IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

A Comprehensive Survey on Big-Data Issues, Challenges and Management Approaches on Cloud Environment

Chetana Tukkoji[1], Dr. Seetharam K[2]

Assistant Professor, Dept of CSE, GITAM University, Bangalore[1, 2]

Abstract: The explosion of data in volume, variety and velocity is catalyzing new business models and reshaping industries. Business leaders can no longer amble their way forward in the age of Big Data. The question that arises now is how to develop a high performance platform to efficiently analyze big data and how to design an appropriate mining algorithm to find the useful things from big data. To deeply discuss this issue, this paper begins with a brief introduction to big data, followed by issues of big data analytics, challenges of big data, and the various tools associated with it and further research directions will also be presented for the next step of big data analytics.

Keywords: Big data analytics, Hadoop, MapReduce, MOA.

I. INTRODUCTION TO CLOUD COMPUTING

Cloud computing is a type of Internet-based computing that provides shared computer processing resources and data to computers and other devices on demand. It is a model for enabling ubiquitous, on-demand access to a shared pool of configurable computing resources (e.g., computer networks, servers, storage, applications and services).

Storage of big data is never a problem due to ubiquitous services offered by cloud. With a massive volume of data being generated, achieving effective storage and processing it becomes a very challenging task with respect to cost and optimization of data. Such preference of big data has raised the attention of research community as solving the key issues of managing big data will eventually lead to innovation and maximization of productivity. Cloud providers typically use a "pay as you go" model. This will lead to unexpectedly high charges if administrators do not adapt to the cloud pricing model [6].

Cloud computing exhibits the following key characteristics:

Agility: as cloud computing may increase users' flexibility with re-provisioning, adding or expanding technological infrastructure resources.

Cost: reductions are claimed by cloud providers. A public-cloud delivery model converts capital expenditures , as it provides the 'pay as you use' model.

Location independence: enable users to access systems using a web browser regardless of their location or what device they use (e.g., PC, mobile phone).

Maintenance: maintaining applications is easier, because they do not need to be installed on each user's computer and can be accessed from different places (e.g., different work locations, while travelling, etc.).

Multi-tenancy: enables sharing of resources and costs across a large pool of users thus allowing for:

- Centralization of infrastructure in locations with lower costs (such as real estate, electricity, etc.)
- Peak-load capacity increases (users need not engineer and pay for the resources and equipment to meet their highest possible load-levels).
- Utilization & efficiency improvements for systems that are often only 10–20% utilized.

Performance evaluation: It is monitored by IT experts from the service provider and consistent and loosely coupled architectures are constructed using web services as the system interface.

Productivity: It may be increased when multiple users can work on the same data simultaneously, rather than waiting for it to be saved and emailed. So that the time can be saved.

Reliability: improves with the use of multiple redundant sites, which makes well-designed cloud computing suitable for business continuity and disaster recovery.

Scalability: the dynamic ("on-demand") provisioning of resources on a fine-grained, self-service basis in real-time (Note, the VM startup time varies by VM type, location, OS and cloud providers), without users having to engineer for peak loads. This gives the ability to scale up when the usage need increases or down if resources are not being used.

Security: can improve due to centralization of data, increased security focused resources, etc., but concerns can persist about loss of control over certain sensitive data and the lack of security for stored kernels.

Big-data Concepts:

The big data are in the form of audio, video, text, satellite image, medical images and many more originated from different business operation. Storage of big data is never a problem due to ubiquitous services offered by cloud. With a massive volume of data being generated, achieving

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

effective storage and processing it becomes a very challenging task with respect to cost and optimization of data. Such preference of big data has raised the attention of research community as solving the key issues of managing big data will eventually lead to innovation and maximization of productivity.

Apparently, networks play a critical role in bridging the different stages, and there is a strong demand to create a fast and reliable inter-connected network for the big data to flow freely on this digital highway. This network highway concerns not only just one segment of data delivery -rather, the whole series of segments for the lifecycle of big data, from access networks, to Internet backbone and to intra and inter-datacenter networks. For each layer of the network, the specific requirements that big data transmission poses should be satisfied. In summary, to attain the full speed for big data transmission and processing, every segment of the network highway should be optimized and seamlessly concatenated.

From the perspective of networking, the standards are required to specify how to transfer big data between different platforms. The transferred big data may contain semi-structured or unstructured data, such as pictures, audios, videos, click streams, log files, and the output of sensors that measure geographic or environmental information. The standards should specify how these data should be encoded and transferred in network systems, so as to facilitate with network QoS management with low latency and high fidelity. Moreover, as novel technologies, like Software Defined Network (SDN), keep emerging in networking systems, corresponding standards are required to specify how these technologies can be exploited for efficient big data transmission. We generally classify big data applications into two categories as Internet applications and mobile wireless network applications with regard to the networking infrastructure they work on. For a single datacenter level, novel network topologies and hardware devices provide cost-effective solutions toward a high performance network that interconnects all servers. For the microscopic level, coordinated all-to-all communications among a subset of servers in a datacenter mitigate network congestion and reduce the completion times of data processing jobs. By analyzing the big data gathered by network monitoring system, those misbehaviors can be identified pro-actively, thus greatly reducing the potential loss.

II. CHARACTERISTICS OF BIG DATA AND ITS ISSUES

Big-data characteristics:

The characteristics of the big data can be explained as below [7][8].

- **Data Volume:** This is the volume of data which completely measures the organizational data. The measurement of data is necessary for accessing of important data. The increase in data volume may lead to the reduction of data quantity, richness, etc.

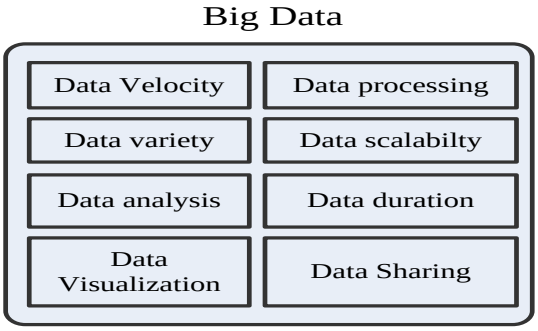


Figure1. Important terms associated with Big Data

- **Data Velocity:** This functions for measurement of data generation, data streaming and data aggregation speed. Today the management of data velocity is like bandwidth issues.
- **Data Variety:** The different kinds of data including audio, image, and video can be considered as data variety in big data. These generated data may be of structured or unstructured or semi-structured type.
- **Data Value:** The measurement of this data value will be helpful in decision making. In big data the value of data, helps in computing.
- **Data Complexity:** The complexity of data is nothing but the interconnection level of data with other elements. The data interconnection may be large or small ripple across the system. This will affect the system behavior in big data.

III. CHALLENGES IN BIG DATA

Recent year's big data has been accumulated in several domains like health care, public administration, retail, bio-chemistry and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents and internet search indexing [9]. Social computing includes social net- work analysis, online communities, recommender systems, reputation systems and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, Thomson Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges [1].

Storage: Clearly not enough hard disks/devices to store the data. Distributed storage is still not enough; manufacturers cannot make enough storage devices in time. Speed in writing to devices, bigger data paths/data-bus.

Integrating and processing of data is more and more[4].

Latency and Bandwidth: Processing big data takes more time.

Taxonomy and Ontology: Data generated from different sources that need to classify. But there is no such standard tools are available.

Security and Privacy: What data is to be secured?

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

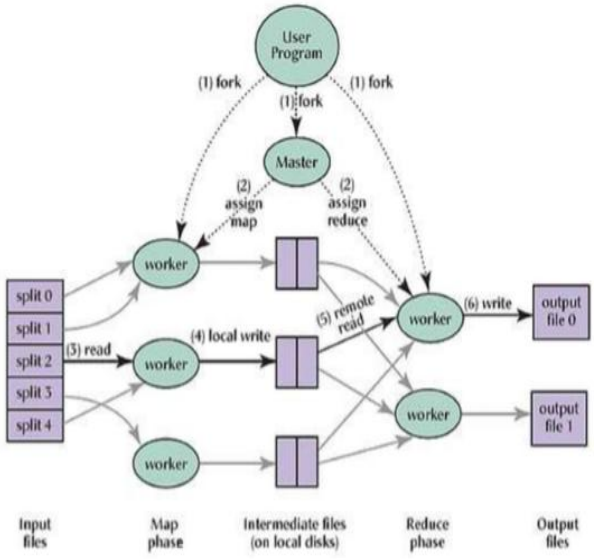
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

IV. BIG DATA MANAGEMENT TOOLS AND TECHNIQUES

MapReduce framework: It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster [5][10]. Mainly it performs on two separate tasks**:**

i). Map: where the data is converted into individual elements.

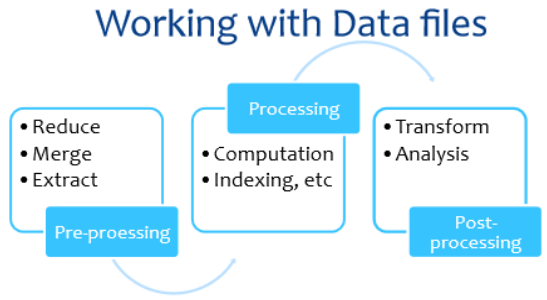
ii). Reduce: combine the result from map into smaller set of tuples for further analysis.



Hadoop: An implementation of MapReduce framework

i). Hadoop is based on the MapReduce framework. It schedules, monitors and re-executes the tasks [3].

ii). Hadoop stack now includes the components for query and storage in addition to MapReduce



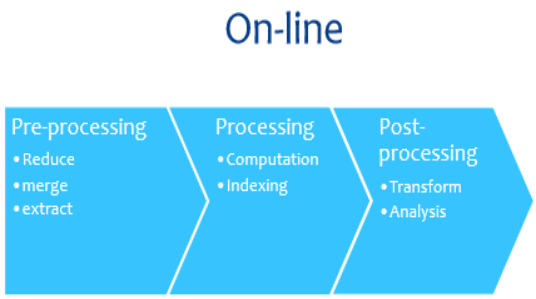
MOA: As a framework for online analytics

i). MOA performs stream mining in real time and large scale machine learning.

ii). MOA can be extended with new mining algorithms and new stream generators or evaluation measures.

iii). The goal is to provide a benchmark suite for the stream mining community.

iv). MOA is also easily used with Hadoop S4 or storm this provides for a more robust and configurable system[2]



V. CONCLUSION

This survey paper gives an idea to know about the issues and challenges of big data in cloud environment. As the big data refers the large volume of data and cloud computing can offer the better scalability for the large data. And also describes the techniques which are useful for analyzing the large data, which is generated from difference sources. In order to handle big data use the standard tools like MapReduce, Hadoop and MOA. The future research scope hints an idea for better management of big data in cloud environment.

REFERENCES

[1] D. P. Acharjya and kauser ahamd P"A Survey on Big Data Analytics: Challenges, Open Research Issues and Tool", (IJACSA), Vol. 7, No. 2, 2016.

[2] http://moa.cms.waikato.ac.nz/

[3] DR. A. N. Nandakumar1, Nandita Yambem "A Survey on Data Mining Algorithms on Apache Hadoop Platform" (IJETAE)Vol 4, Issue 1, January 2014.

[4] Albert Bifet, (2013), "Mining Big data in Real time", Informatica 37, pp15-20

[5] Gong-Qing Wu ―MReC4.5: C4.5 Ensemble Classification with MapReduce‖ chinagrid 2009.

[6] Hashem, Ibrahim Abaker Targio, et al. "The rise of "big data" on cloud computing: Review and open research issues." Information Systems 47 (2015): 98-115.

[7] Géczy, Peter. "Big data characteristics." The Macrotheme Review 3.6 (2014): 94-104

[8] Kaisler, Stephen, et al. "Big data: issues and challenges moving forward." System Sciences (HICSS), 2013 46th Hawaii International Conference on. IEEE, 2013

[9] Puneet Singh Duggal, Sanchita Paul, (2013), "Big Data Analysis:Challenges and Solutions", Int. Conf. on Cloud, Big Data and Trust, RGPV .

[10] Chanchal Yadav, Shullang Wang, Manoj Kumar, (2013) "Algorithm and Approaches to handle large Data- A Survey", IJCSN, 2(3), ISSN:2277-5420(online), pp2277-542