International Journal of Engineering & Technology, 7 (4) (2018) 3607-3612

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.19653
Research paper

Data sources and ingestion big data layers: meta-modeling of key concepts and features

Allae Erraissi [1] *, Abdessamad Belangour [1]

[1] *Laboratory of Information Technology and Modeling LTIM, Hassan II University, Faculty of Sciences Ben M'sik, Casablanca, Morocco*
**Corresponding author E-mail: erraissi.allae@gmail.com*

Abstract

A deluge of data is to be expected in the years to come. Nowadays, huge masses of data is produced every day. For example, if we take only social network users and the Internet of Things, we shall find that they generate large volumes of varied data that have to be transmitted, recorded and processed at high speed. These data are an important source of information that can improve the performance of predictions. Hence, the data are no longer in net structures, easy to consume, but they are represented in different types of structures, namely structured, semi-structured and unstructured data. At the Big Data architecture level, these different data sources are located in the Data Sources layer, which is the starting point for any further processing of Big Data. Indeed, this layer has a direct relationship with the Ingestion layer, which is in charge of validating, transforming, cleaning, reducing and integrating data that can be used later on by the Hadoop ecosystem. In this paper, we applied techniques related to Model Driven Engineering "MDE" to propose a universal Meta-modeling for both Data Sources and Ingestion Big Data layers. These meta-models are platform independent according to Model Driven Architecture pattern, which describes the structures of Data Sources and Ingestion independently from any specific platform.

Keywords: *Meta-Modeling; Big Data; Data Sources Layer; Ingestion Layer; Model Driven Engineering; MDE.*

1. Introduction

In recent years, data has been growing at an unprecedented pace. Our computers, mobile phones, payment tools and the many sensors that now equip our cars, roads, and houses produce a huge amount of data. These data are passed on to thousands of Data Centers, which store, analyze and intersect. Therefore, we entered the Big Data era. According to our earlier research studies [7] [16] [23], we found that vendors offer ready-to-use distributions to handle a Big Data system, namely Cloudera [1], HortonWorks [2], MapR [3], IBM Infosphere BigInsights [4], Pivotal HD [5], Microsoft HD Insight [6], and so on. Each distribution has its own vision for a Big Data system. Up until now, we have done two comparative studies of the five major Hadoop distribution providers of Big Data: the first study [7] was on the components of these distributions, where we proposed five criteria for comparison. The second comparative study [16] [23] was on the architectures of the same Hadoop distributions. While handling this work, we proposed 34 criteria for comparisons. Our main objective is to seek and find the common points and the differences of the solutions proposed by the leaders of the Big Data market namely Cloudera, HortonWorks, IBM BigInsights, Pivotal HD and MapR M3.
However, programmers do not have the meta-models needed to create standard applications that can be compatible with each provider, because each provider has its own policy to design his own Big Data system. In this paper, we shall first present the architecture of a Big Data system, and then we shall use the concepts related to the world of the MDE [8] (Model Driven Engineering). Hence, our main goal is to propose eventually meta-models for the Data Sources and the Ingestion layers in order to standardize the world of Big Data. Developers can use these meta-models to build a common distribution.

2. Related work

According to the research studies we have done on the Big Data world, we found that several distributions could manipulate a Big Data system.
Correspondingly, in our previous works [7] [16] [23], we gave a detailed comparison of the top five big data solution providers. These comparatives studies along with the evaluation made by Forrester Wave [25] on the same Hadoop distributions helped us to define the key concepts of the two layers: Data Sources and Ingestion. We also rely on two other comparative studies made by Robert D. Schneider [26] and V. Starostenkov [27] on the three HortonWorks, Cloudera, and MapR distributions. Yet, this paper is an extended version of a work that we have already published in the proceedings of a conference [29].

3. Model driven engineering

Models have long been used in science and engineering as a fundamental tool for managing complexity. Modeling separates concerns by abstracting specific aspects of reality for specific purposes [10]. This approach has become relatively popular in recent years to deal with analytical and design concerns, relying in particular on modeling languages of the UML family [9].
Model-Driven Engineering (MDE) [17] intend to improve the development of complex systems by focusing on more abstract concerns than conventional programming through models. Thus, this engineering provides a methodological framework for real-time embedded system developers, which currently focuses on the development of abstract models, rather than concepts related to

algorithmic and programming. In fact, this figure below shows the levels of abstraction of the Meta-modeling in the form of modeling pyramid of the OMG:

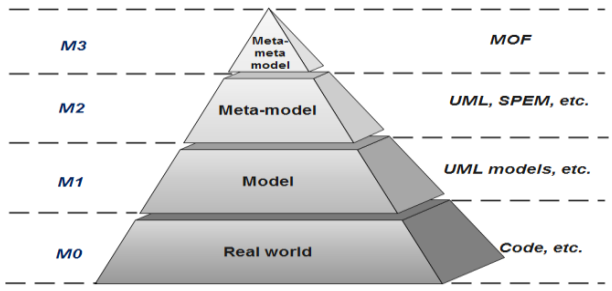

Fig. 1: Modeling Pyramid of the OMG [11].

The modeling pyramid of the OMG consists of:

- The M0 level: The real world is represented at the base of the pyramid; it is the reality one wants to model.
- The M1 level: Models representing this reality constitute the M1 level. A model is an abstraction of a system, modeled as a set of facts constructed with a particular intention. It can represent either a whole system (structure, behavior and non-functional properties) or just one aspect of the system by obscuring these other aspects.
- The M2 level: The notion of models in the MDE explicitly refers to the definition of the languages used to construct them. The language in which this model is expressed must, therefore, be explicitly defined. Researchers have called this modeling language a meta-model. A meta-model is a model that defines a modeling language [24]. It can precisely specify the concepts of a modeling language and establish the relationships between these concepts.
- The M3 level: Meta-Object Facility (MOF) [11] is a standard meta-model definition language to avoid the inordinate increase of various and incompatible meta-models. As a model, the meta-model must be defined from a modeling language called meta-meta-model. To limit the number of levels of abstraction, the meta-meta-model must then have the property of meta-circularity, that is, the ability to describe itself.

4. Architecture of a big data system

As we have mentioned in our first article [7], all the components of the Big Data architecture should be in place before proceeding to analyze all aspects of a large data stream. Indeed, with this correct configuration, you will boost performance by assembling all the needed components and, hence, succeed in processing large volumes of data.

This figure below describes the necessary components of the architecture of a Big Data system. You can choose open source frameworks or packaged licensed products to take full advantage of the features of the different Big Data components [12].

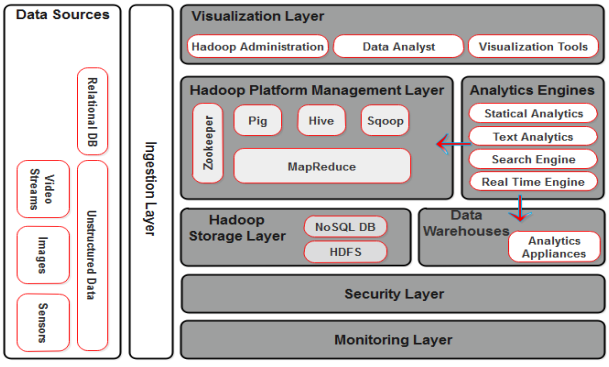

Fig. 2: Architecture of Big Data [12].

The architecture of a Big Data system consists of the following layers:

- Visualization
- Hadoop Platform Management (MapReduce [18], PIG [13], HIVE [14], Sqoop [15] and others)
- Hadoop Storage (NoSQL [19], HDFS [20])
- Security
- Monitoring
- Ingestion
- Data Sources

5. Meta-modeling of data sources layer

4.1. Data sources layer

The volume of information captured under the Big Data is considerable and the sources and formats of this data are very varied. There are two sources of information: internal or external. Data from internal sources are those that companies have collected and stored in a database. Because of this organized management, we are talking about structured data. Conversely, when information is not structured, we call it disparate data. It is mainly information from external sources, that is to say, all the data that one will be able to find and retrieve on the Web, but also all that information that companies collect in a "painless" way for the surfer. With regard to the various data formats, it is sufficient to look at the types of content collected, text, audio files, videos or images, etc. All of these documents are information that is part of Big Data.

Whether they are internal or external sources, all the data that an organization can collect will be useful to the management of that organization. Indeed, this information gathering through big data is an essential tool and an opportunity for any company to better target consumers 'needs and hence satisfy them. Accordingly, the company, which carries out continuous and efficient data management, will be able to address its current and potential customers. In fact, it is an organization's task to combine and decipher all the information it has gathered from different sources, to draw conclusions therefrom, and to perform a very precise segmentation of its clients. All of these actions will then allow her to make the appropriate decisions according to the situations that will arise to her. Thanks to the emergence of the Internet and Information and Communication Technologies in general, today we have at hand a kind of huge database that is Big Data. Data from Big Data is, therefore, a tool in its own right, a raw material that companies regularly use to get all the information they need to make the right decisions in their business.

There are three types of data sources within the Big Data:

- Unstructured data is a generic designation that describes any data external to a structure type. Unstructured textual data is generated by e-mail, Word documents, PowerPoint presentations, or collaboration or instant messaging software. Non-textual data is generated via media such as MP3 audio files, JPEG images, or Flash video files.
- Structured data: By structured data, we mean all kind of data that occupy a fixed field within a file. Developers have reformatted structured data in a way that all its elements have been recognizable thanks to a specific structure. This structure allows everyone to manipulate it according to the various combination for exploiting the information in a good way.
- Semi-structured data is an intermediate form. They are not organized in a complex way that allows sophisticated access and analysis, however, some information may be associated with them, such as metadata tags, which allow the addressing of the elements contained therein.

The real problem with the Big Data definition begins in the data sources layer, where the data sources are large. They have a different speed and they are varied. These data will be included in the Big Data sets to be analyzed at the end. This figure illustrates the different types of data sources.

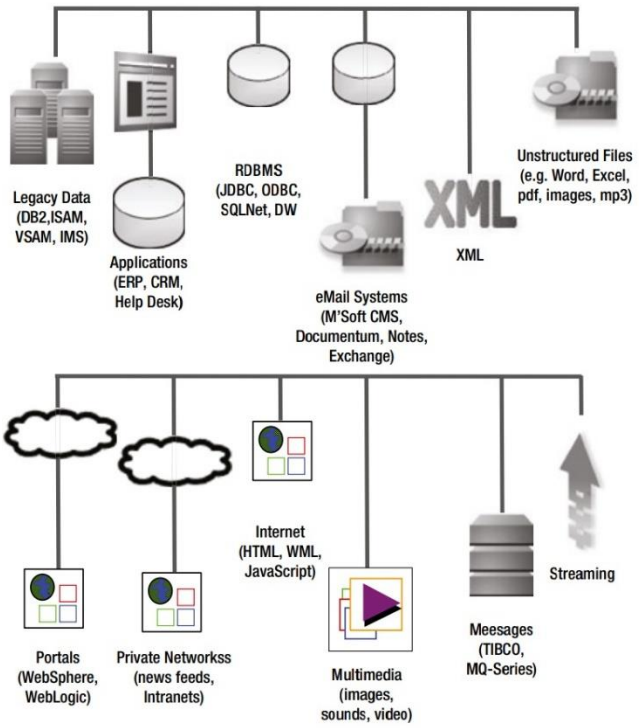

Fig. 3: The Variety of Data Sources.

Over the last ten years, each industry has experienced an explosion in the amount of incoming data [12] due to increased subscriptions, audio data, mobile data, contextual details, social networks, meter data, Meteorological data, operating data, device data and data usage, etc. For example, this figure groups the data sources for the "New Age" which have seen an increase in the volume, the speed, and the variety of the telecommunication domain.

Table 1: New Age Data Sources

New Age Data Sources		
High Volume Sources	Variety of Sources	High-Velocity Sources
Switching devices data	Image and video feeds from social Networking sites	Call data records
Access point data messages	Transaction data	Social networking site conversations
Call data record due to exponential growth in user base	GPS data, E-mail, SMS	GPS data
Feeds from social networking sites	Call center voice feeds	Call center - voice-to-text feeds

4.2. Meta-model for data sources layer

As already mentioned in the previous paragraph, we shall propose a meta-model for the Data Sources layer in Big Data that defines the Meta-modeling of the three types of data sources. These are structured, unstructured and semi-structured data.
Figure 4 is a representation of the meta-model that we propose. In fact, we defined these three types of data sources by the inheritance relationship of the meta-class DataSource, which has as meta-attribute "DataSourceID" of type Int. Accordingly, The meta-class DataSource contains the three types: Structured, SemiStructured, and Unstructured.

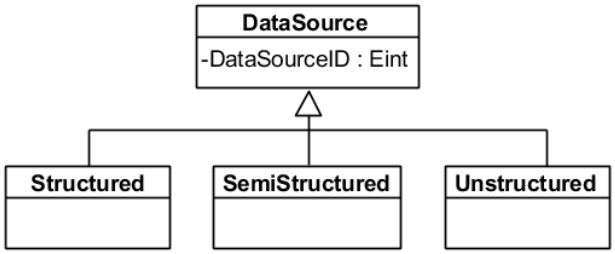

Fig. 4: Meta-Model for the Three Types of Data Sources.

4.2.1. Structured meta-class

First, there is the structured data whose set of possible values is determined and known in advance. For example, in a database gathering, the results of an opinion survey, the age or the socio-professional category of the individuals surveyed are structured data, because the age groups or the list of possible socio-professional categories are determined a priori. In Computer Science, structured data will be stored in relational databases or in databases dedicated to XML [30], called native XML.

4.2.1.1. Relational databases

A relational database is a collection of data organized in the form of defined tables from which we can access and assemble data without having to rearrange the tables in the database. The concept of the relational database was invented by E.F Codd then at IBM in 1970 [31]. This figure below presents the meta-model that we propose for relational databases:

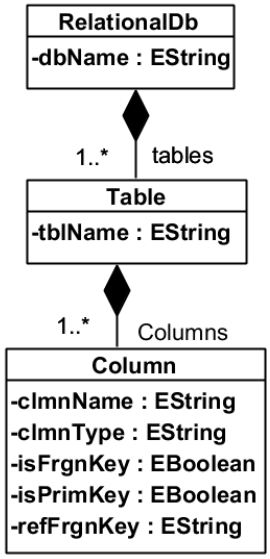

Fig. 5: Meta-Model for Relational Data Bases.

The proposed meta-model for relational databases consists of three meta-classes that are RelationalDb, Table, and Column. Firstly, the main meta-class RelationalDb has the meta-attribute 'DbName' of type String. A relational database can contain multiple tables. On that account, the RelationalDb meta-class in our meta-model consists of the meta-class Table that has the String 'tblName' meta-attribute. Finally, a table is made up of one, or more than one column. The meta-class Column has as meta-attributes "clmnName, clmnType, refFrgnKey" of type String, and "isFrgnKey, isPrimKey" of type Boolean. The meta-attributes "isFrgnKey, isPrimKey" make it possible to know if the column is a primary key or a foreign key in the table.

4.2.1.2. XML files

XML or eXtensible Markup Language is a generic markup language [30]. An XML file is a text file with a particular structure.

XML is a notation, that is to say, a way of writing information. It allows specifying the structure of the content of a document rather than the way to present it. Its markup design makes it possible to structure and represent information in the form of a tree. The meta-model proposed below by the W3C [32], describes an XML document:

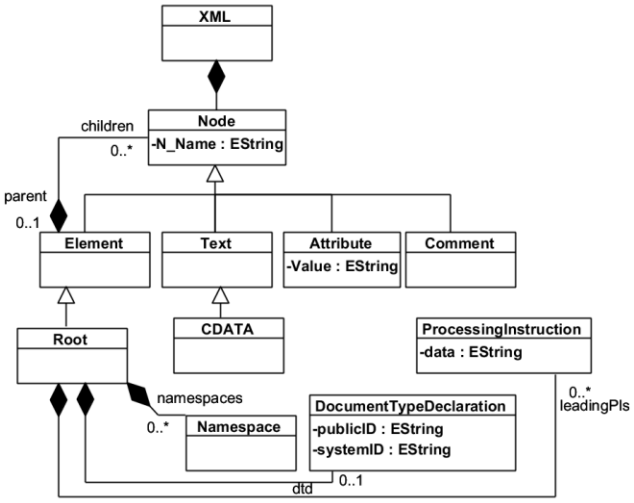

Fig. 6: Meta-Model for XML [32].

4.3. Semi structured meta-class

Semi-structured data are data that have not been organized into a specialized repository, as it is the case in a database. However, it contains related information, such as metadata, which are easier to process than raw data. The following figure shows the meta-model that we propose to define the semi-structured data in Big Data:

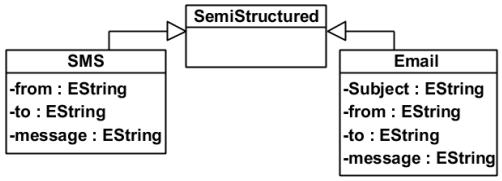

Fig. 7: Meta-Model for Semi-Structured Data.

4.3.1. Unstructured meta-class

Unstructured data is data represented or stored without a predefined format. They are typically plain text, but they can equally contain numbers, dates, and facts. Hence, unstructured data is a generic designation that describes any data external to a type of structure.

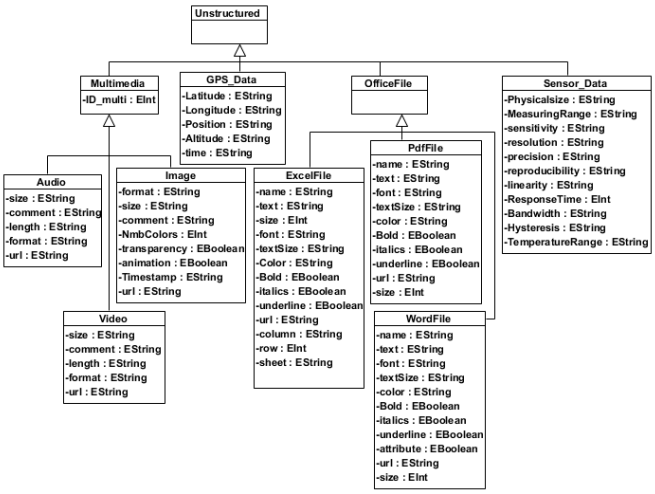

Fig. 8: Meta-Model for Unstructured Data.

The meta-model we proposed has the main meta-class named Unstructured that represents unstructured data. At the Big Data level, we find many data that have not a predefined format, for example, multimedia data, which consist of images audio, and videos. Accordingly, we have represented the multimedia data in our meta-model by the inheritance relationships of the Multimedia meta-class. We also presented GPS data and data captured by the sensors by the two meta-classes GPS_Data and Sensor_Data. Finally, we defined an OfficeFile meta-class to specify the Office files structure (Excel, Word, PowerPonit).

4.3.2. The generic meta-model for data sources layer

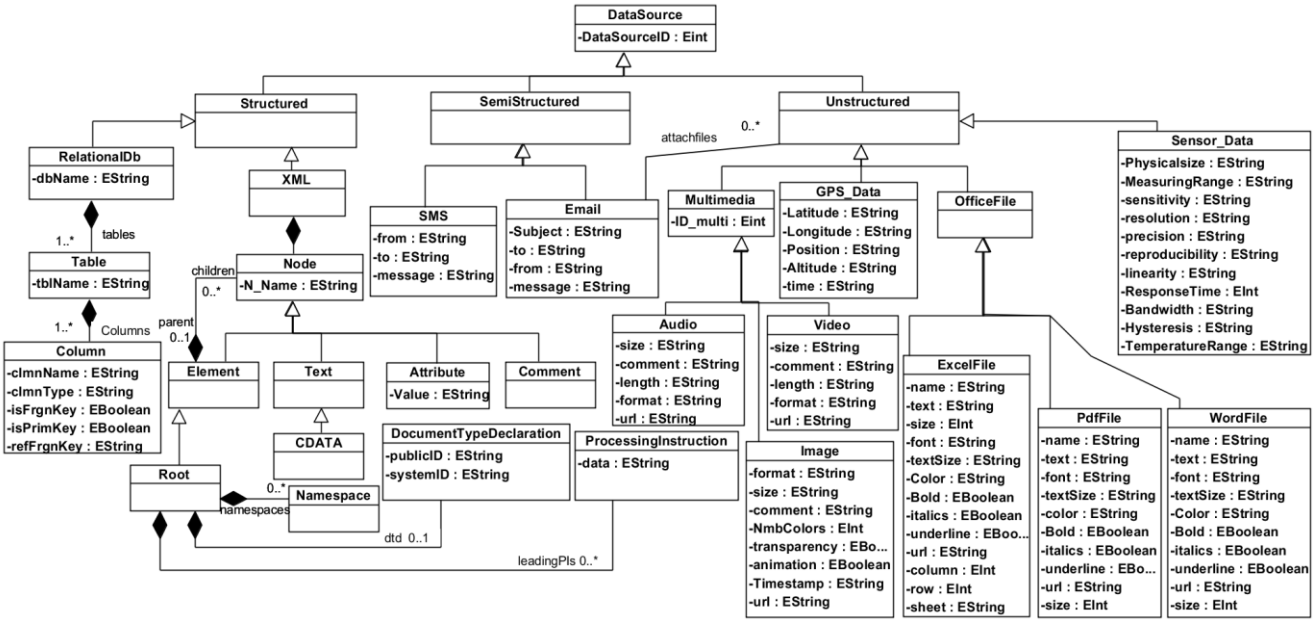

Fig. 9: Meta-Model of Data Sources Layer.

All data sources shown in this figure have to be channeled into the company after validation and proper cleaning. It is the job of the ingestion layer to provide the functionality of being rapidly scalable for the huge data flow.

6. Meta-modeling for ingestion layer

Data ingestion is the process of importing and obtaining data for storage in a database or immediate use. Ingesting something is "taking something or absorbing something".
This layer is responsible for separating the noise from the relevant information. The ingestion layer should be able to handle the enormous volume, high speed, and variety of data. It should have the ability to validate, clean, transform, reduce, and integrate data so that the Hadoop ecosystem can use them later. The new layer must be scalable, flexible, reactive and regulatory in the Big Data architecture. If the manager for software development has not well planned the detailed architecture of this layer, the entire technology stack will be fragile and unstable.

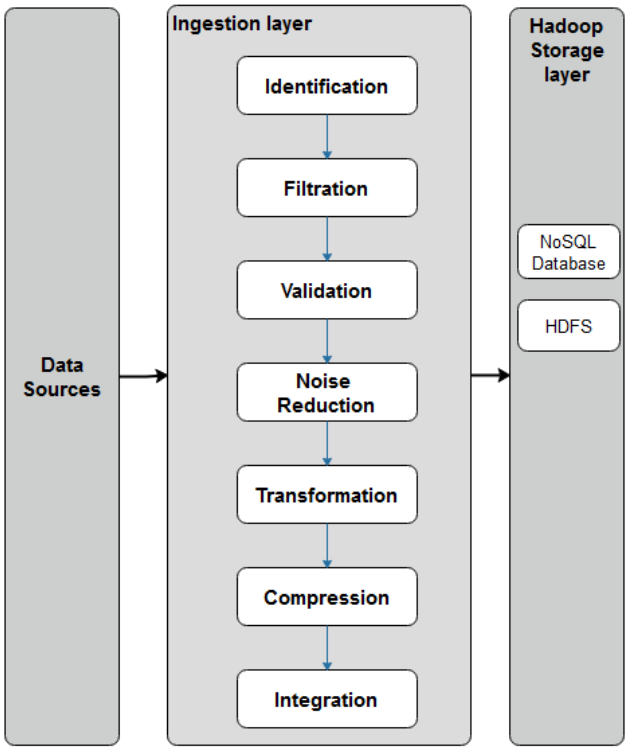

Fig. 10: Steps of Data Ingestion.

The ingestion layer loads the relevant final, noise-free information to the distributed Hadoop storage layer. It should have the ability to validate, clean, transform, reduce, and integrate data into the Big Data system for further processing. The building blocks of the ingestion layer should include the following:

- Identification of different known data formats or assignment of default formats to unstructured data.
- Filtering incoming information corresponding to the company.
- Continuous data validation and analysis.
- Noise reduction using cleaning methods to remove noise and minimize disturbances.
- The transformation may involve the convergence of data fragmentation, de-normalization or synthesis. Compression involves reducing the size of the data, keeping the data relevant in the process. It should not affect the analysis results after compression.
- Integration involves integrating all data into the storage layer, which consists of the HDFS and NoSQL database.

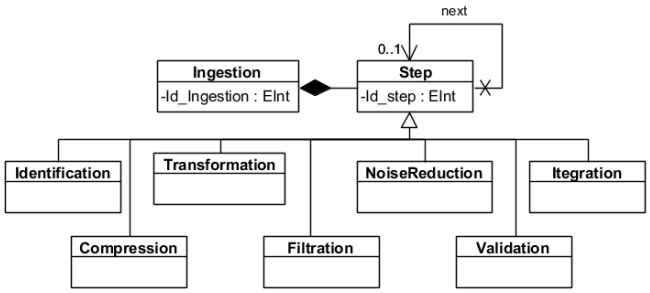

Fig. 11: Meta-Model of Ingestion Layer.

7. Relation between data sources and ingestion layers

According to the Big Data architecture, we find that there is a direct link between the data sources and the Ingestion layers. Then all sorts of data must pass through by the steps constituting the layer of Ingestion before being used by the other layers of the Big Data system. We have expressed this link with the following meta-package diagram that represents the two meta-packages IngestionPkg and DataSourcesPkg and the dependency relationship that links them:

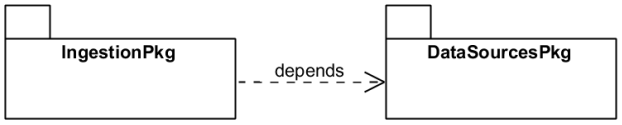

Fig. 12: Ingestionpkg and Data Source SPKG Meta-Packages.

Within these two meta-packages, we express the binding relation by a Mono-directional association that links the meta-class "Data-Sources" of the first meta-model with the meta-class "Ingestion" of the second Meta-model. Thus, Figures 13 and 14 show this relationship:

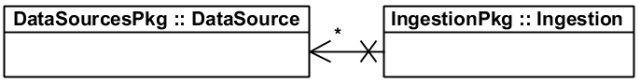

Fig. 13: Association between Data Source and Ingestion Meta-Classes.

After the creation of the two meta-models for the Data Sources and Ingestion layers, in the next step, we shall work on the creation of models respecting these meta-models. Then we shall define the transformation rules between these meta-models using the transformation language ATL (Atlas Transformation Language) [21] [22].

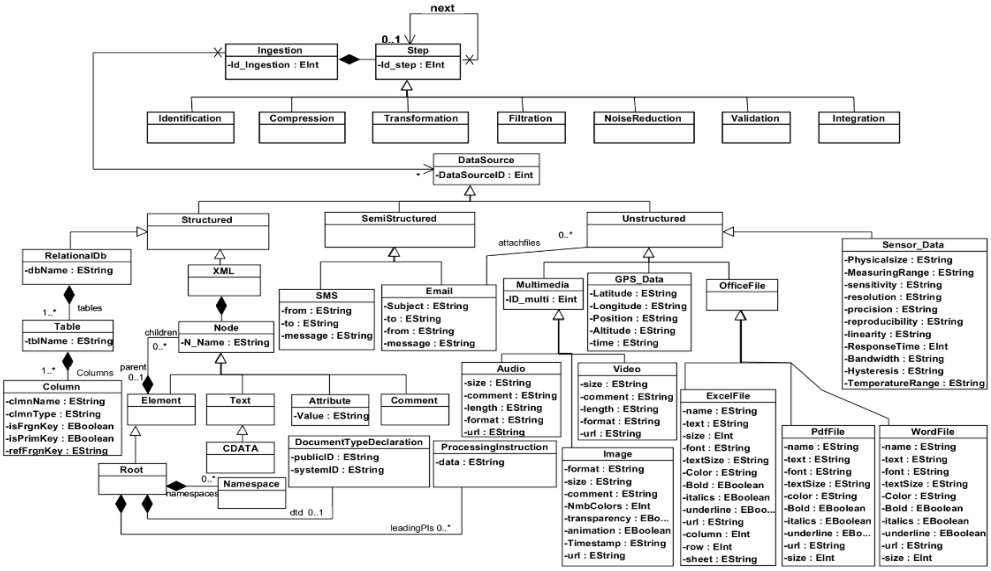

Fig. 14: Generic Meta-Model of Data Sources and Ingestion Big Data Layers.

8. Conclusion and perspectives

The volume of information captured in Big Data has become very large and the sources and formats of this data are very varied. As mentioned above, there are different types of data sources that can be structured, unstructured or semi-structured. Indeed, these data sources, as their name indicates, are a source of essential information, and an opportunity for any organization to better target consumers and satisfy their urgent demands. However, the real problem in Big Data begins in the data sources layer, where the data sources are large, varied and have a different speed. All these data sources must be channeled into the company after validation and proper cleaning. It is the job of the ingestion layer to provide the functionality of being rapidly scalable for the huge data flow. In this article, we applied the modeling techniques (Model Driven Engineering) by proposing meta-models for the two layers Data Sources and Ingestion, with the aim of standardizing concepts at the Big Data level.

In MDA (Model Driven Architecture) approach, the generic meta-model proposed for Data Sources and Ingestion Big Data layers can be used as an independent cross-platform Domain Specific Language [28].

References

[1] Menon, R. (2014). Cloudera Administration Handbook

[2] HortonWorks Data Platform HortonWorks Data Platform: New Book. (2015).

[3] Dunning, T., & Friedman, E. (2015). Real-World Hadoop.

[4] Quintero, D. (n.d.). Front cover implementing an IBM InfoSphere BigInsights Cluster using Linux on Power.

[5] Pivotal Software, I. (2014). Pivotal HD Enterprise Installation and Administrator Guide.

[6] Sarkar, D. (2014). Pro Microsoft HDInsight. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-6056-1.

[7] Allae Erraissi, Abdessamad Belangour, Abderrahim Tragha. "A Big Data Hadoop Building Blocks Comparative Study." International Journal of Computer Trends and Technology. Accessed June 18, 2017. http://www.ijcttjournal.org/archives/ijctt-v48p109.

[8] Royer, Jean-Claude, and Hugo Arboleda. Model-Driven and Software Product Line Engineering. [1] Edition. London, UK : Hoboken, NJ, USA: Wiley-ISTE, 2012.

[9] Rumpe, Bernhard. Modeling with UML: Language, Concepts, and Methods. First ed. 2016 edition. New York, NY: Springer, 2016.

[10] Jean Bézivin. La transformation de modèles. INRIA-ATLAS & Université de Nantes, 2003. Ecole d'Eté d'Informatique CEA EDF INRIA 2003, cours #6.

[11] ISO/IEC/JTC 1/SC 32. ISO/IEC 19502:2005, Information Technology - Meta Object Facility. Multiple. Distributed through American National Standards Institute, 2007.

[12] Sawant, N., & Shah, H. (Software engineer). (2013). Big data application architecture & a problem-solution approach. Apress.

[13] Gates, Alan, and Daniel Dai. Programming Pig: Dataflow Scripting with Hadoop. Two edition. O'Reilly Media, 2016.

[14] Capriolo, Edward, Dean Wampler, and Jason Rutherglen. Programming Hive: Data Warehouse and Query Language for Hadoop. [1] Edition. Sebastopol, CA: O'Reilly Media, 2012.

[15] Ting, Kathleen, and Jarek Jarcec Cecho. Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database. [1] Edition. Sebastopol, CA: O'Reilly Media, 2013.

[16] Allae Erraissi, Abdessamad Belangour, and Abderrahim Tragha, "A Comparative Study of Hadoop-based Big Data Architectures," Int. J. Web Appl. IJWA, vol. 9, no. 4, 2017.

[17] Roessler, Wolfgang. [(Model Driven Engineering for Safety Relevant Embedded Systems)]. VDM Verlag, 2013.

[18] Blokdyk, Gerardus. MapReduce Complete Self-Assessment Guide. CreateSpace Independent Publishing Platform, 2017.

[19] Raj, Pethuru, and Ganesh Chandra Deka. A Deep Dive into NoSQL Databases: The Use Cases and Applications. S.l.: Academic Press, 2018.

[20] Alapati, Sam R. Expert Hadoop Administration: Managing, Tuning, and Securing Spark, YARN, and HDFS. Boston, MA: Addison Wesley, 2016.

[21] "ATL: Atlas Transformation Language Specification of the ATL Virtual Machine."

[22] "ATL: Atlas Transformation Language ATL Starter's Guide," 2005.

[23] Allae Erraissi, Abdessamad Belangour, and Abderrahim Tragha, "Digging into Hadoop-based Big Data Architectures," Int. J. Comput. Sci. Issues IJCSI, vol. 14, no. 6, pp. 52–59, Nov. 2017. http://ijcsi.org/papers/IJCSI-14-6-52-59.pdf.

[24] OMG. Meta Object Facility (MOF) 2.0 Core Specification. January 2006

[25] Read, W., Report, T., & Takeaways, K. (2016). The Forrester WaveTM: Big Data Hadoop Distributions, Q1 2016.

[26] R. D. Schneider, "HADOOP BUYER'S GUIDE," 2014.

[27] V. Starostenkov, R. Senior, and D. Developer, "Hadoop Distributions ".

[28] Pastor, Oscar, and Juan Carlos Molina. Model-Driven Architecture in Practice: A Software Production Environment Based on Conceptual Modeling. 2007 edition. Berlin ; New York: Springer, 2007.

[29] Allae Erraissi, Abdessamad Belangour, and Abderrahim Tragha. "Meta-Modeling of Data Sources and Ingestion Big Data Layers." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, May 26, 2018. https://papers.ssrn.com/abstract=3185342.

[30] Key, Sam. XML Programming Success in a Day: Beginner?s Guide to Fast, Easy, and Efficient Learning of XML Programming. CreateSpace Independent Publishing Platform, 2015.

[31] E. F. Codd, "A Relational Model of Data for Large Shared Data Banks," 1970.

[32] J. D. Allen and Unicode Consortium., the Unicode standard 5.0. Addison-Wesley, 2007.