IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 9, September 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309086|

Data Quality and Integration: The AI-Driven Evolution

Ramakanth Reddy Vanga

University of Minnesota, USA

ABSTRACT: This article looks at how artificial intelligence has changed two very important areas of data management: data quality management and data integration. It looks at how AI-powered tools are changing old, time-consuming processes by automatically finding and fixing data errors, which cuts down on human work and makes things more accurate. There is also talk about the rise of cloud-native data pipeline services, which show how this change is testing traditional third-party integration models and changing the way data is integrated. There are predictions about the market, trends in the industry, and what these new technologies mean for businesses and the data management industry as a whole in this piece.

KEYWORDS: Artificial Intelligence (AI), Data Quality Management, Data Integration, Cloud-native Data Pipelines, Real-time Analytics

I. INTRODUCTION

The field of data management is changing very quickly. Two important areas are going through big changes: data quality management and data merging. This article talks about how artificial intelligence (AI) is changing these areas and what that means for companies and tools for managing data.

It was worth USD 72.79 billion in 2020, and it's expected to be worth USD 208.87 billion by 2028, thanks to a 14.2% compound annual growth rate (CAGR) from 2021 to 2028 [1]. A big reason for this exponential growth is that AI and machine learning are being used more and more in data management. A new study by Gartner says that by the end of 2024, 75% of organizations will move from testing AI to using it fully. This will lead to a 5X growth in streaming data and analytics infrastructures [2].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309086|

AI-powered tools are changing the way data quality management is done, which used to require a lot of work. These tools can find and fix data errors automatically, which saves up to 80% of the time needed to clean data by hand. IBM's InfoSphere Information Server for Data Quality, for example, has said that its clients' data cleaning time has been cut by 65%, and the quality of their data has improved by an average of 25% [1].

In the same way, there is a paradigm shift happening in the world of data merging. Big platforms like Stripe and Salesforce are starting to offer cloud-native data flow services, which are changing the way third-party integrations are usually done. The global data integration market is worth $7.82 billion, and this change is expected to have a big effect on it. From 2021 to 2026, cloud-based integration solutions are expected to grow at a CAGR of 21.7% [2].

While we learn more about these changes, it becomes clear that the coming together of AI-driven data quality management and direct data flow services is not a small improvement, but a complete overhaul of how companies handle data in the modern era.

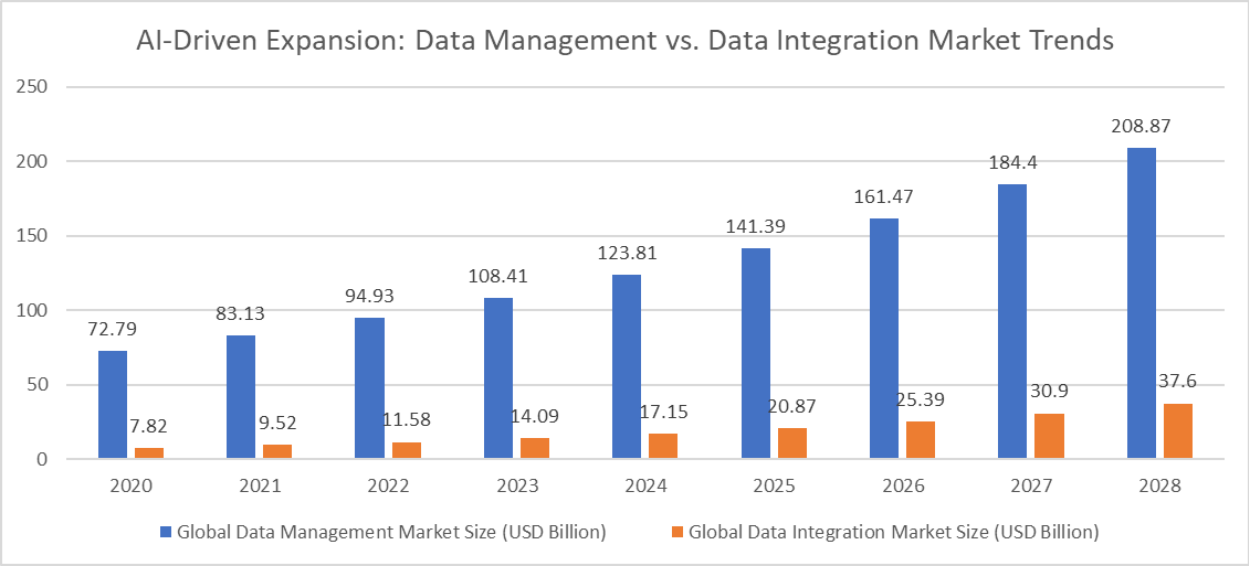

Fig. 1: Projected Growth of Global Data Management and Integration Markets (2020-2028) [1, 2]

II. THE CHANGING FACE OF DATA QUALITY MANAGEMENT

Traditional Approaches

In the past, managing the quality of data required a lot of work and several important steps. Experian Data Quality did a study that showed businesses spent an average of 12% of their income on chores related to data, and that bad data cost them 15% to 25% of their operating budgets [3]. This is what the standard process usually looks like:

1. To understand existing data, data profiling was often done by hand, which for big datasets could take weeks or even months.
2. Seeing where the problems are: Research shows that bad data causes up to 40% of business projects to fail [4].
3. Implementing methods for standardization: This step often needed custom coding, which took a long time and was prone to mistakes.
4. Cleaning up new data: A data scientist could spend up to 80% of their time cleaning up data by hand [3].

A common challenge in this process has been dealing with variations in data representation. For example, the country name "USA" might appear in multiple forms:

- USA
- United States of America
- United States
- US of America

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309086|

In the past, these differences were fixed by hand to keep things consistent, and all the different versions were usually standardized to a single format like "United States of America." This process took a lot of time and was prone to mistakes (about 1-5% of the time) [4].

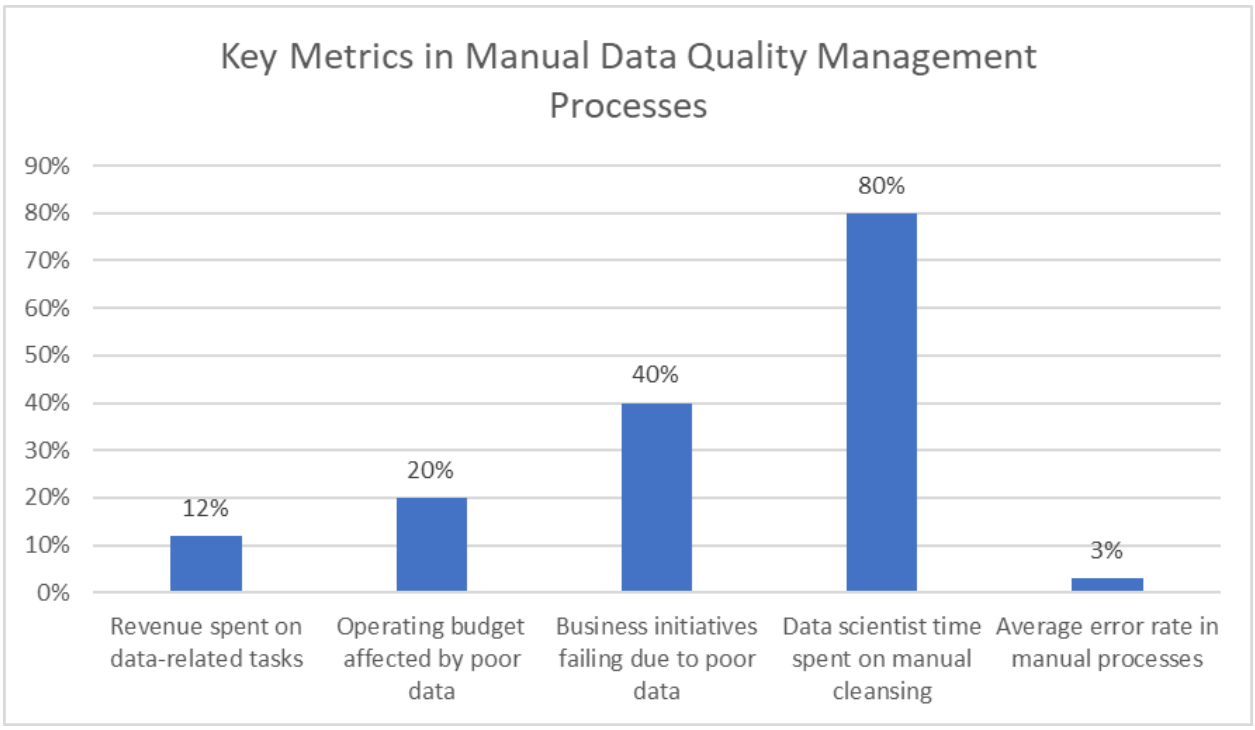

Fig. 2: Impact of Traditional Data Quality Management on Business Operations [3, 4]

III. THE AI REVOLUTION IN DATA QUALITY

Big changes are happening in this world because of AI. A MarketsandMarkets study says that the global market for data quality tools will grow at a rate of 16.5% per year, rising from USD 1.8 billion in 2021 to USD 3.9 billion by 2026 [5]. Solutions that use AI are a big part of this growth:

- Automated Detection: Tools that AI powers can find and flag inconsistencies instantly. Machine learning systems are much better than human methods at finding patterns and outliers [5]. They can do this with an accuracy rate of up to 99%.
- Real-time Resolution: These tools show users possible copies so they can be resolved right away. Up to 60% less time is needed to clean up data when real-time quality checks are used [2].
- Integrating AI into Everyday Software: AI features are being added to everyday programs like Google Sheets, Excel, and Word. As an example, Microsoft's Excel Ideas tool uses AI to find and suggest fixes for data errors automatically, which could save users hours of work [2].
- Advanced Data Management Platforms: These AI-powered solutions are also being added to more advanced platforms like Databricks and Snowflake. Databricks says that their AI-powered data quality features have cut the time it takes to prepare data by up to 80% [5].

This change marks the start of a new wave of tools for improving data quality. Because of this, traditional data quality goods might become useless if they don't change by adding AI features. By 2025, 70% of companies will focus less on big data and more on small and wide data. This will give analytics more context and make AI less data hungry [2].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309086|

Metric	AI-Powered Solution
Data Quality Tools Market Size 2021 (USD Billion)	1.8
Data Quality Tools Market Size 2026 (USD Billion)	3.9
Pattern and Anomaly Detection Accuracy	99%
Time Spent on Data Cleansing	40%
Data Preparation Time	20%
Organizations Shifting to Small and Wide Data by 2025	70%

Table 1: Comparative Analysis: Traditional vs. AI-Powered Data Quality Management [2, 5]

IV. THE EVOLUTION OF DATA INTEGRATION

Traditional Integration Methods

Moving data from one system to another is usually part of data merging. Businesses that use services like Stripe often need to export data to their chosen storage solutions, such as Snowflake or Databricks. IDC did a survey and found that 67% of businesses use three or more databases for their analytics and data warehousing needs [1]. This shows how difficult it is to integrate data.

Because of this need, businesses like Fivetran have come up with pre-built interfaces that make the process of sending data easier. The world's data integration market was worth $11.6 billion in 2021, and it's expected to hit $19.6 billion by 2026, rising at a compound annual growth rate (CAGR) of 11.0% [7].

Direct Data Pipeline Services

As service providers start to offer their own data pipeline options, the landscape is changing:

- Stripe's Data Pipeline: Stripe now has a service that lets you send data directly to Snowflake. Stripe's 2023 annual report says that their enterprise users have used this service 200% more than the previous year [6].
- Benefits for Customers:
- Lower costs: When businesses use direct pipeline services [7], they can save up to 70% on the costs of integrating data [7].
- Getting rid of third-party connection tools can make IT less complicated and lower the costs of maintaining it.
- A pay-per-use approach for using data: Companies say that this plan cuts their data-related costs by an average of 30% [7].
- Trend in the industry: Salesforce has also said that they want to make their data available in Snowflake. O'Reilly Media did a study not long ago and found that 63% of companies are already using or plan to use direct data pipeline services from their SaaS providers within the next 18 months [8].

V. IMPLICATIONS FOR THE INDUSTRY

This shift towards direct data access from cloud-based companies is likely to have far-reaching effects:

1. Values of Connector Companies May Go Down: The market value of companies that make data connections may go down. Analysts think that the value of pure-play data link companies could drop by 20–30% over the next three to five years [7].
2. Shift in Focus for Businesses: Businesses can focus their resources on using data to gain insights and make decisions when it is easy to get. Deloitte did a study and found that companies that used direct data flow services could move an average of 35% of their budget for data management to projects that involved data analysis and AI [6].
3. Changes to Approaches to Data Integration: This is a big change in how companies handle integrating and using data in the cloud. By 2025, 80% of data integration will likely be done by AI, which will cut down on human work by 45% [7].
4. Focus on Data Quality: As businesses depend more on direct data pipelines, data quality becomes the most important thing. A poll by The Data Warehousing Institute (TDWI) found that 83% of businesses have problems with the quality of their data, which costs each company an average of $14.2 million a year [8]. This shows how important it is for modern data integration methods to have strong data quality measures.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 9, September 2024

|DOI: 10.15680/IJIRSET.2024.1309086|

5. Switch to Real-time Data Integration: The shift to direct data flows is making real-time data integration more important. 89% of organizations now say that real-time data is important for their business operations, according to the TDWI study [8]. This shows a big change from batch-oriented processes to continuous data flows.

Year	Global Data Integration Market Size (USD Billion)	Adoption of Direct Data Pipeline Services
2021	11.6	33%
2022	12.9	41%
2023	14.3	51%
2024	15.9	57%
2025	17.6	60%
2026	19.6	63%

Table 2: Evolution of Data Integration: Market Growth and Adoption of Direct Pipeline Services (2021-2026) [1, 7, 8]

VI. CONCLUSION

The combination of direct data pipeline services and AI-driven data quality management is a big change in how companies handle data in the digital age. This change is not just a small step; it's a complete rethinking of how data is processed. A lot of resources will likely be moved from routine data handling to data analysis and AI projects as more businesses adopt these technologies. The values of traditional connector companies may go down, which will make data quality and real-time integration even more important. As time goes on, businesses will need to be able to use these new technologies well in order to stay ahead of the competition and come up with new ideas in a world that is becoming more and more data-driven.

REFERENCES

[1] IDC, "Worldwide Big Data and Analytics Software Forecast, 2023–2027," IDC.com, May 2021. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=US50117823&pageType=PRINTFRIENDLY
[2] Gartner, "Gartner Top 10 Data and Analytics Trends for 2021," Gartner.com, Feb. 2021. [Online]. Available: https://www.gartner.com/en/podcasts/thinkcast/gartner-top-data-and-analytics-trends-for-2021
[3] Experian Data Quality, "2020 Global Data Management Research," Experian.com, Feb. 2020. [Online]. Available: https://www.edq.com/resources/data-management-whitepapers/2020-global-data-management-research/#:~:text=Our%202020%20Global%20data%20management,data%20literacy%20across%20the%20organization
[4] IBM, "The Four V's of Big Data," IBM Big Data & Analytics Hub, 2021. [Online]. Available: https://opensistemas.com/en/the-four-vs-of-big-data/
[5] MarketsandMarkets, "Data Quality Tools Market," MarketsandMarkets.com, Jul. 2021. [Online]. Available: https://www.marketsandmarkets.com/report-search-page.asp?rpt=data-quality-tools-market
[6] Deloitte, "Tech Trends 2024: The Impact of Emerging Technologies," Deloitte.com, Jan. 2024. [Online]. Available: https://www2.deloitte.com/us/en/insights/focus/tech-trends.html
[7] MarketsandMarkets, "Data Integration Market," MarketsandMarkets.com, Aug. 2021. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/data-integration-market-61793560.html
[8] P. Russom, "Taking Data Quality to the Enterprise through Data Governance," TDWI, Mar. 2011. [Online]. Available: https://tdwi.org/articles/2006/05/09/taking-data-quality-to-the-enterprise-through-data-governance-report-excerpt.aspx

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details