Enhancing Enterprise Systems with Big Data

An IBM Redbooks® Point-of-View publication by the IBM Business Analytics team

By Craig Statchuk, IBM Technical Architect, and **Dan Rope**, IBM Visualization Researcher

Highlights

Big data is growing rapidly and driving a need within each enterprise for new tools and methodologies. This new data, when it is used in a comprehensive system, offers the potential for unique insights that drive new actions across the entire enterprise.

Here are three areas that take advantage of big data:

- Marketing intelligence and media awareness
- Corporate governance, risk management, and regulatory compliance
- Operational efficiency

Taking advantage of big data

The rapid growth in big data highlights a need for new tools and methodologies that can augment your best practices. You can reap tangible benefits by correlating the existing data you trust with a wide range of new sources. Big data for business analytics focuses on managing and understanding the transition to a trusted view of new data. Finding new sources of information is not necessarily difficult; the hard part is finding the best correlations with the data you use today.

Big data has an inherent focus on new sources of information, originating from a wide range of providers, in many formats. The attributes of volume, velocity, variety, and veracity (Figure 1) are used to distinguish big data from more traditional enterprise data.

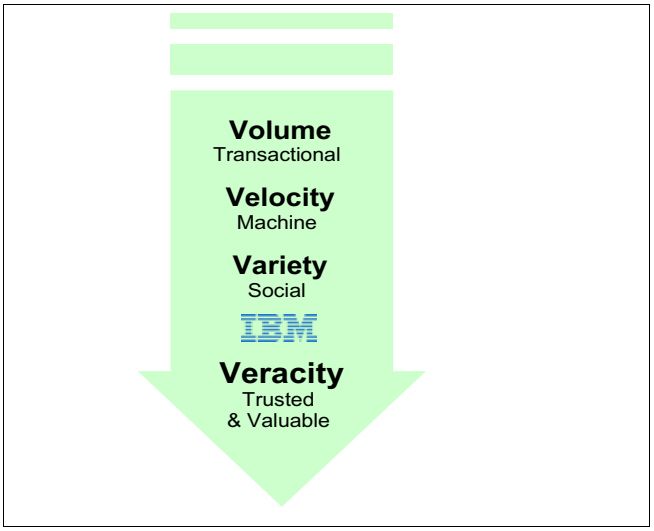

Figure 1 Volume, velocity, variety, and veracity

Here are examples of these attributes:

- Volume

  Data that is provisioned in an unprocessed state exemplifies this type of data. This data helps run the business when the data is rolled up into categories that consist of correctly aligned data. This type of data also provides a starting point for further discovery and analysis.

- Velocity

  In motion data from automated sources represents velocity. Although this data is useful in signaling- or event-driven systems, this data can be difficult to relate to other parts of a business.

- Variety

  This infinitely variable source of knowledge can lead to new business insight or nothing at all. Finding the relevant “needles” in this giant “haystack” is the ongoing challenge with this constantly evolving data source.

- Veracity

  Veracity represents the valuable information that is trusted throughout an enterprise. It is both relevant and flexible. Its lineage can be traced to source data. Groups of items roll up into summaries as needed. Attributes such as cardinality and uniqueness are maintained to assist with enterprise integration.

Start by thinking about the data

First, think about how new information sources can change the way you work:

- Get ready to iterate.

  Your first try is never your best. Expect to learn as you iterate through different analytics.

- Seek higher value.

  You need to understand the value of data. Curated and cleansed information is often harder to acquire but generally reflects value in the domain from which it was collected. Aligning new data with existing business dimensions can increase relevance and importance.

- “Time is money” is now more relevant than ever.

  Traditional metrics can track things such as slow product delivery times. If your customers are unhappy about a late delivery, you might need to look outside your current data sources to discover the problem in a timely manner. Do not wait for reports to show dissatisfaction through a basic sales metric. Find out now so you can do something about this dissatisfaction.

Repetitively experimenting with new sources of data and then aligning the results with what you already know is a key technique for building trusted repositories of new data.

Creating an integrated solution

As an example scenario, we describe a fictitious electric car company that is expanding their marketing, compliance, and operational systems. This activity is part of a comprehensive solution that evolves with their rapidly growing business.

The following three big data solutions are highlighted:

- Marketing intelligence and media awareness
- Corporate governance, risk management, and regulatory compliance
- Operational efficiency

Marketing intelligence and media awareness

In this example scenario, products are introduced and supported through various channels, including mobile platforms, the web, telephone, and showrooms. Each channel provides metrics that are related to marketing, sales, and other operational elements in the business. There is a strong desire by the company to integrate data from these diverse sources to gain new insights, understanding, and efficiency.

Expectations for the marketing intelligence system are equally wide ranging, with the following identified benefits:

- Multi-channel analysis, including new inputs as they are introduced. Initial monitored inputs include online activities, showroom visits, and call center transcript analysis.
- The ability to identify unique customers across channels in different scenarios. For example, identifying the same customer on Twitter and a customer support website.
- Faster access to data as campaigns are run.
- Post campaign analysis over numerous attributes and dimensions.
- Identify profitable customers and predict future candidates.
- Personalization of offers and incentives.
- Better return on marketing investments.

Understanding target customer personas is a primary marketing concern. A definition can be difficult for many reasons:

- Limited sales history
- Many new markets and demographics
- Confusing social media data that does not always align with preconceived notions of customer desires

Like many enterprises, this company lacks a comprehensive view of customers. Their existing databases have incomplete metrics or are not flexible enough to deal with the new types of data that are constantly introduced. Customer data is insufficient in several areas:

- Reasons for buying
- Other products that are considered, including key differentiators
- Past buying preferences from all available sources.

Areas of new interest include the following ones:

- Understanding customer activity on social media
- Effectiveness of corporate usage through new media channels
- Gauging the effectiveness of offers in pre- and post-sales scenarios

Media awareness includes management of media that is deemed crucial during early phases of product adoption. Timely analysis and response is viewed as a competitive advantage. Key areas of interest include the following ones:

- Levels of social media activity
- Identification of influencers
- Positive and negative statements
- Competitor monitoring

Governance, risk, and compliance

As part of a governance and risk assessment initiative for the company, numerous external factors are tracked against a model to mitigate negative business:

- Governance

  Policies that focus on business and finance objectives, including overseeing a possible initial public offering (IPO) soon.

- Risk management

  Analyzing cooperative and disruptive forces, such as economic, political, and competitive factors. Models use input from news media, stock quote services, consumer confidence ratings, and other financial metrics.

- Regulatory compliance

  Monitoring that requires tracking many external metrics, including ad campaigns with an assessment of statements that are attributed to officers of the company.

Operational efficiency

Although not part of the original vision, the company quickly determined that the business could benefit from the big data methodologies that are embraced elsewhere in the organization. Here are the expected benefits from this initiative:

- Supply chain optimization with integrated monitoring and modeling of telemetry from new vehicles in the field. Coverage and thresholds are dynamic because they could never be fully modeled in the past. The expectation is better responsiveness and agility to unforeseen problems.
- Predictive algorithms to highlight future problem areas that are based on experiences of early customers. Results are fed to logistics and inventory systems to potentially divert new parts into the customer support chain when they are urgently required.

Exploring solution capabilities

Traditional database systems can be limited by their predefined structures. New dimensions, measures, or attributes are not always easy to accommodate in an enterprise-compliant manner. Although it is essential for operations and regulatory compliance, the ability to augment and correlate data sets is increasingly important.

For example, generic marketing campaigns can waste media dollars. They might reach a broad audience but miss offering specific customers a more personalized message. Segmenting, tracking, and scoring across various sources provide insight for targeted campaigns. The solution supports cross-channel analytics that are used to follow customer activity from an initial website visit to test driving and to final sale. Finding correlations between these activities is seen as a competitive advantage.

New types of data

New types of data introduce measures of uncertainly that are reflected in three primary attributes:

- Quality
- Relevance
- Flexibility

Uncertain data means that quality, relevance, and flexibility can work in opposition (Figure 2).

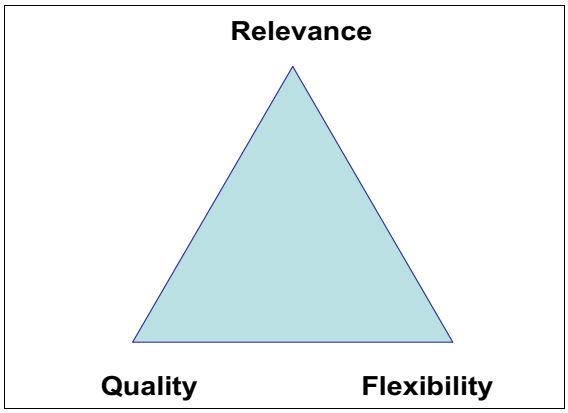

Figure 2 Uncertain data

Introducing the landing zone area

A landing zone area provides staging for new data as it is pulled from telemetry systems, external sites, and other parts of the enterprise. Data is stored in a combination of IBM® InfoSphere® BigInsights™ Hadoop Distributed Files Systems (HDFS) and more traditional database systems. These traditional systems are optimized for large data volumes and fast access with IBM DB2® with Blu Acceleration. The result is a system with these characteristics:

- Greater flexibility because data can be transformed to meet the needs of a wider variety of applications
- Ability to support iterative analytics that ultimately improve data quality
- Access to existing warehouses and data marts to improve relevance

Associating new data with existing data

Integration with business metadata helps build trust in new data sources (Figure 3).

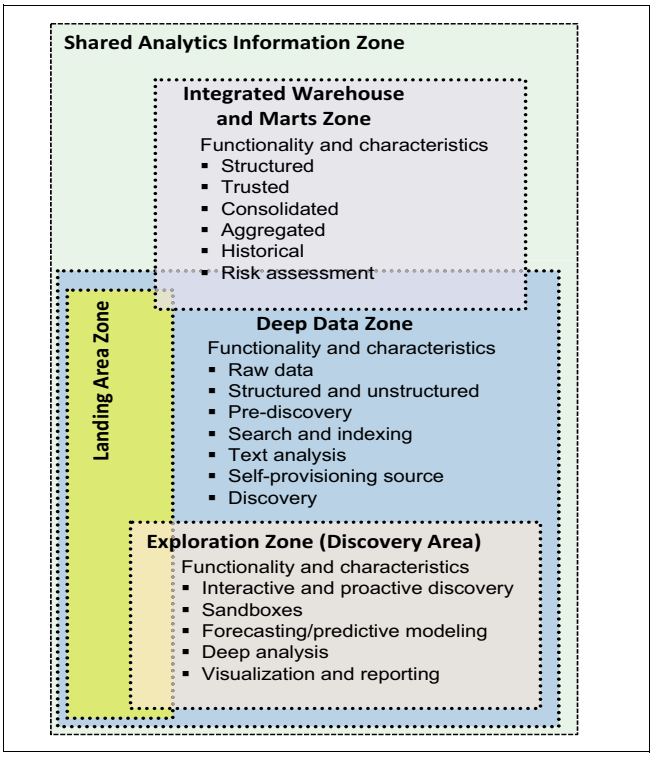

Figure 3 Integrating data

New associations are built to optimally align data with categories and topics that are used by the business. Reporting warehouses are extended with links to new data that is prepared in IBM InfoSphere BigInsights. Relational access is provided through a Hive HDFS interface called IBM Big SQL. IBM SPSS® Modeler is used to align data with new classifications based on probability models. The results are used to augment existing SQL-based reporting data marts. Other tools are used to support the solution:

- IBM SPSS Analytic Catalyst is used to find new correlations between multiple data columns.
- IBM InfoSphere Data Explorer provides faceted search finding associations in unstructured data.

IBM SPSS software solutions

IBM SPSS software helps analysts organize, understand, and ultimately ask relevant questions about new data. The process begins with a notion of the business problem. One such example is knowing what drives a customer toward a vehicle purchase and the important factors that later determine satisfaction.

Customer data is collected with as many attributes as possible. Entity analytics is used with standard data merging techniques to create a wide data set rich in customer demographics, website activities, and survey results. Third-party data is used to supplement internal data where possible. The results are a large variety of data sources that work together to form a complete perspective.

Data is loaded into an IBM InfoSphere BigInsights HDFS in the Exploration Zone (Figure 3 on page 4). IBM SPSS Analytic Server then provides predictive analytic functions. This topology ensures that analytics operate in close proximity to source information and avoids large data transfers. IBM SPSS Analytic Catalyst directly uses this data for insight discovery by relying on the predictive capabilities that are provided by IBM SPSS Analytic Server to find statistically important relationships.

The process is completed by selecting target fields of primary interest, which are explained in terms of other fields that are present. For example, it is desirable to understand why a customer makes a purchase. It is also important to know factors that are related to overall satisfaction. Such collections of data are designated as *target fields*. IBM SPSS Analytic Catalyst finds the fields that best explain the target field variations and uses related predictive models to help explain a statistical relationship.

In our example scenario, IBM SPSS Analytic Catalyst can find a purchase decision that is related to income level, age, and the amount of time that is spent on the website. Also, it can show correlations with previous hybrid vehicle buyers. Satisfaction is found to be driven primarily by annual mileage, charger model, number of total vehicles that are owned, and the financing options that are used. Based on these insights, the company can take tactical actions that might increase purchase decisions or improve satisfaction. For example, they could offer to bundle a higher-end charger with a lease financing option because the insight reveals that better chargers have a real impact on satisfaction. The company might also decide to invest in more vehicle charging stations, as this could improve satisfaction for those customers that have longer driving distances, or the company might choose to modify the website based on the age of likely purchasers.

Predictive modeling and deployment

Advertising campaigns need accurate and cost-effective statistical models to predict future buyers. To achieve this goal, a predictive model for purchase decisions is identified with IBM SPSS Analytic Catalyst. The IBM SPSS Modeler data mining workbench is then used to make iterative improvements to the model. For example, a predictive model helps when you do not know purchase history for the new audience for a marketing campaign. A data analyst can work with IBM SPSS Modeler to build an optimal model to score the likelihood that a potential customer becomes a purchaser. That model can then be reused to score the universe of candidates to receive electric car marketing materials, selecting only those with a greater than 50% probability of making a purchase. Model deployment operationalizes predictive modeling capabilities and is where real business value is realized. IBM SPSS software supports various deployment solutions:

- IBM WebSphere® software using IBM SPSS Collaboration and Deployment Services
- IBM InfoSphere Streams for real-time scoring
- Database scoring support for IBM Netezza, IBM DB2, and other solutions for high-performance batch scoring

Reporting and analysis

Enterprise reporting and analysis are key elements in an integrated big data solution. Here are the key capabilities:

- Building ad hoc data marts with IBM Cognos® Insight and IBM Cognos TM1® to investigate new correlations.
- Exporting results to IBM Cognos 10 Reporting
- Updating enterprise master data management definitions to reflect new correlations and associations.

Data that is related to outcomes such as customer satisfaction typically need further analysis. IBM InfoSphere BigInsights is used to provision new data sources in a worksheet that contains key fields and measures. This data, along with the results of analytic exploration and discovery, are exported to an IBM Cognos Report Studio compliant data warehouse. This data can then be used for generating new reports and dashboards. IBM Cognos TM1 is used for additional ad hoc analysis and experimentation along newly identified reporting dimensions.

Supporting GRC

Big data is a key element in the governance, risk, and compliance (GRC) solution. IBM OpenPages® defines various governance policies by specifying inputs from news and financial sites along with data that is mined from various social media providers:

- Feeds that are extracted from sources such as Twitter and LinkedIn
- Customer comments on support websites
- Summaries that are collected from leading news organizations
- Government supplied financial indicator data

In this scenario, metrics that are related to the perceived state of the automobile market and business conditions are gathered with social media analytics software from IBM.

Related information is augmented and merged into risk assessment applications that are processed in near real time. Rules are interrogated and assessed with IBM Algorithmics® software. Alerts and exception conditions are managed with IBM ILOG® software. Compliance reporting is performed with IBM Cognos software-based analysis applications that are deployed on desktop and mobile devices.

Operational efficiency

The company's operational systems are improved with a complete view of the entire business. Data in disparate warehouses, spreadsheets, and unstructured sources can all be correlated to uncover new insight.

An example is vehicle telemetry that is collected and filtered by IBM InfoSphere Streams creating live summaries of customer vehicle performance. Battery efficiency, vehicle safety, and warning systems are all monitored. External sources, such as social media, are monitored for early warnings that are related to subjects as diverse as satisfaction, reliability, or product usage. IBM InfoSphere Streams is used with IBM Algorithmics as a centralized event and external data clearing house. IBM SPSS Analytic Server is used for predictive model building and for deployment of high volume data. Problems can be relayed to any part of the business for more timely and effective response.

Here are some examples of operational efficiency:

- Spare parts prioritization that is based on live vehicle telemetry and inventory levels
- Input to predictive systems allowing anomalies to be forecasted with greater accuracy
- Automated problem prioritization and escalation
- Ongoing policy compliance assessment that is based on external data
- Data warehouse augmentation through updates from outside sources, including social media providers
- Timely and relevant information made available to the entire enterprise

What's next: How IBM can help

IBM Business Analytics can help your organization extract significant value from big data. IBM provides various ways to blend traditional information with big data. IBM provides new types of analytic techniques and smarter visualizations to ensure that you can capitalize on the unique characteristics of big data.

For more information about this topic, go to the following website:

http://www.ibm.com/software/analytics/solutions/big-data/

Resources for more information

For more information about the concepts that are highlighted in this document, see the following resources:

- *Analytics in a Big Data Environment*, REDP-4877
- *Artificial Intelligence: Learning Through Interactions and Big Data*, REDP-4974
- *Context-Based Analytics in a Big Data World: Better Decisions*, REDP-4962
- *Smarter Analytics: Information Architecture for a New Era of Computing*, REDP-5012
- IBM big data platform

  http://www.ibm.com/software/data/bigdata/

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to: *IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document, REDP-5055-00, was created or updated on October 24, 2013.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (or), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at `http://www.ibm.com/legal/copytrade.shtml`

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Algorithmics®
BigInsights™
Cognos®
DB2®
IBM®
ILOG®
InfoSphere®
OpenPages®
Redbooks®
Redbooks (logo)
SPSS®
TM1®
Velocity™
WebSphere®

The following terms are trademarks of other companies:

Netezza, and N logo are trademarks or registered trademarks of IBM International Group B.V., an IBM Company.

Other company, product, or service names may be trademarks or service marks of others.