Self-Service Data Prep

Moving to a Connected Prep Model

❋ birst

CONTENTS

OVERVIEW

Modern business intelligence (BI) and analytics platforms have transformed the way people work with data. They've made it easier for business analysts to visually explore data and discover insights to help their organizations make better decisions. But despite these advancements, there remains a large section of the business population underserved by existing analytics solutions. This large majority of business people lack the ability to work with data on their own, without support from an expert.

There are a number of reasons for this problem:

- BI and analytics solutions available today are not designed for the majority of business people. Even modern data visualization products, known for their ease of use, are more appropriate for data-savvy analysts and data scientists.

- While these solutions have simplified the process of exploring data visually, they still do not adequately enable non-technical users to assemble (i.e. combine, cleanse, and transform) a coherent dataset from multiple disparate sources of data.

- Conventional self-service data prep tools encourage the creation of data silos, which undermine the decision-making process and force people to work in a vacuum, completely disconnected from the rest of the organization.

Data fuels the critical business decisions an organization needs to make. Without the ability to connect to, integrate, cleanse and transform various data sources, self-service analytics remains limited to a specialized audience, leaving the rest of the organization relying on the knowledge of experts. Empowering business users with self-service data preparation capabilities is a critical step for data-driven organizations looking to extend the power of analytics to a larger audience.

In this paper, we will look at the need for self-service data preparation, the progression of self-service data prep from an individual model to a connected model, and the main use cases for self-service data prep.

THE CASE FOR SELF-SERVICE DATA PREP

Organizations today are becoming more and more data driven. Executives and managers understand that employees empowered with the ability to assemble and analyze data can identify opportunities to increase revenue, reduce costs, improve services, retain customers and, ultimately, achieve a competitive advantage. With the rise of self-service analytics, organizations are looking to extend and expand analytics to more users and functional departments. One problem is that the growing number of data sources decision makers need to access, along with their complexity, is hindering the expansion of analytics.

With data being generated and residing in so many different systems, users find that traditional data sources – typically managed by the IT/BI organization – like a data warehouse or an ERP application are no longer enough to answer questions about business processes that span multiple functions. Today data can reside in cloud-based applications (e.g. Salesforce, Marketo, NetSuite), 3rd party data from

martech vendors, social media, partners, publicly available data and more. For instance, understanding lead-to-cash ratio requires accessing and combining data from three different systems: marketing automation, CRM, and ERP. Bringing this data together traditionally means that business users must rely heavily on their organization's IT/BI staff.

It is the IT/BI organization where one will typically find the skills required for data preparation – the process of integrating, cleansing and transforming data. These skills include an understanding of fundamental concepts around information management – such as tables, keys, relationships, and joins –, familiarity with dimensional modeling and big data platforms, as well as knowledge of database querying languages like SQL and programming skills such as Python and Perl.

However, as the volume and complexity of data grows, data preparation processes are getting longer, more complex, and demand too much from IT. End users find themselves waiting days, weeks and even longer to receive integrated datasets from IT. This leaves users working with out-of-date data – or no data at all – and can result in mistakes and poor decisions. Self-service data preparation can reduce the burden on the IT/BI organization and empower end users with the ability to access and prepare the data they need when they need it.

Organizations embracing self-service data prep tools are reaping the benefits and gaining a competitive advantage over their peers.

More Data Sources - Self-service data prep simplifies the work of integrating and preparing data and allows users to connect to and work with new types of data sources. Users can easily integrate these new sources with IT controlled data sources to arrive at new insights. This gives organizations a more complete understanding of the business.

Reduced Time to Business Insight - Rather than waiting weeks or months for IT, self-service data preparation tools significantly reduce the time it takes for users to get analytic-ready data faster, and generate valuable insights in less time. In turn this improves strategic, operational, and financial decisions.

Self-Sufficiency - Self-service data prep expands the use of analytics by making end users more self-sufficient. By being able to integrate and prepare data for analysis on the fly, business users are able to react to changing business conditions more quickly and increase data-driven decision making.

Knowledge Transfer - Decision makers and business analysts can enrich their own insights as well as those of their colleagues. As end users connect to and analyze more data sources, their prepared data and the insights generated from it can be leveraged across the business. This increases collaborative analytics and harnesses the collective intelligence of the organization.

Yet, the vast majority of organizations using self-service data prep tools today have not moved beyond the individual data prep and consumption model. They are seeing the benefits of self-sufficiency and reduced time to insights across many more data sources, but are struggling to bring together the insights generated across their organization in a manner that is connected and doesn't contribute to the creation of data silos.

ADVANCING FROM INDIVIDUAL SELF-SERVICE DATA PREP TO CONNECTED SELF-SERVICE DATA PREP

IT Led Data Prep

Traditionally, organizations had IT departments that were responsible for data preparation. IT was tasked with integrating, cleansing, and transforming data into suitable reports. Self-service was confined to users selecting pre-built reports from a predefined catalog.

As self-service analytic capabilities improved, users obtained the freedom to change and extend the reports themselves. But they were still limited by what and how much data was made available to them by IT. Users could not easily add new data to their reports and resorted to extracting data to Excel and manually combining the data into one large spreadsheet.

Data Prep for the Data Savvy

The emergence of visual data discovery tools extended users' ability to avoid the pain of spreadsheets and requesting reports from central IT teams. However, users still relied on having the right data to work with in the first place.

Visual data discovery tools gave business users even more capabilities to analyze data on their own, yet these same users still had to rely on specialized IT/BI teams and data savvy analysts to prepare the data for them. Today some data discovery products include built-in data prep capabilities, but they require specialized knowledge of dimensional modeling, SQL, and understanding of the structures of the data sources and how to transform it for proper integration.

Individual Self-Service Data Prep

Today, data preparation is becoming truly self-service. Data prep capabilities have matured enabling non-technical users to explore new data sources by automating as much of the data prep work as possible. Many self-service data prep technologies use machine learning, natural language processing, and other advanced techniques to guide and assist users with data preparation. Graphical user interface provide users with drag and drop capabilities to integrate data sources while algorithms analyze the data structures and proposes specific columns to join the data. With these advancements in self-service data prep, users can now take any data set, prepare it, and then use it for analysis.

Yet as great as self-service data prep is today, there is one major flaw. In the current approach to self-service data prep, the finished product is a physical extract of data. Working with data extracts presents several drawbacks that include additional maintenance burden, as well as security and governance risks. Just as importantly, data extracts increase the possibility of people working with conflicting definitions of the data and contribute to the proliferation of data silos.

Connected Self-Service Data Prep

Companies are starting to rethink the process of self-service data prep and have begun to pursue a more connected model. In this new model, the output of the data prep process is no longer a data silo. Instead, prepared data is connected to the larger organization. End users are able to share their data

prep process and subsequent datasets with others in the organization and vice versa. The data prep process for new data sources and the analytics become fully integrated with the larger organizational data ecosystem, allowing business users to pick and choose from the data sources they need. This leads to the enrichment of data, reduction in time to insight, increased productivity, and knowledge transfer between everyone in the organization.

USE CASES FOR SELF-SERVICE DATA PREP

Big Data

Organizations today are using platforms such as Hadoop to create data lakes as an operational data store. This allows them to collect, store and process massive amounts of data without having to structure the data into a schema.

There is a treasure trove of valuable information within these data lakes. By combining this data with centrally managed IT data sources (data warehouse, ERP, etc) as well as external data sources, organizations can form a complete picture of their operations, manage their supply chain from procurement to fulfillment, and so on. The possibilities are endless.

While data lakes can be made available for analytics, the consistency and quality of this unstructured data is suspect. Self-service data prep tools work on top of big data platforms such as Hadoop to help users find the right data, cleanse, transform and standardize this data, and integrate it with other data sources for analysis.

Reusable Data

With most self-service data prep tools, the data preparation process done by individual users is a one-off event that has to be repeated when there's a new requirement. Additionally, there is another pitfall with many self-service data prep tools: the need to retrace your steps.

Imagine a meeting where everyone is staring at two reports from two different departments. Reviewing the same metric, each department has a different value. This leads to arguments over which report is correct, where the data came from, how was it calculated, and so forth. Understanding how the data was prepared suddenly becomes critically important.

Self-service data prep tools that track each step of the workflow and enable sharing of these reusable workflows with the organization help promote greater trust in the information.

Reusable data prep allows end users to:

- Avoid repeating the entire data prep process when new requirements arise. Simply viewing and editing the process where needed should be sufficient.

- Retrace the data prep steps to clearly demonstrate the choices that were made along the way.

- Standardize and record data prep steps to share and spread amongst the organization to enable collaboration and consistency.

External Data: Public & Third Party Data

With the rise of self-service analytics, data hungry business users are looking beyond their company's data sources for insights. Government census data could be used to pick the next store location. Data from martech vendors is needed to justify marketing spend. Weather data can improve shipping schedules. The amount of insight that can be generated from external data sources is vast. Yet, these insights are merely directional. More strategic insights can be uncovered by integrating these data sources with internal data sources.

Government census data combined with store data (profitability, visitor traffic, surrounding attractions, etc) can pinpoint the best next location and reduce risk. Data from martech vendors together with sales and lead data show which vendor is bringing in the best ROI and not just the most leads. Weather data added to warehouse inventory data and sell-through data can save on expedited shipping costs. With self-service data preparation tools, business users can not only access external data sources but also integrate with internal data sources to produce strategic insights.

Partner Data

Self-service data prep shouldn't just be used to extend the insights within one's organization. In the modern global economy, organizations are working with hundreds of partners to reach and satisfy customers worldwide.

Partners come in all shapes and sizes. Distributors and value added resellers are in charge of sell-through in territories where organizations aren't well established. Raw goods suppliers and OEM partners are helping the organization produce goods for the end customer. Marketing agencies are helping organizations reach and connect to the right audience. And there's many more partners.

What all these partners have is data that is specific to the respective organization they are working with. By giving self-service data prep capabilities to end users, those working hand-in-hand with partners are able to access and combine partner data with internal company data and make strategic decisions that make their organizations more efficient.

Not only that, self-service data prep capabilities and internal datasets can be shared with partners. This enables them to do their own analysis to better service the needs of their partner organization. With self-service data prep, insights can be extended to an organizations network of partners.

Embedded Self-Service Data Prep

How do you make an enterprise business application sticky? You infuse it with value, offer a rich and intuitive user interface, and encourage information sharing and self-service.

By embedding self-service data prep capabilities into your business application, organizations can enable their customers to combine their own data with the information and analytics you provide. Sharing of data prep processes and the insights from their own analysis helps customers to create a network effect - connecting each individual's analysis with others.

With embedded self-service data prep, companies can provide their customers and external users:

Data Integration - Users can integrate their data sets with the centralized application data analytics that you provide.

Data Enrichment and Visualization - Business users can create and share their own business logic with the existing data, or enrich, transform, and extend the data to create powerful visualizations.

Data Collaboration- Each user can network their personalized data blend with others to collaborate and make decisions around consistent and trusted data.

Connecting It All Together

Self-service data prep doesn't just allow users to prep and analyze external and or unmanaged data sources. The key to success is in being connected.

In all these use cases, self-service data prep brings together IT controlled data, external data, and user managed edge data. Users are not only able to form a complete data picture to help answer critical and time sensitive questions, but also be connected to the prep processes and insights generated by their colleagues. This increases productivity (much less re-work or duplicative prep), creates trust throughout the organization, and expands the collective intelligence of all users as they enrich their own insights.

BIRST CONNECTED DATA PREP

Birst's Connected Data Prep enables non-technical business users to access and prepare data with a user-friendly, visual experience that eliminates the need for complicated scripting. With Connected Data Prep, users can network their analytics with data from colleagues, other departments, or the IT organization – leading to enriched insights for smarter, more trusted decisions.

Birst Connected Data Prep has simplified the ETL (extract, transform and load) process for the modern day business user.

Connect - Access data from any data source with one click data extraction and get recommendations of what columns to use from sources. Birst employs modern RESTful connectors making the experience fast, reliable and reusable.

Prepare - Cleanse, merge and refine data using an intuitive, visual experience that instantly shows how data is changing with each transformation applied. Drag-and-drop transformations eliminate the need for complicated scripting, and graphical data lineage ensures users can trust what the data means and where it comes from. Prep history is saved to make the process reusable and sharable throughout the organization.

Relate - Birst creates a network of analytics that connects every part of the organization. Connected Data Prep allows users to take their prepared data and add it to the network, sharing their insights with colleagues.

Birst Connected Data Prep:

- One-click data extraction using smart connectors for a faster, reliable and repeatable experience.

- Intelligent join recommendations for faster, deeper insights.

- Visual data prep shows how data is being transformed in real time.

- Drag and drop transformations without the need for scripting.

- Data and transformation lineage gives clarity into where analysis comes from.

CONCLUSION

Data fuels the critical business decisions an organization needs to make. With the proliferation of data sources and the rise of self-service analytics, companies are looking to self-service data prep to extend and expand analytics to more users and functional departments.

Whether it is big data, public data, external data, partner data, or embedded analytics, self-service data prep is helping business users connect to and integrate data sources as well as cleanse and transform data. Organizations embracing self-service data prep tools are enabling business users to react to changing business conditions and reduce the time to business insights.

As self-service data prep technology continues to advance, look for tools that connect the whole organization through information. Connected self-service data prep capabilities enable business users to share and reuse data prep routines and resulting datasets with others in the organization via a network of trusted analytics. This network of analytics can touch every part of the organization, injecting intelligence into every business process by empowering end users to enrich their own insights as well as those of their colleagues. Ultimately this results in more collaborative analytics and unlocks the collective intelligence of the organization.