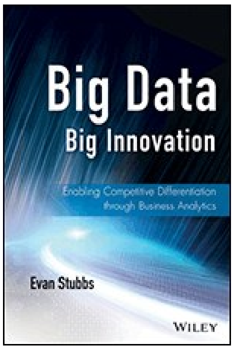

Big Data
Big Innovation
Enabling Competitive Differentiation through Business Analytics
Evan Stubbs
WILEY

Big Data, Big Innovation: Enabling Competitive Differentiation through Business Analytics

by Evan Stubbs

John Wiley & Sons (US). (c) 2014. Copying Prohibited.

Chapter 5: Organizational Design

Overview

It's one thing to know you need something. Knowing where it should go is something else. Not to understate how hard it is to change culture, there's obviously more to success than just setting the right direction. Flourishing through the age of uncertainty requires an *excellent* ability to analyze, predict, and act.

The worst thing to do is to run out and hire people just because someone thinks there's a gap. Instead, the focus should be on making the existing people successful by giving them the right support. They need the right structure, the right focus, and a management mandate to make things happen.

Getting the design right helps tremendously. When correct, it creates economies of scope and scale. These enable structural cost advantages that in some cases can actually create differentiation in their own right. In reinventing an organization, there are four things that should be considered:

1. What should it look like?
2. What should it focus on?
3. What services can it offer?
4. What data does it need?

This chapter will answer these questions and lay the foundation for effective organizational design.

What should it look like?

Every organization carries baggage. While there's sometimes the unique opportunity to build things from scratch, more often we need to work with what's already there. Getting things right means understanding *where we are* as well as working out where we *should* be. Because of this, moving from generating insight to creating value from big data usually requires change.

At its simplest (and least intrusive), this may involve splitting an existing team into two. One group might focus on change and transformation, the other on generating insight. At the other end of the spectrum lie organizations that totally reinvent themselves around using big data and business analytics as a competitive differentiator.

Sometimes, small steps make a lot of sense. Other times, realizing a big vision requires a big step. When resources are scattered across the organization, structural change can be daunting. Luckily, there's an easy way to consider the benefits and disadvantages of various models. Every group needs a manager, and every group needs to cover its costs somehow. Through these, the major options are presented in Figure 5.1.

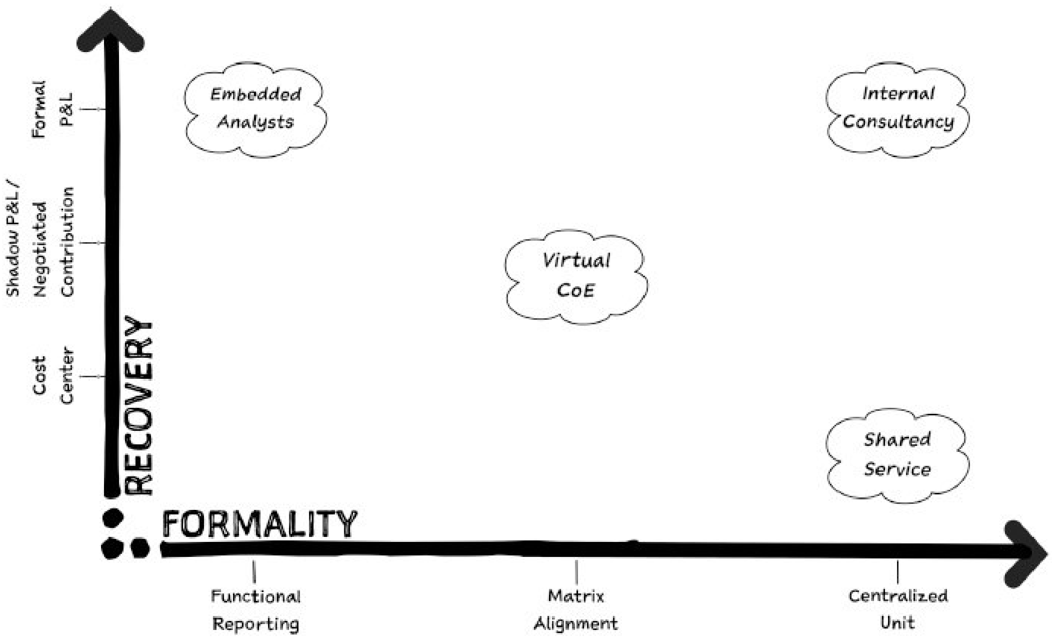


Figure 5.1: Structural Choices

Group Formality

The least intrusive method is functional reporting. Analysts are embedded within business units. Often, they're not even formally recognized; they're just the person who knows how to code in Excel. There *are* advantages to this model. For one, it gives business units freedom to invest in the areas they think will provide the greatest value. Unfortunately, it also increases costs in the long run. If and when multiple business units decide that they need access to similar skills, they usually end up hiring similar people. Local efficiencies, while flexible, rarely lead to global efficiencies.

The midway point involves establishing a virtual group. Usually, this has people report to multiple groups through matrix management. Moving everybody into one group is seen as a step too far, too soon. Instead, analysts report to two masters. Analytical resources and headcount continue to reside in functional lines of business (such as marketing or fraud). These business lines manage the analysts' day-to-day workload.

The organization also maintains a secondary line of reporting through to a centralized group, responsible for strategic direction. Rather than internalizing all analytical activities, this centralized group instead focuses its much smaller headcount on aspects usually overlooked by functional lines of business, often including:

- Establishing a common enterprise value measurement framework
- Defining and delivering on an enterprise analytical architecture
- Evangelizing analytics across the organization
- Supporting change management and organizational enablement

This model offers some good advantages. For one, it creates relatively minimal disruption. Additionally, it fills out the organization's capabilities by investing in areas functional units are unlikely to care about. Enterprise transformation, if successful, is valuable. However, it's a rare business unit that's happy to pay for group benefits.

The model does come with disadvantages. The complexity of managing such a model is nontrivial. Resource contention often becomes a major issue when group and functional demands conflict. Directing investment is also challenging; in practice, the centralized group often ends up acting as a diplomat or negotiator, trying to persuade lines of business to "get along" and invest in the right areas.

The most structured model involves establishing a centralized group. The biggest advantage of this model is that without it, it's almost impossible to achieve real economies of scale or scope. This applies to many functions; it can include areas as diverse as:

- Data science and experimentation

- Analytical data management
- Advanced analytics and predictive modelling
- Quality control
- Optimization
- Business intelligence and dashboarding
- Insight operationalization

Of course, this comes at a cost. Scale creates bureaucracy. It also needs an owner, someone who's willing to make sure the group's creating value. An underleveraged group is just more cost and in most cases has a limited life.

A prime example is an analytical center of excellence, covered in greater detail later in this chapter. Often headed up by a chief analytics officer or chief data scientist, it usually exists as a separate functional line of business in its own right. These centers may report directly to the CEO or fall within another line of business with a focus on shared services such as IT. Usually, they're created to pull all analytical resources into one group, tasked with supporting the business. The group maintains its own headcount, budget, and cost center or, more ideally, profit-and-loss statement (real or shadow).

Each approach offers different advantages and drawbacks; none of them is better than any other. For example, establishing a formal model often requires major organizational change and investment.

This formality comes with benefits. For one, it's easier for the rest of the organization to engage with the centralized group. The virtual model requires significantly less investment and preserves existing powerbases. It does, however, increase management complexity and can make engagement challenging. The functional model allows flexibility but prevents scale.

Usually, the right model is dictated by management commitment to business analytics and organizational politics. For many organizations, the virtual model is simply a steppingstone to drive cultural change. For some, the virtual model doesn't go far enough; nothing less than true centralization will do. As will be covered in Chapter 8, the real answer often involves *multiple* models.

Cost Recovery

The second consideration is how analysts cover their costs. One approach is to act as a shared service. Resources are offered "free of charge." In this model a board, leadership team, or other governance committee manages activities and investment. Ideally, this oversight group focuses on value. They prioritize effort based on quantitative and value-based considerations such as strategic objectives or the quantum of return expected.

This is the easiest model to adopt. Given that the teams responsible for generating answers rarely own the outcomes, it's financially hard to measure these groups' profitability. Unfortunately, it also frequently discourages long-term investment; because the group isn't linked to revenue or profit, it's seen as a cost center. Without significant cultural commitment the broader organization is usually reluctant to invest.

An alternative model is to operate using a shadow profit-and-loss statement. Usually, this is based on negotiation and is approved by other lines of business. While not necessarily appearing in the general ledger, the group has a management mandate to demonstrate return on investment. Costs are registered but successes are credited against the group through a shadow tracking system. By doing this, the group can still demonstrate financial outcomes and success despite not having direct control or ownership over revenue streams. Admittedly, there's a heavy emphasis on negotiation and the perpetual temptation to game the system. However, this midpoint at least allows the leadership to track the value of business analytics.

The most sophisticated approach is to establish a formal profit-and-loss statement. Under this model the group is charged with demonstrating internal profitability. Rather than offering services for free, the group uses group allocations or internal resource request-based pricing to charge out its time to other business units. Key performance indicators are often defined as a blended model, balancing total return on investment against maintaining an agreed realization level.

Tactically, the group needs to remain solvent. Strategically, the group needs to be able to demonstrate how its actions have delivered economic returns. In many ways, this model requires the group to act as a chargeable internal consultancy, actively seeking out business and needing to demonstrate return on investment.

This model is a challenging one. The biggest advantage it offers is direct accountability. Unfortunately, it also drives profit-maximizing behaviors. Leaders of the group will naturally chase their biggest customers, neglecting areas of the business that aren't interesting. While it often ensures cost neutrality, in the absence of a broader cultural commitment it rarely leads to organizational transformation.

Each approach offers different advantages and disadvantages. The biggest advantage of the shared service center approach is ease of engagement. Because resources are free and activities are prioritized through a well-defined process, business units have fewer barriers to trying to leverage business analytics. Equally, though, this often increases the complexity of demonstrating return on investment from business analytics. At its worst, demonstrating success becomes a lobbying process. The business analytics team spends the majority of their time convincing other business units to publicly support the business analytics group regardless of outcome.

Running a separate profit-and-loss statement limits this bad behavior. Return on resources is easily demonstrable based on utilization and project success. However, this upfront cost can act as a significant barrier to business unit experimentation, especially in climates of constrained budgets. When budgets are tight, most business units will resist having to pay to do things differently. If this approach isn't supported by a corresponding culture, the group runs a very real risk of self-optimizing and only working with those business units that are most willing to pay, undermining the whole point of an enterprise approach to business analytics.

What should it focus on?

Embedding analysts in business units is a valid option. For one, it's easy—it doesn't require any broader strategy. Just hire the person and set her to work.

Unfortunately, it does little to help with the trends discussed in Chapter 2. While it does allow a great deal of flexibility, it does little to encourage reuse, human capital development, or economies of scale and scope. Left alone without management support or a mandate to work otherwise, people will normally work independently.

The rise of rōnin will eventually force most organizations to think about trying to centralize and reuse their analytical capabilities. Technology is infinitely reproducible; people are not. That's not to say that embedded analysts are a bad thing. As a hiring model, it's an excellent augmentation to centralized approaches. There just aren't enough analysts in the market to realize every opportunity through continually hiring new people.

The decision to set up a central group (in some form) is a logical conclusion. It does, however, inevitably lead to the question of what it should focus on. Every group needs a purpose.

It's helpful to consider a shared group's function along three lines: (1) They can help build knowledge; (2) they can help deliver; or (3) they can help transform the organization. Organizations that decide to centralize their capabilities often call the result Communities of Practice, Competency Centers, or Centers of Excellence.

It's important to recognize that these definitions aren't absolute; you say tom*a*to, I say tom*ah*to. Definitions vary and, as yet, standard names do not exist. They're used here to highlight how structure and focus can vary even when there's a defined departmental or enterprise-wide capability.

Communities of Practice

The lowest-touch model is a community of practice. Communities of practice tend to focus on helping practitioners share and learn from each other. Right or wrong, they're usually the starting point for a leadership team that has realized the value of business analytics but is concerned with making structural changes.

Their primary objective is to nurture skills. They try to cross-pollinate knowledge between those who would, in the absence of the community of practice, rarely cross paths. They do this through regular meetings or conferences. Their focus is to try to get people to network and to share their experience.

Their main attraction is their low-impact nature; they require no structural change whatsoever. It's usually more like a shared club where attendance is encouraged but optional. They develop social capital and promote collaboration.

At best, this is only a halfway house. Their biggest weakness is that they rarely drive any behavioral changes. Awareness is one thing, change another. More often than not, people revert to their comfort zone after attending the get-togethers. This isn't because of a lack of enthusiasm or a resistance to change. It's simply because it's easier to keep doing what one is doing. Many change agents often overlook this limitation and assume that because all the right people have been brought together, change is inevitable.

Despite being a somewhat halfhearted approach, communities of practice still have their benefits. In the absence of anything else, they help to develop awareness of the importance of business analytics. While they tend to be focused on specific applications such as risk, marketing analytics, or business intelligence, they help by mitigating one of the biggest constraints in any large organization—functional separation. Because of this, the function of a community of practice is often eventually blended into a more structured model such as a competency center.

Competency Centers

Competency centers go further. They change the structure of the organization, drawing similar resources into one group. Their model is usually to act as a shared service center to support the broader organization.

Their main attraction is consolidation and enabling economies of scale. Unlike "insights teams," these centers have a narrow focus, usually defined functionally. They may specialize in predictive modeling. They often specialize in business intelligence. They may focus on machine learning. Integral to this focus is a longer-term strategy that outlines how the group will move toward best practice. They often go beyond ad hoc support to include actual delivery. While they don't own the outcome, they'll usually be responsible for making sure their work makes it into production.

There are significant advantages to this model. By drawing common skills into one group, the organization starts developing economies of scale through specialization. Clarity of focus also helps other groups get engaged. When skills are scattered, it's often hard for people to take advantage of latent capabilities. It's far easier for other groups to get engaged when there's a single team to contact.

It's also a tangible demonstration of strategic intent. Creating a defined group does wonders to clarify what the enterprise sees as a potential competitive advantage. It gives the organization a hook to latch onto and experiment with. Even if others don't necessarily understand the domain, they at least know it's there for them to take advantage of.

Despite their advantages, competency centers are still limited. For one, they maintain a siloed delivery approach. Because of their functionally defined focus, their engagement tends to align with traditional business applications and ignores developing enterprise-wide competencies.

For example, a risk competency center has the potential to add tremendous value across the business. In addition to traditional scoring and simulation activities, they could add real value through driving risk-based pricing and augmenting financial planning to incorporate boundary testing. Unfortunately, this rarely happens. In the absence of specific direction, the team will usually gravitate toward traditional risk management processes such as managing operational risk or identifying behavioral or application risk. When one's goal is being utilized rather than driving change, it's easier to sell to current rather than potential customers. Because of this, much of the cross-functional potential of business analytics is lost.

Another disadvantage is that because their domain is taken as a given, the team usually pays little attention to evangelism. A mandate is both a blessing and curse. On one hand, it establishes responsibility. On the other, it's easy to assume that the rest of the organization will be just as interested and supportive.

In practice, this is rarely the case. By definition, any sophisticated area of expertise is niche. Not everyone in the organization will understand it, let alone value it. Business analytics is fundamentally about change and driving change requires proactivity. While it's not inevitable, competency centers often overlook the importance of evangelism and sales. Instead, they fill the role of a pure shared service center, responding to work requests as they file in. They end up being great at supporting business as usual and "known unknowns," but transformation and tackling the "unknown unknowns" usually just becomes too hard.

Centers of Excellence

The most capable model is a center of excellence. It blends the best of a community of practice with a competency center. It centralizes resources into a shared services model while also taking responsibility for improving the broader organization's knowledge and capabilities in their targeted domain.

However, it also goes beyond this by adding:

- An explicit focus on (and a resource structure to support) communication and evangelization
- Ownership over defining a common value measurement framework
- Responsibility for actively finding opportunities to apply business analytics across the enterprise

At first glance they often look similar to competency centers. Both centralize skills and both support cross-functional business units within the organization. There is a difference, though. Where competency centers are usually fairly reactive, centers of excellence are highly proactive. Where competency centers are content to respond to business requirements, centers of excellence will actively find and deliver incremental value across the business. They provide support. However, they see their primary role as being an agent of change.

A second major difference is that they tend to have a broader focus than most competency centers. This is by no means guaranteed; some maintain a very narrow focus. When tasked correctly, though, there's a key driver that discourages too much specialization—a focus on value creation rather than centralizing skills.

Competency centers are traditionally defined functionally. They draw similar skills together to capitalize on economies of scale. By contrast, centers of excellence are focused on value creation and return on investment. They drive both economies of scale *and* scope. Because of this, they tend to require access to more skills than an equivalent competency center.

They normally maintain the full set of competencies needed in business analytics. They'll maintain people knowledgeable about data management, data science, value measurement visualization, and even how to embed analytics into operational processes. And, these all build on top of specialist skills such as risk management, predictive modeling, or other domains. In contrast, a competency center will often specialize in only a small subset of those competencies, dictated largely by those most required by their targeted focus. When correctly designed, centers of excellence represent extremely skilled, powerful, and valuable groups.[1] The biggest disadvantage of centers of excellence is that they require a certain degree of scale to be successful. As will be covered in Chapter 7, it's almost impossible to find one person with all the skills needed to be successful in business analytics. Teams are the norm, and these teams need to have a certain degree of coverage across core roles and responsibilities if they're to be successful. This coverage requires investment, and most organizations need to have achieved a certain level of comfort before they're willing to take the plunge.

[1] For a good overview on designing centers of excellence and how this links to organizational performance, see Aiman Zeid, *Business Transformation: A Roadmap for Maximizing Organizational Insight* (Hoboken, NJ: John Wiley & Sons, 2014).

What Services Can It Offer?

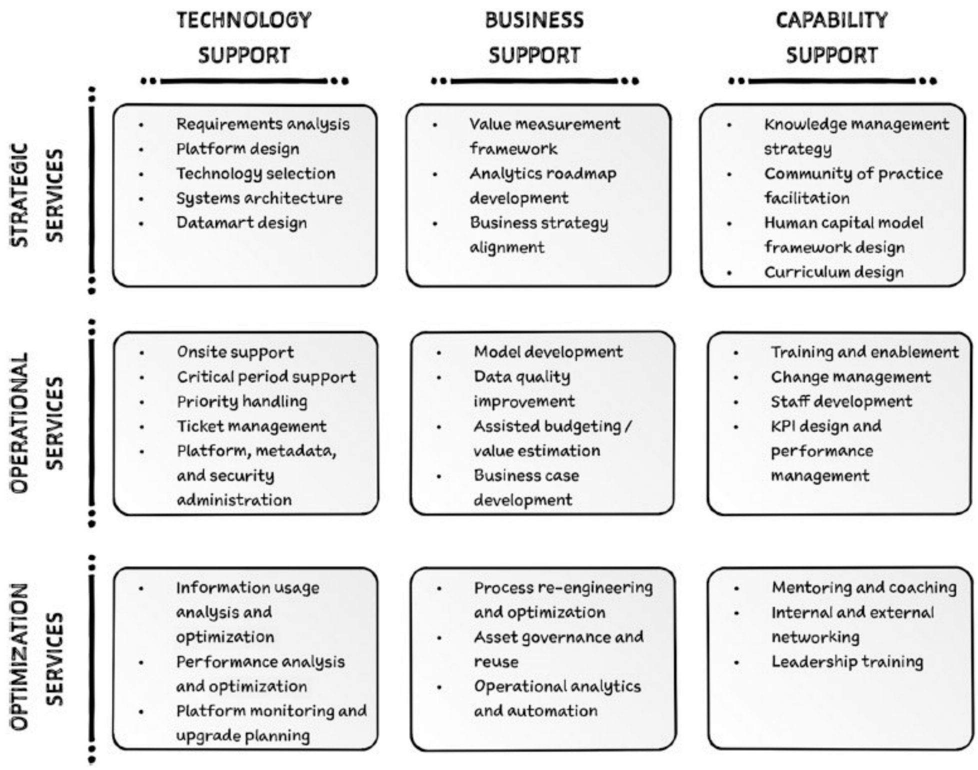
Having focus gives a group purpose. What it doesn't do is explain to the rest of the organization how to get engaged. To take advantage of a capability, people need to know what it is, how to use it, and how to get involved.

One approach is to simply throw a bunch of smart people together and hope for the best. While this works surprisingly well in smaller organizations, it rarely scales. The better approach is to make engagement easy through well-defined services.

Services are simply defined combinations of people, processes, and technology offered to customers with known outcomes. To create value, they support some form of business process. By reducing it to an offer with a clear value proposition, the group makes it easy to explain *why* and *how* other groups can take advantage of their capabilities.

Good support services cover the full gamut of platform support right through to identifying and delivering initiatives. Some are concerned with setting direction. Others are more focused on "keeping the engine running." Still others are focused on making existing things better. Operational excellence in business analytics brings all of these together in a way that drives a culture of continuous improvement and quality.

In defining these services, it's useful to consider the service design model described in Figure 5.2.



	TECHNOLOGY SUPPORT	BUSINESS SUPPORT	CAPABILITY SUPPORT
STRATEGIC SERVICES	• Requirements analysis • Platform design • Technology selection • Systems architecture • Datamart design	• Value measurement framework • Analytics roadmap development • Business strategy alignment	• Knowledge management strategy • Community of practice facilitation • Human capital model framework design • Curriculum design
OPERATIONAL SERVICES	• Onsite support • Critical period support • Priority handling • Ticket management • Platform, metadata, and security administration	• Model development • Data quality improvement • Assisted budgeting / value estimation • Business case development	• Training and enablement • Change management • Staff development • KPI design and performance management
OPTIMIZATION SERVICES	• Information usage analysis and optimization • Performance analysis and optimization • Platform monitoring and upgrade planning	• Process re-engineering and optimization • Asset governance and reuse • Operational analytics and automation	• Mentoring and coaching • Internal and external networking • Leadership training

Figure 5.2: Service Design

Every business analytics service does one of three things. It helps set direction, helps deliver, or helps identify opportunities for improvement. These "activity" services provide support across the major aspects of the business—strategic planning, operational execution, and continuous optimization.

Every one of these services is focused on one of three things: technology, the business, or developing capability. The intersection of each of these helps define specific support services that help create internal or external value.

Not every organization has complete coverage across all of these areas. And, not all of these need to be offered by the same group. For example, technology support services are often managed by IT. Capability support may be managed by human resources (HR). Business support may be managed by a center of excellence. What's important is identifying gaps and closing them.

Strategic Services

Strategic support services help organizations define their direction and establish an execution plan. They help by providing the organization with the skills and support to create a roadmap, develop new capabilities, and provide governance or funding models.

Broadly, they focus on:

- Defining direction and funding execution
- Encouraging a consistent approach
- Developing skills and knowledge

At the strategic level, technology support services focus on developing a technology and data roadmap that map against current and future organizational requirements. Through understanding the organization's strategic direction, they aim to align and fund the organization's technology and data architecture. In addition to defining technology and data roadmaps, they also often provide a clear governance framework through which ongoing upgrades and feature requests can be captured, prioritized, and funded (often through a formal steering committee or the like).

Business support services aim to establish consistency in approach across initiatives. Through defining and encouraging adherence to common processes, they aim to make execution more efficient, consistent, transparent, and effective. Typical focus areas include defining and establishing a common value and effort measurement framework and helping to define the high-level milestones that every business analytics activity will follow. They're often focused on helping the business to develop an analytics roadmap that links tactical value into strategic differentiation.

Finally, capability support services revolve around fostering skills and cross-pollinating knowledge. Through establishing the right cultures and processes, they aim to shift skills from individuals to the organization, creating a self-sustaining culture that values business analytics. Common focus areas include establishing and running communities of practice, helping to define skills development roadmaps, developing knowledge management strategies, and developing curriculums that blend technical and domain training.

Operational Services

Operational support services help organizations deliver value and meet business outcomes. Rather than defining the "to-be" state and facilitating the change needed to get there, they focus on supporting current activities and ensure the business can do its day-to-day business effectively. They help by providing the business with the support it needs to do a variety of administrative and operational activities.

Broadly, they focus on:

- Monitoring existing activities
- Supporting operational execution
- Delivering operational outcomes

In this category, technology support services focus on ensuring the organization's technologies and data repositories perform against expectations and requirements. Through monitoring and resolving technology and data-related issues and requests as they're made, they aim to ensure high availability and prevent platform-related delays. Common services include onsite support, maintaining software currency and managing platform upgrades, priority and trouble ticket handling, metadata and security administration, and critical period support.

Business support services aim to clarify and assist with navigating and executing operationally related business analytics processes. Through providing specialist skills, they aim to help the organization move through each of the major phases associated with creating value from business analytics.[*] Common examples include helping to develop business cases, implementing value measurement and performance management frameworks, and defining change management and communication plans.

These also extend to data science. Through providing resources with specific competencies, they provide "overflow" support to overcome resource-related bottlenecks. Common services include providing data quality, model development, and analytical data management skills. The group might, for example, create a "model factory" to streamline and simplify the creation of predictive models.

Finally, capability support services help develop human capital and knowledge. They focus on training and enablement, support change management and cultural development, and link performance management to outcomes rather than activities.

Optimization Services

Optimization support services close the loop by helping to drive continuous improvement. They focus on identifying potential improvement opportunities and helping to maximize leverage of existing resources and/or assets.

Broadly, they focus on:

- Profiling existing approaches
- Identifying opportunities for improvement
- Assisting with process and asset reengineering

Technology support services in this context focus on profiling and improving the use of technology and data-related assets. Through profiling current usage patterns and identifying bottlenecks and inefficiencies, they aim to help the organization uncover common usage patterns across individuals and groups and facilitate improvements. Common services include helping to rationalize and streamline data management activities, benchmarking platform performance to identify and resolve bottlenecks, and identifying inappropriate or inefficient uses of technology.

Business support services aim to rationalize and streamline business analytics processes. Through mapping and analyzing existing processes across multiple business units, they aim to uncover best practices and replicate them across the organization. Common services include optimizing analytical processes (including information management, model development, and model deployment) and driving best practices in asset management.

Finally, capability support services assist by helping an organization to mature its abilities and become more proficient in business analytics. Through profiling current competencies across the organization and linking these to an agreed strategic plan, they aim to help the organization define specific actions to increase sophistication and experience in business analytics. Common services include facilitating mentoring plans and facilitating internal and external networking to promote cross-pollination of new ideas.

[*] Defining the value, communicating the value, delivering the value, and measuring the value.

What Data Does It Need?

Every organization needs to capture and manage the data that it creates. Regardless of whether it's a small business, a multinational enterprise, or a government agency, they all create and leverage data as part of their day-to-day operations. Bills need to be paid, customers need to be billed, resources need to be managed, services and/or products need to be delivered, and outcomes need to be tracked.

These largely transactional activities help the business operate. They also contribute to big data. There's value in the data, but thinking strategically requires the ability to step back from this transactional point of view and take a more holistic view at how the business operates. Rather than looking at whether an individual order has been fulfilled, decision makers might be interested in reviewing whether the average time needed to fulfil an order is competitive.

Taking this more strategic perspective requires the organization to view its data differently. This often involves consolidating information from multiple operational systems and transforming it such that the data is centered around the item of interest. For example, the organization might be interested in understanding overall customer experience and satisfaction levels. To determine this, they would normally be interested in how each customer interacted with the organization, how effective that interaction was, and how frequently the customer chose to interact in a particular way.

At the lowest level, this information is captured in systems that manage transactional interactions. To build this understanding, analysts might need to pull together data from its contact center, its online platform, as well as its order management system. These systems revolve around the transactions they manage. Respectively, they are concerned with issue tracking, content delivery, and order tracking. While each would capture information about the customer to different degrees, the comprehensiveness of this information will vary substantially. Getting to a strategic point of view involves drawing out the information of interest across the organization as a whole (the customer, in this case) and placing it front and center.

Conceptually, this may seem simple. What usually makes this process a bit complicated is that each of these systems usually has its own way of tracking interactions. For architectural and technical reasons, customer identification numbers may not match between systems. At a very simplistic level, one system may use the customer's full name and address as an identification, one may use the customer's identification number, and one may use the customer's online login details. Consolidating this information into a single view requires mapping tables that link this information together.

The rationale behind an enterprise data warehouse is usually that this information needs to be stored somewhere. Operational systems are normally designed to support a specific function rather than offer architectural flexibility, making them a poor landing point for the consolidated and aggregated data. Additionally, creating and storing these linkages requires processing power, capacity that existing operational systems may not have available. Rather than try to force an existing system to fit, most organizations choose to design a system that's fit for the purpose. And so, they establish a warehouse and start merging all the organization's information into a single environment.

This is a nontrivial task and takes years. And, that's assuming it ever really ends. Most organizations constantly generate new data as fast as their ability to capture information increases. Where they may start simply tracking which pages were viewed on their website, they may eventually get to a point where they track the mouse movements made by every customer across each page. With the amount of effort and expense organizations invest in creating this single, high-quality source of information, it's unsurprising that they try to encourage and sometimes force business analytics teams to use the warehouse and avoid interacting with the upstream source systems.

Unfortunately, this isn't always possible. Enterprise warehouses inevitably make a great starting point (and sometimes, if rarely, an ending point), but there are many situations where they simply do not contain the information the team needs to drive quality outcomes. In these situations, the team needs to source their own information and create their own information stores that go outside of the organization's agreed enterprise warehouse data model. Needless to say, this creates a great deal of tension—to the architectural team, it appears that the analytics team is duplicating large amounts of data. Even worse, data and systems architects often heavily underscope the amount of storage space needed by the business analytics team.

Understanding why this is the case involves understanding the limitations of a traditional warehouse when viewed through a business analytics lens. A team is only as good as the data it can source. And, analytical data often differs from typical warehouse data in four ways:

1. Granularity
2. Temporality
3. Comprehensiveness
4. Statistical completeness

Granularity

Advanced forms of business analytics require granular information, often to the degree of the original transactional measures. When this isn't available, many techniques become impossible.

Warehouses are expensive. They require high-performance technologies and large amounts of time to set them up and make them effective. This performance comes at a cost—highly available and redundant storage doesn't come cheap. Because of this, the designers need to compromise. The fastest (and most logical) way to reduce costs is to design the warehouse based on common requirements rather than comprehensiveness. It would be great to capture *all* of the organization's data in one location. However, most people just need a subset of the data that's theoretically available. A common starting point is simply moving from a product-centric point of view to a customer-centric point of view through creating a single view of customer.

To contain development, maintenance, and storage costs, the design team will limit source data capture to only what's necessary to achieve the required aggregations. They will then discard that same source data once they've met their requirements. This approach works well for relatively unsophisticated applications of business analytics such as reporting and dashboarding.

Unfortunately, it fails to work for more advanced forms of business analytics like predictive modeling and optimization. These rely on the use of statistics to identify patterns within large amounts of data and identify defining characteristics and relationships between elements. Usually due to cost constraints, this information is rarely kept in the warehouse. Capturing and retaining it can make a massive difference in the costs borne by the business.

An average-sized telecommunications company, for example, can generate a few terabytes of person-to-person transactional call information every month. All the majority of the organization usually needs, however, are some simpler measures such as the total number of calls each customer made over the last billing period. While the source data may be on the order of terabytes, the final derived information for all customers could be as small as hundreds of megabytes. Given the cost of highly performing and redundant storage, this represents a major cost difference. Because of this, the warehouse rarely contains the granular transactional information the business analytics team needs. The trick to ensuring *granular* analytical data is to make sure the original transactional data is available in some form if and when it's needed.

Temporality

One of the most powerful aspects of business analytics lies in its ability to identify dynamic characteristics rather than just static measures. Models can link behavioral and environmental changes to desired or unwanted outcomes and by doing so give the organization the ability to predict these outcomes ahead of time. Doing this requires having a record of information over time, and when this isn't available, understanding dynamic relationships becomes impossible.

The high cost of enterprise storage leads many architects and designers to try to contain costs in other ways. Another thing organizations often exclude is transactional history. Most business applications need either only current information or highly summarized year-on-year comparisons. While what happened four years ago may be of importance from an accounting perspective, it's rarely important in an operational context.

Excluding this historical data from the warehouse makes a great deal of sense given this intended use. If an organization holds a year's worth of information in the warehouse at the most granular level, holding five years' worth of information would require five times the storage. In assessing the cost and benefits of doing so, most designers conclude that retaining *all* data simply doesn't make financial sense.

Unfortunately, this information is tremendously valuable when it comes to more sophisticated applications of business analytics. Statistical modeling relies on identifying patterns through repetition. It's a simplistic example, but it's impossible to uncover a trend with two data points. Logically, at least three data points are needed to identify whether factors such as seasonality play a role in driving outcomes, and ideally more.[*]

Most warehouses are designed without these applications in mind and therefore lack sufficient history to enable more sophisticated forms of business analytics. Those that do have sufficient history usually lack granularity. The data usually still exists in source and financial systems; it's simply a case where the team is forced to go elsewhere for the information they need. The trick to ensuring *temporal* analytical data is to start collecting it early; once it's gone, it's gone.

Comprehensiveness

Another major advantage of advanced forms of analytics lies in its ability to incorporate vastly more information that we can mentally process. Models that include thousands of predictors are not unheard of. The best models leverage a wide variety of predictors to help link vastly different behavioral characteristics to target outcomes. Unfortunately, this breadth is rarely fully represented in the warehouse, largely due to the cost it would imply. When this information isn't available, the organization limits its ability to discover and exploit these relationships.

Humans are complex creatures—our behaviors surface in a variety of ways. If we're unhappy with our telephone provider, we may start testing other services such as online voice-over-IP offerings. Our phone usage might gradually decline over time as we favor videoconferencing services where possible. We may call their contact center to complain about our service, find out whether there are other services that might be a better fit, or enquire about our contractual commitments. And, as our contract comes up for renewal, we might start browsing through plans on the company's website, benchmarking plans against competitors.

Each of these actions is a leading indicator of churn—taken as a whole, they flag a customer at high risk of cancellation. Often, this panoramic data can be the difference between knowing what's going to happen and just making a guess. Statistical modeling would help the company not only quantify the degree to which each of these actions increases the odds of cancellation but also create a probability of cancellation for every single customer. Doing so, however, requires having the right data in the first place.

This true comprehensiveness is rarely available in the warehouse. Projects need constraints if they're to be delivered and warehouses are no exception. Trying to boil the ocean and include *all* the organization's data in the warehouse is usually uneconomical. To accommodate for this, the architectural team scopes their warehouse on current business requirements. Unfortunately, analytics is usually a voyage of discovery—it's hard to know what will be useful until one tests one's models with actual data. Inevitably, this means that there will be data of potential value to the analytics team that isn't included in the warehouse.

To be effective, the team needs to extract data from source systems, transform and cleanse it, and store it somewhere. This increases storage requirements and, if the reasons behind this are misunderstood, often creates a great deal of concern among the warehousing team. After all, the warehouse is usually meant to be the single source of truth. The trick to ensuring *comprehensive* analytical data is to give data scientists an area where they can incorporate the data they *need* rather than just the data that's *available*.

Completeness

Mathematical modeling can be a complex field. Many approaches are constrained by a variety of data requirements—some algorithms only work for binary (yes/no) outcomes and some need specific input data characteristics to work. When data isn't formatted or stored correctly in the warehouse, analysts may need to duplicate much of this data simply to enable them to do their jobs.

Having a single source of quality data is fundamental to running a business; it's impossible to make good decisions on bad data. Organizations often talk about this in terms of having accurate, complete, consistent, timely, and auditable data. This, however, creates subtle complexities. A great example is in tracking whether or not customers have opted in for email. Logically, one would think that there are only two possible answers: *yes* or *no*. In practice, there's a third option—they may not have answered the question yet. Representing this in data can become a rather complex question. On one hand, should "yes" and "no" be represented as text or as a number (1 and 0 respectively)? On the other, should a nonresponse be 0 or null (the absence of data)? Each of these is an accurate and complete representation of the data; it's simply a case of changing the storage mechanism. Despite being seemingly trivial, these decisions can have massive impacts on how easily teams can apply sophisticated forms of analytics.

Statistical modeling requires having numerical measures. If the fields are stored as free-text ("yes" and "no"), they need to be converted into a numerical representation before they can fed into a model. Usually, the field needs to be converted into a binary (1/0) representation where the 1s represent the occurrence of an event (having opted in, in this case). This involves extra effort and a complete replication of the field, necessitating more storage. This becomes even worse when the field in question has many levels (options). A retailer that classifies sales by subcategory may have hundreds of discrete values within this field, including "Female fashion—Skirts," "Furniture— Bedroom," and so on. To allow modeling, each of these is usually converted into its own binary field, exploding what was one field into hundreds.

Making things even more complicated is that some algorithms carry various data restrictions. A regression, for example, requires every field input into the model to be populated with a value of some form. Any records that are missing a value in any field are excluded. This creates a significant dilemma—in many situations, incomplete data is the norm. Some records will be missing because incorrect information was entered or because it simply wasn't captured at all. Having accurate data requires the organization to maintain these missing values—having a null field under "opted into email" may still be seen as being accurate and complete even if the field isn't populated. Unfortunately, this prevents all those fields from being used within regression and logistic regression models. To apply a broad set of algorithms, the analytics team need to repopulate these fields with "best-guess" values that are representative of the rest of the data while (hopefully) still preserving auditability by tracking which fields were original and which fields were statistically populated. This process is called *imputation*, and there are a variety of techniques that minimize the amount of statistical bias introduced by the replacement values.

Applying them usually involves duplicating even more fields; it's rare that the warehousing team will allow the analytics team to do wholesale field replacements in the single record of truth. This usually creates tension and substantially increases the amount of storage needed by the analytics team. Not doing so, however, substantially limits the ability of the team to generate accurate predictions when using relatively sophisticated techniques. The trick to ensuring *complete* analytical data is to educate and ensure that the organization's IT support group understands the difference *as well as* the reasons duplication is sometimes necessary.

[*]And, it must be said that three data points will create an extremely poor level of confidence!

Note

1. For a good overview on designing centers of excellence and how this links to organizational performance, see Aiman Zeid, *Business Transformation: A Roadmap for Maximizing Organizational Insight* (Hoboken, NJ: John Wiley & Sons, 2014).