



Using Entity Analytics to Greatly Increase the Accuracy of Your Models Quickly and Easily



Redguides
for Business Leaders

Dr. Lisa Sokol
Jeff Jonas

- Learn how Entity Analytics provides value to your business
- See how IBM SPSS Modeler Premium supports Entity Analytics
- Gain insight into real-time Entity Analytics





Overview of Entity Analytics

Analysts routinely face steep challenges as they attempt to integrate diverse enterprise-wide data. This statement is especially true when this data contains natural variability (for example, Bob versus Robert), unintentional errors (such as a transposed month and day in a date of birth) and professionally fabricated lies (such as a fake identity). Incorrect or incomplete integration can negatively affect any analytic solution that is built by using the data.

By implementing Entity Analytics, analysts can overcome some of the toughest data preparation challenges with unprecedented ease. By using Entity Analytics, analysts can generate higher quality, more accurate analytic models that result in better business outcomes. This activity can be accomplished regardless of whether the goal is detecting and preempting risk or recognizing and responding to opportunity.

One critical data preparation activity involves recognizing when multiple references to the same entity are the same entity (within the same and across data sources). For example, it is essential to understand the difference between three transactions carried out by three different people versus one person who carried out all three transactions.

Given the determination when entities are the same (resolved), even deeper understanding is achieved by recognizing when these resolved entities are related to each other (such as sharing a home address). Going far beyond simplistic match or merge technologies of the past, Entity Analytics delivers something new: true *context accumulation*. Context accumulation is the incremental process of relating new data to previous data and remembering these relationships. You can understand something better by taking into account the information around it. This process results in improved data accuracy.

For example, a stand-alone puzzle piece can be difficult to evaluate for importance when you stare at the piece by itself (shown in Figure 1 on page 2). However, by first comparing the puzzle piece to the whole puzzle, to see how it relates to the previously seen puzzle pieces, you can better understand the bigger picture and make a better prediction.

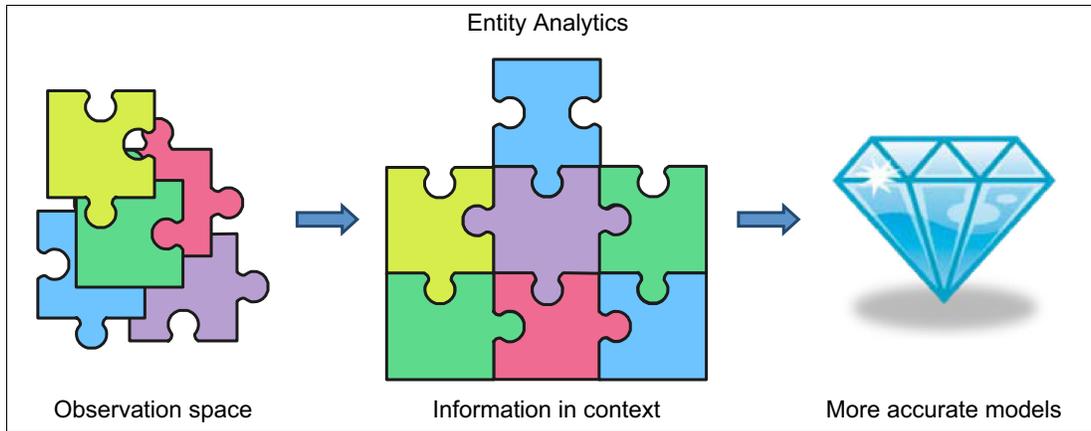


Figure 1 Entity Analytics accumulating context over diverse data

This IBM® Redguide™ publication explains how Entity Analytics can help analysts through the models that they create to drive better business outcomes. This guide provides an overview of the IBM SPSS® Modeler Premium product that incorporates Entity Analytics capabilities. It also provides examples of how Entity Analytics makes determinations and helps to solve business problems. In addition, this guide takes you through a lending scenario for a fictitious bank to illustrate usage of Entity Analytics.

About IBM SPSS Modeler Premium

IBM SPSS Modeler Premium is a high-performance predictive and text analytics workbench that helps you gain unprecedented insight from your data. It provides a broad set of analytic capabilities, including the following capabilities:

- ▶ Visualization and exploration of data
- ▶ Data manipulation
- ▶ Cleaning and transformation of data
- ▶ Creation and evaluation of predictive models
- ▶ Deployment of results in the form of production (runtime) models or scores

Entity Analytics functionality in IBM SPSS Modeler Premium

SPSS Modeler Premium contains Entity Analytics capabilities that analysts can use to quickly associate identity, behavior, and action data with their respective entities in real time or batch, with extraordinary ease. These Entity Analytics capabilities in SPSS Modeler Premium represent a breakthrough technology, the first of its kind that is commercially available. Even better, these capabilities are easy to use so that you can immediately start taking advantage of them.

Historically, analysts spent up to 80% of their time preparing and cleaning data for analysis. By using Entity Analytics, users can now build much more accurate models, based on cleaner data in a shorter time frame. Users of Entity Analytics gain the following distinct advantages:

- ▶ More accurate picture

The more identifiers that accumulate for an entity, the more accurate the Entity Analytics technology becomes.

- ▶ Better models
 - Information in context (understanding how the data relates) delivers higher quality models.
- ▶ Better outcomes
 - Higher quality models applied to context-enhanced transactions produce better decisions (for example, risk score calculations).

For example, one common regulatory practice is to require banks to report all cash transactions over \$5,000. To comply, banks must be able to understand the difference between five seemingly unrelated \$1,000 cash deposit transactions versus one person who transacts a \$5,000 cash deposit. If a bank cannot accurately quantify the cumulative (historical) transactions for that individual, it is unable to determine whether the \$5,000 threshold was crossed.

Entity Analytics (shown in Figure 2) provides an easy means (by using context accumulation) to associate the transactions to the correct entity, despite the lack of a common key. (The accounts do not share a tax ID number.) As a result, when the transactions are in context, the scoring models operate on the \$5,000 number, not a seemingly unrelated number of \$1,000 transactions.

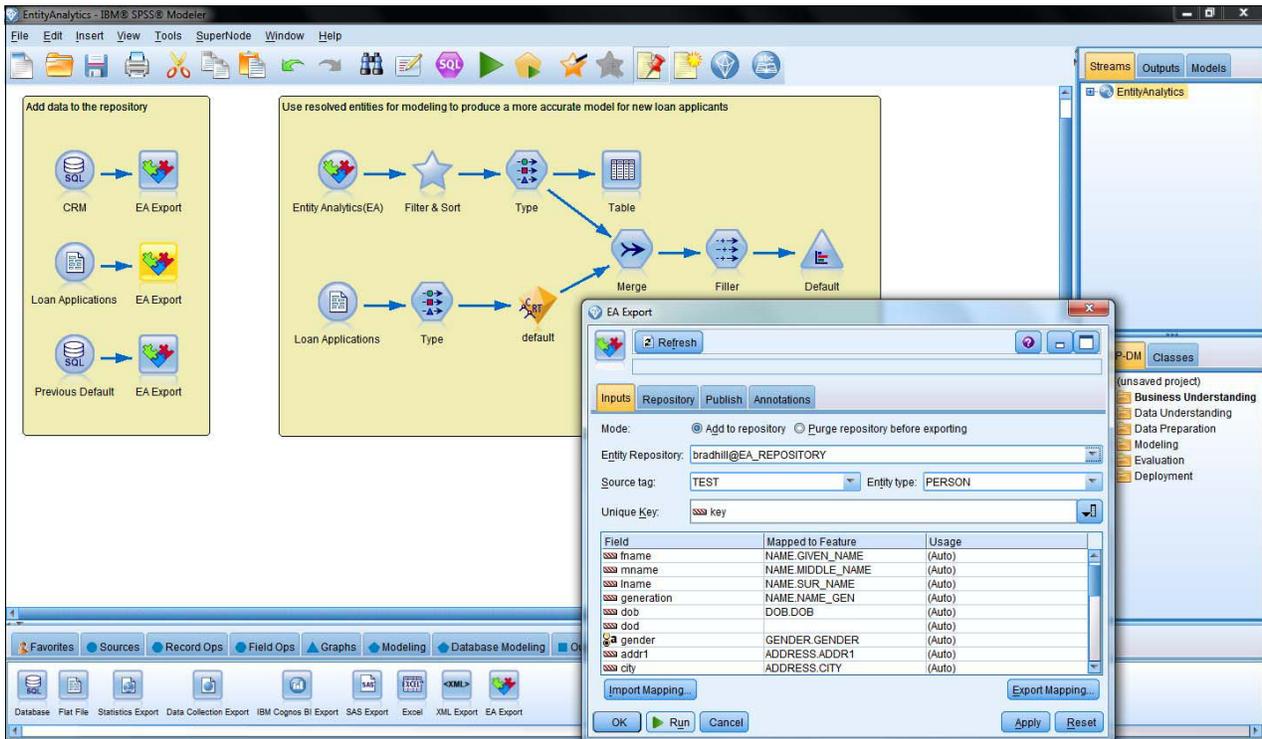


Figure 2 Example of an SPSS Modeler Premium stream that uses Entity Analytics

In SPSS Modeler Premium, Entity Analytics can be used in the following ways:

- ▶ The Entity Analytics export node performs the context accumulation. This node determines whether two entities (such as, individuals, corporations, or vehicles) are the same. This activity is accomplished despite the fact that they were recorded separately and, to a degree, recorded differently. If entities are determined to be the same, the identifiers (such as name, address, or phone) and the measurements (such as average balance or credit limit) of the entity are accumulated for that entity. This node automatically applies sophisticated fuzzy matching techniques. For example, it takes advantage of an internal library based on more than 800 million people names to deliver a world-class

culturally aware name comparison. As entities are resolved, understanding about each entity improves. The export node is frequently used to integrate historical data with new incremental data.

- ▶ By using the Entity Analytics source node (which reads resolved identities), an analyst can access the *in-context* information. This node is frequently used to analyze historical information (in context) and when creating data views to support new model building.
- ▶ The streaming Entity Analytics node is used to apply new records (in batch or real time) to the historical information. It instantly recognizes when entities are the same or related. This ability is roughly akin to being given a piece of data (a puzzle piece) and asking what other related pieces exist (the associated puzzle pieces). Given that this node discovers new data about an entity, this new knowledge can be passed on to a process that rationalizes the data about an entity.

For example, the newly discovered new data that you have in another bank account can be used to update a wealth variable. Down-stream pattern assessment processes can assess that entity again with patterns and models of interest, to determine whether the entity now meets an interesting threshold. The pattern or score assessment will be more accurate, thanks to the enhanced context.

Bank lending scenario that demonstrates using Entity Analytics

To see how Entity Analytics works, consider this hypothetical example that involves a typical bank process of giving loans to customers.

Predictive analytics can be used to help a bank determine which customers are likely to pay back their loans versus defaulting on their loans. To determine the likelihood of an individual paying back a loan, models are created by using available data from various sources such as the following examples:

- ▶ Historical customer data (such as income, debt, or previous defaults)
- ▶ Past lending outcomes (such as credit limit, average payment amount, or delinquencies)
- ▶ Other frequently used data points

Figure 3 shows an example of bank customer lending data. The first two rows are historic customer data. The third row contains data for a customer who is applying for a new loan. The last column in the table indicates whether the customer has a pending loan application.

Customer Number	Income	Credit Debt	Other Debt	Debt To Income	Prev Default	Pending
102	8000	5359	2009	92.1	Y	N
343	9000	6000	3000	100	Y	N
642	31000	1362	4001	17.3	N	Y

Figure 3 Bank lending data

By using historical data, SPSS Modeler Premium can generate a predictive model that can assess new loan requests for the likelihood of repayment. An example of a generated scoring rule might be: "If an individual has a debt to income ratio greater than 24.6 and previous

defaults, they are likely to default on a future loan.” In the example shown in Figure 3, entity #642 is applying for a loan. This person claims to have no previous defaults and to have a low debt-to-income ratio. By using the previously defined rule as an evaluator, the individual might be approved to receive a loan.

If you look closely at Figure 3 on page 4, you can imagine the difference between the three data points about three different customers versus these three data points about the same customer. Suppose that customer #642 is the same person as #102 and #343. Would you consider this customer (who has the pending credit application) to be a credit risk if you knew with some confidence that this person defaulted twice in the past?

If customers used their true names, addresses, and identifiers consistently and provided all details comprehensively and unambiguously, determining that this information is the same customer might be trivial. Unfortunately, because of unintentional data quality issues and periodic criminal intent, determining that this information represents the same customer is easier said than done. Fortunately, with Entity Analytics, users can quickly and easily perform context accumulation to detect exactly this situation and more.

Figure 4 shows that entities #102, #343, and #642 share sufficient identifiers to make a strong claim that these entities are the same customer.

Entity 102		Entity 343		Entity 642		Resolved Entity	
Name	Beth L. Doe-Smith BL Doe	Full	Liz Doe	Full	Elizabeth Lisa Doe	Name	Elizabeth Lisa Doe Liz Doe Beth L Doe-Smith BL Doe
Addr1	123 Main Street 777 Park Road	Addr1	33 Red Dr Mamaroneck NY 10354	Addr1	33 Reed Dr White Plains NY 10354	Addr1	123 Main Street 777 Park Road 33 Red Dr 33 Reed Dr
City	New York	City	Mamaroneck	City	White Plains	City	New York, White Plains, Mamaroneck
State	NY	State	NY	State	NY	State	NY
Phone	9587331234	Phone	958-733-1234 959-698-2234	Phone	959-698-2234	Phone	958-733-1234 959-698-2234
DOB	6/21/1954	Income	\$9,000	DOB	6/21/1954	DOB	6/21/1954
Income	\$8,000	Credit Debt	\$6,000	Credit Debt	\$1,362	Defaults	Yes
Credit Debt	\$5,359	Other Debt	\$3,000	Other Debt	\$4,001	Income	\$48,000
Other Debt	\$2,009	Debt to Income	100	Debt to Income	17.3	Credit Debt	\$12,722
Debt to Income	92.1	Prev Default?	True	Prev Default?	False	Other Debt	\$9,009
Prev Default?	True	Pending Loan	False	Pending Loan	True	Debt to Income	113.5
Pending Loan	False					Prev Default?	True
						Pending Loan	True

Figure 4 Common attributes across diverse records that are used to construct context

Usage of these collected facts in the *Resolved Entity* column highlights essential context to help properly score the pending loan application for entity #642. By using the entity data created by the Entity Analytics source node, an analyst can sum the true credit debt. The analyst can determine that the resolved entity has a credit debt of \$12,722 and has a debt-to-income ratio of 113.5. When the scoring algorithm is applied to the resolved entity, the score indicates that entity #642 should not receive the loan. This example demonstrates the true value of Entity Analytics, which is to provide more accurate decisions, faster.

Real-time Entity Analytics

By using Entity Analytics in SPSS Modeler Premium, companies can analyze transactions in real time to make optimal decisions in context. Based on all the “big picture” information, the models can predict outcomes more accurately for instant decision making, such as real-time fraud detection.

Imagine a fraud investigator who just stumbled on a new address that is related to an ongoing internal criminal investigation. With this information, just seconds later, Entity Analytics alerts this investigator that an employee in the investigator’s credit department has the same address. Through its context accumulation process, Entity Analytics related the new data (new address) to previous data (investigation, customers, and employees), delivering this extraordinary *insider threat* insight and so much more.

Summary

By using the Entity Analytics feature in IBM SPSS Modeler Premium, analysts can pull diverse enterprise data together into context. Organizations can then use this information in context to improve model quality, make better decisions, and ultimately achieve greater success, regardless of whether the objective is mitigating risk or recognizing opportunity.

An organization that can make sense of what it knows and do something about it faster than the competition is more competitive. With this exciting new technology, organizations of all sizes can gain this competitive edge today.

IBM Business Analytics software delivers actionable insights that decision-makers need to achieve better business performance. IBM offers a comprehensive, unified portfolio of business intelligence, predictive and advanced analytics, financial performance and strategy management, governance, risk and compliance, and analytic applications.

With IBM software, companies can identify trends, patterns, and anomalies; compare *what if* scenarios; predict potential threats and opportunities; identify and manage key business risks; and plan, budget, and forecast resources. With these deep analytic capabilities, our customers around the world can better understand, anticipate, and shape business outcomes.

Other resources for more information

For more information about SPSS Modeler Premium, see the product page at:

<http://www.ibm.com/software/analytics/spss/products/modeler/premium.html>

The team who wrote this guide

This guide was produced by a team of specialists from around the world who are working with the International Technical Support Organization (ITSO).

Dr. Lisa Sokol is an architect in the IBM Government Services CTO office. Her primary areas of interest are assisting government communities in dealing with the decision overload problem and using analytics to discover actionable information buried within large amounts of data. She has designed several systems that detect and assess threat risk that is relative to fraud, terrorism, counter intelligence, and criminal activity. She has a doctorate in operations research from the University of Massachusetts.

Jeff Jonas is an IBM Fellow and Chief Scientist of the IBM Entity Analytics Group. Before joining IBM, he led his company, Systems Research and Development, through the design and development of several unique systems. He designs next-generation technology that helps organizations better take advantage of their enterprise-wide information assets. He travels the globe and talks about innovation, national security, and privacy with government leaders, industry executives, leading global think tanks, privacy advocacy groups, and policy research organizations. He is a member of the Markle Foundation Task Force on National Security in the Information Age and a member of the US Geospatial Intelligence Foundation (USGIF) board, the EPIC Advisory Board, and the Privacy International Advisory Board. He is also a Senior Associate at the Center for Strategic and International Studies (CSIS) and a Distinguished Engineer of Information Systems (adjunct) at Singapore Management University.

Thanks to the following people for their contributions to this project:

LindaMay Patterson
ITSO, Rochester, MN

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new IBM Redbooks® publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:
<http://www.redbooks.ibm.com/rss.html>

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 grant 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.

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.

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.

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-4913-00, was created or updated on September 13, 2012.



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®
Redbooks®

Redguide™
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.