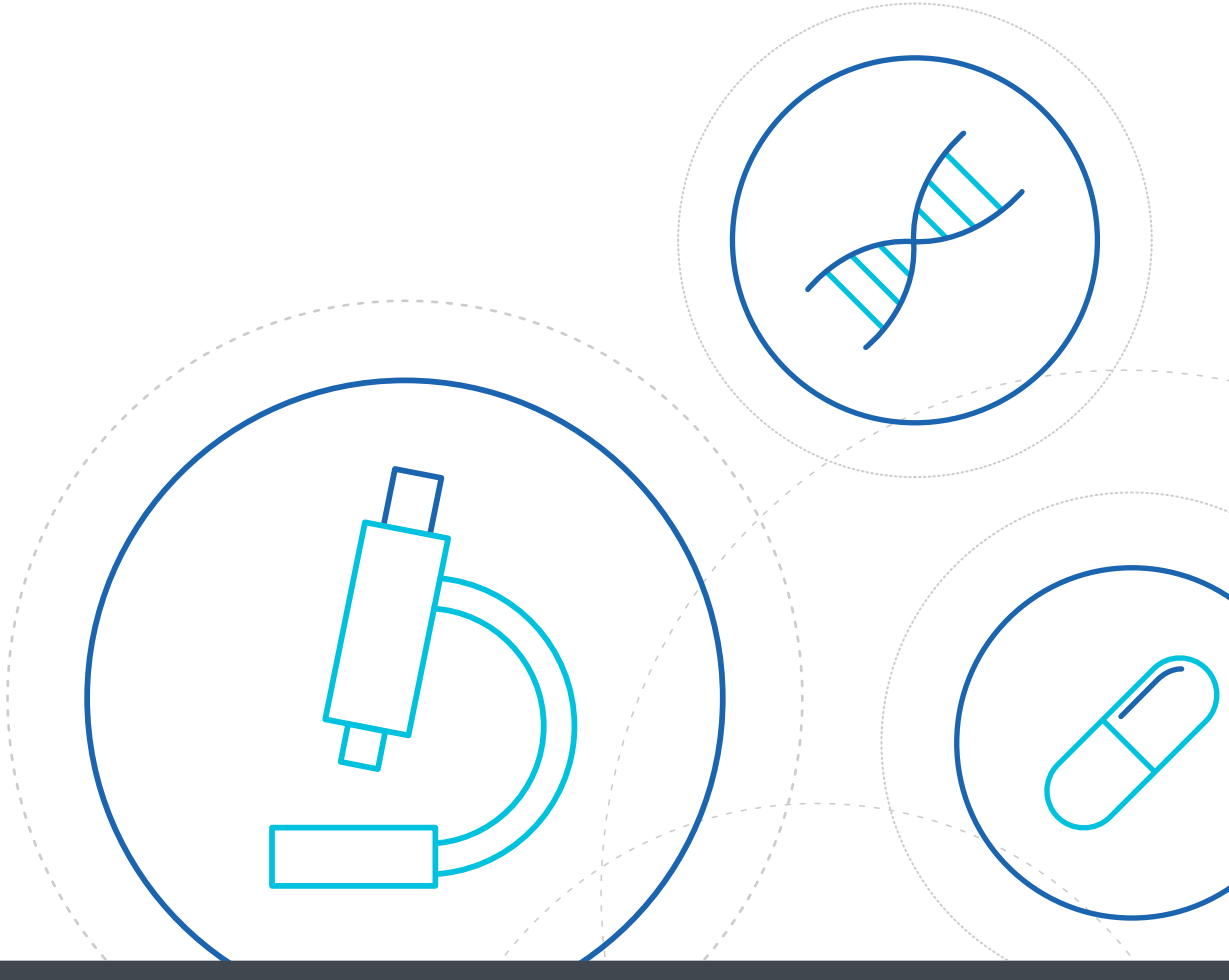
TRIFACTA

575 Market St, 11th Floor
San Francisco, CA 94105

www.trifacta.com
844.332.2821

WHITEPAPER

The Opportunity for Data Wrangling in Life Sciences and Biopharmaceuticals



Finding Growth in Technology Investments

“To address these marketplace dynamics and opportunities for growth, leading pharmaceutical and life science organizations are turning to enhanced data analytics initiatives to help drive innovation, find efficiencies and unlock profits.”

Given its sophisticated and high-stakes undertakings, the life sciences industry is well positioned to receive outsized benefits from investments in data analytics. Added to the usual uncertainty from shifting compliance requirements and more stringent FDA and international regulatory guidelines for new drug development, life sciences organizations are also focusing on new issues that could affect growth in this mature industry. Bringing new types of products to market, such as biotech and genomics drugs, in a changing regulatory environment is incredibly difficult. Factor in healthcare reform initiatives such as the Patent Protection and Affordable Care Act (PPAC), market-related elements such as health service purchasing policies, variable public/stakeholder acceptance of new drugs and the evolution of relevant data sources. All contributing to a complex, continually changing industry with accelerated business cases for analytics investments.

One of the most important drivers of change in the pharmaceutical industry in this decade has been the patent expirations of three dozen of the world’s top brand-name drugs. Ensuing competition from cheaper generics resulted in a significant drop in top drug companies’ annual U.S. sales. Manufacturers found needed savings by shrinking R&D departments, which had traditionally contributed to a large portion of the industry’s spending. To compensate, companies showed signs of restructuring away from vertically integrated organizational structures to a more diversified value chain including emerging companies, contract research groups supporting pre-clinical trial and clinical trials and contract manufacturing groups. While the ecosystem has evolved to support more difficult diseases and better accountability in care, the drug development process has also become more complex, longer and requires coordination among a much broader group of stakeholders. These stakeholders include partnerships with health plans and their payers, data aggregators, health systems/hospitals and integrated health networks, academic institutions, government organizations (e.g. CMS, HHS and NIH) and many others.

These complexities are compounded by an emerging field of ‘personalized medicine’, sometimes referred to as precision or individualized medicine. Personalized medicine uses diagnostic tools to identify specific biomarkers, often genetic, to help assess which medical treatments and procedures will be best for each patient. Care targeted at the individual promises an opportunity to better target disease causation, progression and response to drugs, ultimately improving health outcomes for patients and creating greater efficiency within the healthcare system.

To address these marketplace dynamics and opportunities for growth, leading pharmaceutical and life science organizations are turning to enhanced data analytics initiatives to help drive innovation, find efficiencies and unlock profits.

Opportunity in Innovation Through Data Wrangling

> "Becoming a data-driven organization in this industry requires mastery of utilizing traditional data sources, along with a host of new data feeds, and shared repositories of data in multiple visual and text formats."

Data is the driver of innovation and growth for the modern pharmaceutical and life sciences organization. The big data revolution and corresponding technologies are offering business analysts a new way of working with large scale, raw, disparate data. Leading firms are beginning to explore floods of new data sources from connected devices, patient monitoring and digital health and patient outcome records. This analytic capability, although nascent, is starting to unlock insights into how to improve the efficiency of clinical trials, bring drugs to market faster, reduce operating inefficiencies and make sales teams more successful.

There has also been a shift in data sharing among all of the stakeholders in the value chain. Biopharmaceutical companies are participating in public-private research initiatives to harness the power of big data and genome sequencing, to improve the success rate for discovering new medicines. These shared data repositories have enabled manufacturers to implement patient-centered approaches across all stages of translational and clinical research in therapeutic areas where targeted therapies are less available.

Becoming a data-driven organization in this industry requires mastery of utilizing traditional data sources, along with a host of new data feeds, and shared repositories of data in multiple visual and text formats. Business success requires clinical trial data, census data and treatment therapy data, along with enterprise data sources like ERP, CRM and relational databases to be aggregated into a single view. This challenges associated with this opportunity grow along with the massive expansion of unstructured data, semi-structured data and streaming feeds. The digitization of everything has created scores of new data sources. What was previously confined to historical data, now includes sources such as social media feeds and streaming data from patient monitoring devices.
To consume massive amounts of streaming and unstructured/ semi-structured data, new technologies such as Hadoop are being brought in to aid with storage and processing of data at scale. Hadoop environments provide a "data lake" or "hub" to store any type of data set in its native format, regardless of scale or structure, and process it for a variety of analytic purposes.

Common Challenges in Leveraging Data in Life Sciences

The pharmaceuticals and life sciences industry faces various challenges in utilizing available data.

“The availability of new and shared data has provided opportunities in disease understanding and treatment, gene therapy and drug design and diagnostics. However, it has also added significant new complexities, which obstruct usable data from being utilized for valuable analytics.”

Security and Privacy Concerns

Regulatory and compliance environments mandate specific ways personally identifiable health information can be handled. Organizations must mask and take elevated security precaution with such information. Further, from a competitive perspective, R&D data must be properly encrypted and secured to ensure that the company’s immense investments of time and money are protected.

Data Interoperability

Organizations must unite high volumes of data from sources that were never designed to link together. Unstructured and semi-structured outputs, such as patient monitoring devices and social media streams, are stored in disparate repositories such as Hadoop and NoSQL databases. The influx of data interchange formats from health device APIs and electronic medical records also provide non-standard data formats that stress traditional business intelligence (BI) solutions. Given organizational growth is often driven through M&A strategy within the industry, large and mid-sized organizations continue to struggle with data interoperability issues, as they seek to ingest and digest the large numbers of conflicting and siloed data sets specific to each acquired company. The availability of new and shared data has provided opportunities in disease understanding and treatment, gene therapy and drug design and diagnostics. However, it has also added significant new complexities, which obstruct usable data from being utilized for valuable analytics.

More Efficient Use of Skills and Expertise

Life sciences organizations strive to be data-driven, relying on highly skilled statisticians, programmers, computational chemists and bioinformatics specialists, most with domain specific expertise. Continually hiring talent with the right mix of statistics, programming and science-based domain expertise is difficult. Once they are on board, even then, their skills are often marginalized as they spend a majority of their time in low-level cleaning tasks or not being able to gain access to the data they need. Some estimate that data cleaning and preparation tasks constitute 50-80% of the development time and cost[1] in data warehousing and analytics projects.

Performing exploratory analytics on large, disparate data sets at scale requires a different set of skills, technologies and techniques not commonly found in existing R&D business and data analytics teams.

Barriers to Data Collaboration Within and Across Organizations

Life science organizations are not structured as holistic organizations. Data is structured across a variety of disparate systems, traditionally walled off from each other, with no easy way to transfer data between them.

In total, the industry faces a challenge in that many new sources of data must be parsed, cleaned and transformed at enterprise scales within and across organizations before they can even be leveraged to provide business value.

Use Case Spotlight:
Enabling Pharmacovigilance

> "In total, the industry faces a challenge in that many new sources of data must be parsed, cleaned and transformed at enterprise scales within and across organizations before they can even be leveraged to provide business value."

Trifacta in Action

Leveraging Diverse Data to Enable Pharmacovigilance

Pharmacovigilance has become a critical phase in clinical development programs, following the multi-billion dollar fines and withdrawals of many blockbuster drugs. The safety monitoring regulations have become stringent; and, concerns about safety can have serious implications on a company's stature and reputation. Pharmacovigilance has led to an increase in regulatory surveillance requiring safety data collection and analysis, which has in turn increased costs to the organization.

Big data environments combining Hadoop, data wrangling solutions such as Trifacta and downstream analytic tools, are enabling Life Sciences organizations to add a new dimension to traditional safety analysis. The value proposition of data collection for safety analysis purposes is threefold: unlock value by shortening the time to identify adverse events (AE), maintain compliance and reduce costs.

In the following adverse effect (AE) analysis example, an analyst can take the following data sets from a Hadoop data lake, access, transform and blend them together using Trifacta to track/predict adverse events from a particular drug. The goal is to get to faster and better insights from unconventional new data sources. A sample selection of big data sources may include:

- Health agency databases such as FDA Adverse Event Reporting System (FAERS) and Uppsala Monitoring Center (VigiBase) data
- Real world data from cohort studies often provided as both structured and unstructured data
- Social media (e.g., Twitter, Facebook)

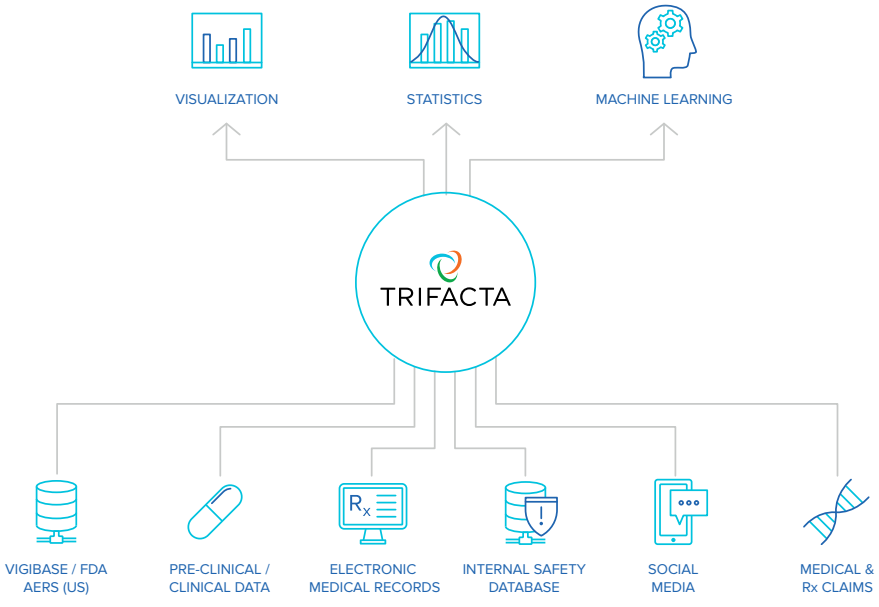


FIGURE 1: DATA WRANGLING TO INCORPORATE DIVERSE DATA FOR SAFETY ANALYSIS

Important Life Science Use Cases Powered by Data Wrangling:

- Product and service enhancements for personalized health care
- Regulatory compliance/internal reporting
- Customer lifetime value analysis

In order to utilize the FAERS data sets (data the FDA collects, harmonizes and posts online) as an example, there are a number of obstacles to first overcome. Common data wrangling challenges include: large numbers of data entry errors and typos, inconsistent units, a large number of missing fields, matching active ingredients across different drugs, periodic format changes, duplicate name variations and duplicate entries. Social media sites such as Twitter have also helped identify adverse event data. While a few individual experiences may not be clinically useful, thousands of drug-related posts could potentially reveal serious and unknown adverse events. Combining a FAERS data set with a social media stream like Twitter would entail a significant amount of work and time to transform and make the combined data usable.

FIGURE 2: SAMPLE FAERS DRUG DATA SET

By using Trifacta to identify early warnings of drug problems, a pharmaceutical company can find real business value in the final analysis by getting to quality data faster. Using other approaches, this process of preparing diverse big data for analysis can often take months, but with Trifacta it can now be performed in a matter of hours.

Empowering Life Sciences Organizations to Create Actionable Data through Data Wrangling

Trifacta presents Life Sciences analytics teams with a new approach to dealing with the challenges of working with the scale, complexity and diversity of today's data. Trifacta's approach to data wrangling utilizes the latest techniques in machine learning, data visualization and human-computer interaction to allow IT teams, data scientists, data analysts and business analysts to become more productive in wrangling data themselves—allowing them to build and manage data products and transformation scripts more effectively and on-demand.

Researchers use Trifacta's Data Wrangling Solution:

- To create product and service enhancements for drug development and delivery. Life Sciences organizations are using Trifacta to curate and standardize biomarker data for individuals in clinical trials. Trifacta enables customers to assess combinations of phenomic and genomic data to aid in more personalized drug development.
- To understand customer lifetime value. Vast data from health assessments, wellness activities, pharmacy claims and online customer information can be unified to aid in upsell and cross-sell activity, customer lifetime value and retention efforts.

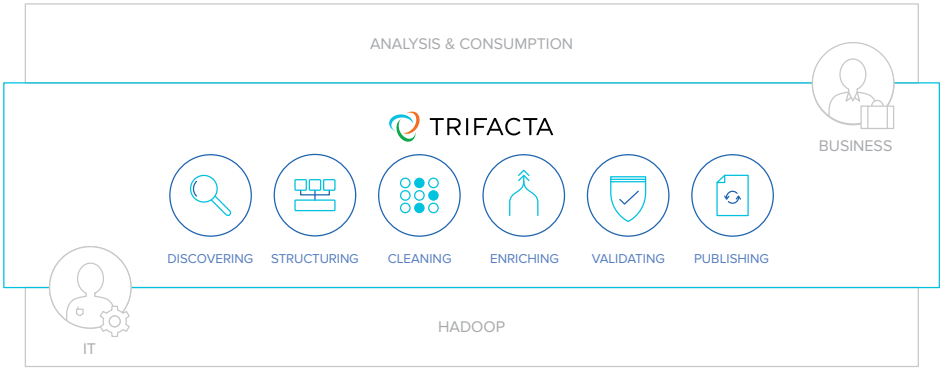


FIGURE 3: HOW TRIFACTA FITS IN THE DATA ORGANIZATION

Benefits of Trifacta:

- **Accelerate Innovation:** Trifacta removes legacy inefficiencies and coding requirements for analytic data preparation by providing a superior approach for analyst teams to directly discover and transform data.
- **Empower Business Users:** Trifacta enables non-technical users to directly access and manipulate raw, complex, data on-demand and on a self-service basis. Business users can be up and running, working directly with raw, multi-faceted data sets in a matter of hours.
- **Faster, More Accurate Outcomes:** It is well known that data analysts and data scientists spend up to 80%[2] of their time preparing data for analysis. By reducing the time to discover, structure, clean and enrich diverse data sources in Hadoop, Trifacta enables analysts to utilize more data in analytic exploration and experimentation resulting in more accurate results at a faster pace.
- **New Analytics Drive New Opportunities:** Given the cost and challenges associated with preparing these complex sources, traditional data sources are often underutilized or completely removed from analytic initiatives. With automated routines for anomaly/irregularity detection and visual summaries of the content of these complex sources, Trifacta enables analysts and researchers to explore more data and discover more possibilities to drive innovation and results.

A Disruptive New Approach to Preparing Data

Experience a new way of working with diverse data—empowering analysts to interact with data in ways they never thought possible.

Interactive Exploration: Trifacta presents the user with automated visual representations of the data based upon the inferred data type of each attribute of the data. These profiles require no specification by the user and Trifacta automatically presents each data type in the most compelling visual representation—geographic elements are presented as maps; time-oriented elements are presented according the common hierarchies such as day, month,

About Trifacta

Trifacta, the leading data wrangling solution for exploratory analytics, significantly enhances the value of an enterprise's big data by enabling users to easily transform and enrich raw, complex data into clean and structured formats for analysis. Leveraging decades of innovative work in human-computer interaction, scalable data management and machine learning, Trifacta's unique technology creates a partnership between user and machine, with each side learning from the other and becoming smarter with experience. Trifacta is backed by Accel Partners, Greylock Partners and Ignition Partners.

For Additional Questions, Contact Trifacta

www.trifacta.com
844.332.2821

Experience the Power of Data Wrangling Today

www.trifacta.com/start-wrangling

year, etc. Every profile is completely interactive—allowing the user to simply select certain elements of the profile to prompt transformation suggestions.

Predictive Transformation: Upon pulling up a data set within Trifacta users are presented with a visual representation of the data set they are working with. These visual representations are interactive—enabling the user to click, drag or select over the specific elements or attributes of the data they would like to manipulate. Every interaction within Trifacta leads to a prediction—the system evaluates the data you are working with and the specific interaction applied against the data to then recommend a ranked list of suggested transformations for the user to evaluate or even edit depending upon what they're trying to do.

As users browse through the different suggested transformations presented to them, the system will also present a preview of how each transformation, when applied to the data, will impact the data itself. This iterative feedback loop is always occurring throughout the use of Trifacta—constantly taking inputs from the data and the user to intelligently recommend ways to manipulate the data and giving the user the ability to validate their work with previews of each transform.

Intelligent Execution: Every transformation step defined by the user in the application is logged in Trifacta's domain specific language called Wrangle, allowing the application to take the finished script the user is defining in Trifacta and compile that down into the appropriate execution framework based upon the scale of the data the user is working with and the type of transformation. Depending upon the data, Trifacta can compile down to Pig, Spark and Trifacta's own execution engine for jobs that can run on a single machine. This is all done behind the scene—abstracting the user from the underlying execution framework.

Collaborative Data Governance: Although the core focus of Trifacta is enabling the people who know the data best to be able to access and transform it themselves, we recognize that organizations require having centralized processes for determining who has access to data, how metadata and lineage are tracked, how transformation jobs are operationalized and how data sets and transformation scripts are shared with other users. Instead of creating a completely separate governance framework in Trifacta, we have built support for the existing enterprise standard frameworks on Hadoop for security, user authentication, access controls, job scheduling and so forth. This enables Trifacta customers to simply implement existing governance policies in Hadoop instead of creating a new, entirely separate governance framework for Trifacta.

Sources

[1] New York Times, For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights, August 2014

[2] Forrester, 3 Ways Data Preparation Tools Help You Get Ahead Of Big Data, February 2015