Configurable and Extensible Multi-flows for Providing Analytics as a Service on the Cloud

Deepak P Prasad M. Deshpande Karin Murthy
IBM Research – India, Bangalore, INDIA
{deepak.s.p, prasadsh, karin.murthy}@in.ibm.com

Abstract—Compared to traditional analytics deployment models, cloud-based solutions for business analytics provide numerous advantages such as reduction of a large upfront infrastructural cost and the efforts to setup an in-house analytics team. Such advantages of cloud-based service delivery make it particularly attractive for small and medium businesses. In spite of these advantages, analytics penetration has been low particularly in developing regions such as India and China due to many other factors. In this paper, we propose pre-packaged configurable workflows for analytics as a means of endearing cloud-based analytics to customers, with a special focus on small and medium businesses in developing regions. We introduce the concept of configurable multi-flows that make it easy for non-technical personnel to use and customize without being aware of the technical details of the various operators involved in the workflow. Multi-flows comprise of an overlap of multiple possible workflows and are easily extensible to include more variations to support the evolving needs of customers non-disruptively and incrementally. We detail a case-study of the Retail sector where an extensive survey of retail businesses in India revealed that configurable pre-packaged workflows may indeed help improve market penetration. We then identify common analytics needs of retail customers, and detail how such tasks can be expressed as configurable multi-flows. Further, we describe a fully functional implementation of our system that supports configurable multi-flows for analytics. Finally, we illustrate the ease-of-use of configurable multi-flows with the use of multiple screenshots.

Keywords—analytics; cloud; configurable; retail; smb

I. INTRODUCTION

Analytics is now used widely in most businesses in order to derive insights from past business activities to enhance and drive business planning for the future. The spectrum of business analytics activities ranges from identifying which products are frequently bought together [2], to decisions that involve a lot of investment such as whether to open a store at a newly identified location. Recently, most large businesses have formed specialized analytics teams to provide them analytical support for decision making, or even to the extent of automating some of the decision making [1]. In order to perform large scale analytics, companies also need to procure computational resources, and licenses for the data analytics solutions. Commercial data analytics solutions, along with the hardware, can easily cost up to \$100k¹. Even

single user licenses for such software with most common functionalities could cost as much as up to \$10k². Under this traditional model, the investment to start using analytics support for the business is two-fold:

- Setting up a business analytics team comprising of statisticians and business analysts. Typical business analytics teams comprise of high-end statisticians [3].
- Procuring expensive hardware (e.g., blade servers) and licenses of analytics toolkits to enable the business analytics team to develop custom analytics solutions.

This translates to a large entry barrier for businesses to start using analytics solutions, effectively limiting the use of analytics solutions to large enterprise customers.

Recent developments in utility computing [4] have given an impetus to reducing this entry barrier. In particular, cloud computing [5] has been leveraged by various companies to provide analytics solutions that eliminate the need for the customer to install costly hardware and software. Birst³ and GoodData⁴ are examples of companies that mainly provide reporting functionalities over utility computing-type models. In addition to such functionalities, companies like in2clouds⁵ provide more sophisticated analytics such as predictive modeling, over the cloud. Despite the reasonable success of such efforts in improving the penetration of analytics in developed geographies like the Americas and Europe, small and medium enterprises in developing regions such as India are yet to deploy analytics on a wide scale. We observe that an interplay of many factors lead to a general low penetration of analytics solutions in developing regions:

- **Rapidly Growing Market:** A rapidly growing market means that many small and medium enterprises can multiply their profits by widening their reach (e.g., opening more stores). Such a method of increasing profits by scaling up is often perceived as a much more deterministic way of increasing net profit and one that requires less know-how as compared to more sophisticated methods like deploying analytics to improve profit margins.
- **Availability of Cheap Labor:** The availability of cheap labor opens up other possibilities for service providers to improve their business, for example, offering labor-intensive services such as pick-up and delivery where

¹ <http://brenocon.com/blog/2009/02/comparison-of-data-analysis-packages-r-matlab-scipy-excel-sas-spss-stata/>

² <http://www.systat.com/SwitchToSystat.aspx>

³ <http://www.birst.com>

⁴ <http://www.gooddata.com>

⁵ <http://www.in2clouds.com>

appropriate. Availability of such easy methods of improving customer retention and profits decreases the importance that is given to performing sophisticated business analytics.

- **Lack of Digitized Data:** Developing regions such as India and China have a large population that still does not use electronic modes of payment such as credit and debit cards. Even an optimistic report⁶ on India projects that credit card penetration will reach only up to as less as 3% by 2014. Such lack of electronic payments and consequent absence of a customer identifier with each transaction makes it hard to group transactions made by the same customer, especially, in domains such as retail. These inevitably make customer profiling hard, thus making analytics less accurate.
- **Non-standard Data Formats:** A wide variety of data entry systems (e.g., POS systems) is available to small and medium businesses that want to use digitization of transaction data. Businesses choose their data entry and management systems by convenience rather than based on whether the system is adhering to standard domain-specific data formats (such as IBM BDW⁷). Often, custom software developed by small-scale software vendors is deployed, leading to heterogeneity of formats of the data stored; such heterogeneity makes it hard to design a generic analytics solution with wide applicability across many small and medium businesses.

These and other reasons make it necessary to lower the entry barrier even more, in order for an analytics solution to generate interest among small and medium businesses in developing regions. Such a market is, however, attractive due to the sheer size as well as its rapid growth. Analytics can help these growing businesses to cope with increasing competition, particularly in urban areas. Moreover, addressing challenges that are unique to such developing markets such as data heterogeneity and widely varying customer needs are likely to lead to solutions that are usable across other emerging developing markets such as Africa and Oceania. In this paper, we present an approach to improve the delivery of analytics services on the cloud with specific focus to improving penetration to small and medium businesses in developing regions. The specific novel features of our approach include:

- **Pre-packaged Domain Specific Workflows:** We identify common analytics tasks for specific domains (e.g., retail, finance) and provide domain-specific pre-packaged workflows to enable the customer to get started on using the solution. However, the wide variance in data formats and customer needs mean that pre-packaged workflows alone cannot be readily usable.
- **Workflows as Configurable Multi-Flows:** We model the pre-packaged workflows as configurable multi-flows, wherein the operators in the workflow can be configured easily. Multi-flows address the data

heterogeneity problem by overlapping multiple flows into a single workflow. A specific workflow for a given schema can be instantiated from the multi-flow. An instantiated workflow utilizes only a subset of the mesh of operators in the multi-flow, based on the needs. Multi-flows also enable providing several related functionalities in the same flow. We will delve into more details in a later section. This addresses the ability to rapidly tune analytics solutions to address new needs of customers.

- **Extensibility:** We provide features to easily extend workflows to support varying requirements for customers. Such inherent support for extensibility makes it easy to incrementally provide support to evolving needs of customers.

II. DOMAIN-SPECIFIC PACKAGED EXTENSIBLE CONFIGURABLE WORKFLOWS

In this section, we describe the core components of our solution: packaged extensible and configurable multi-flows. We start by providing examples on what analytics workflows look like, including the various functionalities involved in the workflow.

Figure 1. Sample SPSS Workflows

Figure 1 shows two simple workflows in SPSS (called streams, in SPSS parlance). Each workflow comprises of several operators, each of which depicts various functionalities such as transformations on data, or data sources, among other things. The first workflow in Figure 1 is a simple visualization workflow. The first operator extracts an SQL database table for which the second operator derives certain attributes; if the sales on each month are attributes in the table, the aggregate sales in the last quarter could be a derived attribute. The third operator represents a filter attribute that discards some attributes before passing it on to the chart creator node. The second workflow illustrates an analytical functionality, namely the usage of a C5.0 decision tree. An input data source, baskrule_input.csv, is fed to the C5.0 decision tree model, whose output predictions are then collected in the output node. Though these examples illustrate simple workflows, workflows in any analytics software may get arbitrarily complex with hundreds of operators and sophisticated non-linear flows. Broadly, analytics workflows consist of the following phases:

Data Preparation: The input data to an analytics workflow is often not in a format that can be readily used. Thus, a

⁶ http://www.rncos.com/Press_Releases/Usage-of-Credit-Cards-Surging-in-India.htm

⁷ <http://www-01.ibm.com/software/data/industry-models/banking-data/>

mesh of data preparation operators is often used to translate the data into a format that is consumable by downstream analytics operators. The operators in Figure 1 to derive new attributes or filter out other attributes are examples of data preparation operators.

Analytics Operators: Analytics operators are those that pertain to analytics functionalities such as clustering input data into groups, building models such as C5.0 (the model in Figure 1) and using pre-built models, among other analytical operations. These are often seen as the core of the workflow since they represent the kind of analytical functionalities that the workflow has been created for.

Data Export and Reporting: The data export and reporting operators transform the output of the analytics operators and may produce visualization of the results, or insert the transformed data into databases or other data storage systems. Often, the output of the analytics operators is not readily consumable by the data export and reporting operators, for example, a chart operator may require further data preparation operators.

In the following sub-sections, we now detail how each of the novel features that we propose have been implemented in our system, in the following sub-sections.

A. Multi-Flows

We use the term multi-flows to refer to workflows that can be instantiated to form different instances of workflows according to varying customer needs. Consider the example multi-flow represented in Figure 2 which has four input data sources, represented by A, B, C and D. The possible outputs are indicated by the nodes M, N and O. Depending on the output desired, one can choose to “ignite” a particular part of the multi-flow. If one chooses to get just the output of type M, the black colored part of the flow alone would be activated.

Figure 2. Example Instantiated Workflow

On similar lines, one could choose multiple output nodes, which would activate larger parts of the multi-flow. In general, multi-flows are meshes of operators consisting of intertwined workflows, one or more of which can be simultaneously activated as per the need at hand. Multi-flows are also useful to handle heterogeneity of data sources by providing a mechanism to instantiate different flows

based on the input format. This is explained in further detail in the extensibility section. Providing such multi-flows instead of providing a library of separate workflows is advantageous due to several reasons, some of which are as follows:

Computational Expense: If multiple related tasks are to be executed on the same data, it is useful not to repeat the basic tasks that are common across multiple tasks. For example, the result of the analytics node H is shared across J and K in Figure 2. Such sharing helps minimizing the computational expense, which in turn minimizes the total cost of running the multiple tasks, when users are billed by usage, as is common in utility computing scenarios.

Single View of Related Tasks: Multi-flows provide the user with a single view of all related tasks, in a single flow. Such a presentation helps the user in exploring the different functionalities on offer, by making small variations in the flow. Consider a multi-flow designed to identify association rules (i.e., correlated items that are often bought together) from a set of transactions, that also includes functionalities for contrasting between two association rule sets. A user who starts off by identifying associations from winter transactions will be implicitly prompted to choose the contrast workflow to contrast the winter associations against all-season associations to derive greater insights into customer buying patterns. While providing extra value to the customer, packaging related functionality can be used by the service provider as an up-sell opportunity.

B. Configurability

Many applications may require configurability for the operators in the workflow. These may range from choosing the number of segments in a customer segmentation task, specifying a list of items to be excluded in an association rule mining task, or even intricate configuration such as specifying the pruning conditions in a decision tree learner. We attach each parameterized operator in a multi-flow with input boxes to enable the user to provide values for the parameters to be used in the instantiation. To simplify matters, we pre-load such input boxes with the most likely values of those parameters so that a user who is not interested in fine-tuning the workflow can readily use it with the default values.

Table 1.Examples of Configuration Parameters

Operator		Parameters
Clustering ⁸		Number of clusters to be generated in the output
Predictive Learner	Model	Type of the predictive model to be learnt, e.g., generative or discriminative
Predictive Applier	Model	Type of output, whether single-class or multi-class
Association Mining ⁹	Rule	Support and Confidence

⁸ http://en.wikipedia.org/wiki/Cluster_analysis

⁹ http://en.wikipedia.org/wiki/Association_rule_learning

C. Extensibility

Consider an analytics task of predicting whether a customer is likely to be a high-value customer or a low-value customer, using the demographic attributes that the customer provides in the form she fills when signing up for the service. In order to build a classifier for this task, labeled data that provides insights into what low-value customer (and high-value customer) demographics separately look like, is necessary. We illustrate a simple workflow in Figure 3. Retail chains that use loyalty grading may already have *labeled user data*, i.e., customer IDs tagged with labels such as *high-value* and *low-value*. Others who do not use such loyalty grading would have to resort to rule-based categorization on *aggregated transaction data* such as labeling those who purchase goods worth at least \$100 a month on an average as *high-value*, and others as *low-value*. Demographic data is then merged with the labeled data using the customer id (or similar attribute); a classifier can now learn how the demographic attributes can be used to discriminate between *low* and *high* value customers. We enable the potential use of different data sources, as described in the example, by means of *attach-points* (or entry points) in the workflow. These are points at which appropriate data sources can be *attached* as illustrated in Figure 3.

Figure 3. Classifier Model Building

While the workflow is flexible enough to accept aggregated transaction data in the absence of labeled user data, a new customer who doesn't even have the ability to aggregate transaction data may not be able to use it. This is because, the only allowed entry point for transaction data assumes that transaction data are aggregated to have attributes such as total value of transactions per customer on a monthly basis (upon which rule-based categorization can be applied). In a typical analytics scenario, such a template workflow may have to be customized for the customer, for example, to replace the entry point for aggregated transaction data with one for raw transaction data that is piped to an aggregation operator before being passed on to the rule-based categorization node.

Our extensible framework allows the workflow to be extended to include an *additional* provision to allow for raw

transaction data. Such an extension to the workflow is illustrated in Figure 4. Apart from satisfying the needs of the new customer, this extension enhances the workflow even for existing customers who may now want to delegate the aggregation task to the workflow if that is found to be a better choice. In this particular example, the person who extends the workflow by including the additional attach-point need not be an expert in analytics and only needs to know the input expected at the entry to the *rule-based categorization* node. In particular, despite this workflow being concerned with building classification models, the person extending the model need not have any know-how of classification models; ***the abstraction of attach-points enables people with lower skill levels to modify workflows meaningfully.***

Figure 4. Extension for Raw Transaction Data Input

Our notion of workflow extensibility, as illustrated in the example above, is aimed at being able to seamlessly incorporate additional provisions into the workflow, and thus enhancing the applicability of the workflow for a wider range of customers. Instantiating a workflow from such a multi-flow involves selecting a attach point based on the data available with the user. On selecting an attach point, the input format gets fixed on that path from source to target and other points on the path get disabled. For example, if a user has aggregated transaction data, he would select that attach point and other points on that path will get disabled as shown in Figure 5. The advantage of a multi-flow is that a single multi-flow can represent many possible flows. Let m denote the number of source to target paths and n_i denote the number of attach-points on path i . Then a single multi-flow represents $\prod_{i=0}^m n_i$ possible workflow instantiations.

Figure 5. Instantiating an extensible multi-flow

III. SYSTEM OVERVIEW

Figure 6 shows a high-level overview of the system, by means of a block diagram, with the arrows indicating flow of data. The system broadly consists of a web application through which clients interact with the system; the web application maintains a set of multi-flows and is connected to several analytics engines and data management systems that perform analytics functionalities and manage data respectively. We will outline the typical operation of the system, and then go on to describe the responsibilities of the various components.

Figure 6. High-level System Architecture

A. Typical System Operation

A typical operation of the system involves several steps, which can be grouped into four distinct phases:

Data Upload and Setup: The customer who wants to run one of the workflows on her data first needs to upload the data into one of the data systems that the application has access to. Towards this, she can make use of the data upload functionality in the web application to upload the data into the chosen data system. Once the data upload is done, the data source shows up as one of the available data sources in the application. Then, the user can choose one of the many multi-flows that are available in the database of multi-flows depending on the kind of analytics task she wants to accomplish. Data sources among the ones already uploaded are then chosen and attached to the appropriate attach-points in the multi-flow. Further, particular flows that need to be executed are chosen from the multi-flow. This completes the setup, and the user is then ready to issue the command to run the workflow.

Workflow Execution: Once the user sets up a particular multi-flow for execution and issues the execute command, the web application looks up the mappings of operators to engines that are stored within its configuration files and figures out the specific commands to be issued to each of the analytics engines. It may be noted that a particular workflow may contain operators available within multiple analytics engines. For example, a workflow that consists of aggregating transaction data followed by identifying association rules may have its mapping set in a way that the

aggregation operator is tied to the SPSS aggregation functionality and the association rule mining task maps to the SAS association rule mining operator. The web application co-ordinates the execution of the workflow by issuing commands to appropriate analytics engines and collecting results to pass on to subsequent downstream operator commands. The analytics engines have access to data management systems and can fetch the data as required directly from the appropriate system.

Reporting Results: Upon completion of workflow execution, the results are made available for download from the web application. In case the results are voluminous as is typical in scenarios such as running a classification model over thousands of objects, the workflow may be configured to dump the results in one of the data management systems available to it. Upon the completion signal, the user can then navigate the system and access the analytics results from the appropriate system.

Administrative Tasks: Apart from the tasks that form part of the typical system operation, the system offers several intuitive provisions for administrative tasks such as adding new multi-flows to the system, extending multi-flows by adding new operators, changing the operator-analytics engine mapping and so on.

B. Component Responsibilities

We now detail the responsibilities and functions of the various components/actors in the block diagram in Figure 6.

Clients: Clients are users of the system who interact directly with the web application. Clients may use the system by performing one of the many tasks listed below:

1. Uploading data into the system, to be stored in one of the many data systems available
2. Choosing one of the multi-flows available in the system
3. Configuring a chosen multi-flow by fine tuning the parameters of the various operators
4. Choosing specific paths in the multi-flow that are of interest to a task at hand
5. Attaching data sources to one or more of the several data attach-points in a chosen multi-flow
6. Giving a command to execute one or more flows within a multi-flow
7. Downloading results of the execution of an analytics workflow

Web Application: The web application is the primary way through which users interact with the system. It presents the user (client) with options to perform the activities necessary to run a specific workflow on the user's data. This involves presenting appropriate multi-flows from the database, allowing the user to choose parts of it for execution and providing options to attach data sources. The web application takes control when the user hits the submit button and subsequently co-ordinates the execution of the flow; this involves issuing commands to appropriate

analytics engines for specific tasks, collecting results and piping them to the next operator, eventually completing the workflow execution in totality. The results are then presented to the user in the web application, either as a file to download, or as instructions on how to access the same from one of the data systems.

Analytics Engines: The applications comprise various instances of analytics engines, each of which has a set of functionalities. These could include instances of commercial analytics engines such as SPSS and SAS, and toolkits of academic origin such as Cluto¹⁰ or Weka¹¹. An analytics engine takes a command from the web application to run analytics tasks on data that reside in one of the many data systems, and accordingly executes the command. The results too are dumped into the location in a data system as indicated in the command, or at a default location.

Data Systems: Data systems have the primary responsibility of storing data. Data is fed to them from the web application when the user uploads data, or by the analytics engines to store intermediate/final results of analytics tasks; data systems are made readily accessible to the analytics engines, as indicated in Figure 6.

IV. CASE STUDY: RETAIL

We built a full-fledged operational prototype of the system focused on the retail domain. Towards this, we surveyed many retail vendors in India to validate our assumptions about analytics penetration and deterrents of usage of analytics in developing regions. Based on the survey, we analyze the common analytics needs of retail vendors in India and demonstrate that such commonly required analytics functionality can indeed be captured across a few configurable multi-flows. We illustrate how configurable multi-flows can be leveraged to develop a cloud-based analytics solution for the retail market.

A. Survey of Indian Retail Customers

Our survey targeted several retail majors in India; we contacted the managers of six major Indian retail chains that sell groceries and food among other household items, and several other smaller stores. Our questionnaire focused on eliciting information to answer the following questions:

- Can analytics help address the top business priorities of Indian retail players?
- Is analytics penetration low among Indian retail players? If so, why?
- Are cloud-based analytics solutions likely to help improve penetration?

- Can common analytics needs of retail customers be captured across a small number of multi-flows?

Analytics and Top Business Priorities: Based on the survey, we list the top-4 business priorities among Indian retail players in Table 2. Though cost reduction, as expected, figures high in the list, it was interesting to note that most retailers were as much or more interested in exploring new ways to reach customers. Out of these four, the top two were found to be areas where analytics is highly relevant; cost reduction can be directly addressed by analytics that supports and improves decision making whereas analytics can also help effectively utilize new ways to reach customers by enabling highly-targeted marketing.

Top Business Priorities in Indian Retail	
1.	New ways to reach customers
2.	Cost Reduction / Improved efficiency
3.	Improved customer service
4.	Increased employee productivity

Table 2. Survey Results on Business Priorities

Reasons for Low Analytics Penetration: Most of the survey respondents univocally admitted that they were yet to use any sophisticated analytics tools. In Table 3, we tabulate the top reasons (along with percentage of respondents who put forth the reason) behind non-usage of analytics.

Obstacles to Analytics Usage	
Lack of know-how about how analytics can improve business	38%
Competing priorities	34%
Lack of skills internally	28%
Ability to get digitized data	24%
Concerns about Information Sharing	23%

Table 3. Survey Results on Obstacles to Analytics Usage

The survey results mostly confirm that the primary entry barriers include lack of analytics skills among the workforce (and indirectly, the reluctance to setup an analytics team), competing priorities and lack of digitization of data.

Can Cloud Analytics help? Of the top-5 reasons in Table 3, cloud-based analytics-service can help alleviate the top-3 reasons. Cloud-based analytics makes the lack of know-how and in-house expertise a non-issue. Also, the reduced cost may help analytics edge out some of the competing priorities. Thus, cloud-based analytics is helpful in addressing some of the primary obstacles to analytics penetration. While overcoming the obstacle of digitization of data is a necessary pre-requisite to use any sort of analytics, cloud-based analytics will only aggravate the skepticism about information sharing. To address the concerns around data privacy and the skepticism about sharing data, privacy-preserving approaches to data sharing in cloud computing need to be developed (e.g., [6-8]). These can enhance the customers' confidence to share data and

¹⁰ <http://glaros.dtc.umn.edu/gkhome/views/cluto/>

¹¹ <http://www.cs.waikato.ac.nz/ml/weka/>

thus enable a cloud-based analytics solution to make headway into the SMB market.

Common Analytics Needs: Given a low-analytics penetration and ignorance about analytics capabilities, in addition to us seeking answers to open-ended questions such as “What are the top-3 tasks that you would want analytics support for?”, we ended up first educating the respondents about common capabilities analytics can bring to the table. Based on such responses, we categorized the top analytics functionalities into four broad tasks, as listed in Table 4. We will illustrate in a later section that each such task may be modeled in its own multi-flow.

Most Demanded Analytics Functionality
1. Customer Segmentation
2. Targeted Marketing
3. Demand Forecasting
4. Association Rule Learning

Table 4. Survey Results on Analytics Functionalities

B. Multi-flows in the Retail Prototype

Based on the survey findings, we built a fully functional prototype of the cloud analytics system with configurable and extensible multi-flows. We implement multi-flows for the following functionalities:

- Unsupervised Customer Segmentation
- Supervised Customer Segmentation
- Association Rule Learning

We delve into details of these workflows in the following sub-sections. We will present a screenshot of each of the workflows, and describe the finer details of the overall system, in a subsequent section.

1) Unsupervised Customer Segmentation

This flow (Fig. 8) takes customer demographic data and aggregated transaction data and segments users into multiple groups, the number of groups being specified by the user. Customers with similar aggregated transaction attributes (e.g., similar monthly spends) and similar demographics are likely to be placed in the same group. We use the popular K-Means clustering algorithm [9] for segmenting customers into the specified number of groups (i.e., clusters).

2) Supervised Customer Segmentation

This flow (Fig. 9) is very similar to the example illustrated in Section II C. The idea is to be able to predict a customer’s likely value in terms of well-known labels (e.g., silver, gold, platinum customers etc.) based on just the demographic data of the customer; demographic data is the only data available at the time the customer signs-up for a loyalty program. Such prediction involves functionality to learn a predictive model based on either labeled customer data, or rules that are applied on aggregated transaction data. The learned model can then be used on the demographics of a new customer to predict on the likely value of the customer.

Figure 7. Unsupervised Segmentation Workflow

Figure 8. Supervised Segmentation Workflow

Figure 9. Association Rule Learning Workflow

3) Association Rule Learning

The task of learning correlations between products in purchase patterns¹² is a popular analytics task on market basket data since it was proposed in 1994 [2]. This involves automatically learning correlations such as $\{onions, potatoes\} \rightarrow \{burger\}$, which indicates that customers who buy onions and potatoes together are likely to also buy burgers. Our workflow (Fig. 9) supports fine-tuning of association rule mining with provisions to restrict rule learning to transactions containing specific products, or transactions of customers that satisfy a specific condition on demographic attributes.

C. Configurability and multi-flows in the Retail Prototype

Configurability: There are two configurable operators among the multi-flows in our prototype. For the K-means clustering operator, the number of clusters is a configurable parameter. As can be seen in Figure 7, an edit box enables the user to choose the number of clusters for the segmentation. The association rule learning operator in Figure 9, apart from having provisions to specify common

¹² http://en.wikipedia.org/wiki/Association_rule_learning

parameters such as the confidence and support, allows for the following fine tuning:

- *Antecedent Product*: Often, the intent of learning association rules is to figure out what kind of products can be bundled together to create a combo offer. In this case, one of the products (e.g., a new product whose sales need to be boosted) is already known, and the customer can choose to search for only those rules that have the selected product as the antecedent. This is enabled by a drop-down box titled *Antecedent* in Figure 9. The default option is to search for general rules including all products among antecedents.
- *Gender*: In cases where gender-based association rules are to be discovered, the user of the system has the option to restrict association rule learning only to those transactions that involve a customer of a specific gender, as against the default case where customers of all genders are included.
- *Age Bounds*: The user can also specify age bounds of customers, such that only transactions by customers within the selected age range are included during association rule learning.
- *Transaction or Demographic Rules*: Finally, the user can specify the type of rules mined. The options provided are to discover classical product correlations using the transaction data or to mine for demographic rules [10] which are rules that indicate correlations between demographic attributes; the rule $\{age:<30\} \rightarrow \{location:indiranagar\}$ indicates that most young customers of the store are from the Indiranagar area.

Transaction metadata based fine tuning of association rule learning, as supported by the gender and age bounds attributes in our prototype, may be used in applications to get fine-grained insights into buying behavior [11].

Multi-Flow: Due to the provision of different attach points for data, each of the three implemented scenarios are multi-flows; for example, choosing to attach aggregated transaction data automatically discards the aggregation operator from the flow. Figure 10 illustrates two distinct flows as two specific black rectangles in the supervised segmentation scenario. The bigger workflow, from the left, involves demographic data and transaction data that are merged to learn a model to classify customers. The workflow on the right accepts pre-classified and merged data that can be directly used to learn the classification model. We let the user select a workflow by his/her choice of datasets to attach.

We automatically infer which workflow from the multi-flow needs to be executed based on the data source attach-points chosen. At the beginning, all four drop-down boxes for attaching data sources (see Figure 10) are enabled; the moment the user chooses to attach a dataset at the “*labeled customer training data*”, the system infers the choice of the

right-side workflow and disables the three other drop-down boxes. Similarly, if the first data source attachment is to the “*aggregated transaction data*”, it infers that aggregation is no longer necessary and disables the “*transaction data*” drop-down box since there is no workflow that uses both these drop-down boxes. Such relations, and which drop-down boxes to disable on each action, are automatically inferred from a configuration file that comprehensively lists the workflows contained in the multi-flow. If the user, at any point, needs to revert a selection and get back to a state where all data source drop-downs are enabled, she can click on the “*reset*” button.

Figure 10. Two Flows

D. Other features of the system

In order to provide an end-to-end solution for cloud-based analytics with configurable and extensible multi-flows, our system implements several functionalities. Those features are implementation-specific and only tangentially related to the main theme of the paper. However, to complete the description of the system, we give a brief overview.

1) Analytics Engines and Data Systems Used

Though our implementation is flexible and configurable to add analytics engines and data systems as necessary by changing a few configuration files, we use an SPSS engine and a DB2 instance as the only analytics engine and data system respectively.

Figure 11. Data Sources Pane

2) Data Uploading

We have implemented functionalities for uploading data from a disk file; we accept files in the CSV format (comma-separated values) where the first line is used to denote the names of the attributes followed by rows of data. On the server side, uploaded files can be kept in file format (within the server’s file system), or can be materialized as a relational table in the attached DB2 system; indexes may be created as needed on those tables to aid analytical

processing. We call these “file data sources” and “table data sources” respectively. Our data sources pane has functionalities to create either table or file data sources by uploading a chosen CSV file, as shown in Figure 11.

Figure 12. View Creation Functionality (left) in operation (right)

3) Data Source Attachment

Each data attach-point has a specific *minimum schema* expectation to be able to make use of the data source attached. The minimum schema expectation makes some attributes mandatory. For example, a demographic data source is required to have a “CustomerID” attribute that uniquely identifies each customer. It may have any number of additional attributes; for example, in the unsupervised segmentation workflow, any additional attribute of a customer is used by the clustering algorithm to measure the similarity between customers. Often, the minimum schema attribute names may not exactly match the customer’s attribute names. For example, the customer may use “CustID” to uniquely identify a customer whereas the attach-point may be expecting “CustomerID”. We have incorporated a view creation facility that helps the customer to create a non-materialized view by mapping the attributes in her data source to those expected by the minimum schema. Once such a view is created, it appears as another data source, and can be chosen from the drop down list of data sources. Figure 12 shows how view creation functionality gives the user the option to pick an attribute from her source to the minimum schema attribute, in this case ID, CUSTID, TRANSACTIONDATE, and TOTALPRICE.

4) Downloading Results

We implemented the workflow execution as an asynchronous operation, where the user finishes the setup (uploading data, attaching data sources, choosing the workflow to be executed) and clicks on “submit”. When the results are ready, the web application alerts the user and provides a link to download the results file.

V. RELATED WORK

Literature related to our work may be broadly classified under three heads: (1) Cloud Computing for Services, (2) Analytics Services, and (3) Workflows for Analytics.

Cloud Computing for Services: Cloud computing is the umbrella term used to denote the delivery of computing as a service rather than as a product. Cloud computing provides shared resources, software and information to computing devices as a utility over the network. It has become widely popular over the last few years and has been assessed to be no less influential than any of the previous trendsetting technologies [13]. Though many obstacles to the usage of cloud computing have been identified in [12], the same report emphasizes that the utility of cloud computing is likely to outweigh its disadvantages. Cloud computing provides many advantages such as elasticity and transference of over-provisioning and under-provisioning risks. Forecasts [14] suggest that cloud computing would grow at a very fast rate over the next few years. Cloud computing has been used for a variety of tasks such as parallel data mining [15] and graph analytics [16].

Analytics Services: Though popular analytics engines such as SAS, SPSS and R started off as stand-alone applications that need to be installed on a server to be used, recent trends suggest that commercial analytics engines are not far off from delivering full-fledged analytics capabilities on the cloud. Predixion Software¹³ and Zementis¹⁴ offer scoring of any PMML¹⁵ model over the cloud, whereas SAS has started to offer many of their functionalities over the cloud [17]. However, to the best of our knowledge, there has been a very limited work on providing powerful analytics functionalities over the cloud that can readily be used without requiring considerable expertise in analytics and the specific tool to be used.

Workflows for Analytics: The main novelty of the solution proposed in this paper rests in the usage of configurable and extensible multi-workflows to make it easy to use analytics despite limited technical know-how. Analytics workflows themselves are not novel and are the primary means of using SPSS (with the SPSS Modeler component). However, a typical SPSS workflow is designed to be used by personnel who understand the semantics of typical data mining operations. In particular, SPSS provides a fleet of functionality that is powerful enough to be used across domains, and is not specific to the scenario in which it may be deployed. For example, the semantics of the CustID field is unknown to a generic clustering operator, and under a default setting, the clustering operator would use the CustID field also for clustering leading to the undesirable results of

¹³ <http://www.predixionsoftware.com/predixion/>

¹⁴ <http://www.zementis.com>

¹⁵ http://en.wikipedia.org/wiki/Predictive_Model_Markup_Language

grouping customers with numerically close CustIDs together. We eliminate the need to explicitly specify the exclusion of such attributes, by folding in the semantics of attributes such as CustID (among others) into the pre-packaged workflows. In a pre-packaged workflow for customer segmentation, the fact that CustID needs to be excluded in clustering would already be already pre-configured. The customer can thereby focus on configuring other parameters that are likely to affect business such as number of clusters in the output or the rules to use in designating customers as high-value or low-value. The concept of multi-flows, where a single template workflow can instantiate many possible workflows, has not been developed in prior literature to the best of our knowledge.

VI. CONCLUSIONS

Analytics, despite its apparent utility in improving business has not been adopted widely. The main deterrents include the large upfront cost of procuring computing infrastructure and the cost of setting up an in-house analytics team of statisticians and business analysts. Such costs are so high that many small and medium businesses cannot afford traditional analytics solutions, especially in the developing regions of the world due to a spectrum of other deterrent factors that we have outlined. Many such deterrents may be effectively eliminated by the delivery of analytics services on the cloud. In this paper, we propose **configurable pre-packaged multi-flows** (flows that span multiple workflows) to deliver analytics services on the cloud towards helping to improve analytics penetration.

Multi-flows bundle together related workflows enabling users to explore and exploit related functionality by making minor configuration changes. We illustrate how changes to workflow configurations may be made easily without any in-depth knowledge of the various operators involved. Our multi-flows are easily **extensible** to non-disruptively include new functionalities in response to the changing needs of a customer (or a changing customer base).

We described a survey of Indian retail players that helped us identify (and validate our assumptions of) the primary obstacles to analytics penetration in the sector; we also analyzed the major analytics needs of such businesses. We illustrated that many such common analytics needs may be effectively captured within a small number of workflows. We developed a fully-functional system that implements the chosen multi-flows for the retail sector. We detailed the various features of the system, the multi-ness of the flows in the system, and example configuration settings that can be used to fine tune specific multi-flows. Through generous usage of screenshots, we illustrated the ease of use of our system; we expect that such enhanced convenience of usage will aid penetration of analytics into new customer bases.

Future Work: We are exploring further enhancements to improve ease of use of analytics solutions on the cloud

including an automatic mapping of attributes in data sources to attach-points by exploiting the data type and distribution. Also, making analytics widespread would necessarily involve addressing some of the problems that are applicable to cloud computing in general. These include concerns about data privacy in the cloud and the loss of control of the computation [18]. Developing techniques to effectively and provably anonymize of data without affecting the results of specific operators such as clustering or classification is an interesting direction to pursue.

ACKNOWLEDGMENT

We thank Abhishek Subramanian, Neethu Suresh, Shubhendu Aggarwal and Varun Srivastava for efforts in implementing the system and performing market surveys.

VII. REFERENCES

- [1] Todd Morrison, "Retail chain uses in-house analytics team to run SAP BusinessObjects", TechTarget Search SAP, 4th March, 2011 <http://searchsap.techtarget.com/news/2240033086/Retail-chain-uses-in-house-analytics-team-to-run-SAP-BusinessObjects>
- [2] Rakesh Agrawal, and Ramakrishnan Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", VLDB 1994, Chile
- [3] "How to structure a business analytics team", <http://mybusinessanalytics.blogspot.in/2011/09/business-analytics-team-structure.html> Retrived March 2, 2012
- [4] MA Rappa, "The utility business model and the future of computing services", IBM Systems Journal, 43(1), 2004
- [5] Brian Hayes, "Cloud Computing", Comm. Of the ACM, 51(7), 2008
- [6] Vibhor Rastogi, Dan Suciu, Sungho Hong, "The boundary between privacy and utility in data publishing", VLDB 2007
- [7] Zhou et. al., "Security and privacy in cloud computing: A Survey", International Conf. on Semantics, Knowledge and Grid, SKG 2010
- [8] Marten Van Dijk, Ari Juels, "On the impossibility of cryptography alone for privacy preserving cloud computing", 5th USENIX Conference on Hot Topics in Security, 2010
- [9] J.B. MacQueen, "Some methods for classification and analysis of multivariate observations", 5th Symp. On Mathematics, Statistics and Probability, 1967
- [10] G. Adomavicius and A. Tuzhilin. "Expert-Driven Validation of Rule-Based User Models in Personalization Applications." Data Mining and Knowledge Discovery, vol. 5, nos. 1/2, January/April 2001
- [11] Mahashweta Das, Deepak P, Prasad Deshpande, Ramakrishnan Kannan, "Fast Rule Mining Over Multi-Dimensional Windows", SDM 2011, Mesa
- [12] Michael Armbrust, "Above the Clouds: A Berkeley View of Cloud Computing", Technical Report, UC Berkeley, 2009
- [13] Daryl Plummer, David Cearley, David Smith, "Cloud Computing Confusion Leads to Opportunity", Gartner Report, 2008
- [14] Adam Holt et. al., "Cloud Computing takes off", Morgan Stanley Blue Paper, May 23, 2011
- [15] Joos-Hendrik Böse, Artur Andrzejak, Mikael Höggqvist, "Beyond online aggregation: parallel and incremental data mining with online Map-Reduce", MDAC 2010 Workshop on Massive Data Analytics over the Cloud
- [16] Padmashree Ravindra, Vikas V. Deshpande, Kemafor Anyanwu, "Towards scalable RDF graph analytics on MapReduce", MDAC 2010 Workshop on Massive Data Analytics over the Cloud
- [17] "SAS® for Customer Experience Analytics available via cloud", <http://www.sas.com/news/preleases/CXACloud.html>, accessed 13th Mar, 2012
- [18] "Stallman: Cloud computing is stupidity", http://news.cnet.com/8301-1001_3-10054253-92.html, accessed 14th March, 2012