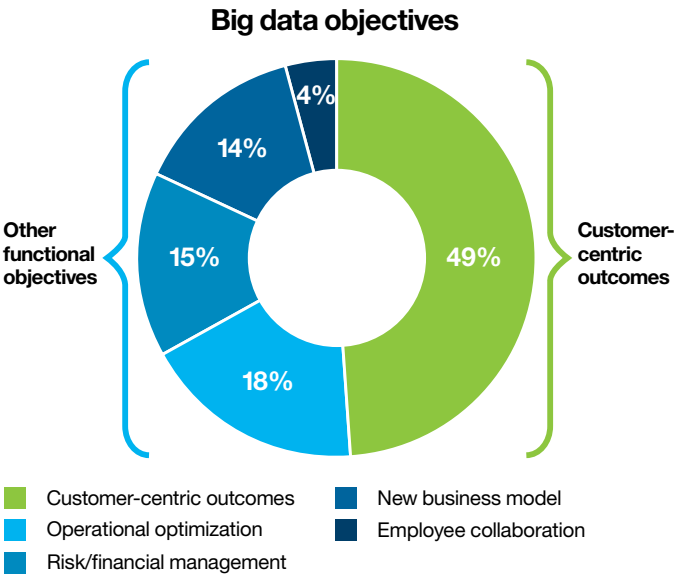
Addressing customer analytics with effective data matching

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

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, online videos or Instagram feeds
- Semi-structured information derived from call-center notes

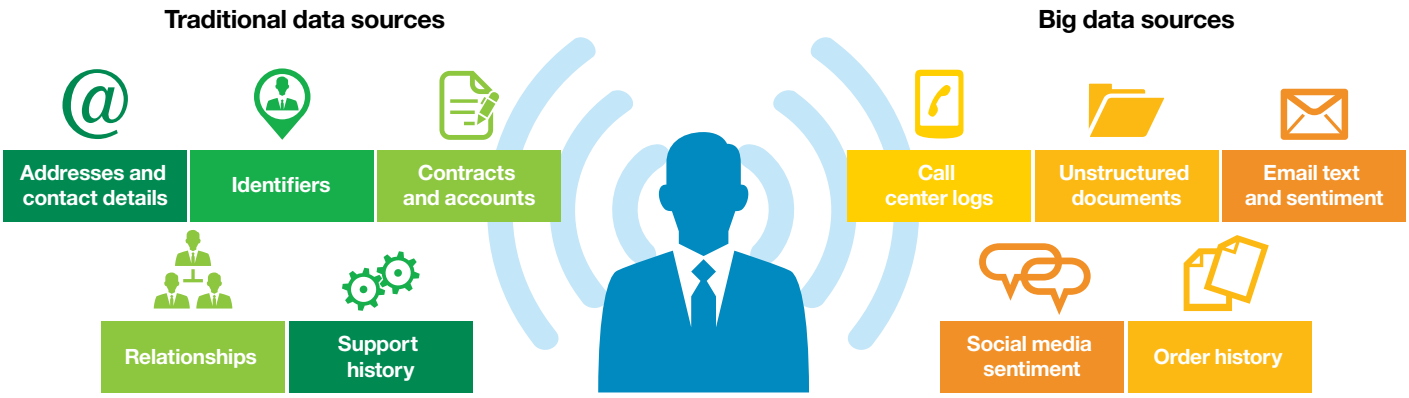


Figure 2. Building a complete view of a customer requires tapping a wide variety of data ranging from addresses and contact details to social media sentiment.

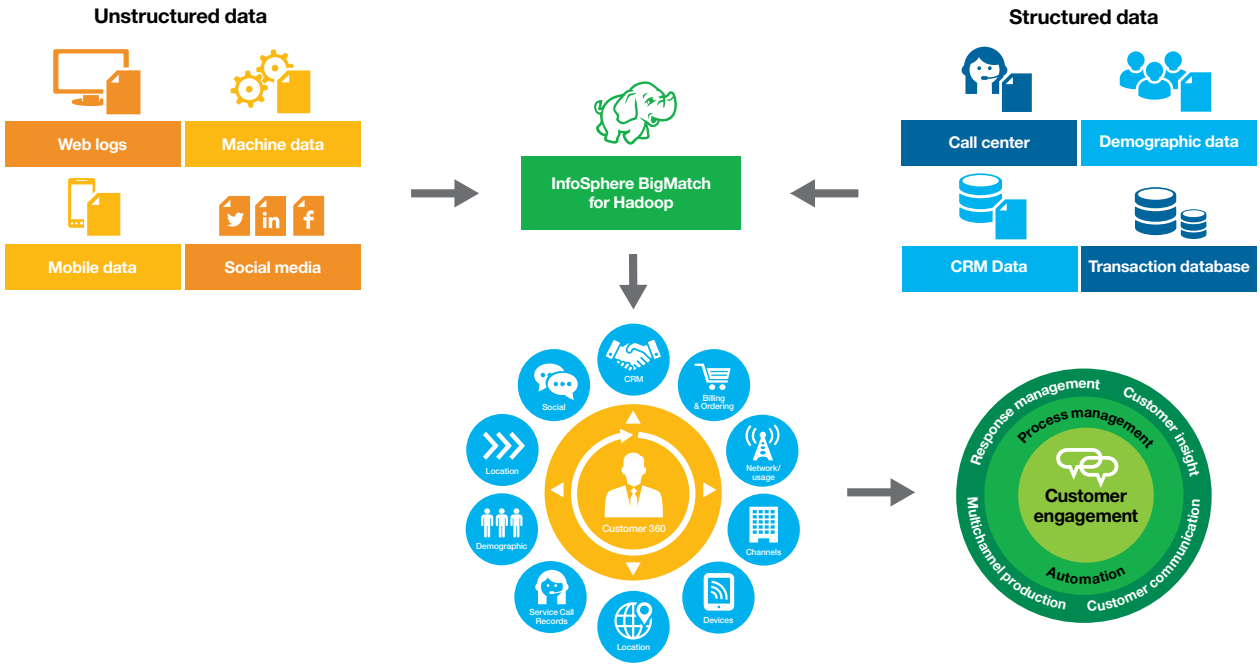


Figure 3. The combination of structured and unstructured data from internal and external sources creates a more complete view of a customer.

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 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

Big Data refers to the large quantities of human or machine-generated data that are generated at high speeds from multiple sources and in multiple forms. The key factor driving Big Data at this point is customer centric initiatives. Business value couldn't be achieved if the data is not managed strategically. To deliver business value, one can't afford to just dump data onto Hadoop, so a data-first approach is the best option. Effective data management and matching capabilities can provide more accurate actionable insights in this context.

and any structure, including totally unstructured. Applying a proper matching and deduplication mechanism on these entire data sets for correctly resolving customer data is always the Core of Confidence for the organization. Establishing and understanding the lineage of the customer data is the most critical factor for making any decision on customer engagements or any sort of campaign management program. Organizations can gain more accurate insights about their customer preference information, transaction trends, brand sentiments, order history and other such information. This can help the organization make more optimized and targeted decisions for customers.

When matching the customer system and Big Data sources are not accurate or if the capability of finding a potential match is limited, there can be implications to the organization of all sizes. For example, the wrong information could be sent to a customer, which was not intended to be sent, leading to potential legal complications. If this happens within the health care domain, a faulty link between two patients could misguide a physician about the proper health information of a patient, resulting in wrong diagnostics or medication prescribed to them. If there is any limitation in finding a potential match, the organization might miss sending the proper communication to intended customers on time, causing a loss for the organization and customer.

It is critical to have a robust and scalable matching algorithm to make this strategy a successful one. This process is essential to establish a solid, trustworthy foundation of information for customer analytics.

In this context, Big Match, built on Probabilistic Matching Engine (PME), and Hadoop Technology together enable the Core of Confidence at Big Data scale. This describes data with volumes in the hundreds of millions. Traditional matching engines can be overwhelmed by this much data, therefore, a Big Data matching solution is much needed. This probabilistic matching engine uses a pre-configured algorithm and distributed processing to analyze data. IBM InfoSphere Big Match for Hadoop uses statistical learning algorithms to provide a scalable solution to search, match and link customer data (see Figure 4). These algorithms are based on IBM Initiate technology, compiling more than 10 years of experience in identifying, relating and matching customers in different industries.

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.

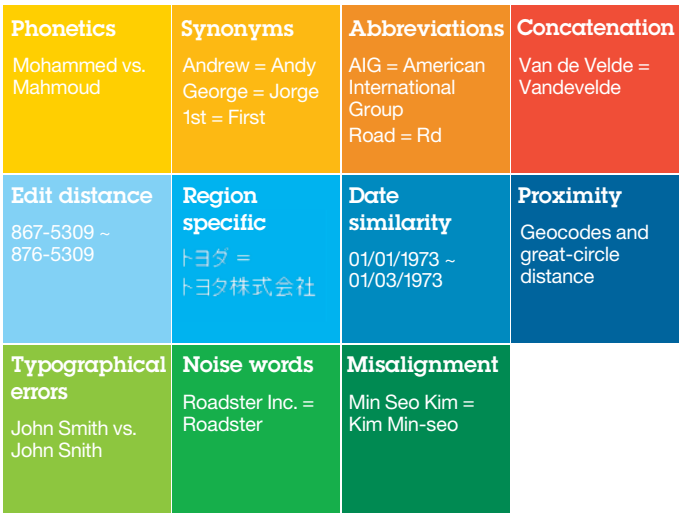


Figure 4. Examples of probabilistic matching

Top business and technical challenges to be addressed

Using the InfoSphere Big Match for Hadoop solution helps solve the critical business challenges currently faced by industries.

A complete 360-degree view of customers using information from all possible sources. IBM Big Match is going to get feed from unstructured documents, call center data, social and public media and other data sources. Using all this information in a controlled manner, the IBM Big Match learning algorithm can create a 360-degree customer profile for any customer engagement and analytics.

More accurate prospect identification. Using the structured and unstructured information from different sources, an organization can identify their potential prospects along with their complete 360-degree details. This prospect information will build confidence in the organization for a better prospect marketing strategy and prospect data analysis.

Trustworthy social analytics. Once the 360-degree view of the customer and prospects are built, social analytics of this trusted data can be validated by IBM Big Match to verify that the results are aligned with already known facts about those customers and prospects.

Apart from the above business challenges, there are many technical challenges that can be resolved with IBM Big Match.

non-compliant with normal enterprise quality control processes. The matching solution needs to be capable of handling this variation of data from different sources. 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.

Huge volume of incoming data. Analytical decisions shouldn't be impacted as the data volume increases exponentially. Matching algorithm and solution should be scalable enough to handle the volume and deliver the same expected results with similar efficiency. 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.

Applying standardization to Big Data volumes and velocity. Similar data coming in different formats might not get standardized before matching process. InfoSphere Big Match for Hadoop uses a data derivation approach to handle non-standardized data. It 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 with InfoSphere Big Match for Hadoop

So far, we have seen how InfoSphere Big Match for Hadoop can address customer analytics challenges, both from business and technical perspectives. This sets up a platform for better customer insights from internal and external data before it goes

to other systems for any specific detailed processing. It supports both real time and batch interface for data feed in any format from different sources. This capability helps to accelerate end-to-end activity.

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

- **Standardization and normalization of input data:** As it accepts the data in its native format, the first step is to normalize and standardize the data as per the defined rules. The derived data is stored in a separate layer, leaving the source data untouched.
- **Search for potential matches:** Search is optimized by a defined critical data set and bucketing strategy within the algorithm. The critical data set and bucketing strategy depends on the source data pattern and quality. This mechanism works faster because the search happens only within the matching buckets and not across the entire data set. Using this optimized search mechanism, a PME algorithm locates the potential match within the derived pool.
- **Score using probabilistic statistics:** Using the search mechanism above, this algorithm associates similar records and assigns a score to indicate the level of similarity. Effectively, linking customer data from different sources takes place through the robust matching and scoring algorithm.
- **Matching category and score:** The results are based on pre-configured weight thresholds that could 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.

InfoSphere Big Match for Hadoop: The highlights

InfoSphere Big Match for Hadoop leverages the probabilistic matching engine from IBM InfoSphere Master Data Management Server, as well as additional functionality like text analytics and big data accelerators.

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 captures a customer's every touch point with the organization. This gives 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 or visit: **ibm.com/**us-en/marketplace/infosphere-big-match-for-hadoop

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
October 2017

[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

Please Recycle

IMW14789-USEN-01