THE ROLE OF BIG DATA IN CLOUD USING IOT APPLICATIONS

KBKS.Durga [1], K.Nirosha[2], B. Durga Sri[3]
[1]Assistant Professor, St. Martin's Engineering College
[2,3]Assistant Professor, MLR Institute of Technology

Abstract

The term massive knowledge arose beneath the explosive increase of world knowledge as a technology that's ready to store and process massive and varied volumes of knowledge, providing each enterprises and science with deep insights over its clients/experiments. Cloud computing provides a reliable, fault-tolerant, on the market and scalable surroundings to harbor massive knowledge distributed management systems. Among the context of this paper we tend to gift an outline of each technologies and cases of success once integration massive knowledge and cloud frameworks. Though massive knowledge solves a lot of our current issues it still presents some gaps and problems that raise concern and want improvement. Security, privacy, measurability, knowledge governance policies, knowledge non uniformity, disaster recovery mechanisms, and alternative challenges are nevertheless to be self-addressed. Alternative considerations are associated with cloud computing and its ability to affect Exabyte's of data or address exaflop computing with efficiency. This explains an overview of each cloud and massive knowledge technologies describing this problem with these technologies

Keywords: Cloud computing, Big Data, Reliable,, IOT, Knowledge;

I.INTRODUCTION

In recent years, there has been AN increasing demand to store and method a lot of and a lot of knowledge, in domains like finance, science, and government. Systems that support massive knowledge, and host them victimization cloud computing, are developed and used with success (Hashem et al., 2014).

Whereas massive knowledge is answerable for storing and process knowledge, cloud provides reliable, fault-tolerant, accessible and scalable surroundings so massive knowledge systems will perform (Hashem et al., 2014). Big data, and specifically massive knowledge analytics, square measure viewed by each business and scientific areas as the way to correlate knowledge, notice patterns and predict new trends. thus there's an enormous interest in leverage these 2 technologies, as they will give businesses with a competitive advantage, and science with ways that to combination and summarize knowledge from experiments like those performed at the massive fundamental particle accelerator (LHC).

To be able to fulfill the present needs, massive knowledge systems should be offered, fault tolerant, ascendable and elastic. In this we have a tendency to describe each cloud computing and massive knowledge systems, that specialize in the problems however to be addressed. we have a tendency to notably discuss security issues once hiring an enormous knowledge vendor: knowledge privacy, knowledge governance, and knowledge heterogeneity; disaster recovery techniques; cloud knowledge uploading methods; and the way cloud computing speed and quantifiability poses a tangle relating to exaflop computing.

Despite some problems however to be improved, we have a tendency to gift 2 examples that show however cloud computing

and massive knowledge will work well along. Our contributions to the present progressive is completed by providing an outline over the problems to enhance or have however to be addressed in each technologies.

The remainder of this research is organized as follows: Section two provides a general summary huge or massive of information and cloud computing; Section three discusses and presents 2 examples that show however big information and cloud computing work well along and particularly however hiring an enormous information merchandiser could also be an honest alternative so organizations will avoid IT worries; Section four discusses the many problems to deal with in cloud computing and large information systems; and Section five presents the discussion, conclusions and future work.

II. BIG DATA & CLOUD COMPUTING

The idea of massive information became a serious force of innovation across each lecturers and companies. The paradigm is viewed as a trial to know and obtain correct insights from huge information sets (big data analytics), providing summarized info over large information masses. As such, this paradigm is regarded by companies as a tool to know their purchasers, to induce nearer to them, realize patterns and predict trends. What is more, huge information is viewed by scientists as a mean to store and method large scientific datasets. this idea may be a hot topic and is predicted to still grow in quality within the coming back years.

Although massive knowledge is usually related to the storage of giant numerous knowledge it conjointly issues ways in which to method and extract information from it (Hashem et al., 2014). The 5 completely different aspects wont to describe massive knowledge (commonly cited because the 5 "V"s) square measure Volume, Variety, Velocity, price and truthfulness (Sakr & Gaber, 2014):

Volume: Describes the scale of datasets that a giant information system deals with. Process and storing huge volumes of knowledge is quite troublesome, since it concerns: measurability so the system will grow; availableness, that guarantees access to knowledge and ways that to perform operations over it; and information measure and performance.

Variety: Issues numerous kinds of information from various sources that huge information frameworks have to be compelled to affect.

Velocity: Considerations the various rates at that knowledge streams could get in or out the system Associate in a system provides an abstraction layer so massive knowledge systems will store knowledge severally of the incoming or outgoing rate.

Value: Considerations verity price of knowledge (i.e., the potential price of the info concerning the data they contain). Brobdingnagian amounts of knowledge are input unless they supply price.

Veracity: Refers to the trait of the info, addressing knowledge confidentiality, integrity, and convenience. Organizations ought to make sure that knowledge likewise because the analyses performed on the info are correct.

Cloud computing is another paradigm that guarantees in theory unlimited on-demand services to its users. Cloud's ability to virtualized resources permits abstracting hardware, requiring very little interaction with cloud service suppliers and sanctionative users to access terabytes of storage, high process power, and high handiness in a very pay-as-you-go model (Gonzalez Martinez et al., 2015). Moreover, it transfers value and responsibilities from the user to the cloud supplier, boosting tiny enterprises to that obtaining started within the IT business represents an outsized endeavor, since the initial IT setup takes an enormous effort because the company needs to contemplate the whole value of possession (TCO), as well as hardware expenses, software system licenses, IT personnel and infrastructure maintenance. Cloud computing provides a simple thanks to get resources on a pay-as-you-go model, giving measurability and handiness, which means that firms will simply negotiate resources with the cloud supplier PRN. Cloud suppliers typically supply 3 totally different basic services: Infrastructure as a Service (IaaS); Platform as a Service (PaaS); and software system as a Service (SaaS):

- IaaS delivers infrastructure, which suggests storage, process power, and virtual machines. The cloud supplier satisfies the requirements of the consumer by virtualizing resources per the service level agreements (SLAs);
- PaaS is constructed atop of IaaS and permits users to deploy cloud applications created victimization the programming and run-time environments supported by the supplier. It's at this level that massive knowledge package area unit implemented.

- SaaS is one amongst the foremost famed cloud models and consists of applications running directly within the cloud provider;

These 3 basic services area unit closely related: SaaS is developed over PaaS and ultimately PaaS is made atop of IaaS.

From the overall cloud services alternative services like information as a Service (DBaaS) (Oracle, 2012), BigData as a Service (BDaaS) and Analytics as a Service (AaaS) arose. Since the cloud virtualizes resources in associate on-demand fashion, it's the foremost appropriate and compliant framework for large processing that through hardware virtualization creates a high process power atmosphere for large knowledge.

III. BIG DATA IN THE CLOUD

Cloud computing delivers of these through hardware virtualization. Thus, huge knowledge and cloud computing square measure 2 compatible ideas as cloud allows huge knowledge to be out there, ascendable and fault tolerant.

Business regards huge knowledge as a valuable business chance. As such, many new corporations like Cloud era, Horton works, Teradata and plenty of others, have began to specialize in delivering huge knowledge as a Service (BDaaS) or information as a Service (DBaaS). Corporations like Google, IBM, Amazon and Microsoft additionally give ways in which for customers to consume huge knowledge on demand. Next, we have a tendency to gift 2 examples, Nokia and RedBus, which debate the booming use of massive knowledge inside cloud environments.

Nokia

The increasing quality of wireless networks and mobile Nokia was one in every of the primary firms to know the advantage of massive information in cloud environments (Cloud era, 2012). Many years past, the corporate used individual DBMSs to accommodate every application demand. However, realizing the benefits of desegregation information into one application, the corporate set to migrate to Hadoop-based systems, desegregation information among an equivalent domain, investment the utilization of analytics algorithms to induce correct insights over its purchasers. As Hadoop uses artifact hardware, the value of storage was cheaper than a conventional RDBMS (Cloud era, 2012).

Since Cloud era Distributed Hadoop (CDH) bundles the foremost widespread open supply comes within the Apache Hadoop stack into one, integrated package, with stable and reliable releases, it embodies a good chance for implementing Hadoop infrastructures and transferring IT and technical issues onto the vendors' specialized groups. Nokia regarded huge information as a Service (BDaaS) as a plus and trusty Cloud era to deploy a Hadoop setting hat copes with its necessities in a very short time-frame. Hadoop, and especially CDH, powerfully helped Nokia to fulfill their desires (Cloud era, 2012).

RedBus

RedBus is that the largest company in Asian nation specialized in on-line ticket and edifice booking. This company wished to implement a strong knowledge analysis tool to achieve insights over its bus booking service (Kumar, 2006). Its datasets might simply stretch up to two terabytes in size. the applying would need to be able to analyze booking and inventory knowledge across many bus operators serving over 10000 routes. What is more, the corporate required avoiding fitting and maintaining a posh in-house infrastructure?

At first, RedBus thought-about implementing in-house clusters of Hadoop servers to method

knowledge. but they shortly realized it'd take an excessive amount of time to line up such an answer which it'd need specialized IT groups to take care of such infrastructure. The corporate then regarded Google big Query because the excellent match for his or her desires, permitting them to know what number times shoppers tried to search out Associate in Nursing on the market seat however were unable to try to it due bus overload;

- Examine decreases in bookings;
- Quickly establish server issues by analyzing knowledge associated with server activity;
- Moving towards huge knowledge brought RedBus business blessings;

Know what number time's shoppers tried to search out Associate in Nursing on the market Google big Query armed RedBus with period knowledge analysis capabilities at 2 hundredth of the value of maintaining a posh Hadoop infrastructure (Kumar, 2006).

As supported by Nokia and RedBus examples, switch towards huge knowledge allows organizations to achieve competitive advantage. To boot, BDaaS provided by huge knowledge vendors permits corporations to depart the technical details for giant knowledge vendors and specialize in their core business desires. Seat however was unable to try to it due bus overload;

IV. BIG DATA ISSUES

Although massive knowledge solves several current issues concerning high volumes of knowledge, it's perpetually dynamical space that's forever in development which still poses some problems. During this section we have a tendency to gift a number of the problems not nonetheless self-addressed by massive knowledge and cloud computing.

As the quantity of knowledge grows at a fast rate, keeping all knowledge is physically cost-ineffective. Therefore, companies should be able to produce policies to outline the life cycle and also the expiration date of knowledge (data governance). Moreover, they ought to outline United Nations agency accesses and with what purpose clients' knowledge is accessed. As knowledge moves to the cloud, security and privacy become a priority that's the topic of broad analysis.

Big knowledge DBMSs generally cope with legion knowledge from many sources (variety), and per sec heterogeneousness is additionally a drag that's presently beneath study. Different problems presently being investigated area unit disaster recovery, a way to simply transfer knowledge onto the cloud, and Exaflop computing. In this we give summary over these statements

Security

Cloud computing and large information security could be a current and important analysis topic (Popović & Hocenski, 2015). This drawback becomes a difficulty to firms once considering uploading information onto the cloud. Queries like United Nations agency is that the real owner of the info, wherever is that the information, United Nations agency has access to that and what reasonably permissions they need area unit exhausting to explain. Firms that area unit going to do business with a cloud supplier ought to remember and raise the subsequent questions

a) Who is that the real owner of the info and United Nations agency has access to it?

The cloud provider's purchasers buy a service and transfer their information onto the cloud. However, to that one among the 2 stakeholders will information extremely belong? What is more, will the supplier use the client's data? What level of access must it and with what functions will use it? Will the cloud supplier get pleasure from that data? In fact, IT groups liable for maintaining the client's information should have access to information clusters. Therefore, it's within the client's best interest to grant restricted access to information to attenuate information access and guarantee that solely approved personal access its information for a sound reason.

These queries appear straightforward to retort to, though they must be processed before hiring a service. Most security problems sometimes return from inside the organizations, therefore it's cheap that firms analyze all information access policies before closing a contract with a cloud supplier.

b) Where is that the data?
Sensitive information that's thought-about legal in one country could also be embezzled in another country, therefore, for the sake of the consumer, there ought to be associate degree agreement upon the placement of information, as its information could also be thought-about embezzled in some countries and result introspection.

The problems to those queries area unit primarily based upon agreements (Service Level Agreements – SLAs), however, these should be rigorously checked so as to totally perceive the roles of every neutral and what policies do the SLAs cowl and not cowl regarding the organization's information. This is often generally one thing that has got to be negotiated.

Concerning limiting information accesses, (Tu et al., 2013) and (Popa et al., 2011) came up with a good thanks to write in code information and run analytical Queries over encrypted information. This way, information access isn't any longer a tangle since each information and queries area unit encrypted. all the same, encoding comes with a price, which frequently suggests that higher question process times.

Privacy
The harvest of data and therefore the use of analytical tools to mine information raises many privacy issues. Guaranteeing information security and protective privacy has become very tough as info is unfold and replicated round the globe. Analytics usually mine users' sensitive info like their medical records, energy consumption, on-line activity, market records etc. This info is exposed to scrutiny, raising issues regarding identification, discrimination, exclusion and loss of management (Tene & Polonetsky, 2012). Historically, organizations used varied strategies of de-identification (anonymization or cryptography of information) to distance data from real identities.

Although, in recent years it absolutely was evidenced that even once information is anonymized, it will still be re-identified and attributed to specific people (Tene & Polonetsky, 2012). Some way to resolve this drawback was to treat all information as in person acknowledgeable and subject to a regulative framework.

Although, doing thus may discourage organizations from victimization de-identification strategies and, therefore, increase privacy and security risks of accessing information.

Privacy and knowledge protection laws ar premised on individual management over info and on principles like knowledge and purpose decrease and limitation. With all, it's not clear that minimizing info assortment is often sensible approach to privacy. Nowadays, the privacy approaches once process activities appear to be supported user consent and on the information that people deliberately offer. Privacy is doubtless a problem that wants any improvement as systems store Brobdingnagian quantities of private info daily.

Heterogeneity
Big information issues massive volumes of information however additionally completely different velocities (i.e., information comes at completely different rates reckoning on its supply output rate and network latency) and nice selection. The latter comprehends terribly giant and heterogeneous volumes of information returning from many autonomous sources. Selection is one amongst the "major aspects of huge information characterization"

Data involves massive information software system at completely different velocities and formats from varied sources. This is often as a result of completely different data collectors like their own schemata or protocols for information recording, nature of various applications also end in various information representations (Wu et al., 2014). Addressing such a good sort of information and completely different speed rates may be an onerous task that massive information systems should handle. This task is aggravated by the actual fact that new forms of file area unit perpetually being created with none reasonably standardization. Though, providing uniform and general thanks to represent and explore advanced and evolving relationships from this information still poses a challenge.

Data Governance

The belief that storage is reasonable, and its price is probably going to say no additional, is true relating to hardware costs. However, a giant knowledge database management system will conjointly concern alternative expenses like infrastructure maintenance, energy, and code licenses (Tallon, 2013). of these expenses combined comprise the whole price of possession (TCO), that is calculable to be seven times beyond the hardware acquisition prices.

Regarding that the TCO will increase in direct proportion to the expansion of huge knowledge, this growth should be strictly controlled. Recall that the "Value" (one of huge knowledge Vs) stands to make sure that solely valuable knowledge is keep, since large amounts of knowledge area unit useless if they comprise no price.

Data Governance came to handle this drawback by making policies that outline for the way long knowledge is viable. The thought consists of practices and structure polices that describe however knowledge ought to be managed through its helpful economic life cycle. These practices comprise 3 totally different categories:
1. Structural practices establish key IT and non-IT call manufacturers and their various roles and responsibilities concerning information possession, price analysis and price management

2. Operational practices encompass the method information governance policies square measure applied. Typically, these policies span a spread of actions like information migration, information retention, access rights, value allocation and backup and recovery (Tallon, 2013).

3. Relative practices formally describe the links of the Congress of Industrial Organizations, business managers and information users in terms of information sharing, price analysis, and education, coaching and strategic IT designing.

Although, there area unit limits to what proportion price information governance will bring, as on the far side a precise purpose stricter information governance will have harmful effects.

Disaster Recovery

Data could be a terribly valuable business and losing information will definitely end in losing price. Just in case of emergency or dangerous accidents like earthquakes, floods and fires, information losses ought to be token. To fulfill this demand, just in case of any incident, information should be quickly out there with token period and loss. However, though this is often a awfully vital issue, the analysis during this explicit space is comparatively low (Subashini & Kavitha, 2011), (Wood et al., 2010), (Chang, 2015). For big firms it's imperative to outline a disaster recovery arrange – as a part of information governance arrange – that not solely depends on backups to reset data however conjointly during a set of procedures that permit fast replacement of the lost servers (Chang, 2015).

From a technical perspective, the work represented in (Chang, 2015) presents an honest methodology, proposing a "multi-purpose approach that permits information to be fixed to multiple sites with multiple methods", making certain a recovery share of just about 100%. The study conjointly states that sometimes, information recovery ways use what they decision a "single-basket approach", which implies there's only destination from that to secure the fixed information.

As the loss of knowledge can doubtless end in the loss of cash, it's vital to be able to respond expeditiously to dangerous incidents. with success deploying huge information DBMSs within the cloud and keeping it perpetually out there and fault-tolerant could powerfully rely upon disaster recovery mechanisms.

Other problems

The current state of the art of cloud computing, big data, and large information platforms specifically, prompts another issues. at intervals this section we tend to discuss information transference onto the cloud; Exaflop computing, that presents a serious concern nowadays; and measurability and physical property problems in cloud computing and large data:

a) Transferring information onto a cloud may be terribly slow method and companies usually prefer to physically send arduous drives to information centers in order that data are often

uploaded. However, this can be neither the foremost practical nor the safest answer to transfer information onto the cloud. Through the years there has been a shot to boost and build economical information transferring algorithms to attenuate upload times and supply a secure thanks to transfer information onto the cloud (Zhang, et al. 2013), however, this method still remains a serious bottleneck.

b) Exaflop computing (Geller, 2011), (Schilling, 2014) is one amongst today's issues that's subject of the many discussions. Today's supercomputers and clouds will upset computer memory unit information sets, however, handling Exabyte size datasets still raises countless issues, since high performance and high information measure is needed to transfer and method such Brobdingnagian volumes of information over the network. Cloud computing might not be the solution, because it is believed to be slower than supercomputers since it's restrained by the existent information measure and latency. High performance computers (HPC) square measure the foremost promising solutions, but the annual price of such a laptop is tremendous. Moreover, there square measure many issues in coming up with exaflop HPCs, particularly concerning economical power consumption. Here, solutions tend to be a lot of GPU based mostly} rather than central processing unit based. There also are issues associated with the high degree of correspondence required among hundred thousands of CPUs.

Analyzing Exabyte information sets needs the development of huge data and analytics that poses another drawback nevertheless to resolve.

c) Scalability and physical property in cloud computing and especially concerning huge knowledge management systems could be a theme that desires any analysis because the current systems hardly handle knowledge peaks mechanically. Most of the time, measurability is triggered manually instead of mechanically and also the progressive of automatic climbable systems shows that the majority algorithms area unit reactive or proactive and often explore measurability from the attitude of higher performance. However, a correct climbable system would enable each manual and automatic reactive and proactive measurability supported many dimensions like security, employment rebalance (i.e.: the necessity to rebalance workload) and redundancy (which would modify fault tolerance and availability). Moreover, current knowledge rebalance algorithms area unit supported bar chart building and cargo exploit (Mahesh et al., 2014).

Research challenges

Issues	Existent solutions	Advantages	Disadvantages
Security	Based on SLAs and data Encryption	Data is encrypted	Querying encrypted data is time-consuming
Privacy	-De-identification -User consent	Provides a reasonable privacy or transfers responsibility to the user	It was proved that most de-identification mechanisms can be reverse engineered
Heterogeneity	One of the big data systems' characteristics is the ability to deal with different data coming at different velocities	The major types of data are covered up	It is difficult to handle such variety of data and such different velocities
Data Governance	Data governance documents	-Specify the way data is handled; -Specify data access policies; -Role specification; -Specify data life cycle	-The data life cycle is not easy to define; -Enforcing data governance policies so much can lead to counterproductive effects
Disaster recovery	Recovery plans	Specify the data recovery locations and procedures	Normally there is only one destination from which to secure data
Data Uploading	-Send HDDs to the cloud provider -Upload data through the Internet	Physically sending the data to the cloud provider is quicker than uploading data but it is much more unsecure	Physically sending data to the cloud provider is dangerous as HDDs can suffer damage from the trip. - Uploading data through the network is time-consuming and, without encryption, can be insecure
High Data processing (Exabyte datasets)	-Cloud computing -HPCs	Cloud computing is not so cost expensive as HPCs but HPCs are believed to handle Exabyte datasets much better	HPCs are very much expensive ant its total cost over a year is hard to maintain. On the other hand, cloud is believed that cannot cope with the requirements for such huge datasets

Table 1: Summarizes issues of Big Data

As mentioned in Section three, cloud and large information technologies work o.k. along. Although the partnership between these 2 technologies is established, each still cause some challenges. Table 1 summarizes the problems of massive information and cloud computing these days. The primary column specifies problems whereas the second describes the present resolutions and therefore the remaining present the benefits and downsides of every solution. Concerning the present issues, we tend to outline a number of the doable advances within the next few years.

Security and Privacy will be resolved victimization encryption. However, a replacement generation of systems should make sure that information is accessed quickly which coding doesn't have an effect on process times thus badly;

Big information selection will be addressed by victimization information standardization. This, we believe, is that the next step to reduce the impact of heterogeneity;

Data governance and information recovery plans square measure tough to manage and implement, however as massive information become a factual technology,

corporations square measure beginning to perceive the requirement of such plans.;

New and secure QoS (quality of service) primarily based information uploading mechanisms is also the solution to ease information uploading onto the cloud; the foremost concern depends upon developing totally automatic reactive and proactive systems that square measure capable of coping with load needs mechanically.

V.CONCLUSION

With information increasing on daily base, huge information systems and particularly, analytic tools became a serious force of innovation that has how to store, method and acquire info over computer memory unit datasets. Cloud environments powerfully leverage huge information solutions by providing fault-tolerant, ascendible and out there environments to huge information systems. Although massive information systems area unit powerful systems that change each enterprises and science to urge insights over information, there are a unit some considerations that require more investigation. Further effort should be used in developing security mechanisms and standardizing information varieties. Another crucial component of huge information is quantifiability, that in business techniques area unit largely manual, rather than automatic.

More analysis must be used to tackle this downside. Relating to this specific space, we tend to area unit aiming to use labile mechanisms so as to develop an answer for implementing snap at many dimensions of huge information systems running on cloud environments. The goal is to analyze the mechanisms that labile computer code will use to trigger quantifiability at completely different levels within the cloud stack. Thus, accommodating knowledge peaks in AN automatic and reactive approach. Within this paper we offer an outline of huge knowledge in cloud environments, light its benefits and showing that each technologies work o.k. along however conjointly presenting the challenges sweet-faced by the 2 technologies.

REFERENCES

[1] Chang, V., 2015. Towards a big data system disaster recovery in a Private cloud. Ad Hoc Networks, 000, pp.1–18.

[2] Communications of the ACM, 54(8), p.16. González-Martínez, J. a. et al., 2015. cloud computing and education: A state-of-the-art survey. Computers & Education, 80, pp.132–151.

[3] Hashem, I.A.T. et al., 2014. The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, pp.98–115.

[4] Mahesh, A. et al., 2014. Distributed File System For Load Rebalancing In cloud Computing. , 2, pp.15–20.

[5] Majhi, S.K. & Shial, G., 2015. Challenges in big data cloud Computing And Future Research Prospects: A Review. The Smart Computing Review, 5(4), pp.340–345.

[6] Morgan Kaufmann, B., 2013. Chapter 5 – data governance for big data analytics: considerations for data policies and processes, in: D. Loshin (Ed.), big data Analytics. , pp.pp. 39–48.

[7] CryptDB : A Practical Encrypted Relational DBMS Design, pp.1–13.

[8] Popović, K. & Hocenski, Z., 2015. cloud computing security issues and challenges. , (January), pp.344– 349.

[9] Available at: http://www.industrytap.com/exaflop-computing-will-save-world-can-afford/15485 [Accessed May 26, 2015].

[10] Subashini, S. & Kavitha, V., 2011. A survey on security issues in service delivery models of cloud computing. Journal of Network and Computer Applications, 34(1), pp.1–11.

[11] Tallon, P.P., 2013. Corporate governance of big data: perspectives on value, risk, and cost. Computer 46, pp.pp. 32–38.

[12] Tu, S. et al., 2013. Processing analytical queries over encrypted data. Proceedings of the VLDB Endowment, 6(5), pp.289–300.

[13] Wood, T. et al., 2010. Disaster recovery as a cloud service: Economic benefits & deployment challenges. 2nd USENIX Workshop on Hot Topics in cloud Computing, pp.1–7.

[14] Wu, X. et al., 2014. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), pp.97–107.