Data Pipeline Management in Practice: Challenges and Opportunities

Data Pipeline Management in Practice: Challenges and Opportunities

Aiswarya Raj Munappy[1], Jan Bosch[1], and Helena Holmström Olsson[2]

[1] Department of Computer Science and Engineering, Chalmers University of Technology, Hörselgången 11, 412 96 Gothenburg, Sweden
{aiswarya,jan.bosch}@chalmers.se
[2] Department of Computer Science and Media Technology, Malmö University, Nordenskiöldsgatan, 211 19 Malmö, Sweden
helena.holmstrom.olsson@mau.se

Abstract. Data pipelines involve a complex chain of interconnected activities that starts with a data source and ends in a data sink. Data pipelines are important for data-driven organizations since a data pipeline can process data in multiple formats from distributed data sources with minimal human intervention, accelerate data life cycle activities, and enhance productivity in data-driven enterprises. However, there are challenges and opportunities in implementing data pipelines but practical industry experiences are seldom reported. The findings of this study are derived by conducting a qualitative multiple-case study and interviews with the representatives of three companies. The challenges include data quality issues, infrastructure maintenance problems, and organizational barriers. On the other hand, data pipelines are implemented to enable traceability, fault-tolerance, and reduce human errors through maximizing automation thereby producing high-quality data. Based on multiple-case study research with five use cases from three case companies, this paper identifies the key challenges and benefits associated with the implementation and use of data pipelines.

Keywords: Data pipelines · Challenges · Opportunities · Organizational · Infrastructure · Data quality · Issues.

1 Introduction

Data is being increasingly used by industries for decision making, training machine learning(ML)/deep learning(DL) models, creating reports, and generating insights. Most of the organizations have already realized that big data is an essential factor for success and consequently, they use big data for business decisions [9] [13]. However, high-quality data is critical for excellent data products [3]. Companies relying on data for making decisions should be able to collect, store, and process high-quality data. Collecting data from multiple assorted sources to producing useful insights is challenging [1]. Moreover, big data is difficult to configure, deploy, and manage due to its volume, velocity, and variety [12].

The complex chain of interconnected activities or processes from data generation through data reception constitutes a data pipeline. In other words, data pipelines are the connected chain of processes where the output of one or more processes becomes an input for another [19]. It is a piece of software that removes many manual steps from the workflow and permits a streamlined, automated flow of data from one node to another. Moreover, it automates the operations involved in the selection, extraction, transformation, aggregation, validation, and loading of data for further analysis and visualization [11]. It offers end to end speed by removing errors and resisting bottlenecks or delay. Data pipelines can process multiple streams of data simultaneously [14].

Data pipelines can handle batch data and intermittent data as streaming data [14]. Therefore, any data source will be compatible with the data pipeline. Furthermore, there is no strict restriction on the data destination. It does not require data storage like a data warehouse or data lake to be the end destination. It can route data through a different application like visualization or machine learning or deep learning model.

Data pipelines in production should run iteratively for a longer duration due to which it has to manage process and performance monitoring, validation, fault detection, and mitigation. Data flow can be precarious, because there are several things that can go wrong during the transportation of data from one node to another: data can become corrupted, it can cause latency, or data sources may overlap and/or generate duplicates [5]. These problems increase in scale and impact as the number of data sources multiplies and complexity of the requirements grows.

Therefore, data pipeline creation, management, and maintenance is a complicated task which demands a considerable amount of time and effort. Most of the companies do this maintenance manually by appointing a dedicated person to guard the data flow through the pipeline. This study aims to investigate the opportunities and challenges practitioners experience after the implementation of the data pipeline at their organization.

The contribution of this paper is three-fold. First, it identifies the key challenges associated with data pipeline management. Second, it describes the opportunities of having a dedicated data pipeline. These challenges and opportunities are validated through a multi-case study with three leading companies in telecommunication and automobile domains. Furthermore, the paper provides a taxonomy of data pipeline challenges including infrastructural, organizational, and technical ones.

The remainder of this paper is organized as follows. In the next section, we present the background of the study. Section III discusses the research methodology adopted for conducting the study. Section IV introduces the use cases and section V describes the opportunities created by the pipelines. Section VI details the challenges faced by practitioners while managing data pipelines. Section VII outlines the threats to validity. Section VIII summarizes our study and the conclusions.

2 Background

Several recent studies have recognized the importance of data pipelines. Raman et. al [19] describes Big Data Pipelines as a mechanism to decompose complex analyses of large data sets into a series of simpler tasks, with independently tuned components for each task. Moreover, large scale companies like Google, Amazon, LinkedIn, and Facebook have recognized the importance of pipelines for their daily activities. Data errors and their impact on machine learning models are described in [6] by Caveness et. al. They also propose a data validation framework that validates the data flowing through the machine learning pipeline.

Chen et. al describes the real-time data processing pipeline at Facebook [8] that handles hundreds of Gigabytes per second across hundreds of data pipelines. The authors also identify five important design decisions that affect their ease of use, performance, fault tolerance, scalability, and correctness and also demonstrate how these design decisions satisfy multiple use cases on Facebook. LinkedIn also has a similar real-time data processing pipeline described by Goodhope et. al in [10]. Data management challenges of deep learning is discussed by Munappy et. al through a multiple case study conducted with five different companies and classifies the challenges according to the data pipeline phases [16]. Lambda architecture proposed by N. Marz et. al and Kappa architecture [18] solves the challenge of handling real-time data streams [14]. Kappa architecture that considers both online and offline data as online is a simplified version of lambda.

Most of these studies illustrate the significance of data pipelines and the opportunities it can bring to the organizations. However, the challenges encountered in the industrial level during the development and maintenance of the data pipelines in production is still not completely solved.

3 Research Methodology

The objective of this study is to understand the existing data pipeline as well as the challenges experienced at the three case companies and to explore the opportunities of implementing a data pipeline. Specifically, this study aims to answer the following research question:
RQ: What are the practical opportunities and challenges associated with the implementation and maintenance of Data Pipelines at the industry level?

3.1 Exploratory Case Study

A qualitative approach was chosen for the case study as it allows the researchers to explore, study, and understand the real-world cases in its context in more depth [23]. Since the concept of data pipelines is a less explored topic in research, we have adopted a case study approach [21]. Moreover, the case study approach can investigate contemporary real-life situations and can provide a foundation for the application of ideas and extension of methods. Each case in the study pertains to a use case that makes use of data. Table 1 details the selected five use cases from three companies.

Table 1. Outline of use cases and roles of the interviewees

Company	Use cases	Interviewed Experts	
		ID	Role
A	Data Collection Pipeline	R1	Senior Data Scientist
A	Data Governance Pipeline	R2	Data Scientist
		R3	Analytics System Architect
		R4	Software Developer
A	Data Pipeline for Machine learning Applications	R5	Data Scientist
		R6	Senior Data Scientist
		R7	Software Developer
		R8	Senior Data Scientist
B	Data Collection Pipeline	R9	Senior Data Engineer
		R10	Data Engineer
		R11	Data Engineer
		R12	Data Analyst and Superuser
C	Data Quality Monitoring Pipeline	R13	Director of data analytics team
		R14	ETL developer
		R15	Software Developer
		R16	Product Owner for data analytics team

3.2 Data Collection

Qualitative data was collected by means of interviews and meetings [22]. Based on the objectives of the research, to explore and study the applications consuming data in the companies, an interview guide with 43 questions categorized into nine sections was formulated. The first and second sections focused on the background of the interviewee. The third and fourth sections focused on the data collection and processing in various use-cases and the last section inquired about data testing and monitoring practices and the impediments encountered during the implementation and maintenance of data pipelines. All interviews were conducted virtually via videoconferencing due to the COVID-19 pandemic. Each interview lasted 40 to 60 minutes. The interviews were recorded with the permission of respondents and were transcribed later for analysis. The first author is an action researcher for the past one year/six months at company A and B respectively who attend weekly meetings with data scientists and data analysts. The data collected through these means are also incorporated.

3.3 Data Analysis

The contact points at the companies helped with analyzing the parts of the pipeline as well as the infrastructure used for building that pipeline. These notes together with the codes from transcripts were further analyzed to obtain an end-to-end view of different use cases. The audio transcripts were investigated for relations, similarities, and dissimilarities. The interview transcripts and meeting notes were open coded following the guidelines by P. Burnard [2]. After careful analysis of collected data and based on the inputs from the other two authors, the first author who is an action researcher at two of the companies developed the findings of the study which were then validated with the interviewees from

the companies by conducting a follow-up meeting. For further validation and to collect feedback from a different team, the findings were also presented before another panel including super users, managers, software developers, data engineers, and data scientists at all three companies who were not involved in the interviews. The results were updated according to the comments at each stage of the validation which in turn helped to reduce the researcher bias.

4 Use cases

In this multi-case study, we explore data pipelines in real-world settings at large-scale software intensive organizations. Company A is within the telecommunication industry with nearly 100,000 employees who distributes easy to use, adoptable, and scalable services that enables connectivity. Further, we investigate Company B from automobile domain with 80,000 employees manufacturing its own cars responsible for collecting data from multiple manufacturing units as well as repair centers. Company C with 2,000 employees focus on automotive engineering and depends on Company B and does modular development, advanced virtual engineering and software development for them. In this section, we present five use cases of data pipelines studied from these three case companies A, B and C.

Case A1: Data Collection Pipeline

The company collects network performance data(every 15 minutes) as well as configuration management data(every 24 hours) in the form of data logs from multiple sources distributed across the globe which is a challenging activity. Data collection from devices located in another country or customer network requires compliance with legal agreement. The collected data can have sensitive information like use details which needs responsible attention. Furthermore, data generated by sources can be of different formats and frequencies. For instance, data generation can be continuous, intermittent or as batches. Consequently, the data collection pipeline should be adaptable with different intensities of data flow.

When data collection pipeline is implemented, these challenges should be carefully addressed. Fig. 1 shows the automatic data collection pipeline that collects data from distributed devices. In this scenario, the device is placed inside a piece of equipment owned by customers. However, the device data is extracted by filtering the customer's sensitive information. Base stations have data generation devices called nodes as well as a device for monitoring and managing the nodes. Data collection agents at the customer premise can interact either with nodes directly. However, access service is used for authentication. The data thus collected is transmitted through a secure tunnel to the data collection toolkit located at the company premise which also has access service for authentication. Data collection toolkit received the data and store it in the central data storage from where the teams can access the data using their data user credentials.

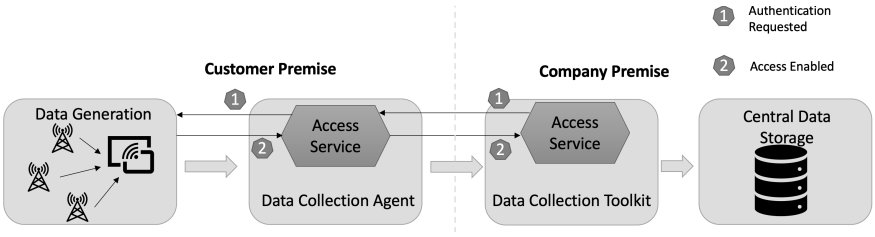

Fig. 1. Data Collection Pipeline

Case A2: Data Governance Pipeline

Fig. 2 illustrates the data pipeline that serves a subset of teams in the company who are working with data whenever they need it (With the term 'data', we mean the link from which the original data can be downloaded). This data pipeline gets two types of data dumps: internal and external which is the performance management data collected in every 15 minutes from the devices deployed in the network. The internal data dump is the data that is ingested by the teams inside the company and external data dump is the data collected directly from the devices in the fields. The data ingestion method varies according to the data source and the ingested data is stored in the data storage for further use. The data can be encrypted form which needs decryption before storing it. Data archiver module sends encrypted data dump to the third-party services for decryption. Decoded links from the third party are transferred to data storage. Therefore, data from distributed sources are made available in a central location. Teams can request data from any stage of the pipeline. The monitoring mechanism in the pipeline is manually carried out by the 'flow guardian' who is responsible for fixing the issues in the pipeline.

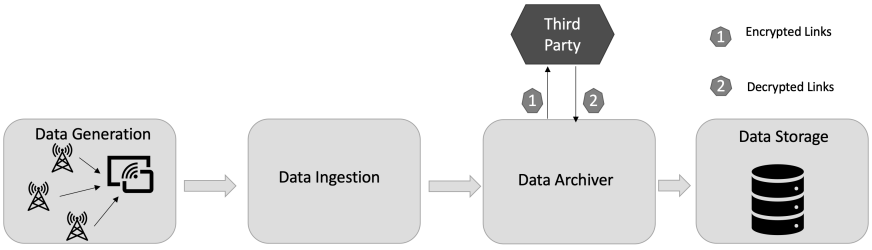

Fig. 2. Data Governance Pipeline

Case A3: Data Pipeline for Machine learning Applications

Data for this pipeline is obtained from the devices that are sent to the repair center. Data pipelines for machine learning applications has four main steps namely ingest, store, transform and aggregate. Data generated by the source is gathered at a special zone in the field. The data ingestion module connected to those zones in the field collects data and ingest into the pipeline as batches. When new compressed files are found in the periodic checks, the transaction is logged and downloads it. These new files are then loaded into the archive directory of the data cluster. The data stored in the cluster cannot be used directly by the machine learning applications. Moreover, the data logs collected from different devices will be of different formats. Data transformation checks for the new files in the archive directory of the data cluster and when found, it is fetched, uncompressed and processed to convert it to an appropriate format. The converted data is then given as input to the data aggregation module where the data is aggregated and summarized to form structured data which is further given as input to the machine learning applications. Fig. 3 illustrates the data pipeline for machine learning applications

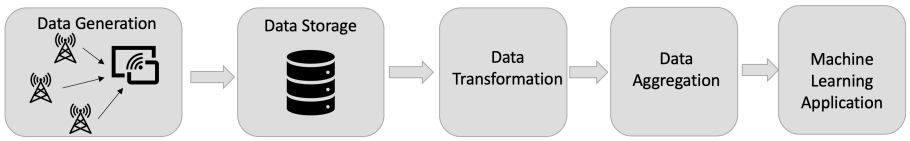

Fig. 3. Data Pipeline for Machine Learning Applications

Case B1: Data Collection Pipeline

The Company B collects and stores three types of data and distributes it for teams as well as co-working organizations distributed around the globe. Plant data, delivery data, warranty data and repair data are the different types of data that are collected from sources such as manufacturing plants, service centers, delivery centers and warranty offices. The company B collects product data from distributed manufacturing plants every 24 hours. These manufacturing units will generate data for each product built there. However, not all the data generated by the plants are collected by the data collection agent of company B. Group Quality IT platform in the company demands the data that needs to be collected from the plants. Also, the data requested by the delivery centers are also collected and stored in the company's data warehouse. Fig. 4 illustrates the data collection pipeline working in company B. The data collected from different sources are in different formats and volume. Therefore, data transfer mechanism as well as data storage is different for all data sources. The data is ingested from the

primary storage and then transformed into a uniform format and stored in a data warehouse which then acts as a supplier for teams as well as other organizations who demand for data. For instance, the delivery centers needs data about the products that are manufactured in the plants.

Fig. 4. Data Collection Pipeline

Case C1: Data Quality analysis Pipeline

The company C receives data collected and stored by company B and creates data quality reports which is used by data scientists team for analysing the product quality. For instance, the report can be used to understand the model that is sent to repair centers frequently. When the data quality is not satisfactory, investigation is initiated and actions are taken to fix the data quality issues. Company B sends data through private network to company C, and they store it in a data storage from where data scientists access it for creating reports and training machine learning models. Fig. 5 shows the data pipeline for data quality analysis at Company C.

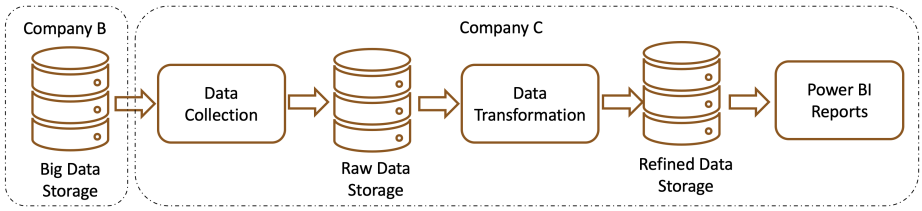

Fig. 5. Data Quality Analysis Pipeline

5 Challenges to Data Pipeline Management

Based on our research, we see that organizations benefit from developing and maintaining data pipelines because of the automation it provides. On the other hand, there are certain challenges faced by practitioners while developing and managing data pipelines. This section describes the challenges of data pipelines

derived through the interpretation of interviews based on the use cases described in section 4. After careful analysis of the challenges obtained from the interviews, we formulated a taxonomy for the classification of challenges namely Infrastructure Challenges, Organizational Challenges and Data Quality Challenges which are described in detail below.

5.1 Infrastructure Challenges

Data pipelines are developed to solve complex data infrastructure challenges. However, data pipeline management has to deal with some infrastructural challenges listed below.

Integrating new data sources: Data pipelines collect data from multiple distributed devices and make it available in a single access point thus solving data accessibility problem. However, the data sources increase rapidly in most of the business scenarios. Therefore, data pipelines should be able to integrate the new data source and also accommodate the data from that new source which is often difficult due to many reasons. Based on the empirical findings from the case study, three common reasons are listed below.

- The data source can be entirely different from the existing sources.
- Format of the data produced by the source might not be compatible with the data pipeline standards.
- Addition of the new source may introduce overhead on the data handling capability of the pipeline.

All the use cases except case C1 described in section 4 experience the challenge of integrating new data sources.

Data pipeline scalability: The ability of a data pipeline to scale with the increased amount of ingested data, while keeping the cost low is a real challenge experienced by the data pipeline developers. When the data produced by the source increases, the data pipeline loses the ability to transfer the data from one node to another leading to the data pipeline breakage and loss of data.

Increased number of nodes and connectors in upstream: Data pipelines are a chain of nodes performing activities connected through connectors that enable data transportation between the two nodes. Essentially, the nodes can have more than one capability. However, for the easy detection of faults, each of the nodes should be preferably assigned a single capability. Thus, the number of nodes and connectors increases in the upstream in relation to the data product yielded from the pipeline. This in turn increases the complexity of the data pipeline and decreases ease of implementation. The fragility and complexity of the data pipeline lead to inevitable delays in adding new types of activity data, which resulted in sticking new activities into inappropriate existing types to avoid human effort, or worse, not capturing activities at all. Practitioners R9, R10, R11 and R14 working on case B1 and C1 raised this challenge.

> "With the increased number of components in the data pipeline which in turn makes it difficult to understand and maintain. It is difficult to

attain the right balance between robustness and complexity" - Senior Data Scientist (R6)

Trade-off between data pipeline complexity and robustness: To build a robust data pipeline, we should have two essential components called fault detection and mitigation strategies. Fault detection identifies faults at each of the data pipeline stages and mitigation strategies help to reduce the impact of the fault. Including these two components increases the complexity of data pipelines. Moreover, it requires the data pipeline developers to anticipate the faults that can occur at each stage and define mitigation actions such that the data flow through the pipeline is not hampered. Some of the common faults can be anticipated and mitigated. However, it is not possible to identify all possible faults and define mitigation actions for those. Senior data scientists working on Case B1 and C1 and data scientist, R5 working on case A3 pointed out this as an important challenge.

Repeated alarms: Sending alarms are the most common and simple mitigation actions automatically taken by the data pipelines. Some faults take time to get fixed and during this time, the person or the team responsible for fixing the issues will get repeated alarms for the same issue. In the worst scenario, this can even lead to new alarms left unnoticed. Sending alarms is a mechanism adopted by all five data pipelines described in section 4. However, data engineers and software developers who participated in the study want to have an alternate mitigation strategy to replace the repeated automatic alarms in the data pipeline such as sending the notification only once and then waiting for a fix for some time.

"Sending notifications is less appreciated by the teams as we get totally submerged in alarms during some days and some notifications are repeatedly sent and it is hard to identify new ones from the huge pile" - Senior Data Engineer (R9)

5.2 Organizational Challenges

This section gives a brief overview of the organization level challenges to data pipeline management.

Dependency on other organizations: Data pipelines can be spread between more than one company like case IV and V. Therefore, co-operation and collaboration are required from all the participating companies to maintain a healthy data pipeline. In most cases, external companies will have very minimal knowledge of what is happening in the other part of the pipeline. For instance, to deliver high-quality data product, company C requires support from company B as they are the suppliers of data.

Lack of communication between teams: Data pipelines are meant to share data between various teams in the organization. However, each team builds pipelines for their use case and thus at least some initial activities are repeated in several data pipelines leading to redundant storage of data. Moreover, if any of the steps fails, the responsible person gets a notification from different teams at the same time for the same issue. Cases A1, A2, and A3 are collecting the same data and storing it in their databases. Data pipeline in case A3 can fetch data

stored by data pipeline A2 instead of collecting raw data from the data sources. However, practitioners working on these use cases were completely unaware of these repeated activities in their pipelines.

Increased responsibilities of Data Pipeline owner: All faults in the data pipeline cannot be fixed automatically. Some faults demand either partial or complete human intervention. Therefore, a flow guardian or data pipeline owner is assigned for each of the pipelines who pays attention to data pipeline activities and takes care of the faults requiring a manual fix. Further, it is hard to assess what code might break due to any given change in data. With the increased use of data pipelines, the responsibilities of the flow guardian or data pipeline owner also increase. Practitioner R11 is assigned responsibilities of a flow guardian, and he has to manually monitor the data pipelines and initiate an investigation and fix whenever a problem is encountered. As Company C is also dependent on Company B, responsibilities are shared between R10 and R11. R10 takes care of request from Company C and R11 attends to the problems with company B.

> "Nobody wants to take up the responsibility of flow guardian. We feel that it consumes a lot of time and effort" - Director of Data Analytics Team (R13)

DataOps-DevOps Collaboration: When seeking to obtain better results from machine learning models require better, more focused data, better labeling, and the use of different attributes. It also means that data scientists and data engineers need to be part of the software development process. DataOps is concerned with a set of practices for the development of software and management of data respectively. Both concepts emphasize communication and collaboration between various teams of the same organization. DataOps combines DevOps with data scientists and data engineers to support development. The challenge of managing and delivering massive volumes of discordant data to those who can use it to generate value is proving extremely hard. Moreover, people working with data are less interested in learning new technologies and tools while it is not a hassle for DevOps users.

5.3 Data Quality Challenges

This section gives a detailed list of the data quality challenges due to improper data pipeline management.

Missing data files: Data files can be lost completely or partially during the transmission from one node to another. Fault detection mechanism can identify the exact point of disappearance. However, obtaining the missing files once again is a complicated task. Missing data files are only detected at the end of the data pipeline and in some cases, this results in poor quality data products. All the use cases experience the challenge of missing data files at different stages of data pipelines and one of the practitioners, R4 identified that 38,732 files had gone missing at a particular stage of the data pipeline over five months.

> "Data quality is a challenge that is being discussed over years. But, at industry level we still struggle to achieve desired level of data quality" - Senior Data Scientist(R1)

Operational errors: Data pipelines encounter operational errors which hampers the overall functioning. Operational errors are very common in non-automated data pipelines. Some parts of the data pipelines cannot be completely automated. Human errors at these steps are the reasons for operational errors. For instance, data labeling in a data pipeline cannot be automated completely due to the unavailability of automated annotation techniques that are compatible with all types of datasets. Practitioner R12, R13, R4, and R3 raised the problem of operational errors and their impact on their respective data pipelines.
Logical changes: Data drifts and change in data distribution results in the data pipeline failures due to the incompatible logic defined in the data pipeline. Therefore, the data pipeline needs to be monitored continuously for change in data distributions and data shifts. Besides, data pipelines should be updated frequently by changing the business logic according to the changes in data sources. Practitioner R12, R13, and R16 explained the struggles of working with outdated business logic in their data pipelines.

6 Opportunities

The previous section illustrated the challenges of data pipelines when implemented in real-world. However, there are many opportunities the data pipeline offers through automating fault detection and mitigation. In this section, we survey some of the most promising opportunities of data pipelines and how practitioners working on data are benefited by the implementation of it.

6.1 Solve data accessibility challenges

Data generated by assorted multiple devices are collected, aggregated, and stored in central storage by data pipelines without human intervention. As a result, data teams within and outside the organization can access data from that central storage if they have proper data access permissions. Accessing data from devices located on the customer premises is a difficult and tedious task. Most often, the devices will be distributed around the globe and teams has to prepare legal agreements complying with the rules of that specific country where the device is located for accessing data. When the data is stored after aggregation, data loses its granularity, and as a result, teams working with fine-grained data has to collect data separately. With data pipelines, teams can access data from any point of the data pipeline if they have necessary permissions. This eliminates repeated collection and storage of the same data by multiple teams.

6.2 Save time and effort of human resources

Automation of data-related activities is maximized through the implementation of data pipelines thereby reducing the human intervention. When a data pipeline has inbuilt monitoring capability, faults will be automatically detected

and alarms will be raised. This reduces the effort of data pipeline managers and flow guardians. As the data pipeline is completely automated, requests by teams will be answered quickly. For instance, if the data quality is not satisfactory to the data analyst, he can request the data from the desired store in the data pipeline, and he receives it without delay. On the other hand, if the workflow is not automated, the data analyst has to investigate and find out where the error has occurred and then inform the responsible person to send the data again which eventually delays the entire data analysis process. Moreover, the effort of the data analyst is also wasted while investigating the source of data error.

> *"We spent time cleaning the data to meet the required quality so that it can be used for further processing such as training or analytics. With the data pipeline, it is easy to acquire better quality data."* - Analytics System Architect(R3)

6.3 Improves traceability of data workflow

Data workflow consists of several interconnected processes that make it complex. Consequently, it is difficult to detect the exact point that induced error. For instance, if the end-user realizes that part of the data is missing, it might be lost during data transmission, while storing the data in a particular schema or due to unavailability of an intermediate process. The end-user has to guess all the different possibilities of data loss and has to investigate all the possibilities to recover the lost data. This is a time-consuming task especially when the data workflow is long and complex. Company C has reported that they have experienced this problem several times and as they are getting data from company B, it took a lot of time for them to rectify the error, and sometimes they won't be able to recover the data. After implementing data pipelines, the process of detecting faults is automated thereby increasing traceability.

> *"Everyone in the organization is aware of the steps and with data pipelines, you will have full traceability of when the pipeline slowing down, leaking, or stops working."* - Data Scientist(R5)

6.4 Supports heterogeneous data sources

Data pipelines can handle multiple assorted data sources. Data ingestion is a process through which data from multiple sources are made available to the data pipeline in a uniform format. Data Ingestion is the process of streaming-in massive amounts of data in our system, from several external sources, for running analytics and other operations required by the business.

> *"Data streams in through several sources into the system at different speeds and sizes. Data ingestion unifies this data and decreases our workload. Data ingestion can be performed as batches or real-time."* - Data Engineer(R10)

6.5 Accelerates Data life cycle activities

The data pipeline encompasses the data life cycle activities from collection to refining; from storage to analysis. It covers the entire data moving process, from where the data is collected, such as on an edge device, where and how it is moved, such as through data streams or batch-processing, and where the data is moved to, such as a data lake or application. Activities involved in the data pipeline are automatically executed in a predefined order and consequently, human involvement is minimized. As the activities are triggered by themselves, the data pipeline accelerates the data life cycle process. Moreover, most of the data pipeline activities are automated thereby increasing the speed of data life cycle process and productivity.

6.6 Standardize the Data Workflow

The activities in a data workflow and their execution order are defined by a data pipeline which gives the employees in the organization an overall view of the entire data management process. Thus, it enables better communication and collaboration between various teams in the organization. Further, data pipelines reduce the burden on IT teams thereby reducing support and maintenance costs as well. Standardization through data pipelines also enables monitoring for known issues and quick troubleshooting of common problems.

> *"Data pipelines provide a bird's eye view of the end to end data workflow. Besides, it also ensures a short resolution time for frequently occurred problems."* - Product Owner(R16)

6.7 Improved Data Analytics and Machine Learning Models

Organizations can make use of carefully designed data pipelines for the preparation of high quality, well-structured, and reliable datasets for analytics and also for developing machine learning as well as deep learning models. Besides, data pipelines automate the movement, aggregation, transformation, and storage of data from multiple assorted sources. Machine learning models are highly sensitive to the input training data. Therefore, quality of training data is very important. Data pipelines are traceable since the stages are predefined yielding better quality data for the models. Moreover, data pipelines ensure a smooth flow of data unless it fails in one of the steps.

6.8 Data Sharing between teams

Data pipelines enable easy data sharing between teams. Practitioners R4, R8, and R9 mentioned that the data collected from devices in the field are undergoing the same processing for different use cases. For instance, data cleaning is an activity performed by all the teams before feeding the data to ML/DL models. Therefore, there is a possibility of the same data going through the same sequence of steps within different teams of the same organization. Further, data storage

also is wasted in such cases due to redundant storage. With the implementation of data pipelines, the teams can request data from a particular step in some other data pipeline and can process the subsequent steps in their data pipeline. However, the data pipeline should be able to serve the requests in such cases.

6.9 Critical Element for DataOps

DataOps is a process-oriented approach on data that spans from the origin of ideas to the creation of graphs and charts which creates value. It merges two data pipelines namely value pipeline and innovation pipeline. Value pipeline is a series of stages that produce value or insights and innovation pipeline is the process through which new analytic ideas are introduced into the value pipeline. Therefore, data pipelines are critical elements for DataOps together with Agile data science, continuous integration, and continuous delivery practices.

7 Threats to Validity

External validity: The presented work is derived from the cases studied with teams in the domains of automobile and telecommunication. Some parts of the work can be seen in parts of the company differently. All the terminologies used in the company are normalized and the implementation details are explained with necessary level of abstraction [15]. We do not claim that the opportunities and challenges will be exactly the same for industries from a different discipline. *Internal Validity:* To address internal validity threat, the findings were validated with other teams in the company who were not involved in the study. Further validation can be done by involving more companies, which we see as future work [21].

8 Related Works

This section presents the most related previous studies on data pipeline development and maintenance.

P. O'Donovan et. al describes an information system model that provides a scalable and fault tolerant big data pipeline for integrating, processing and analysing industrial equipment data [17]. The authors explain the challenges such as development of infrastructures to support real-time smart communication, cultivation of multidisciplinary workforces and next-generation IT departments. However, the study is solely based on a smart manufacturing domain. A survey study by C.L.Philip Chen et. al discusses about Big Data, Big Data applications, Big Data opportunities and challenges, as well as the state-of-the-art techniques and technologies to deal with the Big Data problems [7]. A Big Data platform Quarry is proposed by P. Jovanovic et. al [11] manages the complete data integration lifecycle in the context of complex Big Data settings, specifically focusing on the variety of data coming from numerous external data sources.

Data quality challenges and standards/frameworks to assess data quality are discussed in many works [20] [5] [4]. Although there exists significant number of data quality assessment and mitigation platforms, the industrial practitioners experience data quality issues which indicates that the problem is not solved.

9 Conclusions

The multi-case study indicates challenges and opportunities involved in implementing and managing data pipelines. The challenges are categorized into three namely infrastructural, organizational, and data quality challenges. Nevertheless, the benefits data pipeline brings to the data-driven organizations are not frivolous. A data pipeline is a critical element that can also support a DataOps culture in the organizations. The factors inhibiting Data pipeline adoption were mostly concerned with human aspects e.g. lack of communication and resistance to change; and technical aspects e.g. the complexity of development. Suitability of completely automated data pipelines might be questioned for certain domains and industry sectors, at least for now. However, a completely automated data pipeline is beneficial for the domains that can adopt it. Frequent updates are advantageous, but the effects of short release cycles and other data pipeline practices need to be studied in detail. Understanding the effects on a larger scale could help in assessing the real value of data pipelines.

The purpose and contribution of this paper is to explore the real-time challenges of data pipelines and provide a taxonomy of the challenges. Secondly, it discusses the benefits of data pipelines while building data-intensive models. In future work, we intend to further extend the study with potential solutions to overcome the listed data pipeline challenges.

References

1. Batini, C., Rula, A., Scannapieco, M., Viscusi, G.: From data quality to big data quality. In: Big Data: Concepts, Methodologies, Tools, and Applications, pp. 1934–1956. IGI Global (2016)
2. Burnard, P.: A method of analysing interview transcripts in qualitative research. Nurse education today **11**(6), 461–466 (1991)
3. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data science journal **14** (2015)
4. Carlo, B., Daniele, B., Federico, C., Simone, G.: A data quality methodology for heterogeneous data. International Journal of Database Management Systems **3**(1) (2011)
5. Carretero, A.G., Gualo, F., Caballero, I., Piattini, M.: Mamd 2.0: Environment for data quality processes implantation based on iso 8000-6x and iso/iec 33000. Computer Standards & Interfaces **54**, 139–151 (2017)
6. Caveness, E., GC, P.S., Peng, Z., Polyzotis, N., Roy, S., Zinkevich, M.: Tensorflow data validation: Data analysis and validation in continuous ml pipelines. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. pp. 2793–2796 (2020)

7. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: A survey on big data. Information sciences **275**, 314–347 (2014)
8. Chen, G.J., Wiener, J.L., Iyer, S., Jaiswal, A., Lei, R., Simha, N., Wang, W., Wilfong, K., Williamson, T., Yilmaz, S.: Realtime data processing at facebook. In: Proceedings of the 2016 International Conference on Management of Data. pp. 1087–1098 (2016)
9. Davenport, T.H., Dyché, J.: Big data in big companies. International Institute for Analytics **3** (2013)
10. Goodhope, K., Koshy, J., Kreps, J., Narkhede, N., Park, R., Rao, J., Ye, V.Y.: Building linkedin's real-time activity data pipeline. IEEE Data Eng. Bull. **35**(2), 33–45 (2012)
11. Jovanovic, P., Nadal, S., Romero, O., Abelló, A., Bilalli, B.: Quarry: A user-centered big data integration platform. Information Systems Frontiers pp. 1–25 (2020)
12. Kaisler, S., Armour, F., Espinosa, J.A., Money, W.: Big data: Issues and challenges moving forward. In: 2013 46th Hawaii International Conference on System Sciences. pp. 995–1004. IEEE (2013)
13. Marr, B.: Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results. John Wiley & Sons (2016)
14. Marz, N., Warren, J.: Big Data: Principles and best practices of scalable real-time data systems. New York; Manning Publications Co. (2015)
15. Maxwell, J.A.: Designing a qualitative study. The SAGE handbook of applied social research methods **2**, 214–253 (2008)
16. Munappy, A., Bosch, J., Olsson, H.H., Arpteg, A., Brinne, B.: Data management challenges for deep learning. In: 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). pp. 140–147. IEEE (2019)
17. O'Donovan, P., Leahy, K., Bruton, K., O'Sullivan, D.T.: An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. Journal of Big Data **2**(1), 25 (2015)
18. Pathirage, M.: Kappa architecture - where every thing is a stream. http://milinda.pathirage.org/kappa-architecture.com/, (Accessed on 09/28/2020)
19. Raman, K., Swaminathan, A., Gehrke, J., Joachims, T.: Beyond myopic inference in big data pipelines. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 86–94 (2013)
20. Redman, T.C.: Data's credibility problem. Harvard Business Review **91**(12), 84–88 (2013)
21. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empirical software engineering **14**(2), 131 (2009)
22. Singer, J., Sim, S.E., Lethbridge, T.C.: Software engineering data collection for field studies. In: Guide to Advanced Empirical Software Engineering, pp. 9–34. Springer (2008)
23. Verner, J.M., Sampson, J., Tosic, V., Bakar, N.A., Kitchenham, B.A.: Guidelines for industrially-based multiple case studies in software engineering. In: 2009 Third International Conference on Research Challenges in Information Science. pp. 313–324. IEEE (2009)