IBM Analytics
White Paper

IBM InfoSphere Identity Insight

Technical deep dive

IBM

Contents

Introduction

Many organizations are unable to gain a complete picture of the individuals with whom they do business. This lack of insight can result in missed sales opportunities, substandard customer service and operational inefficiency. But these problems are minor compared to the embarrassment and brand integrity consequences if an organization is caught doing business with a money launderer, a fraudster or known criminal.

Private corporations are not alone in the struggle to close gaps in understanding. The situation may be similar or even more serious for public agencies:

- Social service agencies must assess the extent to which citizens are receiving benefits and ensure they qualify for services
- Law enforcement agencies require access to a complete view of suspects, victims and other related parties
- Intelligence analysts must assess risks within enormous data streams

When agencies lack the ability to turn information into insight, they undermine their ability to prevent fraud, abuse and other threats to public safety. IBM® InfoSphere® Identity Insight provides a platform to assist in the fight against threat and fraud. Its identity and relationship disambiguation technology—based on context accumulation principles, combined with innovative, complex event processing—helps organizations recognize and mitigate the incidence of fraud, deception and collusion.

This white paper provides a comprehensive view of InfoSphere Identity Insight technology, including an overview of context accumulation and related principles, along with a specific data processing example. It also highlights application examples across industries to show how this technology is providing valuable insight into identities: who people really are, who they know and what they do.

What is context accumulation?

Information about people and organizations, by its very nature, is dynamic. People move, get new phone numbers, get married, take new jobs, open bank accounts and so on. To identify potential threats or fraud, organizations must recognize how current activity (new data) relates to historical activity (historical data). By doing so, they are able to understand the whole picture of who a person is. This whole picture is the context.

Context is simply the ability to better understand something by taking into account the things around it. Computers also do this with data in a process called context accumulation. Computers take in new observations in the form of data records from many different sources. Like an observation from a single sense, such as sight, the data record by itself is useless and there is a limited amount of information that can be gleaned from it. However, when that record is taken in context it is more meaningful (see Figure 1).

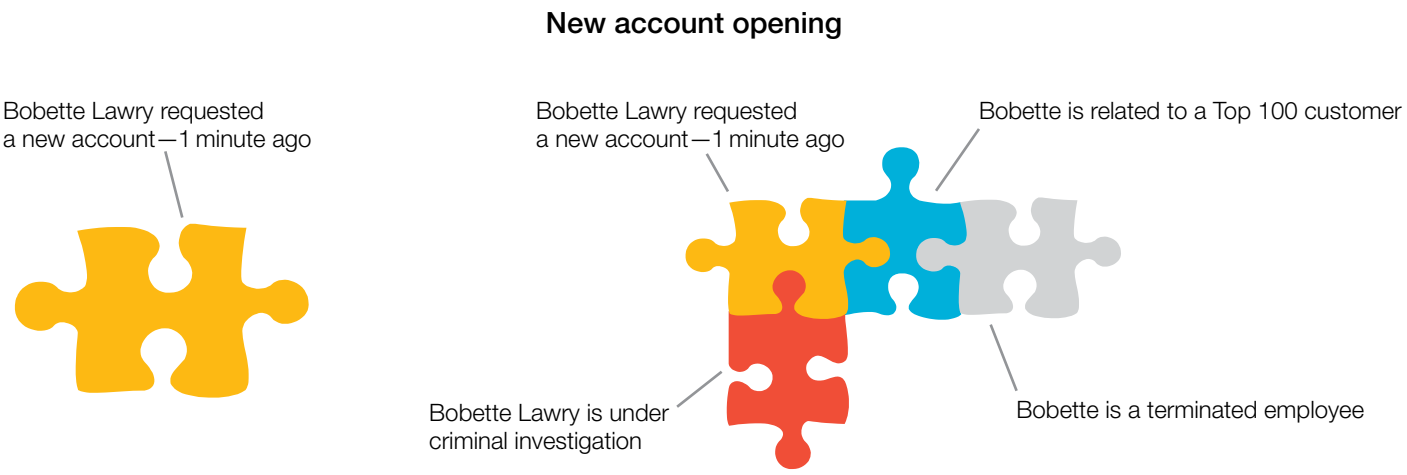

Figure 1. A single piece of information gives limited insight. In context, it is more actionable and more complex analytics can be performed on how multiple pieces of data are related.

Context accumulation as implemented in InfoSphere Identity Insight has several unique properties:

- **Resolving and relating real-world entities:** When comparing data sources that are populated by observations from different processes it is rare to have a unique key, such as a customer identification number, to link them together. InfoSphere Identity Insight uses an entity model where the entities are described by attributes and attributes in common are used to link them together.
- **Entity-centric learning:** Entities are made of multiple data records. Each data record brings its own unique attributes to the entity, and all attributes combined from all the data records equals the entity. InfoSphere Identity Insight then uses entity-centric learning when comparing a new data record against context. An incoming record is compared to all the attributes of the entity even if those attributes were contributed from different records. This will be illustrated further in the next section.
- **Self-correcting:** Determining an identity involves making assertions based on context and using those assertions (insights) for downstream analysis. When new information determines a previous assertion is no longer valid, InfoSphere Identity Insight will automatically correct the assertion.
- **Fully attributed historical recall:** Each data record comes from a unique source. InfoSphere Identity Insight remembers what the original record looked like and where it came from. This is necessary not only for the self-correcting behavior, but also to completely understand the analytics of InfoSphere Identity Insight and make the results actionable.

Analytical processing on entities and events

InfoSphere Identity Insight performs three core functions:

- **Detecting same:** InfoSphere Identity Insight reviews an incoming record and, using context accumulation, determines if the record describes an entity that is the same as one it already knows about. If the record should be part of an existing entity, it is added to the entity and its attributes are appended to the entity. If InfoSphere Identity Insight determines the record is unique—meaning there isn't enough in common with other records to say it is part of an existing entity—InfoSphere Identity Insight will establish it as a new entity. As this process continues, the software also evaluates the new records to determine if the record reverses an earlier assertion.
- **Detecting related:** With entities known, InfoSphere Identity Insight builds the network of interrelated entities. If the entities are people, it builds the social network out to multiple degrees of separation.
- **Detecting patterns:** This is the heart of the analytics performed in InfoSphere Identity Insight. Detecting the same and related is not unlike preparing the entities and entity map for analytics. InfoSphere Identity Insight is simply "connecting the dots." Detecting patterns tells you which dots have connected. This function should not be confused with pattern discovery: InfoSphere Identity Insight will not find a pattern that it has not previously been told to look for.

These core functions are performed as a series of data processing steps as data records are brought into InfoSphere Identity Insight. A customizable data model allows any attributes in the schema to be used for matching and analysis. It has built-in expert algorithms to compare data elements that are pre-configured into a default schema. This schema may work just fine out of the box, or it may be tailored to the business problem. Tailoring can include defining a new entity model with new attributes, as well as the rules that are applied at each step in the process (see Figure 2).

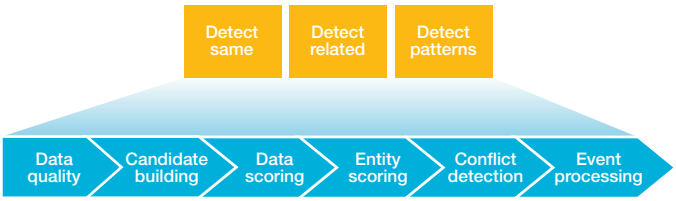

Figure 2. InfoSphere Identity Insight core functions can be expressed as a set of data processing steps, beginning with data quality, that all incoming records follow.

Data quality

This first step cleanses and enhances incoming data values to represent them in the best format for matching (see Figure 3). The data quality step leverages InfoSphere Global Name Management functions, such as:

- Name parsing
- Culture and gender identification
- Name variant generation
- Address parsing and formatting
- Attribute standardization
- Reference lookups

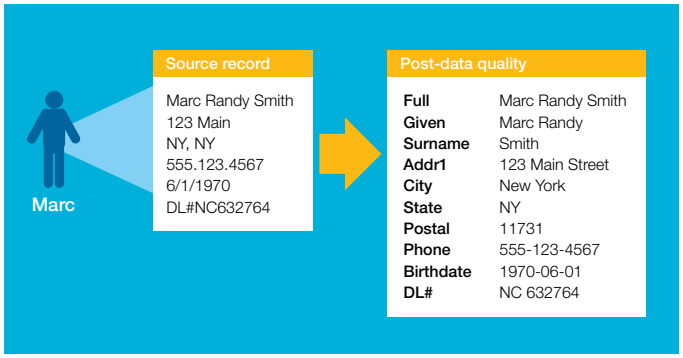

Figure 3. In the data quality step, a source record for Marc is formatted and improved to optimize later data processing.

InfoSphere Identity Insight has an extensive library of data quality management rules that organizations can apply. The data quality step can also include configuring calls to external data quality tools (such as IBM InfoSphere QualityStage®) to further augment the incoming data.

Applying these rules helps create standardized and enriched data values across all data sources to provide a consistent representation of information. It is also used to build intelligent keys for use in the next step, candidate list building.

Candidate list building

Candidate list building casts a net across the known entity registry to find entities that are most likely to resolve or relate to the incoming record (see Figure 4). Any attribute can be used for building candidates, including addresses, unique numbers, non-unique numbers, attributes, emails, names and even biometrics. Close matches are provided by specific rules for building the candidate list as well as by specialized keys that may be built during the data quality step.

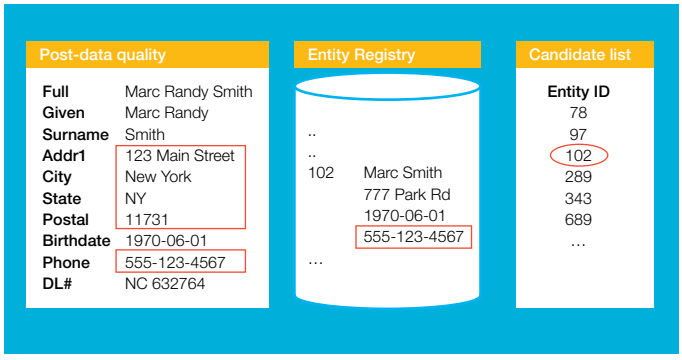

Figure 4. The candidate list building step determines which entities similar to Marc should be processed in more detail in later stages.

The candidate list building step also controls probabilistic value maintenance related to "generic values." InfoSphere Identity Insight may determine that an attribute has become so common it is no longer useful for candidate list building and scoring. Continuing to use it begins to force an "over-resolve" situation where matching occurs too frequently.

For example, InfoSphere Identity Insight may ingest records for the phone number value 555-1234. As more records are processed, it determines that the value 555-1234 has become so common it is no longer useful in the matching process. This value will become "generic," meaning InfoSphere Identity Insight determines it can no longer use this attribute. The software logs the situation and alerts an administrator to rectify the problem.

Over-resolve situations can be avoided by up-front analysis of the source data. A quick study of the source data can determine that the phone number field is populated incorrectly with the value 555-1234 and the problem can be rectified at the source, prior to reaching InfoSphere Identity Insight. Since it is not always possible to perform this analysis up front, InfoSphere Identity Insight includes the generic processing functions as a backup check.

Candidate list size relates directly to analysis performance and comprehensiveness. You should build a candidate list that encompasses the proper entities and isn't too small. If the list is too small, an entity can be missed, triggering a false negative response. If a candidate list is too large, it requires an undue amount of data processing to compare each entity in the candidate list to the incoming record.

Data scoring

The data scoring step tests the candidate list to evaluate which data fields match those on the incoming record (see Figure 5). It uses data type–specific algorithms to compare the fields of the input record against each entity on the candidate list. All entities are tested, and only records that pass the confirmation tests move to the entity scoring process. The full set of attribute histories for existing entities are used in comparisons.

This step uses IBM Global Name Management technology for multicultural name analysis. Like the data quality step, the data scoring step provides a mechanism to integrate call-out scoring algorithms for custom requirements. A specific example of call-out scoring is biometrics. A biometric matching engine is better suited to determine a match score between the biometric attribute in the incoming record and the biometric attributes in the entity. The data scoring step also has specific techniques to handle fuzzy matches, which are close but not exact matches.

	Entity 102	
Incoming record value	Existing entity value	Data score
Marc Randy Smith	Marc Smith	*Strong name match*
123 Main Street	777 Park Rd	*No match*
1970-06-01	1970-06-01	*Exact match*
555-123-4567	555-123-4567	*Exact match*
NC 632764	—	—

Figure 5. Data scoring compares the attributes of the Marc record to attributes of an entity on the candidate list.

Entity rule	Identity resolve
Name + Unique number	
Name + Address	
Name + DOB + Other number	Entity 102
Close Name + Other number	

Name	Marc Randy Smith
	Marc Smith
Addr1	123 Main Street
	777 Park Road
City	New York
State	NY
Postal	11731
Phone	555-123-4567
Birthdate	1970-06-01
DL#	NC 632764
	NC 632674

Figure 6. The entity scoring rules determine the Marc record should be a part of the entity.

Entity scoring

The entity scoring step uses the data scores to evaluate resolution and relationship rules and determine whether the incoming record is an existing or new entity (see Figure 6). It also determines relationships between the entities and evaluates the incoming record against all records and all attributes in the entity.

Pre-configured rules are included for the most common data elements required in a solution, but users may customize rules for the unique business requirements of the solution.

A key part of this process is not just the resolution—or matching of records to entities—but also the ability to un-resolve and re-resolve. Un-resolve and re-resolve make up the self-correcting behavior of InfoSphere Identity Insight. An incoming record may demonstrate that a previous assertion of a same or related match is not correct. InfoSphere Identity Insight will automatically un-resolve the record or records from the entity and re-resolve them as appropriate.

Conflict detection

Conflict detection is the one of the analytics processes performed on the entity map. It analyzes the roles on entities that are resolved or related to validate whether there are conflicts. Users define the role types and the nature of conflicting relationships where roles may be assigned to any data source or specified for an individual input record. These alerts can also be configured on relationships and out to multiple degrees of separation. For example, for a watch list check, InfoSphere Identity Insight could be configured to alert when a person is related to someone on a watch list out to three degrees of separation.

InfoSphere Identity Insight can find different patterns, such as analyzing the role of an identity to determine whether the individual possesses roles that are in conflict (for example, a customer and an individual on a watch list). You can extend role conflicts across the relationships in the social network, such as finding out if a customer knows a person on a watch list. When conflicts are discovered, InfoSphere Identity Insight raises alert notifications to investigators with complete detail descriptions.

Occasionally, a search of InfoSphere Identity Insight may not yield results. This can trigger a persistent query, which is a query that itself becomes a piece of data. Following the resolution process described previously, if a persistent query resolves or relates to an entity, the query is answered. This could happen immediately or later when new incoming records resolve to the query. Because the persistent query is treated as data, a persistent query can resolve to another query.

Consider two analysts looking for the same or similar entities. When the queries resolve to each other, the analysts "discover" that someone else is looking for the same thing and they can choose to collaborate.

Complex event processing

Complex event processing introduces the notion of activity into the entity network. For example, if the entities in the network are people, they are doing things together. Complex event processing helps determine what those people are doing that may be worth noting. This step adds value by operating off the resolved and related entity map.

The Complex Event Processing engine in InfoSphere Identity Insight analyzes a series of transactions or events to test for conformance and find items of importance that match predefined patterns. Users can model and correlate multiple user-defined event types, as well as establish rule thresholds on a variety of conditions across one or more flexible life spans. Like conflict detection, this step can raise notifications to investigators with details on the triggering events.

Real-world applications of InfoSphere Identity Insight

Both public and private enterprises have used InfoSphere Identity Insight to revolutionize their ability to address threat and fraud. The software is helping organizations solve a variety of business challenges across industry sectors.

Financial services

Financial institutions are continually challenged by individuals who seek to defraud their business. These threats emerge from a variety of internal and external channels, and come in a variety of forms—common point of purchase, bust-out fraud, identity theft, impossible geography, card not present and many others. Governments have placed the burden on banks to identify and disrupt these schemes in order to protect their customers. Increasingly strict regulatory mandates require financial enterprises to diligently scrutinize the identity of the individuals with whom they do business and validate that the business is legitimate.

Traditionally, financial organizations applied technologies within a particular transaction channel to deal with a very specific subset of these challenges. This strategy has proven difficult to execute, costly to manage and overrun with false positives.

Instead, financial institutions should establish a rich analytical platform that they can leverage within and across each channel. This platform should discover attempts at misrepresentation of identity, find networks of suspicious individuals who are collaborating, and dissect complex event scenarios that may indicate nefarious activity. The platform must also allow the organization to customize the rules and thresholds to provide maximum flexibility to the business.

Social services

State and local governments are seeking to optimize their programs to better meet the needs of the population. Information pertaining to a citizen and his or her family is currently spread across many systems—welfare, adult and aging, child programs and so on. These information silos prevent social workers from understanding the extent to which citizens qualify for benefits and raise significant hurdles in finding that information.

At the same time, government agencies are coming under increased pressure to drive out costs related to waste, fraud and abuse in the system.

Social service organizations cannot depend solely on new policies to address these issues. Agencies must streamline their efforts by building a complete picture of a citizen by connecting information from across the organization into a single, integrated view. This integration also gives them the ability to harness information they already have in order to recognize unusual and suspicious activity. Agencies can then efficiently connect citizens with the social services they need, while also realizing significant cost savings by preempting the abuse schemes of scam artists.

Tax agencies

With increasing budgetary and staffing restrictions, closing the "tax gap" has never been more important. Such initiatives, however, are complicated by the agencies' need to respond to increasingly sophisticated tax evasion techniques.

These intensifying challenges mandate a fresh approach to managing tax compliance and fraud investigations. Yet, many agencies continue to use traditional random audit selection, tax collection and enforcement methods they know to be outdated. These methods rely upon data and data relationships that are invalid and insufficient to promote maximum compliance—and often put the burden of an onerous examination on honest taxpayers. Auditors need better tools to determine how noncompliance occurs, while also protecting taxpayers' privacy.

Tax agencies must better leverage information—much of which they have already collected—to find the most egregious offenders. The first step is to centralize information about taxpayers, tax service providers, tax processing organizations, employers and financial institutions. Once this data is brought together, agencies can then deploy built-for-purpose analytical tooling in order to:

- Provide tax compliance collectors with a solution that supports and empowers them to make better real-time decisions
- Proactively detect and mitigate fraud to manage risk and follow up on tax avoiders
- Segment taxpayers to detect who may be a higher risk, to better focus the agencies' limited resources
- Detect and determine if taxpayers are properly registered and eliminate duplicate registrations
- Automate manual steps to verify taxpayer names, identities and relationships

Law enforcement

Law enforcement agencies have found that the data they have already collected is an invaluable tool for solving new investigations. However, like many organizations, the number of data silos has led to significant challenges in using this information effectively. Warrants, arrests, street checks, gang data, intelligence reports and ballistics are among many data sources that may supply the key to solving an investigation.

Finding the relevant information, however, is fraught with problems. In most cases, these sources of data do not share common identifiers for any one individual, precluding the department from accessing a complete view of suspects.

Police agencies are often surprised to find that even their record management systems lack the capabilities to solve data issues. While these systems may "suggest" existing records that are a match, they do not enforce the choice and they do little to manage the data conditions that arise when duplicates are created.

Law enforcement agencies are in desperate need of context-accumulating analytics to integrate the disparate data sources that are critical to investigative efforts. Such analytics provide the ability to locate individuals, present a complete history of their interactions, and perform link analysis to uncover potential leads on active and inactive cases. Since each new piece of data to enter the system is analyzed in the context of all known data, investigators are less likely to miss opportunities to share information. For example, the Rochester, Minnesota police department is using a mobile app to run queries on InfoSphere Identity Insight, such as entering license plate numbers during a traffic stop. The app performs analytics to find information about who may be associated with the vehicle, and then filters through predefined rules on serious offenders, outstanding warrants and probation listings. The results are delivered back to the officers in seconds, giving them an opportunity to adjust their response to the situation.[1]

Intelligence

National security challenges have never been greater. Coping with the ever-growing volumes of data makes the nature of the problem even more complicated. Harnessing the information at hand to detect not only the obvious and nonobvious threats, but also the adversaries' efforts to conceal their identities and activities, is no trivial task. Detecting the presence and actions of these individuals is virtually impossible without context accumulation.

InfoSphere Identity Insight technology was originally developed by Systems Research and Development, which was later acquired by IBM. By 2001, the venture capital arm of the United States CIA (In-Q-Tel) recognized the applicability to various intelligence programs, and twice granted funding to the company to further advance the technology. As a result, InfoSphere Identity Insight plays a unique role in national security and intelligence missions.

For more information

To learn more about how IBM InfoSphere Identity Insight can help your organization combat threats and fraud, contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/us-en/marketplace/infosphere-identity-insight

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] Ungerleider, Neal. "This small city's police department builds an app, nabs big data to find and fight bad guys." FastCompany. March 26, 2014. www.fastcompany.com/3027641/this-small-citys-police-department-builds-an-app-nabs-big-data-to-find-and-fight-bad-guys

Please Recycle

IMW14796-USEN-00