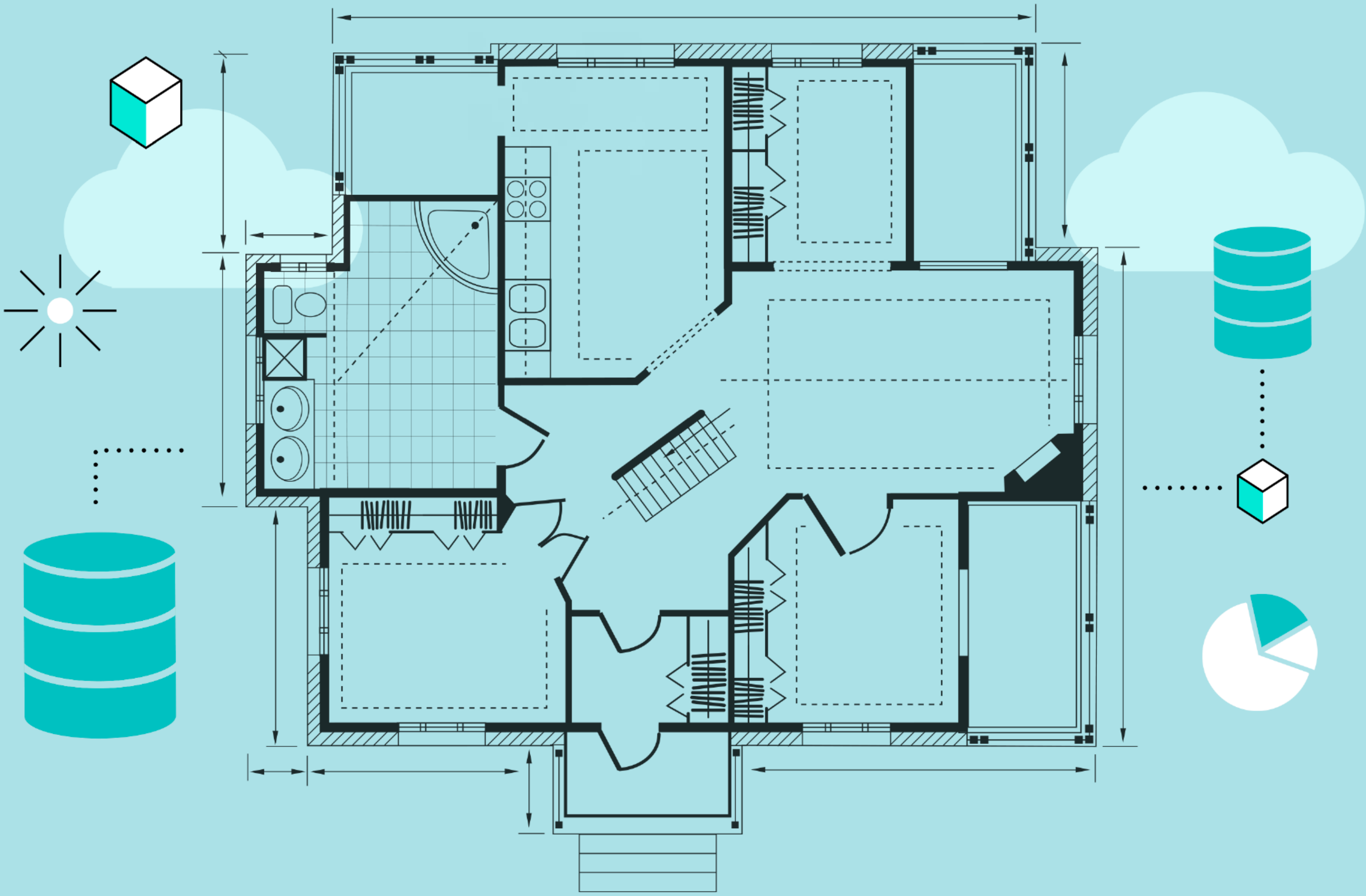
EXPLORIUM

External Data Platforms **as Part of the Modern Data Stack**



Introduction

Modernizing data architectures is top of mind for many IT, data, and analytics leaders at organizations across every industry today. Providing users with the right data, at the right time, and in the right form is a requirement to improve an organization's decision making processes.

'Logical data warehouse', 'data fabric', 'data hub', 'lakehouse', and 'data mesh' are just a few of the names for new modern data architecture paradigms that are currently being promoted as the way forward. These new architectures are focused on internal data sources across disparate data warehouses, data lakes, databases, data marts, and files located both on-premises and the cloud.

Organizations need to implement robust data architectures with the right capabilities that align with business strategy, facilitate insights around impactful use cases, and are flexible to accommodate new priorities. There is no "one-size-fits-all," and over the last few decades, data architectures have gone through several evolutions. Evolving strategies and objectives coupled with new technologies have redefined which components are required in a data architecture.

In light of new technologies, it's essential to plan for external data utilization when modernizing the data architecture within an organization, especially for analytics and machine learning use cases. Overlooking external data and its value is a missed opportunity as it provides important context not always captured in internal data: economic trends, consumer preferences, weather, reviews, social media trends, competitive intelligence, and more. Organizations that are leveraging external data in their analytics programs are outperforming their competitors and have improved customer acquisition, operational efficiencies, and risk & compliance management.

The management consulting firm McKinsey said "A well-structured plan for using external data can provide a competitive edge". Adding an external data platform as a component of your modern data stack is a competitive advantage and will boost the efficacy and performance of your analytics teams.

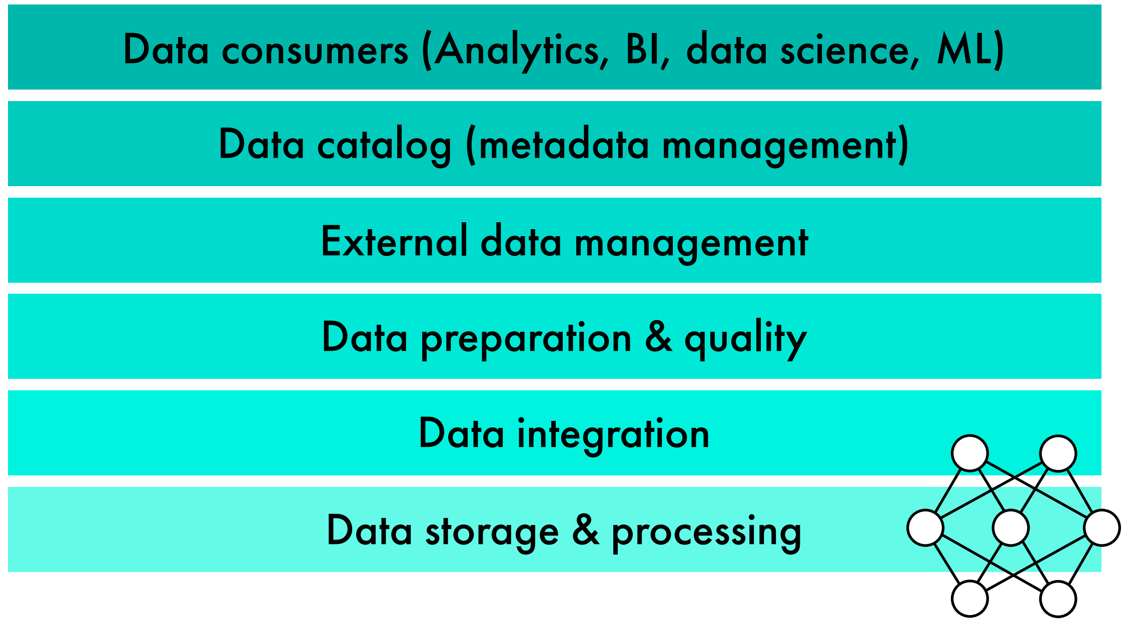
This paper will outline the core components of modern architectures and why an external data platform is a new requirement.

Table of contents

Modern Data Stack

There are many technical papers and analyst reports (Gartner, Forrester, IDC, etc.) that detail the required components of a modern data stack that you can read to better understand the specific capabilities. For the purposes of our paper we will stay high level and describe what Explorium believes are the required layers for a modern data stack. It's also important to highlight that you likely have some of these layers in operation today.

Required layers in a modern data stack


Data consumers (Analytics, BI, data science, ML)
Data catalog (metadata management)
External data management
Data preparation & quality
Data integration
Data storage & processing

Data storage & processing

This is a foundational requirement for the modern data stack, organizations need to store and process the data they collect from their various systems. There are a few ways to do this today. Data warehouses and data lakes are the two most popular but each has distinct use cases where one may be preferred over the other. They are increasingly deployed in the cloud, offering more accessible and affordable options for storing data over legacy on-premises solutions.

Popular options today:

- AWS Redshift and S3
- Azure Synapse and Data Lake Storage
- Databricks
- Google BigQuery
- Snowflake

Data integration

Data integration is a critical piece of the modern data stack. Many organizations have multiple data warehouses and data lakes, sometimes numbering in the dozens. In many cases, analysts and data scientists need data across these different sources and it needs to be integrated into a single table or view. In the past, ETL (extract, transform, and load) was the most popular style of data integration. But new technologies have been introduced that give organizations flexibility in how they integrate data across their various data warehouses and data lakes today. Data virtualization, message-oriented movement, replication, and streaming data integration have enabled organizations to integrate data in new ways and better support real-time demand for data.

Popular options:

- Denodo
- Fivetran
- Informatica
- Streamsets
- Talend

Data preparation & quality

These could be considered two different requirements but for the sake of brevity we've combined them into another key piece of the modern data stack. Preparing data is a critical step before analysis. Ensuring the data has been transformed, properly formatted, and cleansed before analysis ensures results are accurate and can be trusted.

Popular options:

- Apache Griffin
- Ataccama
- MIOSoft
- Precisely
- Trifacta

External data management

This is a newer component for the modern data stack but has become essential. External data has the power to boost your analytics and machine learning models - driving more value and ROI. By adding important context not found in your internal data you uncover competitive advantages. External data platforms are an enabling technology that automate and streamline the steps required to effectively incorporate external data into your overall data and analytics strategy.

Popular option:

- Explorium

Data catalog

Relatively new, data catalogs provide an inventory of available data assets in your organization by leveraging metadata. It helps data pros collect, organize, access, and enrich metadata to support discovery and governance.

Popular options:

- Alation
- Collibra
- data.world

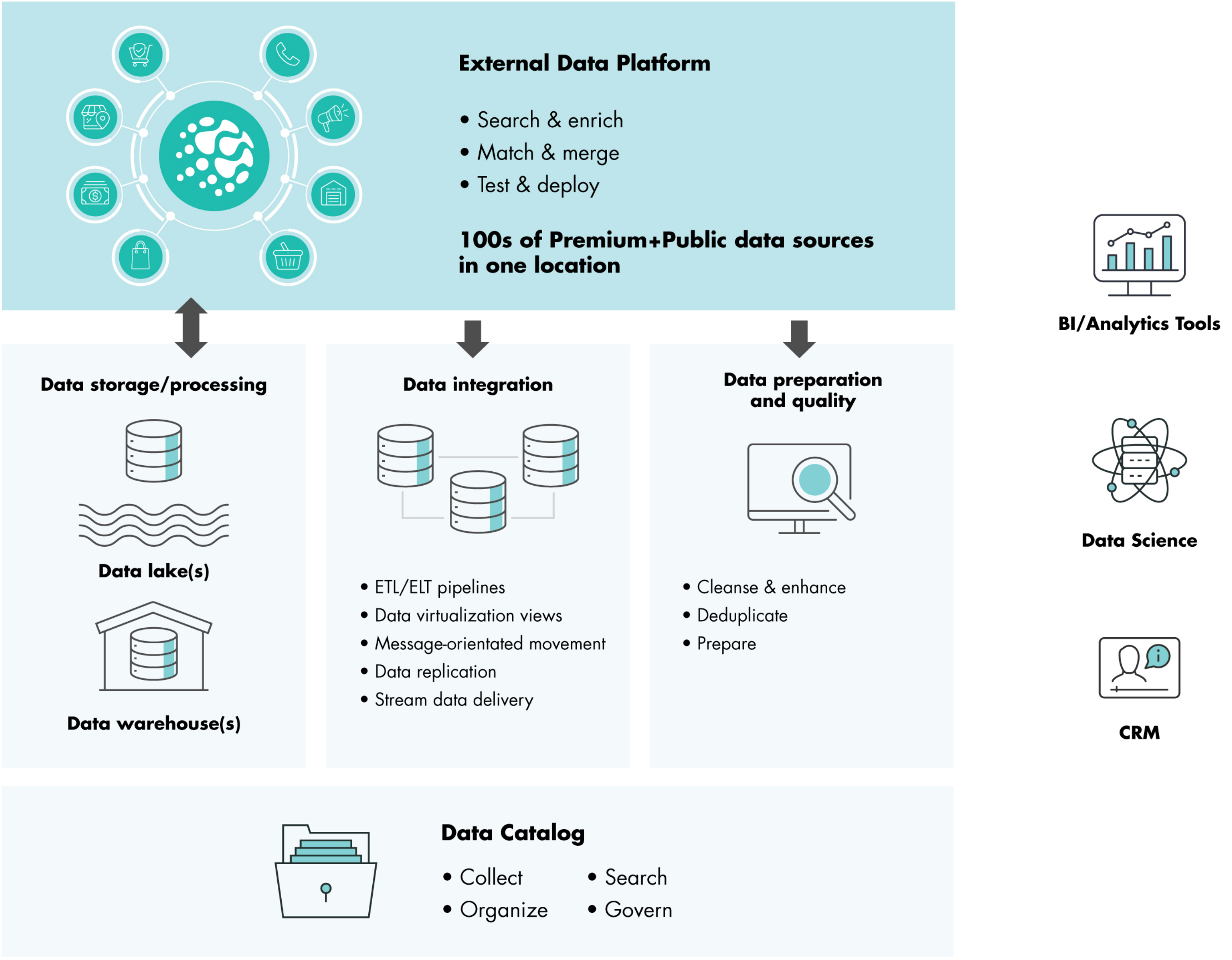
Data consumers

Once you've collected, integrated, cleansed, prepared, and enriched your data, it's time to put it to use. Visualizing and analyzing the data uncovers insights and patterns that are used to improve decision making. It also offers visibility into what is going on. This is also the layer most people in your business will interact with, whether conducting the analysis themselves or viewing the output.

Popular options:

- Apache Superset

- AWS Quicksight
- Dataiku
- Microsoft PowerBI
- Qlik
- SAS
- Tableau



Why an external data platform?

Traditional data acquisition is a lengthy process that doesn't always have analytics in mind.

With data the new competitive battleground, businesses that take advantage of their data will be the leaders; those that do not will fall behind.

But gaining this advantage requires organizations to look beyond their four walls for external data sources that bring important context and a deeper understanding of market conditions. By incorporating external data sources into analytics and ML programs, you improve their accuracy and impact on your business.

One of the biggest problems with scaling external data is that too much variety creates paralysis. Manually conceptualizing, testing, and performing analysis for each project takes too much time and leads to tunnel vision. Overwhelmed analysts and data scientists may start to overlook available data sources and data opportunities alike. Testing every data asset for every project is simply unfeasible.

There is a straightforward technical fix for this: automated external data platforms that are designed to deliver vastly improved external data discovery. External data platforms work by automating your access to thousands of pre-vetted data signals. Not only have these been curated for quality and reliability, they form a single, collective catalog eliminating the need to match and integrate each one separately.

Even better, an external data platform is far more sophisticated than just a data catalog. These don't just connect you to data sources; they will also help you find your way around them, suggesting the most relevant data points, and providing ways to automatically match and integrate them with your internal data sources.

Key Components of an External Data Platform

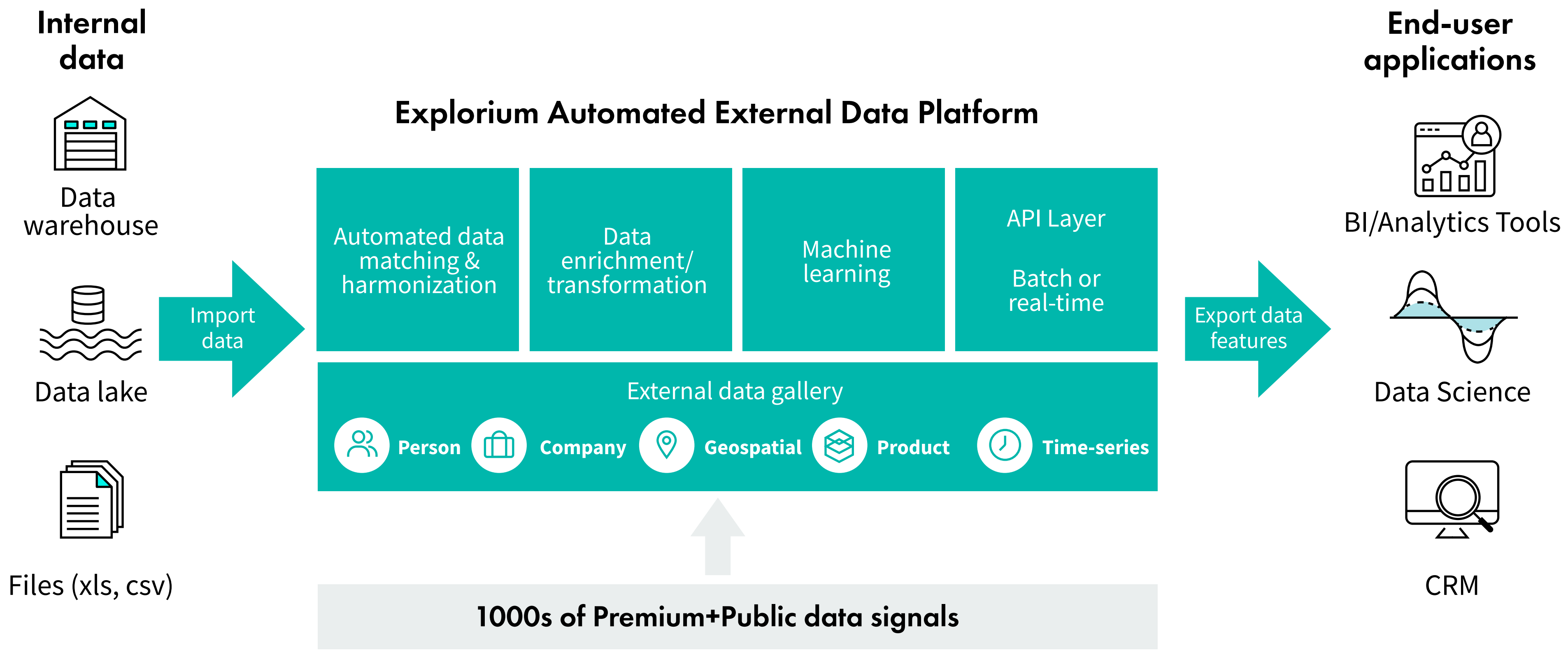
Data Access - Access tens of thousands of company, person, location and other data signals collected and curated from hundreds of data sources.

Automated data matching and harmonization - Match external data with your internal data sets and integrate them for quick consumption.

Data Enrichment and Transformation - Enrich internal datasets with new signals, build new datasets from scratch, transform the dataset for consumption, and create flows/recipes for scheduling the data consumption on an ongoing basis.

Machine Learning - Build models with a built-in data science platform that offers feature discovery and selection with auto-ML capabilities.

Data Orchestration - Integrate your internal data sources with downstream systems. API-first approach makes sure that you access the curated data sets for any analytics or machine learning needs.



Internal data

Data warehouse

Data lake

Files (xls, csv)

Import data

Explorium Automated External Data Platform

Automated data matching & harmonization

Data enrichment/ transformation

Machine learning

API Layer

Batch or real-time

External data gallery

Person · Company · Geospatial · Product · Time-series

1000s of Premium+Public data signals

Export data features

End-user applications

BI/Analytics Tools

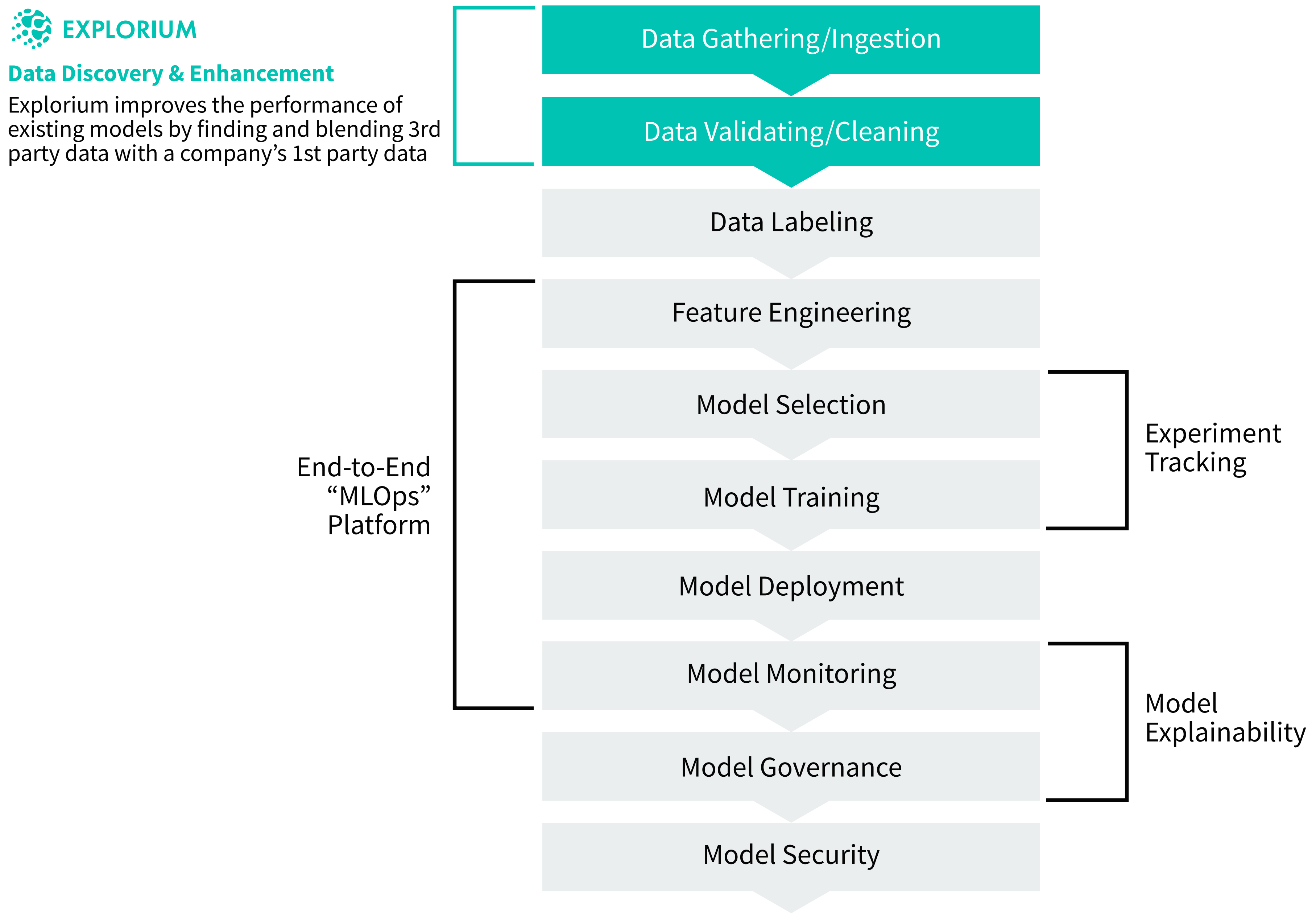
Data Science

CRM

Select an external data platform that connects your people and processes

An external data platform helps your organization to centralize the entire external data acquisition process. It may even help you to automate most of this process, including cleaning up and harmonizing datasets so that they are ready to use in your machine learning projects. However, there can be significant variation in quality from platform to platform, so before you choose one for your business, it's important to pay attention to the details.

The key is to ensure that you are opting for one that's specifically set up with machine learning and analytics in mind. Can you use it to enhance your datasets? Will it suggest useful, relevant details you may not have thought of? What about feature engineering?



Ultimately, your external data platform should remove as many of the roadblocks associated with finding and acquiring data sources as possible. It should also make it easier to streamline your data pipelines and get your model development off to a great start.

The Explorium External Data Platform is designed to deliver rich external data discovery which improves analytics and machine learning models. Explorium automates access to thousands of data signals from a wide variety of proprietary, premium, and public data sources. Not only have these pre-vetted data sources been curated for quality and reliability, they form a single, collective catalog, eliminating the need to match and integrate each one separately. The data is treated as a single resource, enabling the choice of enriching existing data sets or creating new ones.

Explorium removes roadblocks associated with finding and acquiring external data. It makes it easier to build data pipelines and leverage external data throughout your organization's analytics processes. You see the impact of data signals and evaluate uplift in your models before you deploy a new strategy. That's where the true value lies.

EXPLORIUM

About Explorium

Explorium provides the first External Data Platform to improve Analytics and Machine Learning. Explorium enables organizations to automatically discover and use thousands of relevant data signals to improve predictions and ML model performance. Explorium External Data Platform empowers data scientists and analysts to acquire and integrate third-party data efficiently, cost-effectively and in compliance with regulations. With faster, better insights from their models, organizations across fintech, insurance, consumer goods, retail and e-commerce can increase revenue, streamline operations and reduce risks.

Learn more at www.explorium.ai