Formal Methods for Data Analysis

Blagovest Kirilov[1]

Abstract – **In this paper formal methods for data analytics will be discussed upon. The main goal is to calculate an analytical system that can formalise methods for data analysis.**

Keywords – ***Data Analytics, Formal methods***

I. Introduction

Data analytics is getting more and more important for business, science as well as ordinary people. Because of the abundant sources of information that flood the world with it through various channels this leads to an ever-growing demand for analyzing this data and filtering it for effect. Still, although, there exist numerous methods that can execute this task, there is no formal method of defining and using them. This means that the approach to each analysis is different and thus can also lead to a result of wrong data or increased time and resource costs in order to receive the results.

II. Data analytics methods

Information is the core of all success stories as well as failures in the current age of technology. In the ever-changing daily drive there is a requirement for secure information sources, as well as fast analysis of provided information. In the last ten years a new term has been established, namely Big Data. This is the description of techniques and technologies that require new integration methods, in order to find hidden values and data in complex and massive sets of data. They not only analyze and find those, but also sort and order them for best effect.

This information management systems were and are still seen by many businesses as a hindrance rather than a supporting function that gives them new opportunities. Surprising is the fact that even today some businesses refuse to switch to such systems, in the process wasting more than half of their IT budgets just to keep the old systems alive and running, albeit at high cost and high impact. The rising flow of data that flow between interconnected companies, the internet, as well as the public domain, coupled with official state prescriptions, led to companies using such systems for cost reduction on data analytics. In order for this to happen, existing systems and methods are "fixed" to work with the new systems and methods, but without any underlying architecture, process or outcome expectation set-up.

In order to understand the key of this publication on a better scale, basic methods for data analytics and their development in the last ten years is reviewed, as well as a comparison of their key features.

A. Big Data

Even ten years ago big data was in the background of all future-talk on data analytics. The then used method of choice was Business Intelligence, where Business Intelligence was finding trends through descriptive analysis of big data sets with high data density. On the other hand, Big Data is using inductive statistical dependencies for data with lower data density. In addition, Big Data creates forecasts for output data as well as behavior models. Base descriptors of Big Data are the Mass (big datasets), Diversity (data in various formats), Velocity (dataflow) and accuracy (confidence in correct data).

"Actual forecasts say that the global data volume will double itself each year." [1]

This trend, coupled with advances in hardware technology, shows clearly, that data must be handled faster, because their volume is rising constantly. Software and the methods for this are getting more and more crucial in order to achieve a constant flow and analysis result as well as to be scalable. The final goal is to have a way to analyze and present Data in real-time.

B. Base methods for data analysis

One of the widely used methods is the so-called linkage analysis – this method analyses data for opportunities on linking it together, combining it in the so-called life-cycles, which are returned as an output of categorized data.

Another method is Data Mining. Data Mining searches for re-occurring events in data-sets of structured data. Those can also be combined in new patterns.

The last observed method is the predictive analysis. For this method the main difference in comparison to the previous ones is that the data is not structured and coming in in big volumes. This input happens mainly in real-time, and the method looks for patterns in the different data-sets and their data. The end-result is prepared data for further analysis which is combined in separate entities based on the preliminary analysis.

All methods serve their purpose in providing data analysis to different data sets with different results, although there is no clear distinction on when each one of them should be used in reality. As they treat data differently, this also leads to

[1] Blagovest Kirilov is with the Faculty for German Engineer and Business Management Education at the Technical University of Sofia, 8 Kl. Ohridski Blvd, Sofia 1000, Bulgaria.

different outputs and time effort required for each one of them.

C. Comparison of the methods

We can split the methods in three types – descriptive, predictive and prescriptive.

The descriptive methods, such as the linkage analysis, rely on already existing data in order to use milestones for certain events, thus splitting the data in smaller units, which are linked logically to each other, leading to clear outputs. Those methods rely on input data, which are available at an earlier stage, in order to give an output. This is getting harder with the complex flow of input data rising in real-time, because there can be new data types that are created during this real-time flow, who make previous analyses irrelevant. With this, the time required for the execution of those analyzes is directly linked to the amount of incoming data. In addition, if the dataflow is really high, the analysis needs to be repeated and re-arranged every time, which also has a direct effect on the effort. Because of the constant need for state-of-art data-sets the accuracy of the data is also not guaranteed, which has a negative effect on the end-result. Descriptive methods also have a relatively low impact on the optimization of data, because a data reduction is only possible if the only data that is used is historical one, which is not refreshed in real-time.

Predictive methods, for example Predictive Analysis and Data Mining, analyze historical as well as new data-sets, in order to find patterns, that can be used as predictions for future events. Those methods can not say what will happen in the future, they only point at what could happen. This is done by deriving data, which is not gathered, from the one already available. For those methods the time for execution is similar to the one of the descriptive methods, because predictive analyses that are done for future data are not required to be executed again if the future data matches the prediction. Predictive methods are generally used for data reduction purposes, because from them the data can be derived that will flow in in the future, allowing for models for their analysis to be provided, thus having a faster analysis time.

The last type of methods, in this case specific examples of Predictive Analysis and Data Mining, named prescriptive methods, are seen just now as being fully functional. Their goal is to not only provided an analysis on given input data, but also to go a step beyond by presenting possible actions and their effects. The prescriptive methods are a further development of the predictive methods, by saying what exactly needs to be done in order to reach a set goal. They also don't give just one way of achieving this, but multiple branches with different actions, depending on the future data coming in, thus providing actions for every situation that can arise. In order to execute a prescriptive method, a predictive method is required as well as two additional parameters – data, for which actions need to be put in place to influence the results, as well as a system which is documenting the influence of taken actions on the data.

For those systems the data mass is determined not only by the input, but also from the different branches, actions and resulting events, because all of them need to be analyzed in parallel. Thus, it is very important to set up a high data reduction, which can be controlled by the incoming feedback, in order to prevent non-critical and unneeded data set amounts. Furthermore, the accuracy of data is extremely important for such methods, which also leads to a longer execution time compared to the other methods. Still, this is a pointer in the direction of intelligent systems for data analysis giving proven and accurate data as outputs, which also increases their usability. The dataflow doesn't influence the method as far as the rest, because here the future flow can be predicted according to the actions, which is not always the case for say predictive methods.

The following table shows the summary of the most important parameters in relation to the three types of methods for data analysis described prior. It also shows their influence on accuracy and data reduction, their execution time as well as the influence of dataflow, data variance and data quantity on the methods. This calculations also apply for all types of methods mentioned within the prior chapter.

TABLE I
BASE DATA ANALYSIS METHOD COMPARISON

	Execution time	**Influence on**	
		Accuracy	**Data-reduction**
Descriptive Methods	Very high	Medium	Low
Predictive Methods	High	High	High
Prescriptive Methods	Very high	Very High	Very High

	Influenced by		
	Dataflow	**Variance**	**Data quantity**
Descriptive Methods	Low	High	Very High
Predictive Methods	Very High	High	Very High
Prescriptive Methods	High	High	High

As can be seen from the table, there is no clear “winner” from the methods, but each one of them has it’s negative and positive sides. The user needs to decide which one to use. The challenge arises when the user is not aware how to proceed and thus can not rely on a formal method for data analysis in order to cover his needs. This can lead either to increased cost because of additional method set-up or repeated method execution, as well as to wrong data accuracy if the correct method to be used is not chosen. In addition, routine maintenance and future-proofing of such methods bound into client systems is not guaranteed, leading to additional costs and missed opportunities for using the data as a competitive advantage. This, coupled with the complex business environment and customer expectations, can be the decisive factor between staying in business or losing everything. Thus, the need for a formal method for data analysis becomes more and more apparent. It should not only reflect different data inputs, but also differentiate between them and update the output as necessary in order to reflect the best accuracy and dataflow process contunitity.

III. Formal method description

A. Assumptions

In order for the methods to be formalized, they need to be having a couple of base information descriptors provided. The first one is data flow, followed by data quantity as well as data mass. Those are basic definitions of the data inputs, and most methods existing today are also using those descriptors in order to be differentiated between each other[1].

In addition, due to the nature of the methods, they affect the data accuracy, data reduction and execution time of the given input set of data.

B. Method approach

Different methods that can be included in the formalisation are descriptive, predictive and prescriptive methods, as they all can be used upon the same base input data. The important part is the interaction between the data, methods and the resulting output, in order to have formalised rules for the whole process. Thus, an interface will be created based on an use-case that can define the formalisation process, accepting data from different entities in different types and forms, and applying the same basic calculations for the different types of data analysis methods.

The output should be based on this process and always give a result to the users on the receiving side, thus completing the formalisation steps set-up.

The approach taken for the method is described in the next point, where a use-case definition is set-up as a basis for the further rollout and description of the method.

C. Use-case definition

The following use-case is set forward for the method:

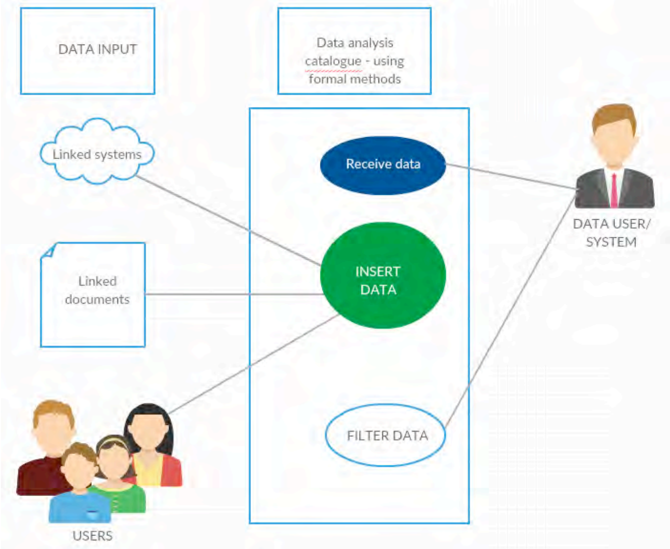

Fig. 1 Use-case for method description

In the above use-case scenario (Fig. 1), all types of data can be collected, regardless of their origin and type. Thus, this ensures the use-case is valid for all scenarios of data analysis.

The data analysis catalogue will pursue the following steps:

- Collect data from the input sources
- Analyze data for base descriptors[2]
- Apply data analysis to collected and pre-analyzed data
- Create output data and publish to end-users

Each of those steps is highlighted in the points below.

D. Data collection

Data collection should work regardless of input method. This is why the interface for the formalization will have input ports accepting all protocols and inputs. Error rates on data recognition should be non-existent, going through several checks to ensure data integrity. Data can be real-time or historical, structured or non-structured, or any combination of the ones above.

E. Data analysis for base descriptors

After collecting the input data, the interface will analyze the data for base characteristics and descriptors to match further algorithms. Thus, data can be sorted and prepared prior to the next step. Here, useless data pieces will be discarded, and the remaining data clustered into different category layers, thus producing a base process for identification, data quantity description, as well as highlighting criteria such as accuracy that can be obtained, mass of data and others.[3]

F. Data analysis application

Following the previous step, the data analysis application will have the following process (including previous steps)[4]:

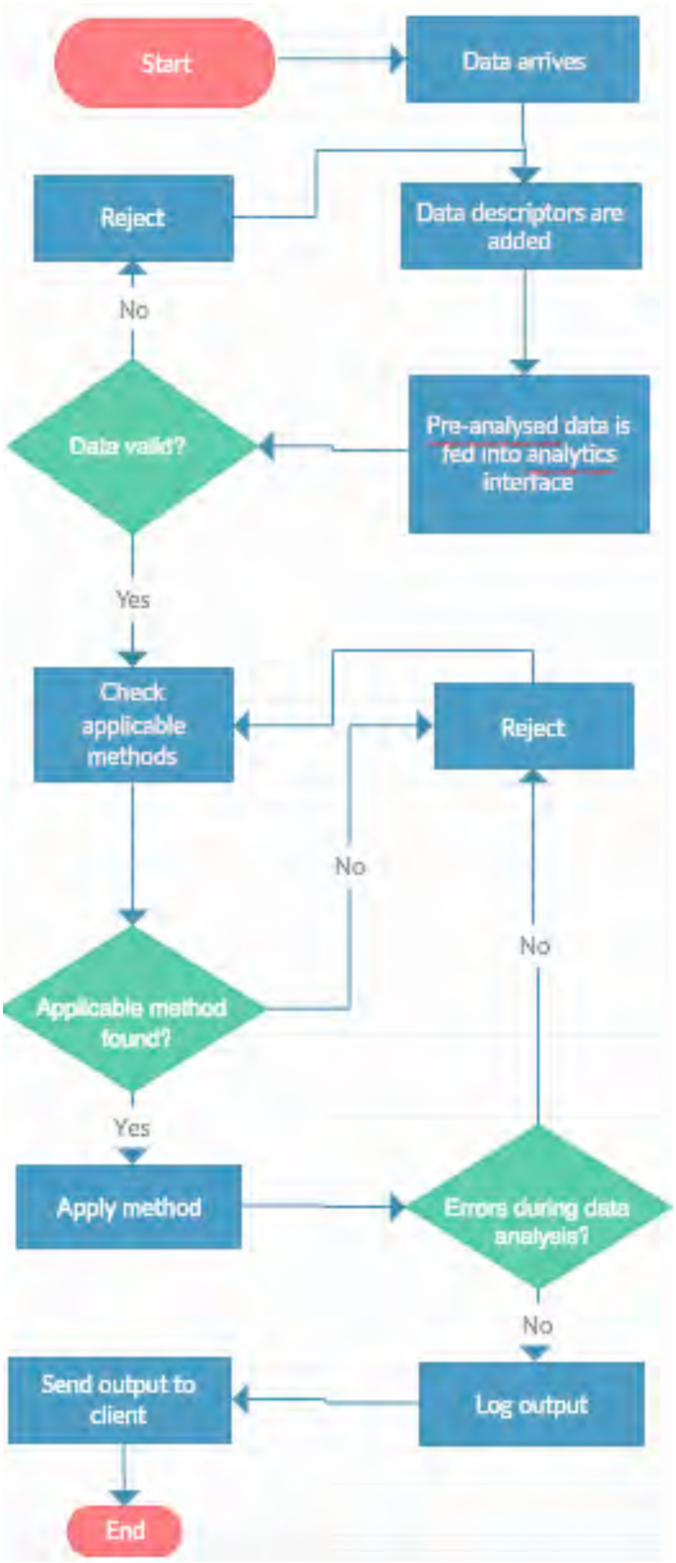

Fig. 2 Process flow for method description

In order for the process to have additional checkpoints, underlying documents for all checks should be created through testing and implementation. Each of those steps should be amended to the process flow as notes and additional documentation steps. In addition, the processing time, requirements as well as error rates should be measurable in order to provide inputs for accuracy and a proof of concept. The interface should be also created in a software implementation that can differentiate and show the process flow in action with base input data.

G. Create output data and publish to end-customers

After the analysis is done and the formal method applied, the resulting data with the analysis should be packaged in different formats, so as to be input in different kinds of databases, linked to existing data for matching and further analysis or saved as a documentation for end-users and data operators. This ensures the process flow described in Fig. 2 will have an end-to-end approach, going from the input of the base data through the analysis and ending with the output of the results as per requirements set-up. This is an additional step after the formal method for data analysis is applied, but ensures practical usability as well as a well designed interface that is created during the previous steps.

IV. Conclusion

In this paper an introduction on methods used to formalize data analysis is laid out on the basis of use-cases[5]. A draft process flow model is outline that gives the base process steps for the formalization. Overlying on this, an interface should be created that will enable the rollout of the formalization for all types of data input, as well as all methods to be used. This model will feature end-to-end functionality, going from the pre-analysis and integrated descriptor checks of input data through applying data analysis methods through formalization and ending in user output. Given the development and importance of the topic for all industries and public areas of importance, this would give a solid foundation for a set-up of formalization to various methods, so as to have scalability, cost-effectiveness and future-proofing going ahead[6].

References

[1] IBM - Analytics: Big Data in production

[2] International Data Corporation (IDC) (2013), From: February 2013

[3] BUBECK, Dr., Uwe; KLEINE BÜNING, Prof. Dr., Hans (2014): Maschinelles Lernen.

[4] Webster, John. "MapReduce: Simplified Data Processing on Large Clusters", "Search Storage", 2004

[5] Hu, Han; Wen, Yonggang; Chua, Tat-Seng; Li, Xuelong (2014). "Towards scalable systems for big data analytics: a technology tutorial". IEEE Access 2: 652–687

[6] From Data Mining to Big Data and Beyond; By Gregory Piatetsky-Shapiro April 18, 2012;