MANTECH Publications

Journal of Engineering Mathematics & Statistics
Volume 3, Issue 2, July-December 2019

Data-Centric Engineering: Principles, Architectures, and Emerging Applications

Ravi Chaudhary[1], Santosh Mishra[2]

Associate Professor[1], Assistant Professor[2]

Department of Systems Engineering

Rungta College of Engineering and Technology, Chhattisgarh

Email ID: *Ravichaudhary998@gmail.com[1], santoshmishraiid@yahoo.com[2]*

Abstract

In recent years, engineering systems have undergone a fundamental shift from model-driven and hardware-centric paradigms toward data-centric approaches. Data-Centric Engineering (DCE) emphasizes the systematic collection, management, processing, and utilization of data as the primary driver of engineering decisions across the system lifecycle. With the rapid growth of sensors, Internet of Things (IoT), cloud platforms, and machine learning techniques, data has become a strategic asset rather than a by-product of engineering processes. This paper presents a comprehensive review of Data-Centric Engineering, discussing its conceptual foundations, enabling technologies, data architectures, and practical applications across multiple engineering domains. Key challenges such as data quality, scalability, interoperability, and ethical concerns are also examined. The study highlights how data-centric thinking reshapes design, operation, and maintenance practices, enabling predictive, adaptive, and resilient engineering systems. The paper concludes by outlining future research directions and emphasizing the need for interdisciplinary collaboration to fully realize the potential of data-centric engineering.

Keywords: *Data-centric engineering, big data, digital engineering, data architecture, predictive analytics, systems engineering*

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

1. INTRODUCTION

Traditional engineering methodologies have long relied on physics-based models, deterministic assumptions, and carefully designed experiments. While these approaches remain essential, they often struggle to cope with the complexity, uncertainty, and scale of modern engineering systems. Advances in sensing, computation, and connectivity have resulted in massive volumes of heterogeneous data generated throughout the lifecycle of engineered systems. This has led to a paradigm shift where data itself becomes the central element in engineering analysis and decision-making.

Data-Centric Engineering (DCE) is an emerging framework that places data at the core of engineering processes, from conceptual design to deployment and maintenance. Instead of starting with fixed models and validating them using limited data, DCE encourages continuous learning from data streams, operational feedback, and real-world observations. This shift is particularly visible in areas such as smart manufacturing, intelligent transportation systems, energy networks, and healthcare engineering.

The motivation for adopting data-centric approaches arises from several factors. First, real-world systems often deviate from idealized models due to environmental variability, human interaction, and system aging. Second, data-driven methods can uncover hidden patterns and relationships that are difficult to derive analytically. Third, the availability of scalable computing infrastructure allows engineers to process and analyze large datasets in near real time.

This paper aims to provide a structured review of Data-Centric Engineering, highlighting its principles, tools, architectures, and applications. The discussion also addresses the limitations and open challenges that must be overcome for broader adoption.

2. CONCEPTUAL FOUNDATIONS OF DATA-CENTRIC ENGINEERING

Data-Centric Engineering (DCE) is grounded in a fundamental rethinking of how engineering systems are represented, analyzed, and improved. Rather than treating data as a secondary product of experiments or operations, DCE places data at the core of engineering reasoning. This

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

conceptual shift affects modeling philosophy, system design, and decision-making processes across disciplines.

2.1 From Model-Centric to Data-Centric Paradigms

Conventional engineering practice has been largely model-centric. In this paradigm, mathematical equations, physical laws, and simplified abstractions are used to represent real-world systems. These models are typically developed during the design phase and remain relatively static throughout the system's operational life. Experimental or operational data is mainly used to estimate parameters, calibrate models, or validate assumptions. While this approach has proven effective for well-understood and stable systems, it often struggles when systems become highly complex, nonlinear, or subject to changing operating conditions.

Data-centric engineering represents a significant departure from this traditional view. In a data-centric paradigm, data is not merely a supporting element but a primary source of knowledge about system behavior. Models are no longer assumed to be complete or fixed representations of reality. Instead, they are continuously updated, refined, or even replaced as new data becomes available. This allows engineering systems to adapt to changes, learn from experience, and respond to uncertainties more effectively.

Importantly, the move toward data-centric methods does not imply the rejection of physical principles or domain expertise. On the contrary, DCE encourages the integration of data-driven methods with established engineering knowledge. Hybrid approaches, such as physics-informed machine learning, combine governing equations with observational data to constrain learning and improve model reliability. By embedding physical laws into data-driven models, engineers can achieve better generalization, reduce data requirements, and improve interpretability. This synergy between data and theory is a defining characteristic of modern data-centric engineering.

2.2 Data as an Engineering Asset

A central idea in Data-Centric Engineering is the recognition of data as a valuable engineering asset. Just as hardware components and software systems require careful design, testing, and

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

maintenance, data must also be engineered systematically. Poorly collected or poorly managed data can be as damaging to an engineering system as faulty hardware or flawed algorithms.

Engineering data assets include raw sensor measurements, processed datasets, simulation outputs, operational logs, and contextual information such as environmental conditions. Managing these assets requires deliberate strategies for data acquisition, storage, documentation, and access control. Metadata, which describes the origin, structure, and quality of data, plays a critical role in enabling data reuse and traceability across projects and teams.

Data quality is a major concern in data-centric systems. Noise, missing values, sensor drift, and bias can significantly affect analytical results and lead to incorrect conclusions. In safety-critical applications, such as transportation or energy systems, poor data quality can result in unsafe decisions and system failures. Therefore, DCE emphasizes rigorous data validation, cleaning, and quality assessment procedures as integral parts of the engineering workflow.

Governance and ethical considerations are also part of treating data as an asset. Issues related to data ownership, privacy, security, and responsible use must be addressed through clear policies and technical safeguards. As engineering systems increasingly rely on shared and real-time data, effective data governance becomes essential for trust and long-term sustainability.

2.3 Lifecycle Perspective

Another foundational aspect of Data-Centric Engineering is its lifecycle-oriented perspective. Unlike traditional approaches that emphasize design-time analysis, DCE recognizes that valuable data is generated throughout the entire lifecycle of an engineering system. This data can be leveraged to support continuous improvement and informed decision-making at every stage.

During the early design phase, historical data from similar systems and simulated datasets are used to understand requirements, explore design alternatives, and perform trade-off analyses. Data-driven insights can reveal patterns in past failures, operational constraints, or user behavior that might not be captured by theoretical models alone.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

In the operational phase, real-time and near-real-time data enables continuous monitoring of system performance. Data-centric techniques support anomaly detection, adaptive control, and performance optimization under varying conditions. Instead of relying on fixed thresholds or predefined rules, systems can learn normal behavior from data and detect deviations more accurately.

Maintenance and end-of-life phases also benefit significantly from a data-centric approach. Condition-based and predictive maintenance strategies rely on operational data to estimate remaining useful life and schedule interventions proactively. Finally, data collected over the system's lifetime provides valuable feedback for future designs, closing the loop between operation and engineering development.

By spanning the full lifecycle, Data-Centric Engineering transforms engineering practice from a one-time design activity into an ongoing, learning-oriented process.

3. ENABLING TECHNOLOGIES

The successful implementation of Data-Centric Engineering depends on a set of enabling technologies that support the collection, storage, processing, and analysis of large and complex datasets. These technologies form the technical backbone of data-centric systems and determine how effectively data can be transformed into actionable engineering knowledge.

3.1 Sensing and Data Acquisition

Reliable sensing and data acquisition are the foundations of data-centric engineering. Modern engineering systems are equipped with a wide range of sensors that capture physical, chemical, mechanical, and operational parameters. Examples include temperature and pressure sensors in industrial plants, vibration sensors in rotating machinery, vision sensors in robotics, and environmental sensors in smart cities. The increased affordability and miniaturization of sensors have significantly expanded the scope of measurable system states.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

The integration of Internet of Things (IoT) technologies has further transformed data acquisition by enabling continuous and distributed data collection. IoT-enabled devices communicate through wired or wireless networks, allowing engineers to monitor geographically dispersed assets in real time. This capability is particularly important for large-scale infrastructures such as transportation networks, energy grids, and manufacturing supply chains.

Despite these advances, sensor-generated data is often imperfect. Measurement noise, calibration errors, sensor drift, and occasional failures introduce uncertainty and inconsistencies. In addition, data transmission losses and synchronization issues can affect data reliability. As a result, data preprocessing steps such as filtering, normalization, outlier detection, and sensor fusion are essential components of the data acquisition process. Addressing uncertainty at this early stage is critical, as errors can propagate through subsequent analysis and lead to unreliable conclusions.

3.2 Data Storage and Management

The rapid growth in engineering data volume, velocity, and variety poses significant challenges for data storage and management. Traditional centralized databases are often insufficient for handling continuous data streams and heterogeneous data types. As a result, data-centric engineering systems increasingly rely on scalable and distributed storage solutions.

Distributed file systems and cloud-based databases provide flexible and cost-effective platforms for storing large datasets. Data lakes are commonly used to store raw and semi-processed data in various formats, including time-series data, images, text logs, and simulation outputs. This approach allows engineers to preserve data for future analysis without imposing rigid schemas at the time of collection.

Effective data management goes beyond storage capacity. Metadata management plays a crucial role in ensuring that data can be understood, located, and reused over time. Metadata typically includes information about data sources, measurement conditions, units, timestamps, and quality indicators. Without proper metadata, datasets may lose their value, as users cannot interpret or trust the data.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Data governance and access control are also important considerations. Engineering projects often involve multiple teams and stakeholders, making it necessary to define clear policies for data sharing, version control, and security. Well-designed data management practices enable collaboration, reduce redundancy, and support long-term sustainability of data-centric initiatives.

3.3 Analytics and Machine Learning

Analytics and machine learning are central to extracting value from engineering data. Data-centric engineering employs a wide spectrum of analytical techniques, ranging from basic descriptive statistics to advanced predictive and prescriptive models. Descriptive analytics helps engineers understand historical behavior and identify trends, while diagnostic analytics aims to explain the causes of observed patterns.

Predictive analytics uses statistical and machine learning models to forecast future system behavior. Techniques such as linear and nonlinear regression, time-series analysis, and probabilistic modeling are widely applied in engineering contexts. For more complex and high-dimensional data, machine learning methods such as clustering, support vector machines, and deep neural networks are increasingly used.

Reinforcement learning and optimization-based methods are particularly relevant for control and decision-making problems. These techniques allow systems to learn optimal actions through interaction with the environment, making them suitable for adaptive control, resource allocation, and scheduling tasks.

The choice of analytical method depends strongly on the nature of the data, the availability of labeled examples, and the engineering objectives. In practice, interpretability, computational efficiency, and robustness are often as important as predictive accuracy. Therefore, data-centric engineering emphasizes careful model selection, validation, and continuous monitoring of model performance as new data becomes available.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

3.4 Digital Twins

Digital twins represent one of the most prominent applications of enabling technologies in data-centric engineering. A digital twin is a virtual representation of a physical system that evolves over time by incorporating real-time and historical data. Unlike static simulation models, digital twins are continuously synchronized with their physical counterparts.

Digital twins integrate data streams from sensors, physics-based models, and data-driven analytics into a unified framework. This integration allows engineers to perform what-if analyses, assess system health, and evaluate alternative operating strategies without interfering with the actual system. For example, in industrial settings, digital twins are used to predict equipment degradation and schedule maintenance activities before failures occur.

Beyond maintenance, digital twins support system optimization and design improvement. Insights gained from operational data can be fed back into the design process, enabling iterative refinement of system components and configurations. As data availability and computational capabilities continue to grow, digital twins are expected to play an increasingly central role in realizing the full potential of data-centric engineering.

4. DATA ARCHITECTURES FOR ENGINEERING SYSTEMS

Data architectures define how data is collected, transmitted, stored, processed, and accessed within engineering systems. In Data-Centric Engineering, architecture design is a critical decision, as it directly affects system performance, scalability, reliability, and the timeliness of engineering decisions. Well-designed data architectures enable efficient handling of large and heterogeneous datasets while supporting advanced analytics and real-time applications.

4.1 Centralized and Distributed Architectures

Engineering systems typically adopt either centralized or distributed data architectures, depending on system scale, latency requirements, and computational constraints. In centralized architectures, data from multiple sources is transmitted to a central repository, such as a cloud server or data center, where storage and processing are performed. This approach simplifies data integration,

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

governance, and global analysis, as all data is available in a unified location. Centralized architectures are commonly used in applications where batch analysis and long-term trend assessment are the primary objectives.

However, centralized data architectures can face limitations related to scalability and latency. As the number of data sources increases, network bandwidth and storage requirements grow rapidly. Transmitting large volumes of raw data to a central location can introduce delays, making it difficult to support real-time monitoring and control. In addition, centralized systems may represent single points of failure, which can reduce system robustness.

Distributed and edge-based architectures address these challenges by processing data closer to its source. In such architectures, computation and preliminary analysis are performed at the sensor, device, or local node level before transmitting summarized or relevant information to higher-level systems. This reduces data transfer costs, improves response times, and enables real-time decision-making in latency-sensitive applications. Distributed architectures are particularly suitable for IoT-enabled systems, autonomous vehicles, and smart infrastructure, where immediate local actions are often required.

In practice, many data-centric engineering systems adopt hybrid architectures that combine centralized and distributed elements. These hybrid approaches balance the need for real-time processing with the benefits of centralized data storage and global analysis.

4.2 Data Pipelines

Data pipelines are structured workflows that define how data flows through different stages of processing in a data-centric engineering system. A typical pipeline includes data ingestion, preprocessing, storage, analysis, and visualization or decision support. Each stage plays a distinct role in transforming raw data into meaningful engineering insights.

Data ingestion involves collecting data from sensors, databases, simulation tools, or external sources. This data is often heterogeneous and arrives at different rates, requiring flexible ingestion

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

mechanisms. Preprocessing steps such as data cleaning, filtering, normalization, and feature extraction are then applied to improve data quality and consistency.

Once processed, data is stored in appropriate repositories for short-term or long-term use. Analytical components apply statistical methods, machine learning algorithms, or optimization techniques to extract patterns, make predictions, or generate recommendations. Finally, results are presented through dashboards, reports, or automated control actions that support engineering decisions.

Automation of data pipelines is a key aspect of data-centric engineering. Automated pipelines reduce manual intervention, minimize errors, and ensure that data processing is repeatable and transparent. Figure 1 illustrates a generic data-centric engineering pipeline, highlighting the flow from data sources to decision support.

4.3 Interoperability and Standards

Interoperability is a persistent challenge in data-centric engineering, particularly in large and multidisciplinary projects. Engineering systems often involve diverse data sources, software tools, and organizational stakeholders, each using different data formats and conventions. Without interoperability, data exchange becomes inefficient, error-prone, and difficult to maintain.

Standardized data formats, communication protocols, and application programming interfaces (APIs) play a crucial role in addressing this issue. Standards enable different components of a system to communicate effectively, reducing the need for custom data translation and integration efforts. In addition, shared ontologies and data models provide common definitions and semantics, ensuring that data is interpreted consistently across disciplines.

Adoption of interoperability standards also supports long-term system evolution. As new components and technologies are introduced, standardized interfaces allow them to be integrated with minimal disruption. In the context of data-centric engineering, interoperability is not merely

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

a technical requirement but a strategic enabler of collaboration, reuse, and innovation across engineering domains.

5. APPLICATIONS OF DATA-CENTRIC ENGINEERING

5.1 Smart Manufacturing

In manufacturing, DCE enables predictive maintenance, quality control, and process optimization. Machine data, sensor readings, and production logs are analyzed to detect anomalies and predict equipment failures. This reduces downtime and improves product quality.

Table 1: Data-Centric Applications in Manufacturing

Application Area	Data Sources	Benefits
Predictive maintenance	Sensor and machine logs	Reduced downtime
Quality monitoring	Inspection and process data	Improved consistency
Process optimization	Production parameters	Higher efficiency

5.2 Transportation and Mobility

Transportation systems generate massive data from vehicles, infrastructure, and users. Data-centric approaches support traffic prediction, route optimization, and safety analysis. Autonomous vehicles are a prominent example, relying on real-time data fusion and learning.

5.3 Energy Systems

In energy engineering, data-centric methods are used for load forecasting, fault detection, and renewable energy integration. Smart grids leverage data from meters, sensors, and weather sources to balance supply and demand dynamically.

5.4 Healthcare Engineering

Healthcare systems increasingly depend on data from medical devices, imaging systems, and electronic health records. Data-centric engineering supports personalized treatment, equipment

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

management, and operational efficiency, although privacy concerns are particularly critical in this domain.

6. CHALLENGES AND LIMITATIONS

Despite its promise, Data-Centric Engineering faces several challenges.

6.1 Data Quality and Uncertainty

Engineering data is often noisy, incomplete, and biased. Ensuring data quality requires robust preprocessing, validation, and uncertainty quantification. Without these steps, data-driven models may produce unreliable results.

6.2 Scalability and Complexity

As systems grow in scale, managing and analyzing data becomes increasingly complex. Computational costs, storage requirements, and system integration issues can limit practical deployment.

6.3 Interpretability and Trust

Many data-driven models, especially deep learning, are perceived as black boxes. In safety-critical engineering applications, lack of interpretability can hinder trust and regulatory approval.

6.4 Ethical and Security Concerns

Data-centric systems raise ethical issues related to privacy, surveillance, and data ownership. Cybersecurity threats also pose significant risks, as compromised data can lead to incorrect decisions or system failures.

7. FUTURE RESEARCH DIRECTIONS

Future work in Data-Centric Engineering is expected to focus on hybrid modeling approaches, combining data-driven and physics-based methods. Advances in explainable artificial intelligence will improve model transparency and acceptance. Additionally, standardized data frameworks and improved governance policies will be critical for large-scale adoption.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

Another promising direction is autonomous data-centric systems that can adapt data collection strategies based on current uncertainty and objectives. This closes the loop between sensing, analysis, and decision-making.

8. CONCLUSION

Data-Centric Engineering represents a transformative shift in how engineering systems are designed, operated, and maintained. By placing data at the center of engineering workflows, DCE enables adaptive, predictive, and resilient systems capable of handling real-world complexity. While significant challenges remain in data quality, scalability, and ethics, ongoing advances in sensing, analytics, and digital infrastructure continue to strengthen the data-centric paradigm. As engineering problems become increasingly complex and interconnected, data-centric approaches will play a vital role in shaping the future of engineering practice and research.

REFERENCES

1. Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18–23.
2. Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary Perspectives on Complex Systems*, 85–113.
3. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
4. Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. *Sage Publications*.
5. Wang, L., Törngren, M., & Onori, M. (2015). Current status and advancement of cyber-physical systems in manufacturing. *Journal of Manufacturing Systems*, 37, 517–527.
6. Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70, 263–286.
7. Jazdi, N. (2014). Cyber physical systems in the context of industry 4.0. *IEEE International Conference on Automation, Quality and Testing, Robotics*, 1–4.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved

8. Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers. *Computers & Fluids*, 200, 104420.
9. Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1), 51–59.
10. Reichert, M., & Weber, B. (2012). *Enabling Flexibility in Process-Aware Information Systems*. Springer.

Page 138-151 © MANTECH PUBLICATIONS 2019. All Rights Reserved