Design and Deployment of Data Developer Toolkit in Cloud Manufacturing Environments[1]*

Nataliya Shakhovska[† 1,2], Dmytro Liaskovsky [† *1], Andy Augousti [†1,3], Solomiia Liaskovska[†1,3], Yevgen Martyn[†4]

[1] *Department of Systems of Artificial Intelligence, Lviv Polytechnic National University,Lviv,79905, Ukraine*
[2] *College of Engeneering, Design and Phisical Sciences, Brunel University London, Kingston Lane, Uxbridge, Middlesex UB8 3PH United Kingdom*
[3] *Faculty of Mechanical Engineering, Kingston University, Kingston Upon Thames, KT1 1LQ, London, United Kingdom*
[4] *Department of Project Management, Information Technologies and Telecommunications, Lviv State University of Life Safety, Lviv,79007, Ukraine*

Abstract

The consequences of the "digital revolution" permeate every sphere of business and science, including manufacturing. Data developers and analysts are at the forefront of these changes, providing companies with tools to extract valuable insights from raw production data. The primary step for companies on this path lies in the strategic adoption and utilization of cloud technologies. However, data developers often encounter challenges when implementing analytical processes in the cloud, especially regarding scalability, resource allocation, data accessibility, and security, which are particularly crucial in the context of manufacturing.

Scalability issues arise due to the changing volume of production data: what may be acceptable today could become overwhelming tomorrow. This unpredictability makes planning analytical processes complex and uncertain, especially when dealing with growing volumes of production data.

Resource allocation poses another challenge that data developers face in the manufacturing environment. Cloud services can be costly, especially for new companies in the manufacturing sector. Additionally, using public clouds means that manufacturing applications and data may reside on servers alongside data from other organizations, posing risks to data confidentiality and security.

Access to production data in the cloud may be restricted or slow due to data transmission over networks, especially when dealing with large volumes of data. This problem can lead to delays in data analysis and processing, ultimately affecting the productivity of manufacturing enterprises.

Manufacturing data is crucial for understanding and optimizing production processes, but processing it in a cloud environment can also pose challenges related to security, speed, and availability.

Keywords

Cloud services, manufacturing, applications, cloud data storage modelling

SMARTINDUSTRY-2024: International Conference on Smart Automation & Robotics for Future Industry, April 18 - 20, 2024, Lviv, Ukraine

[* *] Corresponding author.

[†] These authors contributed equally.

nataliya.b.shakhovska@lpnu.ua (N. Shakhovska) ; dliaskovskyi@gmail.com (D. Liaskovskyi); augousti@kingston.ac.uk (A. Augousti); solomiam@gmail.com (S. Liaskovska); evmartun@gmail.com (Y. Martyn);

0000-0002-6875-8534 (N. Shakhovska); 0009-0005-1621-3953 (D. Liaskovsky); 0000-0003-3000-9332 (A. Augousti); 0000-0002-0822-0951(S. Liaskovska); 0000-0001-9095-7057 (E.Martyn);

1. Introduction

The issue of scalability arises due to the unpredictable nature of data volumes: today they may be minimal, but tomorrow they could skyrocket to unmanageable levels. This unpredictability complicates the planning of analytical processes, making them intricate and uncertain. Additionally, there's the challenge of resource allocation. Cloud services can be costly, particularly for startups. Moreover, if you utilize public clouds, your applications and data may share servers with other organizations, posing risks to data security and confidentiality.

Accessing data in the cloud may also be hindered by limited or slow data transmission over the network, especially when dealing with large volumes of data. This bottleneck can result in delays in data analysis and processing, impacting work efficiency.

Furthermore, there's a significant security concern. Data is among a company's most valuable assets, and mishandling or leaks of this data can severely damage the company's reputation and financial standing.

However, the most significant challenge for data developers and analysts lies in selecting the right infrastructure. Despite understanding the task at hand, they may not always know which infrastructure option is best suited to solve it. This uncertainty can lead to errors in selection, unnecessary expenses, and time wastage.

The objective of this article is to help comprehend these challenges and offer effective solutions for overcoming them. We will analyze the features and capabilities of major cloud providers such as AWS, Azure, and GCP, as well as explore the potential of open-source software such as Zeppelin, Jupyter, and R-Studio.

2. Materials and Methods

To creating virtual copies of mechanical engineering objects and modeling the interaction processes of system parameters, it is also important to focus on the role of data in manufacturing. Data from production processes, which are analyzed and processed, play a key role in making strategic decisions and improving the efficiency of manufacturing processes. This data may include information about equipment status, production quality, material costs, product quality levels, and much more. Modelling Cloud Data Environment and Analysis of Transmission Methods

The Internet of Things (IoT) also plays a crucial role in data collection in manufacturing. Sensors and connected devices can gather a large amount of data about equipment status and processes, allowing for the timely identification of problems, avoiding breakdowns, and optimizing the operation of production lines.

Analyzing and processing this large volume of data (Big Data) requires the use of cloud and artificial intelligence (AI) technologies. Cloud solutions provide computing power for processing large volumes of data, while artificial intelligence helps identify patterns and make forecasts based on this data.

The integration of cloud-based environments becomes crucial for effective data collection, analysis, and utilization in manufacturing. Leveraging specialized IT platforms and business process management systems becomes indispensable. These tools not only streamline the flow of information but also facilitate prompt responses to fluctuations in manufacturing processes, ensuring optimal outcomes. For the purposes of cloud data storage modeling, it is better to use its scalable version introduced by Petrov.:

$$S_m = \langle F, D(t), G, C, L \rangle . \quad (2)$$

Then the cloud data storage model is presented as:

$$S_{cloud} = \langle D, D_{free}, S_{ms} \rangle , \quad (3)$$

where

$$D_{free} \subseteq D$$

subset of available storage devices,

$$S_{ms} = \{S_{m1}, S_{m2}, ..., S_{ml}\}$$

set of scalable storage repositories.

In this case, scalable devices are devices from the set of general devices that do not include the subset of available devices.

$$D_i(t) = D \setminus D_{free} .$$

Scalable devices do not share storage devices.

$$\forall t, i, j, i \neq j \Rightarrow D_i(t) \cap D_j(t) = \varnothing .$$

The cloud storage model has been enhanced as an algebraic system.

$$C_{dw} = \langle S_{cloud\_m}; Y; L \rangle ,$$

$$S_{cloud\_m} = \langle D, D_{free}, S_{ms}, PR \rangle ,$$

$$Y = \{I_{cc,} I_{mpp}, I_{mpd}\} ,$$

where

I_{cc} – the method of selecting a gateway based on query complexity.,

I_{mpp} – the method of multiprotocol streaming data transmission.,

I_{mpd} – the method of multiplexing different data sources for simultaneous transmission.,

PR – data transmission protocol.

Load predicate

$$f_i \in St \cup SemSt \cup UnSt$$

It can be represented by structured, semi-structured, and unstructured data.

In order to organize the provision of any service in cloud technologies and access to cloud storage, in particular, it is necessary to have the appropriate storage. That is, a server or a network of servers through which the storage access service is provided to clients.

The most common way of providing a service to a client is to handle their requests. To handle a request means to receive the request and send a response to the party that created it.

To confirm the existence of the self-similarity property for various data streams in a multiservice network, it is necessary to measure certain characteristics of different types of network traffic. For this, statistical data on streams and data traffic are required, as well as research on the combined stream and variable characteristics of the cloud storage server must be conducted.

3. The development of an analytical platform for data scientists requires connectivity to the following cloud data storage solutions to ensure effective data analysis.

For the development of an analytical platform for data scientists, it is important to connect to the following cloud data storages to ensure effective data analysis.

Amazon Web Services (AWS):

Amazon Simple Storage Service (Amazon S3): S3 is often used as a primary location for big data analytics. It helps store and analyze any amount of data and interacts with a range of AWS analytical services, including Amazon Athena, Amazon Redshift, and AWS Glue.

Amazon Redshift: This is a fully-managed clustered data warehouse service that provides fast, simple, and flexible analysis of all your data using your familiar SQL client.

Amazon DynamoDB: This is a NoSQL database management service that offers fast and predictable performance with seamless scalability.

Microsoft Azure:

Azure Blob Storage: This service provides scalable, reliable, and secure object storage for unstructured data.

Azure Data Lake Storage: This is a secure, scalable, large-scale storage solution for big data.

Azure Synapse Analytics (formerly SQL Data Warehouse): This is an analytics service that seamlessly integrates big data storage with a distributed computational resource.

Google Cloud Platform (GCP):

- **Google Cloud Storage:** This service allows individual users to store large amounts of data on Google Cloud.
- **Google BigQuery:** This is a serverless data warehouse that automatically scales as you store and analyze data.
- **Google Cloud Firestore/Google Cloud Datastore:** These are non-relational databases designed for web-scale applications, depending on your requirements.

Connection parameters to these storages can also be configured with various data analysis libraries to ensure quick access to data.

4. Modeling and results

In today's world, where companies are actively moving towards 'digitalization,' the analysis of internal and external data is becoming increasingly common. Until recently, this data might not have received the necessary attention. However, issues with the diversity and sheer volume of incoming data, ensuring their security, including 'sensitive' information, and the lack of computational resources and tools for working with data pose significant obstacles to effective analytics. This limits the capabilities of experts in data science and machine learning. Considering the possibilities and limitations of cloud environments, engineers, data developers, and scientists need to monitor all aspects of deployment, support, scaling, and payment of infrastructure, secure data access, and the necessity of sharing code and models.

Given these and a number of other reasons, the idea arises to create a tool for accelerating the work of data scientists in the form of self-service. This tool should help quickly deploy powerful analytical 'sandboxes' in the cloud without involving DevOps. It should provide users with the ability to add computational resources as needed, use a convenient interface to install additional libraries and dependencies, collaborate in a team without worrying about the security of the environment and data.

This environment should be compatible with major cloud providers such as Amazon, MS Azure, and Google Cloud. It should allow data scientists to join projects at the analysis stage, speeding up the adoption of analytical decisions without waiting for the final infrastructure to become available and the architecture to be agreed upon.

4.1. Key features of the service for the work of data scientists.

Let's consider main features of the service for the work of data scientists:

Integration with relevant analytical tools such as Jupyter, Zeppelin, RStudio, TensorFlow, Spark, and others.

Support for programming languages such as Python, Scala, R, Java, providing flexibility and the ability to use the latest technologies.

Simplified installation of libraries and frameworks, allowing the environment to be adapted to specific tasks.

Integration with Spark clusters or Cloud Data Engines (such as EMR on AWS, Data Proc on GCP, HDInsight on MS Azure) to ensure high performance when processing large volumes of data.

Full integration with Azure Data Lake for an even broader range of capabilities.

Ensuring data protection during storage and transmission.

Use of LDAP, Cloud Identity Management Services, SSO for user authentication.

The ability to use personal and shared data storage on AWS S3, Azure BlobStorage, Azure Data Lake.

Preparation of financial reports on the use of cloud resources, allowing for cost control.

Working with Spot Instances, Low Priority, and Preemptible VMs for efficient budget use on AWS, Azure, GCP infrastructures.

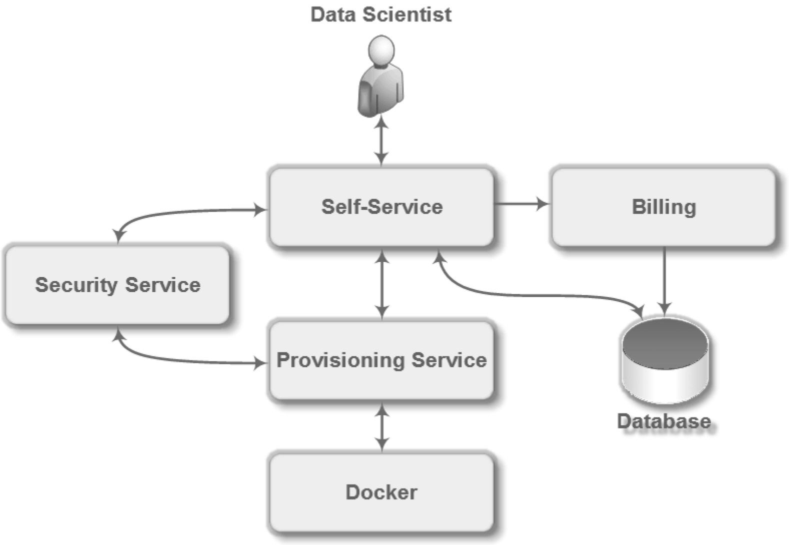

Figure 1: Interconnections of services for the data scientist's work platform

Self-service node (SSN); Edge node; Notebook node (Jupyter, Rstudio, Zeppelin, etc.) Data engine cluster. Such a platform should provide flexibility and speed in determining analytical decisions, as well as empower developers to focus their efforts on data analysis rather than environment setup.

4.2. Self-service node

Creating a Self-Service Node (SSN) is the initial stage of deploying the service. It serves as the main node point from which the environment setup begins. It includes the following key services and components:

1. Web UI: This is the user web interface that allows managing all system components. This interface provides users with a simple and intuitive method to interact with the system.
2. DB: This is the database where some system settings, user personal settings, and system metadata are stored. This database is an important part of the infrastructure as it provides a centralized storage location for data.
3. Docker: Used for creating and managing containers in which various parts of the infrastructure are deployed. Docker containerization technology provides flexibility and speed in deploying services.
4. Jenkins: This is an automation tool, installed on the SSN node, and can be used to manage infrastructure as an alternative to Web UI. Jenkins provides the ability to automate a range of processes related to the software development lifecycle.

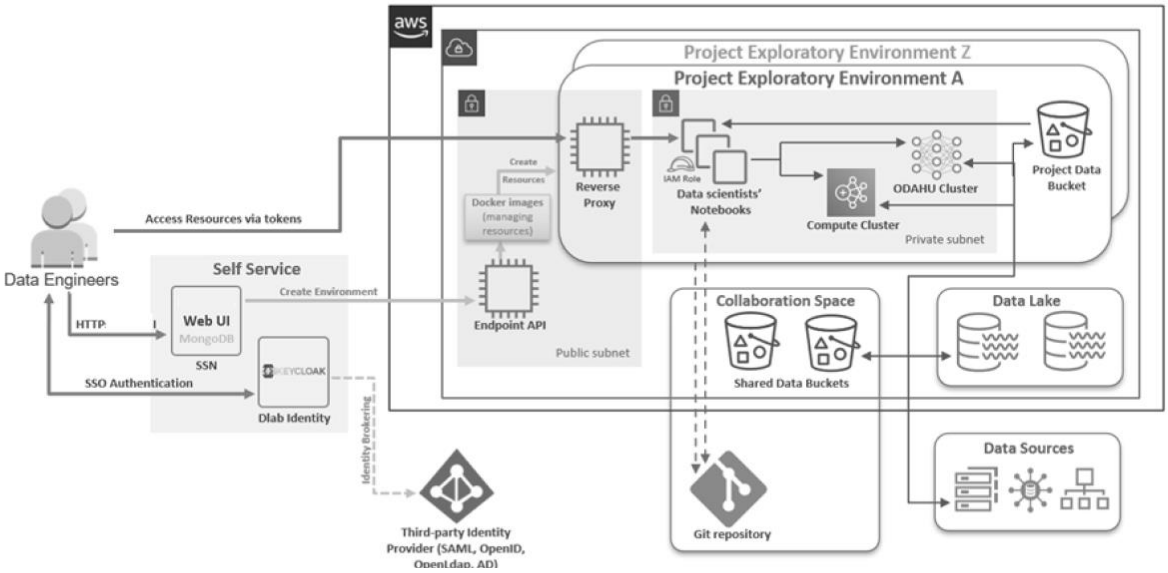

Figure 2: Creating a Self-Service Node (SSN)

Therefore, SSN serves as an important element of the system environment, connecting key components and services.

4.3. Self-service node and Notebook Node

Creating an Edge Node is the next step after entering the system. It serves as a proxy server and SSH gateway for the user, providing users with access to the Notebook via HTTP and SSH through a pre-installed HTTP web proxy server.

Installing the Notebook Node is the next stage. It serves as a server with pre-deployed applications and libraries for data processing, cleansing, transformation, mathematical modeling, Machine Learning

The analyst installs the following tools on the Notebook Node:

1. Jupyter
2. RStudio
3. Zeppelin
4. TensorFlow + Jupyter
5. Deep Learning + Jupyter

Apache Spark is installed for each of the aforementioned analytical tools.

4.4. Data Engine Cluster

After configuring the Notebook Node, users can create the following clusters for it:

Data engine: This is a standalone Spark cluster.

Data engine service: This is a cloud cluster platform (EMR for AWS, HDInsight for MS Azure, or Google Dataproc).

It simplifies the use of Hadoop and Apache Spark for processing and analyzing large volumes of data. Adding a cluster is not mandatory but is only done when additional computational resources are required for tasks.

5. Conclusion

The increasing trend towards digitalization has highlighted the importance of data analysis, both internally and externally, for companies across various industries. However, challenges such as data diversity, volume, security concerns, and limited computational resources hinder effective analytics. These obstacles underscore the need for a solution that empowers data scientists to accelerate their work without being bogged down by infrastructure complexities.

The concept of a self-service tool emerges as a viable solution, enabling rapid deployment of analytical environments in the cloud without DevOps involvement. Such a tool should offer scalability, flexibility, and security, allowing users to focus on analysis rather than environment setup. Compatibility with major cloud providers ensures accessibility and collaboration, expediting the adoption of analytical decisions.

Key features of this service include seamless integration with popular analytical tools and programming languages, simplified library installation, and integration with high-performance computing resources like Spark clusters. Data protection measures, user authentication protocols, and flexible data storage options further enhance its utility and security.

In essence, this self-service environment empowers data scientists to unleash the full potential of their analyses, driving innovation and efficiency in the era of digital transformation.

References

[1] Chao YU, Qing LI, Kui LIU, Yuwen CHEN, Hailong WEI, Industrial Design and Development Software System Architecture Based on Model-Based Systems Engineering and Cloud Computing, Annual Reviews in Control, Volume 51, 2021, Pages 401-423, ISSN 1367-5788, doi:10.1016/j.arcontrol.2021.04.011.

[2] Jafar A. Alzubi, Ramachandran Manikandan, Omar A. Alzubi, Issa Qiqieh, Robbi Rahim, Deepak Gupta, Ashish Khanna, Hashed Needham Schroeder Industrial IoT based Cost Optimized Deep Secured data transmission in cloud, Measurement,Volume 150, 2020, 107077, ISSN 0263-2241, doi:10.1016/j.measurement.2019.107077.

[3] S. Corbellini, E. Di Francia, S. Grassini, L. Iannucci, L. Lombardo, M. Parvis, Cloud based sensor network for environmental monitoring, Measurement, Volume 118, 2018, Pages 354-361, ISSN 0263-2241, doi:10.1016/j.measurement.2017.09.049.

[4] Víctor Garrido-Momparler, Miguel Peris, Smart sensors in environmental/water quality monitoring using IoT and cloud services, Trends in Environmental Analytical Chemistry, Volume 35, 2022, e00173, ISSN 2214-1588, doi:10.1016/j.teac.2022.e00173.

[5] Mahidur R. Sarker, Amna Riaz, M.S. Hossain Lipu, Mohamad Hanif Md Saad, Mohammad Nazir Ahmad, Rabiah Abdul Kadir, José Luis Olazagoitia, Micro energy harvesting for IoT platform: Review analysis toward future research opportunities, Heliyon, vol. 10, Issue 6, 2024 ISSN p.p. 2405-8440 doi:10.1016/j.measurement.2017.09.049.

[6] Meric Yilmaz Salman, Halil Hasar, Review on environmental aspects in smart city concept: Water, waste, air pollution and transportation smart applications using IoT techniques,

Sustainable Cities and Society, vol 94, 2023, 104567, ISSN 2210-6707, doi:10.1016/j.scs.2023.104567.

[7] Abdulmohsen Almalawi, Fawaz Alsolami, Asif Irshad Khan, Ali Alkhathlan, Adil Fahad, Kashif Irshad, Sana Qaiyum, Ahmed S. Alfakeeh, An IoT based system for magnify air pollution monitoring and prognosis using hybrid artificial intelligence technique, Environmental Research, vol. 206, 2022, 112576, ISSN 0013-9351, doi: 10.1016/j.envres.2021.112576.

[8] B. T. Cao, M. Obel, S. Freitag, P. Mark, and G. Meschke, 'Artificial neural network surrogate modelling for real-time predictions and control of building damage during mechanised tunnelling', *Advances in Engineering Software*, vol. 149, p. 102869, Nov. 2020, doi: 10.1016/j.advengsoft.2020.102869.

[9] Zakir Hossain, A. K. M., Hassim, N. B., Alsayaydeh, J. A. J., Hasan, M. K., &amp; Islam, M. R. (2021). A tree-profile shape ultra wide band antenna for chipless RFID tags. International Journal of Advanced Computer Science and Applications, 12(4), 546-550. doi:10.14569/IJACSA.2021.0120469.

[10] Investigation of Anomalous Situations in the Machine-Building Industry Using Phase Trajectories Method, 2022 doi:10.1007/978-3-031-03877-8_5.

[11] N. Shakhovska, V. Yakovyna, and N. Kryvinska, 'An Improved Software Defect Prediction Algorithm Using Self-organizing Maps Combined with Hierarchical Clustering and Data Preprocessing', in *Database and Expert Systems Applications*, vol. 12391, S. Hartmann, J. Küng, G. Kotsis, A. M. Tjoa, and I. Khalil, Eds., in Lecture Notes in Computer Science, vol. 12391. , Cham: Springer International Publishing, 2020, pp. 414–424. doi: 10.1007/978-3-030-59003-1_27.

[12] Adam Wong Yoon Khang, Shamsul J. Elias, Nadiatulhuda Zulkifli, Win Adiyansyah Indra, Jamil Abedalrahim Jamil Alsayaydeh, Zahariah Manap, Johar Akbar Mohamat Gani, 2020. Qualitative Based QoS Performance Study Using Hybrid ACO and PSO Algorithm Routing in MANET. Journal of Physics, Conference Series 1502 (2020) 012004, doi:10.1088/1742-6596/1502/1/012004.

[13] L. Mochurad, K. Shakhovska, and S. Montenegro, 'Parallel Solving of Fredholm Integral Equations of the First Kind by Tikhonov Regularization Method Using OpenMP Technology', in *Advances in Intelligent Systems and Computing IV*, vol. 1080, N. Shakhovska and M. O. Medykovskyy, Eds., in Advances in Intelligent Systems and Computing, vol. 1080. , Cham: Springer International Publishing, 2020, pp. 25–35. doi: 10.1007/978-3-030-33695-0_3.

[14] L. Mochurad, O. Kotsiumbas, and I. Protsyk, 'A Model for Weather Forecasting Based on Parallel Calculations', in *Advances in Artificial Systems for Medicine and Education VI*, vol. 159, Z. Hu, Z. Ye, and M. He, Eds., in Lecture Notes on Data Engineering and Communications Technologies, vol. 159. , Cham: Springer Nature Switzerland, 2023, pp. 35–46. doi: 10.1007/978-3-031-24468-1_4.

[15] S.-A. Mitoulis *et al.*, 'Conflict-resilience framework for critical infrastructure peacebuilding', *Sustainable Cities and Society*, vol. 91, p. 104405, Apr. 2023, doi: 10.1016/j.scs.2023.104405.