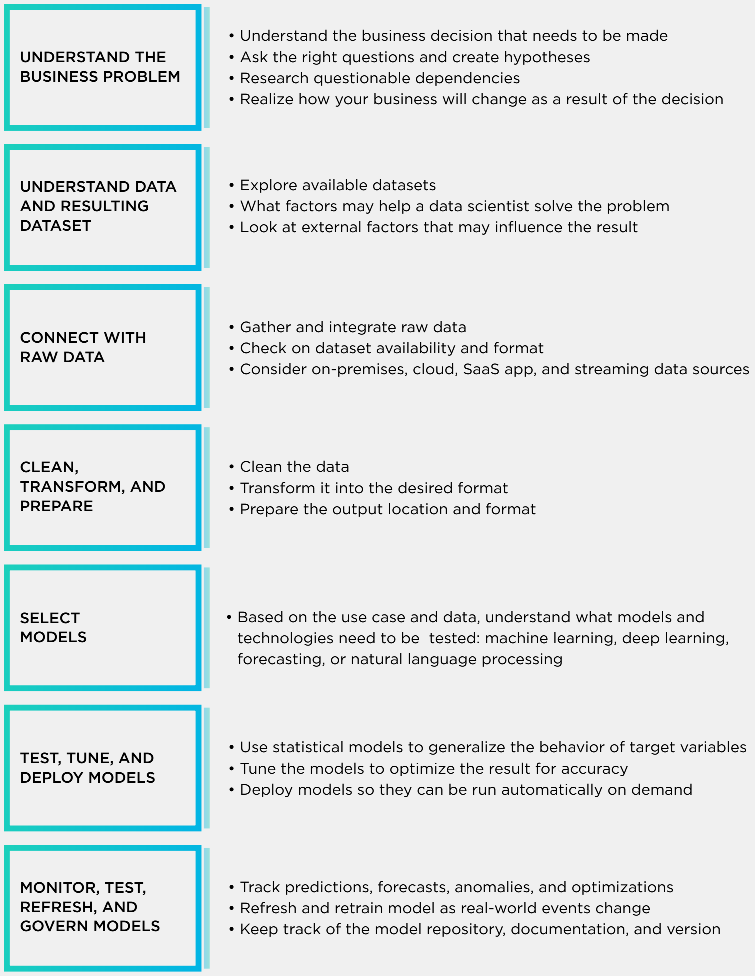
TIBCO whitepaper

Reducing Time to Data Preparation and Analytics for Machine Learning

Fast track data from preparation to machine learning and analytics

The Process

As data is continues to grow at an alarming rate, the race is on for companies to harness insights developed from it. To realize and monetize the value of data science, organizations infuse predictive findings, forecasting, and optimization strategies into business and operational systems. A typical process for a data scientist or citizen data scientist starts with understanding a business problem and moving through multiple steps before getting the final result. Here is a snapshot of what the process looks like:



UNDERSTAND THE BUSINESS PROBLEM

- Understand the business decision that needs to be made
- Ask the right questions and create hypotheses
- Research questionable dependencies
- Realize how your business will change as a result of the decision

UNDERSTAND DATA AND RESULTING DATASET

- Explore available datasets
- What factors may help a data scientist solve the problem
- Look at external factors that may influence the result

CONNECT WITH RAW DATA

- Gather and integrate raw data
- Check on dataset availability and format
- Consider on-premises, cloud, SaaS app, and streaming data sources

CLEAN, TRANSFORM, AND PREPARE

- Clean the data
- Transform it into the desired format
- Prepare the output location and format

SELECT MODELS

- Based on the use case and data, understand what models and technologies need to be tested: machine learning, deep learning, forecasting, or natural language processing

TEST, TUNE, AND DEPLOY MODELS

- Use statistical models to generalize the behavior of target variables
- Tune the models to optimize the result for accuracy
- Deploy models so they can be run automatically on demand

MONITOR, TEST, REFRESH, AND GOVERN MODELS

- Track predictions, forecasts, anomalies, and optimizations
- Refresh and retrain model as real-world events change
- Keep track of the model repository, documentation, and version

Figure 1. Steps to the final result.

The Issue

Today, modern organizations are solving business problems with data science. Data scientists are embedding the outcomes in day-to-day processes for better business decisions. But this is not the case in every organization; some are not prepared to keep up with what data science models need.

Let's look at some of the top issues responsible for data science failures. The chart below shows average percentages of time wasted by type of activity. Notice that the top issues are not related to finding and fitting the right model— but finding, connecting, preparing, and publishing a single source of data that can be used by all users.

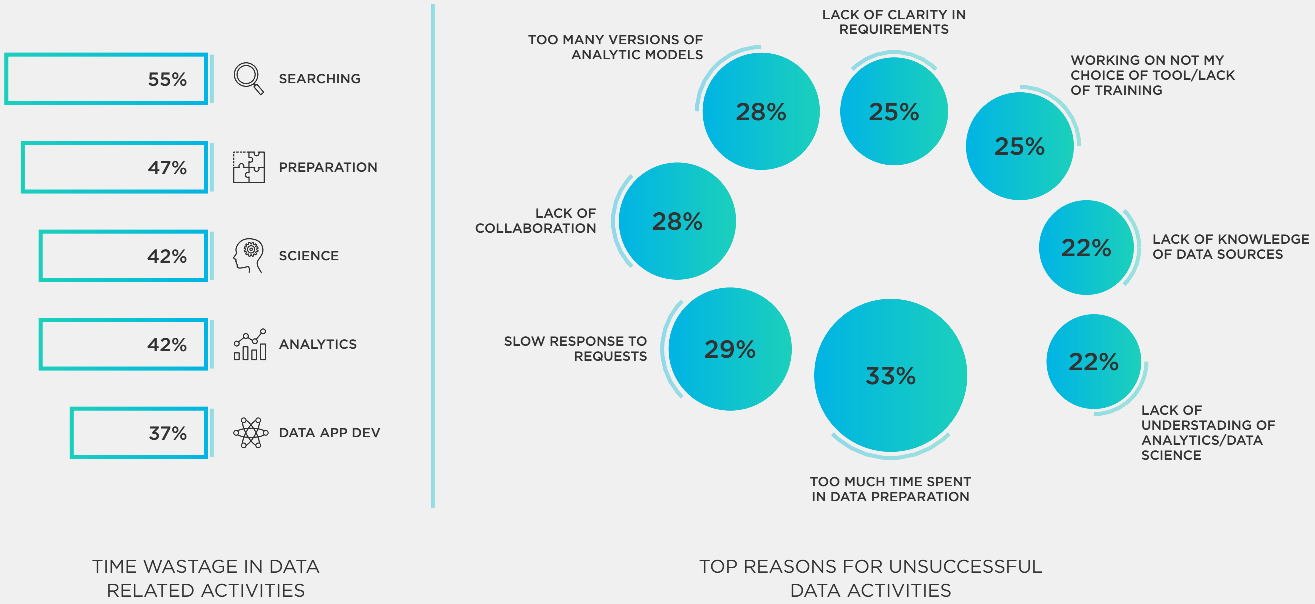


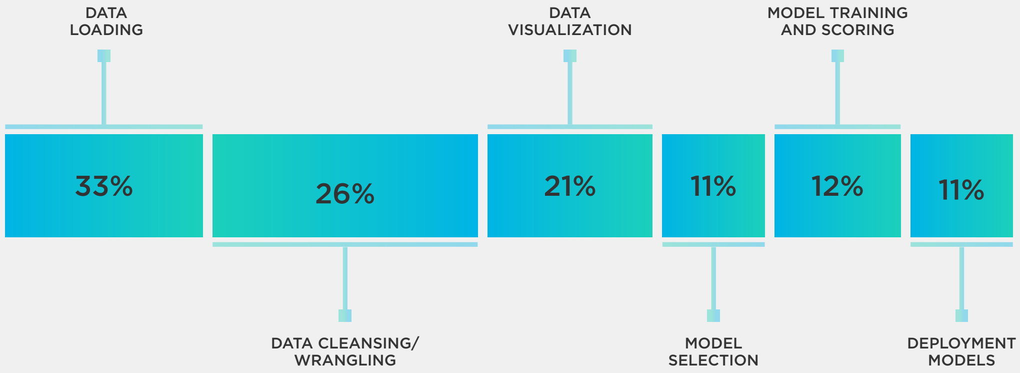
Figure 2. Time wasted in data-related activities and top reasons for unsuccessful analytics activities.

44%

of the time is being wasted weekly because data workers are unsuccessful in their activities

> *"We were disappointed but not surprised to see that data wrangling still takes the lion's share of time in a typical data professional's day.*
>
> *This has negative impact on overall job satisfaction."*
>
> *—2020 State of Data Science, From Hype to Maturity, Anaconda*

The Anaconda State of Data Science 2020 report states that data preparation (loading, wrangling, and cleansing) still takes the most time. The chart below shows an estimate of how data scientists spend their time on a typical business use case.



The Right Approach

Here is the right approach to eliminate some of the issues described in Figure 1.

1. Use a logical data warehouse.

You need data architecture flexible enough to handle random business requests for a logical or even physical data warehouse. A logical data warehouse provides connections to all data sources irrespective of where the data resides. It connects, transforms, and provides a unique access layer for all users, which eliminates miscommunication and over preparation of data because all data is quickly available, connected, prepared, and curated.

2. Monitor the end-to-end data flow.

It's important for data analysts and developers to see the data pipeline as a whole because it helps answer questions on where data is coming from, what transformations are happening, what views are being created, and what has been published in a logical or physical data warehouse. Machine learning models should be part of the data flow detailing in and out data streams.

3. Use any tool.

Users should be able to work in the tool of their choice. Based on their needs, analysts and data scientist use five to seven tools. For every tool replacement that is enforced, users have a learning curve, which slows time to results and restricts technical capability for a period of time.

4. Create and deploy models faster.

Any organization starting to use machine learning to solve business challenges should first crawl, then walk, and then run. Many organizations hire citizen data scientists who have some knowledge of data models and want to use click and run tools to test, train, and deploy machine learning.

5. Increase agility.

With the speed of datascience today, organizations will not have time and money to treat every use case as a large project. Before starting to engage in data science activities, take a step back and analyze the agility of the data architecture in fulfilling data connection, data preparation, and machine learning needs.

6. Increase response time.

With the global expansion of organizations, data pipelines run around the world. Connecting to and delivering data to multiple locations on an as-needed basis for creating data marts, machine learning, analytics, or any other purpose needs to be highly performant.

The Right Solution

How can organizations expand from where they are today to using a flexible data pipeline that eliminates the most common data access, data preparation, data cleansing, and machine learning issues? We can solve this and do more with a hybrid approach with TIBCO WebFOCUS and TIBCO Data Virtualization software.

TIBCO Data Virtualization

Data virtualization provides a modern, unified data layer that enables users to access, combine, transform, and deliver datasets with breakthrough speed and cost-effectiveness. Data virtualization technology gives users fast access to data housed throughout the enterprise—including in traditional databases, big data sources, and cloud and IoT systems—at a fraction of physical warehousing and extract, transform, and load (ETL) time and cost.

TIBCO Data Virtualization software allows you to connect, collaborate, track, cleanse, secure, accelerate, catalog, and deliver the data in the desired format so that it can be used by any user irrespective of use case or their chosen analytics tool.

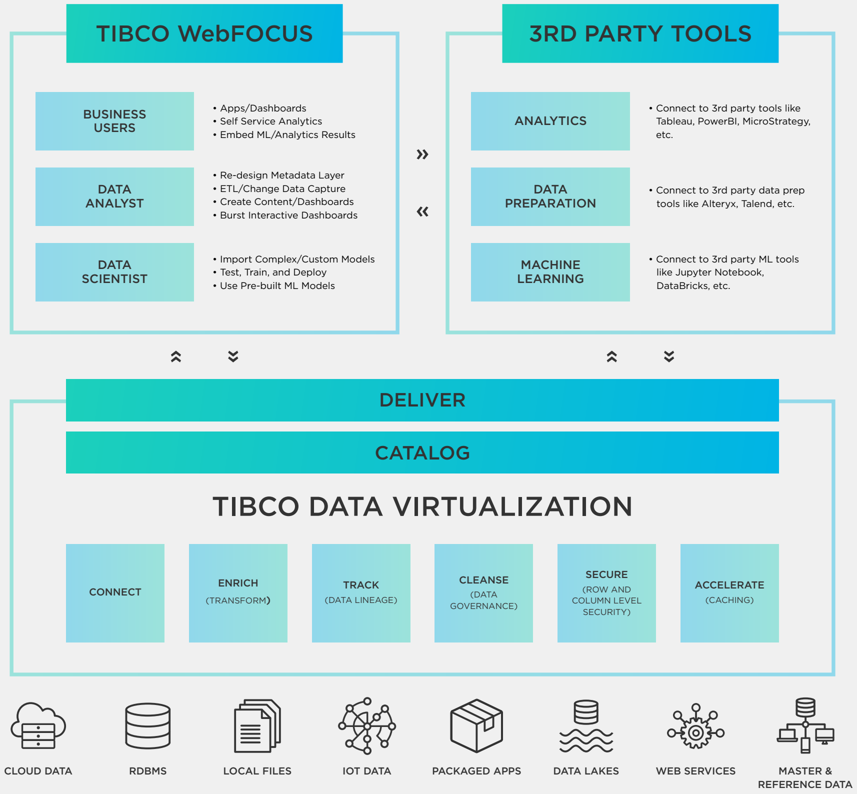
- **Connect:** Connect to any data source irrespective of where the data resides, in any format and data type (SQL, NOSQL, or streaming). Discover data automatically and get confident in the ability to form data joins.
- **Enrich:** Transform the data to make it ready to create a dataset. Join, create new fields, filter, pivot, or bin the data as needed.
- **Track:** Track data origin, what happens to it, and where it moves over time. Trace the root cause of errors in data analytics using data lineage.
- **Cleanse:** Track errors in data and create rules to clean the data in the logical layer.
- **Secure:** Apply security to rows, columns, and on final views. Simulate data access based on user roles.
- **Accelerate:** Cluster the virtualization layer to increase performance. Cache the data for faster reusable access.
- **Catalog:** Lookup or search available datasets using data cataloging capabilities through an intuitive web user interface.
- **Deliver:** Deliver the prepared data in a virtual layer and publish data as SQL or APIs.

TIBCO WebFOCUS

The sophisticated WebFOCUS enterprise business intelligence and analytics platform scales to meet requirements to satisfy every user, including executives, business users, data analysts, data scientists, and external entities such as partners or customers.

The data fabric from TIBCO Data Virtualization software can be ingested by WebFOCUS software to perform operations like reshuffle data, create content and dashboards, distribute content in multiple formats, run machine learning models, and even embed the results in multiple applications.

The WebFOCUS platform provides the capability to run pre-built models via a drag and drop user experience, compare the models, and export the results. It's suitable for citizen data scientists who do not want to get involved in machine learning code. WebFOCUS software can also import highly complex and customized Python/R models or simply connect to curated datasets from a Jupyter Notebook. The output can be embedded into WebFOCUS analytics applications and dashboards or extended to external applications. WebFOCUS analytics applications can scale for use by millions of simultaneous users, and it also allows organizations to take a logical data warehouse in a TIBCO Data Virtualization layer and convert it into a physical data warehouse through a multi-threaded ETL process.



TIBCO WebFOCUS

BUSINESS USERS
- Apps/Dashboards
- Self Service Analytics
- Embed ML/Analytics Results

DATA ANALYST
- Re-design Metadata Layer
- ETL/Change Data Capture
- Create Content/Dashboards
- Burst Interactive Dashboards

DATA SCIENTIST
- Import Complex/Custom Models
- Test, Train, and Deploy
- Use Pre-built ML Models

3RD PARTY TOOLS

ANALYTICS
- Connect to 3rd party tools like Tableau, PowerBI, MicroStrategy, etc.

DATA PREPARATION
- Connect to 3rd party data prep tools like Alteryx, Talend, etc.

MACHINE LEARNING
- Connect to 3rd party ML tools like Jupyter Notebook, DataBricks, etc.

DELIVER

CATALOG

TIBCO DATA VIRTUALIZATION

CONNECT

ENRICH (TRANSFORM)

TRACK (DATA LINEAGE)

CLEANSE (DATA GOVERNANCE)

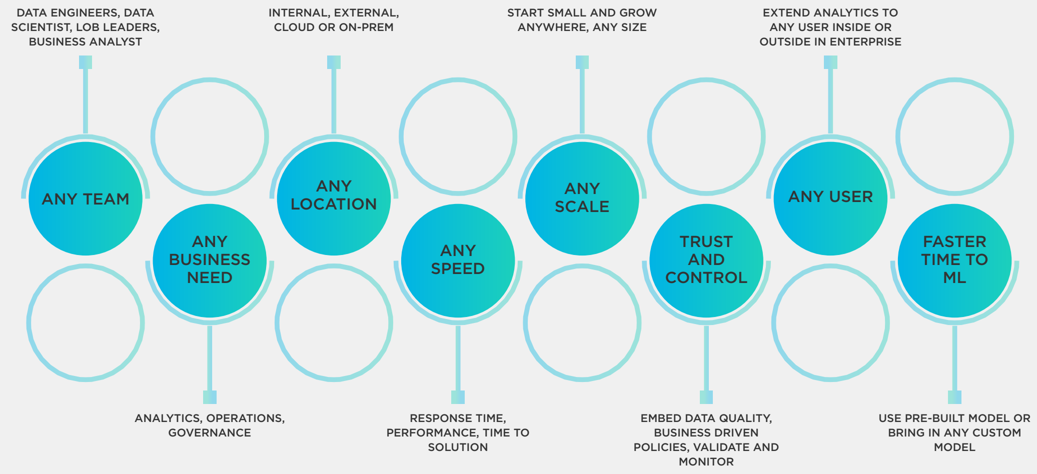
SECURE (ROW AND COLUMN LEVEL SECURITY)

ACCELERATE (CACHING)

The Impact

The solution enabled by TIBCO Data Virtualization and TIBCO WebFOCUS software eliminates the issues discussed in this whitepaper. Additional benefits include:

1. **Simplified data access layer:** Simplify the data access point for all users by providing them access to a unified virtual layer for all their data needs.
2. **Faster time to data**: Provision new data requests quickly as requirements change.
3. **Faster time to analytics:** Modern, easy-to-use WebFOCUS charts and reports deliver a streamlined content-creation experience.
4. **Business friendly access to data:** Maintain consistent master data and deliver data as business-friendly views with clear business definitions across the organization.
5. **Reduced IT cost:** Without needing to maintain physical data warehouses or work on multiple departmental tools with their own process and analytic tools, you can save time and cost.
6. **Governance and security:** Enforce access control across all your data and comply with regulations, including those that require encryption and masking.
7. **Scale on demand:** Your organization can scale as the demand for data and analytics increases. The solution manages multiple data pipelines for every business unit with data silos running through each pipeline. The content created by WebFOCUS software can be distributed to millions of users as needed.


DATA ENGINEERS, DATA SCIENTIST, LOB LEADERS, BUSINESS ANALYST
ANY TEAM
ANY BUSINESS NEED
ANALYTICS, OPERATIONS, GOVERNANCE
INTERNAL, EXTERNAL, CLOUD OR ON-PREM
ANY LOCATION
ANY SPEED
RESPONSE TIME, PERFORMANCE, TIME TO SOLUTION
START SMALL AND GROW ANYWHERE, ANY SIZE
ANY SCALE
TRUST AND CONTROL
EMBED DATA QUALITY, BUSINESS DRIVEN POLICIES, VALIDATE AND MONITOR
EXTEND ANALYTICS TO ANY USER INSIDE OR OUTSIDE IN ENTERPRISE
ANY USER
FASTER TIME TO ML
USE PRE-BUILT MODEL OR BRING IN ANY CUSTOM MODEL

Global Headquarters
3307 Hillview Avenue
Palo Alto, CA 94304
+1 650-846-1000 TEL
+1 800-420-8450
+1 650-846-1005 FAX
www.tibco.com

TIBCO Software Inc. unlocks the potential of real-time data for making faster, smarter decisions. Our Connected Intelligence platform seamlessly connects any application or data source; intelligently unifies data for greater access, trust, and control; and confidently predicts outcomes in real time and at scale. Learn how solutions to our customers' most critical business challenges are made possible by TIBCO at www.tibco.com.

17Aug2021