Cloud Data Mining Architecture

Nicolás Aguilera Contreras
Ingeniería de sistemas
Escuela Colombiana de Ingeniería Julio Garavito
Bogotá, Colombia
nicolas.aguilera-c@mail.escuelaing.edu.co

Nicolás Ortega Limas
Ingeniería de sistemas
Escuela Colombiana de Ingeniería Julio Garavito
Bogotá, Colombia
nicolas.ortega-l@mail.escuelaing.edu.co

Daniel Felipe Walteros Trujillo
Ingeniería de sistemas
Escuela Colombiana de Ingeniería Julio Garavito
Bogotá, Colombia
daniel.walteros@mail.escuelaing.edu.co

***Abstract*—We, as humans, have created more than 4.4 zettabytes of data in 2014, for the previous year we managed to generate around 44 zettabytes of data [1].**

The demand to analyze these large quantities efficiently has become a necessity to keep up with new trends in this world. This is the reason for the growth on the Data analysis ground in recent years. These amounts of data can be translated into information that becomes relevant to understand problems facing an organisation, and to explore data in meaningful ways. Data in itself only represents merely facts and figures. Data analysis organises, interprets, structures and presents the data into useful information that provides context for the data.

***Index Terms*—Data Mining, Distributed Systems Algorithms, Cloud Architectures**

I. Introduction

We live in a world where more and more data is collected and analyzed. Hyper connectivity has become one of our pillars as a society, allowing us to create an entire new industry. Companies, which are successful in analyzing and pushing more personalized content, generally run high on the new financial metrics and encourage the creation of more value.

Is there any way that humanity can use this amount of data? Well, there are alternatives surrounding data usage, one of them referred to as Data mining. Basically, it consists in the practice of automatically searching large stores of data to discover patterns and trends that can describe something. So, it tends to go beyond simple analysis. And being combined with mathematical algorithms and techniques we can turn raw data into useful information that allows us to discover patterns, likely outcomes and diminish the probability of failure in decision making.

AWS recently launched many services, such as AWS Lambda or Amazon API Gateway; these alternatives have made significant enhancements to existing analytics offerings, allowing easy scalability, low power consumption and structural security improvement, therefore with the utilization of AWS Cloud Computing we will show you how to integrate these tools efficiently.

In this paper we are going to present some use cases of recent works developed by other students, then we will proceed to explain what cloud data mining is and finally we are going to describe a general architecture design, an AWS Implementation and their benefits and disadvantages.

We based our proposal on a scheme for cloud data mining, deciding to build an architecture in AWS that does not use any common tool such as Hadoop and performs the data mining process to consult events registered in a digital platform.

Our architecture is handled through the Amazon API Gateway Service, this service connects three AWS Lambda functions. The Create Log function is the one that allows us to store all the information of the events, the Mining Node function reads the events converting them into useful information and the Main Node function is in charge of obtaining the results of the events.

We can state that our finality is to show that with the utilization of AWS Cloud Computing we can implement Data mining models using simple API Gateway alternatives allowing easy scalability, low power consumption and structural security improvement.

II. State of the art

Taking into account the current knowledge about data studied through the analysis of several tools available now. We thought that it might provide an overview of what has been done in this field so far and what should be further investigated.

One of the Projects that we found most interesting was the one developed by Xing Gao and Qianwen Li from the University of Pittsburgh referencing how can Data mining be used to establish the relationship between an audiences characteristics and the type of music that they like. [2]

So how did they do it? They started by dividing their work into steps, being the first data collection. They collected the data from a subset of the EMI One Million Interview Dataset. This subset of data allowed them to work on a first version of the public profiles.Including their age, gender, region, working status, and their attitude towards music in general.

Their second step was to clear and aggregate data. There they took their sample and started to filter and make it usable for the study. They stated the similarities of the artists-based information by counting the frequency of each of the 82 words most commonly used by the users.

The third step consisted of the clustering part. They had a total of 50 artists to analyze, so the job was to find any similarity among these artists based on the scores of the 82 words, which reflect user's opinion regarding every artist, they used SVD (Singular value decomposition [3]) creating a 82 x 50 artists-words matrix, and then proceed with the K-means method for clustering. [4]

Then they combine the user-artist-rating information together with the user's profile, so that they could visualize how the user's characteristics may affect their ratings on each of the artists. The visualization of the normalized matrix showed Artist 42 to 47 generally got higher ratings than Artist 10 to 30, which means that these artists are more liked by the audience. So now they could recommend these artists to the users belonging to those groups.

After establishing all those methods, they had enough data to build a recommendation system. They based it by choosing two groups to be the training matrix. Consequently after implementing their solution, they computed the error between the prediction and the real results. Getting the conclusion that the Root-mean-square deviation of the User based collaborative filtering [5] had better performance than Item based collaborative filtering [6].

This well performed project allows us to understand that it can be developed from a student environment to use unsupervised data mining methods to understand a raw dataset and present the relationship among user profiles and related information that can be found. In this case the artist likeness. Encouraging us to believe and work towards the development and contribution to the data mining community.

III. Data Mining

A. What is Data Mining

The term data mining does not really represent the mining of large amounts of data. In fact, the most appropriate name for this concept would be "knowledge mining from data" but it is too long to be used. However, in most industries, media and investigative works , the term data mining refers to knowledge discovery from data or KDD. This concept is a process that refers to an iterative sequence with the following steps :

1) Data cleaning : to remove noise and inconsistent data.
2) Data integration : where multiple data sources may be combined
3) Data selection : where data relevant to the analysis task are retrieved from the database
4) Data transformation : where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5) Data mining : an essential process where intelligent methods are applied to extract data patterns.
6) Pattern evaluation : to identify the truly interesting patterns representing knowledge.
7) Knowledge presentation : where visualization and knowledge representation techniques are used to present mined knowledge to users.

In conclusion, data mining refers not only to the process of extracting a pattern from large amounts of data but also to the knowledge of it. [7]

B. What types of data that can be mined

1) Database Data: Also known as a database, it consists of a collection of interrelated data that are managed and accessed by software programs. On the other hand, a relational database is a collection of tables in which each of them has a unique name. Each table contains a set of attributes and a long set of tuples. Each tuple represents an object identified by a key and described by a set of attribute values. [7]

2) Data warehouses: A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. [7]

3) Transactional Data: Each record in a transactional database represents one transaction. These usually include a unique id and a list of items requested in the transaction. These transactions can include additional information in their tables. [7]

C. Which technologies are used?

1) Statistics: Statistical models can be the target of data mining. Statistical models are a set of mathematical functions

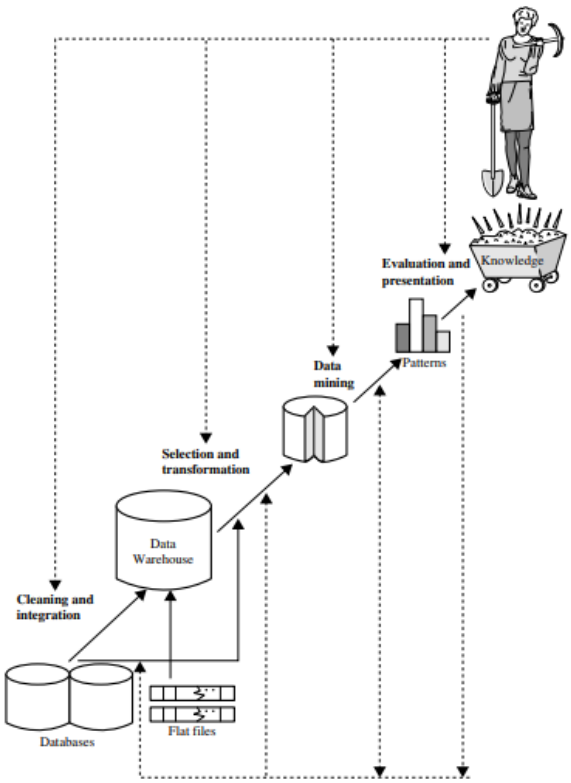

Fig. 1. Knowledge discovery from data process [7]

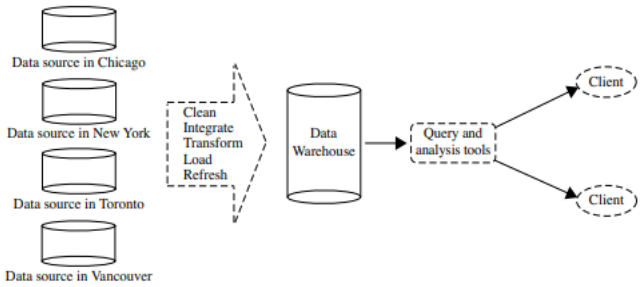

Fig. 2. Example of a framework of a data warehouse for an enterprise [7]

that describe the behavior of objects in a target class and their associated probability distributions. In this way, the data mining process can help identify noisy or missing values in the data.

2) Machine learning: Machine learning is responsible for investigating how computers can learn based on data. In this way, one of the great objectives in the last years of this discipline is to have the ability to recognize complex patterns and make intelligent decisions based on them. This is why this discipline is related to Data Mining and there are some classic problems related to it that have the goal of learning a function that maps an input to an output based on an example of input-output pairs. [7]

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Fig. 3. Fragment of a transactional database for an enterprise [7]

- Supervised learning
- semi-supervised learning
- unsupervised learning
- active learning.

3) Information retrieval (IR): It is the science of looking for information in documents. These documents can be text or they can be on the web. This science assumes that the information is unstructured and that the queries are for key-words. There is an increasing amount of information on the internet. Therefore, it is increasingly important to use data mining methods for the correct search, analysis and delivery of this information on the web. [7]

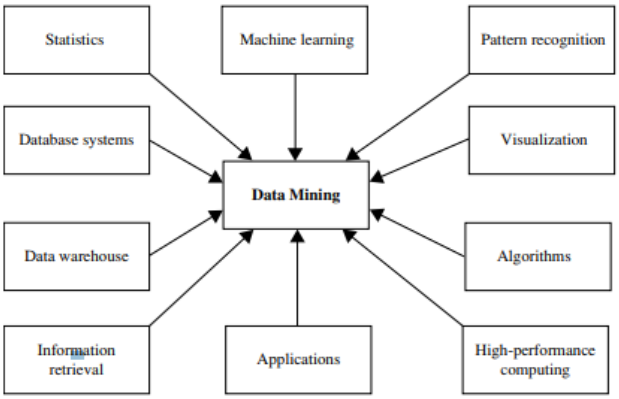

Fig. 4. Data mining adopts techniques from many domains [7]

IV. Cloud Computing

Cloud computing is the ability to offer services through the high capacity that the internet currently gives us, allowing access to software resources internationally, as it is a software application that serves diverse customers. [8]

The ability to be located in various locations and to scale without having to pay for support systems is what differentiates cloud computing from simple outsourcing. [9]

Cloud computing offers companies a pool of well-maintained, secure, easily accessible, on-demand computing resources, such as servers, data storage and application solutions.

This gives companies greater flexibility over their data and information, which can be accessed anywhere, anytime, and is essential for companies with locations around the world or in different work environments. With minimal management, all the software elements of cloud computing can be sized on demand and all that is required on the customer's side is an Internet connection.

Most cloud providers use a shared responsibility model in order to draw a duty boundary point with customers who contract their services. [10]

In this model, the provider is responsible for protecting the infrastructure that runs all the services provided in its cloud; this infrastructure is made up of the hardware, software, networks and facilities that run the services.

On the other hand, the customer's responsibility is limited by the cloud services model they use, as only then can they determine the scope of configuration work as part of their security responsibilities.

Cloud computing offers three service models:

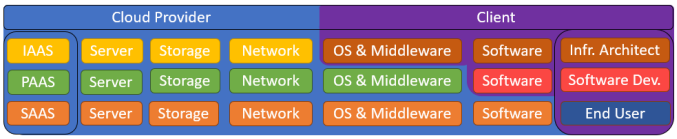

Fig. 5. Cloud Services Model

Where the customer's responsibility consists of the administration and security configuration of the elements not included in each model.

V. Cloud Data Mining Computing

The data mining system based on cloud computing is built with multiple services where each one has a transparent interface to be able to have different types of users, this interface is built based on the development of the system. [11]

The user does not need to know how the system works to use it and since it is the cloud there is no need to worry about the computing and storage capacity of the system, selecting the right algorithm to process the data and collect the results, everything should work properly.

Since it is in the cloud, the cost of this implementation is on demand, which would mean a reduction compared to the cost of supporting the number of servers required for this process, also the storage of this data is not a hardware constraint as it is also deployed in the cloud.

The most essential thing in Cloud Data Mining is the data mining algorithm, although there are several types of algorithms, the most commonly used ones can be classified into the following three categories:

- Classification Algorithm:

  They are suitable for relational data that is tuple structured, their purpose is to use the datasets to mine new information, analyze that information and with it find the classification principle; this principle can be used to classify the data after incorporating more.

- Cluster Analysis:

  Also used for relational data that is structured in tuple form, its purpose is to find a pattern to distribute meaningful data from potential data.

- Association Rules:

  Its objective is to find association rules that relate several sets in a large amount of data, they are suitable for data about transactions and are more effective if their value is Boolean or numeric.

The basic process of the data mining algorithm is described in the next Figure.

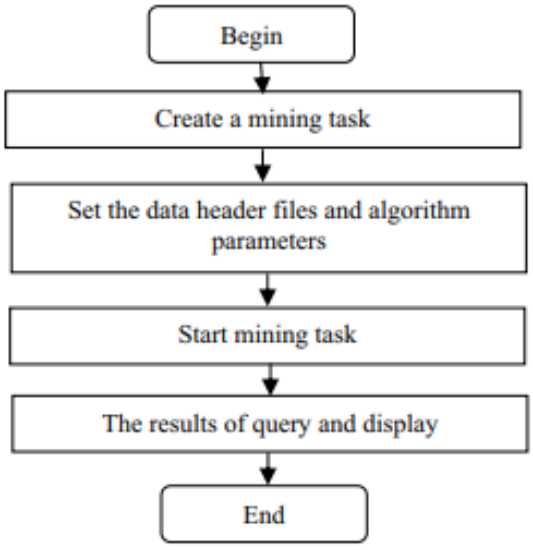

Fig. 6. Data Mining Algorithm Process [11]

VI. System Architecture of Cloud Data Mining Computing

The main feature of this system is the efficient processing of data in a distributed environment, for this reason a MapReduce strategy is used [12], this consists of two stages.

The Map stage, the information is segmented into multiple sets where each one is assigned a Map task, this task

generates the desired result that in our case would make use of the data mining algorithm.

Then, in the Reduce stage, all the results of the Map tasks are taken and by means of the Reduce function the results are combined into one.

The idea of this mechanism is that it is fault tolerant, this implies that if one or more nodes obtain an error when calculating its task the process will distribute its function to another node.

The architecture of this system consists of two types of nodes, the MainCtrlNode which is in charge of storing the information, handling the system metadata, supervising the work and storing the various functions used in the data mining algorithm; and the WorkNode which has the segment of information supplied and its own supervisor. [11]

To illustrate the idea we have the next Figure.

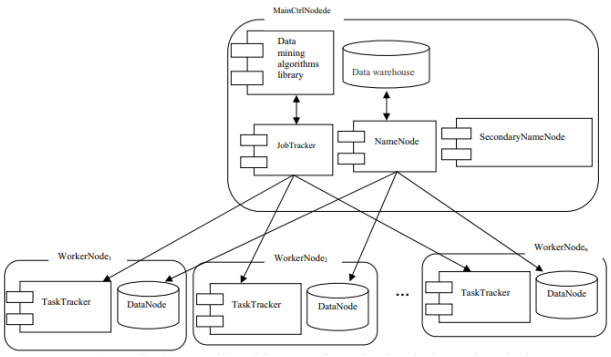

Fig. 7. Full Cloud Data Mining Architecture [11]

VII. Cloud Data Mining In AWS

To demonstrate the great advantages of Cloud Data Mining, a simple implementation of the architecture was performed on the largest cloud provider, Amazon Web Services (AWS).

A. AWS Services Used In The Architecture

- Amazon API Gateway

  a service that essentially allows developers to do a set of options to control and protect services related to APIs at any scale. Basically, it consists of an APIs that serves as a ”gateway” for applications to access backend services data, business logic, and functionality.

  It also allows their users to create RESTful APIs and WebSocket APIs with API Gateway to permit real-time two-way communication applications [13].

  AWS API Gateway also supports Web applications, as well as containerized and serverless workloads. It does this by invoking exposed API methods through the frontend HTTP and WebSocket endpoints, that could be:

  - API Gateway REST API, At the backend HTTP endpoint it integrates a collection of HTTP resources and methods, Lambda functions, or other AWS services.
  - API Gateway HTTP API, It consists of a collection of routes and methods that are integrated with backend HTTP endpoints or Lambda functions.
  - API Gateway WebSocket API, It consists of a collection of WebSocket routes and route keys that are integrated with backend HTTP endpoints, Lambda functions, or other AWS services.

- AWS Lambda

  AWS Lambda is a serverless computing platform provided by AWS that enables you to run code without provisioning or managing servers in real time. If you need to run an application all the time, you will be charged for the time the functions are running. The code that runs on AWS Lambda is called a Lambda function.

  With this tool you can run code for nearly any kind of backend utility or provider besides having to operate administration tasks. It works in a very simple way first you need to add your code as a .ZIP file or container picture and Lambda will automatically allocate the compute execution strength and execute code primarily based on incoming request or tournament for any scale of traffic.

  You do not have to scale your Lambda functions, AWS Lambda scales them automatically on your behalf. Every time an event notification is received for a function automatic scaling starts, launching Amazon EC2 instances on demand [14].

  The first time a function is invoked, AWS Lambda creates an instance of the function and runs its handler method to process the event. When the function returns a response, it stays active and waits to process additional events. If you invoke the function again while the first event is being processed, Lambda initializes another instance, and the function processes the two events concurrently. As more events come in, Lambda will continuously keep doing this assuring high scalability, for example we can see at an eBay handler scalability proposal by Michael Mendlawy in figure 8, They designed an architecture that allowed them to handle as many requests as we want using AWS Lambda. It is divided into three layers of Lambda invocations:

- Orchestration Lambda function.
- Page Group Lambda function.
- Page Lambda function.

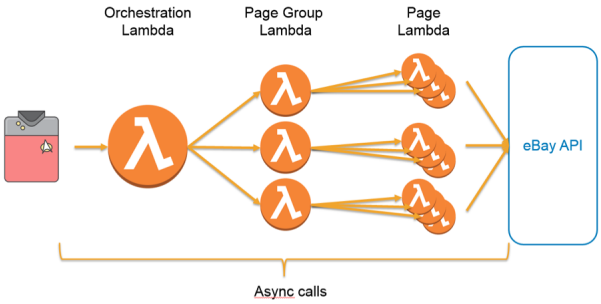

Fig. 8. AWS Lambda Scaling Layers [14]

This can be performed with the use of AWS Elastic Beanstalk and Auto Scaling, where Lambda rapidly locates vacant resources inside its compute fleet and will run the code. Since the code is stateless, Lambda can start as many copies of your feature as wanted besides prolonged deployment and configuration delays. There are no limits to scaling a function. Lambda will dynamically allocate the resources needed for incoming events.

- Amazon DynamoDB

  Amazon DynamoDB is a fast and flexible NoSQL database service for any scale. It provides a key-value database that is fully managed, durable, and multi-active. Also It has built-in backup, restore and security. It even has the capacity to handle more than 10 billion requests per day and 20 million requests per minute. [15]
  Advantages
  - Performance at scale
    Amazon DynamoDB replicates your data across multiple AWS regions. In this way, it provides response times of 1 millisecond at any scale. If a response time of microseconds is required, DynamoDB Accelerator provides fully managed in-memory cache. [15]

  - No server management
    DynamoDB automatically reduces or grows tables to adjust capacity and increase performance. This way, there is no software to install and no servers to manage. [15]

  - Ready for business use
    DynamoDB allows ACID (Atomicity, Consistency, Isolation and Durability) transactions. In addition, it allows the creation of backups at the Terabyte level without affecting performance. Additionally, this data can be exported to Amazon S3 for proper business analysis. [15]

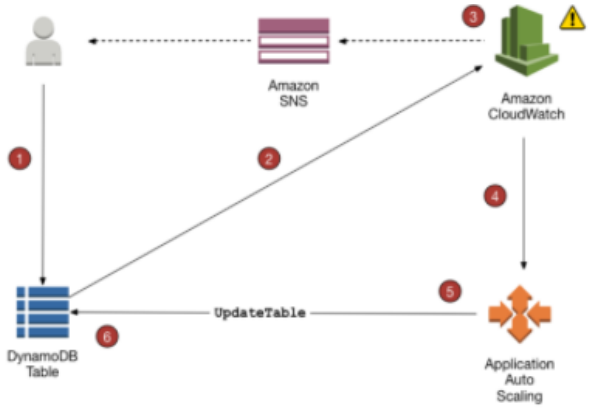

Fig. 9. DynamoDB auto-scaling [16]

B. How does this architecture work?

All the access to the architecture is handled through the Amazon API Gateway Service, this service connects only to three AWS Lambda functions.

- The Create Log function is the one that allows us to store all the information of the events that we want to log inside DynamoDB, the endpoint used to access this function must be linked to the platform on which the events are going to be monitored.

- The Mining Node function reads the events stored in DynamoDB and groups them by date and type of event, converting the information into a more understandable structure suitable for web analysts.

- The Main Node function obtains the results of the events so that the company's analysts can determine modifications or creations in their market strategies based on the activity of their digital platforms; as the amount of these events is very large, the Main Node function distributes the work using the Map Reduce algorithm using instances of the Mining Node function.

To illustrate this architecture we have the next Figure.

For large amounts of information, many instances are required to perform the Mining Node function process, by using AWS Lambda for this work the auto scaling is not only supported by AWS, also no configuration is required to perform it.

To incorporate this strategy in any company, only two integrations are required:

- Integrate the Web Platform

  Whether it is a mobile app or a web page, for each event performed by the users that you want to monitor, you must send the information through a JSON

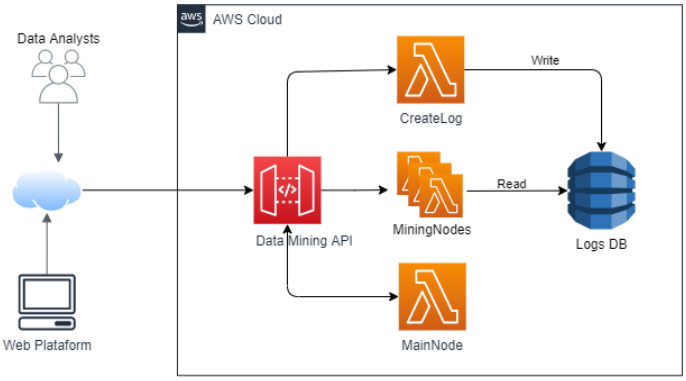

Fig. 10. AWS Cloud Data Mining Architecture

object to the API Gateway, this service will store all the events in DynamoDB through the Lambda CreateLog function.

- Integrating the Analysis

  The enterprise analysts obtain the information of all the events registered by means of a request to the API Gateway, this will return a JSON object indicating the number of events occurring per day.

C. Advantages

- Lambda Functions consume energy only while they are used, for this reason their cost only depends on the number of times they are invoked; this is a great saving compared to an On Premise architecture.
- Since all the communication for the architecture works through the Amazon API Gateway Service, all the internal structure guarantees security.
- The Data allowed in Amazon DynamoDB guarantees maintenance of the information with a high tolerance to failures, depending on the region where it is implemented ensures up to 99.999% availability.
- To incorporate new security systems such as restricting access to certain endpoints, the code of any lambda function must not be modified, only the changes must be made in the Amazon API Gateway Endpoint.
- By default, AWS Lambda functions store the information of their executions in AWS Cloud Watch, so there is a record of the access of each one, this guarantees traceability and allows to manage their use internally in the company.

D. Disadvantages

- With Amazon DynamoDB it is not possible to store built-in data structures as it is possible with MongoDB, for example. Also it's only possible to create 256 tables per Availability Zone Finally it is not possible to do joins between tables, so the events that can track cannot be very complex.
  Also, the table properties are updated by default every 6 hours, which means that without paying extra you cannot get real time data. [17]
- When using a serverless service like AWS Lambda developers will always rely on vendors for debugging and monitoring tools, that means that you cannot debug out the AWS Lambda Console. [18]
- Loss of control in Amazon Lambda can occur in different ways: in configuration, performance, troubleshooting and security, since total control of the administration of the software on which the code is executed is totally transferred to the provider that develops and operates the functions. [18]
- Amazon API Gateway will be a single point of entry and failure between two parts of an architecture. In this way, the API Gateway will handle a multitude of scenarios that can affect the speed and reliability of the system. [19]
- Amazon Lambda functions have a default timeout of 3 seconds. In this way, tasks that involve large amounts of data will tend to exceed the execution time limits. Thus, developers will be forced to rewrite the code in a different architecture. [20]

VIII. Results

To analyze the execution times and perform the debugging process of the architecture implemented in AWS, execution scripts were made using the Postman tool [21], thanks to this, the scripts can be exported in JSON format, which allows any member of the team to test the endpoints with the only requirement of having the tool and Internet Connection.

All services were deployed in the AWS us-east-2 region, locating the servers in Ohio, USA; the long distance between this point and Bogota (Colombia) from where the tests were performed gives us a more approximate result for any type of company than using the sa-east-1 region (Sao Paulo, Brazil) which would be the most efficient point for our team.

Using the Postman application runner [22], we tested the requests for the API Gateway Endpoints of approximately 1000 events registered in the database.

We compared the execution times with up to 30 concurrent requests using the Mining Node with the entire database

(representing the traditional way of obtaining the data) and the Main Node using Data Mining.

The following results were obtained:

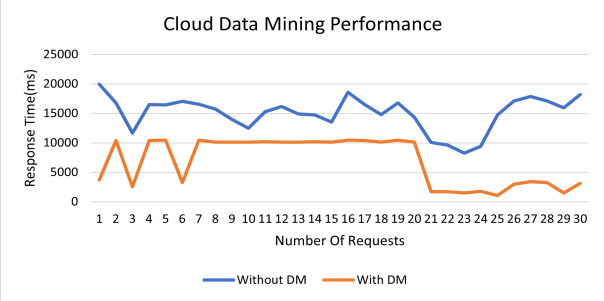

Fig. 11. Execution Results

IX. Conclusions

- Due to Amazon Dynamo's default configurations, it is only possible to ensure the analysis of registered events up to 6 hours before the request without incurring in higher costs; this implies an additional cost if you want to perform digital analysis in real time.
- The integration of this service through API Gateway endpoints allows the platform to be analyzed without having to alter much code.
- Since the creation of events is also deployed in an API Gateway endpoint, the incorporation of new events only requires altering the script within the platform, without the need to alter the architecture code.
- Because AWS Lambda functions are auto-scaled with respect to demand, it can be ensured that the number of services running is the minimum and therefore the cheapest.
- The implementation of Lambda functions with the use of languages such as javascript allows us to modify the functionalities in a short time, reducing development times when incorporating new functionalities.
- As the entire architecture is in the AWS ecosystem, it is necessary for performance to have all services initially deployed in the same region.
- The use of distributed systems in the cloud for data mining allows us to reduce co-
https://www.overleaf.com/project/60246697825128abe8235008sts
in services and hardware since providers only charge for processing.
- According to the results analyzed, with a record of at least 1000 events there was a performance of 46% with respect to the traditional approach of event processing; although this percentage is reduced to a greater amount of data, it is a very significant savings for the digital analysts of a company to better estimate their results and adapt their marketing plans.

References

[1] F. Toomey, "Data, the speed of light and you." https://techcrunch.com/2015/11/08/data-the-speed-of-light-and-you/, NOV 2015. Accessed on 2021-02-12.

[2] Yurulin, "Final report group9." http://www.yurulin.com/class/spring2014_datamining/final_samples/FinalReport_Group9.pdf. Accessed on 2021-02-12.

[3] MIT, "Singular value decomposition (svd) tutorial." https://web.mit.edu/be.400/www/SVD/Singular_Value_Decomposition.htm. Accessed on 2021-02-12.

[4] D. M. J. Garbade, "Understanding k-means clustering in machine learning." https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1, SEP 2018. Accessed on 2021-02-12.

[5] Wikipedia, "Root-mean-square deviation." https://en.wikipedia.org/wiki/Root-mean-square_deviation, FEB 2021. Accessed on 2021-02-12.

[6] E. Alegría, "Sistemas recomendadores con recommenderlab." https://rpubs.com/elias_alegria/intro_recommenderlab, JUN 2020. Accessed on 2021-02-12.

[7] M. K. Jiawei Kan and J. Pei, *Data mining concepts and techniques*, ch. 1. Morgan Kaufmann, 2012.

[8] S. P. Mirashe and D. N. Kalyankar, "Cloud computing," 2010.

[9] SalesForce, "Cloud computing." https://www.salesforce.com/mx/cloud-computing/. Accessed on 2021-02-12.

[10] Amazon, "Shared responsability model." https://aws.amazon.com/es/compliance/shared-responsibility-model/. Accessed on 2021-02-12.

[11] L. H.-w. Z. L.-j. Ren Ying, Lv Hong and W. Li-na, "Data mining based on cloud-computing technology," 2016.

[12] P. A. Tyson Condie, Neil Conway and J. M. Hellerstein, "Mapreduce online," 2010.

[13] Amazon, "Amazon api gateway concepts." https://docs.aws.amazon.com/apigateway/latest/developerguide/api-gateway-basic-concept.html. Accessed on 2021-03-21.

[14] Amazon, "From 0 to 100 k in seconds: Instant scale with aws lambda." https://aws.amazon.com/es/blogs/startups/from-0-to-100-k-in-seconds-instant-scale-with-aws-lambda/. Accessed on 2021-03-21.

[15] Amazon, "Amazon dynamodb features." https://aws.amazon.com/dynamodb/features/. Accessed on 2021-03-21.

[16] Amazon, "Managing throughput capacity automatically with dynamodb auto scaling." https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/AutoScaling.html/. Accessed on 2021-03-21.

[17] ydarias, "Study on dynamodb." https://github.com/ydarias/dynamodb-test. Accessed on 2021-03-21.

[18] aplyca, "Limitaciones de serverless." https://www.aplyca.com/es/blog/limitaciones-de-serverless. Accessed on 2021-03-21.

[19] dashbird, "What's the benefit of using an api gateway." https://dashbird.io/knowledge-base/api-gateway/pros-and-cons-of-using-an-api-gateway/. Accessed on 2021-03-21.

[20] DZone, "The pros and cons of aws lambda." https://dzone.com/articles/the-pros-and-cons-of-aws-lambda#:~:text=Con%3A%20More%20Complex%20Call%20Patterns&text=AWS%20Lambda%20functions%20are%20timeboxed,distributed%20fashion%20on%20your%20data. Accessed on 2021-03-21.

[21] Postman, "Postman main site." https://www.postman.com/. Accessed on 2021-05-08.

[22] Postman, "Using the collection runner." https://learning.postman.com/docs/running-collections/intro-to-collection-runs/. Accessed on 2021-05-08.