



## Big data implementation to optimize the cost for the improvement of interaction among the public and government creativities

By

Dr. Sardar Ali Professor in EEE

Nalla Malla Reddy Engineering College, Autonomous (Affiliated to JNTUH and Approved by AICTE), Divyanagar, Ghatkesar, Hyderabad-500088, Telangana State, India.

### ABSTRACT

This proposed big data project aims to implement the optimizing cost and time to improve the communication and interaction among the private, public and government in a continuous manner to improve the creativities on all time for the public interest. Our aim and objective are that the big data project and plan will be maintained work life balanced at the global levels for a large scale of digital well being. It is a great challenge to integrate the field of multi source data for the application of big data and data analytics along with high performance computing. At present biosciences are looking for significant data analysis to maximize efficiencies and health related issues. The big data has more potential to improve the internal efficiencies and operations through real time operations. The life data acquisition and retention can be immediately build up and analyzed the preventive action based on generated decision making to apply various tools such as, representation, snap view, Microsoft power BI, data wrapper, Plotline, sense, Excel, Java, and Python.

A brief analytic study between Distributed system and Hadoop distributed file system was surveyed in this study for better digital services. Taking comparative techniques for design, architecture, development, and deployment of DFS and HDFS, allows us to use to deduce that both DFS and HDFS are considered two of the most used distributed file systems for dealing with huge clusters where big data lives. This study will help understand the design system and highlight the common features between DFS and big data management is collecting, storing, processing, distributing, retrieving, analyzing and maintenance of multi source data for the better and faster analytics applies to data acquisition, retention and data engineering HDFS.

### 1. INTRODUCTION

Traditional System for Data Collection and Processing: In 3rd generation computing systems and traditional file processing systems include **manual systems and also computer-based file systems that were linked to particular application programs like COBOL, REXX, and C Language.** That type of file processing makes used with our 3GLtrational programming. They share several characteristics that might be not applicable at these moments.

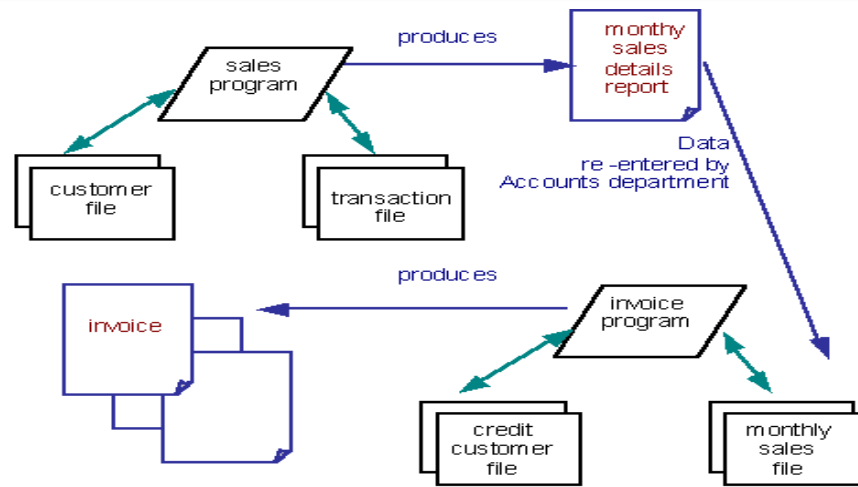


Fig 1: Traditional file system

## 2. PROBLEM IN TRADITIONAL FILE SYSTEM

There is a big difference and un-balanced between traditional system & Big Data application, replication, propagation, and synchronization. The hardware, software, operating system, and file systems are completely isolated with each other in the traditional system. There are many more issue was with traditional system like scalabilities, replication, concurrency, reliability, benchmarking, security, fault tolerance, and high availabilities.

Data redundancy and inconsistency. Difficulty in shearing and accessing data. Data isolation – multiple files and formats. Integrity problems in high end computing system. Since each application has its own data file, the same data may have to be applied.

## 3. DATA INCONSISTENCY:

Data redundancy leads to data inconsistency especially when data is to be required. The poor data security and privacy are the most threatening problem in file Processing System. There is very less security, availability, and reliability in file Processing System as anyone can easily modify and change the data stored in the files. All the users must have some restriction of accessing data up to a level.

**DATA REDUNDANCY:** is possible that, the same information may be duplicated in different files like

- Data Inconsistency
- Difficulty in Accessing data
- Data sharing issue & security issue
- Integrity Problems& CIA Issue
- Atomicity Problems
- Concurrent Access Anomalies
- Security Problems

## 4. DISTRIBUTED FILE SYSTEM

The distributed file system (DFS) is a file system with data are stored, shareable, accessible on a server and the clients are get that data base on their request. The data is right to use and practiced as if it was stored on the consumer location. The DFS produces its extra appropriateness to assign Task and databases between the dissimilar customers on a organism in a arranged and agreed manner. There are many more DFS file system are available like: NFS, UFS, DFS, P2P, JFS, Global File System, and Google File System.



The DFS be valid the Windows Server folder imitation check to reproduction alters between imitated the source and goals. The stakeholders can modify files (Writes) stored on one target, and the file replication repair spreads the alters to the further assigned objectives. The repair protects the mainly up to date transform to a article or folders / CD.

The allocated folder scheme (AFS) is a folder scheme which is dispensed on a variety of folder servers and sites widen above the environmental region. It authorizes courses to contact and store up separated data in the similar manner as in the local folders. It also authorizes the consumer right to use folders from any scheme. It permits network customers to allocate information and folders in a controlled and authorized approach. Although, the servers have complete control over the data and provide users access control (ACM). At that time multi-factors authentication was not developed and balanced.

The distributed file systems are maintaining for the massive amounts of data is stored and process at their own site. The emergence of Hadoop File Systems(hdfs), Global File System, Google File Systems and Network File Systems(NFS) have changed the course of how data is managed in servers and has its individual allegations on shade calculating and Big Data organization. every folder scheme tenders its individual benefits and confronts in tenures of presentation, burden-patience, reliability, scalability and accessibility. This release method an unlock dispute on how these can be acquired up for the operation. The choice of a feature available with each one of them has their own metrics that differentiates them from other heterogeneous file systems. This article seems at a relative reading on the folder schemes and suggests the measure behind the option of assortment of a exact folder scheme and exact stage. The study also discovers the benefits of employing a folder scheme and the advantages and disadvantages connected with them.

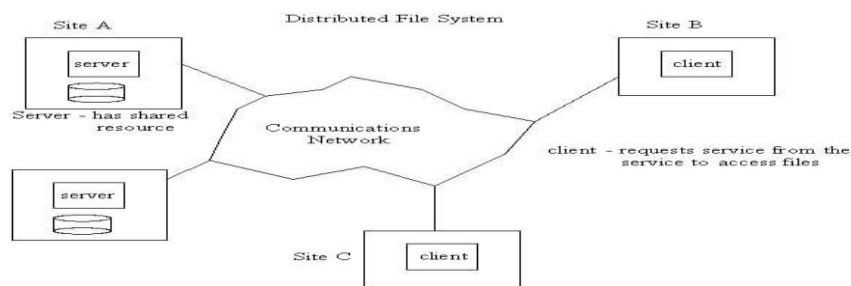


Fig 2: distributed file systems

The dispensed organism will unavoidably amplify over time when additional mechanisms are included to the system, or two systems are connected jointly. A high-quality DFS should be proposed to balance quickly as the scheme's number of knobs and consumers enlarge.

- In dispensed folder scheme knobs and links requires to be protected thus, we can declare that security is at hazard.
- There is an option of defeat of communications and data in the system though progress from one knob to a new.
- Database link in case of dispensed folder scheme cause difficulties.
- As well treating of the database is not simple in dispensed folder scheme as contrast to a lone consumer scheme.
- There are possibilities that burdening will obtain if all knobs attempts to send data at one time



## 6. PROBLEM STATEMENTS

- Dispensed crisis come about all the reasonable stages of a dispensed scheme, not presently at small-point bodily mechanisms. Allocated crisis obtain poorer at upper points of the arrangement, due to persistent. dispensed insects frequently show up long after they are organized to a organism. issued insects can increase across an whole scheme.
- The two basic design concerns are disintegration, the division of the database into separations called fragments, and division, the most selected allocation of portions.
- In a dispensed database scheme, we require to contract with four types of failures: Dealt failures, location, failures, Social media failures and line communiqué failures. A few of these are due to hardware and others are owing to software.
- The previous contain an unanticipated out of order connection in a dispensed database, and the latter include unauthorized access, unauthorized alteration, traffic analysis and complex input.
- dispensed schemes have not any bodily allocated recollection, all computers in the issued system have their individual specific bodily memory. As computer in the dispensed scheme do not allocate the ordinary memory, it is not possible for anybody to know the worldwide status of the full dealt out systems.

### IMPLEMENTATION OF IMPROVING TECHNIQUES:

The Following techniques are to be implemented to achieve the improved performance.

- Support the storage of massive data.
- Detect and quickly respond to hardware failures.
- Streaming data access.
- Simplified consistency model.
- High fault tolerance.
- Low cost Commercial hardware.

## 7. TECHNICAL CHALLENGES

- **To develop data exchange and interoperability architecture to provide multi-disciplinary domain.**
- **To develop the AI/ML-based Analytical platform for integrating multi-sourced data.**
- **To propose a Predictive and Prescriptive Multi-Modeling Platform for physicians to optimize the semantic gap for an accurate analysis.**

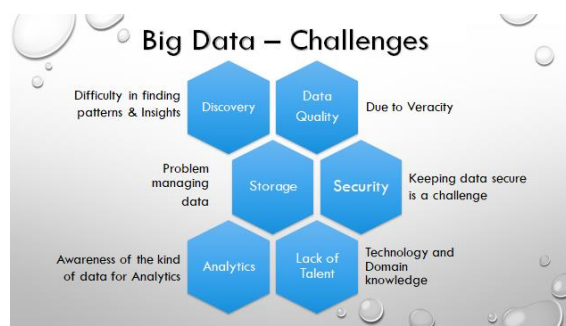


Fig 3: The big data challenges



### BENEFITS OF BIG DATA.

- Gather: From different mixed resource.
- Storage (HDFS): This very big amount of data, Hadoop employs HDFS (Hadoop dispersed folder structure) which utilizes product hardware to form groups and stock up data in a dispersed style.
- It functions on mark one time, interpret several instances code.
- routing: Map decrease concept is related to data dispersed over system to find the obliged output.
- Analyze: Pig, Hive can be utilized to analyze the data.
- Cost: Hadoop is unlock resource, thus the cost is not a problem
- The Hadoop dispersed folder scheme is a dispersed file scheme proposed to run on low-priced product hardware.
- Distributed file systems -Attributes.
  - Highly fault-tolerant, Scalabilities, Reliabilities, high available, replication, synchronization and is designed to be deployed on low-cost hardware.
  - gives tall through output access to use the data and is suitable for purpose to have the big data position.

### 8. HDFS REPLICATION, PROPAGATION AND SYNCHRONIZATION

The data acquisition, retention, storage, and backup are working all together to maintain the stabilities, availabilities, reliability, and scalabilities of big data for all the time and every times for heterogeneous stakeholders( Biomedical Science, Education, Hotel Management, Hospital, Farming, Travel & Tourism, Sales & Marketing)

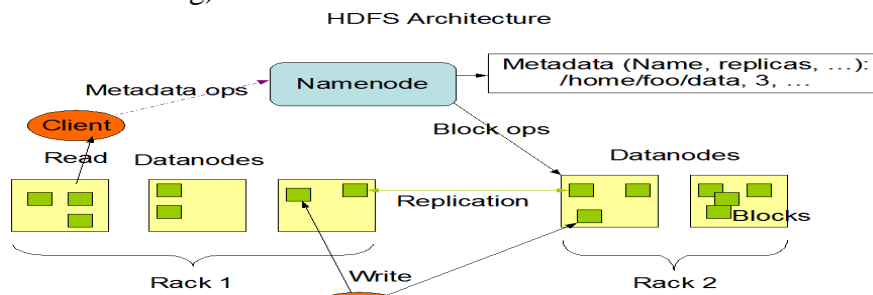


Fig 4: HDFC Architecture

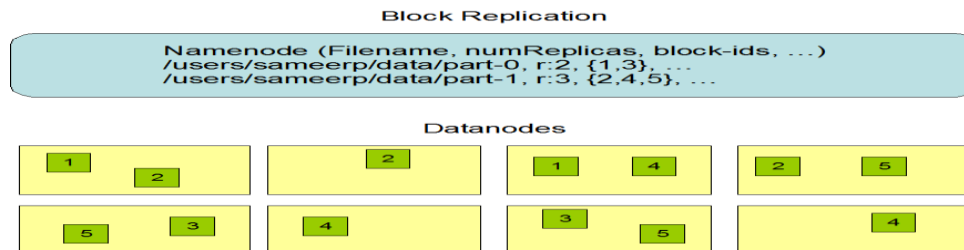


Fig 5: Block Replication



**PERMUTATION & COMBINATIONAL TECHNIQUE:**

HDFS Data Distribution

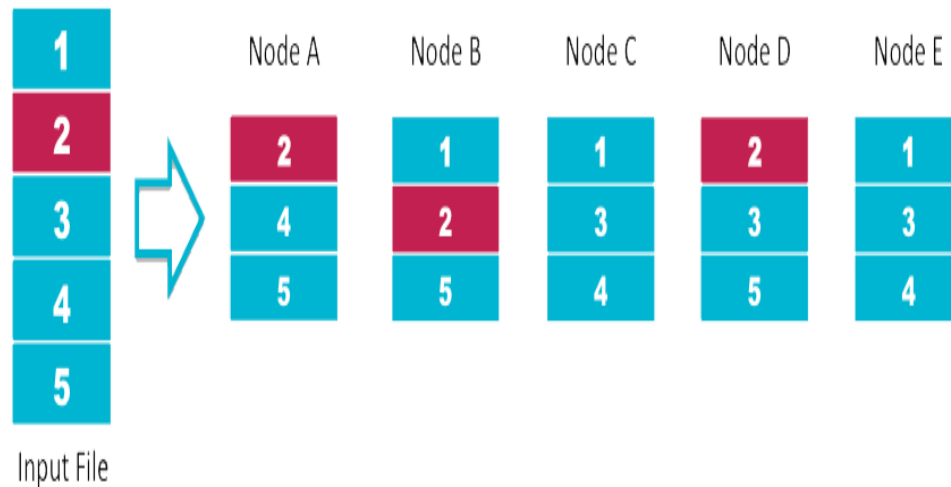


Fig 6: HDFS Data Distribution

Permutation and combination are the methods employed in counting how many outcomes are possible in various situations. Permutations are understood as arrangements and combinations are understood as selections. As per the fundamental principle of counting, there are the sum rules and the product rules to employ counting easily.

Suppose there are 14 boys and 9 girls. If a boy or a girl has to be selected to be the monitor of the class, the teacher can select 1 out of 14 boys or 1 out of 9 girls. She can do it in  $14 + 9 = 23$  ways (using the sum rule of counting). Let us look at another scenario. Suppose Sam usually takes one main course and a drink. Today he has the choice of burger, pizza, hot dog, watermelon juice, and orange juice. What are all the possible combinations that he can try? There are 3 snack choices and 2 drink choices. We multiply to find the combinations.  $3 \times 2 = 6$ . Thus Sam can try 6 combinations using the product rule of counting. This can be shown using tree diagrams as illustrated below.

In order to understand permutation and combination, the concept of factorials has to be recalled. The product of the first  $n$  natural numbers is  $n!$ . The number of ways of arranging  $n$  unlike objects is  $n!$ .

**PERMUTATIONS**

A permutation is an arrangement in a definite order of a number of objects taken some or all at a time. Let us take 10 numbers: 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9. The number of different 4-digit-PIN which can be formed using these 10 numbers is 5040.  $P(10,4) = 5040$ . This is a simple example of permutations. The permutations of 4 numbers taken from 10 numbers equal to the factorial of 10 divided by the factorial of 6.

**Combination**

A combination is all about grouping. The number of different groups which can be formed from the available things can be calculated using combinations. Let us try to understand this with a simple example. A team of 2 is formed from 5 students (William, James, Noah, Logan, and Oliver). This is the combination of 'r' persons from the available 'n' persons is given as  $nCr = \frac{n!}{r!(n-r)!}$ . The combinations can happen in the following 10 ways by which the team of 2 could be formed. The difference of 10 and 4. The permutations is easily calculated using  $nPr = \frac{n!}{(n-r)!}$ .



## 9. IMPACT OF TECHNOLOGY

At this moment, it is a great challenge to integrate the field of Biomedical science, Big data, Data Science along with high-performance computing. At present, bio-sciences are seem to be influence data revelation to exploit effectiveness and nearby health-communicated discovering. big data has the possible to get better inner efficiencies and functions all through automatic progression computerization. The vast quantity of real-time data can be instantly examined and built into health care progression for automated decision making. The Table au, Info graphics and Motion Graphics, Java Scripts, Java, Pythons, .NET, charts, graphs, and maps, Spark, and Zoho tools are more helpful for Visualizations.

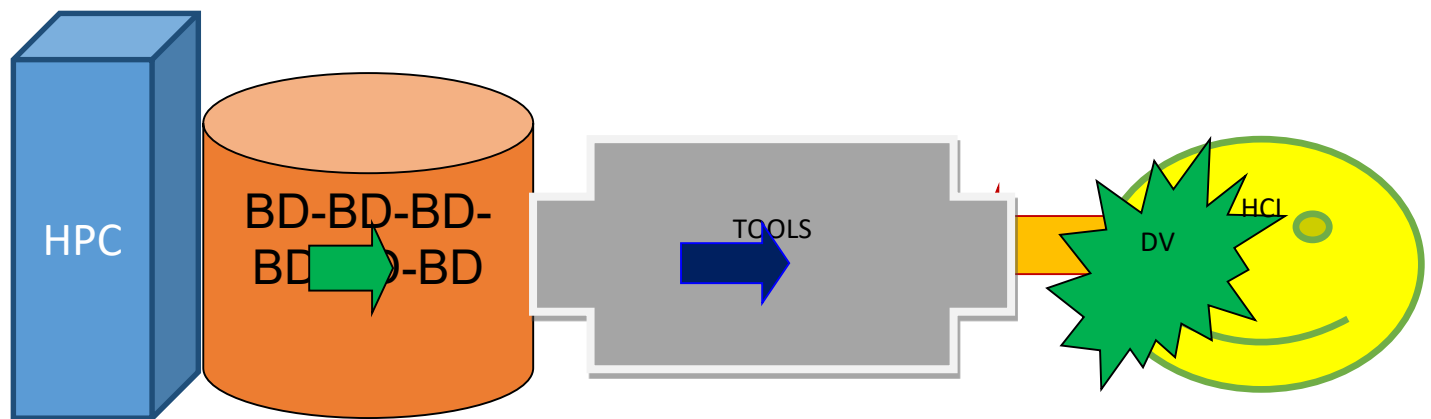


Fig 7: Impact of Technology

Impact Analysis on high tech society

Data Visualizations and Big Data provide various facilities like identifying new opportunities, new patterns, new disease, new business, understanding patient behavior, and requirement, becoming more agile and outperforming the competition in the visualization and analytics field. The DV can now deliver insights that enable businesses to better understand their patients' requirements, improve healthcare techniques, and identify issues and opportunities that emerge in real-time(Big Data and High-performance computing).

Advantage: Highly supportive for Data Science & Analytic

- Scalable: It can reliably, high available, replicable, store and process pet bytes datasets.
- cheap: It distributes the data and processing across clusters of commonly available computers.
- Resourceful: By allocating the data, it can develop it in parallel on the knobs where the data is situated.
- Dependable: It routinely continues numerous copies of data and robotically diverts computing missions based on breakdowns.
- High Available
- Replication
- Archived

## 10. CONCLUSION / RESEARCH SCENARIOS


Hadoop dispensed folder scheme – HDFS is the globe's mainly consistent storage organism. HDFS is a folder scheme of Hadoop proposed for accumulating a extremely huge folders running on a group of



product hardware. It is suggested on the code of storage of a fewer number of huge folders rather than the vast number of little folders.

Hadoop HDFS gives a error-lenient storage sheet for Hadoop and its new modules. HDFS imitation of data assists us to reach this characteristic. It stores data dependably, still in the case of hardware breakdown. It gives tall during output right to use for function data by giving the data entrée in equivalent

#### REFERENCES

- [1]. Debating big data: A literature review on realizing value from big data. Panel Wendy, Arianne Günther, Mohammad H. Rezazade Mehrizi, Marleen Huysman, Frans Feldberg. "The Journal of Strategic Information Systems" Volume 26, Issue 3, September 2017, Pages 191-209
- [2]. " *The Smart cities with big data: Reference models, challenges, and considerations*". Panel Chiehyeon Lim , Kwang-Jae Kim , Paul P. Maglio. *Cities* , Volume 82, December 2018, Pages 86-99.
- [3]. "*Big Data in the Public Sector Applications and Benefits*" Manisha Sahu . Aug 11, 2021, courtesy 
- [4]. Big data analytics for policy making Report "*A study prepared for the European Commission DG INFORMATICS (DG DIGIT)*". This study was carried out for the European Commission by "Deloitte"
- [5]. "*Digital Transformation: Exploring big data Governance in Public Administration*". Alexander Yukhno. *Public Organiz Rev.* 2022 Dec 13 : 1–15.doi: 10.1007/s11115-022-00694-x [Epub ahead of print]