Paxata®

10 Hard Questions to Make Your Choice of Self-Service Data Prep Easy

PAXATA | EBOOK

INTRODUCTION

Self-service data prep may sound easy, intuitive, and exactly what you need for that daunting data transformation and preparation project. Yet, given the vast differences among products and solutions available today, it is easy to get lost.

This guide offers ten real business scenarios that will elicit your attention, help you cut out the noise, and focus on what truly fits your needs and requirements. In each scenario, we provide real life examples and use cases so that you can envision your own environment in the description.

1 IS REAL TIME EXPLORATION OF DATA AN IMPORTANT PART OF YOUR DATA PREPARATION?

Not all data prep solutions are created equal. Some follow a workflow style where predetermined rules are applied to prepare data for analysis. This requires knowing the questions to ask of the data beforehand. For example, in a supply chain scenario, you should prescribe the tool to replace any supplier location occurrence of "Phinix" with "Phoenix". But one can quickly see the limitations of this approach. In this example, would you have considered "Finix"?

As opposed to workflow style data prep products, interactive data prep solutions allow you to explore and validate your data by seeing all of its contents – in real time.

BUSINESS SCENARIOS

Data lakes that were populated with no attention to data quality

Data with high potential of anomaly e.g. weight that could be recorded in ounces, pounds, grams, kilograms, and more

Where context is important (e.g. location data can be London, England or London, Ontario)

Data with a wide distribution of values (e.g. budget distribution or product type distribution)

Data with gaps, peaks, and outliers (e.g. an outlier year of 2107)

Onboarding new client data

2 DO YOU HAVE A LOT OF UNKNOWN DATASETS, SUCH AS THIRD-PARTY DATA AND FORM-FILLS?

When data is sourced from in-house systems of records, the user typically possesses some level of knowledge about it. However, very little is known about data that comes from external sources. For example, onboarding new client or supplier data varies in type and complexity.

The same situation occurs when data comes from form fields such as those in survey software, marketing automation, logistics and scheduling apps, ERP, and others. Data in these situations typically have a wide variety of misspellings.

Similarly, in application migrations or consolidation of legacy systems that were created years ago, one lacks full knowledge of the information architecture.

In order to better understand the data context and semantics, one must think about exploring, profiling, and analyzing the data values and content before integrating or migrating it. Built-in, smart algorithms that can detect semantic and syntactic context of the data, potential joins and different spellings, can accelerate the time to value in these scenarios.

BUSINESS SCENARIOS

Onboarding and blending second- and third-party data with first-party data

Curating external sources of data to create a data product

Application migration or application integration

Data quality and validation of free form text fields (e.g. address cleanup for marketing campaigns)

3 WOULD INACCURATE DATA JEOPARDIZE YOUR REPUTATION OR REVENUE?

Some data preparation tools often limit the user to a small sample of data. In this case, one is left to hope that all of the possible anomalies and outliers are included in the small, allocated sample of data. While this is a viable solution in some use cases, such as prototyping or requirement gathering that will subsequently be validated against the entire body of data, it is a not a solution that is suitable for sensitive situations.

In scenarios where the accuracy of data compromises the outcome, the data prep solution must provide the flexibility to choose the sample size and be able to explore and prepare the entire data. For example, in financial crimes compliance reporting, failure to detect fraudulent transactions which happen to be outliers and are therefore not included in the data sample can severely damage the financial institution's reputation.

BUSINESS SCENARIOS

When the full body of data is needed to ensure accuracy (e.g., financial crimes compliance or clinical trials)

When data is your product (e.g. information market places, information as a service, or product catalogs)

4 IS GOVERNANCE A KEY PIECE OF YOUR DATA PREPARATION AND REPORTING?

In many cases, a data prep project leads to downstream reporting and analytics that are used across executive and multi-functional teams, or externally by government bodies. In these cases, factors such as how the data is sourced, the transformations that it has gone through, and usage analytics on who is consuming it are all important elements in gaining trust and complying with regulations.

Data prep solutions that self-document every user action and each machine-learning operation applied to the data ensure complete auditability.

The key is to provide end-to-end traceability of data. For example, in cases where the information from the data preparation ends up in downstream systems such as business intelligence applications or public portals, it is vital to show a full lineage.

BUSINESS SCENARIOS

Weekly sales and marketing reports

Executive dashboards

Evidence-based medicine in healthcare

Compliance reporting, such as anti-money laundering

5 IS VERSIONING AND HAVING A SNAPSHOT OF YOUR DATA PREP PROJECTS AND DATASETS CRITICAL?

In some cases, data prep projects are a one-time event. However, you will certainly accumulate new data, expand information presented in a business dashboard, extend source systems, or need to record snapshots of your prepared data for point-in-time analysis.

A basic example is when one set of data is prepared for bookings before it is augmented with another dataset to conduct bookings versus revenue analysis. In this scenario, the user may want to store and version "bookings" before blending it with other data sets and storing it as "revenue versus booking."

Having a complete history of a data preparation project at any given point is not a capability offered in all data prep solutions. Only those with a database backend equipped to store a snapshot of data in time are able to achieve this.

BUSINESS SCENARIOS

Regulatory and audit-heavy programs

Data quality monitoring

Single source to multi-source progression of data blending

Daily / weekly / quarterly reporting

6 DO YOU HAVE A HIGH RATIO OF BUSINESS SMES TO DATA ENGINEERS?

Historically, data preparation has been within the purview of IT teams. This is partly a legacy issue, as the tools and techniques that were created in the 1990s were created for developers with technical skills. Today, this paradigm is shifting.

The reality is that for every data engineer or technical resource in a given organization, there are potentially 1000+ business analysts, all of whom want more information to perform their jobs better. Therefore, treating data preparation as an IT task inherently creates a bottleneck.

The context of the data remains within the line of business. For instance, a supply chain manager would be aware that “Govis Pharmaceuticals” and “Novis Pharmaceuticals” are in fact the same vendor, and the different entries are purely recording mistakes. IT would not have knowledge of that context.

So, for scenarios where business knowledge is critical in preparing data, or scenarios where the number of analysts outweighs the number of technical resources, a business-oriented data preparation tool makes the most sense.

BUSINESS SCENARIOS

Weekly sales reporting

Supply chain / inventory management

IoT device usage data analysis

Call center data prep

7 DO YOU NEED TO CREATE GOVERNED BUSINESS USER SANDBOXES FOR YOUR DATA LAKE?

In addition to scenarios where a business team wants to take a hands-on approach to its data preparation as described in section 6, there are scenarios where IT desires to provide a governed or contained environment for its business users.

This could be a data lake type of scenario. In order to unlock the value of this data, IT needs to section off parts of the lake to business teams.

Unfortunately, data lakes provide limited data exploration capabilities for business users. They require familiarity with SQL or programming languages. Additionally, the performance of data lakes is often poor, and querying directly from the lake creates latency and a less-than-ideal user experience.

A sophisticated data prep solution that provides an easy-to-use interface for business users to interact with data – ideally, at the speed of thought – is critical to improving the data lake's business value.

BUSINESS SCENARIOS

Data lake exploration

Business use case development and prototyping

Ad hoc analysis of product usage (e.g. IoT device usage)

8 ARE YOU OPERATING IN A MULTI-REGION, MULTI-DEPARTMENT, OR MULTI-CLIENT ENVIRONMENT?

Enterprise-ready data prep solutions have advanced multi-tenancy capabilities created to meet the needs of diverse business entities and their SLAs.

For example, a global sales and marketing organization might need to create individual tenants for data prep projects across corporate and regional teams. While each tenant has full separation of data and functional privileges (e.g., creating a project versus only viewing it), some individuals may need to access multiple tenants.

Reverse the scenarios. There might be a variety of types and sources of authentication servers. For example, a US team may use a series of LDAP servers, whereas the UK team has its own separate, albeit SAML, authentication framework, but all teams across the globe must access a shared tenant.

A basic data prep tool cannot serve these many layers of complexity of authentication and authorization, while an enterprise one can.

BUSINESS SCENARIOS

Marketing operations accessing sales and support tenants for insights

All individuals across various LDAPs / SAML authentication servers accessing a common tenant (e.g. AML tenant)

Never-expiring service accounts to access tenants

Consultants and OEMs who provide data prep services to their customers accessing all of their customer tenants using a single ID and password

9 DO YOU HAVE A MULTI-CLOUD STRATEGY?

Many companies today are avoiding the vendor lock-in situations that occurred a couple of decades ago with the titans of the enterprise software industry. These companies are considering a hybrid environment – using a mixture of various cloud environments and in-house systems.

Taking an agnostic approach to data prep software that can run across multiple cloud and on-premises environments not only reduces vendor lock-in, it also increases the interoperability that is required for fast migrations of workloads from one environment to the other.

It ensures that data prep can live closer to the data gravity where it is at rest, and also enables integration and movement of data across cloud and on-premises sources.

BUSINESS SCENARIOS

Regional / localized cloud services investments

Hybrid cloud and on-premises environments to segregate sensitive and non-sensitive data

Choosing the right cloud infrastructure provider for right application or workload

10 DO YOU HAVE A VARIETY OF STAKEHOLDERS WHO WANT TO PARTICIPATE IN DATA PREPARATION PROJECTS?

Today, data projects are no longer in the hands of one developer or one technical resource. A data project is often conducted by multiple parties, including analysts, decision-makers, reporting and analytics teams, data science groups, and others.

Therefore, the traditional approach of creating data projects and publishing results for others to see and provide feedback to, slows the process; it is a legacy, waterfall approach.

A modern data prep application provides a real time, inline, multi-user experience, with simultaneous editing and immediate feedback for all parties involved to be part of the preparation project. As matter of fact, the style of collaboration should be similar to that leveraged in Google sheets or Google docs types of applications.

Workflow style types of data prep solutions don't offer such a fluid experience. If multi-user collaboration is a critical requirement for you, look for modern data prep solutions that offer this seamless user experience.

BUSINESS SCENARIOS

Cross-team projects (e.g. sales, support, and marketing)

Data preparation projects with large groups of analysts

Real time business and IT collaborations

Information as a service projects and exchanges between a vendor who is preparing data for its clients

CLOSING THOUGHTS

Choosing the right data prep solution is not easy. While on the surface everything sounds the same, you now know that not all tools are created equal. You need to bear in mind your business scenarios, types of users, and internal and external requirements before selecting the right data prep solution that is tailored to your organization's specific needs.

Companies around the globe rely on Paxata to get smart about information. Paxata is the pioneer that intelligently empowers all business consumers to transform raw data into ready information, instantly and automatically, with an enterprise-grade, self-service data preparation application and machine learning platform. Our Adaptive Information Platform weaves data into an information fabric from any source and any cloud to create trusted insights. Business consumers use clicks, not code to achieve results in minutes, not months. With Paxata, Be an Information Inspired Business.

Paxata is headquartered in Redwood City, California with offices in New York, Ohio, Washington D.C., and Singapore.

Paxata Headquarters 305 Walnut Street Redwood City, CA 94063 1-855-9-PAXATA paxata.com