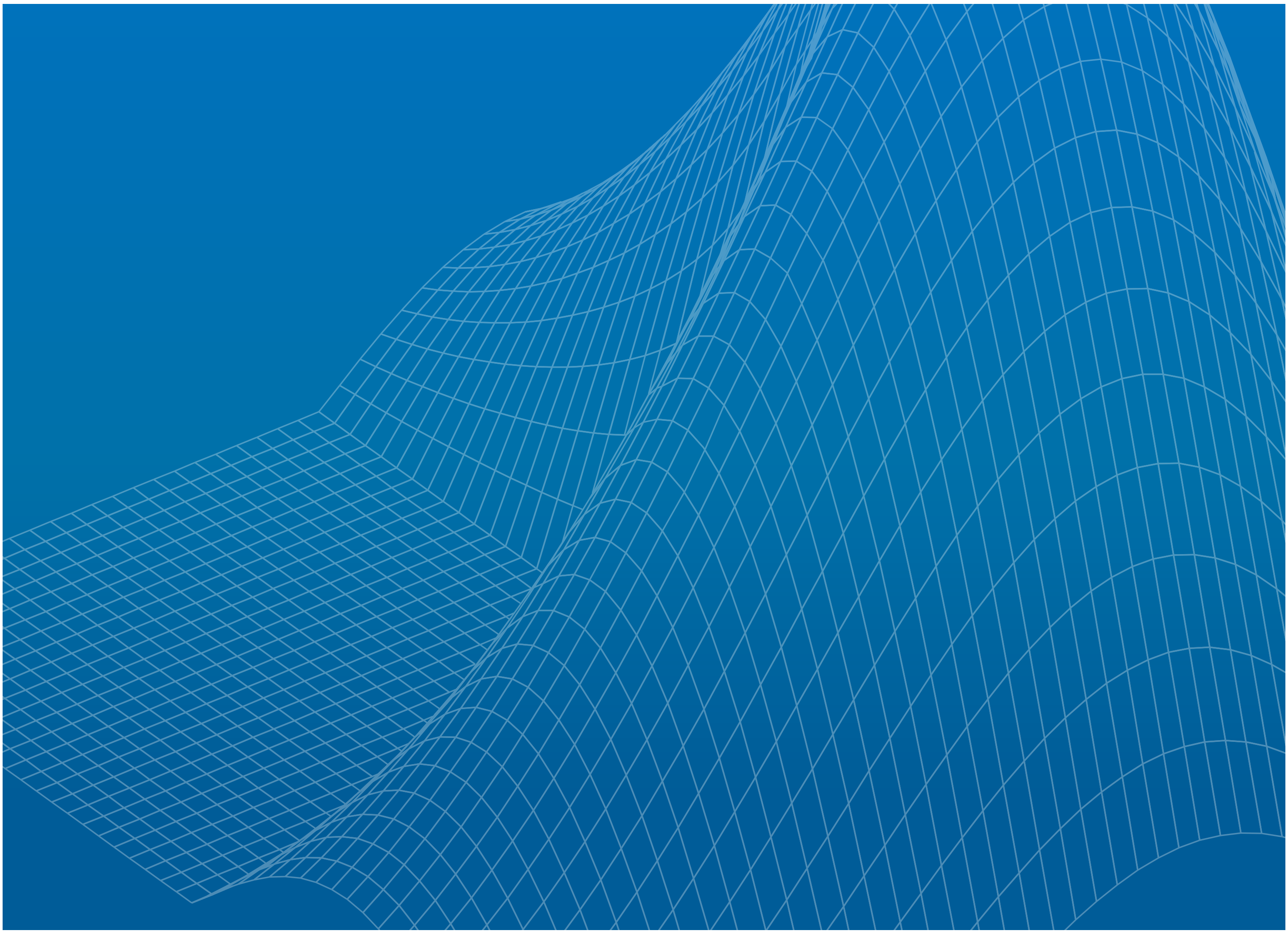


The Manager's Guide to Solving the Big Data Conundrum

WHITE PAPER

As business and engineering data proliferates, more businesses are turning away from using spreadsheets to analyze their big data.

This white paper discusses how you can harness the power of your big data by replacing or extending Microsoft® Excel® with MATLAB® to:

- Gain insights into business and engineering data
- Perform advanced, complex analyses without a specialist
- Ensure you can trust the results of the computations
- Visualize how different algorithms work without having to program
- Execute high-speed processing of large data sets

Big Data's State of Play

If you are responsible for projects or teams whose purpose is to spin raw data into gold, you have probably read a big data blog post or two. It is easy to feel intimidated by all the new terminology like semantic processing, data lakes, deep learning, neural networks, recursive networks, graph theory, and many others. But don't miss the forest for the trees: Data analysis is not an all-or-nothing discipline, and it is okay to start with what is immediately achievable, and make plans from there. Data offers businesses opportunities to gain actionable insights, and as the data grows so do the possibilities and complexities.

Business data, such as market trends, revenue forecasting, and transactional data, is often gathered and analyzed by one team, while engineering data, such as data from sensors, embedded systems, or cameras, is handled by another team.

With the rise of big data, machine learning, neural networks, and other advanced data analytics capabilities, business and engineering data has become a potential gold mine, if you know where, and how, to look.

Business data is often accessed through programs like Excel. Different sources and types of data are often saved in different spreadsheets, a necessity as data grows and causes simple tasks to take longer and longer. As a result, the data analyst has to open multiple windows to get the full view of the data, and integration becomes a significant difficulty.

It is not enough to have this data; it must also be navigable to find actionable insights. Some businesses spend a lot of resources trying to find a person who is well versed in business and industry knowledge and has coding skills, a combination that is predictably difficult to hire. One alternative to spending a lot of time and money hiring and ramping up a new specialist is to equip your existing data analysts, who already have the required business knowledge, with the tools they need to find actionable insights in the business and engineering data.

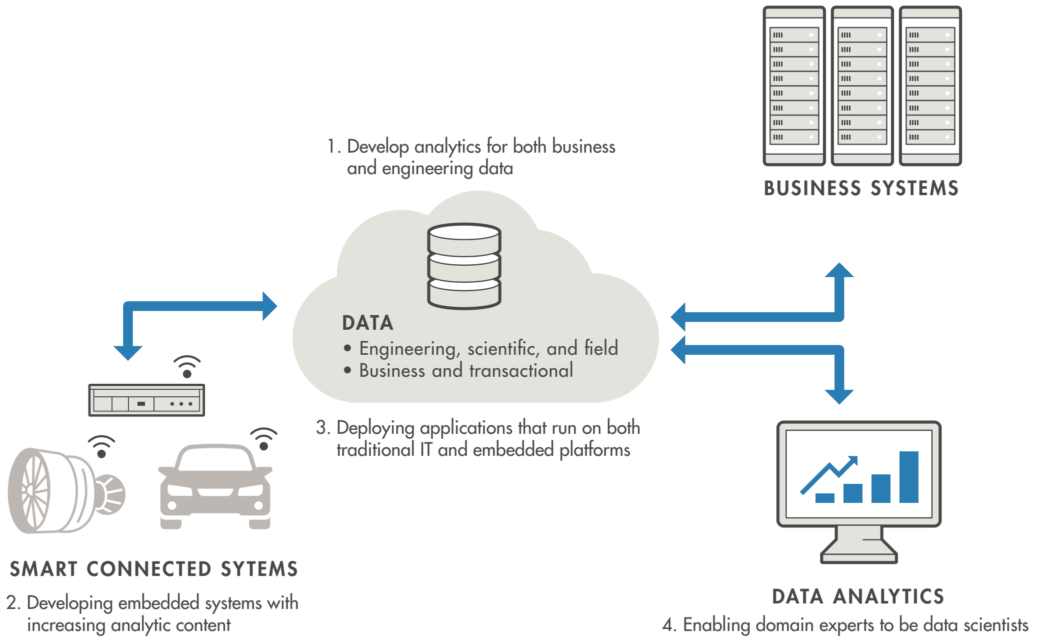

1. Develop analytics for both business and engineering data
BUSINESS SYSTEMS
DATA
• Engineering, scientific, and field
• Business and transactional
3. Deploying applications that run on both traditional IT and embedded platforms
SMART CONNECTED SYTEMS
2. Developing embedded systems with increasing analytic content
DATA ANALYTICS
4. Enabling domain experts to be data scientists

Figure 1. Using data analytics to gain insight from business and engineering data.

Extending Excel

For many years, Excel has been the go-to tool for recording and analyzing business and transactional data. If a data analyst is working with a lot of data, but not "big" data, and only wants to store and analyze the data, then in general practice there is nothing wrong with using Excel. As data grows and the desire to do more with the data arises, Excel should no longer be the default choice for a few reasons:

- The number of rows per sheet is limited (it is a little over 1 million). Once data passes that number, the analyst has the added complexity of working in multiple sheets.
- Visual analysis is difficult. The analyst must sift through all the raw data, spot flaws (especially difficult with non-numeric data), and visualize.
- Sharing data across teams is tricky. Deciding who can edit data, how to version control changes, and ensure data is current can be harder than it seems.

If the business is not ready to move past Excel despite these limitations, ***MATLAB*** applications can help extend Excel to work with big data. You can:

- Import Excel data into MATLAB
- Access MATLAB from Excel
- Package MATLAB code as Excel add-ins

Data can come from many sources, and businesses can easily underestimate the time it's going to take to aggregate that data into one location. Engineering, scientific, and field data is increasingly crucial to making the best business decisions, yet building home-grown applications capable of analyzing all the information available can yield questionable results in terms of time savings, performance, and scalability. By importing Excel data into MATLAB, you can add multiple Excel spreadsheets to work in one window.

MATLAB is also accessible from within Excel. ***Spreadsheet Link***™ connects Excel with the MATLAB workspace, enabling you to access the MATLAB environment from an Excel spreadsheet. An analyst can exchange data between MATLAB and Excel, taking advantage of the familiar Excel interface while gaining access to MATLAB algorithms in image processing, data analytics, and control engineering.

With ***MATLAB Compiler***™, you can package math, graphics, and user interfaces created in MATLAB as Excel add-ins to perform analyses with Excel. These Excel add-ins can be distributed royalty-free to users who do not have MATLAB, and they require no Microsoft® Visual Basic® for Applications (VBA) programming.

You can make sure multiple users get the latest version of your analytics automatically by deploying add-ins to MATLAB Production Server™ using MATLAB Compiler SDK™.

In many industries, businesses are developing embedded systems that have increasingly analytic content. This data is often in a different format and therefore unlikely to seamlessly integrate with the wider business data, which leads businesses to look for solutions that can replace or extend what they have in place.

Applications developed by analysts and engineers often need to run on both traditional IT and embedded platforms, adding another computer language to the analysts' requirements. Building effective, flexible, extendable applications to link all these systems should be done with an application that has robust support and documentation. These applications make code, data, and models sharable across teams.

MATLAB has comprehensive, professional documentation written by engineers and scientists. Reliable, real-time technical support staff answers questions quickly, and users can tap into the knowledge and experience of over 100,000 community members and MathWorks engineers on ***MATLAB Central.***

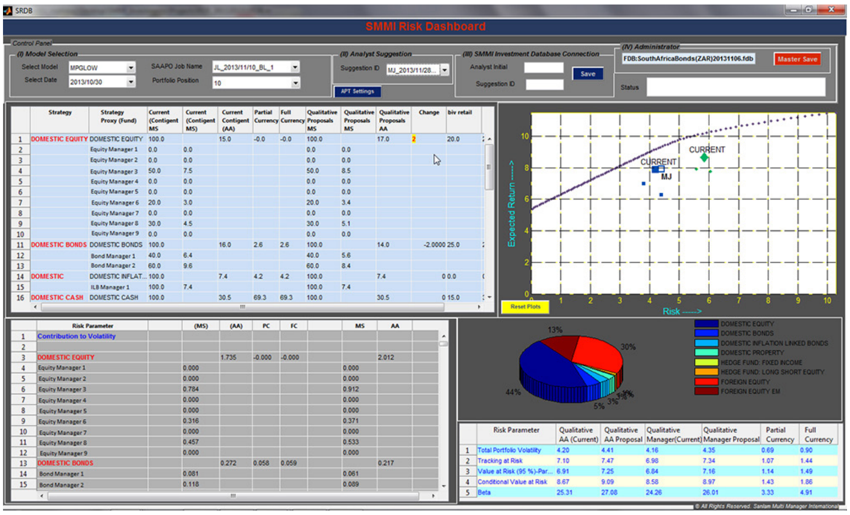
Storing and Querying Big Data Sets

The size of the data, variance in data types, and skills of an IT team will often determine how and where raw data will be stored. Whether data is kept in on-premises SQL Server® databases or in Hadoop clusters hosted in the cloud, the data must be accessible.

Engineers can build applications to analyze and visualize the data in MATLAB, which has toolboxes specifically for this task. MATLAB applications integrate with web, database, and application servers so that all the data can be accessed in one place. This gives analysts all the information they need to make the best business recommendations.

REAL-WORLD EXAMPLE

Sanlam Multi-Manager International Develops Dashboard for Quantitative Risk Analysis



SRDB

SMMI Risk Dashboard

Control Panel

(I) Model Selection

Select Model: MPGLOW

Select Date: 2013/10/30

SAAPO Job Name: JL_2013/11/10_BL_1

Portfolio Position: 10

(II) Analyst Suggestion

Suggestion ID: MJ_2013/11/28...

(III) SMMI Investment Database Connection

Analyst Initial:

Suggestion ID:

(IV) Administrator

FDB:SouthAfricaBonds(ZAR)20131106.fdb

Status:

	Strategy	Strategy Proxy (Fund)	Current (Contigent MS	Current (Contigent MS)	Current Contigent (AA)	Partial Currency	Full Currency	Qualitative Proposals MS	Qualitative Proposals MS	Qualitative Proposals AA	Change	biv retail
1	DOMESTIC EQUITY	DOMESTIC EQUITY	100.0		15.0	-0.0	-0.0	100.0		17.0	2	20.0
2		Equity Manager 1	0.0	0.0				0.0	0.0			
3		Equity Manager 2	0.0	0.0				0.0	0.0			
4		Equity Manager 3	50.0	7.5				50.0	8.5			
5		Equity Manager 4	0.0	0.0				0.0	0.0			
6		Equity Manager 5	0.0	0.0				0.0	0.0			
7		Equity Manager 6	20.0	3.0				20.0	3.4			
8		Equity Manager 7	0.0	0.0				0.0	0.0			
9		Equity Manager 8	30.0	4.5				30.0	5.1			
10		Equity Manager 9	0.0	0.0				0.0	0.0			
11	DOMESTIC BONDS	DOMESTIC BONDS	100.0		16.0	2.6	2.6	100.0		14.0	-2.0000	25.0
12		Bond Manager 1	40.0	6.4				40.0	5.6			
13		Bond Manager 2	60.0	9.6				60.0	8.4			
14	DOMESTIC	DOMESTIC INFLAT...	100.0		7.4	4.2	4.2	100.0		7.4	0	0.0
15		ILB Manager 1	100.0	7.4				100.0	7.4			
16	DOMESTIC CASH	DOMESTIC CASH	100.0		30.5	69.3	69.3	100.0		30.5	0	15.0

	Risk Parameter		(MS)	(AA)	PC	FC		MS	AA	
1	Contribution to Volatility									
2										
3	DOMESTIC EQUITY			1.735	-0.000	-0.000			2.012	
4	Equity Manager 1		0.000					0.000		
5	Equity Manager 2		0.000					0.000		
6	Equity Manager 3		0.784					0.912		
7	Equity Manager 4		0.000					0.000		
8	Equity Manager 5		0.000					0.000		
9	Equity Manager 6		0.316					0.371		
10	Equity Manager 7		0.000					0.000		
11	Equity Manager 8		0.457					0.533		
12	Equity Manager 9		0.000					0.000		
13	DOMESTIC BONDS			0.272	0.058	0.059			0.217	
14	Bond Manager 1		0.081					0.061		
15	Bond Manager 2		0.118					0.089		

	Risk Parameter	Qualitative AA (Current)	Qualitative AA Proposal	Qualitative Manager(Current)	Qualitative Manager Proposal	Partial Currency	Full Currency
1	Total Portfolio Volatility	4.20	4.41	4.16	4.35	0.69	0.90
2	Tracking at Risk	7.10	7.47	6.98	7.34	1.07	1.44
3	Value at Risk (95 %)-Par...	6.91	7.25	6.84	7.16	1.14	1.49
4	Conditional Value at Risk	8.67	9.09	8.58	8.97	1.43	1.86
5	Beta	25.31	27.08	24.26	26.01	3.33	4.91

© All Rights Reserved. Sanlam Multi Manager International

Sanlam Multi-Manager International (SMMI) analysts had identified inefficiencies in how they assigned weightings to asset classes and asset managers, tracked inputs over time, and shared analysis results. The team had previously relied on Microsoft Excel spreadsheets for analysis tasks, but complex calculations were taking too long, and tracking and maintaining multiple spreadsheets was a burden. SMMI initially considered developing a custom solution using VBA, but decided that it would be too slow for the rigorous quantitative calculations they needed to perform.

The analysts wanted to replace their spreadsheet-based solution with an application that accelerated analysis, simplified data management, and integrated risk input modeling, portfolio optimization, and data visualization.

SMMI found they could speed up analysis tasks by switching to MATLAB. Their results:

- Calculation time reduced from minutes to seconds. "It took up to five minutes to open the large, complex spreadsheets we used before, and another 20 seconds to recalculate after we made changes," says Jason Liddle, risk manager at SMMI. "With MATLAB we get the results almost instantaneously. Plus, we can save the results of our analysis to a database, which is much easier to access and manage than sets of spreadsheets."
- Quantitative analysis tools widely deployed. "After we deployed the quantitative risk dashboard created with MATLAB and MATLAB Compiler, our analysts were much better prepared to have productive, qualitative debates," notes Mathew John, investment analyst at SMMI. "MATLAB Compiler enabled us to scale the solution we had developed and make it available to our entire investment team."
- Development time shortened by months. "If we had chosen C++ or VBA instead of MATLAB, it probably would have taken us four times longer," says Liddle. "MATLAB makes it easy to test new ideas using built-in portfolio optimization functions and quickly turn an initial prototype into a production application."

>> Read the full story

MATLAB applications are a reliable solution that can shorten the time it takes to build an application, regularly analyze the data, and scale up, so analysts can arrive at conclusions faster even if the data grows unexpectedly.

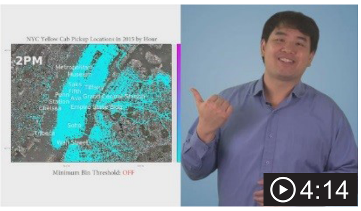
Analysts need a solution they can modify to consistently suit their needs. Deploying code in some applications is a treacherous and time-intensive task. However, deploying MATLAB applications and components provides many advantages:

- Domain experts can maintain ownership of ideas, algorithms, and applications.
- Software developers have the flexibility to integrate a common algorithm with different programming languages and platforms.
- Organizations gain efficiency by avoiding time-consuming and error-prone recoding and by easily adopting algorithm improvements throughout the application's life cycle.

Scale Up Calculations from Desktop to Cluster

Among the challenges of processing big engineering data sets are that they are often too large to fit into available memory and take too long to analyze on a single processor. In MATLAB, tall arrays address these challenges by enabling you to process data in many small chunks that individually fit into memory and to process many chunks in parallel.

Tall arrays provide a way to work with data that has millions or billions of rows. You do not need to write special code to take into account the huge size of the data. Tall arrays let you work with large data sets intuitively using the same functions that you use to work with in-memory MATLAB arrays. MATLAB handles all of the data chunking and processing in the background and automatically minimizes the number of passes through the data.



MATLAB Tall Arrays in Action

Use the power and simplicity of tall arrays in MATLAB to access, explore, and process big data, as well as data that doesn't fit in memory (including data stored on Spark and Hadoop).

REAL-WORLD EXAMPLE

Cornell Bioacoustics Scientists Develop a High-Performance Computing Platform for Analyzing Big Data

Researchers analyzing acoustic data must contend with noise from weather, other animals, and nearby machinery and vehicles. The variability of animal sounds across individuals within a species is a further complication. These two factors—noise and variability—increase the number of false positives and negatives, reducing the detection algorithms' accuracy.

At the Cornell University Laboratory of Ornithology, processing the hundreds of terabytes of data that Bioacoustics Research Program (BRP) scientists gather presents another challenge. A typical project involves processing years of raw acoustic data—up to 10TB—recorded on multiple channels. Each channel may capture hundreds of millions of events—sounds that stand out when the data is viewed as a spectrogram. Algorithms tested on small, high-quality samples are often considerably less accurate when applied to larger, noisier data sets.

BRP analysis tools must serve a wide range of research initiatives, environments, and shifting requirements. "Answers to our initial research questions often lead to brand-new avenues to explore, and we need to be able to handle these sudden changes in our requirements," says Dr. Christopher Clark, senior scientist and director of BRP.

To develop a high-performance computing (HPC) platform for analyzing big data, BRP turned to MATLAB. Their results:

- Years of development time saved. "A study of projected costs showed that if we had to do this on our own, it would take three years, $1 million, and a lot of outside help to develop the kind of HPC platform we needed," says Dr. Peter Dugan, lead data scientist for BRP. "With Parallel Computing Toolbox and MATLAB Distributed Computing Server, we developed the platform in under three months."
- Analysis time reduced from weeks to hours. "It took one of our algorithms 19 weeks to process 90 days of data," says Dr. Dugan. "Using Parallel Computing Toolbox and MATLAB Distributed Computing Server, we completed the same analysis on our cluster in 8 hours."
- Previously unprocessed data analyzed in days. "One data set captured 100,000 hours of sound. It was so large that we had previously processed less than 1% of it, estimating that it would take a year or more to process the rest," says Dr. Dugan. "With our MATLAB HPC platform, we processed the data six times, using different detection algorithms, in two days."

>> Read the full story

Parallel Computing Toolbox™ can immediately speed up your tall array calculations by using the full processing power of multicore computers to execute applications with a parallel pool of workers. If you have a cluster, including Hadoop clusters, then you can scale up your calculations further using *MATLAB Distributed Computing Server*™. One of the benefits of developing your algorithms with tall arrays is that you only need to write the code once. You can develop your code locally, then scale up to run in parallel on multicore computers or clusters and clouds, without having to rewrite your code.

Traditional data analysis techniques typically involve moving data into a computational environment to be analyzed. This approach—bringing the data to the analytics—may not be feasible for big engineering data. Support for Hadoop in MATLAB enables teams to bring analytics to the data, and leave big data where it is stored.

Using MATLAB to hook into Excel and Hadoop means that engineers don't need to be fully conversant with multiple software packages.

REAL-WORLD EXAMPLE

Analyzing Test Data from a Worldwide Fleet of Fuel Cell Vehicles at Daimler AG

To understand vehicle usage patterns, track fuel consumption, plan hydrogen refueling infrastructure, and understand how driving patterns affect vehicle performance, one team from Daimler uses MATLAB to query a central database.

> *"One measure of our success at providing useful results is the increasing demand for our services from engineers and managers at Daimler. We are swamped with requests. Being able to access the database, perform multiple analyses, plot results, and produce insightful reports in the integrated MATLAB environment is a big advantage. It means that we can easily add resources to our team.*
>
> *"Our engineers only need to know one software package—MATLAB—instead of multiple applications, and we do not have to spend time integrating diverse tools. Instead, we generate helpful results.*
>
> *"We continue to refine our current analyses and develop new ones to provide further insights and deeper understanding of fuel cell vehicle performance and infrastructure."*
>
> — Tim McGuire, Taylor Roche, and Andreas Weinberger, Mercedes-Benz RDNA, Inc.

>> Read the full article

Letting Your Team Work as a Team

Resources are so often the defining factor in software projects. How many engineers or analysts do you have? How much training will be needed? Which skills will these people need to produce actionable results? How much productivity is lost while analysts are performing repetitive tasks?

Manual tasks such as refreshing spreadsheets, aggregating new data, and sharing algorithms, models, and code can be a significant time sink for data analysts. Instead of wasting hours of time on repetitive tasks, analysts can increase their productivity by moving from spreadsheets to a platform developed using MATLAB.

REAL-WORLD EXAMPLE

Trient Develops Financial Analytics Platform to Support Its Investment Team

Each day, Trient Asset Management, an independent and fundamentally driven investment company providing fund and portfolio management, traverses and analyzes many gigabytes of historical data while continuing to clean and aggregate new data arriving from markets around the world. In the past, Trient analysts used Microsoft Excel for data management and analysis, but spreadsheets made it difficult to handle high data volume and velocity.

Spreadsheets had additional drawbacks. First, individual analysts encountered difficulties sharing their work with the wider team. Second, spreadsheets restricted the scope of the analyses they could perform. Third, making even a small change, such as adding a single new asset to be tracked, required hours of manual adjustments. As a result, highly skilled analysts spent too much time on basic tasks such as moving data and recalculating spreadsheets.

Trient replaced their spreadsheets with a financial analytics platform developed using MATLAB. Their results:

- Manual monthly calculations automated and run daily. "When we used spreadsheets, the many manual steps took so much time that we only ran calculations on a subset of data once a month," says Ariel Fischer, head of Systems Development and Infrastructure at Trient. "With our MATLAB based platform, we now run our automated analysis daily on many gigabytes of data."
- Calculation speeds increased tenfold with multicore processing. "We dramatically reduced calculation times using Parallel Computing Toolbox," notes Fischer. "The speed increased almost linearly with the number of cores; our calculations are now about 10 times faster when we run them on a 12-core processor."
- New asset class screening models implemented in hours. "With object-oriented programming in MATLAB, we created a framework and components that are easily reused, even by analysts who are not expert programmers," says Fischer. "After setting up the initial framework and implementing the first screening model, it took only a few hours to implement an existing model by reusing the framework in MATLAB."

>> Read the full story

A key benefit of using MATLAB is that analysts have all the data they need, without having to build special applications in other tools and languages to access the data repositories. Engineers and analysts can work together more efficiently in a single, integrated environment like MATLAB. Teams aren't siloed, building a single application and then going through the arduous task of integrating it with a legacy system. Instead, they can build apps to share code and models across teams.

MATLAB analytics can be packaged as deployable components compatible with a wide range of development environments, making them accessible across the business. This means that a change in a model or code made by one team does not need to be manually reproduced by every other team using the same model. This is especially useful for businesses that have adopted Model-Based Design. Engineers can share standalone MATLAB applications or run MATLAB analytics as a part of web, database, desktop, and enterprise applications. For low-latency and scalable production applications, you can manage MATLAB analytics running as a centralized service that is callable from many diverse applications.

Summary

MATLAB offers a unique chance for businesses to leverage their big data across their departments. By using MATLAB to analyze data, businesses can go beyond Excel to import their data into a single source, perform complex analyses, and share those applications and data to other teams.

Next Step

>> Learn more about integrating MATLAB into enterprise-scale applications