CHALLENGES OF BIG DATA APPLICATIONS IN CLOUD COMPUTING

Manoj Muniswamaiah, Dr. Tilak Agerwala and Dr. Charles Tappert

Seidenberg School of CSIS, Pace University, White Plains, New York

ABSTRACT

Big Data applications are used for decision making process for gaining useful insights hidden from large volume of data. They make use of cloud computing infrastructure for massive scale and complex computation which eliminates the need to maintain dedicated hardware and software resources. The relationship between big data and cloud computing is presented with focus on challenges and issues in data storage with different formats, data transformation techniques applied, data quality and business challenges associated with it. Also, some good practices which helps in big data analysis has been listed.

KEYWORDS

Big data; cloud computing; data transformation; data analysis; data warehousing

1. INTRODUCTION

The volume and information captured from devices and multimedia by organizations is increasing and has almost doubled every year. This big data generated is characterized to be huge, can be structured or unstructured which requires pre-processing and cannot be easily loaded into regular relational databases. Healthcare, finance, engineering, e-commerce and various scientific fields use these data for decision making and analysis. The advancement in data science, data storage and cloud computing has allowed for storage and mining of big data [1].

Cloud computing has resulted in increased parallel processing, scalability, virtualization of resources and integration with data storages. Cloud computing has also reduced the infrastructure cost required to maintain these resources which has resulted in the scalability of data produced and consumed by the big data applications. Cloud virtualization provides the process to share the resources and isolation of hardware to increase the access, management, analysis and computation of the data [1].

The main objective of this paper is to provide challenges and issues of big data applications in cloud computing which requires data to be processed efficiently and provide some good design principles.

2. BIG DATA

Data which is difficult to store, manage and analyse through traditional databases is termed as "Big Data". It requires integration of various technologies to discover hidden values from the data that is varied, complex and requires heavy computing. The characteristics of big data are.

Natarajan Meghanathan et al. (Eds) : CCSEA, CLOUD, SIPRO, AIFU, SEA, DKMP, NCOM - 2019
pp. 221-232, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90918

1) Volume - Collection of data from different sources which would allow users to data mine the hidden information and patterns found in them.
2) Velocity - Data been streamed in real time from sources such as IoT devices. It is the speed at which data is transferred and consumed for collection and archiving.
3) Variety - Data collected in either structured or unstructured format from sensors and social networks. Unstructured data include text messages, audio, blogs.
4) Variability - Data flow can be highly inconsistent and varies during peak period and their ingestion into the data stores.
5) Value - Represents the hidden value discovered from the data for decision making.
6) Veracity - It refers to the reliability of the data source. Its importance is in the context and the meaning it adds to the analysis.
7) Validity - It refers to the accuracy of the data been collected for its intended use.
8) Vulnerability - It represents the security aspects of the data been collected and stored.
9) Volatility - How long the data needs to be stored historically before it is considered irrelevant.
10) Visualization - In-memory tools which are used to plot data points representing as data clusters or tree map [2].

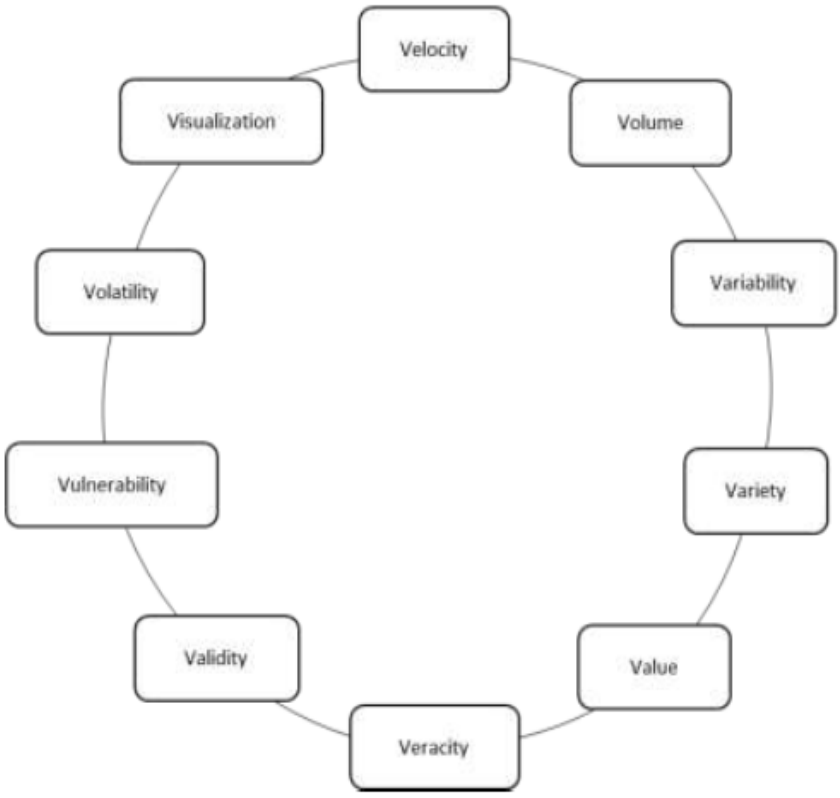

Figure 1: V's of Big Data

3. BIG DATA CLASSIFICATION

Analysis Type - Whether the data is analysed in real time or batch process. Banks use real time analysis for fraud detection whereas business strategic decisions can make use of batch process.

Processing Methodology - Business requirements determine whether predictive, ad-hoc or reporting methodology needs to be used.

Data Frequency - Determines how much of data is ingested and the rate of its arrival. Data could be continues as in real-time feeds and also time series based.

Data Type - It could be historical, transactional and real-time such as streams.

Data Format - Structured data such as transactions can be stored in relational databases.

Unstructured and semi-structured data can be stored in NoSQL data stores. Formats determine the kind of data stores to be used to store and process them.

Data Source - Determines from where the data is generated like social media, machines or human generated.

Data consumers - List of all users and applications which make use of the processed data [3].

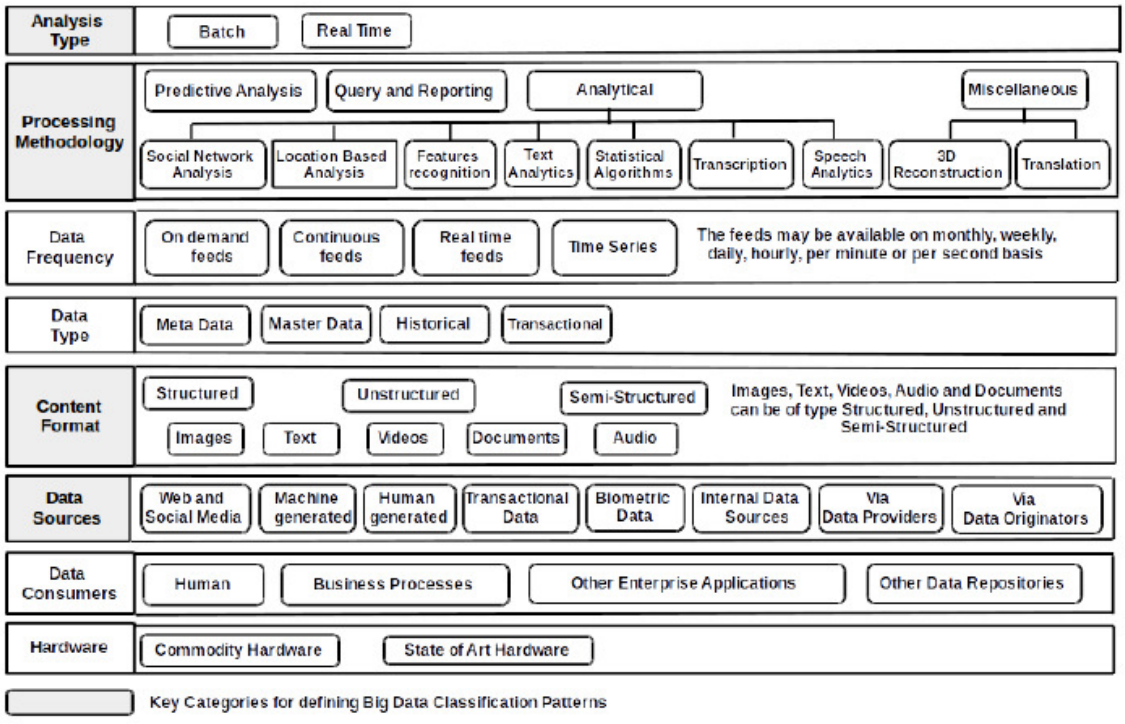

Figure 2: Big Data Classification

Big data is classified based upon its source, format, data store, frequency, processing methodology and analysis types as shown in Figure 2.

4. CLOUD COMPUTING

Cloud computing has become default platform for storage, computation, application services and parallel data processing. It allows organizations to concentrate on core business without having to worry about the infrastructure, maintenance and availability of the resources. Figure 3 shows the differences between on premise and cloud services. It shows the services offered by each computing layer and differences between them.

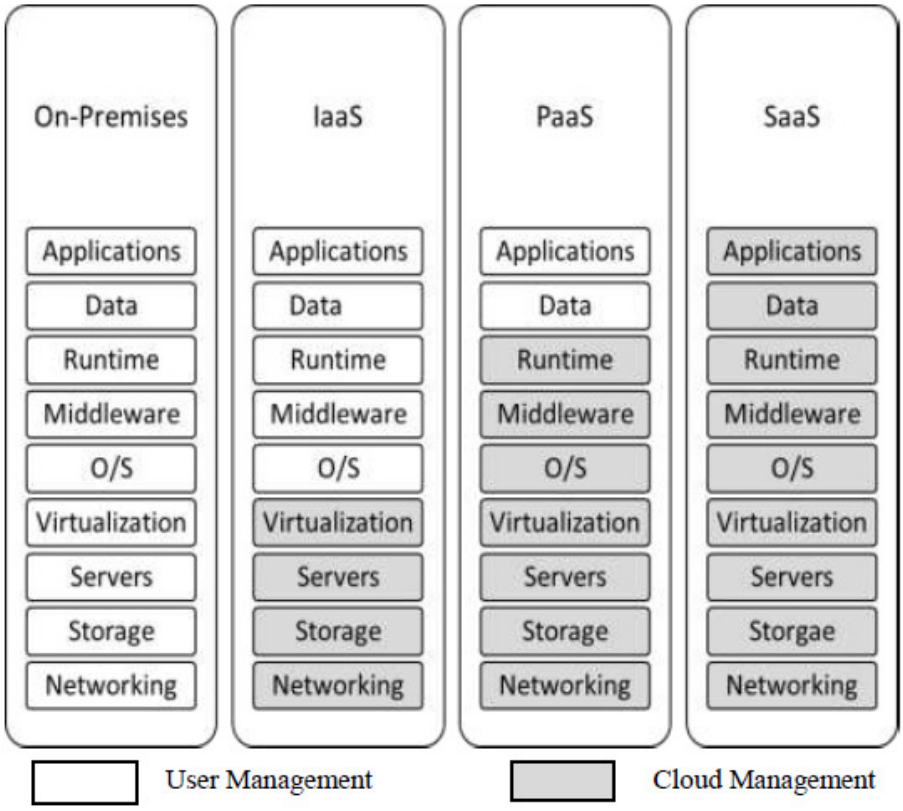

Figure 3: Summary of Key Differences

SaaS: Software as a Service

Software as a service represents the most commonly used business option in cloud services. It uses the internet to deliver applications to users. It does not require installations on client side as they run directly through web. In SaaS vendor manages all the servers, middleware and storage of the data. It eliminates users to install, manage and upgrade softwares.

PaaS: Platform as a Service

Platform as a Service model is been used by developers to build applications. It allows business to design and create applications that are integrated in to PaaS software components. These applications are scalable and highly available since they have cloud characteristics.

IaaS: Infrastructure as a Service

Infrastructure as a Service cloud computing model provides servers, storage, operating systems to organizations through virtualization technology. IaaS provides same capabilities as data centers without having to maintain them physically [4]. Figure 4 represents the different cloud computing services been offered.

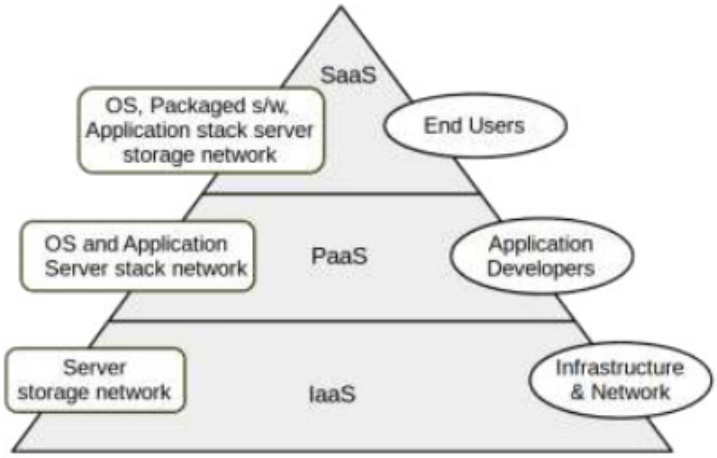

Figure 4: Primary Cloud Computing Services

5. RELATIONSHIP BETWEEN THE CLOUD AND BIG DATA

Cloud computing and big data go together, as cloud provides the required storage and computing capacity to analyse big data. Cloud computing also offers the distributed processing for scalability and also expansion through virtual machines to meet the requirements of exponential data growth. It has resulted in the expansion of analytical platforms. This has resulted in service providers like Amazon, Microsoft and Google in offering big data systems in cost efficient manner.

Cloud computing environment has several providers and user terminals. Data is collected using big data tools later it is stored and processed in cloud. Cloud provides on-demand resources and services for uninterrupted data management. The most common models for big analytics is software services such as (SaaS), Platform service like (PaaS) and Infrastructure service like (IaaS). Recently Cloud analytics and Analytics as a Service (AaaS) are provided to clients on demand. Analytics as a Service (AaaS) provides services for a fast and scalable way to integrate data in semi-structured, unstructured and structured format, transform and analyse them.

Virtualization simulates a virtual computing environment that can run operating system and applications on it. Virtualization reduces the workload and unifies them in to a physical server which helps in consolidation of multi-core CPUs in to one physical node. This reduces and improves resource utilization and power consumption as compared to the multi-node setup. Virtualized big data applications like Hadoop provide benefits which cannot be provided using physical infrastructure in terms of resources utilization, cost and data management. Virtual data includes wide range of data sources and improves the data access from heterogeneous environments. It also enables high-speed data flow over the network for faster data processing.

Information privacy and security are one of the important aspects of big data in cloud as data is hosted and processed on the third party services and infrastructure. Service level agreements must be maintained between providers and consumers in order to bring confidence in users. Security of big data in the cloud is important because data needs to be protected from malicious intruders, treats and also how the cloud providers securely maintain huge disk space [5].

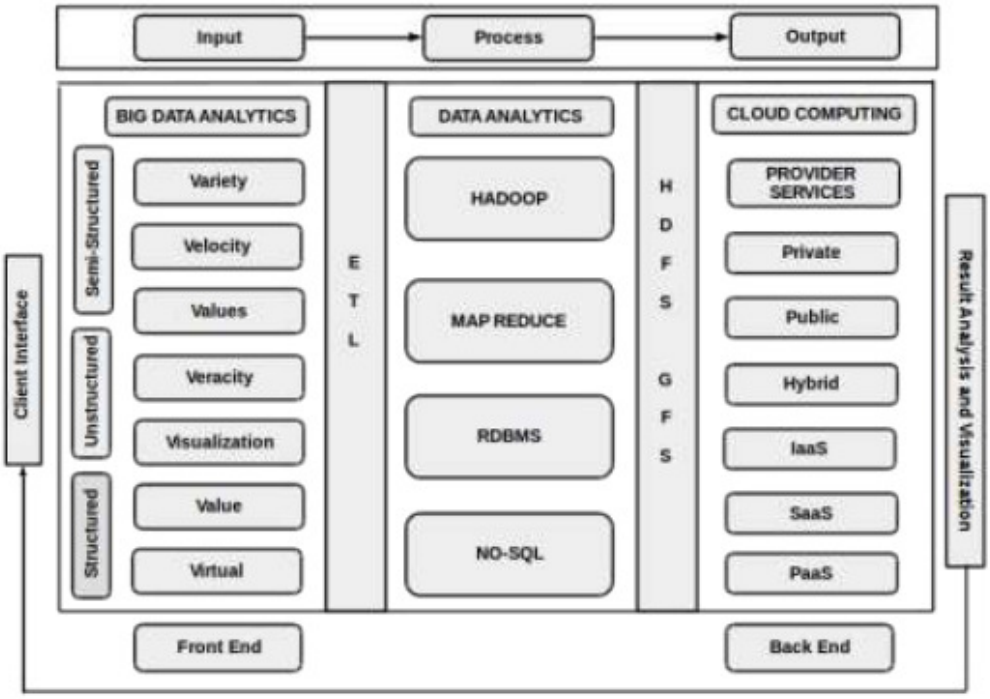

Figure 5: Big Data and Cloud Computing

The relationship between big data and cloud computing follows input, processing and output model as shown in Figure 5. The input is the data obtained from various data sources and are processed and stored using Hadoop and data stores. Processing steps includes all the tasks required to transform input data. Output is the result obtained after data been processed for analysis and visualization. Internet of Things (IoT) is one of the common factors between Cloud computing and big data. Data generated from IoT devices needs to be analysed in real time. Cloud providers allow data to be transmitted over internet or via lease lines. It provides a pathway for the data to navigate, store and be analyzed. Cloud computing provides common platform for IoT and big data. IoT is the source of the data and big data is an analytical technology platform of the data as depicted in the Figure [6].

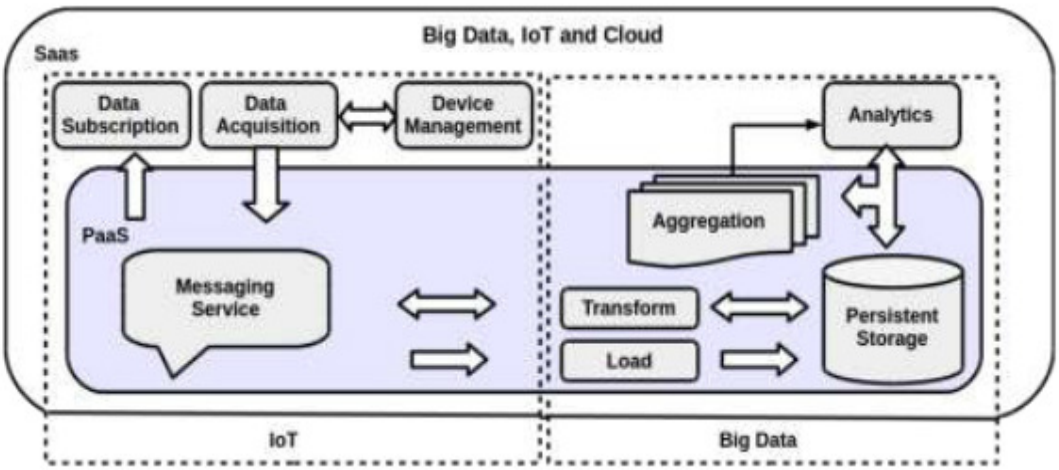

Figure 6: Overview of IoT, Big Data processing and Cloud Computing

6. CASE STUDIES

There are several case studies of big data on cloud computing.

A. Redbus

Redbus is an online travel agency for bus ticket booking in India. Redbus decided to use Google data infrastructure for data processing and analysis in order to improve customer sales and management of the ticket booking system [6].

B. Nokia mobile company

Nokia mobile phones are been used by many people for telecommunication. Nokia gathers large amount of data from mobile phones in petabyte scale for business decision strategies using Hadoop data warehouse for analytics [6].

C. Tweet Mining in Cloud

Noordhuis et al. [6] used cloud computing to gather and analyse tweets. Amazon cloud infrastructure was used to perform all the computations. Tweets were crawled and later page ranking algorithm was applied. The data crawled had nearly 50 million nodes and 1.8 billion edges.

7. DATA STORES

Modern databases needs to handle large volume and different variety of data formats. They are expected to deliver extreme performance and scale both horizontally and vertically. Database architects have produced NoSQL and NewSQL as alternatives to relational database. Below are characteristics of relational database, NoSQL and NewSQL [7].

Characteristics of Databases	Relational Database	NoSQL Database	New SQL Database
ACID property	✓	×	✓
Analytical and OLTP support	✓	×	✓
Data analysis	✓	×	✓
Requires Schema	✓	×	×
Data format support	×	✓	×
Distributed parallel processing	✓	✓	✓
Scalability	×	✓	✓

8. HADOOP TOOLS AND TECHNIQUES

Big data applications use various tools and techniques for processing and analyses of the data below table represents some of them [8].

Tools/Techniques	Description	Developed by	Written in
HDFS	Redundant and Reliable massive data storage	Introduced by Google	Java
Map Reduce	Distributed data processing framework	Introduced by Google	Java

YARN	Cluster resource management framework	Apache	Java
Storm	Stream based task parallelism	Twitter	Clojure
Spark	Stream based data parallelism	Berkeley	Scala
Map Reduce	Java API.	Introduced by Google	Java
Pig	Framework to run script language Pig Latin	Yahoo	Java
Hive	SQL-like language HiveQL	Facebook	Java
HCatalog	Relational table view of data in HDFS	Apache	Java
HBase	NoSQL column oriented Google's	BigTable	Java
Casandra	NoSQL column oriented	Facebook	Java
Flume	Import/Export unstructure or semi-structure data into HDFS. Data ingestion into HDFS.	Apache	Java
Sqoop	Tool designed for efficiently transferring bulk structured data (RDBMS) into HDFS and vies versa.	Apache	Java
Kafka	Distributed publish-subscribe messaging system for data integration	LinkedIn	Scala
Ambari	Web based cluster management UI	Hortonworks	Java
Mahout	Library of machine learning algorithms	Apache	Java
Oozie	Define collection of jobs with their execution sequence and schedule time	Apache	Java
Sentry	Role based authorization of data stored on an Apache Hadoop cluster.	Cloudera	Java
Zookeeper	Coordination service between hadoop ecosystems.	Yahoo	Java

9. RESEARCH CHALLENGES

Big data can be stored, processed and analysed in many different ways. The data generated has many attributes which results in different dimensions of data to come in to play. This gives rise to challenges in processing big data and business issues associated with it. Volume of the data been generated worldwide doubles almost every year. Retail industries do millions of translations per day and also have established data warehouses to store data to take advantages of machine learning techniques to get the insight of data which would help in the business strategies. Public administration sector also uses information patterns from data generated from different age levels of population to increase the productivity. Also, many of the scientific fields have become data

driven and probe into the knowledge discovered from these data. Although cloud computing is been used for processing of big data applications there are several challenges in data storage, data transformation, data quality, privacy, governance [9].

Data Capture and Storage

Data gathered from various sensor devices, machine logs and networks keeps increasing every year. It has changed the way we store data and their access mechanism. Previously, hard disk drives (HDD) had poor I/O performance but solid-state drives (SSD) may alleviate I/O performance to some extent but not completely.

Data Transmission

Cloud data stores are used for data storage however, network bandwidth and security poses challenges.

Data Curation

It involves data archiving, management and retrieval process. Structured data is stored in data warehouse and data marts which requires pre-processing of data before loading data and also can be queried using Standard Query Languages. Unstructured data is stored in NoSQL data stores which are schema free, support replication, distributed storage and consistency. There are various NoSQL data stores such as key-value, columnar, document and graph data stores which are specific to type of data which gets stored in them.

Scalability

Scalability is mainly manual and is static. Most of the big data systems must be elastic to handle data changes. At the platform level there is vertical and horizontal scalability.

Elasticity

Elasticity accommodates data peaks using replication, migration and resizing techniques. Most of these are manual instead being automated.

Availability

Availability refers to systems been available to users. One of the key aspect of cloud providers is to allow users to access one or more data services in short time even during security breach.

Data integrity

Data needs to be modified only by the authorized user or parties. Since the users may not be able to physically access the data, the cloud should provide mechanisms to check for the integrity of data.

Security and Privacy

Based on the service level agreement the data can be encrypted. But querying encrypted data would result in time consumption. User privacy can be de- identified, it's also been proved that de-identification can be reverse engineered.

Heterogeneity

Big data systems need to deal with different formats of data coming from various sources. Handling unstructured data during peak hours and processing them for analysis becomes a challenge.

Data Governance

Data governance specify the way data needs to be handled, data access policies have its life cycle. Defining the data cycle is not easy task and also its policies could lead to counter productiveness.

Data Uploading

Data is usually been uploaded through internet which is unsecure but results in time consumption if they are encrypted and transmitted.

Data Recovery

Specifies the procedures and locations from where the data can be recovered. Generally there is only one destination from where the data is securely recovered.

Data Visualization

Data Visualization is used to represent knowledge graphically for better intuition and understanding. It helps to analyse the data quickly.

10. BIG DATA BUSINESS CHALLENEGES

Utilities: Power consumption prediction

Utility companies use smart meter to measure gas and electricity consumption. These devices generate huge volumes of data. A big data solution needs to monitor and analyse power generation and consumption using smart meters.

Social Network: Sentiment analysis

Social networking companies such as Twitter needs to determine what users are saying and topics which are trending in order to perform sentiment analysis.

Telecommunication: Predictive analytics

Telecommunication provides need to build churn models which depends on the customer profile data attributes. Predictive analytics can predict churn by analysing the subscribers calling patterns.

Customer Service: Call monitor

Call center big data solutions use application logs to improve performance. The log files needs to be consolidated from different formats before they can be used for analysis.

Banking: Fraud Detection

Banking companies should be able to prevent fraud on a transaction or a user account. Big data solutions should analyse transactions in real time and provide recommendations for immediate action and stop fraud.

Retailers: Product recommendation

Retailers can monitor user browsing patterns and history of products purchased and provide a solution to recommend products based on it. Retailers need to make privacy disclosures to the users before implementing these applications [3].

11. GOOD PRINCIPLES

Below are some of the good design principles for big data applications

Good Architectural Design

Big data architecture should provide distributed and parallel processing through cloud services. NoSQL can be used for high performance and faster retrieval of data. Lambda and Kappa architectures can be used for processing in real-time and batch processing mode.

Different Analytical Methods

Big data applications need to take the advantage of data mining, machine learning, distributed programming, statistical analysis, in-memory analytics and visualization techniques offered through cloud.

Use appropriate technique

No one technique can be used to analyse data. We must use appropriate technology stack to analyse the data.

Use in-memory analytics

It is not advisable to move data around. In-memory database analytics can be used to execute analytics where data resides. In-memory analytics also provides real-time processing of data.

Distributed data storage for in-memory analytics

The data needs to be partitioned and stored in distributed data stores to take the advantage of in memory analytics. Cloud computing infrastructure offers this distributed data storage solutions which must be adopted.

Coordination between tasks and data is required

To achieve scalability and fault-tolerance coordination between data and its processing tasks is required. Specialized cluster management frameworks as a Zookeeper can be used [10].

12. CONCLUSION

In the big data era of innovation and competition driven by advancements in cloud computing has resulted in discovering hidden knowledge from the data. In this paper we have given an overview of big data applications in cloud computing and its challenges in storing, transformation, processing data and some good design principles which could lead to further research.

REFERENCES

[1] Konstantinou, I., Angelou, E., Boumpouka, C., Tsoumakos, D., & Koziris, N. (2011, October). On the elasticity of nosql databases over cloud management platforms. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2385-2388). ACM.

[2] Abadi, D. J. (2009). Data management in the cloud: Limitations and opportunities. IEEE Data Eng. Bull, 32(1), 3-12.

[3] Luhn, H. P. (1958). A business intelligence system. IBM Journal of Research and Development, 2(4), 314-319

[4] https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/

[5] https://www.ripublication.com/ijaer17/ijaerv12n17_89.pdf

[6] Sakr, S. & Gaber, M.M., 2014. Large Scale and big data: Processing and Management Auerbach, ed.

[7] Han, J., Haihong, E., Le, G., & Du, J. (2011, October). Survey on nosql database. In Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on (pp. 363-366). IEEE.

[8] Zhang, L. et al., 2013. Moving big datato the cloud. INFOCOM, 2013 Proceedings IEEE, pp.405–409

[9] http://acme.able.cs.cmu.edu/pubs/uploads/pdf/IoTBD_2016_10.pdf

[10] Labrinidis and Jagadish 2012, A. Labrinidis and H. Jagadish, Challenges and Opportunities with Big Data, In Proc. of the VLDB Endowment, 5(12):2032-2033, 2012