Big Data Technology for Scientific Applications

George Suciu
R&D Department
BEIA Consult International
Bucharest, Romania
george@beia.ro

Muneeb Anwar
R&D Department
BEIA Consult International
Bucharest, Romania
ma@beia.ro

Ioana Rogojanu
R&D Department
BEIA Consult International
Bucharest, Romania
ioana.rogojanu@beia.ro

Adrian Pasat
R&D Department
BEIA Consult International
Bucharest, Romania
adrian.pasat@beia.ro

Alina Stanoiu
R&D Department
BEIA Consult International
Bucharest, Romania
alina.stanoiu@beia.ro

***Abstract*—Big Data refers to volumes which exceed the capacity of current online processing and storage systems. Public data and online data search are hampered because data sets remain scattered. Data storage is a technology issue which seems to be solvable soon by cloud computing, but at this moment a large-capacity and low-cost storage solution represent a long-term challenge which requires new paradigms and research. Big Data is not only narrowed to data viewpoint, but it has surfaced as a stream that includes combined technologies, tools, and real-world applications. The paper exhibits a simple, general and short introduction of Big Data from Internet of Things (IoT) devices. The purpose of this paper is to analyze different technologies used for storing and managing Big Data for research purposes.**

***Keywords*—Big Data, data storage, public data, IoT, Raspberry Pi, Pycom.**

I. Introduction

The era of Big Data has involved a great number of available datasets which are dynamic and heterogeneous. The difficulty level for transforming a normal user into someone who can explore the data is more burdensome nowadays due to the great amount of data.

Data processing has become a major research topic of modern science, involving several challenges related to data visualization, interaction, storage and personalization. Data visualization tools provide ways for users to explore and analyze datasets.

These data are generated from e-mails, search queries, posts, sensors, geographic information systems (GIS), applications and stored in databases. With the advances in Cloud Computing and Internet of Things (IoT), Big Data is experiencing an exponential growth [1]. Big Data is a range of data, which cannot be collected and concocted by a single machine. Big Data does not refer to the data only large in size.

The most well-known rendition of Big Data simultaneously is a four Vs concept: Volume, Velocity, Variety, and Veracity. So, data controls large volume, comes with high velocity, from a variety of sources and formats and having surpassing conjecture is referred to as Big Data. Big Data has tremendous volume. Velocity- involves the speed of production and processing of data i.e. proportion of penetrating streaming data into the system is rapid. Variety- applies to a distinctive form of data, i.e. unorganized or semi-structured data (text, sensor data, audio, video, click stream, log file, XML) introduced from different sources. Veracity- points ambiguity of data, i.e. quality of data being captured. Data like posts on social networking sites are imprecise.

The incredible advancement of social networks, IoT, the combined objects, and mobile technology is prompting an extraordinary increase of data which all corporations are defined with. The technologies broadly produce significances of data which has to be gathered, characterized, deployed, deposited, analyzed and so on. From the structure of the Internet of Things, data mining for big data is significant for IoT to advance the intelligence assistance in several applications [2]. The conception and availability of big data, mainly effluent data, have cherished privacy interests among the common public and these concerns are expected to grow and diversify [3].

This paper aims to provide a brief introduction of Big Data concepts. Also, it would be show a way of an efficient way for storing data coming from sensors connected to different devices.

The rest of the paper is organized as follows: Section II presents related work, Section III describes the devices used to gather data from sensors and how to manage the amount of data, Section IV describes a method which allows the users to create a relevant database for scientific research, while Section V concludes the paper.

II. Related Work

This section will review several modern technologies created for Big Data analysis.

Most traditional visualization systems can't handle the size of actual available datasets. They are restricted to deal with small size datasets, which can be analyzed with conventional techniques [4]. The 5G (fifth generation) of communication technologies promotes the IoT technologies in numerous applications, mainly in healthcare. It enables 100 times greater wireless bandwidth with energy conservation and maximum storehouse utilization by implementing big data analytics [5]. Oracle implements the authority of both Spark and Hadoop able to be combined with the present data of industries which were earlier using Oracle applications and Database. The co-operation performed is deeply performant, secure and adequately automated. Oracle Big Data contributes to the integrated collection of products to prepare and analyze several data sources to obtain further insights and take advantage of unknown relationships [6].

An entire IoT system should be efficient to advance great data method for storing, processing, and examining data. This kind of system is Hadoop MapReduce [7]. It is an innovative data analysis and processing tool. Apache Spark [8] is a complex data partition system. With in-memory ability, it challenged to be permanent than MapReduce by hundred times. Apache Flume [9] is a classified, secure service for gathering, aggregating and transferring vast volumes of streaming data. Apache Kafka [10] is a highly classified,

publish-subscribe messaging method. With Kafka, data can be utilized by various applications. Orion Context Broker [11] implements a publish-subscribe tool for enrolling connection elements and enduring them by updates and inquiries. Apache Flink [12] is a cascade real-time stream processing engine that produces data dissemination, communication, and fault-tolerance. Nevertheless, there are presently numerous innovative strategies based on Platform as a Service (PaaS) contributing more manageable IoT services with data processing capability motivated from Big Data methods and the opportunity to contrive storage cloud locally and globally. The purpose of enlarging the PaaS appearance to IoT is to introduce a platform devoted to IoT developers that can overcome the time-to-market for an application by lowering the development costs [13].

Big Data techniques facilitate the processing of the immense amount of data generated by sensors. These techniques provide creating actionable information and knowledge out of raw data [14]. To this end, the local and global clouds approach the other connection challenge: when the Internet is not accessible, the user can still reach some IoT functionalities from the local Cloud.

As NoSQL databases become popular for Big Data storage, there are several time series databases available such as Graphite [15], which has an operation-ready monitoring mechanism that operates identically well on inexpensive hardware or Cloud base. Organizations use Graphite to trace the administration of their websites, applications, business services, and networked servers. It designated the start of a recent generation of monitoring tools, causing it accessible than ever to store, recover, share, and envision time-series data. Graphite obtains the most significant ecosystems of data combinations and tools, so one can only use a compilation agent or language ties.

Furthermore, Grafana [16] is an advanced open source platform for time scale analytics. It enables to ask, visualize, warn on and follow the metrics concerning their storage place. Grafana is likely to design, examine, and share dashboards with the partners, because of its data-driven features.

III. IoT Cloud

In this section, we will present the IoT devices used to gather data from sensors such as Raspberry Pi, Pycom and Libelium, as well how to manage the huge amount of data gathered.

For data visualization, we used Grafana [17]. Each Grafana table is a query to its supported Data Source. The supported Data Sources, as shown in Fig. 1, are Graphite [18], Elasticsearch [19], CloudWatch, InfluxDB, OpenTSDB, Prometheus. For our experiments, we have used InfluxDB as Data Source. Grafana has no limitation for the number of displayed metrics, but the more tables and graphs the user has meant more load on the InfluxDB server.

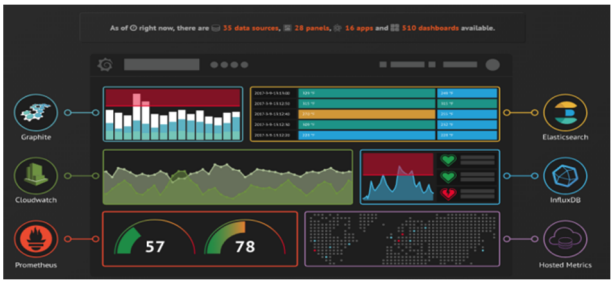

Fig. 1. Grafana's supported Data Sources

A. Raspberry Pi

Raspberry Pi [20] is a single board computer which enables the communication with several devices and users for a large variety of application. Users can create tolerably wondrous smart devices which can collect, process and display data by connecting different types of sensors to Raspberry Pi; this includes sensors like photoresistors, temperature, humidity and so on [21].

In Fig. 2 we present the graph for a temperature sensor connected to Raspberry Pi. In Fig. 3 we show the data history.

Fig. 2. Graph based on data gathered from a temperature sensor

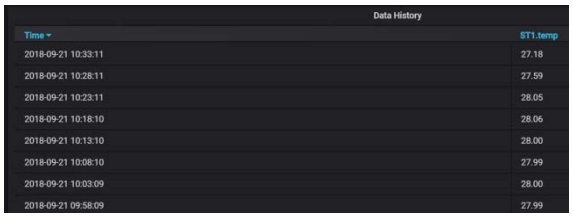

Fig. 3. Data history for Raspberry Pi

B. Pycom

Pycom devices are created for the IoT. They can all connect to the Internet. Pycom is a Micro-Python enabled microcontroller expansion board sustaining various networks including LoRa. Different sorts of sensors can be combined to the Pycom board [22]. LoRa has authorized novel IoT applications in several domains, including health. IoT handles a set of protocols to communicate within devices. LoRa is a low-power wide-area network (LPWAN) protocol and is used for proceeding end-devices (sensors) by low energy consumption [23]. Fig. 4 shows the dashboard created for PyTrack in Grafana, the three axes of the accelerometer and the latitude and longitude. In this case, the pictures were taken with the device indoor, so the latitude and longitude can't be defined.

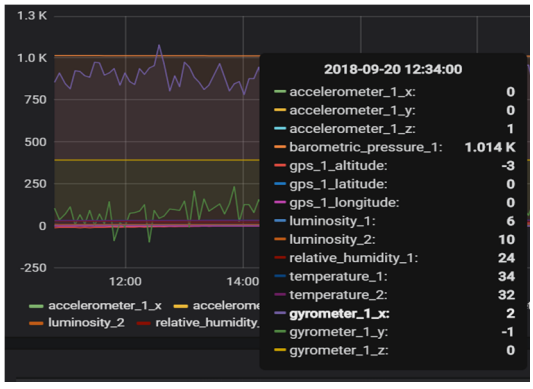

Fig. 4. Graph based on data gathered from Pycom's sensors

PyCom PyTrack [24] was chosen for our experiments providing sufficient coverage. Firstly, the sensors transmit data to the PyCom PyTrack board, then to the gateway and finally to the Cloud Platform. Also, the transferred data is shown on a Grafana dashboard as data history, as seen in Fig. 5.

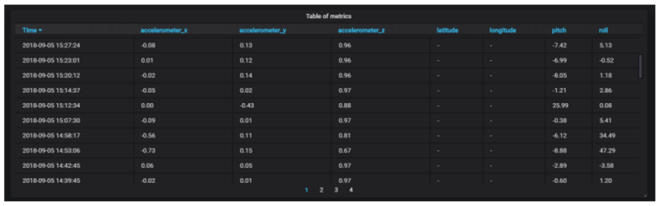

Fig. 5. Data history for PyCom

C. Libelium

Libelium stations [25] are utilized for analyzing the level of pollution - air pollution, agricultural, noise, water, meteorology, water management systems, measurement of renewable energy potential, plant disease management, etc. Libelium stations can support different types of sensors, such as temperature, relative humidity, solar radiation, wind speed, wind direction, barometric pressure, leaf wetness, soil temperature, precipitation, luminosity, and ultraviolet radiation. The data from these sensors are stored in the Libelium gateway, called Meshlium [26].

In Fig. 6 we have connected three sensors (CO, Temperature, Volatile organic compound (VOC)) to a Libelium board.

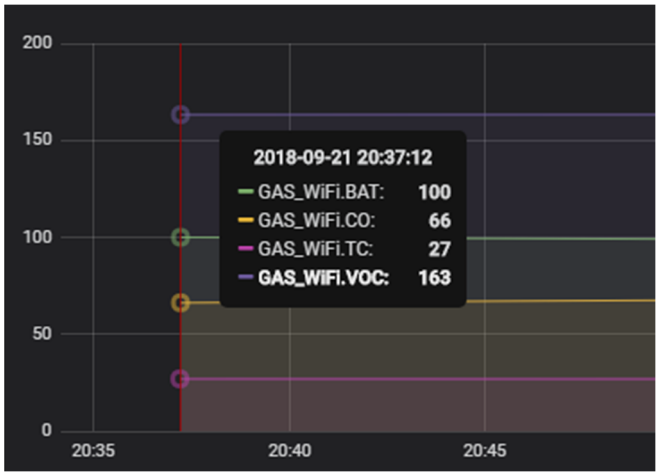

Fig. 6. Graph based on data gathered from Libelium sensors

IV. Big Data for Academic Research

Due to its notable potential in generating business value, Big Data has become the focus of academic and corporate investigation [27]. Environmental studies are often related to locations and regions of interest (ROI). Data resources, which include the IoT metadata and geographic data need to be discoverable through web-based access to authors, keywords, affiliation, spatial context, cover data.

In Fig. 7, we present the results of researching for all available journals and articles which are dealing with Big Data. In order to retrieve the articles, we searched by keywords using a Python script and Scopus library [28].

We can see that the number of articles is increasing since 2014. For 2018, the number of articles is still low because there are still articles which haven't been published yet.

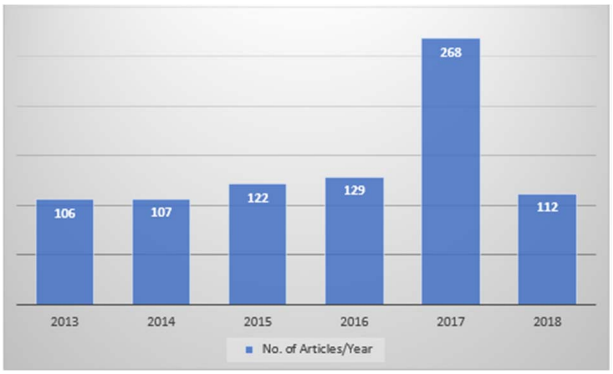

Fig. 7. The distribution of articles related to big data

Based on this graph, we can provide the distribution of articles for each country and identify ROIs.

This method is also useful for creating a relevant database for scientific research. Using Scopus library, we have researched for available online articles which have "big data" as a keyword, but users can search by any keyword, or they can add more keywords and make a more specific database by restricting the topics even more. One of the main advantages of using Scopus, is that it is also available online [29].

V. Conclusion

In this paper, we have presented how the users can manage with the huge amount of data gathered from different sensors which are connected to three different devices. Also, we have analyzed the corpus of different journals in order to identify the importance of Big Data for scientific research. This method is also useful for making a relevant database for scientific research and it also allows the user to identify the ROIs.

As future work, we plan to use other Data Sources and check for differences in order to improve the actual method for Grafana visualization. Also, we will make a relevant database using a Python script and Scholarly library which retrieves all articles available on Google Scholar with the purpose to compare the two libraries.

Acknowledgment

This work has been supported in part by UEFISCDI Romania and MCI through projects WINS@HI, ESTABLISH, 3DSafeguard, and TelMonAer, and funded in part by European Union's Horizon 2020 research and innovation program under grant agreement No. 777996 (SealedGRID project) and No. 787002 (SAFECARE project).

References

[1] Barkham, R., Bokhari, S., &Saiz, A. "Urban Big Data: City Management and Real Estate Markets," GovLab Digest: New York, USA, 2018.

[2] Zhang, Q., Yang, L. T., Chen, Z., & Li, P. "High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT," Information Fusion, vol. 39, pp. 72–80,2018.

[3] Mooney, S. J., Pejaver, V. "Big Data in Public Health: Terminology Machine Learning, and Privacy,". Annual Review of Public Health, vol. 39(1), pp. 95–112, 2018.

[4] Bikakis, N. "Big Data Visualization Tools" arXiv preprint arXiv:1801.08336, 2018.

[5] Dey, N., Hassanien, A. E., Bhatt, C., Ashour, A. S., &Satapathy, S. C. (Eds.). "Internet of Things and Big Data Analytics Toward Next-Generation Intelligence,". Studies in Big Data, vol. 30, 2018.

[6] Oracle. (2017). Big data features [Online]. Retrieved January 22, 2017 from https://cloud. oracle.com/en_US/big-data/features

[7] http://hadoop.apache.org.

[8] http://spark.apache.org.

[9] https://flume.apache.org.

[10] http://kafka.apache.org.

[11] http://catalogue.fiware.org.

[12] http://flink.apache.org.

[13] Suciu, G., Vulpe, A., Martian, A., Halunga, S., &Vizireanu, D. N. (2016). Big data processing for renewable energy telemetry using a decentralized cloud M2M system. Wireless Personal Communications, 87(3), pp. 1113-1128.

[14] Ochian, A., Suciu, G., Fratu, O., & Suciu, V. "Big data search for environmental telemetry. In Communications and Networking," IEEE International Black Sea Conference, pp. 182-184, 2014.

[15] https://graphite.readthedocs.io

[16] https://grafana.com/

[17] https://grafana.beia-telemetrie.ro/

[18] https://graphiteapp.org/

[19] https://www.elastic.co/

[20] https://www.raspberrypi.org/

[21] Schmidt, M. Raspberry Pi; Pragmatic Bookshelf: Raleigh, NC, USA, 2014.

[22] Jalaian, B., Gregory, T., Suri, N., Russell, S., Sadler, L., & Lee, M. "Evaluating LoRaWAN-based IoT devices for the tactical military environment," IEEE 4th World Forum on Internet of Things (WF-IoT). doi:10.1109/wf-iot.2018.8355225, 2018.

[23] Hayati N.Suryanegara M., "The IoT LoRa System Design for Tracking and Monitoring Patient with Mental Disorder ", IEEE International Conference on Communication, Networks and Satellite (Comnetsat), pp. 135-139, 2017.

[24] https://pycom.io/product/pytrack/

[25] http://www.libelium.com

[26] D. S.Gangwar, S.Tyagi, "Challenges and Opportunities for Sensor and Actuator Networks in Indian Agriculture," 8th International Conference on Computational Intelligence and Communication Networks, 2016.

[27] Wamba, S. F., Akter, S., Edwards, A., Chopin, G., &Gnanzou, D. "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study,". International Journal of Production Economics, vol. 165, pp. 234-246, 2015.

[28] https://github.com/scopus-api/scopus

[29] https://www.scopus.com/