Addressing customer analytics with effective data matching

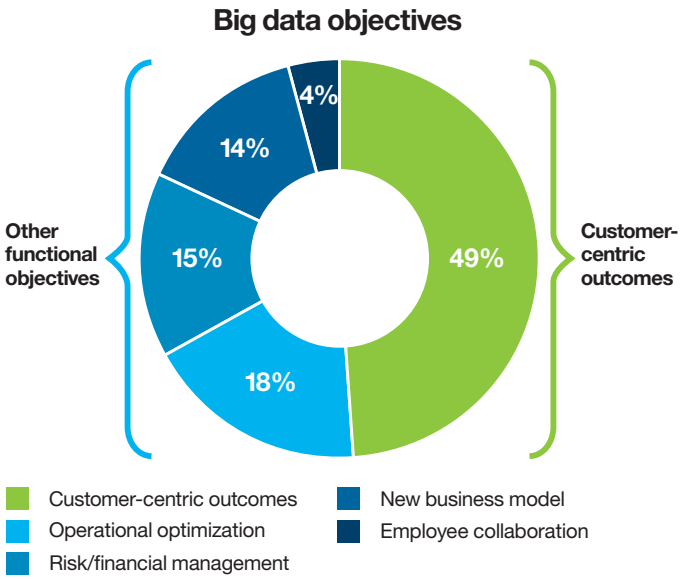
Analyze multiple sources of operational and analytical information with IBM InfoSphere Big Match for Hadoop

IBM

Developing customer behavior insight with big data and analytics

With the advent of big data, organizations worldwide are attempting to use data and analytics to solve problems previously out of their reach. Many are applying big data and analytics to create competitive advantage within their markets, often focusing on building a thorough understanding of their customer base.

High-priority big data and analytics projects often target customer-centric outcomes such as improving customer loyalty or improving up-selling. In fact, an IBM Institute for Business Value study found that nearly half of all organizations with active big data pilots or implementations identified customer-centric outcomes as a top objective (see Figure 1).[1] However, big data and analytics can also help companies understand how changes to products or services will impact customers, as well as address aspects of security and intelligence, risk and financial management, and operational optimization.



Top functional objectives identified by organizations with active big data pilots or implementations. Responses have been weighted and aggregated.

Total respondents n = 1,061

Source: IBM Institute for Business Value. "Analytics: The Real-World Use of Big Data."

Figure 1. According to an IBM Institute for Business Value survey, nearly half of respondents' big data efforts target customer-centric outcomes.

The era of big data is opening up much larger volumes and new, unstructured varieties of data for analysis, all of which informs a full view of a customer. But as organizations begin to execute on big data and analytics projects, many quickly run into a road-block: How do they correlate the complete, accurate customer information necessary to perform a particular analysis? And how do they do it without moving data from source to source (which increases the risk of errors or data loss)?

For example, creating a complete picture of a single customer requires locating and combining data from both traditional and big data sources (see Figure 2), including:

- A terabyte of records containing the last 12 months of orders
- The last month of website log data
- Three to six months of social media information, such as tweets, Vine videos or Instagram feeds
- Semi-structured information derived from call-center notes

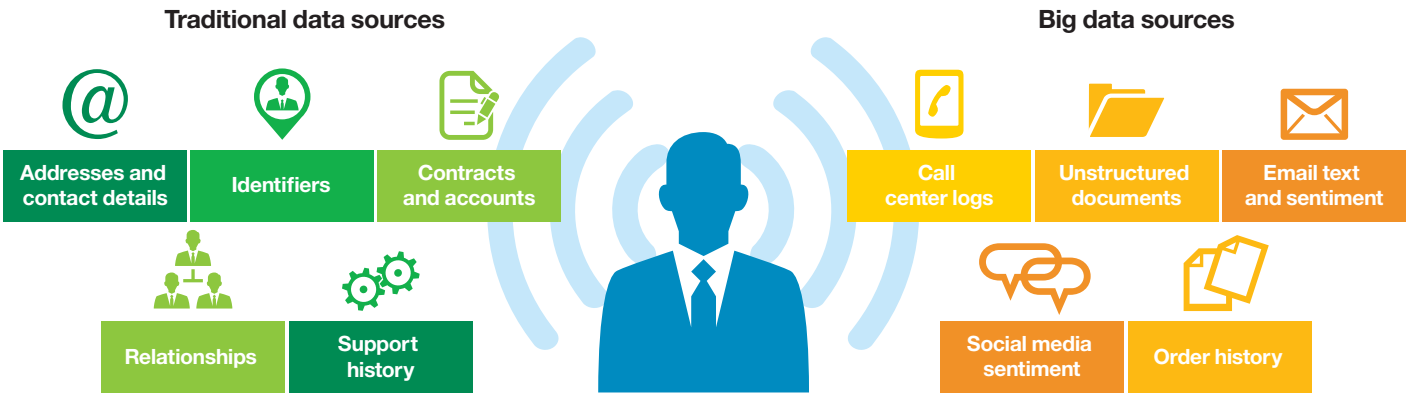


Figure 2. Building a complete view of a customer requires tapping a wide variety of data ranging from addresses and contact details to social media sentiment.

Matching technology is a valuable tool for connecting the dots across multiple data sources to establish a deeper understanding of an organization's customers. This, combined with the ability to move processing closer to the data and minimize data movement, allows for an order-of-magnitude performance improvement over traditional relational database management system (RDBMS) processing.

This white paper discusses the need for effective matching in big data environments, the top three challenges of customer analytics in a big data environment and how an effective matching engine that is optimized for big data, such as IBM® InfoSphere® Big Match for Hadoop, helps organizations connect information in a scalable, accurate manner.

The importance of a big data matching solution

Realizing positive customer-centric outcomes requires insights derived from data that already exists within an organization as well as external and unstructured data (see Figure 3). Establishing the lineage of this data gives organizations confidence in their information—a critical factor when making the decisions necessary to differentiate customer engagements. It also helps systems accurately process the most appropriate data elements for the task at hand, such as providing insight into customer actions across a region or helping teams learn more about an individual customer.

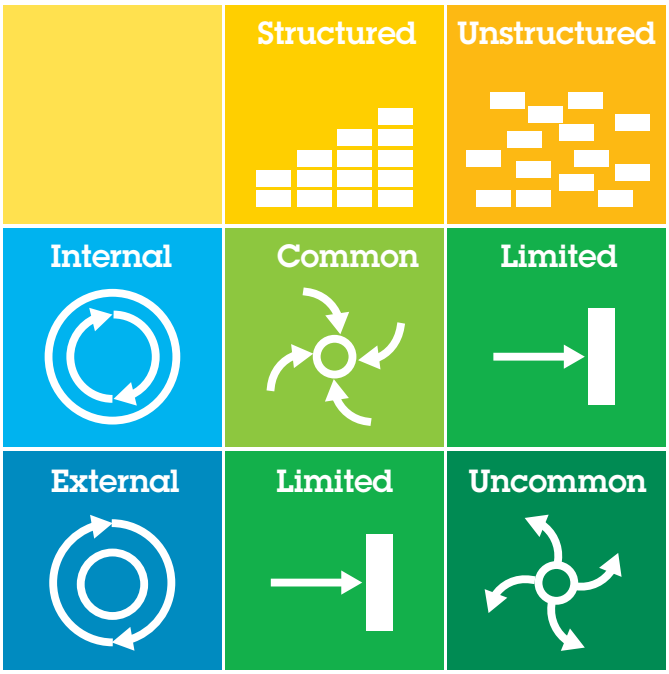


Figure 3. The combination of structured and unstructured data from internal and external sources creates a more complete view of a customer.

Matching technologies play a critical role in enhancing data accuracy by streamlining the process of sorting out incorrect, duplicate and related data. This process is essential to establish a solid, trustworthy foundation of information for customer analysis. Matching engines may be partnered with related capabilities such as text analytics to help extract relevant information and connect it to customer profiles, building an enhanced view of the customer.

A recent Forrester survey shows that governing customer data is deemed more important than managing other forms of data, with 79 percent of respondents saying customer data should be somewhat or highly governed.[2] This interest in using trusted data to know customers better and in greater detail is driving interest in what IBM calls the enhanced 360-degree view of the customer. This approach focuses on using a wide range of structured, unstructured, internal and external data to assemble a complete view of the customer.

The challenges of accurate matching become much more complex as a company combines traditional customer records with big data sources. The consequences of this inaccuracy can range from unfortunate to catastrophic: mislinking two records in a mailing list means that a potential customer may not receive promotional material, while faulty record links at a healthcare organization may give a physician incorrect information in developing a course of treatment and result in patient harm.

As the numbers of records start to move into the hundreds of millions, with more records arriving daily, a traditional matching engine rapidly becomes overwhelmed. Ideally, a big data matching solution uses distributed probabilistic matching that lends itself to the distributed paradigm used by big data environments. This probabilistic matching capability does not rely on an operational server, but instead uses pre-configured algorithms and distributed processing to analyze data. The algorithms standardize and bucket the data, and then compare data elements to uncover differences and similarities (see Figure 4). The matching technology compares the data elements and assigns similarity scores, or weights, that take into account the probability of certain data elements matching based on the distribution of data in the actual data set. This is what makes the matches truly probabilistic.

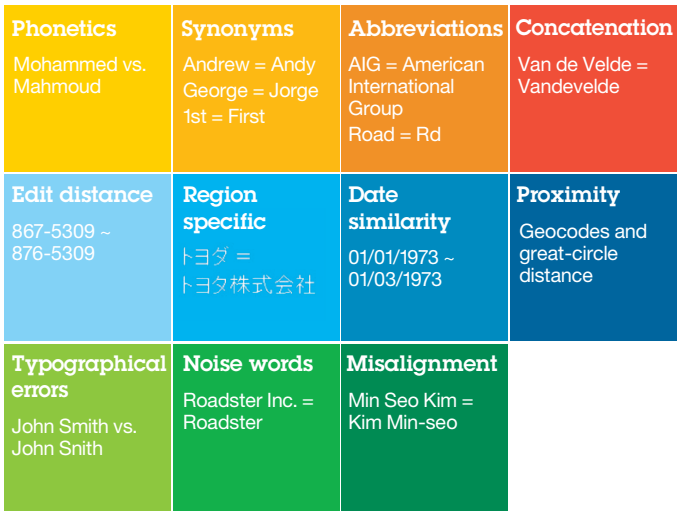


Figure 4. Examples of probabilistic matching

Identifying three top customer analytics challenges

Not all matching solutions are optimized for customer analytics or big data. As companies embrace the era of big data, they must change the way they access, analyze and process information. The following represent three top customer analytics challenges in a big data environment.

1. **Ungoverned data:** Customer data comes from internal and external sources and may be structured or unstructured. It is also often ungoverned—sources like social data are not always subject to normal enterprise quality-control processes. Organizations need a matching solution that adapts to the data quality variations from a variety of sources.

2. **Unprecedented volume:** Organizations cannot compromise on the quality of their analytical decisions even as data volumes increase exponentially. They need scalable solutions to process and analyze larger amounts of customer data and efficiently deliver results to the systems and people who need it.
3. **Nonstandardized data:** Because data flows in a big data architecture usually follow an extract, load and transform (ELT) pattern, there is no budget or time to undertake expensive data standardization operations prior to processing the data. Organizations need a matching tool that does not require costly, time-consuming data preprocessing.

Challenge #1: Ungoverned data

Data sets that include customer data naturally contain governed and ungoverned data, whether the source is internal or external. Internal systems often support specific business processes such as marketing, sales and customer service, and information on the same customer may be included in each system with minor or even major variations if the data is not properly governed. Data from uncontrolled, external sources such as social media channels may have even larger data quality uncertainties. These sources may be updated frequently, and it is difficult to standardize and cleanse them because of their unstructured, constantly changing nature.

As a result, big data and analytics technologies should include capabilities focused on improving matching in order to deliver a trustworthy, true view of customer information. InfoSphere Big Match for Hadoop includes quality control capabilities that enable the solution to take in any type of data in any form—clean or dirty—and deliver the same matching outcome. To help further ensure accuracy, InfoSphere Big Match for Hadoop supports matching techniques that can be configured to solve specific business problems and are designed to minimize both false negatives and false positives.

Challenge #2: Unprecedented volume

In a recent Unisphere Research survey, nearly one-third of respondents said that they manage over 100 applications, and more than 10 percent said they manage more than 1,000 applications.[3] Each of those applications may hold important data about customer activities that need to be associated with the organization's customer records to gain better insights.

Social media, loyalty programs, mobile devices and customer-facing websites also help drive data volumes to big data scale. Organizations want to capture nuggets of customer interaction from each touchpoint, knowing that the cumulative effect will enhance demographic and individual customer profiles. Powerful matching capabilities are essential to cope with this volume and variety of data without slowing business processes or delaying the timing of customer offers. In InfoSphere Big Match for Hadoop, a combination of distributed probabilistic matching, big data accelerators and text analytics extract relevant information and help connect it to customer profiles at the speed of business.

Challenge #3: Nonstandardized data

Standardization transforms similar data received in various formats to a common format. Normally, this process includes actions such as standardizing all alphabetic characters, removal of punctuation, anonymous value checks and data ordering. However, it is impractical to apply this kind of standardization when dealing with big data volumes and velocity—there's just too much data coming in too quickly.

With big data architectures promoting an ELT model over traditional methods, a big data matching solution must support nonstandardized data and enable data resolution across multiple sources prior to transforming data.

InfoSphere Big Match for Hadoop uses a data derivation approach to handle nonstandardized data, which has a significant advantage: an organization can accurately aggregate data,

understand how it relates to existing data assets and then decide which data sets merit further investment—and do so without undertaking an expensive preprocessing effort. Next, InfoSphere Big Match for Hadoop links the optimized derived data sets to associate related records by confidence level, and computes those associations. Consuming applications can use these customer relationship insights to make more effective decisions.

Connecting the customer data dots with InfoSphere Big Match for Hadoop

InfoSphere Big Match for Hadoop software addresses all three customer analytics challenges by breaking down barriers between internal and external data sources and operational applications. It serves as an essential platform for gaining insights from data prior to sending it to other systems for more detailed processing. It can also accommodate data in its native format and in real time or as batch files from individual records, internal data sources, business partners and other third-party sources, which helps speed up processing.

InfoSphere Big Match for Hadoop algorithms achieve a high level of matching accuracy by using a multistep process:

- **Optimize data for statistical comparisons:** The matching engine normalizes and compacts data, creating a separate derived data layer while the source data remains intact.
- **Find all the potential matches:** Searching widely in the derived pool to locate potential matches, the matching engine uses multiple search keys for each record to avoid errors.
- **Score using probabilistic statistics:** InfoSphere Big Match for Hadoop produces links and scores to indicate which records likely represent the same entity and the strength of that likelihood.
- **Assign a match result:** These results are based on preconfigured weight thresholds that can be modified according to the data that needs to be processed. Potential matches that do not meet the thresholds are not linked as entities. Search application programming interfaces (APIs) enable users to pass a threshold as a parameter, allowing them to look at a broader range of results depending on the business problem they are trying to address.

In this way, InfoSphere Big Match for Hadoop enables the fast, accurate and efficient linking needed to derive deeper customer insight from big data.

InfoSphere Big Match for Hadoop: The highlights

InfoSphere Big Match for Hadoop leverages capabilities from IBM InfoSphere BigInsights™, including text analytics and big data accelerators, as well as the probabilistic matching engine from IBM InfoSphere Master Data Management Enterprise Edition.

InfoSphere Big Match for Hadoop features include:

- Persisted store to keep track of matching IDs
- Easy scale-up by adding data nodes to the environment
- Bulk extract utility for using the same matching results in multiple consuming systems
- API-based support, including Java and REST-based APIs

Building a robust analytics foundation

Customer-centric initiatives drive many big data investments. Matching customer data that comes from multiple internal and external sources and varies in quality is critical to understanding your customer.

Your organization can perform that matching quickly and easily with InfoSphere Big Match for Hadoop. It helps resolve big data matching challenges with tried-and-tested techniques—and does so in a way that offers quick time to value, leveraging industry best practices.

By providing a robust foundation of information for customer analytics, InfoSphere Big Match for Hadoop helps organizations build an enhanced 360-degree view of their customers. It enables organizations to create a robust profile that encompasses a customer's every touch point with the organization, giving business leaders and knowledge workers the ability to innovate and act with confidence to drive strategic, customer-centric initiatives.

For more information

To learn more about InfoSphere Big Match for Hadoop, please contact your IBM representative or IBM Business Partner.

Additionally, IBM Global Financing can help you acquire the software capabilities that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize a financing solution to suit your business and development goals, enable effective cash management, and improve your total cost of ownership. Fund your critical IT investment and propel your business forward with IBM Global Financing. For more information, visit: **ibm.com**/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
September 2014

[1] IBM Institute of Business Value in collaboration with Saïd Business School at the University of Oxford. "Analytics: The real-world use of big data." October 2012.

[2] Commissioned study conducted by Forrester Consulting on behalf of IBM. July 2013.

[3] 2014 Unisphere Research Study. *Data Governance Moves* Big *Data From Hype to Confidence*. https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov25649&S_CMP=iigar10

IMW14789-USEN-00