The Introduction of Big Data in Cloud Computing

Austin Gruenberg, Computer Science

Minnesota State University Moorhead

Abstract:

One of the fastest-growing technologies that many people are unaware of is the world of cloud computing. Having started in 2006, it is a relatively new technological advancement in the computer industry. The major branch of cloud computing that I decided to focus on was big data. I decided to research this topic to better understand what its current uses are, to see what the future holds for Big Data and cloud computing and because it is a growing, significant piece of technology being used in our society today. Big data and cloud computing are very important industries and have a bright future in the tech industry.

Cloud Computing:

Cloud computing is an on-demand delivery service that allows consumers to have access to IT resources over the internet. The resources that are available worldwide are physically present within large, major data centers set up by companies that are most commonly built by the major corporations, such as Google and Amazon, that are providing these cloud services to consumers around the world. The main difference between cloud systems and other systems is that the cloud is not tied to any dedicated servers. Cloud computing is thought to have begun in mid-2006, but it is recorded to have been brought up and discussed as early as the 1960s.

There are four main models of cloud computing being used at this time. These include infrastructure as a service, platform as a service, software as a service, and workplace as a service. Infrastructure as a service is the ability for the consumer to control processing and storage, for example, being able to deploy and run operating systems and virtual storage. Platform as a service is where consumers can use the cloud infrastructure to access new or existing applications. Workplace as the software is when companies/businesses install and organize applications that allow their users to access all the necessary software. This has become prevalent with the introduction of Microsoft's Office 365 being utilized within businesses and even schools. The most used software as a service is Office 365, which provides all of Mircrosoft's application in one area to be accessed. Lastly, data as a service gives users the ability to store information with disk space within the cloud. A major cloud application that was utilized in my high school was Google Drive, which is a cloud-based storage solution that can be easily accessed, for free, by anyone with a Google account.

Cloud Computing is not a perfect service, however. If you do not have access to a stable internet connection, you will not have success using a cloud application. Also, programs ran over the cloud will normally run slower than when they were running on a local computer (Arutyunov 2012). Even though cloud computing is not a fault proof, the possibilities that it provides the world with are worth the drawbacks. Its implications for consumers do not compare to the possibilities that it has in the big data industry.

Big Data:

The continuous increase in volume and detail of the data that is being gathered by organizations is becoming overwhelming to deal with using traditional data analytical methods. This is increase in data volume has created the term “big data” in recent years and has continued to grow since its introduction. Big data is commonly categorized by the 4V’s, namely, volume, variety, velocity, and value. Volume refers to the amount of all types of data that is generated by multiple sources and continue to expand. Variety refers to the different types of data that is collected and where it is received from. Velocity refers to the speed of the data being transferred. Lastly, value refers to the process of discovering hidden values from massive datasets with various types ands rapid generation. Big data is primarily classified into different categories to help better understand their characteristics. These main categories include data sources, content format, data stores, data staging, and data processing. Each of these categories have their own defining characteristics that make them stand apart from one another (Hashem 2015). A diagram showcasing the different categories of big data can be seen in Figure 1.

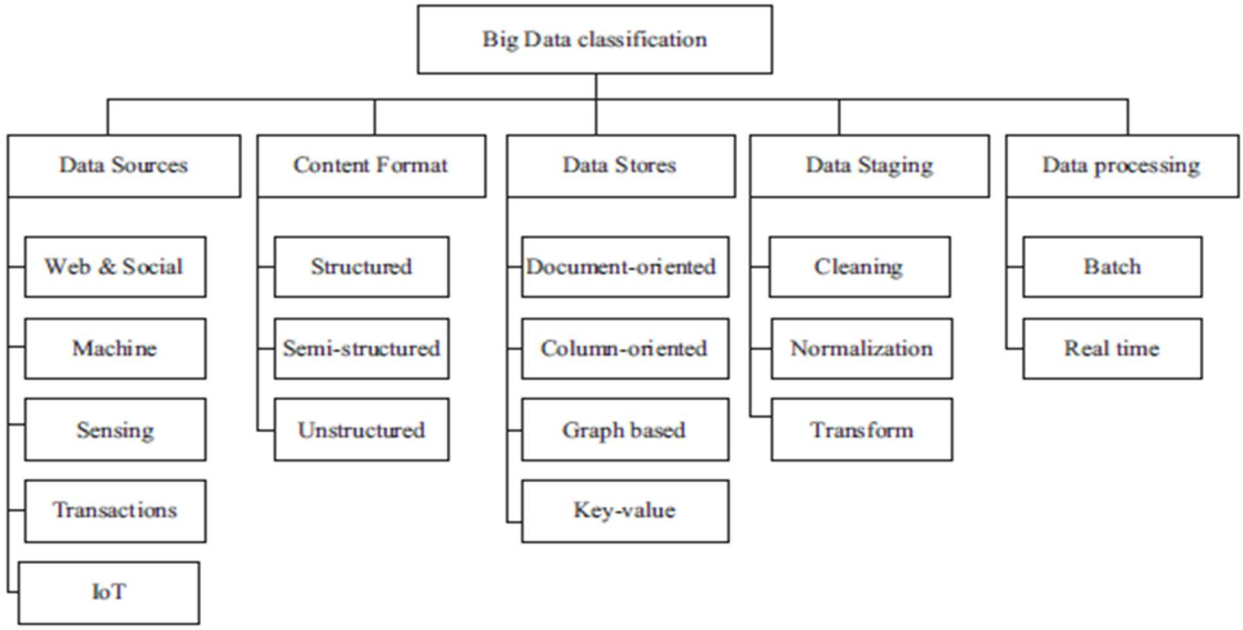

Fig. 1. Big data classification

How Big Data and Cloud Computing Work Together:

Both cloud computing and big data are their industries that have their purpose for being utilized today. However, they are extremely powerful and essential for companies to pair both technologies together. Cloud computing is so revolutionary to the big data industry simply because of its ability to significantly reduce costs and make the information available to anyone, anywhere. This reduction of costs comes from the limitless storage capacity of the cloud, as well as the scalable processing power that cloud computing provides.

The most common application that takes advantage of the cloud for use with big data is Hadoop. A list of common Hadoop commands can be seen below in Figure 2. Hadoop has 4 different libraries being Common, MapReduce, Distributed File System (HDFS), and Yarn. Common is mainly utilized for regular workloads and can provide common utilities across all platforms. MapReduce is the part of Hadoop that processes the data that is being stored within its HDFS. Yarn is just the updated version of MapReduce and contains all the same utilizations of its counterpart.

Pairing Hadoop's ability to store and analyze the big data, using the cloud computing storage and processing power to handle it, with some form of artificial intelligence, AI, to analyze the data being retrieved and make appropriate analyses and confident decisions to better improve the companies' data being collected. This is the strength that these industries provide when working together and are currently being implemented around the world rapidly.

1. -put uploads files from the local file system to HDFS

(hadoop fs -put Sample2.txt /user/agruenberg/newDirectory)

2. -cat allows the user to open a file and see the contents.

(hadoop fs -cat /user/agruenberg/newDirectory/Sample1.txt)

3. -cp will copy all files form one HDFS location to another HDFS location.

(hadoop fs -cp /location1 /location2)

4. -mv moves the files form one HDFS location to another HDFS location

(hadoop fs -mv /location1 /location2)

5. -ls is used to list all available files and subdirectors in the default directory.

(hadoop fs – ls)

6. -mkdir creates a new directory under the specified location with the designated location.

(hadoop fs -mkdir /user/agruenberg/newDirectory)

7. -copyFromLocal allows you to copy a local filesystem to an HDFS location.

(hadoop fs -copyFromLocal Sample1.txt /user/agruenberg)

Fig. 2. Example Hadoop Commands

Real World Implication:

One of the ways that cloud computing and big data are working together to provide a useful solution is with the creation of a wide area monitoring of power grids. A smart grid requires applications such as data mining, intelligent diagnosis, and higher storage requirements. Cloud computing allows for these smart grids to utilize more storage resources through virtualization and parallel processing to give the users a new form of the services they need. At the time of this article, this cloud-based architecture was only being proposed and not in actual

use, but I felt that it shows a lot of good information about how these industries can work together soon. Hadoop is the most commonly used cloud compatible, open-source system and is the "backbone" of this implementation. The cloud platform provides more reliable data storage and management for the big data mining information by utilizing the HDFS software that Hadoop provides. Through analysis of the information received, intelligent forecasting and decision making for dispatching operators can be done thanks to the addition of this new software (Liu 2019).

Effects of the Pandemic:

Through the last year, many businesses suffered. However, this was not the case for everyone. Major IT giants, such as Microsoft, Amazon, and Google, had their businesses improve through the pandemic because of big data. The lockdowns and restrictions that were put in place around the world due to the pandemic sped up the demand for cloud technologies. Amazons web service and Azure cloud computing were the biggest applications to grow and be utilized in the last year, which led to a company growth of 37% and 23% respectively to allow work to be more efficient with everyone at their homes. The coverage of 5G networks is continuing to grow, and with that, many cloud service providers are partnering with mobile operators to allow their services to even more readily available to everyone (Polites 2020). This continued development of 5G is only one reason why the future of this growing industry is so bright.

A Day in the Life:

I was given the privilege to meet with a current big data analyst that works for the Game Show Network data team. As someone trying to learn more about this industry, I found it very interesting and beneficial to learn about what it's like to work within the field today. I asked them just a few questions via email and received some excellent answers about what they do for GSN, what an average day for them is like, what important practices/applications are in the industry, and simply what they enjoy the most about working in the big data industry.

Just like any data analyst, the main purpose for them at GSN is to collect, store, and analyze data created by their studios to produce the greatest amount of money from its users. At GSN, there are five different video game studios owned under that name, so the data team is required to work with each studio's data and craft/modify features and applications to yield better engagement, more engagement, and higher profits. The data team at GSN uses the data warehouse Vertica and the S3 data lake to make the data quarriable. Working with these applications and working on improving internal infrastructure to help other analysts is the main premise of the job but is not all that they do on the data team.

The day-to-day experience varies for everyone at any company and the same holds true at GSN. Another goal of the data team is to save the company money by producing more cost-efficient methods of working with the data. For this reason, the most recent project being developed was creating a new

data pipeline from an old, more costly version. This pipeline is an Amazon Kinesis stream and a Java service that they built to convert the data to forms that can be queried within Vertica. Another common daily task is working on adding new services or analysts. This can lead to lots of time being spent explaining the software and helping analysts learn the new products that you've been developing. CircleCI is the software they use to build and deploy new services, but all the code in the infrastructure is written themselves in CloudFormation. The next project for the GSN data team is replacing CircleCI with in-house software that is based around a Kubernetes solution.

The biggest advice that I could get from meeting with the employee was to always be prepared to learn while working in any computer-related field. Whether you're on the clock or not, you should always be looking to use new applications to improve your skills and better your overall knowledge. Since the cloud and big data are constantly evolving, these transitions and constant learning curves can become frustrating but are generally worth the effort and you begin learning patterns of the new tech.

When I was preparing questions to ask the data analyst, I was most interested in hearing about why they enjoy their job in working within the big data industry. Their response explained that the satisfying feeling after completing a project from start to finish and see it giving others success is the main reason they have fallen in love with this industry. They also consider themselves as lifelong learners and love to be allowed to learn new things every day they go to work.

Future of the Industry:

Cloud computing is still considered to be at the beginning of its development. As it is now, cloud computing is just a trend with many incomplete models and technologies. In order for its future to be successful, it must fix its issues at hand and begin to define a new direction for development in the future. Also, cloud computing is expected to transition from being just a resource being purchased and utilized by major company's to becoming a common service for all users to be able to purchase and decide what they see is the best use for it (Arutyunov 2012).

One of the groups that the industry hopes to start taking advantage of cloud computing is developing countries and small businesses. Both struggle to keep up with their competition due to the high cost it takes to maintain current technology services. However, cloud computing allows them both to exploit high-end applications that were unavailable to them because of the absurd costs. This is one of the major benefits that cloud computing can offer and is why it continues to grow in popularity (Marston 2011).

As time continues, the amount of data created by the billions of people on the internet and interacting with businesses will exponentially grow over time. This growth further shows the need for cloud storage to deal with big data for the distant future. The need for data analysts will also be in high demand with the increase in data and is a good choice for new students to study and get involved with early.

Conclusion:

By conducting my research on big data and cloud computing, I was able to learn a lot about what both industries do and what their affect will have on the future. In my project, I discussed how these technologies can work together in multiple ways to improve a variety of different businesses. Eventually, all businesses will be able to automatically improve themselves by implementing these services with AI to analyze their data and produce productive conclusions. Hadoop is one of the main applications within the big data industry that easily allows the use of cloud services to store and manipulate companies' big data and has worked hard to make these services more mainstream. The example of the power grid system being monitored, and automatically dispatching operators based on the large amount of video being analyzed at once, is a perfect example of what the possibilities of these industries entail. Not only are the implications of these industries impactful, but they also were able to keep businesses, and people, virtually together within the biggest health crisis in the last century. Having the opportunity to meet with a current employee working within these fields was a great experience and provided solid information as to what it is actually like to work in these fields. This last year alone proved that these industries are not going anywhere and have very bright futures in the world of technology.

Works Cited

1. Arutyunov, V. “Cloud Computing: Its History of Development, Modern State, and Future Considerations.” Scientific and technical information processing 39, no. 3 (July 2012): 173–178.
2. Fuguang, Yao. “Research on Campus Network Cloud Storage Open Platform Based on Cloud Computing and Big Data Technology.” Journal of intelligent & fuzzy systems 38, no. 2 (February 6, 2020): 1215–1223.
3. Gupta, Agrawal. “Soft Computing Techniques for Big Data and Cloud Computing.” Soft computing (Berlin, Germany) 24, no. 8 (April 2020): 5483–5484.
4. Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. “The Rise of ‘big Data’ on Cloud Computing: Review and Open Research Issues.” Information systems (Oxford) 47 (2015): 98–115.
5. "Leading the future for the connected world." *Express Computers*, November 12, 2020, NA. *Gale OneFile: Computer Science* (accessed February 1, 2021). https://link.gale.com/apps/doc/A641267509/CDB?u=mnalll&sid=CDB&xid=6b7c9d58.
6. Liu, Liang. “A Cloud-Computing and Big Data Based Wide Area Monitoring of Power Grids Strategy.” IOP conference series. Materials Science and Engineering 677, no. 4 (December 1, 2019).
7. Marston, Sean, Zhi Li, Subhajyoti Bandyopadhyay, Juheng Zhang, and Anand Ghalsasi. “Cloud Computing — The Business Perspective.” Decision Support Systems 51, no. 1 (2011): 176–189.
8. Polites, Jared. Benzinga: Data And Cloud Technologies Helped IT Giants Stay Afloat This Year. Newstex Finance & Accounting Blogs. Chatham: Newstex, 2020.
9. Sacolick, Isaac. "Are you ready for multicloud? A checklist." *InfoWorld.com*, December 14, 2020, NA. *Gale OneFile: Computer Science* (accessed February 1, 2021). https://link.gale.com/apps/doc/A645012792/CDB?u=mnalll&sid=CDB&xid=0c91f625.
10. VOUK, Miaden A. “Cloud Computing : Issues, Research and Implementations.” CIT. Journal of computing and information technology 16, no. 4 (2008): 235–246.
11. Xiaona, Ma. “Informatization Strategies of Education and Teaching Management in the Era of Cloud Computing and Big Data.” Journal of physics. Conference series 1738, no. 1 (January 1, 2021).
12. Xu, Xun. “From Cloud Computing to Cloud Manufacturing.” Robotics and computer-integrated manufacturing 28, no. 1 (2012): 75–86.