Title: 2023 State of Data + AI
Author: Databricks
Publication Year: 2023

Introduction:

Background

In recent years, the data analytics landscape has rapidly evolved with innovations like cloud computing, big data, open source frameworks, and software-as-a-service tools. These advances have enabled access to vastly greater data resources for all types of organizations and entities. However, many still rely on legacy data warehouses and pipelines that create information silos, lack agility, and cannot harness unstructured data at scale for cutting-edge AI/ML applications. As a result, taking advantage of modern innovations in artificial intelligence remains challenging for most organizations.

Current Status

While many data leaders recognize the urgent need to modernize their technology stacks and develop new skills to enable smarter applications of AI and machine learning, they struggle to create holistic strategies and align updated processes, workforce capabilities, and tools. Although advances like low/no-code machine learning, natural language processing, and large language models have further accelerated what is possible with AI to drive competitive advantage, organizations across sectors face a widening capability gap between this potential and their ability to actually build, deploy, and manage models at scale.

Purpose of Publication

This annual report by Databricks aims to uncover global data and AI adoption trends based on usage patterns from its extensive customer ecosystem. It seeks to help data leaders assess their strategy against real-world benchmarks and inform their roadmap.

Methods:

Usage data was aggregated from over 9,000 Databricks customers on the lakehouse platform, spanning industries and company sizes. Usage was measured by number of customers adopting products. Growth trends over the past year were analyzed across machine learning applications, product categories, data formats, and migrations.

Key Findings/Highlights:

1. Rapid Growth in SaaS LLM APIs Usage:
   Companies' usage of Large Language Model APIs, like ChatGPT, surged by 1310% from late 2022 to mid-2023.
2. Dominance of NLP:
   Nearly half (49%) of Python data science activities focus on Natural Language Processing, marking its top position in application usage.
3. Elevated Model Development:
   A 411% yearly rise in models entering production and a 54% increase in ML experimentation highlights dynamic AI advancements.
4. Shift to Advanced Data Use Cases:

While Business Intelligence remains dominant, companies are exploring deeper, advanced data scenarios.

5. Popularity of dbt and Data Integration:
   The data tool 'dbt' shines with a 206% annual growth, with data integration on Databricks Lakehouse growing by 117%.
6. Lakehouse Attraction:
   A majority (61%) of migrations to Lakehouse are from traditional and cloud data storage solutions.
7. Diverse Applications of DS/ML:
   With the rise of Large Language Models, businesses across sectors harness data science for growth, insights, and improved experiences.
8. Diverse Python Tools:
   Specialized Python libraries cater to a wide array of DS/ML needs, from speech recognition to time series analysis.
9. Significance of Simulations and Optimization:
   30% of DS/ML applications are geared towards simulations and optimizations, revealing a trend towards data-driven problem-solving.

Challenges and Trends:

Trends:

1. Emergence of AI's Potential: The launch of ChatGPT marked a significant realization of AI's capabilities, drawing attention to its vast potential in reshaping industries and strategies.
2. Proliferation of LLMs: A dramatic 1310% growth in the adoption of SaaS LLM APIs indicates the transformative impact and value businesses see in LLMs like ChatGPT.
3. High Demand for NLP and LLMs: NLP is not only growing rapidly but also paving the way for tasks that were previously challenging for traditional code, such as content summarization and sentiment extraction. LLMs, falling under this category, are expected to further revolutionize applications like chatbots, research assistance, and content generation.
4. Data Migration Trends: Organizations are increasingly migrating to new architectures, driven by the limitations and scalability challenges of legacy data platforms.
5. Rise of Data Warehousing on Lakehouse: There's a marked increase in data warehousing on the Lakehouse Platform, especially with the use of Databricks SQL, indicating a shift towards unified data management solutions.

Challenges:

1. Managing Rapid AI Advancements: The unparalleled pace of AI discoveries implies that organizations must constantly adapt, ensuring their strategies and tools align with the latest innovations.
2. Ethical and Practical Challenges with LLMs: LLMs can generate vast amounts of human-like text, raising concerns about authenticity, potential misinformation, and the risk of misuse in various domains.

3. Data Migration Complexities: Migrating to new data architectures can be risky, expensive, and time-consuming. Ensuring smooth transitions without data loss or integrity issues is crucial.
4. Handling Diverse Data Formats: The increasing variety of data formats, especially with the growth of semi-structured and unstructured data, presents challenges in data integration, management, and analysis.
5. Harnessing Advanced ML and AI Use Cases: As companies delve deeper into advanced ML and AI applications, there's a need to ensure that these tools and models are effectively harnessed to derive tangible value..

Implications:

- AI-Driven Transformation:
  Post-ChatGPT's launch, businesses are poised for an AI-led transformation, offering competitive advantages to early adopters.
- Language Models' Impact:
  The surge in LLM adoption signals a shift towards human-centric AI applications, enriching user experiences across sectors.
- Data-Centric Decisions:
  NLP's dominance and data integration trends highlight a move towards data-informed strategies and operations.
- Unified Data Solutions:
  Migration trends to platforms like Lakehouse suggest a future with streamlined data management, enhancing operational efficiency.
- Evolving Skillsets:
  Rapid AI advancements necessitate skillset evolution, pointing to potential shifts in tech job roles and training.
- Ethical AI Concerns:
  LLMs' capabilities might trigger increased ethical scrutiny, potentially leading to new regulatory guidelines for responsible AI use.
- AI as a Strategic Pillar:
  The emphasis on AI in organizational strategies indicates its rising importance as a key investment area.

Future Directions:
Data leaders need roadmaps to transform architectures, skills, and processes to capitalize on AI innovations and accelerate time-to-insight.

Conclusion:
The report shows sophisticated data and AI strategies taking hold across companies as they adopt innovations like lakehouses, data integration tools, and NLP/LLMs that will drive the next generation of data-fueled competition.