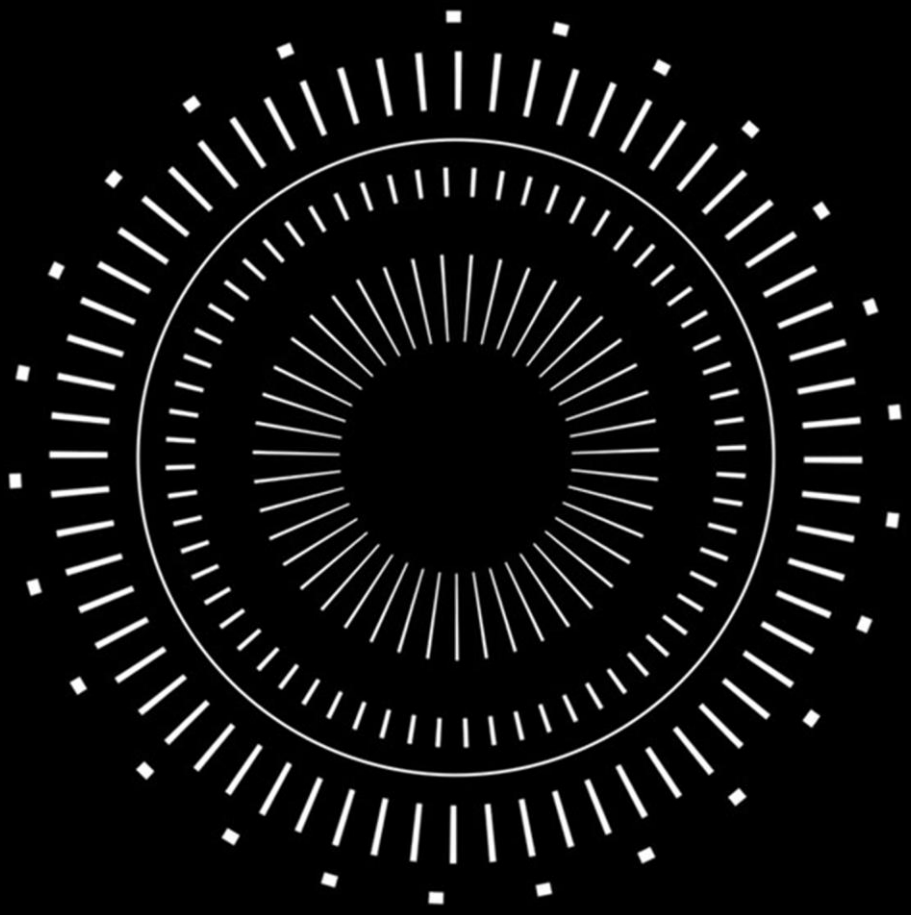
Deloitte.



Navigate data management challenges to enable AI initiatives

Smart data management is the foundation of organization-wide usage of Artificial Intelligence.

Leading organizations are able to fully leverage the power of Artificial Intelligence and generate value by enabling data professionals to have access to well-organized high quality data from across the entire organization. But how can this be achieved?

By Naser Bakhshi, Amisha Khera & Alessio Bilato

The Deloitte AI Loop (DAIL)

The Deloitte AI Loop provides a framework that mimics the human approach in the space of artificial intelligence. Based on our experience in bringing cognitive solutions to our clients, we have lined out DAIL as a blueprint for all aspects that should be covered in a successful AI solution, as we explained in the introductory blog.
This is the second article of the DAIL series, focusing on the SENSE component, consisting of tools, technology and infrastructure to measure, capture and monitor data from business processes, behavior and the environment.

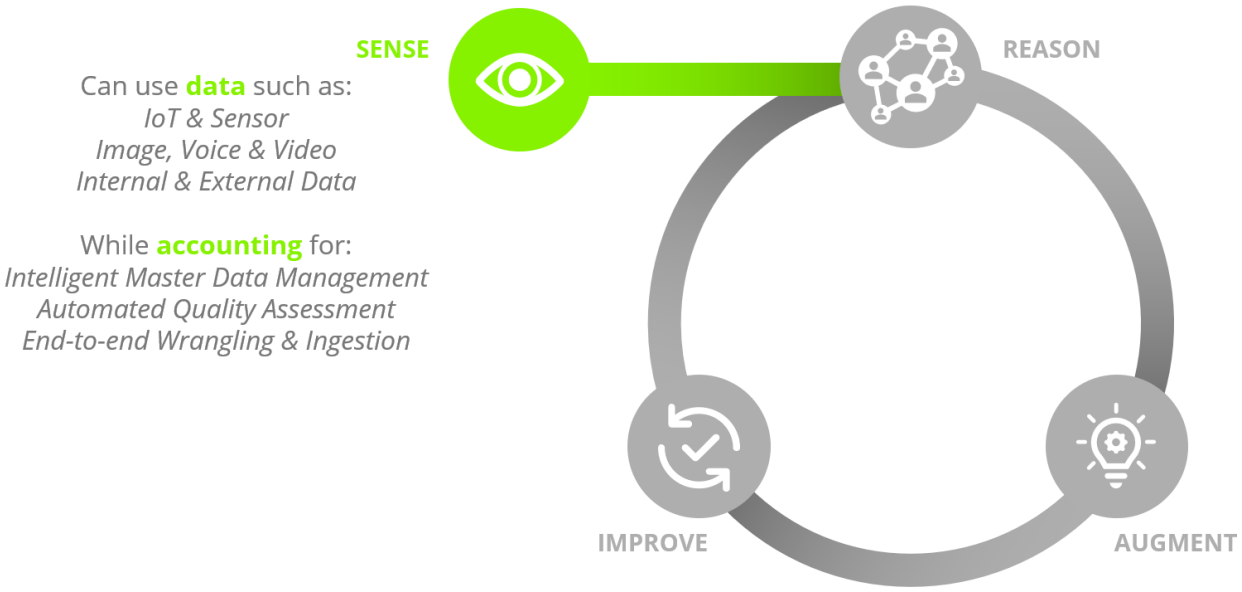


Figure 1

Value of SENSE

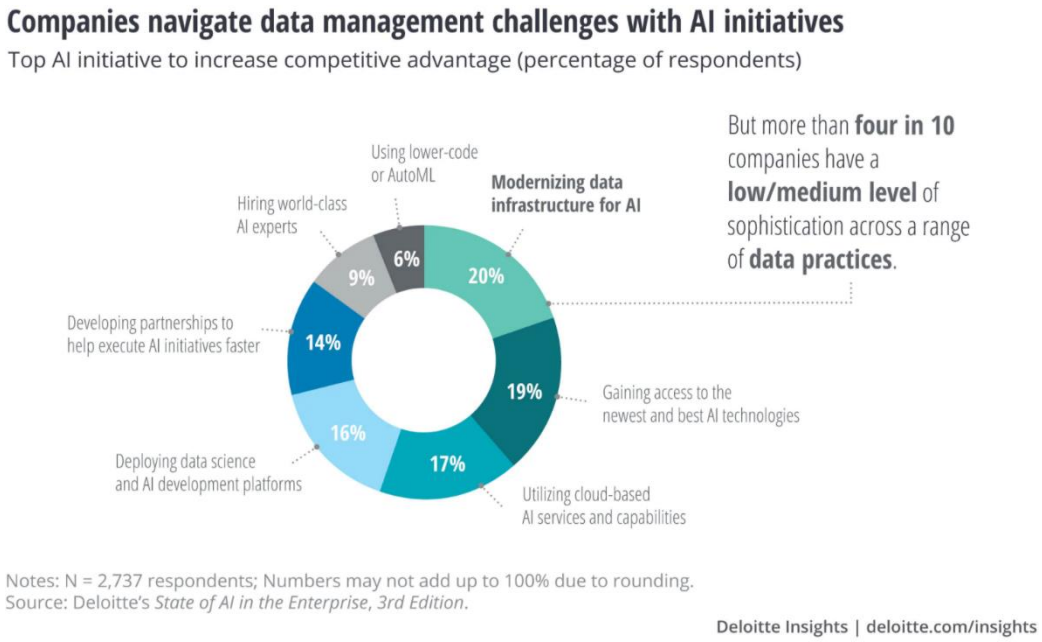


Figure 2

In Deloitte's latest State of AI in the Enterprise survey of more than 2700 executives, respondents indicated that the modernization of their data infrastructure is their top initiative in order to increase competitive advantage from AI.

This is aligned with our experience showing that key barriers to adoption of AI-driven initiatives originate from challenges in data infrastructure, governance and management. Common challenges are:

1. **Data is spread across different business systems across organization**. Organizations are adopting specialized systems to manage an increasing data volume, but these are often tailored to specific business requirements or initiatives within a single line of business, not aligned to the enterprise-wide data architecture. A scattered data environment limits visibility and usability of data assets for data scientists.

2. **Data scientists spend the majority of their time in an AI project to address two factors from the dataset to drive quality of the prediction models:**
    a. **Data quality:** high quality data is necessary, otherwise machine learning algorithms will mistakenly learn to mimic and then produce inconsistent data results. This results in weak performance at the least and harmful outputs at the worst. Data duplication, incompleteness and inconsistency decrease the value of the resulting AI applications and increase the amount of time data scientists spend in trying to fix it.

b. **Data context**: data capturing the context of the problem is as important as data capturing the outcome. Data scientists often work very closely with SMEs (subject matter experts) to make sure the contextual data is also provided to the model. For example, for a manufacturing use case, to perform automated quality control, it is not sufficient to only know the final quality label. Relevant supporting evidence, such as color, weight, dimensions, hardness, and other testing parameters, also need to be provided to the data, which normally comes from the manufacturing engineers or quality inspectors.

To achieve higher data quality and more data context, it's recommended that the organization invests the time and effort upfront in establishing uniform data foundation with integrated capabilities of data security, data management and data governance. A cloud-based data lake provides a central platform to bring together siloed data sources, to enable AI models to SENSE the environment where the organization operates.

The next section describes Deloitte's approach for enabling SENSE.

Enabling SENSE

Given the challenges, what are the best approaches companies can take to accelerate preparation of high quality data for analytics and fast-track deployment of AI projects into production?

There are four guiding principles to enable SENSE in organizations:

1. **Invest in creation of centrally governed data lakes.** Data lakes are data architectures that allow scalable, secure and easy storage of data in virtually any format, both in its original and processed form. With the proper governance in place, they enable centralized and curated storage of data assets, creating a single source of truth across the organization.

   Figure 3 illustrates one possible governance structure for a data lake ecosystem in which different zones offer appropriate governance controls:
   a. Zone 1 is owned by IT and stores copies of the raw data through the ingestion process.
   b. Zone 2 is where business users can create their own data sets based on raw data from zone 1 as well as external data sources. This zone would be trusted for group (i.e., business unit or division) use, and could be controlled by group-sharing settings.
   c. Zone 3 data sets, maintained by IT, are vetted and stored in optimal formats before being shared with the broader organization. Only data in zone 3 would be trusted for broad organizational uses.

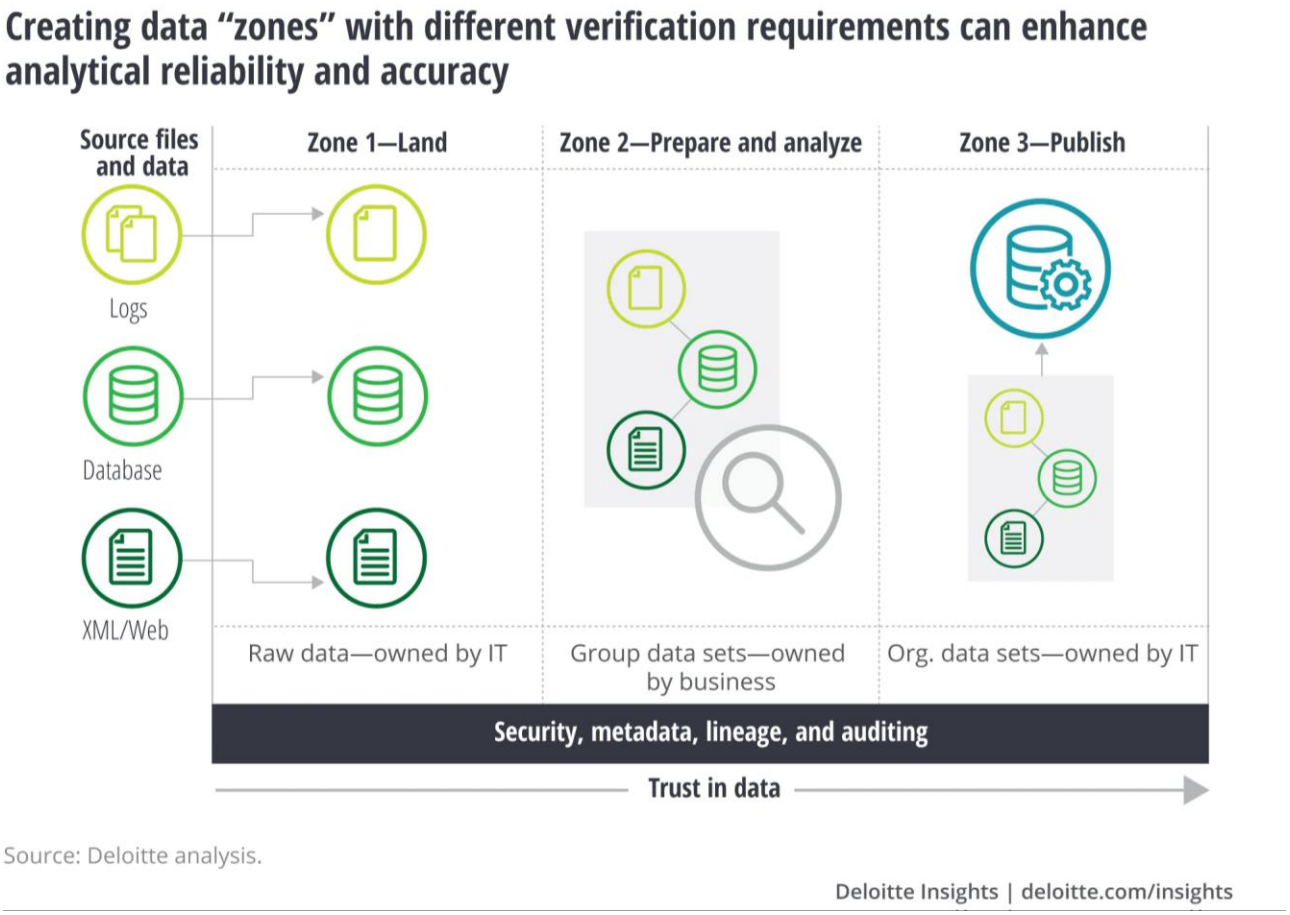


Figure 3

2. **Automate data acquisition from internal and external data sources**
   Data is loaded into data lakes through ingestion pipelines, which should be tailored and optimized to each data source. As a best practice, data acquisition from internal and external data sources should be entirely automated for consistency, speed and efficiency gains.

   One example of how this can be achieved is through real-time processing enabled by event-oriented serverless architectures, which streamline data flow between sources and target. A common application of serverless architecture is in the IoT domain, where streaming data from sensors is merged with batch process data, and analyzed in a single analytics engine to offer an integrated real-time visualization to business users.

   Serverless architectures (e.g. lambda, kappa) implemented on public cloud providers like AWS, Google Cloud and Microsoft Azure, serve low latency views based on data ingestion and processing of massive quantities of data for both traditional batch data and real-time streaming data, ensuring scalability and high availability of the solution.

3. **Implement smart data management ensuring data quality.**
   To enable user access to high quality data, a central catalogue of all data assets needs to be established, where data complies to the organization's governance controls and is checked for quality before being incorporated in AI projects or applications.

   Modern data management solutions like Informatica Axon and Informatica Data Quality; or Collibra in combination with Syncsort Trillium, can automate the tasks to profile the quality of the data using machine learning at scale, and help accelerate data quality remediation projects to correct the data quality at the source system. These solutions also offer capabilities to give automated suggestions to fix data quality and complete the transformations with a simple click of a button. Some examples of automated suggestions for data are gender identification, standardization, matching, and deduplication.

4. **Standardize data preparation and transformation platform**
   Using a standard platform and coding language for implementing data preparation and transformation pipelines, provides the ability to share plans, work, and insights among teams and individuals to improve reusability and shareability of vetted data pipelines and to expedite data preparation.

   As an example, we helped a global retail client to develop standard data pipelines using Python and Azure Databricks to expedite big data preparation tasks and provision scalable infrastructure ensuring security and monitoring of data processing jobs running at scale. This resulted in operational efficiencies within the project while minimizing resource usage and costs.

While enterprises can still get started with AI without the listed capabilities, having these practices in place ensures a solid foundation for scalability and organization-wide adoption. In addition, AI models developed on data of higher quality typically deliver deeper insights and more sustainable results, ultimately generating value for the organization.

What to expect next?

This blog is part of a series in which we deep dive in the different components of DAIL and describe them in a more in-depth fashion. Next up, we will discuss the REASON which will bring us from the data we have collected to the models we employ.

About the authors

Naser Bakhshi

Naser is a Director in the Deloitte Dutch Analytics service line. In the world of ever-growing data and continues new technological advances, his goal is to bring to bring the best, the newest and most innovative analytics solutions and unlock actionable insights for a fact-based decision making. Naser combines enabling technologies such as Machine Learning, Cognitive Analytics and intelligent self-learning systems to develop and embed advanced decisions support systems in the supply chain and in customer domain. Moreover, he's an experienced transition manager with a strong feeling for organizational change and sensitivity helping organizations to define analytics strategy and build analytical capabilities.

Amisha Khera

Amisha is a Specialist Lead in the Dutch Deloitte Data and Analytics Modernization team. Amisha specializes in data platform architecture, mainly involved in cloud vendor projects, driving innovation with the goal to bring state of the art solutions to clients. Amisha has over 9 years of experience in cloud architecture, data engineering and applying AI and analytics at clients in the Consumer and Energy and Resources industries. Before joining Deloitte, Amisha was part of several big data technology start-up companies based in US and Japan.

Alessio Bilato

Alessio is a Consultant in the Data and Analytics Modernization practice of Deloitte Consulting with 3 years of experience in analytics, innovation strategy and cyber security, focusing on the Industry 4.0 and digital factories space. He helps enterprises connect data, design and strategic topics to develop new services and find effective solutions to business and social challenges.

References

https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/about-deloitte/deloitte-cn-dtt-thriving-in-the-era-of-persuasive-ai-en-200819.pdf

https://www2.deloitte.com/us/en/insights/industry/technology/ai-and-data-management.html

https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html

Deloitte.