ABSTRACT

Improved capabilities are described for identifying a first classification scheme associated with product attributes of a first grouping of products, identifying a second classification scheme associated with product attributes of a second grouping of products, and receiving a record of data relating to an item, the classification of which is uncertain. It may also involve receiving a dictionary of attributes associated with products and assigning the item to at least one of the classification schemes based on probabilistic matching among the attributes in the classification schemes, the attributes in the dictionary of attributes and the known attributes of the item.
6,200 START

6,202 IDENTIFY A FIRST CLASSIFICATION SCHEME ASSOCIATED WITH PRODUCT ATTRIBUTES OF A FIRST GROUPING OF PRODUCTS

6,204 IDENTIFYING A SECOND CLASSIFICATION SCHEME ASSOCIATED WITH PRODUCT ATTRIBUTES OF A SECOND GROUPING OF PRODUCTS

6,208 RECEIVING A RECORD OF DATA RELATING TO AN ITEM, THE CLASSIFICATION OF WHICH IS UNCERTAIN

6,210 RECEIVING A DICTIONARY OF ATTRIBUTES ASSOCIATED WITH PRODUCTS; AND ASSIGNING THE ITEM TO AT LEAST ONE OF THE CLASSIFICATION SCHEMES BASED ON PROBABILISTIC MATCHING AMONG THE ATTRIBUTES IN THE CLASSIFICATION SCHEMES, THE ATTRIBUTES IN THE DICTIONARY OF ATTRIBUTES AND THE KNOWN ATTRIBUTES OF THE ITEM

STOP

FIG. 3
SIMILARITY MATCHING OF PRODUCTS
BASED ON MULTIPLE CLASSIFICATION
SCHEMES

CROSS-REFERENCE TO RELATED
APPLICATIONS

[0001] This application is a continuation of U.S. patent
application Ser. No. 12/023,310 filed on Jan. 31, 2008, which
is a continuation-in-part of U.S. patent application Ser. No.
12/021,263 filed on Jan. 28, 2008. These applications claim
the benefit of the following U.S. provisional applications:
App. No. 60/886,798 filed on Jan. 26, 2007; App. No. 60/886,
801 filed on Jan. 26, 2007; App. No. 60/887,573 filed on Jan.
31, 2007; App. No. 60/891,508 filed on Feb. 24, 2007; App.
No. 60/891,936 filed on Feb. 27, 2007; and App. No. 60/952,

[0002] Each of the above applications is incorporated by
reference herein in its entirety.

BACKGROUND

[0003] 1. Field
[0004] The disclosure relates to methods and systems for
analyzing data, and more particularly to methods and systems
for aggregating, projecting, and releasing data.
[0005] 2. Description of Related Art
[0006] Currently, there exists a large variety of data
sources, such as census data or movement data received from
point-of-sale terminals, sample data received from manual
surveys, panel data obtained from the inputs of consumers
who are members of panels, fact data relating to products,
sales, and many other factors associated with the sales and
marketing efforts of an enterprise, and dimension data relat-
ing to dimensions along which an enterprise wishes to under-
stand data, such as in order to analyze consumer behaviors,
predict likely outcomes of decisions relating to an enter-
prise’s activities, and to project from sample sets of data to a
larger universe. Conventional methods of synthesizing,
aggregating, and exploring such a universe of data comprise
techniques such as OLAP, which fix aggregation points along
the dimensions of the universe in order to reduce the size and
complexity of unified information sets such as OLAP stars.
Exploration of the unified information sets can involve run-
time queries and query-time projections, both of which are
constrained in current methods by a priori decisions that must
be made to project and aggregate the universe of data. In
practice, going back and changing the a priori decisions can
lift these constraints, but this requires an arduous and
computationally complex restructuring and reprocessing of
data.

[0007] According to current business practices, unified
information sets and results drawn from such information sets
can be released to third parties according to so-called “releas-
ability” rules. These rules might apply to any and all of the
data from which the unified information sets are drawn, the
dimensions (or points or ranges along the dimensions), the
third party (or members or sub-organizations of the third
party), and so on. Given this, there can be a complex interac-
tion between the data, the dimensions, the third party, the
releasability rules, the levels along the dimensions at which
aggregations are performed, the information that is drawn
from the unified information sets, and so on. In practice,
configuring a system to apply the releasability rules is an
error-prone process that requires extensive manual set up and
results in a brittle mechanism that cannot adapt to on-the-fly
changes in data, dimensions, third parties, rules, aggrega-
tions, projections, user queries, and so on.

[0008] Various projection methodologies are known in the
art. Still other projection methodologies are subjects of
embodiments. In any case, different projection methodologies
provide outputs that have different statistical qualities.

[0009] Information systems are a significant bottle neck for
market analysis activities. The architecture of information
systems is often not designed to provide on-demand flexible
access, integration at a very granular level, or many other
critical capabilities necessary to support growth. Thus, infor-
mation systems are counter-productive to growth. Hundreds
of market and consumer databases make it very difficult to
manage or integrate data. For example, there may be a sepa-
rate database for each data source, hierarchy, and other data
characteristics relevant to market analysis. Different market
views and product hierarchies proliferate among manufactur-
ers and retailers. Restatements of data hierarchies waste pre-
cious time and are very expensive. Navigation from among
views of data, such as from global views to regional to neigh-
borhood to store views is virtually impossible, because there
are different hierarchies used to store data from global to
region to neighborhood to store-level data. Analyses and
insights often take weeks or months, or they are never pro-
duced. Insights are often sub-optimal because of silo-driven,
narrowly defined, ad hoc analysis projects. Reflecting the ad
hoc nature of these analytic projects are the analytic tools and
infrastructure developed to support them. Currently, market
analysis, business intelligence, and the like often use rigid
data cubes that may include hundreds of databases that are
impossible to integrate. These systems may include hundreds
of views, hierarchies, clusters, and so forth, each of which is
associated with its own rigid data cube. This may make it
almost impossible to navigate from global uses that are used,
for example, to develop overall company strategy, down to
specific program implementation or customer-driven uses.
These ad hoc analytic tools and infrastructure are fragmented
and disconnected.

[0010] In sum, there are many problems associated with the
data used for market analysis, and there is a need for a flexi-
able, extendable analytic platform, the architecture for which
is designed to support a broad array of evolving market analysis
needs. Furthermore, there is a need for better business intel-
ligence in order to accelerate revenue growth, make business
intelligence more customer-driven, to gain insights about
markets in a more timely fashion, and a need for data projec-
tion and release methods and systems that provide improved
dimensional flexibility, reduced query-time computational
complexity, automatic selection and blending of projection
methodologies, and flexibly applied releasability rules.

SUMMARY

[0011] In embodiments, systems and methods may involve
using a platform as disclosed herein for applications
described herein where the systems and methods involve
identifying a first classification scheme associated with prod-
uct attributes of a first grouping of products. It may also involve identifying a second classification scheme associated with product attributes of a second grouping of products. It may also involve receiving a record of data relating to an item, the classification of which is uncertain. It may also involve receiving a dictionary of attributes associated with products and assigning the item to at least one of the classification schemes based on probabilistic matching among the attributes in the classification schemes, the attributes in the dictionary of attributes and the known attributes of the item.

[0012] In embodiments, the probabilistic matching may be iterated until a statistical criterion is met. The product attribute may be a nutritional level, a brand, a product category, based at least in part on a SKU, and the like. In addition, the product attribute may be a physical attribute, where the physical attribute is a feature, a scent, a packaging type, a product launch date, a display location, and the like.

[0013] These and other systems, methods, objects, features, and advantages of implementations will be apparent to those skilled in the art from the following detailed description of the preferred embodiment and the drawings. Capitalized terms used herein (such as relating to titles of data objects, tables, or the like) should be understood to encompass other similar content or features performing similar functions, except where the context specifically limits such terms to the use herein.

BRIEF DESCRIPTION OF THE FIGURES

[0014] The invention and the following detailed description of certain embodiments thereof may be understood by reference to the following figures:

[0015] FIG. 1 illustrates an analytic platform for performing data analysis.

[0016] FIG. 2 depicts a similarities matching process.

[0017] FIG. 3 depicts a logical flow of a similarity matching process using multiple classification schemes.

DETAILED DESCRIPTION

[0018] Referring to FIG. 1, the methods and systems disclosed herein are related to improved methods for handling and using data and metadata for the benefit of an enterprise. An analytic platform 100 may support and include such improved methods and systems. The analytic platform 100 may include, in certain embodiments, a range of hardware systems, software modules, data storage facilities, application programming interfaces, human-readable interfaces, and methodologies, as well as a range of applications, solutions, products, and methods that use various outputs of the analytic platform 100, as more particularly detailed herein, other embodiments of which would be understood by one of ordinary skill in the art and are encompassed herein. Among other components, the analytic platform 100 includes methods and systems for providing various representations of data and metadata, methodologies for acting on data and metadata, an analytic engine, and a data management facility that is capable of handling disaggregated data and performing aggregation, calculations, functions, and real-time or quasi-real-time projections. In certain embodiments, the methods and systems enable much more rapid and flexible manipulation of data sets, so that certain calculations and projections can be done in a fraction of the time as compared with older generation systems.

[0019] In embodiments, data compression and aggregations of data, such as fact data sources 102, and dimension data sources 104, may be performed in conjunction with a user query such that the aggregation dataset can be specifically generated in a form most applicable for generating calculations and projections based on the query. In embodiments, data compression and aggregations of data may be done prior to, in anticipation of, and/or following a query. In embodiments, an analytic platform 100 (described in more detail below) may calculate projections and other solutions dynamically and create hierarchical data structures with custom dimensions that facilitate the analysis. Such methods and systems may be used to process point-of-sale (POS) data, retail information, geography information, causal information, survey information, census data and other forms of data and forms of assessments of past performance (e.g., estimating the past sales of a certain product within a certain geographical region over a certain period of time) or projections of future results (e.g., estimating the future or expected sales of a certain product within a certain geographical region over a certain period of time). In turn, various estimates and projections can be used for various purposes of an enterprise, such as relating to purchasing, supply chain management, handling of inventory, pricing decisions, the planning of promotions, marketing plans, financial reporting, and many others.

[0020] Referring still to FIG. 1 an analytic platform 100 is illustrated that may be used to analyze and process data in a disaggregated or aggregated format, including without limitation, dimension data defining the dimensions along which various items are measured and factual data about the facts that are measured with respect to the dimensions. Factual data may come from a wide variety of sources and be of a wide range of types, such as traditional periodic point-of-sale (POS) data, causal data (such as data about activities of an enterprise, such as in-store promotions, that are posited to cause changes in factual data), household panel data, frequent shopper program information, daily, weekly, or real time POS data, store database data, store list files, stubs, dictionary data, product lists, as well as custom and traditional audit data. Further extensions into transaction level data, RFID data and data from non-retail industries may also be processed according to the methods and systems described herein.

[0021] In embodiments, a data loading facility 108 may be used to extract data from available data sources and load them to or within the analytic platform 100 for further storage, manipulation, structuring, fusion, analysis, retrieval, querying and other uses. The data loading facility 108 may have the a plurality of responsibilities that may include eliminating data for non-releasable items, providing correct venue group flags for a venue group, feeding a core information matrix with relevant information (such as and without limitation statistical metrics), or the like. In an embodiment, the data loading facility 108 eliminate non-related items. Available data sources may include a plurality of fact data sources 102 and a plurality of dimension data sources 104. Fact data sources 102 may include, for example, facts about sales volume, dollar sales, distribution, price, POS data, loyalty card transaction files, sales audit files, retailer sales data, and many other fact data sources 102 containing facts about the sales of the enterprise, as well as causal facts, such as facts about activities of the enterprise, in-store promotion audits, electronic pricing and/or promotion files, feature ad coding files, or others that tend to influence or cause changes in sales or
other events, such as facts about in-store promotions, advertising, incentive programs, and the like. Other fact data sources may include custom shelf audit files, shipment data files, media data files, explanatory data (e.g., data regarding weather), attitudinal data, or usage data. Dimension data sources 104 may include information relating to any dimensions along which an enterprise wishes to collect data, such as dimensions relating to products sold (e.g., attribute data relating to the types of products that are sold, such as data about UPC codes, product hierarchies, categories, brands, sub-brands, SKUs and the like), venue data (e.g., store, chain, region, country, etc.), time data (e.g., day, week, quad-week, quarter, 12-week, etc.), geographic data (including breakdowns of stores by city, state, region, country or other geographic groupings), consumer or customer data (e.g., household, individual, demographics, household groupings, etc.), and other dimension data sources 104. While embodiments disclosed herein relate primarily to the collection of sales and marketing-related facts and the handling of dimensions related to the sales and marketing activities of an enterprise, it should be understood that the methods and systems disclosed herein may be applied to facts of other types and to the handling of dimensions of other types, such as facts and dimensions related to manufacturing activities, financial activities, information technology activities, media activities, supply chain management activities, accounting activities, political activities, contracting activities, and many others.

[0022] In an embodiment, the analytic platform 100 comprises a combination of data, technologies, methods, and delivery mechanisms brought together by an analytic engine. The analytic platform 100 may provide a novel approach to managing and integrating market and enterprise information and enabling predictive analytics. The analytic platform 100 may leverage approaches to representing and storing the base data so that it may be consumed and delivered in real-time, with flexibility and open integration. This representation of the data, when combined with the analytic methods and techniques, and a delivery infrastructure, may minimize the processing time and cost and maximize the performance and value for the end user. This technique may be applied to problems where there may be a need to access integrated views across multiple data sources, where there may be a large number of dimensions to vary, and where there may be a need to rapidly and accurately handle dynamic dimensionality requests, with appropriate aggregations and projections, where there may be highly personalized and flexible real-time reporting 190, analysis 192 and forecasting capabilities required, where there may be a need to tie seamlessly and on-the-fly with other enterprise applications 184 via web services 194 such as to receive a request with specific dimensionality, apply appropriate calculation methods, perform and deliver an outcome (e.g., dataset, coefficient, etc.), and the like.

[0023] The analytic platform 100 may provide innovative solutions to application partners, including on-demand pricing insights, emerging category insights, product launch management, loyalty insights, daily data out-of-stock insights, assortment planning, on-demand audit groups, neighborhood insights, shopper insights, health and wellness insights, consumer tracking and targeting, and the like.

[0024] A decision framework may enable new revenue and competitive advantages to application partners by brand building, product innovation, consumer-centric retail execution, consumer and shopper relationship management, and the like. Predictive planning and optimization solutions, automated analytics and insight solutions, and on-demand business performance reporting may be drawn from a plurality of sources, such as InfoScan, total C-scan, daily data, panel data, retailer direct data, SAP, consumer segmentation, consumer demographics, FSP/loyalty data, data provided directly for customers, or the like.

[0025] The analytic platform 100 may have advantages over more traditional federation/consolidation approaches, requiring fewer updates in a smaller portion of the process. The analytic platform 100 may support greater insight to users, and provide users with more innovative applications. The analytic platform 100 may provide a unified reporting and solutions framework, providing on-demand and scheduled reports in a user dashboard with summary views and graphical dial indicators, as well as flexible formatting options. Benefits and products of the analytic platform 100 may include non-additive measures for custom product groupings, elimination of restatements to save significant time and effort, cross-category visibility to spot emerging trends, provide a total market picture for faster competitor analysis, provide granular data on demand to view detailed retail performance, provide attribute driven analysis for market insights, and the like.

[0026] The analytic capabilities of implementations may provide for on-demand projection, on-demand aggregation, multi-source master data management, and the like. On-demand projection may be derived directly for all possible geographies, store and demographic attributes, per geography or category, with built-in dynamic reusability controls, and the like. On-demand aggregation may provide both additive and non-additive measures, provide custom groups, provide cross-category or geography analytics, and the like. Multi-source master data management may provide management of dimension member catalogue and hierarchy attributes, processing of raw fact data that may reduce harmonization work to attribute matching, product and store attributes stored relationally, with data that may be extended independently of fact data, and used to create additional dimensions, and the like.

[0027] In addition, the analytic platform 100 may provide flexibility, while maintaining a structured user approach. Flexibility may be realized with multiple hierarchies applied to the same database, the ability to create new custom hierarchies and views, rapid addition of new measures and dimensions, and the like. The user may be provided a structured approach through publishing and subscribing reports to a broader user base, by enabling multiple user classes with different privileges, providing security access, and the like. The user may also be provided with increased performance and ease of use, through leading-edge hardware and software, and web application for integrated analysis.

[0028] In embodiments, the data available within a fact data source 102 and a dimension data source 104 may be linked, such as through the use of a key. For example, key-based fusion of fact 102 and dimension data 104 may occur by using a key, such as using the Ablilite Key software product offered by Acxiom, in order to fuse multiple sources of data. For example, such a key can be used to relate loyalty card data (e.g., Grocery Store 1 loyalty card, Grocery Store 2 loyalty card, and Convenience Store 1 loyalty card) that are available for a single customer, so that the fact data from multiple sources can be used as a fused data source for analysis on desirable dimensions. For example, an analyst might wish to
view time-series trends in the dollar sales allotted by the
customer to each store within a given product category.

[0029] In embodiments the data loading facility may com-
prise any of a wide range of data loading facilities, including
or using suitable connectors, bridges, adaptors, extraction
engines, transformation engines, loading engines, data filter-
ing facilities, data cleansing facilities, data integration facili-
ties, or the like, of the type known as any of ordinary skill in
the art. In various embodiments, there are many situations
where a store will provide POS data and causal information
relating to its store. For example, the POS data may be auto-
matically transmitted to the facts database after the sales
information has been collected at the stores POS terminals.
The same store may also provide information about how it
promoted certain products, its store or the like. This data may
be stored in another database; however, this causal informa-
tion may provide one with insight on recent sales activities so
it may be used in later sales assessments or forecasts. Simi-
larly, a manufacturer may load product attribute data into yet
another database and this data may also be accessible for sales
assessment or projection analysis. For example, when mak-
ing such analysis one may be interested in knowing what
categories of products sold well or what brand sold well. In
this case, the causal store information may be aggregated with
the POS data and dimension data corresponding to the prod-
cuts referred to in the POS data. With this aggregation of
information one can make an analysis on any of the related
data.

[0030] Referring still to FIG. 1, data that is obtained by
the data loading facility 108 may be transferred to a plurality of
facilities within the analytic platform 100, including the data
Mart 114. In embodiments the data loading facility 108 may
contain one or more interfaces 182 by which the data loaded
by the data loading facility 108 may interact with or be used
by other facilities within the platform 100 or external to the
platform. Interfaces to the data loading facility 108 may include
human-readable user interfaces, application program-
ning interfaces (APIs), registries or similar facilities
suitable for providing interfaces to services in a services
oriented architecture, connectors, bridges, adaptors, bind-
ings, protocols, message brokers, extraction facilities, trans-
formation facilities, loading facilities and other data integra-
tion facilities suitable for allowing various other entities to
interact with the data loading facility 108. The interfaces 182
may support interactions with the data loading facility 108 by
applications 184, solutions 188, reporting facilities 190, ana-
lyses facilities 192, services 194 or other entities, external to
or internal to an enterprise. In embodiments these inter-
faces are associated with interfaces 182 to the platform 100,
but in other embodiments direct interfaces may exist to the
data loading facility 108, either by other components of the
platform 100, or by external entities.

[0031] Referring still to FIG. 1, in embodiments the data
Mart facility 114 may be used to store data loaded from the
data loading facility 108 and to make the data loaded from the
data loading facility 108 available to various other entities in
or external to the platform 100 in a convenient format. Within
the data Mart 114 facilities may be present to further store,
manipulate, structure, subset, merge, join, fuse, or perform a
wide range of data structuring and manipulation activities.
The data Mart facility 114 may also allow storage, manipula-
tion and retrieval of metadata, and perform activities on meta-
data similar to those disclosed with respect to data. Thus, the
data Mart facility 114 may allow storage of data and metadata
about facts (including sales facts, causal facts, and the like)
and dimension data, as well as other relevant data and meta-
data. In embodiments, the data Mart facility 114 may com-
press the data and/or create summaries in order to facilitate
faster processing by other of the applications 184 within the
platform 100 (e.g. the analytic server 134). In embodiments
the data Mart facility 114 may include various methods, com-
ponents, modules, systems, sub-systems, features or facilities
associated with data and metadata.

[0032] In certain embodiments the data Mart facility 114
may contain one or more interfaces 182 (not shown on FIG.
1), by which the data loaded by the data Mart facility 114 may
interact with or be used by other facilities within the platform
100 or external to the platform. Interfaces to the data Mart
facility 114 may include human-readable user interfaces,
application programming interfaces (APIs), registries or similar
facilities suitable for providing interfaces to services in a services
oriented architecture, connectors, bridges, adaptors, bind-
ings, protocols, message brokers, extraction facilities,
transformation facilities, loading facilities and other data integra-
tion facilities suitable for allowing various other entities
to interact with the data Mart facility 114. These interfaces
may comprise interfaces 182 to the platform 100 as a whole,
or may be interfaces associated directly with the data Mart
facility 114 itself, such as for access from other components
of the platform 100 or for access by external entities directly
to the data Mart facility 114. The interfaces 182 may support
interactions with the data Mart facility 114 by applications
184, solutions 188, reporting facilities 190, analyses facilities
192, services 194 (each of which is describe in greater detail
herein) or other entities, external to or internal to an enter-
prise.

[0033] In certain optional embodiments, the security facili-
ty 118 may be any hardware or software implementation,
process, procedure, or protocol that may be used to block,
limit, filter or alter access to the data Mart facility 114, and/or
any of the facilities within the data Mart facility 114, by a
human operator, a group of operators, an organization, soft-
ware program, bot, virus, or some other entity or program.
The security facility 118 may include a firewall, an anti-virus
facility, a facility for managing permission to store, manipu-
late and/or retrieve data or metadata, a conditional access
facility, a logging facility, a tracking facility, a reporting facili-
ty, an asset management facility, an intrusion-detection facili-
ty, an intrusion-prevention facility or other suitable security
facility.

[0034] Still referring to FIG. 1, the analytic platform 100
may include an analytic engine 134. The analytic engine 134
may be used to build and deploy analytic applications or
solutions or undertake analytic methods based upon the use
of a plurality of data sources and data types. Among other
things, the analytic engine 134 may perform a wide range of
calculations and data manipulation steps necessary to apply
models, such as mathematical and economic models, to sets of
data, including fact data, dimension data, and metadata. The
analytic engine 134 may be associated with an interface 182,
such as any of the interfaces described herein.

[0035] The analytic engine 134 may interact with a model
storage facility 148, which may be any facility for generat-
ing models used in the analysis of sets of data, such as economic
models, econometric models, forecasting models, decision
support models, estimation models, projection models, and
many others. In embodiments output from the analytic engine
134 may be used to condition or refine models in the model
storage 148; thus, there may be a feedback loop between the two, where calculations in the analytic engine 134 are used to refine models managed by the model storage facility 148.

[0036] In embodiments, a security facility 138 of the analytic engine 134 may be the same or similar to the security facility 118 associated with the data mart facility 114, as described herein. Alternatively, the security facility 138 associated with the analytic engine 134 may have features and rules that are specifically designed to operate within the analytic engine 134.

[0037] As illustrated in FIG. 1, the analytic platform 100 may contain a master data management hub 150 (MDM1). In embodiments the MDM1 150 may serve as a central facility for handling dimension data used within the analytic platform 100, such as data about products, stores, venues, geographies, time periods and the like, as well as various other dimensions relating to or associated with the data and metadata types in the data sources 102, 104, the data loading facility 108, the data mart facility 114, the analytic engine 134, the model storage facility 148 or various applications, 184, solutions 188, reporting facilities 190, analytic facilities 192 or services 194 that interact with the analytic platform 100. The MDM1 150 may in embodiments include a security facility 152, an interface 158, a data loader 160, a data manipulation and structuring facility 162, and one or more staging tables 164. The data loader 160 may be used to receive data. Data may enter the MDM1 from various sources, such as from the data mart 114 after the data mart 114 completes its intended processing of the information and data that it received as described herein. Data may also enter the MDM1 through a user interface 158, such as an API or a human user interface, web browser or some other interface, of any of the types disclosed herein or in the documents incorporated by reference herein. The user interface 158 may be deployed on a client device, such as a PDA, personal computer, laptop computer, cellular phone, or some other client device capable of handling data. In embodiments, the staging tables 164 may be included in the MDM1 150.

[0038] In embodiments, a matching facility 180 may be associated with the MDM1 150. The matching facility 180 may receive an input data hierarchy within the MDM1 150 and analyze the characteristics of the hierarchy and select a set of attributes that are salient to a particular analytic interest (e.g., product selection by a type of consumer, product sales by a type of venue, and so forth). The matching facility 180 may select primary attributes, match attributes, associate attributes, block attributes and prioritize the attributes. The matching facility 180 may associate each attribute with a weight and define a set of probabilistic weights. The probabilistic weights may be the probability of a match or a non-match, or thresholds of a match or non-match that is associated with an analytic purpose (e.g., product purchase). The probabilistic weights may then be used in an algorithm that is run within a probabilistic matching engine (e.g., IBM QualityStage). The output of the matching engine may provide information on, for example, other products which are appropriate to include in a data hierarchy, the untapped market (i.e. other venues) in which a product is probabilistically more likely to sell well, and so forth. In embodiments, the matching facility 180 may be used to generate projections of what types of products, people, customers, retailers, stores, store departments, etc. are similar in nature and therefore they may be appropriate to combine in a projection or an assessment.

[0039] As illustrated in FIG. 1, the analytic platform 100 may include a projection facility 178. A projection facility 178 may be used to produce projections, whereby a partial data set (such as data from a subset of stores of a chain) is projected to a universe (such as all of the stores in a chain), by applying appropriate weights to the data in the partial data set. A wide range of potential projection methodologies exist, including cell-based methodologies, store matrix methodologies, iterative proportional fitting methodologies, virtual census methodologies, and others. The methodologies can be used to generate projection factors. As to any given projection, there is typically a tradeoff among various statistical quality measurements associated with that type of projection. Some projections are more accurate than others, while some are more consistent, have less spillage, are more closely calibrated, or have other attributes that make them relatively more or less desirable depending on how the output of the projection is likely to be used. In embodiments of the platform 100, the projection facility 178 takes different information from the MDM1 150 or from another source and provides a set of projection weightings along the applicable dimensions, typically reflected in a matrix of projection weights, which can be applied at the data mart facility 114 to a partial data set in order to render a projected data set. The projection facility 178 may have an interface 182 of any of the types disclosed herein.

[0040] As shown in FIG. 1, an interface 182 may be included in the analytic platform 100. In embodiments, data may be transferred to the MDM1 150 of the platform 100 using a user interface 182. The interface 182 may be a web browser operating over the Internet or within an intranet or other network, it may be an analytic engine 134, an application plug-in, or some other user interface that is capable of handling data. The interface 182 may be human readable or may consist of one or more application programming interfaces, or it may include various connectors, adaptors, bridges, services, transformation facilities, extraction facilities, loading facilities, bindings, couplings, or other data integration facilities, including any such facilities described herein or in documents incorporated by reference herein.

[0041] As illustrated in FIG. 1, the platform 100 may interact with a variety of applications 184, solutions 188, reporting facilities 190, analytic facilities 192 and services 194, such as web services, or with other platforms or systems of an enterprise or external to an enterprise. Any such applications 184, solutions 188, reporting facilities 190, analytic facilities 192 and services 194 may interact with the platform 100 in a variety of ways, such as providing input to the platform 100 (such as data, metadata, dimension information, models, projections, or the like), taking output from the platform 100 (such as data, metadata, projection information, information about similarities, analytic output, output from calculations, or the like), modifying the platform 100 (including in a feedback or iterative loop), being modified by the platform 100 (again optionally in a feedback or iterative loop), or the like.

[0042] In embodiments one or more applications 184 or solutions 188 may interact with the platform 100 via an interface 182. Applications 184 and solutions 188 may include applications and solutions (consisting of a combination of hardware, software and methods, among other components) that relate to planning the sales and marketing activities of an enterprise, decision support applications, financial reporting applications, applications relating to strategic planning, enterprise dashboard applications, supply chain management...
applications, inventory management and ordering applications, manufacturing applications, customer relationship management applications, information technology applications, applications relating to purchasing, applications relating to pricing, promotion, positioning, placement and products, and a wide range of other applications and solutions.

[0043] In embodiments, applications 184 and solutions 188 may include analytic output that is organized around a topic area. For example, the organizing principle of an application 184 or a solution 188 may be a new product introduction. Manufacturers may release thousands of new products each year. It may be useful for an analytic platform 100 to be able to group analysis around the topic area, such as new products, and organize a bundle of analyses and workflows that are presented as an application 184 or solution 188. Applications 184 and solutions 188 may incorporate planning information, forecasting information, “what if?” scenario capability, and other analytic features. Applications 184 and solutions 188 may be associated with web services 194 that enable users within a client’s organization to access and work with the applications 184 and solutions 188.

[0044] In embodiments, the analytic platform 100 may facilitate delivering information to external applications 184. This may include providing data or analytic results to certain classes of applications 184. For example and without limitation, an application may include enterprise resource planning/backbone applications 184 such as SAP, including those applications 184 focused on Marketing, Sales & Operations Planning and Supply Chain Management. In another example, an application may include business intelligence applications 184, including those applications 184 that may apply data mining techniques. In another example, an application may include customer relationship management applications 184, including customer sales force applications 184. In another example, an application may include specialty applications 184 such as a price or SKU optimization application. The analytic platform 100 may facilitate supply chain efficiency applications 184. For example and without limitation, an application may include supply chain models based on sales out (POS/FSP) rather than sales in (Shipments). In another example, an application may include RFID based supply chain management. In another example, an application may include a retailer co-op to enable partnership with a distributor who may manage collective stock and distribution services. The analytic platform 100 may be applied to industries characterized by large multi-dimensional data structures. This may include industries such as telecommunications, elections and polling, and the like. The analytic platform 100 may be applied to opportunities to vend large amounts of data through a portal with the possibility to deliver highly customized views for individual users with effectively controlled user accessibility rights. This may include collaborative groups such as insurance brokers, real estate agents, and the like. The analytic platform 100 may be applied to applications 184 requiring self monitoring of critical coefficients and parameters. Such applications 184 may rely on constant updating of statistical models, such as financial models, with real-time flows of data and ongoing re-calibration and optimization. The analytic platform 100 may be applied to applications 184 that require breaking apart and recombining geographies and territories at will.

[0045] In embodiments, a matching facility 180 may be a similarity facility 180 may be associated with the MDMH 150. The similarity facility 180 may receive an input data hierarchy within the MDMH 150 and analyze the characteristics of the hierarchy and select a set of attributes that are salient to a particular analytic interest (e.g., product selection by a type of consumer, product sales by a type of venue, and so forth). The similarity facility 180 may select primary attributes, match attributes, associate attributes, block attributes and prioritize the attributes. The similarity facility 180 may associate each attribute with a weight and define a set of probabilistic weights. The probabilistic weights may be the probability of a match or a non-match, or thresholds of a match or non-match that is associated with an analytic purpose (e.g., product purchase). The probabilistic weights may then be used in an algorithm that is run within a probabilistic matching engine (e.g., IBM QualityStage). The output of the matching engine may provide information on, for example, other products which are appropriate to include in a data hierarchy, the untapped market (i.e. other venues) in which a product is probabilistically more likely to sell well, and so forth. In embodiments, the similarity facility 180 may be used to generate projections of what types of products, people, customers, retailers, stores, store departments, etc. are similar in nature and therefore they may be appropriate to combine in a projection or an assessment.

[0046] In embodiments, the MDMH 150 may accommodate a blend of disaggregated and pre-aggregated data as necessitated by a client’s needs. For example, a client in the retail industry may have a need for a rolling, real-time assessment of store performance within a sales region. The ability of the MDMH 150 to accommodate twinkle data, and the like, may give the client useful insights into disaggregated sales data as it becomes available and make it possible to create projections based upon it and other available data. At the same time, the client may have pre-aggregated data available for use, for example a competitor’s sales data, economic indicators, inventory, or some other dataset. The MDMH 150 may handle the dimension data needed to combine the use of these diverse data sets.

[0047] Referring to FIG. 1 and FIG. 2, several types of data management challenges may be grouped into a class of problems known as similarity; which may be described as the problem of finding the commonalities or similarities within data structures. While there may be situations where data comparisons, fusions, combinations, aggregations and the like can be made directly (e.g. through an explicit reference to the same item name) there are many situations where there exist two things that are characterized by various attributes but where the attributes do not match. In such situations, in certain embodiments, a similarities facility 180 may be used to determine if the characteristics (e.g. the identified attributes in a fact or dimensions database) of the two things are close enough for the things to be called “similar.” In embodiments, the similarities facility 180 may use a probabilistic based methodology for the determination of similarity. The probabilistic methodology may use input (e.g. user input or API input) to determine if the two things are similar within a certain degree of confidence and/or the probabilistic methodology may produce an indication of how likely the two things are to be similar. There are many processes that can be performed on the data once two or more things are determined to be similar. For example, data associated with the two things may be aggregated or fused together such that the information pertaining to the two things can be used as a projection of one whole. In certain embodiments, the similarity information may be used to generate new attributes or other descriptions to
be associated with the thing being analyzed. For example, once a certain product is identified as similar to a certain class of products, data indicating such may be associated with the certain product. New attribute data may be associated with an item and the information may be loaded into a dimensions database such that data associated with the item, and the item itself, can be used in projections or assessments.

A similarities facility 180 according to some implementations may be used to assess the similarity of products, items, departments, stores, environments, real estate, competitors, markets, regions, performance, regional performance, and a variety of other things. For example, in the management of retail market information, the similarity problem may manifest itself as a need to identify similar stores for purposes of projecting regional performance from a sample of as the need to identify the competing or comparison set of products for any given item; as the need to identify similar households for purposes of bias correction; as the need to properly place new or previously unclassified or reclassified items in a product hierarchy; or for other projections and assessments. In another example, again from the retail industry, automated item placement may pose a problem. Often the current solution is labor intensive and may take from eight to twelve weeks for a new product to get properly placed within a store, department, shelf or otherwise. This delay inhibits the effectiveness of the analysis of new product introductions and in the tactical monitoring of category areas for a daily data service. However, these application sets need the product list to be up to date. In addition, the management of custom retail hierarchies may require that items from all other retailers in the market be placed inside that hierarchy, and often the structure of that hierarchy is not based on existing attributes. This may mean that the logic of the hierarchy itself must be discovered and then all other items from the market must be ‘similarly’ organized. In the current environment, this process takes months and is prone to error. In embodiments, issues of similarity are automated using techniques such as rules-based matching, algorithmic matching, and other similarities methods. In the present example, the similarities facility 180 may be used to assess the similarity of the new product with existing products to determine placement information. Once the similar matches are made, new attributes may be added to the new product description such that it is properly grouped physically in the store and electronically within the data structures to facilitate proper assessments and projections.

The similarity of entities may be associated with the concept of grouping entities for the purpose of reporting. One purpose may be to codify the rules placing entities into a specific classification definition. Some examples of such specific classification definitions may be item similarity for use in automatic classification and hierarchy, venue similarity for use in projections and imputations, consumer similarity in areas like ethnicity and economies used for developing weighting factors, or the like.

There may be certain matching requirements used by the similarities facility 180 in the determination of similarity. For example, scenarios for matching similarity may involve determining similar items related to a single specified item, where the similarities engine is programmed to identify all the items in the repository that are similar to it with respect to some specific criteria (e.g. one or more attributes listed in the dimensions database). The similarities engine may also or instead be programmed to analyze a list of items, where all the items in the repository are similar to each of the items in the list with respect to some specific criteria. Likewise, the similarities facility 180 may analyze an item within a list of items, where group items are placed into classifications of similar items with respect to some specific criteria; or the like.

In embodiments, the similarities facility 180 may use a probabilistic matching engine where the probabilistic matching engine compares all or some subset of attributes to determine the similarity. Each of the attributes may be equally considered in the probabilistic evaluation or certain attributes may be weighted with greater relevance. The similarities facility 180 may select certain attributes to weight with greater relevance and/or it may receive input to determine which attributes to weight. For example, the similarities engine may automatically select and disproportionately weight the attributes associated with ‘scent’ when assessing the products that it understands to be deodorants. In other embodiments, a user may determine which attributes or fields to disproportionately weight. For example, a user may determine a priority for the weighting of certain attributes (e.g. attributes within a list of attributes identified in a dimensions database), and load the prioritization, weighting or other such information into the similarities facility 180. The similarities facility 180 may then run a probabilistic matching engine with the weights as a factor in determining the extent to which things are similar.

An advantage of using probabilistic matching for doing similarity may be related to an unduplication process. The matching for unduplication and similarity may be similar. However, they may be based on different sets of attributes. For unduplication, all attributes may be used since the system may be looking for an exact match or duplicate item. This may work by matching a list against itself. In similarity, the system may be looking for items that are similar with regard to physical attributes, not the same exact item.

For two entities to be similar, they may have to be evaluated by a specific similarity measure. In most cases, they may have to be in the same domain, but this also may depend on the similarity measure that is used. The similarity measure that the system may use is the probabilistic matching of physical attributes of items, such as a deodorant keycat (or deodorant “key category”), where a keycat is a block of items that have a similar set of attributes. In this case, since the item may be a domain, and venue is a domain, an item may not be looked at as being similar to a venue.

The concept of similarity may be based on the similarity of the values of attributes that describe specific entities. This may involve developing similarity measures and the metadata to support them. The process may include deciding the purpose of the similarity; selecting the set of entities to be used in the similarity process; analyzing the characteristics (attributes and values) of each entity in the set of possible similar entities; deciding which attributes will be used as part of the similarity measure; deciding on the priority and weight of the set of attributes and their values; defining the similarity measure; defining all the probabilistic weights needed to perform the similarity measure; defining the thresholds, such as automatic definite match, automatic definite non-match, or undecided match with human intervention; or the like.

The measure used may be the probabilistic matching of certain physical attributes. This may be associated with automatic record linkage. Types of matching may include individual matching that may be used for the single item scenario, where one file contains a single item and a second file contains a master list of deodorants; many-to-one matching that may be a one-to-many matching of individual
records from one file to a second file; single file grouping that may be for grouping similar items in a single list; geographic coding that may insert missing information from one file into a second file when there is a match, useful for adding new attribute information from external examples to the repository that does not currently exist; unduplication service that may identify duplicate items in a single list; or the like.

[0056] Weighting factors of each attribute and of the total composite weight of the match may be important aspects to take into account in the matching process. Since the system may use probabilistic matching, the frequency analysis of each of the attribute values may be taken into account. The weight of the match of a specific attribute may be computed as the logarithm to the base two of the ratio of m and u, where the m probability is the probability that a field agrees given that the record pair being examined is a matched pair, which may be one minus the error rate of the field, and the u probability is the probability that a field agrees given that the record pair being examined is an unmatched pair, which may be the probability that the field agrees at random. The composite weight of the match, also referred to as match weight, may be the sum of each of the matched attribute weights. Each matched attribute may then be computed. If two attributes do not match, the disagreement weight may be calculated as: log₂ ((1−m)/(1−u)). Each time a match is accomplished, a match weight may also be calculated. Thresholds may be established to decide if this is a good match, a partial match, or not a match at all.

[0057] When doing probabilistic matching, different types of attributes may be needed, such as block attributes and match attributes. In this instance, block attributes may divide a list of items into blocks. This may provide a better performance on the matching. The block attributes may have to match before the block of items is examined. Two keycats may have similar attributes with different sets of attribute values. A value may be the information stored in a value fact of an attribute. There may be different types of attributes, such as global attributes across all keycats, keycat specific attributes, or the like. A category may also be used as a block. A category may be a classification of items made up of one to many different full or partial keycats.

[0058] Global attributes may be used across a plurality of keycats. Block attributes for the item domain may come from either the set of global or keycat specific attributes. Global attributes for the item domain may include keycat, keycat description, system, generation, vendor, item, week added, week completed coding, week last moved, price, UPC description, UPC 70 character description, US Item, maintained, brand code, company code, company name, major brand name, immediate parent company, top parent company, brand type, minor brand name, major brand short description, or the like.

[0059] Specific keycat attributes may not be common across keycats. There may be more descriptive attributes that a consumer would look at. Some of these attributes may be common across multiple keycats. Most of the keycat specific attributes may be used as match attributes. However, for each keycat, the most important keycat specific attribute may be used as a block attribute. The match attributes may be where the true match for similarity occurs. Examples of deodorant keycat attributes may include total ounces, total count, base ounces, store location, per unit ounce, product type, package, flavor, scent, strength, additives, form, or the like. The block and match attributes may be selected from a list of deodorant specific attributes, such as per unit ounce, product type, package, flavor, scent, strength, additives, form, or the like.

[0060] The similarity process for deciding if items in an item domain are similar may use a probabilistic matching engine, such as the IBM Web Sphere QualityStage (QS). The process steps may include: extraction of the items from the item dictionary now and the new repository in the future, conversion of all volumetric attributes and defined attributes for the keycat into specific columns in a table using the long description as values, formatting the information into a fixed length column ASCII file, setting up a new project, entering the data file, mapping the data file columns, setting up and running an investigate stage to develop a frequency analysis of the values for each of the attributes that will be used in the match stage, setting up the block attributes from the list of deodorant specific attributes, or the like. Another process step may be associated with setting up and running a character concatenate investigate stage for each of the attributes, such as per unit ounce, product type, package, flavor, scent, strength, additives, form, and the like, that may be used in the matching process.

[0061] It should be appreciated that the probabilistic matching engine methodology is but one of the many methods that may be used within the similarity facility. Others methods may include, but are not limited to, time series similarity methods, attribute-based methods, spillage-based methods, information theory methods, classification trees, or some other similarity methodology, combination of similarity methodologies, or plurality of methodologies.

[0062] In embodiments, referring to FIG. 3, systems and methods may involve using a platform as disclosed herein for applications described herein where the systems and methods involve identifying a first classification scheme associated with product attributes of a first grouping of products. It may also involve identifying a second classification scheme associated with product attributes of a second grouping of products. It may also involve receiving a record of data relating to an item, the classification of which is uncertain. It may also involve receiving a dictionary of attributes associated with products and assigning the item to at least one of the classification schemes based on probabilistic matching among the attributes in the classification schemes, the attributes in the dictionary of attributes and the known attributes of the item.

[0063] In embodiments, the probabilistic matching may be iterated until a statistical criterion is met. The product attribute may be a nutritional level, a brand, a product category, based at least in part on a SKU, and the like. In addition, the product attribute may be a physical attribute, where the physical attribute is a flavor, a scent, a packaging type, a product launch date, a display location, and the like.

[0064] The elements depicted in flow charts and block diagrams throughout the figures imply logical boundaries between the elements. However, according to software or hardware engineering practices, the depicted elements and the functions thereof may be implemented as parts of a monolithic software structure, as standalone software modules, or as modules that employ external routines, code, services, and so forth, or any combination of these, and all such implementations are within the scope of the present disclosure. Thus, while the foregoing drawings and description set forth functional aspects of the disclosed systems, no particular arrangement of software for implementing these functional aspects
should be inferred from these descriptions unless explicitly stated or otherwise clear from the context.

Similarly, it will be appreciated that the various steps identified and described above may be varied, and that the order of steps may be adapted to particular applications of the techniques disclosed herein. All such variations and modifications are intended to fall within the scope of this disclosure. As such, the depiction and/or description of an order for various steps should not be understood to require a particular order of execution for those steps, unless required by a particular application, or explicitly stated or otherwise clear from the context.

The methods or processes described above, and steps thereof, may be realized in hardware, software, or any combination of these suitable for a particular application. The hardware may include a general-purpose computer and/or dedicated computing device. The processes may be realized in one or more microprocessors, microcontrollers, embedded microcontrollers, programmable digital signal processors or other programmable device, along with internal and/or external memory. The processes may also, or instead, be embodied in an application specific integrated circuit, a programmable gate array, a programmable array logic, or any other device or combination of devices that may be configured to process electronic signals. It will further be appreciated that one or more of the processes may be realized as computer executable code created using a structured programming language such as C, an object oriented programming language such as C++, or any other high-level or low-level programming language (including assembly languages, hardware description languages, and database programming languages and technologies) that may be stored, compiled or interpreted to run on one of the above devices, as well as heterogeneous combinations of processors, processor architectures, or combinations of different hardware and software.

Thus, in one aspect, each method described above and combinations thereof may be embodied in computer executable code that, when executing on one or more computing devices, performs the steps thereof. In another aspect, the methods may be embodied in systems that perform the steps thereof, and may be distributed across devices in a number of ways, or all of the functionality may be integrated into a dedicated, standalone device or other hardware. In another aspect, means for performing the steps associated with the processes described above may include any of the hardware and/or software described above. All such permutations and combinations are intended to fall within the scope of the present disclosure.

While implementations have been disclosed in connection with the preferred embodiments shown and described in detail, various modifications and improvements thereon will become readily apparent to those skilled in the art. Accordingly, the spirit and scope of the present invention is not to be limited by the foregoing examples, but is to be understood in the broadest sense allowable by law.

All documents referenced herein are hereby incorporated by reference.

1-12. (canceled)

13. A computer program product for placing a new item within a number of product hierarchies, the computer program product comprising computer executable embodied in a non-transitory computer readable medium that, when executing on one or more computing devices, performs the steps of:

- identifying a first hierarchy of products, the first hierarchy providing a first classification scheme associated with product attributes of a first grouping of products in a first category, wherein each product in the first hierarchy is included in a block of the first hierarchy having known block attributes, and each product in the first hierarchy has known match attributes;
- identifying a second hierarchy of products, the second hierarchy providing a second classification scheme associated with product attributes of a second grouping of products in a second category that classifies a different type of product than the first hierarchy of products, wherein each product is included in a block of the second hierarchy having known block attributes, and each product in the second hierarchy has known match attributes;
- using a computer, receiving a record of attributes of an unclassified item;
- assigning the item to a location in at least one of the first or second hierarchies based on probabilistic matching among the attributes of products in the hierarchies, and known attributes of the item, thereby classifying the item in a location in the first hierarchy or the second hierarchy;
- after assigning the item to the location, generating new attribute data based on a product type corresponding to the location of the item and existing attribute data and associating the new attribute data with the item.

14. The computer program product of claim 13, further comprising code that performs the step of iterating the probabilistic matching until a statistical criterion is met.

15. The computer program product of claim 13, wherein the known attributes of the item include a nutritional level.

16. The computer program product of claim 13, wherein the attributes of the item include a brand.

17. The computer program product of claim 13, wherein the attributes of the item include product category.

18. The computer program product of claim 13, wherein the attributes of the item include an attribute based at least in part on a SKU.

19. The computer program product of claim 13, wherein the attributes of the item include physical attribute.

20. The computer program product of claim 19, wherein the physical attribute is a flavor.

21. The computer program product of claim 19, wherein the physical attribute is a scent.

22. The computer program product of claim 19, wherein the physical attribute is packaging type.

23. The computer program product of claim 19, wherein the physical attribute is a product launch date.

24. The computer program product of claim 19, wherein the physical attribute is a display location.

25. The computer program product of claim 13, wherein the item to one of the hierarchies includes assigning the item to one of the hierarchies based on probabilistic matching to the attributes in a dictionary of attributes for one of the hierarchies.

26. A system comprising:

- a database storing a first hierarchy, the first hierarchy providing a first classification scheme associated with product attributes of a first grouping of products in a first category, wherein each product in the first hierarchy is included in a block of the first hierarchy having known block attributes, and each product in the first hierarchy has known match attributes of products, and a second
hierarchy of products, the second hierarchy providing a second classification scheme associated with product attributes of a second grouping of products in a second category that classifies a different type of product than the first hierarchy of products, wherein each product is included in a block of the second hierarchy having known block attributes, and each product in the second hierarchy has known match attributes; and

a memory; and

a processor configured by code stored in the memory to perform the steps of:

receiving a record of attributes of an unclassified item;
assigning the item to a location in at least one of the first or second hierarchies based on probabilistic matching among the attributes of products in the hierarchies, and known attributes of the item, thereby classifying the item in a location in the first hierarchy or the second hierarchy; and

after assigning the item to the location, generating new attribute data based on a product type corresponding to the location of the item and existing attribute data and associating the new attribute data with the item.

27. The system of claim 26, wherein the known attributes of the item include a nutritional level.

28. The system of claim 26, wherein the attributes of the item include a brand.

29. The system of claim 26, wherein the attributes of the item include product category.

30. The system of claim 26, wherein the attributes of the item include an attribute based at least in part on a SKU.

31. The system of claim 26, wherein the attributes of the item include a physical attribute.

32. The system of claim 31, wherein the physical attribute is at least one of a flavor, a scent, a packaging type, a product launch date, and a display location.

* * * * *