The Analytics of Clouds and Big Data Computing

Dr.E.Kesavulu Reddy

Assistant Professor, Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh-India-517502

Abstract: *Knowledge Discovery in Data (KDD) aims to extract non obvious information using careful and detailed analysis and interpretation. 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. Cloud computing is a versatile technology that can support a wide range of applications. The implementation of data mining techniques based on Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse which can reduces the costs of infrastructure and storage. Data Mining can retrieve the useful and potential information from the cloud. Big Data is usually defined by three characteristics called 3Vs (Volume, Velocity and Variety). It refers to data that are too large, dynamic and complex. In this context, data are difficult to capture, store, manage, and analyze using traditional data management tools. This paper 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.*

Keywords: *Data Mining, Data Management, Cloud Computing, Big Data*.

I. Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviour, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Today, one of the biggest challenges that institutions/organizations face is the explosive growth of data andto use this data to improve the quality of managerial decisions. Many institutions/organizations have their ownnetwork existing within their own locations. These networks can be expanded to enhance the access to theinformation resources using Data Mining and Cloud Computing technology [1]. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1].

Thecore functionalities of data mining are applying various methods and algorithms in order to discover and extractpatterns of stored data [2]. The field of data mining have been prospered and posed into new areas of human life with variousintegrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization,Artificial Intelligence and Computation capabilities etc. The various application areas of data mining are LifeSciences (LS), Customer Relationship Management (CRM), Web Applications, Manufacturing, CompetitiveIntelligence, Teaching Support, Climate modeling, Astronomy, and Behavioural Ecology etc.

A.Data Mining Parameters

- Association - Looking for patterns where one eventis connected to another event.
- Sequence or path analysis - Looking for patterns whereone event leads to another later event
- Classification - Looking for new patterns
- Clustering - Finding and visually documentinggroups of facts not previously known
- Forecasting - Discovering patterns in data that can leadto reasonable predictions about the future. This area of datamining is known as predictive analytics.

II. What Is Cloud Computing

Cloud computing is a general term for anything thatinvolves delivering hosted services over the Internet. Theseservices are broadly divided into three categories: Infrastructure-as-a- Service (IaaS), Platform-as-a-Service (PaaS) and Software-as- a-Service (SaaS). The name cloudcomputing was inspired by the cloud symbol that's oftenused to represent the Internet in flowcharts and diagrams.The term "cloud" is used as a metaphor for the Internet, basedon the cloud drawing used in the past to represent thetelephone network, The actual term "cloud" borrows fromtelephony in that telecommunications companies, who untilthe 1990s offered primarily dedicated point-to-point datacircuits, began offering Virtual Private Network(VPN)services with comparable quality of service but at a muchlower cost.

Cloud computing is becoming one of the next industry buzzwords. It joins the ranks of terms including: grid computing, utility computing, virtualization, clering, etc. Cloud computing overlaps some of the concepts ofdistributed, grid and utility computing, however it does haveits own meaning if contextually used correctly. Theconceptual overlap is partly due to technology changes,usages and implementations over the years. The cloud is avirtualization of resources that maintains and manages itself.Cloud computing really isaccessing resources and services needed to perform functionswith dynamically changing needs. An application or servicedeveloper requests access from the cloud rather than aspecific endpoint or named resource.

A. Cloud Services

There are three types of cloud services infrastructure as aService, platform as a Service, Software as a Service. Inwhich SaaS is king of all the services.

B.IaaS

Delivers computer infrastructure as a utilityservice, typically in a virtualized environment.Provides enormous potential for extensibility andscale. Major players in this filed are Amazon's EC2,Google App Engine etc.

C.PaaS

It provides a platform or solution stack on acloud infrastructure. Sits on a top of the IaaSarchitecture and integrates with development andmiddleware capabilities as well as database, messaging and queuing functions. Examples areForece.com offered by Salesforce.com

D.SaaS

It provides the application over the Internetor Intranet via a cloud Infrastructure. Built onunderlying IaaSand PaaS Layer. Examples are theelectronic mails that we are using today.

III. Data Management

The common phases of a traditionalanalytics workflow for Big Data shown below in the figure 1. 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.

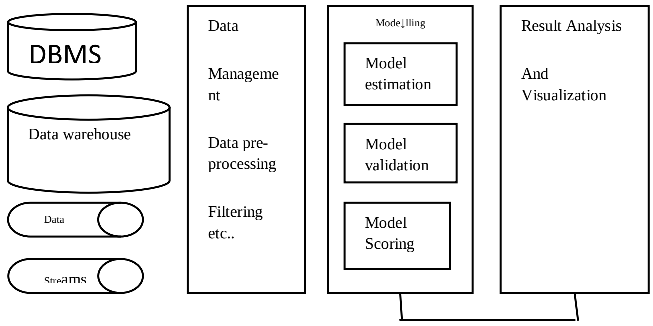

Fig.1. Overview of the analytics workflow for big data.

Analytics solutions can be classified as descriptive, predictive, or prescriptive.Descriptive analytics uses historical data to identify patterns and create management reports, it is concerned with modeling past behaviour. Predictive analytics attempts to predict the future by analyzing current and historical data. Prescriptive solutions assist analysts in decisions by determining actions and assessing their impact regarding business objectives, requirements, and constraints[3].

Performing analytics onlarge volumes of data requires efficient methods to store, filter, transform,and retrieve the data. Some of the challenges of deploying data managementsolutions on Cloud environments have been known for some time [14, 15,16], and solutions to perform analytics on the Cloud face similar challenges.Cloud analytics solutions need to consider the multiple Cloud deploymentmodels adopted by enterprises, where Clouds can be for instance. Private deployed on a private network, managed by the organizationitself or by a third party. A private Cloud is suitable for businessesthat require the highest level of control of security and data privacy. Insuch conditions, this type of Cloud infrastructure can be used to sharethe services and data more efficiently across the different departmentsof a large enterprise.

Public Cloud offers high efficiency and shared resources withlow cost. The analytics services and data management are handledby the provider and the quality of service (e.g. privacy, security, andavailability) is specified in a contract. Organizations can leverage theseClouds to carry out analytics with a reduced cost or share insights ofpublic analytics results.Hybrid cloud combines both Clouds where additional resources from a publicCloud can be provided as needed to a private Cloud. Customers candevelop and deploy analytics applications using a private environment,thus reaping benefits from elasticity and higher degree of security thanusing only a public Cloud.Considering the Cloud deployments, the following scenarios are generallyenvisioned regarding the availability of data and analytics models [16]

A. Big Data

With the advent of social network Web sites, users create records of their lives bydaily posting details of activities they perform, events they attend, placesthey visit, pictures they take, and things they enjoy and want. This datadeluge is often referred to as Big Data [4, 5, 6]; a term that conveys thechallenges it poses on existing infrastructure in respect to storage, management,interoperability, governance, and analysis of the data. Big Data is characterized by what is often referred to as a multi-V model. Variety represents the data types, velocity refers tothe rate at which the data is produced and processed, and volume defines theamount of data. Veracity refers to how much the data can be trusted given the reliability of its source [2], whereas value correspond the monetary worth that a company can derive from employing Big Data computing. Althoughthe choice of Vs used to explain Big Data is often arbitrary and varies acrossreports and articles on the Webvariety, velocity, and volume [18, 19] are the items most commonlymentioned.

In [23] present architecture that integratesmonitoring and analytics.Increasingly often, data arriving via streams needs to be analysed and compared against historical information. Different data sources may usetheir own formats, which makes it difficult to integrate data from multiplesources in an analytics solution. As highlighted in existing work [24], standardformats and interfaces are crucial so that solution providers can benefit fromeconomies of scale derived from data integration capabilities that address theneeds of a wide range of customers.

B.Data Storage

Several solutions were proposed to store and retrieve large amounts ofdata demanded by Big Data, some of which are currently used in Clouds. Internet-scalable systems such as the Google File System (GFS) [25] attemptto provide the robustness, scalability, and reliability that certain Internetservices need. Other solutions provide object-store capabilities where filescan be replicated across multiple geographical sites to improve redundancy, scalability, and data availability. One key aspect in providing performance for Big Data analytics applicationsis the data locality. This is because the volume of data involved inthe analytics makes it prohibitive to transfer the data to process it. Thiswas the preferred option in typical high performance computing systemsin the context of Big Data, this approach of movingdata to computation nodes would generate large ratio of data transfer time to processing time. Thus, a different approach is preferred, where computationis moved to where the data is. The same approach of exploring data localitywas explored previously in scientific work flows [26] and in Data Grids [27].In the context of Big Data analytics, MapReduce presents an interestingmodel where data locality is explored to improve the performance of applications.Among the drawbacks of Cloud storage techniques and MapReduce implementations,there is the fact that they require the customer to learn anew set of APIs to build analytics solutions for the Cloud. To minimize thishurdle, previous work has also investigated POSIX-like file systems for dataanalytics.

Although a large part of the data produced nowadays is unstructured,relational databases have been the choice most organisations have made tostore data about their customers, sales, and products, among other things.As data managed by traditional DBMS ages, it is moved to data warehousesfor analysis and for sporadic retrieval.Data processing and analytics capabilitiesare moving towards Enterprise Data Warehouses (EDWs), or arebeing deployed in data hubs [17] to facilitate reuse across various data sets.In respect to EDW, some Cloud providers offer solutions that promise toscale to one pentbyte of data or more. Amazon Redshift [31], for instance, offers columnar storage and data compression and aims to deliver high queryperformance by exploring a series of features, including a massively parallelprocessing architecture using high performance hardware, mesh networks, locallyattached storage, and zone maps to reduce the I/O required by queries.Amazon Data Pipeline [21] allows a customer to move data across differentAmazon Web Services, such as Elastic MapReduce (EMR) [33] and DynamoDB[34], and hence compose the required analytics capabilities.

Another distinctive trend in Cloud computing is the increasing using ofNoSQL databases as the preferred method for storing and retrieving information.Han et al. [35] presented a survey of NoSQL databases with emphasis ontheir advantages and limitations for Cloud computing. The survey classifiesNoSQL systems according to their capacity in addressing different pairs ofCAP (consistency, availability, partitioning). The survey also explores thedata model that the studied NoSQL systems support.

C.Data Integration Solutions

Research published a technical report that discusses some of theproblems that traditional Business Intelligence (BI) faces [30], highlightingthat there is often a surplus of soiled data preparation, storage, and processing.Authors of the report envision some data processing and Big Dataanalytics capabilities being migrated to the EDW, hence freeing organizationsfrom unnecessary data transfer and replication and the use of disparate dataprocessingand analysis solutions.EDWs or Cloud based data warehouses, however, create certain issues inrespect to data integration and the addition of new data sources. Standardformats and interfaces can be essential to achieve economies of scale andmeet the needs of a large number of customers [24]. Some solutions attemptto address some of these issues [20, 36].

In [36] provides Software as a Service (SaaS) solution that offers analyticsfunctionalities on a subscription model; and appliances with the businessanalytics infrastructure, hence providing a model that allows a customer tomigrate gradually from an on-premise analytics to a scenario with Cloudprovidedanalytics infrastructure. To improve the market penetration ofanalytics solutions in emerging markets such as India, in [37] propose a multi flow solution for analytics that can be deployed on the Cloud.The multi flow approach provides a range of possible analytics operators andflows to compose analytics solutions; viewed as work flows or instantiations ofa multi flow solution.

D.Data Processing and Resource Management

MapReduce [38] is one of the most popular programming models to processlarge amounts of data on clusters of computers. In [63] is the mostused open source MapReduce implementation, also made available by severalCloud providers [33, 40, 41, 42]. In [33] enables customers to instantiate Hadoop clusters to process large amounts of data using theAmazon Elastic Compute Cloud (EC2) and other Amazon Web Services fordata storage and transfer.Daytona [40], a MapReduce runtime for Windows Azure, leverages thescalable storage services provided by Azure's Cloud infrastructure as thesource and destination of data. It uses Cloud features to provide load balancingand fault tolerance. The system relies on a master-slave architecture where the master is responsible for scheduling tasks and the slaves for carrying out map and reduces operations. A hybrid Cloud is used tospeed up the application execution. Other characteristics of the applicationare security features and cost-effective exploration of Cloud resources.

E. Challenges in Big Data Management

In this section, we discussed current research targeting the issue of BigData management for analytics. There are still, however, many open challengesin this topic.

- Data storage: How to efficiently recognize and store important informationextracted from unstructured data? How to store large volumesof information in a way it can be timely retrieved? How to storeinformation in a way that it can be easily migrated/ported betweendata centres/Cloud providers?

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

- DataProcessing and Resource Management: New programming modelsoptimized for streaming and/or multidimensional data; new backendengines that manage optimised file systemse.g. Map Reduce, work flows, and bag-of-tasks on a single solution/abstraction. How to optimise resource usage and energy consumption when executing theanalytics application?

IV. Open Challenges

There are many research challenges in the yield of Big Data visualization.First, more efficient data processing techniques are required in order to enablereal-time visualization. Some techniquesthat can be employed with this objective, such as reduction of accuracy ofresults, coarsely processing of data points, compatible with the resolution ofthe visualization device, reduced convergence, and data scale confinement.Methods considering each of these techniques could be further researchedand improved. A cost-effective device for large-scale visualization is another hot topic foranalyticsvisualization, as they enable finer resolution than simple screens.
Visualization for management of computer networks and software analytics are also areas that are attracting attention of researchers and practitionersfor its extreme relevance to management of large-scale infrastructure(such as Clouds) and software, with implications in global software development,open source software development, and software quality improvements.

V. Conclusion

The Big Data trend is being seen by industries as a way of obtainingadvantage over their competitors: if one business is able to make sense of theinformation contained in the data reasonably quicker, it will be able to getmore costumers, increase the revenue per customer, optimize its operation,and reduce its costs. Nevertheless, Big Data analytics is still a challengingand time demanding task that requires expensive software, large computationalinfrastructure, and effort.Cloud computing helps in alleviating these problems by providing resourceson-demand with costs proportional to the actual usage. Cloud infrastructure offers such elastic capacity to supply computationalresources on demand, the area of Cloud-supported analytics is still in its early days.Cloud computing plays a key role for Big Data; not only because it providesinfrastructure and tools, but also because it is a business modelthat Big Data analytics can follow (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)).

Authors Information

Dr. E. Kesavulu Reddy

I am Dr. E. Kesavulu Reddy and work as an Assistant Professor in Dept. of.Computer Science, Sri Venkateswara University College of Commerce Management and Computer Science, Tirupati (AP)-India. My research areas of interest in the field of Computer Science are Elliptic Curve Cryptography- Network Security, Data Mining, and Neural Networks.

References

[1] F. Schomm, F. Stahl, G. Vossen, Marketplaces for Data: An Initial Survey, SIGMOD Record 42 (1) (2013)
[2] P. S. Yu, On Mining Big Data, in: J. Wang, H.Xiong, Y. Ishikawa,J. Xu, J. Zhou (Eds.), Web-geInformation Management, Vol. 7923,Lecture Notes in Computer Science, Springer-Verlag, Berlin,Heidelberg,2013.
[3] X. Sun, B. Gao, Y. Zhang, W. An, H. Cao, C.Guo, W. Sun, Towards Delivering Analytical Solutions in Cloud: Business Models andTechnical Challenges, in: Proceedings of the IEEE 8th InternationalConference on e-Business Engineering (ICEBE 2011),pp 347-351, IEEE ComputerSociety, Washington,USA, 2011,
[4] A. McAfee, E. Brynjolfsson, Big Data: The Management Revolution, Harvard Business Review, pp 60- 68,2012.
[5] B. Franks, Taming The Big Data Tidal Wave:Finding Opportunitiesin Huge Data Streams with Advanced Analytics, 1st Edition, Wileyand SAS Business Series, Wiley, 2012.
[6] G. Bell, T. Hey, A. Szalay, Beyond the Data Deluge, Science 323 (5919),pp 1297-1298,2009.
[7] T. H. Davenport, J. G. Harris, R. Morison, Analytics atWork: Smarter Decisions, Better Results, Harvard Business Review Press, 2010.
[8] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD Process for Extracting Useful Knowledge fromVolumes of Data, Communicationsof the ACM 39 (11), pp 27-34,1996.
[9] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, 2011.
[10] E. A. King, How to Buy Data Mining: A Framework for AvoidingCostly Project Pitfalls in PredictiveAnalytics, DMReview 15 (10).
[11] T. H. Davenport, J. G. Harris, Competing on Analytics: The New Science of Winning, Harvard Business Review Press, 2007.
[12] R. L. Grossman, What is Analytic Infrastructure and Why Should You Care?, ACM SIGKDD Explorations Newsletter 11 (1),5-9,2009.
[13] D. J. Abadi, Data Management in the Cloud: Limitations and Opportunities,IEEE Data Engineering Bulletin 32 (1),3-12,2009.
[14] S. Sakr, A. Liu, D. Batista, M. Alomari, A Survey of Large Scale Data Management Approaches in Cloud Environments, IEEE CommunicationsSurveys Tutorials 13 (3),311-336,2011.
[15] D. S. Katz, S. Jha, M. Parashar, O. Rana, J. B. Weissman, Survey and Analysis of Production Distributed Computing Infrastructures, CoRR abs/1208.2649.
[16] P. R. Krishna, K. I. Varma, Cloud Analytics: A Path Towards Next Generation Affordable BI, hite paper, Infosys ,2012.
[17] D. Jensen, K. Konkel, A. Mohindra, F. Naccarati, E. Sam, Business Analytics in the Cloud,White paper IBW03004-USEN-00, IBM, April2012.
[18] P. Russom, Big Data Analytics, TDWI best practices report, The Data Warehousing Institute(TDWI) Research ,2011.
[19] P. Zikopoulos, C. Eaton, P. Zikopoulos, Understanding Big Data: Analyticsfor Enterprise ClassHadoop and Streaming Data, McGraw-Hill Companies, Inc., 2012.
[20] PivotLinkAnalyticsCLOUD.
[21] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T.Do, O. Dousse, J. Eberle, M.Miettinen, The Mobile Data Challenge:Big Data for Mobile Computing Research, 2012.
[22] A. Iosup, A. Lascateu, N. Tapus, CAMEO: Enabling social networksfor Massively Multiplayer Online Games through Continuous Analyticsand Cloud Computing, in: Proceedings of the 9thAnnual Workshopon Network and Systems Support for Games pp 1-6, 2010.
[23] C. Wang, K. Schwan, V. Talwar, G. Eisenhauer, L. Hu, M. Wolf, A Flexible Architecture Integrating Monitoring and Analytics for Managing Large-Scale Data Centers, in: Proceedings of the 8th ACM International Conference on Autonomic Computing (ICAC 2011), pp 141-150,New York, USA, 2011
[24] D. Fisher, R. DeLine, M. Czerwinski, S. Drucker, Interactions with Big Data AnalyticsInteractions 19 (3),pp 50-59,2012.
[25] S.Ghemawat, H. Gobio_, S.-T. Leung, The Google File System, in: Proceedings of the 9th ACM Symposium on Operating Systems Principles,pp 29-43, ACM, New York, USA, 2003.
[26] E. Deelman, A. Chervenak, Data management challenges of data intensive scientific work flows, in: Proceedings of the 8th IEEE International Symposium on Cluster Computing and the GridIEEE Computer Society, pp 687-692,2008.
[27] S. Venugopal, R. Buyya, K. Ramamohanarao, A taxonomy of datagrids for distributed datasharing, management and processing, ACM Computing Surveys 38(1),pp1-53,2006.
[28] R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar,M. Shah, R. Tewari, Cloud Analytics: Proceedings of the Conference onHotTopics in Cloud Computing), USENIX Association, Berkeley, USA, 2009.
[28] R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar,M. Shah, R. Tewari, Cloud Analytics: Proceedings of the Conference onHot Topics in Cloud Computing), USENIX Association, Berkeley, USA, 2009.
[29] F. Schmuck, R. Haskin, GPFS: A Shared-Disk File System for Large Computing Clusters, in: Proceedings of the 1st Conference on File and Storage Technologies (FAST'02), Monterey, Pp 231-244, USA, 2002.
[30] J. Kobielus, In-Database Analytics: The Heart of the Predictive Enterprise,Technical report, Forrester Research, Inc., Cambridge, USA, Nov, 2009.
[31] Amazon red shift.
[32] Amazon data pipeline.
[33] Amazon Elastic MapReduce (EMR).
[34] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels, Dynamo: Amazon's Highly Available Key-Value Store, SIGOPS Operating Systems Review 41 (6) (2007) 205{220.
[35] J. Han, H. E, G. Le, J. Du, Survey on NoSQL database, in the 6th International Conference on Pervasive Computing and Applications (ICPCA2011), IEEE,pp 363-366,South Africa, 2011.
[36] Birst Inc., http://www.birst.com.
[37] P. Deepak, P. M. Deshpande, K. Murthy, Configurable and Extensible Multi-flows for ProvidingAnalytics as a Service on the Cloud, in the Proceedings of the 2012 Annual SRII Global Conference (SRII 2012),pp 1-10,2012.
[38] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters Communications of the ACM 51 (1).
[39] Apache Hadoop, http://hadoop.apache.org.
[40] R. S. Barga, J. Ekanayake, W. Lu, Project Daytona: Data Analytics as a Cloud Service, in: A.Kementsietsidis, M. A. V. Salles (Eds.),
Proceedings of the International Conference of Data Engineering (ICDE 2012), IEEE Computer Society,pp 1317-1320. 2012.
[41] Info chimps cloud overview.
[42] Windows Azure HD Insight. Kementsietsidis, M. A. V. Salles (Eds.), Proceedings of the International Conference of Data Engineering (ICDE 2012), IEEE Computer Society,pp 1317-1320. 2012.
[41] Info chimps cloud overview.
[42] Windows Azure HD Insight.