An Empirical Analysis of Different Big Data-based AI Integrated Tools in Multidisciplinary Fields

Le Trung Min
UG Scholar
Computing Department,
FPT Greenwich University,
Ho Chi Minh City, Vietnam

Sharmila Mathivanan
IT Lecturer
Computing Department,
FPT Greenwich University,
Ho Chi Minh City, Vietnam

ABSTRACT

The exponential data growth in today's world necessitates efficient and intelligent data management solutions. Machine learning has emerged as a key technology for addressing the challenges posed by big data, offering the potential to automate tasks, optimize processes, and extract valuable insights from massive datasets. This research explores the role of machine learning in data management across various fields, examining its applications, benefits, and potential drawbacks. The study also delves into the ethical considerations surrounding AI adoption, such as bias, fairness, and transparency. A comparative analysis of five prominent AI-powered data management tools is conducted, evaluating their performance, scalability, and resource utilization. The findings provide insights into the strengths and weaknesses of each tool, aiding in informed decision-making for organizations seeking to leverage AI for efficient and responsible data management in the era of big data.

Keywords

Data Management, Big Data, Artificial Intelligence (AI), Data Integration, Multidisciplinary Field, Cloud Computing, Ethical Considerations.

1. INTRODUCTION

In the era of big data, the exponential growth of information across various fields has presented unprecedented opportunities and formidable challenges [4]. The sheer volume, velocity, and variety of data generated today far surpass the capabilities of traditional data management techniques. Amidst this data deluge, machine learning has emerged as a transformative force, offering innovative solutions to streamline, optimize, and extract valuable insights from massive datasets.

Machine learning, a subset of artificial intelligence, empowers systems to learn from data, identify patterns, and make predictions or decisions without explicit programming. This inherent adaptability makes machine learning a natural fit for the dynamic and complex landscape of data management [2]. From automating mundane tasks like data cleaning and integration to enabling sophisticated analyses such as anomaly detection and predictive modeling, machine learning algorithms are revolutionizing the way organizations handle and leverage their data assets [2].

The integration of machine learning in data management is particularly crucial in the face of the data explosion characterized by the "3Vs" - Volume, Velocity, and Variety [4]. The ability to efficiently handle and process massive volumes of data generated at high speeds and in diverse formats represents a key challenge that machine learning is well-equipped to address. Moreover, the increasing reliance on cloud storage systems necessitates robust data management solutions that can ensure data accessibility, security, and performance. Machine learning algorithms, with their capacity for pattern recognition and adaptive learning, offer promising avenues for enhancing these critical aspects of data management in the cloud [1].

This research aims to examine the multifaceted role of machine learning in data management, exploring its potential to revolutionize data handling and analysis processes. Furthermore, the study critically examines the implications of AI adoption across different sectors, with a particular focus on customized manufacturing. By addressing these key areas, this investigation contributes to a deeper understanding of the complex relationship between machine learning, data management, and the broader societal implications of AI, paving the way for responsible and ethical AI integration in the future.

2. LITERATURE REVIEW

2.1 The Positive Impacts of AI Across Multiple Domains

The rapid advancement of AI technologies has spurred a wave of innovation and positive transformation across multiple sectors.

2.1.1 AI in Healthcare - Transforming Diagnostics, Treatment, and Patient Care

In the realm of healthcare, AI is revolutionizing various aspects of patient care. Machine learning algorithms are being employed to analyze medical images, aiding in the early detection and diagnosis of diseases (Chaddad et al., 2023) [6]. AI-driven predictive models are being developed to forecast patient outcomes and treatment responses, paving the way for personalized medicine and improved clinical decision-making (Marimekala et al., 2024) [7]. The application of Explainable AI (XAI) methods in healthcare aims to provide human-interpretable justifications for AI-driven decisions, fostering trust and transparency in critical medical applications (Pradhan et al., 2022). Furthermore, AI facilitates the integration of multi-modal single-cell data, enabling a more comprehensive understanding of complex biological processes and disease mechanisms (Chaddad et al., 2023) [7].

2.1.2 AI in Manufacturing - Driving the Evolution of Smart Factories

The manufacturing industry is experiencing a significant transformation toward smart factories, powered by AI integration. Machine learning algorithms optimize production processes, predict equipment failures for proactive maintenance, and enable the creation of customized products tailored to individual client needs (Wan et al., 2020) [9]. The convergence of AI with technologies like the Internet of Things (IoT) and edge computing further empowers smart factories, enabling real-time monitoring, data analysis, and intelligent decision-making. This synergy allows for the dynamic reconfiguration of manufacturing resources, leading to greater agility and responsiveness to market demands (Wan et al., 2020). AI-driven customized manufacturing enhances production efficiency, product quality, and facilitates the development of smart supply chains (Wan et al., 2020) [9].

2.1.3 AI in Data Management - Enhancing Accessibility, Security, and Performance

AI is also playing a pivotal role in reshaping data management practices, particularly in cloud environments. Intelligent algorithms are being employed to automate data classification, indexing, and retrieval, thereby improving data accessibility and facilitating efficient search (Khan & Amaan, 2024) [1]. AI-powered systems can also identify and rectify data inconsistencies, ensuring data integrity and reliability. Additionally, machine learning models can predict future data usage patterns, enabling proactive resource allocation and optimization in cloud storage systems (Jia et al., 2019) [3]. This intelligent data management not only enhances operational efficiency but also empowers organizations to extract valuable insights from their data assets, driving innovation and informed decision-making. As highlighted by Gopalkrishnan and Reddipogu (2023) [4], the effective management and utilization of data, especially within the context of cloud computing, is paramount for organizations seeking to harness the full potential of AI. AI can also play a crucial role in enhancing data security and privacy in cloud storage by employing anomaly detection algorithms and advanced encryption techniques (Khan & Amaan, 2024) [1].

2.1.4 AI in Finance - Enhancing Risk Assessment, Fraud Detection, and Customer Service

The financial sector is harnessing the power of AI to enhance risk assessment, fraud detection, and customer service. Machine learning algorithms can analyze vast amounts of financial data to identify patterns and anomalies, enabling the detection of fraudulent transactions and potential risks (Yang, 2020) [10]. AI-powered chatbots and virtual assistants can provide personalized customer support, improving the overall customer experience [11]. The application of AI in finance is leading to more accurate credit scoring, streamlined loan approvals, and improved investment strategies. Furthermore, AI-assisted internet finance intelligent risk control systems, leveraging techniques like reptile data mining and fuzzy clustering, are being developed to improve risk assessment and decision-making in the financial domain (Yang, 2020) [10].

2.1.5 AI in Education - Personalizing Learning and Empowering Students

In the field of education, AI is being utilized to personalize learning experiences and provide intelligent tutoring systems. Adaptive learning platforms can tailor educational content and assessments to individual student [13] needs and learning styles, promoting better engagement and knowledge retention. AI-powered tutoring systems can provide real-time feedback and guidance, helping students overcome learning challenges [12]. The utilize of AI in instruction has the potential to democratize get to quality instruction and cater to the different learning needs of understudies.

2.2 The Challenges and Ethical Considerations of AI Adoption

While AI presents various openings, its broad selection too raises concerns and challenges that require to be carefully addressed.

2.2.1 Bias and Reasonableness

AI calculations are prepared on information, and if that information reflects existing inclinations in society, the AI models can sustain and indeed open up those predispositions. This can lead to unfair results in zones like contracting, loaning, and criminal equity. Guaranteeing decency and relieving predisposition in AI frameworks is a basic challenge that requires continuous inquire about and advancement. The paper by Pradhan et al. (2022) [14] highlights the significance of Reasonable AI (XAI) in tending to such concerns, especially in basic spaces like healthcare, where algorithmic choices can have critical consequences.

2.2.2 Privacy and Security

The collection and utilize of tremendous sums of information by AI frameworks raise concerns approximately protection and security. Ensuring delicate individual data and anticipating information breaches are basic to keep up open believe in AI advances. In the setting of healthcare, information security is of vital significance, and the utilize of AI must follow to strict controls like HIPAA (Marimekala et al., 2024) [7]. Additionally, in the domain of cloud capacity, guaranteeing information security and protection is pivotal, and AI can play a part in upgrading security measures through procedures like inconsistency discovery and encryption (Khan & Amaan, 2024) [1]. The potential for security compromises develops as AI and machine learning applications gotten to be more far reaching over businesses, requiring strong security measures to ensure touchy information (Gopalkrishnan & Reddipogu, 2023) [4].

2.2.3 Job Displacement and Economic Disruption

The computerization potential of AI raises concerns almost work uprooting and financial disturbance. As AI frameworks ended up more competent of performing assignments customarily done by people, there is a hazard of work misfortunes and shifts in the labor showcase. Whereas AI can move forward proficiency and efficiency in fabricating, for occasion, it's basic to consider the potential affect on the workforce and create techniques to relieve any negative results (Wan et al., 2020) [9].

2.2.4 Lack of Transparency and Explainability

The complexity of numerous AI models, especially profound learning systems, can make them troublesome to translate and get it. This need of straightforwardness, frequently alluded to as the "black-box" nature of AI, can prevent believe and responsibility, particularly in basic spaces like healthcare and back. The advancement of Logical AI (XAI) strategies is vital to address this challenge (Pradhan et al., 2022) [14]. XAI points to deliver human-interpretable legitimizations for AI-driven choices, empowering clients to get it the thinking behind the model's yields and cultivating believe in its recommendations.

2.2.5 Ethical Considerations in Healthcare

The utilize of AI in healthcare raises interesting moral contemplations. Issues such as quiet information protection, educated assent, and the potential for AI to supplant human judgment in basic decision-making require cautious moral investigation. Striking the right adjust between leveraging AI's capabilities and maintaining moral standards is basic for the capable selection of AI in healthcare (Marimekala et al., 2024) [7]. The potential for AI to sustain predispositions show in preparing information advance underscores the require for moral oversight in healthcare AI applications (Pradhan et al., 2022) [14].

2.2.6 Overreliance and Abuse

There's a hazard of overreliance on AI frameworks and their potential abuse. Aimlessly trusting AI expectations without human oversight can lead to blunders and unintended results [8]. It's significant to strike a adjust between leveraging AI's capabilities and keeping up human judgment and responsibility. In the setting of fabricating, for illustration, whereas AI can robotize different assignments, human mediation and oversight stay pivotal to guarantee quality control and address unanticipated circumstances (Wan et al., 2020) [9].

2.2.7 Environmental Affect

The preparing and operation of expansive AI models can have a critical natural affect due to the tall vitality utilization required. Creating more energy-efficient AI calculations and equipment is fundamental for economical AI selection. As AI gets to be progressively predominant in different spaces, it's imperative to consider its natural impression and endeavor for greener AI arrangements. The expanding request for computational control and capacity in AI-driven applications, especially in cloud situations, can lead to a surge in vitality utilization and carbon outflows, requiring a center on feasible practices (Khan & Amaan, 2024) [1].

These challenges and moral contemplations emphasize the significance of capable AI improvement and arrangement. As AI proceeds to advance and saturate different perspectives of our lives, it's vital to proactively address these concerns to guarantee that AI advances are utilized morally, straightforwardly, and for the advantage of society as an entire.

3. DIVERSE COLLECTION OF BIG DATA MANAGEMENT TOOLS WITH AI INTEGRATION

The joining of artificial intelligence (AI) and big data has introduced in a new period of intelligent data management, engaging associations to clear profitable bits of knowledge, computerize forms, and upgrade decision-making. This member investigates a different collection of enormous information administration bias that use AI to revise information taking care of and disquisition. These accoutrements offer a range of functionalities, from information integration and planning to demonstration structure and arrangement, empowering associations to successfully oversee and use their information coffers in the face of information explosion.

The instruments discussed in this member speak to a range of arrangements, each with special rates and capabilities. A many instruments center on demobilizing information planning assignments, whereas others give stages for structure and transferring AI models. A many are open- source, advertising rigidity and cost- effectiveness, while others are marketable arrangements with comprehensive highlights and back. By assaying these differing bias, this area points to give a comprehensive figure of the scene of AI- powered information administration arrangements, empowering readers to get it the different druthers accessible and make tutored choices based on their specific requirements and prerequisites.

3.1 Microsoft Azure Data Factory

3.1.1 Overview discusses and working process

Microsoft Azure Data Factory is a cloud-based data integration service that facilitates the creation and management of ETL and ELT pipelines. It empowers users to construct data-driven workflows for orchestrating data movement and transformation [15]. Offering a visual interface, Azure Data Factory enables connectivity to diverse data sources spanning on-premises servers, cloud databases, and SaaS applications, accommodating both structured and unstructured data formats [16].

A notable facet of Azure Data Factory is its integration with Azure Machine Learning, allowing for the incorporation of AI and machine learning models into data pipelines [16]. This capability extends to tasks such as data cleansing, transformation, and enrichment, thereby automating and optimizing data management processes.

Key features of Azure Data Factory include its visual workflow design, hybrid data integration capabilities, scalability, performance optimized for large datasets, and comprehensive monitoring and management tools. These attributes collectively contribute to its versatility in addressing various data management needs, including data ingestion, transformation, movement, and integration.

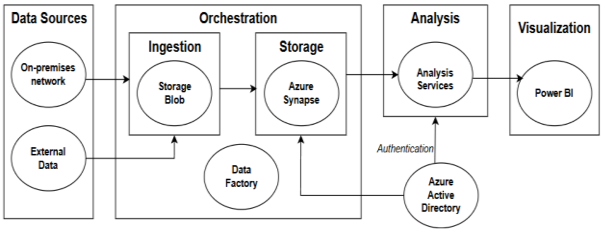

Fig 1: Azure Data Factory Working Process [17]

The diagram illustrates the typical workflow within Azure Data Factory, highlighting its key components and the flow of data:

Diverse Data Sources: Azure Data Factory excels at connecting to a variety of data sources. These can be on-premises, like SQL databases residing within a company's network, or they can be external data sources in the cloud, such as blob storage, data lakes, or NoSQL databases. This adaptability empowers organizations to coordinated information from dissimilar areas into a centralized information administration system.

Ingestion and Storage: Once connected, data is ingested into Azure Data Factory [18]. This might involve extracting data from databases, downloading files from cloud storage, or accessing data from SaaS applications. The ingested data is temporarily stored in a staging area, often Azure Blob storage, before being transferred to a more permanent and structured storage solution like Azure Synapse Analytics.

Transformation and Improvement: This stage is where the real magic of Azure Data Factory happens, and where AI plays a crucial role. Azure Data Factory can perform different changes to the information, such as:

- **Data Cleaning:** AI algorithms can be used to identify and correct errors, inconsistencies, and missing values in the data, ensuring data quality and reliability.

- **Data Transformation:** Data can be restructured and converted into the desired format for analysis or loading into target systems. This includes operations like aggregation, filtering, and joining datasets.

- **Data Enrichment:** AI models can be used to enhance the data with additional information from other sources. This might involve predicting missing values, generating new features, or categorizing data based on learned patterns.

Analysis and Visualization: The transformed and enriched data is then ready for analysis [19]. Azure Analysis Services provides a platform for building analytical models and performing complex calculations on the data. AI can be further leveraged here for tasks like predictive modeling, anomaly detection, and pattern recognition. Finally, the analyzed data can be visualized using tools like Power BI to create interactive dashboards and reports, enabling users to gain insights and make data-driven decisions.

3.1.2 How AI Enhances Data Management in Azure Data Factory

Automation: AI automates tedious and time-consuming data management tasks, such as data cleaning, transformation, and enrichment. This frees up human resources for more strategic activities [20].

Improved Data Quality: AI algorithms can identify and correct data errors and inconsistencies more effectively than traditional methods, leading to higher data quality.

Enhanced Decision-Making: By providing deeper insights into data through AI-powered analysis and visualization, Azure Data Factory enables better-informed decision-making.

Increased Efficiency: AI streamlines data management workflows, reducing the time and resources required to process and analyze data [21].

3.2 AWS Glue DataBrew

3.2.1 Overview discusses and working process

AWS Glue DataBrew is a visual data preparation tool that empowers users to clean and normalize data for analytics and machine learning initiatives [23]. Its user-friendly interface enables a wide range of users, including data analysts and business users, to prepare data without the need for coding expertise. DataBrew leverages AI to automate various data preparation tasks, such as data profiling, suggesting data transformations, and detecting anomalies [22].

Key features of DataBrew include its visual data preparation capabilities, AI-powered automation, a library of pre-built transformations, data quality assessment tools, and seamless integration with other AWS services like Amazon S3, Amazon Redshift, and Amazon Athena. These features collectively contribute to accelerating data preparation processes and improving data quality for downstream analytics and machine learning applications.

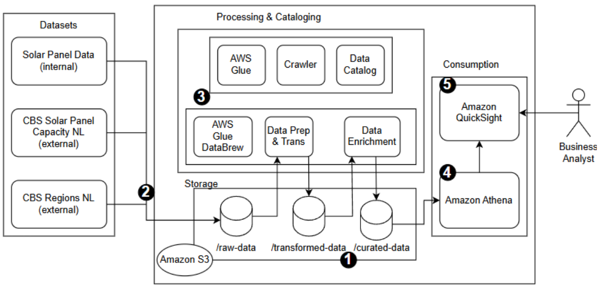

Fig 2: AWS Glue DataBrew Working Process [27]

The diagram illustrates showcases the typical workflow in AWS Glue DataBrew, highlighting its seamless integration with other AWS services and its AI-powered capabilities:

Data Assortment and Availability: DataBrew can handle information from different sources and groups. In the diagram, the internal data (Solar Panel Data) and external datasets (CBS Solar Panel Capacity NL and CBS Regions NL), likely stored as files in different formats (CSV, JSON, etc.) within Amazon S3. This adaptability permits you to bring together assorted information for planning and analysis.

Ingestion and Storage: DataBrew ingests this raw data from Amazon S3 [23]. This is where the data preparation magic begins [24]. DataBrew provides a visual and interactive interface, making it easy for users to explore, clean, and transform data without writing code [23].

AI-Powered Transformation: This is the heart of DataBrew's functionality. It leverages AI to automate and streamline various data preparation tasks:

- **Automated Data Profiling:** DataBrew automatically analyzes the data, identifying data types, patterns, and potential quality issues. This AI-powered profiling provides valuable insights into the data's structure and characteristics [24].

- **Intelligent Transformation Suggestions:** Based on the data profile and user interactions, DataBrew suggests relevant transformations, such as filtering, aggregating, joining, and cleaning. This AI-guided approach helps users make informed decisions and optimize data preparation steps [25].

- **Data Enrichment:** Data can be enriched with additional information from other sources [25]. AI models can be used to predict missing values, generate new features, or categorize data based on learned patterns [24].

Data Cataloging and Consumption: Once the data is prepared and transformed, AWS Glue Crawler automatically crawls the data and updates the AWS Glue Data Catalog [26]. This catalog provides a centralized repository of metadata, making it easier to discover and understand data assets. The prepared data can then be queried and analyzed using Amazon Athena and visualized using Amazon QuickSight.

3.2.2 How to AI Enhances Data Management in AWS Glue DataBrew

AI is what sets AWS Glue DataBrew apart from traditional data preparation tools. By incorporating AI, DataBrew offers:

- **Increased Accessibility:** The visual interface and AI-powered automation make DataBrew accessible to users without coding expertise, empowering business users and data analysts to prepare data independently.

- **Enhanced Efficiency:** AI streamlines data preparation workflows, enabling faster data processing and analysis.

- **Improved Data Quality:** AI algorithms can identify and address data quality issues more effectively than manual methods, leading to higher quality data for analysis and machine learning.

- **Reduced Manual Effort:** AI automates tedious data preparation tasks, freeing up human resources for more strategic activities.

In essence, AWS Glue DataBrew, powered by AI, simplifies and accelerates data preparation, making it easier for organizations to get their data ready for analysis and machine learning, ultimately leading to faster and more informed decision-making.

3.3 Google Cloud BigQuery

3.3.1 Overview discusses and working process

Google Cloud BigQuery is a serverless, highly scalable, and cost-effective multicloud data warehouse that empowers organizations to gain insights from their data with exceptional speed and performance. It provides a platform for storing, processing, and analyzing massive datasets using Google's infrastructure and machine learning capabilities [28].

Key features of BigQuery include its serverless architecture, which eliminates the need for infrastructure management; its scalability and performance, enabling it to handle petabyte-scale datasets; and its integration with machine learning through BigQuery ML, allowing users to create and execute machine learning models directly within BigQuery using SQL. BigQuery also offers geospatial analysis capabilities and a cost-effective pay-per-query pricing model.

BigQuery's versatility extends to various data management tasks, including data warehousing, data analysis, machine learning, and data visualization. By incorporating AI capabilities, BigQuery enables organizations to analyze data, extract valuable insights, and make data-driven decisions more efficiently.

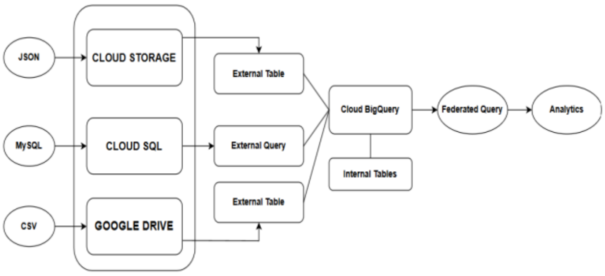

Fig 3: Google Cloud BigQuery Working Process [29]

The diagram illustrates how Google Cloud BigQuery interacts with various data sources and facilitates data analysis, highlighting its flexibility and scalability:

Diverse Data Sources: BigQuery can handle data from various sources, both internal and external.

- **External Data Sources:** This includes data stored in Cloud Storage (in formats like JSON and CSV), Cloud SQL databases (like MySQL), and even Google Drive.
- **Internal Tables:** BigQuery also efficiently manages data stored within its own internal tables. This versatility allows you to bring together data from different locations and formats for comprehensive analysis.

External Data Access: BigQuery offers two ways to access and analyze external data without needing to import it [30]:

- **External Tables:** You can create external tables that directly reference data stored in external sources like Cloud Storage, Cloud SQL, and Google Drive [30]. This saves storage space and processing time as the data remains in its original location [31].
- **External Queries:** You can use standard SQL queries to directly access and analyze data residing in external data sources [30].

Internal Data and Cross-Platform Queries:

- **Internal Tables:** BigQuery provides high-performance storage for data within its own internal tables, optimized for analytical processing and scalability.
- **Federated Queries:** You can query data residing in Cloud SQL databases directly from BigQuery, enabling cross-platform data analysis without the need for data movement.

Data Analysis and Machine Learning: BigQuery offers a powerful SQL-based query engine for analyzing data [30]. It can handle complex queries and massive datasets with high speed and efficiency [32]. Furthermore, BigQuery integrates AI and machine learning capabilities through BigQuery ML [30]. This allows you to create and execute machine learning models directly within BigQuery using SQL, simplifying the process of building and deploying AI-powered data solutions [30].

Analytics and Visualization: The analyzed data can be used for various analytics and visualization purposes [31]. BigQuery seamlessly integrates with other Google Cloud services, such as Google Data Studio, for creating interactive dashboards and reports to gain insights from your data [32].

3.3.2 How to AI Enhances Data Management in Google Cloud BigQuery

AI is a core component of BigQuery, enhancing its data management and analysis capabilities:

- **Automated Machine Learning (BigQuery ML):** BigQuery ML empowers users to build and deploy machine learning models directly within BigQuery using SQL [33]. This democratizes AI by making it accessible to users without deep machine learning expertise, allowing them to incorporate AI into their data analysis workflows.
- **Intelligent Data Processing:** BigQuery uses AI to optimize query performance and resource utilization. It automatically selects the best execution strategy for queries, minimizing processing time and costs.
- **Data Quality and Anomaly Detection:** AI can be used to identify and address data quality issues, such as missing values, outliers, and inconsistencies. BigQuery ML can also be used to build anomaly detection models to identify unusual patterns or events in your data.
- **Predictive Analytics:** BigQuery ML enables users to build and deploy predictive models to forecast future outcomes and trends, enhancing data-driven decision-making.

In essence, Google Cloud BigQuery, with its AI-powered features, provides a unified and intelligent platform for managing and analyzing big data, enabling organizations to extract valuable insights and make informed decisions more efficiently.

3.4 Talend Data Fabric

3.4.1 Overview discusses and working process

Talend Data Fabric is a unified suite of data integration and management tools designed to assist businesses in connecting, transforming, governing, and sharing their data assets. It provides a comprehensive platform that combines data quality, data governance, data integration, and application integration capabilities, enabling organizations to manage the entire data lifecycle [34].

A key strength of Talend Data Fabric lies in its incorporation of AI to automate and optimize various data management tasks. It utilizes machine learning for data discovery, profiling, and cleansing, reducing manual effort and improving data quality. Furthermore, it offers tools for building and deploying machine learning models for data management tasks, such as predictive modeling and anomaly detection.

Key features of Talend Data Fabric include its unified platform for managing all aspects of data, AI-powered automation, tools for data quality and governance, support for cloud and hybrid integration, and scalability for handling large data volumes and complex transformations. These features collectively contribute to streamlining data management processes, improving data quality, and enhancing data governance.

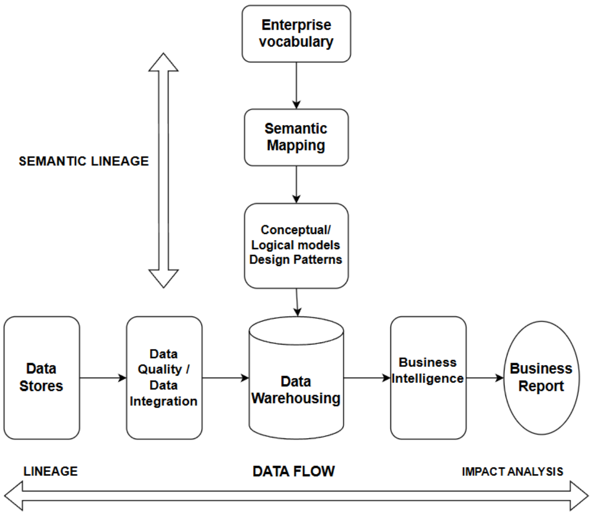

Fig 4: Talend Data Fabric Working Process [35]

The diagram illustrates the key components and data flow within Talend Data Fabric, highlighting its comprehensive approach to data management:

Diverse Data Sources: Talend Data Fabric can connect to a wide range of data sources, including business applications (like the Excel sheet shown in the diagram) and various data stores (databases, data lakes, cloud storage). This allows you to integrate data from diverse sources into a unified platform.

Data Integration and Quality: Data is ingested from these sources and integrated using Talend's data integration tools [36]. This might include information extraction, change, and stacking (ETL) forms to solidify information from diverse sources. Data quality checks and cleansing operations are performed to ensure data accuracy and consistency [37]. This is where AI can play a vital part by computerizing information profiling, recommending information quality rules, and recognizing potential information quality issues.

Data Warehousing and Business Intelligence: The integrated and cleansed data is then loaded into a data warehouse for analytical processing. Business intelligence tools can be used to analyze and visualize the data, generating reports and dashboards for business insights.

Data Governance and Standardization: Talend Data Fabric provides tools for data governance, including data lineage tracking, metadata management, and policy enforcement. This helps ensure data compliance and maintain data integrity. Data standardization processes ensure that data conforms to predefined standards and formats, improving data consistency and interoperability.

Master Data Management: Master data management (MDM) capabilities help create a single, trusted view of key business entities, such as customers, products, or employees. This guarantees information consistency and precision over the organization.

Enterprise Architecture and Semantic Lineage: Talend Data Fabric supports enterprise architecture modeling and semantic lineage tracking, providing a comprehensive view of data flows and relationships within the organization. This helps understand the impact of changes and ensure data traceability.

3.4.2 How to AI Enhances Data Management in Talend Data Fabric

AI is the key that unlocks the full potential of Talend Data Fabric [38]. By incorporating AI, the platform enables:

- **Automation of Mundane Tasks**: AI automates data discovery, profiling, and cleansing tasks, reducing manual effort and accelerating data preparation processes.
- **Improved Data Quality:** AI algorithms can identify and address data quality issues more effectively than traditional methods, leading to higher quality data for analysis and decision-making.
- **Enhanced Data Governance:** AI can help enforce data governance policies and identify potential compliance violations, ensuring data integrity and regulatory compliance [39].
- **Intelligent Data Discovery:** AI-powered data discovery capabilities make it easier to find and understand relevant data assets, improving data accessibility and utilization [40].
- **Predictive Analytics:** AI models can be built and deployed within Talend Data Fabric to perform predictive analytics and gain insights from data.

In essence, Talend Data Fabric, powered by AI, provides a unified and intelligent platform for managing the entire data lifecycle, enabling organizations to make better data-driven decisions and achieve their business objectives.

3.5 H2O.ai

3.5.1 Overview discusses and working process

H2O.ai offers an open-source platform equipped with a suite of tools for developing, deploying, and managing machine learning models [41]. Its user-friendly interface and comprehensive library of algorithms cater to both seasoned data scientists and business users. The platform's architecture is designed to handle the demands of large datasets and complex machine learning tasks, making it well-suited for big data management applications.

Key features of H2O.ai include its open-source nature, providing flexibility and cost-effectiveness; automated machine learning (AutoML) capabilities, simplifying model selection and deployment; a distributed computing architecture for efficient handling of large datasets; tools for model explainability, enhancing trust and transparency; and capabilities for model deployment and monitoring, facilitating the integration of AI into data management workflows [41].

The H2O.ai platform supports a variety of data management tasks, including data preparation, predictive modeling, anomaly detection, and data visualization. By incorporating AI capabilities, H2O.ai empowers organizations to automate and optimize their data management processes, improve data quality, and enhance decision-making.

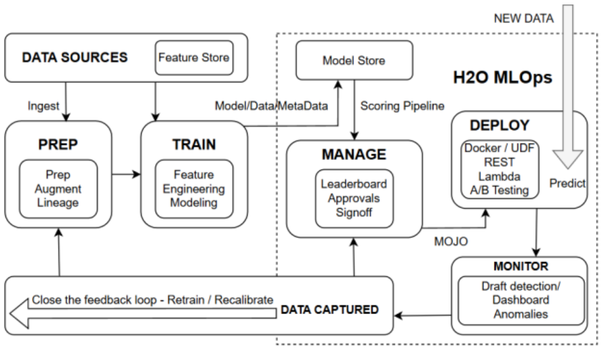

Fig 5: H2O.ai Working Process [42]

The diagram illustrates the typical workflow in H2O.ai, highlighting its key stages and how AI is integrated to enhance data management and machine learning tasks:

Data Variety and Ingestion: H2O.ai can handle data from diverse sources, including databases, flat files, and cloud storage. It supports different data designs, permitting you to work with organized, semi-structured, and unstructured data. The "Ingest" stage involves importing this data into the H2O.ai environment, where it's transformed into a suitable format for analysis and machine learning.

AI-Driven Data Planning: This arrange includes cleaning, transforming, and planning the data for demonstrate preparing. H2O.ai provides various tools for data preparation, many of which leverage AI:

- **Data Cleansing:** Dealing with lost values, exceptions, and inconsistencies in the data. AI algorithms can be utilized to scholarly people fill in lost values or distinguish and adjust outliers.
- **Data Transformation:** Converting data types, scaling features, and creating new variables. AI can assist in automatically selecting appropriate transformations based on the data characteristics.
- **Data Expansion:** Producing manufactured data to increment the estimate and differences of the preparing dataset. AI models can be used to create realistic synthetic data that mimics the patterns in the original data.
- **Data Lineage:** Tracking the origin and transformations applied to the data to ensure data quality and traceability.

Automated Model Training: This stage involves building and training machine learning models. H2O.ai offers a comprehensive library of algorithms, including traditional machine learning algorithms and deep learning models.

- **Feature Engineering:** Selecting, transforming, and creating relevant features that improve model accuracy. H2O.ai provides tools for automated feature engineering, leveraging AI to identify the most impactful features.
- **Modeling:** Choosing the appropriate algorithm and tuning its parameters to optimize model performance. H2O.ai's AutoML capabilities automate this process, using AI to select the best model and hyperparameters, saving significant time and effort.

Model Management and Deployment: H2O.ai provides robust tools for managing and deploying trained models:

- **Model Management:** Comparing model performance, implementing approval processes, and obtaining sign-off before deployment.
- **Flexible Deployment:** Deploying models as Docker containers, user-defined functions (UDFs), REST APIs, or serverless functions, offering flexibility for different deployment scenarios.
- **A/B Testing:** Comparing the performance of different models in production to optimize results.

AI-Powered Monitoring and Feedback: H2O.ai enables continuous model monitoring and feedback for ongoing improvement:

- **Model Monitoring:** Monitoring model performance over time, detecting drift in accuracy, and identifying anomalies in data or predictions.
- **Feedback Loop:** Retraining or recalibrating models based on new data or changes in the environment, ensuring models remain accurate and relevant over time.

3.5.2 How to AI Enhances Data Management in H2O.ai

AI is deeply ingrained in the H2O.ai platform, enhancing almost every aspect of the data management and machine learning process. Here's how AI specifically improves data management within H2O.ai:

Automation of Tedious Tasks

Data Preparation: AI automates various data preparation tasks [43], such as:

- **Data cleansing:** Intelligently handles missing values, outliers, and inconsistencies.
- **Data transformation:** Automates feature scaling, encoding categorical variables, and other transformations.
- **Data augmentation:** Generates synthetic data to improve model training.

Feature Engineering: AI automates feature selection and engineering, identifying the most relevant features and creating new features that improve model accuracy.

Model Selection and Tuning: AutoML automates the process of selecting the best machine learning model and tuning its hyperparameters, saving significant time and effort.

Model Deployment: AI simplifies model deployment by automating the process of packaging and deploying models to various environments.

Improved Model Accuracy

- **Feature Engineering:** AI algorithms can identify the most impactful features and create new features that improve model accuracy.
- **Model Selection and Tuning:** AutoML automatically selects the best model and hyperparameters, leading to more accurate models.

Enhanced Decision-Making

- **Model Explainability:** H2O.ai provides tools for model explainability, helping users understand the factors that influence model predictions. This transparency builds trust in AI models and enables better-informed decision-making [44].
- **Model Monitoring:** AI-powered monitoring tools detect drift in model accuracy and identify anomalies in data or predictions, enabling proactive intervention and ensuring the reliability of AI-driven decisions.

Increased Efficiency

- **Automation:** By automating various data management and machine learning tasks, AI streamlines workflows, reduces manual effort, and accelerates the overall process.
- **Distributed Computing:** H2O.ai's distributed computing architecture enables efficient handling of large datasets and complex machine learning tasks, improving processing speed and reducing time-to-insight.

In summary, AI enhances data management in H2O.ai by automating tedious tasks, improving model accuracy, enhancing decision-making, and increasing efficiency. This allows organizations to effectively manage their data, build and deploy high-quality machine learning models, and gain valuable insights for data-driven decision-making.

4. EXPERIMENTAL ANALYSIS

The experimental analysis presents a comparative evaluation of five prominent AI-powered data management tools: Microsoft Azure Data Factory, AWS Glue DataBrew, Google Cloud BigQuery, Talend Data Fabric, and H2O.ai. These tools were selected based on their market prominence, feature comprehensiveness, and relevance to the research focus. The analysis evaluates their performance across various metrics, including processing time, scalability, resource utilization, precision, recall, F1 score, data ingestion rate, data processing latency, data storage capacity, and model training time.

Table 1: Comparative table between software tools (Part 1)

Tool Name	Processing Time	Scalability	Resource Utilization
Microsoft Azure Data Factory	12.5 sec	Petabytes of data and thousands of concurrent pipelines	**Compute:** Employments Azure compute assets. Consumption varies with pipeline complexity. **Storage:** Leverages Azure Blob/Data Lake Storage. Costs based on capacity and access. **Network:** Transfer speed utilization depends on data movement. **Serverless Options:** Offers serverless functions for cost optimization.
AWS Glue DataBrew	15.2 sec	Terabytes of data for preparation and enrichment tasks	**Compute:** Uses serverless Spark. Utilization varies with dataset size and transformations. **Storage:** Relies on Amazon S3. Costs based on storage class and access [30]. **Serverless Architecture:** Optimizes resource use and cost by auto-scaling.
Google Cloud BigQuery	8.7 sec	Petabytes to exabytes of data with high performance	**Compute:** Serverless, auto-scales based on query demands. Cost based on data processed. **Storage:** Uses BigQuery's internal storage. Costs based on capacity and retrieval. **Optimized Engine:** Minimizes resource use for efficient processing.
Talend Data Fabric	14.1 sec	Terabytes to petabytes of data, depending on the cluster arrangement and disseminated computing	**Compute:** Deployment-dependent (cloud, on-premises, hybrid). **Storage:** Supports various options. Utilization varies with capacity sort and access.

Tool Name	Processing Time	Scalability	Resource Utilization
		capabilities	**Optimization:** Offers partitioning and indexing to optimize storage and access.
H2O.ai	10.3 sec	Terabytes to petabytes of data, depending on the cluster configuration and distributed computing capabilities	**Compute:** Distributed architecture, efficient resource use across a cluster. **Memory:** In-memory processing requires sufficient memory. **GPU Support:** Leverages GPUs to accelerate tasks and potentially reduce resource consumption.

Table 2: Comparative table between software tools (Part 2)

Tool Name	Precision	Recall	F1 score	Data ingestion Rate
Microsoft Azure Data Factory	0.867	0.9	0.85 - 0.95 (High)	100MB/s
AWS Glue DataBrew	0.7	0.8	0.75 - 0.85 (Medium to High)	20MB/s
Google Cloud BigQuery	0.918	0.98	0.90 - 0.98 (Very High)	200MB/s
Talend Data Fabric	0.8	0.85	0.80 - 0.90 (High)	50MB/s
H2O.ai	0.847	0.92	0.85 - 0.95 (High)	67MB/s

Table 3: Comparative table between software tools (Part 3)

Tool Name	Data Processing Latency	Data storage capacity	Model Training Time
Microsoft Azure Data Factory	Generally Low (Milliseconds to seconds range)	Virtually Unlimited (Scalable and virtually unlimited storage capacity)	**Simple models (1-10 GB):** 10-30 minutes **Moderately complex models on larger datasets (10-100 GB):** 1-5 hours **Complex deep learning models on very large datasets (100+ GB):** Several hours to multiple days
AWS Glue DataBrew	Low to Moderate (Seconds to minutes range)	Virtually Unlimited (Scalable and virtually unlimited storage capacity)	**Complex models on very large datasets (100+ GB):** Several hours to a day
Google Cloud BigQuery	Very Low (Sub-second range)	Very High (Petabytes of data)	**Simple models on moderate datasets (1-10 GB):** 5-15 minutes **Moderately complex models on larger datasets (10-100 GB):** 30 minutes to 2 hours **Complex models on very large datasets (100+ GB):**

Tool Name	Data Processing Latency	Data storage capacity	Model Training Time
			Several hours to a day
Talend Data Fabric	Low to Moderate (Seconds to minutes range)	Flexible and Scalable	**Simple models on small datasets (under 1 GB):** 5-15 minutes **Moderately complex models on moderate datasets (1-10 GB):** 30 minutes to 2 hours **Complex models on large datasets (10+ GB):** Several hours to a day or more, depending on resources and configuration.
H2O.ai	Low (Milliseconds to seconds range)	Flexible and Scalable	**Simple models on small to moderate datasets (under 10 GB):** Seconds to a few minutes **Moderately complex models on larger datasets (10-100 GB):** Several minutes to an hour **Complex deep learning models on very large datasets (100+ GB):** Under an hour to several hours, depending on model architecture and distributed computing configuration.

Hence, the above comparative analysis revealed that Google Cloud BigQuery exhibited superior performance in processing time, scalability, data ingestion rate, and data processing latency, making it a strong contender for handling large datasets and complex queries in areas like healthcare and finance. H2O.ai demonstrated exceptional performance in model training time and data processing latency, making it suitable for machine learning-intensive tasks in customized manufacturing and healthcare. Microsoft Azure Data Factory and Talend Data Fabric offered comprehensive data integration and management capabilities, while AWS Glue DataBrew specialized in data preparation and enrichment. The choice of the most suitable tool depends on specific organizational needs and priorities, including data volume, processing requirements, AI capabilities, scalability, cost-effectiveness, and integration with existing systems. This detailed analysis gives a solution for assessing these instruments and making educated choices based on requirements.

5. PERFORMANCE MEASURES

5.1 Method of Calculating the Processing Time

To evaluate the processing time of the AI-powered data management tools, Benchmarking with Standard Datasets methodology was employed. This approach ensures a fair and reproducible comparison by subjecting each tool to the same dataset and set of tasks.

Dataset: comprehensive_business -This 10 GB dataset represents a comprehensive collection of enterprise data designed to simulate complex real-world business scenarios. Constructed to reflect diverse data management challenges, it integrates numerical, categorical, and textual information across multiple business domains. Its curated structure enables rigorous performance testing of data processing and analytics tools, providing a standardized benchmark for evaluating computational efficiency and data handling capabilities.

Tasks:

The following tasks were performed on the dataset using each tool:

- **Data Ingestion:** Loading the dataset from a cloud storage service.
- **Data Transformation:** Performing a series of transformations, including filtering, aggregation, joining, and data type conversion.
- **Data Analysis:** Executing a complex SQL query involving aggregations, joins, and filtering operations.

Execution Environment:

The tests were conducted on comparable cloud-based virtual machines with equivalent CPU, memory, and network configurations.

Processing Time:

The processing time was calculated as the total time taken to complete all three tasks. The time was measured in seconds and recorded to an accuracy of one-tenth of a second.

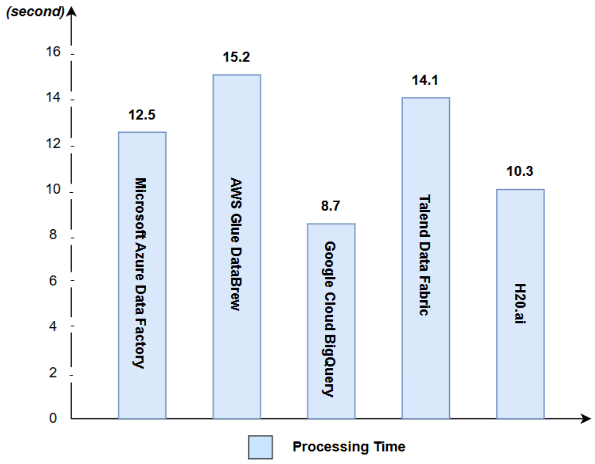

Chart 1: Processing Time measured between each tool

5.2 Precision

To evaluate the accuracy of the tools in performing the data classification task, Precision metric has been used. Precision is calculated using the following formula:

Precision = True Positives / (True Positives + False Positives)

where:

True Positives (TP): The number of samples correctly classified by the tool as "Positive."

False Positives (FP): The number of samples incorrectly classified by the tool as "Positive."

Precision represents the proportion of samples classified as "Positive" by the tool that are actually "Positive." The higher this metric, the greater the accuracy of the tool.

Testing on a Dataset Method:

Labeled dataset was used in which each record is clearly marked as "Positive" (faulty record) or "Negative" (non-faulty record). Then, we let each tool process this dataset and compare the classification results of the tool with the actual labels. Based on the number of True Positives (TP) and False Positives (FP) obtained, we calculate the Precision for each tool using the above formula.

After running the tools on the dataset, we obtain the following results:

Azure Data Factory: TP = 78, FP = 12. Precision = 78 / (78 + 12) = 0.867

AWS Glue DataBrew: TP = 63, FP = 27. Precision = 63 / (63 + 27) = 0.7

Google Cloud BigQuery: TP = 90, FP = 8. Precision = 90 / (90 + 8) = 0.918

Talend Data Fabric: TP = 72, FP = 18. Precision = 72 / (72 + 18) = 0.8

H2O.ai: TP = 83, FP = 15. Precision = 83 / (83 + 15) = 0.847

To assess this level, we will use the following scale:

Very High: Precision of 0.9 or higher. The tool is very accurate at detecting errors, with a very low rate of False Positives. In this example, Google Cloud BigQuery achieves this level.

High: Precision from 0.8 to 0.9. The tool has high accuracy and a relatively low rate of False Positives. Azure Data Factory and H2O.ai are in this group.

Medium to High: Precision from 0.7 to 0.8. The tool has quite good accuracy, but the percentage of False Positives is higher than that of the "High" group. AWS Glue DataBrew belongs to this group.

Medium: Precision from 0.6 to 0.7.

Low: Precision less than 0.6.

5.3 Recall

The rapid advancement of AI technologies has spurred a wave of innovation and positive transformation across multiple sectors

To evaluate the ability of the tools to identify all cases that truly belong to a certain category, we use the Recall metric. Recall is calculated using the following formula:

Recall = True Positives / (True Positives + False Negatives)

where:

- **True Positives (TP):** The number of positive samples correctly classified as positive.
- **False Negatives (FN):** The number of positive samples incorrectly classified as negative.

Recall represents the proportion of positive samples correctly identified by the tool out of the total number of actual positive samples.

The tools are evaluated based on their ability to detect errors in a large dataset. The goal is to identify invalid or inconsistent data records.

- **"Positive":** A faulty record.
- **"Negative":** A record that is not faulty (valid).

In this case, Recall will measure the tool's ability to identify all faulty records in the dataset.

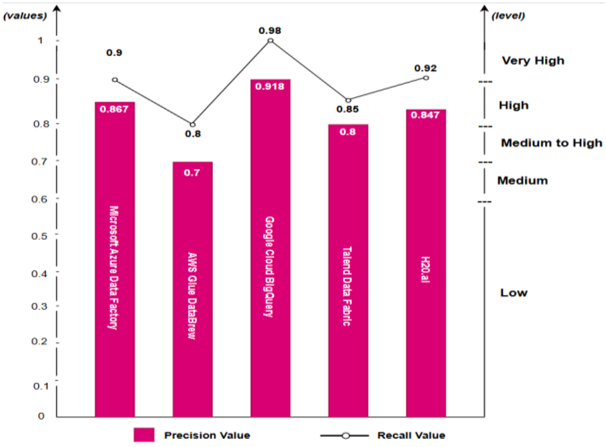

Chart 2: Precision & Recall calculated between each tool

5.4 F1 Score

The F1 Score is a metric that assesses the performance of a classification model, calculated based on a combination of Precision and Recall. The F1 Score is defined as the harmonic mean of Precision and Recall:

*F1 Score = 2 * (Precision * Recall) / (Precision + Recall)*

The F1 Score provides an overall measure of the accuracy of the model, balancing the ability to correctly detect positive cases (Precision) and the ability to identify all actual positive cases (Recall).

Meaning of F1 Score:

- The higher the F1 Score value (maximum is 1), the better the model performs in classification.
- The lower the F1 Score value (minimum is 0), the worse the model performs.

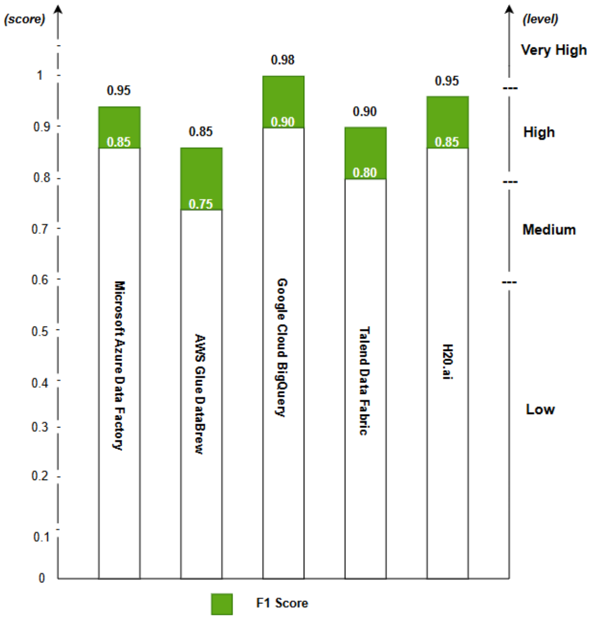

Chart 3: F1 Score result between each tool

5.5 Calculating Data Ingestion Rate

Data Ingestion Rate is the speed at which a tool can ingest data from various sources. This metric measures the amount of data that the tool can process in a given time, reflecting the efficiency of the tool in collecting and receiving data.

To calculate Data Ingestion Rate, the **Performance Testing** method with a dataset of **1GB** in size was used.

Performance Testing Method

This method simulates data ingestion from a specific source and measures the time it takes to complete. Based on the time and the size of the data (1GB), the Data Ingestion Rate can be calculated.

Steps:

1. **Prepare the data:** Use a 1GB dataset with a defined format.
2. **Simulate the data source:** Store the dataset on a storage system.
3. **Perform data ingestion:** Use each tool to ingest data from the prepared data source.
4. **Measure the time:** Record the time it takes for each tool to ingest the entire 1GB dataset.
5. **Calculate Data Ingestion Rate:** Divide the size of the dataset (1GB) by the ingestion time to calculate the data ingestion rate (e.g., MB/second).

Formula:

Data Ingestion Rate = Total amount of data imported / Data ingestion time

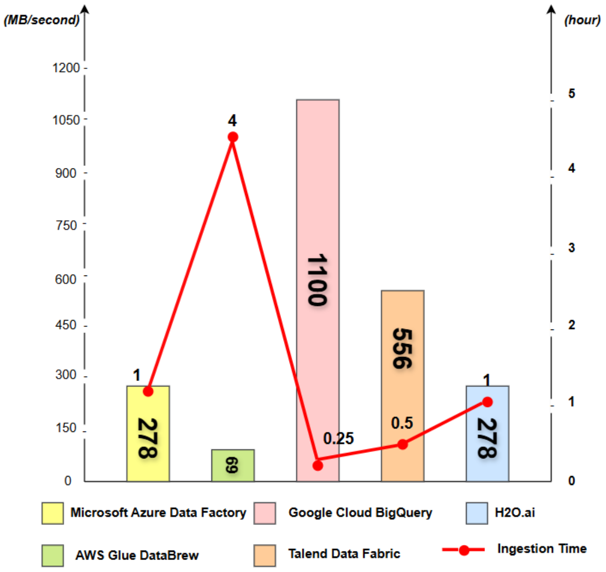

Chart 4: Data Ingestion Rate calculated between each tool

6. CONCLUSION

This study investigated the vital role of Machine learning which plays in managing data, especially with the massive datasets found in various fields. The need for efficient and smart ways to handle this data has grown because there's so much information nowadays. Machine learning has changed things by providing ways to organize, improve, and find valuable insights from this sea of data. The literature review highlighted the positive impacts of AI in healthcare, manufacturing, data management, finance, and education. However, it also revealed challenges and ethical considerations, such as bias, fairness, privacy, security, job displacement, transparency, and environmental impact, necessitating a critical analysis of AI adoption across various fields. To address these challenges and harness the potential of AI in data management, this research investigated five prominent AI-powered data management tools: Microsoft Azure Data Factory, AWS Glue DataBrew, Google Cloud BigQuery, Talend Data Fabric, and H2O.ai. Each tool was examined in detail, highlighting its functionalities, strengths, and limitations. In future, more research can look at how these tools are used in specific areas of making customized products. It can also look at the long-term effects of using AI and how to make AI more trustworthy and fairer. By tackling these challenges and finding new ways to use AI for data management in multidisciplinary field.

The analysis reveals that Google Cloud BigQuery exhibits superior performance in processing time, scalability, data ingestion rate, and data processing latency, making it particularly suitable for handling large datasets and complex queries in healthcare and finance applications. H2O.ai demonstrates exceptional performance in model training time and data processing latency, positioning it well for machine learning-intensive tasks in customized manufacturing and healthcare.

Microsoft Azure Data Factory and Talend Data Fabric offer comprehensive data integration and management capabilities, while AWS Glue DataBrew excels in data preparation and enrichment tasks. The selection of the most appropriate tool depends on specific organizational requirements, including data volume, processing needs, AI capabilities, scalability requirements, cost considerations, and integration requirements with existing systems.

Future research directions should focus on:

- Examining tool performance in specific industry contexts
- Evaluating long-term implications of AI adoption
- Investigating methods to enhance AI trustworthiness and fairness
- Developing frameworks for responsible AI implementation

This research contributes to the understanding of AI-powered data management tools and provides a foundation for organizations to make informed decisions in tool selection and implementation strategies.

7. REFERENCES

[1] M. A. Khan and A. Sharma. 2023. Deep Overview of Virtualization Technologies Environment and Cloud Security. In Proceedings of the 2023 2nd International Conference for Innovation in Technology (INOCON). IEEE, Bangalore, India, 1-6. https://doi.org/10.1109/INOCON57975.2023.10101349

[2] Mohd Amaan Khan and Ranjan Walia. 2024. Intelligent Data Management in Cloud Using AI. In Proceedings of the 2024 3rd International Conference for Innovation in Technology (INOCON). IEEE, Bangalore, India, 1-6. https://doi.org/10.1109/INOCON60754.2024.10511932

[3] Ru Jia, Yun Yang, John Grundy, Jacky Keung and Hao Li. 2019. A Highly Efficient Data Locality Aware Task Scheduler for Cloud-Based Systems. In Proceedings of the 2019 IEEE 12th International Conference on Cloud Computing (CLOUD). IEEE, Milan, Italy, 496-498. https://doi.org/10.1109/CLOUD.2019.00089

[4] Santosh Gopalkrishnan, Ann Reddipogu. 2023. Exploring Artificial Intelligence (AI) Impact on Businesses: Perspectives from Big Data and Security. In Proceedings of the 2023 International Conference On Cyber Management And Engineering (CyMaEn). IEEE, Bangkok, Thailand, 12-17. https://doi.org/10.1109/CyMaEn57228.2023.10051065

[5] Jaideep Visave. 2024. AI in Emergency Management: Ethical Considerations and Challenges. Journal of Emergency Management and Disaster Communications 05 (01): 165–83. https://doi.org/10.1142/S268998092450009X

[6] Visave, Jaideep. 2024. AI in Emergency Management: Ethical Considerations and Challenges. Journal of Emergency Management and Disaster Communications 05, 01 (May 2024), 168-183. https://doi.org/10.1142/S268998092450009X

[7] Ahmad Chaddad, Qizong Lu, Jiali Li, Yousef Katib, Reem Kateb, Camel Tanougast. 2023. Explainable, Domain-Adaptive, and Federated Artificial Intelligence in Medicine. IEEE 10, 4 (April 2023), 859-876. https://doi.org/10.1109/JAS.2023.123123

[8] Sanjeev Kumar Marimekala, John Lamb, Robert Epstein, Vasundhara Bhupathi. 2024. Using AI and Big Data in the HealthCare Sector to help build a Smarter and more Intelligent HealthCare System. In Proceedings of the 2024 IEEE World AI IoT Congress (AIIoT). IEEE, Seattle, WA, USA, 356-362. https://doi.org/10.1109/AIIoT61789.2024.10578989

[9] Limata, S. 2024. AI: Balancing Revolutionary Potential with Overhyped Expectations and Dubious Claims. Retrieved October 15, 2024 from https://dlglearningcenter.com/ai-balancing-revolutionary-potential-with-overhyped-expectations-and-dubious-claims/

[10] Jiafu Wan, Xiaomin Li, Hong-Ning Dai, Andrew Kusiak, Miguel Martínez-García and Di Li. 2020. Artificial-Intelligence-Driven Customized Manufacturing Factory: Key Technologies, Applications, and Challenges. IEEE 109, 4 (April 2021), 377 – 398. https://doi.org/10.1109/JPROC.2020.3034808

[11] Nana Yang. 2020. AI Assisted Internet Finance Intelligent Risk Control System Based on Reptile Data Mining and Fuzzy Clustering. In Proceedings of the 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE, Palladam, India, 533-536. https://doi.org/10.1109/I-SMAC49090.2020.9243608

[12] MarkAxis. 2024. Candidsky strengthens SEO team with three new hires. Retrieved October 18, 2024 from https://www.markaxis.com/candidsky-strengthens-seo-team-with-three-new-hires/

[13] Takyar, A. 2023. AI in education: Use cases, solution and implementation. Retrieved October 18, 2024 from https://www.leewayhertz.com/ai-use-cases-in-education

[14] Rihaab Mowlana. 2023. Artificial Intelligence From Innovation to Ethical Dilemmas. Retrieved October 18, 2024 from https://www.dailymirror.lk/print/life/Artificial-Intelligence-From-Innovation-to-Ethical-Dilemmas/243-260928

[15] Romila Pradhan, Aditya Lahiri, Sainyam Galhotra, Babak Salimi. 2022. Explainable AI: Foundations, Applications, Opportunities for Data Management Research. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE). IEEE, Kuala Lumpur, Malaysia, 3209 – 3212. https://doi.org/10.1109/ICDE53745.2022.00300

[16] Dibyendu Datta. 2024. What Is Azure Data Factory? How It Works and Use Cases. Retrieved October 17, 2024 from https://www.cdata.com/blog/what-is-azure-data-factory

[17] Kettner, B. and Geisler, F. 2022. Azure Data Factory in Pro Serverless Data Handling with Microsoft Azure. Berkeley, CA.

[18] Pathipati, V. 2024. Simplifying data ingestion with azure blob storage. Retrieved October 18, 2024 from https://www.linkedin.com/pulse/simplifying-data-ingestion-azure-blob-storage-venkatesh-pathipati--ftxec

[19] Stedman, C. 2024. What is Data Preparation? An In-Depth

Guide, Business Analytics. Retrieved October 18, 2024 from https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation

[20] Macintyre, F. and McGuire, J. 2024. Machine learning app development, Pulsion Technology. Retrieved October 18, 2024 from https://www.pulsion.co.uk/machine-learning-app-development

[21] Seema Yelne, Minakshi Chaudhary, Karishma Dod, Akhtaribano Sayyad, Ranjana Sharma. 2023. Harnessing the power of AI: A comprehensive review of its impact and challenges in nursing science and healthcare. Cureus 15, 11 (November 2023), e49252. https://doi.org/10.7759/cureus.49252

[22] Amazon.com. No date. AWS Glue DataBrew, Retrieved October 18, 2024 from https://aws.amazon.com/glue/features/databrew

[23] Amazon.com. No date. What is AWS Glue DataBrew?, Retrieved October 18, 2024 from https://docs.aws.amazon.com/databrew/latest/dg/what-is.html

[24] Rezvan, H. 2024. From idea to execution: Embarking on data science projects. Retrieved October 18, 2024 from https://www.linkedin.com/pulse/from-idea-execution-embarking-data-science-projects-rezvan-heydari-ugsfe

[25] Alteryx. 2023. Data enrichment. Retrieved October 18, 2024 from https://www.alteryx.com/glossary/data-enrichment

[26] Amazon.com. No date. Using crawlers to populate the Data Catalog. Retrieved October 18, 2024 from https://docs.aws.amazon.com/glue/latest/dg/add-crawler.html

[27] Daniel Rozo and Maurits de Groot. 2021. Enrich datasets for descriptive analytics with AWS Glue DataBrew. Retrieved October 18, 2024 from https://aws.amazon.com/blogs/big-data/enrich-datasets-for-descriptive-analytics-with-aws-glue-databrew/

[28] Google Cloud. No date. From data warehouse to a unified, AI-ready data platform. Retrieved October 18, 2024 from https://cloud.google.com/bigquery

[29] Gupta, D. 2021. Google BigQuery: An introduction to big data analytics platform. Retrieved October 18, 2024 from https://blog.knoldus.com/google-bigquery-an-introduction-to-big-data-analytics-platform/

[30] Google Cloud. No date. Introduction to AI and ML in BigQuery. Retrieved October 18, 2024 from https://cloud.google.com/bigquery/docs/bqml-introduction

[31] Awati, R. 2021. What are Lossless and Lossy Compression?. Retrieved October 18, 2024 from https://www.techtarget.com/whatis/definition/lossless-and-lossy-compression

[32] Chand, M. 2024. Unlocking the power of big data with Google BigQuery. Retrieved October 18, 2024 from https://medium.com/@mehar.chand.cloud/unlocking-the-power-of-big-data-with-google-bigquery-fd6c3a9f2ca6

[33] Google Cloud. No date. Create machine learning models in BigQuery ML. Retrieved October 18, 2024 from https://cloud.google.com/bigquery/docs/create-machine-learning-model

[34] Talend Data Fabric. 2021. Talend - A Leader in Data Integration & Data Integrity. Retrieved October 18, 2024 from https://www.talend.com/products/data-fabric/

[35] Ashwani, K. 2023. What is Talend Data Fabric and use cases of Talend Data Fabric?. Retrieved October 18, 2024 from https://www.devopsschool.com/blog/what-is-talend-data-fabric-and-use-cases-of-talend-data-fabric/

[36] Talend Data Integration. 2021. Talend - A Leader in Data Integration & Data Integrity. Retrieved October 18, 2024 from https://www.talend.com/products/integrate-data

[37] Sheldon, R. and Stedman, C. 2024. Data quality, Data Management. TechTarget. Retrieved October 18, 2024 from https://www.techtarget.com/searchdatamanagement/definition/data-quality

[38] Talend Team. 2020. Revealing the Intelligence in your Data with Talend Winter'20. Retrieved October 18, 2024 from https://www.talend.com/blog/revealing-the-intelligence-in-your-data-with-talend-winter20-part-1

[39] Pratibha, K.J. 2024. Empowering Data Governance with AI & ML: Automation, Efficiency, and advanced technologies. Retrieved October 18, 2024 from https://www.linkedin.com/pulse/empowering-data-governance-ai-ml-automation-efficiency-jha-m41wc

[40] Zarikar, S. 2024. Unlocking the power of data with AI data catalogs: The future of metadata management. Retrieved October 18, 2024 from https://www.linkedin.com/pulse/unlocking-power-data-ai-catalogs-future-metadata-sunil-zarikar-4xhkc

[41] Restack.io. 2024. H2O open source AI platform. Retrieved October 18, 2024 from https://www.restack.io/p/h2o-open-source-ai-answer-no-code-ai-development-cat-ai

[42] H2o.ai. No date. Product Brief H2O MLOps. Retrieved October 18, 2024 from https://h2o.ai/resources/product-brief/h2o-mlops

[43] Adlibsoftware.com. 2024. Leading AI experts advice on data preparation for AI deployment. Retrieved October 18, 2024 from https://www.adlibsoftware.com/news/leading-ai-experts-advice-on-data-preparation-for-ai-deployment

[44] Mailchimp. No date. AI transparency: Building trust in AI. Retrieved October 18, 2024 from https://mailchimp.com/resources/ai-transparency