KANINI
Excellence Redefined

Data Discovery

The Means to the Right End.

Table of Contents

Introduction

Data is Everything!

This phrase goes around the software industry whenever a new program or project begins. But this is the most clichéd and misused phrase as people say this without attempting to understand the true impact of good data.

It's not that they don't want to, but it's because data initiatives are so intertwined with the business as well as technology that it takes a lot of meticulousness to build efficient data platforms. Many a time, projects get initiated with more planning on the "Digitalization" and poor planning on the "Data Transformation". This causes havoc at various levels. Any project that focuses just on digitalization, ignoring data is tantamount to committing self-destruction.

This brings up the question - are these problems with building data platforms because of lack of awareness or due to the feeling analogized to "when we are in the jungle, we are left with no other choice, but to take one day at a time", i.e., "data problems are anyway uncontrollable, so why worry about planning ahead?"

Well, the answer is a bit of everything! Data platforms are not just a one-time investment. They need to be carefully built and monitored. But it is worth the effort. Why?! Data is useful only when it is of high quality. The cost of bad data is higher than the cost invested in building a data platform that prevents the creation of bad data and its diffusion across the business systems. The repercussions of "bad data" are exponentially higher for a business.

> Every year, poor data quality costs organizations an average of $12.9 million. Data quality is directly connected to the quality of decision-making. By 2025, 60 percent of data quality processes—as opposed to being separate, independent tasks—will be autonomously embedded and integrated into crucial business workflows.
>
> - *Gartner*

A Robust Data Platform

Need of the Hour

In today's data-driven business world, organizations usually connect to several data sources. But, unfortunately, they wouldn't have an idea as to how many sources they are currently pulling data from. Obviously, such businesses fail to achieve their objectives due to poor data quality.

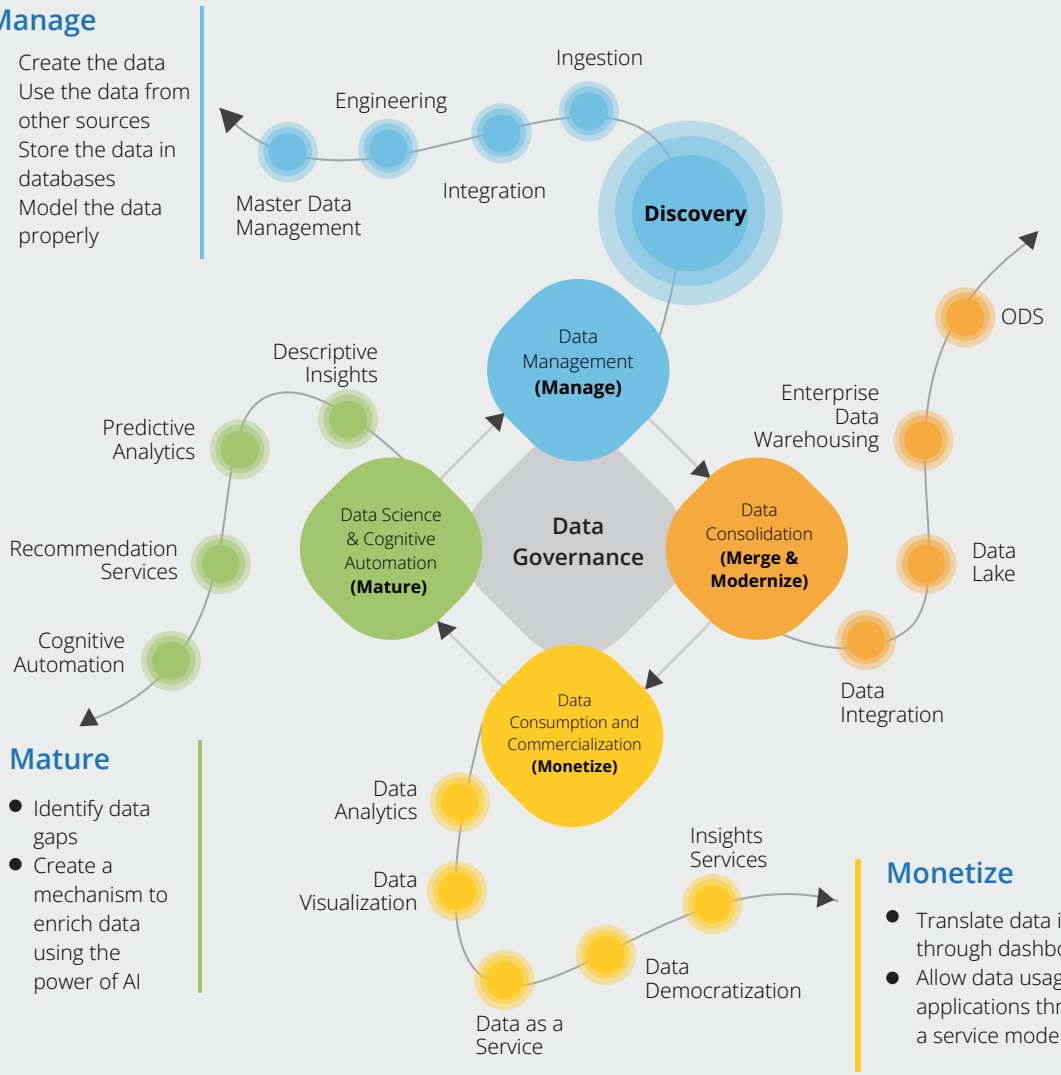
As a famous adage goes, "You cannot build a great building on a weak foundation." Similarly, to avoid poor data quality, an organization should consider building a robust and reliable data platform. Its primary focus should be on conducting a solid / thorough "data discovery exercise" to avoid poor data quality. Just like a superstructure needs a strong foundation to be built, a data platform cannot be effective without a proper understanding of how data travels from raw data to data insights.

Every data initiative needs to go through a cyclic exercise in a data platform, constituting four key elements – Data Management, Data Consolidation, Data Consumption & Commercialization, Data Science & Cognitive Automation – across four functional elements such as Manage, Merge & Modernize, Monetize, and Mature (the 4Ms) respectively.

4 Key Pillars of a Data Platform

Manage

- Create the data
- Use the data from other sources
- Store the data in databases
- Model the data properly



Merge

- Integrate the data from various data sources
- Store the integrated data in a data lake or a big data platform
- Brings data from different sources into the data lake
- Tech Stack: MuleSoft/ Informatica ETL/Spark, etc.

Modernize

- Use the data in the data lake for aggregation and consolidation via ETL
- Save the aggregated data in the data warehouse
- Tech stack: Snowflake, Redshift, Synapse, etc.

Mature

- Identify data gaps
- Create a mechanism to enrich data using the power of AI

Monetize

- Translate data into insights through dashboards
- Allow data usage by other applications through data as a service model

Every data platform needs to be revisited periodically, as and when the business gets diversified or when there is a change in business re-strategy. Data journeys in a data platform are always defined by the business objectives. Out of the four elements of the data platform, data management is the most important as it involves both creation and management of data repositories and begins with “data discovery”.

What is Data Discovery?

In today’s world of data proliferation, every business is becoming a data-driven organization. Their success is highly dependent on effectively gathering, managing, and analyzing data to provide actionable insights enabling them to make smarter decisions. For this, they need to thoroughly understand where the data is coming from - the varied sources of data within the organization.

Data discovery starts with the end goal in mind. It is the process of diving deep into all your data locations and gathering resources from disparate systems to reveal new meaning (insights) of the data. Without knowing the data locations and data points you want to bring together, representing different relationships within the data (Data visualization), is not possible.

Data discovery starts with searching for answers to the right questions.

Key Questions

- For a given business problem, what all could be the business workflows?
- For each of the data flows, what are the various source systems that send the data?
- For the different data flows, does the data change and transform?
- Should the data be pulled from the sources or should the source systems push the data?
- What is the frequency of data pull or push?
- What is the tech stack of the source systems?
- What is the source data schema?
- Does the source schema need to get transformed?
- Are there any relationships between sources?
- Is there a need for integrating the sources and extracting the data into one common repository?
- What is the data staleness rate of each source data?
- What is the fault tolerance of each source system?

Why is Data Discovery Important?

Organizations rely on data to make crucial decisions every day, but they need to ensure the quality of their data analyses. Your data science team should sort through millions of data points in multiple formats and from various sources to uncover all the value sealed within the data. Data discovery helps companies turn this data into useful insights.

A robust data discovery exercise is a culture that transforms the organization into a well-oiled engine. Just like an engine that needs good care and timely oil replacements, a data platform engine works well with the right data oil.

> A good data discovery exercise transforms the entire organization to build a robust data-driven culture.

Benefits of Data Discovery to Your Organization

- Focusing on data discovery helps the organization to understand which sources of data need to be kept, or revisited and decommissioned, or repurposed.
- Data discovery helps an organization to find out the rate of data decay for each source and take corrective and preventive actions.
- Data discovery allows the organization to understand and plan the data delivery from each source.
- Good and repetitive data discovery exercise will help an organization to avoid blind spots and reduce losses.
- It also helps an organization to truly merge, monetize, and mature its data well.
- When people in the organization become more data aware, the organization thrives, not just survives.

Misconceptions about Data Projects Hampering Data Discovery

The importance of data discovery is quite understated in most modern-day data projects because:

- Technology has grown so much that there is a common thought that data problems can be solved by technology as long as we have a skilled technology team constituting data architects. It is a misconception that we need not worry much about the veracity of the data discovery process and the data architecture.
- It's a common apprehension that data discovery is a time-consuming process that eats away the time taken for designing and development of a data platform or for the problem to be solved.

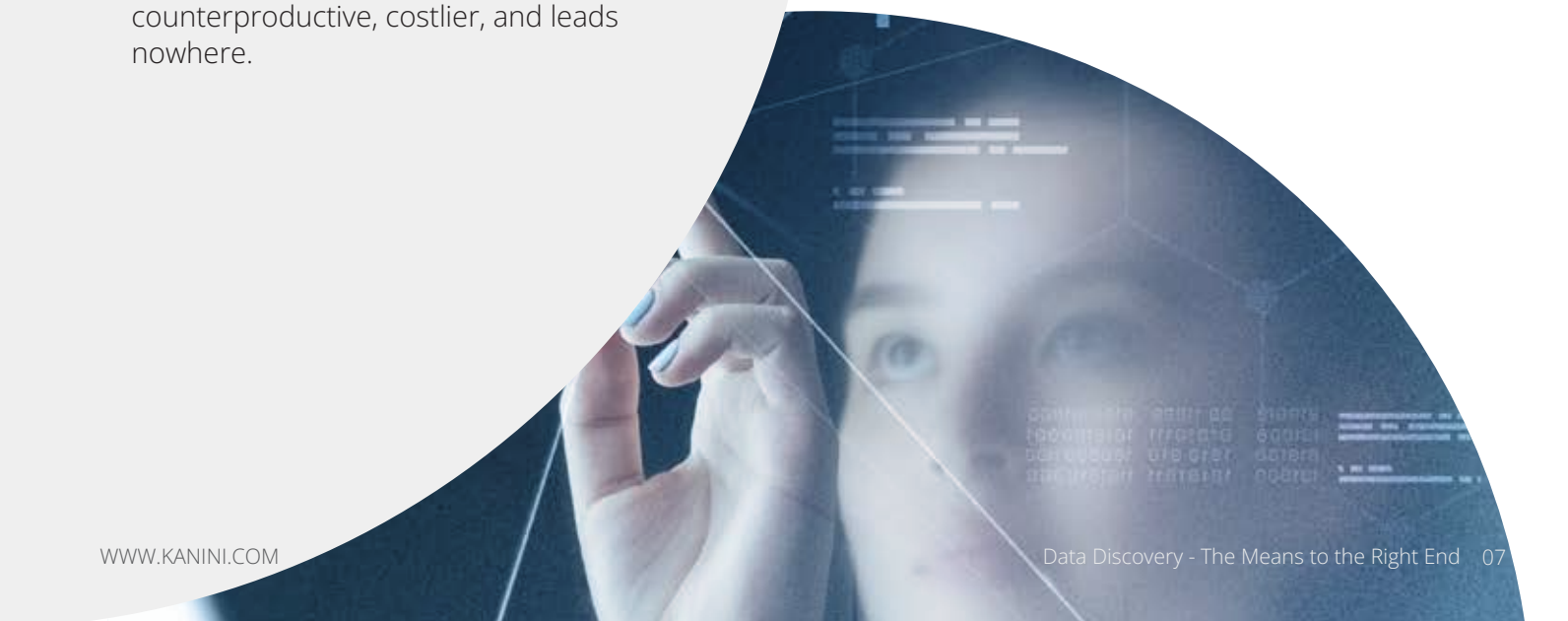
These misconceptions not only kill the pace of a data project but also take a route that is counterproductive, costlier, and leads nowhere.

Data Discovery

A Three-pronged Approach

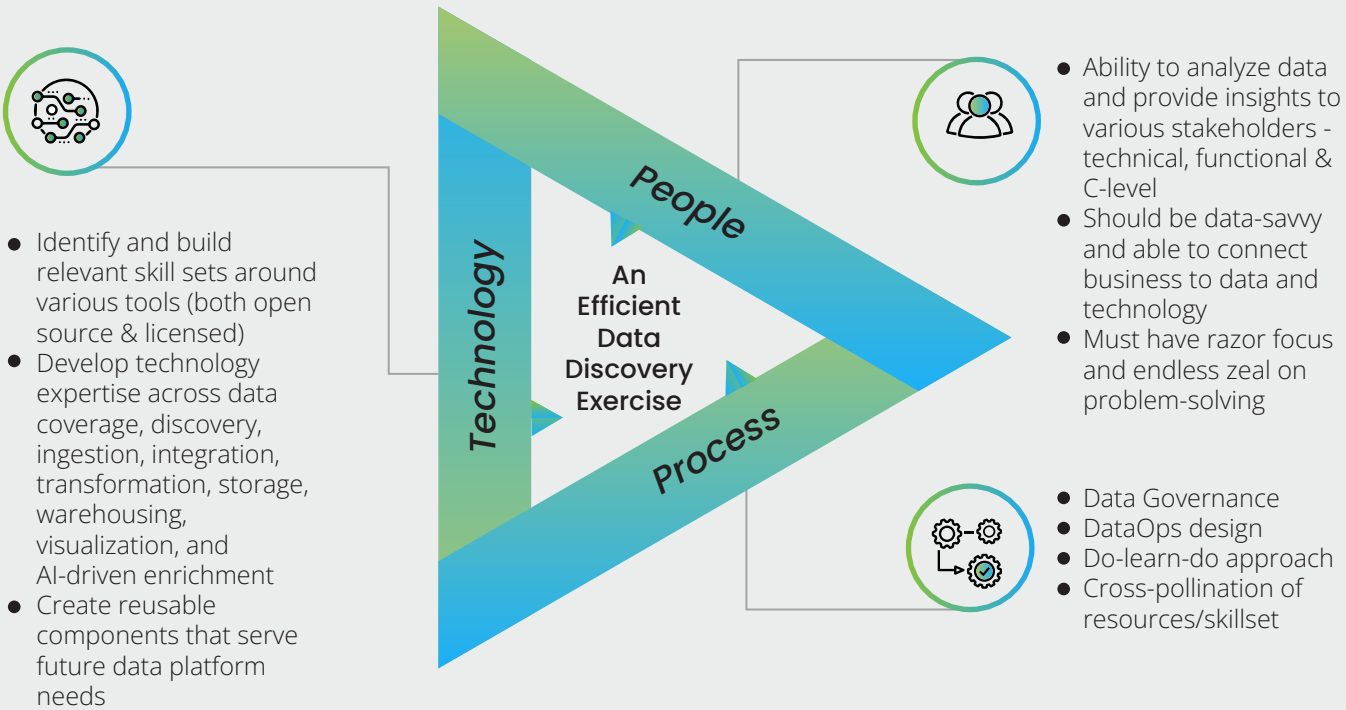
Let's imagine that your organization is collecting several terabytes of data at various stages of the customer journey. To obtain business insights from this raw and unstructured data, your team must spend a lot of time sorting and cleaning the data. In addition, you might have been constrained by the capabilities of your systems. To manage all this efficiently, a good data discovery exercise entails a combination of people with relevant skills, processes that streamline data sorting, cleaning, and technology – both software and hardware.

However, bringing technology, process, and people together is easier said than done. How much is the investment you need to set up the right infrastructure? What goes into accomplishing the right structure and talent in your data team? How can you handle the operational challenges that come with managing people, process, and technology in your data project? Let us find the answers:



A Three-pronged Approach for Data Discovery

People, Process, & Technology



01

People

Building the right talent and team composition

In the data discovery exercise, the team consists of data analysts, data engineers, data architects, data scientists, and machine learning engineers. All these roles contribute to different segments of the data pipeline and require a specific skillset. For the success of any data project, the right team composition consisting of a variety of data professionals and talent is critical.

For instance, organizations which are at the beginning of building their data platform will require more data engineers to implement and manage it and set up repositories. As they progress towards the next steps of their data roadmap, more data architects will be required.

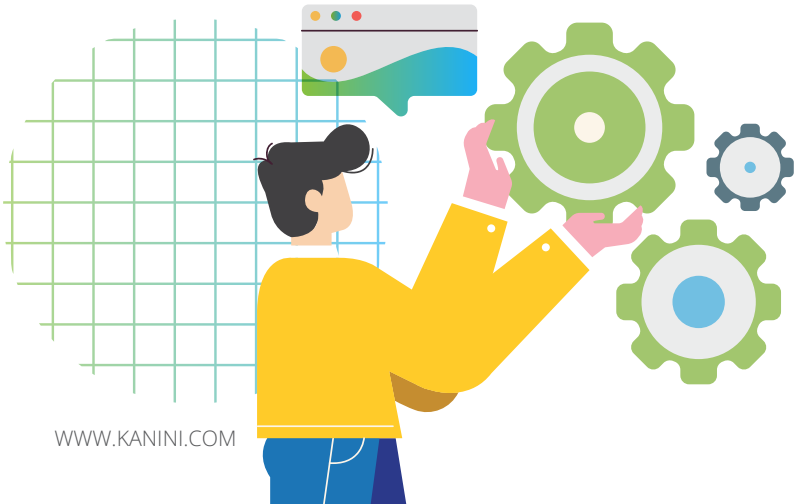
Likewise, technology architects and business analysts also play a critical role in the data discovery journey. The role of a data architect is also to work with the business analyst and the tech architect to automate wherever possible in the data discovery process.

Data discovery is an ongoing process to identify how and to where the data is flowing across the organization and understand the risk. Hence, it requires multiple people playing different roles throughout the data journey. So, who are the people who should be involved in this exercise? Is it a data architect or a business analyst or a technology architect? The answer is... a mix of all three.

The business analyst should clearly set the expectation of the project outcomes and the business process flows.

The data architect should work on finding all the relevant data sources, data flows resulting in data ingestion, integration, transformation, persistence and analytics.

The technology/solutions architect should be able to provide the right tools and techniques for helping the data architect and business analyst.



It is important to note that these skill sets are typically not cross-functional. This means that a data engineer would not necessarily have the ability to perform the role of a data scientist or a data architect. Therefore, the talent acquisition team must understand how different roles and expertise contribute to the end-to-end data pipeline. This will help them map the right candidates to the right jobs and build strong data teams.

Process

Establish a data governance team ensuring an efficient data discovery

Data discovery involves continuous research and experimentation. It is a tedious process. The more time you spend on data discovery, the greater are the chances of building a strong data platform. All this while data projects are not funded for life. So, the processes need to be tailor-made to exactly suit the needs of the end outcome. It's the responsibility of a data architect to work with the team to help identify the process at the very beginning.

Every "i" must be dotted, and every "t" must be crossed – pay meticulous attention to detail. The data discovery process must be documented very clearly. Details about the source systems, their business flows, data flows, the key business fields and their impact on the end outcome should be captured clearly in the process documentation.

Data is like a river. No one can clearly paint the course of the river as the river can easily take its own course. For a smooth process to be established, organizations should set up a data governance team that comprises data stewards, architects, and SMEs. Data architects or data stewards are NOT just SQL developers or ETL developers. They are the ones who should envision the journey of data and build control systems throughout the journey, so that the right data reaches the right stakeholder for the right reasons. A data architect should be able to predict the course of what can go wrong and build control systems and control it.

Be ready to learn, unlearn and relearn to drive success to your data projects.

The data governance team will have the onus of understanding and documenting the below:

Governance aspect of data.

How is master data going to be managed?

What can be master data?

What fields should go into a data warehouse?

What should go into a database?

How should these fields be archived?

Who should have the access to the data?

What is the data provisioning strategy?

How should data be anonymized (if there is a need)?

How should data be democratized?

Who and what applications should see the data?

Be aware and on top of data decay, data leaks.

The characteristics of a good data governance team for proper data discovery are:

- Patience
- Mindset of "Leave No Stone Unturned"
- Identify repetitive tasks and push them for automation
- Be ready to learn/unlearn and relearn
- Travel the entire cycle of data platform build as opposed to the "Build-Operate-Transfer" mindset.

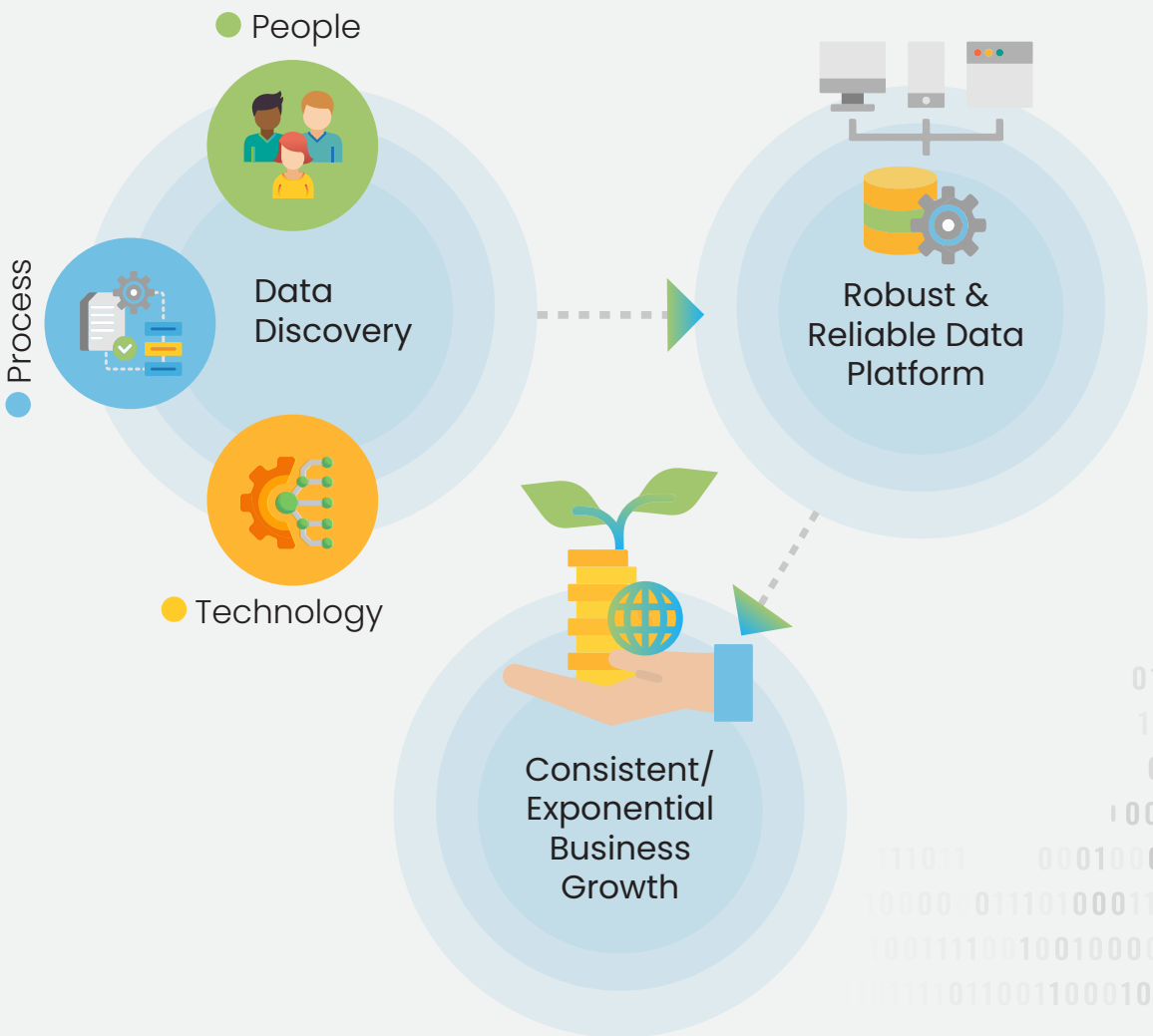
03 Technology

Set up the architecture and technology capabilities right

As the data architect envisions the next steps in a data project, they can also decide the right tools and technologies that they need to effectively control, manage, and enrich data. This can be done with automatic data cataloging. Devising a short-term road map, maybe for three to five years, for data projects ensures that organizations have a good understanding of how to build the right data capabilities that are required for the current stage of their data project and data discovery.

It also helps them to purchase the data assets that the business needs, avoiding unnecessary investments. For instance, organizations that are just embarking on their data journey may not require the same tech stack as the organizations which are in the advanced state. Hence, organizations must perform quick research and experiments to understand the requirements for the next stage of their data project, entailing data discovery.

Value Proposition of Data Discovery



Conclusion

The importance of data and analytics has paved its way across industries to derive real-time and actionable insights. Organizations embarking on data journeys and arming themselves with the right foundation are gaining a competitive edge and making better business decisions.

Every organization that wishes to tread its path to success needs to focus on its data platform in line with its digital platform. Data projects can only be successful when the data platforms are constantly revisited and tested, ensuring their robustness. Data discovery is the first step in a data project and becomes an ongoing process. Data discovery, as a culture, must be inculcated into the DNA of the organization.

The data analytics landscape is continuously evolving! Organizations are leveraging the in-depth experience of trusted IT vendors with data & analytics thought leadership to handle the operational challenges of managing their data platforms across people, process, and technologies. We at KANINI can help you move ahead with your data projects in a more efficient manner and make them successful. For more information, reach us at transformations@kanini.com.

About KANINI

KANINI Software Solutions is a digital transformation enabler, providing cutting-edge software services and solutions that help enterprises drive innovation and business growth. We specialize in ServiceNow Consultation and Implementation, Product Engineering, Data Analytics & AI, and Cloud Enablement – all delivered through flexible engagement models. Our customers trust us for accelerating their digital transformation journeys and enabling them to be future-ready.

KANINI offers a diverse range of Data Analytics services and technology solutions across Data Science, Data Engineering, Data Visualization, AI & ML. Find more about our Data Analytics services & solutions here: https://kanini.com/ai-analytics/

Contact us at (615) 465-8287 | transformations@kanini.com
USA | India | Mexico | Singapore | Ukraine | UAE | Colombia