Version 4 Release 2

IBM i2 Analyze
Correlation Guide

IBM

Note

Before you use this information and the product that it supports, read the information in "Notices" on page 9.

This edition applies to version 4, release 2, modification 0 of IBM® i2® Analyze and to all subsequent releases and modifications until otherwise indicated in new editions. Ensure that you are reading the appropriate document for the version of the product that you are using. To find a specific version of this document, access the Configuring section of the IBM Knowledge Center, and ensure that you select the correct version.

© Copyright International Business Machines Corporation 2018.
US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Information Store data correlation

This documentation provides an overview of correlation support that is available during data ingestion into the Information Store in IBM i2 Analyze. Later sections describe how to configure correlation with an example use case.

Intended audience

This documentation is intended for users who want to correlate data as it is being ingested into the Information Store.

Users must understand how to ingest data into the Information Store. For more information about ingesting data into the Information Store, see Information Store data ingestion.

Users must understand the i2 Analyze data model, and i2 Analyze record structure. For more information, see Data in i2 Analyze record.

Important: Before you use Information Store data correlation in your deployment of i2 Analyze, you must install the i2 Analyze 4.2.0.1 Fix Pack or Enterprise Insight Analysis 2.2.0.1 Fix Pack. For more information about downloading and installing the Fix Packs, see Release Material.

Overview of correlation

Correlation is the process of associating data based on strong identifiers. For the process of ingesting data into the Information Store, i2 Analyze can use correlation identifiers that you provide to determine how to process and represent data in i2 Analyze records.

Correlation in i2 Analyze

During the ingestion process for the Information Store, correlation can be used to determine when data being ingested should be associated with existing records, and is represented by a single i2 Analyze record. In i2 Analyze, this operation is known as *merge*.

During the correlation process, an identifier is used to determine how each row of data should be associated. You present the identifier to the Information Store with the other staging data during ingestion.

Correlation uses

You might want to use correlation when you are ingesting data that originates from disparate sources, which have common properties, or have the potential to represent the same real-world objects. For example, if you have two data sources that contain information about people.

Another scenario where you might use correlation, is when the data that you are ingesting is in the form of event driven models (crime or complaint reports) where the same actors (people, locations, phones, and vehicles) might be referred to frequently.

Correlation can be used in these scenarios to combine multiple source records into single i2 Analyze records for link analysis.

© Copyright IBM Corp. 2018

You can limit the use of correlation to data from a specific source, of certain item types, or per row of data ingested into the Information Store. You do not have to provide a correlation identifier for all the data that you ingest into the Information Store.

Correlation method

In i2 Analyze, correlation identifiers and implicit discriminators are used to determine how the Information Store processes data during ingestion.

When you ingest data into the Information Store, you can provide a correlation identifier type and key value that are used to construct the correlation identifier for each row of data in the staging table. The type and key values that you provide are used to process data that is determined to represent the same real world object. Implicit discriminators are formed from parts of the i2 Analyze data model in the Information Store. Even if correlation identifiers match, if values for elements of the i2 Analyze data model are not compatible, that data cannot be represented by the same i2 Analyze record. For more information about correlation identifiers and implicit discriminators, see "Correlation identifiers" on page 2.

During the ingestion process, i2 Analyze compares the correlation identifiers of the data to be ingested and existing data in the Information Store. The value of the correlation identifiers determine the operations that occur. For more information about the correlation operations that can occur, see "Correlation operations" on page 4.

Example data sets are provided that demonstrate the correlation behavior available in this release of i2 Analyze. For more information, see "Correlation example" on page 7.

Important: Before you use Information Store data correlation in your deployment of i2 Analyze, you must install the i2 Analyze 4.2.0.1 Fix Pack or Enterprise Insight Analysis 2.2.0.1 Fix Pack. For more information about downloading and installing the Fix Packs, see Release Material.

Correlation identifiers

The role of a correlation identifier is to indicate that data is for a specific real-world object. If multiple pieces of data are about the same specific real-world object, they should have the same correlation identifier. At ingestion time, the correlation identifier of incoming data informs the Information Store how to process that data. Depending on the current state of the i2 Analyze record that is associated with the data, a match with the correlation identifier on an inbound row of data determines the outcome of the association. The record that the data is associated with is determined by the origin identifier of the data.

You specify the values for the correlation identifier in the staging table that you are ingesting the data from. The correlation identifier is made up of two parts, the *correlation identifier type* and the *correlation identifier key*.

type

The *type* of a correlation identifier specifies the type of correlation key that you are using as part of the correlation identifier. If you are generating correlation keys using different methods, you might want to distinguish them by specifying the name of the method as the correlation identifier type. If your correlation keys are consistent regardless of how they were created, you might want to use a constant value for the correlation identifier type.

The value of the correlation identifier type might be seen by analysts.

Note: The length of the value for the *type* must not exceed 100 bytes. This is equivalent to 100 ASCII characters.

key

The *key* of a correlation identifier contains the information necessary to identify whether multiple pieces of data represent the same real-world object. If multiple pieces of data represent the same real-world object, they should have the same correlation identifier key.

Note: The length of the value for the *key* must not exceed 1000 bytes. This is equivalent to 1000 ASCII characters.

To prepare your data for correlation by i2 Analyze, you might choose to use a matching engine or context computing platform. Matching engine and context computing platforms can support the identification of matches that enable you to identify when data that is stored in multiple sources represents a single entity. You can provide these values to the Information Store at ingestion time. An example of such a tool is IBM InfoSphere Identity Insight. InfoSphere Identity Insight provides resolved entities with an entity identifier. If you are using InfoSphere Identity Insight, you can populate the correlation identifier type to record this, for example `identityInsight`. You might populate the correlation identifier key with the entity identifier, for example 1234. This generates a correlation identifier of `identityInsight.1234`. For more information about IBM InfoSphere Identity Insight, see Overview of IBM InfoSphere Identity Insight.

Alternatively, as part of the data processing to add data to the staging tables, you might populate the correlation identifier with values from property fields that distinguish entities. For example, to distinguish People entities you might combine the values for their date of birth and an identification number, and you might specify the type as `manual`. This generates a correlation identifier of `manual.1991-02-11123456`.

The complete correlation identifier is used for comparison, therefore only data with correlation identifiers of the same types are correlated.

For more information about specifying a correlation identifier during the ingestion process, see Information Store staging tables.

Implicit discriminators

In addition to the correlation identifier that is created from the type and key values that you provide, *implicit discriminators* are also used during the matching process. If the correlation identifier matches, the following implicit discriminators are also compared. The implicit discriminators must be compatible to enable correlation to occur.

Item type

The item type of the data that you are ingesting must be the same as the item type of the existing i2 Analyze record that is matched by the correlation identifier. If the item types are not the same, then no correlation operations occur.

Security dimension values

The security dimension values of the data that you are ingesting must be the same as the data that is matched by the correlation identifier. If the security dimension values are not the same, then no correlation operations occur.

Link direction and ends

For link data, the link direction and ends of the data that you are ingesting must be the same as the data that is matched by the correlation identifier. If the link direction and ends are not the same, then no correlation operations occur.

The direction and ends of a link are inspected, and direction is respected. For example, a link from A to B of direction 'WITH' matches with a link from B to A of direction 'AGAINST'. A link from A to B of direction 'WITH' does not match with a link from A to B of direction 'AGAINST'.

Note: To use correlation, you must understand the incoming data and confirm that it meets a number of conditions.

After you ingest data with correlation identifiers, the correlation identifiers and implicit discriminators that are ingested into the Information Store must not change. To ensure that your correlation identifiers and implicit discriminators do not change, the following must be true for the lifetime of that data in the Information Store:

- The method that you use to generate the correlation identifiers must not change. For example, if you used IBM InfoSphere Identity Insight to create the correlation identifiers, you must continue to do so.
- The property types and values that the method for generating correlation identifiers uses must not change. For example, if your correlation identifiers for people include data from their date of birth, then the value for their date of birth must not change and you must continue to use the date of birth property.
- The security dimension values for the data must not change. For example, if you ingest data with the security dimension values of `HI`, `OSI`, `CON`, you must always ingest that data with those security dimension values.
- If it is a link record, the link ends must not change. For example, if you ingest a link between entities A and B. Whenever you reingest that link, it must always have entities A and B as its link ends.
- If it is a link record, the link direction must not change. For example, if you ingest a link with direction of `W`, you must always ingest that data with direction `W`.

If you try to ingest data that causes the correlation identifier or implicit discriminators of a record to change, the data is rejected during the ingestion process. For more information, see Troubleshooting the ingestion process.

If you later discover that the data does not meet these conditions, you must remove all of the data with correlation identifiers from the Information Store and reingest the data without using correlation.

Correlation operations

When the Information Store receives a correlation identifier, the way the system responds depends on the values of the correlation identifiers, and the state of the records that are associated with them.

The following section explains the *merge* operation that the system can respond with when it receives a correlation identifier. This response is in addition to the insert and update operations that are part of the standard ingestion process.

Merge

When one or more pieces of data are determined to represent the same real-world object, the data is merged into a single i2 Analyze record. For the Information Store to merge data, the correlation identifiers must match and the implicit discriminators must be compatible.

A merge operation can occur in the following scenarios during ingestion:

- New data in the staging table contains the same correlation identifier as an existing record in the Information Store. The new data has an origin identifier that is not associated with the existing record.
- Multiple rows of data in the staging table contain the same correlation identifier.

After a merge operation, the following is true for the merged record:

- The record has a piece of provenance for all of the source information that contributed to the merged record.

- The property values for the merged i2 Analyze record are taken from the provenance that has the most recent value for SOURCE_LAST_UPDATED.

  Note: If the existing provenance for a record was not ingested with a value for SOURCE_LAST_UPDATED, the provenance that has a value is used. Otherwise, the property values to use are determined by the order of the origin id keys associated with the record. To ensure data consistency, update your existing records with a value for the SOURCE_LAST_UPDATED column before you start using correlation, and continue to update the value accordingly.

During ingestion, the number of merge operations that occur is reported in the MERGE_COUNT column of the ingestion report.

The following diagram demonstrates the merge operation.

In the example of a merge operation, data in the staging table is merged into an existing i2 Analyze entity record because the correlation identifiers match.

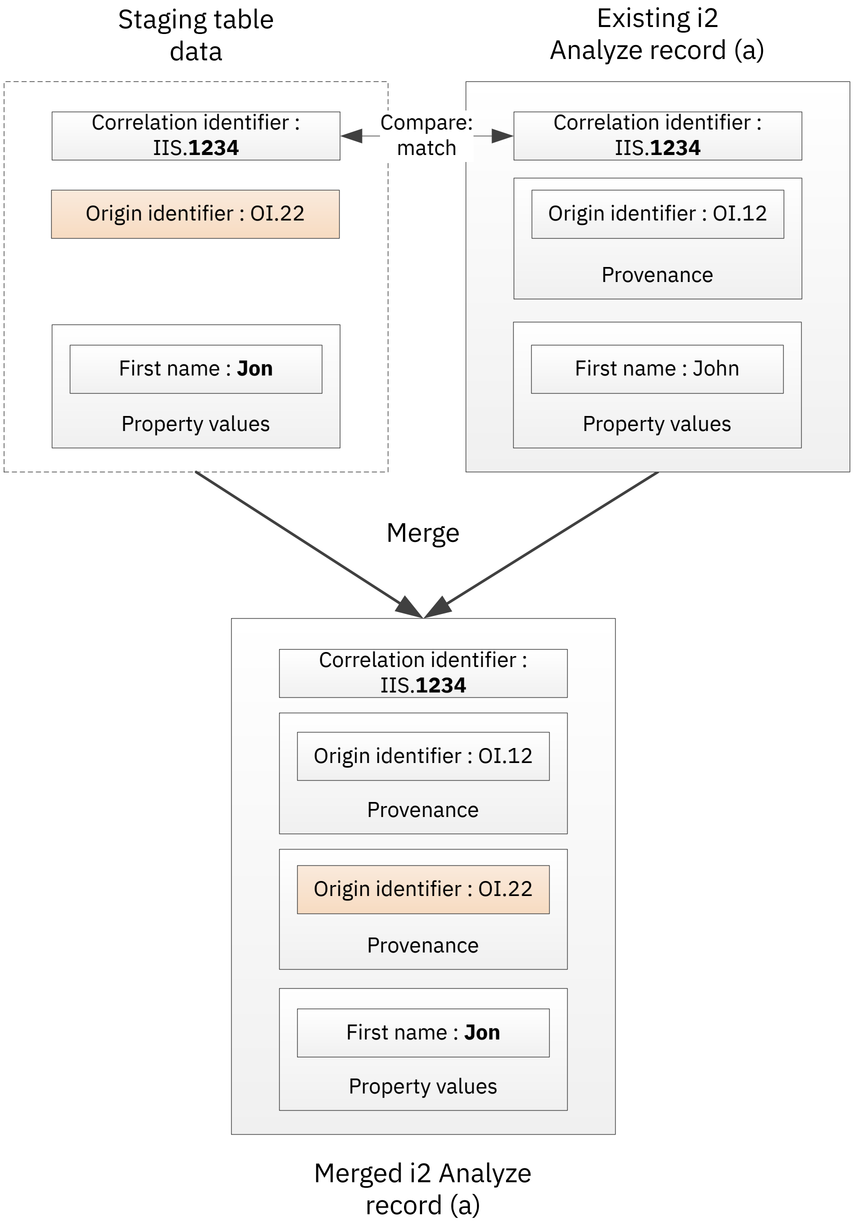
Staging table data
Existing i2 Analyze record (a)
Correlation identifier : IIS.**1234**
Compare: match
Correlation identifier : IIS.**1234**
Origin identifier : OI.22
Origin identifier : OI.12
Provenance
First name : **Jon**
Property values
First name : John
Property values
Merge
Correlation identifier : IIS.**1234**
Origin identifier : OI.12
Provenance
Origin identifier : OI.22
Provenance
First name : **Jon**
Property values
Merged i2 Analyze record (a)

Figure 1: Incoming staging data merges with an existing i2 Analyze record.

In the diagram, the correlation identifiers of data in the staging table and the existing i2 Analyze record match, which causes a merge operation. The existing i2 Analyze record is not associated with the origin identifier of the incoming data. In this example, it is assumed that the staging table data is more recent than the existing data. As part of the merge, the property values from the staging table row are used. This results in a change to the value for the first name property from "John" to "Jon". The merged i2 Analyze record now contains provenance for the origin identifier `OI.12` and one for the new data, `OI.22`.

Delete circular links

As a result of correlation operations, circular links might occur in the Information Store. The Information Store does not support circular links.

If the staging table causes existing records that are linked to each other in the Information Store to merge, as part of that merge the i2 Analyze link record would become circular. The existing link record is deleted from the Information Store. This is shown in the `DELETE_RECORD_COUNT` column of the ingestion report.

Correlation example

Example data sets are provided that have been passed through a matching engine, each row of data has been populated with a correlation identifier. You can ingest the example data, and then inspect the items in the Information Store from Analyst's Notebook Premium to demonstrate the correlation behavior.

Before you begin

Important: Before you use Information Store data correlation in your deployment of i2 Analyze, you must install the i2 Analyze 4.2.0.1 Fix Pack or Enterprise Insight Analysis 2.2.0.1 Fix Pack. For more information about downloading and installing the Fix Packs, see Release Material.

About this task

Use the `ingestExampleData` toolkit task to ingest example data sets that demonstrate the correlation behavior in i2 Analyze. For more information about the operations that occur, and how the i2 Analyze records are effected, see "Correlation operations" on page 4.

`law-enforcement-data-set-2`

This data set contains data with correlation identifiers. After you ingest this data set, the Information Store database contains i2 Analyze records with correlation identifiers. No correlation operations occur when you ingest this data set.

`law-enforcement-data-set-2-merge`

This data set contains data that causes a number of merge operations to occur with some of the i2 Analyze records that were created from the first data set.

You can find the example data sets in the `toolkit\examples\data` directory.

Procedure

1. In a command prompt, navigate to the `toolkit\scripts` directory.
2. Ingest the first example data set.
   a) Run the following command to ingest the first example data set:

```
setup -t ingestExampleData -e law-enforcement-data-set-2
```

   The `examples\data\law-enforcement-data-set-2` directory contains a set of CSV files that contain data that has been passed through a matching engine and processed to meet the requirements of the staging tables. Each row of data contains a correlation identifier type and a unique key value. The `LoadCSVDataCommands.db2` file is called to populate the staging tables with the example data.

The Information Store now contains i2 Analyze records with correlation identifiers. However, no correlation operations occurred during the ingestion.

b) In Analyst's Notebook Premium, search for "Julia Yochum" and add the returned entity to the chart.

3. Ingest the `law-enforcement-data-set-2-merge` data to demonstrate merge operations.

a) Run the following command to ingest the second example data set:

```
setup -t ingestExampleData -e law-enforcement-data-set-2-merge
```

The `examples\data\law-enforcement-data-set-2-merge` directory contains a set of CSV files that contain data that has been passed through a matching engine and processed to meet the requirements of the staging tables. Each row of data contains a correlation identifier type and key value. Each row contains a unique value for the `SOURCE_ID`, which represents the origin identifier.

In this scenario, the matching engine has identified that some of the data represents the same real-world objects as data in the first data set. As part of this process, the correlation identifiers match with some of the existing data in the database, but the origin identifiers are different. These matches cause a number of merge operations to occur during the ingestion.

For example, you can see on line 2 of the `person.csv` file, the correlation identifier key is `person_0` which matches the correlation identifier of the record for person "Julia Yochum". This match causes a merge between the existing record and the incoming row of data. As part of the merge, the property values from the incoming row of data are used for the record, this results in the full name changing to "Julie Yocham".

In the ingestion reports, you can see the number and type of correlation operations that occurred during the ingestion. For more information about understanding the ingestion reports, see Understanding ingestion reports.

b) In Analyst's Notebook Premium, select the chart item representing Julia Yochum and click **Get changes**.

You can see that the name has changed due to the merge operation described above.

What to do next

After you have investigated the correlation behavior, you can clear the example data from your system. For more information, see Clearing data from the system.

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.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM United Kingdom Limited Hursley House Hursley Park Winchester, Hants, SO21 2JN UK

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

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 IBM Corp. 2018

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.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java™ and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Other names may be trademarks of their respective owners. Other company, product, and service names may be trademarks or service marks of others.