An Overview of Data Collection on HDFS for DA

Padma Lochan Pradhan, JSPM University Pune, India

Amol Rajmane, JSPM University Pune, India

Abstract

The author has to focus, take care and emphasize the step-by-step process of data collection from traditional to the real-time field of visualization. The author has to define, design, development, and deployment the source data sets in the cognitive field. In this paper, the author is focusing on the various stages of data collection and utilization in the area of data science to integrate Visualizations, Data Science, and Big Data with high-performance computing. How the role and responsibilities are supporting the system and integrating with heterogeneous subsystems and technologies to visualize the data in multi-disciplinary areas. We have to take action plans effectively and efficiently to integrate, automate, transparent, and communicate the cognitive data acquisition that resolves the complex data analysis in heterogeneous fields. Data Analytics in data science can be applied to every aspect of business processes and optimize the operations management, cost, time, and risk simultaneously. The cognitive science improves the business automated process by applying Big Data & Visualization. The analyses methods of improving the provision of clinical services, enhancing disease prevention, and measuring the effectiveness of various treatment options (AIML).

Keywords: Hadoop Distributed File System(HDFS); Data Acquisition (DAQ); Data Analytic (DA); Data Visualization (DV); Big Data; High-performance computing (HPC).

Introduction

Data collection is the process of gathering and identifying information about changes in relevant processes to answer research questions, evaluate hypotheses, and evaluate consequences The fundamental purpose of data collection is to measure and record evidence to improve and facilitate data analysis. Since the direction of data collection is to current state data collection become be of the outstanding quality and value. The assessment of related data is called a database[2-3]. A database is a data structure that stores knowledge. Most databases have multiple tables, and each table can have multiple fields. The essential purpose of data collection is to assemble input precisely and performance to ensure accuracy and facilitate data analysis.

A Data acquisition (DAQ) is a collection of software and hardware that allows to measure or substantially control something in the real world. A data purification system is necessary DAQ hardware, sensors and actuators, communicate hardware, and a computer running DAQ[7-8] software. Data Collection, or more frequently known as DAQ, is the process of digitizing advice about the world around us so this it can be viewed, analyzed, and stored on a computer. A simple example is the process of measuring the temperature in a room as a digital value using a sensor such as a thermonuclear. Today's data collection systems may include data analysis and reporting software, network connectivity, and remote control and monitoring options[15].

A Data acquisition is the process of checking signals that measure every physical nature and reorganize them into digital form for rectification by computers and related software.

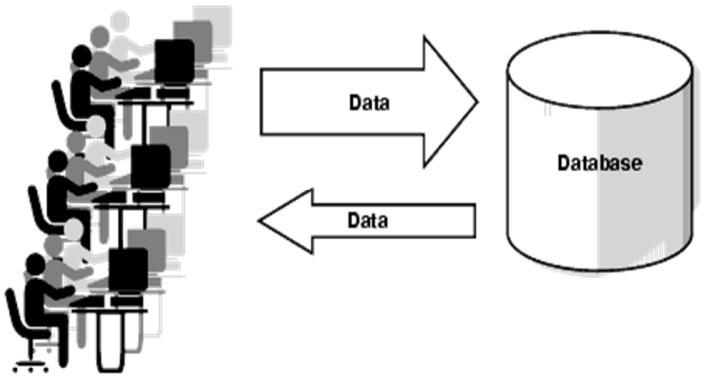
For research in many fields, data collection is often a specialized process, with researchers developing and using spe cific measures designed to collect data. However, in business and research, the data collected must be accurate for a nalysis and research to be valid. Data collection in business happens on many levels. As transactions are processed a nd data is entered, IT systems routinely collect information about customers, employees, sales and other aspects of b usiness and services[12-14]. The company also conducts surveys and monitors customer feedback on social media. Data scientists, other analysts, and business users collect relevant data from internal processes and, if necessary, exte

rnal information for later analysis. The second job is the first step in data preparation, collection, and preparation for business intelligence (BI) users and analytic application[7-8]

Traditional Data Collection
A data collection tool is a device or tool used to collect data, such as a survey or computer-assisted interview. Also, here are some data collection techniques applied by data collection tools.

1. Interviews
2. Questionnaires
3. Case Studies
4. Usage Data
5. Checklists
6. Surveys
7. Observations
8. Documents and records
9. Focus groups
10. Oral history

Figure: 1 Traditional Data Collection



Traditional Problem statement (EDP, MIS, and ERP):

- Inconsistent data collection standards.
- Context of data collection.
- Data collection is not core to business function.
- Complexity of data collection
- Lack of training in data collection.
- Lack of quality assurance processes.
- Changes to definitions, policies and maintaining data comparability.

Related Work

When wireless technology was introduced in the 1990s, information has changed a lot over time, especially in its am ount, collection, type, and analysis. Information has become what it is today, affecting every business. The future of knowledge and the future of business research are also intertwined. Therefore, it is important to understand its histor

y, especially as changes occur in the next data12-14].
Other information about astrological studies and time study led to the later discovery. Finally, as more information is discovered, so does the need for tools to collect, analyze and store information, even in the early days of informatio n history.

However, the data transfer has not yet come to an end. We bet you can guess what changed the world of information as we knew it in the 1990s the Internet. More information, more types of information, and new ways of collecting, u sing and analyzing information are all products of the Internet. But since we all know this story, we can jump straigh t to the important part - the future of knowledge[23].
Databases allow us to store information about a topic in a unified way. In addition to data storage, merging, extractin g and saving data-related content**.**

REAL TIME DATA COLLECTION-HDFS

The Real-time data is information that is delivered immediately after collection. There is no delay in the timeliness of the information provided. Real-time data is often used for navigation or tracking There are five benefits on real-time data collection.

- It gives deep analysis insights immediately.
- It improves action, reaction and optimize the resources.
- It democratizes the stakeholder data for faster decision system.
- It provides a 360-degree view of customer's interest.

  It encourages to make business processes more faster and efficient.

Literature Survey and Data Collection

This proposed literature survey provides basic data regarding the first step of data identification and classification for storage, and retrieval process on various database system for the purpose of application operation, maintenance, services. The file system demand and supply are the two parts of the same coin. The demand is directly proportional to the quality of data and information. The necessity of study and analysis in any data collection process has increased because of the changes in logic, structure, and the type of technology applied to services that generate quality of data. Finally, the business increases along with technology for business and society, which creates impress information and spreads over its complex society. We have to focus on performance, security, enhancement, integration, verification and validation the data collection process. This survey article proposes and resolves the heterogeneous data collection system by applying data collection, classification, frequent pattern, re-organization, change management, version control, and access control mechanism up to the highest level of the operation and services for specific requirement for the business and technology in a continuous process.

Table 1:Data Collection & Analysis

Tools	Lunched in Year	Developed By	Utilities	Remarks
Excel	1985	Microsoft Corporation	Store, restore, retrieve, Display Graphical Analysis	We can't create a table or delete a block of cells if the workbook is shared Problem in Tera Data set. Hardware & Software Performance Issue Problem in Automated Report Processing Only possible Structure data

FoxPro	1984	Fox Software	Store, restore, retrieve, Display Graphical Analysis not available	I/O Operation Failure error(OS) NTFS-Read Only Updation problem Problem in Tera Data set. Hardware & Software Performance Issue Problem in Automated Report Processing. Only possible Structure data
Dbase	1979	C. Wayne Ratliff	Store, restore, retrieve, Display Graphical Analysis not available	Temporary File Is Too Large Foxpro Error. Problem in Tera Data set. Hardware & Software Performance Issue Problem in Automated Report Processing Problem in Heterogeneous Data sets Only possible Structure data
DBMS-Oracle, Sybase, My SQL,: homogenous data	1970	Edgar F. Codd	Store, restore, retrieve, Display Graphical Analysis not available	Problem with process improvement. table does not use a partitioned primary index: Problem in Tera Data set. Hardware & Software Performance Issue Hardware & Software Performance Issue Only possible structure data
OLTP	1970	IBM- American Airlines		Data is reach & Information is poor
OLAP	1993	Edgar F. Codd Stefan Urbanek		ETL Tools is Costly & Time consuming.
Big-Data	1990	John Mashey		Unstructured, Semi-structure, and Structure
Future Data AIML	1997	Cognitive Technology IBM-Garry Kasparov,	Analyze emerging patterns, spot business opportunities and take care of critical process-centric issues in real time	Robotic process automation to speed up the data analysis. DV Accurate Data Analysis Optimize the Risk & Maximize the DSS
Apache Software Foundation Hadoop Distributed File System	HDFS	Apache License Cross-platform Doug Cutting, Mike Cafarella	2004 HDFS Initiated 2004-2007	Open source Google FS clone. Distributed Fault Tolerance system Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Hadoop was originally designed. HDFS is far better than the GFS.

Problem Statements:

We can conclude that the existing literature survey, there are many problems are reflected and unsolvable till now. Therefore, MIS, OLTP, OLAP and current decision support system are problematic with the following points. The OLTP and MIS data quality are poor, Uncertainty, disintegration, unstructured, un classification, unordered, and static system is not properly identified and classified as per top management desired manner

We can conclude that the existing literature survey, there are many problems are reflected and unsolvable till now. The data collections and management information Systems are poor data quality, Uncertainty, disintegration, unstructured, un-classification, unordered, and static system is not properly identified and classified as per requirement of the top management.

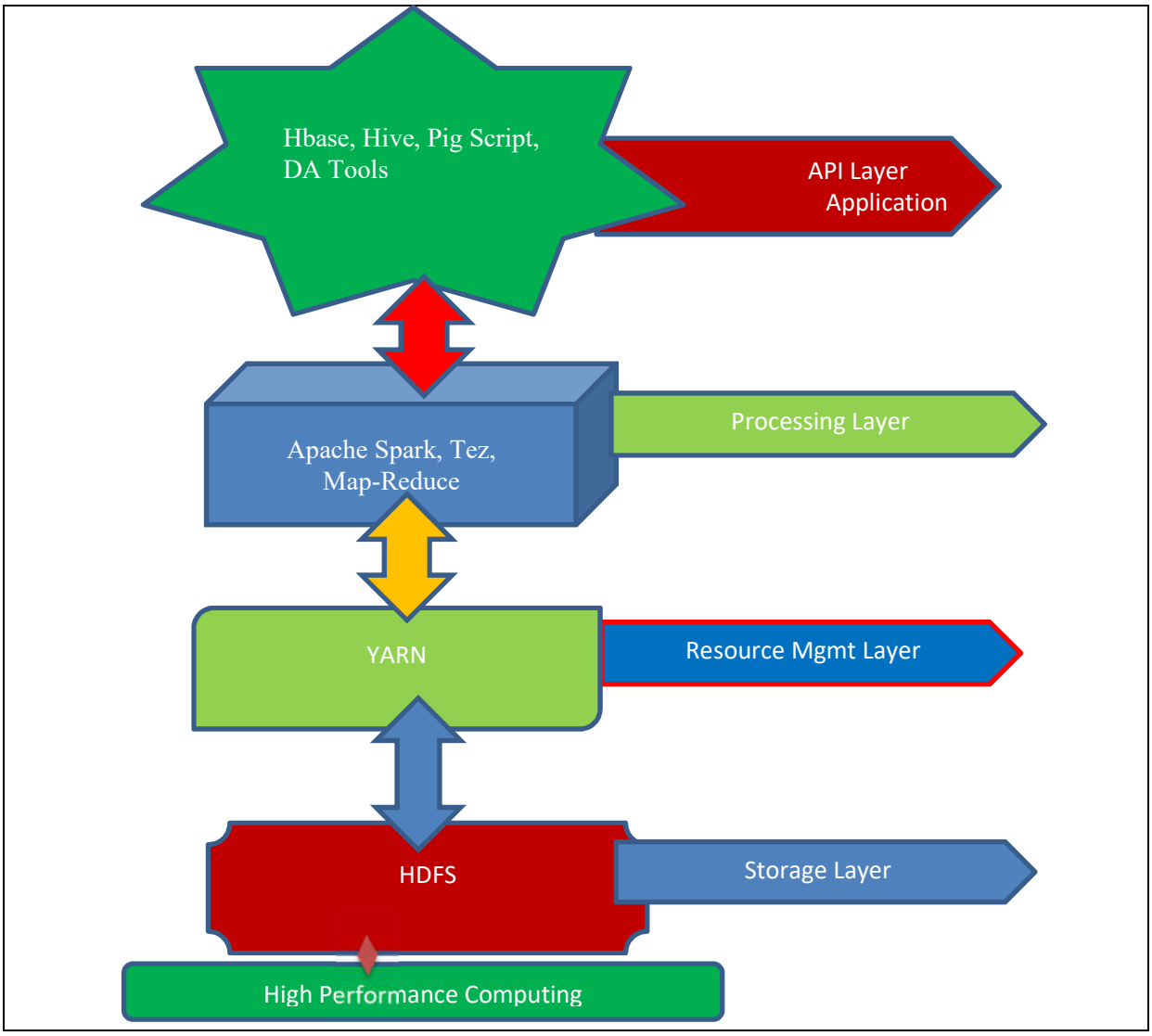
- The scalability, integrity, mapping, and reliability Issue in data and information.
- Problem in Identification & Classification of heterogeneous data
- Inconsistency data and information degrade the performance of the large network as well as HPC.
- Issue in efficiency and effectiveness of data as well as information.
- Dirty data over a OLTP(Unstructured & unification issue)
- Data is rich & information is poor.
- Large reference Integrity is one of the key dependencies, but theoretically, association, dependency has no life but in composition & aggression have life.
- OLTP, EDP, MIS, EIS, DSS are not supporting frequent multi pattern data analysis.
- The scalability, integrity, mapping, and reliability Issue in data and information.
- Problem in Identification, Classification, frequent pattern of Objects & Devices which is not available of traditional database.
- Inconsistency data and information degrade the performance of the large network as well as large operating system.
- RDBMS does not support many to many relationships.
- Data is rich & information is poor.
- SQL Support batch processing and can generate MIS & OLTP only, therefore problem in decision support system.

RESEARCH SCENARIOS

Hadoop Distributed File system –At present HDFS is the world's most reliable storage system. HDFS is a File System of Hadoop designed for storing very large files running on a cluster of commodity hardware. It is designed on the principle of storage of less number of large files rather than the huge number of small files.

Hadoop HDFS provides a fault-tolerant storage layer for Hadoop and its other components. HDFS Replication of data helps us to attain this feature. It stores data reliably, even in the case of hardware failure. It provides high throughput access to application data by providing the data access in parallel.

Figure: 2 Block Diagram of HDFS



Apache Hadoop
Apache Hadoop is a Java-based open source framework that deals with some big data.
Apply an array/clusters of objects that controls the collection of values
is the solution for data processors.
Provides fast access to information and actions & Reaction.
It apply Hadoop Distributed File System (HDFS) which allows heterogeneous types of data to be stored.
It is where data **is** stored and processed.
Hadoop for cluster management, parallel process and data storage.

Implementation of HDFS

Hadoop Distributed File System (HDFS) is a distributed file system designed to run on low-cost hardware. Applying new easy-to-use tools can solve complex problems faster than trying to create custom solutions. The market is full of vendors offering Hadoop as a Service(HASS) or as a standalone tool.

PERMUTATION & COMBINATIAONAL TECHANIQUE:

Figure 3: Nodes and Data Mgmt. on HDFS

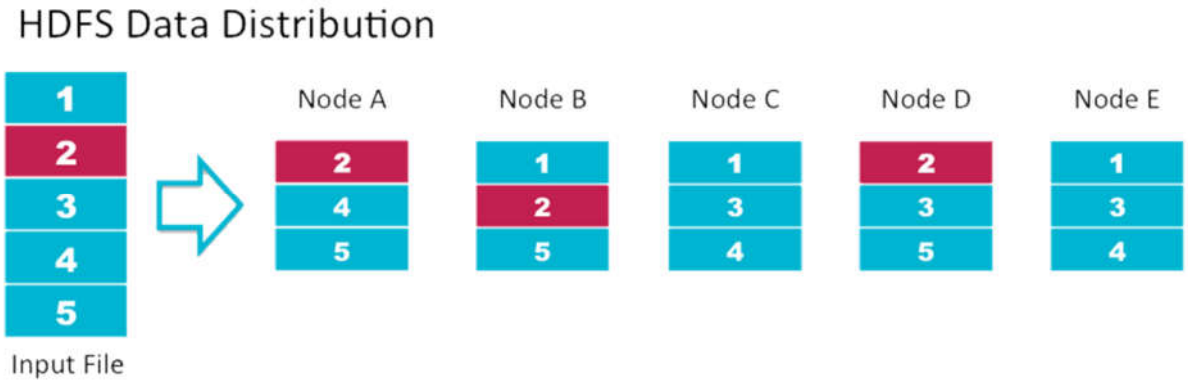
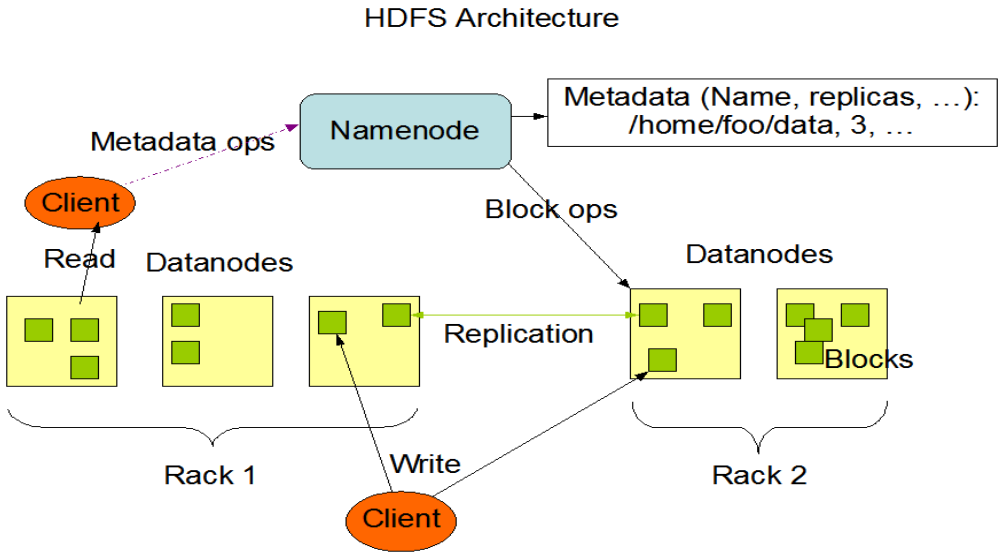


Figure: 4 Block Diagram of HDFS Internal Operation & Services



High Reliability and Fault Tolerance (Internal Operation & Services)

The heartbeat is a TCP handshake signal. Data nodes on each slave send continuous heartbeats to the master. The working-time of the heartbeat is three seconds. If the Name Node does not receive a signal for more than ten minutes, it writes the Data Node and the data block is assigned o the different nodes.

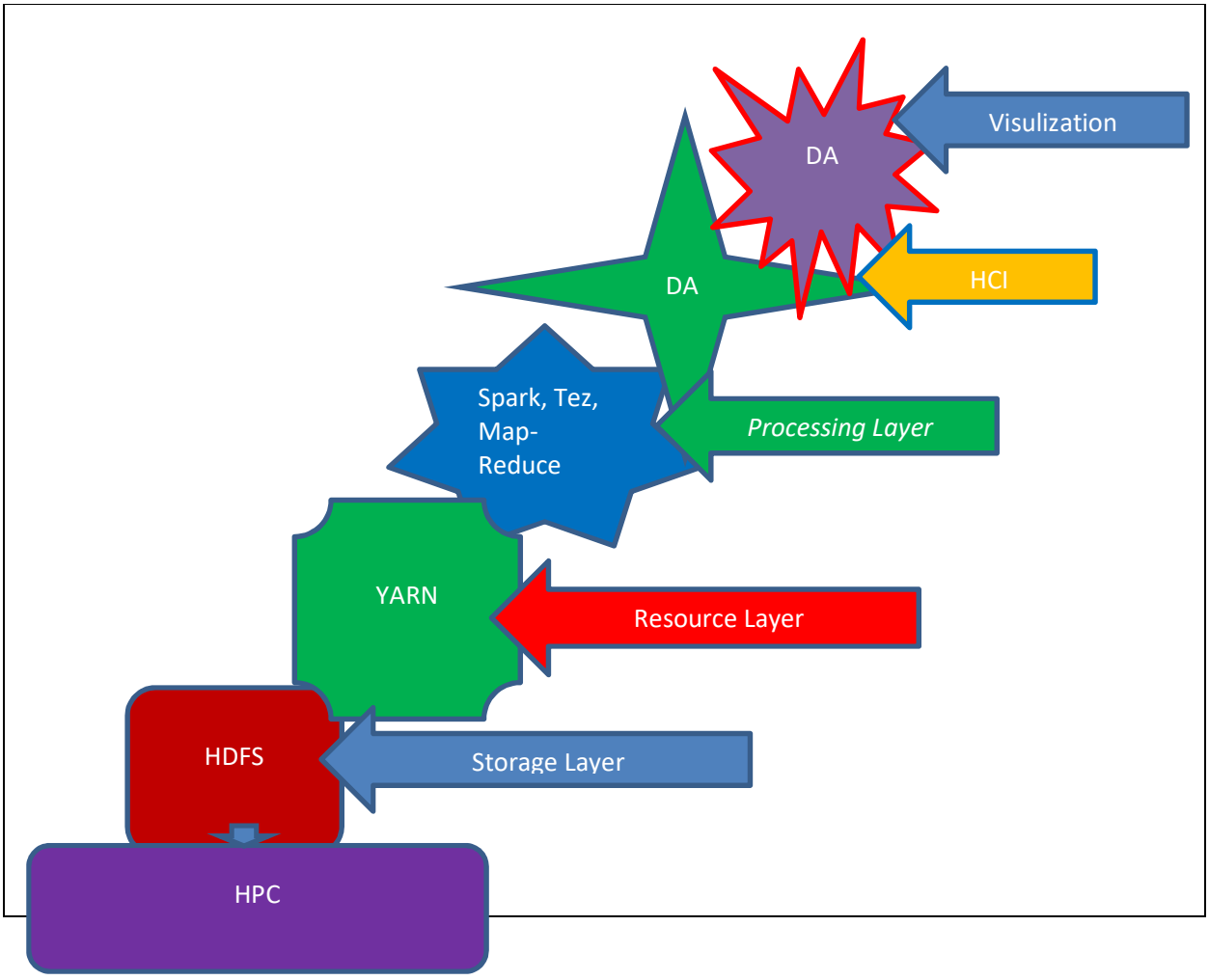
Do not try to reduce the load on the Name Node by reducing the heartbeat frequency. Even in large groups it is impo rtant to "know" the name. Without a steady and regular heartbeat, the name node is severely affected and cannot effe ctively protect the clusters (Resource Mgmt).

To avoid negative consequences, maintain rack sensitive locations and keep copies of data blocks on server racks. If you lose a server rack, other copies remain available with minimal impact on data processing.

Operation & Service of HDFS.

- **Collection: From Various Heterogeneous Source.**
- **Storage (HDFS):** This huge amount of data, Hadoop uses HDFS (Hadoop Distributed File System) which uses commodity hardware to form clusters and store data in a distributed fashion.
- **Function:** It works on Write once, Read many times principle (WORM).
- **Operation and Processing:** Map Reduce paradigm is applied to data distributed over network to find the required output.
- **Analysis Tools:** Pig, Hive can be used to analyse the data.
- **Cost Effective:** Hadoop is open source so the cost is no more an issue.

Figure: 5 Integrated Diagram for Big Data Operation & Services



RESULT ANALYSIS

DV can also be used as a form of mental rehearsal. Through process visualization, increase our selective attention.

- The data science is the sub sets of Big Data and the Visualization is the sub sets of Data Science.
- Data Science examples Such as; Identification, classification, pattern Matching, and prediction of disease, Optimizing shipping and logistics routes in real-time, detection of frauds, healthcare recommendations, automating digital ads, etc. Data Science helps these sectors in various ways.
- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals =VA(Visual Analytics). To identify and share real-time trends, pattern, outliers, and new insights about the information represented in the data. Data visualization is the graphical representation of information and data. The data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs. It easier to identify trends, patterns, and outliers within large data sets. The data visualization is the presentation of data in a pictorial or graphical format. The Visualization is a technique that allows you to set the parameters to make your future vision of reality. In creative visualization, we can direct our brain to focus on what matters are involved and to engage in a process called selective attention. The Visualization is a technique that quick support in current and future DSS.

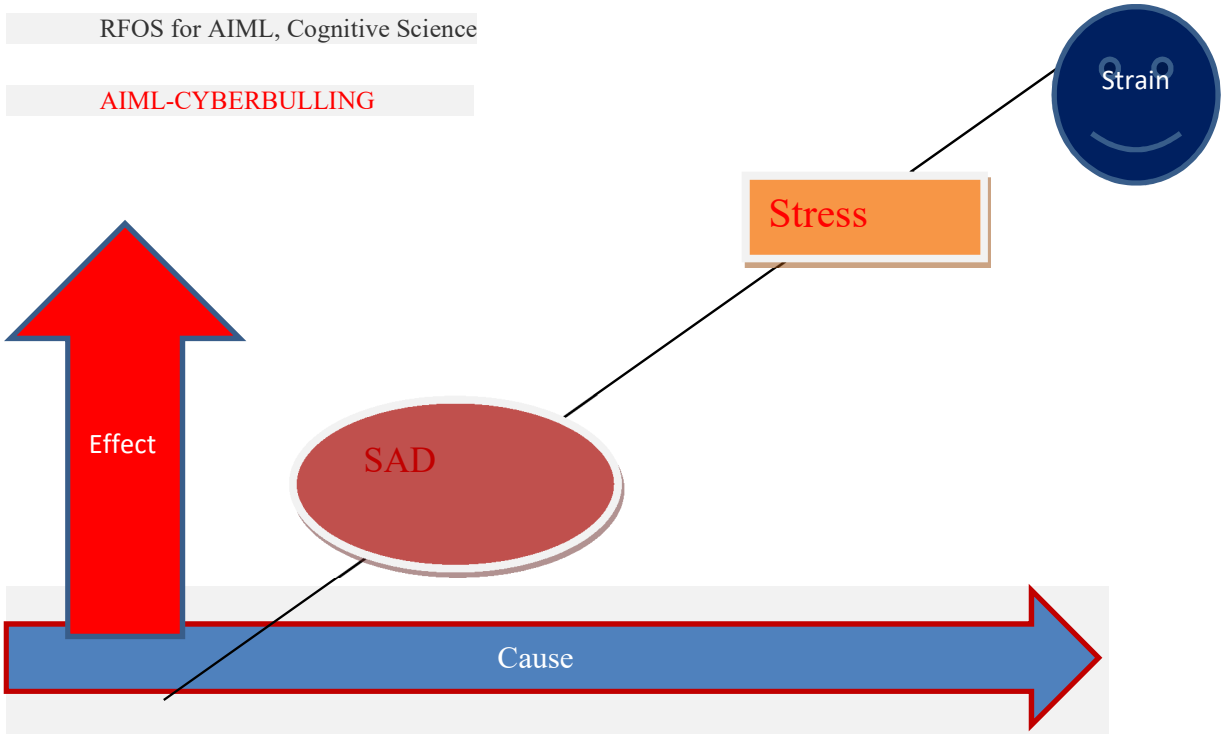
Discussion

The Data Visualization is a useful technique that helps us to reach our goals and live our dreams. It fulfilling the vision and mission simultaneously of the stake holders. It works by getting our mind and body ready for what we want to do and what happening – just like exercise, the more we do it, the stronger it becomes.

Figure: 6 Block DiagramAIML CYBER BULLING

RFOS for AIML, Cognitive Science

AIML-CYBERBULLING



- Studies of mind, function, operation, services, and behavior, To see the visualization report
- Mental Action, Reaction, Function, Behaviors, and Processes changed due to visualization Result.
- Philosophy and psychology change to view the clinical report.
- Hard and soft Science develop that is called cognitive science due to change of N - number of technology along with cyber bulling.
- Studies of mind, function, and behavior changes, to see the data visualization report.
- It can be used to build our self-confidence by seeing and think(AI/ML).

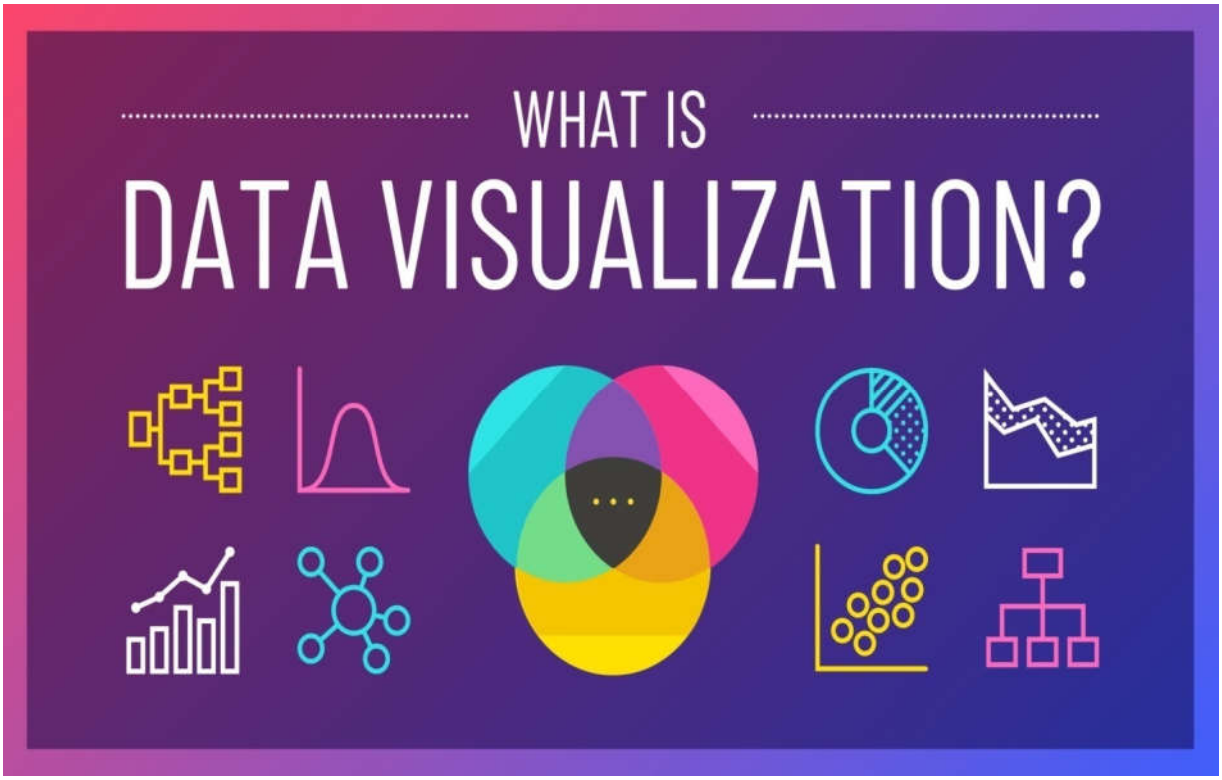
Mental stress, strain, sad, and depression happened due to clinical report

The Data Visualization can be used to motivate our focus on right time and right way, and work toward, our future ideal/safe state.

The Data Visualization **lets you comprehend vast amounts of data at a glance and in a better way**. It helps to understand the data better to measure its impact on the business and communicates the insight visually to internal and external audiences.

Figure 7: DV Real life Example: DV-Structural Learning

Figure 8 Deep VA



Conclusion

Now have an in-depth understanding of Apache Hadoop and the individual elements that form an efficient ecosystem for better, faster, efficient, and effective data collection. At present major industry is implementing Hadoop & Big data to be able to cope with the explosion of data volumes, and a dynamic developer community has enhanced Hadoop evolve and become a large-scale, general-purpose computing platform. The HDFS involves the collection, storage, processing, and investigation of data sets for decision science processes for the causes and effect of public interest through HCI, Cognitive Science to seeing & things (AI/ML). The Data Science researchers look for ways to standardize, normalize, optimize the cost and time for top management. Therefore, HDFS technologies are the software utility designed for collecting storing, analyzing, processing, and extracting huge datasets from the structure, semi-structure unstructured large data sets (Big Data) which can't be handled with the traditional data processing software. Many Companies required big data processing technologies to analyze the massive amount of real-time data. The HDFS makes more economical, benchmarking, distributed fault-tolerant, scalabilities, reliabilities, high available, replication, synchronization and is designed to be deployed on low-cost hardware, software, and network for heterogeneous data sets. Finally, the authors are concluded that the HDFS is far better than the GFS. GFS is better than DFS.

References.

[1] A. K. Gupta (2012). *Management Information System.* New Delhi, India: S Chand Publishing.

[2] Alex and Stephen.(2013). Dataware housing Data Mining and OLAP, New Delhi, India, McGrow Hill.

[3] Abraham & Sudarshan(2013). Database System Concept, New Delhi, India, McGrow Hill.

[4] Andrew Haigh.(2011). Object Oriented Analysis & Design, New Delhi, India: Tata McGraw Hill.

[5] Anil Maheshwari.(2017).Data Analytics Paperback, McGraw Hill Education.

[6] Bernard, K. (2007). *Discrete Mathematical structures*. New Delhi, India: Person Education India (PHI).

[7] Darcey Kobs.(2021). Data Science for Beginners: Comprehensive Guide to Most Important Basics in Data Science, Alex Published.

[8] Gary & James.(2009). Database Managent & Design, New Delhi, India: Person Education India (PHI).

[9] Elaine, Rich. (2012). Artificial Intelligence. New Delhi, India: Tata McGraw Hill.

[10] Elmasri & Navathe.(2007). Fundamental of Database System, New Delhi, India: Person Education India (PHI).

[11] F. Anowar and S. Sadaoui.(2020). Detection of auction fraud in commercial sites. Journal of theoretical and applied electronic commerce research.

[12] Foster Provost, Tom Fawcett .(2013).Data Science for Business, Publisher(s): O'Reilly Media, Inc. ISBN: 9781449361327

[13] Joel Grus.(2015).Data Science from Scratch Paperback, Edition 1st, O′Reill

[14] Jiawei & Kamber.(2011). Data Mining Concept & Technology, Elsvier

[15] Michael Minelli .(2018). Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Business, Gildan Media · Narrated by Ryan Burke.

[16] Michael & Rumbaugh.(2011). Object Oriented Methodology & Design with UML, New Delhi, India: Person Education India (PHI).

[17] Nilsson. (2002. Principle of Artificial Intelligence. New Delhi, India: Narosa Publication House.

[18] N. D. Vohra (2007). *Quantitive Techniques in Management*. New Delhi, India: Wiley Publishing Inc.

[19] Nasir Abbas.(2014).Memory-Type Control Charts for Monitoring the Process Dispersion, Quality and Reliability Engineering International. Wiley, Vol. 30, 623–632, 2014.

[20] Nice den Boer, "A Risk Control Strategy Corporate Computer Operations," Computer Audit, Vol., 4, pp. 18-30, January 1995.

[21] Neha & Manoj.(2018). "Analysing the impact of online Brand trust on Sales Promotions and Online Buying Decision", The IUP journal of Marketing Management Vol. XVII No.3 Aug 2018.

[22] N. Meenakshi K. E. Rajakumari S. Hariharasitaraman.(2021). Data Science and Machine Learning Paperback – Notion Press, India

[23] Rajiv Chopra.(2022) Data Science with Artificial Intelligence, Machine Learning and Deep Learning - Simplified Q & A, Khanna Book Publishing, India.

[24] Richard B Chase, Robert, Nicholas, Nitin.(2006), Operations Management. New Delhi, India: Tata McGraw Hill. 2006.

[25] Raj Sahil.(2017). Management Information System | Second Edition | By Pearson Paperback.

[26] Jeremy L. Boerger. (2021). Rethinking Information Technology Asset Management Paperback – Import, 5 April 2021.
[27] S. M. Shats F. Dong and H. Xu.(2010). Reasoning under uncertainty for shill detection in online auctions using dempster-shafer theory. *International Journal of Software Engineering and Knowledge Engineering*.

[28] Shon, H. (2002). *Security Management Practices*. New Delhi, India: Wiley Publishing Inc.
[29] S. Maital & D.V.R Seshadri (2008*). Innovation Management*. New Delhi, India: Response Books.
[30] Shu Qing Liu.(2014). "Research on the quality stability evaluation and monitoring based on the pre-control chart*, "* International Journal of Quality & Reliability Management, Volume 31 Issue 9, pp.966 – 982, .
[31] Seymour Lipschutz & Varsh Patil.(2010).Discreate Mathematics, New Delhi, India: Tata McGraw Hill.
[32] Sudarshan.(2019). Database System Concepts, Seventh Edition. McGraw-Hill, India
[33] Sales Rodrigues, K. A. (2022). Book Review: An Introduction to Nonparametric Statistics. *Journal of Behavioral Data Science*, *2*(1), 124–127. https://doi.org/10.35566/jbds/v2n1/p8

[34] Trivedi. (2009). *Artificial Intelligence*. New Delhi, India: Khanna Book Publishing.

[35] Turban, Aronson, Liang, Sharda (2009). *Decision Support and Business Intelligence Systems*. New Delhi, India: Person Education India (PHI).

[36] Tong Xinand Ban Xiaofang. (2014}. ,"A Hierarchical Information System Risk Evaluation Method Based on Asset Dependence Chain, ," Intl. Journal of Information & Network Security,Vol. 3, No.3, 2014.

[37] Tzvi Razand David Hillson (2025). "A Comparative Review of Risk Management Standards,"Risk Management,7, 53–66; doi:10.1057/palgrave.rm.8240227, 2005.

[38] Thomas S. Coleman, (2011). "A practical guide to risk management," Research foundation of CFA Institute, ISBN 978-1-934667-41-5, July 2011.
[38]V. K. Jain, (2018). Data Science and Analytics Paperback Khanna Publishing, India

[39] Weber, Ron. (2014). *Information System Control and Audit*. New Delhi, India: Person Education India (PHI).
[40] Waggoner, P., & Kennedy, R. (2022). The Role of Personality in Trust in Public Policy Automation. *Journal of Behavioral Data Science*, *2*(1), 106–123. https://doi.org/10.35566/jbds/v2n1/p4/