Big Data - Concepts

18 December 2023

1. Introduction to Big Data

Lecture

- Big Data - Beyond the Hype,
- Big Data Skills and Sources of Big Data,
- Big Data Adoption,
- Research and Changing Nature of Data Repositories,
- Data Sharing and Reuse Practices and Their Implications for Repository Data Curation,
- Overlooked and Overrated Data Sharing,
- Data Curation Services in Action,
- Open Exit: Reaching the End of The Data Life Cycle,
- The Current State of Meta-Repositories for Data
- Curation of Scientific Data at Risk of Loss: Data Rescue And Dissemination

HA&DR - High Availability and Disaster Recovery
SPOF - Single Point of Failures

Environment Setup at present:

Oracle Virtual Box , one instance of VM - Ubuntu Server 20.04 LTS (installed with SSH) , 4GB RAM, 2 NIC (one connected to NAT and another to Host-Only network), 60GB Diskspace (partitioned and installed with LVM). We have allocated 2 Vcpu's

Big Data

- Refers to extremely large and complex datasets that traditional data processing tools and methods are unable to handle effectively.
- These datasets typically involve vast amounts of information generated at high velocity and come in various formats, including structured, semi-structured, and unstructured data.
- 3V's - primary characteristics of big data are often referred
  - Volume: Big data involves large amounts of data. This can range from terabytes to petabytes and beyond. The sheer volume of data is a significant challenge for storage, processing, and analysis.
  - Velocity: Big data is generated at a high speed. Data streams in real-time or near-real-time, requiring systems to process and analyze information quickly to derive meaningful insights.
  - Variety: Big data comes in various formats, including structured data (e.g., databases), semi-structured data (e.g., XML, JSON), and unstructured data (e.g., text, images, videos). Managing and analyzing diverse data types is a key aspect of big data.

In addition to the three Vs, two more Vs are sometimes added to further characterize big data:

  - Variability: Big data can exhibit variability in terms of data flow and structure. This variability can pose challenges for data processing and analysis.
  - Veracity: Refers to the quality and reliability of the data. Big data sources may include errors, inconsistencies, and inaccuracies, making it crucial to ensure the veracity of the data before making decisions based on it.

Organizations leverage big data technologies and analytics to extract valuable insights, improve decision-making processes, and gain a competitive advantage.

- Technologies like Apache Hadoop, Apache Spark, NoSQL databases, and machine learning algorithms are commonly used in big data processing and analysis.
- Big data analytics involves the use of advanced analytics techniques, such as data mining, machine learning, and predictive modeling, to extract meaningful patterns and insights from large datasets.

NoSQL --> Not Only SQL
RDBMS --> OLTP (online Transaction Processing) / OLAP (online Analytical Processing) --> SQL (structured query language)
RAID --> Redundant Array of independent disks
DATA --> Applications --> OS File systems ---> Storage

Bigdata ---> data produced by the different devices and applications are as follows:

- Black box data --> a component of helicopter, airplane and jets, etc.., IT captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- Social media data
- Stock exchange data
- Search engine data
- Transport data
- Power grid data and many more where the data generation falls under the category of 3V's

The data will be of 3 types:

- Structured data ---> ALL RDBMS data
- Semi Structured Data --> XML data, JSON, YAML
- Unstructured data ---> word, pdf, text, media, logs ,etc

Big data technologies	latency	concurrency	access pattern	queries	"data scope"	"end user"	technology
Operational big data	1 ms - 100 ms	1000 - 100,000	write and reads	selective	operational	customer	NoSQL
Analytical big data	1 min - 100 min	1 - 10	reads	unselective	retrospective	ata scientists	Map Reduce, MPP database

Challenges we face with big data ---> capturing the data, curation, storage, searching, sharing, transfer, analysis, presentation
To address the above challenges organizations normally take the help of enterprise servers.

Big data skills:

1. Database ---> SQL (MySQL and PostgreSQL, Mariadb eg.,) and NoSQL databases (MongoDB, Casandra etc..,)
2. Big Data Technologies :
    a. Hadoop Ecosystem: Familiarity with components like HDFS, MapReduce and Hive
    b. Apache Spack: for fast and general purpose cluster computing
3. Data analysis and visualization:
    a. Data analytics tools : Apache Spack, Apache Flink and Hadopp for processing large datasets.
    b. Data visualization: tools like Tableau, Power BI or Matplotlib for presenting insights virtually
4. Programming languages:
    a. Python --> data analysis, Machine learning and scripting tasks
    b. Java --> commonly used for distributed computing framework like apache hadoop
5. Machine learning, statistical and mathematically skills
6. Data engineering:
    a. ETL (Extract , Transform, Load)--> designing and implementing ETL processes for moving and transforming data
    b. Data warehousing concepts and technologies
7. Cloud computing platforms(AWS, Azure or GCP) and BDaas (Big data as a service)

Sources of big data ---> Social media platforms, IoT, Web and E-Commerce, Business Transactions, Healthcare, Machine and sensor data, scientific research, weather and climate data, genomic data, government and public services etc.,,

- While the adoption of big data brings numerous benefits, it also presents challenges, including data privacy concerns, security issues, and the need for skilled professionals.
- Successful adoption requires a comprehensive strategy, investment in the right technologies, and a commitment to fostering a data-centric culture within the organization.

Research and Changing Nature of Data Repositories

The field of big data is evolving rapidly, and researchers are actively exploring new approaches and technologies to address the challenges posed by the increasing volume, velocity, and variety of data. Here are some key trends and research areas related to the changing nature of data repositories in the context of big data:

1. Distributed storage system (HDFS)
2. Data Replication and consistency Models (data integrity and availability in an distributed env)
3. Real-time data processing (handling the straining and storing of data -- Apache kafka { Distributed publish-subscribe messaging system} and Apache flink {platform for scalable batch and stream data processing})
4. Data privacy and security (protecting sensitive information and allowing secure multi-party computation)
5. Data Management -- Machine learning
6. Serverless computing for data repositories
7. Hybrid and multi-cloud data repositories
8. Analytics and graph databases

Data Sharing and Reuse Practices and Their Implications for Repository Data Curation

Data sharing and reuse practices are critical aspects of scientific research and play a significant role in advancing knowledge and fostering collaboration. Effective data sharing requires proper curation of data repositories to ensure that shared data is discoverable, accessible, and usable by others. Here are some key practices and their implications for repository data curation:

1. Data licensing and permissions
2. Metadata standards and documentations
3. Versioning (track changes and updates to datasets over time) and persistent identifies (provide a stable references for citing datasets)
4. Data Quality Assurance (data qualify checks and ensuring data integrity before sharing) (validation check, error correction and Documenation of data cleaning process to enhance the reliability of shared datasets.
5. Data Access Polices
6. Standardized data formats
7. Interoperability and data linkages
8. Long-term preservation strategies

9. Tracing and providing data usage metrics
10. Community engagement and feedback

By implementing these practices, data repositories contribute to a culture of responsible data sharing and reuse.
Effective curation ensures that shared data remains valuable, reliable, and accessible, benefiting both the original data creators and the wider research community.

T o be contined ..,

2. Introduction to Hadoop

Lecture

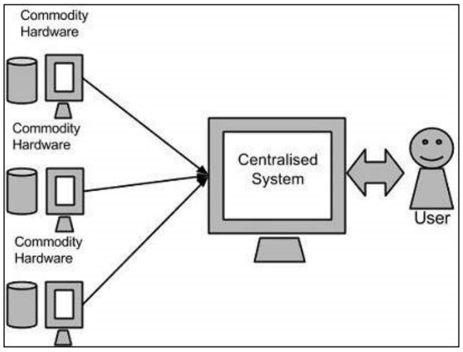
- A Brief History of Hadoop,
- Evolution of Hadoop,
- Introduction to Hadoop and its components
- Comparison with Other Systems,
- Hadoop Releases
- Hadoop Distributions and Vendors

Hadoop as a Big data Solutions

Tradition approach for handling big data and processing it:-

USERS ---> Applications ---> centralized systems --> Relation data base

Google's solutions

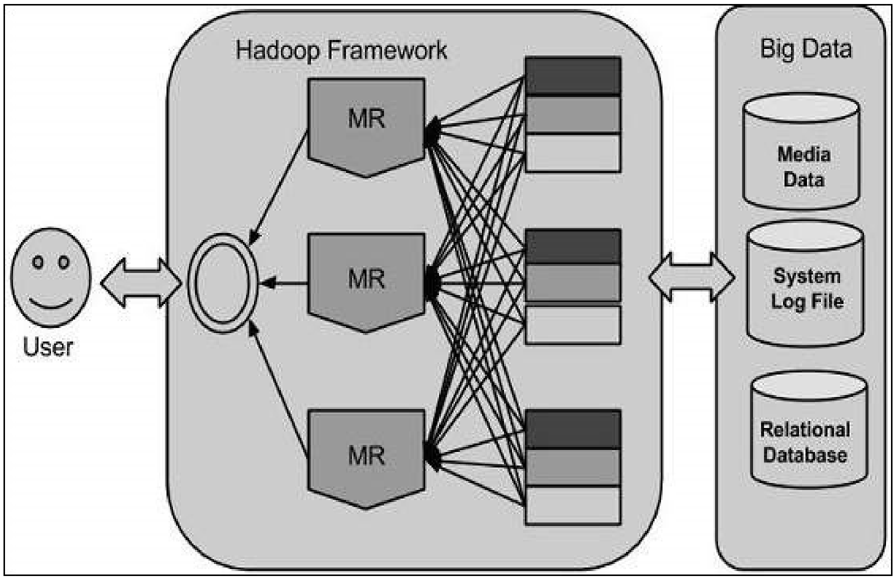


They used the algorithm called MapReduce. Algorithm divides the task into small parts and assigns them to many computers and collects the results from them which when integrated form the result datasets

Hadoop

Dough Cutting and his team developed an open source project called **HADDOP** by using the solutions provided by Google.

Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

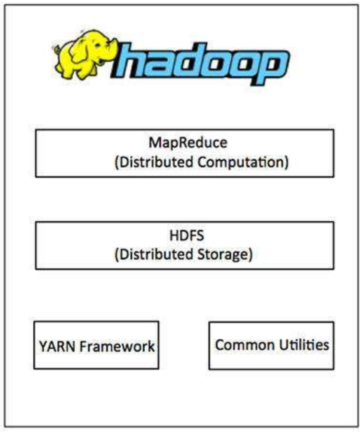


Hadoop is an Apache Open Source framework written in java which allows distributes processing of large datasets across clusters of computers using simple programming models.

Hadoop Architecture :-

At is core, Hadoop has two major layers:

- Processing / Computation layer (MapReduce)
  - Mapreduce --> parallel programming model for writing distributed applications (devise at google for efficient processing of large amounts of data - multi-terabyte data-sets) on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner (i.e, eliminating all type of Single Point of Failures)
  - MapRecude program runs on Hadoop which is an Apache Open source Framework
- Storage Layer (Hadoop Distributed File system)
  - HDFS - based on Google file systems (GFS) and provides a distributed file system that is designed to run on commodity hardware.
  - It is highlly fault-tolerant and designed to be deployed on a low-cost hardware.
  - High throughput access to application data and is suitable for applications having large datasets



The Hadoop framework application works in an env that provides distributed storage and computation across clusters of computers.
Designed to scale up form single server to thousands of machines, each offering local computation and storage.

- Hadoop YARN --> This is a framework for job scheduling and clusters resource management
- Hadoop common utilities --> java libraries and utilities required by other Hadoop modules