Dynamically reconfigurable data pipeline in the edge network

Isuru Nuwanthilaka
Department of Computer Science
University of Moratuwa
Sri Lanka
isurun.21@cse.mrt.ac.lk

Gayashan Amarasinghe
Department of Computer Science
University of Moratuwa
Sri Lanka
gayashan@cse.mrt.ac.lk

***Abstract*—This literature review explores the existing research in the domain of data science pipeline as the ground work for the research titled dynamically reconfigurable data pipeline in the edge network.**

***Index Terms*—data science, pipeline, architecture, edge**

I. Introduction

Data pipeline is playing a significant role in the data science landscape. It mainly focuses on extraction, data preparation, cleaning , transforming, training models and visualization. So it requires effective architecture to support these functionalities. In this literature review, existing approaches of data pipelines are explored.

II. Problem

Most of the data pipelines are one-for-all models. So its efficiency might be reduced for some pipeline tasks. In this research, how to dynamically reconfigure the architecture of the data pipeline using the resources available in the edge devices is discussed.

Therefore the following three main questions are being analysed.

1. Dynamic reconfiguration of the pipeline architecture.

2. Optimal/near-optimal utilisation of edge resources to support proposed architecture.

3. Evaluation of algorithms that are benefited from this approach (performance evaluation).

III. Motivation

Nowadays, edge devices are resource rich compared to a few years back. Also there are tools/libs popping up which enable effective use of machine learning (ML) (ex: tensorflow mobile sdk) in the edge devices. Now the developers and the community tend to execute ML tasks on the edge devices which opens us to another research area of ML pipelines in edge network. So this research is carried out to explore the possibilities of devising reconfigurable architectures for ML pipelines at edge.

IV. Literature Survey

The literature is presented under relevant categories based on the context of the each research.

A. Data Pipeline Challenges

1) Manufacturing process data analysis pipelines: a requirements analysis and survey [1]: In this paper researchers have surveyed about the smart manufacturing pipelines. Smart manufacturing is how the digitization of all the manufacturing activities in the physical process occurs. There are several key activities like data ingestion, communication and visualization but researchers have focused more on storage and analysis phases.

Mainly, two questions are being surveyed in this paper.

1. Key requirements for a big data analysis pipeline.

2. Available tools and solutions for pipelines in academic literature.

To answer the first question they have identified key phases and their requirements.

a) Ingestion: This is the main component for entering data to the big data analysis platforms. This opens up several other operations like identification, validation,filtering etc tasks. Integrations to few data sources play a significant role in this domain. It needs to implement several connectors to various platforms/systems and also needs to support data formats. As data generated at different locations it may be in different quality, formats and rates. Some systems may generate real-time data and big data platforms need those data in real time or near real time. Some applications may need those data in batch-wise. So all these issues arise in the data ingestion phase.

b) Communication: Communication channels are much needed component in the data pipelines as it is the medium where data is routed among the different tools in the big data platforms. In some cases, ingestion and communication bind as one tool in the system. The communication should be reliable and secure. In various environments, the communication architectures may differ. For example, in some IoT related systems need publisher-subscriber pattern, some may need pull or push patterns. Also these communication channels should be able to handle the utilization of bandwidth efficiently.

c) Storage: In the days of 1990s, the product processes stored only few mega bytes, but now the situation is lot more changed and 10s of giga bytes are being generated daily in the manufacturing pipelines. So handling this amount of data should be well thought design. Also some data should be stored for years as they may be required for legal or privacy policies. So the real problem is the security and cost of these long term data storage in the modern day cloud platforms. It is needed the compression algorithms and different databases to operate on these data types and their relations. Apart from that, these databases or storage facilities need to provide high speed read and writes as most of the applications directly depend on data.

d) Analysis: The paper describes the analysis requirements mentioned in Qin[2].Support for heterogeneous data types, scaling to 10000s of variables, analysis of imperfect data generated by various heterogeneous sources, support for time-series analysis, data mining and machine learning are mentioned as key requirements.

e) Visualization: Big data visualization differs from traditional data representations due to volume,variety and velocity. Even though the data sets are large,to get accepted by the engineers those representations should be interactive,dynamic and well performing. So handling concurrency, real time visualizations are becoming new challenges.

Based on the requirements mentioned above and some other requirements detailed in the paper they have come up with a literature review for the latest tools and solutions in the industry. First they have defined a criteria of what they are going to look at.

In the paper they have mentioned a set of tools according to the topics. In addition, 6 pipelines have no tool setup for ingestion, 19 have no tools for communication and 13 for visualization.

2) Examining the Challenges in Development Data Pipeline [3]: Today's developing word is more relying on data driven decisions.Therefore, there are many significant and emerging requirements in the data analysis process. In this research, authors have explored the various challenges in several phases in data pipeline development. They have conducted several interviews about the development and from the feedback they got ,they have identified the common challenges like correcting text fields,extracting textual data from different documents like PDF, merging hierarchical data etc.

Authors have identified three main phases ; data collection, data cleansing and data analytics.

a) Data collection: In the data collection phase they have segmented it into two sub topics - data entry and tools. There can be many challenges associated with the data entry. Main way of data entry is human data entry. So human errors are unavoidable.Due to crowd sourcing, there are errors in the data sets - labeling issues, textual errors etc. In some cases, data must me extracted from PDF files, so there can be extraction errors. Those extractions can be erroneous due to bad writings, tool imperfections etc. Sometimes, those data entry people are less skilled and lacking the training, so these human errors are expected. Other challenge is tool support. When it comes to big data tools should be efficient due to volume, variety and velocity. Most of the time these tools have to transform paper form into digitization.Also, in tools, they have described latencies of dashboards and their inability to cater large data sets in visualization.

b) Data cleansing: This is another very important phase. This section include two sub topics;corrections and structuring. Here authors have mentioned the challenges fixing the issue introduced specially in the data entry and prepossessing for analytics. One challenge is replacing values or masking so the privacy of the data is preserved. But this should be done consciously as it may make the data useless. Apart from that, removing duplicates, unit conversions, fix spelling mistakes, clean up open texts are some key challenges. In the structuring authors have mentioned merging data from different sources, restructuring data formats, code conversion/terminology mapping, splitting the aggregations as challenges.

c) Data analytics: In this section, two important sub topics are mentioned;calculations and visualizations.In the calculations the main challenges are to find the missing data, identify the data outliers, calculate accuracy for test data and fine tune values to remove the biases in data sets and training. Authors have mentions visualizations challenges as eyeball data for data accuracy, calculate daily/weekly/real-time reports generation.

As for future work, this paper suggests interviewing more diversified resource persons and tech-savvy organizations.

3) Data Pipeline Selection and Optimization[4]: In this paper, author discusses the data pipeline hyper-parameter optimization problem as a standard optimization problem. Data preprocessing acts as a significant component in the data pipeline. They have focused on selecting data pipelines and optimizing it. Their main contributions to the field from this paper are,

1. Impact of the data pipeline configurations on classification problems.

2. Define data pipeline selection and optimization problem.

3. Analyse the impact of Sequential Model Based Optimization (SMBO) for the above mentioned problem.

4. Discuss about optimal configurations.

In the future work they have identified that this could be evolved on an online version using streaming concepts.

4) Data Life cycle Challenges in Production Machine Learning: A Survey[5]: The authors have presented research challenges in the domain of production machine learning taken from their experience working at Google. The main focus is on three main areas - data understanding, data validation and cleaning , and data preparation.

They have talked about different dimensions of production ML.

1. User interacting with ML platforms - different roles of people involving in data ML pipelines.

2. Data lifecycle in an ML pipeline - prepare,train and evaluating, validating ,clean,serve.

3. Data understanding. a. Sanity check - data in expected shape b. Analyses for launch and iterate - feature based, data lifecycle

4. Data validation and cleaning. a. Alert tradeoffs b. Alert categories

5. Data preparation.

6. Feature engineering and selection.

7. Data enrichment.

In the last section they have identified that the following are the lessons learned.

1. There are many data management challenges beyond data flow optimizing.

2. Making realistic assumptions when developing ML products.

3. Identification of users diversified needs.

4. Keep smooth integration to development workflow.

B. Data Pipeline Architectures

1) Putting Data Science Pipelines on the Edge[6]: In this research paper, authors have come up with a new architecture named as "Just in Time architecture for data science (JITA-4DS)". Their main concern is that one-for-all architectures are not efficient enough to cater the requirements of different data science pipelines.In addition to that existing IT architectures are not well designed to provide on-demand infrastructures.Those one-fits-all architectures assumes,

1. High reliability and the high availability of the network

2. High energy and economically consuming resources

So considering the agile nature of the pipelines they are trying to provide an algorithm and resource management techniques for reconfiguring disaggregated data centers. Their concern is about the configurations on the cloud infrastructure.

This research has a significant impact on the proposed research as it discusses mainly about dynamic reconfiguration on cloud leaving opportunities on the edge network.

The work related to this paper is an extended version of two types of approaches.

1. Disaggregated data centers as a alternative for one-fits-all architectures

2. Data science pipelines's execution platforms in the cloud.

Currently they are addressing the challenges in the resource management of simple environments and simulation of different levels of JITA-4DS.

2) Modelling Data Pipelines[7]: In this paper , authors mainly propose a new conceptual model of data pipeline. In the first part of the paper they are discussing the importance of data products- APIs, dashboards etc. Addition to that, data collection in distributed set up, semi structured data sources and ETL/ELT pipelines are identified as important. As a result of the proposed model they try to improve the efficiency of the pipeline to reduce the latency in the development of data products.

This paper has three main parts. In the first part they conduct interviews with professionals and organizations to identify challenges in data collection pipeline for data analytics, building data pipelines for data governance and ML pipeline areas. Then, in the second part they analyse feedback and propose a new conceptual model. That model consists mainly of data sources, data collection,data lakes,data ingestion,data processing,data staging,data warehousing,data labeling,data preprocessing and ML/DL modeling phases.

In the last part, authors have discussed the validity of this model taking several cases and have given suggestions to improve.

3) Scalable data pipeline architecture to support the industrial internet of things Moneer[8]: Industrial internet of things(IIOT) is a growing field. With the expansion of data sources, data formats ,scaling up of data processing and distribution - scalable data pipelines have become challenging.In this research paper they are proposing an architecture for scalable data pipeline to data processing and distribution data from various different sources.

When designing the architecture they have identified few important facts.One concern is formatting data into standard data types. With this approach they can scale data sources using the same interface to the pipelines. Next they focus to synchronize time stamps across multiple sources so it orders the events accordingly.Also they try to enable deployments on embedded systems and legacy systems. Apart from these, managing continuity of the data streams, enable data persistency, route data flows into multiple endpoints are some of other concerns in designing the architecture.

In the following section authors have proposed an architecture considering the above mentioned concerns. This include a middleware which is an architectural pattern that integrates hardware and software to enable two functionalities; moving data from different data sources and distributed in scalable manner.

Front part of the architecture consists of different connectors to MTC components and ad-hoc sources, then they talk through MQTT clients to the middleware. Middleware part consists of MQTT broker and Apache Kafka. Those connect to data applications and different flavours of databases.

To evaluate the architecture, they have used Hurco VMX24 device and MTConnect devices. They have not specifically mentioned the result data in the paper.

4) Feedback Driven Improvement of Data Preparation Pipelines[9]: Data preparation is the laborious work which requires huge effort. This adds upfront cost to the data analysis.Some data preparation tasks are automated and it is important to have the ability to refine the decision of preparation on the feedback from the users of data products.In this research author explore the approaches to revise the results of diverse data preparation components based on the feedback about the correctness of the value in the data products.

Data preparation takes 80% of data scientist's usual work where they like to spend more time on data analysis part.In preparation phase, according to researchers, it has several steps ; discovery,profiling,matching,mapping,format transformation and entity resolution.

In the next section they have mentioned VADA[10] architectural aspects which is relevant to their proposing feedback driven architecture. Authors have mentioned important aspects like input data sources,target schema,user context and data context.

Proposed approach consists of mainly two steps.

1. Given the feedback, identify a collection of hypotheses which explain the feedback.

2. Aspects of a hypothesis in the above collection should be reviewed matching to the current preparation pipeline and fine tune.

In the next section , they have come up with mathematical models and pseudo codes of the algorithms to solidify their proposed approach on selecting hypotheses.

5) An Edge-Based Framework for Enabling Data-Driven Pipelines for IoT Systems[11]: Internet of Things (IoT) generates large data nowadays as the number of devices get multiplied day by day. In this research, authors proposing a new architecture named R-Pulsar which tries to enable the cloud capabilities in the edge devices. They propose it as an edge-based framework for data pipelines.They have focused on IoT devices because by 2030 there will be 500 Bn Cisco devices only and nearly half of the data come from the sensors means edge network. They have mentioned that precision medical devices/applications for continuous monitoring of medical instruments, healthcare sensors related to lifestyle and urban mobility applications will be transformative application this era.

As the programming abstraction , they have discussed what,when and where data needs to be processed.They have taken smart city and disaster response application as cases.There are two existing models in this field; edge data analytic and cloud-edge programming models. In the edge data analytics they have mentioned about some tools like GeeLytics which is designed to dynamical stream processing, AWS Greengras, IBM Watson IoT and Azure IoT. For cloud and edge programming models they have taken about MapReduce technique, Apache Hadoop, Apache Spark as tools. Most of the tools are designed to be in cloud as it is resource rich and scalable.

The R-Pulsar is a framework designed to be a software stack that extends cloud capabilities to edge. This uses a distributed architecture which implemented in several layers.

a) Location aware overlay network layer: IoT data comes from different locations. So this temporal and spacial information related to data should be processed properly according to the locations. Nodes in the architecture considered as Rendezvous Point (RP). Those RPs are grouped according to there geological placements. This layer acts as a lookup to route the messages to relevant RP group.

b) Content-based Routing Layer: This layer builds upon the previous layer. This is responsible for delivery of the messages based on the content. They have used two separate methods,

1. Routing using simple keyword tuples - use IDs

2. Routing using complex keyword tuples - use wildcards, ranges

c) Serverless Messaging Layer: Serverless computing is a cloud computing model which abstracts the management of infrastructure and make the developer's life easy. In the context , this layer is responsible for post the AR message with profile routed from the previous layer. This layer has two main components ; matching engine and profile manager.

d) Memory-mapped Streaming Analytics Pipeline: This layer is implemented such that it collects data from multiple data sources, process data and make them available to use. This consists of few other layers. Data collection layer gathers data from different source and make available to the pipeline. Service like Apache Kafka, Google Pub/Sub, Amazon Firehose and Mosquitto are mentioned as possible tools for this layer. Next the data processing layer is used to process and perform computation on the collected data. This R-Pulsar is validated using Apache Edgent. Other layer is data storage and query layer which enable SQL-like querying. R-Pulser relies on RockDB which is a key-value database.

e) Rule-based Programming Abstraction: This rules define when data must be sent to cloud.Rule engine consists of IF-THEN which gives developer the ability to conditionally call the processing tasks.

The evaluation is done using three environments ; Raspberry Pi system, android system and cloud system. The system specification is mentioned in the paper. In the result section, authors have mentioned that they have seen 20% increase in performance and faster completion of the tasks with bandwidth reduction of 82%. In future work they suggest to support to advanced storage strategies ,reduce the cost associate with data movements and improve the energy efficiency.

6) On the Design and Architecture of Deployment Pipelines in Cloud- and Service-Based Computing – A Model-Based Qualitative Study[12]: In this research,authors have proposed a formal architecture to DevOps and continuous delivery (CD) pipelines. Even though there is a huge rise regarding this, there is no formal discussion about the design and architecture of DevOps and CD pipelines. So authors have derived a formal architecture and verified that it increase the precision and the efficiency compared to the informally modelled pipelines.

They have collected in-depth study of 25 informal deployment practices from the industrial practitioners. To create the ground work , they have collected details on three main questions; What are recurring practices for designing deployment pipeline structures? What are the relevant environments in deployment pipelines? and What are the architectural elements relevant for building a deployment pipeline infrastructure?

In the following sections of the paper, they have modeled mathematically the architecture of the devops pipeline.

7) Edge Based Data-Driven Pipelines (Technical Report)[13]: This research report investigates about the edge on-device stream processing platform which enables the serverless computational capabilities across the cloud and edge uniformly. This work has evaluated the previously discussed R-Pulser framework[11].

8) Review of social media analytical process and Big Data pipeline[14]: Social media data analytic is another rising research area. With the current expansion of technologies previous mere consumers are turned into social data producer leading to big data use cases. Now it is very important to extract social insights to different fields like business, marketing , advertising, tourism and many more.

As the basis for this research, authors have formed up two main questions.

1. What types of challenges researchers have in analysing social media data?

2. How could big data technologies can be integrated to solve those challenges?

Next they have talked about Big Data and Big Data Vs; volume, velocity,variety. The survey has done database-oriented manner and the inclusion rule was to paper to be in between 2008 and 2018.Next they have summarised their findings in a field-wise manner.

In the following sections, authors have described five distinct steps for social media analytics; acquisition and recording, information extraction and cleaning, data integration aggregation and representation,query processing data modeling and analysis, big Data interpretation.

Also, authors have summarized the challenges they observed related to the big data 4V's.Considering those, they have come-up with social media analytic steps. Overall they have studied the joint interaction between social media analytic pipelines and big data.

9) Data Pipeline Architecture for Serverless Platform[4]: In this paper, authors propose a novel data pipeline architecture to decompose monolithic data pipeline into independently deployable pipeline components so that it can be scaled up/down , monitored and managed properly to fulfil the incoming large data sets. This architecture is proposed for the cloud environment. They have come up with a model using TOSCA language which is recently developed standard for cloud based applications.

To implement the model they have used Apache NiFi and suggested Amazon data pipeline as an alternative tool. Their abstracted PipelineBlock model consists of three main groups;SourcePB,MidwayPB and DestinationPB.For every pipeline , PB is created and it acts as a black box. SourcePB used to get data from local and remote data streams. MidwayPipelineBlock is for intermediate data processing and analytics. DestinationPB is used to publish data similar to SourcePB but different direction. TOSCA model has relevant node types to match these components.

Authors have identified Viarota as a potential used case for this model which is a mobile and web-based cloud application.In future work they have mentioned about expanding the TOSCA node types to support data movements.

10) Pipeline architecture for mobile data analysis[15]: This paper presents a cloud architecture which provides support to collect and analyse mobile data. With their pipeline architecture users can extract data like multimedia data, location information and barcodes. Also this system provide easy UI components so anyone coming to this system needs not to have programming knowledge.Authors have identified that even thought the advancements in hardware and software, still people struggle to create apps with less lead time or less resources. They have come up with this new project *Maritaca* as a possible solution for above issue. It provides the flexibility to create apps easily. Initially researchers have explored the existing tools like App Inventor, Nokia Data Gathering, Open Data Kit (ODK),DoForms,Mafuta Go,Fulcrum etc but as per the authors none of them provide data analysis support.

The Maritaca project consists of two main components;mobile component and server component. Mobile component is an Android application and it is an engine of interpreter design pattern - it interprets the xml description. Next one is the server component. It has several items. *Form editor* is drag and drop supported web UI written in HTML and AJAX. *Analytics editor* is also a web UI which allows to query the collected data. *Cassandra database* is used as a scalable database. *Hadoop file system* is used to unstructured data and multimedia file across distributed file system.*Solr Engine* enables searching in apps. These components provides several features.

In the future work, authors have mentioned that they are going to expand its infrastructure with cloud resources without compromising its performance. Addition to that they hope to implement it on iOS as an extension to current Android development.

C. Data Pipeline Security

1) Integration of security standards in devops pipelines[16]: In the past few years, devops pipelines adopted to the industrial control systems (ICS) in order to automate the product swim-lanes and reduce the lead time to the market. So it can improve the customer expectations and increase the development time rather than focusing on operations. But in the modern technology world there are lots of security gaps in the pipelines so in this paper, authors introduce methods to implement security standards to the devops pipelines without affecting lead time much.

In the paper two main contributions added to the field,

1. Explore the systematical approach to devops pipelines to compliance with IEC 62443-4-1 standard

2. Describe the automation capabilities for 4-1 standards with the available tools in the industry.

In the ICS domain it is important to adopt IEC 62443-4-1 standard. With security concern, now new area called 'DevSecOps' emerged. In the IEC 62443-4-1 standard ,it has proposed eight practices;security management(SM),specification for security requirement (SR),secure design (SD), secure implementation (SI), security validation and verification testing (SVV), management of security related issues (DM),security update management (SUM),security guidelines (SG). This work presents how each practice applied to devops phases.

To apply above security practices, authors suggest three main activities,

a) Describe standard requirements as activities: This mainly focuses on detailed analysis of 4-1 standard and its requirements. After identifying the standard activities they define the 4-1 standard process models.

b) Determine 4-1 standard automation capabilities: Even though they are trying to adopt standards they might not fit properly in the automation environments thus needed to be evaluate the automation capabilities without affecting to the lead time.

c) Map activities in to pipeline stages: In this phase they are going to map the 4-1 standard pipeline specifications to Security Standard Complaint (S^2C) DevOps pipeline specifications.

Finally they have evaluated their qualitative study at Siemens AG which is main contributor to 4-1 standard in ICS. They have interview persons responsible for security of pipelines and have tried to answer the questions related to precision and usefulness.

In future work, they suppose to build tools to increase the percentage of 4-1 standard requirements that can be fully automated in the ICS domain.

2) Security Support in Continuous Deployment Pipeline[17]: Continuous delivery pipelines (CDP) lack security tactics. In this paper author have researched on the security tactics that can be used in the CDP without affecting to the existing behavior. They have done penetration test to the CDP components and have observed that it increases the security of the components like repository,continuous integration server and main server.They have identified several security risks in the repository, main server and CI server. For examples, uncontrolled access categorized as a repository risk, poor authentication mechanisms and uncontrolled access mentioned as main server (AWS) risk and starting with previously infected version or replay identified as CI server (Jenkins) risks.

In the next section they have proposed possible security tactics which come after the thorough analysis of previous security risks.

1. Limiting the access of the repository branches so only selected group can commit/merge code to default build branch.

2. Securing main server access with private key along with SSH. AWS supports this.

3. Use role base access control (RBAC) for main server. AWS has IAM ecosystem.

4. Setting up clean VM environments for build phase in jenkins CI server.

5. Use jenkins role plugin to control access so that the users can have limited access to delete, create, modify pipelines.

In the final section they have summarized the quantitative and qualitative results of this proposed system. Using tools like OWASP Zed Attack Proxy (ZAP) they have scanned security vulnerabilities in the system and has done some penetration tests.

V. Conclusion

In this short essay, we have explored the existing work related to data science pipeline in different perspectives;

1. Data pipeline challenges and requirements.
2. Pipeline design and architectures.
3. Data and delivery pipeline security.

By looking at the literature, few key knowledge gaps are identified. One important area is data pipeline architecture which seems to be having less research done as per our knowledge. We are inspired from the research work of 'Putting Data Science Pipelines on the Edge[6]' and 'An Edge-Based Framework for Enabling Data-Driven Pipelines for IoT Systems[11]'. Following them, we propose a research regarding the possibilities of re-configurable architectures on the edge for data pipelines.

References

[1] A. Ismail, H. L. Truong, and W. Kastner, “Manufacturing process data analysis pipelines: a requirements analysis and survey,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–26, 2019.

[2] S. J. Qin, “Process data analytics in the era of big data,” *AIChE Journal*, vol. 60, no. 9, pp. 3092–3100, 2014.

[3] F. Pervaiz, A. Vashistha, and R. Anderson, “Examining the challenges in development data pipeline,” *COMPASS 2019 - Proceedings of the 2019 Conference on Computing and Sustainable Societies*, pp. 13–21, 2019.

[4] C. Dehury, P. Jakovits, S. N. Srirama, V. Tountopoulos, and G. Giotis, *Data pipeline architecture for serverless platform*, vol. 1269 CCIS. Springer International Publishing, 2020.

[5] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, “Data lifecycle challenges in production machine learning: A survey,” *SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018.

[6] A. Akoglu and G. Vargas-Solar, “Putting Data Science Pipelines on the Edge,” pp. 1–13, 2021.

[7] A. Raj, J. Bosch, H. H. Olsson, and T. J. Wang, “Modelling Data Pipelines,” *Proceedings - 46th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2020*, pp. 13–20, 2020.

[8] M. Helu, T. Sprock, D. Hartenstine, R. Venketesh, and W. Sobel, “Scalable data pipeline architecture to support the industrial internet of things,” *CIRP Annals*, vol. 69, no. 1, pp. 385–388, 2020.

[9] N. Konstantinou and N. W. Paton, “Feedback driven improvement of data preparation pipelines,” *Information Systems*, vol. 92, 2020.

[10] N. Konstantinou, M. Koehler, E. Abel, C. Civili, B. Neumayr, E. Sallinger, A. A. Fernandes, G. Gottlob, J. A. Keane, L. Libkin, and N. W. Paton, “The vada architecture for cost-effective data wrangling,” in *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, (New York, NY, USA), p. 1599–1602, Association for Computing Machinery, 2017.

[11] E. G. Renart, D. Balouek-Thomert, and M. Parashar, “An edge-based framework for enabling data-driven pipelines for IoT systems,” *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2019*, pp. 885–894, 2019.

[12] U. Zdun, E. Ntentos, K. Plakidas, A. El Malki, D. Schall, and F. Li, “On the design and architecture of deployment pipelines in cloud-and service-based computing-a model-based qualitative study,” *Proceedings - 2019 IEEE International Conference on Services Computing, SCC 2019 - Part of the 2019 IEEE World Congress on Services*, pp. 141–145, 2019.

[13] E. G. Renart, D. Balouek-Thomert, and M. Parashar, “Edge Based Data-Driven Pipelines (Technical Report),” 2018.

[14] H. Sebei, M. A. Hadj Taieb, and M. Ben Aouicha, “Review of social media analytics process and Big Data pipeline,” *Social Network Analysis and Mining*, vol. 8, no. 1, 2018.

[15] A. F. Conceição, J. V. Sánchez, B. G. Dos Santos, D. Vieira, and V. Rocha, “Pipeline architecture for mobile data analysis,” *International Conference on Information Networking*, no. February 2015, pp. 492–496, 2014.

[16] F. Moyón, R. Soares, M. Pinto-Albuquerque, D. Mendez, and K. Beckers, “Integration of Security Standards in DevOps Pipelines: An Industry Case Study,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12562 LNCS, pp. 434–452, 2020.

[17] F. Ullah, A. J. Raft, M. Shahin, M. Zahedi, and M. A. Babar, “Security support in continuous deployment pipeline,” *ENASE 2017 - Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering*, pp. 57–68, 2017.