FUZZY MATCHING WITH ARBUTUS

AN ARBUTUS SOFTWARE SOLUTION PAPER

CONTENTS

EXECUTIVE SUMMARY

Auditors have used fuzzy matching tools and technology for decades to find similarities, duplicates, or anomalies in data. Even so, these capabilities are still not generally available in most tools, and are seldom easy to use.

Today, most organizations have data scattered across countless unrelated systems, meaning the importance of performing fuzzy logic comparisons of data is greater than ever.

Fuzzy tools can help with:

- Data harmonization or de-duplication
- Fraud investigations
- Matching accounts from different systems for security purposes
- Identifying similarities for any analytic purpose, such as testing data quality

Arbutus technology provides auditors and business analysts with powerful and intuitive data analysis tools featuring robust fuzzy testing capabilities built right in. Common fuzzy comparison algorithms, like Soundex and Levenshtein, are just the start, as Arbutus technology puts easy-to-use fuzzy matching tools in the hands of its users.

Arbutus can also perform fuzzy comparisons on and between disparate data, including mainframe legacy data, source data files not in a data mart or warehouse, ERP-based data, and even web sources.

Look for this symbol: ❯ throughout for various filters, commands or functions in Arbutus that will help you perform sophisticated data analysis using fuzzy logic.

FUZZY MATCHING WITH ARBUTUS

Whether you are working with a large, legacy data file, a traditional flat file, or an everyday spreadsheet, if your task is to find similarities or duplicates in data – like names or addresses – "fuzzy matching" is almost always the best way to go.

Fuzzy technology can help with:

- Data harmonization or de-duplication
- Fraud investigations
- Matching accounts from different systems for security purposes
- Identifying similarities for any analytic purpose, such as testing data quality

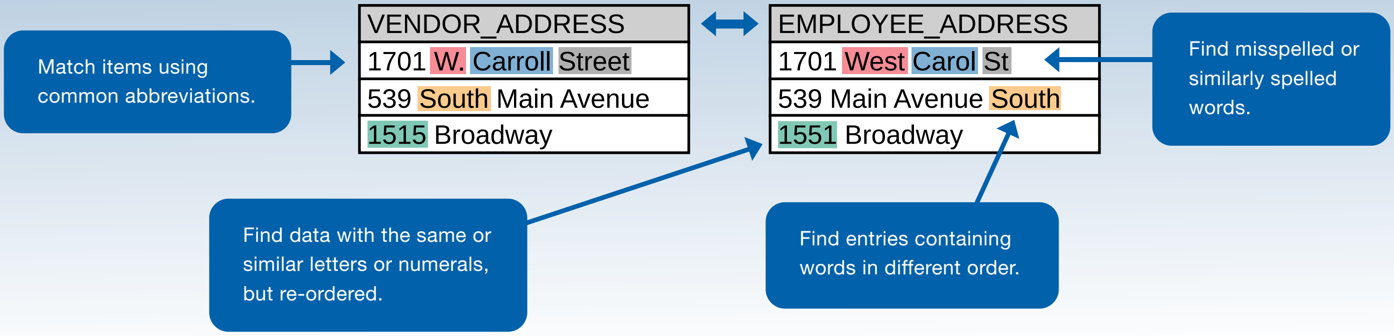
Various software solutions perform degree of difference comparisons using algorithms like SOUNDEX and LEVENSHTEIN, but few tools make it easy. Even worse, when the data is located on disparate or legacy systems, most tools with fuzzy logic capabilities can't even access the data.

If you work at a smaller organization, you face even greater challenges, as you may not have access to appropriate tools and IT resources. You may even need to know rudimentary SQL programming to attempt these comparisons in MS Access and Excel, which are far from ideal tools for the task.

With Arbutus, common difference comparison algorithms, like SOUNDEX or LEVENSHTEIN, are built right into functions such as SOUNDSLIKE, DIFFERENCE, and NEAR. You can also use Arbutus to prepare data for meaningful and timely fuzzy logic comparisons of data when you need to match data between unrelated systems, or test for data quality.

FRAUD TEST: FUZZY LOGIC AT WORK IN ARBUTUS

Are there employees at your organization posing as company vendors? This practice is prohibited at most organizations. In the fuzzy test below, the Address fields from two databases have been compared to find potential matches – and fraudsters. This test required no programming and was set up and deployed in under 5 minutes.



Match items using common abbreviations.

VENDOR_ADDRESS
1701 W. Carroll Street
539 South Main Avenue
1515 Broadway

EMPLOYEE_ADDRESS
1701 West Carol St
539 Main Avenue South
1551 Broadway

Find misspelled or similarly spelled words.

Find data with the same or similar letters or numerals, but re-ordered.

Find entries containing words in different order.

DIFFERENCE COMPARISONS

Arbutus offers a variety of built-in tools to perform 'exact' as well as 'fuzzy difference' comparisons, including:

- **SOUNDEX ALGORITHM** compares English names that are pronounced the same, or close, but spelled differently. Particularly useful with data transcribed from conversations or phone calls.
    - **SOUNDEX** generates the four-digit SOUNDEX code for any name, for comparison of overall similarity or difference.
    - **SOUNDSLIKE** encapsulates and extends the SOUNDEX algorithm, converting and comparing two strings for equality. For example, the filter:

      SOUNDSLIKE(name, “Smith”)

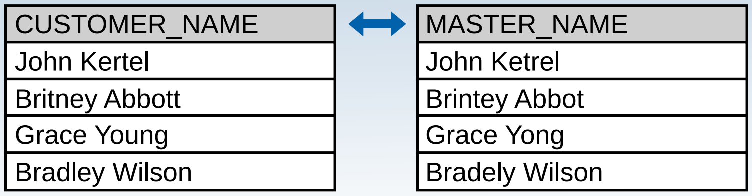
      ...will match “Smythe”, “Smithe” and “Schmidt”. Other examples include Catherine/Kathyrn, Lee/Leigh/Li, Fisher/Fischer, Don/Dawn, or Johanson/Johanssen.
- **LEVENSHTEIN ALGORITHM** compares string data, such as names or addresses.
    - **DIFFERENCE** calculates the LEVENSHTEIN distance for two strings based on a degree of similarity to be determined between the two strings. This can be used to identify a wide range of accidental or intentional data errors.
    - **NEAR** automatically uses the comparison method most appropriate for the data being compared (LEVENSHTEIN for strings, for example). NEAR comparisons can be applied to any of the fundamental data types so that you can easily identify data elements that, while not exact, are close.

FUZZY MATCH USING THE LEVENSHTEIN ALGORITHM

The following short function will identify data entries that differ by two or fewer characters, so you can find similarities in data residing in two or more related or unrelated data tables:

NEAR(customer.name, master.name, 2)

Using this function on “CUSTOMER_NAME” and “MASTER_NAME” yielded the following results, some of which might merit further investigation:



CUSTOMER_NAME	⟷	MASTER_NAME
John Kertel		John Ketrel
Britney Abbott		Brintey Abbot
Grace Young		Grace Yong
Bradley Wilson		Bradely Wilson

DIFFERENCE COMPARISONS (CONT'D)

- **FREE TEXT COMPARISONS** allow you to search for one or more string values in your data. Just like a Google search, "Smart Search" by Arbutus allows you to search for single or multiple terms, literals, or any other text, either in selected fields or in the entire record. "Smart Search" makes free text searching your data files as easy as any web search.
- **FORMAT** compares or categorizes data by its own characteristics, making it useful for data quality management applications. However, instead of comparing character to character, you can compare strings where digits match any digit and alpha data matches any other alpha.

Grouped comparisons allow you to match to sets or ranges of values:

- **BETWEEN** provides an easy way to specify a range of values to match
- **MATCH** allows you to specify a list of values to be matched against
- **COMPLEX COMPARISONS** allow any of the above techniques to be used in any combination, to achieve exact or fuzzy requirements. For example, these three could be combined:

  ```
  ZIP1=ZIP2 and Near(date1, date2,3) and
  (name1=name2 or amount1=amount2)
  ```

- **EXACT AND RELATIVE COMPARISONS** – Fuzzy comparisons often incorporate regular comparisons. With Arbutus, you can compare any data types for exact or relative comparisons. You can also compare data elements directly, regardless of how or where they are physically stored. This is because Arbutus internally standardizes data automatically, making comparing disparate data seamless.
  - **DATE AND TIME DATA** - Dates and times may be stored in any valid date, time or datetime format, or in any character set. Even datetimes stored in character or numeric data types can be flagged as datetimes. Arbutus automatically converts all datetimes internally to a standard datetime format and all datetime comparisons are made using this internal format (not the manner in which datetimes are physically stored).

DIFFERENCE COMPARISONS (CONT'D)

- **NUMERIC DATA** - When reading source data, Arbutus automatically standardizes all numeric data types (Packed, Zoned, Binary, etc.) internally, so that comparisons and mathematics between disparate numeric data types is as simple as:

  field1>field2
  or
  field5=field3+field4

  Even bit-level and non-byte-aligned numeric data is supported.
- **CHARACTER DATA** - Both ASCII and EBCDIC character data are directly supported and can be compared and combined without regard for the source character set.
- **MIXED DATA** - Arbutus provides functions that allow data stored using fundamentally different data types to be directly converted and compared.

DATA HARMONIZATION

When data isn't neatly arranged, Arbutus can help standardize or harmonize it. The following functions can be used in any combination, or combined with any of the comparisons described above, to suit your needs:

- **UPPER** and **LOWER** standardize the case of a string for consistent comparison. For example, "John Smith" can be automatically converted to "JOHN SMITH".
- **TRIM**, **LTRIM,** and **ALLTRIM** remove leading and/or trailing blanks to improve data quality, so "John Smith " becomes "John Smith".
- **COMPACT** removes extra blanks between words. Like TRIM, it improves comparability, as "John   Smith" becomes "John Smith".
- **INCLUDE** and **EXCLUDE** are functions that specify characters to be kept or removed (e.g., blank spacing, punctuation, foreign characters, etc.) to ensure punctuation or formatting does not reduce comparability. For example:

  Include(phone, "0~9")

  ...will convert "(888) 123-4567" into "8881234567".

DATA HARMONIZATION (CONT’D)

- **REPLACE** is useful for standard abbreviations (e.g., ST. or ST for STREET) as well as to correct common data entry errors (like I for 1). To use another phone example, some entries with country codes might be entered as “+1 888 123 4567” or “+44 1 234 567”.

  Replace(phone, “+1 “, ””, “+”, “”)

  ...automatically removes any “+1” North American prefixes, as well as the “+” from any other country codes.

- **NORMALIZE** combines various harmonization techniques described above, as it automatically:
  - replaces non-blank non-alphanumerics (such as punctuation) with blanks
  - trims leading or trailing blanks and compacts contiguous blanks
  - replaces foreign characters with English equivalents
  - capitalizes the result (such as JOHN for John)

  You can also apply any number of standardized data substitutions or removals (e.g., William/Wm., New Jersey/NJ, Boulevard/BLVD).

STANDARDIZE DATA IN SECONDS

Address fields often contain unnecessary descriptions, such as Unit, #, Suite or Apt, or extra punctuation, like commas or periods. The NORMALIZE filter automatically cleans up data like this:

ADDRESS
Suite 45, 123 W Main Street, Miami, Florida, USA
#45 - 123 WEST MAIN ST, MIAMI FL US
Apt.,45,123 W. Main Street, Miami,Florida,U.S.A..

→

NORMALIZED_ADDRESS
45 123 W MAIN ST MIAMI FL USA
45 123 W MAIN ST MIAMI FL USA
45 123 W MAIN ST MIAMI FL USA

- **ARRANGE** rearranges characters in a string into descending order. This is a special purpose test that is particularly useful in identifying transposition errors, such as (888) 132-4567, or words in different orders.

- **SUBSTRING** selects a portion of a string for comparison. Continuing with our phone number example, if you had already harmonized the phone numbers to “8881234567” then:

  Substring(phone, 1, 3)

  ...would extract just the area code for comparison.

PERFORMING FUZZY COMPARISONS

Once you have your comparisons identified and your data harmonized, the next step is to perform the comparisons.

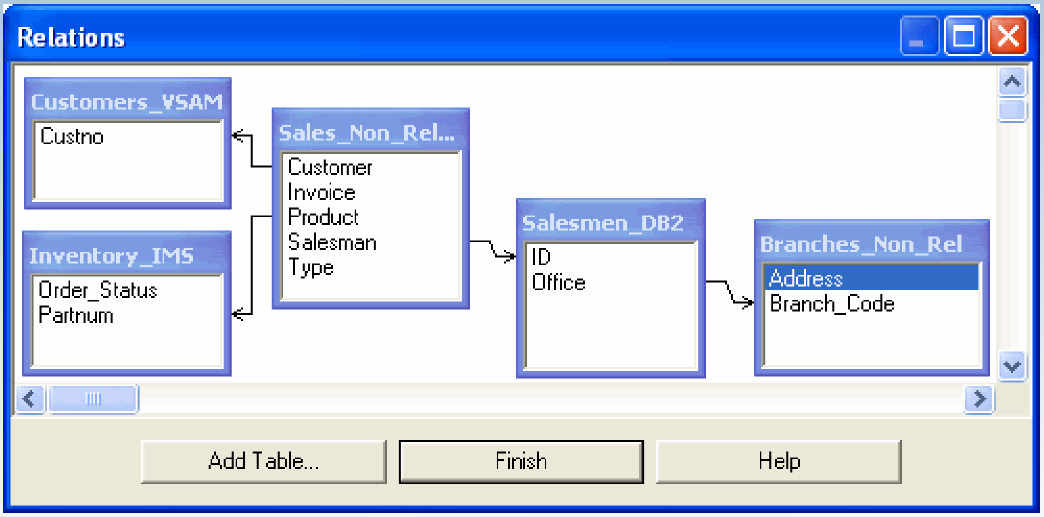
When you are making **comparisons within a file**, you may be comparing data within the same record, or comparing data between records. In either case, the data on either side of the comparison may be a simple field, the result of an expression, or the result of a multi-value virtual field that takes on its value based on defined criteria.

If you are making **comparisons between files that share a common key**, either directly or indirectly, Arbutus offers two different options:

- **RELATIONS** virtually combines tables based on a common key. (*see graphic below*)
- **JOIN** combines data from multiple tables to create a new, physical table. JOIN allows matching based on one or more common keys, and will let you output data from matched records, unmatched records, or a combination of both. (*see graphic on next page*)

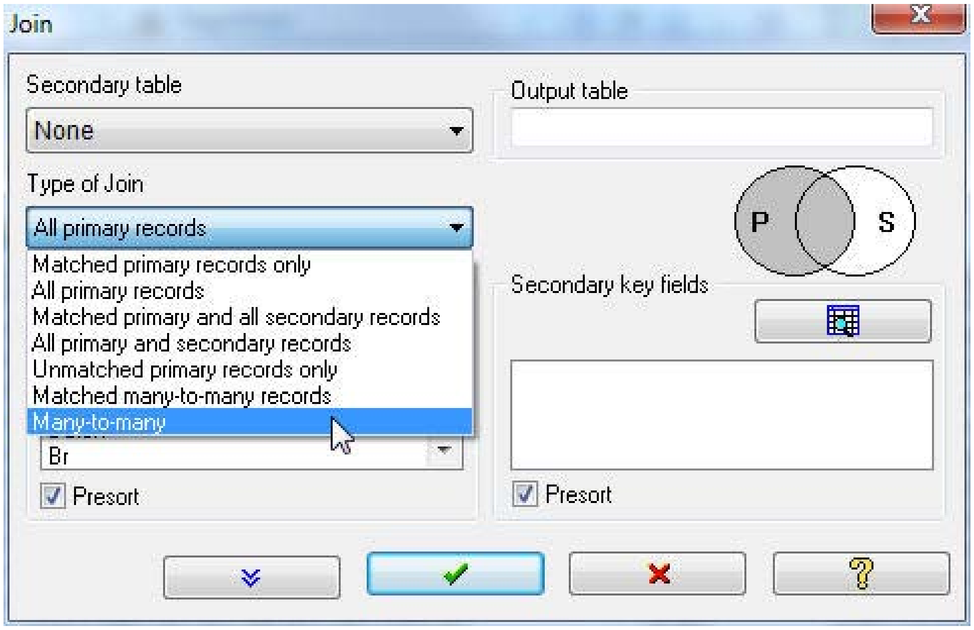
RELATIONS: VIRTUAL RELATIONSHIPS BETWEEN FILES OR TABLES

With Arbutus, even disparate data can be prepared for fuzzy testing. Using the click-and-drag interface of Arbutus, you can easily define **RELATIONS** between any tables in a star-schema or snowflake-schema arrangement. You can also define multiple, independent relationships.



COMBINE AND COMPARE DATA FROM MULTIPLE TABLES

JOIN offers a broader choice of options than RELATIONS from the primary to the secondary table. Like RELATIONS, a JOIN can be performed in a few clicks, and even offers the option of selecting your JOIN type by clicking on a Venn diagram.


Join

Secondary table: None

Output table:

Type of Join: All primary records

Br

☑ Presort

Secondary key fields

☑ Presort

Matched primary records only
All primary records
Matched primary and all secondary records
All primary and secondary records
Unmatched primary records only
Matched many-to-many records
Many-to-many
P
S

Arbutus also offers ways to make **comparisons between files that do not share a traditional key**:

- **MATCHED MANY-TO-MANY JOIN** combines records based on a common data element that would not normally be considered a traditional key, such as a ZIP code, city, or state. A MATCHED MANY-TO-MANY JOIN can greatly reduce the complexity of the JOIN results and significantly improve performance when joining large sets of data.
- **MANY-TO-MANY JOIN** combines all records in one file with all records in a second file, in the same way as a standard SQL join. You can then apply any of the harmonization and comparison techniques described above to limit the size of the resulting data set. This technique can be used to compare data from any source, including disparate sources. Keep in mind a MANY-TO-MANY JOIN often yields a complex result, so it is almost always used as a last resort.

COMPARING DISPARATE OR LEGACY DATA

When comparing data, you are rarely lucky enough to have it reside in a single, homogeneous environment. Your data may reside in relational systems like Oracle, ERP systems such as SAP, legacy mainframe environments of often heinous complexity and size, or even on web pages. Often, it will span several of these.

With Arbutus servers, you can compare all of this data.

Arbutus Servers are native server-based technologies optimized to process large or complex files in a highly efficient manner. All Arbutus servers display their native data in an instantly recognizable tabular format.

- The Arbutus Windows Server supports virtually all Windows-based data, and directly reads relational sources (via ODBC), including SAP and even web URLs.
- The Arbutus zSeries and iSeries servers allow you to directly read native databases (such as IMS, DB2, and ADABAS) along with native data (like VSAM and QSAM) regardless of their internal complexities.

What makes Arbutus servers unique for data comparison is that all Arbutus servers can freely exchange their data. As a result, data from these disparate platforms can be easily transferred to a single platform – typically the Arbutus Windows Server – and compared. This allows you to compare data sources you might not have otherwise considered.

EXTENDING THE REACH OF COMMON APPLICATIONS

If you already use common tools such as Microsoft Excel or Crystal Reports, Arbutus' powerful data access and harmonization capabilities can extend their reach.

Arbutus can provide access to all of your data via LegacyLink, which is an ODBC driver that acts like a "data pipe", providing any ODBC-compliant Windows-based application with read-only access to all of your data.

Imagine reading harmonized mainframe legacy data directly from Excel or Crystal Reports. With Arbutus, that dream becomes a reality.

THE LAST WORD

Auditors often struggle with some form of programming or with a work-around to the technical limitation of their current tools. Many times, fuzzy logic-based testing is not done due to these barriers. Arbutus technology helps auditors and other business users find new and better results from their analyses using fuzzy technology tools.

- Visit www.ArbutusSoftware.com/form-eval
  to request a 30-day free trial of Arbutus Analyzer.
- Request a 20-minute web demo to see Analyzer in action.
  Contact us today via email or call toll-free.

ARBUTUS

#270 - 6450 Roberts Street
Burnaby, British Columbia
Canada V5G 4E1

Toll-free: 877.333.6336
Direct: 604.456.6336
Fax: 604.437.7872
info@ArbutusSoftware.com
www.ArbutusSoftware.com

Based on 25 years of innovation excellence, Arbutus delivers the very best in purpose-built audit analytics technology to meet the exacting demands of today's business environment. Auditors, business analysts, and fraud investigators rely on Arbutus to enhance their testing, analysis, and compliance capabilities.

The data universe is wide and varied. One of our core strengths as a technology firm is the ability to easily work with all types of data, both legacy and non-legacy. Arbutus solutions allow auditors, IT, and business professionals to overcome many of their current constraints in areas such as data migration, data quality, fraud detection, and data analysis.

Arbutus Audit Analytics, our flagship product suite, is a proven solution used by auditors, business professionals, IT, and management all over the world. With outstanding customer service, strong product support, and flexible licensing, Arbutus Software offers the best value for advanced data conversion, migration, and analysis solutions.

Arbutus Software Inc. is a privately held company based in Greater Vancouver. For more information about our company or products, please contact us.