S&P Global
Market Intelligence **Business Impact Brief**

Bringing Automation and Structure to the Data Lake

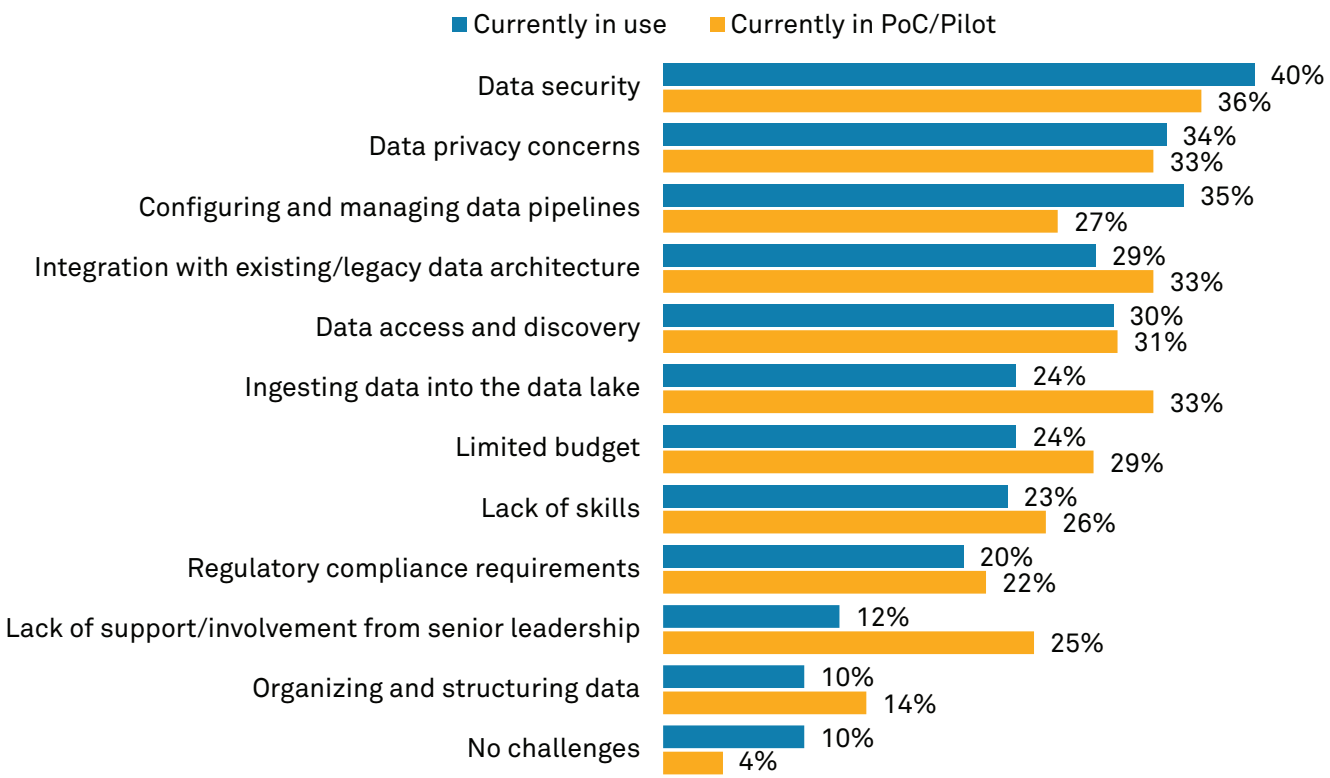
The 451 Take

The data lake has become an established component of many enterprise data and analytics strategies. Nearly three-quarters (71%) of enterprises are currently using or piloting a data lake environment or plan to do so within the next 12 months, according to the results of 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms 2021 survey.

Data lakes have the potential to improve analytic efficiency and innovation by providing a platform for low-cost storage of multi-structured data from multiple applications. Data stored in a data lake can potentially be explored and analyzed by users in multiple roles using multiple analytics approaches – for example, data analysts using business intelligence tools, as well as data scientists using machine learning and predictive analytics techniques.

While the data lake was initially an on-premises environment based on Apache Hadoop, data lakes are today predominantly deployed in the cloud with a focus on cloud object storage services as the underlying repository. Well over half (58%) of respondents from organizations that are already in deployment or proof of concept with data lake environments have data lakes in the public cloud, according to 451 Research's recent Voice of the Enterprise survey results, while 39% have data lakes in an on-premises datacenter.

Challenges Generating Insight from Data Lakes



Q: What are the most significant challenges your organization faces in generating insight from its data lake environment? Please select all that apply.

Base: Respondents whose organization currently has a data lake in use or in a proof-of-concept or pilot stage (n=302)

Source: 451 Research's Voice of the Enterprise: Data & Analytics, Data Platforms 2021

The overall goal of the data lake is to accelerate enterprise decision-making and deliver value from data. 451 Research's survey results also indicate that data lakes are delivering on their promise of providing business agility. More than half (51%) of enterprises that are already in deployment or proof of concept with data lakes cited improved business agility as a benefit, followed by improving and developing products and services (47%), and enhancing customer service and engagement (43%).

However, challenges remain. Data lakes are strategic initiatives with all the associated complexity that comes along with that. Many early initiatives suffered from a lack of appropriate data engineering processes and data management, as well as limited governance functionality, making them relatively inaccessible for general business or self-service users. These former limitations are now being addressed with new functionality designed to marry data management, data governance and security capabilities with the low-cost agile benefits of the data lake. 451 Research survey results highlight the need for this new functionality. In terms of challenges to generating insight from data lakes, data security is the most significant, followed by data privacy concerns, and configuring and managing data pipelines (see figure above).

Business Impact

Data lake challenges evolve with experience. Those in the proof-of-concept/pilot stage of data lake adoption are more concerned about ingesting data into the data lake (33%) than they are about making use of it (data access and discovery – 31%, and configuring and managing data pipelines – 27%). Configuring and managing data pipelines is a more significant concern for enterprises in production (35%), while security and data privacy challenges are also more significant for enterprises already in production.

Delays from data to decision remain a barrier to sales and business agility. Although the vast majority of enterprises are attempting to accelerate data-driven decision-making, the time taken to generate value from data remains a challenge. Nearly a third (31%) of survey respondents reported that it takes their organization more than six days to generate valuable business insight from raw data – a situation not helped by the complexity of assembling raw data from multiple sources.

Data lakes enable support for mixed workloads on the same data. While data lakes are primarily associated with business intelligence, two-thirds of survey respondents are using their data lakes to support mixed workloads. More than a quarter of enterprises that are already in deployment or proof of concept with data lakes are running business intelligence, data science and operational workloads on their data lakes, while 21% are running both business intelligence and data science workloads.

The benefits of data lakes also evolve as enterprises move from proof of concept to production. A higher proportion of respondents currently in production with a data lake cited improved business agility as a benefit (57%) compared to those in pilot or proof of concept (43%). This appears to validate that improved business agility is a benefit that increases based on deployment maturity. Those in production with data lakes are also more likely to cite the benefits of automating business processes and reducing risk exposure.

Looking Ahead

The future of the data lake is a combination of the benefits of low-cost cloud data storage and structured data management, data governance and security capabilities designed to address some of the challenges cited above. These are capabilities that have previously been associated with the data warehouse rather than the data lake but are now being incorporated into what is sometimes called the 'data lakehouse' or a 'modern data lake.'

Interest is growing fast, especially among enterprises that have already adopted a data lake. 451 Research's survey results found that 93% of enterprises already in production or proof of concept with a data lake are currently using or piloting a data lakehouse environment, or plan to do so within the next 12 months.

Whatever you call the resulting environment, new products and functionality are available to accelerate the data pipelines in a data lake by automating data ingestion, data staging and data normalization. Additionally, new table formats provide the basis for more advanced data security and governance. In combination, these features enable enterprises to focus their attention on analyzing the data and accelerating business insight, rather than ingesting, staging and curating data.