Revision to the Paper ID b5d6583f-1abe-4325-9459-588bcec2cdad Maximizing Data Quality While Ensuring Data Protection in Service-Based Data Pipelines

Antongiacomo Polimeno, Chiara Braghin, Marco Anisetti, Claudio A. Ardagna

We would like to first thank the associate editor and reviewers for their valuable and helpful reviews. We revised the paper according to their comments. In this letter, we provide 3 different sections (one for each of the reviewers), containing our replies and the rationale for our modifications.

Please note that in the reviewers' comments, sections and page numbers refer to the old version of the paper, whereas in our replies they refer to the revised version of the paper. For reviewers' convenience, major modifications have been highlighted in blue in the revised paper and in this reply letter.

As a general remark, following the reviewers' comments on the need to *i)* clarify the relevance of the reference scenario and the novelty of the proposed approach, *ii)* clarify the utility of our quality metrics and the soundness of corresponding experimental evaluation, and *iii)* compare our approach with a baseline modeling solutions in the state of the art, we substantially revised the paper as follows.

- We revised Section 2 to better position our framework, emphasizing that our approach is agnostic to specific orchestrators and the DAG-based modeling serves as an abstraction for service selection and pipeline instantiation, rather than as a means to model an executable pipeline to be fed in the service orchestrators. Additionally, we revised Section 2.1 providing a more detailed description of our reference scenario clarifying its relevance in real-world settings, with reference to the adopted data and the considered application domain (i.e., law enforcement).
- We extended Section 5.1 to clarify the rationale behind the selection of the two data quality metrics used in our experimental evaluation, and their link to the quality dimension completeness. We also clarified that our approach is agnostic to quality metrics and dimensions.
- We extended Section 6.1 with the definition of a baseline solution modeling the approaches in literature. We compare the quality and performance of our heuristic with the ones of the baseline. In addition, for solutions where a direct experimental comparison was impossible, we extended Section 7 (related work) with a discussion and a table (Table 3) comparing our heuristic and main (industrial) solutions according to different evaluation criteria.
- We added new Section 6.4 providing a discussion of the main lessons learned emerging by our experimental results.
- We extended Section 8 with a discussion on our future directions at all layers of our approach: quality metrics and dimensions, wider evaluation and real-world assessment, integration of our approach with existing QoS approaches, and dynamic service selection.
- We revised Figure 2 to better illustrate our framework, and added Figure 3 providing the pseudocode of the service selection algorithm.

In the process of reviewing the paper, we substantially refined and improved the introduction, better positioning our approach with respect to existing state of the art and the research area on data quality and protection. We then revised the overview of the state of the art and the terminology to make it consistent throughout the paper.

We believe our changes now clarified the positioning of our methodology with respect to the state of the art, its novelty and contribution, and the advantages it brings. We also believe they increase the readability of the paper, clarifying our methodology, implementation choices, and experimental results.

1 Response to Reviewer 1

Comment: The paper is well-written and coherent with many good aspects. The division of work is also well-motivated. The experimental results are fair (no lessons learned).

Response: *Thanks for your comment. We added the new Section 6.4 discussing the main lessons learned emerging from our experimental results as follows.*

> "The experimental results we obtained yield several valuable insights that merit further discussion. Three key observations emerged as follows.
>
> **Trade-off Between Execution Time and Quality.** As expected, the execution time improvement provided by our heuristic introduces a loss of quality with respect to the exhaustive approach. This loss causes an increase in the quality variance, especially when the window size ($|w|$) is small compared to the vertex count. A fine-grained tuning of heuristic parameters is needed to balance computational efficiency and data quality.
>
> **Impact of Window Size on Quality.** Our heuristic well approximates the exhaustive approach. While larger window sizes generally lead to better performance, there exists a breakpoint where the balance between window size and performance is optimal. Beyond this point, the incremental gain in data quality modeled by our quality metrics does not justify the additional computational effort introduced by larger window sizes. We also note that lower window sizes are more unstable, especially when the *wide* setting is used, meaning that the data quality varies significantly among different configurations. This effect stabilizes at larger window sizes, approximately half the size of the pipeline (e.g., $|w|=l/2$).
>
> **Sliding Window Approach versus Global Awareness.** The intrinsic nature of our sliding window heuristic can sometimes lead to a local optimum, as the window size limits the candidate services for each pipeline stage to a restricted subset, which may prevent reaching the global optimum. This aspect is maximized when using the baseline representing the state of the art, where the sliding window heuristic is configured with a window size of $|w|=1$. Additionally, as dependencies between services increase, the likelihood of finding a sub-optimal solution rises. Our experiments show that *i)* increasing the window size helps mitigate this issue and *ii)* a broader decision-making scope becomes essential as service dependencies grow more complex."

Comment: The main limitation, from my point of view, is the modeling with DAGs. In this case, you assume that one service ends, but generally we add deadlines to better control the execution time. Please, make a positioning.

Response: *We do agree with the reviewer that our modeling built on DAGs cannot properly model the control flow of the pipeline. However, the objective of our pipeline template and instance is to describe an abstraction of the service-based pipeline structure used to model the* data flow *among services, and not the executable version of the service pipeline itself.*

Our DAG-based structure is then used as the basis of our service selection process in Sections 3 and 4 that selects the set of services to be integrated within the pipeline to maximize data quality while ensuring data protection requirements.

Upon the service selection ends and the pipeline instance is generated, the most appropriate solution to transform the pipeline instance in an executable pipeline can be selected. We note that our approach is agnostic to the specific executable environment (orchestrator), though this last part is out of the scope of this paper. We also note that for orchestrators like Apache Airflow, which use DAGs to manage and visualize complex pipelines, our DAG-based abstraction represents an executable representation of the service pipeline.

We clarified this aspect when presenting our system model in Section 2.1 as follows.

> "We consider a service-based environment where a service-based data pipeline (service pipeline in the following) is designed to analyze data. Our service pipeline is enriched with metadata specifying data protection requirements and functional specifications, and models the data flow among component services, without posing any restrictions on the control flow."

We also clarified in the introduction that our goal is the selection of those services that will compose the service pipeline (where the DAG only represents its abstract view), while the assembly of the executable instance of the pipeline is out of the scope of this paper as follows.

> "Based on the aforementioned considerations, we propose a data governance framework for service-based data pipelines. The primary objective of this framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding privacy and security requirements.[1]"

Comment: The authors do not discuss the performance guarantee for the heuristics.

Response: *Our experiments extensively evaluated the performance (execution time) and quality of our heuristic, comparing them against the ones retrieved by a baseline modeling the state of the art and the exhaustive approach. The goal of our experiments is summarized in Section 6 as follows.*

> "We experimentally evaluated the performance and quality of our methodology (heuristic algorithm in Secontion 5.3), and compared it against the exhaustive approach in Section 5.2 and our baseline modeling solutions in the state of the art in Section 6.1. In the following, Section 6.1 presents the simulator and experimental settings used in our experiments; Section 6.2 analyses the performance of our solution in terms of execution time; Section 6.3 discusses the quality of the best pipeline instance generated by our solution according to the metrics M_J and M_{JSD} in Section 5.1. "

We also extended Section 6.1 to clarify that our simulator enables the evaluation of the performance and quality of our sliding-window heuristic against both a baseline modeling the state of the art (a greedy approach where each node in the pipeline is instantiated independently) and the exhaustive approach providing the optimum pipeline instance.

> "Our simulator also supports the comparison of the performance and quality of our sliding-window heuristic with *i)* a baseline modeling solutions in the state of the art and *ii)* the exhaustive approach (i.e., the theoretical optimum). We modeled our baseline as a greedy approach that, for each node of the service pipeline, selects the best service that maximizes the data quality, while addressing data protection requirements in annotation λ. The reason is that, to the best of our knowledge, existing (industry) solutions and standards do not support service-based data pipelines and are therefore unable to instantiate the service pipeline according to the pipeline structure and service dependencies. We therefore defined our baseline as the sliding window heuristic configured with window size |w|=1. We implemented the exhaustive approach calculating the theoretical optimum as the sliding window heuristic configured with window size |w|=l, to illustrate the potential efficiency of our heuristics within realistic computational limits."

Comment: Add references for GDPR and HIPAA
What do you mean by PCA at the bottom of page 12?

[1] We note that, the assembly of the selected services in an executable pipeline is out of the scope of this paper. Our approach however is agnostic to the specific executable environment.

Bottom of page 25: I do not understand the discussion. Could you clarify?

Response: *All done, thanks!*

Comment: Table 1 is not necessary. The accompanying text is enough.

Response: *Thanks for your comment. However, we decided to keep Table 1 in the paper because, though the accompanying text might be enough in Section 2, it could increase the overall readability of the examples in the remainder of the paper. Please let us know if you still consider Table 1 unnecessary and we will remove it.*

Comment: The authors also need to enforce the Future works. We have no idea in the current text.

Response: *Following the reviewer's comment, we added our future work in Section 8 as follows.*

> "The wide success and adoption of cloud-edge infrastructures and their intrinsic multitenancy radically change the way in which distributed systems are developed, deployed, and executed, redefining IT scalability, flexibility, and efficiency. Multitenancy in fact enables multiple users to share resources, such as computing power, storage, and services, optimizing their utilization and reducing operational costs.
>
> The increasing ability of collecting and managing huge volume of data, coupled with a paradigm shift in service delivery models, has also significantly enhanced scalability and efficiency in data analytics. Data are treated as digital products, which are managed and analyzed by multiple services orchestrated in pipelines. This shift is fostering the emergence of new platforms and environments, such as data marketplaces and data spaces, where data in critical domains (e.g., law enforcement, healthcare, transportation) can be pooled and shared to maximize data quality and trustworthiness, and distributed data management systems supporting data storing, versioning, and sharing for complex analytics processes.[2]
>
> The flip side of a scenario where service-based data pipelines orchestrate services selected at run time and are delivered in the cloud-edge continuum is the increased complexity in data governance. Data are shared and analyzed by multiple services owned by different providers introducing unique security challenges. On one side, the pipeline owner and data providers have different security requirements, access policies, and data sensitivity that vary according to the specific orchestrated services; on the other side, orchestrated services (data consumers) have different profiles that impact on the amount of data they can access and analyze."

---

[2] https://joinup.ec.europa.eu/collection/elise-europeanlocation-interoperability-solutions-e-government/glossary/term/data-marketplace, https://internationaldataspaces.org/, https://digitalstrategy.ec.europa.eu/en/library/staff-working-documentdata-spaces

2 Response to Reviewer 2

Comment:

The paper "Balancing Data Quality and Protection in Distributed Big Data Pipelines" presents a framework addressing an important issue—balancing data quality and protection.

However, the balance between quality and protection (like anonymization) has been explored quite a bit already in areas like privacy-preserving data mining, differential privacy, and secure multiparty computation. Established frameworks (e.g., k-anonymity, l-diversity, differential privacy) already address this trade-off, and this paper doesn't seem to push those concepts forward significantly.

Response: *We do agree with the reviewers that other solutions exist in the literature (e.g., k-anonymity, l-diversity, differential privacy) that permit to configurably manage anonymity/privacy and their impact on the overall data quality, for instance, by tuning the value of k for k-anonymity. However, the contribution and novelty of our proposal lie in the definition of a data governance framework for service-based data pipelines. The primary objective of this framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding privacy and security requirements.*

More in detail, our work focuses on different requirements that make existing frameworks and methodologies suboptimal or inapplicable, and therefore supports greater flexibility and broader applicability. In the following, we summarize how existing solutions and our approach address the requirements target of our paper.

i) **Requirement:** *support a scenario built on service-based data pipelines, where pipelines orchestrate services selected at run time and are delivered in the cloud-edge continuum.*
Existing approaches *traditionally consider monolithic systems/services. They do not support the orchestration of service-based data pipelines through a service selection process driven by data quality and protection requirements.*
Our approach *proposes a data governance framework that supports the instantiation of service-based data pipelines with services owned and operated by different parties.*

ii) **Requirement:** *provide a service selection process that maximizes the data quality across the entire pipeline, while ensuring data protection requirements.*
Existing approaches *are traditionally applied to single datasets managed by a single data consumer (e.g., a service or a software). As a consequence, they do not support the quality and privacy of data across the entire lifecycle of the service-based data pipeline.*
Our approach *supports the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding privacy and security requirements. Service selection is done according to the data flow in the pipeline.*

iii) **Requirement:** *being agnostic to specific data protection frameworks.*
Existing approaches *focus on vertical solutions strongly bound to specific data protection frameworks (e.g., anonymity).*
Our approach *is built on policies that implement data transformations to address data protection requirements. Data transformations are agnostic to the specific data protection framework over which the transformation is built. We note that existing techniques can be easily integrated into our approach to support service selection and policy enforcement.*

iv) **Requirement:** *automatically tune data protection requirement enforcement according to the service profile.*
Existing approaches *do not support this requirement being built for scenarios considering monolithic systems/services.*
Our approach *provides a solution based on policies that enforce data protection requirements according to the service profile including, for instance, the service owner, the data sensitivity, the type of service.*

According to the above discussion, to better clarify the contribution and novelty of the paper, We substantially revised the introduction describing:

- *the considered scenario:*

  > "The wide success and adoption of cloud-edge infrastructures and their intrinsic multitenancy radically change the way in which distributed systems are developed, deployed, and executed, redefining IT scalability, flexibility, and efficiency. Multitenancy in fact enables multiple users to share resources, such as computing power, storage, and services, optimizing their utilization and reducing operational costs.
  >
  > In this scenario, the increasing ability of collecting and managing huge volume of data, coupled with a paradigm shift in service delivery models, has also significantly enhanced scalability and efficiency in data analytics. Data are treated as digital products, which are managed and analyzed by multiple services orchestrated in data pipelines. This shift is also fostering the emergence of new platforms and environments, such as data marketplaces and data spaces, where data in critical domains (e.g., justice, healthcare, transportation) can be pooled and shared to maximize data quality and trustworthiness, and distributed data management systems support data storing, versioning, and sharing for complex analytics processes.[3]
  >
  > The flip side of a scenario where data pipelines orchestrate services selected at run time and are delivered in the cloud-edge continuum is the increased complexity in data governance. Data are shared and analyzed by multiple services owned by different providers introducing unique security challenges. On one side, the pipeline owner and data providers have different security requirements, access policies, and data sensitivity that vary according to the specific orchestrated services; on the other side, orchestrated services (data consumers) have different profiles that impact on the amount of data they can access and analyze."

- *a brief discussion of the shortages of existing solutions:*

  > "Adequate measures such as encryption, access control mechanisms, and data anonymization techniques have been implemented to protect data against unauthorized access and ensure compliance with regulatory requirements such as GDPR [1] or HIPAA [2]. However, data quality is also crucial and must be guaranteed, as the removal or alteration of personally identifiable information from datasets to safeguard individuals' privacy can compromise the accuracy of analytics results.
  >
  > So far, all research endeavors have been mainly concentrated on exploring these two issues separately: on one hand, *data quality*, encompassing accuracy, reliability, and suitability, has been investigated to understand the implications in analytical contexts [3,4]. On the other hand, *data security and privacy* focused on the protection of confidential information and adherence to rigorous privacy regulations [5-8]. Although extensively studied, these investigations often prioritize enhancing the quality, security, and privacy of source data rather than ensuring data quality, security, and privacy throughout the entire processing pipeline, or the integrity of outcomes derived from data. "

- *the requirements and questions driving our work:*

  > "A valid solution requires a holistic approach that integrates technological solutions, organizational policies, and ongoing monitoring and adaptation to emerging threats and regulatory changes across the entire pipeline lifecycle. The implementation of robust access control mechanisms or privacy techniques, ensuring that only authorized users can access specific datasets (or a portion thereof) is just a mandatory but initial step. Additional requirements are emerging. First, data protection requirements should be defined at each stage of the pipeline, potentially integrating techniques like data masking and anonymization (e.g., k-anonymity, l-diversity, differential privacy) to safeguard sensitive information, thereby preserving data privacy while enabling high quality data sharing and analysis. Second, data lineage should be prioritized, fostering a comprehensive understanding and optimization of data flows and transformations within complex analytical ecosystems. Third, data protection and data quality requirements should

[3] https://joinup.ec.europa.eu/collection/elise-europeanlocation-interoperability-solutions-e-government/glossary/term/data-marketplace, https://internationaldataspaces.org/, https://digitalstrategy.ec.europa.eu/en/library/staff-working-documentdata-spaces

drive the process that builds a pipeline with maximum data quality, while addressing data protection requirements.

When evaluating a solution meeting the above criteria, the following questions naturally arise:

i) How does a data protection solution affect data quality in the pipeline? How can we minimize this impact thus maximizing the overall data quality?

ii) Should data protection be implemented at each pipeline step rather than filtering all data at the outset?

iii) In a scenario where service-based data pipelines are built by selecting the best services among various candidate services, how might these choices be driven by quality requirements?"

- *and the contribution of our work:*

> "The primary contributions of the paper can be summarized as follows: *i)* we define a data governance framework that implements an algorithm for the selection of data processing services enriched with metadata that describe both data protection and functional requirements; *ii)* we propose a parametric heuristic tailored to address the computational complexity of the NP-hard service selection problem that maximizes the quality of data, while addressing data protection and functional requirements; *iii)* we evaluate the performance and quality of the algorithm through experiments conducted using a real, open dataset from the domain of law enforcement. Performance and quality are compared against a baseline modeling current approaches in literature."

Comment: One suggestion for the authors is to develop new, more specific metrics to evaluate data quality when privacy-preserving transformations are applied. This would bring something unexplored to the table and advance the conversation around how we balance quality and privacy in data pipelines.

Response: *Following up on the previous reply to this reviewer, the contribution and novelty of our proposal lie in the definition of a data governance framework for service-based data pipelines. The primary objective of this framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding privacy and security requirements.*

We clarified the contribution of our paper in the introduction as follows.

> "Based on the aforementioned considerations, we propose a data governance framework for service-based data pipelines. The primary objective of our framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding security and privacy requirements.[4] To this aim, each element of the pipeline is *annotated* with *i)* data protection requirements expressing transformation on data and *ii)* functional specifications on services expressing data manipulations carried out during each service execution. Though applicable to a generic scenario, our data governance approach starts from the assumption that maintaining a larger volume of data leads to higher data quality; as a consequence, its service selection algorithm focuses on maximizing data quality in terms of data completeness by retaining the maximum amount of information when applying data protection transformations.
>
> The primary contributions of the paper can be summarized as follows: *i)* we define a data governance framework that implements an algorithm for the selection of data processing services enriched with metadata that describe both data protection and functional requirements; *ii)* we propose a parametric heuristic tailored to address the computational complexity of the NP-hard service selection problem that maximizes the quality of data, while addressing

[4] We note that the assembly of the selected services in an executable pipeline is out of the scope of this paper. However, our approach is agnostic to the specific executable environment.

> data protection and functional requirements; *iii)* we evaluate the performance and quality of the algorithm through experiments conducted using a real, open dataset from the domain of law enforcement. Performance and quality are compared against a baseline modeling current approaches in literature."

Quality metrics and dimensions have been extensively analyzed in the literature and the definition of new metrics and dimensions is therefore outside of the scope of this paper. Rather, our methodology is independent of the specific set of quality metrics and dimensions. In our paper, we focus on quality dimension completeness and two related metrics (Section 5), one qualitative and one quantitative, that have the peculiarity of being simple and unsophisticated. This choice put our experimental evaluation in a scenario that was not affected by the effectiveness of the adopted metrics and dimensions. We clarified this choice in the introduction as follows:

> "Though applicable to a generic scenario, our data governance approach starts from the assumption that maintaining a larger volume of data leads to higher data quality; as a consequence, its service selection algorithm focuses on maximizing data quality in terms of its completeness by retaining the maximum amount of information when applying data protection transformations."

and in Section 5 as follows:

> "Different definition of quality exists (e.g., [3,4]) according to different dimensions such as completeness, timeliness, and accuracy, to name but a few. Quality metrics measure the data quality preserved at each step of the data pipeline according to the selected quality dimensions, and can be classified as *quantitative* or *qualitative*. Quantitative metrics monitor the amount of data lost during data transformations to model the quality difference between datasets X and Y. Qualitative metrics evaluate changes in the properties of datasets X and Y. For instance, qualitative metrics can measure the changes in the statistical distribution of the two datasets.
>
> *In this paper, for simplicity but no lack of generality, we consider quality dimension completeness and use two metrics, one quantitative and one qualitative, to compare the input dataset X and the dataset Y obtained by enforcing data protection requirements (i.e., our policy-driven transformation described in Section 4) on X at each step of the pipeline. We note that the choice of considering a single dimension, and simple and unsophisticated metrics puts our experimental evaluation in Section 6 in a scenario that is not affected by the effectiveness of the adopted metrics. We also note that a complete taxonomy of possible dimensions and metrics is outside the scope of this paper and will be the target of our future work."*

Comment: The sliding window heuristic for service selection is another area that feels a bit underdeveloped. While it improves efficiency, it's a commonly used approach for solving service selection problems, especially those based on quality of service (QoS). The authors could look into more advanced techniques—perhaps something like reinforcement learning, which could dynamically select services based on real-time feedback. That would add more innovation and could improve pipeline efficiency while maintaining the balance between quality and privacy.

Response: *As clarified in the previous replies to this reviewer, our goal is to define a data governance framework for service-based data pipelines. The primary objective of this framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding privacy and security requirements.*

Our solution was inspired by QoS approaches but applied in a different scenario and aimed at different objectives. Here, we aimed to maximize the quality of the data processed by the services in the pipeline, rather than the quality of services.

We note that, in our solution, data transformations are enforced by our framework according to the services' profile and data protection policies only; the specific service execution behavior is therefore not affecting the data transformation process and, in turn, the composition of the pipeline instance. In other words, the performance of our solution is not degrading according to the behavior of each service, but only depends on its profile. As a consequence, the adaptation of a pipeline instance is triggered by a change in one of our building blocks, that is, the service profile, the policies, and the functional requirements.

In general, the idea of making our proposal adaptive and dynamically selecting services based on (real-time) feedback is very interesting and will be the target of our future work. We clarified this in the text as follows:

> "The paper leaves space for future work. [...] Moreover, we plan to explore adaptive techniques based on machine learning for dynamic service selection, which increases the stability of data quality and privacy in varying operational conditions.[...]"

Comment: It would also be helpful to see this framework tested in **real-world scenarios**, perhaps in industries like healthcare or financial services, where data quality and privacy are critical.

Response: *We do agree with the reviewer that testing the framework in multiple real-world scenarios is valuable and important. For this reason, our reference scenario in Section 2.2 considers a real-world scenario based on real and open data in law enforcement, a domain that involves PII (Personally Identifiable Information) and necessitates robust security measures enforced by stringent security/privacy policies and quality requirements.*

Following the reviewer's comment, we spotted the need to clarify the relevance of our reference scenario in Section 2.2, as reported in the following.

> "Our approach targets application domains involving sensitive data, such as Personally Identifiable Information (PII), that must be securely shared and protected across diverse and complex analytical processes involving multiple stakeholders. It is applicable across industrial use cases based on cloud-edge infrastructures, where data from third-party (IoT) devices are injected and shared via the cloud, as well as in data ecosystems across sectors such as healthcare, finance, law enforcement and justice.
>
> Our reference scenario draws on commonly used dataspaces, such as dataspace on public administration, focusing specifically on the law enforcement domain. Using open data, we selected a scenario that includes real sensitive records of individuals detained in Connecticut Department of Correction facilities while awaiting trial.[5] Various stakeholders may use these data for different objectives: public health agencies to monitor inmate health trends, judicial bodies to track case processing efficiency, advocacy groups to identify disparities in detention, policymakers to analyze the impacts on the criminal justice system, social services to prepare post-release support, researchers to study the broader social effects of pre-trial detention, and correctional departments to compare admission trends across facilities.
>
> To streamline the use case, we focused on a subset of this real-world scenario, envisioning three Department of Correction (DOC) partners - Connecticut, New York, and New Hampshire - sharing data according to their privacy policies. In this scenario, a user from the Connecticut DOC seeks to compare admission trends in Connecticut's facilities with those in New York and New Hampshire to evaluate, for instance, possible discrimination and unfair treatment of individuals awaiting trial. Additionally, the policy requires that all service execution remains within the Connecticut DOC environment, mandating data protection measures if data transmission extends beyond Connecticut's borders.
>
> Our reference scenario aligns with the latest regulations on data governance (e.g., the European

[5] https://data.ct.gov/Public-Safety/Accused-Pre-Trial-Inmates-in-Correctional-Faciliti/b674-jy6w

> AI Data Governance Act[6]) and artificial intelligence (e.g., the EU AI Act[7]). In particular, the EU AI Act identifies law enforcement and administration of justice as high-risk domains where proper data governance, risk management, and quality management systems must be employed in AI training and operation following the requirements on data quality and protection set in this paper."

It is also important to note that the relevant application domains are not limited to law enforcement, but can be easily extended to other real-world domains including healthcare and finance. This will be the topic of our future work, as discussed in the refined Section 8 as follows.

> "The paper leaves space for future work. [...] Third, we will evaluate our methodology in different real-world production scenarios with the scope of evaluating its practical usability and utility, bridging the gap between theoretical and practical efficiency. [...]"

Comment: Comparing the performance of this framework against existing industry solutions could highlight its contributions, especially if it can outperform those in key areas like maintaining data quality while ensuring privacy.

Response: *Thanks to the reviewer's comment, we noticed we did not explicitly define the baseline (modeling approaches in literature) in the paper.*

To the best of our knowledge, there are no industry solutions that can be directly compared with our approach and its specific target requirements. Solutions in the state of the art, balancing data quality and protection, do not support distributed environments based on service-based data pipelines and can only maximize data quality locally and step-by-step. We therefore modeled our baseline as an approach that instantiates each node in the pipeline independently, with no information on the pipeline structure, by selecting the corresponding services following a greedy approach. This approach consists of our sliding window heuristic configured with window size $|w|=1$.

We extended Section 6.1 with the definition of our baseline:

> Our simulator also supports the comparison of the performance and quality of our sliding-window heuristic with a baseline modeling solutions in the state of the art as well as the exhaustive approach (i.e., the theoretical optimum). We first modeled our baseline as a greedy approach that, for each node of the service pipeline, selects the best service that maximizes the data quality, while addressing data protection requirements in annotation λ. The reason is that, to the best of our knowledge, existing (industry) solutions and standards do not support service-based data pipelines and are therefore unable to instantiate the service pipeline according to the pipeline structure and service dependencies. We therefore defined our baseline as the sliding window heuristic configured with window size $|\mathrm{w}|=1$. We then implemented the exhaustive approach calculating the theoretical optimum as the sliding window heuristic configured with window size $|\mathrm{w}|=l$, to illustrate the potential efficiency of our heuristics within realistic computational limits.

and compared its quality and performance against the ones measured on our sliding-window heuristic (Sections 6.2 and 6.3).

In addition, to complement the experimental comparison in Section 6, we extended Section 7 (related work) with a discussion and a table (Table 3) comparing our heuristic and main (industrial) solutions according to different evaluation criteria as follows.

> The pipeline template proposed in this work addresses these challenges by enabling to express the security policies at the right level of granularity, considering individual services in the pipeline. It can also be easily mapped onto specific platforms, such as Apache-based systems, as we have demonstrated in [30]. Table 3 provides a comparative analysis with relevant existing

[6] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[7] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Table 3: Comparative analysis with relevant existing approaches. Feature support is classified according to ✓(fully supported), ∼(partially supported or limited in scope), ✗(not supported)

Solution	F1	F2	F3	F4
Microsoft Presidio [31]	✓, can integrate within cloud-edge pipelines	∼, focuses on data redaction	✓, compatible with diverse techniques	∼, pre-built PII detectors with configurable policies
Apache Ranger [32]	∼, mostly limited to cloud settings	✗, provides access control rather than service optimization	✓, integrates with various techniques	✓; high expressiveness with fine-grained policy control
Google Cloud DLP [33]	✓, primarily within Google Cloud	∼, focuses on redaction and anonymization	✓, works across data types	∼, flexible templates for data masking and redaction policies
AWS Macie [34]	∼, suited for AWS cloud infrastructure	∼, prioritizes data protection	✓, AWS-centric	∼, supports predefined PII types but less customizable
IBM Guardium [35]	✓, supports hybrid cloud and on-prem setups	✗, focuses on monitoring and access control	✓, adaptable to multiple frameworks	✓, extensive policy-based access control and monitoring
Apache Sentry [36]	∼, Hadoop ecosystems	✗, static access control	✗, closely tied to Hadoop	∼, supports column and row-level access control
Our paper	✓, suitable for cloud-edge environments	✓, selection of services that optimize quality while ensuring protection	✓, data-protection techniques agnostic	✓, high expressiveness with fine-grained policy control

approaches, highlighting how few industrial solutions compare to our framework according to the following critical features:

- **F1 – Service-Based Pipeline Support in the Cloud-Edge Continuum:** The ability to effectively operate within distributed environments spanning cloud and edge infrastructure.
- **F2 – Quality-Aware Service Selection Ensuring Data Protection:** The capacity to optimize service selection processes, maintaining data quality across the pipeline and ensuring robust data protection measures.
- **F3 – Framework-Agnostic Data Protection:** The degree to which each solution is bound to specific data protection techniques.
- **F4 – Policy Expressiveness:** The degree to which each solution supports fine-grained specification of policies or privacy measures.

According to Table 3, most of competitor solutions have full support for F3, while no solution has full support for F2. All solutions provide partial or full support for F1 and F4, with F4 fully supported by just two of the competitors. Microsoft Presidio aligns most closely with our approach, as it supports cloud-edge integration, offers compatibility with diverse techniques, and includes configurable policies for PII detection. However, our tool uniquely supports the optimization of data quality alongside privacy through a service selection feature and across the entire pipeline lifecycle. Additionally, unlike other solutions that are cloud-specific, our tool maintains compatibility across hybrid environments, addressing both cloud-edge and on-premise scenarios.

Comment: Lastly, I'd suggest exploring a more adaptive approach to the privacy-utility trade-off. Developing a model that adjusts privacy levels based on data sensitivity, the user's role, or the type of analytics being performed would make this work more dynamic and innovative, offering a fresh take on how we handle the balance between data protection and quality.

Response: *Thanks for your comment. Our solution based on policies permits to adjust data transformations (privacy levels in your comments) according to the service profile (including the type of service, the type of analytics, the user role, and the attributes of the service owner) as well as the data sensitivity (specified in the object of the policy). For the reviewer's convenience, in the following, we report the relevant portion of the text in Section 3:*

> *"Data Protection Annotation λ expresses data protection requirements in the form of access control policies. We consider an attribute-based access control model that offers flexible fine-grained authorization and adapts its standard key components to address the unique characteristics of a big data environment. Access requirements are expressed in the form of policy conditions that are defined as follows.*
>
> ***Definition 1 (Policy Condition)*** *A* Policy Condition pc *is a Boolean expression of the form* (attr_name *op* attr_value)*, with op$\in\{<,>,=,\neq,\leq,\geq\}$,* attr_name *an attribute label, and* attr_value *the corresponding attribute value.*
>
> *Built on policy conditions, an access control policy is then defined as follows.*
>
> ***Definition 2 (Policy)*** *A policy p$\in$P is 5-uple <subj, obj, act, env, T^P> that specifies who (*subject*) can access what (*object*) with action (*action*), in a specific context (*environment*) and under specific obligations (*data transformation*).*
>
> *More in detail, subject subj specifies a service s_i issuing an access request to perform an action on an object. It is a set $\{pc_i\}$ of* Policy Conditions *as defined in Definition 1. For instance, (classifier="SVM") specifies a service providing a SVM classifier. We note that subj can also specify conditions on the service owner (e.g., owner_ location="EU") and the service user (e.g., service_ user_ role="DOC Director").*
>
> *Object obj defines the data governed by the access policy. In this case, it is a set $\{pc_i\}$ of* Policy Conditions *on the object's attributes. For instance, {(type="dataset"), (region="CT")} refers to an object of type dataset and whose region is Connecticut.*
>
> *Action act specifies the operations that can be performed within a big data environment, from traditional atomic operations on databases (e.g., CRUD operations) to coarser operations, such as an Apache Spark Direct Acyclic Graph (DAG), Hadoop MapReduce, an analytics function call, and an analytics pipeline."*

---

3 Response to Reviewer 3

Comment: The introduction (and even the title) is a bit misleading, since it suggests that data quality and protection are conflicting factors that pipelines must balance. Then, as the paper is read it is clear that both aspects are not considered equally.

Response: *We do agree with the reviewer that the title and the introduction did not correctly convey the message and contribution of the paper. We therefore changed the title from "Balancing Data Quality and Protection in Distributed Big Data Pipelines" (Original Title) to "Maximizing Data Quality While Ensuring Data Protection in Service-Based Data Pipelines" (Current Title).*

We then substantially revised the introduction to correctly reflect the content and contribution of the paper describing:

- *the considered scenario:*

  > "The wide success and adoption of cloud-edge infrastructures and their intrinsic multitenancy radically change the way in which distributed systems are developed, deployed, and executed, redefining IT scalability, flexibility, and efficiency. Multitenancy in fact enables multiple users to share resources, such as computing power, storage, and services, optimizing their utilization and reducing operational costs.
  >
  > In this scenario, the increasing ability of collecting and managing huge volume of data, coupled with a paradigm shift in service delivery models, has also significantly enhanced scalability and efficiency in data analytics. Data are treated as digital products, which are managed and analyzed by multiple services orchestrated in data pipelines. This shift is also fostering the emergence of new platforms and environments, such as data marketplaces and data spaces, where data in critical domains (e.g., justice, healthcare, transportation) can be pooled and shared to maximize data quality and trustworthiness, and distributed data management systems support data storing, versioning, and sharing for complex analytics processes.[8]
  >
  > The flip side of a scenario where data pipelines orchestrate services selected at run time and are delivered in the cloud-edge continuum is the increased complexity in data governance. Data are shared and analyzed by multiple services owned by different providers introducing unique security challenges. On one side, the pipeline owner and data providers have different security requirements, access policies, and data sensitivity that vary according to the specific orchestrated services; on the other side, orchestrated services (data consumers) have different profiles that impact on the amount of data they can access and analyze."

- *a brief discussion of the shortages of existing solutions:*

  > "Adequate measures such as encryption, access control mechanisms, and data anonymization techniques have been implemented to protect data against unauthorized access and ensure compliance with regulatory requirements such as GDPR[1] or HIPAA[2]. However, data quality is also crucial and must be guaranteed, as the removal or alteration of personally identifiable information from datasets to safeguard individuals' privacy can compromise the accuracy of analytics results.
  >
  > So far, all research endeavors have been mainly concentrated on exploring these two issues separately: on one hand, *data quality*, encompassing accuracy, reliability, and suitability, has been investigated to understand the implications in analytical contexts [3,4]. On the other hand, *data security and privacy* focused on the protection of confidential information and adherence to rigorous privacy regulations[5-8]. Although extensively studied, these investigations often prioritize enhancing the quality, security, and privacy of source data rather than ensuring data quality, security, and privacy throughout the entire processing pipeline, or the integrity of outcomes derived from data."

[8] https://joinup.ec.europa.eu/collection/elise-europeanlocation-interoperability-solutions-e-government/glossary/term/data-marketplace, https://internationaldataspaces.org/, https://digitalstrategy.ec.europa.eu/en/library/staff-working-documentdata-spaces

- *the requirements and questions driving our work:*

  "A valid solution requires a holistic approach that integrates technological solutions, organizational policies, and ongoing monitoring and adaptation to emerging threats and regulatory changes across the entire pipeline lifecycle. The implementation of robust access control mechanisms or privacy techniques, ensuring that only authorized users can access specific datasets (or a portion thereof) is just a mandatory but initial step. Additional requirements are emerging. First, data protection requirements should be defined at each stage of the pipeline, potentially integrating techniques like data masking and anonymization (e.g., k-anonymity, l-diversity, differential privacy) to safeguard sensitive information, thereby preserving data privacy while enabling high quality data sharing and analysis. Second, data lineage should be prioritized, fostering a comprehensive understanding and optimization of data flows and transformations within complex analytical ecosystems. Third, data protection and data quality requirements should drive the process that builds a pipeline with maximum data quality, while addressing data protection requirements.

  When evaluating a solution meeting the above criteria, the following questions naturally arise:

  i) How does a data protection solution affect data quality in the pipeline? How can we minimize this impact thus maximizing the overall data quality?

  ii) Should data protection be implemented at each pipeline step rather than filtering all data at the outset?

  iii) In a scenario where service-based data pipelines are built by selecting the best services among various candidate services, how might these choices be driven by quality requirements?"

- *and the contribution of our work:*

  "Based on the aforementioned considerations, we propose a data governance framework for service-based data pipelines. The primary objective of our framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding security and privacy requirements.[9] To this aim, each element of the pipeline is *annotated* with *i)* data protection requirements expressing transformation on data and *ii)* functional specifications on services expressing data manipulations carried out during each service execution. Though applicable to a generic scenario, our data governance approach starts from the assumption that maintaining a larger volume of data leads to higher data quality; as a consequence, its service selection algorithm focuses on maximizing data quality in terms of data completeness by retaining the maximum amount of information when applying data protection transformations.

  The primary contributions of the paper can be summarized as follows: *i)* we define a data governance framework that implements an algorithm for the selection of data processing services enriched with metadata that describe both data protection and functional requirements; *ii)* we propose a parametric heuristic tailored to address the computational complexity of the NP-hard service selection problem that maximizes the quality of data, while addressing data protection and functional requirements; *iii)* we evaluate the performance and quality of the algorithm through experiments conducted using a real, open dataset from the domain of law enforcement. Performance and quality are compared against a baseline modeling current approaches in literature."

[9] We note that the assembly of the selected services in an executable pipeline is out of the scope of this paper. However, our approach is agnostic to the specific executable environment.

Comment: The introduction is not really compelling by showing the application area of such an approach.

Response: *Following our previous reply to this reviewer, we substantially refined the introduction in Section 1 clarifying the objectives and contributions of the paper. In the process of refining the entire introduction, we also clarified the application area of our approach by detailing:*

- *the considered scenario with reference to service-based data pipelines, data spaces, and relevant application domains:*

  > "The wide success and adoption of cloud-edge infrastructures and their intrinsic multitenancy radically change the way in which distributed systems are developed, deployed, and executed, redefining IT scalability, flexibility, and efficiency. Multitenancy in fact enables multiple users to share resources, such as computing power, storage, and services, optimizing their utilization and reducing operational costs.
  >
  > The increasing ability of collecting and managing huge volume of data, coupled with a paradigm shift in service delivery models, has also significantly enhanced scalability and efficiency in data analytics. Data are treated as digital products, which are managed and analyzed by multiple services orchestrated in pipelines. This shift is fostering the emergence of new platforms and environments, such as data marketplaces and data spaces, where data in critical domains (e.g., law enforcement, healthcare, transportation) can be pooled and shared to maximize data quality and trustworthiness, and distributed data management systems supporting data storing, versioning, and sharing for complex analytics processes.[10]
  >
  > The flip side of a scenario where service-based data pipelines orchestrate services selected at run time and are delivered in the cloud-edge continuum is the increased complexity in data governance. Data are shared and analyzed by multiple services owned by different providers introducing unique security challenges. On one side, the pipeline owner and data providers have different security requirements, access policies, and data sensitivity that vary according to the specific orchestrated services; on the other side, orchestrated services (data consumers) have different profiles that impact on the amount of data they can access and analyze."

- *and the contribution of our work:*

  > "Based on the aforementioned considerations, we propose a data governance framework for service-based data pipelines. The primary objective of this framework is to support the selection of data processing services within the pipeline, with a central focus on the selection of those services that maximize data quality, while upholding privacy and security requirements.[11] To this aim, each element of the pipeline is *annotated* with *i)* data protection requirements expressing transformation on data and *ii)* functional specifications on services expressing data manipulations carried out during each service execution. Though applicable to a generic scenario, our data governance approach starts from the assumption that maintaining a larger volume of data leads to higher data quality; as a consequence, its service selection algorithm focuses on maximizing data quality in terms of its completeness by retaining the maximum amount of information when applying data protection transformations.
  >
  > The primary contributions of the paper can be summarized as follows: *i)* we define a data governance framework that implements an algorithm for the selection of data processing services enriched with metadata that describe both data protection and functional requirements; *ii)* we propose a parametric heuristic tailored to address the computational complexity of the NP-hard service selection problem that maximizes the quality of data, while addressing data protection and functional requirements; *iii)* we evaluate

[10] https://joinup.ec.europa.eu/collection/elise-europeanlocation-interoperability-solutions-e-government/glossary/term/data-marketplace, https://internationaldataspaces.org/, https://digitalstrategy.ec.europa.eu/en/library/staff-working-documentdata-spaces

[11] We note that the assembly of the selected services in an executable pipeline is out of the scope of this paper. Our approach however is agnostic to the specific executable environment.

> the performance and quality of the algorithm through experiments conducted using an open dataset from the domain of justice. Performance and quality are compared against a baseline modeling current approaches in literature."

To better clarify the application area of our approach, we also revised the reference scenario in Section 2, please see our reply to the next comment.

Comment: The use case presented in Section 2.2 is fictitious, so is there a real need and use case for such an approach? Maybe in the context of Dataspaces?

Response: *Thanks to the reviewer's comment, we clarified the relevance and goal of our reference scenario, as well as the need for our approach in this letter and in Section 2.2.*

In particular, we believe our reference scenario is not fictitious according to the following three reasons: i) *it models scenarios involving multiple stakeholders and sensitive data that must be securely shared and protected within diverse, complex analytical processes, spanning from intricate industrial cases to large-scale dataspaces;* ii) *it pertains to the law enforcement domain, a field that frequently involves PII (Personally Identifiable Information) and necessitates robust security measures enforced by stringent security/privacy policies and quality requirements;* iii) *it is built on a real-world context and real, open data. We note that the scenario had the only objective to enhance the readability of the case study and the experimental results, and to simplify the testing of our algorithm. We believe that adding further complexity would not have contributed additional value to the paper.*

To address the reviewer's comment, we clarified the significance of the reference scenario at the beginning of Section 2.2 as follows.

> "Our approach targets application domains involving sensitive data, such as Personally Identifiable Information (PII), that must be securely shared and protected across diverse and complex analytical processes involving multiple stakeholders. It is applicable across industrial use cases based on cloud-edge infrastructures, where data from third-party (IoT) devices are injected and shared via the cloud, as well as in data ecosystems across sectors such as healthcare, finance, law enforcement and justice.
>
> Our reference scenario draws on commonly used dataspaces, such as dataspace on public administration, focusing specifically on the law enforcement domain. Using open data, we selected a scenario that includes real sensitive records of individuals detained in Connecticut Department of Correction facilities while awaiting trial.[12] Various stakeholders may use these data for different objectives: public health agencies to monitor inmate health trends, judicial bodies to track case processing efficiency, advocacy groups to identify disparities in detention, policymakers to analyze the impacts on the criminal justice system, social services to prepare post-release support, researchers to study the broader social effects of pre-trial detention, and correctional departments to compare admission trends across facilities.
>
> To streamline the use case, we focused on a subset of this real-world scenario, envisioning three Department of Correction (DOC) partners - Connecticut, New York, and New Hampshire - sharing data according to their privacy policies. In this scenario, a user from the Connecticut DOC seeks to compare admission trends in Connecticut's facilities with those in New York and New Hampshire to evaluate, for instance, possible discrimination and unfair treatment of individuals awaiting trial. Additionally, the policy requires that all service execution remains within the Connecticut DOC environment, mandating data protection measures if data transmission extends beyond Connecticut's borders.
>
> Our reference scenario aligns with the latest regulations on data governance (e.g., the European AI Data Governance Act[13]) and artificial intelligence (e.g., the EU AI Act[14]). In particular, the EU AI Act identifies law enforcement and administration of justice as high-risk domains where

[12] https://data.ct.gov/Public-Safety/Accused-Pre-Trial-Inmates-in-Correctional-Faciliti/b674-jy6w

[13] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[14] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

> proper data governance, risk management, and quality management systems must be employed in AI training and operation following the requirements on data quality and protection set in this paper."

In this context, it is important to note that the application domains of our approach are not limited to the one of law enforcement, but can easily extended to other real-world domains including healthcare and finance. This will be the topic of our future work as discussed in the refined Section 8 as follows.

> "The paper leaves space for future work. [...] Third, we will evaluate our methodology in different real-world production scenarios with the scope of evaluating its practical usability and utility, bridging the gap between theoretical and practical efficiency."

In addition, the need for approaches similar to the one in our paper is recognized by latest regulations on data governance (e.g., the European AI Data Governance Act[15]*) and artificial intelligence (e.g., The EU AI Act*[16]*). In particular, Article 27 of the EU AI Act states: 'Privacy and data governance means that AI systems are developed and used in accordance with privacy and data protection rules, while processing data that meets high standards in terms of quality and integrity.'*

*More in detail, the EU AI Act identifies the requirements for high-risk AI domains (see Article 6: Classification Rules for High-Risk AI Systems) under which law enforcement and administration of justice are listed. In particular, Annex III: High-Risk AI Systems Referred to in Article 6(2) and, specifically, point 6(2) states that high-risk law enforcement scenarios include: "*AI systems intended to be used by or on behalf of law enforcement authorities, or by Union institutions, bodies, offices or agencies, in support of law enforcement authorities to evaluate the reliability of evidence in the course of the investigation or prosecution of criminal offences;*."*

The EU AI Act prescribes in the above domains the use of advanced data governance and quality management systems like the one in our paper. In particular, Article 10: Data and Data Governance - Point 6(c) states:

1. "High-risk AI systems which make use of techniques involving the training of AI models with data shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5 whenever such data sets are used.
2. Training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular:
   - [...]
   - (b) data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection;
   - [...]
   - (e) an assessment of the availability, quantity and suitability of the data sets that are needed;
   - (f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations;"

and Article 17: Quality Management System states: "[...] (f) systems and procedures for data management, including data acquisition, data collection, data analysis, data labelling, data storage, data filtration, data mining, data aggregation, data retention and any other operation regarding the data that is performed before and for the purpose of the placing on the market or the putting into service of high-risk AI systems.*"*

We believe the above discussion and the text added in the paper clarified the goal and relevance of

[15] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[16] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

our reference scenario.

Comment: Some sections like 3.3 are quite verbose and dense. A simplification of explanations or a supporting image would really help on understanding specially when a set of steps are presented.

Response: *Thanks to your comment, we noticed that Figure 2 and Figure 3 in the original paper were not adding much to the discussion. We, therefore, added a new Figure 2 visually presenting Example 3.1 and all its artifacts in detail, and modified Figure 3 accordingly to present the corresponding pipeline instance in Example 4.1.*

Comment: Similarly, Section 4 textually describes an algorithm. Why not present directly a pseudocode algorithm leveraging the formal concepts presented before like in page 23?

Response: *Thanks for your comment. We added the pseudocode of the algorithm in Figure 4 and used it as a reference in the text.*

```
INPUT
$G^{\lambda,\gamma}$ : Pipeline Template
$S^c$: Candidate Services

OUTPUT
$G'$: Pipeline Instance

Instantiate_Pipeline($G^{\lambda,\gamma}$, $S^c$)

/* Initialize the pipeline instance*/
$G'$ = {};
/* Traverse the pipeline template using BFS*/
for each $v$ in $G^{\lambda,\gamma}$
    $v'$ = Generate_Vertex($v$);
    $G'$ = $G' \cup v'$;
    $S'$ = Filter_Services($S^c[v]$, $v$.policies);
    selectedService = Select_A_Service($S'$);
    $v'$.service = selectedService;
endfor;
return $G'$;


Filter_Services($S^c[v]$, policies)
/* Filter candidate services based on policies*/
$S'$ = {};
for each service $s$ in $S^c[v]$:
    if $s$.profile satisfies any policy:
        $S'$ = $S' \cup s$;
    endif;
endfor;
return $S'$;
```

Comment: The proposed quality metrics in 5.1.1 and 5.1.2 are not really data quality metrics (e.g., coverage, completeness, timeliness, ...). The metrics adopted are more related to profiling the data and their distributions.

Response: *We do agree with the reviewer that, as discussed in literature (e.g., [3,4]), data quality is a multidimensional measure. However, for simplicity but no lack of generality, in our paper we focused on metrics that provide a global indication of the quality in terms of the sole dimension completeness. Specifically, our work assumes high-quality and complete data, and tracks data degradation (in terms of completeness) across the entire pipeline lifecycle. Our metrics track i) quantitative degradation of completeness and ii) statistical degradation in terms of the statistical distribution of data.*

We note that our framework is generic enough to support multiple dimensions and metrics, driving service selection and composition. These dimensions and metrics can be easily replaced and adapted to track other aspects of the data such as coverage and timeliness. However, the main goal of our paper was the definition of a methodology for service selection that guarantees maximum quality, while ensuring a minimum level of data protection, and is independent of the specific quality dimensions and metrics. We clarified our choice in Section 5.1 as follows.

> "Ensuring data quality is mandatory to implement data pipelines that provide accurate results and decision-making along the whole pipeline execution. Different definition of quality exists (e.g.[3,4]) according to different dimensions such as completeness, timeliness, and accuracy, to name but a few. Quality metrics measure the data quality preserved at each step of the data pipeline according to the selected quality dimensions, and can be classified as *quantitative* or *qualitative*. Quantitative metrics monitor the amount of data lost during data transformations to model the quality difference between datasets X and Y. Qualitative metrics evaluate changes in the properties of datasets X and Y. For instance, qualitative metrics can measure the changes in the statistical distribution of the two datasets.
>
> In this paper, for simplicity but no lack of generality, we consider quality dimension completeness and use two metrics, one quantitative and one qualitative, to compare the input dataset X and the dataset Y obtained by enforcing data protection requirements (i.e., our policy-driven transformation described in Section 4) on X at each step of the pipeline. We note that the choice of considering a single dimension, and simple and unsophisticated metrics put our experimental evaluation in Section 6 in a scenario that is not affected by the effectiveness of the adopted metrics. We also note that a complete taxonomy of possible dimensions and metrics is outside the scope of this paper and will be the target of our future work."

Our future work will consider the multidimensional nature of data quality as discussed in Section 8.

> "The paper leaves space for future work. First, we will extend our methodology with a taxonomy of possible quality dimensions and metrics supporting the definition of a multidimensional data quality that considers multiple dimensions such as, for instance, completeness, timeliness, and accuracy. Multiple dimensions and metrics will be adopted and weighted according to user priorities or task-specific requirements to better address the inherent multidimensional nature of data quality. This extension will enable more sophisticated monitoring and optimization mechanisms throughout the entire data lifecycle. [...]"

Comment: The experimental section is solely based on simulations over synthetic data. A real workload (pipelines, policies and data) should be presented to evaluate the practical effectiveness of the approach.

Response: *As already clarified in our replies to this reviewer, our reference scenario in Section 2.2 and therefore the simulations in our experiments consider a real-world scenario in the domain of law enforcement. The law enforcement domain involves PII (Personally Identifiable Information) and necessitates robust security measures enforced by stringent security/privacy poli-*

cies and quality requirements. Our reference scenario and experimental simulations are based on real and open data available at https://data.ct.gov/Public-Safety/Accused-Pre-Trial-Inmates-in-Correctional-Faciliti/b674-jy6w and enriched by randomly generating first and last names to introduce additional sensitive data, increase data diversity, and increase the variety of security policies.

In addition, our choice of prioritizing a wide and horizontal experimental simulation over specific and vertical testing with real workloads, pipelines, and policies was exactly to evaluate the practical effectiveness of our approach in a generic and domain-agnostic scenario. By randomly generating policies and pipelines we tested tens of possible different configurations, which would be difficult to reproduce with specific real workloads, also minimizing the bias introduced by the workload choice.

It is also important to note that the application domain of our approach is not limited to law enforcement, but can be easily extended to other real-world scenarios including healthcare and finance. This will be the topic of our future work, as discussed in the refined Section 8 as follows.

> "The paper leaves space for future work. [...] Third, we will evaluate our methodology in different real-world production scenarios with the scope of evaluating its practical usability and utility, bridging the gap between theoretical and practical efficiency."

Comment: No alternative or baseline method from the state of the art is evaluated and compared against yours. Most likely there is no direct competitor that can be used out-of-the-box, but if I did not have your method how would I implement this approach? This is your baseline, and at least you should show improvements over the baseline.

Response: *Thanks to the reviewer's comment, we noticed we did not explicitly define the baseline (modeling approaches in literature) in the paper.*

As the reviewer suggests, to the best of our knowledge, there are no competitor solutions that can be directly compared with our approach and its specific target requirements. Solutions in the state of the art, balancing data quality and protection, do not support distributed environments based on service-based data pipelines and can only maximize data quality locally and step-by-step.

We therefore modeled our baseline as an approach that instantiates each node in the pipeline independently, with no information on the pipeline structure, by selecting the corresponding services following a greedy approach. This approach consists of our sliding window heuristic configured with window size |w|=1.

We extended Section 6.1 with the definition of our baseline:

> "Our simulator also supports the comparison of the performance and quality of our sliding-window heuristic with a baseline modeling solutions in the state of the art as well as the exhaustive approach (i.e., the theoretical optimum). We first modeled our baseline as a greedy approach that, for each node of the service pipeline, selects the best service that maximizes the data quality, while addressing data protection requirements in annotation λ. The reason is that, to the best of our knowledge, existing (industry) solutions and standards do not support service-based data pipelines and are therefore unable to instantiate the service pipeline according to the pipeline structure and service dependencies. We therefore defined our baseline as the sliding window heuristic configured with window size |w|=1. We then implemented the exhaustive approach calculating the theoretical optimum as the sliding window heuristic configured with window size |w|=l, to illustrate the potential efficiency of our heuristics within realistic computational limits. "

We also note that a possible alternative baseline can enforce data protection requirements at ingestion time, applying the union of all applicable security policies and corresponding transformations. This approach however does not apply to a scenario built on service pipelines, which results in the anonymization of the entire dataset with small window sizes, and is therefore not discussed in the paper.

Table 3: Comparative analysis with relevant existing approaches. Feature support is classified according to ✓(fully supported), ∼(partially supported or limited in scope), ✗(not supported)

Solution	F1	F2	F3	F4
Microsoft Presidio [31]	✓, can integrate within cloud-edge pipelines	∼, focuses on data redaction	✓, compatible with diverse techniques	∼, pre-built PII detectors with configurable policies
Apache Ranger [32]	∼, mostly limited to cloud settings	✗, provides access control rather than service optimization	✓, integrates with various techniques	✓; high expressiveness with fine-grained policy control
Google Cloud DLP [33]	✓, primarily within Google Cloud	∼, focuses on redaction and anonymization	✓, works across data types	∼, flexible templates for data masking and redaction policies
AWS Macie [34]	∼, suited for AWS cloud infrastructure	∼, prioritizes data protection	✓, AWS-centric	∼, supports predefined PII types but less customizable
IBM Guardium [35]	✓, supports hybrid cloud and on-prem setups	✗, focuses on monitoring and access control	✓, adaptable to multiple frameworks	✓, extensive policy-based access control and monitoring
Apache Sentry [36]	∼, Hadoop ecosystems	✗, static access control	✗, closely tied to Hadoop	∼, supports column and row-level access control
Our paper	✓, suitable for cloud-edge environments	✓, selection of services that optimize quality while ensuring protection	✓, data-protection techniques agnostic	✓, high expressiveness with fine-grained policy control

Finally, to complement the experimental comparison in Section 6, we extended Section 7 (related work) with a discussion and a table (Table 3) comparing our heuristic and main competitor solutions according to different evaluation criteria as follows.

The pipeline template proposed in this work addresses these challenges by enabling to express the security policies at the right level of granularity, considering individual services in the pipeline. It can also be easily mapped onto specific platforms, such as Apache-based systems, as we have demonstrated in [30]. Table 3 provides a comparative analysis with relevant existing approaches, highlighting how few industrial solutions compare to our framework according to the following critical features:

- **F1 – Service-Based Pipeline Support in the Cloud-Edge Continuum:** The ability to effectively operate within distributed environments spanning cloud and edge infrastructure.
- **F2 – Quality-Aware Service Selection Ensuring Data Protection:** The capacity to optimize service selection processes, maintaining data quality across the pipeline and ensuring robust data protection measures.
- **F3 – Framework-Agnostic Data Protection:** The degree to which each solution is bound to specific data protection techniques.
- **F4 – Policy Expressiveness:** The degree to which each solution supports fine-grained specification of policies or privacy measures.

According to Table 3, most of competitor solutions have full support for F3, while no solution has full support for F2. All solutions provide partial or full support for F1 and F4, with F4 fully supported by just two of the competitors. Microsoft Presidio aligns most closely with our approach, as it supports cloud-edge integration, offers compatibility with diverse techniques, and includes configurable policies for PII detection. However, our tool uniquely supports the optimization of data quality alongside privacy through a service selection feature and across the entire pipeline lifecycle. Additionally, unlike other solutions that are cloud-specific, our tool maintains compatibility across hybrid environments, addressing both cloud-edge and on-premise scenarios.