ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.4

Big Data Computing and Clouds: Trends and Future Directions

P. Arpitha[1], Dr.P.V.Kumar[2]
Research Scholar[1], Professor[2]
Department of Computer Science
Rayalaseema University, India[1]
VRS, Osmania University, India[2]

Abstract:
Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. 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),[1][2] which can be rapidly provisioned and released with minimal management effort. Cloud computing and storage solutions provide users and enterprises with various capabilities to store and process their data in either privately owned, or third-party data centers[3] that may be located far from the user–ranging in distance from across a city to across the world. Cloud computing relies on sharing of resources to achieve coherence and economy of scale, similar to a utility (like the electricity grid) over an electricity network. This paper discusses approaches and environments for carrying out analytics on Clouds for Big Data applications. It revolves around four important areas of analytics and Big Data, namely (i) data management and supporting architectures; (ii) model development and scoring; (iii) visualisation and user interaction; and (iv) business models. Through a detailed survey, we identify possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.

Keywords: Big data, cloud computing, data management

I. INTRODUCTION:

Society is becoming increasingly more instrumented and as a result, organizations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured and unstructured data are important as they can help organizations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimise the use of resources. This paradigm is being popularly termed as Big Data. The most often claimed benefits of Clouds include offering resources in a pay-as-you-go fashion, improved availability and elasticity, and cost reduction. Clouds can prevent organizations from spending money for maintaining peak-provisioned IT infrastructure that they are unlikely to use most of the time. Whilst at first glance the value proposition of Clouds as a platform to carry out analytics is strong, there are many challenges that need to be overcome to make Clouds an ideal platform for scalable analytics. In this article we survey approaches, environments, and technologies on areas that are key to Big Data analytics capabilities and discuss how they help building analytics solutions for Clouds. We focus on the most important technical issues on enabling Cloud analytics, but also highlight some of the non-technical challenges faced by organizations that want to provide analytics as a service in the Cloud. In addition, we describe a set of gaps and recommendations for the research community on future directions on Cloud-supported Big Data computing.

2. BACKGROUND AND METHODOLOGY:

Organizations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity and [web site tracking, sensors, finance, accounting, among other reasons. With the advent of social network Web sites, users create records of their lives by daily posting details of activities they perform, events they attend, places they visit, pictures they take, and things they enjoy and want. This data deluge is often referred to as Big Data and a term that conveys the challenges it poses on existing infrastructure with respect to storage, management, interoperability, governance, and analysis of the data. In today's competitive market, being able to explore data to understand customer behavior, segment customer base, offer customized services, and gain insights from data provided by multiple sources is key to competitive advantage. Although decision makers would like to base their decisions and actions on insights gained from this data , making sense of data, extracting non obvious patterns, and using these patterns to predict future behavior are not new topics. Knowledge Discovery in Data (KDD) aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining more specifically, aims to discover previously unknown interrelations

among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions. Fig. 1 depicts the common phases of a traditional analytics workflow for Big Data. Data from various sources, including databases, streams, marts, and data warehouses, are used to build models. The large volume and different types of the data can demand pre-processing tasks for integrating the data, cleaning it, and filtering it. The prepared data is used to train a model and to estimate its parameters. Once the model is estimated, it should be validated before its consumption. Normally this phase requires the use of the original input data and specific methods to validate the created model. Finally, the model is consumed and applied to data as it arrives. This phase, called model scoring, is used to generate predictions, prescriptions, and recommendations. The results are interpreted and evaluated, used to generate new models or calibrate existing ones, or are integrated to pre-processed data.

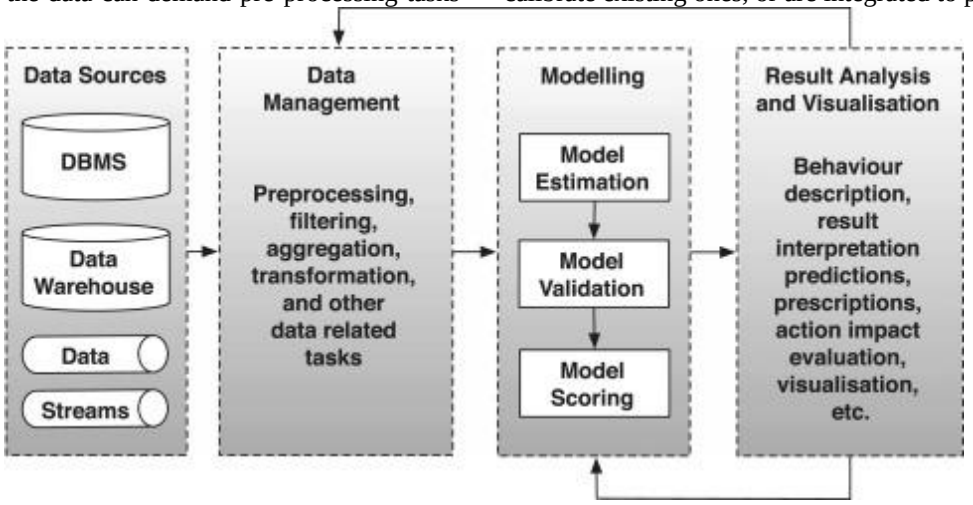


Figure.1. Overview of the analytics workflow for Big Data.

Figure options

Analytics solutions can be classified as descriptive, predictive, or prescriptive as illustrated in Fig. 2. Descriptive analytics uses historical data to identify patterns and create management reports; it is concerned with modelling past behaviour. Predictive analytics attempts to predict the future by analysing current and historical data. Prescriptive solutions assist analysts in decisions by determining actions and assessing their impact regarding business objectives, requirements, and constraints.

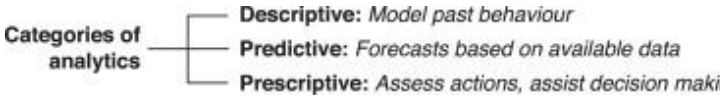


Figure.2. Categories of analytics.

3. DATA MANAGEMENT:

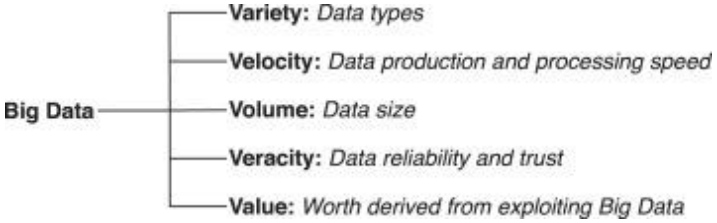
***Private*:** deployed on a private network, managed by the organisation itself or by a third party. A private Cloud is suitable for businesses that require the highest level of control of security and data privacy. In such conditions, this type of Cloud infrastructure can be used to share the services and data more efficiently across the different departments of a large enterprise.

***Public*:** deployed off-site over the Internet and available to the general public. Public Cloud offers high efficiency and shared resources with low cost. The analytics services and data management are handled by the provider and the quality of service (*e.g.* privacy, security, and availability) is specified in a contract. Organisations can leverage these Clouds to carry out analytics with a reduced cost or share insights of public analytics results.

***Hybrid*:** combines both Clouds where additional resources from a public Cloud can be provided as needed to a private Cloud. Customers can develop and deploy analytics applications using a private environment, thus reaping benefits from elasticity and higher degree of security than using only a public Cloud. Considering the Cloud deployments, the following scenarios are generally envisioned regarding the availability of data and analytics models (i) data and models are private; (ii) data is public, models are private; (iii) data and models are public; and (iv) data is private, models are public. Jensen et al. advocate on deployment models for Cloud analytics solutions that vary from solutions using privately hosted software and infrastructure, to private analytics hosted on a third party infrastructure, to public model where the solutions are hosted on a public Cloud.

3.1. Data variety and velocity

Big Data is characterised by what is often referred to as a multi-V model, as depicted inFig. 3. Variety represents the data types, velocity refers to the rate at which the data is produced and processed, and volume defines the amount of data. Veracity refers to how much the data can be trusted given the reliability of its source



3.2. Data storage

Several solutions were proposed to store and retrieve large amounts of data demanded by Big Data, some of which are currently used in Clouds. Internet-scale file systems such as the Google File System (attempt to provide the robustness, scalability, and reliability that certain Internet services need.

3.3. Data integration solutions

Forrester Research published a technical report that discusses some of the problems that traditional Business Intelligence (BI) faces, highlighting that there is often a surplus of siloed data preparation, storage, and processing.

3.4 Challenges in big data management

In this section, we discuss current research targeting the issue of Big Data management for analytics. There are still, however, many open challenges in this topic. The list below is not exhaustive, and as more research in this field is conducted, more challenging issues will arise.

Data variety:

How to handle an always increasing volume of data? Especially when the data is unstructured, how to quickly extract meaningful content out of it? How to aggregate and correlate streaming data from multiple sources?

Data storage:

How to efficiently recognise and store important information extracted from unstructured data? How to store large volumes of information in a way it can be timely retrieved? Are current file systems optimised for the volume and variety demanded by analytics applications? If not, what new capabilities are needed? How to store information in a way that it can be easily migrated/ported between data centers/Cloud providers?

Data integration:

New protocols and interfaces for integration of data that are able to manage data of different nature (structured, unstructured, semi-structured) and sources.

Data Processing and Resource Management:

New programming models optimized for streaming and/or multidimensional data; new backend engines that manage optimized file systems; engines able to combine applications from multiple programming models (*e.g.* MapReduce, workflows, and bag-of-tasks) on a single solution/abstraction. How to optimise resource usage and energy consumption when executing the analytics application?

- **Focusing on the business** -Since all the services will execute over the internet, a government does not have to bother about technical issues and other problems associated with physical storage and backup. A government can thus focus more on their core services.

- **Performance** - It delivers reliable performance irrespective to the geographical location of the user. Another key feature could be the automatic updating of services and applications.

- **Security** - Cloud Computing offers optimum security which protects you against any unauthorized access, modification and loss of data.

- **Flexibility** - Even if part of the cloud environment fails or stops working, the other resources continue to work until the problem is fixed.

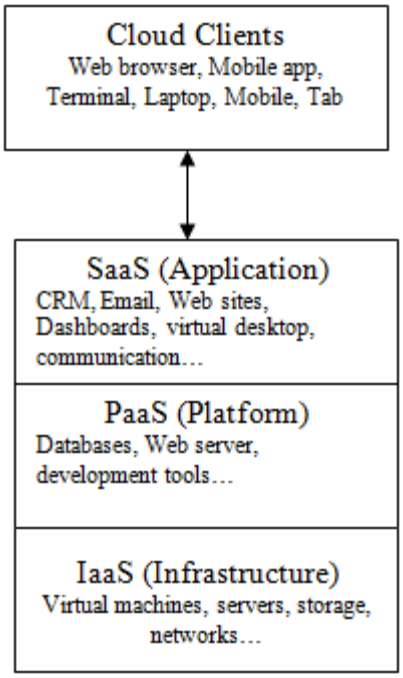


Figure.3. Architecture of Cloud Computing

4. CLOUD AT A GLANCE:

Cloud Flavours

- Public cloud where resources are dynamically provisioned on a fine-grained, selfservice• basis over the Internet from an off-site third-party provider [2]
- Private cloud provides computing on private networks. These capitalize on data• security, corporate governance and reliability.
- Hybrid cloud environment is combination of both Public and Private Cloud. This is• more typical of cloud computing for most enterprises.

Figure.4. Cloud Computing Architecture

Cloud Services:

Cloud-computing providers offer different type of services according to different models.

- IaaS refers to online services that abstract user from the detail of infrastructure like

• physical computing resources, location, data partitioning, scaling, security, backup etc.

PaaS provider typically develops toolkit and standards for development and channels• for distribution and payment. In the PaaS models, cloud providers deliver a computing platform, typically including operating system, programming-language execution environment, database, and web server [2].

➢ SaaS is where a user no longer owns the software but instead uses it when required• using cloud computing. The software remains the property of the service provider and the user pays for access either by annual subscription or on a pay-per-use basis.

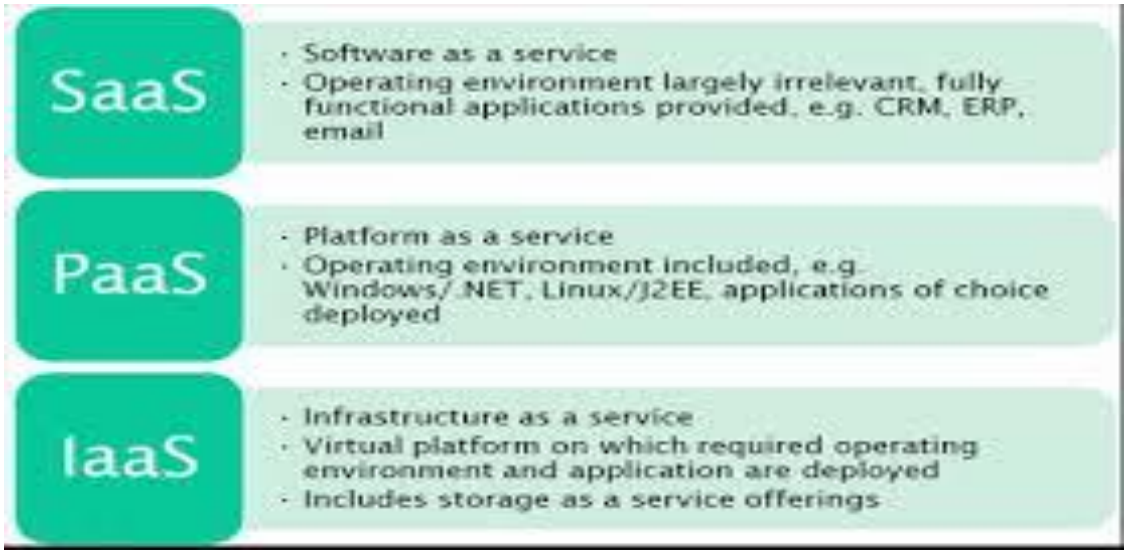


Figure.5. Cloud Services

5. VISIONS OF DIGITAL INDIA AND CLOUD COMPUTING:

Vision enabled by Cloud computing which are coincides with visions of digital India:

➢ Hybrid cloud / public cloud for information and help in online forms. Data submission and manipulation using Authentication, Authorization and Audit using private cloud. Use on-premise and off-premise aggregator (using cloud solutions such as IBM Caste Iron).
➢ Common Service Centre information and private space on public cloud.
➢ Individual documents and certificates using hybrid cloud.
➢ Collaborative digital platform using public cloud Service Enablement.
➢ Highly secured data of classified nature should be hosted within the firewall and available on secured private cloud. This will have multiple security checks on secured network.
➢ Open information and services from the Government bodies can use public cloud provided by one of the vendors such as Amazon Web Services, Microsoft Azure or Google Cloud.
➢ For few services that would require aggregation, customization and integration with other cloud providers, hybrid cloud could be used. Aggregation is managed wherein secured data within government private cloud is manipulated/stripped-off to provide data to an application on a public cloud.

To enable Digital India it is necessary to evaluate the type of services that will be provided to citizens. Digital India should have strategy wherein the Government will be providing information and services to internal and external stakeholders. This requires having a strong architecture principle and policy to host data and services to relevant cloud delivery model. [3]

6. FUTURE SCOPE OF CLOUD COMPUTING:

The summary of a few reports from popular research analyst firms are as below:

Cisco Consulting Services (CCS): The study surveyed 4,226 IT leaders in 18 industries across 9 key economies, including 600 from India. Wherein about 83% of respondents in India were "very satisfied" and another 13% "somewhat satisfied" with cloud, representing a total 96% positive rating.

Gartner: Cloud computing will become the bulk of new IT spend by 2016. From 2012 through 2017, across all segments of the cloud computing market, cloud services revenue is projected to have a Compound Annual Growth Rate (CAGR) of 33.2%, with SaaS and IaaS growth rates projected to be 34.4% and 39.8% respectively.Also, with higher rates of SaaS adoption, by 2017, USD 4.2 Bn will be spent on cloud services in India, USD 1.8 Bn of which will be spent on software as a service (SaaS)

CII: Public cloud computing in India is forecast to grow 36% in 2013 to total USD 443 Mn, up from USD 326 Mn in 2012. The Indian IT-BPO vendors can develop their social media, mobility, analytics and cloud computing (SMAC) strategies and cross the USD 225 Bn mark by 2020.

VMware, Inc. Cloud index: A study, conducted by Forrester Research across 12 Asia Pacific countries, reveals that nearly 89% respondents in India believe that Cloud Computing, or 'as-a-service' approach, is relevant to their organisation and nearly 79% say they currently have a cloud-related initiative in place within the organisation, or are planning to implement cloud, or 'as-a-service' approach, in the next 12 months.

Microsoft IDC study: Cloud computing will generate some 14 Mn new jobs worldwide by 2015, and India alone will create over 2 Mn jobs.

EMC and Zinnov Management Consulting: India would require at least 100,000 professionals in private cloud alone by 2015.[4]

7. RISK IN CLOUD COMPUTING:

Risks in policies, like data privacy, security, intellectual property rights, cybercrime etc., which are considered critical for the future of cloud computing.[5]

8. CONCLUSION:

The vision of Digital India aims to transform the country into a digitally empowered society and knowledge economy. The Digital India is transformational in nature and would ensure that Government services are available to citizens electronically. It would also bring in public accountability through mandated delivery of government's services electronically, a Unique ID and e-Pramaan based on authentic and standard based interoperable and integrated government applications and data basis. To became India digitally empowered technology plays an important role. Therefore cloud computing technology having highest future demand for smooth functioning, security and fulfilling the goals set for Digital India.

9. REFERENCES:

[1] http://searchcloudcomputing.techtarget.com/definition/cloud-computing

[2] https://www.urbanpro.com/a/cloud-computing-future-scope

[3] http://pib.nic.in/newsite/PrintRelease.aspx?relid=108926

[4]http://in.pcmag.com/networking-communications-software/38970/feature/what-is-cloud-computing

[5] https://en.wikipedia.org/wiki/Digital_India

[6]. http://www.huffingtonpost.in/2015/07/02/digital-india-modi_n_7711622.html