Building a solid foundation for big data analytics

6 best practices for capitalizing on the full potential of big data

IBM

With information streaming in from more sources than ever before, organizations face the daunting challenge of gaining insights from new data sources and types, including structured and unstructured sources. In addition to the growing volume and variety of data, there's also the issue of velocity, which refers to the speed at which data is coming in and how fast you need to be able to take advantage of it. In this environment, it is not only access to information, but the ability to analyze and act upon this information in a timely manner that creates competitive advantage.

Building applications for handling big data requires laser-like focus on solutions that allow you to deliver scalable, reliable and flexible infrastructure for fast-growing analytics environments. The "right" infrastructure—one that is optimized for performance, flexibility and long-term value—can contribute to successful business results by:

- Accelerating analytics and speeding up time to business insights
- Sharing compute resources for optimum utilization and reduced capital costs
- Enabling advanced storage virtualization, along with policy-based tiering, to balance storage cost and performance
- Reducing complexity, simplifying management and enabling workers to focus on strategic priorities

If big data already plays a dominant role within your business, your organizational changes and successes have likely been dramatic. However, you can't rest on your laurels—big data and analytics is an investment that requires attention and discipline to maintain. Continual evaluation and assessment will keep processes running smoothly, while new technologies may open up additional access and analysis paths. For example, Apache Spark has emerged as a leading computing framework for analytics, supplanting MapReduce in many cases. And for good reason: organizations able to harness Spark can run analytics substantially faster as they process ever-growing volumes of data.

If data is still playing a limited, traditional role at the edges of individual departments, depending on your business needs and goals, you may want to consider moving big data analytics to the center of your business. Once you've decided to give big data a more central role, you need a clearly defined infrastructure strategy to help ensure your analytical initiatives are built on a solid foundation. Following are six best practices that your implementation teams can follow to maximize your return on investment and increase your chances of success.

1: Identify the business opportunity

Big data isn't about technology; it's about business outcomes and performance. Therefore, it's essential that you have a reasonable assurance that your business needs and demands will be met—before investing your first dime. The logical starting point for any new initiative should stem from a set of business needs, questions or opportunities that have measurable results, such as improved customer satisfaction, increased profit margins or faster decision-making.

Rather than focusing on technology, your initiative should begin by addressing how it will help your organization achieve clearly defined business objectives. Focus first on high-value opportunities with the greatest potential for positive impact. Identify capabilities that are most critical for the first use cases, including the software and services you need to realize them.

The sheer volume of data and options can be overwhelming—and distracting if team members get too enamored with the technology. To avoid this trap, consider using a technology blueprint, which can help you define the scope of big data within the organization by identifying the key business challenges to which it will be applied. The blueprint creates a common understanding of how you intend to use big data to improve your business and meet your defined objectives. It also takes into consideration existing and planned future data, technology and skills, and outlines where to start and how to develop a plan that aligns with your business and IT strategies.

For some organizations, an expert guide is essential to get started in the right direction. Often, external consultants can help organizations see what's possible from an independent, unbiased perspective; understand the impacts of possible scenarios; and help articulate a strategy that incorporates buy-in from all parts of the enterprise.

2: Think big, but start small

There are many unknowns when working with data that your organization has never used before. Which elements of the data hold value? What are the most important metrics the data can generate? For example, an exercise to extract cost savings from finance processes may be a quick win that persuades others in the business of the overall potential of big data and the need to begin generating ideas for such initiatives themselves. Since the costs and time required to achieve success can be hard to estimate, it makes sense to start with the low-hanging fruit.

Pursuing big data in small, incremental steps offers a lower-risk way to see what big data can do for your organization and to test your readiness to use it. You can then apply the lessons learned to larger initiatives across the business.

Start by defining a few relatively simple analytics that won't take much time or data to run. For example, to drive additional revenue, an online retailer might implement a product recommendation application that automatically generates personalized product suggestions based on a customer's current and historical shopping interests, along with general trends and business rules.

Using agile and iterative implementation techniques allows you to see what the data can do. More importantly, this approach yields results that are easy to test to see what type of value the analytics provide. For optimal results, your architects and developers may have to learn how to work with the latest technologies, such as Spark. Fortunately, advanced service orchestration and resource management capabilities can help ease this task, allowing designers to deploy new frameworks efficiently and effectively, increasing performance and scale and eliminating resource silos.

Successful prototypes also make it far easier to get the support required for the larger effort. Best of all, the full effort will now be less risky because the data is better understood and the value is already partially proven. This targeted approach allows analysts to zero in on what they need and ignore the rest. They can create test and control groups to whom they can send the follow-up offers, and then they can help analyze the results.

During this process, they'll also learn a great deal about the data and how to use it. Avoid getting too complicated too quickly, reducing scope to the simplest, most valuable objectives, and be prepared to scale once a solution catches on.

Reaping the rewards of real-time analytics

One of the world's largest telecommunications companies with hundreds of millions of subscribers needed real-time big data analytics capabilities to gauge the performance of its network, monitor usage patterns, detect potential fraud, quickly settle customer disputes and analyze subscriber credit risk. The lack of real-time insight was impacting the company's decision-making and delaying customer response, leading to dissatisfaction and customer churn.

The company sought a high-performance analytics solution that could concurrently query call detail record (CDR) data and process the data in parallel by loading it through multiple nodes on a file system. With the company's CDR data sizes continuing to increase, it needed a system that would enable it to support 1,000 concurrent records per second with a query response time of under a few seconds. It needed the infrastructure to be able to scale out linearly and remain cost-effective as its CDR sizes grew.

The company implemented a proof-of-concept study to evaluate the performance capabilities of an IBM big data solution consisting of IBM hardware and software, including IBM® PowerLinux™ servers, an IBM Spectrum Symphony distributed computing framework and an IBM Spectrum Scale™ file system. Results showed the system significantly outperformed a competitive option, while using only half as many servers.

Compared to a competitive option, which used open source solutions running on x86 hardware, the IBM solution delivered:

- **More than four times the data loading performance per core**
- **Efficient scaling up to 900 concurrent users**
- **Query performance of two to three times better across the range of CDR data sets**

In light of the success of the smaller-scale proof of concept, the customer went on to implement a full-scale solution, deploying 30 servers to support thousands of concurrent tasks and processing in excess of 100 billion records per day.

3: Experiment with different approaches

Analytics professionals typically excel at measuring the right information, but often get bogged down by the volume and variety of data. Instead of setting up formal processes to capture, process and analyze all of the data all of the time, capture a subset of defined data—perhaps a month's worth of data for about one product for one division.

Encourage staff to re-examine every process and use case and develop the curiosity needed to infuse analytics in other areas—not just updating traditional data-driven applications, but locating areas where big data can fundamentally change the way the business operates. For example, could real-time (or near-real-time) analytics help drive new business value? Initially, real-time processing may not seem like a big distinction; however, such capabilities have been used to create entirely new lines of business. For example, using big data analytics, communications service providers (CSPs) can analyze location data from millions of mobile devices to reach more customers with personalized, targeted products, services and marketing—creating new revenue opportunities.

Evaluate different application frameworks and make sure they are well-suited to your analytical workloads. For example, would Hadoop be the best fit for your needs or should you go with Spark? For applications involving iterative machine-learning algorithms and interactive analytics, Spark offers several performance advantages, including interactive querying of data, stream processing and sensor data processing. Spark is also developer-friendly, with libraries, easy-to-use application programming interfaces (APIs) for a variety of programming languages and a rich set of high-level tools to support big data analytics. Plus, Spark can run in a Hadoop environment if an organization already has that in place.

Also, consider existing expertise and toolsets. For example, if you already have in-house Hadoop expertise, is it better to leverage that? Or does the value of Spark merit investment in new capabilities and updated workflows?

4: Optimize your computing resources

One of the big challenges in distributed analytical environments is that many application frameworks have their own workload schedulers and operate on the assumption that they are running on their own dedicated infrastructure. Spark is a good example. Setting up ad hoc Spark clusters can lead to inefficient use of resources, management challenges and security issues that all impede deployment of Spark in a production environment. The result can be multiple underutilized and costly infrastructure silos, each dedicated to its own set of applications. This can be particularly true of experimental setups where people tend to bring in numerous application frameworks and set up a separate cluster for each, because that seems like the easiest (or maybe only) way to go.

The rapid pace in which popular frameworks like Spark are updated can result in various groups running different versions of the framework. Management software with multi-tenant capabilities allows you to run several instances and different versions at the same time. Instead of setting up numerous server clusters for different applications, you can use high-performance grid management software that allows applications to share a pool of servers without interfering with one another. Such software can dynamically allocate resources across the cluster as needed. The software can also reduce a lot of the manual steps in deploying applications and can shift workloads around automatically to provide fault tolerance and keep utilization rates high.

With the ability to share resources across applications, users and lines of business, you can optimize existing hardware, mitigate cluster sprawl and defer the need for incremental capital investment. In addition, with reporting, usage accounting and capacity planning information now consolidated, administrators can more efficiently monitor resources to meet service-level objectives, and new resources can be quickly brought online as needed. In addition, running multiple framework instances on shared infrastructure helps maximize resource utilization and contain costs.

5: Don't forget about data management

Big data analytics workloads place extraordinary demands on high-performance computing infrastructure. Supporting these workloads can be especially challenging for organizations that have unpredictable spikes in resource demand, or need access to additional compute or storage resources to support a growing business.

One emerging solution is software defined infrastructure (SDI), a model in which cloud computing resources are managed and controlled by software rather than human operators. The goal of SDI is to yield an application-aware infrastructure that captures workload requirements and deployment best practices, provides policy-based automation across data center environments and optimizes the workloads as they're running.

Software defined storage (SDS), a subcomponent of SDI, involves separating data management from physical storage through a virtualization layer. This approach allows storage optimization through automated provisioning and tiering systems, and more cost-effective scalability on heterogeneous storage systems that can be geographically distributed. SDS places the emphasis on storage-related services rather than storage hardware. Unlike open source Hadoop Distributed File System (HDFS) storage offerings that require a piecemeal assembly of components, a POSIX-compliant solution has several compelling advantages, including:

- Time-based scheduling policies
- Centralized GUI for managing resource plans
- Dynamic updates of resource plans
- Pre-emption with sophisticated reclaim logic
- Proven multi-tenancy with service-level agreement (SLA) management and quality of service

Drive more value with scalable storage

Nuance Communication's advanced, self-optimizing natural speech-recognition and processing technologies are helping to define the next generation of human-computer interaction: intelligent systems. To deliver on its promise of facilitating human-computer interactions, Nuance must ingest, store and analyze huge—and ever-growing—volumes of speech data in real time.

To meet its needs, Nuance deployed IBM Spectrum Scale, a hardware-agnostic, software defined file storage solution with a single global namespace. Built from a collection of storage devices that contain the file system data and metadata, the system enables applications to access files through standard POSIX file system interfaces.

Nuance ingests voice data as hundreds of millions of sound files every day, storing them first as Swift objects in a private OpenStack cloud. "We chose IBM Spectrum Scale because no other file system could offer the same nondisruptive scalability," says Bob Oesterlin, senior storage engineer at Nuance. "We can add storage or clients in real time, rebalance existing storage where we see hotspots, take devices offline for maintenance—all with no impact on performance. This gives us the ability to grow and take advantage of new technology."

Nuance uses the IBM Spectrum Scale policy engine heat map to automatically move the most frequently accessed files to the fastest disk systems, and to demote unused files to slower, less-costly storage. IBM Spectrum Scale tiering helps Nuance utilize its capacity efficiently, and it is self-optimizing over time. "Using tiered storage accelerates operations by 20 percent out of the box," says Oesterlin. "Also, we can keep using older systems for longer, and drive more value from past investments."

With IBM Spectrum Scale optimizing the placement of files for cost-effective performance, Nuance can deliver exceptional responsiveness to external clients and to its internal R&D community without incurring unsustainable costs. "IBM Spectrum Scale is one of those pieces of software that is an unsung hero," says Oesterlin. "The role it plays in our high-performance computing (HPC) environment is vital."

The goal of software defined storage is to provide administrators with flexible management capabilities through programming. Without the constraints of a physical system, a storage resource can be deployed much more efficiently. Its administration can be simplified with a unified user interface and automated policy-based management for a shared storage pool that is housed across a variety of hardware.

What to look for in a cluster management solution

- Ability to support multiple instances of different framework versions
- Consolidated architecture for simplified deployment and monitoring
- Sophisticated resource scheduling for improved time to results
- Highly efficient utilization for better cost containment
- Enterprise-class security through role-based access control

6: Build on your initial successes

Once your initial use cases are established, the next step is to scale big data and analytics within the initial test environment, and then eventually into other areas of the organization. Rather than starting from scratch each time, manage your initial successes and results cohesively so you can easily reference and reuse your best methods and approaches. You may want to follow a "rinse and repeat" cycle where you return to step one to identify new opportunities, experiment with different approaches and evaluate emerging technologies. Ideally, you will be able to cycle through the steps more rapidly by applying lessons learned from previous efforts.

When new investments are made, it's essential to build infrastructure that can be easily scaled and shared as teams and demands change. New deployment options allow you to plan how your big data initiatives will run inside and outside your organization, including:

- On premises
- In the cloud (public or private)
- As a service
- Hybrid (on premises and in the cloud)

The right deployment method can make or break a big data initiative. Evaluate your options carefully based on the business challenges you're addressing and the needs of the intended users. You will progress in your own style and at your own speed. Investments will expand and contract depending on business conditions and imperatives, but the increasingly strategic use of information—and developing it into business-critical knowledge—remains the central constant.

Put your big data plans on a firm foundation

New-generation applications and open source frameworks such as Hadoop and Apache Spark enable you to analyze new and diverse digital data streams. This analysis can reveal new sources of economic value, provide fresh insights into customer and market trends, or optimize operations to reduce cost and increase productivity. No matter what industry you're in, those results can translate into tangible benefits for your business. But reaping these benefits requires broad, strategic planning followed by discrete actions. You need the right tools to capture and organize a wide variety of data types from different sources, and then quickly analyze it within the context of all your enterprise data. By following a defined implementation strategy and applying the right technologies, you can capitalize on the full potential of big data and gain an important competitive edge.

Why IBM?

IBM offers a comprehensive portfolio of software defined infrastructure solutions designed to help your organization deliver IT services in the most efficient way possible, optimizing resource utilization to speed time to results and reduce costs. These offerings help maximize the potential of your infrastructure to accelerate your analytics, HPC, Hadoop, Apache Spark and cloud-native applications at any scale, extract insight from your data and get higher-quality products to market faster.

For more information

To learn more about IBM big data analytics solutions, contact your IBM sales representative or IBM Business Partner, or visit:

- **ibm.com**/spectrum-computing
- **ibm.com**/systems/spectrum-computing/conductor-with-spark.html

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
June 2016

Please Recycle

DCW03063-USEN-03