International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 08 | August - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Cloud Computing: Innovation Opportunities and Challenges.

DILPREET KAUR , Msc IT

Abstract

Big Data has emerged in the past few years as a new paradigm providing abundant data and opportunities to improve and/or enable research and decision-support applications with unprecedented value for digital earth applications including business, sciences and engineering. At the same time, Big Data presents challenges for digital earth to store, transport, process, mine and serve the data. Cloud computing provides fundamental support to address the challenges with shared computing resources including computing, storage, networking and analytical software; the application of these resources has fostered impressive Big Data advancements. This paper surveys the two frontiers – Big Data and cloud computing – and reviews the advantages and consequences of utilizing cloud computing to tackling Big Data in the digital earth and relevant science domains.

Introduction

Big data and cloud computing are two distinctly different ideas, but the two concepts have become so interwoven that they are almost inseparable. It's important to define the two ideas and see how they relate. Big data refers to vast amounts of data that can be structured, semi structured or unstructured. It is all about analytics and is usually derived from different sources, such as user input, IoT sensors and sales data. Big data also refers to the act of processing enormous volumes of data to address some query, as well as identify a trend or pattern. Data is analyzed through a set of mathematical algorithms, which vary depending on what the data means, how many sources are involved and the business's intent behind the analysis. Distributed computing software platforms, such as Apache Hadoop, Data bricks and Cloud era, are used to split up and organize such complex analytics.

The evolution of Big Data, especially its adoption by industry and government, expands the content/meaning of Big Data. The original volume-based definition now encompasses the data itself, relevant technologies and expertise to help generate, collect, store, manage, process, analyze, present and utilize data, as well as the information and knowledge derived. For example, the Big Earth Data Initiative designates Big Data as an investment opportunity and a ‘calling card’ for advancing earth science and digital earth using Big Data and

relevant processing technologies. For the geospatial domain, Big Data has evolved along a path from purely data to a broader concept including data, technology and workforce. The focus is the geographic aspects of Big Data from Social, Earth Observation (EO), Sensor Observation Service (SOS), Cyber Infrastructure (CI), social media and business. For example, EO generates terabytes (TB) of images daily; climate simulations by the IPCC (Intergovernmental Panel on Climate Change) produce hundreds of peta-bytes (PB) for future climate analyses; and SOS produces even more from sensor web and citizen as sensors. Social and business data are generated at a faster pace with specific geographic and temporal footprints.

The Sources of Big Data

The bulk of big data generated comes from three primary sources: social data, machine data and transactional data. In addition, companies need to make the distinction between data which is generated internally, that is to say it resides behind a company's firewall, and externally data generated which needs to be imported into a system.

Whether data is unstructured or structured is also an important factor. Unstructured data does not have a pre-defined data model and therefore requires more resources to make sense of it.

- **Earth sciences**

The advancement of sensing and computing simulation technologies enabled collection and generation of massive data sets every second at different spatiotemporal scales for monitoring, understanding and presenting complex earth systems. For example, EO collects TB of images daily with increasing spatial, temporal and resolutions. Geospatial models also generate large spatiotemporal data *via* numerical simulations of the complex earth systems. Climate science is an exemplar representing the Big Data shift across all digital earth domains in using big spatiotemporal data to monitor and describe the complex earth climate system. For example, the IPCC AR5 alone produced 10,000 TB of climate data, and the next IPCC will engage hundreds of PB. It is critical to efficiently analyze these data for detecting global temperature anomalies, identifying geographical regions with similar or disparate climate patterns, and investigating spatiotemporal distribution of extreme weather events. However, efficiently mining information from PB of climate data is still challenging.

- **Internet of Things**

Advanced sensors and their hosting devices (e.g. mobile phones, health monitors) are connected in a cyber-physical system to measure time and location of humans, movement of automobiles, vibration of machine, temperature, precipitation, humidity and chemical changes in the atmosphere. The Internet of Things captures this new domain and continuously generates data streams across the globe with geographical footprints from interconnected mobile devices, personal computers, sensors, RFID tags and cameras. Big Data generated from the various sensors of IoT contains rich spatiotemporal information. The advance of IoT and Big Data technologies presents an array of applications including better product-line management, more effective and timely criminal investigation, boosting agriculture productivity and accelerating the development of smart cities with new architecture.

- **Social Science**

Social networks, such as Twitter and Facebook, generate Big Data and are transforming social sciences. As of the time of writing, Twitter users around the world are producing around 6000 tweets per second which corresponds to 500 million tweets per day and around 200 billion tweets per year. Economists, political scientists, sociologists and other social scholars use Big Data mining methods to analyze social interactions, health records, phone logs, government records and other digital traces While such mining methods benefit governments and social studies, it is still challenging to quickly extract spatiotemporal patterns from big social data to, for example, help predict criminal activity, observe emerging public health threats and provide more effective intervention.

- **Astronomy**

Astronomy is producing a spatiotemporal map of the universe by observing the sky using advanced sky survey technologies. Mapping and surveying the universe generates vast amounts of spatiotemporal data. For example, Sloan Digital Sky Survey (SDSS) generated 116 TB data. New observational instruments scheduled for operation in 2023 will generate 15 TB data nightly and deliver 200 PB data in total to address the structure and evolution of the universe Besides observational data, the Large Hadron Collider is investigating how the universe originated and operates at the atomic leveland produces 60 TB of experimental data per day and 15 PB data per annum. Subsequent to collection, the foremost challenge in this new astronomical arena is managing the Big Data, making sense of the information efficiently, and finding interesting celestial objects

and processing in an effective manner. The astronomy big data not only records information on how universe evolves, but also can be used to understanding how Earth evolves and protecting Earth from outer space impact, such as planetary defence.

- **Business**

Business intelligence and analytics are enhanced with Big Data for decisions on strategy, managing optimization and competition. Business actions generate large volume, high velocity and highly unstructured data sets. These data contain rich geospatial information, such as where and when a transition occurred. To manage and process these data, the full spectrum of data processing technologies has been developed for the distributed and scalable storage environment. However, it remains a challenge to efficiently construct spatiotemporal statistical models from business data to optimize product placement, analyze customer transaction and market structure, develop personalized product recommendation systems, manage risks and support timely business decisions.

- **Industry**

In the fourth industrial revolution, products and production systems leverage IoT and Big Data to build ad-hoc networks for self-control and self-optimization. Big Data poses a host of challenges to Industry 4.0, including the following:

(i) Seamless integration of energy and production.
(ii) Centralization of data correlations from all production levels.
(iii) Optimization of performance of scheduling algorithms.
(iv) Storage of Big Data in a semi-structured data model to enable real-time queries and random access without time-consuming operations and data joins.
(v) Realization of on-the-fly analysis to help organizations react quickly to unanticipated events and detect hidden patterns that compromise production efficiency.

Cloud computing could be leveraged to tackle these challenges in Industry 4.0 for networking, data integration, data analytics and intelligence for Cyber-Physical Systems and resiliency and self-adaptation .In addition to the reviewed six sources, Big Data challenges may also come from other relevant domains such as medical research, public health, smart cities, security management, emergency response and disaster recovery.

Unlocking Real Value From Data

Real business value comes from an ability to combine this data in ways to generate insights, decisions and actions. CloudMoyo helps companies develop a comprehensive, cohesive and sustainable analytics strategy, which gives them the tools to differentiate themselves via actionable insights and supports employees and the business itself. A number of factors point to the value of the niche that companies like CloudMoyo are fulfilling. A recent study found that two-thirds of companies with the most advanced technology in this area cannot hire enough people to run these capabilities**.** Added to that, analytics is resource-intensive.

Large companies struggle to allocate enough resources, but for smaller companies, it's inconceivable that they can dedicate all that is needed for effective analysis. In both these cases, outsourcing is an invaluable advantage to have.

While there is a generally acknowledged understanding that big data can provide a competitive advantage, those who are partnering with sophisticated third-party providers stand a much better chance of benefitting from high-quality, affordable insights. The era of big data is well and truly upon us, and it's no longer a question of whether enterprises should engage with big data, but how. Technology giant Cisco predicts that the amount of data produced in 2020 will be 50 times what it is today. No wonder then that companies feel overwhelmed and desperately in need of solid advice from specialists who understand their business and can combine it with technology to deliver results.

The pros of big data in the cloud

The cloud brings a variety of important benefits to businesses of all sizes. Some of the most immediate and substantial benefits of big data in the cloud include the following.

Scalability

A typical business data center faces limits in physical space, power, cooling and the budget to purchase and deploy the sheer volume of hardware it needs to build a big data infrastructure. By comparison, a public cloud manages hundreds of thousands of servers spread across a fleet of global data centers. The infrastructure and software services are already there, and users can assemble the infrastructure for a big data project of almost any size.

Agility

Not all big data projects are the same. One project may need 100 servers, and another project might demand 2,000 servers. With cloud, users can employ as many resources as needed to accomplish a task and then release those resources when the task is complete.

Cost

A business data center is an enormous capital expense. Beyond hardware, businesses must also pay for facilities, power, ongoing maintenance and more. The cloud works all those costs into a flexible rental model where resources and services are available on demand and follow a pay-per-use model.

Accessibility

Many clouds provide a global footprint, which enables resources and services to deploy in most major global regions. This enables data and processing activity to take place proximally to the region where the big data task is located. For example, if a bulk of data is stored in a certain region of a cloud provider, it's relatively simple to implement the resources and services for a big data project in that specific cloud region -- rather than sustaining the cost of moving that data to another region.

Resilience

Data is the real value of big data projects, and the benefit of cloud resilience is in data storage reliability. Clouds replicate data as a matter of standard practice to maintain high availability in storage resources, and even more durable storage options are available in the cloud.

The cons of big data in the cloud

Public clouds and many third-party big data services have proven their value in big data use cases. Despite the benefits, businesses must also consider some of the potential pitfalls. Some major disadvantages of big data in the cloud can include the following.

Network dependence

Cloud use depends on complete network connectivity from the LAN, across the internet, to the cloud provider's network. Outages along that network path can result in increased latency at best or complete cloud

inaccessibility at worst. While an outage might not impact a big data project in the same ways that it would affect a mission-critical workload, the effect of outages should still be considered in any big data use of the cloud.

Storage costs

Data storage in the cloud can present a substantial long-term cost for big data projects. The three principal issues are data storage, data migration and data retention. It takes time to load large amounts of data into the cloud, and then those storage instances incur a monthly fee. If the data is moved again, there may be additional fees. Also, big data sets are often time-sensitive, meaning that some data may have no value to a big data analysis even hours into the future. Retaining unnecessary data costs money, so businesses must employ comprehensive data retention and deletion policies to manage cloud storage costs around big data.

Security

The data involved in big data projects can involve proprietary or personally identifiable data that is subject to data protection and other industry- or government-driven regulations. Cloud users must take the steps needed to maintain security in cloud storage and computing through adequate authentication and authorization, encryption for data at rest and in flight, and copious logging of how they access and use data.

Challenges to Big Data in the Cloud environment:

Just as Big Data has provided organizations with terabytes of data, it has also presented an issue of managing this data under a traditional framework. How to analyze the large sum of data to take out only the most useful bits? Analyzing these large volumes of data often becomes a difficult task as well. In the high speed connectivity era, moving large sets of data and providing the details needed to access it, is also a problem. These large sets of data often carry sensitive information like credit/debit card numbers, addresses and other details, raising data security concerns.

- **Datastorage**

  Storage challenges are posed by the volume, velocity and variety of Big Data. Storing Big Data on traditional physical storage is problematic as hard disk drives (HDDs) often fail, and traditional data protection mechanisms are not efficient with PB-scale storage. In addition, the velocity of Big Data requires the storage systems to be able to scale up quickly which is difficult to achieve with traditional storage systems. Cloud storage services offer virtually unlimited storage with high fault tolerance which provides potential solutions to address Big Data storage challenges. However, transferring to and hosting Big Data on the cloud is expensive given the size of data volume. Principles and algorithms, considering the spatiotemporal patterns of data usage, need to be developed to determine the data's analytical value and its preservation datasets by balancing the cost of storage and data transmission with the fast accumulation of Big Data.

- **Data Management**

  It is difficult for computers to efficiently manage, analyze and visualize big, unstructured and heterogeneous data. The variety and veracity of Big Data are redefining the data management paradigm, demanding new technologies to clean, store, and organize unstructured data. While metadata are essential for the integrity of data provenances, the challenge remains to automatically generate metadata to describe Big Data and relevant processes. Generating metadata for geospatial data is even challenging due to the data's intrinsic characteristics of high-dimensionality and complexity. Besides metadata generation, Big Data also poses challenges to database management systems (DBMSs) because traditional RBDMSs lack scalability for managing and storing unstructured Big Data. While non-relational databases such as MongoDB and HBase are designed for Big data, how to tailor these NoSQL databases to handle geospatial Big Data by developing efficient spatiotemporal indexing and querying algorithms is still a challenging issue.

- **Data Processing**

  Processing large volumes of data requires dedicated computing resources and this is partially handled by the increasing speed of CPU, network and storage. However the computing resources required for processing Big Data far exceed the processing power offered by traditional commuting paradigms. Cloud

computing offers virtually unlimited and on-demand processing power as a partial solution. However, shifting to the cloud ushers in a number of new issues. First is the limitation of cloud computing's network bandwidth which impacts the computation efficiency over large data volumes. Second is data locality for Big Data processing. While 'moving computation to data' is a design principle followed by many Big Data processing platforms, such as Hadoop, the virtualization and pooled nature of cloud computing makes it a challenging task to track and ensure data locality, and to support data processing involving intensive data exchange and communication.

- **Data Analysis**

  Data analysis is an important phase in the value chain of Big Data for information extraction and predictions. However, analyzing Big Data challenges the complexity and scalability of the underlying algorithms. Big Data analysis requires sophisticated scalable and interoperable algorithms and is addressed by welding analysis programs to parallel processing platforms to harness the power of distributed processing. However, this 'divide and conquer' strategy does not work with deep and multi-scale iterations that are required for most geospatial data analysis/mining algorithms. Furthermore, most existing analytical algorithms require structured homogeneous data and have difficulties in processing the heterogeneity of Big Data. This gap requires either new algorithms that cope with heterogeneous data or new tools for pre-processing data to make them structured to fit existing algorithms. In geospatial domain, optimizing existing spatial analysis algorithms by integrating spatiotemporal principles to accelerate geospatial knowledge discovery is challenging and has become a high priority research field of 'spatiotemporal thinking, computing and applications.

Current status of tackling Big Data challenges with cloud computing

While the Big Data challenges can be tackled by many advanced technologies, such as HPC, cloud computing is the most elusive and important. This section reviews the status of using cloud computing to address the Big Data challenges.

Scalability

Scalability on distributed and virtualized processors is and has been a bottleneck for leveraging cloud computing to process Big Data.

(i) Co-existing VMs decrease the disk throughput.

(ii) Performance on physical clusters is significantly better than that on virtual clusters.

(iii) Performance degradation due to separation of the services depends on the data-to-compute ratio.

(iv) Application completion progress correlates with the power consumption, and power consumption is application specific.

Various cloud performance benchmarks and evaluations demonstrated that balancing the number and size of VMs as a function of the specific applications is critical to achieve optimal scalability for geospatial Big Data applications. Accordingly, different strategies improve scalability and achieve cost-effectiveness while handle Big Data processing tasks with scalable and elastic service to disseminate data.

Quality of Service

Quality of Service (QoS) describes the overall performance and is particularly important for Big Data applications and cloud computing in scheduling applications on the distributed cloud. If data services and cloud data centers are geographically distributed, it is essential to monitor the QoS globally for Big Data implementation and cloud computing. However, more efforts should be devoted to handling multiple QoS requirements from different users in the process of resource and task scheduling within a single or multiple cloud environments.

Scheduling

Job scheduling effectively allocates computing resources to a set of different tasks. However, scheduling is a challenge in automatic and dynamic resource provisioning for Big Data proposed several research directions for cloud resource scheduling, including real-time, adaptive dynamic, large-scale, multi-objective, and distributed and parallel scheduling. As one of the most popular frameworks for Big Data processing, Hadoop Map Reduce is optimized (e.g. task partitioning, execution) to accommodate better Big Data processing Progress on scheduling was made in geospatial fields as well. However, research effort is need on developing more sophisticated scheduling algorithms that can leverage the spatial relationships of data, computing resources, application and users to optimize the task execution process and the resource utilization of the underlying computing infrastructure.

Innovation support

These research initiatives respond to the 10 aspects envisioned to produce the next generation of valuable technology-enabled businesses as identified by the McKinsey report. For example, to support distributed co-creation across the computer network, the McKinsey report proposed advancements in distributed storage, interdisciplinary collaboration, workflow sharing and mobile computing as well as to collocate spatiotemporally resources and creators. Experiments are increasingly dependent on simulations facilitated by Big Data and cloud computing to address challenges in engineering design of complex systems, including health care, logistics and manufacturing in the digital earth context. This requires a well-trained workforce, collaboration cross-domains, data mining analytics and spatiotemporal collocation of various data, processing and domain resources. Investing in resources to promote the public good requires support from all research directions; spatiotemporal collocation is a cardinal component to achieve all 10 innovations with methodologies, tools and solutions. For example, multi-scale collaborations require multi-spatiotemporal levels of collaboration across different domains supported by distributed storage.