Improving Financial Information Processing at Banco de México: A case for a Big Data Architecture[1]

Mario Alejandro Gaytán González*
Banco de Mexico, Mexico City, Mexico – agaytan@banxico.org.mx

Jesús Gerardo Cruz Hernández
Banco de Mexico, Mexico City, Mexico – jesus.cruz@banxico.org.mx

ABSTRACT
Access to granular microdata by financial authorities has open a vast world of possibilities for analysis and supervision and a long run reduction of cost for financial intermediaries. Nonetheless, while users require greater speed and flexibility for consulting this information, its large volume and variety of formats impose a burden on traditional infrastructures based on relational databases, and therefore on the processes of collection, revision, and generation of information. Banco de México has a financial system information model that relies heavily on granular microdata and processing this information has always posed several challenges. However, new technologies have improved the prospects from the whole range of activities from data processing to data analysis for granular microdata. This paper presents, the use of a small prototype of a Big Data architecture to perform some proofs of concept applied to accessing and processing of financial information, in particular it is shown: i) the improvement in the speed of access to the information related to futures and forwards for two in-house applications: a User Querying Tool and an Interactive Portal for Financial Information (PIIF, due to its Spanish acronym) and ii) the improvement in the collection and validation process and the cross comparison with other sources of information when frequent reprocessing is needed.
Keywords: data processing; cooperation; financial information; big data architecture

1. INTRODUCTION
Since the Second half on the 1990s, Banco de México has developed a model of financial system information based on granular microdata and for financial markets transactions daily data. Granular microdata has provided greater flexibility to respond to new information requirements without increasing the regulatory costs of reporting of financial institutions. Nonetheless, the technological architecture has been based on relational databases to process, store and disseminate the information among different users (financial authorities and the general public). Although this architecture has been from a long time sufficient to serve different users' needs, the increase in the number of requirements, reporting institutions and volume of financial operations, their expected trend and an increasing demand for access to the highest level of granularity by an increasing number of users have proof that the current architecture is unsustainable in the middle run. When severe stress episodes have compromised processing times of the architecture both for the reporting institutions and for the generation of the information resources and statistics the usual response has been infrastructure growth. Nonetheless, new technological architectures that deal with Big Data have brought new insights to solve problems regarding volume of information that is stored, the variety of this information and the velocity to process the information. In the context of central banking granular information Big Data offers several opportunities regarding mainly analytics (predictive modeling, machine learning, and data mining), nonetheless in this paper we present uses for improving the processing and delivery of information.
The remainder of this paper is structured as follows: the next section gives an overview of the current situation of the handling of granular data at Banco de México, its activities and its overall technological architecture, stressing the difficulties that are faced to accomplish the processing and delivery of information. Section 3 presents the architecture of a small Big Data prototype that was used to carry out the proof of concept of the business cases presented in Section 4. In turn, section 4 describes 2 business cases about how Big Data could support: i) the improvement of processing and querying of information using in-house tools to visualize the required information and ii) the improvement in the collection and validation processes of financial information. Section 5 presents

[1] The views and conclusions presented in this paper are exclusively the responsibility of the authors and do not necessarily reflect those of Banco de Mexico.

results of the evaluation of one proof of concept giving more technical description of the proposed prototype that was used. Finally, section 6 discusses the roadmap ahead and concludes.

2. CURRENT SITUATION OVERVIEW

The financial system information model in Banco de Mexico has a very strong component of daily granular microdata of financial operations performed by banks and brokerage houses. Currently, Banco de México collects and process daily information on money market operations (repos, reverse repos, securities lending and spot transactions with securities), security by security database, and securities holdings, FX market operations, derivatives (swaps, futures and forwards, and options), interbank loans and deposits and time deposits. This model of granular information implies large and growing volumes of data. In recent years, this volume has grown as a result of new or improved reports, increasing number of financial operations and increasing number of reporting institutions. The process of this information is depicted in Figure 1.

Figure 1: Model for the process of information at Banco de México.

The availability of highly detailed information allows a strong validation process for information quality. This validation process includes: Format rules, business rules of the information, crosschecks for consistency of information on specific operations across the different counterparties, crosschecks of consistency of information across different regulatory reports for the same institution, daily information is crosschecked with monthly reports and regulatory regimes. This validation process and the high opportunity of the requests generates that financial institutions tend to retransmit their information several times until the desired quality is achieved, making this a very exhaustive process.

After the first successful process of complete validation, several query and BI tools are used by final users to consume useful information in the form of reports and graphical dashboards.

A traditional architecture based on relational databases and data warehouse sometimes becomes under critical stress to convey good response times because the time required to process large volumes of information and in turn the time to deliver the relevant information to final users.

2.1 TECHNOLOGICAL ARCHITECTURE

In order to perform the core model for collecting, processing and consuming information, Banco de México has an architecture that is depicted in Figure 2.

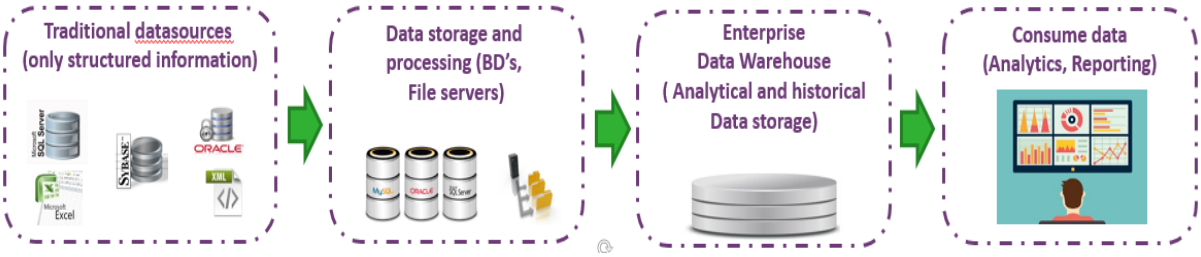

Figure 2: General schema of the current architecture

This architecture is based in traditional relational databases with the following specifications: Regular relational database, CPU with 4 cores with a processor of 3724 MHz and RAM of 36 GB.

In this architecture, the introduction of new data sources with different formats mostly imply totally new data storage to store the new data with its own format. Secondly, it is also necessary that IT and business users have a clear understanding about the information, how should be processed and the way to be consumed since the very beginning. It is frequent that a considerable number of requirements come to light after the information has been received, which is costly. With the current architecture user applications must redirect to either operational data and historical data for analysis purposes.

3. PROPOSED ARCHITECTURE DESCRIPTION

Given the constraints described above for processing financial information, Banco de México is evaluating a new technological architecture that allows on one hand the storage of big volume of data regardless data format in such a way is properly accessible, and on the other hand that the proposed architecture ensures that the management and processing of data is performed in the least time possible.

A solution based on Big Data describes an environment with a big amount of data (Volume) where data could be structured or unstructured and in several formats (Variety) and must be processed at the least time possible (Velocity). The importance of Big Data in any organization do not relies on the amount of data residing in the organization but in how data is managed. It is expected that a solution based on Big Data would allow to perform in a better way complex tasks such as:

- Cross checks using big volume of information.
- Analyze and validate data, both structured and unstructured, on demand and in real time.
- Use complex and iterative statistical processing over large data sets.

The general schema of an architecture based on Big Data is depicted in Figure 3.

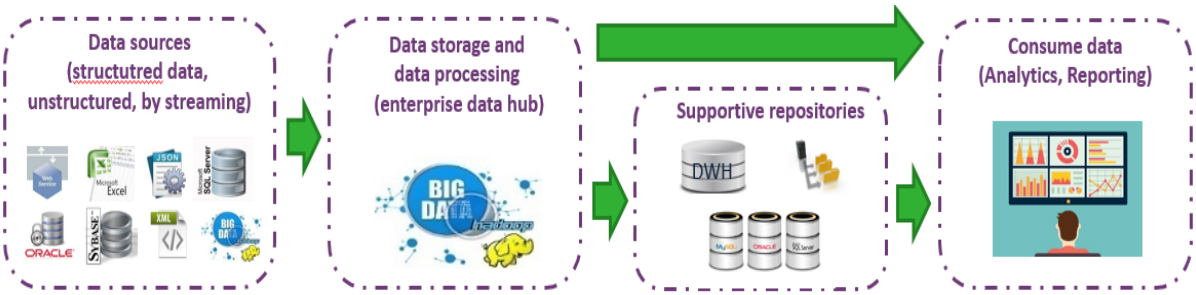

Figure 3: General schema of an architecture based on Big Data

This kind of architecture allows storage regardless data format, hence the transformation to any other format will now depend on what is the best storage format to consume and process data. The processed data, according to different business needs can be get, refined, standardized and processed in real time, additionally is not mandatory to have a very detailed model to store data from the beginning since the model could have an initial structure to receive data and can be adapted to the changes on the business needs.

Another important aspect of this architecture is that both daily operational data and big amount of historical data for analytical purposes can be stored in the same platform without affecting the performance to access and query data. In the same sense, this architecture offers a number of diverse technologies to query and process data depending on the business needs, for instance Cloudera Impala **[1]**, Spark and Spark SQL **[2]**, Pig **[3]**, Hive **[4]** among others. It deserves to point out that under this architecture it is not necessary to move data to other applications but the processing is kept on the data repository.

It was evaluated whether financial information at Banco de México satisfied at least 2 out of 3 Vs[2] for Big Data:

- Volume: The amount of data coming from financial institutions as well as the information that is generated internally it is expected to growth considerably. Most data requested by Bando de México is granular micro-data.
- Variety: The information managed by DISF comes in different formats such as CSV, XLS, XML TXT, even there are some limitations in processing non structured formats like PDF.
- Velocity: The performance to process and consume information is being compromised due to limitations on the current infrastructure, hence response time is increasing importantly.

To be capable to verify these benefits, a prototype based on Big Data technologies were implemented and 2 business cases were proposed which are described in the next section.

[2] According to several Big Data practitioners, a candidate for using a Big Data architecture must satisfy at least 2 out of 3 Vs (Volume, Variety and Velocity). Even Mark van Rijmenam **[5]**, considers important other 4 Vs for completeness.

4. PROOFS OF CONCEPT: TWO BUSINESS CASES FOR THE USE OF BIG DATA.

Considering the proposed architecture in the previews section, were chosen two business cases where it is though that the use of Big Data would be beneficial given the current inefficiencies of the present architecture. The business cases are explained in the next part of this paper.

4.1 BUSINESS CASE 1: PROCESSING AND QUERYING DATA.

Information must have value to the end users. An important aspect of this value is the velocity on which information is delivered to be analyzed and used for decision making. In this sense, this case of use is intended for the improvement in the response time required to query, transform and consume information related to Futures and Forwards. This process generates aggregated figures on daily reported transactions of futures and forwards traded by credit institutions and brokerage houses and reported in a daily basis. Specifically, this process aggregates amounts operated in futures and forwards transactions by institution and transaction date, assigning predetermined characteristics (e.g. underlying asset, currency, counterparty residence, counterparty type, etc). Therefore, this process "enriches" data of these derivatives operations using other sources of granular information. The generated information is needed to get it ready in a daily basis in the least time possible for decision making. The business case 1 is graphically described in Figure 4.

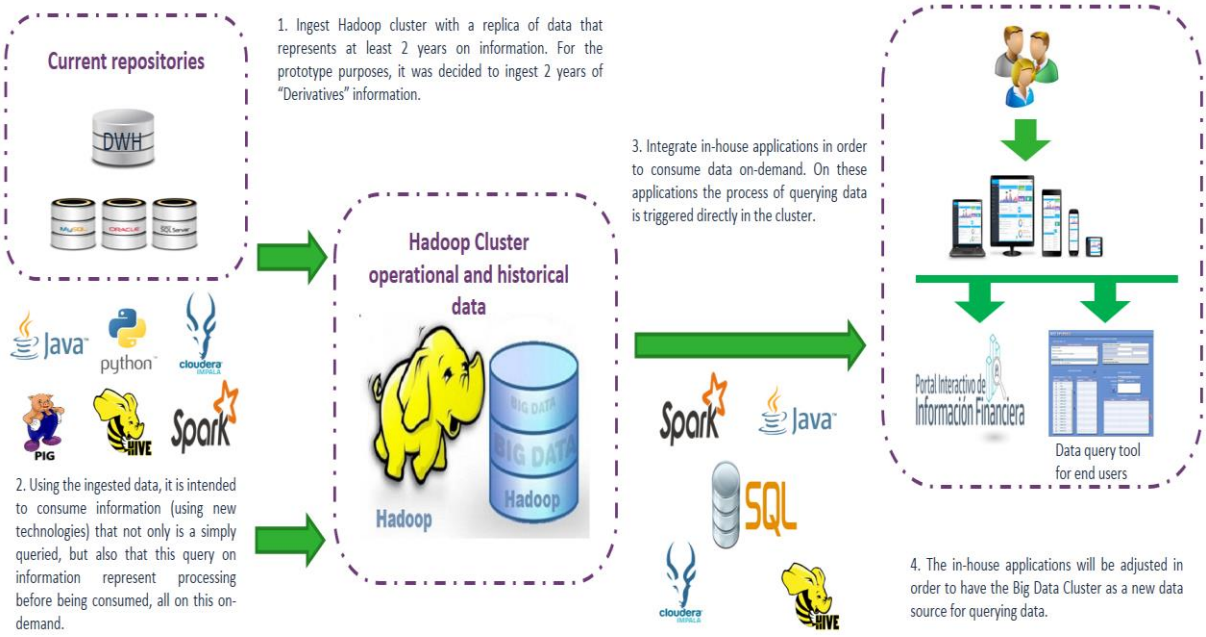

Figure 4: Graphical description of business case of querying and processing data

4.2 BUSINESS CASE 2: COLLECTION AND VALIDATION OF INFORMATION.

The core process of processing financial information includes collecting, validating and transforming information for a later consumption. It is very important that collected information is validated as fast as possible regardless the volume of data. Currently, this process had presented some inefficiencies for processing big amount of data, which in turn represents delays in accepting a final version of data without validation errors. This situation represents time consumed by the user of the financial entity. In addition, sometimes the reporting institution must send data several times because of either problems to pass data validation or changes in the information. This constant re-sending of data can imply large use of resources under the current architecture.

In this sense, this business case is intended for the improvement of the first two phases of the core model depicted in Figure 1. The business case 1 is graphically described in Figure 5.

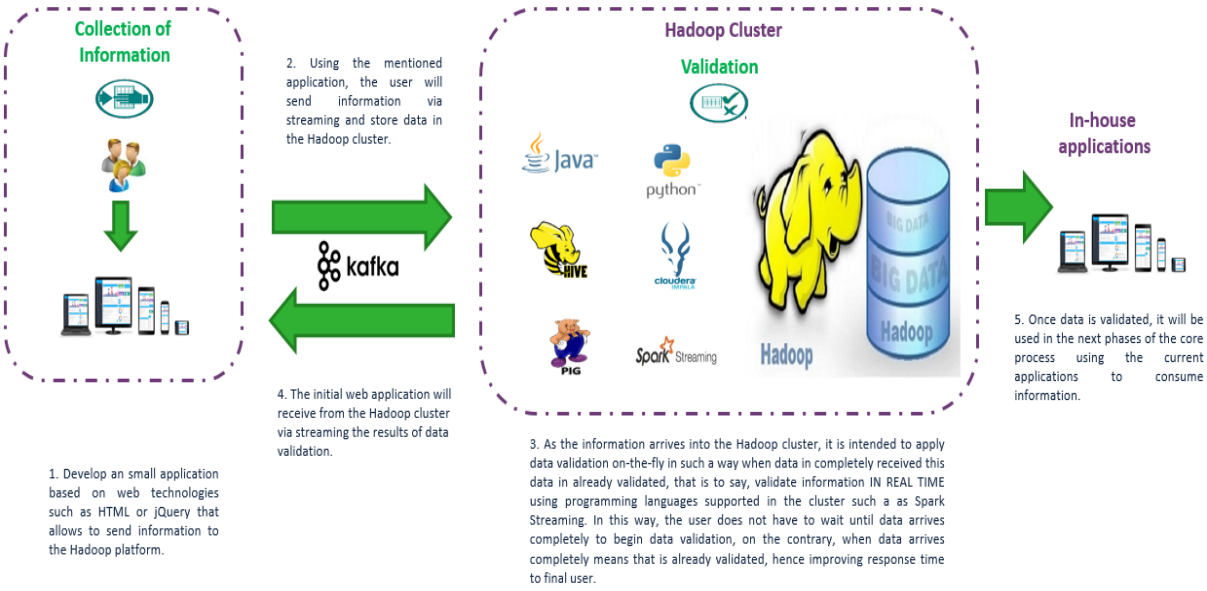

Figure 5: Graphical description of business case of collecting and validating information

For the purposes of this paper, just the business case 1 was implemented. The results are explained in the next section.

5. EVALUATION AND RESULTS

The purpose of this section is to evaluate the performance and recourse management results of the Big Data Cluster used as a prototype to give some indicators about its behavior in several conditions that will be explained accordingly.

5.1 EXPERIMENTAL SETUP

All the experiments were run on the same cluster prototype consist in 5 CPU's, where one on them is the MASTER and the remaining are SLAVES. The characteristics of each node in the proposed Big Data cluster are the following: Hadoop vendor: Cloudera CDH version 5.8, Processor INTEL i5-6500™ of 3.2 GHZ, Master's RAM of 16 GB (10 GB effective to the cluster), Master's hard disk unit of 256 GB (93 GB effective to the cluster), Slave's RAM of 16 GB (10 GB effective to the cluster), Slave's hard disk unit of 256 GB (150 GB effective to the cluster), Microsoft Windows 10 Professional, 4 cores per node, Virtual Box 5.1.14 as a virtual machine, Linux vendor CentOS 6.8, Linux kernel 2.6.32, Hadoop 2.6.0, Cloudera Impala 2.7 and Apache Spark 1.6.

Given this specification, the cluster in total has 40 GB of RAM and 600 GB of storage capacity (possible a bit more if Master node participates in the process). The storage configuration in each node was assessed on the best file format that it performs best on, that in this case were Parquet **[6]** and using Snappy **[7]** as a compression method.

In addition to the mentioned technology within the Big Data Cluster, it was needed to have a tool that has transparent interaction with the Hadoop Cluster regarding ETL functionalities that adapts easily to the available infrastructure in Banco de México as an orchestration element as well as for ingesting and querying data. In this sense, Pentaho Data Integration **[8]** was chosen. In this sense, Cloudera **[9]** was chosen due to it is one of the most trustworthy and popular distributions of Hadoop, besides is open source and exist the community version with represents no extra costs, which was very convenient for prototyping purposes.

5.2 SINGLE USER PERFORMANCE

There were tested 5 different cases in order to compare the cluster's performance in the situation of scaling horizontally, that is to say, adding or removing nodes to the cluster, hence there were tested using from 1 single node and adding another node until we used the whole 5 nodes within the cluster.

In Figure 6 compares the performance where a single user submitted 3 times the process or querying information to the cluster. It is noticed that the more nodes in the cluster, and in turn more memory available in the cluster, the better performance in the processing of data.

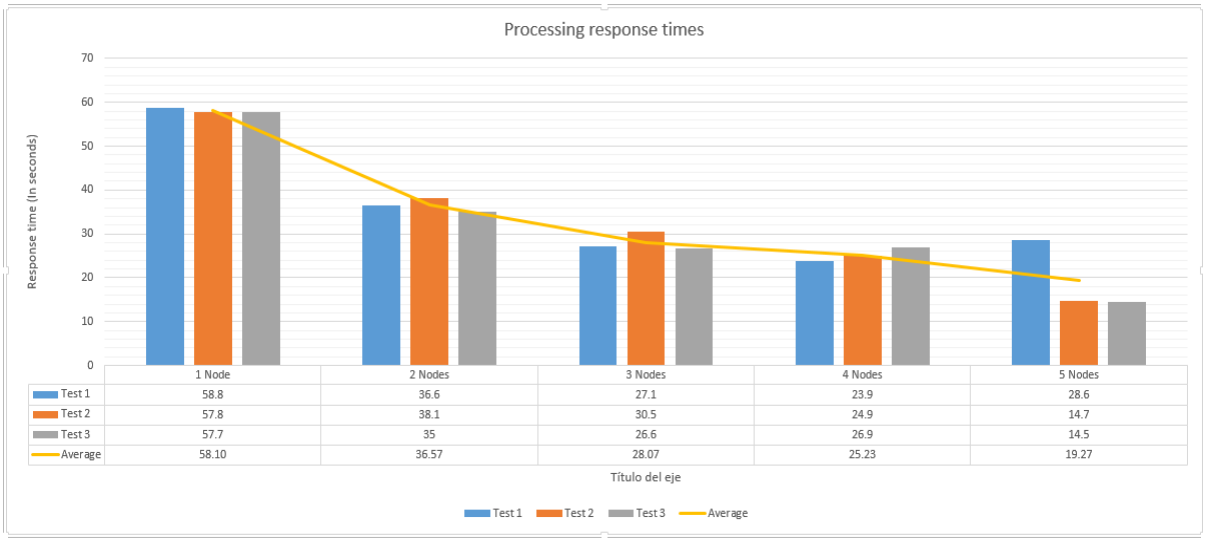

Figure 6: Comparison of processing response times with different number of active nodes within the cluster

It is very important to emphasize that these results are based in a prototype, that is to say, the resources available for testing are minimal.

5.4 MULTI-USER PERFORMANCE

It was measured performance on the same business case in the case of concurrency. In Figure 7 it is shown the performances in terms of response time in the scenario from 1 to 6 users concurrently submitting the process of the information required. These results are very interesting since even though were 6 users, the time that is consumed is not as considerable as the time consumed in the current architecture. Multi-users scenario is more closed to real world applications.

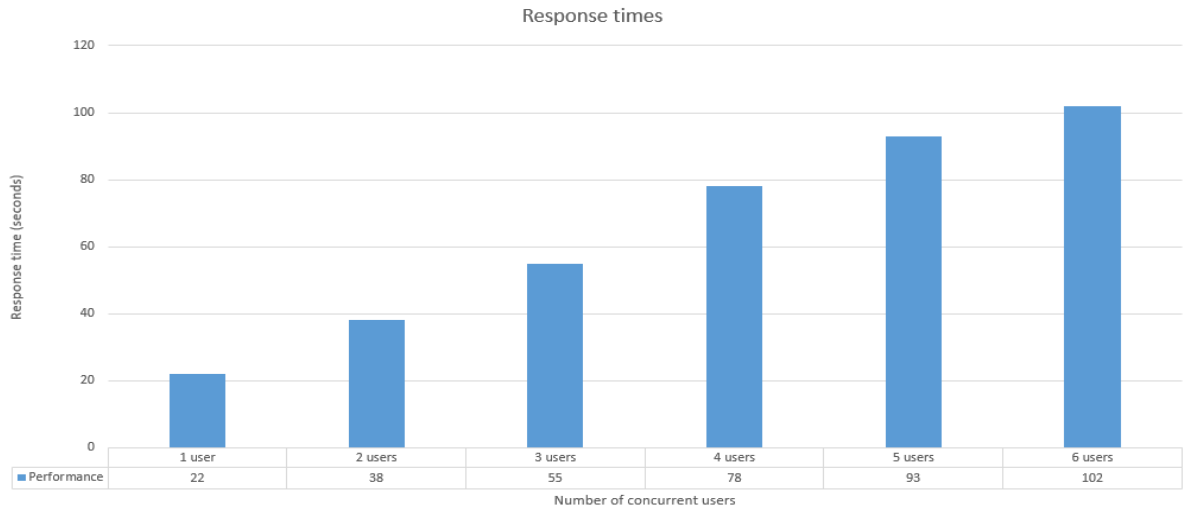

Figure 7: Comparison of multiuser response time

5.5 COMPARING AGAINS THE CURRENT TECHNOLOGY

As we mentioned in section 2.1, the total amount of Memory (RAM) that is available in the production database that is used in the Directorate of Financial System Information is about 36 GB and in the case of the prototype implemented for the purpose of this paper it is about 40 GB. Given the characteristics in terms of RAM memory used in both scenarios, it is possible to carry out a feasible comparison.

In the traditional architecture based on a relational database, the performance in terms of response time for the same business case is:

- In the best scenario (with only a single user on the database): 180 seconds.

- In a scenario with very hard use of the database (in peak hours and more than 2 users): > 1 hour

Considering that both the current production architecture and the prototype implemented in this paper, also the metrics of the current production environment and the performance results shown in sections 5.2 and 5.3 for the same business case, it can be said that the prototype is performing so much better having similar conditions. The performance improvement is notorious since was about 86% faster (180 seconds in current situation against 22 up to 25 seconds in the same single user case with this prototype).
Taking in count these results and also that in a production environment conformed by a cluster with a Master and 6 nodes where each node has a 2TB of storage space and 256 GB of RAM, it is feasible to predict that the response time to the same use case could be definitely very fast, in the order of just a few seconds even with concurrent users.

6. ROADMAP AND CONCLUDING REMARKS

In this paper, we gave an overview about the current architecture implemented for processing information at Banco de México. Even though this architecture has been suitable to accomplish with the core process, nowadays is insufficient due to both constant growth on data and the variety of this data in terms of format. Additionally, the demand for more information to be consumed for final users has represented an increase in the processing power which has had an impact on the response time. The use of Big Data as a core architecture has shown to be a good option to overcome current problems for processing data.
The results show that the use of Big Data bring a new world of possibilities to improve several parts of the core model for processing information. Even though the results of the prototype had an impact on the relevant use case regarding the improvement in the performance for querying and processing data, there is much work left to be done. Our roadmap, generally speaking, fall into the next categories: find out which parts of the process are still candidates to refactor and make improvements as much as possible in order to keep operating with the existing infrastructure, keep going ahead with additional research to realize what could be solved with a Big Data architecture in order to increase our knowledge about Big Data, and finally get started to develop a strategy to have a Big Data – centric organization.

6.1 PERFORMANCE ENHANCEMENTS IN CURRENT ARCHITECTURE

During the process to implement the use case about querying and processing information (use case 1), some opportunities for improvement came out. Even though the current infrastructure has some limitations, there is some possibilities to perform improvements in the current implementation that could help to minimize the response time and solve some problems about the use of relational databases. It is a fact that Big Data will bring new ways of solution to the present problems, but on the way is still possible to carry out several improvements to keep working with the less number of drawbacks. Additionally, to keep going with the implementation of new use cases will help not only to analyze use cases and find points of improvement, but also to acquire more knowledge about Big Data and as a help to be ready once Big Data is an official architecture to work with.

6.2 ADDITIONAL ENHANCEMENTS AND IMPLEMENT MAJOR USE CASES.

The advantage of Big Data and its technological ecosystem is that a problem could be solved using different technologies or a combination of technologies depending on the specific problem. The first use case was implemented using just cluster resources with Impala and using Pentaho to translate some calculation that represents no valuable time for the performance measurement. There is still some other approaches that worth to carry out like the one using Spark and Spark SQL in order to find out what are the benefits using this approach and appreciate if there is any improvement if performance or not. As it was explained in section 4 and 4.2, there is still another use case on the road which in turn will be implemented.
Additionally to these two use cases, as we mentioned before, there are several opportunities to take advantage of a Big Data architecture. In the medium term it is planned to use Big Data also for improving our current User Service Desk in which we would like to take advantage of another interesting application of Big Data related with Machine Learning **[10]**. In a daily basis, our personnel receive a lot of requests for providing help and support using a manual Service Desk. A lot of services are provided through this desk, several requests are very similar making a number of the responses to this support services very predictive and repetitive. It is intended to use machine learning techniques to identify patterns in the historic "Response or solution to the service requested" repository and make that requests to be served automatically according to patterns in the request, making more efficient the support response time to requesters. The more requests arrive, the more effective are machine learning techniques, and the faster is the response.

6.2 STRATEGY TO BECOME A BIGDATA–CENTRIC ORGANIZATION.

The effort used during the implementation of the Big Data prototype has been very valuable to realize, as organization, which are our strengths and being very consent about our weaknesses and what are the advantages we have with some parts of our current infrastructure and what are the drawbacks. In order to make valuable the effort done so far, it is very important to get started to plan a strategy to make all this knowledge available and beneficial to the rest of business users, and in turn, to Banco de México.

According to Mark van Rijmenam **[11]** it is very important to take in count 4 stages to have a successful Big Data strategy. In this sense, senior management and top stakeholders now have a clear understanding about Big Data and the meaning and advantages of having this kind of infrastructure. During the development of this paper we have described at least three possible uses cases and implemented one as a proof of concept which helped to show the benefits regarding to have an architecture based on Big Data. Proofs of concept are very important since these proofs allows to every stakeholder involved realize about the real picture of what is really implemented and its results.

Now with the results in place, it is very important to share this results with more people, firstly within the Directorate of Financial System Information, and then when appropriate to other areas in Banco de México in order to generate more interest in this venture and start not just more proofs of concepts but more larger projects.

6.3. CONCLUSIONS

In this paper we presented a prototype based on Big Data technologies that was intended to implement a number of proofs of concepts related to some business cases that currently show some performance problems. During the evaluation was realized that the performance results using a Big Data cluster were positive and even through was just one business case with a proof of concept, the positive achievements are just the beginning to keep going with more implementation to realize what more is possible to do with Big Data.

The major benefit of an architecture based on Big Data is that it provides more value to the information offering the possibility to process more amount of data to be delivered to more users in less time given that the response time in this architecture is faster than traditional architectures based on traditional relational data bases.

REFERENCES

[1] “Cloudera Impala”, Cloudera, 2017, https://github.com/cloudera/Impala

[2] “Apache Spark”, The Apache Software Foundation, 2017, http://spark.apache.org/

[3] “Apache Pig”, The Apache Software Foundation, 2016, https://pig.apache.org/

[4] “Apache Hive”, The Apache Software Foundation, 2015, https://hive.apache.org/

[5] “Why The 3V’s Are Not Sufficient To Describe Big Data”, Mark van Rijmenam, August 6, 2014, https://datafloq.com/read/3vs-sufficient-describe-big-data/166

[6] “Apache Parquet”, The Apache Software Foundation, 2014, https://parquet.apache.org/

[7] “Snappy, a fast compressor/decompressor”, GitHub Inc., 2017, https://github.com/google/snappy

[8] “Pentaho Data Integration”, Pentaho Corporation, 2017, http://www.pentaho.com/product/data-integration

[9] “Cloudera”, Cloudera, 2017, https://www.cloudera.com/

[10] “What is machine learning and which applications has nowadays?” (Available only in Spanish), Mayo 19, 2016, http://www.intelygenz.es/que-es-machine-learning-y-que-aplicaciones-tiene-dia-a-dia/

[11] “The Big Data Roadmap to a Winning Big Data Strategy”, Mark van Rijmenam, May 14, 2015, https://datafloq.com/read/the-big-data-roadmap/195