CITO Research
Advancing the craft of technology leadership

A New Strategic Approach for Data and Analytics

COAUTHORED BY:

Dan Woods
Chief Analyst, CITO Research

Chris Twogood
*VP Product & Services Marketing,
Teradata Corporation*

CONTENTS

CITO Research

Advancing the craft of technology leadership

The Déjà Vu of Big Data Debt

Haven't we been here before? A new world of technology presents itself, with exciting possibilities. We rush to take advantage of its capabilities, but at some point we find out we could have done a better job.

The world of big data is just such a scenario. Many companies are confusing adopting big data technology with creating a coherent big data strategy and in the process are creating big data debt. Big data debt is a concept inspired by the notion of technical debt, a term coined by Ward Cunningham, inventor of wikis and other innovations and further developed by Martin Fowler.

Big data debt ensues when we create one-off, application centric solutions

Big data debt ensues when we create one-off, application centric solutions. Instead of focusing on solving one immediate problem, we should instead create a coherent and comprehensive architecture that solves many problems. We should look at each project as an investment in building that architecture and to ensure we get maximum return over the long haul.

But how? What does that architecture look like? How does it incorporate new data and new analytics? How does it exploit the transformed economics of data processing brought about by Apache™ Hadoop®?

Our remedy for big data debt is the data and analytic centric approach, a set of concepts and guidelines that allow us to invest in big data and get increasing long-term returns rather than spending money to get a short-term payback that creates a long-term debt.

What Has Changed?

Let's look at the current situation facing the world of data and analytics as if we are investors seeking to maximize return. The first question is: What has changed? Why is the situation we are facing different from what we faced in the past? There are four main differences:

❶ **We are facing a flood of data.** The BI infrastructure of the past was built when data was scarce, but now we face a world awash in data. Annual data generation rates will increase 4,300% by 2020.[1] The current amount of data is a trickle compared with what is coming.

❷ **The signal in the data varies widely.** In the past, our enterprise applications were the source of most of our data. This data was painstakingly collected, highly curated, and had a very high signal. Now we have data that is all over the map, from high signal, high cost data to massive haystacks of low-signal data that have a few valuable needles of signal lurking within.

[1] IDC, cited here, http://www.dcggroup.com/news/big-data-hoarding-the-elephant-in-the-room.

❸ **The economics of storing data have changed.** Hadoop and cloud storage systems have transformed the economics of storing vast amounts of data. Hadoop combines the power of open source with low cost commodity hardware to allow companies to store huge amounts of data. Amazon Web Services S3 and OpenStack's Swift make vast amounts of cloud storage affordable.

❹ **New types of data processing and analytics are available.** New big data technologies, such as graph and MapReduce, have made processing data cheaper than ever. Sifting through and transforming huge amounts of data doesn't cost a fortune. In addition, new types of analytics for text and sentiment analysis and dozens of machine learning algorithms have become far easier and cheaper to use.

The new forms of data, new technology, and new analytics all have huge potential, yet companies are struggling to make sense of the new world of big data and the existing world of high-signal data. Right now most businesses find that these three statements ring true:

- It is difficult to harness all the data available to create business value
- It takes too long to get the answers we need
- Our current environment is too costly and complex

The systems for data and analytics that most companies have were not built to handle the amount and variety of data now available. As a result, for many years, the tendency has been to solve one problem at a time. But in many companies, the one problem at a time approach, which we call the application centric approach, has led to massive duplication of data and separation of data into many silos. You may not notice the pain from the application centric approach immediately because the cost of all these separate solutions is spread over many budgets. What becomes clear at a global level is that costs keep rising and staff still seems frustrated about getting the answers they need.

The way out of this situation is simple in theory. When trying to satisfy an organization's thirst for data and analytics, don't just solve the problem at hand but consider how you can solve it and create a system that stores the data so that analytics can be applied easily today and in the future. The goal is that all data can be accessed at any time and all the relevant analytics can be applied to any of the data. **This is the data and analytic centric approach.** But applying the data and analytic centric approach requires new thinking to master the increased complexity of big data.

The benefit of the data and analytic centric approach is that it addresses the data deluge and allows progress to be made with respect to the three challenges mentioned earlier, enabling companies to move closer to harnessing all data to create business value, getting the answers they need in a timely fashion, and simplifying their environment.

The data and analytic centric approach represents an important shift toward a company wide mindset. Data is viewed as a strategic asset that becomes vital to the day-to-day operations of the business, which is grounded and stabilized by the use of a well-understood collection of data. This reliance on high quality data not only increases the clock speed of the company, but also gradually induces a cultural shift. Important analyses, decisions, and actions always begin with the same question: What does the data tell us?

Data is a strategic asset that becomes vital to the day-to-day operations of the business

Falling into the App Trap

An application centric approach is perfectly natural. Many IT organizations pride themselves on being responsive and providing speedy service. The application centric response is simply to build a straightforward solution—but this creates an unsustainable foundation and data anarchy.

For example, suppose the sales department needs a better forecasting application. You assemble the data in a data silo, do the needed transformations, work to create a model, test it, and, if all goes well, the sales department is happy. Then the head of manufacturing needs an app to help with predictive maintenance. You use the same approach. Then marketing calls, then finance, and on and on.

An application centric approach builds analytics silos. These silos are modeled after the operational source. This provides value to answer a set of questions against that silo's set of data but it does not support answering questions across multiple platforms.

The application centric approach creates lots of replicated capabilities, lots of ETL and software to maintain, lots of servers filling up data centers, consuming power and administrator time, and perhaps worst of all, lots of little collections of data all over the place.

Here's another way to think of it. Every organization has many special purpose spreadsheets that are created to solve specific problems. There is nothing wrong with this approach with respect to each individual problem being solved. Looking at the big picture, what we find is a multitude of disconnected, hard to maintain spreadsheets that don't contribute to a strategic company wide view and don't make it easier to answer more questions. All too often, big data projects follow the same pattern. They create the equivalent of many massive special purpose spreadsheets that can process terabytes of data, but don't create a large asset.

The cost of such sprawl is only the most obvious problem. When it comes time to ask larger questions, those that require connecting dots across the enterprise, you cannot bring everything together.

From a financial point of view, the application centric approach just builds mounting levels of big data debt without a corresponding return. The kinds of things we seek from modern business intelligence solutions, such as self-service and reuse, are not possible.

As the cost and inflexibility rises, the results are predictable. You find that people use spreadsheets to solve problems that spreadsheets were not designed to solve. For example, using a spreadsheet to integrate data may mean that the spreadsheet takes hours to process the data or that all requests for new applications are highly intermediated. The analysts, developers, and DBAs—the only people who can actually build and change the apps—are overburdened and become a serious bottleneck.

The worst part of it is that the arrival of new or better data, instead of bringing a sense of excitement, creates a sense of dread because the only way to make use of that data is to dive into a morass of hardwired applications. More data doesn't add up to more value. That's not why we spend so much time and money on data.

A data and analytic centric approach requires a company wide mindset

How a Data and Analytic Centric Approach Increases Big Data Potential

A data and analytic centric approach requires a company wide mindset. The idea is to satisfy the individual requests of the organization in a way that preserves as much as possible the following ideal:

An analyst or end-user should be able to access all the data with all the analytics available.

So, if a request came in from the sales department for a forecasting application, you wouldn't just build it in a hardwired fashion as fast as possible. Instead, you would think about the data involved and determine how to store it so that it could be useful for this application and for others in the future. You would think about how the data involved should be connected to other existing data sets, but also how it might need to be joined with others in the future, considering how it might be integrated. Perhaps more fields would be preserved just in case they were needed.

When it comes to building a predictive model, the same approach would hold. If it were possible to make parts of the model into reusable components, you would do so.

Would such a data and analytic centric app be more expensive to create than a hard-wired app? Perhaps, if you were just comparing the cost of one app. But now let's look at what happens as the second, third, fourth, or tenth app is created. Gradually a data and analytics infrastructure is created to support the apps your company needs. Instead of the cost of lots of small replicated silos, you have a central infrastructure that is a starting point for a new app.

Will every app be able to get a head start? Yes, the cost of building subsequent apps in a data and analytic centric way is much lower than with the application centric approach. When new data arrives, it can be incorporated into the infrastructure. Agility is built into the system. You can connect the dots across all the data the enterprise has to offer. More data will mean more value, which is as it should be.

The success of the data and analytic centric approach rests on three principles:

1. Put data and analytics at the center
2. Create an agile environment to drive innovation on demand
3. Simplify your infrastructure

If you adhere to these principles, you can get as close to applying any analytic to all of the data you have.

The Need for Multi-Faceted Data Integration

The foundation of a data and analytic centric architecture is a multi-faceted approach to data integration. In the past, integration of data happened only one way (data modeling using third normal form and other methods) and in one location (the data warehouse). Now we have more data to integrate that is far different than anything we have faced before. We need multiple approaches to integration so that we can have the right degrees of integration to support our work.

> *The foundation of a data and analytic centric architecture is a multi-faceted approach to data integration*

In one sense, integration is about balancing an investment across a portfolio of data. Data that is high-value and high-use, in other words, which has high business value density, should be carefully modeled so that all the most frequently asked questions can be quickly answered. Data that is useful but used less often should be available for use without much effort.

Analysts celebrate the central importance of integration because it makes data accessible when needed and increases clarity and communication.

> *"Enterprises are spending more time integrating big data sources than on any other big data management function, including loading, transforming, processing, or securing. Integrating a dozen structured, unstructured, and semi structured big data sources with varying velocities and extreme volumes can be very challenging. Although enterprise architects and their companies can integrate big data sources by writing code and leveraging big data application programming interfaces and connectors, such an approach is time-consuming and requires skilled developers. The bottom line? In order to succeed in big data initiatives, look at big data integration solutions that can help automate and simplify the integration process, reducing the time-to-value."*
>
> Forrester Research, "Market Overview: Big Data Integration — Your Big Data Strategy Is Not Complete Without Big Data Integration," Dec 2014

The importance of data integration across multiple repositories is massive, which is why analysts are so focused on it, as the following quote reveals:

> *"Fragmented data—Many large enterprises already have petabytes of data stored in various big data repositories, and this is likely to grow to exabytes in coming years. However, not all of this data is in Hadoop; enterprises still store high-value data in databases, data warehouses, and legacy systems to support their low-latency data platform for MDM, transactional applications, and other critical real-time applications. This hybrid data management architecture leaves data spread across many repositories, creating silos that become difficult to integrate."*
>
> Forrester Research, "Market Overview: Big Data Integration — Your Big Data Strategy Is Not Complete Without Big Data Integration," Dec 2014

One of the explicit themes of these quotes is that big data is a threat to integration and getting value from data, just as the willy-nilly use of spreadsheets has been from their inception. To stave off these threats, the accessibility and common language achieved through integration needs not only to be present in established forms of using data but also for the new repositories and forms of data processing that have been created to handle big data. The solution is to apply a three-pronged approach to data integration that applies tightly coupled, loosely coupled, and non-coupled forms as needed.

Data Integration in Video Games

Consider how a video game company uses data. The company has high value data about customers, such as daily and monthly active users, sales of games, who is buying what where, who is calling tech support, and all of the associated financial information. This data has high signal or high business value density. Most of this data comes from established enterprise applications.

The company also has game telemetry data that answers questions such as who plays which game, how they play the game, and how much time they spend playing. This data, which is stored in a big data repository like Hadoop, is also high value and must be integrated with the data from the traditional applications so key questions can be answered. For video game companies, both of these types of data must be tightly integrated.

An even more detailed level of game telemetry data captures each move in every game, recording every time a player presses a button, shoots a gun, or throws a pass, along with the results. A game designer might want to ask a question such as “Is it too hard or too easy to hit a target from 100 meters with a sniper scope?” Answering this question requires diving into a mass of detailed game telemetry data and doing some work to get an answer. This data is stored to answer such questions, but is loosely coupled.

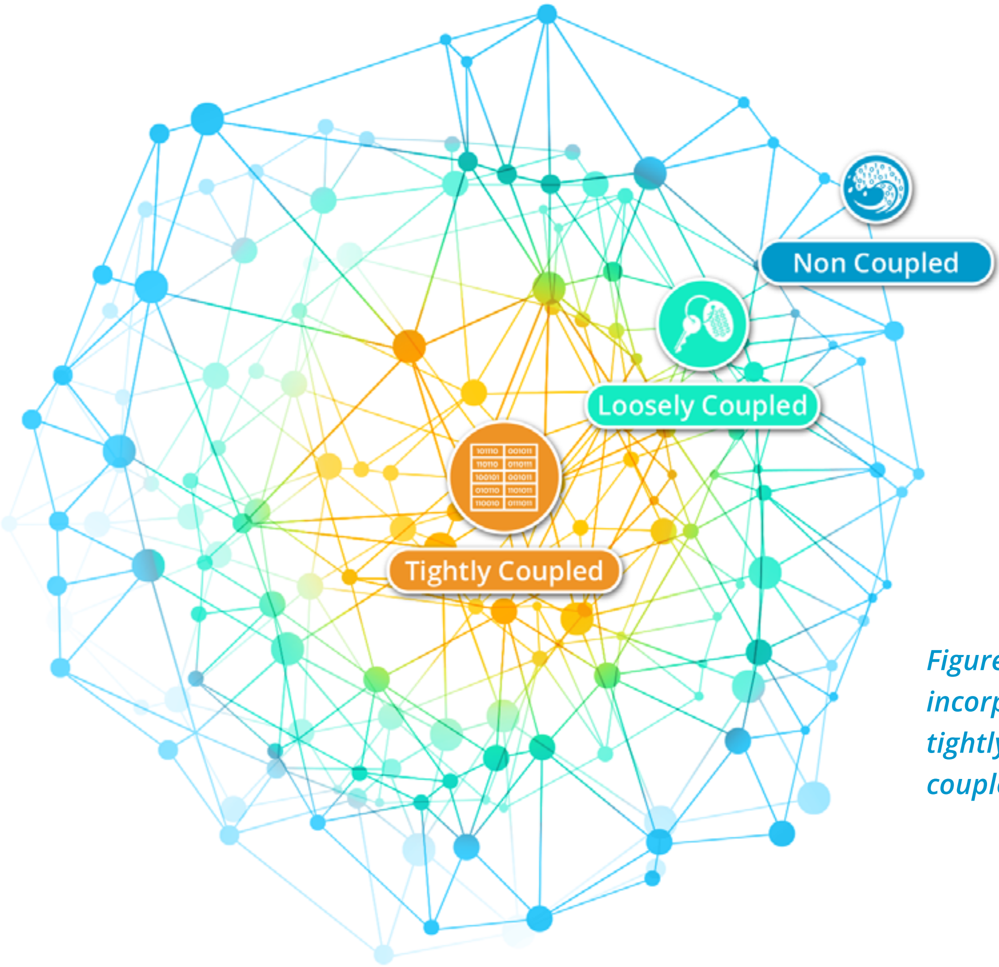

Figure 1. Data integration incorporates data that is tightly coupled, loosely coupled and non-coupled

Tightly Coupled Data

The form of integration that has been in use for decades in the data warehouse is tightly coupled integration. It's a way of creating a fine-grained, integrated data model that is well understood and heavily used. Tightly coupled integration includes both the information we use most often every day, and the forms of modeling used to create a data warehouse and optimize its performance, ensure consistent results, make navigation across data easy, and enforce security when needed.

In a tightly coupled integration, the model to contain the data is carefully designed. The data poured into the model is thoroughly vetted to ensure that it is high quality.

Tightly coupled integration requires upfront effort, but the cost per use is low because usage is so high. The investment in tightly coupled integration makes sense because it will make answering questions easy for thousands of people who access the data millions of times.

The danger in applying tightly coupled integration is going too far and making the model too big from the outset. This approach can "over model" a problem, which leads to delays and higher costs. It is better to under model and put data that has suspected value in a loosely coupled form until it's clear that tighter coupling is needed. This approach provides agility and reduces costs.

Loosely Coupled Data

Loosely coupled data refers to data that clearly has some valuable elements but also has parts that are either low value, changing frequently, or not clearly understood. Big data repositories are often the right location for loosely coupled data. The goal is to invest enough to allow easy access to the information needed.

It is better to under model and put data that has suspected value in a loosely coupled form until it's clear that tighter coupling is needed

Let's say we are doing an analysis of the high value customer segment, but want to get an idea of how active customers are. We want to create segments for high value/active in the last 90 days, high value/active in the last 12 months, and high value/not active.

Because we have been paying attention to the right place to store data, we have stored data such as web logs, social media interactions, call center records, and any other evidence of interactions in our big data repository. While each of these forms of data has widely varying structure and lots of different fields, we have done some preprocessing so that we can access each form of data using a customer identifier. Then, by scoring the intensity of each interaction and counting how many interactions took place in the time period for each segment, we can come up with a segmentation of the high value customers we are looking for.

In the loosely coupled paradigm, you don't create a comprehensive model. For each of the data sets involved, just enough of the data is modeled to make it useful for the analysis or application at hand. You create just enough consistency across all the data sets involved so that they can be joined. It also might make sense to pre-process and create fields that you suspect might be useful in the future for other analyses.

Examples of data that typically should be loosely coupled include web logs, social media data, call detail records, and the like.

Non-coupled Data

With non-coupled data, no modeling has been done. In other words, in loosely coupled integration we find the keys that can be used to make sense of the data and the fields that interest us. In non-coupled integration, we store data that might be needed in the future in some type of low-cost storage. We store the data in its raw original fidelity. Regardless of where it resides, we make the data available for exploration by analysts.

Non-coupled data can be harvested for signals that can be of use on their own or perhaps delivered into a loosely coupled or tightly coupled repository.

From Non-Coupled to Loosely Coupled: Uncovering Secrets of On-Time Performance

Suppose we had the engine sensor readings for a fleet of trucks. We might explore the data to find some metrics that are common to drivers with the best on-time performance. However, sensors don't tell you who the driver of the truck is for a given trip; that's stored in a human resources table. Another data set has a Driver ID associated with a Trip ID. But there's still no Trip ID (yet) in the sensor data, so someone has to create that key on the fly to associate the sensor data with a specific trip and driver. An analyst might derive Trip ID by looking at sensor readings such as date, time, and location of beginning and end of trip. By associating drivers to trips in this way, we are taking non-coupled data and making it loosely coupled. In this way we could analyze the data to discover the driving secrets of the best drivers in order to train other drivers about how to have a better on-time record.

How Each Type of Integration Works Together

Tightly coupled, loosely coupled and non-coupled data work together in a data and analytic centric approach. They work together dynamically and data migrates based on usage. By adding keys to non-coupled data (making it loosely coupled), it is easier to join data across tightly and loosely coupled data sets on demand when analysts wants to drive deeper insight without having to model all the data. If the analysis is recurring or widely used, it might be more cost effective to make the data tightly coupled, putting it in an enterprise data warehouse that supports many simultaneous users.

We need three types of integration because the amount of investment to make in each type of data varies. The higher the business value density, the more investment is justified because the data will be used so often. The goal is to use the right approach to modeling and integration so that the data provides maximum value at minimum cost. Finding the right way to use a data set is a skill that improves over time.

You know you have the balance right if there is an increase in agility. Business agility (the ability to answer questions quickly) arises from both tightly coupled approaches (data re-use, query performance, ease of use) and loosely/non-coupled approaches (ingest data immediately and start exploring, define new ways of structuring data in ways others did not conceive). If you have all the data where it needs to be, when you build that next application, you should be able to find much of what you need and get the right data assembled much faster than in the past.

Once you have data in the proper place with the right amount of structure, the challenge then becomes determining what type of analytics will unlock its secrets.

The Need for Analytic Flexibility

The analytic centric message is parallel to what we have discussed about data centricity. Just as we have faced an explosion of data, we are also living in a golden age of analytics. Advanced analytics have been with us for a while, but they have rarely been easy. What is truly new about the current golden age is the massive expansion of awareness about the power of analytics, the productization of many different types of analytics through open source and other means, and the availability of an unprecedented amount of data.

Just as we have faced an explosion of data, we are also living in a golden age of analytics

Right now, the application centric approach is leading many companies to build the equivalent of GPS-style black boxes for analytics. Analytic silos are being created to perform a certain type of analysis on a specific data set. For example, at firms that spend lots of time analyzing social media, one can build a fabulous system for graph analytics or create a great black box for sentiment analysis.

But if you build lots of black boxes, you will be forever in the business of moving data in and out of them. In a tangible way, these analytics black boxes repeat on a larger scale with more complexity the same mistake that occurs when spreadsheets are used to solve too many problems. A far better approach is to build an infrastructure that allows any analytic to be applied to all of the needed data.

The Mindset for Analytics

We must apply the same mindset to analytics that we did to data. Instead of hardwiring analytics into an application or creating a black box, we must make people performing analytics more productive in following ways:

- **Apply the right analytic to the right data integration type.** Instead of a black box, we want analytics that can reach out for all the needed data and then apply the analytic. This approach minimizes data movement and data duplication.

- **Leverage multiple analytic techniques to get insights**. Expand access to analytics through applications that follow the loosely coupled principle of creating just enough structure to answer frequently asked questions.

- **Provide self-service analytics for all skill levels.** You will not achieve victory if everyone must be an R programmer to do analytics. Instead, offers analytics systems that support a spectrum of users, from data scientists to business analysts.

Just as we worked to integrate data in a multi-faceted way that made it ready for use by future applications, we must also focus on ways to make the analytics we create ubiquitous, re-usable and easier to use. Again, for an individual application, this approach may be more expensive, but for a portfolio, the savings come in many ways. When the time needed to write and execute an analytic is reduced, productivity increases both for beginners and experts.

Analytics productivity comes from having pre-built functions, enabling rapid data preparation, and making it easy to explore data in its native repository so that you don't have to move it. In addition, the more work that can be shifted to repositories that are easier to use, the more will get done. You don't want to create an environment where Java programming is required when you can get the same job done with SQL or other widely available skills.

Figure 2 shows how data is leveraged by analytics, pushed into key analytic processes, and finally embedded in business decisions and business processes. Note that the figure puts business process on the left, because in fact to be data and analytic centric and thus data-driven, you must start by thinking about your business process first.

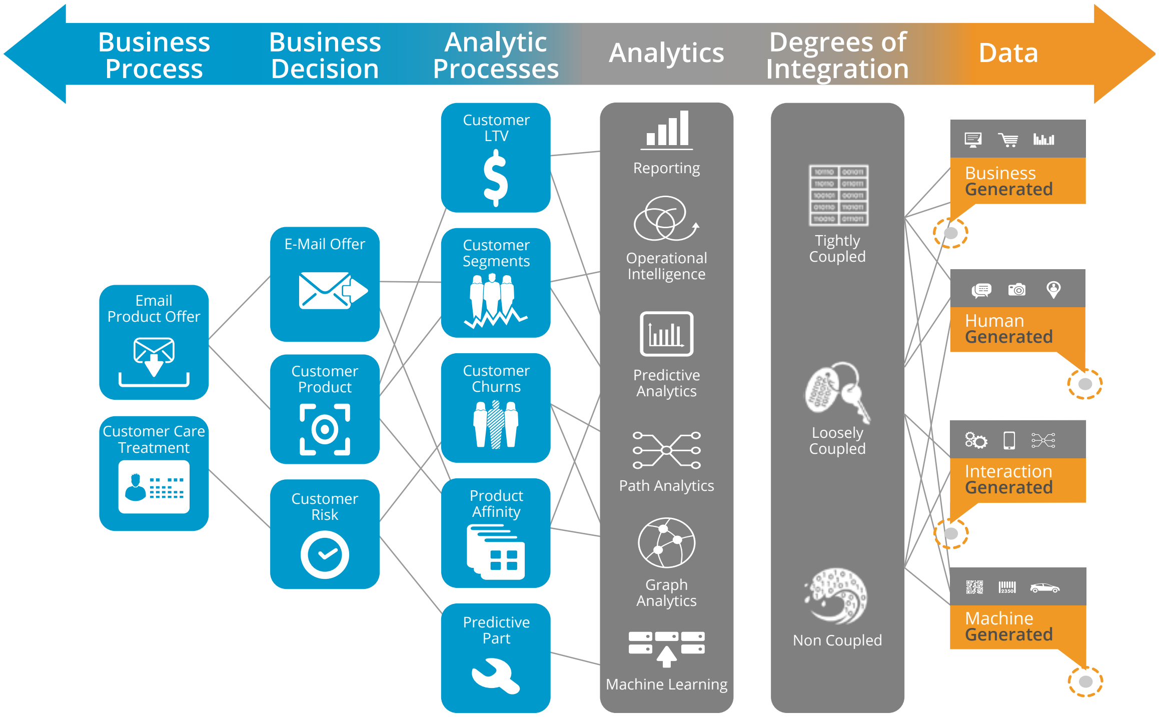

Figure 2. The results of a data and analytic centric approach: Data-driven business

Putting Analytics to Work

How do we apply analytics in a world in which we have integrated data and reusable and carefully designed analytics? The ideal, of course, is that we are able to apply analytics to data at the right degree of integration. As we said earlier, integration is the key, but push-down processing also plays a role.

Building on our example about the high-value customer analysis, suppose that we looked at not only how recently high-value customers were active but also added two other elements. We could analyze the reactions of high-value customers, looking for evidence that they were unhappy and possibly at risk of churning. We could also do a graph analysis of other customers related to the high-value customers because churn tends to affect others who are closely related. Then, based on the results, anti-churn methods could be applied to unhappy high-value customers and those related to them. With a data and analytic centric approach, such a query could be constructed by an end-user.

An Optimized Portfolio for Reduced Cost and Complexity

The essence of the data and analytic centric approach is to understand the nature of the problems your organization is trying to solve and then invest appropriately. People have wasted money in the past by over-modeling and over-integrating, which can lead to an architecture that is hard to change. Right now, people are falling prey to the practice of under-modeling and under-integrating, which increases both costs and complexity, creating a mass of big data debt that will have to be repaid later.

If you understand the data that is most important and the way it can be used at scale, the data that has the most business value density, you can save by investing in tightly coupled integration. But until you are sure, you can keep things loosely coupled or non-coupled. This is the true spirit of the agile approach: waiting until experience demonstrates what is really needed and then building it incrementally.

By adhering to the principles we have discussed, it is possible to become data and analytic centric. The best result is the assurance that you have made the right investments, investments that will yield long-term value for your organization.

Conclusion

Adopting a data and analytic centric approach requires a deeper analysis of user needs, one that takes into account both the immediate business purpose being served as well as future needs that may arise.

Adopting a data and analytic centric approach requires a deeper analysis of user needs

By seeking to understand the personality of data and how it will be used, designers of a data and analytic centric platform can not only use the well-established tightly coupled approach, but also employ loose and non-coupling in order to embrace data at varying levels of maturity.

Constructing a data and analytic centric platform achieves three key objectives:

- **Reduce time to meet new needs.** As the data and analytic centric infrastructure grows, the nervous system becomes more powerful and can be quickly adapted to meet new needs.
- **Decrease complexity and cost of infrastructure.** Because application centric silos are avoided, data and analytics can be used broadly and at a lower cost.

- **Increase productivity.** More people can benefit from analytics because time has been invested upfront in making analytics easier for people with various skillsets. Additionally, new data can be integrated at a lower cost since time and money are not being wasted over-modeling data that will not be used.

Remember: this is a new way of thinking. It is possible to learn from the experience of others and get help from advisers to increase the speed of progress toward a data and analytic centric nervous system and also to reduce risk. It is essential to have a clear and coherent roadmap and architecture to get the best results. Experienced advisers can accelerate such planning.

For more information, see http://bigdata.teradata.com/ ▸

This paper was created by CITO Research and sponsored by Teradata

Teradata helps companies get more value from data than any other company. Teradata's leading portfolio of big data analytic solutions, integrated marketing applications, and services can help organizations gain a sustainable competitive advantage with data. Visit teradata.com. (http://www.teradata.com)

Chris Twogood

Vice President of Product and Services Marketing, Teradata Corporation

I believe that Big Data can help us provide better customer experience, drive lower operational costs and improve margins. This leads to increased customer loyalty and enables companies to improve business advantage. I have been with Teradata for over 25 years helping customers understand how they get value from data. I drive Teradata's Marketing focus around how companies become data-driven, and this includes leveraging best of breed solutions such as Teradata, Aster, Hadoop and how that all comes together within a Unified Data Architecture.

Dan Woods

Chief Analyst and Founder, CITO Research

I do research to understand and explain how technology makes people more effective in achieving their goals. My strong belief is that we are at the threshold of a golden age of IT, in which the promise of gaining value from technology will be fulfilled. I have written or coauthored more than 20 books about business and technology, including APIs: A Strategy Guide. I write about data science, cloud computing, and IT management in articles, books, and on CITO Research, as well as in my column on Forbes.com.