International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Mar-2024)

http://dx.doi.org/10.12785/ijcds/150183

Using Cloud Services to Improve Weather Forecasting Based on Weather Big Data Scraped From Web Sources

Abderrahim El Mhouti[1], Mohamed Fahim[1], Asmae Bahbah[2], Yassine El Borji[3], Adil Soufi[4] and Mohamed Erradi[2]

[1]*ISISA, FS, Abdelmalek Essaadi University, Tetouan, Morocco*
[2]*S2IPU, ENS, Abdelmalek Essaadi University, Tetouan, Morocco*
[3]*SOVIA, ENSAH, Abdelmalek Essaadi University, Tetouan, Morocco*
[4]*FSTH, Abdelmalek Essaadi University, Tetouan, Morocco*

Received 11 Jun. 2023, Revised 29 Jan. 2024, Accepted 10 Feb. 2024, Published 1 Mar. 2024

Abstract: Big Data (BD) scraping systems are among the recommended approaches for large-scale web data extraction. However, these systems for collecting large amounts of data face many challenges, including processing, storage, and data extraction reliability. Due to its potentials, cloud computing is becoming a viable solution to support BD scraping systems. This paper tenders a cloud based-web scraping framework for weather BD extraction and analysis. The aim is to extract weather BD from web sources, analyze this data and use it for visualization and forecasting purposes, and this by enabling elastic and on-demand resources. The framework is implemented using Selenium and Amazon Web Services and tested with Morocco weather data. The suggested cloud-based scrapper's performance and scalability analysis reveals that it provides more efficiency in terms of data collecting and analysis, as well as forecast quality, due to its capacity to leverage cloud resources.

Keywords: Cloud computing, Weather Big Data, Weather forecasting, Web scraping

1. Introduction

Over the last years, and with the emerging evolution of web technologies, many sections of society are now interested in the concept of Big Data (BD). In this context, weather institutions are not excluded since the analysis of weather BD leads to better results in weather forecast and assists forecasters to predict the weather with more precision [1]. The "weather Big Data" available today on the web are becoming increasingly important. Visualizing this data plays a crucial role in understanding what exactly the data means while processing them and inferring results [2]. The beneficiaries of that information are not only people who plan their weekends or holidays, but also organizations in charge of health or civil security [3], and companies in the insurance, energy, transport or agriculture [4].

Today, BD statistics are used extensively by weather organizations all throughout the world. As a result, data collection and conversion to structured data are now possible due to the rise of numerous BD approaches and tools. The technique of web scraping represents one approach to BD with significant potential [5]and is one of the key technologies used to access the BD issued from web sources. It makes it possible to convert web-generated unstructured data into data well structured for archiving and analysis [6].

However, Big Data scraping systems for collecting large amounts of data face many challenges, which concern mainly issues of storage capacity, intensive computing capacity and also the reliability of data extraction [5][7]. In this sense, rapid advancements in cloud computing services make it a suitable platform to support BD scraping systems.

This paper focuses on the use of cloud computing services in the web scraping process of weather BD. The aim is to offer not only weather BD visualization and high-resolution weather forecasts, but also elastic and flexible weather data analysis.

Thus, this paper proposes to set up a weather BD analysis and visualization framework based on web scraping technique and cloud computing services. The proposed framework is implemented and experimented using Morocco weather data issued from a set of web sources. For this, we used the Selenium tool to provide the scraping pro-

E-mail address: a.elmhouti@uae.ac.ma, m.fahim@uae.ac.ma, a.bahbah@uae.ac.ma, y.borji@uae.ac.ma, a.soufi@uae.ac.ma, a.eradi@uae.ac.ma

cess, Amazon Web Services' EC2 (Elastic Compute Cloud) to implement the cloud architecture, and DynamoDB as a NoSQL database to store the weather BD.

The analysis of the performance and the scalability of the developed cloud-based scrapper shows that the latter offers better efficiency in terms of data collection and data analysis and in terms of the quality of forecasts, thanks to the possibility of benefiting from cloud resources. Unlike traditional web scrapers, the proposed cloud-based scraper offers better data extraction efficiency through its human-mimicking web automation methodology. Furthermore, weather BD processing can be moved to cloud servers that provide secure and resizable computing capacity. As for him, the huge amount of weather BD can be easily stored in cloud storage.

Compared to other related leading contributions, the importance of the proposed Framework is reflected in its ability to manage and interpret hundreds of thousands of weather inputs (frequently changing) to create a forecast for an area at a given time. Indeed, several contributions estimate that weather forecasting systems based on a centralized architecture only process petabytes of data per day, which limits the update frequency of forecast engines, and which in turn limits the way which organizations can rely on weather information.

Because the weather is changing rapidly, a system that can quickly process and model data and provide updated forecasts is needed. It is in this sense that the novelty of this work lies in the fact of building the proposed scraper forecast engine on an elastic and scalable AWS cloud infrastructure to quickly manage the data that feeds it, and reliably, accurately and consistently generate high-resolution weather forecasts. Thus, the major contribution of the present work manifests itself in increasing the frequency and the quality of weather forecast modelling to provide people and organizations with information (statistical and descriptive data models) to help them better track weather forecasts.

The remaining elements of this paper are organized as follows: the 2nd section presents the background of weather forecasting using web scraping and cloud-based BD analytics. Section 3 investigates deals on the related works. The 4th section looks at the design of the proposed framework. This section explores also the cloud-based web scraping framework experimentation and discusses and analyses the obtained results. The 5th section concludes this work and addresses the future works.

2. Research Background

In the area of weather forecasting, implementing innovative approaches to collecting and analyzing weather BD from web sources will yield significant results and will assist forecasters to predict the weather more accurately. To achieve this goal, several techniques and technologies are put forward to collect, manage and analyze the enormous volume of weather information from various web sources for weather forecasting purposes. However, the existing solutions face a number of challenges including the detection of noisy data, the availability of hardware and software resources and the management of large data sets themselves.

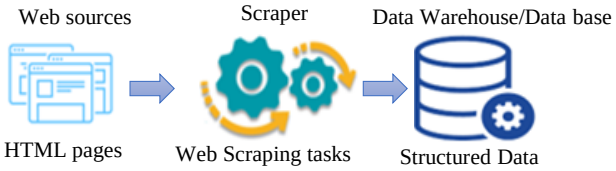

Figure 1. Web scraping sequencing

This paper is a contribution to research efforts in the field of weather forecast based on weather BD from web sources. With weather forecasting as the application, the paper suggests a cloud-based web scraping platform for BD analysis and visualization. The data collection method used web scraping technology, which enables real-time web data extraction. High performance weather computing is made possible by the cloud architecture of the framework.

As part of this work, we proposed to adopt an approach combining several techniques, in particular web scraping as collection technique, BD as data concept and cloud computing as infrastructure. By adopting such BD analysis approach for weather forecasting purposes, the challenges associated with traditional information management methods and technologies can be overcome.

First, since most weather data are provided in real time on websites, we proposed to use web scraping technology as a data collection technique. BD scraping is a method for online crawling and mass data collection from various web sources. It automates the gathering of data and the transformation of data scraped into various forms, such as Excel, txt, CSV, JSON HTML, etc. The collected data are grouped in a database or data warehouse. Fig. 1 describes the Web scraping process.

In BD scraping process, the extraction mode can be manual, Semi-automatic or Automatic. In addition, web scraping for BD requires advanced approaches and technologies such as regular expressions, XPath or DOM (Document Object Model) [8].

The most common tools used to scrap data from web sources are HTTrack, Selenium, cURL, Wget, Scrappy, Node.js, Mozenda, Import.io and PhantomJS. Below, in Table 1, a collection of Web scraping tools of "low level" is provided.

Others related issues of data extraction systems are ETL (Extraction, Transformation, Loading) methods, which gather data from multiple business processes and transfer it to a database or data warehouse. ETL processes take care

TABLE I. EXAMPLES OF WEB SCRAPING SOLUTIONS

Scraping Tool	Language	Description
Rvest	R	R Includes A package that which allows to recover data from Web sources.
Goutte	PHP	There is a PHP library that serves for Web scraping/web crawling.
JQuery	JavaScript	It is a JavaScript library that facilitate, using asynchronous requests (AJAX), the Web scraping in the client's favor.
Scrapy	Python	There is a library for Python used for Web crawling/scraping purposes.
Selenium Web-Driver	Java	It is an API (Application Programming Interface) used for programing actions on the interface, and for checking the answers, and also scrape the data.

of operations that happen in the architecture of the data warehouse backstage.

The web scraping method is used in various activities including BD purposes. Examples of research work about the adoption of web scraping technique for the extraction of weather and other types of data are discussed in the "related works" section.

Secondly, the use of web scraping as a technique to collect weather data from web sources allows to generate a great amount of variated data constituting a set of weather BD. The BD concept describes large-scale data tools and infrastructures that handle increasing demands in processing data volume, variability and velocity. It is related to the variety of various data formats with both batch and stream processing in many domains [9]. Besides the size of data, most studies broaden the definition of BD to include five essential elements (5V of BD): volume, velocity, variety, value, and veracity [10][11].

Many BD systems have emerged in recent years. Hadoop [12], Spark [13]and Flink are designed and developed as multipurpose BD systems, whereas a number of others, like GraphLab [14]and SciDB [15] are created to handle particular types of workloads.

Finally, a cloud computing architecture is required to deliver quick and accurate predictions on a flexible basis, using the newest hardware and software systems, and in a financially responsible manner, even if the availability of weather data sets is important to generate the forecast. In cloud computing, common Internet protocols and networking standards are used to access this technology, which utilizes virtualized resources [16]. Cloud computing is also a concept for providing practical, ubiquitous and on-demand network access to a pool of computing resources that can be quickly supplied and released with little administration labor or service provider contact [17].

The following main traits set the cloud computing notion apart from existing computer paradigms: broad network access, on-demand self-service, resource pooling, rapid scalability and elasticity and measured service [18]. The cloud paradigm offers a wide range of services. Software, Platform and Infrastructure are the three main fundamental service models (or layers). On the other hand, a cloud environment can be deployed in four different ways [16]: private, public, hybrid or community cloud.

In the following, we discuss some research works proposing weather scraping systems and other types of BD, while focusing on the importance of cloud services for such systems.

3. Related Works

In today's digitized world, many information is put online and the size of online data is becoming huge day by day [19]. This vast quantity of data known as Big Data has become an integral part of all industries and business sectors, especially, weather forecasting as an indispensable and important procedure in people's daily life. In this context, the adoption of web scraping technique for weather BD extraction has given rise to numerous applications. Many weather institutions and academics have shown interest in these applications and various cases of weather BD extraction using web scraping technique have been discussed.

This section discusses a description of closely related works and existing issues in order to highlight the novelty and motivation of this work.

Indeed, there are several research works dealing with the weather data collection and analysis for forecasting and decision support purposes. However, in terms of the technological concepts adopted, the approaches proposed differ from one work to another. Some works are based on the concept of weather BD, other works adopt web scraping as a data collection technology, while few of these works use cloud computing technology. Finally, some of these works combine two or three of these technological concepts.

Among these works, in [20], the authors have used web scraping method to extract weather data in cities selected in South Sumatra. The data gathered creates a data warehouse used for more research on data mining of South Sumatra weather forecasts and eventually evolved into a weather-based decision support tool.

In their study presented in [21], the authors rely on

publicly available weather data and web scraping to extract a large amount of traffic accident data. Thus, a collection of weather and traffic accident data are used as an example to show how data mining techniques are employed in order to determine the link between various weather parameters and traffic accidents.

Another study presented in [22] aims to introduce a new web scraping-based approach to generate large and quality meteorological databases from freely searchable web weather sources and apply them to spatial landslide assessments in near real time. In this study, the authors automated the process from real-time meteorological data collection to geostatistical analyses to assess the spatial pattern of precipitation. The findings of this work are immediately applied to physics-based regional landslide sensitivity models.

In the work presented in [23], through an interactive dashboard, the authors have made contributions to an open-source API for web scraping. The aim is to extend the SASSCAL (Southern African Science Service Centre for Climate and Land Management) functionality for weather data extraction, analysis of these statistical data and their visualization. The application is implemented using the R environment and deploys scraping process and data processing techniques in order to support access to weather data of SASSCAL. The developed application lowers the chance of human error and the effort required by researchers to produce the relevant data sets.

The study presented in [24] constructed an analytical BD prediction system for weather temperature using MapReduce algorithm. The proposal was made because the volume of the sensors and the speed of the data in each of the sensors make the processing of the data long and complex. The proposed model facilitates sensor data management using scaling.

In addition to weather, there are many fields including security, transportation, energy, entertainment, finance and emergency services, that rely on web scraping for quick and efficient collection and analysis of available BD to easily make quality decisions. As a first example, in their study presented in [25], the authors have proposed to analyze BD in real time using scraping method in the Apache Spark environment. Flipkart mobile sentiment data is analyzed in real time. Machine learning tools are used to predict the ratings of different types of mobile phones.

E-commerce is also a vast field of application of web scraping technologies to analyze BD. Reference [26] used web crawling and scraping methods to collect HTML data from an e-commerce website in order to determine when a product has been updated. Also, the work presented in [27] studies the relevance of web scraping in online marketing and e-commerce. The authors outline the benefits of scraping method and give a real-world example that e-commerce companies and internet marketers can use.

Additionally, web scraping is being employed more frequently these days for marketing BD research. More and more ”traces” are being left on the internet, which reveal our tastes, routines, etc. For marketing analysis (particularly in online advertising), these data can certainly be used. Web scraping concept has become a crucial use in marketing and also in data science, according to [28]. In addition, the authors emphasize the use of social media and open data as scraping targets and offer illustrations of how to apply website content classification in a market research scenario.

Reference [29] show how to use the technique of web scraping to take advantage of big data concerning the opinions of employees of an organization which are constantly provided on different websites. The aim is to help the management of the organization to adjust or modify business decisions using the wishes, dissatisfactions or needs of the employees. In this research, the authors first provide best practices for proposed data collection and transformation for analysis. Then they demonstrate how two datasets including employee reviews were extracted using scraping techniques, how the data was analyzed using text extraction techniques to uncover business insights, and how the results were compared.

Finally, the web is a compelling data resource for the academic research field. Thus, web scraping is an ideal solution to extract and analyze data in this field. In this context, for data-based industries, the study conducted in [5] made an effort to address both scraping and the viability of BD applications in a single architecture based on the cloud. The adopted cloud architecture allows storage and compute resources to be managed elastically on demand taking advantage of Amazon Web Services. In this work, the authors recognize that, for scraping to work effectively with dynamic and interactive web pages, it must simulate human behavior, and this by adopting a scalable cloud architecture allowing to execute several scrapers in parallel. This statement was also mentioned by a work around the design of a web scraping system for massive data extraction [30].

There are other research studies that focus on the evaluation of web scraping tools [31] and the BD extraction for use in psychological research [32].

So, the review and analysis of related work shows that there are several contributions in the area of weather data collection and its use for forecasting and decision support. In this study, we also found that web scraping, as a technology increasingly used for data collection from web sources, has been the subject of several scientific papers and research analyses. This web data extraction technology is being used in various fields, including weather data analysis and visualization.

In addition, several works have attempted to use web scraping to collect BD for various purposes. However, the research of the state of the art that was done shows that,

when it comes to performing web scraping on a large scale, it becomes an almost impossible activity with a single computer because it demands a lot of time and storage space to accomplish its task due to the limitations imposed by the Internet network.

Indeed, there are a significant difference between usual data scraping and BD scraping. In large scale web scraping systems, more advanced technologies and approaches are required. Thus, web scraping of BD needs the deployment of multiple instances of scraper in a cluster and the provision of sufficient hardware and software resources. While achieving such an architecture with conventional internally housed servers is challenging, we think it is rather easy to design with cloud services. It is in this same sense that this work proposes to combine the three technological concepts (web scraping, big data and cloud computing) to build a web scraping system for weather BD analysis based on a cloud computing architecture. Such a system, based on an elastic and scalable cloud infrastructure, will allow to quickly manage and analyses the data that feeds it, and reliably, accurately and consistently generate relevant weather forecasts.

4. A Cloud-Based Web Scraping Framework For Weather BD Analysis And Visualisation

As explored in the previous sections, when it comes to large amounts of data, traditional web scraping systems survive with many concerns that revolve principally around the processing resources optimization and storage requirements. It is in this sense that we propose a cloud-based web scraping framework for weather BD analysis and visualization to improve weather forecasting.

Within this section, the paper explores the design study of the proposed scraper, deals with its implementation and presents and discusses the results of its experimentation.

A. *Scraper architecture*

The proposed cloud-based scraper is based on a cloud architecture and a set of modules that ensure BD extraction, analysis and visualization.

On the one hand, for scraping process to work effectively with interactive and dynamic web pages, all scraping resources are deployed on a cloud environment (Amazon Web Services). Thus, with the resources of a scalable cloud architecture, it is easier to run scraping process in parallel and the resources can be scaled as needed if the load becomes too heavy. Utilizing cloud services has the benefit of allowing the scraper to work in a manner that allows it to scrape several URLs simultaneously and use as many resources as necessary. The proposed system uses Amazon cloud services such as SQS and S3 to store retrieved content, schedule jobs and generate more resources.

On the other hand, the proposed scraper is based on the ETL process. Web scraping process includes 3 main phases: useful data collection, transformation of data and then loading of data . Extracted data are kept in a data warehouse for storage in various formats. Fig. 2 illustrates the proposed web scraper architecture.

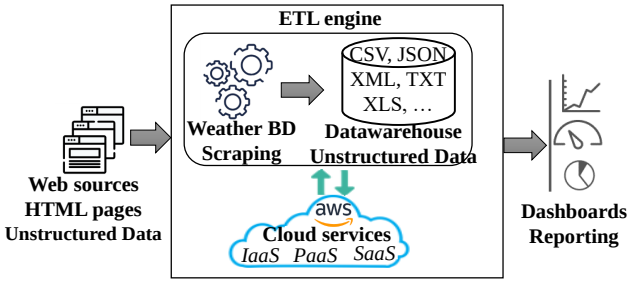

Figure 2. The proposed cloud-based web scraper architecture

The raw weather BD collected are transformed into interpretable information in three stages. The first stage involves extracting the raw data. This involves retrieving data from a number of resources on the Internet, by identifying the source and using a practical web scraping tool that eliminates the need to write scripts.

The second stage is the data analysis stage, which involves cleaning the data and checking its accuracy, as the quality of the data can directly affect the result of the analysis. In this stage, the data will be sent to users structured in various formats (CSV, JSON, XML, TXT, etc.) and used for visualization purposes in dashboards.

The third stage is the structured data storage stage. This is a key stage because it enables the data to be preserved. Data storage varies in terms of capacity and speed. That's why we use cloud storage services.

Once analyzed, this data can be used for weather forecasting purposes and can be viewed via an interactive dashboard using Qlik Sense data analytics tool.

B. *Implementation and experimental setup*

To experiment the proposed scraper, we implemented the proposed scraper using Java as development language, Selenium tool for test automation for the web and Qlik Sense for data visualization.

On the other hand, the cloud architecture is implemented by using an Amazon web server to manage the scraping and DB tools. Thus, SQS (Simple Queue Service), EC2 (Elastic Compute Cloud) and S3 (Simple Storage Service) are adopted to enable elastic and on-demand management of processing and storage resources. Amazon Machine Image (AMI), which produces an instance of the EC2 virtual machine, is stored via the S3 service. The rationale for choosing EC2 is that it enables scalable application deployment using virtual machine instances that can be created, started, and stopped as needed, making the proposed model flexible, efficient and with elastic resources. Fig. 3 illustrates the software architecture of the suggested web scraper cloud

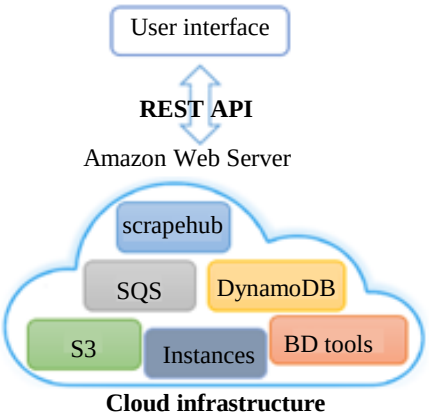

Figure 3. Cloud infrastructure implementation

infrastructure.

In the proposed framework, the user interface is connected to the Amazon web server deployed in the cloud using the REST API (Representational State Transfer). The latter is a software architecture that defines a pattern for server and client communications over the network. As for scrape-hub, it is used to enable the processing of scraping requests as well as event management and monitoring. This tool, which uses Amazons SQS to maintain a list of URLs to scrape, also takes care of allocating the necessary number of instances for the calculation based on the URL block. These URLs are divided into block numbers and stored in the S3 bucket, waiting to be retrieved by an instance of the scraper engine. The latter is a program that uses the URL as input to retrieve the page's content from the virtual machine hosting the scraping engine called an EC2 instance.

In addition, the cloud-based architecture also incorporates the functionality required for BD applications. For example, for storing the retrieved weather data, DynamoDB service is used. DynamoDB is a document-oriented, cross-platform NoSQL database that is free and open source. Once the scraping process is complete, the output can be downloaded in the following formats: CSV, TSV, text, XML, or HTML. The data extracted is used by BD applications for analysis and visualization.

To automate the testing of web applications, Selenium tool has been used. Selenium is considered one of the most widely used web scraping tools due to the fact that it provides web drivers replicating a real user using a browser. The advantage of Selenium is that it allows to navigate on the pages. So, if the user sees the data in his browser, he could scrape it via Selenium. Additionally, it is cross-platform, open source, free and supports a variety of browsers and programming languages. It also permits the usage of web browsers running on distant machines.

To test the implemented scraper, the data used was extracted from several web source from several web

TABLE II. WEATHER DATA SET

Year	Mont	Day	Hour	Temp	Preci	Wind	Hum	Press	Visib	...
…	…	…	…	…	…	…	…	…	…	
2022	1	1	0	11	0	11		1030	10	
2022	1	1	1	13	0	24	74	1029,4	10	
2022	1	1	2	13	0	22		1029	10	
2022	1	1	3	13	0	26		1028	10	
2022	1	1	4	12,2	0	19	77	1028,9	10	
2022	1	1	5	12	0	19		1028	10	
2022	1	1	6	12	0	17		1028	10	
2022	1	1	7	11,2	0	15	81	1029,3	10	
2022	1	1	8	11	0	13		1029	10	
2022	1	1	9	11	0	17		1030	10	
2022	1	1	10	12,8	0	17	77	1031,1	28	
2022	1	1	11	15	0	17		1031	10	
2022	1	1	12	17	0	17		1031	10	
2022	1	1	13	18,3	0	17	56	1031	40	
2022	1	1	14	19	0	7		1030	10	
2022	1	1	15	19	0	6		1029	10	
2022	1	1	16	18,1	0	11	62	1029,6	30	
2022	1	1	17	17	0	11		1029	10	
2022	1	1	18	15	0	9		1030	10	
2022	1	1	19	13,8	0	6	78	1030,4	28	
2022	1	1	20	13	0	6		1030	10	
2022	1	1	21	12	0	9		1030	10	
2022	1	1	22	11,2	0	9	85	1030,9	17	
…	…	…	…	…	…	…	…	…	…	

sources, including in particular sources presenting meteorological data on Morocco. Among these sources, the http://www.meteoma.net website which represent a weather information site that provides meteorological information for more than 270 Moroccan towns and villages. This source provides safe and generally protected family content on temperature (maximum and minimum), wind (direction and speed), humidity, etc. To ensure data scraping from the web source, the HTML code of the web pages must be extracted to analyze where the tags and data are. This is ensured using web browser inspector.

Generally, the data generated are unstructured at first, which becomes a difficult task to analyze. Table 2 illustrates an extract of the raw data collected.

The weather data extracted are treated using Talend Open Studio and then they are sent for storage to the Data Warehouse which uses fault tolerant file systems because there is data replication and if one of the nodes fails the whole configuration does not crash. The number of records used is set to more than 8500 and the block size is more than 26 GB.

Thus, once the unstructured data are sent to the data warehouse, the data can be used by analysis tools. The input weather data includes values for temperature (max and min), location, time, wind (direction and speed), humidity, etc. The data file (Table 2) is divided into data blocks using the input split method to control the block size. On the AWS is the mapper process that receives each divided data block (record) and processes it. In this work, we used a NoSQL database based on the key-value model. In the proposed Framework, key-value pairs are made by a mapping process where the key represents the location

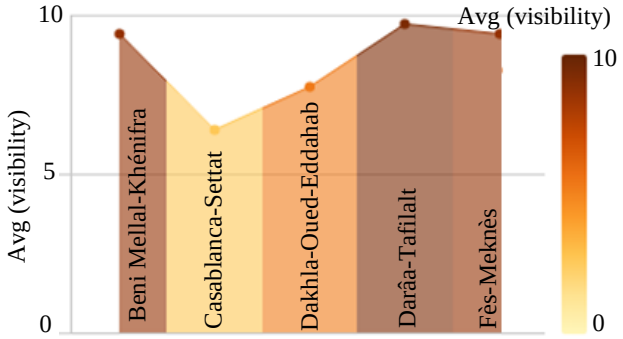

Figure 4. Descriptive visualisation of weather data

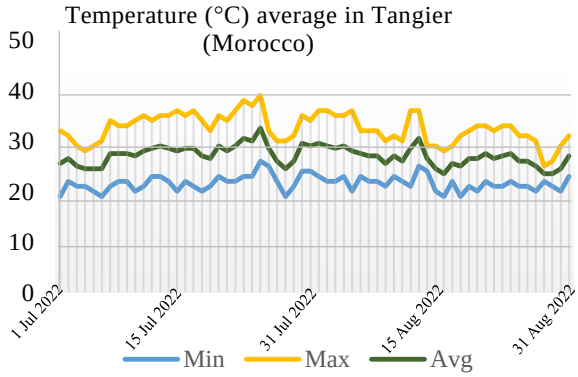

Figure 5. Temperature forecasts for the city of Tangier during the months of July and August

and the value represents the time. The key-value pairs are organized into lists according to key type. The sorted outputs are then used as inputs to simplify the jobs, which in turn simplifies the dataset volume's values. In addition, the AWS architecture controls all other tasks, such as resources and scheduling.

C. *Results and discussion*

The dashboard module of the proposed cloud-based web scraper produces many types of weather data visualizations allowing to generate precise and high-resolution weather forecasts. The sheet in Fig. 4 is an example of descriptive of weather data which includes 5 regions with maximum visibility.

Similarly, Fig. 5 illustrates the results of temperature forecasts for the city of Tangier during the months of July and August during the year 2022.

Thanks to these dashboard's data trends, it is feasible to get accurate estimations and draw statistical judgments about the weather and its repercussions throughout the season. These statistics, for instance, make it possible to predict locations and periods when temperatures are likely to be highest or lowest (Fig. 5). Additionally, they aid in

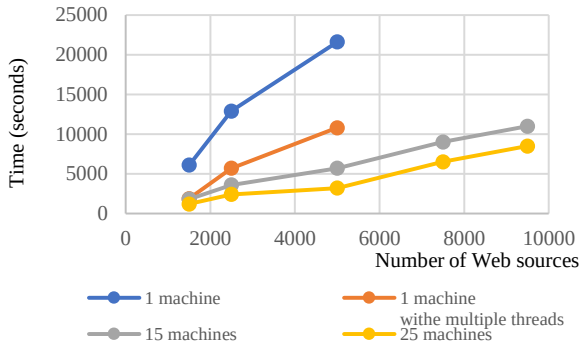

Figure 6. Average completion scraping time

determining visibility in a certain location (Fig. 4).

In contrast, we examined the time required for the scraper to finish the scraping operation in order to assess its performance/scalability. The number of online sources collected as a function of the scraping time necessary to complete the scraping in seconds is shown in Fig. 6.

This graph illustrates the scraping time required for the designed cloud-based framework with the increase in the number of web sources to over 7000 sources. In this graphic, the displayed curves represent the scraping time in a cloud implementation with 15 machines and with 25 machines, and the scraping time in a non-cloud implementation with one machine and with one machine using multiple threads.

We specify here that, for feasibility reasons, the virtual machines used for the comparison, as shown in Fig. 6, have a large processing capacity which is in good agreement with our ability to support the expenses in terms of pricing for AWS cloud-based machines.

Following the evaluation of the time required for the scraper to complete the scraping, we discover that as we approach 5000 online sources, data collecting gets progressively complex with a single computer, which can be rationalized by the fact that servers limit the quantity of data that one IP address can access in a given period of time. The same is true of the machine that uses concurrent threads. The efficiency of the scraper declines as the number of web sources rises, despite the computation being significantly faster than the sequential node. Meanwhile, the proposed cloud-base scraper addresses this issue with the use of multiple scraping machines distributed across the cloud. We can easily see that using more cloud machines gives less computing time, which is an essential resource.

Indeed, in cloud implementations with 15 machines and with 25 machines, the data collection time is approximately linear. This means that the proposed architecture works

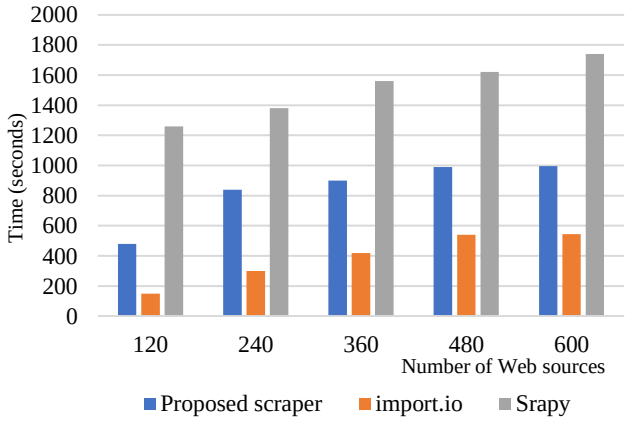

Figure 7. Comparaison of the proposed scraper with others scraper's in cloud

normally with the progressive increase of the number of web sources to be scanned. It can be deduced from the above experience that the cloud-based architecture is well scalable with an increase in the number of web sources to be collected. Furthermore, the scraping process is likely to be faster using an extended cloud architecture than the one we used in this study.

Then, in order to examine the performance of the designed scraper, we compared it with some open-source cloud-based web scraping solutions. In this comparison, a single machine instance was used for the experiment. Fig. 7 above shows a performance comparison of the proposed scraper compared with Import.io and Scrapy, and this in terms of scraping time against a varied number of web sources.

The suggested Framework is distributed and has no restrictions on the number of instances of computing resources that can be employed, in contrast to current scraping models. DynamoDB enables flexible resource use in the cloud. Indeed, although the collection time of the proposed scraper is higher than that of Import.io, the proposed scraper offers more reliability and efficiency for scraping because the use of cloud services allows it to exploit the resources with a lot of flexibility. The Selenium tool allows freedom in the design of the scraping model because it is better at simulating human behavior. At the level of data extraction, the scraping performance of the Selenium tool is comparable to that of other scrapers currently on the market and it remains among the best performing scraping tools in terms of functionality [31].

Thus, the experimentation conducted shows that the proposed Framework is able to handle an increasing scraping workload (number of web sources) in a normal way. This architecture fully adopts the underlying resources to handle the increased scraping workload. The software technologies deployed on the cloud also support the scalability of the system.

On the other hand, since data extraction is only the first step of the DB, the proposed scraping framework can be coupled with other BD applications. Indeed, the proposed architecture is compatible with other BD applications. It permits web scraping to get DynamoDB data and the development of models for the analysis of such data.

By comparing the forecast precision of the proposed system with that of the normal weather forecasts of the meteorological authorities, we deduce that the use of this type of technology for large-scale weather data analysis has the potential to significantly improve weather forecasting. Indeed, local meteorological authorities usually present data in more or less raw formats (tables, figures, percentages, etc.). Therefore, compared to what is presented by these local weather authorities (e.g., the site where the data are collected), and given the potentialities of the proposed system, the latter allows to efficiently analyze the data, to represent them in several statistical data forms and to provide updated weather forecasts.

Finally, the legitimacy and ethical use of web scraping methods are often issues that concern researchers in this field. For this, web scraping strategies must take into account two essential dimensions: copyright and unauthorized seizure. Regarding this work, note that we have followed an appropriate procedure in terms of copyright and ethical issues. Thus, for all the web sources we used, we asked the owner of the source to authorize us to use its data.

5. Conclusion And Future Work

To overcome the limitations of traditional web scraping systems used in the context of BD, this work proposed to build a web scraper based on a cloud architecture, allowing the weather BD analysis and visualization and improving weather forecasting. Data processing and modelling workloads run on Amazon EC2, a service that provides secure, scalable computing capacity in the cloud.

The analysis of the performance of the designed cloud-based scrapper shows that the latter is effective on data extraction due to its ability to run a large number of parallel instances in the cloud. The sophisticated cloud computing technology allows the proposed web scraper to operate more efficiently. The cloud services enable elastic computation and on-demand access to the storage resources in a distributed setting. Thus, the web scraping system is developed as a Java application, deployed in a cloud environment, permitting the analysis of data, its management and visualization via a Dashboard and the use of this data thereafter for forecasting purposes in relation to the weather.

The suggested framework serves as a reference for academics who want to build similar systems in the weather domain or other areas that produce large amounts of data. In this sense, the proposed architecture is compatible with other open-source BD applications.

In terms of the perspectives of this work, we have

planned to evaluate the performance and scalability of the proposed architecture using a large number of nodes and web resources. Based on a set of measurements performed on the prototype and the first results, we present an analysis of the performance and scalability of the scraper, from low-level features to large-scale applications. Also, for the scraper to be able to produce accurate weather forecasts, it is suggested that the data collection process be continued using web scraping tools, data mining, and machine learning methods. With the inclusion of layers of machine learning, more timely and valuable insights can be generated for weather-dependent organizations.

References

[1] M. Fathi, M. H. Kashani, S. M. Jameii, and E. Mahdipour, "Big data analytics in weather forecasting: A systematic review," *Archives of Computational Methods in Engineering*, vol. 29, pp. 1247–1275, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237828595

[2] S. A. Hirve and C. P. Reddy, "Data visualisation using augmented reality for education system," *Int. J. Comput. Appl. Technol.*, vol. 68, pp. 292–297, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251501901

[3] C.-H. Lee, S. Lin, C.-L. Kao, M.-Y. Hong, P. Huang, C.-L. Shih, and C.-C. Chuang, "Impact of climate change on disaster events in metropolitan cities -trend of disasters reported by taiwan national medical response and preparedness system." *Environmental research*, vol. 183, p. 109186, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211229953

[4] V. Ramachandran, R. Ramalakshmi, B. P. Kavin, I. Hussain, A. H. Almaliki, A. A. Almaliki, A. Y. Elnaggar, and E. E. Hussein, "Exploiting iot and its enabled technologies for irrigation needs in agriculture," *Water*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247144743

[5] R. S. Chaulagain, S. Pandey, S. R. Basnet, and S. Shakya, "Cloud based web scraping for big data applications," *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 138–143, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:25361975

[6] E. Vargiu and M. Urru, "Exploiting web scraping in a collaborative filtering- based approach to web advertising," *Artif. Intell. Res.*, vol. 2, pp. 44–54, 2012. [Online]. Available: https://api.semanticscholar.org/CorpusID:29662240

[7] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *International Journal of Advances in Soft Computing and its Applications*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245584401

[8] P. Vassiliadis and A. Simitsis, "Extraction, transformation, and loading," in *Encyclopedia of Database Systems*, 2009.

[9] R. Kune, P. K. Konugurthi, A. Agarwal, C. R. Rao, and R. Buyya, "The anatomy of big data computing," *Software: Practice and Experience*, vol. 46, pp. 105 – 79, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:6074976

[10] I. K. Nti, J. A. Quarcoo, J. Aning, and G. K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects," *Big Data Min. Anal.*, vol. 5, pp. 81–97, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246337082

[11] M. Trifu and M. L. Ivan, "Big data: present and future," *Database Systems Journal*, vol. 5, pp. 32–41, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:167395609

[12] T. White, "Hadoop: The definitive guide," 2009.

[13] M. A. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Symposium on Networked Systems Design and Implementation*, 2012.

[14] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein, "Distributed graphlab : A framework for machine learning and data mining in the cloud," 2012.

[15] P. G. Brown, "Overview of scidb: large scale array storage, processing and analysis," *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:14544985

[16] G. Bhure and S. Bansod, "E-learning using cloud computing," *International Journal of Information and Computation Technology*, 2014. [Online]. Available: https://www.ripublication.com/irph/ijict_spl/ijictv4n1spl_07.pdf

[17] P. Mell and T. Grance, "The nist definition of cloud computing," 2011.

[18] L. Wang, J. Tao, M. Kunze, A. C. Castellanos, D. Kramer, and W. Karl, "Scientific cloud computing: Early definition and experience," *2008 10th IEEE International Conference on High Performance Computing and Communications*, pp. 825–830, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:13391099

[19] B. H. AlOwaimer and S. Mishra, "Analysis of web browser for digital forensics investigation," *Int. J. Comput. Appl. Technol.*, vol. 65, pp. 160–172, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235190098

[20] Y. Fatmasari, N. Kunang, and S. D. Purnamasari, "Web scraping techniques to collect weather data in south sumatera," *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pp. 385–390, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:57762402

[21] M. Novkovic, M. Arsenovic, S. Sladojevic, A. Anderla, and D. Stefanovic, "Data science applied to extract insights from data -weather data influence on traffic accidents," *INFOTEH-JAHORINA*, vol. 16, pp. 387–392, 2017. [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=rcGi9XYAAAAJ&citation_for_view=rcGi9XYAAAAJ:W7OEmFMy1HYC

[22] E. Canli, M. Mergili, T. Glade, and B. Loigge, "Generating web scraped high-quality weather databases for near-real-time derivation of spatial landslide susceptibility," 2018.

[23] T. S. Thapelo, M. Namoshe, O. Matsebe, T. Motshegwa, and M.-J. M. Bopape, "Sasscal websapi: A web scraping application programming interface to support access to sasscal's weather data," *Data Science Journal*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237719804

[24] K. A. Ismail, M. A. Majid, J. M. Zain, and N. A. A. Bakar, "Big data prediction framework for weather temperature

based on mapreduce algorithm," *2016 IEEE Conference on Open Systems (ICOS)*, pp. 13–17, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:36761617

[25] P. Ganguly, G. Parihar, and M. Sivagami, "Real-time big data analysis using web scraping in apache spark environment: Case study—mobile data analysis from flipkart," *Artificial Intelligence and Technologies*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:245307088

[26] I. E. Onyenwe, E. G. Onyedinma, C. A. Nwafor, and O. Agbata, "Developing products update-alert system for e-commerce websites users using html data and web scraping technique," *ArXiv*, vol. abs/2109.00656, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237386196

[27] K. Henrys, "Importance of web scraping in e-commerce and e-marketing," *SSRN Electronic Journal*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:234950503

[28] M. Herrmann and L. Hoyden, "Applied webscraping in market research," 2016.

[29] L. G. Tanasescu, A. Vines, A.-R. Bologa, and C. A. Vaida, "Big data etl process and its impact on text mining analysis for employees' reviews," *Applied Sciences*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251152442

[30] S. G. Upadhyay, V. Pant, S. Bhasin, and M. K. Pattanshetti, "Articulating the construction of a web scraper for massive data extraction," *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–4, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:38507855

[31] E. Persson, "Evaluating tools and techniques for web scraping," *Master's Thesis, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology*, 2019. [Online]. Available: http://kth.diva-portal.org/smash/get/diva2:1415998/FULLTEXT01.pdf

[32] R. N. Landers, R. C. Brusso, K. J. Cavanaugh, and A. B. Collmus, "A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research." *Psychological methods*, vol. 21 4, pp. 475–492, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:9679303

Prof. A. El Mhouti teaches computer science at the Faculty of Science at the Abdelmalek Essaadi University in Tetouan, Morocco. In 2015, he graduated with a PhD in computer science from the same institution. Big Data analytics, machine learning, deep learning, cloud computing and technology enhanced learning are some of its research interests. There are several articles in Mr. El Mhouti's area of expertise.

Prof. M. Fahim teaches computer science at the Faculty of Science and Technologies at the Abdelmalek Essaadi University in Tetouan, Morocco. In 2019, he graduated with a PhD in computer science from Morocco's Moulay Ismail University. Big Data analytics, NLP, machine learning, and educational technology are some of its research interests. In his area of expertise, Mr. Fahim has written and published several articles.

Mrs. A. Bahbah earned in 2015 a master's degree in applied mathematics. She is a PhD candidate at the Abdelmalek Essaadi University in Morocco's. She focuses on the computerized modeling of students' and teachers' approaches to solving mathematical problems. She participates in a number of conferences both domestically and abroad..

Prof. Y. El Borji teaches computer science at the National School of Applied Sciences belongs to the Abdelmalek Essaadi University in Tetouan, Morocco. In 2016, he graduated with a PhD in computer science from the same university. Integrated data models, gamification, serious games, mixed reality, simulations, machine learning and data sharing formats are some of its study interests.

Prof. Adil Soufi teaches computer science at the Faculty of Science and Technologies belonging to the Abdelmalek Essaadi University in Tetouan, Morocco. A PhD in computer science was earned by Mr. Soufi from the same institution. Machine learning, online learning, modeling, and epidemic model fitting are some of his research specialties. In his research fields, he has published a number of articles.

Prof. Mohamed Erradi teaches and trains in educational technology and multimedia engineering at Ecole Normale Supérieure belonging to the Abdelmalek Essaadi University in Morocco. A PhD in physics and chemical sciences was earned by Mr. Erradi from the Mohammed V University in Morocco. Educational technologies, technology enhanced learning and e-pedagogy are some of his research domains. He has published several articles in these areas of research.