. Integrating data quality services (dqs) in big data ecosystems: Challenges, best practices, and opportunities for decision-making. *Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems*, 4(11), pp.1-12.
- [8] Juhasz, Z., 2021. Quantitative cost comparison of on-premise and cloud infrastructure based EEG data processing. *Cluster Computing*, 24(2), pp.625-641.
- [9] Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E. and Stoica, I., 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.
- [10] Junaid, S.B., Imam, A.A., Balogun, A.O., De Silva, L.C., Surakat, Y.A., Kumar, G., Abdulkarim, M., Shuaibu, A.N., Garba, A., Sahalu, Y. and Mohammed, A., 2022, October. Recent advancements in emerging technologies for healthcare management systems: a survey. In *Healthcare* (Vol. 10, No. 10, p. 1940). MDPI.
- [11] Mishra, M.S., KK, S.D. and MK, B.N., 2019. People & process dimensions of automation in business process management industry. *International Journal of Engineering and Advanced Technology*, 8(6), pp.2465-2472.
- [12] Banerjee, S., 2023. Challenges and solutions for data management in cloud-based environments. *International Journal of Advanced Research in Science, Communication and Technology*, pp.370-378.
- [13] Dikhanbayeva, D., Shaikholla, S., Suleiman, Z. and Turkyilmaz, A., 2020. Assessment of industry 4.0 maturity models by design principles. *Sustainability*, 12(23), p.9927.
- [14] Zahid, R., Altaf, A., Ahmad, T., Iqbal, F., Vera, Y.A.M., Flores, M.A.L. and Ashraf, I., 2023. Secure data management life cycle for government big-data ecosystem: Design and development perspective. *Systems*, 11(8), p.380.
- [15] Haleem, A., Javaid, M., Singh, R.P., Rab, S. and Suman, R., 2021. Hyperautomation for the enhancement of automation in industries. *Sensors International*, 2, p.100124.
- [16] Muvva, S., 2021. Cloud-Native Data Engineering: Leveraging Scalable, Resilient, and Efficient Pipelines for the Future of Data. *ESP Journal of Engineering & Technology Advancements*, 1(2), pp.287-292.
- [17] Akindemowo, A.O., Erigha, E.D., Obuse, E., Ajayi, J.O., Adebayo, A., Afuwape, A.A. and Adanyin, A., 2021. A Conceptual Framework for Automating Data Pipelines Using ELT Tools in Cloud-Native Environments. *Journal of Frontiers in Multidisciplinary Research*, 2(1), pp.440-452.
- [18] Tubis, A.A., 2023. Digital maturity assessment model for the organizational and process dimensions. *Sustainability*, 15(20), p.15122.

[19] Liu, Y., Peng, J. and Yu, Z., 2018, August. Big data platform architecture under the background of financial technology: In the insurance industry as an example. In Proceedings of the 2018 international conference on big data engineering and technology (pp. 31-35).

[20] Henning, S. and Hasselbring, W., 2022. A configurable method for benchmarking scalability of cloud-native applications. Empirical Software Engineering, 27(6), p.143.

[21] Bykov, V.Y. and Shyshkina, M.P., 2018. The conceptual basis of the university cloud-based learning and research environment formation and development in view of the open science priorities. Інформаційні технології і засоби навчання, (68,№ 6), pp.1-19.

[22] Niewiadomski, P., Stachowiak, A. and Pawlak, N., 2019. Knowledge on IT tools based on AI maturity–Industry 4.0 perspective. Procedia Manufacturing, 39, pp.574-582.