Oct 31st, 10:45 AM - 12:30 PM

Big Data Analytics on Cloud: challenges, techniques and technologies

Aleksander Biberaj
University of Tirana

Olimpion Shurdi
University of Tirana

Bledar Kazia
Canadian Institute of Technology (CIT), Tirana, Albania

Renalda Kushe
University of Tirana

Alban Rakipi
University of Tirana

Big Data Analytics on Cloud: challenges, techniques and technologies

Aleksander Biberaj[1] , Olimpion Shurdi[2] , Bledar Kazia[3] , Renalda Kushe[4] , Alban Rakipi[5]
[1]Faculty of Information Technology, Polytechnic University of Tirana, Tirana, 1001, Albania
[2]Faculty of Information Technology, Polytechnic University of Tirana, Tirana, 1001, Albania
[3]Canadian Institute of Technology (CIT), Tirana, Albania
[4]Faculty of Information Technology, Polytechnic University of Tirana, Tirana, 1001, Albania
[5]Faculty of Information Technology, Polytechnic University of Tirana, Tirana, 1001, Albania

Abstract
These days it is known that Big Data Analytics is taking a huge attention from researchers and also from business. We all are witness of the data growth that every institution, company or even individuals store in order to use them in the future. There is a big potential to extract useful data from this Big Data that is stored usually in Cloud because sometimes there is not enough local space to store big amounts of data. There is a huge number of sectors where Big Data can be helpful including economic and business activities, public administration, national security, scientific researches in many areas, etc. This data in order to be used must get processed, usually by using Big Data Analytics Techniques. It is for sure that the future of business and technology will be relied on Big Data Analytics. This paper aims to show how big data is analyzed especially when it is deployed on cloud as well as the challenges, techniques and technologies that are used and can be used, in order to analyze Big Data on Cloud. We discuss and implement different methodologies of Big Data Analytics on Cloud.

Keywords: Big Data, Big Data Analytics, Cloud Computing, Data Storage.

1. Introduction
Since the beginning of digital era, the data generated, stored and deployed has been obviously increased. This is because all businesses and institutions have linked their continuity with the data. Big Data can be represented in different forms like Blogs, Websites, Data Warehouses, Streaming etc. Now Big Data has become a trend [1]. It is necessary to choose an efficient and measurable tool for storing and analyzing these data. Big Data Analytics comes out from a union between two Data fields, Big Data and Data Analytics. The specialists that deal with Big Data Analytics (Analysts), help companies use these data to derive the right and better conclusions but also help different users to make quicker and better decisions. Among the best known techniques for Big Data Analytics can be mentioned text analytics, machine learning, predictive analytics and Data mining. This techniques are applied in the unexploited data that are already stored somewhere. The objective of this paper is to show, based on experiments, what is the most efficient technique for Data Analytics on Cloud

1.1 Big Data
If we refer to the Big Data definition we can say that these are structured, semi-structured or unstructured data that together make a complex data infrastructure [2]. In order to manage such a complex data architecture, strong technical and managerial solution should be chosen. The best way to explain Big Data is the multi-V model that is illustrated in Fig.1

Schema.1 *Multi-V Model Big Data*

As you can see from the previous figure the main characteristics of Big Data are *Value, Volume, Variety, Velocity, and Variability* [3]. Variety is related to the different types of data that can be found in Big Data sets as the Velocity is the rate in which the data is produced. If we talk about Volume it is clear that is related to the size of data. On the other hand the Variability and Value are respectively connected with data reliability and data worth.

Another way to explain Big Data is HACE Theorem represented by Wu, Zhu, Wu and Ding [4]. This theorem is related to Data Mining, which is one of the most used techniques in Big Data Analytics on Cloud. This theorem proposes a Big Data processing model by using data mining techniques. There are two main characteristics that this theorem involves. The first one is related to the volume of data that comes from different sources and the second one has to do with data decentralization and distribution.

1.2 Cloud Computing

NIST (National Institute of Standards and Technology) is an all-around acknowledged organization throughout the world for their standardization in the field of Information Technology. NIST characterizes the Cloud Computing architecture by describing five fundamental characteristics, three cloud service models and four cloud deployment models [5]. Cloud computing has ended up a standout amongst the most talked about innovation in late year and it is characterized as another sort of registering that depends on sharing figuring assets as opposed to having neighborhood servers or individual gadgets to handle applications [5]. Cloud computing means putting your information away and accessing the information and projects over the Internet rather than your PC's hard commute. The Cloud which is an illustration for the web as the administrations is given in the stage that is accessible for all.

Cloud computing is a term for the conveyance of facilitated administrations over the Internet. Cloud computing is equivalent another figuring sort; lattice registering where unused handling cycles of CPUS's over all PCs in a system are bridled to take care of issues seriously as opposed to utilizing stand-alone machine.

The cloud computing, as indicated by Peter and Timothy, can be partitioned into five primary gatherings [5].

1. **On-demand self-service:** A customer can change server time and system stockpiling and other PC equipment and programming settings straightforwardly from the web without really connecting with the supplier to change. Along these lines, they have a reasonable picture of what is finished.
2. **Measured service:** Cloud frameworks control how much the customers utilized the administrations they gave to customers. The basic administrations are capacity, CPU use, data transmission, and GPU utilization. Utilized assets are being checked and afterward a report is prepared so that the client get a diagram of the use and pay.

3. **Broad network access:** The system access is wide as far as it can be utilized over a ton of stages and gadgets that empowers simple correspondence between diverse hubs (e.g. Smartphone's, tablets, work-stations and portable PCs).
4. **Rapid elasticity:** The supplier's capacities can be provisioned and discharged flexibly and naturally. The capacities can be changed to fit the buyer requests and needs, and should be possible whenever.
5. **Resource pooling:** The supplier's assets are pooled and arranged at one area to give administrations to different purchasers through loads of virtual and physical assets that can be appointed and reassigned to fit the customers. The client cannot get too specifically to the supplier's assets. In any case, they have some information of where it is topographically. The assets of suppliers are capacity, system, transfer speed and servers.

There are four diverse deployment models random to what service model that is utilized there are deployment models that can be connected to every one of them [5].

1. **Private cloud:** This kind of cloud structure is to be utilized to one and only association furthermore it is overseen by them. The framework could be taken care of by an outsider or themselves relying upon the administration understanding that exist.
2. **Public cloud:** Public cloud is accessible to the overall population frequently for free or with a payment. The administration can be given by an administration, organizations partnerships, or unions. A decent example is the free online stockpiles; iCloud, drobox, and Google drive are some surely understood cases.
3. **Community cloud:** This sort of cloud is utilized by groups that comprise of numerous association or clients. It might be possessed by the group or by an outsider to serve the group. The supplier's contribution relies on upon the administration display that is relevant to requests of the group. This is well alternative when the organizations are framing associations over the assets they have.
4. **Hybrid cloud:** Hybrid cloud is a Private's blend, open and group cloud models. In spite of the fact that they are just limited together, we have half and half cloud as an arrangement model. A sample could be that an open cloud exists inside of an association for all workers and inside of this cloud is a private cloud available for supervisors.

2. Big Data Analytics on Cloud

These days there are a lot of technologies that are used to adopt big data in cloud. Since big data is deployed in cloud, there is a huge need to use this data in a proper way. This leads to using the term of Big Data Analytics. All this process will help data driven industries to develop and also to forecast what they expect. Big Data analytics on Cloud could provide us a lot of opportunities like:

- Increasing Operational Efficiency.
- Informing Strategic Direction.
- Identifying and developing new products and services.
- Enhanced Customer Experience.
- Identifying New markets.
- Faster go to market.

All these opportunities help the companies in decision-making process in order to establish new developments. But there are a lot of challenges to perform Big Data Analytics on Cloud:

- Data Capture and Storage
- Data Transmitting
- Data Curation
- Data Analysis
- Data Visualization

In order to successfully Analyze big data on cloud, we have to undergo all this steps, based on the type of data we are analyzing. We have to keep in mind that the data in not always structured, so before analyzing data it is better to structure them and also the information retrieval is not always linear, it can also be non-linear. For example, if we execute a query in a 100GB dataset and in a 10GB dataset, the time for the query to execute can be higher on 10 GB dataset than in 100 GB dataset.

2.1 Techniques and Technologies

There are a lot of techniques used in Big Data analytics. These techniques are divided in big groups as:

- Big Data Techniques
- Big Data Tools
- Stream Processing Big Data Tools
- Big Data Tools based on Interactive Analysis

Big Data Techniques: These techniques are used to efficiently process large volume of data in a shortest time possible. In this category there are involved a lot of disciplines like Data Mining, Statistics, Neural Networks, Social Network Analysis, Pattern Recognition etc. Some of the most important techniques are Optimization Methods, Statistics, Data Mining, Machine Learning, Visualization Approaches etc. All these methods apply on Big Data in order to analyze them and get what we want from them.
Big Data Tools: Here there are included the tools for batch processing or said in other words the tools used to distribute the data. Some of the most important tools are Apache Hadoop/MapReduce, Dryad, Apache Mahout, Jaspersoft BI suit, Pentaho Business Analytics, Skytree Server, Tableau, Karmasphere Studio and Analyst, Talend Open Studio.
Stream Processing Big Data Tools: In this group there are included tools that are used for real time data processing. The most import real time processing tools are Storm, S4, SQL-Stream s-Server, Splunk, Apache Kafka, Sap Hana. Some of these tools are a bit old but there are a lot of technologies used these days and these tools are needed to maintain them.
Big Data Tools based on Interactive Analysis: Interactive Analysis is a concept that puts the data into an interactive environment. This concept allows users to analyze the information that they want. The most important tools for this kind of analysis are Google's Dremel and Apache Drill.

3. Experiments

We performed some experiments related to Big Data Analytics on Cloud. For performing these experiments we selected 2 public Datasets:

1. Airline_Ontime_Data – 8GB
2. Dataset Created by us with false data – 100GB

Both this datasets were tested on 3 of the most important Cloud Service providers as Google (Big Query), Amazon (Amazon Web Services) and Microsoft (Azure).
First we had to import the datasets in all of the environments and that perform the same queries for the same dataset in the 3 of the Cloud Provider Platforms. We measure the performance of the query in all the platforms and then we compare them with each other.
In order to produce the Dataset with false data, we used Python so we could generate as many data as we want (100 GB in our case). The reason why we decided to produce a dataset with false data is because we wanted to test the non-linear information retrieval.
Based on the queries that we will perform, not all the data will be analyzed. This will depend on the tables from where we want to get the data from.
The Airline_Ontime_Data has in total two tables from where we will get the data from and the Dataset created by us has 100 tables and each of them is 1GB. We will see in the result section that even the tables have the same amount of data, when we want to get all the data from one of the tables, we will have different performance in time.
The Query performed for the first Dataset is:

```
Select data,arrival_time,arrival_delay
From Airline_Ontime_Data.Flights
Where airline_code="19977"
```

The query for the second database is:

```
Select *
From test_datase.1GB1 (test_datase.1GB2……)
```

4. Experimental Results

The results for the first query were very impressive. The query on Google platform (Big Query) was executed in 24.3s. The same query on Microsoft Platform (Azure) was executed in 24.7s and in Amazon (AWS) it was executed in 24.6s. The data returned was in the same amount but the time was different. We can see from the results that we got the best performance in Google Big Query. Please see the graph below in order to understand better the results:

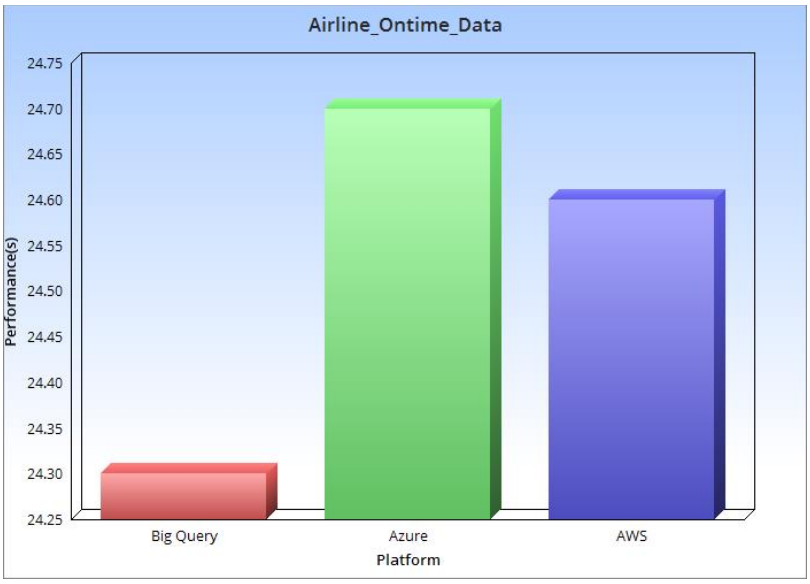

Figure 1 First Test Results

As for the second dataset, we executed the query for 1G, 15GB and 100GB in each platform. The results that we got were as following.

1GB Table

The query on Google platform (Big Query) was executed in 5.2s. The same query on Microsoft Platform (Azure) was executed in 5.8 s and in Amazon (AWS) it was executed in 5.5s. The data returned was in the same amount but the time was different. We can see from the results that we got the best performance in Google Big Query. Please see the graph below in order to understand better the results:

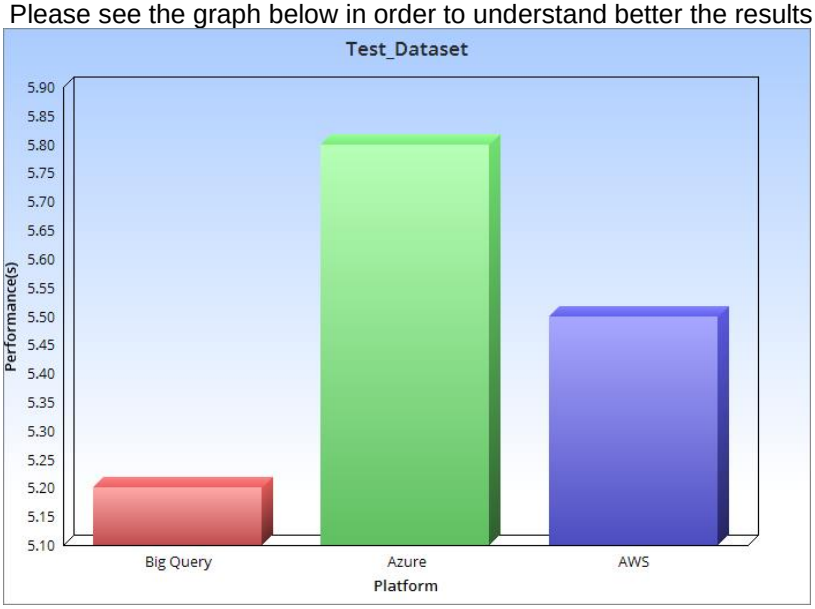

Figure 2 Second Test Results

15GB Data (Join 15 tables)

The query on Google platform (Big Query) was executed in 31.1s. The same query on Microsoft Platform (Azure) was executed in 31.4s and in Amazon (AWS) it was executed in 31.6s. The data returned was in

the same amount but the time was different. We can see from the results that we got the best performance in Google Big Query. What is different from other test cases here is that we got a better performance in Azure than in AWS. Please see the graph below in order to understand better the results:

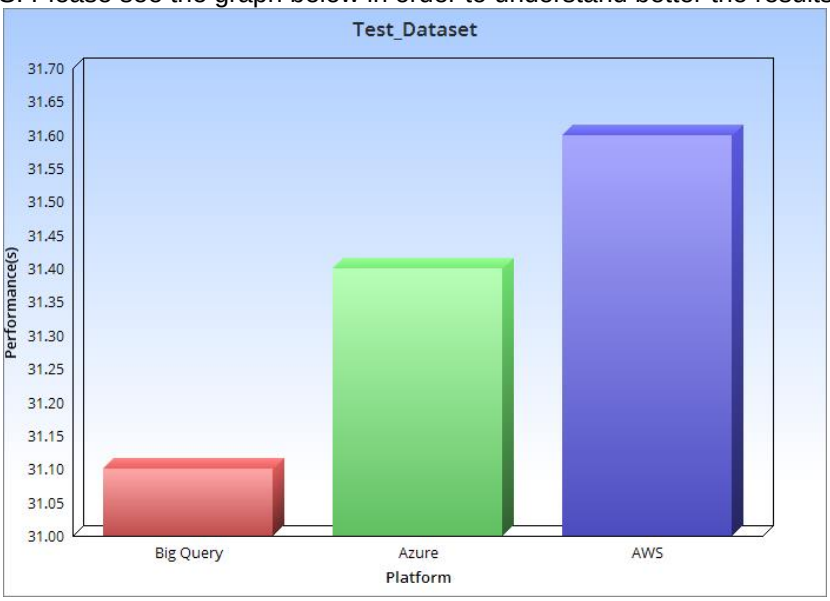

Figure 3 Third Test Results

100GB Data

The query on Google platform (Big Query) was executed in 27.8.1s. The same query on Microsoft Platform (Azure) was executed in 32.5s and in Amazon (AWS) it was executed in 31.9s. The data returned was in the same amount but the time was different. We can see from the results that we got the best performance in Google Big Query. Here we can see that even the data that needs to be returned is greater than in previous test, the time that is needed is less than in the others. The reason for this, as mentioned above, is that the information retrieval in this case in non-linear. Please see the graph below in order to understand better the results:

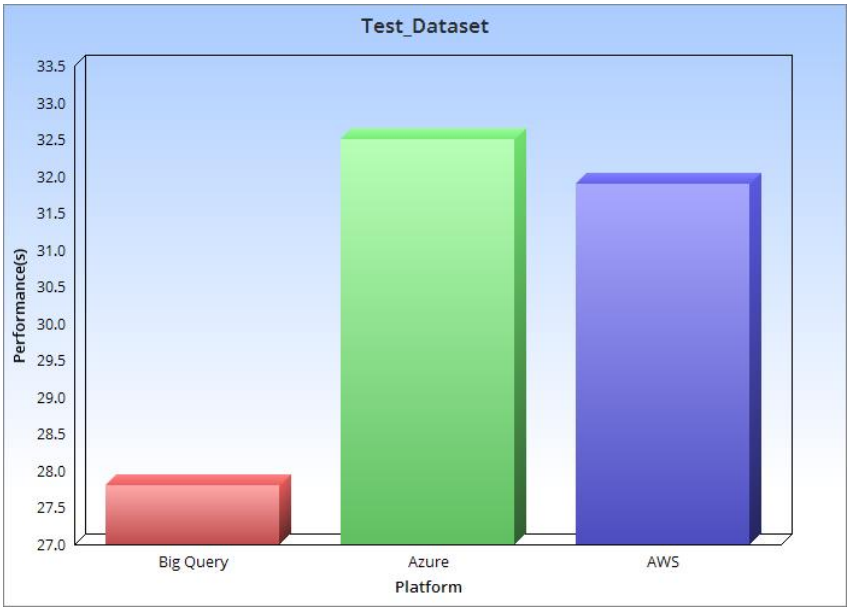

Figure 4 Fourth Test Results

5. Conclusions

Based on the experiments that we did and also on the research in the area of Big Data Analytics on cloud, we have to say that we reached in the results that before applying any analysis on the data, we must undergo some important steps in order to make the data ready for analysis. The first important step is to structure the date is the data is not structured and then apply all techniques mentioned above step by step.

While testing the Big Data Analysis on all platforms, we found out that the best result in all experiments was achieved in Google Big Query platform. This because this platform is specific to Big Data Analytics because it uses Google's Dremel for Interactive Analysis, which is the type of the analyze we did. In other platforms also we got good results and we have to mention that based on the information returned all platforms were 100% accurate.

Even the amount of data in larger, we can achieve a smaller time while we execute the same query. The reason for this, as we mentioned if because the information retrieval is not always linear. It can be non-linear as well. We can see this in the test that we did with 100GB data, where we achieved smaller time than 15GB data.

References

[1] Samiya Khan1, Kashish Ara Shakil and Mansaf Alam. Cloud-based Big Data Analytics – A Survey of Current Research and Future Directions. Conference Paper (December 2015) Retrieved from: https://www.researchgate.net/publication/281144737

[2] Manekar, A. and Pradeepini, G. (2015). A Review on Cloud-based Big Data Analytics. ICSES Journal on Computer Networks and Communication (IJCNC), May 2015, Vol. 1, No, 1. Retrieved from: http://www.i-cses.com/ijcnc/Archive/V1N1/IJCNC-V1N1-P0001.pdf

[3] Assuncao, M. D., Calheiros, R. N., Bianchi, S. and Netto, M. A. S. (2015). Big Data Computing and Clouds: Trends and Future Directions. J. Parallel Distrib. Computing, 79-80 (2015) 3-15. Retrieved from: http://www.buyya.com/papers/BDC-Trends-JPDC.pdf

[4] Wu, X., Zhu, X., Wu, G. and Ding, W. (2013). Data Mining with Big Data. Retrieved from: http://www.cs.umb.edu/~ding/papers/TKDE2013.pdf

[5] Mell Peter, Grance Timothy (2009). The NIST definition of cloud computing. Retrieved February 25 2012 from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[6] Fig.1 http://www.nieuws.social/strategie_nieuws/avg-data-en-de-ontwikkeling-van-een-nieuw-soort-kapitaal/