State Of The Art And Literature Review for Project 4:
Intelligent analytics – automatic report generation

By

ANISSA KEROUAZ

BOUBKER ENNAJY

JAMAL EDDINE OBEIDAT

NAOUAR EL BOUMASHOULI

AMINE ACHOUHAM

Training: Data Engineering and Data analytics: DataOps Specialization,

RABAT, November 2025

Summary

Table of pages

List of figures

General Introduction:

This document presents the state-of-the-art and literature review for our project, carried out as part of the Arkx Talent Factory bootcamp within the JobInTech program, in collaboration with CDG's DXC Technology. Conducted under the Data Engineering and Data Analytics: DataOps Specialization, our project, Project 4: Intelligent Analytics – Automatic Report Generation, focuses on developing an intelligent platform that automatically generates dynamic, interpretable, and shareable reports from data collected through existing pipelines.

Recent advances in data engineering and automation have dramatically transformed how organizations leverage information. Companies today collect massive amounts of data from diverse sources—including processing pipelines, transactional systems, and online platforms—all aimed at improving decision-making and optimizing performance. Despite this, turning data into actionable insights remains challenging, particularly when producing accurate, timely, and understandable reports.

Manual report generation is often repetitive and time-consuming, with a high potential for errors. Teams must aggregate heterogeneous datasets, calculate performance metrics (KPIs), and write explanatory summaries, which slows down analytical workflows and makes it harder to respond to evolving data. These challenges highlight the need for smarter, more automated reporting solutions.

Our project aims to address these challenges by designing a platform capable of automatically generating daily and weekly reports on pipeline performance, including success rates, processed volumes, and average latency, as well as data quality metrics such as completeness, uniqueness, and anomaly rates. The platform is also designed to detect trends and anomalies, inserting contextual alerts whenever significant changes occur.

A key innovation explored in the literature and incorporated in our approach is the use of large language models (LLMs) for AI-driven report generation. By leveraging LLMs, the platform can transform complex analytical data into concise, human-readable summaries, enabling stakeholders to quickly interpret insights without manually parsing raw metrics. This approach represents a major advancement over traditional reporting methods and aligns with recent trends in automated data interpretation.

From a technical perspective, the project will rely on tools for data collection, processing, governance, AI-driven report generation using LLMs, and data visualization to communicate insights effectively.

This document is structured to first present the context and challenges of automated reporting, then review relevant literature and technical approaches, and finally highlight insights and trends that inform the design of our platform.

Chapitre1 : Introduction

1. Motivation :

Modern data pipelines generate huge amounts of information every day. Relying on manual reporting is slow, error-prone, and often leaves important trends or anomalies unnoticed. By automating reporting with AI, we can create a system that not only summarizes data but also explains it, providing insights that are easy to understand and act upon.

2. Problem Statement:

Current reporting processes face several challenges:

- Manual report creation takes time and is prone to errors
- Insights are often delayed, limiting timely action
- Reports are difficult to interpret, making it hard to spot trends or anomalies quickly
- There's a lack of integration between raw pipeline data, quality metrics, and visualization

These issues highlight the need for a solution that combines automation, analysis, and AI-driven explanation in one seamless platform.

3. Objectives of the Project:

Our platform aims to:

- Automate performance and data quality reports for operational pipelines, including metrics like success rates, processing volumes, latency, and anomaly scores.
- Detect trends and anomalies automatically, highlighting deviations before they cause problems.
- Provide AI-driven textual summaries, translating data into human-readable insights for better decision-making.

Chapter 2: State of the Art / Literature and Tools Review

Traditional reporting relies heavily on dashboards and BI tools like Power BI or Grafana. While these tools are powerful, they usually require manual setup, and they don't provide intelligent interpretation or dynamic textual summaries.développement, mettant en lumière les choix techniques et esthétiques qui sous-tendent chaque aspect de mon projet.

Automation in data workflows brings speed, reliability, and consistency. Following DataOps principles reduces human error, enforces data quality, and ensures that reporting is both accurate and reproducible.

1. Automatic Data Understanding and Schema Inference:

The ability to automatically interpret raw data has become essential in today's data processing ecosystems. Heterogeneous file formats, such as CSV, JSON, Parquet, Excel, and text logs, are becoming more and more common in organizations. These file formats are frequently characterized by ambiguous types, inconsistent structures, missing values, or mixed semantic meanings. These inputs need to be converted into a standardized representation with a well-defined schema before any analytical or reporting workflow can start. Large-scale or real-time data pipelines cannot be integrated with manual inspection due to its slowness and error-proneness.

The goal of schema inference is to directly determine the structure and data types of a dataset from its raw input. Contemporary frameworks guess column types, identify patterns, and draw attention to anomalies using heuristics, statistics, or AI-based techniques. The need for more flexible and intelligent solutions, like the one this project suggests, is prompted by the fact that current tools still have trouble handling jumbled or unclear data.

2. Multi-Format Data Ingestion Approaches:

As we explored existing tools for loading data without prior knowledge of its schema, we found that some BI platforms implement this capability, but with a limitation: the user must manually specify the schema details (such as file format, header presence, separator type for CSV files, etc.).

Another solution comes from Python libraries like Pandas, Polars, and Dask. These tools provide powerful functions for loading a wide range of data formats (the choice between them will be discussed later). However, what they lack is

automation—how can we integrate their capabilities into a single system that performs schema detection and loading automatically?

We must also consider chunked processing. Since datasets can be extremely large, it may be impossible to load them entirely into memory. Handling data in chunks is therefore essential to ensure scalability and reliable performance.

3. Multi-Format Data Ingestion Approaches:

As we explored existing tools for loading data without prior knowledge of its schema, we found that some BI platforms implement this capability, but with a limitation: the user must manually specify the schema details (such as file format, header presence, separator type for CSV files, etc.).

Another solution comes from Python libraries like Pandas, Polars, and Dask. These tools provide powerful functions for loading a wide range of data formats (the choice between them will be discussed later). However, what they lack is automation—how can we integrate their capabilities into a single system that performs schema detection and loading automatically?

We must also consider chunked processing. Since datasets can be extremely large, it may be impossible to load them entirely into memory. Handling data in chunks is therefore essential to ensure scalability and reliable performance.

4. Automatic Schema Detection Techniques:

Schema inference presents additional difficulties. Identifying column types, counting rows and columns, calculating the number of unique values per column, and other tasks are all necessary even if we are able to successfully load the data into our system. We looked into a number of prior Python libraries to solve this.

Summary statistics and automated data profiling are offered by pandas-profiling / ydata-profiling. But in our tests, we discovered that it returned None for every column type, indicating that it did not infer the schema at all.

Additionally, pandas_datatypes carries out basic profiling and type detection. Although it has a significant limitation—that it does not infer the categorical type—it generally operated as intended. The accuracy of the resulting schema is decreased because any categorical column is instead inferred as a simple string.

5. Foundations of data analysis and reporting:

What are statistics?

Statistics means the use of data to investigate trends and relationships hidden in data, used by governments and scientists. It requires a collection of data, then we provide a summary using descriptive statistics. There's also inferential statistics to formally test hypotheses and make estimates about the population. Finally, we interpret the results and communicate findings. The crucial factor that pushes us to use this approach is when we have an ordered population and we don't want to lose the order or the potential pattern hidden behind it.

Sampling – extracting a representative subset

Collecting data from the entire population is complicated, time consuming, and expensive. Instead, we can use a technique called sampling. Sampling, or the extraction of a sample from data, is a technique that focuses on the use of a part of the data that represents the full data. The sample preserves the distribution of numeric variables, diversity of the categorical variables, the outliers and minority classes, and other aspects.

Choosing sample size

Decide on your sample size either by looking at other studies in your field or using statistics. A sample that's too small may be unrepresentative of the sample, while a sample that's too large will be more costly than necessary, for EDA or exploratory data analysis, it is highly recommended to use up to 10% of the dataset, this keeps visuals and summary statistics accurate enough

Types of sampling:

Simple random sampling:

Every member of the population has an equal chance of being included, this method requires a complete list of the population, from which members are chosen randomly.

The SRS is most suitable when the population is relatively homogeneous, meaning that all individuals have similar characteristics or attributes or when we have a limited resource.

Stratified sampling:

Focuses on dividing the population into distinct subgroups or strata based on certain characteristics, such as age, gender, or income level. Samples are then

randomly selected from each stratum in proportion to their representation in the population. This method ensures that each subgroup is adequately represented in the sample, we can use this type of sampling when the population exhibits variability in characteristics of interests or we have a specific subgroup to analyze.

Cluster sampling

This technique involves dividing the population into clusters or groups, such as households or schools, and then randomly selecting some clusters to include in the sample, this method is efficient when it's difficult or impractical to obtain a complete list of individuals in the population, as clusters can be sampled more easily.

Systematic sampling

The technique selecting members from a larger population at a regular interval, determined by dividing the population size by the desired sample size. After randomly selecting a starting point within the first interval, the researcher selects every nth individual

descriptive statistics & visual exploration

Once the data collected is ready, the next step is to use descriptive statistics that summarize the data. There are various ways and techniques to inspect the data, such as the organization of the numerical features in frequency distribution visualizations, such as:

Histogram: graphical representation of the distribution of a numeric variable. It divides the data into bins and counts how many observations fall into each bin.

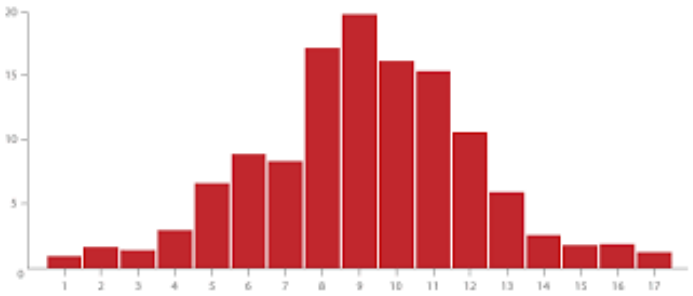

Figure 1:histogram

- This highlights how the values of a variable are spread, whether the distribution is systematic or skewed, whether there are outliers, and so much more.

- ***Skewed***: shows whether the data is symmetric or leans toward one side.

Positive skew (right-skewed): Long tail on the right Many small values, few very large

Negative skew (left-skewed): Long tail on the left Many large values, few very small ones.

- Outliers are observations that are significantly different from other values, like an extremely large or small data points.

There are various ways to inspect data, like displaying data from a key variable in a bar chart or boxplot.

Bar chart:

A visualization used to display and compare the frequencies, counts, or proportions of categorical data. Each category is represented as a bar, and the height or length of the bar indicates its value.

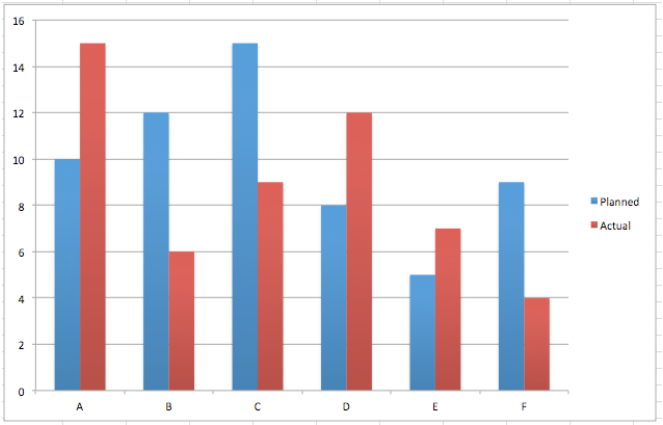

Figure 2: bar chart

The provided chart highlights how categories are distributed in the dataset, we can extract which categories are the most common, determine whether the data is balanced or not, which categories are rare.

Another way to inspect the data is by visualizing the relationship between variables using heatmap and scatter plot.

Heatmap

A ***heatmap*** is a graphical representation that uses color to show the strength of relationships between variables.

In data analysis, the most common type is a ***correlation heatmap***, which visualizes how numerical features relate to each other, helps to quickly identify strongly correlated columns and feature that may impact the target variable and much more.

- ***Correlation*:** measures the strength and direction of the relationship between two numerical variables.

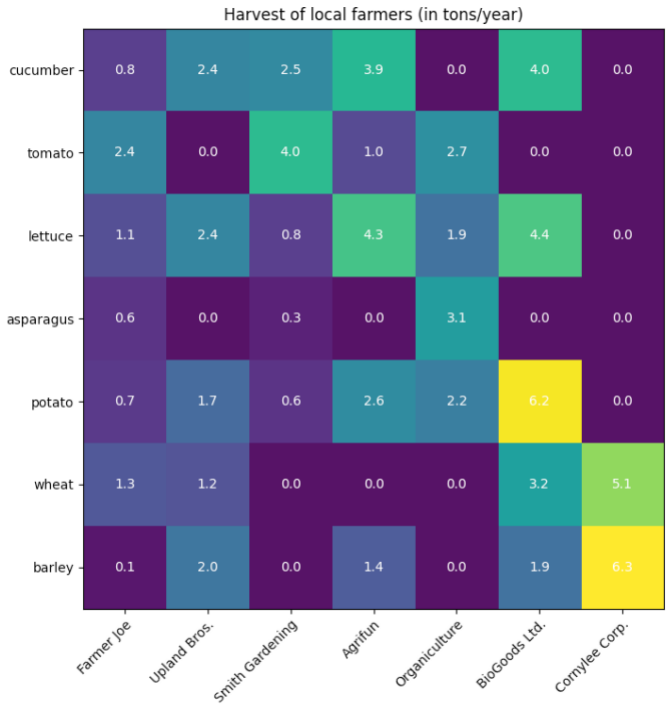

Figure 3:heatmap

Scatterplot:

A visualization used to show the relationship between two numerical variables, it helps analysts see patterns, trends, clusters, and potential correlations between features. It highlights the presence of outliers or clusters upon the dataset.

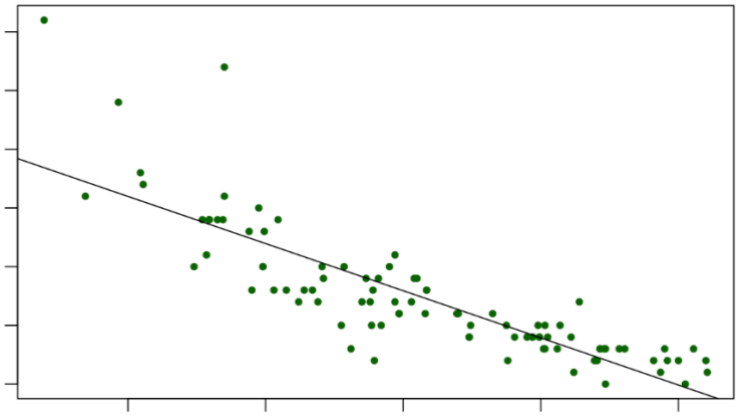

Figure 4: scatterplot

In this section, we reviewed the foundational concepts behind statistical analysis and exploratory data reporting. We covered how descriptive statistics summarize data, how sampling allows working efficiently with large datasets while preserving key properties, and how visualizations help uncover patterns, relationships, and potential issues.

By understanding these concepts, we ensure that our intelligent reporting pipeline is built on sound methodological principles. This foundation not only improves data quality and insight generation but also guides further analyses and decision making in a systematic and reliable way.

6. AI in Reporting :

Artificial intelligence, particularly Large Language Models (LLMs) and multi-agent systems, has recently emerged as a powerful tool for automating and enhancing reporting. These AI-driven systems go beyond static dashboards, offering not just visualization but semantic understanding, reasoning, and natural language synthesis.

Recent research shows that AI can be used to:

- Detect anomalies and trends automatically:

AI models can continuously monitor pipeline metrics or dataset features and identify deviations from expected patterns. For example, if the volume of processed data drops significantly or a key metric shows unusual variance, AI

agents can flag this and include explanations in reports. This reduces reliance on manual monitoring and ensures timely alerts.

- Translate structured data into natural language summaries

Using LLMs, structured datasets such as tables or CSV files can be transformed into human-readable narratives. These narratives can describe distributions, correlations, outliers, or key performance indicators. For instance, Data2Text Agents (Liu et al., 2023) demonstrate how tabular data can be automatically summarized into coherent textual insights, bridging the gap between raw metrics and actionable intelligence.

- Generate charts and visualizations with reasoning and context

Beyond plotting data, AI agents can select the most appropriate visualization type based on dataset characteristics, trends, and the questions being answered. The METAL framework (Li et al., 2025) illustrates iterative multi-agent processes where agents not only generate charts but review, critique, and refine them, producing high-quality, context-aware visualizations.

Notable Works and Insights:

- DAgent (Xu et al., 2025): Demonstrates a multi-agent architecture with Planner, Tool, and Memory modules. The Planner decomposes analytical goals into steps, the Tool executes queries and visualizations, and Memory maintains context across tasks. This stepwise reasoning inspired our Supervisor, Assistant, and Metadata Agents, enabling coherent, multi-step report generation.
- METAL (Li et al., 2025): Focuses on collaborative chart generation, using iterative feedback between agents to refine visualizations. This idea informs our Assistant Agent's visualization feedback loop, ensuring that plots not only represent data accurately but also highlight meaningful trends.
- AutoAgents (Chen et al., 2024): Introduces dynamic creation and coordination of specialized agents. Tasks are assigned to agents based on the data's characteristics and analytical goals, while a supervising agent oversees the process. This approach inspired our Supervisor Agent, enabling flexible delegation and dynamic reasoning.
- Data2Text Agents (Liu et al., 2023): Combine statistical reasoning with natural language generation to produce narrative summaries from structured data. Our Metadata Agent draws directly from this work,

converting column types, distributions, and relationships into textual insights.

- VizML (Hu et al., 2019): Applies machine learning to predict suitable visualizations from dataset features. This principle guides our Assistant Agent in recommending visualizations that best communicate trends, anomalies, and relationships.
- AutoReport (Nguyen et al., 2021) and Table2Text (Roberts et al., 2020): Demonstrate automated KPI extraction and table-to-text reporting. These studies support our approach to report generation, ensuring that numerical outputs are translated into fluent, actionable narratives.

Limitations and Gaps in Current Approaches:

Despite these advances, existing solutions still face notable challenges:

- Many systems remain semi-automated, requiring human configuration for tasks such as chart selection, KPI definition, or summary framing.
- Most approaches lack end-to-end reasoning from raw data ingestion to fully synthesized narrative reports. AI models may generate text or charts, but rarely combine both with semantic understanding and anomaly detection in one pipeline.
- Integration gaps exist between structured analysis, visualization, anomaly detection, and textual explanation. Few systems unify these components into a coherent workflow that is both reproducible and interpretable.
- Adaptability and dynamic agent collaboration are limited; most solutions cannot automatically assign subtasks based on dataset characteristics or adjust strategies in real time.

How Our Platform Addresses These Gaps:

Our proposed platform combines LLM-powered agents with automation and multi-agent orchestration, addressing the gaps identified above:

- Full end-to-end reasoning: From raw pipeline data ingestion to semantic analysis, visualization, and textual reporting.
- Dynamic task decomposition: Supervisor Agent allocates tasks based on dataset characteristics and analytical objectives.

- Iterative refinement: Agents provide feedback loops for charts, summaries, and anomaly explanations.
- Semantic grounding: Metadata Agent ensures reports are interpretable and contextualized for humans, not just generated programmatically.
- Scalability and flexibility: Multi-agent system adapts to different datasets, metrics, and reporting frequencies without manual reconfiguration.

By integrating insights from DAgent, METAL, AutoAgents, Data2Text, VizML, and AutoReport, our platform represents a next-generation approach to automated reporting, combining AI reasoning, collaboration, and explainability in one cohesive system.

Chapter 3: Proposed Solution:

1. System Concept :

Our platform takes raw pipeline data and transforms it into **intelligent, readable reports**. At the heart of the system are **agents powered by LLMs**, working together to understand data, plan analyses, execute calculations, and generate summaries in natural language. This ensures that the reports are not just accurate, but also **meaningful and easy to interpret**.

2. Functional Objectives:

The platform will:

- Automatically generate daily or periodic performance and data quality reports
- Detect trends and anomalies without manual intervention
- Produce AI-driven textual summaries for clear insights
- Enable report distribution via email or dashboards

3. Proposed Architecture (High-Level) :

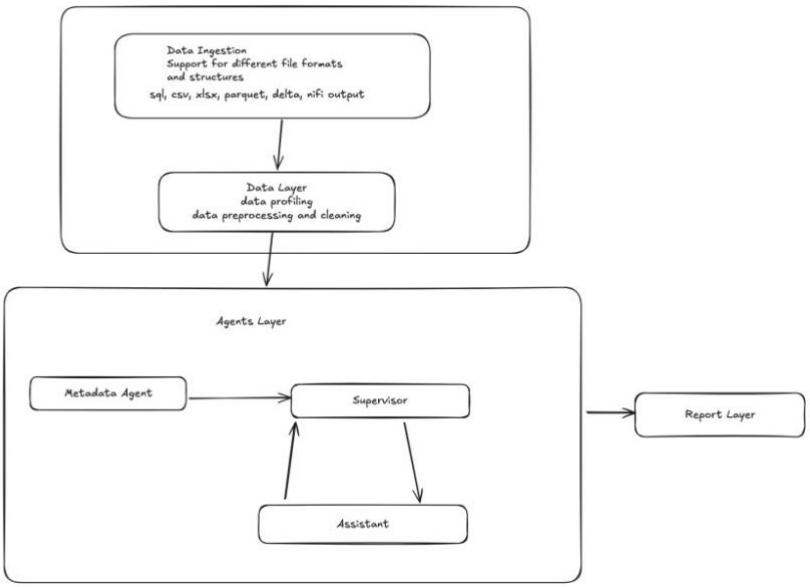

Figure 5:Project High Level Architecture.

Architecture Description:

- Data Collection: Support multiple file formats and connections
- Data Profiling: Python tools analyze schema, distributions, correlations, and anomalies

Agent-Based Analysis:

- Metadata Agent: understands the dataset and creates summaries
- Supervisor Agent: plans analysis and assigns tasks

- Assistant Agent: executes calculations, generates charts, and refines outputs
- Report Generation: narrative summaries and visualizations are compiled into PDF or HTML reports
- Visualization & Distribution: reports are shared with stakeholders and integrated into dashboards

4. Technology Stack :

Component	**Tools / Framework**
Metadata Management	OpenMetadata
Analytics & AI	Python (Pandas, Polars), LLMs (OpenAI GPT, Llama 3, Mistral), LangChain
Visualization	Power BI, Grafana, Matplotlib, Plotly, Seaborn
Reporting	Jinja2, ReportLab, Markdown, WeasyPrint
Execution	Docker, Jupyter kernel

Why this stack? It balances flexibility, scalability, and ease of integration, while LLMs provide reasoning and natural language generation.

Expected Benefits:

- Reduce manual reporting effort significantly
- Detect trends and anomalies faster and more reliably
- Provide clear, actionable insights in human-readable reports
- Improve decision-making and operational efficiency

Chapter 4: Conclusion:

1. Summary of Findings from Research:

Through our research and literature review, several key insights have emerged about the current landscape of automated reporting and intelligent analytics:

The importance of automation and DataOps principles: Modern data pipelines generate vast amounts of data daily, and traditional manual reporting is no longer sufficient. Automation ensures that reports are generated consistently, quickly, and accurately, while also reducing the risk of human error. DataOps practices improve workflow reliability and reproducibility, forming a strong foundation for intelligent reporting systems.

The transformative potential of AI and LLMs: Large Language Models and multi-agent systems have demonstrated the ability to not only automate computations and visualizations but also to reason about data, detect anomalies, and generate human-readable narratives. Research works such as DAgent, METAL, AutoAgents, and Data2Text Agents show that specialized agents working collaboratively can produce structured, contextualized, and interpretable outputs from complex datasets.

Limitations of current solutions: While dashboards, BI tools, and existing multi-agent frameworks provide some level of automation, they often remain semi-automated, lack end-to-end reasoning, and fail to integrate semantic understanding, anomaly detection, and narrative reporting in a single workflow. Most tools cannot dynamically adapt to new datasets or automatically adjust their reasoning and visualization strategies.

Gap and opportunity: There is a clear opportunity to develop a system that bridges these gaps, combining dynamic agent collaboration, AI-driven reasoning, semantic understanding, visualization, and natural language summarization. Our proposed platform leverages these insights, aiming to deliver a truly intelligent reporting system capable of autonomous analysis, contextual explanation, and actionable insight generation.

2. Contribution of the Proposed Platform:

Our AI-Driven Automated Report Generation System addresses the limitations identified in existing solutions by integrating:

- **Metadata profiling and semantic interpretation** through the Metadata Agent, providing a meaningful understanding of the dataset.
- **Strategic planning and task delegation** via the Supervisor Agent, ensuring that analysis is structured, coherent, and contextually relevant.
- **Execution, visualization, and iterative refinement** through the Assistant Agent, producing high-quality charts, tables, and explanatory summaries.

- **Natural language synthesis** for generating readable reports that translate numerical and statistical outputs into actionable insights.
- **Scalable, end-to-end orchestration**, enabling integration with operational pipelines (e.g., NiFi) and automated distribution of reports.

This combination ensures that the system is not just **automated**, but **intelligent**, capable of reasoning about the data, detecting trends or anomalies, and explaining results in a way that is meaningful for decision-makers.

3. Contribution of the Proposed Platform:

The deployment of this platform is expected to deliver multiple benefits:

- Efficiency and consistency: Reports can be generated without manual intervention, saving time and reducing errors.
- Improved decision-making: Timely detection of trends and anomalies enables faster response to operational issues.
- Enhanced interpretability: Semantic summaries and visualizations provide context, making complex data understandable.
- Scalability and adaptability: The agent-based architecture allows the platform to evolve as new datasets, metrics, or analytical requirements emerge.

Conclusion

In this review, we explored how organizations currently handle reporting and the ways AI is reshaping this process. Traditional dashboards and BI tools are useful,

but they often fall short in providing timely, interpretable, and actionable insights. Research on multi-agent systems and LLMs shows that AI can not only automate analysis but also explain data in a way that humans can understand. These insights give a clear foundation for building smarter reporting solutions that save time, highlight trends and anomalies, and help decision-makers act with confidence.