Blueprints for Big Data Success

Succeeding with four common scenarios

Introduction

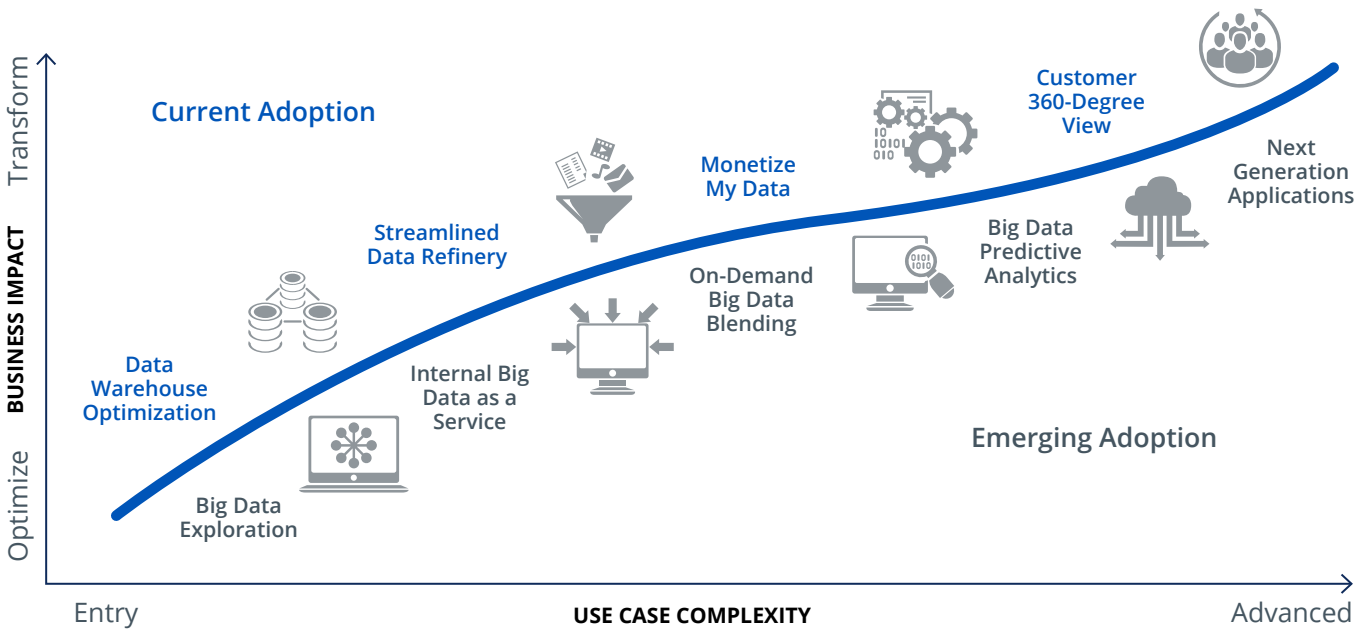
By now it's become fairly clear that big data represents a big shift in the enterprise technology landscape. IDC estimates that the amount of useful data worldwide will increase 20x between 2010 and 2020, while 77% of the data relevant to enterprises will be unstructured through 2015.[1] As these volume 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.

Though the big data opportunity is growing rapidly, research indicates that the top two big data challenges that organizations face are determining how to get value out of big data and defining a big data strategy, respectively.[2] 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 complexity (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.

2 Gartner "Big Data Adoption in 2013 Shows Substance Behind the Hype," 2013.



Included below is a brief definition for 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). 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 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 3rd 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 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

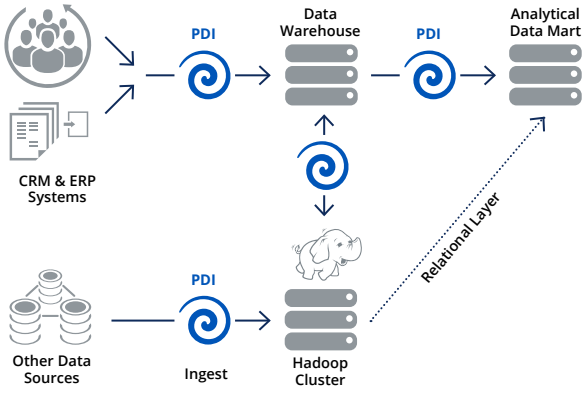
WHAT IS IT AND WHY ARE COMPANIES INVESTING IN IT?

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 (TB) vs. approximately $5,000 to $10,000 per TB or more for fully load data warehouse storage including required hardware, servers, etc.[3] As such, IT organizations will transfer less frequently used data from their DW to Hadoop to save on data storage costs, while satisfying 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 CRM and 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]

Pentaho 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.

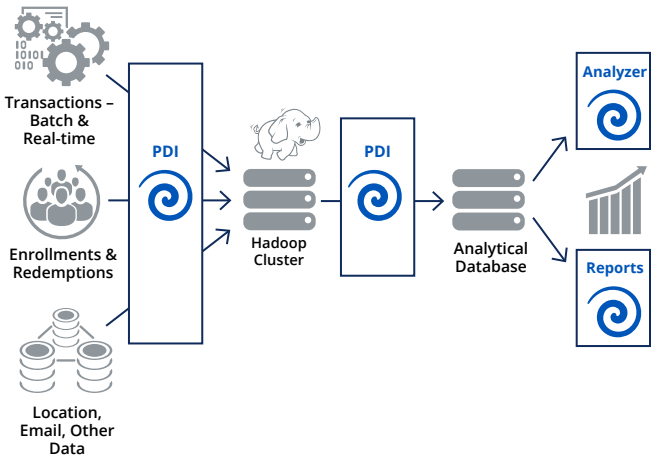
This use case is often a logical extension of the cost savings and operational enhancements of DW Optimization. At this point, a greater amount and variety of data is being loaded into Hadoop – making Hadoop more than an archive, but a source of valuable multi-source business information, just waiting 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 analyst becomes more important. An integrated platform is needed for data connectivity and business intelligence – its 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, multi-dimensional 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

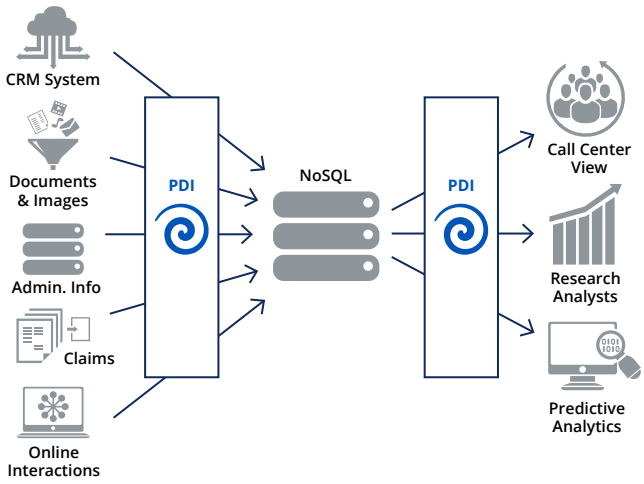
WHAT IS IT AND WHY ARE COMPANIES INVESTING IN IT?

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 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/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 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 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

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

from the right sources to the right people in the right fashion. Analytics must be presented to users in a way they 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 store(s) 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 BI, including:

- Intuitive and customizable dashboards for executives
- Sophisticated and responsive ad hoc slicing/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 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.

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.

Monetize My Data

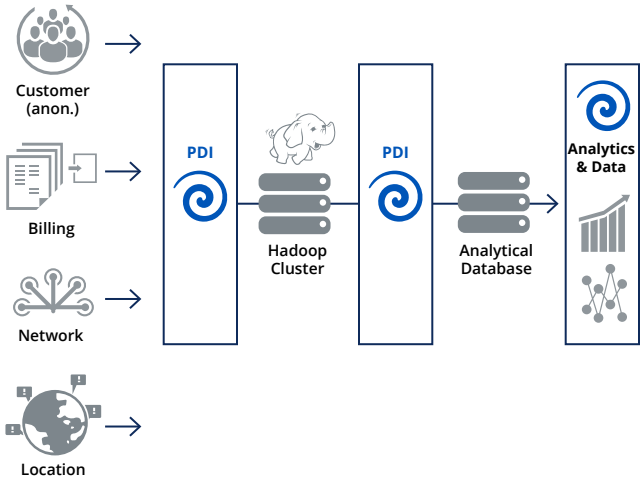
WHAT IS IT AND WHY ARE COMPANIES INVESTING IN IT?

The cost efficiencies of big data and its capacity to handle a variety of data structures open up a variety of ways to enhance an enterprise's existing core business. However, it also offers the potential to add new strategic revenue streams that are in many ways separate from the core business. Monetize My Data is one such case – it enables selling the data itself.

As enterprises collect a greater variety and volume of data through their day-to-day operations, its potential value to 3rd parties increases. In this use case, the data is organized, enriched, and de-identified (to make anonymous the individuals and entities that the data is collected from) before being sold, often to external marketing buyers. For instance, a telecom company could aggregate location data from mobile handsets during different times of day, combine with demographic data, and sell the resulting data sets to a retail company to support store siting decisions. The result is a new source of revenue for the telecom company, as its data is helping the 'brick and mortar' retailer get more intelligent about effectively targeting its audience.

WHAT DOES IT LOOK LIKE?

In the example below, a telecom company combines demographic and human mobility data to offer a specialized analytics service to third parties, leveraging geospatial visualizations to understand retail buying potential. The use case leverages both Hadoop and an Analytical database.



KEY PROJECT CONSIDERATIONS

Gartner predicts that 30% of businesses will be monetizing data assets directly by 2016[5] – a real opportunity exists, and leveraging the right big data tools and approaches can help unlock this potential. In the Monetize My Data use case, Hadoop as a data processing platform provides for lower costs and higher margins relative to higher-priced legacy data warehousing solutions (at least 5x to 10x cheaper per TB as outlined under "Data Warehouse Optimization"). Profitability and time to value are further enhanced with Pentaho's no-coding big data integration and business analytics functionality. At the same time, delivering analytics as a service to third parties may require embedding reporting and visualizations into a branded web application. Pentaho's open architecture and visual flexibility make it a natural fit for such an approach.

5 Gartner, "Gartner Predicts 30 Percent of Businesses Will Be Monetizing Their Information Assets Directly by 2016," 2013.

Learn more about Pentaho Business Analytics

pentaho.com/contact
+1 (866) 660-7555.

Global Headquarters

Citadel International - Suite 340
5950 Hazeltine National Drive
Orlando, FL 32822, USA
tel +1 407 812 6736
fax +1 407 517 4575

US & Worldwide Sales Office

353 Sacramento Street, Suite 1500
San Francisco, CA 94111, USA
tel +1 415 525 5540
toll free +1 866 660 7555

United Kingdom, Rest of Europe, Middle East, Africa

London, United Kingdom
tel +44 (0) 20 3574 4790
toll free (UK) 0 800 680 0693

FRANCE
Offices - Paris, France
tel +33 97 51 82 296
toll free (France) 0800 915343

GERMANY, AUSTRIA, SWITZERLAND
Offices - Munich, Germany
tel +49 (0) 322 2109 4279
toll free (Germany) 0800 186 0332

BELGIUM, NETHERLANDS, LUXEMBOURG
Offices - Antwerp, Belgium
tel (Netherlands) +31 8 58 880 585
toll free (Belgium) 0800 773 83

ITALY, SPAIN, PORTUGAL
Offices - Valencia, Spain
toll free (Italy) 800 798 217
toll free (Portugal) 800 180 060

Be social with Pentaho: