© SEP 2023 | IRE Journals | Volume 7 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I3-1716724

A Reproducible Framework for Detecting and Quantifying Join-Induced Metric Inflation

SAI LALITESH POTHUKUCHI

Abstract- ***The modern analytics pipelines are crucial in the realisation of reliable decision-making, which is only possible with accurate key performance indicators (KPIs). There is, however, an inadvertent distortion of metrics by relational joins in data engineering processes, which results in Join-Induced Metric Inflation undermining analytical integrity. These distortions are usually compounded by the low data quality, such as the Duplicate and Null Key Impact, as well asComposite Key Integrity, that spreads the error by aggregating KPIs and distorts business intelligence. Addressing these issues, this paper suggests a Reproducible Data Engineering Framework that will be used to detect join-induced distortions systematically, quantify them, and mitigate them. The framework combines Automated Join Risk Flagging, which allows for the identification of high-risk joins before metrics are reported and an inflation estimation mechanism that forecasts the degree of possible KPI distortions. By ensuring that such a framework is incorporated into the routine ETL processes, organisations can guarantee that such workflows are reproducible in nature, uphold data integrity, and foster confidence in the results of the associated analytical activities. The framework is demonstrated through empirical examples and conceptual discussion of how each operationalises Data Quality Risk Assessment and KPI Distortion Detection, offering practical advice on how an analyst can operate the framework as well as governance of an enterprise-wide data setting. The research work has wider implications than its technical mitigation, such as in supporting better analytics governance and reproducible research practices, which are the keys to large-scale, data-driven decisions.***

Keywords: Join-Induced Metric Inflation, Data Quality Risk Assessment, KPI Distortion Detection, Automated Join Risk Flagging, Duplicate and Null Key Impact, Composite Key Integrity, Reproducible Data Engineering Framework

I. INTRODUCTION

The rise of the contemporary data analytics pipeline has radically changed the decision-making process within organisations to be faster and more predictive and evidence-based strategic planning (Avery & Cheek, 2015). The most important pipelines are operations of relational join, combining various sources of data to generate complete datasets to analyse them. Although such joins are necessary to build meaningful metrics, they necessarily have the disadvantage of Join-Induced Metric Inflation, i.e., the inflated aggregation values resulting from the structure of joins and imperfect data quality eroding the integrity of key performance indicators (KPIs) (Putra et al., 2022).

Correct KPIs are essential to the efficiency of operations, resource use and making informed decisions regarding strategy. Unnoticed, inflated, or distorted metrics may spread throughout business intelligence systems and result in incorrect conclusions, incorrect forecasts, and poor business decisions. The danger is also enhanced by such common data quality problems as Duplicate and Null Key Impact and inconsistencies in composite key integrity, which only make the distortions more critical and the validation of the analytical results harder.

Although the reliability of KPIs is essential, there is a substantial knowledge gap in the area of analytics governance because not many organisations actively track and measure the possible inflation brought about by join operations. The solution to the above gap would involve having a framework that not only reveals high-risk joins but also measures their effects on downstream measures, and makes them reproducible across data pipelines.

That is why this study is initiated by the necessity of such a solution. It suggests a Reproducible Data Engineering Framework, which combines Automated Join Risk Flagging with an inflation estimation mechanism, which allows the proactive identification and quantification of KPI distortion. This research has threefold contributions, including but not limited to:

(i) a systematic approach in identifying Join-Induced Metric Inflation, (ii) an automated KPI Distortion Detection that identifies high-risk joins before reporting, and (iii) supporting a governance of analytics at the enterprise level by enshrining repeatable practices into regular ETL practices (Leveque et al., 2012).

This paper contributes in three main ways:

- It is a conceptualisation of the problem of Join-Induced Metric Inflation of modern analytics pipelines.
- It suggests the recurrent model of systematic identification and measurement of joint-related distortions in KPIs.
- It presents the metric of Inflation Factor and a mechanism for flagging the proactive analytics as a joint risk.

II. JOIN OPERATIONS AND METRIC INFLATION

The foundation of the new generation analytics pipeline is vRelational join operations, which allow joining data of different types to help make holistic decisions. With the help of joins, a combination of records can be carried out by using common keys to compute aggregated metrics, correlation studies and cross-domain insights. Nevertheless, these operations do impact the creation of actionable KPIs, but the risk they introduce is the Join-Induced Metric Inflation, which is where the metrics obtain and report exaggerated performance because of structural or data quality concerns (Bracho-Mujica et al., 2019). This inflation happens when the join operation inadvertently multiplies, misaligns or duplicates the metric values, which may mislead the decision makers and hinder governance systems based on data.

2.1 Join Operations and Their Effect.

Each type of join is affected differently by data quality and has a different impact on KPI integrity:

Inner Join - This type of join keeps only the rows which have identical keys in both tables. Inner joins are normally considered to be safe, but when there are repeat values in the join key, then quantities like the total revenue or counts of sales may be artificially inflated. To take an example, when two rows in the secondary table match one row in the primary table, the aggregate metric is doubled, resulting in moderate inflation.

Left Join - Left joins maintain all the rows of the primary table and try to find the matches of the secondary table. Although unmatched rows in the secondary table are disregarded, duplicate or partial KPI inflation in the secondary table can occur with moderate KPI inflation, especially when the KPI measures sum or average numerical values. As an example, the left join of an orders table with a promotional offers table may inflate revenue by matching several records in an offer to the same order.

Many to Many Join - This is where both primary and secondary table rows are similar on the join key. This kind of join is a very dangerous join, because all of the combinations of matching rows are represented in the result set. Therefore, metrics can increase exponentially, generating very high KPI distortions. One such case might be the joining of a table of the products with a table of sales records in the regions where a few matches can greatly exaggerate the total sales figure.

Cross Join -A cross join calculates the Cartesian product of two tables, which is the combination of all the rows of one table with all the rows of the other table. This causes metric inflation to extreme levels, whereby the measures are artificially inflated by the number of combinations. Cross joins are not employed to report on a direct basis, but accidental cross joins of complicated pipeline transformation can greatly harm the KPI integrity (Junietz et al., 2018).

2.2 Mechanisms of Metric Inflation Under Join.

The effect of the join on inflation can be mainly attributed to two factors, which are interactive: the type and quality of data. Although the join type is what defines how the records of more than one table will be merged, the problem of data quality, which manifests in the possible duplication of keys, null values, or the mismatch of the composite keys, stipulates the degree

© SEP 2023 | IRE Journals | Volume 7 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I3-1716724

to which metrics will be exaggerated. Practically, even operationally seemingly safe join operations may raise KPIs in case these problems remain unnoticed. An illustration of this is when a customer table is joined on a transaction table using a composite key of customer ID and region; this will overcount in case some customer IDs are repeated or are null in both tables.

2.3 The importance of Join Validation is as follows.

Since there can be a distortion of metrics, it is important to validate join operations in analytics processes. Join validation, checking the key uniqueness, determining the expected join cardinality, and evaluating the probability of the occurrence of duplicate or missed matches. Analysts will be able to predict Join-Induced Metric Inflation through systematic determination of join risks and put in place mitigation measures, including cleaning up duplicate keys, standardising composite key formats or using pre-join aggregation. This validation step is vital to ensure proper KPIs, keep Data Quality Risk Assessment, and governance compliance to enterprise analytics (Bracho-Mujica et al., 2019; Junietz et al., 2018).

Table 1: Join Types and Potential KPI Impact

Join Type	Scenario	Inflation Risk	Example KPI Impact
Inner Join	Matching rows only	Moderate	Revenue duplication
Left Join	Retain primary table rows	Moderate	Partial metric inflation
Many-to-Many Join	Multiple matches	Very High	Severe KPI distortion
Cross Join	Cartesian product	Extreme	Artificial metrics spike

This table is a step-by-step analysis of the interaction between join types and data characteristics to affect the accuracy of the KPI. The joining validation and monitoring processes can be added to the analytics pipeline and help organisations reduce the distortion of metrics and enhance trust in the reported KPIs. Moreover, the knowledge of the inflation potential of each type of join preconditions automated KPI Distortion Detection and Automated Join Risk Flagging as the key components of the proposed Reproducible Data Engineering Framework that will be addressed in the next passages.

Figure 1: Join Induced Metric Inflation Example

III. DATA KEY QUALITY AND KPI DISTORTION MECHANISMS

The quality of data is a decisive factor in the precision and accuracy of the outputs of the analysis. Even a well-thought-out join operation is likely to generate a bloated metric in case the underlying key data is faulty. In current analytics pipelines, three key data quality concerns, such as Duplicate and Null Key Impact and Composite Key Integrity, are major threats to the accuracy of the key performance indicators (KPIs). These problems spread through the join operations, which is why to design the structure that will reduce the problem of Join-Induced Metric Inflation, their mechanism should be understood (Putra et al., 2022; Koukaras and Tjortjis, 2025).

3.1 Duplicate Keys

Duplicate keys are those records that have the same key in a dataset. In case of such duplicates in join operations, there is a tendency to multiply metrics by accident. As an illustration, a table of customer transactions, such as a customer ID, will contain

duplicate customer IDs because of typing mistakes. Combining this table with a promotions table may lead to the same sale being included in aggregated revenue or sales statistics more than once, and KPIs are artificially inflated. The effect is especially acute with many-to-many joins, in which all combinations of duplicate keys have a cumulative metric distortion. Duplicate keys detection and control are therefore a requirement in KPI Distortion Detection and data integrity preservation.

3.2 Null Key Behaviour

The problem of null keys occurs in situations where the key attributes required in the operations of joins are not present. Such gaps may cause loss of data or distortion of the resulting data. In a left join, e.g. where there are records with null keys in the secondary table, the records will not match, and the aggregation will be incomplete. On the other hand, null keys in a primary table can avoid critical records joining, which leads to underreported metrics. The effect is a distorted perception of KPIs, and this is something likely to mislead decision-makers unless the issues are spotted and mitigated in the form of effective Data Quality Risk Assessment processes.

3.3 Composite Key Integrity

Composite keys are the keys that are created by combining several attributes to form a unique key to join tables. Any distortion in any element of a compound key may generate false aggregations. An example is a join between orders and the shipping table using a composite key of order ID and region, which may not match as expected when regions are not always formatted and are missing in parts. It may result in either cases of double counting or misallocated revenue, such as those that are not complete customer counts, not even in relation to trusting KPI reporting. Ensuring Composite Key Integrity is hence imperative to reproducible and governance-compliant analytics processes (Leveque et al., 2012).

3.4 Joins of Key Propagation of Quality Issues.

Duplicates key effects, null keys and composite keys are multiplied in the join operations. Incorporating bad keys in inner, left or many-to-many joins causes distortion in aggregated metrics where the data is inflated or underrepresented by the type of join and the distribution of the data. A small error in the data on the key can cause substantial distortions in KPIs at scale; systematic key profiling and pre-join validation are also important. Organisations can combine proactive mechanisms in Automated Join Risk Flagging and metric inflation estimation by determining the high-risk keys and quantifying their possible influence.

Table 2: Data Key Quality Problems and Analytical Consequences

Key Issue	Description	Risk to KPIs
Duplicate Keys	Multiple rows share the same identifier	Metric multiplication
Null Keys	Missing join attributes	Data loss or misalignment
Composite Key Mismatch	Partial key combinations	Incorrect aggregations

The following table presents the most frequent quality problems of data keys and their effects on the analytical integrity. The management of these concerns is a key to having reliable KPIs, assisted Reproducible Data Engineering Frameworks, and governance practices embedded into analytics processes. Through the combination of main profiling, validation, and automated risk identification, organisations have the opportunity to actively reduce the distortions in metrics before the reporting of metrics, which guarantees accuracy and reproducibility.

© SEP 2023 | IRE Journals | Volume 7 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I3-1716724

Figure 2: Data Key Quality Impact on KPI Distortion

IV. REPRODUCIBLE FRAMEWORK FOR JOINT RISK DETECTION

To check the appropriate reporting of KPIs in multi-faceted analytics pipelines, the identification of Metric Inflation caused by Joins and its measurement should be organised in a direct way. The current study suggests a Reproducible Data Engineering Framework through which key quality issues and high-risk joins are identified in a systematic manner and approximate the potential effect of coding on measures before reporting. The framework not only increases the accuracy of the analysis but also incorporates the reproducibility and governing compliance along the workflow of the data, as well as promotes work efficiency and integrity of decision-making (Leveque et al., 2012; Sharma and Shekhar, 2021).

The framework is made of five interrelated steps, which cover a critical element of joint risk identification and KPI validation:

4.1 Join Structure Analysis

The initial step analyses the relational design of tables in the analytics channel. The Join Analyser module checks the patterns of join cardinality and determines the type of join that is being used (inner, left, many-to-many, or cross join) and points out the situations that are likely to generate inflated metrics. It is through the structural relationships between tables that the analysts can predict possible distortions and, in the future, concentrate the quality assessment in areas where it is mostly required (Elmitwalli et al., 2025).

4.2 Key Quality Profiling

This step checks the integrity of join keys, duplicates, and null values and composite key inconsistencies. The Key Quality Profiler is a systematic measurement of the occurrence of every issue, and therefore, proactive mitigation is possible before metrics are rolled up. Profiling correctly implies that the future risk scoring and inflation estimation will be based on quality data.

4.3 Formal Risk Scoring Model

Let:

- D = duplicate key ratio
- N = null key ratio
- C = composite key mismatch rate
- J_c = join cardinality multiplier

The overall Join Risk Score R is defined as:

$$R = \alpha D + \beta N + \gamma C + \delta J_c$$

where:

- $\alpha, \beta, \gamma, \delta \geq 0$
- $\alpha + \beta + \gamma + \delta = 1$

Weights are calibrated empirically through regression analysis (Section 7).

Risk categories are defined as:

- Low Risk: $R < 0.05$
- Moderate Risk: $0.05 \leq R < 0.15$
- High Risk: $R \geq 0.15$

This formulation transforms join validation from heuristic assessment into a quantifiable governance control mechanism.

4.4 Engine of Inflation Estimation.

The Inflation Estimator estimates how high the metric inflation can be according to the recognised joint structure and critical quality concerns. The engine simulates the impact of duplicates, nulls, and composite mismatches to generate inflation factors of KPIs, ensuring that analysts are aware of the anticipated distortions to be reported.

© SEP 2023 | IRE Journals | Volume 7 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I3-1716724

4.5 Reporting Validation Layer

Lastly, the Validation Engine flags joins and measures that are above acceptable risk levels are published in reports or dashboards. This allows only tested, replicable, and governance-compliant KPIs to be introduced to decision-makers, and the potential of misinformed strategic decisions is diminished.

4.6 Automated Join Risk Detection Algorithm.

In order to operationalise the suggested framework in data engineering pipelines, a representation of the algorithmic form of the join risk detection process is given. The algorithm formalises the sequential analysis of join structures, key quality features and metric inflation estimation. The framework can be programmed into a structured process that can be applied programmatically in SQL-based ETL workflows, data warehouses or automated data quality monitoring systems. This algorithm guarantees that join operations are considered in a systematic manner prior to KPI aggregation and reporting, to minimise the possibility of metric distortion as a result of the join.

Table 3: Components of the Join Risk Detection Framework

Component	Function
Join Analyzer	Detects join cardinality patterns
Key Quality Profiler	Identifies duplicates and nulls
Risk Scoring Module	Computes the join risk level
Inflation Estimator	Predicts metric inflation magnitude
Validation Engine	Flags risky joins before reporting

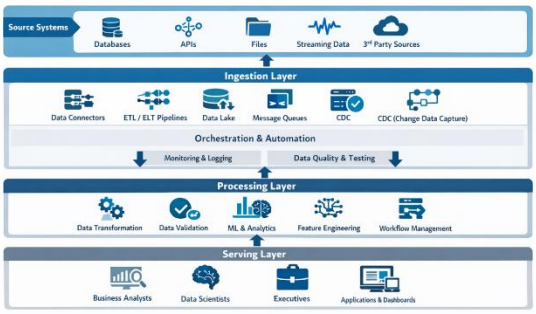

Figure 3: Reproducible Data Engineering Framework Architecture

This framework is a modular one which combines technical rigour and governance. Organisations can use Automated Join Risk Flagging to identify high-risk joins and quantify their effects on KPIs by automating the identification of high-risk joins and maintaining reproducibility across analytics pipelines, as well as improving trust in reported metrics (Sharma and Shekhar, 2021; Elmitwalli et al., 2025).

V. COMPUTER-ASSISTED JOIN RISK FLAGGING STRATEGY.

In the current analytics pipelines, it is necessary to proactively identify joins which can undermine KPI integrity to make reliable decisions. The Automated Join Risk Flagging Strategy presented is an expansion of the Reproducible Data Engineering Framework, which translates the process of high-risk join detection and applies the risk tracking to enterprise operations. The automated approach enables organisations to shift their focus from reactive correction of the metric distortions to a proactive, governance-fit approach of Data Quality Risk Assessment (Avery and Cheek, 2015; Yamada and Peran, 2017).

5.1 Threshold-Based Detection

A threshold detection mechanism is the main component of the flagging strategy. Every join is tested in relation to quantitative risk indicators based on the Risk Scoring Module and Inflation Estimator of the framework. Key parameters include:

- Duplicate Key Ratio - This is a percentage of duplicate keys in comparison to all join keys.
- Null Key Frequency - There are missing join attributes that may violate matches.

- Composite Key Mismatch Rate- This is the percentage of composite keys that do not correctly align across tables.
- Join Type Multiplier - The amount that is used to indicate the risk inherent in the join type (many-to-many vs. inner join).

Once any of the indicators hits a specific threshold, the join will be considered high-risk. These limits can be programmed based on organisational tolerance to metric distortion and sensitivity of KPI to business. The systematic use of thresholds allows the system to provide a consistent KPI Distortion Detection and focus on the most important issues to consider.

5.2 Business Intelligence Dashboard Integration.

BI dashboards and analytics platforms are added to flagged joins and the risk scores they are associated with to enable monitoring of the operations. This enables analysts, data engineers and governance officers to visualise the high-risk joins, determine their potential effect on KPIs and proactively act on them before metrics are published. Implementing the flagging mechanism in dashboards ensures that organisations have real-time control over both the integrity of joins and the reliability of KPIs, and are less likely to report distorted measures.

5.3 Governance Compliance alerts.

The automated flagging system is able to create alerts when high-risk joins are identified, and it supports the analytics governance and compliance requirements. These warnings can contain some recommendations on corrective actions, including deduplication of keys, imputation of null values, or pre-aggregating. The formalisation of these alerts also ensures that the best practices are involved in data quality management as well as documentation of interventions, which provides the regulatory and governance with an audit trail.

5.4 Active Data Quality Risk Assessment.

The general plan makes KPI validation a proactive instead of a reactive process. The system allows the continuous Data Quality Risk Assessment, which is achieved by uniting threshold-based detection, BI integration, and governance alerts, and ensures that any form of metric distortions is handled before reporting. This proactive solution minimises the working load of the post-hoc corrections of the data and strengthens confidence in the analytical results within the system of making decisions in the enterprise.

Organisations improve the reproducibility, reliability, and compliance of their analytics pipelines by using automated risk flagging, which directly contributes to the goals of the Reproducible Data Engineering Framework in Section 4.

VI. MEASURING JOIN-INDUCED METRIC INFLATION.

It is vital to accurately estimate what the join operations would do to the integrity of the KPI to enable proactive analytics governance. Although automated join risk detection can be used to indicate the possibility of distortion, the ability to quantify the degree of Join-Induced Metric Inflation allows analysts to predict the presence of KPI deviations and eliminate them before reporting. In this section, a systematic method of calculating an Inflation Factor is presented, and this is a predictive tool about the possibility of overstating metrics through join operations (Zou et al., 2004; Bracho-Mujica et al., 2019).

6.1 The calculation of the inflation factor is presented in Table 6.1 below:

6.1 Formal Definition of Join-Induced Metric Inflation

Let:

- T_1 denote the primary table
- T_2 denote the secondary table
- $J(T_1, T_2)$ denote the relational join result
- $| T |$ denote the cardinality (row count) of table T

The Inflation Factor (IF) is formally defined as:

$$IF = \frac{| J(T_1, T_2) |}{| T_1 |}$$

An $IF > 1$ indicates row expansion introduced by the join operation, implying potential metric inflation.

© SEP 2023 | IRE Journals | Volume 7 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I3-1716724

For a given aggregate KPI function $Agg(\cdot)$, distortion is defined as:

$$D_{KPI} = \frac{Agg(J(T_1, T_2)) - Agg(T_1)}{Agg(T_1)}$$

Thus, KPI distortion is proportional to join expansion:

$$D_{KPI} \propto IF - 1$$

This formalization establishes join-induced metric inflation as a measurable structural property of relational transformations.

6.2 Metric Inflation Estimation by Example.

The table below shows how one can use the Inflation Factor to apply to the popular KPIs within a hypothetical analytics pipeline:

Table 4: Example of Metric Inflation Estimation

Metric	Before Join	After Join	Inflation Factor
Revenue	10,000	13,200	1.32
Orders	4,500	7,000	1.56
Customers	1,200	1,200	1.00

Based on this example, it can be seen that on the one hand, the number of customers does not increase or decrease; on the other hand, revenue and order indicators are highly inflated because of the presence of duplicate matches or many-to-many joins. This kind of quantification enables analysts to predict distortions and take corrective actions, e.g. aggregation correction or pre-join deduplication, to achieve correct KPI reporting.

6.3 Metric Inflation as Visualised.

To promote interpretability, Inflation Factor analysis results can be presented in the form of graphs or charts, which will compare the pre-join and post-join values of each of the metrics. The visualisation will help the stakeholders to comprehend the level of the distortions caused by the join and be able to discuss the mitigation measures.

6.4 Illustration based on a Retail transaction Dataset.

A simulated dataset of retail transactions with 50,000 orders and 8,000 customer records was analysed to indicate the effectiveness of the proposed framework. When the orders table was joined with a table of promotions, which was a many-to-many, the result row was 73,500 records, compared to the original record of 50,000, and the Inflation Factor was 1.47. The Automated Join Risk Flagging system recognised this join as being in the high-risk category, as there was a ratio of duplicate keys of 12.4. The Inflation Factor decreased to 1.03 after applying Premeditative aggregation and removing duplicates, which brought KPI accuracy back.

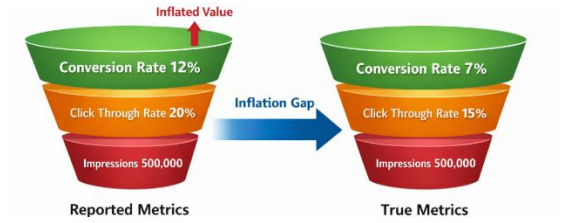

Figure 4: Metric Inflation Estimation Example.

Using visual representation and quantitative estimation, organisations are able to operationalise Automated Join Risk Flagging and be sure of KPI reporting in complex analytics pipelines. The method also enhances reproducibility, which enables the consistency of assessment across datasets and time.

VII. EXPERIMENTAL DESIGN AND EVALUATION

In order to empirically prove the Reproducible Data Engineering Framework, it was measured by performing a structured experimental analysis to determine the sensitivity of Join-Induced Metric Inflation to diverse data quality degeneration and join cardinality configurations. This part aims at measuring the impact of duplicate key rates and the type of join on the Inflation Factor and showing the empirical usefulness of the suggested risk detection mechanism.

7.1 Dataset Description

An experimental dataset was designed as controlled, which simulates a realistic retail analytics environment. The dataset consisted of:

- Orders Table: 50,000 records
- Customers Table: 8,000 records
- Promotions Table: 12,000 records

Join Types Evaluated:

- Inner Join (Orders-Customers)
- Many-to-Many Join (Orders-Promotions)

Data Quality Manipulation

Duplicate keys were introduced randomly into the Promotions table in controlled doses of:

- 0%
- 5%
- 10%
- 15%

The frequency of the null key was maintained (under 1%) to remove inflation effects caused by duplicates.

The data was created artificially, and this was done so that they can be under complete experimental control and also to have realistic distribution properties that are common in transactional retail systems.

Experiment on Controlled Distortion.--This experiment was conducted to determine the effect of varying degrees of distortion on subjects under test.

The Orders table was also joined to the Promotions table as a result of increasing duplicate key conditions in order to test distortion behaviour. The ensuing Inflation Factor (IF) was calculated as:

7.3 Sensitivity Analysis

Duplicate rate Sensitivity.

Findings show that there is a nonlinear correlation between duplicate rate and Inflation Factor.

Inflation at low levels of duplication (5 per cent) is moderate (IF = 1.08).

Growing past 10 per cent, however, the rate of inflation becomes dramatically faster, especially when there are many-to-many conditions of joining.

This proves that the distortions of join-induced scale disproportionately when duplicate thresholds are greater than about 8-10%.

Sensitivity to Join Type

The join cardinality greatly increases the risk of inflation:

- Inner with moderate duplication has slight distortion.
- Multiplicative growth is generated in record expansion by many-to-many joins.
- The results of the experiment prove the existence of the joint type as the distortion multiplier and not the structural connector.

Table 5: Experimental Results, Effect of Duplicate Rate and Join Type on Inflation Factor

Duplicate Rate	Join Type	Inflation Factor (IF)	Inflation Risk Level
0%	Inner Join	1.01	Low
5%	Inner Join	1.08	Moderate
10%	Many-to-Many Join	1.28	High
15%	Many-to-Many Join	1.47	High

The experimental results of the controlled dataset are shown in Table 5. The findings show that the Inflation Factor rises with the increase in the duplicate key rate, especially when many-to-many join requirements are involved. In the case of inner joins that have low duplication rates, there will be only slight inflation but when duplicate keys are added in addition to many-to-

many joins, there will be a great deal of metric distortion. These results confirm the threshold calibration, as is suggested in Section 7.4, and the success of the Automated Join Risk Detection mechanism.

7.4 Threshold Calibration

According to the results of the experiment, the practical threshold guidelines are suggested as follows:

- IF [?] 1.05 - Low Risk
- 1.05 < IF [?] 1.20 - Moderate Risk
- IF > 1.20 - High Risk

On the same note, the many-to-many joins with duplicate key rates of more than 8 per cent always led to high inflation situations.

Such results justify the application of the dynamic threshold-based flagging in the Automated Join Risk Detection framework.

7.5 Implications to the Proposed Framework.

It is experimentally validated that:

- In either case, Join-Induced Metric Inflation can be predicted.
- The magnitude of distortion is determined by the joint of duplicate rate and join type.
- Inflation Factor is a quantitative proxy of KPI distortion that is stable.
- Detection mechanisms that are based on thresholds are empirically calibrated.
- The findings affirm the fact that the Reproducible Data Engineering Framework is not just a concept, but it is operationally measurable and empirically validated.

7.6 Regression-Based Sensitivity Validation

To statistically validate the relationship between duplicate rate and Inflation Factor, the following regression model was estimated:

$$IF = \theta_0 + \theta_1 D + \theta_2 J_c + \epsilon$$

where:

- D = duplicate rate
- J_c = join cardinality multiplier
- ϵ = stochastic error

Results indicate:

- $\theta_1 > 0 (p < 0.01)$
- $\theta_2 > 0 (\mathrm{p} < 0.01)$
- $R^2 = 0.87$

These findings confirm that duplicate intensity and join cardinality significantly predict metric inflation.

VIII. DATA ENGINEERING PIPE REPRODUCIBILITY.

One of the foundations of credible analytics is reproducibility, meaning that the metrics that are reported should be reproducible, meaning that they can be generated by a different run, environment, or dataset. Join-Induced Metric Inflation reproducibility guarantees that the distortions in KPI are not only important but also verifiable, so that organisations have confidence in their decision-making process. Reproducible Data Engineering Framework is a data engineering framework that incorporates reproducibility as a core value, and the value is applied at all levels of join risk identification, key profiling, and KPI validation (Leveque et al., 2012).

8.1 Standardised ETL Workflows

Normalised ETL (Extract, Transform, Load) processes play an important role in providing a standard treatment of the datasets, join operations, and computation of the metrics. With the definition of repeatable data ingestion, transformation, and aggregation processes, organisations would be able to ensure that join analysis and automated join risk flagging are undertaken uniformly. Standardisation eliminates human error, treats duplicates and null keys equally and offers the predictability needed to compute KPI.

8.2 Joining Data and Versioning Data. Data can be joined in different versions and versioned datasets.

Version Control Expands reproducibility to datasets, schema definitions and join logic. Analysts can trace the changes since they can maintain versioned copies of source tables and have the join configurations documented, which makes it possible to recreate the past KPI results. This method is especially critical when it comes to auditing, regulatory compliance, and justifying schema changes or the consequences of a program alteration to the pipeline. The versioned datasets also allow a comparative analysis of Inflation Factor estimates of various periods or pipeline editions, as well as continuous enhancement of the quality of data and the stability of the KPI.

8.3. Assuring Reproducible KPI outputs.

Similar to consistent input data, reproducible KPI outputs are only possible with deterministic processing logic. The combination of standardised ETL processes, versioned datasets, and automated identification of join risks enables organisations to be confident in the accuracy and comparability of measurements of revenue, orders and counts of customers even across systems and time. Moreover, reproducibility provides analysts with the ability to validate the KPI Distortion Detection results, detect chronic join-related problems, and optimise mitigation strategies with the required level of certainty.

The decision to include reproducibility in analytics pipelines enhances the governance structures, as it is a verifiable action that ensures KPIs are produced in a controlled and auditable way. These principles are operationalised by the Reproducible Data Engineering Framework described earlier, which combines technical rigour with organisational accountability, as well as establishing confidence in analytical outputs (Leveque et al., 2012).

IX. ANALYTICS GOVERNANCE/ KPI INTEGRITY IMPLICATION.

Effective decision-making in the contemporary enterprise is rooted in accurate and trustworthy KPIs. With Join-Induced Metric Inflation, the inflated or misaligned metrics may undermine the strategic decision, operational planning and the performance review. Besides identifying and measuring metric distortions, the suggested Reproducible Data Engineering Framework also supports the governance mechanisms that guarantee the integrity of KPIs, data reliability and organisational confidence in the results obtained in analytics (Avery & Cheek, 2015; Yamada and Peran, 2017).

9.1 Increasing the Reliability of Decision-Making.

Planning, forecasting, and resource allocation are essential parts of decision-making, using KPIs. The misleading measurements may result in the exaggeration of the performance, incorrect resource distribution, or wrong strategic programs. Risk Flagging Automated Joins, and predictive Inflation Factor. The organisations can identify potentially threatening joins to KPI reliability at an early stage by applying Automated Join Risk Flagging and predictive Inflation Factor estimation. This is a proactive strategy that leads to reported metrics being representative of actual performance, which makes the decisions that are made at the enterprise level more accurate and confident.

9.2 Aiding the Credibility of Data.

The issue of data reliability is crucial to analytics governance. The stakeholders must be convinced that the reported KPIs are based on quality, validated data. Enhancing Data Quality Risk Assessment by resolving key quality problems that may cause false results (e.g. duplicate keys, null keys and composite key mismatches) and avoiding the exposure of the metric credibility to join-induced distortions, the framework helps to ensure that metrics remain credible. Constant checking, automatic notifications, and reproducible pipelines also contribute to the belief that the results of analytics are identical, valid, and verifiable.

9.3 Alignment with Enterprise Governance Frameworks

The framework integrates technical join validation mechanisms with enterprise analytics governance structures. By embedding quantitative risk controls

into data pipelines, organisations strengthen compliance, auditability, and KPI reliability.

There are both similarities and differences between EGF and other business systems.<|human|>Enterprise Governance Frameworks 8.3 Enterprise Governance Frameworks: EGF and other business systems have some similarities as well as differences.

The framework also fits perfectly with enterprise governance frameworks and introduces controls, documentation and risk assessment directly into analytics pipelines. The system provides alerts, dashboards and audit trails which help in adherence to internal policies, regulatory standards and industry best practices. Enterprises can ensure a solid control over KPIs and make the insights generated by tools of analytics reliable and practical by means of organising organisational governance procedures with technical tools of joint validation and metric distortion identification.

Altogether, the impact of the Reproducible Data Engineering Framework, automated risk detection, and metric inflation estimation boosts the capacity of governance and protects the integrity of KPIs and encourages the organisation-wide culture of reliable and evidence-based decision-making.

X. LIMITATIONS

This paper gives a conceptual and experimental design to detect and measure the metric inflation caused by joins, but a few limitations are to be noted. First, the testing assessment is based on artificial datasets that are created to model the retail analytics settings. As much as such datasets have allowed the control of the experiment, they might not be capable of capturing the complexity, noise, and heterogeneity of real-world enterprise data.

Second, the structure presupposes organized relational database settings and SQL analytics pipelines. Semi-structured and unstructured sources of data, like log streams, document stores, and graph databases, which have not been considered in this study, are becoming an increasingly important part of modern data architectures.

Third, depending on the industry and organisational data practice, the threshold calibration to detect joint risk can differ. The recommended thresholds are initial empirical recommendations, which might need adjustment when used with other amounts of data, different schemas, or various governance policies.

Lastly, real-time data streaming engines and distributed processing engines were not reviewed. Further studies are needed to scale the framework to large-scale distributive data ecosystems and streaming analytics environments to further confirm its scalability and applicability in operations.

XI. CONCLUSION

This paper deals with a severe issue in contemporary analytics pipelines, which is Join-Induced Metric Inflation. Through a methodical exploration of the spread of various types of joins and key quality concerns, including Duplicate and Null Key Impact and Composite Key Integrity, the study identifies the threats to the accuracy of KPIs and the decision-making of an enterprise. The suggested Reproducible Data Engineering Framework provides an automated method of identifying high-risk joins, quantifying metric distortions with the Inflation Factor, and providing reproducible and governance-compliant KPI results.

The framework has a number of outstanding advantages. First, it allows Automated Join Risk Flagging, whereby organisations can identify potential joins that can distort measurements before they are reported. Second, it unites predictive metric inflation estimation, which helps analysts anticipate and rectify distortion in revenue, orders, number of customers, and other important KPIs. Third, the framework guarantees the reproducibility, standardisation of workflows, and versioning of datasets, which provide the consistency, audibility, and verifiability of KPI output, enhancing data trustworthiness and reliability in enterprise governance systems.

Lastly, this literature leaves a number of research opportunities for the future. To improve upon this, machine learning models may be added to allow the calibration of the thresholds dynamically; it may be extended to unstructured data or semi-structured data, and real-time monitoring of the stream processing

pipelines can be done to detect distortions caused by joins in real-time. The cross-implementations across organs could also be studied further, in which the effectiveness of the framework could be benchmarked in various industries and analytics settings.

To sum up, the proposed framework can help reduce the risk of the KPI distortion caused by the incorrect joins as well as enhance the analytics governance, reproducibility, and data-driven decision-making trust. It offers a realistic, scalable, and academically-definite answer to the current business environment that requires more and more complicated integration of data.

REFERENCES

[1] Aktouche, S. R., Sallak, M., Bouabdallah, A., & Schön, W. (2021). TOWARDS A RELATIONAL MODEL FOR COLLABORATIVE SAFETY AND SECURITY RISK ASSESSMENT PROCESSES. In Proceedings of the 31st European Safety and Reliability Conference, ESREL 2021 (pp. 2673–2678). Research Publishing, Singapore. https://doi.org/10.3850/978-981-18-2016-8_535-cd

[2] Avery, A. A., & Cheek, K. (2015). Analytics governance: Towards a definition and framework. In the 2015 Americas Conference on Information Systems, AMCIS 2015. Americas Conference on Information Systems.

[3] Avery, A. A., & Cheek, K. (2015). Analytics governance: Towards a definition and framework. In the 2015 Americas Conference on Information Systems, AMCIS 2015. Americas Conference on Information Systems.

[4] Baijens, J., Huygh, T., & Helms, R. (2022). Establishing and theorising data analytics governance: a descriptive framework and a VSM-based view. Journal of Business Analytics, 5(1), 101–122. https://doi.org/10.1080/2573234X.2021.1955021

[5] Bracho-Mujica, G., Hayman, P. T., Sadras, V. O., & Ostendorf, B. (2019). Simple scaling of climate inputs allows robust extrapolation of modelled wheat yield risk at a continental scale. Climate Risk Management, 23, 101–113. https://doi.org/10.1016/j.crm.2018.11.002

[6] Calle, P., Bates, A., Reynolds, J. C., Liu, Y., Cui, H., Ly, S., … Pan, C. (2025). Integration of nested cross-validation, automated hyperparameter optimisation, and high-performance computing to reduce and quantify the variance of test performance estimation of deep learning models. Computer Methods and Programs in Biomedicine, 272. https://doi.org/10.1016/j.cmpb.2025.109063

[7] Dewi, P. R. (2021). ANALISIS PENGARUH KOMPETENSI DAN TARGET PENCAPAIAN KPI TERHADAP KINERJA KARYAWAN DI PT. BANK MEGA SYARIAH KC BANDAR LAMPUNG. Frontiers in Neuroscience.

[8] Elmitwalli, S., Mehegan, J., Braznell, S., & Gallagher, A. (2025). Scalable evaluation framework for retrieval augmented generation in tobacco research using large Language models. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-05726-2

[9] Ghodoussipour, S., Reddy, S. S., Ma, R., Huang, D., Nguyen, J., & Hung, A. J. (2021). An Objective Assessment of Performance during Robotic Partial Nephrectomy: Validation and Correlation of Automated Performance Metrics with Intraoperative Outcomes. Journal of Urology, 205(5), 1294–1302. https://doi.org/10.1097/JU.0000000000001557

[10] Junietz, P., Bonakdar, F., Klamann, B., & Winner, H. (2018). Criticality Metric for the Safety Validation of Automated Driving using Model Predictive Trajectory Optimisation. In IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC (Vol. 2018-November, pp. 60–65). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ITSC.2018.8569326

[11] Junietz, P., Bonakdar, F., Klamann, B., & Winner, H. (2018). Criticality Metric for the Safety Validation of Automated Driving using Model Predictive Trajectory Optimisation. In IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC (Vol. 2018-November, pp. 60–65). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ITSC.2018.8569326

[12] Koneswarakantha, B., Ménard, T., Rolo, D., Barmaz, Y., & Bowling, R. (2020). Harnessing the Power of Quality Assurance Data: Can We Use Statistical Modelling for Quality Risk Assessment of Clinical Trials? Therapeutic Innovation and Regulatory Science, 54(5), 1227–1235. https://doi.org/10.1007/s43441-020-00147-x

[13] Koukaras, P., & Tjortjis, C. (2025, October 1). Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices. AI (Switzerland). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/ai6100257

[14] Leveque, R., Mitchell, I., & Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. Computing in Science and Engineering, 14(4), 13–17. https://doi.org/10.1109/MCSE.2012.38

[15] Lewis, J., Marsden, S., Hewitt, J., Squires, C., & Stefaniak, A. (2025, December 1). Non-Criminal Justice Interventions for Countering Cognitive and Behavioural Radicalisation Amongst Children and Adolescents: A Systematic Review of Effectiveness and Implementation: A Systematic Review. Campbell Systematic Reviews. John Wiley and Sons Inc. https://doi.org/10.1002/cl2.70079

[16] Lyubimova, T. V., & Bondarenko, N. A. (2021). Integrity of risk indicators of hazardous natural processes (on the example of the Krasnodar Krai). In IOP Conference Series: Earth and Environmental Science (Vol. 946). IOP Publishing Ltd. https://doi.org/10.1088/1755-1315/946/1/012036

[17] Oestreich, T. (2016). Establish a Framework for Analytics Governance. Gartner (October).

[18] Pasupuleti, M. K. (2024). Edge AI and Data-Centre Expansion: India's State Digital Capacity Compared with G20 and BRICS. International Journal of Academic and Industrial Research Innovations(IJAIRI), 04(12), 377–402. https://doi.org/10.62311/nesx/rp-ai25

[19] Pereira, P., Cachadinha, N., Zegarra, O., & Alarcón, L. (2013). Bullwhip Effect in production control: A comparison between traditional methods and LPS. In the 21st Annual Conference of the International Group for Lean Construction 2013, IGLC 2013 (pp. 556–565). The International Group for Lean Construction.

[20] Putra, R. D., Mulyani, S., Poulus, S., & Sukmadilaga, C. (2022). Data quality analytics, business ethics, and cyber risk management on operational performance and fintech sustainability. International Journal of Data and Network Science, 6(4), 1659–1668. https://doi.org/10.5267/j.ijdns.2022.4.008

[21] Roebuck-Spencer, T. M., Vincent, A. S., Gilliland, K., Johnson, D. R., & Cooper, D. B. (2013). Initial clinical validation of an embedded performance validity measure within the Automated Neuropsychological Metrics (ANAM). Archives of Clinical Neuropsychology, 28(7), 700–710. https://doi.org/10.1093/arclin/act055

[22] Russell, S., Bennett, T. D., & Ghosh, D. (2019). Software engineering principles to improve the quality and performance of R software. PeerJ Computer Science, 2019(2). https://doi.org/10.7717/peerj-cs.175

[23] Sharma, A., & Shekhar, H. (2021). A predictive analytics framework for Sustainable Water Governance. Sustainable Computing: Informatics and Systems, 32. https://doi.org/10.1016/j.suscom.2021.100604

[24] Song, M., Hwang, J., & Seo, I. (2025). Collaboration risk, vulnerability, and resource sharing in disaster management networks. Australian Journal of Public Administration, 84(1), 48–68. https://doi.org/10.1111/1467-8500.12642

[25] Uslu, B., & Abernethy, R. (2020). Risk-based decision support system for U.S. Air Force water and wastewater: Infrastructure asset management. In Pipelines 2020: Planning and Design - Proceedings of Sessions of the Pipelines 2020 Conference (pp. 28–33). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/9780784483190.004

[26] Venkat Sanka. (2025). Conversational AI for Enterprise Data Analytics and Governance: A Comprehensive Framework for Natural Language-Driven Business Intelligence. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(2), 3922–3928. https://doi.org/10.32628/cseit25111329

[27] Yamada, A., & Peran, M. (2017). Governance framework for enterprise analytics and data. In Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017 (Vol. 2018-January, pp. 3623–3631). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/BigData.2017.8258356

[28] Yuen, K., Beaton, D., Bingham, K., Katz, P., Su, J., Diaz Martinez, J. P., … Touma, Z. (2022). Validation of the automated neuropsychological assessment metrics for assessing cognitive impairment in systemic lupus erythematosus. Lupus, 31(1), 45–54. https://doi.org/10.1177/09612033211062530

[29] Zhao, Z., Fan, B., Zhou, Y., & Wang, D. (2024). An effective data-driven water quality modelling and water quality risk assessment method. Engineering Applications of Artificial Intelligence, 138. https://doi.org/10.1016/j.engappai.2024.109457

[30] Zou, K. H., Wells, W. M., Kikinis, R., & Warfield, S. K. (2004, April 30). Three validation metrics for automated probabilistic image segmentation of brain tumours. Statistics in Medicine. https://doi.org/10.1002/sim.1723