Big data architecture and patterns, Part 2: How to know if a big data solution is right for your organization

Divakar Mysore (mrdivakar@in.ibm.com)
Senior IT Architect
IBM

Shrikant Khupat (skhupat1@in.ibm.com)
Application Architect
IBM

Shweta Jain (shweta.jain@in.ibm.com)
IT Architect
IBM

08 October 2013

This article describes a dimensions-based approach for assessing the viability of a big data solution. By answering questions that explore each dimension, apply what you know about your own environment to determine whether a big data solution is appropriate. A careful look at each dimension yields clues about whether it's time for your big data services to evolve.

View more content in this series

Introduction

Before making the decision to invest in a big data solution, evaluate the data available for analysis; the insight that might be gained from analyzing it; and the resources available to define, design, create, and deploy a big data platform. Asking the right questions is a good place to start. Use the questions in this article to guide your investigation. The answers will begin to reveal more about the characteristics of the data and the problem you're trying to solve.

Although organizations generally have a vague understanding of the type of data that needs to be analyzed, it's quite possible that the specifics are not as clear. After all, the data might hold keys to patterns that have not been noticed before, and once a pattern is recognized, the need for additional analysis becomes obvious. To help uncover these *unknown* unknowns, start by implementing a few basic use cases, and in the process, collect and gather data that was not previously available. As the data repository is built and more data is collected, a data scientist is

better able to determine the key data and better able to build predictive and statistical models that will generate more insight.

It may also be the case that the organization already knows what it does not know. To address these **known unknowns**, the organization must start by working with a data scientist to identify the external or third-party data sources and to implement a few use cases that rely on this external data.

This article first tries to answer some of the questions typically raised by most CIOs prior to taking up a big data initiative, then focuses on a dimensions-based approach that will help in assessing the viability of a big data solution for an organization.

Does my big data problem require a big data solution?

Big data, a little at a time

For the most part, organizations choose to implement a big data solution incrementally. Not every analytical and reporting requirement requires a big data solution. For projects that perform parallel processing on a large dataset or ad-hoc reporting from multiple data sources, a big data solution may not be necessary.

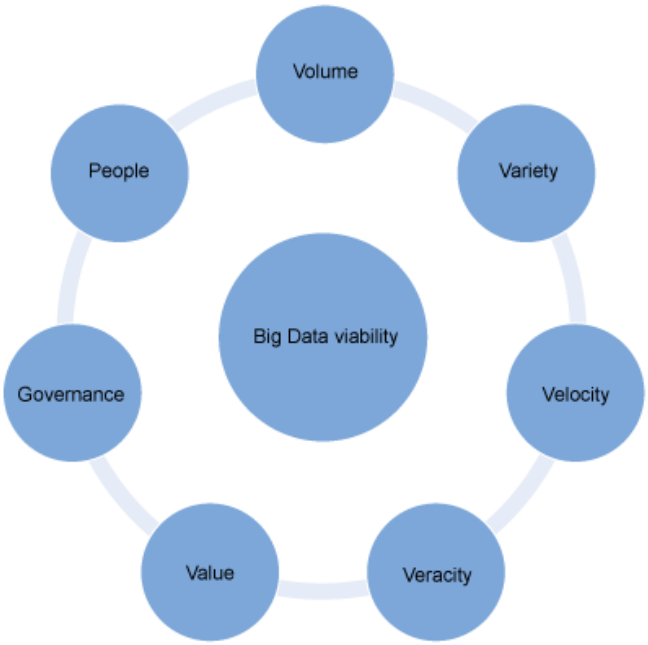
With the advent of big data technologies, organizations are asking themselves: "Is big data the right solution to my business problem, or does it provide me with a business opportunity? Are business opportunities hiding in the big data? Here are some of the typical questions we hear from CIOs:

- What kind of insight and business value are possible if I use big data technologies?
- Is it possible to augment my existing data warehouse?
- How do I assess the cost of expanding my current environment or adopting a new solution?
- What is the impact on my existing IT governance?
- Can I incrementally implement a big data solution?
- What specific skills are required to understand and analyze the requirements to build and maintain the big data solution?
- Do I have existing enterprise data that could be used to deliver business insight?
- The complexity of my data coming in from a variety of sources is increasing. Can a big data solution help?

Dimensions to help assess the viability of a big data solution

To answer these questions, this article proposes a structured approach for evaluating the viability of a big data solution according to the dimensions shown in the following figure.

Figure 1. Dimensions to consider when assessing the viability of a big data solution



- *Business value* from the insight that might be gained from analyzing the data
- *Governance considerations* for the new sources of data and how the data will be used
- *People* with relevant skills available and commitment of sponsors
- *Volume of the data* being captured
- *Variety of data sources,* data types, and data formats
- *Velocity* at which the data is generated, the speed with which it needs to be acted upon, or the rate at which it is changing
- *Veracity* of the data, or rather, the uncertainty or trustworthiness of the data

For each dimension, we include key questions. Assign a weight and priority for each dimension, according to the business context. The assessment will vary by business case and by organization. Consider working through these questions in a series of workshops with the relevant business and IT stakeholders.

Business value: What insights are possible with big data technologies?

Many organizations wonder if the business insights they are seeking can be addressed by a big data solution. There are no definitive guidelines that define the insights that can be derived from big data. The scenarios need to be identified by the organization and they evolve over time. A data scientist is key to determining and identifying the business use cases and scenarios that, if implemented, will bring significant value to the business.

The data scientist must be able to understand the key performance indicators and apply statistical and complex algorithms to the data to get a list of use cases. The use cases vary by industry and business. It's helpful to study the market for what competitors are doing, which market forces are

at work, and primarily, what customers are looking for. The following table shows examples of use cases from various industries.

Table 1. Sample use cases from various industries

Industry	Sample use cases
E-commerce and online retail	E-retailers like eBay are constantly creating target offers to boost customer lifetime value (CLV); deliver consistent cross-channel customer experiences; harvest customer leads from sales, marketing, and other sources; and continuously optimize back-end processes. • Recommendation engines: Increase average order size by recommending complementary products based on predictive analysis for cross-selling. • Cross-channel analytics: Sales attribution, average order value, and lifetime value (for example, how many in-store purchases resulted from a particular recommendation, advertisement or promotion). • Event analytics: What series of steps (the golden path) led to a desired outcome (product purchase or registration, for example)? • "Right offer at the right time" and "Next-best offer": Deploying predictive models in combination with recommendation engines that drive automated next-best offers and tailored interactions across multiple interaction channels.
Retail and customer-focused	• Merchandizing and market-basket analysis • Campaign management and customer loyalty programs • Supply-chain management and analytics • Event- and behavior-based targeting • Market and consumer segmentations • Predictive analysis: Retailers want to predict factors that might be important for a buyer before the product is put on shelves
Financial services	• Compliance and regulatory reporting • Risk analysis and management • Fraud detection and security analytics • CRM and customer loyalty programs • Credit risk, scoring, and analysis • High-speed arbitrage trading • Trade surveillance • Abnormal trading pattern analysis
Fraud detection	Fraud management helps improve customer profitability by predicting the likelihood that a given transaction or customer account is experiencing fraud. Solutions analyze transactions in real time and generate recommendations for immediate action, which is critical to stopping third-party fraud, as well as first-party fraud and deliberate misuse of account privileges. Solutions are typically designed to detect and prevent a wide variety of fraud and risk types across multiple industries, including: • Credit and debit payment card fraud • Deposit account fraud • Technical fraud and bad debt • Healthcare fraud • Medicaid and Medicare fraud • Property and casualty insurance fraud • Workers' compensation fraud • Insurance fraud
Web and digital media	Much of the data we currently work with is the direct consequence of increased social media and digital marketing. Customers generate a trail of "data exhaust" that can be mined and put to use. • Large-scale click-stream analytics • Ad targeting, analysis, forecasting, and optimization • Abuse and click-fraud prevention • Social graph analysis and profile segmentation

	• Campaign management and loyalty programs
Public sector	• Fraud detection • Threat detection • Cyber-security • Compliance and regulatory analysis • Energy consumption and carbon footprint management
Health and life sciences	• Health insurance fraud detection • Campaign and sales program optimization • Brand management • Patient care quality and program analysis • Medical device and pharmaceutical supply-chain management • Drug discovery and development analysis
Telecommunications	• Revenue assurance and price optimization • Customer churn prevention • Campaign management and customer loyalty • Call Detail Record (CDR) analysis • Network performance and optimization • Mobile user location analysis
Utilities	Utilities run big, expensive, complicated systems to generate power. Each grid includes sophisticated sensors that monitor voltage, current, frequency and other important operating characteristics. Efficiency means paying careful attention to all of the data streaming from the sensors. Utilities are now leveraging Hadoop clusters to analyze power generation (supply) and power consumption (demand) data via smart meters. The adoption of smart meters has resulted in a deluge of data flowing at unprecedented levels. Most utilities are ill-prepared to analyze the data once the meters are turned on.
Media	In the cable industry, big data can be used to analyze set-top box data on a daily basis by large cable operators such as Time Warner, Comcast, and Cox Communications. This data can be leveraged to adjust advertising or promotional activity.
Miscellaneous	• Mashups: Mobile user location and precision targeting • Machine-generated data • Online dating: A leading online dating service uses sophisticated analysis to measure the compatibility between individual members, so it can suggest good matches • Online gaming • Predictive maintenance of aircraft and automobiles

Potential customers are generating huge amounts of new data on social networks and review sites. Within the enterprise, transactional data and web logs are growing as customers switch to online channels to conduct business and interact with companies.

Prioritizing the data

Start by creating an inventory of the data that exists within the enterprise. Identify data that exists in the internal systems and applications and data coming in from third parties. If a business problem can be solved with existing data, data from external sources might not be required.

Consider the cost of building a big data solution and weigh it against the value of the new insight to the business.

When this new data is analyzed in the context of the archived data about existing customers, businesses gain insight into new business opportunities.

Big data can offer a viable solution if:

- The value generated by the insight developed from the data is worth the capital cost of investing in a big data solution
- Customer-facing scenarios demonstrate the potential value from the insight

When evaluating the business value to be gained by a big data solution, consider your whether your current environment can be expanded and weigh the cost of this investment.

Can my current environment be expanded?

Ask the following questions to determine if you can augment the existing data warehouse platform?

- Are the current datasets very large — on the order of terabytes or petabytes?
- Does the existing warehouse environment contain a repository of *all* data generated or acquired?
- Is there a significant amount of cold or low-touch data that is not being analyzed to derive business insight?
- Do you have to throw data away because you are unable to store or process it?
- Do you want to be able to perform data exploration on complex and large amounts of data?
- Do you want to be able to do analysis of non-operational data?
- Are you interested in using your data for traditional and new types of analytics?
- Are you trying to delay an upgrade to your existing data warehouse?
- Are you looking for ways to lower your overall cost of doing analytics?

If the answer to any of these questions is yes, explore ways to augment the existing data warehouse environment.

What is the cost of expanding my current environment?

The cost and feasibility of extending an existing data warehouse platform or IT environment vs. implementing a big data solution depends on:

- Existing tools and technology
- Scalability of the existing system
- The processing power of the existing environment
- The storage capability of the existing platform
- Governance and policies in force
- The heterogeneity of existing IT applications
- The technology and business skills that exist in the organization.

It also depends on the volume of data that will be gathered and collected from new data sources, the complexity of business use cases, the analytical complexity of processing, and how expensive it is to get the data and people with the right skill set. Can the existing pool of resources develop new big data skills or can the resources with niche skills be hired externally?

Keep in mind that the effect of a big data initiative on other projects under way. Acquiring data from new sources is costly. It's important to first identify any data that exists internally in the systems and applications and in third-party data being received currently. If a business problem can be solved with existing data, data from external sources may not be required.

Assess the application portfolio of the organization before procuring new tools and applications. For example, a plain vanilla Hadoop platform may not be sufficient for the requirements, and it may be necessary to buy specialized tools. Or in contrast, a commercial version of Hadoop may be expensive for the current use case, but may be needed as a long-term investment to support a strategic big data platform. Consider the cost of the infrastructure, hardware, software, and maintenance required by for big data tools and technologies.

Governance and control on data: What is the impact on existing IT governance?

When deciding whether to implement a big data platform, an organization might be looking at new data sources and new types of data elements where the ownership of the day is not clearly defined. Certain industry regulations govern the data that is acquired and used by an organization. For example, in the case of healthcare, is it legitimate to access patient data to derive insight from the data? Similar rules govern all industries. In addition to issues of IT governance, business processes of an organization may also need to be redefined or modified to enable the organization to acquire, store, and access external data.

Consider the following governance-related issues in the context of your situation:

- **Security and privacy**— In keeping with local regulations, what data can the solution access? What data can be stored? What data should be encrypted during motion? At rest? Who is allowed to see the raw data and the insights?
- **Standardization of data**— Are there standards governing the data? Is the data in a proprietary format? Is some of the data in a non-standard format?
- **Timeframe in which the data is available**— Is the data available in a timeframe that allows action to be taken in a timely fashion?
- **Ownership of data**— Who owns the data? Does the solution have appropriate access and permission to use the data?
- **Allowable uses:** How is the data allowed to be used?

Can I incrementally implement a big data solution?

A big data solution can be incrementally implemented. It's helpful to clearly define the scope of the business problem and to set, in measurable terms, the expected business revenue gain.

For the foundational business case, take care in outlining the scope of the problem and projected benefits from the solution. If the scope is too small, the business benefits will not be realised, and if it's too large, it will be challenging to get the funding and complete the project inside an appropriate timeframe. Define the core functions in the first iteration of the project, so that it's easy to win the confidence of stakeholders.

People: Are the right skills on board and the right people aligned?

Specific skills are required to understand and analyze the requirements and maintain the big data solution. These skills include industry knowledge, domain expertise, and technical knowledge on big data tools and technologies. Data scientists with expertise in modeling, statistics, analytics, and math are key to the success of any big data initiative.

Before undertaking a new big data project, make sure the right people are on board:

- Do you have buy-in from stakeholders and other business sponsors who are willing to invest in the project?
- Are data scientists available who understand the domain, who can look at the massive quantity of data and who can identify ways to generate meaningful and useful insights from the data?

Is there existing data that can be used to get insight?

All organizations have quite a lot of data not being harnessed for business insight. Pockets include log files, errors files, and operational data from applications. Don't overlook this data as a potential source of valuable information.

Is the data complexity increasing?

Look for hints that the complexity of data has increased, especially with regard to volume, variety, velocity, and veracity.

Has the volume of data increased?

You may want to consider a big data solution if:

- The data is sized in petabytes and exabytes, and in the near future, might grow to zetabytes.
- The data volume is posing technical and economic challenges to store, search, share, analyze, and visualize using traditional methods, such as relational database engines.
- The data processing can currently make use of massive parallel processing power on available hardware.

Has the variety of data increased?

The variety of data might demand a big data solution if:

- The data content and structure cannot be anticipated or predicted.
- The data format varies, including structured, semi-structured, and unstructured data.
- The data can be generated by users and machines in any format, for example: Microsoft® Word files, Microsoft Excel® spreadsheets, Microsoft PowerPoint presentations, PDF files, social media, web and software logs, email, photos and video footage from cameras, information-sensing mobile devices, aerial sensory technologies, genomics, and medical records.

- New types of data have emerged from sources that weren't previously mined for insight.
- Domain entities take on different meanings in different contexts.

Has the velocity of the data increased or changed?

Consider whether your data:

- Is changing rapidly and must be responded to immediately
- Has overwhelmed traditional technologies and methods, which are no longer adequate to handle data coming in real time

Is your data trustworthy?

Consider a big data solution if:

- The authenticity or accuracy of the data is unknown.
- The data includes ambiguous information.
- It's unclear whether the data is complete.

A big data solution might be appropriate if there is reasonable complexity in the volume, variety, velocity, or veracity of the data. For more complex data, assess any risks associated with implementing a big data solution. For less complex data, traditional solutions should be assessed.

Is all big data a big data problem?

Not all big data situations require a big data solution. Look for hints in the market. What are competitors doing? What market forces are at work? What are the customers demanding?

Use the questions in this article to help you determine whether a big data solution is appropriate for your business situation and for the business insight you need. If you've decided it's time to embark on a big data project, watch for the next article on defining a logical architecture and determining the key components required for your big data solution.

Resources

Learn

- "Big data architecture and patterns, Part 1: Introduction to big data classification and architecture" defines key concepts for building the architecture for a big data solution.
- Check out the Big Data Hub to find popular links such as "The four V's of big data" and "Top 5 big data use cases."
- What is a data scientist? and what does one do?
- Learn more about big data in the developerWorks big data content area. Find technical documentation, how-to articles, education, downloads, product information, and more.
- Find resources to help you get started with InfoSphere BigInsights, IBM's Hadoop-based offering that extends the value of open source Hadoop with features like Big SQL, text analytics, and BigSheets.
- Follow these self-paced tutorials (PDF) to learn how to manage your big data environment, import data for analysis, analyze data with BigSheets, develop your first big data application, develop Big SQL queries to analyze big data, and create an extractor to derive insights from text documents with InfoSphere BigInsights.
- Find resources to help you get started with InfoSphere Streams, IBM's high-performance computing platform that enables user-developed applications to rapidly ingest, analyze, and correlate information as it arrives from thousands of real-time sources.
- Stay current with developerWorks technical events and webcasts.
- Follow developerWorks on Twitter.

Get products and technologies

- Check out the many IBM big data products available for trial download.
- Download InfoSphere BigInsights Quick Start Edition, available as a native software installation or as a VMware image.
- Download InfoSphere Streams, available as a native software installation or as a VMware image.
- Use InfoSphere Streams on IBM SmartCloud Enterprise.
- Build your next development project with IBM trial software, available for download directly from developerWorks.

Discuss

- Connect with IBM big data experts on Twitter.
- Ask questions and get answers in the InfoSphere BigInsights forum.
- Ask questions and get answers in the InfoSphere Streams forum.
- Check out the developerWorks blogs and get involved in the developerWorks community.
- IBM big data and analytics on Facebook.

About the authors

Divakar Mysore

Divakar Mysore is an IBM-certified senior IT architect with more than 15 years of experience in the IT industry. He has been part of multiple strategic initiatives for global corporations. He has extensive experience as enterprise architect, application architect, system engineer, data modeler, and test architect. He leads the application architecture discipline for the Enterprise Architecture and Technology team in Global Delivery India. He drives technical vitality initiatives on mobile, front office, social and big data.

Shrikant Khupat

Shrikant Khupat is an IBM application architect. He is experienced in defining enterprise class, distributed, disconnected, client-server architectures, and designs. He has exposure to a variety of domains, such as insurance and energy and utilities. He worked on complex solutions involving distributed data processing using Apache Hadoop and unstructured data processing using machine learning languages. His current interests include defining big data architecture and patterns.

Shweta Jain

Shweta Jain is an accredited IT architect with IBM AIS Global Delivery with more than 10 years of industry experience. She specializes in architecting SOA-based integration solutions using industry standards and frameworks. She has experience in architecture, design, implementation, and testing of integration solutions based on the SOA framework, SOMA methodology, and the software development life cycle based on methods. As an integration architect, she is responsible for architecting the BPM/EAI layer using IBM tools, standards, processes, and methodologies; and incorporating industry standards for complex integration and transformation projects. She also enjoys reading and contributing to latest technologies, such as big data.