Context-Based Analytics in a Big Data World: Better Decisions

An IBM® Redbooks® Point-of-View publication

By **Lisa Sokol, Ph.D**, IBM Software Group, US Federal CTO, and **Steve Chan, Ph.D**, Prince of Wales Senior Fellow at MIT and Senior Fellow at Harvard Kennedy School's Center for Public Leadership, Massachusetts Institute of Technology and Harvard University, Cambridge, MA

Highlights

The growing complexity of the big data observational space requires organizations to use innovative approaches in order to find non-obvious relationships. Context-based analytics provides the following key advantages:

- Creating data within the appropriate context delivers higher quality models.
- Higher quality models that are applied to contextually-correct data can lead to better mission decisions and better outcomes.
- Using real-time contextual analytics enables timely entity assessments, while the observations are still occurring.
- Using context analytics with big data allows organization to achieve greater success regardless of whether the objective is mitigating risk or recognizing opportunity.

As the world becomes more instrumented, interconnected, and intelligent, the volume of information that is generated is growing at an exponential rate. The conversation surrounding this information explosion and about big data has centered on the size and management of this data. However, there is also an opportunity to improve critical business insight by taking advantage of the context that is created from big data.

Context, the cumulative history that is derived from data observations about entities (people, places, and things), is a critical component of analytic decision process. Without context, business conclusions might be flawed. By using context analytics with big data, organizations can derive trends, patterns, and relationships from unstructured data and related structured data. These insights can help an organization to make fact-based decisions to anticipate and shape business outcomes.

Without context, conclusions that are drawn from big data might be flawed

Organizations use analytics to learn about entities and then use the derived information to make mission-critical decisions. One important context analytics use case lies in the realm of social networking. The reach of social networking (for example, Facebook) and microblogging services (for example, Twitter) has extended to hundreds of millions of users. US-based Twitter generates more than 200 million text-based messages, or "tweets," per day and, by itself, has attracted more than 170 million active users. This venue and other similar services constitute an observational space by which trends and patterns in attitudes can be ascertained. As these venues become increasingly more valuable, the ability to perform pattern detection to make sense of the new data and to create insight in real time becomes critical.

There are several basic building blocks for insight. *Entities* are defined as people, places, things, locations, organizations, and events. Entities are an important focus of big data analytics. *Context* is defined as a better understanding of how entities relate. *Cumulative context* is the memory of how entities relate over time.

Unlocking big data's potential for reasoning about entities requires us to think differently about analytics. Using cumulative context and its derived information is critical to the success of analytics. Organizations must learn how to apply context to big data, or the conclusions and mission-critical decisions that are made from the analysis might be in error.

Combining a big data analytics environment with context analytics creates a novel analytic infrastructure that can enhance analytics decision-making within the big data space. This type of data varies in volume, velocity, variety (unstructured and structured), and veracity. This new environment can provide the following capabilities:

- Enable discovery and visualization analytics to take advantage of context and, thus, to increase the accuracy of models and patterns.
- Enhance the correctness of model and pattern assessment and discover data that is difficult to find.
- Realize the impact of new data on models, patterns, and situational assessment in real time.
- Make behavioral observations by discovering similar patterns of life.

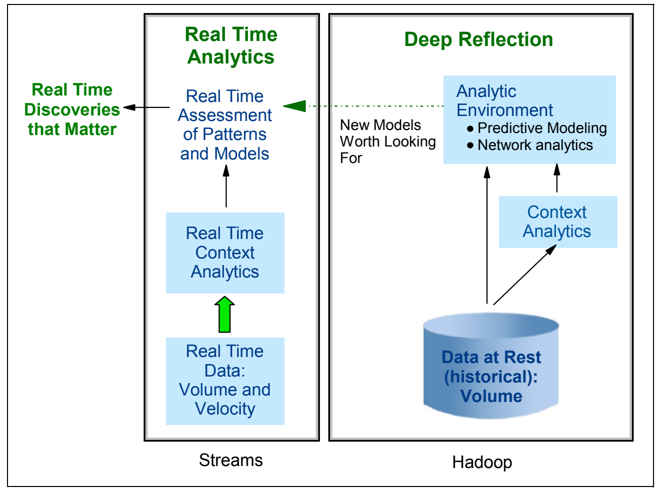

Figure 1 Key enabling infrastructure technologies

The big data analytics environment

Big data analytics requires both a layered technology deck as well as multiple infrastructures. Different analytics perform different functions, and the data itself varies in volume, variety, and velocity (streams where data flows over constantly running “queries”).

Figure 1 illustrates two key enabling infrastructure technologies, IBM Apache Hadoop MapReduce and Streams. Within an IBM Hadoop environment, deep analytics can be performed on very large amounts of historical data and data at rest. IBM InfoSphere® Streams technology enables continuous analysis of massive volumes of streaming data with submillisecond response times. When these infrastructures are combined with traditional enterprise data marts, analytics can take advantage of the full range of data.

Understanding context analytics

Under the umbrella of big analytics, *context analytics* denotes the incremental context accumulators that can detect like and related entities across large, sparse, and disparate collections of data. The data collection includes both current and historical data. The completeness of the data context enables analytics to correctly assess entities of interest.

With each new data observation (for example, actions, behaviors, locations, activities, or attributes), there is the possibility of a new discovery. The discovery might be some additional knowledge about the entity, such as a new phone number for a person. Another type of discovery is that of a new relationship between one of more entities. This relationship can be a direct connection between two entities, or it might be a non-obvious relationship, for example, a relationship that is several connections distant. The net result of this new observation is a deeper understanding of the entities within the observation set.

Fundamentally, the following basic assertions can be made about each new observation:

- It can be connected with an existing entity.
- It can be placed near a like entity.
- It can be considered a new, unassociated entity.

In addition, the new observation can alter prior assumptions or assertions.

Attributes help to uniquely define an entity. For example, attributes that are associated with an automobile might include the automobile maker, model name, model type, color, vehicle identification number (VIN), and owner's name. Some attributes are permanently associated with an entity, such as the automobile's VIN number, and other attributes might change over an entity's lifetime, such as the automobile's owner.

If two entities share sufficient attributes (for example, two people entities who share a mobile phone number, email address, and birth date), the possibility exists that these entities are the same person. *Entity disambiguation* is the process of using context analytics and taking advantage of the available attribute data that is associated with a two or more entities to determine whether the separate entities should be merged into a single entity. Merged entities can be effectively used within the decision evaluation process. For example, you can apply financial analytics to a resolved entity to assess the likelihood of an individual to pay back a loan.

Some data (attribute values, actions, or behaviors) are important in and of themselves. However, more often than not, the surrounding context can help you to understand the importance of the current observation. Every time that you use a credit card to make a purchase, context is used to determine whether the purchase is a "normal" purchase. Context includes both data about the current status of the entity and historical information, such as past actions, behaviors, locations visited or lived in, events attended, and so on. Abnormal purchases are sometimes refused by credit card vendors.

Many analytics approaches are impacted by the attribute data quality, data that has been subjected to errors of inadvertency (for example, a transposed month and day in a date of birth), errors in omission (accidental or intentional) or commission (disinformation), and natural variability in data (for example "Bob" versus "Robert"). These kinds of data errors can be helpful to context-based systems. For example, the natural variability in the data when the month and day in the date of birth are transposed is used within context accumulating systems. When the system tries to recognize like entities in the future, accuracy goes up because the system has learned from the natural variability of the past. In addition, recognizing that "Marek" is sometimes recorded as "Mark" is in fact helpful.

Context accumulation is the incremental process of relating new data to previous data. As part of this reconciliation process, the context analytics environment can discover different types of relationships that are relative to the entities. As an example, consider a traditional relationship where two different entities share an attribute, such as a home address. Context analytics must accumulate a person's environment, activities, connections, and preferences to create a more accurate understanding of the data. Accurate big data analytics requires context about these entities.

Deep reflection: Analytics using data at rest

Deep reflection analytics uses historical data about entities to discover historical patterns, models, predictors, relationships, and trends. Depending on the size of the historical data, this type of analytics can be performed either in a traditional data warehouse or a Hadoop-based environment. The correctness of the deep analytics depends on the following factors:

- Richness of the data set
- Relevancy of the data
- Diversity and quality of the data
- Ability of the analytics to use the information that is generated by the context analytics

Deep analytics provides data ingestion, transformation, and exploration functional aspects of the decision making process. The data transformation function extracts structured data from the unstructured data, for example, extracting sentiment and topics from mobile phone texts. The exploration portion of the analytics typically focuses on the discovery of relationships between outcomes of interest and data variables and the values of these variables. An excellent software platform for the model discovery is IBM SPSS® Modeler. Modeler Premium is a high-performance predictive and text analytics workbench that can help you gain insight from data. It provides a broad set of analytic capabilities, including visualizing and exploring data, manipulating data, cleaning and transforming data, and deploying results.

The primary output of deep reflection analytics is the patterns or models that are discovered within the modeling process. You can use the discovered information, for example, the patterns or scoring models, to assess real-time data and to guide real-time, mission-critical decisions.

Data transformation: Applying the results from context analytics

High-level analytics, such as discovery, scoring, and visualization, require a data set and data values that represent entity behaviors and activities. The discoveries that are made by contextual analytics about the different entities must be passed on to the data that is used by the high-level analytics models.

As an example, banking deposit data includes an individual's banking deposit history, such as the deposit size, deposit components (cash and checks), deposit date and time stamp, and deposit location. Typically in a database, each banking deposit is a separate row. A predictive scoring model requires all of the banking information that is associated with a single entity. Thus, the transformation must use the knowledge that is gained by context analytics that is relative to the entities within the deposit database that are actually the same entity. It uses this knowledge to resolve the different individual deposit transactions that are associated with a single entity into a single conceptual data deposit transaction unit.

Information discovery and visualization environment

Analytics are used to discover actionable insights about entities. Different types of insights require different types of analytics or combinations of analytics within architectural stacks. You can then use these insights to optimize decisions. For example, a customer service agent can use insight to decide whether to give a customer a discount for a new contract. A system can automatically assess whether a potentially suspicious activity should be escalated for investigation. A logistics manager can decide whether a truck is safe to put on the road for the next delivery. Patterns, models, and relationships all benefit from knowledge that can be extracted from the cumulative context. The decision about whether a truck is safe to send out for a delivery is more effective when using both the knowledge of that truck's past usage and repair history and the history of similar trucks and repair depots.

Contextually-derived information can provide value as demonstrated in the following types of analytics:

- Predictive analytics
- Visualization of relationships between entities

Predictive analytics using historical data

Predictive analytics can discover how historical data (variables and values) are related to outcomes of interest. The analytical discovery process uses techniques, such as data mining, text mining, social media analytics, as well as statistical analysis, regression analysis, cluster analysis, and correlation analysis, to learn from an organization's aggregate data set. When the enterprise learns from its historical experience, it can take action to apply what it has learned.

The outputs of a predictive modeling process are models and patterns that contain insight to outcomes of interest. These models and patterns can be deployed against new incoming (real-time) data in a real-time analytics environment. Consider the instantiation of a predictive model that was created to guide a bank loan approval process. Banks traditionally use scoring models that are based upon historical data to determine what about a person (the entity) makes that person a good or bad credit risk relative to loan payback. Suppose that the scoring model data is the result of poorly merged data from several banks. In this case, the data about an individual's wealth history appears in more than one place within the data set. The model score is inaccurate if the assessment uses only a portion of an individual's wealth history.

Context analytics can determine that different data observations are associated with a single individual. That contextual knowledge can be exploited here by correctly merging all of the wealth data for an individual. The loan score for that individual is now correct. When contextual knowledge is used to transform historical data into a more accurate data model, the outcome of the predictive modeling process creates higher quality, more accurate predictive models.

Visualization of historical data

Visualization is used to perform ad hoc analysis on the different aspects of entities. One commonly used visualization technique is network visualization, where a network is defined as nodes (entities) and edges (relationships between entities). Context analytics can determine which nodes in the graph should be merged and can then pass this information to the network analytics software.

For example, context analytics can determine that the three nodes that are shown in Figure 2, Margaret Adams, Peggy Jane Adams, and Maggie Adams are variants of the same identity.

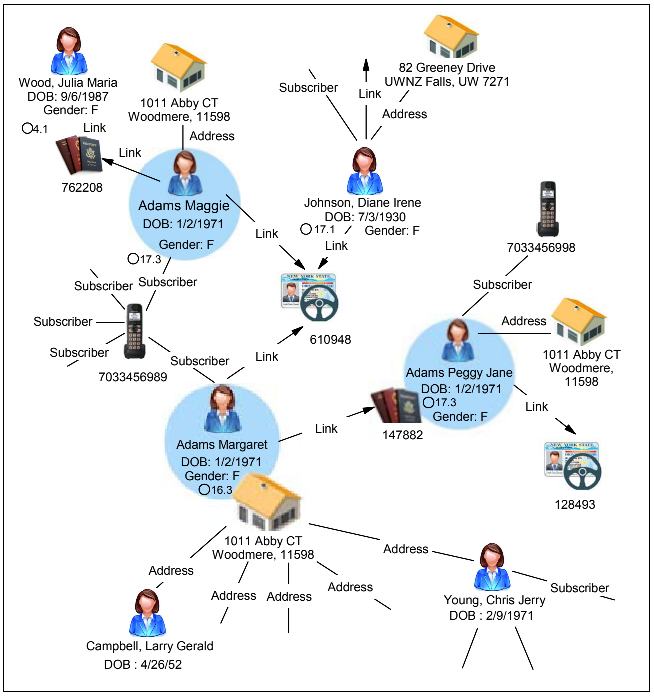

Figure 2 Data displays in more than one data set

Visualization correctness dictates that these three nodes should be merged into the single node, as shown in Figure 3.

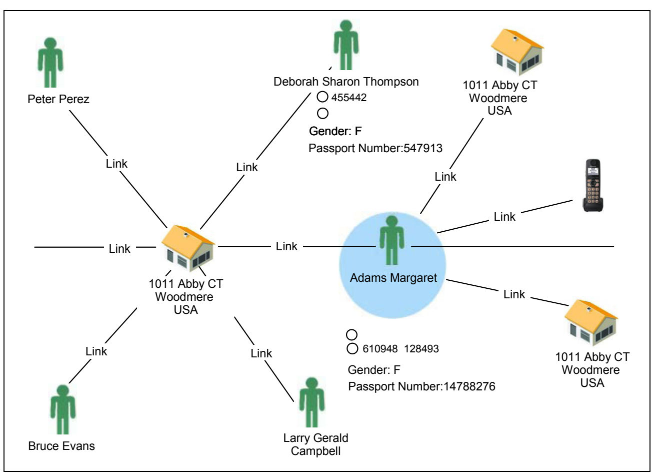

Figure 3 Merging the data into a single node

The application of context analytics to visualization techniques, such as network visualization, is an improvement in node and edge accuracy. A further advantage of using context analytics is that the number of hops between two nodes is the same or reduced and, thus, graph complexity is either the same or reduced. Reducing the graph complexity through the coalescing of aliases also makes it more likely for an analyst to discover relationships that are being intentionally hidden. Analytics, including social network analytics that are built on top of the context-driven network, will also be more accurate. Finally, the usage of context analytics to create a better and more accurate presentation of network relationships means better mission-critical outcomes.

Real-time analytics

The real-time portion of the analytics environment, which is shown in Figure 4, must make sense of observations as they present themselves. The context analytics portion of the environment must constantly reassess the cumulative context and update as appropriate with every new observation.

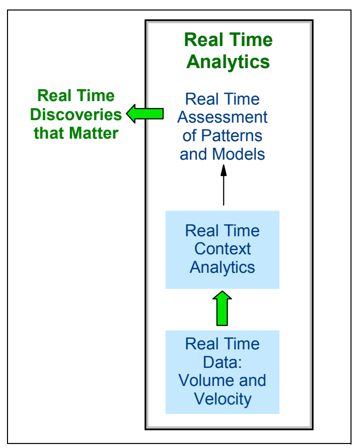

Figure 4 Making sense of observations

This process must assess each new observation (action, behavior, address, phone number, and so forth) to determine the following items:

- Whether the observation data can be added to an existing entity
- Whether it can be placed near a like entity
- Whether it can be considered a new, unassociated entity

This cumulative, cohesive picture of entities enables the analytics to use a combination of internal relevance detection models and situational assessment algorithms to make sense of and to evaluate different aspects of a targeted situation space. Within the social analytics space, for example, “tweets,” can be assessed for real-time discoveries, such as anomalies, trend changes, weak signals, new topics of discussion, and subtle changes in any aspect of the social infrastructure. In addition, real-time analytics (for example, pattern and scoring) must make these assessments fast enough to do something with the information while the observations are happening. As the real-time analytics find discoveries that matter, alerts are sent to users.

The context analytics engine supports both traditional relationship and pattern of life detection. Traditional relationship detection is the discovery that two different entities share an identification attribute, for example, a phone number or a home address.

Many behavioral observations have an associated temporal dimension, or time stamp, as well as a geospatial aspect. A second type of supported relationship detection exploits time and space observations through the discovery of shared rare events. This latter function enables the discovery of individuals with similar patterns of life. Examples of patterns of life include movements that are associated with daily life, such as daily visits to a local coffee shop or weekly visits to libraries or banks. Contextual analytics can discover rare events within a time and space logic to determine when different entities are “hanging out” together. Pattern of life is a significant tool with which to detect weak signals and subtle changes in the fabric of patterns and behaviors. The combination of identifying behaviors within time and space with persistent context creates a new opportunity for relationship detection.

Each time an entity is modified or created, it can be further evaluated in real time against the patterns, rules, and alerts that were developed in the deep analytics environment. The real-time analytics environment can discover whether newly defined entities now match models, if there are interesting changes of parameter values, new evidence for hypothesis confirmation, or surprising and relevant events and insights. The analytics environment can also determine whether the addition of this new data element changes the existing scores or the likelihood of accuracy for analytics models, trends, sentiments, behaviors, scenarios, and situations. Those changes or discoveries deemed relevant and interesting can then be pushed to appropriate users.

What’s next: How IBM can help

Context analytics is a critical capability when it comes to harnessing the big data space for both sense and respond and deep reflection activities. The combination of deep analytics and contextual data enhances the accuracy of the discovery and visualization modeling. The generated predictive models are of a higher accuracy. The network visualizations are cleaner and more correct.

The combination of real-time analytics and contextual data makes it easier to locate related observations. When viewed together, these observations provide useful information. As the definition of a new entity changes (and is persisted), analytics can determine whether the new entity has become interesting (matching a set of relevant rules, changing in overall relevance score, complementing user historical document usage, and so on).

Using context analytics achieves better outcomes. Higher quality models that are applied to context-enhanced transactions can produce better decisions.

IBM provides an extensive portfolio of contextual analytics software products. Although these products are effective in their own right, they can also be integrated into extremely effective analytics stacks. One of these products, IBM SPSS Modeler, offers a high-performance data mining and text analytics workbench that enables organizations to proactively discover patterns of interest, trends, and anomalies. SPSS Premium V15 incorporates a subset of the full contextual analytics capabilities in its modeling palate.

Resources for more information

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

- IBM InfoSphere Sensemaking

  https://www-304.ibm.com/industries/publicsector/fileserve?contentid=235174

- Jeff Jonas, IBM Fellow and Chief Scientist of the IBM Entity Analytics Group, blogs on Sensemaking and Context Analytics

  http://jeffjonas.typepad.com/jeff_jonas/

- *Analytics in a Big Data Environment*, REDP-4877

  http://www.redbooks.ibm.com/abstracts/redp4877.html?Open

- IBM Big Data Analytics

  http://www-01.ibm.com/software/data/infosphere/bigdata-analytics.html

- *Harness the Power of Big Data: The IBM Big Data Platform* (An IBM eBook)

  https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov8257&S_TACT=109HF53W&S_CMP=is_bdebook3

- *Turning Big Data into Actionable Information with IBM InfoSphere Streams*, TIPS0948

  http://www.redbooks.ibm.com/abstracts/tips0948.html?Open

- IBM SPSS Modeler

  http://www.ibm.com/software/analytics/spss/products/modeler/?

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-4962-00, was created or updated on August 20, 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:

IBM®
InfoSphere®
Redbooks®
Redbooks (logo)
SPSS®

The following terms are trademarks of other companies:

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

