

Spectrum™ Technology Platform

Version 2019.1.0

Summary of Features

Spectrum™ Technology Platform is an enterprise data quality solution that improves the completeness, validity, consistency, timeliness, and accuracy of your data through standardization, verification, and enhancement. Ensuring that your data is accurate, complete, and up to date enables your firm to better understand and connect with your customers.

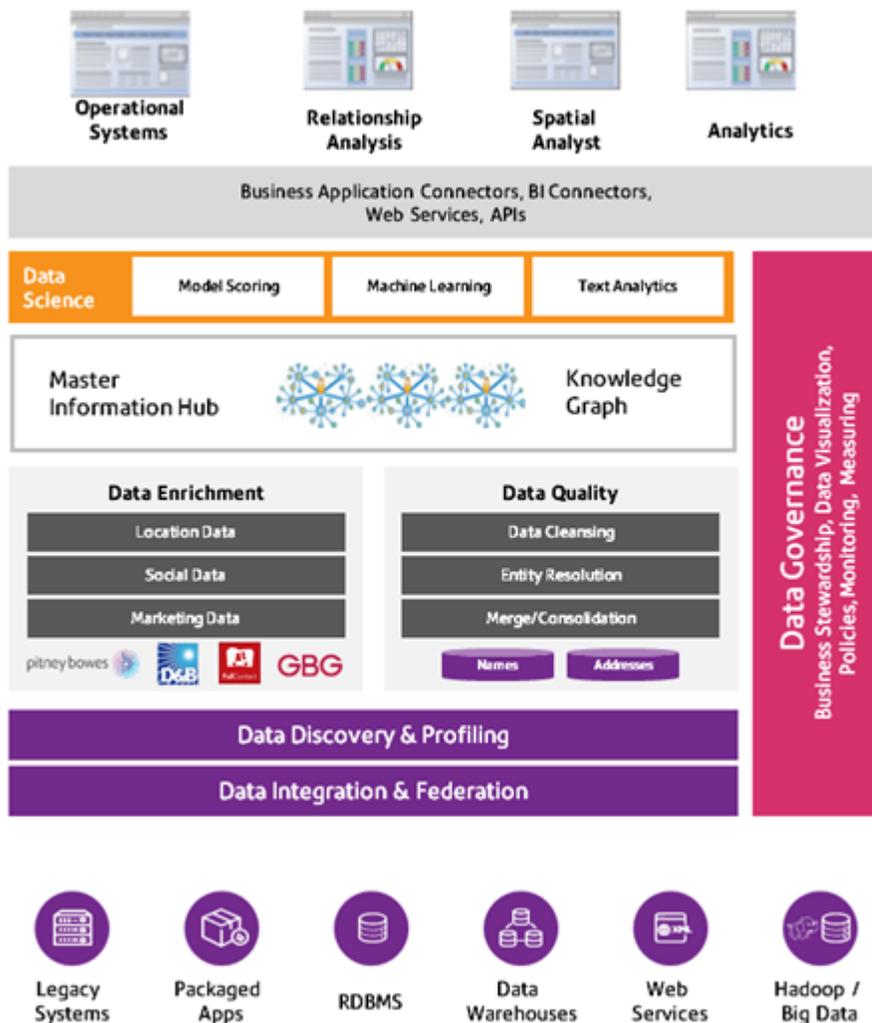
Contents:

Enterprise Data Management Architecture	2
Spectrum™ Technology Platform Architecture	5
Bundles and Modules	9



Enterprise Data Management Architecture

With Spectrum™ Technology Platform, you can build a comprehensive enterprise data management process, or you can use it as a more targeted solution. This diagram illustrates a complete solution that takes data from its source, through data enrichment and data quality processes, feeding a master data management hub which makes a single view of the data available to multiple business applications.



Master Information Hub

The Master Information Hub hub allows for rapid modeling of entities and their complex relationships across roles, processes, and interactions. It provides built-in social network analysis capabilities to

help you understand influencers, predict churn, detect non-obvious relationships and fraudulent patterns, and provide recommendations.

Spectrum™ Technology Platform supports two approaches to the MDM hub. In the master hub approach, the data is maintained in a single MDM database and applications access the data from the MDM database. In the registry approach, the data is maintained in each business application and the MDM hub registry contains keys which are used to find related records. For example, a customer's record may exist in an order entry database and a customer support database. The MDM registry would contain a single key which could be used to access the customer data in both places.

Data Enrichment

Data enrichment processes augment your data with additional information. You can base enrichment on spatial data, marketing data, or data from other detail sources. For example, if you have a database of customer addresses, you could geocode the address to determine the latitude/longitude coordinates of the address and store those coordinates as part of the record. You can then use your customer data to perform a variety of spatial calculations, such as finding the customer's nearest bank branch. Spectrum™ Technology Platform allows you to enrich your data with a variety of information, including geocoding (with the Enterprise Geocoding Module), tax jurisdiction assignment (with the Enterprise Tax Module), geospatial calculations (with the Spatial Module), and travel directions between points (with the Spatial Module).

Data Quality and Data Governance

Data quality and data governance processes check your data for duplicate records, inconsistent information, and inaccurate information.

Duplicate matching identifies potential duplicate records or relationships between records, whether the data is name and address in nature or any other type of customer information. Spectrum™ Technology Platform allows you to specify a consistent set of business match rules using Boolean matching methods, scoring methods, thresholds, algorithms, and weights to determine if a group of records contains duplicates. Spectrum™ Technology Platform supports extensive customization so you can tailor the rules to the unique needs of your business.

Once duplicate records have been identified, you may wish to consolidate records. Spectrum™ Technology Platform allows you to specify how to link or merge duplicate records so you can create the most accurate and complete record from any collection of customer information. For example, you can build a single best-of-breed record from all of the records in a household. The Advanced Matching Module is used to identify duplicates and eliminate them.

Data quality processes also standardize your data. Standardization is a critical process because standardized data elements are necessary to achieve the highest possible results for matching and identifying relationships between records. While several modules perform standardization of one type or another, the Spectrum™ Technology Platform Data Normalization module provides the most comprehensive set of standardization features. In addition, the Universal Name module provides specific data quality features for handling personal name and business name data.

Standardized data is not necessarily accurate data. Spectrum™ Technology Platform can compare your data to known, up-to-date reference data for correctness. The sources used for this process may include regulatory bodies such as the U.S. Postal Service, third-party data providers such as Experian or Dunn and Bradstreet, or your company's internal reference sources, such as accounting data. Spectrum™ Technology Platform is particularly strong in address data validation. It can validate or standardize addresses in 250 countries and territories around the world. There are two modules that perform address validation: the Address Now Module and the Universal Addressing Module.

To determine which one is right for you, discuss your needs with your account executive.

While Spectrum™ Technology Platform can automatically handle a wide range of data quality issues, there are some situations where a manual review by a data steward is appropriate. To support this, the Business Steward Module provides a way to specify the rules that will trigger a manual review, and it provides a web-enabled tool for reviewing exception records. It includes integrated access to third-party tools such as Bing maps and Experian data to aid data stewards in the review and resolution process.

Data Discovery and Profiling

Data discovery is the process of scanning your data resources to get a complete inventory of your data landscape. Spectrum™ Technology Platform can scan structured data, unstructured data, and semi-structured data using a wide array of data profiling techniques. The results of the scan are used to automatically generate a library of documentation describing your company's data assets and to create a metadata repository. This documentation and accompanying metadata repository provide the insight you need before beginning data integration, data quality, data governance, or master data management projects.

For more information about the Spectrum™ Technology Platform Data Discovery Module, contact your account executive.

Data Integration and Federation

Once you have an inventory of your data landscape, you need to consider how you will access the data you need to manage. Spectrum™ Technology Platform can connect to data in multiple sources either directly or through integration with your existing data access technologies. It supports batch and real-time data integration capabilities for a variety of business needs, including data warehousing, data quality, systems integration, and migration. Spectrum™ Technology Platform can access data in RDBMS databases, data warehouses, XML files, flat files, and more. Spectrum™ Technology Platform supports SQL queries with complex joins and aggregations and provides a visual query development tool. In addition, Spectrum™ Technology Platform can access data over REST and SOAP web services.

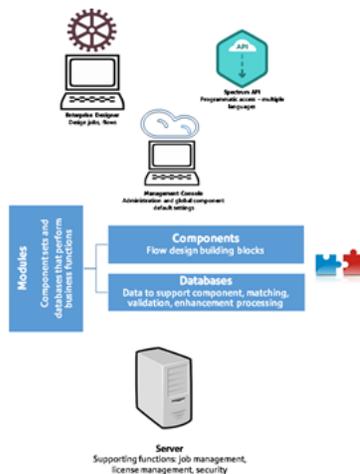
Spectrum™ Technology Platform can trigger batch processing based on the appearance of one or more source files in a specified folder. This "hot folder" trigger is useful for monitoring FTP uploads and processing them as they occur.

Some of these data integration capabilities require a license for the Enterprise Data Integration Module. For more information, contact your account executive.

Finally, Spectrum™ Technology Platform can integrate with packaged applications such as SAP.

Spectrum™ Technology Platform Architecture

Spectrum™ Technology Platform from Pitney Bowes consists of a server that runs a number of modules. These modules provide different functions, such as address validation, geocoding, and advanced parsing, among others. This diagram illustrates the Spectrum™ Technology Platform architecture.



Server

The foundation of the Spectrum™ Technology Platform is the server. The server handles data processing, synchronizes repository data, and manages communication. It provides job management and security features.

Modules

Modules are sets of features that perform a specific function. For example, the Universal Addressing Module standardizes addresses to conform to postal standards. The Enterprise Tax Module determines the tax jurisdictions that apply to a given address. Modules are grouped together to solve common business problems and licensed together as bundles.

Components

Modules are comprised of components which perform a specific function in a flow or as a service. For example, the Enterprise Geocoding module's Geocode US Address component takes an address and returns the latitude and longitude coordinates for that address; the Universal Addressing module's Get City State Province takes a postal code and returns the city and state or province where that postal code is located.

The components that you have available on your system depend on which Spectrum™ Technology Platform bundle you have licensed.

Databases

Some modules depend on databases containing reference data. For example, the Universal Addressing module needs to have access to U.S. Postal Service data in order to verify and standardize addresses in the U.S. Databases are installed separately and some are updated on a regular basis to provide you with the latest data.

Modules have both required and optional databases. Optional databases provide data needed for certain features that can enhance your Spectrum™ Technology Platform process.

Management Console

Management Console is a tool for administering Spectrum™ Technology Platform. You can use Management Console to:

- Define the connections between Spectrum™ Technology Platform and your data
- Specify the default settings for services and flows
- Manage user accounts, including permissions and passwords
- View logs
- View licenses including license expiration information

Management Console | Flows | Services | Resources | System | ? admin

Home > Resources: Data Sources

Data Sources

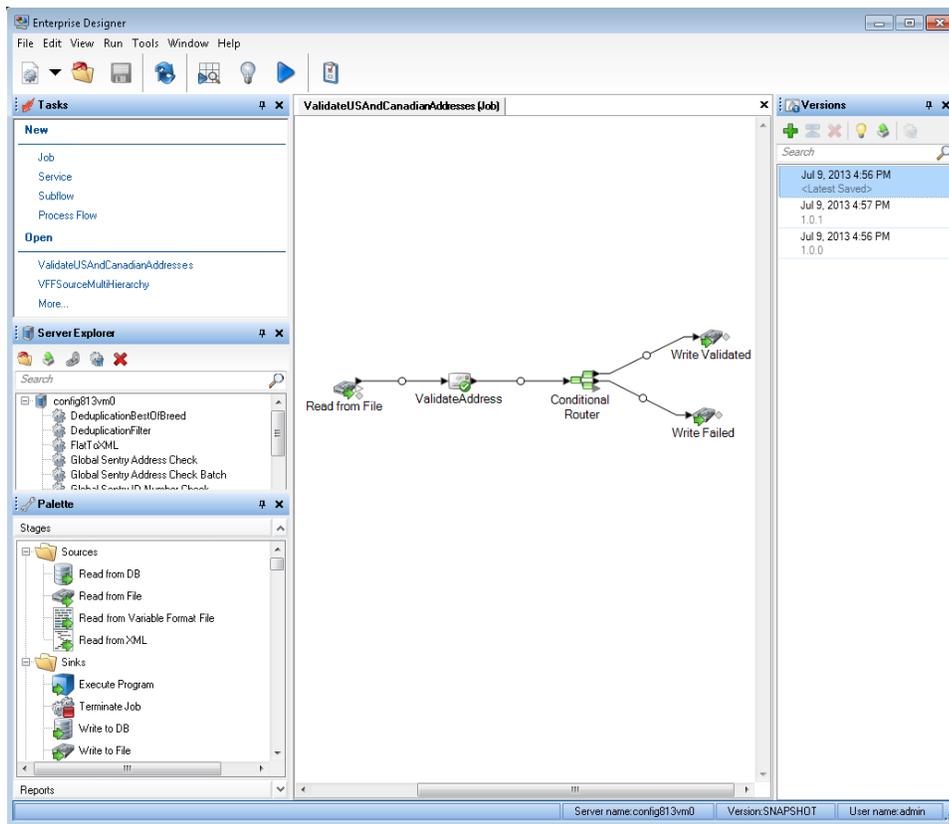
+ [edit] [delete] [share] [filter]

Name	Type
test1	FTP
test2	FTP
test4	Cloud
test5HDFS	HDFS
mdg1teamcity1	FTP

Showing 5 of 5 records | Rows per page 10

Enterprise Designer

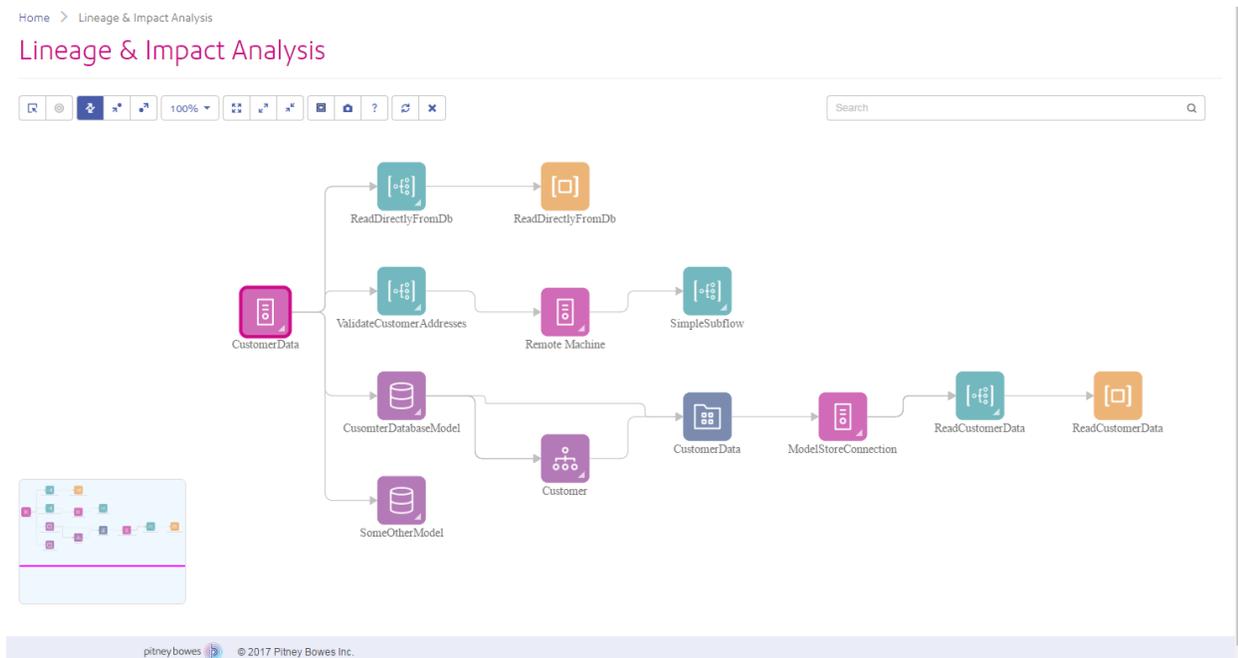
Enterprise Designer is a tool for creating Spectrum™ Technology Platform jobs, services, subflows, and process flows. It provides a familiar drag-and-drop interface to allow you to graphically create complex flows.



Note: Enterprise Designer will be replaced by Flow Designer in a future release. Flow Designer is in Technical Preview status at this time.

Metadata Insights

Metadata Insights gives you the control you need to deliver accurate and timely data-driven insights to your business. Use Metadata Insights to develop data models, view the flow of data from source to business application, and assess the quality of your data through profiling. With this insight, you can identify the data resources to use to answer particular business questions, adapt and optimize processes to improve the usefulness and consistency of data across your business, and troubleshoot data issues.



Web Services and API

You can integrate Spectrum™ Technology Platform capabilities into your applications using web services and programming APIs. These interfaces provide simple integration, streamline record processing, and support backward compatibility of future versions.

The Spectrum™ Technology Platform API is available for these languages:

- C
- C++
- COM
- Java
- .NET

Web services are available via SOAP and REST.

Administration Utility - Command Line Interface (CLI)

The Administration Utility provides command line access to administrative functions. You can run commands interactively or in scripts. Some administrative functions are not available in the Administration Utility. For these functions, you can use Management Console as well as some component applications.

Bundles and Modules

Spectrum™ Technology Platform bundles are designed to serve many business needs. Bundles simplify the purchasing process by providing a preselected collection of modules with the capabilities necessary to meet particular business needs. Modules provide specific processing capabilities such as geocoding, address validation, matching, standardization, and more.

Address Management Bundle

Improve the quality of incomplete and inaccurate addresses based on the best available referential data.

Business Steward Module

The Business Steward Module is a set of features that allow you to identify and resolve exception records. Exception records are records that Spectrum™ Technology Platform could not confidently process and that require manual review by a data steward. Some examples of exceptions are:

- Address verification failures
- Geocoding failures
- Low-confidence matches
- Merge/consolidation decisions

The Business Steward Module provides a browser-based tool for manually reviewing exception records. Once exception records are manually corrected and approved, you can add them back into your Spectrum™ Technology Platform data quality process.

The audit log shows many of the actions that take place within the Business Steward Module. The audit log is a Management Console tool that records user activity. The log may include:

- Adding exceptions in the Write Exceptions stage
- Deleting exceptions in the Read Exceptions stage and the Business Steward Portal Manage Exceptions page
- Assigning exceptions in the Business Steward Portal Exception Manager
- Read exceptions in the Business Steward Portal Exception Editor
- Updating exceptions in the Business Steward Portal Exception Editor
- Revalidating exceptions in the Business Steward Portal Exception Editor

Read more about the audit log in the Spectrum™ Technology Platform Administration Guide for the WebUI.

Components

The Business Steward Module consists of:

- **Exception Monitor**—A stage that evaluates records against a set of conditions to determine if the record requires manual review by a data steward. When records meet those conditions, this stage can send an email notifying recipients of the exceptions.
- **Write Exceptions**—A stage that writes the exception records to the exception repository. Once exception records are in the exception repository they are available for review by a data steward.
- **Exception Dashboard and Editor**—A browser-based tool that displays a dashboard of summary statistics and charts to help you understand the kinds of exceptions that occur in your data. It is also an editor that allows you to modify exception records and approve them for reprocessing.
- **Manage Exceptions**—A browser-based tool that enables you to review and manage exception record activity for all users.
- **Data Quality Performance**—A browser-based tool that provides information on trends within your exception records and enables you to identify key performance indicators and send notifications when certain conditions are met.
- **Read Exceptions**—A stage that reads approved or non-approved exceptions from the exception repository. This stage allows you to reprocess exception records that have been corrected by a data steward.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.
- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.
- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Global Addressing Module

The Global Addressing Module for the Spectrum™ Technology Platform provides enhanced address standardization and validation. The Global Addressing Module benefits include:

- **Better data**—Global Address Validation combines data from multiple data sources into a single Global Addressing Validation database to provide the most extensive and accurate international addressing data possible. For the United States, we offer a USPS® CASS Certified™ solution.
- **Global addressing expertise**—Global Addressing Validation analyzes and compares each input address to the Global Addressing database for the appropriate country. If needed, Global Address Validation corrects and formats the address based on postal standards and user preferences.
- **Automated address suggestion**—Global Type Ahead automatically suggests addresses as you type characters.
- **Address parsing**—Global Address Parser splits address strings into individual address elements with the help of machine learning techniques.
- **Database searching**—US Database Lookup provides the ability to search the US Database directly for address information including house number, street name, city name, and ZIP Code.

Components

The Global Addressing Module consists of:

- **Global Address Parser**—Splits address strings into individual address elements, such as organization name, city, locality, district, and postal code, using machine learning techniques.
- **Global Address Validation**—Analyzes and compares each input address to the Global Addressing database for the appropriate country. If needed, Global Address Validation corrects and formats the address using the postal standards for that country.
- **Global Type-Ahead**—Automatically suggests addresses as you type and immediately returns candidates based on your input.
- **US Database Lookup**—Provides the ability to search the US Database directly for address information.

Universal Addressing Module

The Universal Addressing Module is an address quality module that can standardize and validate addresses, improving the deliverability of mail. The Universal Addressing Module can ensure that your address data adheres to quality standards established by the postal authority. An address that adheres to these standards is more likely to be delivered in a timely manner. In addition, mailers who follow these standards can qualify for significant postage discounts. For information on discounts for U.S. mail, refer to the USPS *Domestic Mail Manual (DMM)* available at <https://postalpro.usps.com/>. For information on discounts for Canadian mail, refer to the Canada Post website at www.canadapost.ca. For information on discounts for Australian mail, refer to the Australia Post website at www.auspost.com.au.

You can use the Universal Addressing Module in batch mode, real-time mode, or as a hosted service, depending on which option you have licensed. The batch version of the Universal Addressing Module is CASS Certified™ by the USPS®.

Components

The Universal Addressing Module consists of the following components. These components can work with U.S., Canadian, Australian, and international addresses as long as you have licensed the

appropriate database (if you are running Universal Addressing in your environment) or hosted service (if you are using Universal Addressing through the Pitney Bowes hosted services).

- **Auto Complete Loqate**— Offers real-time entry of address data and returns instant results based on each character entered into the form, ensuring only accurate data is entered into the database.
- **Get Candidate Addresses**— Returns a list of possible matches for a given address.
- **Get Candidate Addresses Loqate**— Returns a list of possible matches for a given address using a Loqate engine and database.
- **Get City State Province**— Returns the city and state or province for a given postal code.
- **Get City State Province Loqate**— Returns the city and state or province for a given postal code using a Loqate engine and database.
- **Get Postal Codes**— Returns the postal codes for a given city.
- **Get Postal Codes Loqate**—Returns the postal codes for a given city using a Loqate engine and database.
- **Validate Address**—Standardizes and validates addresses using U.S., Canadian, and international postal data.
- **Validate Address Global**— Provides enhanced address standardization and validation for addresses outside the U.S. and Canada. Validate Address Global can also validate addresses in the U.S. and Canada, but its strength is in validating addresses in other countries. If you process a significant number of addresses outside the U.S. and Canada, consider using Validate Address Global.
- **Validate Address Loqate**— Standardizes and validates addresses using postal authority address data. Validate Address Loqate can correct information and format addresses using the format preferred by an applicable postal authority. It also adds missing postal information, such as postal codes, city names, and state or province names.

Analytics Scoring Bundle

Score a model's fit using the Data Science bundle in production IT environments.

Analytics Scoring Module

The Analytics Scoring Module:

- Enables in-dataflow evaluation using predictive models defined in either QMML (Spectrum Miner's proprietary model format) or PMML (industry-standard Predictive Model Markup Language) formats. Dataflows can write and retrieve data from Miner datasets used by the Miner predictive analytics software. This approach to modeling enables data enrichment by adding predicted or scored output from models created by your data insight team with industry-standard data modeling tools. Examples include models that calculate churn risk for existing customers or credit scoring models to determine a consumer's credit rating.
- Allows Java scoring to support scoring a model's fit with the Machine Learning Module.
- Score new data using an existing model in a production IT environment.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.
- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.
- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.
- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.
- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Enterprise Data Integration Module

The Enterprise Data Integration Module provides the features you need to connect to data in multiple sources either directly or through integration with your existing data access technologies. With the Enterprise Data Integration Module, you can use Spectrum™ Technology Platform for data warehousing, data quality, systems integration, and migration.

Components

The Enterprise Data Integration Module consists of:

- **Call Stored Procedure**—A source stage that executes a stored procedure in a database, and returns the results of the stored procedure call as input for the dataflow. Use Call Stored Procedure when you want to get data from a database using a database's stored procedure rather than a query to a table or view.
- **Field Selector**—Allows you to choose which fields to pass to the next stage in the dataflow. You can use Field Selector to remove unwanted fields from a dataflow. For example, if you have created a new field by combining the data from two fields, and you no longer need the two source fields, you can use Field Selector to retain only the new field and remove the two source fields from the dataflow.
- **Generate Time Dimension**—Creates date records, one for each day of the date range you specify. You can then write these records to a time dimension table in a database using the Write to DB stage. The time dimension table can then be used to perform accurate calculations based on a period. For example, sales by quarter, budget spend by quarter, and revenue by day are all analyses that require a time dimension. Time dimension tables also enable you to account for fiscal years or non-standard quarters in the analysis.
- **Query Cache**—Looks up data in a cache based on values in one or more dataflow fields and returns data from matching records in the cache, adding the cache record's data to the record in the dataflow. Looking up data in a cache can improve performance compared to looking up data in a database.

- **Read From DB**—Reads data from a database table or view as input to a dataflow. Read From DB is available for jobs and subflows but not services.
- **Read From HL7**—Reads Health Level Seven (HL7) data from a text file as input to a dataflow. HL7 is a messaging standard used in the healthcare industry to exchange data between systems. For more information about HL7, review the details at www.hl7.org.
- **Read from Spreadsheet**—Reads data from an Excel spreadsheet as input to a dataflow. It supports both .xls and .xlsx file formats.
- **Read From Variable Format File**—Reads data from a file containing records of varying layout, reading each record as a list field. You can specify the tag that indicates the parent record type, and all other record types will become list fields under the parent.
- **Read From XML**—Reads an XML file into a job or subflow. It defines the file's path and data format, including XML schema and data element details.
- **Transposer**—Converts columns to rows. Transposing data is the opposite of pivoting data using the Group Statistics stage, which transforms row data into columns.
- **Write to Cache**—Loads output from a dataflow into a global cache, making the data available for lookup from the Query Cache stage. Using a global cache for data lookups improves performance compared to lookups to databases.
- **Write to DB**—Writes the output of a dataflow to a database.
- **Write to Spreadsheet**—Writes data to an Excel spreadsheet as output from a dataflow.
- **Write to Variable Format File**—Writes records of varying layouts to an output file.
- **Write to XML**—Writes the output of a job or subflow to an XML file.

Data Federation Bundle

Create virtual representations of multiple datasets to explore and manipulate the information as if it were a single dataset. Present a real time combined set of data, without having to move the information from its physical database.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical

data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.

- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.
- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Data Integration Bundle

Bring together data from multiple systems for query, analysis, cleansing and enrichment purposes. Connect to big data sources, both structured and unstructured, and combine that data with other datasets to offer the business a broader set of information from which to gain insights.

Enterprise Data Integration Module

The Enterprise Data Integration Module provides the features you need to connect to data in multiple sources either directly or through integration with your existing data access technologies. With the Enterprise Data Integration Module, you can use Spectrum™ Technology Platform for data warehousing, data quality, systems integration, and migration.

Components

The Enterprise Data Integration Module consists of:

- **Call Stored Procedure**—A source stage that executes a stored procedure in a database, and returns the results of the stored procedure call as input for the dataflow. Use Call Stored Procedure when you want to get data from a database using a database's stored procedure rather than a query to a table or view.
- **Field Selector**—Allows you to choose which fields to pass to the next stage in the dataflow. You can use Field Selector to remove unwanted fields from a dataflow. For example, if you have created a new field by combining the data from two fields, and you no longer need the two source fields, you can use Field Selector to retain only the new field and remove the two source fields from the dataflow.
- **Generate Time Dimension**—Creates date records, one for each day of the date range you specify. You can then write these records to a time dimension table in a database using the Write to DB stage. The time dimension table can then be used to perform accurate calculations based on a period. For example, sales by quarter, budget spend by quarter, and revenue by day are all analyses that require a time dimension. Time dimension tables also enable you to account for fiscal years or non-standard quarters in the analysis.
- **Query Cache**—Looks up data in a cache based on values in one or more dataflow fields and returns data from matching records in the cache, adding the cache record's data to the record in the dataflow. Looking up data in a cache can improve performance compared to looking up data in a database.

- **Read From DB**—Reads data from a database table or view as input to a dataflow. Read From DB is available for jobs and subflows but not services.
- **Read From HL7**—Reads Health Level Seven (HL7) data from a text file as input to a dataflow. HL7 is a messaging standard used in the healthcare industry to exchange data between systems. For more information about HL7, review the details at www.hl7.org.
- **Read from Spreadsheet**—Reads data from an Excel spreadsheet as input to a dataflow. It supports both .xls and .xlsx file formats.
- **Read From Variable Format File**—Reads data from a file containing records of varying layout, reading each record as a list field. You can specify the tag that indicates the parent record type, and all other record types will become list fields under the parent.
- **Read From XML**—Reads an XML file into a job or subflow. It defines the file's path and data format, including XML schema and data element details.
- **Transposer**—Converts columns to rows. Transposing data is the opposite of pivoting data using the Group Statistics stage, which transforms row data into columns.
- **Write to Cache**—Loads output from a dataflow into a global cache, making the data available for lookup from the Query Cache stage. Using a global cache for data lookups improves performance compared to lookups to databases.
- **Write to DB**—Writes the output of a dataflow to a database.
- **Write to Spreadsheet**—Writes data to an Excel spreadsheet as output from a dataflow.
- **Write to Variable Format File**—Writes records of varying layouts to an output file.
- **Write to XML**—Writes the output of a job or subflow to an XML file.

Data Quality Bundle

Cleanse and validate data to improve accuracy with the Data Quality bundle. Improve the customer experience and drive efficiency in customer operations. Prevent bad data from being entered by checking and validating new information at the point of entry.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.

- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.
- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.
- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.

- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Universal Name Module

To perform the most accurate standardization you may need to break up strings of data into multiple fields. Spectrum™ Technology Platform provides advanced parsing features that enable you to parse personal names, company names, and many other terms and abbreviations. In addition, you can create your own list of custom terms to use as the basis of scan and extract operations.

Components

The Universal Name Module consists of:

- **Name Variant Finder**—This component works in either a first name or last name mode in order to query the names database and return name variations. Name Variant Finder can limit the number of results by gender and culture. Spectrum™ Technology Platform includes a base names file. Add-on names files of other cultures are available and can be deployed by copying the appropriate JAR file into the `modules/tables/ext` folder. Arabic names are available as an add-on. The data in this file is by Nomino, Inc.
- **Open Name Parser**—The open name parser breaks down personal and business names and other terms in the name data field into their component parts. These parsed name elements are then subsequently available to other automated operations such as name matching, name standardization, or multi-record name consolidation.

Data Science Bundle

The users of data science and analytics gain customer insights for the greatest return on marketing investment.

Data scientists can apply advanced analytics to understand customers, predict behaviors, and prescribe profitable next actions. Unlocking the potential of data science and analytics will enable you to gain greater insights into your customers, allowing you to spend your marketing budget more efficiently.

The Data Science Module allows you to:

- Easily access and transform data, both structured and unstructured
- Use machine-learning modeling algorithms
- Deploy models to customer engagement, geospatial analytics, and other production systems
- Leverage real-time web scoring services

Data Science includes:

- [Machine Learning Module](#) on page 20
- [Information Extraction Module](#) on page 21
- [Advanced Matching Module](#) on page 13
- Data integration and preparation provided by other Spectrum modules

Machine Learning Module

The Spectrum™ Technology Platform Machine Learning Module provides the ability to group (bin) numeric data and fit supervised and unsupervised machine learning model data in those models.

Note: The Machine Learning Module is supported only on Windows and Linux operating systems.

Note: The Machine Learning Module uses an underlying H2O.ai library for modeling algorithms in K-Means Clustering, Linear Regression, Logistic Regression, Principal Component Analysis, Random Forest Classification, and Random Forest Regression.

Binning

Binning divides records into groups (bins) for a continuous variable without taking into account objective information. You can perform unsupervised binning in one of two ways: using equal-width bins or equal-frequency bins.

K-Means Clustering

K-Means Clustering creates models based on analytical clustering, which segments a set of records into clusters of similar records based on data values.

Linear Regression

Linear Regression performs machine learning by creating models from datasets that use continuous objectives with input variables.

Logistic Regression

Logistic Regression creates models from datasets that use binary objectives with input variables.

Principal Component Analysis

Principal Component Analysis is a statistical process that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables known as principal components.

Random Forest Classification

Random Forest Classification performs machine learning by creating models from datasets that use continuous objectives with input variables.

Random Forest Regression

Random Forest Regression performs machine learning by creating models from datasets that use binary objectives with input variables.

Machine Learning Model Management

Machine Learning Model Management includes Model Assessment, which enables you to manage all machine learning models on your Spectrum™ Technology Platform server, and Binning Management, which enables you to manage all binning on your Spectrum™ Technology Platform server.

Information Extraction Module

The Information Extraction Module has capabilities of advanced text processing and information extraction from any natural language input text.

It has pre-trained models that are used to extract entities from an input text, determine relationship between the entities, and assign the category to which the text belongs.

Features Provided

Entity Extraction

This module extracts entities from an unstructured data and classifies it into types, such as **Location**, **Date**, **Organization**, **ProperNouns**, **Address** and **Person**.

The module ships with some *preexisting entities*. However, it also has the capability to train models based on your requirement. For details on training a model and defining custom entities, see **Custom Entities** in the User Guide.

Relationship Extraction

Relationship Extraction is the process of analysing the unstructured text to identify the relationship between the various extracted entities.

The entity types supported for relationship extraction are:

- Person
- Organisation
- Location

The supported relationship types are:

- AffiliatedWith
- LivesIn
- OrgBasedIn
- LocatedIn
- Negative

Text Categorization

Text categorization, also known as text classification, is the process of assigning custom categories to the unstructured content or plain text, such as email, news articles, and comments on the basis of how much of its content matches the category. Categorization can be done based on subject, author, date, or virtually any classification system defined.

You can create your own categorizer by training a categorizer model with your data and categories. The trainer analyzes the data and stores the information it gains in the training process. It then analyzes the content and determines the category to which the content belongs.

The text categorization feature uses statistical text categorization process. It applies machine learning methods to learn automatic classification rules that are based on human-labeled training documents.

Because you are able to apply the categorization of your choice, you first need to "train" your model to "learn" the categories. After this, you can use that model in the **Text Categorizer** stage to categorize your unstructured data.

Spectrum™ Technology Platform uses administration utility commands to manage text categorization models. For a description of these commands, see **Administration Utility** section of **Administration Guide**.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.
- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.

- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Global Sentry Bundle

The Global Sentry Module matches your records against government watch lists so that you can identify countries, individuals, and organizations in your data that have financial sanctions imposed on them by governments around the globe. The Global Sentry Module addresses the specific requirements of the USA PATRIOT Act and similar laws.

The Global Sentry Module identifies matches between your records and entities on the following government watch lists:

- Australian Government Department of Foreign Affairs and Trade
- Bank of England
- European Union - Common and Foreign Security Policy
- Office of Superintendent of Financial Institutions of Canada
- U.S. Bureau of Industry and Security - Denied Persons List
- U.S. Bureau of Industry and Security - Unverified List
- U.S. Treasury – Office of Foreign Assets Control
- U.S Department of State Directorate of Defense Trade Controls - List of Statutorily Debarred Parties

Global Sentry provides optimized rules for matching and scoring, along with a knowledge base of names from 120 cultures and associated gender and name variant data to identify matches between your data and the watch lists. Matches are performed against sanctioned country names, names, addresses, ID numbers, and other information such as date of birth to provide an overall risk score. This information allows your organization to make a decision on whether or not to block a particular transaction.

The Global Sentry Module uses stages from the Universal Name Module, Data Normalization Module, and Advanced Matching Module.

Universal Name Module

To perform the most accurate standardization you may need to break up strings of data into multiple fields. Spectrum™ Technology Platform provides advanced parsing features that enable you to parse personal names, company names, and many other terms and abbreviations. In addition, you can create your own list of custom terms to use as the basis of scan and extract operations.

Components

The Universal Name Module consists of:

- **Name Variant Finder**—This component works in either a first name or last name mode in order to query the names database and return name variations. Name Variant Finder can limit the number of results by gender and culture. Spectrum™ Technology Platform includes a base names file. Add-on names files of other cultures are available and can be deployed by copying the appropriate JAR file into the `modules/tables/ext` folder. Arabic names are available as an add-on. The data in this file is by Nomino, Inc.
- **Open Name Parser**—The open name parser breaks down personal and business names and other terms in the name data field into their component parts. These parsed name elements are then subsequently available to other automated operations such as name matching, name standardization, or multi-record name consolidation.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.
- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.
- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.
- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.
- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Entity Resolution Bundle

Understand relationships between people, products, locations, and more. Quickly identify an individual who may appear across multiple systems.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.
- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.
- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Business Steward Module

The Business Steward Module is a set of features that allow you to identify and resolve exception records. Exception records are records that Spectrum™ Technology Platform could not confidently process and that require manual review by a data steward. Some examples of exceptions are:

- Address verification failures
- Geocoding failures
- Low-confidence matches
- Merge/consolidation decisions

The Business Steward Module provides a browser-based tool for manually reviewing exception records. Once exception records are manually corrected and approved, you can add them back into your Spectrum™ Technology Platform data quality process.

The audit log shows many of the actions that take place within the Business Steward Module. The audit log is a Management Console tool that records user activity. The log may include:

- Adding exceptions in the Write Exceptions stage
- Deleting exceptions in the Read Exceptions stage and the Business Steward Portal Manage Exceptions page
- Assigning exceptions in the Business Steward Portal Exception Manager
- Read exceptions in the Business Steward Portal Exception Editor
- Updating exceptions in the Business Steward Portal Exception Editor
- Revalidating exceptions in the Business Steward Portal Exception Editor

Read more about the audit log in the Spectrum™ Technology Platform Administration Guide for the WebUI.

Components

The Business Steward Module consists of:

- **Exception Monitor**—A stage that evaluates records against a set of conditions to determine if the record requires manual review by a data steward. When records meet those conditions, this stage can send an email notifying recipients of the exceptions.
- **Write Exceptions**—A stage that writes the exception records to the exception repository. Once exception records are in the exception repository they are available for review by a data steward.
- **Exception Dashboard and Editor**—A browser-based tool that displays a dashboard of summary statistics and charts to help you understand the kinds of exceptions that occur in your data. It is also an editor that allows you to modify exception records and approve them for reprocessing.
- **Manage Exceptions**—A browser-based tool that enables you to review and manage exception record activity for all users.
- **Data Quality Performance**—A browser-based tool that provides information on trends within your exception records and enables you to identify key performance indicators and send notifications when certain conditions are met.
- **Read Exceptions**—A stage that reads approved or non-approved exceptions from the exception repository. This stage allows you to reprocess exception records that have been corrected by a data steward.

Data Hub Stages

Data Hub stages encapsulate Data Hub operations in dataflows. Dataflows are created in Enterprise Designer and executed with the Job Executor or Process Flow Executor client tools. These stages may also be incorporate in service dataflows, which can be accessed as web services or using the Spectrum™ Technology Platform API. For more information, see the *Data Flow Designer Guide*, which you can access on the Enterprise Designer **Help** menu.

Stage	Description
Delete from Hub	The Delete from Hub stage deletes entities and relationships from a model.

Stage	Description
Import to Hub	The Import to Hub stage defines a new model or populate an existing model from two incoming channels of data. It includes an optional outgoing error port that collects records not successfully processed by the dataflow.
Merge Entities	The Merge Entities stage is used to merge two or more entities into one.
Read from Hub	The Read From Hub stage uses a saved query or new query to read data from an existing model.
Query Hub	The Query Hub stage is an intermediate stage that uses incoming data rows to define queries that extract specific entities and relationships from a model.
Split Entity	The Split Entity stage splits one entity into two or more new entities.
Write to Hub	Write to Hub is a sink stage allows you to intuitively create a model using input data to define entities, relationships, and properties. Upon execution, Write to Hub loads the data into the hub.

Geocoding

Analyze geographic data and improve business decisions. Capabilities include:

Forward Geocoding

Enter an address and get outputs such as geographic coordinates, which can be used for detailed spatial analysis and demographics assignment. Capabilities include:

- Standardize Address according to country rules
- Cleanse Address
- Verify / Validate Address
- Match to street data or local geography
- Return detailed address and location codes for additional decision making
- Support for street and address point level
- Built-in address standardization and validation

Reverse Geocoding

Enter a geocode, a point represented by a latitude and longitude coordinate, and receive address information about the provided geocode. Reverse geocoding cascades from most accurate to least accurate possible locations, including point-level address, interpolated address, street centerline, etc.

Interactive Geocoding

Also described as type ahead, autocomplete, or autosuggest, it provides suggestions for users as they enter a query.

The Group By Area capability provides type area based on an area like postcode.

Note: the **GeoComplete Module** (GCM) has been deprecated. Interactive Geocoding is the recommended replacement.

Enterprise Geocoding Module

The Enterprise Geocoding Module performs address standardization, address geocoding, and postal code centroid geocoding. You can enter an address and get outputs such as geographic coordinates, which can be used for detailed spatial analysis and demographics assignment. You can also enter a geocode, a point represented by a latitude and longitude coordinate, and receive address information about the provided geocode.

Components

Enterprise Geocoding Module consists of the following stages. The specific stages you have depend on your license.

- **Geocode Global**—Takes an address in any supported country and returns latitude/longitude coordinates and other information. Geocode Address Global geocodes addresses only from countries you have licensed.
- **Geocode Address World**—Takes an address located in any of the supported countries and returns the city centroid or, for some countries, postal centroid. Geocode World cannot geocode to the street address level.
- **Geocode US Address**—Takes an input address and returns latitude/longitude coordinates and other address information.
- **GNAF PID Location Search**—Identifies the address and latitude/longitude coordinates for a Geocoded National Address File Persistent Identifier (G-NAF PID).
- **Reverse APN Lookup**—Takes an Assessor's Parcel Number (APN), Federal Information Processing Standards (FIPS) county code, and FIPS state code and returns the address of the parcel.
- **Reverse Geocode US Location**—Takes as input a geocode (latitude and longitude coordinate) and returns the address of the location.
- **Reverse Geocode Global** – for all countries except USA.
- **Reverse pbKey Lookup**—Takes a pbKey™ unique identifier as input and returns all standard returns that are provided as part of address matching.

Deprecated Components

The following stages have been deprecated:

- **Geocode Address AUS**— *Use the Global Geocoding Module's Global Geocoding stage.* Takes an address in Australia and returns latitude/longitude coordinates and other information. It also can geocode against a Geocoded National Address File Persistent Identifier point (G-NAF PID).
- **Geocode Address GBR**—*Use the Global Geocoding Module's Global Geocoding stage.* Takes an address in Great Britain and returns latitude/longitude coordinates and other information.

Global Geocoding Module

The Global Geocoding module offers a streamlined geocoding dataflow that is easy to set up and implement in Enterprise Designer. This allows you to integrate multi-country geocoding capabilities and functionality.

Capabilities include:

- Single and multiple line input search for street addresses
- Match mode is one of Standard, Close, or Relaxed
- Preference for geocoding to PO Box or Streets.
- Candidates returned in order based on distance from X/Y origin
- Search points of interest (POI)
- Search All to include both street addresses and POI
- Search POI ¹ categories and subcategories
- Search for brand (such as a specific hotel chain) in categories and subcategories
- Search within a bounding box
- Matched fields reported with returned candidates

Stages

Global Geocoding Module consists of the following stages.

- **Global Geocode** – ALL countries including USA and World
- **Global Reverse Geocode** – ALL countries including USA and World
- **Interactive Geocoding** – ALL countries including USA and World

GeoConfidence Module

The GeoConfidence Module is used to determine the probability that an address or street intersection is within a given area. The module takes an address or intersection's location (determined by Geocode US Address), converts that location to a point, line, or polygon (depending on the precision of the match), then compares that shape with a database of known shapes to see if the two overlap, and the percentage overlap. For example, you could use the GeoConfidence Module to make decisions on a flood zone rating based on how much overlap there is between an address's location

¹ POI and category/subcategory features are available only if you have licensed and installed POI data.

and the flood zone data. Anything greater than a 95% overlap with a 100-year flood zone may indicate that the address is in the flood zone. Conversely, anything less than 95% could cause your business process to send the address to exception processing that might include a manual review.

An address or intersection can be geocoded to a point, an address along a street segment (an array of street segment points), ZIP + 4 centroid, ZIP + 2 centroid, or ZIP Code centroid (polygons). You can use these shapes (points, lines, or polygons) to compare with other shapes to determine overlap, which can be used to determine a risk or probability.

Different geoconfidence polygons are generated depending on the GeoConfidence result returned by the Enterprise Geocoding Module. For more information about the GeoConfidence information returned by the Enterprise Geocoding Module, see the Enterprise Geocoding Module documentation.

The GeoConfidence Module supports U.S. locations only.

Note: GeoConfidence uses services provided by the Enterprise Geocoding and Spatial modules.

Stages

GeoConfidence deploys three dataflows that you can modify in Enterprise Designer. Each dataflow consists of various stages that were installed with the Enterprise Geocoding and Spatial modules.

For information about each stage in the installed dataflows, see the relevant component chapter in the *Spectrum™ Technology Platform User's Guide*.

The names of the dataflows are:

- **GeoConfidenceSurface** This is the dataflow that creates the geoconfidence surface that can be used for further analysis. The input is the GeoConfidence information that is returned from the Enterprise Geocoding Module. Currently, only the Geocode US Address stage can return this information.
- **CreatePointsConvexHull** This is a subflow that is used by the GeoConfidenceSurface template. You should not need to make any changes to this subflow.
- **FloodRiskAnalysis** This is an example dataflow.

Pitney Bowes Geocoding Connector

Previously referred to as "PBLocator", the Pitney Bowes Geocoding Connector allows customers to integrate Pitney Bowes geocoding within third-party systems such as ArcGIS™ Online or ArcGIS™ Pro Desktop.

You'll need one of these Pitney Bowes geocoding solutions to generate our geocodes:

- Global Geocoding Software Developer Kit (SDK)
- Spectrum Global Geocoding Module
- Location Intelligence GeoCode API (available at locate.pitneybowes.com)

Geocoding Datasets

Datasets are dynamic and information about what's new, datasets by country, and bug fixes are summarized in our Data Release Notes. Please refer to the [Data Release Notes](#) area of our Support site for the most current information.

GeoEnrichment Module

Learn an address's ownership, demographics, consumer spending, telecommunication information, and more.

Information Extraction Module

The Information Extraction Module has capabilities of advanced text processing and information extraction from any natural language input text.

It has pre-trained models that are used to extract entities from an input text, determine relationship between the entities, and assign the category to which the text belongs.

Features Provided

Entity Extraction

This module extracts entities from an unstructured data and classifies it into types, such as **Location**, **Date**, **Organization**, **ProperNouns**, **Address** and **Person**.

The module ships with some *preexisting entities*. However, it also has the capability to train models based on your requirement. For details on training a model and defining custom entities, see [Custom Entities](#) in the User Guide.

Relationship Extraction

Relationship Extraction is the process of analysing the unstructured text to identify the relationship between the various extracted entities.

The entity types supported for relationship extraction are:

- Person
- Organisation
- Location

The supported relationship types are:

- AffiliatedWith
- LivesIn
- OrgBasedIn

- LocatedIn
- Negative

Text Categorization

Text categorization, also known as text classification, is the process of assigning custom categories to the unstructured content or plain text, such as email, news articles, and comments on the basis of how much of its content matches the category. Categorization can be done based on subject, author, date, or virtually any classification system defined.

You can create your own categorizer by training a categorizer model with your data and categories. The trainer analyzes the data and stores the information it gains in the training process. It then analyzes the content and determines the category to which the content belongs.

The text categorization feature uses statistical text categorization process. It applies machine learning methods to learn automatic classification rules that are based on human-labeled training documents.

Because you are able to apply the categorization of your choice, you first need to "train" your model to "learn" the categories. After this, you can use that model in the **Text Categorizer** stage to categorize your unstructured data.

Spectrum™ Technology Platform uses administration utility commands to manage text categorization models. For a description of these commands, see **Administration Utility** section of **Administration Guide**.

Information Stewardship Bundle

Understand the state of your data using a dashboard linked to key performance indicators. Deal with exceptions that occur as a result of data processing.

Business Steward Module

The Business Steward Module is a set of features that allow you to identify and resolve exception records. Exception records are records that Spectrum™ Technology Platform could not confidently process and that require manual review by a data steward. Some examples of exceptions are:

- Address verification failures
- Geocoding failures
- Low-confidence matches
- Merge/consolidation decisions

The Business Steward Module provides a browser-based tool for manually reviewing exception records. Once exception records are manually corrected and approved, you can add them back into your Spectrum™ Technology Platform data quality process.

The audit log shows many of the actions that take place within the Business Steward Module. The audit log is a Management Console tool that records user activity. The log may include:

- Adding exceptions in the Write Exceptions stage
- Deleting exceptions in the Read Exceptions stage and the Business Steward Portal Manage Exceptions page
- Assigning exceptions in the Business Steward Portal Exception Manager
- Read exceptions in the Business Steward Portal Exception Editor
- Updating exceptions in the Business Steward Portal Exception Editor
- Revalidating exceptions in the Business Steward Portal Exception Editor

Read more about the audit log in the Spectrum™ Technology Platform Administration Guide for the WebUI.

Components

The Business Steward Module consists of:

- **Exception Monitor**—A stage that evaluates records against a set of conditions to determine if the record requires manual review by a data steward. When records meet those conditions, this stage can send an email notifying recipients of the exceptions.
- **Write Exceptions**—A stage that writes the exception records to the exception repository. Once exception records are in the exception repository they are available for review by a data steward.
- **Exception Dashboard and Editor**—A browser-based tool that displays a dashboard of summary statistics and charts to help you understand the kinds of exceptions that occur in your data. It is also an editor that allows you to modify exception records and approve them for reprocessing.
- **Manage Exceptions**—A browser-based tool that enables you to review and manage exception record activity for all users.
- **Data Quality Performance**—A browser-based tool that provides information on trends within your exception records and enables you to identify key performance indicators and send notifications when certain conditions are met.
- **Read Exceptions**—A stage that reads approved or non-approved exceptions from the exception repository. This stage allows you to reprocess exception records that have been corrected by a data steward.

Enterprise Data Integration Module

The Enterprise Data Integration Module provides the features you need to connect to data in multiple sources either directly or through integration with your existing data access technologies. With the Enterprise Data Integration Module, you can use Spectrum™ Technology Platform for data warehousing, data quality, systems integration, and migration.

Components

The Enterprise Data Integration Module consists of:

- **Call Stored Procedure**—A source stage that executes a stored procedure in a database, and returns the results of the stored procedure call as input for the dataflow. Use Call Stored Procedure when you want to get data from a database using a database's stored procedure rather than a query to a table or view.

- **Field Selector**—Allows you to choose which fields to pass to the next stage in the dataflow. You can use Field Selector to remove unwanted fields from a dataflow. For example, if you have created a new field by combining the data from two fields, and you no longer need the two source fields, you can use Field Selector to retain only the new field and remove the two source fields from the dataflow.
- **Generate Time Dimension**—Creates date records, one for each day of the date range you specify. You can then write these records to a time dimension table in a database using the Write to DB stage. The time dimension table can then be used to perform accurate calculations based on a period. For example, sales by quarter, budget spend by quarter, and revenue by day are all analyses that require a time dimension. Time dimension tables also enable you to account for fiscal years or non-standard quarters in the analysis.
- **Query Cache**—Looks up data in a cache based on values in one or more dataflow fields and returns data from matching records in the cache, adding the cache record's data to the record in the dataflow. Looking up data in a cache can improve performance compared to looking up data in a database.
- **Read From DB**—Reads data from a database table or view as input to a dataflow. Read From DB is available for jobs and subflows but not services.
- **Read From HL7**—Reads Health Level Seven (HL7) data from a text file as input to a dataflow. HL7 is a messaging standard used in the healthcare industry to exchange data between systems. For more information about HL7, review the details at www.hl7.org.
- **Read from Spreadsheet**—Reads data from an Excel spreadsheet as input to a dataflow. It supports both .xls and .xlsx file formats.
- **Read From Variable Format File**—Reads data from a file containing records of varying layout, reading each record as a list field. You can specify the tag that indicates the parent record type, and all other record types will become list fields under the parent.
- **Read From XML**—Reads an XML file into a job or subflow. It defines the file's path and data format, including XML schema and data element details.
- **Transposer**—Converts columns to rows. Transposing data is the opposite of pivoting data using the Group Statistics stage, which transforms row data into columns.
- **Write to Cache**—Loads output from a dataflow into a global cache, making the data available for lookup from the Query Cache stage. Using a global cache for data lookups improves performance compared to lookups to databases.
- **Write to DB**—Writes the output of a dataflow to a database.
- **Write to Spreadsheet**—Writes data to an Excel spreadsheet as output from a dataflow.
- **Write to Variable Format File**—Writes records of varying layouts to an output file.
- **Write to XML**—Writes the output of a job or subflow to an XML file.

Master Data Management Bundle

Create a consolidated view of business data from multiple sources with the Master Data Management bundle. With a consolidated view of business data, your marketing department can gain new insights on individual customers. Your operational teams can create an aggregated view of information

captured from multiple channels, and present it to front line teams at the moment they engage with the customer. Your compliance teams can build a complete view of all business operations with a single entity to comply with industry regulation and defend against fraudulent activities.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.
- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.
- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Business Steward Module

The Business Steward Module is a set of features that allow you to identify and resolve exception records. Exception records are records that Spectrum™ Technology Platform could not confidently process and that require manual review by a data steward. Some examples of exceptions are:

- Address verification failures
- Geocoding failures
- Low-confidence matches
- Merge/consolidation decisions

The Business Steward Module provides a browser-based tool for manually reviewing exception records. Once exception records are manually corrected and approved, you can add them back into your Spectrum™ Technology Platform data quality process.

The audit log shows many of the actions that take place within the Business Steward Module. The audit log is a Management Console tool that records user activity. The log may include:

- Adding exceptions in the Write Exceptions stage
- Deleting exceptions in the Read Exceptions stage and the Business Steward Portal Manage Exceptions page
- Assigning exceptions in the Business Steward Portal Exception Manager
- Read exceptions in the Business Steward Portal Exception Editor
- Updating exceptions in the Business Steward Portal Exception Editor
- Revalidating exceptions in the Business Steward Portal Exception Editor

Read more about the audit log in the Spectrum™ Technology Platform Administration Guide for the WebUI.

Components

The Business Steward Module consists of:

- **Exception Monitor**—A stage that evaluates records against a set of conditions to determine if the record requires manual review by a data steward. When records meet those conditions, this stage can send an email notifying recipients of the exceptions.
- **Write Exceptions**—A stage that writes the exception records to the exception repository. Once exception records are in the exception repository they are available for review by a data steward.
- **Exception Dashboard and Editor**—A browser-based tool that displays a dashboard of summary statistics and charts to help you understand the kinds of exceptions that occur in your data. It is also an editor that allows you to modify exception records and approve them for reprocessing.
- **Manage Exceptions**—A browser-based tool that enables you to review and manage exception record activity for all users.
- **Data Quality Performance**—A browser-based tool that provides information on trends within your exception records and enables you to identify key performance indicators and send notifications when certain conditions are met.

- **Read Exceptions**—A stage that reads approved or non-approved exceptions from the exception repository. This stage allows you to reprocess exception records that have been corrected by a data steward.

Data Hub Module

Data Hub Module identifies and measures relationships between people, groups of people, or almost any other type of entity. It can use your data to identify hidden, latent, or non-obvious connections among people, groups, and organizations. It can also identify shifting patterns of connectedness, distance, centrality, influence status, and importance. With this information, you can predict customer influence and impact and categorize entities as leaders, followers, influencers, outliers, or late adopters. You can also predict churn and offer more personalized customer service by anticipating customer needs and optimizing offerings. You can even use it to detect and prevent fraud, crime, and terrorism.

Data Hub models represent the entities and relationships used to identify connections among people, groups, and organizations, as well as patterns of connectedness, distance, centrality, influence, status, and importance. Data Hub models may be created and edited in [Metadata Insights](#). The Data Hub module itself provides several different client interfaces that may be used to work with Data Hub models:

Data Hub Visualization	This relationship discovery tool presents data in a graphical form, showing the data structure and the relationships among the entities. Users can analyze data visually to gain insights they may miss using traditional table-based numeric analysis.
Data Hub Browser	This discovery tool allows users to search the contents of models by entering natural-language queries and browsing the results.
Data Hub Stages	Data Hub stages encapsulate Data Hub operations in dataflows. Dataflows are created in Enterprise Designer and executed with the Job Executer or Process Flow Executer client tools. These stages may also be incorporated in service dataflows, which can be accessed as web services or using the Spectrum™ Technology Platform API.
Relationship Analysis Client	Used to view data in a graphical manner that identifies and depicts structure and relationships between entities. You can also use the tool to manipulate entities and relationships based on their properties and values.
Command Line Interface	The Administration Utility Command Line Interface (CLI) provides Data Hub commands to manage models, view jobs and job status, run centrality algorithms, and manage model schemas.

Data Hub Stages

Data Hub stages encapsulate Data Hub operations in dataflows. Dataflows are created in Enterprise Designer and executed with the Job Executer or Process Flow Executer client tools. These stages may also be incorporate in service dataflows, which can be accessed as web services or using the Spectrum™ Technology Platform API.

Stage	Description
Delete from Hub	The Delete from Hub stage deletes entities and relationships from a model.
Import to Hub	The Import to Hub stage defines a new model or populate an existing model from two incoming channels of data. It includes an optional outgoing error port that collects records not successfully processed by the dataflow.
Merge Entities	The Merge Entities stage is used to merge two or more entities into one.
Read from Hub	The Read From Hub stage uses a saved query or new query to read data from an existing model.
Query Hub	The Query Hub stage is an intermediate stage that uses incoming data rows to define queries that extract specific entities and relationships from a model.
Split Entity	The Split Entity stage splits one entity into two or more new entities.
Write to Hub	Write to Hub is a sink stage allows you to intuitively create a model using input data to define entities, relationships, and properties. Upon execution, Write to Hub loads the data into the hub.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.
- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table

Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.

- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Enterprise Data Integration Module

The Enterprise Data Integration Module provides the features you need to connect to data in multiple sources either directly or through integration with your existing data access technologies. With the Enterprise Data Integration Module, you can use Spectrum™ Technology Platform for data warehousing, data quality, systems integration, and migration.

Components

The Enterprise Data Integration Module consists of:

- **Call Stored Procedure**—A source stage that executes a stored procedure in a database, and returns the results of the stored procedure call as input for the dataflow. Use Call Stored Procedure when you want to get data from a database using a database's stored procedure rather than a query to a table or view.
- **Field Selector**—Allows you to choose which fields to pass to the next stage in the dataflow. You can use Field Selector to remove unwanted fields from a dataflow. For example, if you have created a new field by combining the data from two fields, and you no longer need the two source fields, you can use Field Selector to retain only the new field and remove the two source fields from the dataflow.
- **Generate Time Dimension**—Creates date records, one for each day of the date range you specify. You can then write these records to a time dimension table in a database using the Write to DB stage. The time dimension table can then be used to perform accurate calculations based on a period. For example, sales by quarter, budget spend by quarter, and revenue by day are all analyses that require a time dimension. Time dimension tables also enable you to account for fiscal years or non-standard quarters in the analysis.
- **Query Cache**—Looks up data in a cache based on values in one or more dataflow fields and returns data from matching records in the cache, adding the cache record's data to the record in the dataflow. Looking up data in a cache can improve performance compared to looking up data in a database.
- **Read From DB**—Reads data from a database table or view as input to a dataflow. Read From DB is available for jobs and subflows but not services.
- **Read From HL7**—Reads Health Level Seven (HL7) data from a text file as input to a dataflow. HL7 is a messaging standard used in the healthcare industry to exchange data between systems. For more information about HL7, review the details at www.hl7.org.
- **Read from Spreadsheet**—Reads data from an Excel spreadsheet as input to a dataflow. It supports both .xls and .xlsx file formats.
- **Read From Variable Format File**—Reads data from a file containing records of varying layout, reading each record as a list field. You can specify the tag that indicates the parent record type, and all other record types will become list fields under the parent.
- **Read From XML**—Reads an XML file into a job or subflow. It defines the file's path and data format, including XML schema and data element details.

- **Transposer**—Converts columns to rows. Transposing data is the opposite of pivoting data using the Group Statistics stage, which transforms row data into columns.
- **Write to Cache**—Loads output from a dataflow into a global cache, making the data available for lookup from the Query Cache stage. Using a global cache for data lookups improves performance compared to lookups to databases.
- **Write to DB**—Writes the output of a dataflow to a database.
- **Write to Spreadsheet**—Writes data to an Excel spreadsheet as output from a dataflow.
- **Write to Variable Format File**—Writes records of varying layouts to an output file.
- **Write to XML**—Writes the output of a job or subflow to an XML file.

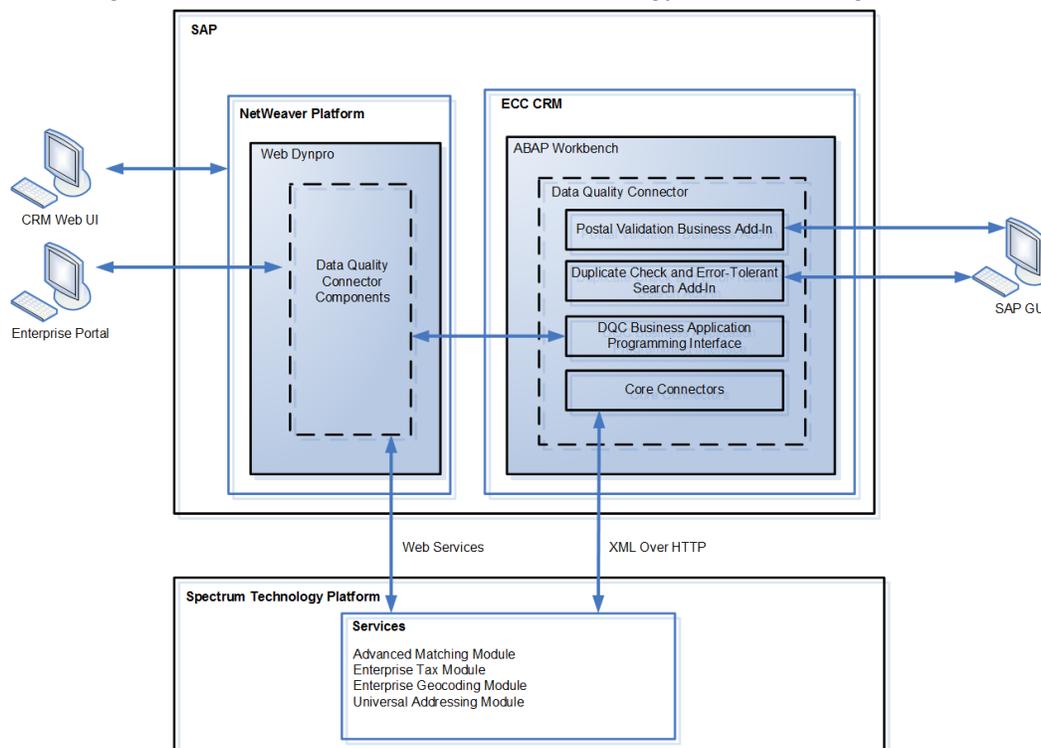
Microsoft Dynamics CRM Module

SAP Module

The Spectrum™ Technology Platform SAP Module optimizes the quality of your customer, vendor and partner records so you can improve operations across the enterprise.

Because SAP is such a far-reaching application, introducing accurate address data provides equally far-reaching benefits. With data quality for over 220 countries, the SAP Module identifies and manages duplicate records, standardizes and validates addresses, auto-populates missing fields, and gives you a single, integrated view of your customers.

This diagram illustrates how Spectrum™ Technology Platform integrates with SAP.

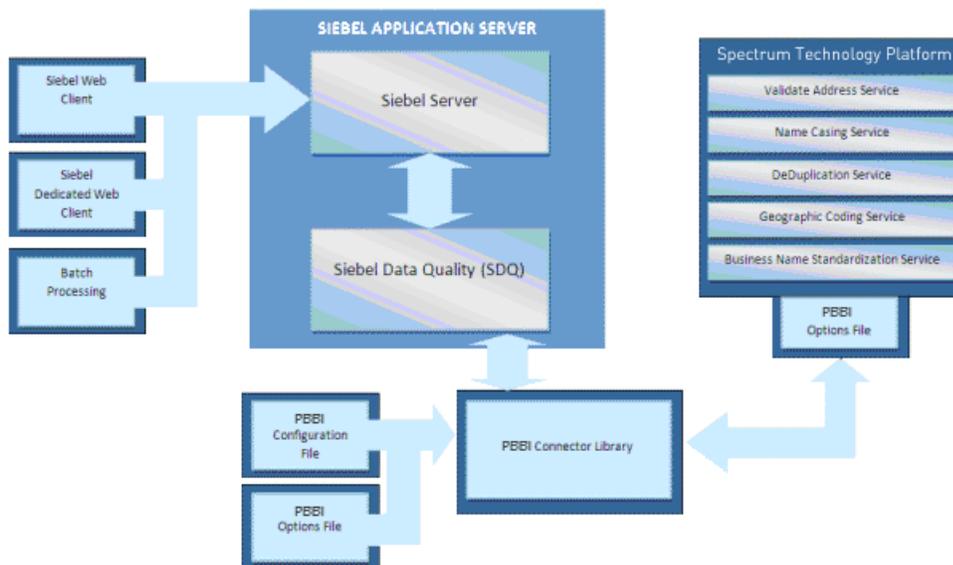


Siebel Module

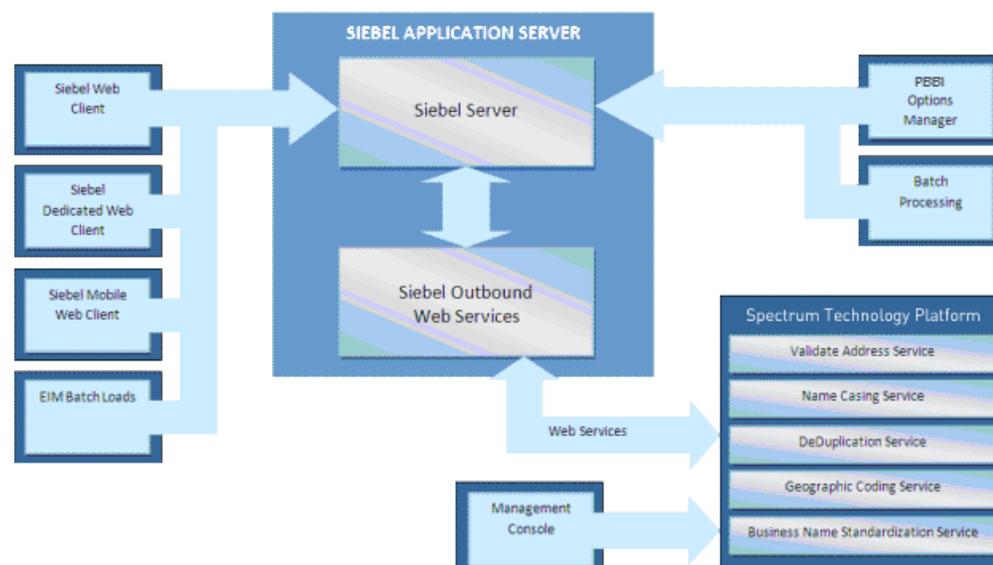
The Spectrum™ Technology Platform Siebel Module identifies and manages duplicate records, standardizes and validates addresses, standardizes and validates names, and auto-populates missing fields. Records in your Siebel system are checked against known, up-to-date reference data from sources such as regulatory bodies (for example, the United States Postal Service), third-party data providers (for example, Dun & Bradstreet) or your company's internal reference sources (for example, accounting data).

The Siebel Module is available in two implementations: Siebel Data Quality (SDQ) and Non-SDQ. The SDQ implementation uses the Siebel Data Quality Universal Connector to connect Siebel to Spectrum™ Technology Platform. The Non-SDQ implementation utilizes a custom connector to connect Siebel to Spectrum™ Technology Platform.

SDQ Architecture



Non-SDQ Architecture



The following table illustrates the differences between SDQ and Non-SDQ implementation.

Table 1: Supported Web Clients

	SDQ	Non-SDQ
Web Client	Yes	Yes
Dedicated Web Client	Yes	Yes
Mobile Web Client	No	Yes
Open UI	Yes	Yes

Data cleansing may be enabled for business addresses, personal addresses, and prospect addresses. De-duplication may be enabled for accounts, contacts, business addresses, personal addresses², and prospects. Settings are global for interactive mode and can be unique for each batch job. Settings for the Siebel Module are managed using options and configuration files for SDQ. For Non-SDQ, Siebel's Options Manager and the Spectrum™ Technology Platform Management Console are used to set options and configuration.

² Personal Address is supported only for Siebel Industry Applications

Universal Addressing Module

The Universal Addressing Module is an address quality module that can standardize and validate addresses, improving the deliverability of mail. The Universal Addressing Module can ensure that your address data adheres to quality standards established by the postal authority. An address that adheres to these standards is more likely to be delivered in a timely manner. In addition, mailers who follow these standards can qualify for significant postage discounts. For information on discounts for U.S. mail, refer to the USPS *Domestic Mail Manual (DMM)* available at <https://postalpro.usps.com/>. For information on discounts for Canadian mail, refer to the Canada Post website at www.canadapost.ca. For information on discounts for Australian mail, refer to the Australia Post website at www.auspost.com.au.

You can use the Universal Addressing Module in batch mode, real-time mode, or as a hosted service, depending on which option you have licensed. The batch version of the Universal Addressing Module is CASS Certified™ by the USPS®.

Components

The Universal Addressing Module consists of the following components. These components can work with U.S., Canadian, Australian, and international addresses as long as you have licensed the appropriate database (if you are running Universal Addressing in your environment) or hosted service (if you are using Universal Addressing through the Pitney Bowes hosted services).

- **Auto Complete Loqate**— Offers real-time entry of address data and returns instant results based on each character entered into the form, ensuring only accurate data is entered into the database.
- **Get Candidate Addresses**— Returns a list of possible matches for a given address.
- **Get Candidate Addresses Loqate**— Returns a list of possible matches for a given address using a Loqate engine and database.
- **Get City State Province**— Returns the city and state or province for a given postal code.
- **Get City State Province Loqate**— Returns the city and state or province for a given postal code using a Loqate engine and database.
- **Get Postal Codes**— Returns the postal codes for a given city.
- **Get Postal Codes Loqate**—Returns the postal codes for a given city using a Loqate engine and database.
- **Validate Address**—Standardizes and validates addresses using U.S., Canadian, and international postal data.
- **Validate Address Global**— Provides enhanced address standardization and validation for addresses outside the U.S. and Canada. Validate Address Global can also validate addresses in the U.S. and Canada, but its strength is in validating addresses in other countries. If you process a significant number of addresses outside the U.S. and Canada, consider using Validate Address Global.
- **Validate Address Loqate**— Standardizes and validates addresses using postal authority address data. Validate Address Loqate can correct information and format addresses using the format preferred by an applicable postal authority. It also adds missing postal information, such as postal codes, city names, and state or province names.

Universal Name Module

To perform the most accurate standardization you may need to break up strings of data into multiple fields. Spectrum™ Technology Platform provides advanced parsing features that enable you to parse personal names, company names, and many other terms and abbreviations. In addition, you can create your own list of custom terms to use as the basis of scan and extract operations.

Components

The Universal Name Module consists of:

- **Name Variant Finder**—This component works in either a first name or last name mode in order to query the names database and return name variations. Name Variant Finder can limit the number of results by gender and culture. Spectrum™ Technology Platform includes a base names file. Add-on names files of other cultures are available and can be deployed by copying the appropriate JAR file into the `modules/tables/ext` folder. Arabic names are available as an add-on. The data in this file is by Nomino, Inc.
- **Open Name Parser**—The open name parser breaks down personal and business names and other terms in the name data field into their component parts. These parsed name elements are then subsequently available to other automated operations such as name matching, name standardization, or multi-record name consolidation.

Metadata Insights Module

Metadata Insights gives you the control you need to deliver accurate and timely data-driven insights to your business. It presents a single view of data residing in disparate sources and allows you seamless movement among modules till task completion.

Metadata Insights empowers you to build a robust Data Management Strategy and provides the foundational technology for solving single view use cases. Capabilities around Business Glossary , Discovery, Profiling, and Lineage enables data democratization and management of data as a strategic asset, thus creating new revenue streams for the organizations.

The insight gained from this application facilitates:

- Identification of data resources for answering particular business questions
- Process optimization to improve usefulness and consistency of data across the enterprise
- Troubleshooting data issues.

These are the six modules of Metadata Insights that work together seamlessly to provide you one of the best data management experiences.

Connect

This module helps you connect to your data that might be residing on disparate sources, such as Salesforce, Apache Cassandra, Hadoop, Dynamo db, SQL server, as well as CSV files. Once you

establish the connection successfully, you can access these data in the **Discover**, **Model**, and **Profile** modules of Metadata Insights as well as the various stages of **Enterprise designer**.

Discover

The Discover module gives you a unified view of all the connections you have configured using the **Connect** module. Use this module to:

- Discover all the assets in your connections
- Catalog and classify the discovered assets
- Tag your assets for an easy access during modeling and profiling
- Filter assets based on the tags, semantic type, data type, connection type or connection name.

Business Glossary

For data to make sense in an organizational context, it needs to be consistent, contextual, and transparent. Users through the organization need to have a common understanding of the definition and lineage of the data they are using. And what can be a better way to achieve this than arriving at the definitions, policies, and rules governing data collaboratively through work flows and multilevel approvals.

The **Define Glossary** module allows you to do all this with an intuitive, user-friendly interface.

- **Ease of use:** Business Glossary helps you search for the required data without having to master the tables, column names, fields, and metadata.
- **Transparency:** The data comes with its full context and lineage, so you can trust it without an iota of doubt.
- **Removes conflicts, confusions:** You have a comprehensive view of all the business terms with the related data and metadata. This avoids all the conflicts in terms of meaning and its usage.
- **Swift, powerful, simple search:** The advanced search features make searching swift, simple, and powerful. You can use any aspect of the data including relationship, tags, and attribute to search it.

These powerful features of Define Glossary make it most suitable for your business needs if you deal with data, and inconsistency in its definition across the enterprise is hindering you. If you are not able to find the right data, at the right time, and reconciling the differences in data definition in various departments is making life chaotic for you, ultimately eroding the trust in your data.

Model

The Modeling view is where you create physical and logical data models and deploy those into a model store, thus creating a layer of abstraction over the underlying data sources on the Spectrum™ Technology Platform server.

Producing data-driven insights requires that technical teams and business teams have a common understanding of the organization's data assets and how those assets will be used to support business decisions. Technical teams understand the design of databases while business teams understand the business objects (such as customer, store, or vendor) that are of interest. Metadata Insights helps bridge this gap by providing physical modeling and logical modeling tools that are

visually rich and independent of each other, enabling you to create both a technical view of data assets and a business view of objects of interest, and to link the two through mapping.

Physical Model

A *physical model* organizes your organization's data assets in a meaningful way. A physical model makes it possible to pull data from individual tables, columns, and views to create a single resource that you can then use to supply data to logical models or to perform profiling.

Logical Model

A *logical model* defines the objects that your business is interested in and the attributes of those objects, as well as how objects are related to each other. For example, a logical model for a customer might contain attributes for name and date of birth. It might also have a relationship to a home address object, which contains attributes for address lines, city, and postal code. Once you have defined the attributes of the objects your business is interested in, you can map physical data sources to the logical model's attributes, thereby identifying the specific data asset that will be used to populate that attribute.

Mapping

The built-in suggestion capabilities of Metadata Insights make mapping of logical to physical models an intuitive, effortless job. During the mapping process, the system suggests all the discovered tables and columns that map with the selected logical entities. For the suggested assets the semantic and data type are also displayed giving you the much needed context.

Data Hub Models

A Data Hub model represents entities and relationships. It is used to identify connections among people, groups, and organizations, as well as patterns of connectedness, distance, centrality, influence, status, and importance. In Metadata Insights, you can create an entirely new Data Hub model, edit an existing Data Hub model, or use a logical model as a template for a new Data Hub model. You can then use Metadata Insights to add and edit entities, relationships, and their properties, including virtual entities and relationships representing data stored in external databases.

Profile

Successful decision making is heavily dependent on reliable, correct, complete, and valid data. Incomplete records, malformed fields, and a lack of context can result in misleading or inaccurate data being delivered to your business users, which can result in flawed decisions.

Data profiling scans your data records from all the data sources - irrespective of its volume and complexity. It identifies problems related to correctness, completeness, and validity in the data, and suggests actions to fix the issues. Thus, it improves the quality and utility of your data with very little manual effort.

Benefits of data profiling

- It is the first step in analyzing your data and predicting how much effort is needed to make it usable
- It improves your trust on the data set you have
- It is one of the mandatory steps for taking control of your organizational data and using it across the enterprise.

Scorecard

Monitoring is one of the indispensable aspects of data governance and good decision making. Constant monitoring of data quality helps you to take measures to improve the consistency, reliability, and accuracy of your data. Scorecards help you measure the quality of your data and assign score cards to it by tracking key matrices defined by you. It represents your data health in a graphical form, making assessment all the more easy and swift for you.

Trends

Trends help you measure the improvement in the quality of your data over a period of time. For example, in a merger between two companies (Company A and Company B), company A needs to migrate customer data in its system from company B. It will run a scorecard on the data at the beginning of the process to see the quality and may thereafter do this scoring after every cleansing. Trends plot a chart based on your recent runs or for a specified date range to give you a clearer picture on the improvement index.

Note: You need to have the **Data Integration** module installed in order to use the profiling feature of **Metadata Insights**.

Lineage and Impact Analysis

The Lineage and Impact Analysis view shows how data flows from data sources to data destinations and through Spectrum™ Technology Platform flows. Lineage and impact analysis are similar concepts that describe different ways of tracing the flow of data.

Lineage shows where data comes from. You can use it to trace the path of data back to its source, showing all the systems that process and store the data along the way, such as Spectrum™ Technology Platform flows, databases, and files.

Impact analysis shows where data goes and the systems that depend on data from a selected data resource. You can use it to view the flows, databases, and files that use a data resource directly or indirectly. Looking at impact analysis is useful if you want to understand how a modification to a database, file, or flow will affect the processes and systems that use the data.

Metadata Insights can show lineage and impact analysis in a single diagram that shows the complete flow of data from source to destination. You can also choose to view lineage only or impact only. By viewing data lineage and impact analysis together you can pinpoint issues in your data processes and plan for upgrades and modifications to your data processes.

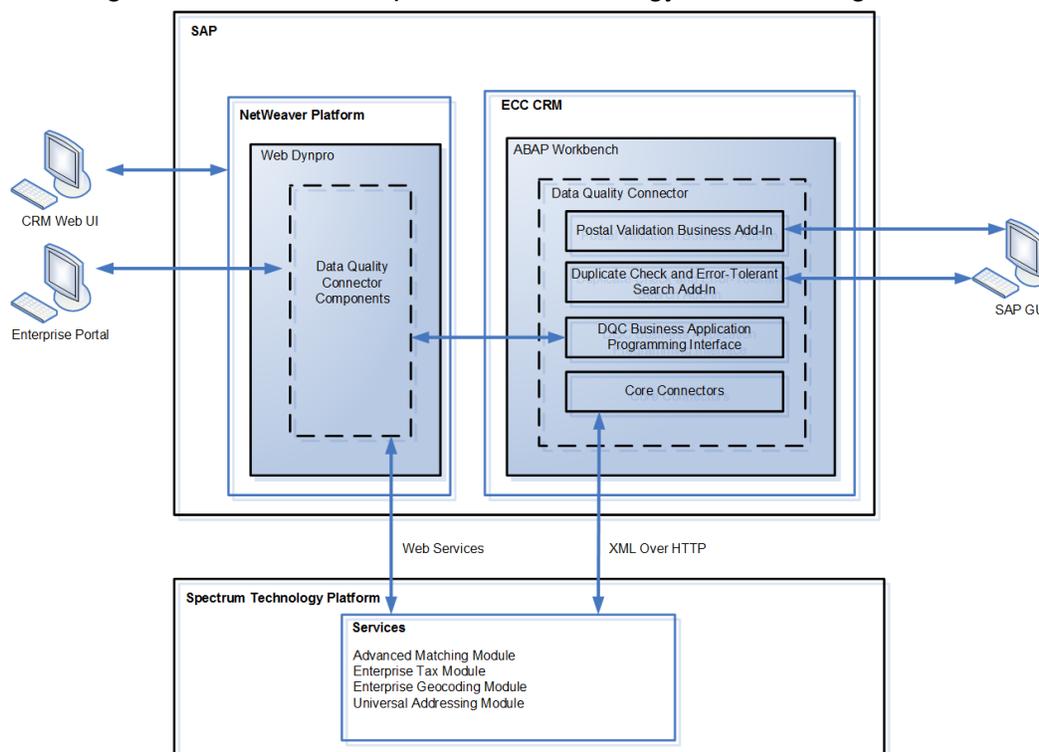
Microsoft Dynamics CRM Module

SAP Module

The Spectrum™ Technology Platform SAP Module optimizes the quality of your customer, vendor and partner records so you can improve operations across the enterprise.

Because SAP is such a far-reaching application, introducing accurate address data provides equally far-reaching benefits. With data quality for over 220 countries, the SAP Module identifies and manages duplicate records, standardizes and validates addresses, auto-populates missing fields, and gives you a single, integrated view of your customers.

This diagram illustrates how Spectrum™ Technology Platform integrates with SAP.



Screeener Solutions Bundle

Screeener manages and versions both internal and external lists, along with their list entries. Screeener can either screen parties (customers and prospects) against those lists or Pitney Bowes services can disseminate lists to a screening system.

Advanced Matching Module

The Advanced Matching Module matches records between and within any number of input files. You can also use the Advanced Matching Module to match on a variety of fields including name, address, name and address, or non-name and address fields, such as social security number or date of birth.

Components

The Advanced Matching Module consists of:

- **Best-of-Breed**—This stage selects a best-of-breed record from the duplicates cluster by selecting a template record, then using that record to build a composite record that becomes the survivor record.
- **Candidate Finder**—This stage finds candidate records that will form the set of potential matches that the Transactional Match stage will evaluate. Candidate Finder is used with Transactional Match.
- **Duplicate Synchronization**—This stage allows you to specify fields from a collection of records to copy (post) to the corresponding fields of all records in the collection.
- **Filter**—This stage allows you to specify the criteria that records must satisfy to be retained or removed from a collection of duplicate records, either for further processing downstream or for your output file.
- **Interflow Match**—This stage identifies matches between similar data records across multiple input streams.
- **Intraflow Match**—This stage identifies matches between similar data records within a single input stream.
- **Match Analysis**—Use Match Analysis to analyze and compare results from matching stages to understand the outcome of the matching process and improve overall matching results.
- **Match Key Generator**—This stage creates a non-unique key shared by all like records. You have the option of grouping records that share the same key for comparison.
- **Private Match**—This stage enables two entities to compare datasets and identify common records, or customers, without compromising sensitive information. Special encryption helps to retain security during matching, and a decryption function shows the output of the matched data.
- **Transactional Match**—This stage matches suspect transactions against a database using the Candidate Finder Stage to query and return potential candidate records.
- **Write to Search Index**—This stage enables you to create a full-text index based on the data that the stage is accepting for processing. You can retain data in a dedicated search index to improve response times when searching the Candidate Finder.

Business Steward Module

The Business Steward Module is a set of features that allow you to identify and resolve exception records. Exception records are records that Spectrum™ Technology Platform could not confidently process and that require manual review by a data steward. Some examples of exceptions are:

- Address verification failures

- Geocoding failures
- Low-confidence matches
- Merge/consolidation decisions

The Business Steward Module provides a browser-based tool for manually reviewing exception records. Once exception records are manually corrected and approved, you can add them back into your Spectrum™ Technology Platform data quality process.

The audit log shows many of the actions that take place within the Business Steward Module. The audit log is a Management Console tool that records user activity. The log may include:

- Adding exceptions in the Write Exceptions stage
- Deleting exceptions in the Read Exceptions stage and the Business Steward Portal Manage Exceptions page
- Assigning exceptions in the Business Steward Portal Exception Manager
- Read exceptions in the Business Steward Portal Exception Editor
- Updating exceptions in the Business Steward Portal Exception Editor
- Revalidating exceptions in the Business Steward Portal Exception Editor

Read more about the audit log in the Spectrum™ Technology Platform Administration Guide for the WebUI.

Components

The Business Steward Module consists of:

- **Exception Monitor**—A stage that evaluates records against a set of conditions to determine if the record requires manual review by a data steward. When records meet those conditions, this stage can send an email notifying recipients of the exceptions.
- **Write Exceptions**—A stage that writes the exception records to the exception repository. Once exception records are in the exception repository they are available for review by a data steward.
- **Exception Dashboard and Editor**—A browser-based tool that displays a dashboard of summary statistics and charts to help you understand the kinds of exceptions that occur in your data. It is also an editor that allows you to modify exception records and approve them for reprocessing.
- **Manage Exceptions**—A browser-based tool that enables you to review and manage exception record activity for all users.
- **Data Quality Performance**—A browser-based tool that provides information on trends within your exception records and enables you to identify key performance indicators and send notifications when certain conditions are met.
- **Read Exceptions**—A stage that reads approved or non-approved exceptions from the exception repository. This stage allows you to reprocess exception records that have been corrected by a data steward.

Data Normalization Module

The Data Normalization Module examines terms in a record and determines if the term is in the preferred form.

Components

The Data Normalization Module consists of:

- **Advanced Transformer**—This stage scans and splits strings of data into multiple fields, placing the extracted and non-extracted data into a new or existing field.
- **Open Parser**—This stage parses your input data from many cultures of the world using a simple, but powerful parsing grammar. Using this grammar, you can define a sequence of expressions that represent domain patterns for parsing your input data. Open Parser also collects statistical data and scores the parsing matches to help you determine the effectiveness of your parsing grammars.
- **Table Lookup**—This stage evaluates a term and compares it to a previously validated form of that term. If the term is not in the proper form, then the standard version replaces the term. Table Lookup can change full words to abbreviations, change abbreviations to full words, change nicknames to full names, and can correct misspellings.
- **Transliterator**—Transliterator converts a string between Latin and other scripts.

Enterprise Data Integration Module

The Enterprise Data Integration Module provides the features you need to connect to data in multiple sources either directly or through integration with your existing data access technologies. With the Enterprise Data Integration Module, you can use Spectrum™ Technology Platform for data warehousing, data quality, systems integration, and migration.

Components

The Enterprise Data Integration Module consists of:

- **Call Stored Procedure**—A source stage that executes a stored procedure in a database, and returns the results of the stored procedure call as input for the dataflow. Use Call Stored Procedure when you want to get data from a database using a database's stored procedure rather than a query to a table or view.
- **Field Selector**—Allows you to choose which fields to pass to the next stage in the dataflow. You can use Field Selector to remove unwanted fields from a dataflow. For example, if you have created a new field by combining the data from two fields, and you no longer need the two source fields, you can use Field Selector to retain only the new field and remove the two source fields from the dataflow.
- **Generate Time Dimension**—Creates date records, one for each day of the date range you specify. You can then write these records to a time dimension table in a database using the Write to DB stage. The time dimension table can then be used to perform accurate calculations based on a period. For example, sales by quarter, budget spend by quarter, and revenue by day are all analyses that require a time dimension. Time dimension tables also enable you to account for fiscal years or non-standard quarters in the analysis.
- **Query Cache**—Looks up data in a cache based on values in one or more dataflow fields and returns data from matching records in the cache, adding the cache record's data to the record in the dataflow. Looking up data in a cache can improve performance compared to looking up data in a database.

- **Read From DB**—Reads data from a database table or view as input to a dataflow. Read From DB is available for jobs and subflows but not services.
- **Read From HL7**—Reads Health Level Seven (HL7) data from a text file as input to a dataflow. HL7 is a messaging standard used in the healthcare industry to exchange data between systems. For more information about HL7, review the details at www.hl7.org.
- **Read from Spreadsheet**—Reads data from an Excel spreadsheet as input to a dataflow. It supports both .xls and .xlsx file formats.
- **Read From Variable Format File**—Reads data from a file containing records of varying layout, reading each record as a list field. You can specify the tag that indicates the parent record type, and all other record types will become list fields under the parent.
- **Read From XML**—Reads an XML file into a job or subflow. It defines the file's path and data format, including XML schema and data element details.
- **Transposer**—Converts columns to rows. Transposing data is the opposite of pivoting data using the Group Statistics stage, which transforms row data into columns.
- **Write to Cache**—Loads output from a dataflow into a global cache, making the data available for lookup from the Query Cache stage. Using a global cache for data lookups improves performance compared to lookups to databases.
- **Write to DB**—Writes the output of a dataflow to a database.
- **Write to Spreadsheet**—Writes data to an Excel spreadsheet as output from a dataflow.
- **Write to Variable Format File**—Writes records of varying layouts to an output file.
- **Write to XML**—Writes the output of a job or subflow to an XML file.

Universal Name Module

To perform the most accurate standardization you may need to break up strings of data into multiple fields. Spectrum™ Technology Platform provides advanced parsing features that enable you to parse personal names, company names, and many other terms and abbreviations. In addition, you can create your own list of custom terms to use as the basis of scan and extract operations.

Components

The Universal Name Module consists of:

- **Name Variant Finder**—This component works in either a first name or last name mode in order to query the names database and return name variations. Name Variant Finder can limit the number of results by gender and culture. Spectrum™ Technology Platform includes a base names file. Add-on names files of other cultures are available and can be deployed by copying the appropriate JAR file into the `modules/tables/ext` folder. Arabic names are available as an add-on. The data in this file is by Nomino, Inc.
- **Open Name Parser**—The open name parser breaks down personal and business names and other terms in the name data field into their component parts. These parsed name elements are then subsequently available to other automated operations such as name matching, name standardization, or multi-record name consolidation.

Smart Data Quality Module

Spectrum™ Technology Platform Smart Data Quality is a *Machine Learning* based solution which helps to create initial match rules and potential match key components for your entity resolution process. With added machine learning capabilities to the *Data Quality* processes, the matching procedure has been significantly simplified and is capable of unlocking maximum potential available in your data.

Matching algorithms and thresholds are learnt in an automated manner based on the user's matching scenario. An initial match rule and potential match key components are generated via the inputs and tagging provided.

To generate match rules and match key components using this system, upload a *sample data* which must be a comprehensive collection of your records containing all possible variations. Subsequently, select the columns on which matching has to be performed, followed by grouping them into rough groups such that record pairs with different variations can be chosen. Further, tag the records according to your matching scenario and obtain potential match key components as well as a match rule learnt from your sample data.

See the task flow and the subsequent sections for a step-by-step guide to generate a match rule and potential match key components.

The Task Flow



1. Start with *Selecting files from the source*. The selected file should contain *Sample Data*; this must be an actual comprehensive representation of your records with all possible variations.
2. After uploading a sample data, you must **Select Columns** from your data on which you wish to perform matching. The columns selected in this step are used for **Generating Groups** in the next step.
3. In this step, you have to specify a value for **Group Strength**. This strength aids in generating groups of similar records. You must review these groups before moving to the next step.
4. After generating and reviewing groups, you must tag the displayed record pairs as **Match**, **Non-Match**, or **Unsure**. These tags assist in generating an accurate match rule and potential match key components.
5. The final step involved is to view and analyze the generated results. Upon reviewing the match rule, you can choose to export it to *match rule repository* in the **Enterprise Designer** and consume it in the *matching stages*. For more information about match rules, see [Match Rules](#).

After reviewing match key components, you can use them in the **Match Key Generator** stage of the **Enterprise Designer**.

Spatial and Routing Bundle

The Spectrum Spatial Module (previously called the Location Intelligence Module) provides organizations with a suite of broadly applicable spatial capabilities, including spatial analysis, geoprocessing, routing, and mapping. You can combine these capabilities with a wide range of data management capabilities in Spectrum™ Technology Platform and an extensive data catalog to solve a diverse variety of business problems.

The Spectrum Spatial Module installs:

- Stages in the data flow designer that provide spatial and routing capabilities and that customize services exposed via flows.
- Spatial Manager, which is a web application for administrators managing Spectrum Spatial.
- OGC, REST, and SOAP routing and spatial services used by application developers for creating solutions.

The Spectrum Spatial Module sells as a bundle with Spectrum Spatial Analyst, which is an out-of-the-box web mapping client to Spectrum Spatial. It is managed by Spatial Manager and uses the Spectrum Spatial Services. The client is extensible using the Angular extensibility framework to create custom components.

Spectrum Spatial helps you analyze data: you can use the web mapping client, write your application using spatial services, or design a data flow using spatial stages. You can extend Spectrum Spatial by creating custom services in data flows, or you can extend the user interface (UI) of the web mapping application using its extensibility framework.

Routing Stages

Routing stages obtain driving or walking directions, calculate drive time and drive distance, and identify locations within a certain time or distance from a starting point.

Components

The Spectrum Spatial Module includes the following routing stages:

- **Matrix** – Calculates the travel time and distances between an array of start and end locations.
- **Route** – Determines the shortest or fastest route between two distinct points (start and end locations). A route can include intermediate points between start and end locations.
- **Travel Boundary** – Determines a drive or walk-time or distance boundary from a given point or location.

Spatial Stages

Spatial stages determine the relationship between a location and areas or points of interest. They perform a variety of spatial calculations, such as area, union, intersection, point in polygon, and more.

Components

The Spectrum Spatial Module includes the following spatial stages:

- **Closest Site** – Determines which sites are closest to a given location.
- **Find Nearest** – Locates the points of interest (POI) that are nearest to a given location.
- **Point In Polygon** – Determines whether a location is within a specified area.
- **Query Spatial Data** – Queries a spatial database using MYSQL functions that can, for example, search for a point within a given polygon.
- **Read Spatial Data** – Accesses spatial data in a variety of commonly-used spatial data formats using MYSQL functions.
- **Spatial Calculator** – Performs string and geometry conversions as well a variety of calculations on spatial data, such as determining the area of a polygon or distance between two points.
- **Spatial Union** – Returns a geometry object that represents the union of two input geometry objects.
- **Write Spatial Data** – Inserts data into a named table, or update data in a named table, as the output of a job. Also deletes data.

Tax Jurisdiction Solutions Bundle

Append accurate tax jurisdiction information to addresses.

Enterprise Tax Module

Note: The Enterprise Tax Module processes only U.S. addresses.

The Enterprise Tax Module determines the tax jurisdiction for an address. The Enterprise Tax Module takes an address, standardizes it, then matches the address to an exact physical location, returning latitude/longitude coordinates with the correct tax jurisdictions for the address. This solution greatly reduces the inaccuracies associated with matching to 9-digit and 5-digit ZIP Codes. The Enterprise Tax Module can also calculate latitude and longitude coordinates for individual address locations, including the use of interpolation and offset.

For companies that deliver goods or services to locations that don't have a recognizable address, the Enterprise Tax Module can accept latitude/longitude coordinates as input to return the tax jurisdictions and corresponding GeoTAX Key values. Examples of this type of location include bill boards, street lights, communications towers, or new housing developments.

The Enterprise Tax Module uses a database of tax jurisdictions provided by TomTom. This data, which is collected through an ongoing research program and updated regularly, provides current jurisdictional boundary information down to the municipal and special tax district levels.

In addition to determining the tax jurisdiction for an address or location, with the optional Pitney Bowes Sales and Use Tax Rate file, the Enterprise Tax Module can return sales and use tax rates for each of the assigned tax jurisdictions as well as the total tax rate for the assigned locations.

There is also the option to use other software, such as Vertex or Sovos, to cross-reference and supply tax rate data.

Components

The Enterprise Tax Module consists of:

- **Assign GeoTAX Info** - Takes an input address and returns census, latitude/longitude, and tax information about the address. Assign GeoTAX Info utilizes Pitney Bowes' GeoTAX technology.
- **Calculate Distance** - Returns the distance between two input coordinates.
- **Reverse GeoTAX Info Lookup** - Returns census and tax information about an input location. This service does not return an address for a location.



3001 Summer Street
Stamford CT 06926-0700
USA

www.pitneybowes.com