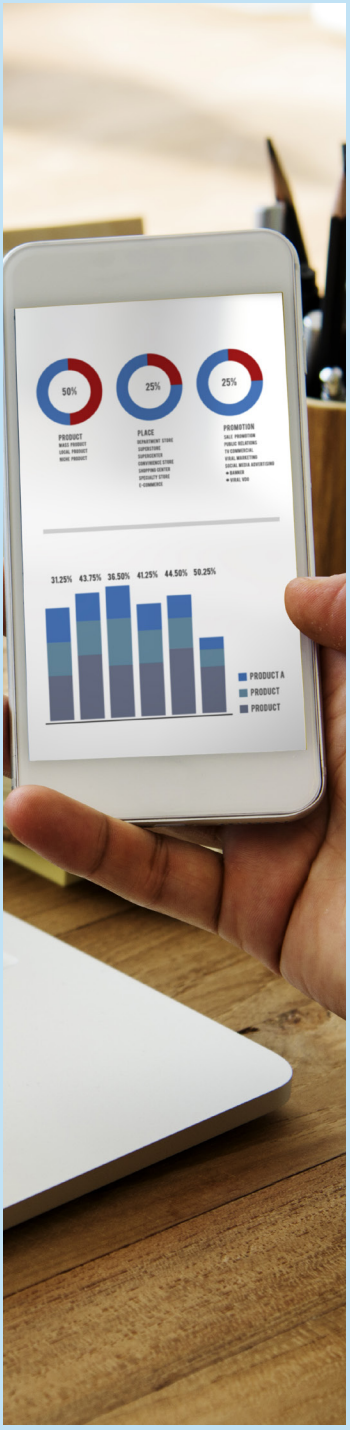
VIEW POINT



SELF-SERVICE FOR DATA PREPARATION AND ADVANCED ANALYTICS OVERVIEW

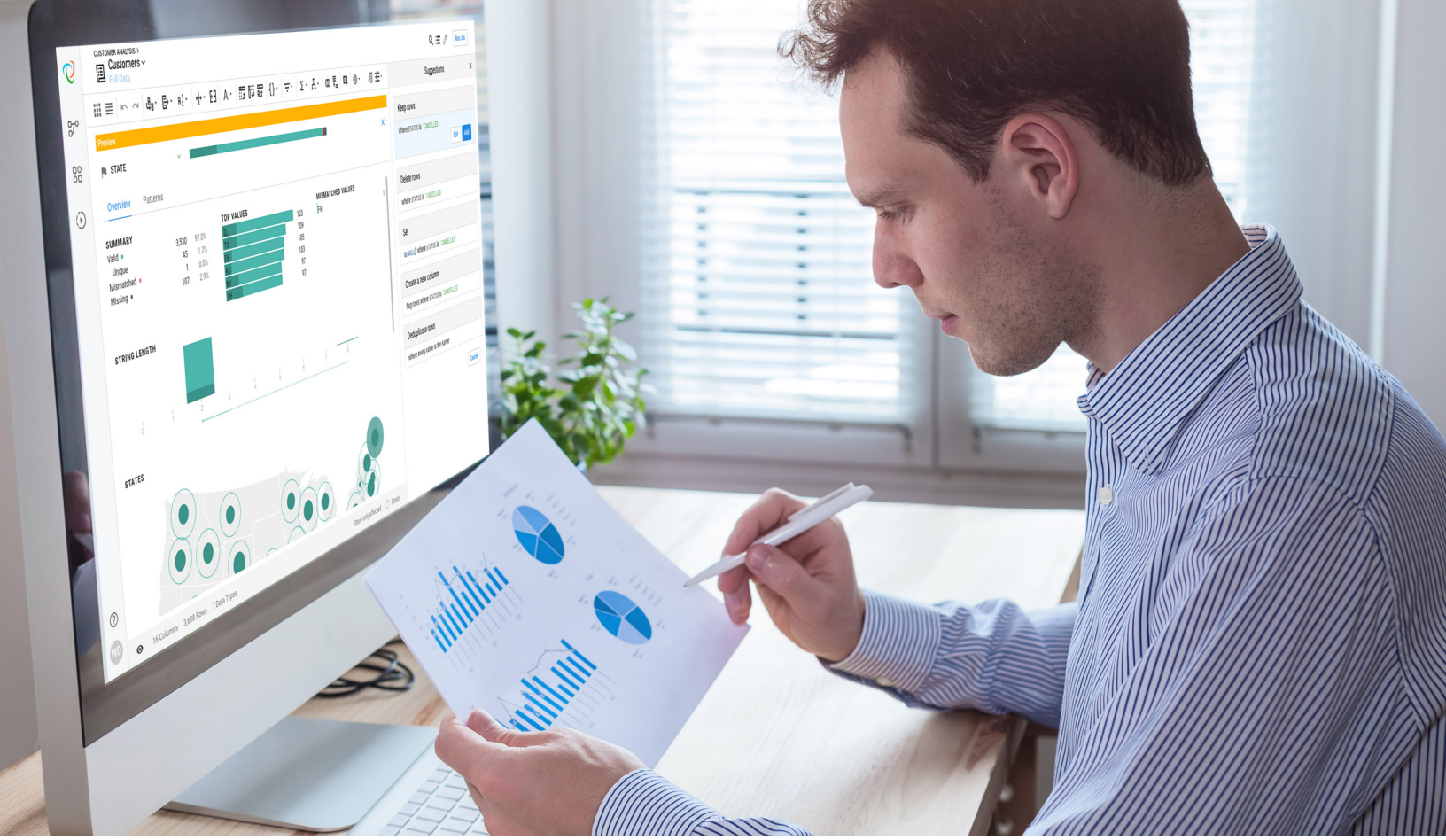
Abstract

Infosys sees organizations struggling with data challenges on a daily basis. We provide advice and implement solutions to help them become more efficient and innovative with their data-driven initiatives. It's our point of view that data preparation must be a self-service business operation. The people who know the data best must be able to prepare data themselves, without needing to rely on IT or other technical resources to do it for them.

The Infosys analytics practice recommends self-service data preparation for modernizing analytics platforms and has seen substantial benefits through its adoption. Infosys partners with Trifacta whose data wrangling platform helps organizations accelerate the process of getting data ready to use.

TRIFACTA®

Infosys®
Navigate your next



The Evolution of Data Preparation

Data preparation has long been known to be some of the most time-consuming, inefficient and painful work in data. It's estimated that data preparation consumes more than 80 percent of any data project. Industry analysts now recognize data preparation as its own unique category of data service. Dresner initiated research in 2014, a Gartner market guide followed shortly thereafter, and Forrester completed the first major data preparation vendor evaluation in 2017.

Preparing data for analysis, for machine learning or for use in a business application is nothing new. But the ways in which data preparation takes place, and the people responsible for data preparation, have changed dramatically in the last few years.

The Traditional Domain of IT

Data preparation was traditionally the exclusive domain of the technical experts in IT. Solutions were initially hand-coded, which was cumbersome and time-consuming. ETL (Extract-Transform-Load) tools sped things up. But neither hand-coding nor ETL tools bridged the gap between IT and the business. They still relied on the limited resources of IT, the only group with the adequate technical skills to prepare data for a specific purpose. What was missing was the context for the data that only business leaders could provide. Context got lost in translation. Additional costs piled up. Delivery delays slowed things down.

The Need for Business Context

As organizations took on more data-driven initiatives, business leaders grew impatient with the IT bottleneck. Some business leaders started doing data preparation work themselves using the one tool they knew best—Excel. But this manual solution introduced approximations and errors in the data, and the outcomes were neither repeatable nor scalable. Other business leaders effectively created shadow IT teams. Work got done, but not without circumventing enterprise IT processes and protocols and putting other mission-critical projects at risk.

The process need to be flipped on its head.

Inverting the Paradigm to Empower the Business

Self-service data preparation reorients the responsibility of preparing data for analysis, for machine learning or for use in a business application toward the individuals with the greatest business context.

Business users must be able to see the impact of every data preparation step as they work so they can validate each step in their process as they're building it, instead of waiting for a job to fully run before they can check results.

Trifacta Data Wrangling Platform

Trifacta data wrangling platform empowers data analysts in business teams and citizen data scientists to access, explore and prepare diverse data themselves without having to go through IT's long and frustrating software development life cycle. They enable more end users to engage in the data preparation process and more data to be used in analysis. As a result, organizations can drive more value from more data, more quickly and efficiently.

Visual + Intuitive

Trifacta data wrangling platform provides an interactive workflow that starts with a visual representation of the data that's common to business users. Not only is the data preparation process visual, it's interactive.

Trifacta spent decades researching and designing intuitive data preparation interface and workflow that makes the work smarter, using machine learning (ML), and more efficient for data workers. Clicking and dragging from a variety of guided menus lets users develop transformations in far less time than building complicated regex statements or formulas. This drag-and-drop approach is more intuitive and efficient than traditional Excel or ETL-based approaches.

Machine Learning-Enabled

Trifacta uses machine learning throughout every step of a data preparation workflow and within the interface to provide best-fit recommendations. Every interaction kicks off ML-driven suggestions that guide users through the process of cleaning and preparing their data. Users have then to accept or revise Trifacta ML suggestions to clean and structure data.

Like any machine learning system, the recommendations provided improve with learning from many data points. With more than 50,000 users in 143 countries using Trifacta Wrangler, its free product, Trifacta has assembled an enormous amount of anonymized and secured data points to make its ML algorithms effective. No other data preparation vendor comes close to embedding the same amount of training data in its products, which is used to constantly improve the suggestions provided to users.

This new intelligent approach to data preparation accelerates analytics processes by up to 90 percent, enabling organizations to focus on creating highly accurate predictive data models, increasing human productivity, accelerating decision-making and delivering tangible competitive advantages.

Aligning Enterprise IT with the Business

While the goal of self-service data preparation is to empower business users and make them more efficient, enterprise IT policies in security, governance, auditability or scalability can't be sacrificed in the process. IT and business groups must be aligned around both agility and governance.

The Trifacta data wrangling platform seamlessly ties into an organization's existing data security and access control framework. There's no separate security framework to manage. The platform tracks and provides transparent lineage of every preparation workflow users develop in the product. The IT team has visibility into everything users are developing within the platform. This information also integrates into any data catalog to provide centralized governance, lineage and audit control.

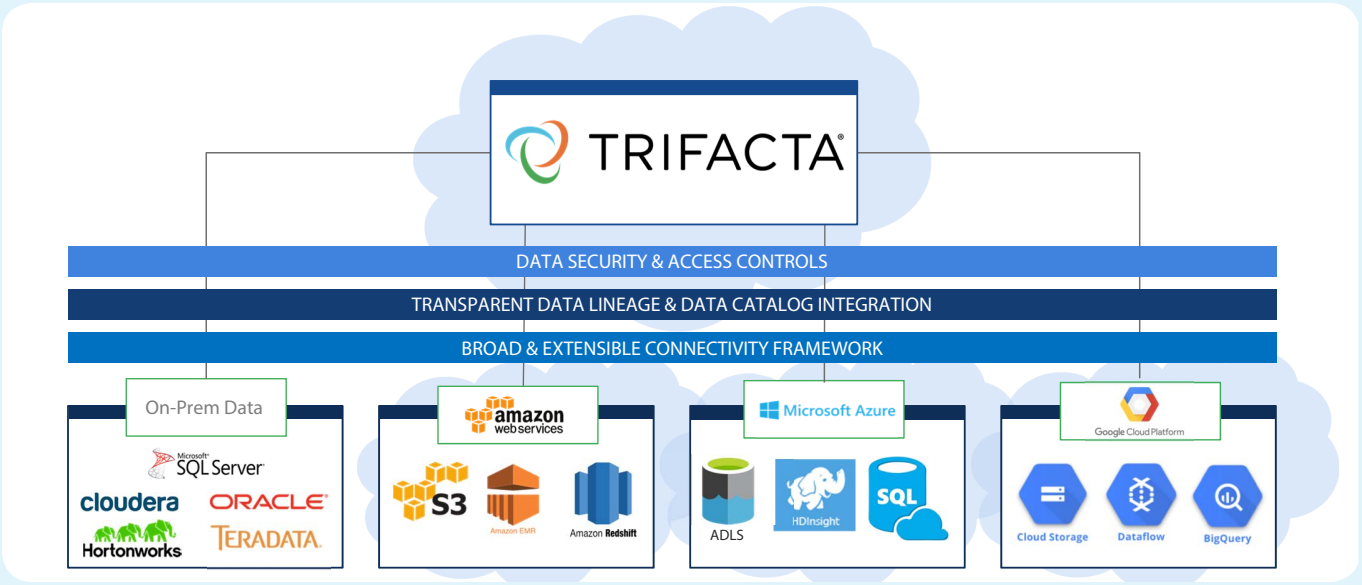
Minimizing Risk

Data proliferation is a nightmare for any organization, both in terms of management costs and security risks. Any data preparation solution should be independent of underlying processing platforms to future-proof investments in a fast-changing technology landscape.

The Trifacta data wrangling platform connects diverse data sources across on-prem and cloud environments and pushes down the processing where the data resides, instead of forcing on data movement. Data preparation work is abstracted from the underlying computation of the transformations being developed, minimizing risk.

Embracing the Hybrid Cloud

We've reached an inflection point in technology where business applications are cloud-first. The same trend is now happening for analytics. Organizations must be ready to embrace cloud and hybrid analytics with big data in mind. Any self-service data preparation solution should support cloud and on-premise deployment with the interoperability to let any organization decide where the data and the processing can be located. Trifacta data wrangling solutions embrace the hybrid cloud. They can run on premise, supporting major Hadoop distributions to process data. But they can also run on AWS, Microsoft Azure or Google, with their specialized versions of big data processing platforms such as EMR, HDInsight or BigQuery.



Google-Approved

Trifacta is the only embedded Google Dataprep solution offered by the Google Cloud Platform. Trifacta solution was subjected to Google's vigorous vetting for scalability, security and usage simplicity, and their selection for the Google Cloud Platform is a testament to the value Google sees in Trifacta. The cloud version of Trifacta Wrangler gives everyone the opportunity to select the approach that fits best their needs.

The Infosys Analytics Workbench

If organizations are going to scale up their use of statistical analysis, impediments must be removed. IT cannot keep data analysts from creating and iterating analytical models, and data scientists and analysts cannot keep business users from reusing analytical models.

The Infosys Analytics Workbench was designed to remove these impediments.

Data analysts have long wished for a platform that handles data exploration, data wrangling and modeling, and publication and consumption of models. There are many tools for wrangling and modeling, but most of these do not suggest data changes to the analyst as they work, nor do they allow for smart data searches within big data. Many of these tools restrict the libraries that can be used for modeling once data preparation is done.

The Infosys Analytics Workbench allows data analysts to:

- Search for data within a big data lake
- Wrangle data with visualization and machine learning support
- Create models in open source R or Python
- Use workflows within the workbench to submit and approve and publish models
- Create a business user-friendly experience for using the model
- Consume analytic models

Proprietary Analytic Model Consumption

Infosys provides a unique and innovative way for business users to consume analytic models without needing technical or IT

support on parameterization or selection. With the Infosys Analytic Workbench business users can:

- Select a model from a gallery
- Select the scope of the model, such as applying a store clustering model to only a region
- Select parameters they want to change in the model, such as removing competition parameters in a store clustering model
- Add inputs such, as a range of the number of clusters they would like to see, and then run the model with these changed inputs

Advanced Search Algorithms

Searching for data within a data lake is not trivial and there are advanced algorithms at work here. Finding customer data in enterprise data lakes may show up surprises such as addresses being present in the warranty data base, phone numbers being present in market research panel data sent by an agency and stored in the market research data base, and so on. Enabling discovery of such data resources is key to a search function in big data-based analytical modeling to ensure the models access all possible data to achieve their objectives.

Self Service Statistical Analysis for Business Users

Business users are largely at a disadvantage when it comes to leveraging self-service analytical models. After all, they are not statistically trained. Since models are built for specific data patterns, how can those remain the same from case to case, and how can one reuse such models? Most business users can understand how self-service works for visualization, for example, with Tableau and other similar tools. Business users can slice and dice visualizations, click and select and move dimensions, and so on. But extending business user self-service capabilities to statistical analysis has thus far proved difficult.

The Infosys Analytics Workbench enables business users to unlock the world of data with self service analytics. The tool works off the metadata and provides a list of potential scope choices that the data scientist or analyst can enable for business users. It offers similar lists of parameters from which the analyst can decide which to "open up" to business users for de-selection. In some cases, data analysts/scientists who have developed the model may disable selection of model parameters by business users. Using the constraints framework, data scientists can trigger business users to input parameters that the model will use—a range of clusters in a clustering model, for example.

Since data is dynamic, a business user's selection of scope, parameters or constraints may lead to the model being statistically invalid. In such cases, the business user need not understand such variables as R square, p values or silhouette index. The business user simply sees a R-Y-G signal indicating the model's status: safe to use, use with caution, or do not use (speak to a data scientist). Once the data scientist/analyst sets up the degrees of freedom for the business user and publishes the model, all subsequent model reuses can happen in self-service mode, without the need for technical intervention or its bottleneck.

Summary

When the Infosys Analytics Workbench embedding **Trifacta data preparation platform** is deployed, we find that demand on data scientists/analysts' time goes down while the usage of analytical models by business users, goes up. Self-service data preparation leads to more statistical analysis, which can mean better business results.

For example, a retail business user may use a model to segment customers or isolate assortment gaps or see which customers are churn or are likely to make certain purchases. They can conduct this statistical analysis without having to wait for limited data scientist/analysis resources to become available; they can configure and perform this analysis themselves, right when it's needed to make better business decisions.

When business users no longer have to depend on a limited set of data scientists/analysts to do statistical analysis, more business decisions will be based on statistical analysis. Self-service predictive analysis will become commonplace sooner rather than later. Excel may evolve to connect with open source statistical analysis tools. More likely, a new generation of self-service tools of statistical analysis and machine learning will be available on every desktop, either on premise, or more likely, on the cloud. The Infosys Analytics Workbench, supplemented by the self-service data preparation capabilities of Trifacta data wrangling platform, is paving the way.

About the Authors

Subhashis Nath

Subhashis Nath is an AVP in the analytics practice and leads the analytics practice for the CRL segment. Prior to this, he has been a business consultant for over a couple of decades working with retail and CPG clients across the US, Europe and APAC. With multiple papers published, as well as patents, Subhashis is a thought leader in the analytics and optimization space especially in the retail and consumer goods segment. Here he talks about his patented innovation on business user self-service analytics.

Bertrand Cariou

Senior Director Partner Marketing, Trifacta

Bertrand Cariou is Senior Director of Partner Marketing at Trifacta. He has more than 25 years of experience in the computing and data management industry holding roles from engineering, consulting, product management, product marketing and company strategy. Bertrand joined Trifacta 4 years ago after 15 years tenure in the largest independent software vendor in data management where he founded their French operations to ultimately led their big data lake strategy globally.

For more information, contact askus@infosys.com

Infosys®
Navigate your next