HITACHI
Inspire the Next

Blueprints for Big Data Success

Succeed With Four Common Scenarios

By Hitachi Vantara

September 2019

Introduction

By now it's become fairly clear that big data represents a big shift in the enterprise technology landscape. IDC estimates that the amount of useful data worldwide will increase 20 times between 2010 and 2020, while 77% of the data relevant to enterprises will be unstructured through 2015.[1] As these volume and variety trends continue, companies are increasingly turning to Hadoop, NoSQL and other tools to tackle information issues not readily addressable with older relational database and data warehouse technologies.

How do you determine how to get value out of your big data? For that matter, how do you define a big data strategy? These are top issues facing today's organizations. In light of these challenges, this piece intends to identify and explain big data use cases generating business results for companies today and shed light on emerging use cases expected in the near future. The following pages address what these use cases are, why companies are investing in them and their common reference architectures.

Mapped out below are 10 key enterprise use cases for big data, categorized according to the ability to generate business impact (Y axis) as well as level of implementation complex-

ity (X axis). Business impact ranges from optimizing current processes to transforming entire business models. Complexity ranges from entry-level implementations relying on fairly standard technologies to advanced cases relying on combinations of technologies, some of which are not largely commercialized. The use cases are marked either "Current Adoption" or "Emerging Adoption." The former indicates more widely implemented use cases that follow fairly repeatable guidelines, while the latter indicates implementations that are less common today, but are expected to appear more often in the future. This paper, discusses in detail the "Current Adoption" use cases.

[1] IDC Digital Universe Study, 2012.
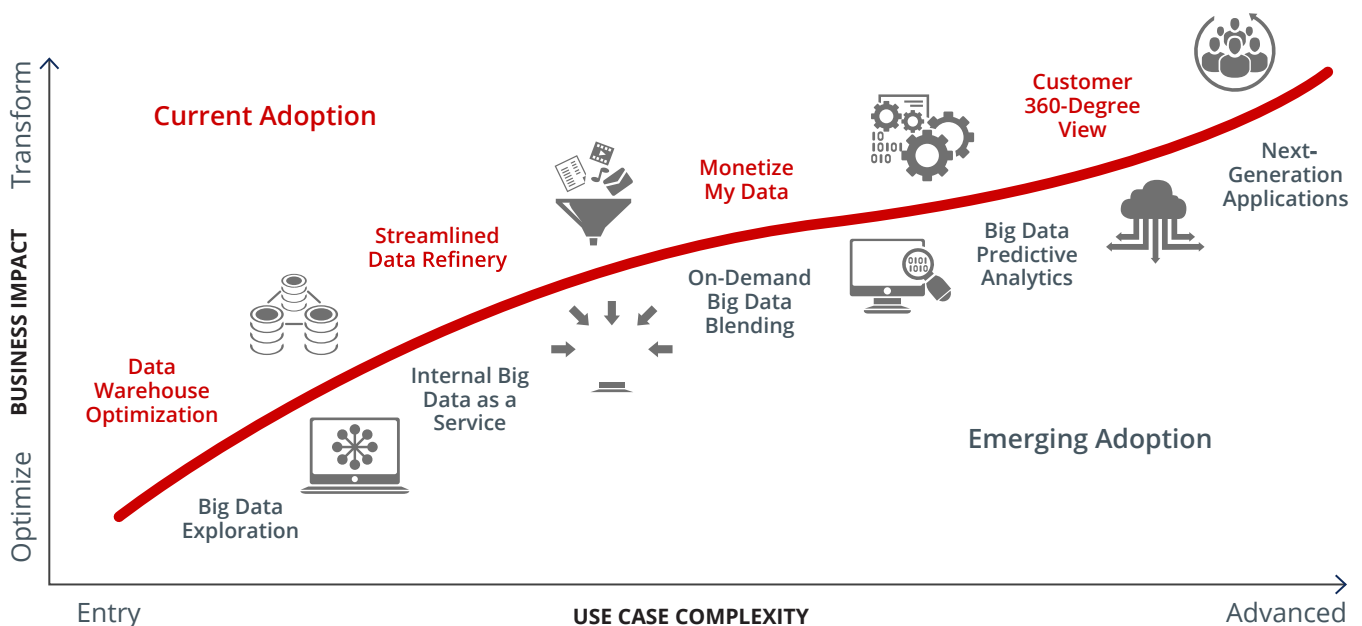
Included Below Is a Brief Description of Each of These Use Cases:

Current Adoption

- **Data Warehouse Optimization —** The traditional data warehouse (DW) is strained by rising data volumes, meaning stakeholders can't get the analytics they need on time. Expanding DW capacity can be costly, so organizations tap big data to offload less frequently used data and improve DW performance.

- **Streamlined Data Refinery —** Here the big data store becomes the landing and processing zone for data from many diverse sources, before it is pushed downstream for low-latency analytics (most likely to an analytical database for rapid queries). Extract, transform, load (ETL) and data management cost savings are scaled up, and big data becomes an essential part of the analytics process.

- **Customer 360-Degree View —** The 360-degree view blends a variety of operational and transactional data sources to create an on-demand analytical view across customer touch points. It also includes providing customer-facing employees and partners with information made available inside everyday line-of-business applications.

- **Monetize My Data —** In this case, enriched and de-identified data sets are delivered as a service to third-party customers. It leverages powerful data processing and embedded analytics to generate a new revenue stream for the enterprise.

Emerging Adoption

- **Big Data Exploration —** Companies are dumping massive data into big data stores, but they aren't always sure what information is in there ("dark data"), or if it can be leveraged in a productive way. To "get their feet wet," analysts will run basic data mining algorithms and work to correlate patterns they find with data from other sources.

- **Harnessing Machine and Sensor Data —** Until recently it has been cost-prohibitive to tap into analytics on high-volume data from devices like sensors, routers and set-top boxes. Today, however, big data has enabled the use case of harnessing this information for data mining and low latency analytics, ultimately empowering organizations to take quick action on operations and service issues.

- **Big Data Predictive Analytics —** Big data offers a new set of tools for optimizing machine-learning algorithms (for training and evaluation) and using them to predict or influence outcomes (scoring). Running predictive analytics in the big data store has applications in fraud detection, recommendation engines and offer optimization.

- **Next-Generation Applications —** While cloud computing and software as a service (SaaS) are not new trends, their next phase will likely hinge on big data. Application providers are innovating around data and analytics architecture to make their products more powerful, intelligent and valuable to customers. An embedded analytics interface inside the end-user application allows the vendor to fully capitalize on this innovation.

- **On-Demand Big Data Blending —** Once big data stores are implemented, teams are often still subject to the time constraints of existing data warehouse infrastructure. Time-sensitive needs may require bypassing the DW altogether. "Just in time" blending avoids the need to stage data, delivering accurate, timely data from all sources to analytics.

- **Internal Big Data as a Service —** Enterprises are tapping into big data as a shared database service, to be provisioned across a number of application development teams for data ingestion and access. The goal is to achieve economies of scale and cost savings relative to a more silo-based approach. ETL and analytics solutions are included as components of the centralized enterprise stack.
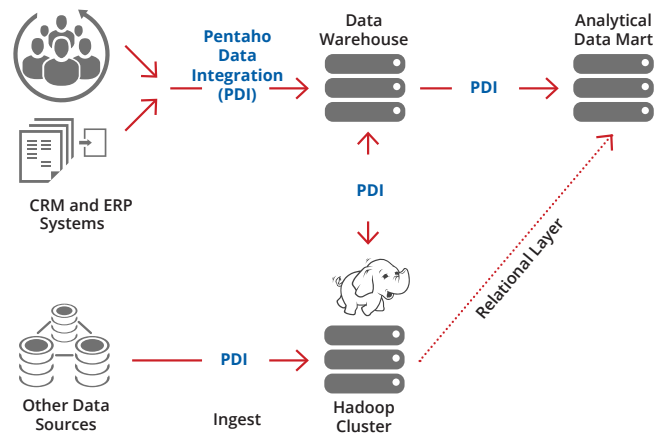
Data Warehouse Optimization

What Is It and Why Are Companies Investing in It?

Data warehouse optimization is one of the most commonly seen business use cases for big data, driven primarily by two pains: cost and operational performance. As the volume of data a company needs to store and access grows, existing data warehouse capacity becomes strained. This leads to deteriorating query performance and access to data for IT and business users. In addition, it creates pressure to buy additional data warehouse storage capacity from incumbent vendors — a very pricey and possibly only temporary solution as data keeps expanding.

As a result, enterprises have looked to big data, specifically Hadoop, to reduce this pressure. Hadoop's distributed computing model provides for powerful processing on commodity hardware, storing data in HDFS (Hadoop Distributed File System) can be an order of magnitude cheaper than traditional data warehouse storage. Specifically, Hadoop storage cost is approximately $1,000 per terabyte versus approximately $5,000 to $10,000 per terabyte or more for fully loaded data warehouse storage, including required hardware, servers and so forth.[3] As such, IT organizations will transfer less frequently used data from their DW to Hadoop to save on data storage costs, while satisfying service level agreements (SLAs) and compliance requirements to deliver data on time.

What Does It Look Like?

In this example, we have an enterprise that is leveraging data from customer relationship management (CRM) and enterprise resource planning (ERP) systems as well as other sources. A Hadoop cluster has been implemented to offload less frequently used data from the existing data warehouse, saving on storage costs and speeding query performance as analysts need to access information from the analytical data mart.

Key Project Considerations

While Data Warehouse Optimization is one of the most common big data use cases seen today, it still requires time, effort and planning to execute. Hadoop is still an emerging technology, and using the "out-of-box" tools accompanying Hadoop distributions requires Java coding expertise to create the routines that actually offload the DW data into Hadoop. Developers and analysts with Hadoop expertise are often difficult for enterprises to hire in sufficient numbers, and can command compensation approximately 50% higher than staff with skills in SQL and other more traditional tools.[4]

The Pentaho platform is valuable in providing an intuitive graphical user interface (GUI) for big data integration that eliminates manual coding and makes Hadoop accessible to all data developers. This accelerates time to value and reduces labor costs. Even if enterprises already have a data integration solution in place, legacy platforms don't have complete no-coding solutions to integrate existing data sources and databases with Hadoop.

3 Information Week, "How Hadoop Cuts Big Data Costs," 2012.
4 O'Reilly, "2013 Data Science Salary Survey," 2013.

Streamlined Data Refinery

What Is It and Why Are Companies Investing in It?

In the face of exploding volumes of structured transaction, customer and other data, traditional ETL systems slow down, making analytics unworkable. The "Data Refinery" solution streamlines most data sources through a scalable big data processing hub, using Hadoop for transformation. Refined data is pushed to an analytical database for low-latency self-service analytics across diverse data.

This use case is often a logical extension of the cost savings and operational enhancements of DW Optimization. At this point, a greater amount and variety of data is being loaded into Hadoop. This makes Hadoop not only an archive, but also a source of valuable multisource business information, just waiting to be queried. As such, this use case is more transformative than DW optimization. The organization can establish usable analytics on diverse sources of data at high volume, thanks to faster queries, rapid ingestion and powerful processing provided by the combination of Hadoop and an analytical database (such as Vertica or Greenplum). By the same token, teams can engineer data sets for predictive analytics more quickly.

What Does It Look Like?

This example below shows a refinery architecture for an electronic marketing firm that delivers personalized offers. Online campaign, enrollment and transactional data is ingested via Hadoop, processed and then sent on to an analytical database. A business analytics front-end includes reporting and ad hoc analysis for business users.

Key Project Considerations

The staff and productivity challenges from DW Optimization still ring true in this case. Not surprisingly, return on investment can be enhanced with tools that eliminate coding and simplify the process of integrating big data stores to various relational systems. Otherwise, this use case is generally a more
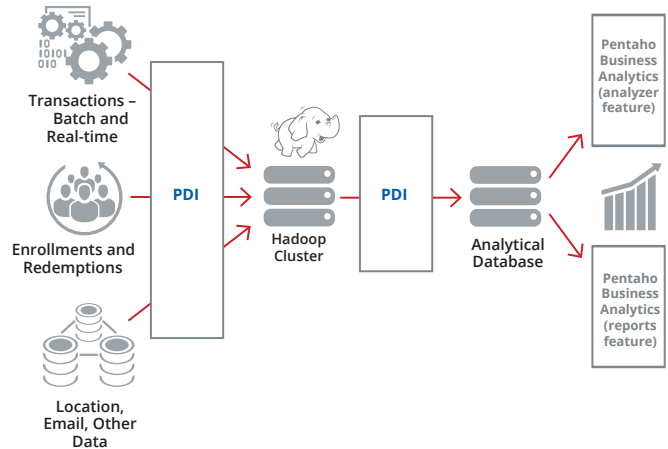
expansive and lengthier integration project, which may involve consolidating many point-to-point system connections into a centralized Hub model. The project becomes more complex to execute as the variety of data types and sources increases. This underlines the importance of selecting data integration and analytics platforms with highly flexible connectivity to a wide variety of current and emerging data systems. Given the emerging importance of analytical insights from Hadoop in this use case, collaboration between data developers and business analysts becomes more important. An integrated platform is needed for data connectivity and business intelligence; it's much more difficult to effectively coordinate when IT and business users are leveraging isolated toolsets.

Finally, an analytical database is normally a key part of this architecture. These databases are optimized for business intelligence, usually through faster query performance, greater scalability, multidimensional analysis "cubes" and/or in-memory functionality. By comparison, traditional transactional databases may not provide for the required level of query performance and analytics functionality.

Customer 360-Degree View

What Is It and Why Are Companies Investing in It?

Companies have long sought to bring a variety of data sources together to create an on-demand analytical view across customer touch points. Leveraging both big data and traditional data sources in a fully integrated environment, organizations can accomplish this and achieve tremendous actionable customer insight. Whereas DW optimization and streamlined data refinery are primarily cost- and efficiency-driven use cases, the customer 360-degree view is clearly aimed at boosting customer lifetime value, especially in competitive consumer markets where churn is a key concern (such as telecommunications, hospitality and consumer financial services). The two main levers for success are raising cross-sell and up-sell revenues and minimizing churn risk.

This use case is enabled on the back end by bringing virtually all customer touch-point data into a single repository for fast queries (most likely NoSQL or Hadoop). It's enabled on the front end by bringing relevant metrics into a centralized location for business users. By blending together previously isolated data, the customer 360-degree view gives sales and services teams a more complete understanding of the buyer, while providing a better picture of how a brand's products and services are perceived. Equipping employees with this insight at the point of their interaction with customers gives them the power to make more productive and profitable decisions on the fly.

What Does It Look Like?

In the example above, a financial services company ingests data from various sources into a single big data store, in this case NoSQL. From there, the data is processed and summarized at the customer unique ID level in order to build the 360-degree view. Accurate and governed customer data is then routed to the appropriate analytics views for each role, including call center staff, research analysts and data scientists.
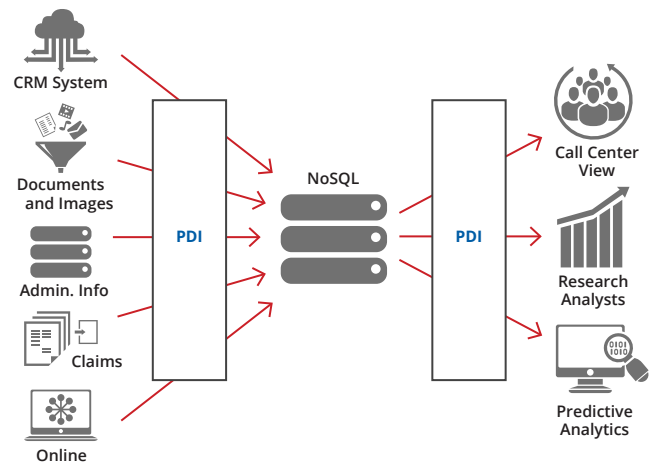
Key Project Considerations

While this implementation can be transformative for businesses, it can also be highly complex and resource-intensive. On top of the big data labor resources challenges and point-to-point integration challenges described in previous use cases, the customer 360-degree view requires significant strategic planning from a business perspective. First, specific revenue-related goals should be tied to the project. Stakeholders must identify both the potential drivers of customer satisfaction and potential opportunities for customer-facing staff to take advantage of that data. At the same time, the relevant business end users must be a part of the planning process, so that information gets delivered from the right sources to the right people in the right fashion. Analytics must be presented to users in a way that they

<div style="text-align:center; color:red; font-size:large;">

The two main levers for success are raising cross-sell/up-sell revenues and minimizing churn risk.

</div>

will be sure to adopt: This means making them easy to access and intuitive to understand, as well as embedding the analytics into crucial operational applications.

From a technical perspective, a NoSQL solution such as MongoDB may be the big data store of choice if an enterprise is looking to route many time-sensitive streams of customer info into a single collection that can be distributed across servers quickly and easy. Hadoop is a better fit where data can be processed in batches and must be stored historically. Often both Hadoop and NoSQL are leveraged in the same architecture. While integrating the big data stores of choice to a wide variety of back-end applications and databases is crucial, a new set of front-end requirements will likely arise. Different consumers of customer analytics will require different types of business intelligence (BI), including:

- Intuitive and customizable dashboards for executives.

- Sophisticated and responsive ad hoc slicing and dicing tools
  for analysts.

- Distributed reporting capabilities for sharing information across teams.

- Data mining and predictive analytics tools for data scientists.

- Analytics that can be embedded into operational software such as CRM and service apps.

It makes sense from a technology compatibility and vendor relationship perspective to look for data and analytics providers that offer most if not all of these capabilities in an integrated platform. At the same time, vendors should have committed to providing big data integration capabilities to seamlessly adjust to changes in technology versions over time. This approach will minimize reconfiguration of a project deployment and make it more resilient. The ability to accommodate evolving user needs and data architectures is crucial in a project as sophisticated as the customer 360-degree view.

> It makes sense from a technology compatibility and vendor relationship perspective to look for data and analytics providers that offer most if not all of these capabilities in an integrated platform.

Learn more about Pentaho Data Integration and Pentaho Business Analytics at hitachivantara.com.

›

Hitachi Vantara at a Glance

Your data is the key to new revenue, better customer experiences and lower costs. With technology and expertise, Hitachi Vantara drives data to meaningful outcomes.

Hitachi Vantara