DataOps for Product Information Management:
A study of adoption readiness

Nguyen Thi Thanh Phuong

Master’s Thesis
Big Data Analysis Program
2021

MASTER’S THESIS	
Arcada University of Applied Science	
Degree Programme:	Big Data Analysis Program
Identification number:	8668
Author:	Nguyen Thi Thanh Phuong
Title:	DataOps for Product Information Management: A study of adoption readiness
Supervisor (Arcada):	Magnus Westerlund
Commissioned by:	

Abstract:

Data Operations (DataOps) is currently being introduced in software-intensive companies, but there are not many companies that have fully adopted DataOps. DataOps is a process-oriented methodology that is people-driven rather than technology-driven. DataOps provide a best practice for data orchestration, automation, and collaboration, that aims to improve productivity and continuous assurance.

The thesis will study the adoption readiness for DataOps in product information management for a cleantech company. The thesis explores and details common problems in data management, such as misinformation, misuse, copying-pasting errors, duplicate, miscommunication, and manual work fatigue. Product information management solutions are often plagued by inconsistencies, user-friendliness issues, multiple document variants, and unclear versioning.

The aim of the thesis is to assist the company to define an effective product information management system. Further, the thesis will detail current challenges and elaborate on how DataOps can be adopted for product data management. The study was conducted through inductive qualitative interviews with eight experts in different teams in the company. The results were obtained by identifying the common points of view among interviewees. The research results are validated by discussing with the experts and by using Natural Language Processing (NLP) modeling for determining commonalities in interview data that provide some objective reasoning.

The insight gained is that product information management should be built using modular standard products and to be built using a numbering scheme that assists in finding the family product number that refers to parent numbers, which may have the same child items or assembly numbers. The dataflow design needs to be shared and implemented across both the vertical and horizontal organization. The scope of work and sales order data can be created manually, the other shared common data should be generated automatically from a central repository or integrated platforms of different teams. The study opens a new opportunity to increase the awareness of data management, tools, and platforms, which can be delivered to end-users through video training, in-class training, feedback forums, and Q&A channels.

Keywords:	DataOps, product information system, product portfolio management, dataflow design, modular product, automation, data orchestration, collaboration
Number of pages:	80
Language:	English
Date of acceptance:	

ACKNOWLEDGEMENTS

This master’s degree thesis has been accomplished in recognition of lifelong learning and my new journey into data analysis.

Firstly, I express my deepest gratitude to Prof. Magnus Westerlund who has trusted my competence and given me a chance to study the specialization program in big data analysis. He is always patient to guide me back on track whenever I was lost during the studying time for this thesis. I would like to give very special thanks to Mr. Andrey Shcherbakov who accesses my thesis and provides me valuable inputs to improve my thesis.

Secondly, I am deeply grateful to Mr. Timo Kallionpää who is my supervisor and has always supported me to complete my work and study my thesis for the data operations of the company at the same time. I would like to express my appreciation to Mr. Juha Säreströniemi who has explained to me the project management and knowledge of the company from his intensive working experience.

This thesis would also not be accomplished without support from the group of experts. I truly appreciate valuable discussions during the interviews with Mr. Graham Mackie, Josh Claassen, Heikki Kolisoja, Jori Arrakoski, Sami Niemelä, and Mrs. Maija Casey;those who have promptly and professionally shared knowledge about the work process of the company case in this study. I sincerely thank Mr. Garold Murdachaew for sharing expertise opinion, colleagues from groups of HR and Spinnaker project team for chatting me about things besides work.

I am sincere gratitude to my best friend Daniela Lopes for spending her time reading my thesis and encouraging me to complete my thesis.

Finally, I greatly appreciate my husband - Thao Pham and my son - Dang Pham, who are inseparable parts of my life, and have given me the strength to fulfill the research work and keep balance between work and life. I am thankful to my families in Vietnam who are in good self-care during the pandemic so that I can concentrate on the thesis.

CONTENTS

Figures

Tables

Abbreviations

BoM	Bill of Material
CRM	Customer Relationship Management
ERP	Enterprise Resource Planning
NLP	Natural Language Processing
PLM	Product Lifecycle Management

1 INTRODUCTION

1.1 Background

Data is nowadays accessed and used for more than a decade by Government bodies, companies, academic institutions, and citizens (Demchenko et al., 2018). Data is a valuable asset for many companies, and high-quality data can help decision-making (Munappy et al., 2020). According to Demchenko et al. (2018), the growing volume of data with economic digitalization can help organizations to capture and extract more valuable information. The paper mentioned good data is satisfied with the principles of Findable - Accessible- Interoperable – Reusable. In addition, the value of data complies with several characteristics, including sharable, accurate, integrated with other information, not depletable, and others (Demchenko et al., 2018).

There is a trend of increasing smart data handling to enable the potential of a data-driven economy and technology integration (Demchenko et al., 2018). A novel automatic way is required to operate data drastically changing in volume, velocity, and variety (Mainali et al., 2021). From the paper of Demchenko et al. (2018), data processing and data analytics need to be supported by scalable distributed data storage and big data infrastructures, such as by cloud-based providers like Amazon Web Services (AWS), Microsoft Azure, and other providers. DataOps or Data Operation is a good practice for data management to increase the speed and accuracy of data analytics in software industries (Rodriguez, de Araújo and Mazzara, 2020). On the one hand, DataOps is recently introduced to shorten end-to-end data analytic life-cycle time (Munappy et al., 2020). On the other hand, DataOps is developed based on DevOps and Agile methodology from the software industry, and Lean Manufacturing (Mainali et al., 2021). The paper of Mainali et al. (2021) showed the goals of DataOps are continuous integration and continuous delivery. There are three important elements of DataOps,automation, orchestration, and collaboration (Munappy et al., 2020). In addition, DataOps focuses on emerging cross-function collaboration to maintain a data pipeline for data-driven decisions efficiency, promote reusability to reduce operational costs, and rapid response to markets (Sahoo and Premchand, 2019). Moreover, DataOps manage the interaction between the data analytics team, the operations team, and the customers to effectively generated value for the business and continuously satisfy customers' needs (Rodriguez, de Araújo and Mazzara, 2020). Furthermore, DataOps is a people-driven practice into an effective team collaboration rather than technology-driven practice (Rodriguez, de Araújo and Mazzara, 2020).

From the study of Rodriguez, de Araújo and Mazzara (2020), DataOps requires an enterprise-grade platform capable of scalability, availability, and reliability, that can help to reduce the time and cost of copying and moving large datasets across different data silos. The study introduced four main components of a DataOps platform, including data pipeline orchestration; assurance automation, quality control, and monitoring; continuous deployment; and deployment of the data science model. In addition, the study explained the orchestrator to build a data pipeline is a software entity that manages the processes, executes data analytics steps like data gathering, access, integration, modeling, and visualization, and handle the exceptions. The assurance automation is to continuously monitor data quality, and automated tests of all the changes 24/7 (Rodriguez, de Araújo and Mazzara, 2020). Continuous deployment allows controlling the movement and continuous configuration of the code into a development environment (Rodriguez, de Araújo and Mazzara, 2020).

Moreover, Rodriguez, de Araújo and Mazzara (2020) discussed the implementation of DataOps in the software industry requires several steps, especially automated tests for every added new feature that does not disrupt any functionality of the system. The necessity of tracking and maintaining data artifacts is managed through a version control system mentioned in the paper. To ensure a smooth workflow, the code is updated and merged back to the trunk after making and testing changes (Rodriguez, de Araújo and Mazzara, 2020). To prevent data from being mixed up new data emerged, every developer should have a subset of the data to work on in their environment (Rodriguez, de Araújo and Mazzara, 2020). To boost productivity, the data analytics pipeline should be divided into smaller and reusable accessible components (Rodriguez, de Araújo and Mazzara, 2020).

On the other hand, manufacturing industries are continuously trying to improve operational efficiencies for creating more business insights that are beneficial and allow for the monetization of data (Sahoo and Premchand, 2019). From the paper of Sahoo and Premchand (2019), DataOps can help to eliminate data silos by connecting different data pipelines, and improve backlog management and data quality for automated and faster data pipelines. The paper brought an opinion of IoT (Internet of Things) devices like sensors are primary sources of data that first send data to a hub or are locally cached before being uploaded in batch into a central system. In addition, embedded products like connected and automated devices may require a high amount of computing power for data-driven decisions from enormous amounts of generated data (Sahoo and Premchand, 2019). DataOps has high potential if used correctly to bring a lot of business value to manufacturing organizations (Sahoo and Premchand, 2019).

However, there are few companies that have succeeded in adopting DataOps practices (Munappy et al., 2020). This thesis considers a case study of a cleantech company providing innovative and sustainable waste treatment solutions. There is a need for increasing spare parts after-sales for the company. The spare parts must be delivered at the right time with the right product. Consequently, it is required to have the right product information for spare parts. Accurate and consistent product data should be available in the system. In addition, it should be easy to find and access data that can help for a fast response to customers. The thesis will investigate a real case to study DataOps principles and evaluate the possibilities of the effectiveness DataOps for product information management.

1.2 The objectives and research questions

The thesis focuses on the descriptive aim to answer the following two research questions:

1. What are the characteristics of an effective product information management system for the studied company case?
2. What are the principles of data operation to manage product information for the studied company case?

The qualitative research will conduct in several interviews to get more insights into the data operation of a company. The qualitative analysis on one hand can be obtained by reflecting theoretical DataOps manifesto principles on the current situation of the studied company case. On the other hand, the interview transcriptions will be analyzed by Natural Language Processing (NLP) machine learning to obtain key words or topics discussed within the company.

1.3 Limitations

There are several potential limitations to the validity of this thesis. The study has been limited to one type of cleantech company that brings a narrow view of DataOps for other industries. The research was limited in the number of interviews. This opens the possibility of continued research in the future to increase the number of interviews with targeted interviews who gain a more proper understanding of DataOps possibilities. Moreover, the study needs to validate with more experts to get all possible aspects of DataOps adopted in the company.

2 LITERATURE REVIEW

2.1 Data management overview

Data is considered an organizational asset, but it differs from other assets in the way of data management (Munappy et al., 2020). Data can be classified by type of data, or data content, or data format, or data protection level, or data storage, or data access (DAMA, 2017). There are several characteristics of data summarised as the following statements: (DAMA, 2017)

- Data is not tangible and no longer treated as a 'by-product' of operational processes
- Data is not consumed when used, but it can be stolen without being gone because it is easy to copy and transport
- Data is not easy to reproduce when it is lost or deleted
- Data is dynamic, at the same time, the same data can be originated or used in different ways by multiple people at multiple places within an organization

Data management is complied with business-driven and IT-driven that aim to create data value. Data value is often temporal (i.g. changes over time) and contextual for one organization and not for others**.** In addition, a monetary value is putting data as another asset. Data management is controlled by data governance through data lifecycle management as the wheel data management framework in Figure 1. Moreover, data management activities include all core activities, especially foundational activities, lifecycle management activities, usage merging from lifecycle management activities, and data governance activities. (DAMA, 2017; Pulvirenti, 2020)

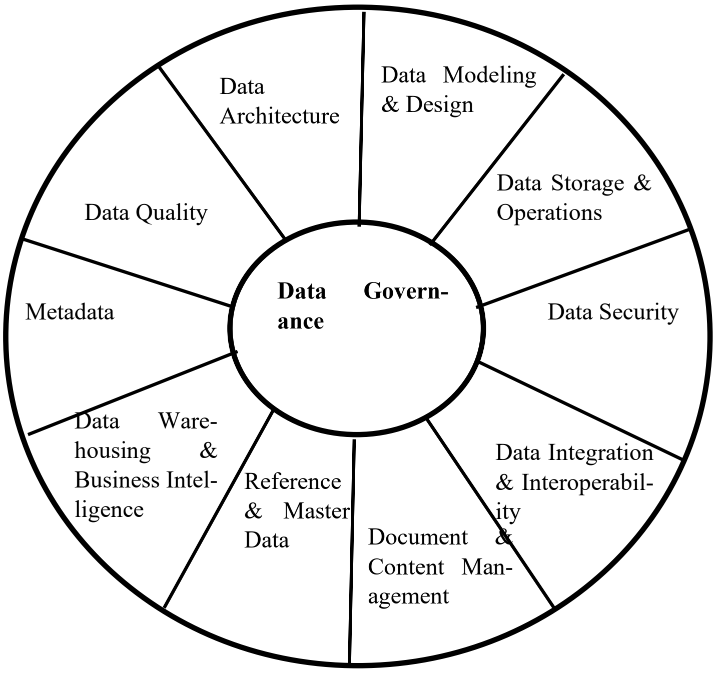

Data Architecture
Data Modeling & Design
Data Storage & Operations
Data Quality
Data Governance
Metadata
Data Security
Data Warehousing & Business Intelligence
Reference & Master Data
Data Integration & Interoperability
Document & Content Management

Figure 1. Data management framework (DAMA, 2017)

Data governance can guide all other data management activities managed properly at all levels according to policies and best practices. Data governance provides principles, policies, strategies, framework, metrics, and oversight. Data governance is developed based on organizational change management and mainly culture change or people behaviours. (DAMA, 2017, Pulvirenti, 2020)

Data modeling is the process of discovering, analyzing, and understanding data requirements for current and future business requirements. The proper data model can lower the support cost and the cost of building a new application. (DAMA, 2017)

Data security activities exist to reduce risk from external or internal threats. Data security aims to protect the proper authentication, authorization, access, audit, and the entitlement of data and information assets in alignment with stakeholder concerns, government regulation, and legitimate business concerns. There are several steps to implement the data security process, including identifying and classifying sensitive data assets, locating sensitive data throughout the enterprise, determining the level of protection, and identifying the interaction with business processes. (DAMA, 2017; Held et. al., 2016)

2.1.1 Data Architecture

Data architecture is fundamental to data management through creating and maintaining organizational knowledge about data and data movement systems. Data architecture can increase data value by identifying opportunities for data usage, cost reduction, or risk mitigation. In addition, data architecture deals with quality-oriented and innovation-oriented in a shorter-term perspective by using unproven business logic and leading-edge technologies. Moreover, data architecture practice is based on using data architecture artifacts or master blueprints including defining data requirements, reviewing data designs, determining data lineage impact, controlling data replication, enforcing data architecture standards, guiding data technology and renewal decisions; collaborating with various stakeholders; and establishing the semantics of an enterprise. There are two types of enterprise data architecture design, especially enterprise data model and data flow design. The enterprise data model contains a conceptual data model with a set of key enterprise data entities in subject area models as illustrated in an example in Figure 2. There are three subject area models, such as product design subject area, commercial offer subject area, and sales subject area. The relationship of each subject area needs to be controlled over subject area borders, which can cause miscommunication or misinformation between subject areas. Each entity in one subject area model should reside in only one subject area, but it can be related to entities in other subject areas. (DAMA, 2017; Mainali, 2020)

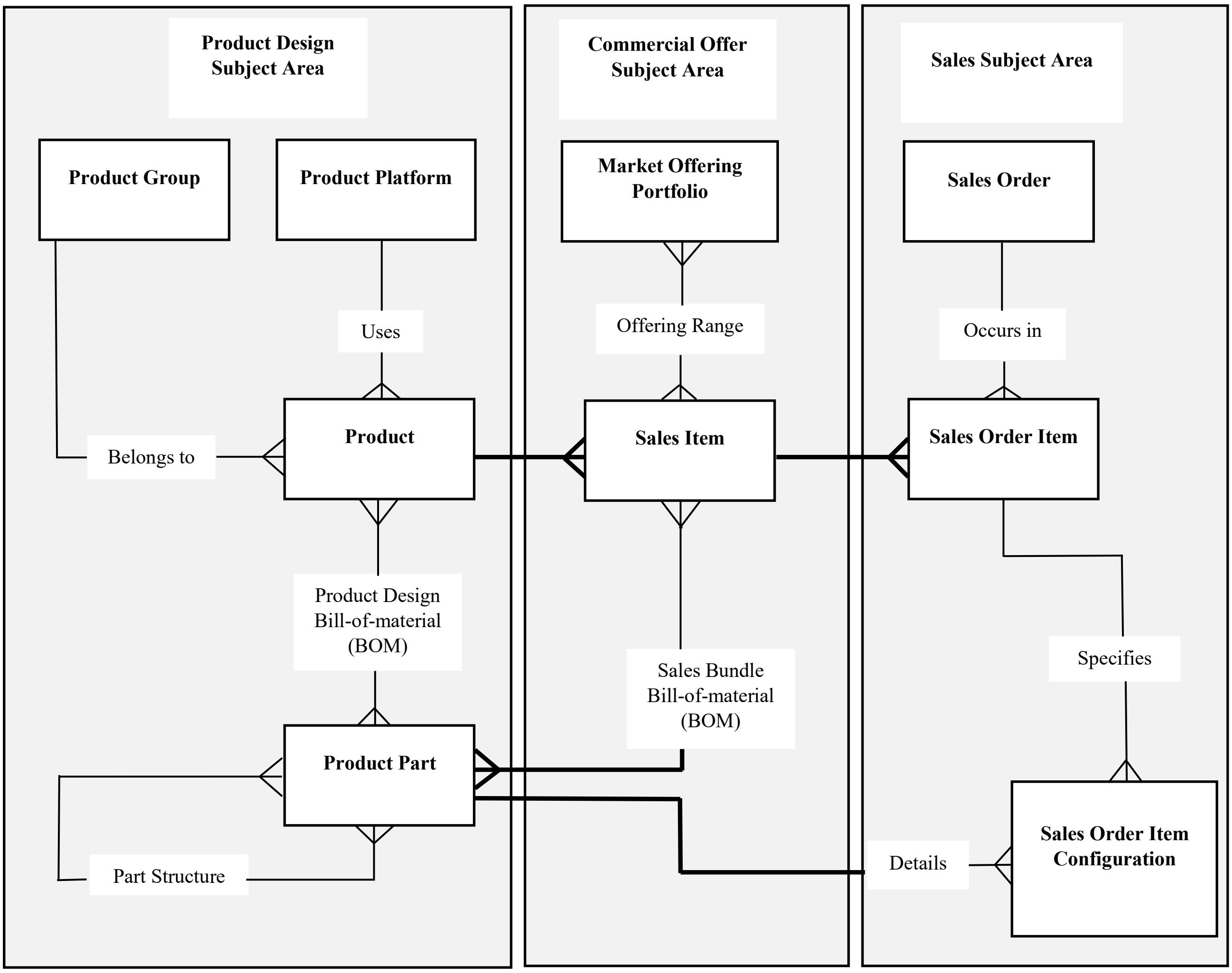

Product Design Subject Area
Commercial Offer Subject Area
Sales Subject Area
Product Group
Product Platform
Market Offering Portfolio
Sales Order
Uses
Offering Range
Occurs in
Belongs to
Product
Sales Item
Sales Order Item
Product Design Bill-of-material (BOM)
Sales Bundle Bill-of-material (BOM)
Specifies
Product Part
Part Structure
Details
Sales Order Item Configuration

Figure 2. The enterprise data model (DAMA, 2017)

Data flow can be designed like the diagram that depicts what kind of data flows between systems as shown in Figure 3. Data flows from the product design management system to manufacturing, to the sales department, to the customer, and an aftermarket system. (DAMA, 2017)

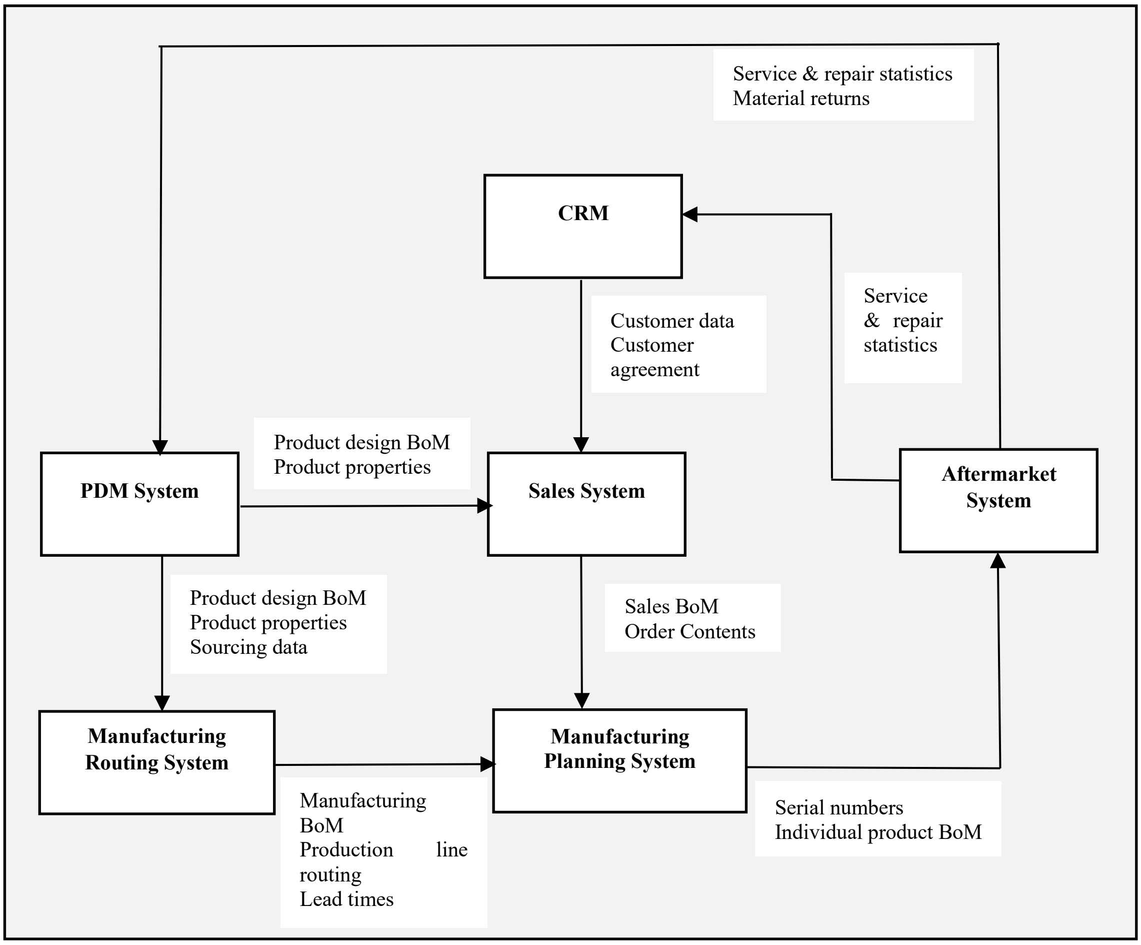


Figure 3. The data flow diagram design (DAMA, 2017)

A business-data-driven roadmap in Figure 4 illustrates the data dependencies of business capabilities. The lowest level of dependency is on the top with Product Management and Customer Management, and the highest level of dependency at the bottom is where Customer Invoice Management depends on Customer Management and Sales Order Management.

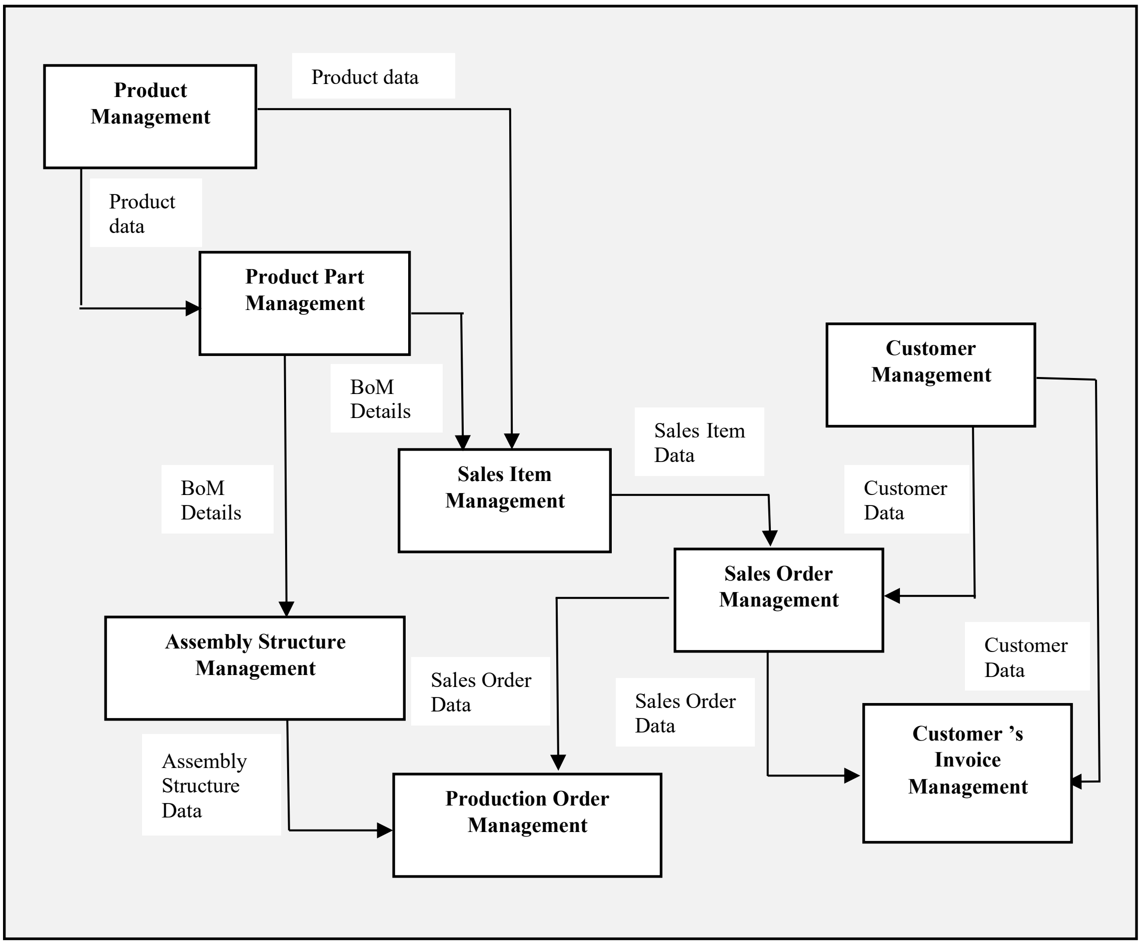


Figure 4. The diagram of data dependencies in data flow (DAMA, 2017)

2.1.2 Data Storage and Operations

Data storage and operations are to ensure business continuity and minimize the risk of disruption. Data storage and operations can manage the availability of data in a database and maximize its value throughout the data lifecycle from creation, acquisition to disposal. There are two sub-activities, including database operations support for activities related to the data lifecycle and database technology support with IT infrastructure. Database architecture types are centralized (i.e., single database) or distributed databases (i.e., federated, or autonomous database). Databases are organized in a way of more/less controlled structure, especially hierarchical, relational, non-relational schema. There are several specialized databases, such as object databases (e.g., computer-assisted design and manufacturing (CAD/CAM)), document databases (e.g., XML database for shopping-cart application for retail websites), or cloud databases. Data storage and operations include the design, implementation, and support of stored data. Another activity is to change data capture with two different methods, including data versioning and reading logs. In addition, data migration is to transfer data between storage types, formats, or computer systems with as little change as possible. Moreover, the principle of data management is to manage the cost of maintaining data that should not exceed its value to the organization. (DAMA, 2017)

2.1.3 Document and content management

Firstly, document and content management is the processes of planning, implementing, and controlling activities for data lifecycle management effectively and efficiently, including controlling the capture (i.e., layout, structure, logic, content), storage, access, retrieval, and use of data and information. Secondly, document and content management ensure integration capabilities between structured and unstructured content. The content management system is to collect, organize, categorize through index by keywords, structure information resources, retrieve contents, store, and maintain links between components or whole documents. Thirdly, document and content management needs to comply with legal and regulatory requirements and customer expectations regarding good records management for business continuity. The records management system in paper documents or electronically stored information addresses automation of retention and disposition (i.e., inactive documents transferred to off-site storage), e-discovery support, and long-term archiving. Digital asset management is to manage the storage, tracking, and use of rich media documents, by using optical character recognition or intelligent character recognition. In addition, the organization needs to include the overall corporate backup and recovery activities. Moreover, the principles of document and content management systems are mainly dependent on everyone’s responsibility in an organization and include accountability, integrity, protection, compliance, availability, retention, disposition, and transparency. Finally, key performance indicators are both quantitative and qualitative measures to review document and content management system performance with tangible benefits (e.g., productivity, cost reduction, information quality, and so on), and intangible benefits (e.g., collaboration, simplification of job routines and workflow) at the strategic and operational levels. (DAMA, 2017)

Reference and master data manage shared data and information assets in data structure and data values across business domains and applications within an organization. Master data management needs to meet organizational goals, provide an authoritative source of managing data quality, reduce risks, and reduce the cost of data integration. The availability and quality of master data can help to leverage transactional data and enterprise structure data for business activities. The master data management can manage business entity resolution (e.g., parties, customers, products and services, financial structures, legal matters, locations) to maintain entity instances consistently across systems. Reference data management is to control any data used to characterize or classify other data in an organization, or any data related to information external to an organization. Reference data includes hierarchies for parent and child relationships and needs updating manually. In addition, reference and master data require governance and stewardship to ensure completed, clear, and understandable data. Moreover, there are several activities for reference and master data management, including monitoring data movement, providing channels to receive and respond to requests for changes, and collaborating between multiple parties in data-sharing agreements. (DAMA, 2017; Held et. al., 2016)

Business intelligence designs a mechanism to describe the relationship between transactional level and operational level reports in an atomic data warehouse. The key principles to drive business intelligence monitoring are transparency and visibility. (DAMA, 2017)

The data warehouse is an integration process to get data from a range of sources into a common data form, location, and model. Data warehouses can reduce data redundancy, improve the consistency of information, and enable effective business analysis and decision-making for many purposes. The data warehouse is related to software programs used to collect, extract, clean, transform, load, and store data from other systems. In addition, data warehouses include processes that interact with Metadata repositories to make data accessible and usable for analysis. There are two types of data integration processes, especially historical loads, and ongoing updates. (DAMA, 2017; Held et. al., 2016)

Metadata is the data used to manage data creation, processing, and use. Metadata is essential for data management including information about data itself (e.g., databases, data elements, data models), the concepts the data represents (e.g., technical and business processes, application systems, software codes), the connection between data and concepts (e.g., data rules and constraints). The reliable metadata can identify private or sensitive data, manage the data lifecycle and movement through systems, meet compliance requirements, and minimize risk exposure. There are three types of Metadata, business, technical, and operational metadata. In addition, Metadata can be classified into several categories, like descriptive, structural, and administrative metadata. However, Metadata has several challenges that need to be addressed, like meeting cultural resistance in an organization, being a low priority in many organizations, or lacking Metadata standards in the exchange of data with operational trading partners. (DAMA, 2017; Held et. al., 2016)

2.1.4 Data Quality

Data quality needs to maintain data reliability and trustworthiness through data management. Data quality management can prevent wasting the effort of collecting, storing, security, and using data, reduce costs, improve efficiency, and mitigate risks. Data quality management focuses on the most important or critical data, such as regulatory and financial reporting, business policy and strategy, and ongoing operations. In addition, misunderstood and misused are common risks of using data. Data quality management requires changing organizational cultures and adopting a quality mindset. High data quality means that data is available, relevant, complete, accurate, consistent, timely, usable, meaningful, and understood. On the other hand, the low-quality data is inaccurate, incomplete, and out-of-date. Poor quality data is costly to any business, any organization, or company because it costs money to produce data. There are 15 dimensions across four categories of data quality, especially intrinsic data quality (i.e., accuracy, objectivity, believability, reputation), contextual data quality (i.e., value-added, relevancy, timeliness, completeness, appropriate amount of data), representational data quality (i.e., interpretability, ease of understanding, representational consistency, concise representation), and accessibility data quality (i.e., accessibility, access security). Data quality management goals are developing a data-governed approach; controlling data quality as part of the data lifecycle with standards, requirements, and specifications; implementing processes to measure, monitor, and report on data quality levels; advocating opportunities for data quality improvement. (DAMA, 2017; Held et. al., 2016)

2.2 DataOps overview

2.2.1 DataOps definitions

According to Munappy et al. (2020), DataOps definitions can be defined the combination of DevOps, Agile methodology, and Lean Manufacturing principles. The paper explained DevOps is to build a collaboration between Development and Operation teams to reduce the development lifecycle and fast delivery of high-quality systems. In addition, Agile methodology is to build a close collaboration with customers and quick response to the change in customer requirements (Munappy et al., 2020). Lean manufacturing is to reduce non-value tasks (Munappy et al., 2020).

According to Ereth (2021), DataOps is a set of best practices, processes, tools, and technologies with introduce automation in the data collection, validation, and certification process (Fig. 5). From the paper of Munappy et al. (2020), DataOps is to promote the culture of collaboration and continuous improvement (i.g. continuous integration and continuous delivery). Moreover, DataOps is an approach to eliminating data silos by connecting different data pipelines, especially value pipelines and innovation pipelines (Munappy et al., 2020). The data pipeline is built to minimize and eliminate the manual process of sequencing data processes throughout the data lifecycle (Mainali et al., 20220). Furthermore, DataOps is implemented based on a people-driven practice rather than a technology-oriented practice (Ereth, 2021).

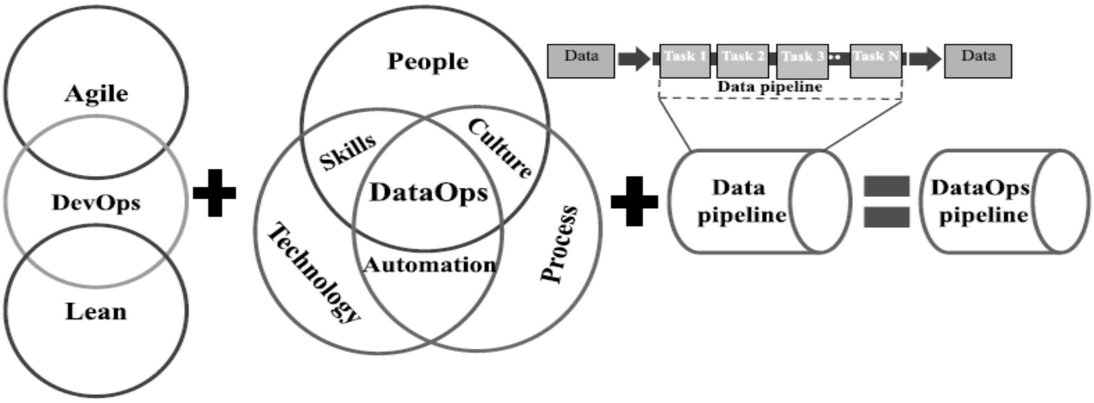


Figure 5. The schematic of DataOps pipeline (Mainali, 2021)

2.2.2 DataOps challenges

There is limited experience in implementing DataOps in a company. Several major challenges need to be addressed to data analytics or a company in general as the following below: (Rodriguez, 2020; Bergh et al., 2019; Mainali et al., 2021)

- A company has too many errors per month. Data errors from internal and external data sources are unavoidable. Data errors are subtle, such as duplicate records, and difficult to trace and resolve quickly. Repeated unfixed data errors can damage the reputation of the company and the customers' trust and prevent data pipelines from flowing correctly.
- A company is too slow to deploy changes into production. The goalposts keep moving, for instance, the user's needs require immediate responses, or the user fosters a continuous series of questions for new requests.
- Data after being collected through multiple devices and platforms, lives in silos, being stored in separate databases. The process of accessing and integrating data from these myriad sources is complex, lengthy, and subject to bottlenecks and blockages.
- Data in operational systems is usually not structured to optimize reads, aggregations, or easily understood by humans, for example, the file names contain descriptive names of the contents, and the tables contain intuitive connections of the contents.
- It requires a lot of effort and time to validate and verify changes, update, maintain, and assure the quality of the data pipeline.
- There are numerous manual processes on a regular basis for some data processes not to be automatized. It causes error-prone manual steps, labor-intensive, time-consuming, and tedious to reduce team productivity.
- A revision control can become a nightmare when copying data into many reports and then having to manage changes manually.

Consequently, many possible ways are to overcome these challenges, such as data under control, efficient automation to reduce duplication or impeccable data quality (Bergh et al., 2019). According to Bergh et al. (2019), DataOps processes and tools can harmonize a measured amount of centralization and automated data orchestration. For instance, it can establish a shared reality report if different teams share the same data, or the standard metrics can be implemented if data is under the control of one group like IT (Bergh et al., 2019). About a data supplier mentioned in the paper, the IT master data management team can be an external third party or an internal group. The master data management is to link all critical data of an organization to a common reference list (Bergh et al., 2019). The broader data users can utilize the useful master data at one centralization-innovation spectrum to offer the most centralization capabilities, the fastest innovation, and the right transition space in the center (Bergh et al., 2019).

2.2.3 DataOps benefits

The goal of DataOps is to minimize the overhead, mitigate non-value add tasks, and free up the team from significant manual effort. With automation, DataOps can enable 24x7 monitoring of the data pipeline and focus on new requirements of customers. With faster cycle time, automated orchestration, higher quality, and better end-to-end data pipeline visibility, DataOps can improve team communication and coordination to prevent data isolation due to different tools and workflows. The complexity of workflows presents like a directed-acyclic graph in Figure 6. (Bergh et al., 2019; Oracle, 2020)

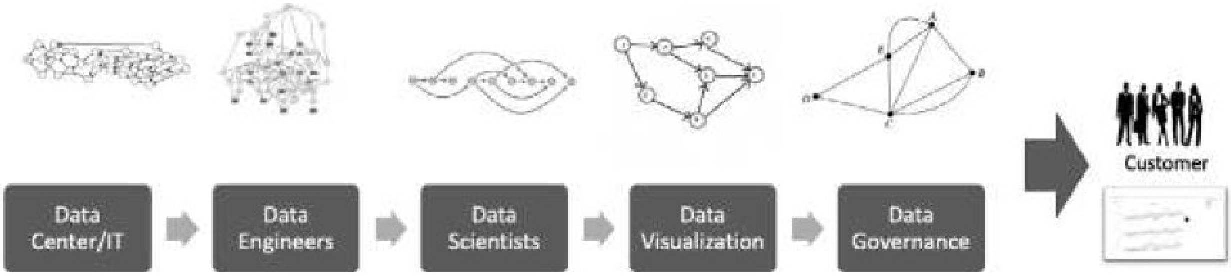


Figure 6. The directed-acyclic graph example about the complexity of workflows in an enterprise (Bergh et al., 2019)

DataOps can unify the data operations pipeline under one orchestrated workflow. Using DataOps can help to build a high relational coordination enterprise with several benefits for the workflow: (Bergh et al., 2019; Mainali et al., 2021)

- Robust: an impact review board ensures the new update does not disrupt critical operations.

- Transparent: a dashboard illustrates the status of new updates and the operational status of the data operational pipeline including automated alerts of issues.
- Efficient: a centralized automated orchestration of the end-to-end data pipeline minimizes manual steps.
- Repeatable: an automated revision control detects error and fault resilience of the data operations pipeline.
- Sharable: a services-oriented architecture encourages reuse for the team.

Moreover, DataOps improves teamwork with these benefits: (Bergh et al., 2019; Sahoo and Premchand, 2019)

- Emigrate easily from one team member to another, or from development to production
- Collaborate and coordinate work for effective teamwork with a compelling direction, strong structure, supportive context, and shared mindset
- Automated orchestration and reduce process variability and errors
- Maintain security with access control
- Re-use pipeline and components when developing new features
- Self-service to move forward without waiting for the official approval
- Gatekeepers allow everyone can have access to the available data
- Transparency for pipeline status and statistics

2.2.4 DataOps Manifesto

There are 18 principles of DataOps developed to get the value of analytics: (Bergh et al., 2019; Mainali et al., 2021; Mainali, 2020)

1. Continuous delivery of valuable analytics insights and satisfy customers
2. The valuable analytics insights are delivered together with accurate data and robust frameworks and systems
3. It is a necessary face-to-face conversation with customers to evolve the customer needs
4. It is a team sport for the analytic teams with diverse tools and skills
5. It needs daily interactions between customers, analytic teams, and operations throughout the project

6. The self-organized teams can get the best analytic insight, algorithms, architectures, requirements, and designs
7. It needs to create sustainable and scalable data analytic teams and processes
8. Self-reflecting at regular intervals is essential for checking operational performance
9. The analytic teams use a variety of individual tools to generate code and configuration that are to access, integrate, model, and visualize data
10. The beginning-to-end orchestration includes data, tools, code, environment, and the analytic team's work for analytic success
11. The version control is required to make reproducible results
12. It is required easy to create, isolate, safe, and disposable technical environments
13. Simplicity is essential to enhance agility by maximizing the amount of work not done
14. DataOps focuses on continuous efficiencies in the manufacture of analytic insight
15. The analytic pipelines can automatically detect abnormalities or security issues
16. It is required to monitor quality and performance for detecting unexpected variation
17. Reusing or avoiding the repetition of previous work can enhance the analytic insight manufacturing efficiency
18. It is essential to improve cycle times by minimizing the time and effort to turn a customer's need into an analytic action and release results

2.3 DataOps case studies examples

This section introduces two case studies of adopting DataOps in B2B software-intensive company for healthcare domain and a multinational network and telecommunications company. These studies demonstrate different steps and stages for an organization implementing DataOps.

2.3.1 DataOps adopted in B2B software -intensive company

Figalist (2021) studied the use of DataOps by various internal and external platform providers in the healthcare, industry, and advertising domain. The study introduced four stakeholders' phases when the companies adopted DataOps mindset. They presented the interaction of other stakeholders with a data scientist or data engineer, such as skepticism - proactive engagement, interest – explanation - inspiration, exploration/experimentation - inclusion, and collaboration - collaboration. In the first phase mentioned in the paper, the skepticism phase of stakeholders and other team members cannot see the value in data-driven ways of working, while the proactive engagement phase is to react to the stakeholders' skepticism by implementing a first proof of concept. In the second phase discussed in the paper, the stakeholders already show interest in adopting DataOps, and data scientist and data engineer need to provide explanations and inspiration during interaction with stakeholders. In next phase, Figalist (2021) explained stakeholders enter the exploration and experimentation phase to start their own ideas and give more feedbacks on existing analyses, whereas data team can begin to strengthen the inclusion stakeholders to involve actively in decisions or express their additional information needs. The last phase is collaboration for both stakeholders and data team acting as a team and commit a solid and permanent collaboration together or participate in regular and continuous feedback sessions (Figalist, 2021). However, stakeholders can start in any phase and do not need go through all the four phases (Figalist, 2021).

Moreover, the study of Figalist (2021) proposed a generic model was computed across three stages, including preparation, execution, and evaluation. The model was adaptable to each individual stakeholder for the adoption of DataOps mindset. The preparation stage of the model is required for available data on user-level from different data sources that are stored in the common data storage. The analysis stage of the model will be implemented on the customer level with the prepared use case-specific data. The execution stage of the model can run analysis after choosing a feasible method with the optimized parameter settings and selected input features. The sufficient quality of the results needs to be evaluated to derive insights before preparing and presenting to the stakeholders (Figalist, 2021). The processes including data collection, data processing, method selection, analysis execution, and analysis evaluation can remain the same across all stakeholder phases (Figalist, 2021). However, Figalist (2021) suggested the collaborative steps such as use case selection, and results presentation, depend on the stakeholders' mindset and organizational maturity of transitioning towards DataOps and data-driven ways of working.

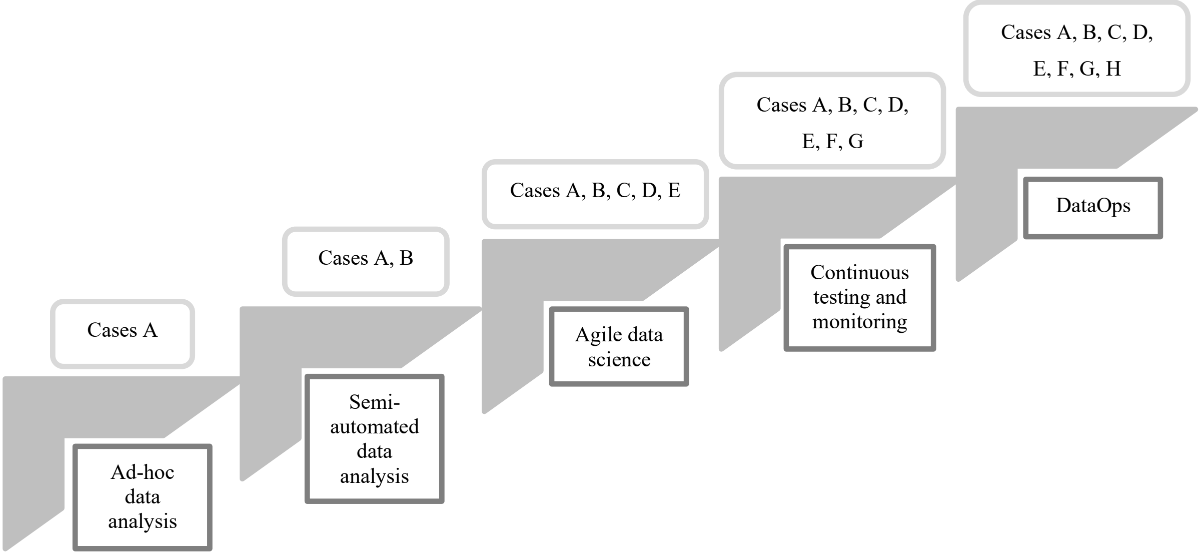
2.3.2 DataOps adopted in a multinational network and telecommunications company

Munappy et al. (2020) considered eight different studied cases (Table 1) for the investigation of DataOps approach in data collection at different sources throughout the semi-structured interviews of data scientists, data analysts, data engineers, . The studied company was Ericsson which provides services, software, and infrastructure in network and telecommunications technology (Munappy et al*,* 2020).

Table 1. The summary of data working processes from different teams in Ericsson company (Munappy et al, 2020)

Cases	**Activities**	**Benefits**
A	Automated data collection for data analytics	-Run test cases 24/7 -Automated alarm when test case fails
B	Building data pipeline	-Minimize human involvement by building data pipeline to get easier insights from raw data -The scheduler can control the execution of different stages of data pipeline
C	Toolkit for network analytics	-Automated monitor, analyse and troubleshooting networks to save man-hours and reduce consultant service -Produce reports for the customers automatically and enable new opportunities
D	Building continuous integration pipelines	-Get the data and the feedback data continuously from customers -Reduce the time of analytics, like predicting if a customer is going to return the product or when the customer needs to buy new products
E	Tracking the software version	-Shorten the cycle time towards the customers by using the feedback loop from the customers -Reduce debug a certain issue -Understand data easier by creating dashboards out of data
F	Testing the software quality	-Monitor features as expected or as designed -Upgrade features
G	KPI analysis software	-Automated collection of data from the nodes of continuous deployment zone for manual formular KPI -Help to turn KPI's analysis into informed business decision
H	Building data pipeline for continuous integration and continuous delivery	-Provide availability of high-quality data -Immediately inform a variation from the usual pattern and find the error to fix

According to this study, a five-stage evolution at different maturity levels can be defined before the introduction of DatOps (Fig.7). Munappy et al *(*2020) introduced each stage with different requirements and challenges to accelerate the production of high-quality data insights as the following summary.



Cases A
Ad-hoc data analysis
Cases A, B
Semi-automated data analysis
Cases A, B, C, D, E
Agile data science
Cases A, B, C, D, E, F, G
Continuous testing and monitoring
Cases A, B, C, D, E, F, G, H
DataOps

Figure 7. The stairway evolution model of DataOps (Munappy et al, 2020).

Notes. Cases from A to H according to Table 1

- Phase 1: Ad-hoc data analysis

  Activities: The reports or insights are created on-demand to answer very specific business analytics questions immediately. The ad-hoc analysis is highly dependent on the templates provided by the IT department and lets users deal with different data sources in a flexible and scalable way.

  Requirements: It needs technology to collect real-time data from multiple data sources.

  Challenges: Data silos storing data collected from different sources can prevent getting a full picture and lead to insufficient decisions making.

- Phase 2: Semi-Automated data analysis

  Activities: Data pipeline for collecting and processing a huge volume of raw data from both internal and external sources can implement data analytics in a more efficient and automatic way. Data technologies to manage the pipeline can be categorized into four technologies developed for data engineering, data preparation, data storage, and data visualization. Data engineering performs two different operations including data collection and data ingestion. Data preparation involves the preparation of metadata links to the path of data storage and the aggregation module all the links for different types of data by encoding or encrypting data. Data storage is to store metadata links in the database where teams can find and download raw data. Data visualization shows the performance of various requirements of the stakeholders, such as data cleaning, data filtering, data processing, and data transformation.

  Requirements: It is required a data process to control and coordinate data technologies and data pipelines.

  Challenges: It lacks data pipeline robustness due to still monitoring and fixing issues manually.

- Phase 3: Agile data science

  Activities: It is required to follow the agile methodology for delivering insights in short sprints and evolving customer requirements. The team works store in a common central repository in order to synchronize.

  Requirements: Data team and customers should interact directly, and customers should be delivered their demands frequently.

  Challenges: The flow of the data pipeline should be monitored continuously, and good quality data should be made available.

- Phase 4: Continuous testing and monitoring

  Activities: With an automated monitoring mechanism, it can help to identify the reason for the unexpected output. With automated alerts, it can help to lead proper measures to the broken pipeline.

  Requirements: It is required to have an automated continuous testing and monitoring mechanism and automated alerts, that can detect the problems immediately and handle pipeline breakage.

  Challenges: There are several ways to manage and orchestrate the pipeline characteristics, also the mechanism of accelerating new data analytic ideas into the existing value pipeline.

- Phase 5: DataOps

  Activities: Automation, orchestration, collaboration are the most important elements of DataOps. The output of the previous task becomes the input for the next task in the data flow. In addition, the execution of steps in the data flow occurs in a specific sequential order. It is essential to organize all the people who work on data as a team to reduce data silos and allow people to know what the other team is doing.

  Requirements: Regarding massive download data from the same place, there is a barrier for the existing pipeline to serve a larger number of data requests. Continuous integration and continuous delivery practices need to be implemented including orchestration, advanced automation, agile practices, monitoring, and controlling the entire data life cycle process.

  Challenges: Organizational restructuring is a major challenge of DataOps because of lacking available skilled team that can learn new tools and technologies. In addition, data silos need to be handled properly.

3 RESEARCH METHODOLOGY

3.1 Exploratory case study

The company for the case study is one of the leading providers of cleantech solutions for the marine industry. The company established in 1975 and presents across four continents that leading to a considerable data volume both in paper and electronic document. The company offers customers from various shipyards different innovative cleantech solutions, which address the megatrends of climate change in energy efficiency and saving freshwater resources. The cleantech solutions are mainly developed for dry and wet waste treatment systems, wastewater treatment systems, freshwater generation systems, and vacuum systems (Fig 8). The company delivers turnkey projects with full services including providing the complete design package and supplying all products manufactured by trustworthy suppliers to the customers. There is a large volume of documents created, exchanged, or disposed from internal and external sources. The type of data varies from technical and business data. However, this thesis will focus on technical product data management through documentation management system including the scope of supply, purchase orders, mechanical, electrical information and detail drawings, product technical description, manual documents, and component or bill of the material list for spare parts.

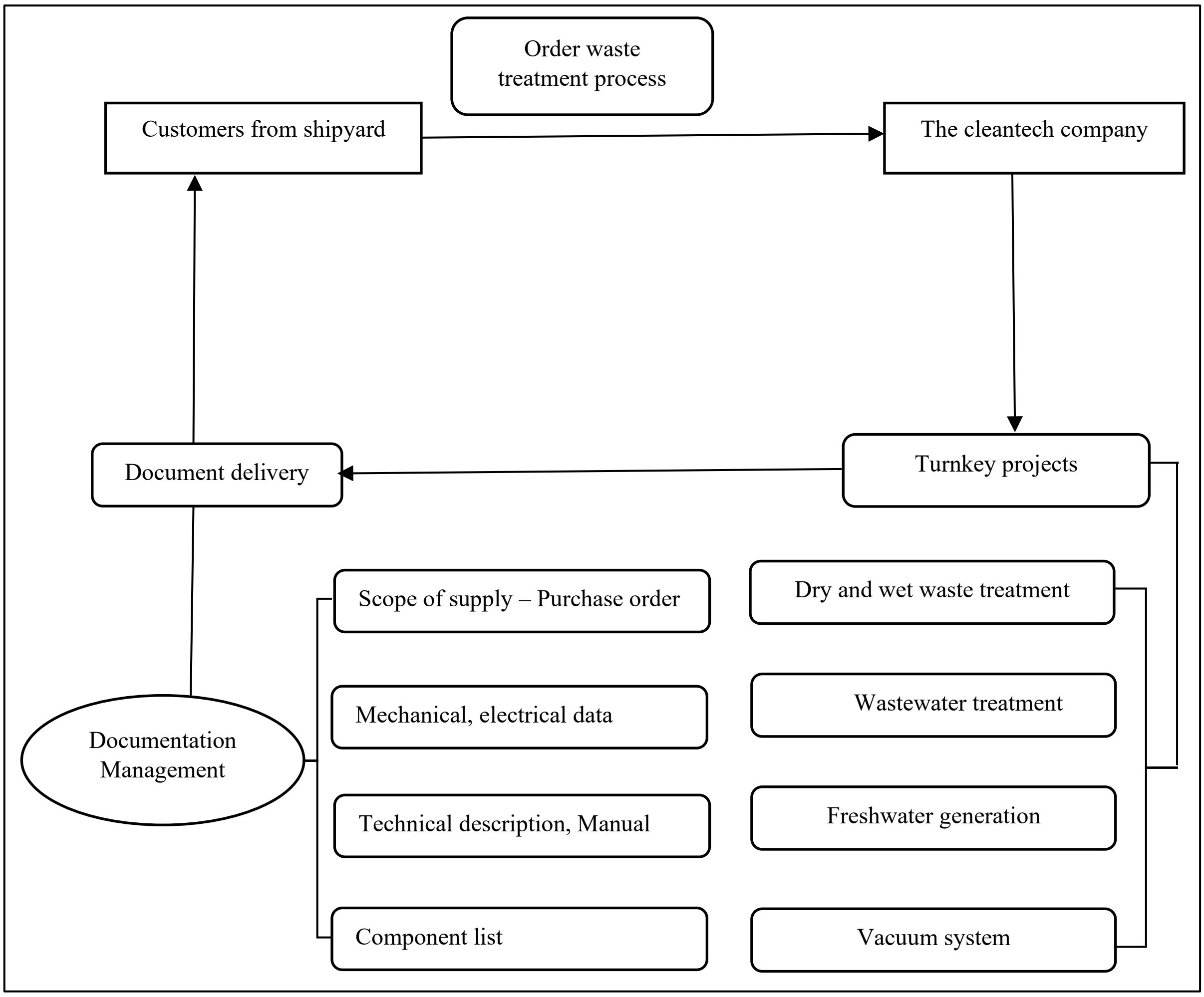
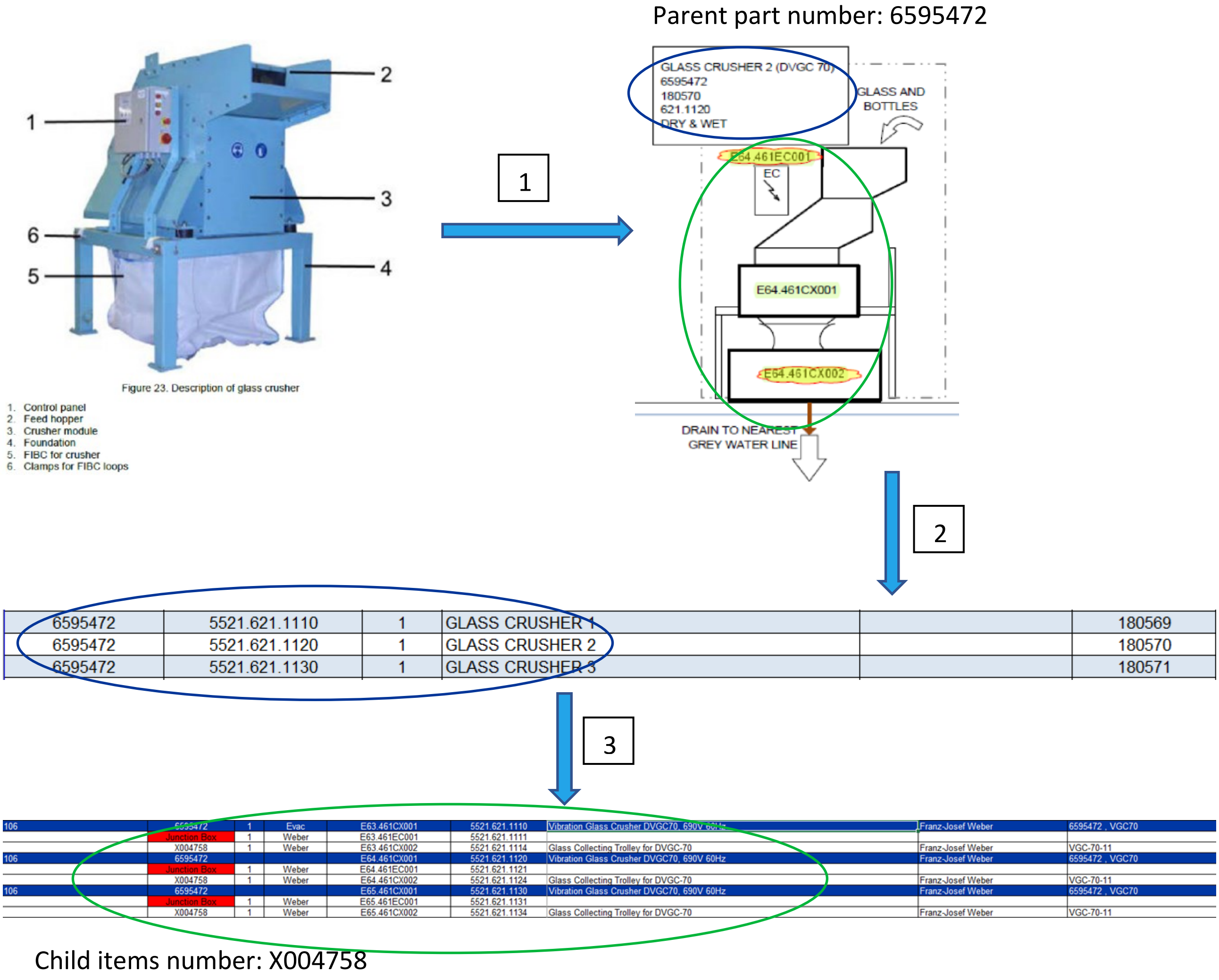


Figure 8. The workflow description of the cleantech company

The product data management system controls data variety, especially product number, product description, manufacturer, and assemblies as sub-parts of the active products. The product data management system can enhance spare parts management systems for customers to maintain customers' processes running properly and continuously. Consequently, it can increase spare parts sales for the company. Therefore, it is highly important to provide accurate and speedy information of product and spare parts. Figure 9 shows one example of data flow for the dry waste treatment process.



6595472	5521.621.1110	1	GLASS CRUSHER 1		180569
6595472	5521.621.1120	1	GLASS CRUSHER 2		180570
6595472	5521.621.1130	1	GLASS CRUSHER 3		180571

106	6595472	1	Evac	E63.461CX001	5521.621.1110	Vibration Glass Crusher DVGC70, 690V 60Hz	Franz-Josef Weber	6595472 , VGC70
	Junction Box	1	Weber	E63.461EC001	5521.621.1111			
	X004758	1	Weber	E63.461CX002	5521.621.1114	Glass Collecting Trolley for DVGC-70	Franz-Josef Weber	VGC-70-11
106	6595472			E64.461CX001	5521.621.1120	Vibration Glass Crusher DVGC70, 690V 60Hz	Franz-Josef Weber	6595472 , VGC70
	Junction Box	1	Weber	E64.461EC001	5521.621.1121			
	X004758	1	Weber	E64.461CX002	5521.621.1124	Glass Collecting Trolley for DVGC-70	Franz-Josef Weber	VGC-70-11
106	6595472			E65.461CX001	5521.621.1130	Vibration Glass Crusher DVGC70, 690V 60Hz	Franz-Josef Weber	6595472 , VGC70
	Junction Box	1	Weber	E65.461EC001	5521.621.1131			
	X004758	1	Weber	E65.461CX002	5521.621.1134	Glass Collecting Trolley for DVGC-70	Franz-Josef Weber	VGC-70-11

Figure 9. One example of data flow in a product data management system

The data flow starts with a technical description of the six main components built for the glass crusher in the left corner of Figure 9. These components are illustrated in 2D ACAD drawing, and 3D model with the parent part number (e.g., 6595472), and each main component has a separated tag number defined by the company. In the second stage, the parent part number is shown in the scope of supply with more information, such as quantity, and serial numbers. In the third stage, the parent part number includes all child items numbers and descriptions from the product design. This data flow is a first simple flow, and the data continues with more detailed information like assemblies, mechanical dimensions, wiring, power requirement, also painting information.

The product data information is used and handled manually through different shared workspace platforms. Figure 10 indicates the overview of the company's process from the beginning with the sale order to the end with project handover and close-out. The project documentation and product information are executed variously in the specific steps of project initiation, project development, project execution, product availability, product delivery, product installation, and project commissioning. There are several data platforms in use, such as Enterprise Resource Planning (ERP), product data platform, and project hub.

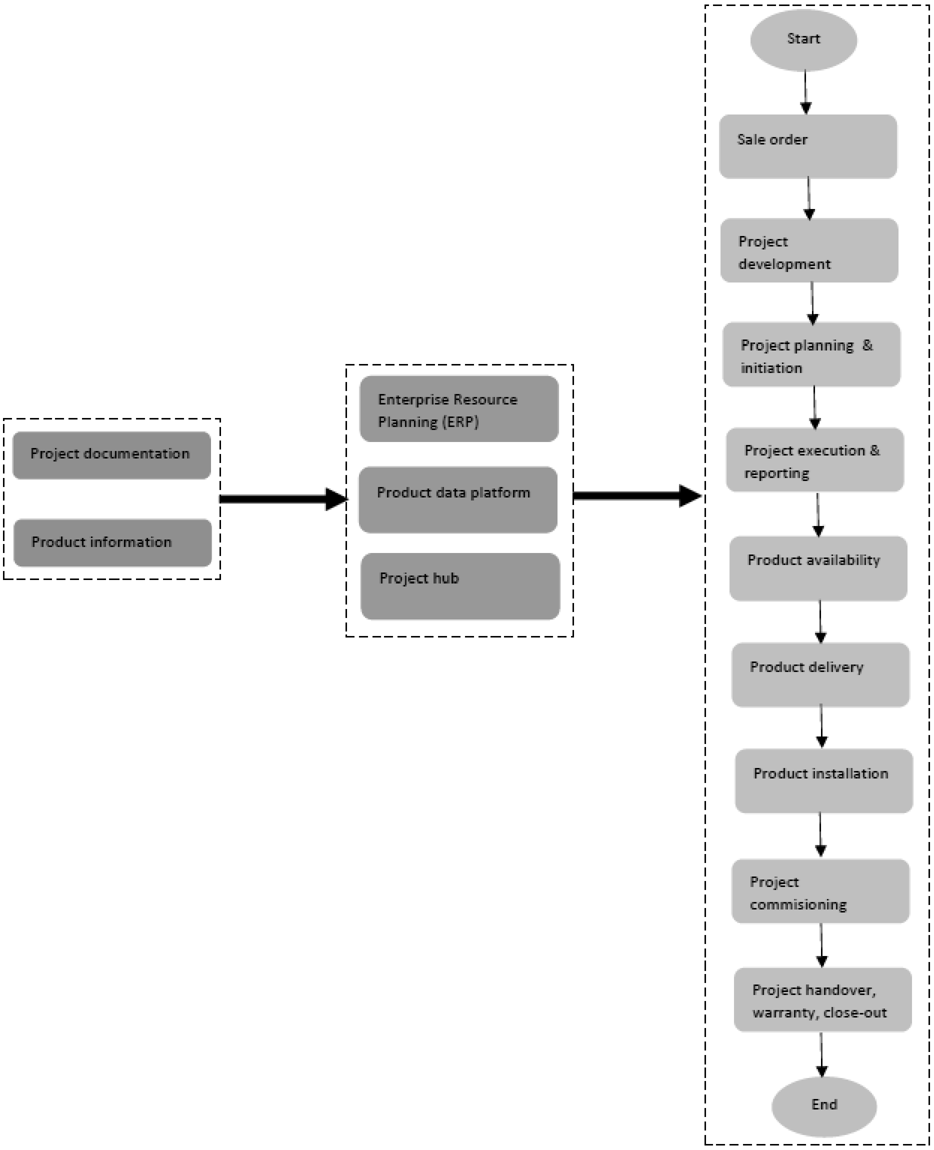


Figure 10. The overview of the company working process

3.2 Selection of method

The thesis will conduct unstructured or semi-structured interviews with several experts that can bring more insights about the data operation process. There are twelve questions prepared in a checklist for interviewees' scores based on their experiences. The twelve questions were selected based on Data Analytics problems and DataOps Manifesto (Appendix 1). Each question will be coded from Q1 to Q12. Each interviewee will score the checklist from one to five, including one is for the answer not relevant, not important, not necessary, disagree; five is for the answer highly relevant, extremely important, truly necessary, totally agree. The score will be summarized and visualized to reflect the important characteristics of the data operation process for the company case on the DataOps Manifesto principles.

In addition, the interview will be continued with unstructured discussion with the interviewees. The method will help to find patterns for new concept development and theoretical discovery with the careful presentation of evidence (Gioia, Corley and Hamilton, 2013). There are at least five questions prepared in advance for each interviewee based on their different positions (Appendix 2). The interview will take around 45 minutes or one hour. The interview will organize through Teams and recorded with the automatic transcription tool from Teams for each interview. The analysis process will start with building the first-order concepts including reading the interview transcription and classifying into small groups based on the content of the interviews. The second-order themes will be continued to categorize the first-order concepts into more abstract concepts. The third stage of the analysis process is named as aggregate dimensions in the paper of Gioia, Corley and Hamilton (2013), by combining abstract concepts into key categories identified in the previous stage. On the other hand, the transcription will be modeled by Natural Language Processing (NLP) machine learning for topic modeling.

3.3 Data collection

Qualitative research requires a highly relevant target group to interview. The interview will be interpreted based on discussions of eight interviewees (Table 2). The interviewees was selected by experience in the research topic or have a relevant position in the company. The desired outcome is to understand about DataOps for an effective product information management process.

Table 2. The list of target interviewees

Roles	**Group**	**Number of interviewees**
Related to Documentation Engineer	1	3
Related to Project Manager	2	1
Related to IT/ICT	3	2
Related to Product Design and Management	4	2

3.4 Data analysis

The important segment for the transcript analysis process is constructing the interview transcription as sentences for the unit of analysis (Fahy, 2001). The sentence segment is to take automatic raw transcripts formatted sentences with the sentence boundaries from the pause between words and by adding appropriate capitalization and punctuation to have meaningful sentences (Coden & Brown, 2001). Therefore, the interview transcription needs to be restructured by removing the words from the author and connecting words from the interviewees into a sentence with acceptable meaning and correct grammar with the Grammarly tool. The text preprocessing in NLP modeling was implemented with the following steps, including lowercasing, punctuation removal, stop words removal, tokenization, stemming, lemmatization (Kulkarni & Shivananda, 2019):

- Lowercasing: to convert all uppercase characters into lowercase characters by using the default lower () function in Python and reducing the size of the vocabulary of the text data.
- Punctuation removal: to remove the punctuation by using string function and to ensure not have different forms of the same word.
- Tokenizing text: to split the text into minimal meaningful units, including sentence tokenizer and word tokenizer.

- Removing stop words: to remove common words that carry no meaning or less meaning compared to other keywords and by using the NLTK library with a set of stop words in the English language.
- Stemming: a process of extracting the root form of a word by removing the prefix or suffix of a word.
- Lemmatization: a process of converting a word to its root from invalid words belonging to the language.

After text preprocessing, converting text to features using the TF-IDF method is to reflect the important words which appeared frequently in all documents. Term frequency (TF) is the ratio of the count of a word present in a sentence, to the length of the sentence for capturing the importance of the word. Inverse Document Frequency (IDF) is the log of the ratio of the total number of rows to the number of rows in a particular document in which the word was presented. The list of top-ranked keywords was selected as collateral information for the research.

In addition, Latent Dirichlet Allocation (LDA) is a topic modeling method to extract hidden topics from a collection of documents (Chen & Wang, 2019). LDA is a flexible generative probabilistic model of a corpus (Nazarko, Frank and Westerlund, 2021). There are three main parameters, including the number of topics, the number of words per topic, and the number of topics per document (Nazarko, Frank and Westerlund, 2021). The LDA model was implemented by using Python Gensim library. The clusters of keywords were generated from LDA model and interpreted the topics based on the experience of the author. Text after processing will be converted into a graph by using graph visualization and analysis tools pyLDAvis.

Finally, the eight interview transcriptions will be continually evaluated by reading and extracting the key terms for the qualitative research. The possible topics will compare with the topics obtained by the machine learning model to get the final key topics about DataOps for the company.

4 CONCLUSIONS

The study aimed to figure out the requirements for effective product portfolio management in the company based on the experts' views during inductive qualitative interviews. It is a good thing as all interviewees agreed on what good data or bad data is. Duplicate, copying errors and wrong data, out-of-date information, unavailable or misused data, are common issues that prevent to delivery of good data to customers. It opens an opportunity to bring the product portfolio to a higher data maturity level that all projects can easily find information and use properly. It is costly to manage product variants. The standard modular product information can be built and used as the family product number to link all parent product variants that have the same child items. Moreover, product identification is not only based on product numbers referring to everything but also based on product descriptions and technical information that customers can understand. To ensure data quality, it is required everybody's responsibility in each stage of the data lifecycle, such as keeping explanations records for any changes.

The study aims to understand about data operations process to handle and manage product data through data lifecycle and project lifecycle combined in product information system. The company has many different tools and platforms, but they do not have a connection or integration together automatically. It is hard to communicate, or feedback about problems from one team to another team. DataOps principles focus on data orchestration, automation, and collaboration. Data orchestration is implemented to collect data from multiple sources for the company with global operations. Data architecture or data flow needs to design appropriately before building the trust in automation and monitoring. Data pipeline can connect data creators and data consumers. Automation in place is to collect data from different teams in the company and from the customers' side. Customers' data operations are transferred from sensors or smart tags to a central hub to predict what parts need to replace or what product will be returned. Transparency and visibility are essential for team communication and collaboration in the company and with manufacturers, suppliers, and customers. Daily interactions are required during the project lifecycle with the involvement of all teams. The project handover documentation can be extracted and transformed automatically into a central database. The reference list and master list can be generated from the common database for all users.

However, the study results need to be validated by experts from the company. Due to limited time and experience, the author has only brought the first step of studying DataOps for the company. Although the study got the results mainly based on the inductive qualitative analysis manually. It opens an opportunity for the improvement of NLP modeling results in text extraction from a transcription of an interview.

REFERENCES

Bergh, C., Benghiat, G. and Strod, E. (2019). *The DataOps cookbook,* s.l.: DataKitchen, Inc.

Chen, X. and Wang, H. (2019) 'Automated chat transcript analysis using topic modeling for library reference services', *Proceedings of the Association for Information Science and Technology*, 56(1), pp. 368–371. doi:10.1002/pra2.31.

Coden, A.R. and Brown, E.W. (2001) 'Speech transcript analysis for automatic search', in *Proceedings of the 34th Annual Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences. HICSS-34*, Maui, HI, USA: IEEE Comput. Soc, p. 9. doi:10.1109/HICSS.2001.926473.

DAMA (2017). *DAMA - DMBOK Data management body of knowledge.* 2nd ed. s.l.:Technics Publications.

Demchenko, Y., Los, W. and de Laat, C. (2018) 'DATA AS ECONOMIC GOODS: DEFINITIONS, PROPERTIES, CHALLENGES, ENABLING TECHNOLOGIES FOR FUTURE DATA MARKETS', (2), p. 10.

Ereth, J. (2018) 'DataOps – Towards a Definition', Proc. of the Conf. "Lernen, Wissen, Daten, Analysen", Mannheim, Germany, August 22-24 2018, CEUR-WS.org, Vol-2191, pp 104-112

Fahy, P.J. (2001) 'Addressing some Common Problems in Transcript Analysis', *The International Review of Research in Open and Distributed Learning*, 1(2). doi:10.19173/irrodl. v1i2.321.

Figalist, I. (2021). From operational data to business insights adopting data-drive practices in B2B software-intensive company. Eindhoven Univeristy of Technology.

Gioia, D.A., Corley, K.G. and Hamilton, A.L. (2013) 'Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology', *Organizational Research Methods*, 16(1), pp. 15–31. doi:10.1177/1094428112452151.

Hannila, H. (2019). *Towards data-driven decision making in product porfolio management.* Univeristy of Oulu.

Held, J., Stonebraker, M., Davenport, T.H., Ilyas, I., Brodie, M. L., Palmer, A. and Markarian, J. (2016). *Getting data right Tacking the challengese of big data volume and variety.* O'Reilly Media, Inc.

Kulkarni, A. and Shivananda, A. (2019) Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Berkeley, CA: Apress. doi:10.1007/978-1-4842-4267-4.

Mainali, K. (2020). *DataOps: Towards understanding and defining data analytics approach.* KTH Royal Institute of Technology.

Mainali, K., Ehrlinger, L., Himmelbauer, J. and Matskin, M. (2021) 'Discovering DataOps: A Comprehensive Review of Definitions, Use Cases, and Tools', *DATA ANALYTICS*, p. 10.

Mostow, J. (1985). Toward better models of the design process. *AI magazine,* Volume 6.

Munappy, A.R. *et al.* (2020) 'From Ad-Hoc Data Analytics to DataOps', in *Proceedings of the International Conference on Software and System Processes*. *ICSSP '20: International Conference on Software and System Processes*, Seoul Republic of Korea: ACM, pp. 165–174. doi:10.1145/3379177.3388909.

Munappy, A. R. (2021). *Data managemtn and data pipeline: An empirical investigation in the embedded systems domain.* Chalmers University of Technology.

Nazarko, G., Frank, R. and Westerlund, M. (2021) 'Topic Modeling of StormFront Forum Posts', p.7.

Oracle. (2020). DataOps: An agile methodology for data-driven organization.

Paranyushkin, D. (2011) 'Identifying the Pathways for Meaning Circulation using Text Network Analysis', p. 25.

Pulvirenti, G. (2020). *A method of evaluation and prioritization of data governance activities in big data projects.*

Rodriguez, M., de Araújo, L.J.P. and Mazzara, M. (2020) 'Good practices for the adoption of DataOps in the software industry', *Journal of Physics: Conference Series*, 1694, p. 012032. doi:10.1088/1742-6596/1694/1/012032.

Sahoo, P.R. and Premchand A. (2019) 'DataOps in Manufacturing and Utilities Industries', *International Journal of Applied Information Systems*, 12, p. 6.

Simpson, T. W. (2004). Product platform design and customization: Status and promise. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing,* Volume 18, pp. 3-20.

Söderström, N. (2021). *Data management process for fast-growing companies.*

APPENDICES

Appendix 1: The list of questions in the checklist for the interview

Questions	Therorical subjects	Codes
Data error from different data sources (internal, external 3rd party) prevent the workflow correctly	Data Analytics challenges	Q1
Data formats are not optimized in a structured way	Data Analytics challenges	Q2
Manual process fatigue causes problems (e.g., produces errors, time consuming, and high potential of employees leaving)	Data Analytics challenges	Q3
Easy movement/migration between team members with many tools, platforms, and environments can build a diverse and dynamic team	DataOps Manifesto 3RD: EMBRACE CHANGE	Q4
Collaborate and coordinate work in terms of compelling direction, strong structure, supportive context, and shared mindset, are necessary for effective teamwork	DataOps Manifesto 10th: ORCHESTRATE	Q5
Automated work can reduce process variability and errors resulting from manual steps	DataOps Manifesto 15th: QUALITY IS PARAMOUNT	Q6
Continually satisfy your customers and evolve customer needs	DataOps Manifesto 1st: CONTINUALLY SATISFY YOUR CUSTOMER	Q7
Minimise the time and effort to turn a customer need into an action	DataOps Manifesto 18th: IMPROVE CYCLE TIMES	Q8
Reduce non-value-added tasks, avoid repetition of previously completed work	DataOps Manifesto 13th: SIMPLICITY	Q9
Use a version control system with built-in error detection and fault resilience	DataOps Manifesto 11th: MAKE IT REPRODUCIBLE	Q10
Encourage reuse work, database from shared workspace	DataOps Manifesto 17th: REUSE	Q11
Implement metrics shared to different departments to track the productivity (e.g., fast response, errors free days)	DataOps Manifesto 16th: MONITOR QUALITY AND PERFORMANCE	Q12

Appendix 2: The list of discussion questions for interviewees

Questions
What is a good product information management system?
What do you think about the idea of building a ‘Customer data platform’ instead of a ‘Product data platform’?
What do you define as non-value-add tasks from your daily work?
How can the company improve collaboration and coordination between teams?
What do you think about role changes after building trust through automation and monitoring?
In which ways can the company improve product information management?
In which way the company has a good data operation process, about people, process, technology
How to build trust from data through automation and minimise human interaction for product information management?

Appendix 7. TF-IDF and KMeans scripts

```
def get_fnames():
   fnames = []
   for root,_,files in os.walk("interview"):
      for fname in files:
         if fname[-4:] == ".txt":
            fnames.append(os.path.join(root, fname))
   return fnames
print("Number of interview transcripts: {}".format(len(get_fnames())))
name_list = get_fnames()
def read_file(fname):
   with open(fname, 'r',encoding="ISO-8859-1") as f:
      # get interview as a single string
      interview = ' '.join([line[:-1].strip() for line in f])
      return interview
documents = []
for i in name_list:
   documents.append(read_file(i))
from sklearn.feature_extraction.text import CountVectorizer
vect = CountVectorizer()
vect.fit(documents)
print("Vocabulary size: {}".format(len(vect.vocabulary_)))
print("Vocabulary content:\n {}".format(vect.vocabulary_))
cv = CountVectorizer(ngram_range=(3, 3)).fit(documents)
print("Vocabulary size: {}".format(len(cv.vocabulary_)))
print("Vocabulary:\n{}".format(cv.get_feature_names()))
```

```
tfidf_vectorizer = TfidfVectorizer(stop_words = 'english', lowercase= True, ngram_range = (3,3), min_df=1, use_idf=True, sublinear_tf=True, max_df=8)

tfidf_matrix = tfidf_vectorizer.fit_transform(documents)

tfidf_matrix.toarray().shape # N_docs x N_terms

features = tfidf_vectorizer.get_feature_names()

for doc_i in range(8):

    print("\nDocument %d, key words by TF-IDF" % doc_i)

    for term, score in sorted(list(zip(features,tfidf_matrix.toarray()[doc_i])), key=lambda x:-x[1])[:10]:

        print("%.2f\t%s" % (score, term))

from sklearn.cluster import KMeans

km = KMeans()

km.fit(tfidf_matrix)

import heapq, numpy as np

def print_clusters(matrix, clusters, n_keywords=10):

    for cluster in range(min(clusters), max(clusters)+1):

        cluster_docs = [i for i, c in enumerate(clusters) if c == cluster]

        print("Cluster: %d (%d docs)" % (cluster, len(cluster_docs)))

        new_matrix = np.zeros((len(cluster_docs), matrix.shape[1]))

        for cluster_i, doc_vec in enumerate(matrix[cluster_docs].toarray()):

            for idx, score in heapq.nlargest(n_keywords, enumerate(doc_vec), key=lambda x:x[1]):

                new_matrix[cluster_i][idx] = score

        keywords = heapq.nlargest(n_keywords, zip(new_matrix.sum(axis=0), features))

        print(', '.join([w for s,w in keywords]))

        print()

print_clusters(tfidf_matrix, km.labels_)
```

Appendix 8. LDA scripts

```
import spacy
def lemmatization(texts, allowed_postags=['NOUN','ADJ','VERB','ADV']):
    nlp = spacy.load('en_core_web_sm', disable=['parser','ner'])
    texts_out = []
    for text in texts:
        doc=nlp(text)
        new_text =[]
        for token  in doc:
            if token.pos_ in allowed_postags:
                new_text.append(token.lemma_)
        final=' '.join(new_text)
        texts_out.append(final)
    return (texts_out)
lemmatized_texts =lemmatization(documents)
print(lemmatized_texts[0][0:99])
def gen_words(texts):
    final = []
    for text in texts:
        new =gensim.utils.simple_preprocess(text, deacc=True)
        final.append(new)
    return (final)
data_words = gen_words(lemmatized_texts)
print(data_words[0][0:20])
import gensim.corpora as corpora
```

```
id2word = corpora.Dictionary(data_words)

corpus =[]

for text in data_words:

   new =id2word.doc2bow(text)

   corpus.append(new)

print(corpus[0][0:20])

lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
id2word = id2word, num_topics = 10, random_state=100,
chunksize=100,  passes=10, iterations=100)

import pyLDAvis

import pyLDAvis.gensim_models as gensimvis

pyLDAvis.enable_notebook()

gensimvis.prepare(lda_model, corpus, id2word, mds ='mmds', R=40)
```