ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

BIG DATA PROCESS AND ITS INVOLVEMENT IN CLOUD COMPUTING

S.MydeenBadhshaB.Sc.[1]P.ThaneemFailzaM.E.[2]

[1]*PG Departmentof Computer Science, SadakathullahAppa College, Tirunelveli*
Tamil Nadu, India
msbathusha@yahoo.com

[2]*PG Departmentof Computer Science, Assistant Professor, SadakathullahAppa College, Tirunelveli*
Tamil Nadu, India
faz.evergreen@gmail.com

Abstract -**The advent of the digital age has led to a rise in different types of data with every passing day. This data is complex and needs to be stored, processed and analyzed for information that can be used by organizations. Recent days, big data is beginning to have a major impact for using such type of complex data. Cloud computing provides an apt platform for big data analytics in view of the storage and computing requirements of the latter. This makes cloud-based analytics a viable research field. However, several issues need to be addressed and risks need to be mitigated before practical applications of this synergistic model can be popularly used. This paper is to clearly elaborate the processing of big data with the challenges and the exposure of big data for cloud computing.**

Key words-**Cloud computing, Big Data**

I. INTRODUCTION

With the advent of the digital age, the amount of data being generated, stored and shared has been on the rise. From data warehouses, webpages and blogs to audio/video streams, all of these are sources of massive amounts of data. The result of this proliferation is the generation of massive amounts of pervasive and complex data, which needs to be efficiently created, stored, shared and analyzed to extract useful information.This data has huge potential, ever-increasing complexity, insecurity and risks, and irrelevance. The requirement of an efficient and effective analytics service, applications, programming tools and frameworks has given birth to the concept of Big Data Processing and Analytics.

Big data analytics has found application in several domains and fields. Some of these applications include medical research, solutions for the transportation and logistics sector, global security and prediction and management of issues concerning the socio-economic and environmental sector, to name a few. Apart from standard applications in business and commerce and society administration, scientific research is one of the most critical applications of big data in the real world. Big data, by definition, is a term used to describe a variety of data - structured, semi-structured and unstructured, which makes it a complex data infrastructure.The complexity of this infrastructure requires powerful management and technological solutions. “Big data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. The most popular definition in recent years uses the “Three V’s”: volume (size of datasets and storage), velocity (speed of incoming data), and variety (data types). Reacting fast enough to deal with data velocity is a challenge for most organizations. With increasing interest and insight in the field, the “Three V’s” have been expanded to “Five V’s”: volume, velocity, variety, veracity (integrity of data), value (usefulness of data) and complexity (degree of interconnection among data structures Figure 1 illustrates the Five V’s model. The different types of data available on a dataset determine variety while the rate at which data is produced determines velocity. Predictably, the size of data is called volume.

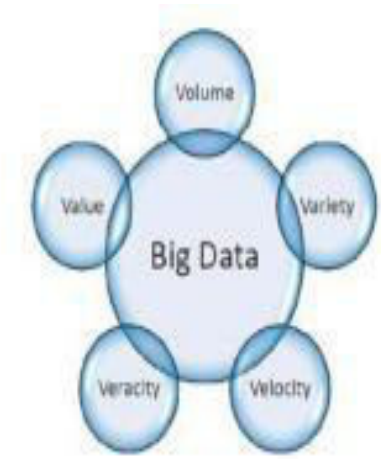

Fig. 1 – Five V's model of big data

The cloud computing environment offers development, installation and implementation of software and data applications 'as a service'. The services are as follows: platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS). Infrastructure-as-a-service is a model that provides computing and storage resource as a service. On the other hand, in case of PaaS and SaaS, the cloud services provide software platform or software itself as a service to its clients.

II. BIG DATA IN CLOUD

Traditional data management tools and data processing or data mining techniques cannot be used for Big Data Analytics for the large volume and complexity of the datasets that it includes. Conventional business intelligence applications make use of methods, which are based on traditional analytics methods and techniques and make use of OLAP, BPM, Mining and database systems like RDBMS.

It was in the 1980s that artificial intelligence-based algorithms were developed for data mining. Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand and Steinberg [25] mention the ten most influential data mining algorithms k-means, C4.5, Apriori, Expectation Maximization (EM), PageRank, SVM (support vector machine), AdaBoost, CART, a ve Bayes and kNN (k-nearest neighbors). Most of these algorithms have been used commercially as well. Alam and Shakil [38] propose architecture for management of data through cloud techniques.

One of the most popular models used for data processing on cluster of computers is MapReduce/Hadoop, the most productive model for Big Data Analytics yet mentions that languages and extensions like HiveQL, Latin and Pig have overpowering benefits for this use [33].Hadoop is simply an open-source implementation of the MapReduce framework, which was originally created as a distributed file system.

For cloud-based big data analytics, several frameworks like Google MapReduce, Spark, Haloop, Twister, Hadoop Reduce and Hadoop++ are available. Figure 2 gives a pictorial representation of the use of cloud computing in big data analytics. These frameworks are used for storing and processing of data. In order to store this data, which may be of any structure,databases like HBase, BigTable and HadoopDB may be used. When it comes to data processing, the Pig and Hive technologies come into the picture.

Some of the recent research breakthroughs and milestones in cloud-based big data analytics are discussed here. Lee [16] elaborates on the advantages and limitations of MapReduce in parallel data analytics. A Hadoop-based data analytics system, created by Starfish [13], improves the performance of the clusters throughout the cycle of data analytics. Moreover, the users are not required to understand the configuration details.

Research efforts have been made to create a big data management framework for the cloud. Khan, Naqvi, Alam and Rizvi [35] propose a data model and provides a schema for big data in the cloud and attempts to ease the process of querying data for the user. Moreover, an important subject of research has been performance and speed of operation. Ortiz, Oneto and Anguita [28] explore the use of a proposed integrated Hadoop and MPI/OpenMP system and how the same can improve speed and performance.

In view of the fact that data needs to be transferred between data centers that are usually located distances apart, power consumption becomes a crucial parameter when it comes to analyzing efficiency of the system. A network-based routing algorithm called GreeDi can be used for finding the most energy efficient path to the cloud data center during big data processing and storage [29].

There are several practical simulation-enabled analytics systems. One such system is given by Li, Calheiros, Lu, Wang, Palit, Zheng and Buyya [17], which is a Direct Acrylic Graph (DAG) form analytical application used for modeling and predicting the outbreak of Dengue in Singapore.

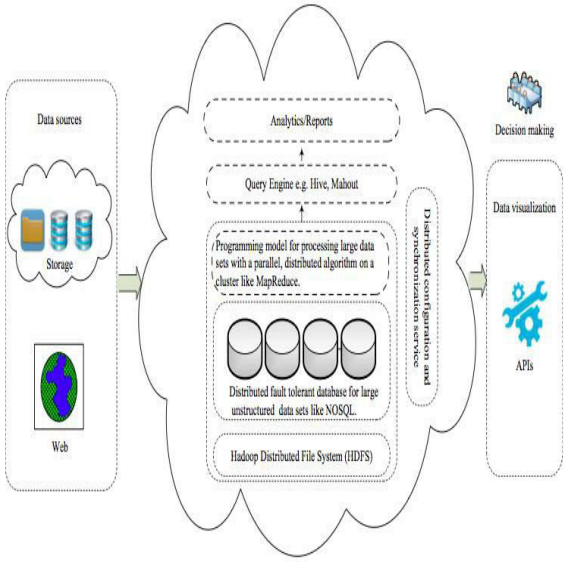

Fig 2: cloud computing with Big Data

III.BIG DATA ANALYSIS PROCESSING PIPELINE

The process of big data analysis is performed in five steps as shown in fig. 3. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

A. Process pipeline

1) Data Acquisition and Recording

Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce petabytes of data today.

Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information

The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data provenance. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. For example, a processing error at one step can render subsequent analysis useless; with suitable provenance, we can easily identify all subsequent processing that dependent on this step. Thus we need research both into generating suitable metadata and into data systems that carry the provenance of data and its metadata through data analysis pipelines.

2) Information Extraction and Cleaning

Frequently, the information collected will not be in aformat ready for analysis. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements (possibly with some associated uncertainty), and image data such as x-rays. We cannot leave the data in this form and still effectivelyanalyze it. Rather we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge. Note that this data also includes images and will in the future include video; such extraction is often highly application dependent (e.g., what you want to pull out of an MRI is very different from what you would pull out of a picture of the stars, or a surveillance photo). In addition, due to the ubiquity of surveillance cameras and popularity of GPS-enabled mobile phones, cameras, and other portable devices, rich and high fidelity location and trajectory (i.e., movement in space) data can also be extracted.

We are used to thinking of Big Data as always telling us the truth, but this is actually far from reality. Existing work on data cleaning assumes well-recognized constraints on valid data or well-understood error models; for many emerging Big Data domains these do not exist.Fig3: Big Data Analysis Pipeline

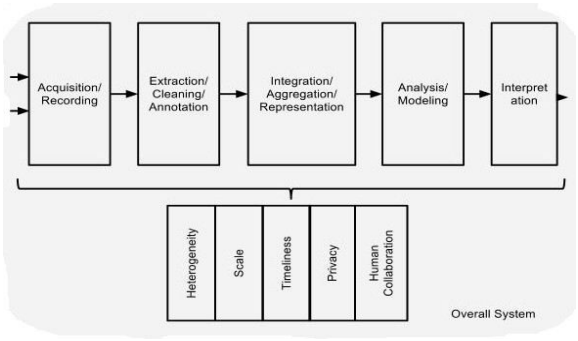

Fig3: Big Data Analysis Pipeline

3. Data Integration, Aggregation, and Representation

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Consider, for example, data from a range of scientific experiments. If we just have a bunch of data sets in a repository, it is unlikely anyone will ever be able to find, let alone reuse, any of this data. With adequate metadata, there is some hope, but even so, challenges will remain due to differences in Even for simpler analyses that depend on only one data set, there remains an important question of suitable database designDatabase design is today an art, and is carefully executed in the enterprise context by highly-paid professionals. We must enable other professionals, such as domain scientists, to create effective database designs, either through devising tools to assist them in the design process or through forgoing the design process completely and developing techniques so that databases can be used effectively in the absence of intelligent database design.

4. *Query Processing, Data Modeling, and Analysis*

Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models.

Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve experimental details and in data record structure.

Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to be expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions.

Big Data is also enabling the next generation of interactive data analysis with real-time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, to populate hot-lists or recommendations, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today.

A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses. Today's analysts are impeded by a tedious process of exporting data from the database, performing a non-SQL process and bringing the data back. This is an obstacle to carrying over the interactive elegance of the first generation of SQL-driven OLAP systems into the data mining type of analysis that is in increasing demand. A tight coupling between declarative query languages and the functions of such packages will benefit both expressiveness and performance of the analysis.

5. Interpretation

Having the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. Thisinterpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we

saw above, there are many possible sources of error: computer systems can have bugs, models almost always have assumptions, and results can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the computer. The computer system must make it easy for her to do so. This is particularly a challenge with Big Data due to its complexity. There are often crucial assumptions behind the data recorded. Analytical pipelines can often involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision-maker diligence -- rather than accept the stated solvency of a financial institution at face value, a decision-maker has to examine critically the many assumptions at multiple stages of analysis.

In short, it is rarely enough to provide just the results. By studying how best to capture, store, and query provenance, in conjunction with techniques to capture adequate metadata, we can create an infrastructure to provide users with the ability both to interpret analytical results obtained and to repeat the analysis with different assumptions, parameters, or data sets.

Furthermore, with a few clicks the user should be able to drill down into each piece of data that she sees and understand its provenance, which is a key feature to understanding the data. That is, users need to be able to see not just the results, but also understand why they are seeing those results.

B. Challenges in Big Data Analysis

Having described the multiple phases in the Big Data analysis pipeline, we now turn to some common challenges that underlie many, and sometimes all, of these phases. These are shown as five boxes in the second row of Fig. 3.

1) Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many (traditional) data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

2) Scale

Of course, the first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But,there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static. The dramatic shift that is underway is the move towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap) into very large clusters. This level of sharing of resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs so that we can meet the goals of each workload cost-effectively, and to deal with system failures, which occur more frequently as we operate on larger and larger

The other dramatic shift that is underway is the transformative change of the traditional I/O subsystem. For many decades, hard disk drives (HDDs) were used to store persistent data. HDDs had far slower random IO performance than sequential IO performance, and data processing engines formatted their data and designed their query processing

methods to "work around" this limitation. But, HDDs are increasingly being replaced by solid state drives today, and other technologies such as Phase Change Memory are around the corner.

3) Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data

There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

4) Privacy

The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

There are many additional challenging research problems. For example, we do not know yet how to share private data while limiting disclosure and ensuring sufficient data utility in the shared data. The existing paradigm of differential privacy is a very important step in the right direction, but it unfortunately reduces information content too far in order to be useful in most practical cases. In addition, real data is not static but gets larger and changes over time; none of the prevailing techniques results in any useful content being released in this scenario. Yet another very important direction is to rethink security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

5) Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Indeed, CAPTCHAs exploit precisely this fact to tell human web users apart from computer programs. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. There is similar value to human input at all stages of the analysis pipeline.

In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. Wikipedia, the online encyclopedia, is perhaps the best known example of crowd-sourced data. We are relying upon information provided by strangers. Most often, what they say is correct. However, we should expect there to be individuals who have other motives and abilities – some may have a reason to provide false information in an intentional attempt to mislead. While mostsuch errors will be detected and corrected by others in the crowd, we need technologies to facilitate this. We also need a framework to use in analysis of such crowd-sourced data with conflicting statements. As humans, we can look at reviews of a restaurant, some of which are positive and others critical, and come up with a summary assessment based on which we can decide whether to try eating there.

IV. CONCLUSION

This is an age of big data and the emergence of this field of study has attracted the attention of many practitioners and researchers. Moreover, most of this data is already on the cloud. Therefore, shifting big data analytics to the cloud framework is a viable option. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFERENCES

[1] Agarwal, D., Das, S. and Abbadi, A. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM 978-1-4503-0528-0/11/0003.

[2] Manekar, A. and Pradeepini, G. (2015). A Review on Cloud-based Big Data Analytics. ICSES Journal on Computer Networks and Communication (IJCNC), May 2015, Vol. 1, No, 1.

[3] Neaga, I. and Hao, Y. (2014). A Holistic Analysis of Cloud Based Big Data Mining. International Journal of Knowledge, Innovation and Entrepreneurship. Volume 2 No. 2, 2014

[4] Chen, C. L. P. and Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.

[5] Alam, M., &Shakil, K. A. (2013). Cloud Database Management System Architecture. UACEE International Journal of Computer Science and its Applications, 3(1), 27-31.