Proc. of the Third Intl. Conf. on Advances in Information Processing and Communication Technology - IPCT 2015

ISBN: 978-1-63248-077-4 doi: 10.15224/ 978-1-63248-077-4-96

Big data Migration for Telecom Business Support Systems

Ioan DRĂGAN **Răzvan Daniel ZOTA**

***Abstract*—Migrating data in Business Support Systems is never an easy job even when changing only one component. Migrating to a cloud solution that has limited flexibility in data structure needs more than a simple data migration approach. Multiplying that several fold, since each customer has its own data source, it requires an automated data recognition and migration methodology. Machine learning is an answer to data recognition, but applying it on data structures still requires some future development.**

***Keywords*— business support systems, bss, bigdata, Hadoop, machine learning, data migration**

I. Problem description – BSS data migration

In one of the simplest forms, business support systems (BSS) represent the "connection point" between external relations (customers, suppliers and partners) and an enterprise's products and services. Moreover, products and services are correlated with corresponding resources, like networking infrastructure, applications, contents and factories [1].

Basically, a BSS has to handle the taking of orders, payment issues, revenues and managing customers, etc. According to eTOM Framework it supports four processes: product management, order management, revenue management and customer management [2].

- Product management supports product development, sales and management of products, offers and bundles addressed to businesses and regular customers. Product management regularly includes offering product discounts, appropriate pricing and managing how products relate to one another.
- Customer management. Service providers require a single view of the customer and need to support complex hierarchies across customer-facing applications also known as customer relationship management. Customer management also covers the partner management and 24x7 web-based customer self-service.
- Revenue management is focused on billing, charging and settlement.
- Order management involves taking and handling the customer order. It encompasses four areas: order decomposition, order orchestration, order fallout and order status management.

Deploying it on a cloud solution raises a number of concerns that need to be addressed even before proceeding to implement such a massive project.

BSS on premises deployments have a mix of vendors, each one of them having different data structures, different databases and even different means of accessing these databases.

Migration from one vendor of a component to another is already a full time project, but migrating all the data to a new platform that has a much less flexible data structure than the source makes the job even harder.

To capture the best and the worst experiences the telecom operators had during their internal migrations, either from one vendor to another or even upgrading the same vendor's product.

Running through the interviews, other common concerns were identified that have to be treated with high priority:

1. Legacy systems' poor data quality
2. Very large amounts of data in legacy systems backup of historical data that has to be migrated or at least kept in a human readable format
3. Network bandwidth consumed to move this data
4. No means of validating migrated data quality against the legacy systems
5. No industry standard data model is present

In order to capture also their previous experience with this kind of data model, we have attempted to identify their previous failures in migrating data from one system to another and quantify them into risks, costs of mitigation or fixing and probability of happening:

1. IT resources are not business process experts. Failing to involve all the business owners will result in data migrated without their business purpose and structure. The cost of fixing this event is very high, considering that business interruptions might occur. The probability of happening is medium.

Ioan DRĂGAN

Bucharest University of Economic Studies

Răzvan Daniel ZOTA

Bucharest University of Economic Studies

Proc. of the Third Intl. Conf. on Advances in Information Processing and Communication Technology - IPCT 2015

ISBN: 978-1-63248-077-4 doi: 10.15224/ 978-1-63248-077-4-96

2. Bad data is migrated. The cost of mitigation is usually low, consisting only in manual corrections, but the risk of happening is very high.
3. Data is migrated in one big batch before the rollout and it is not consistent with the business needs. This bad practice is usually avoided in business critical systems, but the cost of fixing this is not very high if properly mitigated from the beginning.
4. Budget constraints that are usually overrun due to inadequate assessment and scoping in the initiate phase of the project. The probability of happening is very high, but usually budgeted from the beginning.

Transferring data between computer systems or storage systems is never an easy task. Most computer systems have a mix between structured and unstructured data formats that have to be translated into the new business models that the target software provides.
Next we will look at the existing alternatives, traditional and modern migration scenarios. Picking up best out of each we will build up a solution that allows migration at a certain level of complexity with none or minimum human intervention.

II. Extract, transform and load (ETL)

The extract, transform and load methodology is used in database operations, mainly in data warehousing and refers to three simple processes [3]:

- Extracting data from multiple sources and environments with a known structure
- Transforming data – applying a set of fixes transformations over a known structure of the extracted data
- Loading the transformed data into a new database with a fixed or an adaptable structure.

ETL systems can involve a considerable scale of complexity and relying only on a singular system for data integrity might involve significant operational risks.
The data complexity or even data quality in a production system can be easily overseen by developers and we might run into one of the common issues stated in the previous analysis: bad data is migrated.
To mitigate this risks, a data proofing system has to be implemented that validates the data against a set of qualitative rules. The benefits of data profiling are to improve data quality, shorten the implementation cycle of major migration and data warehousing projects, and improve understanding of data for the users [4].
ETL tools are suited for the task of migrating data from one database to another. Using the ETL tools is advisable particularly when moving the data between the data stores which do not have any direct connection or interface implemented.
Applications, even when developed by the same vendors, usually store data in significantly different models which make direct data transfer impossible. The ETL process is a must as the Transformation step is not always straight forward and of course, application migration usually does include storage and database migration as well. ETL tools, in this instance, have the advantage of its ready-to-use connectivity to disparate data sources/targets.

There are a various number of software products available on the market that provide such enablement for ETL, but as mentioned this require extensive development work and integration with each and every system and all data has to be analyzed and mapped accordingly. At most, source data can be ran against a predefined design for an automatic data proofing, but the rest requires manual intervention of data analysts and software developers.

Having a unified method of data migration requires an automatic data discovery and mapping towards the new data model.

III. Big data migrations

Big Data generally refers to the large amounts, at least terabytes, of poly structured data that flows continuously in heterogeneous systems and possibly in multiple organizations, including structured data, video, text, sensor logs, call records and others. [5]
A traditional ETL system extracts data from multiple sources, then performs a set of data proofing and transformations and loads it into a new database for different purposes. When the source data sets are large, fast, and unstructured, traditional ETL can become the bottleneck, because it is too complex to develop, too expensive to operate, and takes too long to execute.

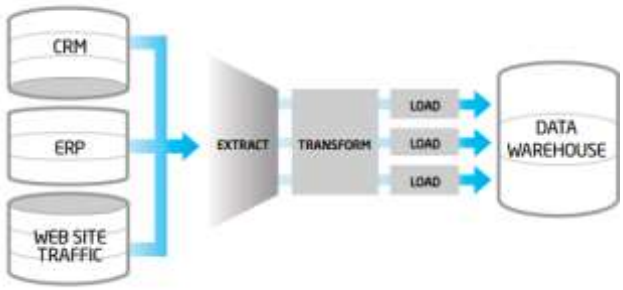

Figure 1. Traditional ETL plaform

Many Hadoop advocates argue that ETL platforms will come to an end since this data processing platform offers an ideal environment to handle data transformation at the needed scalability and cost advantages. [6]
ETL has grown exponentially in the past years, especially with the data input growing and also gathering more and more historical data. By definition, ETL requests data to be moved, which is translated into delay and capacity issues. Transforming data whilst we're transferring it adds more and more delay and processing consumption. This is seen by most of the IT admins as a non value added system that has to be removed or transformed into a single system.

Proc. of the Third Intl. Conf. on Advances in Information Processing and Communication Technology - IPCT 2015

ISBN: 978-1-63248-077-4 doi: 10.15224/ 978-1-63248-077-4-96

This led to changing the environment from a split infrastructure, where the systems that produced data were not the same as the ones consuming data, into a singular system that does all of that and has also the means of reporting it.

Using Hadoop as a data hub in an enterprise data management architecture, we now have a extreme-performance environment to store, transform and consume data, without traditional ETL at a more convenient cost.

Hadoop might be a good option to store and to transform data into the format we need, but it still doesn't solve our primarily issue: how do we extract data from heterogeneous sources and move it into a structured data warehouse for cloud application's usage?

IV. Using machine learning on Hadoop data stores to automate data transformation

Machine learning is a new concept that comes into the picture strictly tied to big data evolution. This new concept provides means of pattern recognition and predictions based on big data analysis.

Machine learning usually addresses tasks that can be grouped into 3 categories [7]:

- Supervised learning: the software is presented a set of example data and the desired output and it will define the transformation method
- Unsupervised learning: no information is given to the learning algorithm, leaving it on it's own to find a pattern or a structure
- Reinforcement learning: the software interacts with a dynamic environment in which it has to perform some tasks. The teacher tells the software if it's the expected result or not.

A key characteristic of Hadoop is called "no schema on-write," which means you do not need to have a pre-defined data schema before loading data into Hadoop. This is valid not only for structured data (such as call detail records, product transactions, customer data), but also for unstructured data like social media data (forums, comments, emails and any other communication means). Regardless of whether the incoming data is structured or not, it can be rapidly loaded into Hadoop without any transformation, where it can be analyzed, transformed and structured. [8]

Looking at our main goal: getting a big amount of structured data, but with different schemas and transforming it into a new, standardized schema, we can combine Hadoop's big data storage and processing capabilities and apply machine learning to have the data transformation mapping into our cloud BSS defined schema.

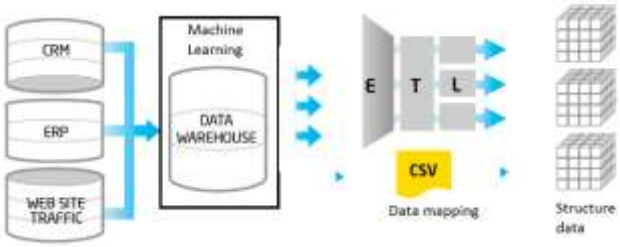

Figure 2. Combined Hadoop ETL and Machine Learning

All of this, can be done in a few simple phases:

- The first step is to collect all the data from various sources and store it in an unstructured format in a Hadoop cluster.
- Next, we set-up a set of test data, import it into the source system and the destination system.
- We run a set of machine learning algorithms to define the needed data transformations. After the mapping rules are created they are added into a list that can be reviewed and managed by an interactive interface. The list is available for us to delete and modify the rules. After the rules are updated we can proceed to the rules checking and execution.
- We run this transformation on the data stored in the Hadoop cluster
- Rerun from second step until we have all data migrated into a structured format

Similar automatic migration tools have been developed for SQL databases that automatically detect fields mapping, but traditional SQL can't be scaled for big data migrations. [9]

The architecture of such systems can be replicated though for Hadoop, which is in many ways similar to traditional databases if the data is structured. Considering the fact that BSS still stores most of its data in SQL databases, we can continue assuming that the data has some level of structures.

Below you can find a model of the proposed Azure Machine Learning implementation.

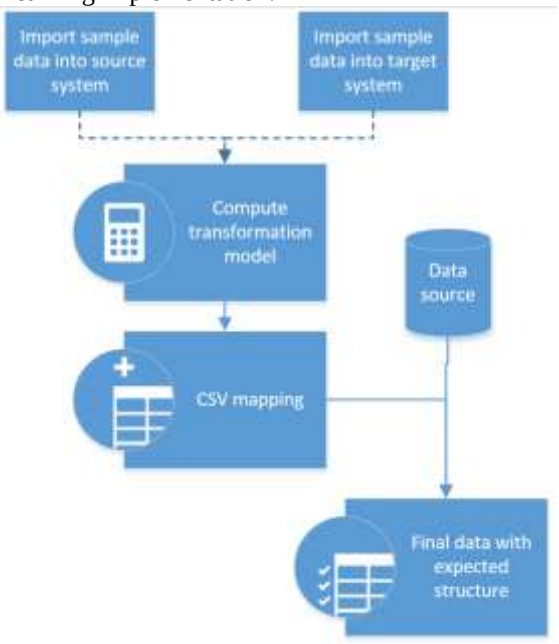

Figure 3. Azure Machine Learning Model

Proc. of the Third Intl. Conf. on Advances in Information Processing and Communication Technology - IPCT 2015

ISBN: 978-1-63248-077-4 doi: 10.15224/ 978-1-63248-077-4-96

In theory, applying machine learning on a set of data seems a simple job, but tests performed with Azure Machine Learning proved that machine learning are not 100% accurate, which should be the case when migrating data.
The tests performed by Azure Machine Learning development team, based on ISO 5725 standard, states almost a 90% accuracy. [10]
Corrections obviously have to be done by human intervention, but even automating 90% of the work is a good step forward.

v. Conclusions

Data migrations are never an easy job, but when scaling to a large set of poly-structured data, sometimes with poor data hygiene, we have to use automatic learning tools minimize the problem to a human solvable one.
Another benefit of self-service data preparation is that IT resources are freed up to focus on developing new application models that could help the business evolve in a predictable and profitable way. The most difficult part of the migration process is pulling a lot of data from a lot of different sources and transforming it into a structured model. Machine learning tools are evolving as we speak and starting from a 90% accuracy is already a good result. Using methodologies that already provide good results and are under development by other entities is a method of uncoordinated group collaboration towards service evolution.
Moving towards cloud applications with new data integration requirements and the growing need to navigate large data stores filled with a wide variety of structures and models are promoting even further the interest in self-service data preparation.
Eventually, cloud deployments present numerous problems since this niche software did not present a financial interest for cloud solutions provides. Solving these problems is actually a matter of processes and convincing telecom operators to invest time and effort and work together with their software or media partners and cloud service providers.
In order to comply with the strict requirements of telecom operators, we have firstly to under-stand their needs, even challenge that are not backed up by business requirements and fill in the gaps in cloud offerings.
Technical requirements for cloud infrastructure are being covered by service evolution as we speak, so our main concerns should be business processes and business requirements that have to be covered with a BSS solution "one size fits all".

References

[1] L. Angelin, U. Ollson, P. Tengroth. Business Support Systems. Internet: http://www.ericsson.com/res/thecompany/docs/publications/ericsson_review/2010/business_support_systems.pdf [Feb, 2010].
[2] eTOM – The Business Process Framework, pages 41-49, GB921B [Mar, 2014]
[3] Wikipedia.com, "Extract, transform, load"
[4] Jack E. Olson (2003), "Data Quality: The Accuracy dimension", Morgan Kaufmann Publishers], (pp. 140–142)
[5] Intel whitepaper, "Extract, Transform, and Load Big Data with Apache Hadoop"
[6] Phil Shelley, CTO Sears Holdings, CEO Metascale, "ETL's Days Are Numbered", InformationWeek 2012
[7] C. M. Bishop (2006), "Pattern Recognition and Machine Learning", Springer ISBN 0-387-31073-8.
[8] Phil Simon (March 18, 2013), "Too Big to Ignore: The Business Case for Big Data", Wiley ISBN 978-1-118-63817-0.
[9] Andreea Marin, Ciprian Dobre, Decebal Popescu, Valentin Cristea, "e-System for Automatic Data Migration", Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2010 12th International Symposium
[10]Azure Machile learning Development team Blog https://azure.microsoft.com/en-us/documentation/articles/machine-learning-azure-ml-customer-churn-scenario/

Acknowledgment

This work was cofinanced from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/187/1.5/S/155656 „Help for doctoral researchers in economic sciences in Romania"

Ioan Dragan has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2011 and IT&C Security Master in 2013. Currently he is a PhD. candidate at Economical Informatics ASE Bucharest.

Răzvan Daniel ZOTA has graduated the Faculty of Mathematics Informatics at the University of Bucharest in 1992. He holds a Ph.D. in Economic Informatics from 2000 and now is professor at the Department of Economic Informatics and Cybernetics from the Bucharest University of Economic Studies. From 2010 he is Ph.D. supervisor in the field of Economic Informatics. His last published books in 2004 are "IT Basics" and "Computer Networks" in ASE Publishing House, Bucharest, Romania. His recent work focuses on business cloud computing, computer networks and applications.