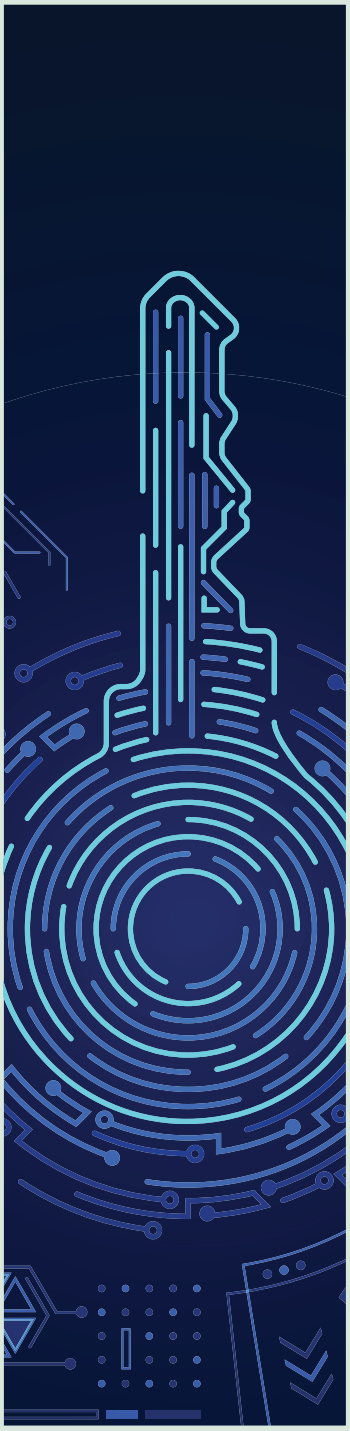
WHITE PAPER



HOW CAN PRIVACY ENHANCING TECHNOLOGY ON AZURE ENABLE IMPROVED DATA ANALYSIS FOR AI

Abstract

Despite organizations investing in Artificial Intelligence(AI) and automation through Machine Learning(ML), many enterprise data projects are unsuccessful as organizations fail to focus on the data needed to solve business challenges.

The rate at which organizations exchange data across internal and disparate external networks has tremendously increased, making it more complex and expensive to derive insights from data.

This whitepaper provides a human-centric framework for an organization's data and a high-level design for a purpose-built privacy-first data lake on Microsoft Azure.

The objective is to enable organizations to identify, secure, share, and monetize data through the **Infosys Enterprise Data Privacy Suite (iEDPS)** in an Azure ecosystem.

Infosys®
Navigate your next

Table of Contents

Table of Figures

Intended Audience

This whitepaper is intended for Azure Architects, Enterprise Data Architects, Data Scientists, and Data Privacy Professionals.

Better Data is a Business Imperative

Today, every enterprise wants to be data-driven. Gartner states $200B is spent annually on data-centric services to maximize the potential of data. 9 of 10 companies that invest in data-centric services fail or have very little evidence of a positive return on investment (ROI). Building a profit-generating data-centric project requires a lot of effort and symbiosis across the organization - from planning to deployment.

Further, businesses must be equipped to regulate their digital data in line with modern legal and privacy requirements, such as GDPR, CCPA, Consumer Data Act, and other personal data protection legislation. These privacy standards and regulations aim to protect consumers from businesses that inappropriately or unethically collect, use, or share their personal information.

These challenges have changed the way businesses manage, protect, and process consumer data. For instance, under GDPR, individuals are empowered to request:

- What personal information is captured?
- How personal information is being used?
- How is personal information deleted on request?

Organizations are investing heavily in AI and automation through machine learning. Despite many enterprise data projects failing, organizations must focus on building data for business use cases. The focus of data experts should be understanding the data and calculating its utility in solving key use cases across the following parameters:

- **Data Utility** – Is this the right data set in the correct format for analysis and solving the use case?
- **Personal Data Breach** – Is any specific end-user data being used without consent? Will this cause a compliance issue resulting in hefty fines or lawsuits?
- **Missing Data** – Does any part of the data is needed to be synthetically generated?

This whitepaper provides an outline to design and conduct business on a privacy-first data lake on Azure to identify, secure, share, delete, and monetize enterprise data on Azure leveraging the Infosys PrivacyNext platform amidst a partner ecosystem.

Key Challenges Faced by Enterprises in Finding the Right Data

Organizations today manage petabytes of data in data warehouses, files, cloud, and physical storage. These large data sets leveraged for decision-making could be structured, unstructured, or historical data not readily available.

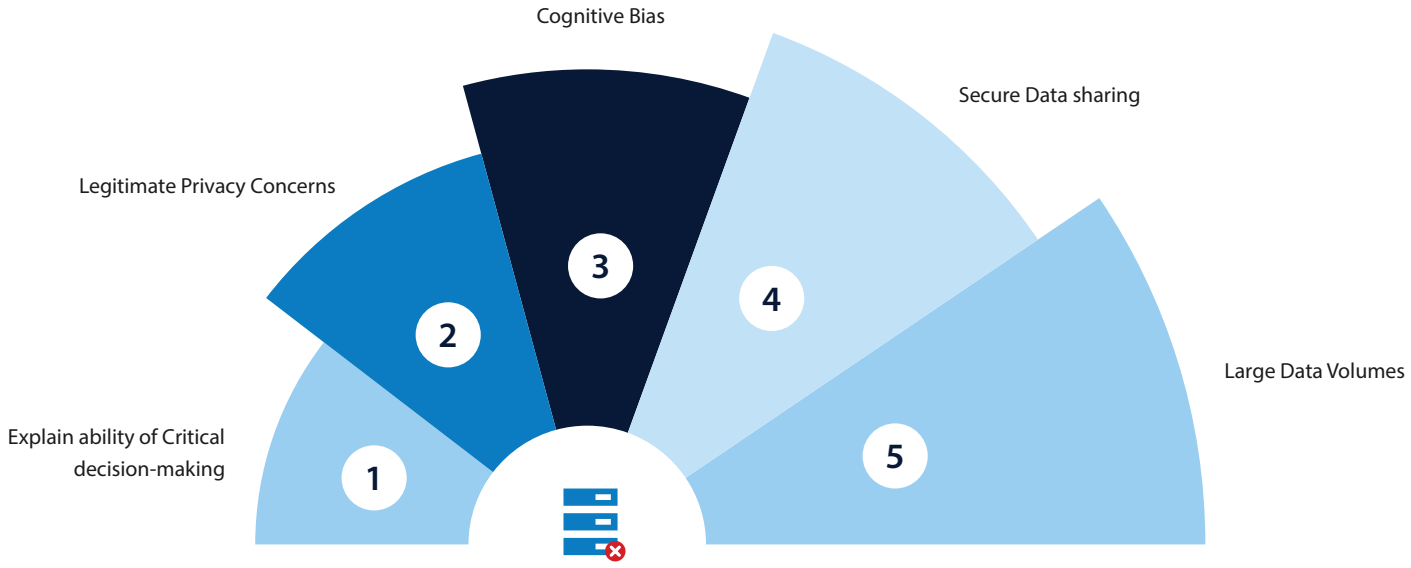


Figure 1: Key data challenges faced by Organizations

Some of the key challenges faced by enterprises are:

- **Explaining Critical Decision-making Data** – The decisions made by a machine must be central to decision-making parameters and aligned with use cases. Organizations struggle to find traces of the data vital to the overall decision-making process.
- **Legitimate Privacy Concerns** – With the rise of regulations such as GDPR, CCPA, Consumer Data Act, and other personal data protection acts, there is a risk of personal data breach leading to lawsuits, brand erosion, and compliance penalties. For instance, a personal data breach and non-compliance with GDPR can cost a company 4% of its annual global turnover.
- **Cognitive Bias** – Machine learning models are always viewed as mysterious black boxes. There are more than 200 identified data bias types. For example, in the Zeigarnik effect, incomplete tasks create a bias in the decision of a call center application chatbot instead of a complete task.
- **Secure Data Sharing** – Many organizations cannot securely share their data with partners, suppliers, or vendors outside their organization. There is a likelihood of personal data getting re-identification and a potential risk of data breach. For instance, banks struggle to share sensitive data with institutions, regulators, partners, or customers.
- **Large Volumes of Data** – Business and technology leaders are unsure how an individual's data has been handled and they tend to get overwhelmed by the enormous amount and variability in data at their enterprise. Moreover, the problem has escalated as the pace at which data is being collected and exchanged across internal and external networks has increased.

Human-Centric Framework to Generate Data – Azure Reference Architecture

From stealth startups to fintech giants and government institutions, teams are feverishly working on their data strategies. For a successful data pipeline on Azure, the key is to build contextual data with high data utility. Machine learning tutorials start with the assumption that data for solving the business problem is already available to the user. Hence, the success of data analysis lies in finding or creating the data to solve the business problem.

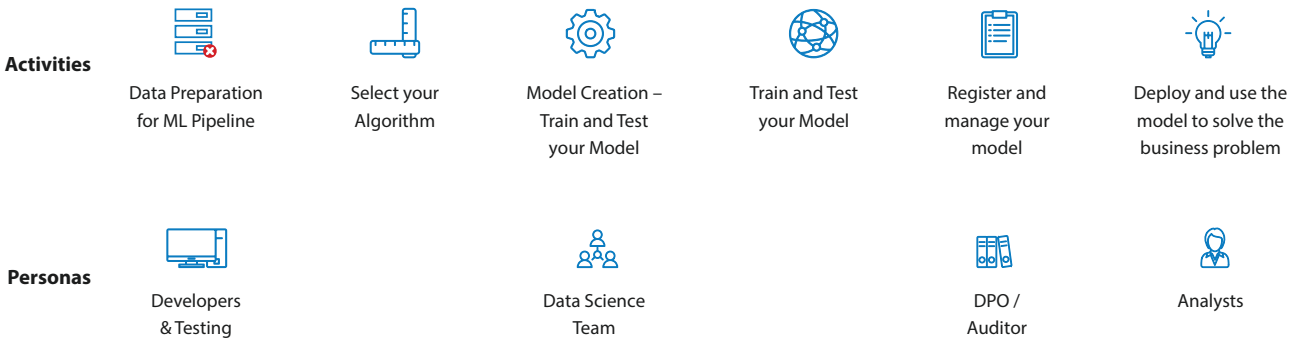


Figure 2: Lack of right data for your Data pipeline

The Coronavirus (COVID-19) pandemic pushed companies into uncharted territory. Conferences switched to virtual meetups, Zoom replaced air travel, and remote connectivity replaced the workplace.

These shifts have changed the way data is used for making informed and critical business decisions. Data moved from independent data centers to the cloud, and with the advent of 5G, to the edge. Artificial Intelligence (AI) algorithms need better data sets to mature and provide optimal outcomes. Data-driven enterprises rely on the context of data being captured correctly. To enable enterprises from managing data to becoming data-driven, Infosys proposes a human-centric framework that consists of four key steps for processing, analyzing, deconstructing, and re-imagining data:

- **Identify Your Data** – Unbundle data from key business processes and user interactions through a Data Discovery Process. Data could be in structured or unstructured forms, or the files can be analyzed for:
    1. Business Context – How useful is this data for solving the business analysis problem?
    2. The Emotion Behind Data – What is the intent behind every user interaction with the system or other users?
    3. The Underlying Cost of Data Privacy – Apply regulatory templates and assess the exposure level of personal data used in decision making.
- **Secure Your Data** – Traditional perimeter-based security and privacy are not enough, as enterprises are increasingly operating across the edge and hybrid cloud ecosystems. Privacy by Design needs to be the standard with proactive protection and built-in post-breach remedial measures.
- **Share Your Data** – Privacy defaults should be applied with the user's consent before data sharing. Moreover, secure data sharing should be enabled through hardware-based security or data anonymization when data is shared across employees, customers, or partners.
- **Monetize, Audit, and Delete Data** – Enterprise-ready data sets will have to be generated for training and learning of AI models. These ready-to-use data sets need regulatory audits and secure disposition post usage.

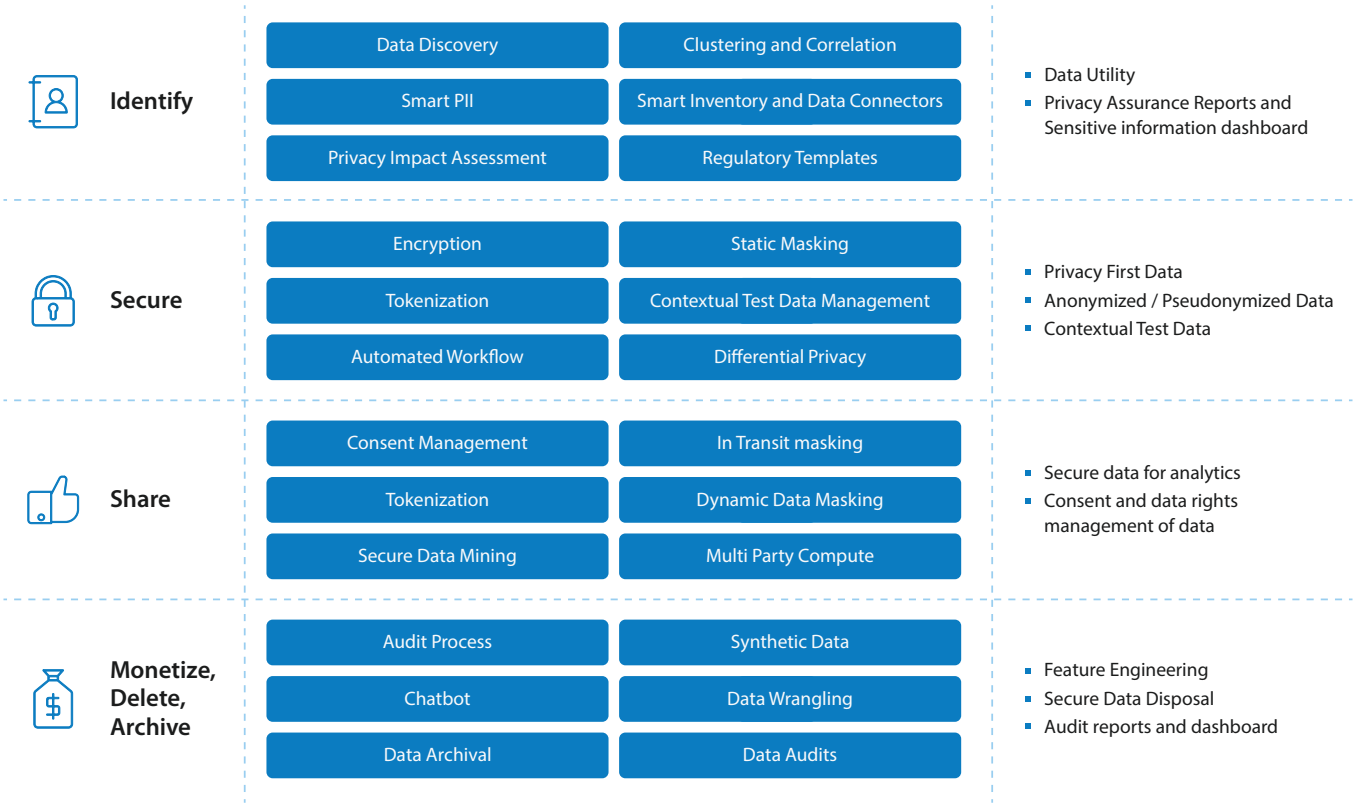


Figure 3: Human Centric Framework for Data

Privacy-First Azure Data Lake for a Data-Driven Enterprise

Increased computing power, less expensive storage capacity, and better network connectivity are turning the current flood of data into an endless wave of customer personal information, sales data, product specifications, and customer preferences. Data is being consumed and generated in different formats such as IoT data, images, files, social media feeds, and via internal collaboration tools such as email, chats, and Teams or Zoom video recordings.

Business leaders are looking at a simplified solution that can be designed to ingest, process, secure, and analyze both structured and unstructured data. Virtual Data Lakes on Azure provide an optimal solution that can operate in conjunction with traditional Enterprise Data Warehouse (EDW) and Cloud-Based Warehouses. Data stored in native formats and ready-to-provision infrastructure on the cloud provide significant cost savings.

Infosys recommends a 4-stage privacy-first data model based on the human-centric framework with segregated zones for building data privacy controls. Enterprises can discover, protect, share, and finally monetize and consume data for AI, automation, and data analysis.

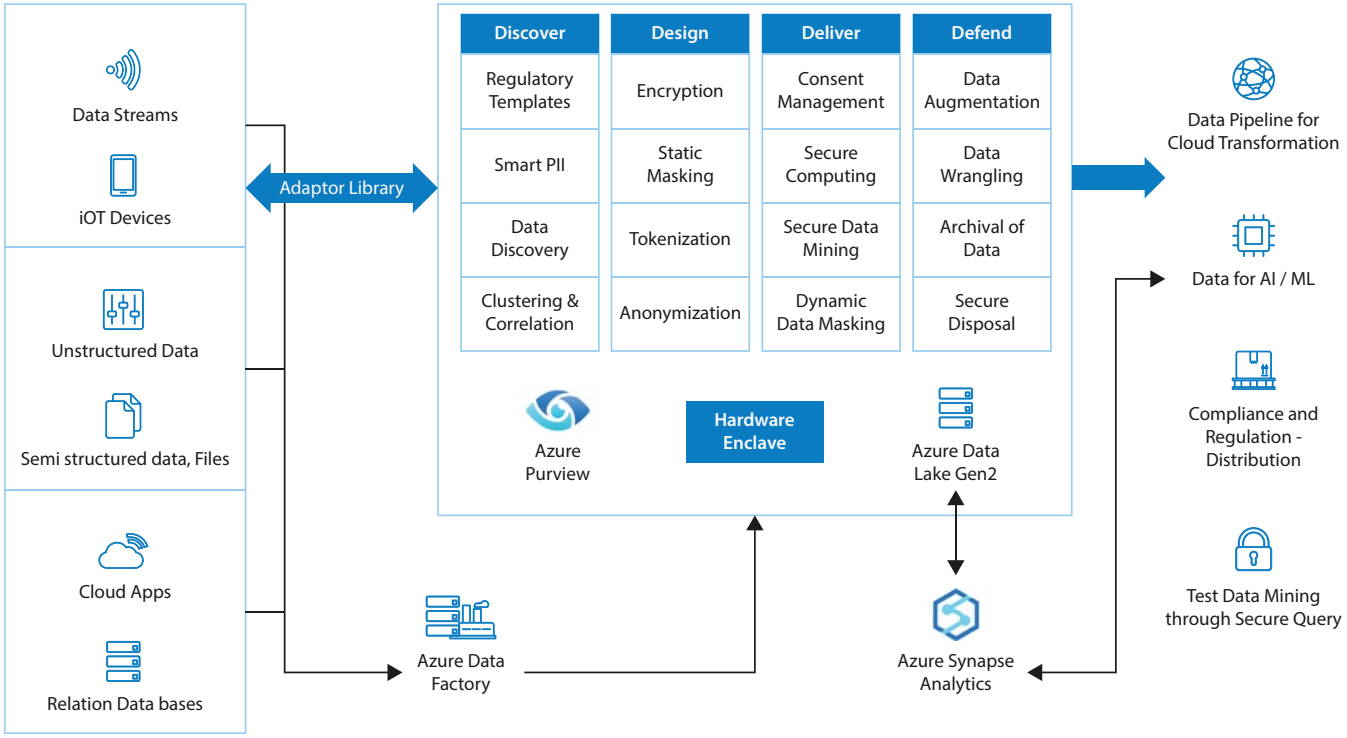


Figure 4: Privacy First Azure Data Lake for a Data Driven Enterprise

Stage	Phase	Value
Stage 1	Discover	The first phase focuses on identifying sensitive information ingested from heterogeneous data sources ranging from data streams, videos, unstructured data, files, app data, and structured data on RDBMS through Azure Data Factory and Infosys' library of adaptors. The data ingested is discovered, where personally identifiable information (PII) and other sensitive information is scanned and identified by the Data Discovery capabilities of the PrivacyNext Platform and Azure Purview.
Stage 2	Design	The ingested data is secured through 180+ algorithms selected based on the business use case and the type of data that needs to be secured. The focus is to anonymize or pseudonymize the data based on the use case.
Stage 3	Deliver	A two-pronged approach to share the data is incorporated – consent management is a critical step where an individual's data rights are at the heart of the sharing process. Next, additional controls like data virtualization, differential privacy, and multi-party computation with dynamic data masking, are applied to share the data.
Stage 4	Defend	The consumption phase is where the data can be synthesized, contextually generated, and wrangled for feature engineering and consumption for a Data Pipeline. This step also focuses on secure data disposal and proactive monitoring of data for data privacy risks.

Purpose-Built Privacy-First Data Lake for AI / ML Data Pipeline

Successful organizations connect silos across the enterprise to leverage data for their AI / ML Pipeline. Let us take an example to illustrate the privacy-first data lake in action. With the advent of 5G and the race to launch multiple high-bandwidth data products, many telecommunications companies are struggling to manage their complex Service Delivery and Assurance processes. Telecommunications companies need to find a balance between the rapidly evolving 5G-based data-driven and virtualization economy, Internet of Things (IoT), and the growing complexities of regulation around Industry 4.0.

To discover the factors influencing customer churn and preferences on a 5G network, telecommunications companies need to collect and analyze user data, usage patterns, billing, and other consumption metrics. Sharing data within the organization or outside for this analysis could cause a potential data breach.

Now, consider a process using the human-centric framework on a privacy-first Azure data lake that enables an organization to:

- Collect data from diverse sources in the organization and administer regulatory templates to identify sensitive data.
- Extract relevant information like customer usage, billing history, and personally identifiable information, securing it through various obfuscation techniques and storing it in the Azure data lake.
- Share the data securely with data analysts for predicting churn by pulling relevant information from the data lake.
- Monetize data by using the privacy-first data that generates data sets for AI/ ML pipeline and wrangles data through partner products such as Trifacta for feature engineering, predicting, and forecasting churn.

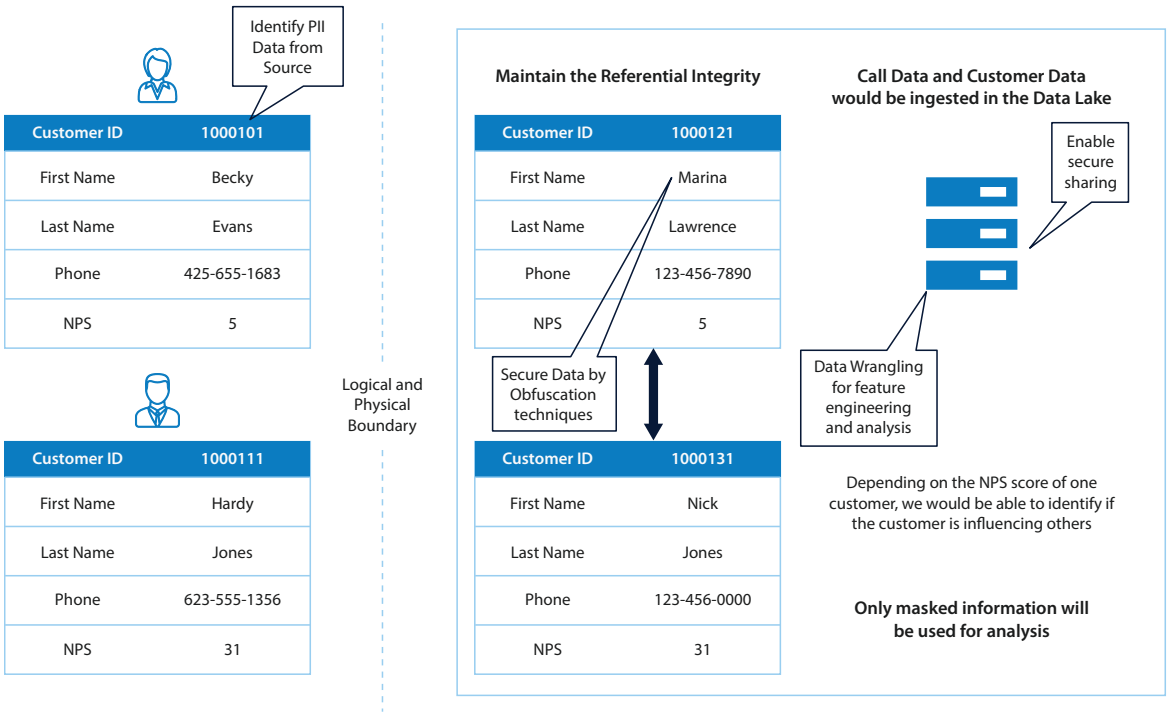


Figure 5: Prediction of NPS on Privacy First Data Lake

Once designed and implemented, a privacy-first data lake from Infosys Enterprise Data Privacy Suite (iEDPS) on Azure enables organizations to reimagine how data is ingested, consumed, and shared securely both inside and outside the enterprise.

iEDPS provides enterprise-class data privacy capabilities and enables organizations to adhere to global regulatory standards such as GDPR, CCPA, HIPAA, PIPEDA, GLBA, ITAR, and various local regulations. Built-in deterministic, selective, dynamic, and static masking features, Data Discovery, and Data Generation capabilities, iEDPS can be deployed on any platform and supports all major databases and file systems.

With 70+ successful customer implementations and a dedicated team focused on data privacy consulting, product engineering, and customer success, iEDPS enables privacy-first organizations and unparalleled protection of enterprise data.

For more information, visit iEDPS on the Azure Marketplace.

About the Authors

Pramod Vasanth is a Principal Cloud Solution Architect in the Microsoft Partner Success Team. Pramod is part of a technical architect team which focuses on defining and executing the strategy to enable Microsoft partners worldwide to increase their cloud technical capacity and capability. Pramod has over 19 years of experiences in working on large and complex projects on Microsoft technology. Pramod focuses on Application and Infrastructure design on Azure.

Anil Dwarakanath is a Principal Cloud Solution Architect within Partner Success Team focused on Advanced Analytics, Cognitive Services, AI and Machine Learning on Azure platform. He also has extensive experience on SQL and No SQL databases, Data Warehousing, Advanced Analytics using Big Data technologies, Visualization and worked with several marquee customers in deploying high scale analytics, AI solutions and ML models. He is based in Redmond. He is also a Microsoft Data Science Insider focused on working on scenarios where customers are looking to embark on AI and ML on Azure.

Nandini Venkatesh Adhini is an Associate Business Analyst at Infosys Center for Emerging Technology Solutions (iCETS). Her primary skillset involves formulating pre-sales strategy for privacy offerings, conducting market research and competitor analysis.

Karthik Nagarajan is an Industry Principal Consultant at Infosys Center for Emerging Technology Solutions (ICETS). He has more than 15+ years of experience in customer experience solution architecture, product development, and business development. He currently works with the product team of Infosys Enterprise Data Privacy Suite, Data augmentation, and CX strategy.

References

- https://iapp.org/news/a/5g-to-raise-privacy-challenges-and-opportunities/
- https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/connected-world-an-evolution-in-connectivity-beyond-the-5g-revolution
- https://www.mckinsey.com/~/media/mckinsey/industries/advanced%20electronics/our%20insights/the%205g%20era%20new%20horizons%20for%20advanced%20electronics%20and%20industrial%20companies/the-5g-era-new-horizons-for-advanced-electronics-and-industrial-companies.pdf

For more information, contact askus@infosys.com

Infosys®
Navigate your next