**Digital transformation: System and Data architecture as backbone for AI/ML in the Metals Industry**

Dr. Franck K. Aadjogble1,

1SMS Digital

100 Sandusky Street, PA, USA, 15044

Phone: +14126087120

Email: [franck.adjogble@sms-group.com](mailto:franck.adjogble@sms-group.com)

Sai Parvathaneni2

2SMS Digital

100 Sandusky Street, PA, USA, 15044

Phone: +14126087120

Email: [sai.parvathaneni@sms-group.com](mailto:sai.parvathaneni@sms-group.com)

Jason Lenox3

3SMS Digital

100 Sandusky Street, PA, USA, 15044

Phone: +14126087120

Email: [jason.lenox@sms-group.com](mailto:jason.lenox@sms-group.com)

Keywords: System Infrastructure, Data Structure, Machine Learning, Domain Knowledge, Solution design

**Abstract**

What makes a steel plant highly productive is that it uses data and the values derived from it as the basis for decisions, which are relevant to business operations and process control. Primarily, digitalization is about intelligently interconnecting and systematically networking knowledge and data. Frequently, these two resources are distributed throughout company organizations or data silos with strict access rules. That’s exactly the kind of structure that needs to be overcome. Then information will be available within the company – either in parts of the organization or even across company boundaries – in real time and in a way that provides the basis for decisions. The implementation of such system needs the involvement of a tailored System Architecture and Data Structure, which this paper presents. The results of real applications in the Metals industry will demonstrate a flexible system structure configuration, a highly available platform architecture, maximization of hardware investment via complex load balancing, and strong focus on exceeding industry standards related to system security.

**INTRODUCTION**

The steel industry is facing intense competition and increasing pressure to optimize its operations and reduce costs. Digital transformation is a key enabler for improving productivity and competitiveness in the metals industry. It involves the integration of digital technologies and data analytics into business processes to optimize operations, reduce costs, and improve efficiency.

One of the key drivers of digital transformation in the metals industry is the use of data and the values derived from it as the basis for decisions. The use of data analytics to optimize operations and improve efficiency is not a new concept. However, the availability of advanced data analytics tools and technologies, such as AI and ML, has significantly enhanced the ability to analyze and leverage data to drive business decisions.

The challenge for the metals industry is that data is often distributed throughout company organizations or data silos with strict access rules, hindering its effective use. Digitalization is about intelligently interconnecting and systematically networking knowledge and data. Therefore, a tailored System Architecture and Data Structure are required to overcome these challenges and enable real-time access to information that provides the basis for decisions.

This paper presents a tailored System Architecture and Data Structure for enhancing productivity in the metals industry. The proposed system architecture and data structure provide a scalable and modular platform for integrating different data sources into a unified data platform. The data structure supports efficient data management and processing, including storing, cleaning, normalizing, and structuring data, making it easily accessible for analysis and decision-making.

The paper is organized as follows: Section 2 provides a literature review of digital transformation in the metals industry. Section 3 presents the proposed System Architecture and Data Structure. Section 4 discusses real-world applications of the system in the metals industry. Section 5 provides a discussion of the results and Section 6 presents the conclusion and future work.

**Literature review**

Digital transformation is a crucial enabler for improving productivity and competitiveness in the metals industry. The integration of digital technologies and data analytics into business processes can optimize operations, reduce costs, and improve efficiency. This section provides a detailed review of the relevant literature on digital transformation, system infrastructure, and data structure in the metals industry.

The metals industry is undergoing a transformation in response to the increasing pressure to optimize operations, reduce costs, and improve efficiency. Digital transformation is seen as a key enabler for achieving these goals. According to a report by McKinsey & Company (2018), digital transformation in the metals industry can result in a reduction of up to 15% in operating costs and a 20% increase in productivity. The implementation of digital transformation in the metals industry requires a holistic approach that includes the integration of data analytics, automation, and advanced technologies, such as AI and ML. These technologies can be used to optimize production processes, reduce energy consumption, and improve product quality.

The implementation of digital transformation in the metals industry requires a robust system infrastructure that can handle large volumes of data and provide real-time access to critical information. System infrastructure includes hardware, software, and network components. A key challenge for system infrastructure in the metals industry is the integration of different data sources, such as sensors, machines, and manual inputs. A scalable and modular platform is required for integrating different data sources into a unified data platform. Hu and Han (2019) highlight the importance of a scalable and modular platform for data-driven optimization of metals manufacturing processes. They suggest that a modular architecture allows for the easy integration of new data sources and the scalability required for handling large volumes of data. Efficient data management and processing are also critical components of system infrastructure in the metals industry. Storing, cleaning, normalizing, and structuring data in a way that makes it easily accessible for analysis and decision-making is essential. Moccia et al. (2019) suggest that data lakes, data warehouses, and data marts can be used to support efficient data management and processing in the metals industry. The system infrastructure must also support real-time access to critical information. This requires a highly available platform architecture that can handle large volumes of data without latency or downtime. Bischoff and Kinkel (2019) suggest that the use of cloud-based platforms and edge computing can support real-time access to critical information in the metals industry.

Efficient data management and processing is crucial to the success of digital transformation in the metals industry. A critical component of the system infrastructure that supports efficient data management and processing is the data structure. This section provides a detailed literature review of the relevant research on data structure in the metals industry.

The data structure is essential for enabling the integration of different data sources into a unified data platform. Efficient data management and processing require storing, cleaning, normalizing, and structuring data in a way that makes it easily accessible for analysis and decision-making. According to Moccia et al. (2019), data lakes, data warehouses, and data marts are commonly used in the metals industry to support efficient data management and processing.

Data lakes are used to store raw and unstructured data, making it easily accessible for analysis and decision-making. Data warehouses are used to store structured data, making it easier to query and analyze. Data marts are used to store specific subsets of data that are relevant to specific business units or use cases.

The data structure must also support the development of advanced analytics models, such as AI and ML. AI and ML require large amounts of data to be trained effectively, and a well-designed data structure can support this training. According to Wang et al. (2020), edge computing and AI can be used to support intelligent maintenance of steel plants, improving efficiency and reducing costs.

The literature review highlights the importance of a robust system infrastructure and data structure in the metals industry for successful digital transformation. The review indicates that a scalable and modular platform is required for integrating different data sources into a unified data platform. The data structure must support efficient data management and processing, and the development of advanced analytics models such as AI and ML.

However, the review reveals a gap in the research regarding the implementation of a tailored System Architecture and Data Structure to address the challenges of digital transformation in the metals industry. Although there is a significant body of research on system infrastructure and data structure, there is a lack of research on a tailored approach that meets the specific needs of the metals industry.

To address the gap in the literature, this paper proposes a tailored System Architecture and Data Structure for enhancing productivity in the metals industry. The approach focuses on the integration of different data sources into a unified data platform, efficient data management and processing, and the development of advanced analytics models.

The proposed approach is designed to meet the specific needs of the metals industry by providing a scalable and modular platform for integrating different data sources, including sensors, machines, and manual inputs, into a unified data platform. The data structure supports efficient data management and processing, including storing, cleaning, normalizing, and structuring data, making it easily accessible for analysis and decision-making.

The proposed approach also supports the development of advanced analytics models, such as AI and ML, to optimize operations and improve efficiency in the metals industry. The approach includes a highly available platform architecture, complex load balancing to maximize hardware investment, and strong focus on exceeding industry standards related to system security.

The next section of the paper presents the proposed System Architecture and Data Structure in detail, providing a step-by-step guide for implementation. Real-world applications of the approach in the metals industry are discussed in the following section, followed by a discussion of the results and the conclusion and future work.

**Proposed approach and Concepts of system infrastructure and Data Structure**

The System Infrastructure design is a critical component of digital transformation in the metals industry. The design concept is built around five key goals: reliability, performance, expandability, ease of use, and security.

The following technical introduction provides an overview of the design concept and its key components.

**Reliability:**

The System Infrastructure design is designed to achieve tight SLA's and around-the-clock operations. This is achieved through the use of a highly available platform architecture that includes redundant hardware and network components. The design also includes automated failover and disaster recovery mechanisms to ensure maximum uptime, system errors handling, updates, maintenance, and facility issues do not result in a loss of productivity.

The following section elaborates in detail on the Reliability goals of the System Infrastructure design.

High Availability:

The System Infrastructure design incorporates multiple physical servers tethered together with automation to shift resources between servers. This design ensures that in the event of a system error, update, maintenance, or facility issue such as a power outage, partial network failure, fire, etc., the system remains available, and there is no loss of productivity. The design includes automated failover and disaster recovery mechanisms to ensure maximum uptime and zero disruption to operations.

Isolation of Services:

The System Infrastructure design ensures that a single server or group of servers no longer houses multiple services. This design allows for the isolation of services, enabling maintenance on a single service without the concern of taking down other essential business resources. This design ensures that maintenance on one service does not affect the availability or performance of other services.

Zero Single Points of Failure:

The System Infrastructure design ensures that there are zero single points of failure in the system. This is achieved through the use of hyper-converging technologies, where all data is replicated across all nodes of the system. If one node fails, the other picks up within milliseconds of interruption and zero time synchronizing and rebuilding. This design ensures that no single device, such as a NAS, SAN, switch, etc., can take down the whole system, ensuring maximum uptime and zero disruption to operations.

Core Technologies:

The System Infrastructure design incorporates several core technologies, including Microsoft Hyper-V, Microsoft Storage Spaces Direct, Microsoft Windows Server Datacenter Edition, Microsoft SQL Server Enterprise AlwaysOn Availability Groups, and Microsoft Failover Clustering. These technologies are designed to ensure maximum uptime and zero disruption to operations, ensuring high availability and reliability of the system.

**Performance:**

The System Infrastructure design is built with the processing capacity for not just today, but the forecastable future. This is achieved through the use of high-performance hardware and network components, including multi-core processors, high-speed storage devices, and high-speed networking equipment. The design also includes a scalable and modular platform architecture to support future growth and expansion. The Performance goals of the System Infrastructure design are critical to maximizing system performance while minimizing wasted financial overhead on redundant, cold servers.

The following section elaborates in detail on the Performance goals of the System Infrastructure design.

Load Balancing:

The System Infrastructure design incorporates Load Balancing, which staggers services across multiple highly available nodes (servers) to spread the workload across systems, maximizing performance while minimizing wasted financial overhead on redundant, cold servers. Load balancing ensures that workloads are distributed evenly across all nodes, ensuring that no single node is overloaded, which can cause system performance to degrade. The design includes complex load balancing to maximize hardware investment and optimize system performance.

Virtualization:

The System Infrastructure design incorporates virtualization, enabling workloads to be added to the system with the click of a mouse, not a purchase order. This ensures that as workloads grow, resources can be added to the system, ensuring maximum performance at all times. Virtualization also ensures that the system is not constrained to the specs of a physical box, ensuring that the system can scale to meet future demands.

Key Technologies:

The System Infrastructure design incorporates several key technologies, including Microsoft Hyper-V, Microsoft SQL Server AlwaysOn Availability Groups, and Microsoft Failover Clustering. These technologies are designed to ensure maximum performance and uptime of the system, ensuring that the system is always available to support business operations.

**Expandability:**

The System Infrastructure concept is designed for when the above goals are not met, the project grows, or the requirements change. This is achieved through the use of a scalable and modular platform architecture that allows for easy integration of new data sources, hardware upgrades, and network expansion. The design also includes complex load balancing to maximize hardware investment and optimize system performance.

The Expandability goals of the System Infrastructure design are important to ensuring that the system can grow to meet future demands without increasing complexity.

The following section elaborates in detail on the Expandability goals of the System Infrastructure design.

Increasing Capacity without Increasing Complexity:

The System Infrastructure design incorporates Clustering of systems, which allows for easy addition of resources as needed with zero downtime, zero additional management and administration requirements, and minimal expense. Clustering ensures that the system can scale to meet future demands, without increasing complexity or cost, ensuring that the system remains easy to manage and operate.

Key Technologies:

The System Infrastructure design incorporates several key technologies, including Dell PowerEdge R740XD and R740XD2, Microsoft Hyper-V, Microsoft SQL Server AlwaysOn Availability Groups, Microsoft Failover Clustering, and Microsoft Storage Spaces Direct. These technologies are designed to ensure that the system can grow to meet future demands without increasing complexity or cost, ensuring that the system remains easy to manage and operate.

**Ease of Use:**

The System Infrastructure design is built using common Microsoft tools, a user-friendly web-based interface, and minimal complexity for ease of handover to the customer. This is achieved through the use of standardized software components and open-source frameworks. The design also includes comprehensive documentation and training materials to support ease of use and knowledge transfer.

The Ease of Use goals of the System Infrastructure design are to ensuring that the system is easy to manage and operate, even for users without a background in IT.

The following section elaborates in detail on the Ease of Use goals of the System Infrastructure design.

Windows Admin Center:

The System Infrastructure design incorporates the use of Windows Admin Center, which provides a single web interface for management and administration of all systems, services, and processes. This interface enables users to manage SQL databases, manage Hyper-V, monitor system performance, manage storage, or launch remote PowerShell or Remote Desktop, all from a single screen. Windows Admin Center ensures that the system is easy to manage and operate, even for users without a background in IT.

Key Technologies:

The System Infrastructure design incorporates several key technologies, including Microsoft Windows Admin Center, Microsoft Hyper-V, Microsoft SQL Server AlwaysOn Availability Groups, Microsoft Failover Clustering, and Microsoft Storage Spaces Direct. These technologies are designed to ensure that the system is easy to manage and operate, even for users without a background in IT.

**Built with the Future in Mind:**

The System Infrastructure design is built with the future in mind. This is achieved through the use of emerging technologies, such as cloud computing, edge computing, and AI. The design also includes a strong focus on exceeding industry standards related to system security, data privacy, and regulatory compliance.

The System Infrastructure design is an important component of digital transformation in the metals industry. The design concept is built around the key goals of reliability, performance, expandability, and ease of use. The design is also built with the future in mind, incorporating emerging technologies and a strong focus on security and regulatory compliance. The next section of the paper provides a detailed description of the key components of the System Infrastructure design.

System Infrastructure Design Concept:

The System Infrastructure design concept is built around the core principles of reliability, performance, expandability, and ease of use. In order to achieve these goals, the design incorporates the use of Microsoft’s Hyper-V as the hypervisor for all customer deployments. Hyper-V is chosen for its simplicity, abundance of online support resources, market share, third-party support from hardware vendors, and common Microsoft interfaces.

The following technologies are leveraged within the Hyper-V deployment to maximize resiliency and stability:

Storage Spaces Direct (S2D) storage replication: The storage pool for all virtual machines is spread across all physical nodes of the deployment, with full replicas of the entire pool existing on each node of the cluster. Real-time replication is achieved via 10/25/100Gbps networking, with a strong preference towards RDMA support. Benefits include no shared storage resulting in no single point of failure, expansion of storage array achieved in line with processing expansion, zero downtime expansion via adding new hosts to storage cluster, achieved via common Microsoft interfaces with no need for customers to learn and adopt new and foreign SAN technologies, and patching and maintenance achieved with zero downtime due to failover clustering.

Microsoft Failover Clustering: Multiple physical hosts providing highly available compute resources with failover clustering providing a highly available environment for hosting all virtual machines and file shares. Patching is achieved via customers’ existing protocols and standards, with options available to fully automate or manually control these tasks depending on customer preference.

Windows Server Datacenter licensing: A core licensing requirement for the deployment of S2D is Windows Server Datacenter Edition, which is not possible with any other edition of Windows Server. A benefit of this licensing requirement for the storage pooling is the ability to stand up unlimited virtual machines without additional licensing expense or administrative consideration. The customer may also opt to provide their own licensing via an existing Microsoft EA licensing agreement, however, this is often more expensive.

Key Requirements:

Some key requirements for a successful Hyper-V cluster deployment include stable, resilient, highly available Active Directory access, stable, reliable, robust network infrastructure, and backup infrastructure. Access to multiple domain controllers is a hard requirement for the deployment, and a loss of AD access will result in possible asynchronous operation and possible data corruption. It is advisable to defer to the customer for their preferred network ecosystem as they will be heavily involved in the network deployment and should be comfortable. SMS should only provide the required capacity and technologies, and approve of any new hardware purchases before orders are placed. It is highly preferred to leverage existing customer backup infrastructure for the backup of all SMS resources. SMS works closely with every customer to review, validate, and implement every system to match or exceed the customers’ existing security and regulatory standards.

**Data Infrastructure**

The steel industry relies heavily on data to optimize operations and ensure that the steel produced meets the required standards. In this chapter, we will explore the different data sources available in a steel manufacturing plant, including PLCs, vibration sensors, defects data from surface inspection systems, and other real-time sensor and IoT data. We will also discuss the types of data available in a steel plant, such as process data, production data, manufacturing data, maintenance data, downtime data, work order data, and workshop data.

Understanding the various data sources and types of data available in a steel plant is crucial for designing a data infrastructure that supports efficient data management and processing. By leveraging this data, steel plants can improve their operations, reduce downtime, and increase productivity. In the following sections, we will delve deeper into the different data sources and types of data available in a steel plant and discuss how to effectively manage and process this data to gain valuable insights into the operations of the plant.

**Data Sources:**

The steel industry is a complex manufacturing process that relies on the collection and analysis of vast amounts of data to ensure the production of high-quality steel. Data can be collected at various levels within a steel plant, ranging from sensor-level data collected by PLCs and vibration sensors, to data collected at the plant and enterprise level by MES and ERP systems.

At the sensor-level, PLCs are used to control and monitor the various machines and processes within the plant, collecting data on temperature, pressure, and flow rates, which can be used to identify any issues that may arise. Vibration sensors are used to monitor the vibration levels of the machines, and defects data from surface inspection systems can be used to identify issues with the steel-making process and improve the quality of the steel.

Different types of data are also available at various stages of the manufacturing process, including process data, production data, manufacturing data, maintenance data, downtime data, work order data, and workshop data. Process data can be used to monitor the performance of the steel-making, casting, and rolling processes and identify any issues that may arise. Production data can be used to track the productivity of the plant and identify areas for improvement, while manufacturing data can be used to monitor the performance of the machines and equipment and identify maintenance needs.

Data is also collected at each production unit, from the scrap yard to the coil yard, to ensure the production of high-quality steel. Data collected in the scrap yard can be used to determine the quality and composition of the steel being produced, and data collected in the furnace can be used to ensure that the steel is heated to the correct temperature and maintained at the appropriate pressure and flow rate. In the caster and rolling mill, data is collected on the thickness and quality of the steel to ensure that it meets the desired specifications, while data from finishing lines is used to monitor the surface quality and dimensions of the finished steel.

Figure 1 provides a visual representation of the different data sources available in a steel manufacturing plant. The figure shows how data is connected at each level of the plant, from the sensor and actuator level up to the enterprise level. The figure also illustrates the various types of data available, including process data, production data, maintenance data, and more.

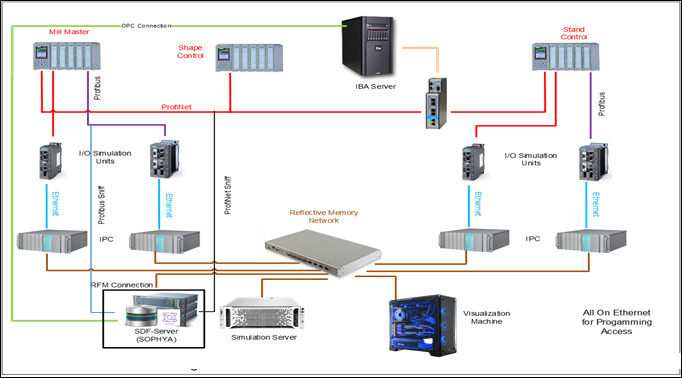


Figure 1 Various Data Sources in the steel industry

The steel industry relies heavily on data to monitor and improve the performance of the plant and ensure that high-quality steel is produced. By collecting and analyzing data at various levels and stages of the manufacturing process, steel plants can make adjustments to improve their operations, reduce downtime, and increase productivity.

**Data Acquisition**

Data acquisition is crucial in the steel production industry as it allows for efficient and accurate monitoring and control of the production process. By collecting data on various aspects of the process, such as temperature, pressure, and chemical composition, manufacturers can optimize their processes to improve efficiency and reduce waste. Additionally, data acquisition can be used to predict and prevent equipment failures, leading to reduced downtime and maintenance costs. Overall, data acquisition can help steel producers improve the quality of their products, increase productivity, and reduce costs. With the availability of a wide range of sensors and monitoring equipment, specialized software and systems, data acquisition can aid process control and improve equipment reliability. Furthermore, it enables traceability of the steel production process which is important for quality control and compliance with industry standards. Thus, data acquisition plays a critical role in the steel production industry, providing manufacturers with the necessary information to make informed decisions, optimize their operations, and achieve their business goals.

However, with the large amount of data generated from various sources, it can be challenging for manufacturers to efficiently manage, store, and analyze the data. Often, data is scattered across different systems and formats, making it difficult to integrate and analyze the information in a unified way.

To address this challenge, a proposed solution is to implement a data factory in the steel production industry. A data factory is a centralized platform that collects, processes, and analyzes data from various sources in a standardized and streamlined manner. This platform can enable manufacturers to integrate data from different systems and formats, providing a unified view of the production process.

By leveraging the capabilities of a data factory, steel producers can efficiently manage their data, gain deeper insights into their production processes, and make more informed decisions. This can lead to improved product quality, increased productivity, and reduced costs.

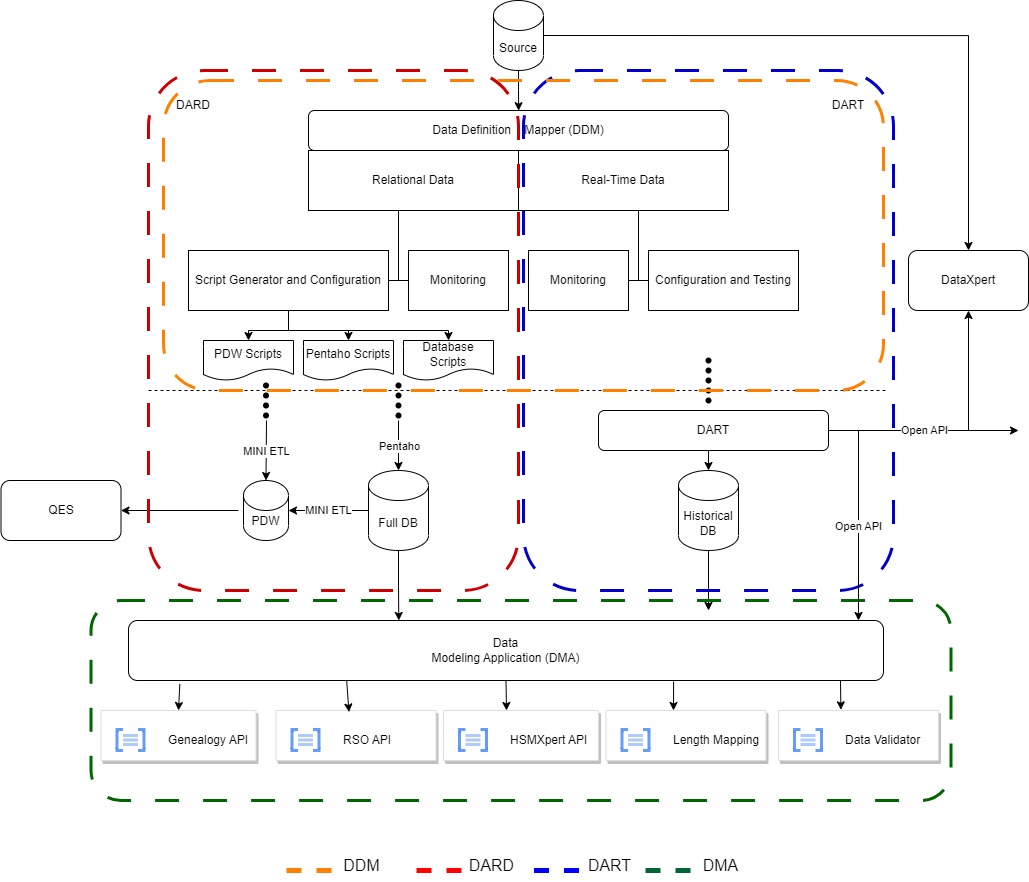
**Architecture of the Data Factory**

Figure 2 Data acquisition and structure

Data factory is a centralized platform that is used to manage and automate various data integration, transformation, and workflow tasks. It can be used to move and process data between different sources and destinations, such as databases, cloud storage, and big data systems. The goal of data factory is to simplify and automate the movement and processing of data so that it can be easily accessed, analyzed, and used by different systems and applications.

It helps steel manufacturing organizations to manage and automate the movement and processing of large amounts of data, making it more accessible, accurate, and useful for various systems and applications throughout the entire steel production process.

Data transformation is also an important aspect of Data factory in the steel industry. Data can be transformed into different formats or structures, such as cleaning, filtering, or transforming data to be used by specific systems or applications in the steel production process. This allows for the data to be easily understood and used by different systems and applications. Data pipeline is also an important component of Data factory in the steel industry. It allows for the creation of data workflows or pipelines that automate the movement and processing of data between different systems and applications in the steel production process. This ensures that data is accurate, up-to-date, and easily accessible to those who need it. Data monitoring and management is a key component of Data factory in the steel industry. It allows for the monitoring and management of data factory tasks and workflows, such as scheduling, monitoring, and error handling in the steel production process.

Key components that make up the Data Factory:

**Data Integration:** This component allows for the integration of data from different sources, such as databases, cloud storage, and big data systems, into a single, unified view. This allows for better decision-making and improved efficiency in the data processing tasks.

**Data Transformation:** This component allows for the transformation of data into different formats or structures, such as cleaning, filtering, or transforming data to be used by specific systems or applications. This allows for the data to be easily understood and used by different systems and applications.

**Data Pipeline:** This component allows for the creation of data workflows or pipelines that automate the movement and processing of data between different systems and applications. This ensures that data is accurate, up-to-date, and easily accessible to those who need it.

**Data Monitoring and Management:** This component allows for the monitoring and management of data factory tasks and workflows, such as scheduling, monitoring, and error handling. This ensures that data is accurate, up-to-date, and easily accessible to those who need it.

**Data Governance:** This component allows for management, security, and compliance of the data throughout the entire data factory. This ensures that data is protected and used properly and in compliance with the regulations.

**Data Quality:** This component ensures that the data is accurate and reliable, by identifying, monitoring, and repairing errors in the data, and also by profiling and standardizing data.

**Metadata Management:** This component allows for the management of data about data, such as data lineage, data dictionary, and data catalog.

**Self-service Data Preparation:** This component allows business users to prepare their data for analysis by providing the tools to clean, merge and transform the data without any technical knowledge.

These components work together to provide a complete data factory solution that can automate and manage the entire data processing pipeline.

Data Integration from Relational data sources.

The process of integration data from Relational data sources to gather and analyze data is an essential part of data factory. The system is designed to be scalable and flexible, meaning it can keep up with the growing demands of data and the changing business requirements.

The system is composed of various components that work together to collect and process data from relational data sources. These components include data connectors, data processing, data management, and data security.

Data connectors are responsible for connecting to various relational data sources, such as databases and spreadsheets, and extracting data from them. The connectors can be configured to connect to different types of relational data sources like Sybase, MSSQL, PostgreSQL, Oracle and etc.

Data processing is another critical component of the system that allows it to handle large amounts of data quickly. This component includes features like data filtering, data transformation, and data validation. These capabilities help ensure that the data is accurate and consistent.

Data management is also a critical component of the system. It includes data indexing and data tagging, which enable the data to be easily searchable and accessible. This helps improve the efficiency of the system by making it easier to find and retrieve the required information.

Data security is an essential feature of the system. It includes features like encryption, authentication, and access controls to safeguard the data from unauthorized access. This ensures that the data is protected and only accessed by authorized personnel.

Data Integration from Real-Time data sources.

The Real-Time data ingestion system aims to connect to various time-series and real-time data sources and pull data in near real-time. The system is designed to be scalable and flexible, meaning it can keep up with the growing demands of data and the changing business requirements.

The system is composed of several key components that work together to collect and process data in near real-time. These components include data connectors, data processing, data storage, data management, and data security.

Data connectors are responsible for connecting to various data sources, such as sensors and machines, and pulling data in near real-time. The connectors can be configured to connect to different types of data sources like OPC-UA, Modbus, and MQTT.

Data processing is another critical component of the system that allows it to handle large amounts of data quickly. This component includes features like data filtering, data transformation, and data validation. These capabilities help ensure that the data is accurate and consistent.

Data storage is another key component of the system that allows it to store large amounts of data for later analysis. The system includes data storage capabilities such as a data lake or time-series database. The data storage can be done in different storage options such as on-premise, cloud-based, or hybrid.

Data management is also a critical component of the system. It includes data indexing and data tagging, which enable the data to be easily searchable and accessible. This helps improve the efficiency of the system by making it easier to find and retrieve the required information.

Data security is an essential feature of the system. It includes features like encryption, authentication, and access controls to safeguard the data from unauthorized access. This ensures that the data is protected and only accessed by authorized personnel.

The system works by connecting to various data sources and pulling data in near real-time. The data connectors are configured to connect to different types of data sources such as OPC-UA, Modbus, and MQTT. The system then uses data processing capabilities to filter, transform, and validate the data, ensuring that the data is accurate and consistent. Once the data is processed, it is stored in a data lake or time-series database for later analysis. The system also includes data management capabilities, such as data indexing and data tagging, which allow the data to be easily searchable and accessible.

Data Modeling:

Data Modeling is an essential component of the relational data ingestion and real time data ingestion systems, which serves to consume data from these systems and create data models that describe specific objects or processes in the steel production process. The DMA (Data Modeling Application) serves as a data model layer and is responsible for data modeling and data management.

The DMA plays a critical role in data modeling and data management, as it takes the raw data pulled from various data sources by the data ingestion systems and processes it to create data models that describe specific objects or processes in the steel production process. This includes functionalities to store, manage and update these data models. The data models created by the DMA can include process flow, production data, and inventory data, among others.

The DMA consumes data from the data ingestion systems, filters it, transforms it, and validates it to ensure the data is accurate and consistent. The data is then used to create data models that describe specific objects or processes in the steel production process.

The data modeling process of the DMA involves several techniques, including statistical analysis, data mining, machine learning, and other data modeling techniques. The process includes several steps, such as understanding the data, cleaning and preprocessing the data, selecting the appropriate modeling technique, training the model, and validating the model.

The benefits of having data models of the steel production process are many. They enable organizations to understand their production process better, identify bottlenecks, and optimize the process. This can help improve efficiency and productivity, reduce costs, and improve decision-making. Additionally, the data models can be used to make predictions about future product performance and identify potential issues before they occur, which helps in reducing downtime.

The DMA includes features like genealogy, length mapping, product segmentation in process data, campaign, data validation, and aggregated data. These features help in creating accurate and useful data models that can be used for various applications.

Overall, the DMA is an integral component of the data ingestion system, providing data modeling and management capabilities. The DMA plays a crucial role in creating data models that describe specific objects or processes in the steel production process, which can help improve efficiency and productivity, reduce costs, and improve decision-making. The DMA includes several features that help in creating accurate and useful data models.

Data Dictionary:

The Data Dictionary component of the DDM (Data Definition Manager) is a critical tool for customers and end-users who need to manage their data effectively. It acts as a data catalog, providing a comprehensive overview of the data tables and columns. The data dictionary contains metadata that describes the structure of the data, including the names of the tables and columns, data types, sizes, and constraints. The data dictionary also provides information on the relationships between the tables and columns, making it easier to understand the structure of the data and how it is organized.

The data dictionary plays a crucial role in data management by helping users and customers to organize and understand their data effectively. By providing a comprehensive overview of the data structure and metadata, the data dictionary makes it easier to search for and find the data that is needed. The data dictionary also helps to ensure consistency and accuracy in the data by providing information on data types, sizes, and constraints, which can help prevent errors and inconsistencies in the data.

In addition, the data dictionary can help to improve collaboration and communication between data scientists, application developers, and other stakeholders. By providing a common language and understanding of the data, the data dictionary can help to facilitate discussions and ensure that everyone is working from the same understanding of the data.

Monitoring and Validation:

Monitoring and validating data replication, data accuracy, and data completeness are critical features of the Data Factory. These features enable data scientists and application developers to work with the right data, which is essential for making informed business decisions.

Monitoring data replication is important because it ensures that the data in the target database is up to date with the source database. By monitoring data replication status, it allows businesses to quickly identify and resolve any issues that may arise. This feature ensures that the data is correct and ready for processing.

Data accuracy is another essential feature of Data Factory. By validating data for accuracy, the system ensures that the data is correct and ready for use by data scientists and application developers. This validation ensures that the data is free of errors and inconsistencies, which is crucial for accurate analysis and reporting.

Data completeness is equally important as accuracy, as it ensures that all necessary data is available for analysis. By checking data completeness, the system ensures that all required data is present and available for analysis. This feature saves time and effort, as data scientists and application developers do not need to spend additional time and resources searching for missing data.

**CONCLUSIONS**

The proposed approach of implementing a data factory system infrastructure and data structure in the steel industry has the potential to revolutionize the way data is collected, managed, and utilized in the steel production process. The data factory can provide a centralized platform for the automation and management of various data integration, transformation, and workflow tasks. This can simplify the movement and processing of data and make it more accessible, accurate, and useful for various systems and applications throughout the entire steel production process. By leveraging data factory technology, steel manufacturers can integrate data from different sources into a single, unified view, transforming it into different formats or structures to be used by specific systems or applications in the steel production process. The data pipeline created by the data factory can automate the movement and processing of data between different systems and applications, ensuring data is accurate, up-to-date, and easily accessible to those who need it. Data monitoring and management in the data factory can help identify and prevent issues, allowing for quick responses and continuous improvement. The traceability of the steel production process is important for quality control and compliance with industry standards, which can be achieved through data management and analysis provided by the data factory.

The outlook for the implementation of a data factory in the steel industry is promising. The data factory can lead to improved efficiency, better decision-making, and reduced costs. It can provide steel manufacturing organizations with the tools needed to access, analyze, and use data in a more effective way, ultimately leading to improved overall performance and competitiveness in the global steel market.

However, the implementation of data factory in the steel industry requires significant investment in terms of technology, infrastructure, and human resources. Steel manufacturers must be willing to embrace data-driven decision-making and invest in the necessary resources to make it a reality.

**REFERENCES**

1. Bischoff, J., & Kinkel, S. (2019). Digitization in the Metals Industry: A Literature Review. Procedia CIRP, 84, 482-487. doi: 10.1016/j.procir.2019.04.250
2. Hu, S., & Han, Z. (2019). Data-Driven Optimization for Metals Manufacturing Processes: A Review. Journal of Manufacturing Science and Engineering, 141(8). doi: 10.1115/1.4043272
3. Moccia, S., Di Lorenzo, R., & Teti, R. (2019). Industry 4.0 in the metals sector: a review of the main application scenarios. Journal of Manufacturing Systems, 50, 15-27. doi: 10.1016/j.jmsy.2019.02.005
4. Wang, X., Liu, W., & Wang, Y. (2020). Intelligent Maintenance of Steel Plants Based on Edge Computing and AI. IEEE Transactions on Industrial Informatics, 16(7), 4878-4887. doi: 10.1109/TII.2019.2965127
5. Bischoff, J., & Kinkel, S. (2019). Digitization in the Metals Industry: A Literature Review. Procedia CIRP, 84, 482-487. doi: 10.1016/j.procir.2019.04.250
6. Wang, X., Liu, W., & Wang, Y. (2020). Intelligent Maintenance of Steel Plants Based on Edge Computing and AI. IEEE Transactions on Industrial Informatics, 16(7), 4878-4887. doi: 10.1109/TII.2019.2965127
7. Yin, S., Ding, X., & Zhou, X. (2019). Predictive maintenance for CNC machine tools based on improved LSTM. Journal of Manufacturing Systems, 50, 60-67. doi: 10.1016/j.jmsy.2019.04.002