1. What is big data and hype?

Big Data refers to large data sets which are analyzed to understand data trends, which is also referred to as data mining, but data science utilizes machine learning algorithms to design and create statistical methods to generate information from big data that can be implemented to enhance business processes

The advances in analysing big data allow us to e.g. decode human DNA in minutes, find cures for cancer, accurately predict human behaviour, foil terrorist attacks, pinpoint marketing efforts and prevent diseases.

One of the most significant big data trends is using big data analytics to power AI/ML automation, both for consumer-facing needs and internal operations. Without the depth and breadth of big data, these automated tools would not have the training data necessary to replace human actions at an enterprise.

2. A.Sources of big data

This source includes all social media posts, videos posted etc. Machine (sensor) – this data comes from what can be measured by the equipment used. Transactional – this comes from the transactions which are undertaken by the organization. This is perhaps the most traditional of the sources.

B. Skills of big data:

Programming languages, quantitative analysis, data mining, data visualization, problem-solving, SQL/NoSQL databases, cloud computing, machine learning, and continuous learning are all essential skills for big data.

- Computer programming with languages like C++, Java, and Python.
- Databases and SQL.
- ETL and data warehousing.
- Talend, IBM DataStage, Pentaho, and Informatica.
- Operating system knowledge for Unix, Linux, Windows, and Solaris.
- Hadoop.
- Apache Spark.
- Data mining and modeling.

Sources of Big data

A significant part of big data is generated from three primary resources:

- **Machine data**
- **Social data, and**
- **Transactional data.**

Big data comes from various sources -- like are transaction processing systems, customer databases, documents, emails, medical records, internet clickstream logs, mobile apps and social networks.

How Does Big Data Analytics Work?

Companies need to work around analytics applications, partner with data scientists and engage with other data analysts to extract relevant and valid insights from big data. In addition, they must have an enhanced understanding of all available data. Finally, the analytics team also needs to clarify what they want to extract from the data.

- Cleansing,
- Profiling,
- Transformation,
- Validation of data sets.

3. Big data Adoption

Big data adoption is a process through which businesses find innovative ways to enhance productivity and predict risk to satisfy customers need more efficiently.

Big Data is a technology-driven movement and its strategic importance requires special focus and attention during its adoption:

Why do companies adapt big data?

Companies use big data in their systems to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits.

Adoption Framework:

A key factor in the success of any new program is the way it is approached from the inception phase itself. Any Big Data program that requires the integration of data with strategic planning is going to be critical and will carry heavy penalties in case of failure. The right

framework to enable the adoption of Big Data analytics within the organization must be adopted. The critical components of this framework include:

- **Data discovery**
- **Analytics discovery**
- **Tools and technology discovery**
- **Infrastructure discovery**
- **Implementation**

Recommendations for Big Data Adoption:

Driven by the need to solve business challenges, in light of both advancing technologies and the changing nature of data, banking and financial markets companies are starting to look closer at Big Data's potential benefits. To extract more value from Big Data, we offer a broad set of recommendations tailored to banks and financial markets firms.

- **Commit initial efforts to customer-centric outcomes.**
- **Define Big Data strategy with a business-centric blueprint.**
- **Start with existing data to achieve near-term results.**
- **Build Analytics capabilities based on business priorities.**
- **Create a business case based on measurable outcomes.**

4. Data Repositories

The data repository is a large database infrastructure, several databases that collect, manage, and store data sets for data analysis, sharing and reporting.

Research and Changing Nature of Data Repositories

The data repository uses structured organization methods, standardized schemas, and metadata to ensure that the data is always the same and easy to find. It has tools for storing, managing, and protecting data, such as compression, indexing, access controls, encryption, and reporting.

There are two main types of data repositories -

Transactional and Analytical. For high-volume day-to-day operational data such as banking transactions, Transactional, or OLTP, systems are the ideal choice

Examples of Data Repositories: The term data repository can be used to describe several ways to collect and store data:

● **A data warehouse** is a large data repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related.

A data warehouse, or enterprise data warehouse (EDW), is a system to aggregate your data from multiple sources so it's easy to access and analyze. Data warehouses typically store large amounts of historical data that can be queried by data engineers and business analysts for the purpose of business intelligence.

When to use a data warehouse

- store all of your historical data in a central repository
- analyze your web, mobile, CRM, and other applications together in a single place
- get deeper business insights than traditional analytics tools by querying data directly with SQL
- provide multiple people access to the same data set simultaneously

Key factors to analyze which data warehouse will best suit your business needs:

1. Data types

There are three types of data structured, unstructured, and semi-structured. Structured data is quantifiable data that can be organized neatly into rows and columns.

Unstructured data is data that can't be easily managed and analyzed. Think written content (like blog posts or answers to open-ended survey questions), images, videos, audio files, and PDFs. If you're looking to store purely unstructured data, you should consider a **data lake** instead of a **data warehouse**.

Semi-structured data is a mix of structured and unstructured data. Take an email, for example. The content of that email is unstructured, but there are quantifiable aspects to the email.

2. Scaling for data storage

Most data warehouses typically allow you to store massive amounts of data without much overhead cost.

3. Scaling for performance

The performance of a data warehouse refers to how fast your queries can run and how you maintain that speed in times of high demand.

4. Maintenance

You likely want your engineers focused on building and maintaining your products instead of worrying about ETL pipelines and day-to-day management of your warehouse—especially if

you have a small team. In that case, you'll want a data warehouse that is self-optimizing like BigQuery, Snowflake, or IBM Db2.

5. Ecosystem

Consider using a data warehouse that is within the ecosystem of the applications you already use. For example, Azure Synapse Analytics is in the ecosystem of Microsoft products, Redshift within AWS, and BigQuery within the Google Cloud ecosystem.

6. Cost

Many factors go into data warehouse pricing, including storage, warehouse size, run time, and queries. For Redshift, you pay per hour based on nodes or per bytes scanned. BigQuery, on the other hand, has both a flat-rate model and a per-query model. Snowflake, IBM Db2, and Azure are all based on storage and compute time.

B.Data Lake:

Data lakes are central data repositories used to store any and all raw data. A data lake has no predefined schema, so it retains all of the original attributes of the data collected, making it best suited for storing data that doesn't have an intended use case yet.

A data lake is a large data repository that stores unstructured data that is classified and tagged with metadata.

● **Data marts** are subsets of the data repository. These data marts are more targeted to what the data user needs and easier to use. Data marts also are more secure because they limit authorized users to isolated data sets. Those users cannot access all the data in the data repository.

● **Metadata repositories** store data about data and databases. The metadata explains where the data source, how it was captured, and what it represents.

● **Data cubes** are lists of data with three or more dimensions stored as a table — as you may find in a spreadsheet.

Benefits of Data Repositories

There is value to storing and analyzing data. Businesses can make decisions based upon more than anecdote and instinct

● Isolation allows for easier and faster data reporting or analysis because the data is clustered together.

● Database administrators have easier time tracking problems because data repositories are compartmentalized

● Data is preserved and archived.

Disadvantages of Data Repositories

There are several vulnerabilities that exist in data repositories that enterprises must manage effectively to mitigate potential data security risks, including:

● Growing data sets could slow down systems. Therefore, making sure database management systems can scale with data growth is necessary.

● A system crash could affect all the data. Backup the databases and isolate access applications so system risk is restrained.

● Unauthorized users can access all sensitive data more easily than if it was distributed across several locations.

5. A.Data curation

Data curation is the process of creating, organizing and maintaining data sets so they can be accessed and used by people looking for information. It involves collecting, structuring, indexing and cataloging data for users in an organization, group or the general public.

There are three steps in the data curation process:

- data identification,
- data cleansing
- data transformation

Big data curation tools **like hubs and Grooper** provide incredible value to existing document-based data workflows.

5B.Data sharing

Data sharing is the process of making the same data resources available to multiple applications, users, or organizations. It includes technologies, practices, legal frameworks, and cultural elements that facilitate secure data access for multiple entities without compromising data integrity.

Sharing information can be done through various communication methods such as face-to-face conversations, instant messaging, email, video conferencing, or company wikis.

Why is data sharing important?

Data sharing allows researchers to build upon the work of others rather than repeat already existing research. Sharing data also enables researchers to perform meta-analyses on the current research data.

Risk of data sharing?

Implementing data sharing is like opening the floodgates to a wide range of potential threats, such as hacking or malware. The more people who have access to the data, the more opportunities there are for unauthorized parties to access it.

Disadvantage of data sharing?

Sharing data and valuable information raises a multitude of risk factors for individuals and organizations. Some of the most common risks that occur are accidental sharing, employee data theft, ransomware, too much data access and more.

Benefits of Sharing Data

- Transparency. Scholarly publications and scientific claims are the descriptions of a research work and its conclusions, but by themselves do not provide full disclosure on how the research has been done. ...
- Collaboration.
- Research Acceleration.
- Reproducibility.
- Data Citation.

Challenges of sharing data?

1. Ensuring data security
2. Maintaining data privacy
3. Data interoperability issues
4. Managing data volume
5. Ensuring data accuracy and integrity
6. Navigating regulatory landscapes
7. Overcoming organizational silos
8. Addressing trust concerns
9. Cost implications of data sharing

5C.Data Reuse:

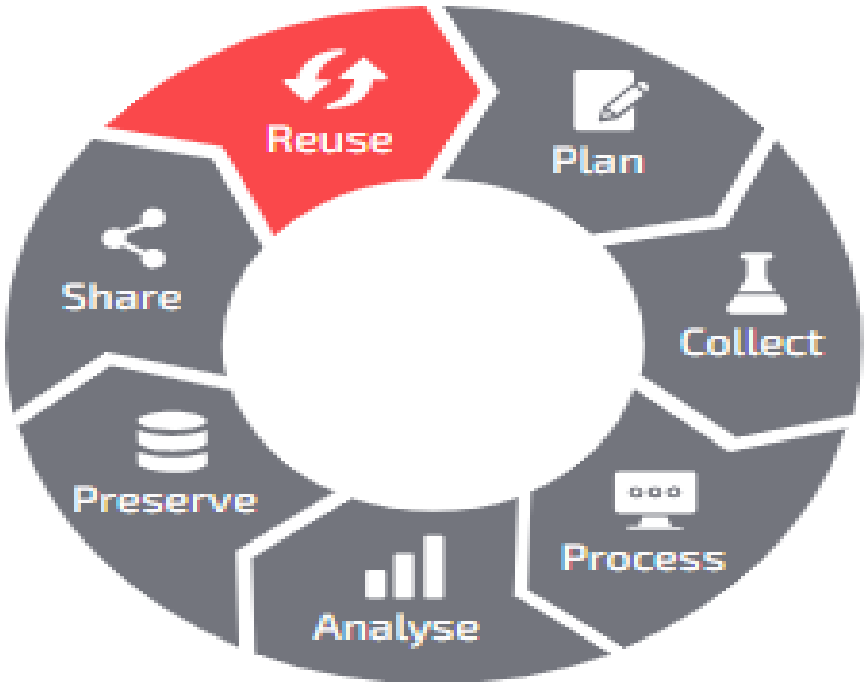
Data reuse means using data for other purposes than it was originally collected for. Reuse of data is particularly important in science, as it allows different researchers to analyse and publish findings based on the same data independently of one another. Reusability is one key component of the FAIR principles.

Why is data reuse important?

By reusing existing data you can:

- Obtain reference data for your research;
- Avoid doing new, unnecessary experiments;
- Run analyses to verify that reported findings are correct, and thereby making subsequent findings more robust;
- Make research more robust by aggregating results obtained from different methods or samples;

Data reuse life cycle:



6. Overlooked and Overrated Data Sharing

Data sharing is a critical aspect of modern society, but it can indeed be both overlooked and overrated depending on the context and the way it is approached.

A.Overlooked Data Sharing:

Privacy Concerns: In some cases, individuals or organizations may overlook data sharing because of legitimate concerns about privacy. Sharing sensitive personal or proprietary data can lead to security breaches and privacy violations.

Regulatory Constraints: In highly regulated industries, such as healthcare and finance, there are strict rules governing data sharing to protect sensitive information. These regulations can discourage data sharing and innovation in these sectors.

Lack of Incentives: Organizations may not be motivated to share data if there is no clear benefit or incentive for them. This lack of motivation can result in underutilization of valuable data resources.

B. Overrated Data Sharing:

Data Overload: The belief that more data is always better is not necessarily true. Collecting and sharing vast amounts of data without a clear purpose can lead to data overload, making it difficult to extract meaningful insights and potentially wasting resources.

Security Risks: Overrating data sharing can lead to sharing sensitive information without adequate security measures in place. This can result in data breaches and compromises in cybersecurity.

Data Monetization: While sharing data can be valuable, the idea of monetizing every piece of data may be overrated. Not all data is equally valuable, and the effort to monetize data can sometimes outweigh the benefits.

7. Data Curation Services in Action:

Data curation services are essential in a variety of fields, including research, healthcare, finance, and more, where data plays a critical role. They help organizations and researchers make the most of their data assets by ensuring data is well-organized, high-quality, secure, and readily available for analysis and decision-making.

Data curation services play a vital role in managing, organizing, and preserving data to ensure its quality, accessibility, and usefulness. Here's a glimpse of how data curation services work in action:

1. **Data Collection and Ingestion:**

Data curation services start by collecting and ingesting data from various sources. This may include databases, documents, web scraping, sensor networks, or any other data repositories.

2. **Data Cleaning and Quality Assurance:**

Once the data is collected, it undergoes a rigorous cleaning process. Data curation experts identify and rectify errors, inconsistencies, and duplicates. This step ensures that the data is accurate and reliable.

3. **Metadata Creation:**

Metadata is crucial for understanding and managing data effectively. Curation services create comprehensive metadata that describes the data's source, structure, and meaning. This metadata makes it easier to search and discover relevant data.

4. **Data Transformation:**

Data often needs to be transformed into a common format or structure to make it more usable. Data curation services may standardize data formats, convert unstructured data into structured data, and transform data for analysis.

5. **Data Organization and Indexing:**

Curation services organize data into logical categories or data sets. They also index the data to create an efficient and user-friendly search and retrieval system.

6. **Version Control and Data Provenance:**

Managing data versions is essential, especially in collaborative or dynamic environments. Data curation services keep track of data provenance and maintain version control to ensure traceability and data lineage.

7. **Access Control and Security:**

Data curation services implement access controls to protect sensitive data. They also ensure data security by applying encryption, authentication, and authorization measures.

8. **Data Preservation and Archiving:**

Data curation services are responsible for preserving data for the long term. This includes archiving data in secure repositories, ensuring data remains accessible even as technology evolves.

9. **Data Sharing and Collaboration:**

These services facilitate data sharing and collaboration by providing tools and platforms for users to access, analyse, and share data securely with others.

10. **Data Discovery and Visualization:**

Data curation services may provide tools and techniques for users to discover and visualize data, making it easier to derive insights from the curated data.

11. **Continuous Monitoring and Updates:**

Data is not static, and data curation is an ongoing process. Services continuously monitor data quality, update metadata, and adapt to changing requirements and technology.

12. **Compliance and Governance:**

Data curation services ensure that data complies with legal and regulatory requirements. They establish governance frameworks to manage data responsibly.

8.Open Exit: Reaching the End of the Data Life Cycle

The "Open Exit" in the context of the data life cycle refers to the responsible and transparent management of data when it reaches the end of its useful life.

The "Open Exit" approach emphasizes transparency, responsibility, and compliance in managing data at the end of its life cycle. It ensures that data is managed in a way that respects privacy, security, and legal requirements, while also promoting sustainable practices and good governance.

1. **Data Identification and Assessment:**

   The first step in the "Open Exit" process is identifying data that has reached the end of its life cycle. This involves assessing the data's value and potential for archival or disposal.

2. **Archiving for Historical or Compliance Purposes:**

   Some data, such as historical records or data required for compliance reasons, may need to be archived rather than immediately deleted. Archiving ensures that the data is preserved in a secure and accessible manner.

3. **Data Disposal:**

   Data that no longer serves any purpose or poses a security risk may be marked for disposal. Secure data disposal methods, such as shredding physical documents or securely erasing digital data, must be employed to prevent data breaches.

4. **Documentation and Records:**

   All actions taken during the data end-of-life process should be well-documented. This includes recording what data was archived, what was disposed of, and the methods used for disposal.

5. **Data Transition:**

   In some cases, data may need to transition from one system or format to another, such as migrating from legacy systems to newer technology. This transition should be carefully managed to ensure data integrity and availability.

6. **Data Privacy and Compliance:**

   Considerations for data privacy and compliance should continue during the end-of-life process. Sensitive data must be protected and disposed of in accordance with legal and regulatory requirements.

7. **Communication and Transparency:**

Transparency is a key aspect of the "Open Exit" concept. Organizations and data custodians should communicate their data management and disposal processes to relevant stakeholders, including users and regulators.

8. **Data Recovery in Emergencies:**

   Even after data has reached the end of its life cycle, there may be instances where data recovery becomes necessary, such as in the event of legal disputes or unforeseen events. Plans for data recovery should be in place.

9. **Continuous Monitoring:**

   Data end-of-life management should be an ongoing process, with regular reviews to ensure that data is appropriately archived or disposed of as circumstances change.

10. **Environmental Responsibility:**

    When disposing of physical data storage media, consider environmental responsibility by recycling or disposing of hardware in an environmentally friendly manner.

Data lifecycle Management

A data lifecycle refers to the different stages a unit of data undergoes, from initial collection to when it's no longer considered useful and deleted. It's a continuous, policy-based process where each phase informs the next.

Data lifecycle management (DLM) refers to the policies, tools, and internal training that helps dictate the data lifecycle. It's essentially the framework for managing how data is collected, cleaned, stored, used, and eventually deleted.

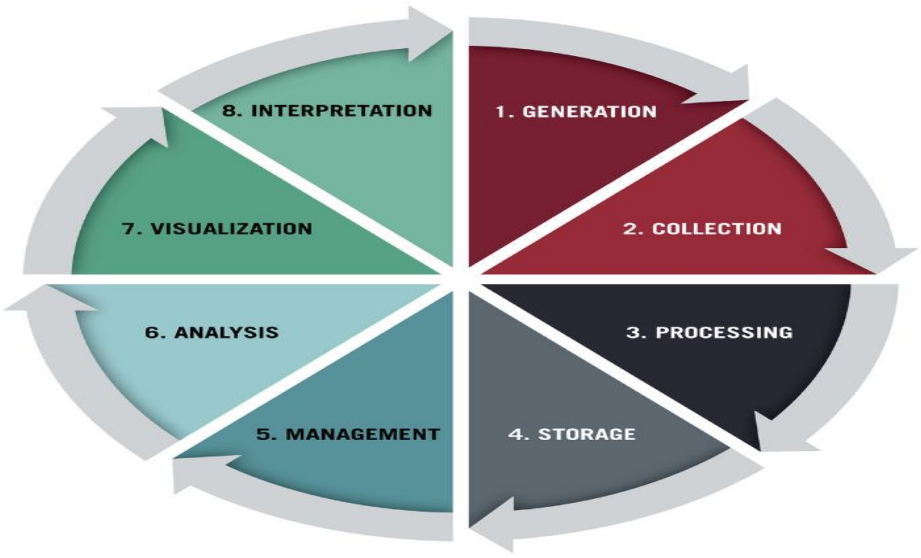
Efficient lifecycle management helps you maximize resources and ensure accuracy.

1. **Generation**

How data will be generated in real time and how user will interact with the API.

2. Collection

The second stage in the data lifecycle is collecting customer data from various internal and external sources. Depending on what you prefer, and whether you populate your database manually or automatically, this stage can also be called data creation, data acquisition, or data entry.



3. Processing

The next step is to process your data so it becomes usable. Data processing falls into three core categories: encryption, wrangling, and compression.

- **Data encryption**: scrambling or translating human-readable data into a format that can only be decoded by authorized personnel.
- **Data wrangling**: cleaning and transforming data from its raw state into a more accessible and functional format.
- **Data compression**: reducing the size of a piece of data and making it easier to store by restructuring or re-encoding it.

4. Storage

Once data is collected, it's time to store it. Efficient data lifecycle management helps create a single source of truth within an organization by storing data in a central repository.

The type of data you collect will determine where it should be stored.

- **For structured data** (like the kind that would fit seamlessly into an Excel sheet), should be stored in a **relational database or a data warehouse**.
- **For unstructured data** (like images, text files, audio, etc.) should be stored in a non-relational database or a data lake.

5. Data Management/Archiving

Archiving involves moving data from all active deployment environments into an archive. At this point, the data is no longer operationally useful, but instead of destroying it, you're keeping it in a long-term storage location.

Archived data is useful for future reference or analysis, but it can also pose a security risk for businesses if their systems get breached. The solution is to use a tool that prioritizes data privacy and security, so unauthorized personnel don't gain access to sensitive data.

6. Analysis

Data analysis involves studying processed or raw data to identify trends and patterns. Some of the techniques you can use at this stage include machine learning, statistical modelling, artificial intelligence, data mining, and algorithms.

7. Visualization:

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Visualize the data and get the fact about the data using graph.

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations.

8. Interpretation

Data interpretation is the process of reviewing data and arriving at relevant conclusions using various analytical research methods. Data analysis assists researchers in categorizing, manipulating data, and summarizing data to answer critical questions.

There are four steps to data interpretation:

1) Assemble the information. 2) Develop findings

3) Develop conclusions 4) Develop recommendations

9. The Current State of Meta-Repositories for Data

A metadata repository is a centralized storage and management system for metadata. Metadata is data about data, such as the structure, origin, usage, and relationships of data elements. A metadata repository enables you to access, update, and share metadata across different platforms, tools, and users.

A Meta data repository (MDR) is a component which manages Meta data. MDRs can manage information about the processes which create, use, or update the data, the hardware components that host these processes or the database system, or other (human) resources which make use of the data.

A metadata repository is a software tool that stores descriptive information about the data model used to store and share metadata. Metadata repositories combine diagrams and text, enabling metadata integration and change.

A metadata repository should contain the following: A description of the data warehouse structure, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.

The role of Meta data repository in data warehouse?

Metadata in data warehouses is the backbone of effective data management. It provides valuable insights into the data stored within the system and enhances data understanding, governance, integration, and accessibility. A data repository is also known as a data library or data archive.

The concept of meta-repositories for data, which refers to repositories that **aggregate and index data** from various sources, was in development and evolution. Here's the general state of meta-repositories for data as of that time:

1. **Data Aggregation and Indexing:** Meta-repositories are platforms or services designed to collect and index data from a wide range of sources, making it easier for researchers, organizations, and individuals to discover and access data from a single location. Examples of these platforms include data.gov, DataCite, and re3data.
2. **Diverse Data Types:** Meta-repositories are increasingly accommodating diverse data types, including structured data, unstructured data, and multimedia data. This diversity helps cater to the needs of various domains and research areas.
3. **Interoperability:** Many meta-repositories strive to ensure interoperability by using standardized metadata schemas and APIs. This makes it easier for users to search, access, and integrate data from multiple repositories into their work.
4. **Open Data and Open Science:** Open data initiatives have gained momentum, and many meta-repositories focus on providing open access to data. Open science principles, which promote transparency and collaboration, often guide the development of these platforms.
5. **Collaborative Data Sharing:** Some meta-repositories encourage collaborative data sharing, allowing users to upload, share, and collaborate on datasets. These platforms often incorporate version control and data provenance features.
6. **Data Quality and Curation:** Maintaining data quality and curation is a priority for many meta-repositories. They may include tools and processes for data curation, validation, and quality assessment.

7. **Data Citation:** Efforts to standardize data citation are becoming more common. Many meta-repositories support the assignment of Digital Object Identifiers (DOIs) to datasets, which makes it easier to cite and reference data in research publications.
8. **Integration with Analysis Tools:** Some meta-repositories offer integration with data analysis tools, enabling users to analyze and visualize data directly within the platform.
9. **Security and Privacy:** Data security and privacy are important considerations. Meta-repositories implement security measures to protect sensitive data and adhere to privacy regulations.
10. **AI and Machine Learning:** Some meta-repositories are exploring the use of AI and machine learning for data discovery and recommendation, helping users find datasets that are relevant to their research.
11. **Sustainability:** Ensuring the sustainability of meta-repositories is a challenge. Many rely on funding and support from governments, institutions, or organizations, and they need sustainable business models.

11. <u>Curation of Scientific Data at Risk of Loss: Data Rescue and Dissemination</u>

Data rescue is the process of securing data at risk of being lost due to deterioration or simple obsolescence of the storage media, natural hazards, theft or vicious destruction, and ensuring that data can be easily accessed and used.

Data rescue, also known as data recovery or data salvage, involves the process of retrieving, preserving, and restoring data that may be at risk of being lost due to various factors, including technological obsolescence, data degradation, or natural disasters. This process is crucial for maintaining historical, scientific, and cultural data for future use. Key components of data rescue include:

1. **Identification of At-Risk Data:** The first step in data rescue is identifying datasets that are at risk of being lost. This could include data stored on outdated media, in obsolete file formats, or in deteriorating physical records.
2. **Data Recovery:** Once at-risk data is identified, efforts are made to recover and preserve it. This may involve digitizing physical records, converting data to modern formats, and ensuring that the data remains accessible for the long term.
3. **Metadata Documentation:** It's essential to document metadata associated with rescued data. Metadata provides context about the data, its source, and any changes made during the recovery process.

4. **Quality Control:** Data quality checks are performed to ensure that the rescued data is accurate and complete. This may involve data validation and verification.
5. **Archiving and Storage:** Rescued data is typically archived in a secure and reliable storage system to prevent future data loss. Data may be stored in repositories or archives designed for long-term preservation.
6. **Access and Distribution:** Making the rescued data accessible to relevant users is a key aspect of data rescue. This may involve creating data catalogs, providing access through online platforms, and ensuring that researchers and the public can use the data.

Data Dissemination:

Data dissemination is the process of sharing data with a wider audience, whether it's for research, decision-making, or public access. The goal is to make data available in a user-friendly and accessible manner while addressing privacy and security concerns. Key components of data dissemination include:

Data Catalogs and Portals: Data dissemination often starts with creating data catalogs or portals that provide detailed information about available datasets. These catalogs help users discover and select relevant data.

Access Controls: Depending on the type of data, access controls and security measures may be in place to protect sensitive information. Access may be restricted to authorized users or made publicly available.

Data Formats: Data may be disseminated in various formats, including raw data files, APIs (Application Programming Interfaces), and interactive data visualization tools to accommodate different user needs.

Metadata and Documentation: Comprehensive metadata and documentation are essential for understanding the data, including its source, purpose, and structure. This information aids in data discovery and interpretation.

Data Sharing Agreements: In some cases, data dissemination may involve agreements, licenses, or terms of use that define how the data can be utilized by external parties.

Data Preservation: Ensuring the long-term preservation of disseminated data is crucial. This includes regular backups, version control, and monitoring for data degradation.

Data Accessibility: Efforts should be made to make data as accessible as possible, including compliance with accessibility standards for individuals with disabilities.

Data rescue and dissemination play a vital role in preserving valuable data assets and making them available for research, analysis, and decision-making. These processes contribute to data transparency, scientific progress, and informed decision-making in various fields.

Steps of data recovery

- Stop using all affected devices. ...
- Record details on what happened. ...
- Decide which data recovery method to use. ...
- Contact an IT professional for support. ...
- Prevent future data incidents.

Why do we need data recovery?

Data recovery is an aspect of backup and recovery and an integral component of your overall disaster recovery plan (DRP). Companies rely on data to inform business decisions and to support day-to-day operations. As such, any data loss can seriously impact continuity, which is why data recovery is so critical.

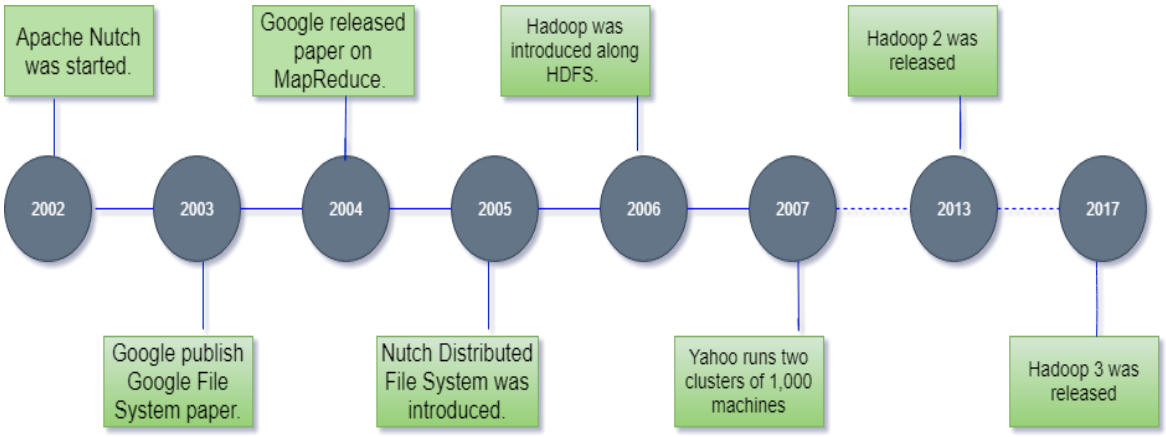
Introduction to Hadoop

Hadoop is an open source framework based on Java that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time. It is the most commonly used software to handle Big Data.

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

History of Hadoop

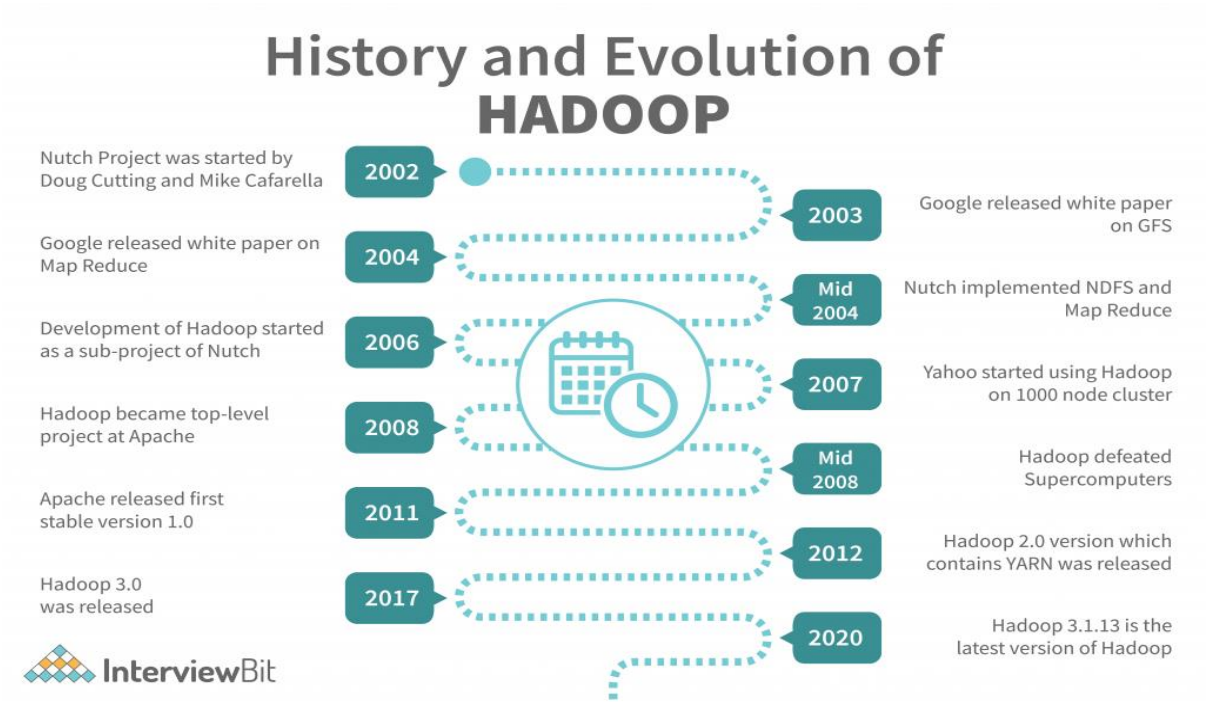
The Hadoop was started by Doug Cutting and Mike Cafarella in 2002. Its origin was the Google File System paper, published by Google.



Year	Event
2003	Google released the paper, Google File System (GFS).
2004	Google released a white paper on Map Reduce.
2006	Hadoop introduced. Hadoop 0.1.0 released. Yahoo deploys 300 machines and within this year reaches 600 machines.
2007	Yahoo runs 2 clusters of 1000 machines. Hadoop includes HBase.
2008	YARN JIRA opened Hadoop becomes the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds. Yahoo clusters loaded with 10 terabytes per day. Cloudera was founded as a Hadoop distributor.
2009	Yahoo runs 17 clusters of 24,000 machines. Hadoop becomes capable enough to sort a petabyte.

	Map Reduce and HDFS become separate subproject.
2010	Hadoop added the support for Kerberos. Hadoop operates 4,000 nodes with 40 petabytes. Apache Hive and Pig released.
2011	Apache Zookeeper released. Yahoo has 42,000 Hadoop nodes and hundreds of petabytes of storage.
2012	Apache Hadoop 1.0 version released.
2013	Apache Hadoop 2.2 version released.
2014	Apache Hadoop 2.6 version released.
2015	Apache Hadoop 2.7 version released.
2017	Apache Hadoop 3.0 version released.
2018	Apache Hadoop 3.1 version released.

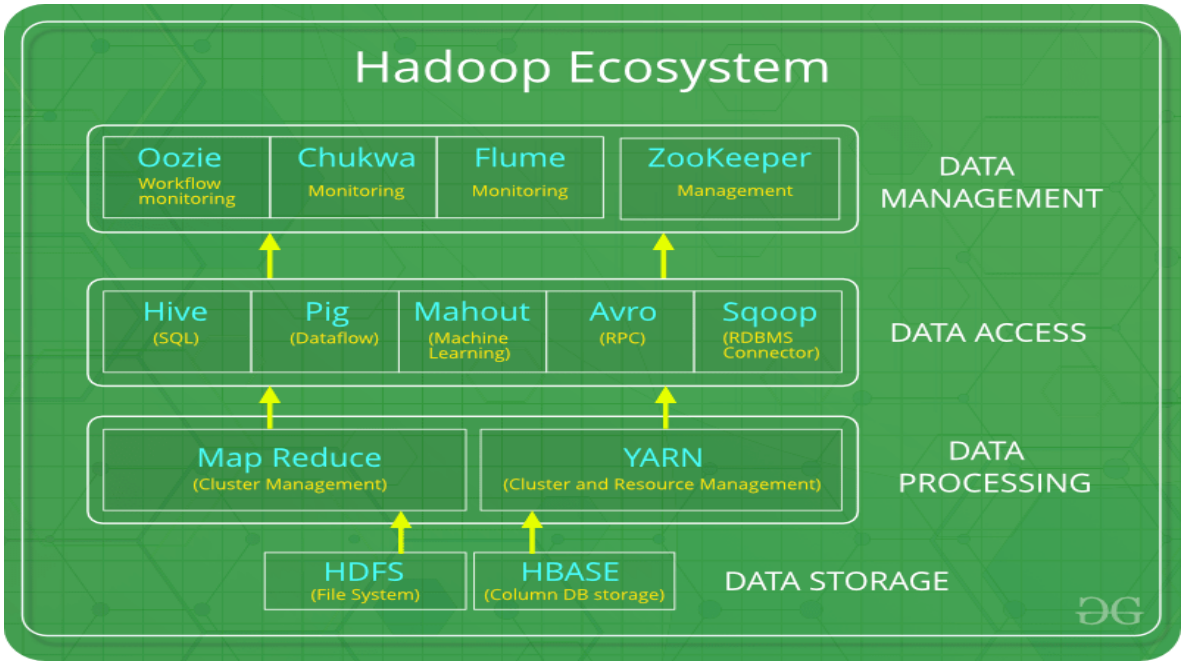
Evolution of Hadoop



Hadoop and its components

Apache Hadoop is a Java-based open-source platform for storing and analyzing big sets of unstructured data.

Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common.



Following are the components that collectively form a Hadoop ecosystem:

- **HDFS:** Hadoop Distributed File System
- **YARN**: Yet another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing
- **PIG, HIVE**: Query based processing of data services
- **HBase:** NoSQL Database
- **Mahout, Spark MLLib**: Machine Learning algorithm libraries
- **Solar, Lucene**: Searching and Indexing
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling

1.HDFS:

HDFS is the primary or major component of Hadoop ecosystem and is responsible for

storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.

HDFS consists of two core components i.e.

A. **Name node**
B. **Data Node**

Name Node is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the data nodes that stores the actual data. These data nodes are commodity hardware in the distributed environment. Undoubtedly, making Hadoop cost effective.

HDFS maintains all the coordination between the clusters and hardware, thus working at the heart of the system.

2.YARN:

Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.

Consists of three major components i.e.

- Resource Manager
- Nodes Manager
- Application Manager

Resource manager has the privilege of allocating resources for the applications in a system whereas Node managers work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager.

3.MapReduce:

By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing's logic and helps to write applications which transform big data sets into a manageable one.

Map Reduce makes the use of two functions i.e. Map() and Reduce() whose task is:

1. ***Map()*** performs sorting and filtering of data and thereby organizing them in the form of group. Map generates a key-value pair based result which is later on processed by the Reduce() method.

2. ***Reduce(),*** as the name suggests does the summarization by aggregating the mapped data. In simple, Reduce() takes the output generated by Map() as input and combines those tuples into smaller set of tuples.

4.PIG:

Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.

- It is a platform for structuring the data flow, processing and analysing huge data sets.
- Pig does the work of executing commands and in the background, all the activities of Map Reduce are taken care of. After the processing, pig stores the result in HDFS.
- Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the JVM.
- Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

5.HIVE:

- With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).
- It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL datatypes are supported by Hive thus, making the query processing easier.

6.Mahout:

Mahout, allows Machine Learnability to a system or application. Machine Learning, as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.

- It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning. It allows invoking algorithms as per our need with the help of its own libraries.

7.Apache Spark:

It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.

- It consumes in memory resources hence, thus being faster than the prior in terms of optimization.
- Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing, hence both are used in most of the companies interchangeably.

8.Apache HBase:

It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.

- At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data

Other Components: Apart from all of these, there are some other components too that carry out a huge task in order to make Hadoop capable of processing large datasets. They are as follows:

Solr, Lucene: These are the two services that perform the task of searching and indexing with the help of some java libraries, especially Lucene is based on Java which allows spell check mechanism, as well. However, Lucene is driven by Solr.

- **Zookeeper:** There was a huge issue of management of coordination and synchronization among the resources or the components of Hadoop which resulted in inconsistency, often. Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.
- **Oozie:** Oozie simply performs the task of a scheduler, thus scheduling jobs and binding them together as a single unit. There is two kinds of jobs .i.e Oozie workflow and Oozie coordinator jobs. Oozie workflow is the jobs that need to be executed in a sequentially ordered manner whereas Oozie Coordinator jobs are those that are triggered when some data or external stimulus is given to it.

Comparison with Other Systems

Hadoop is a popular and widely used system for distributed data storage and processing. However, it's important to understand that Hadoop is just one component of the broader big data ecosystem, and there are other systems and technologies that serve similar or complementary purposes. Here's a comparison of Hadoop with some other systems:

1. **Apache Spark:**
    - Hadoop: Hadoop's MapReduce is primarily designed for batch processing of large datasets.

- **Spark:** Apache Spark is a data processing framework that provides both batch processing and real-time stream processing capabilities. It's known for its speed and in-memory processing, making it suitable for a wide range of use cases.

2. **NoSQL Databases (e.g., MongoDB, Cassandra):**
   - **Hadoop:** Hadoop's HDFS (Hadoop Distributed File System) is designed for distributed storage but is not a database system. HBase, a component of the Hadoop ecosystem, is a NoSQL database that can be used with Hadoop.
   - **NoSQL Databases:** NoSQL databases are designed for efficient storage and retrieval of structured or semi-structured data. They are suitable for various applications, such as web applications and IoT data storage.
3. **Distributed Data Warehouses (e.g., Amazon Redshift, Google BigQuery):**
   - **Hadoop:** Hadoop focuses on data processing and requires additional tools for data warehousing capabilities.
   - **Distributed Data Warehouses:** These systems are optimized for analytical queries and are typically used for business intelligence and reporting tasks. They are well-suited for ad-hoc SQL queries.
4. **Kafka (for Real-time Data Streams):**
   - **Hadoop:** Hadoop is not a real-time data streaming platform.
   - **Kafka:** Apache Kafka is a distributed streaming platform used for real-time data ingestion and processing. It can work alongside Hadoop for real-time data pipelines.
5. **Machine Learning Frameworks (e.g., TensorFlow, PyTorch):**
   - **Hadoop:** While Hadoop can be used for pre-processing and feature extraction for machine learning, it is not a machine learning framework.
   - **Machine Learning Frameworks:** These are designed for developing and training machine learning models. They can be integrated with Hadoop for analysing and applying machine learning models to big data.
6. **Flink and Storm (for Real-time Stream Processing):**
   - Hadoop: Hadoop's traditional MapReduce is not suitable for real-time stream processing.

- **Flink and Storm:** These systems are designed for real-time stream processing and complex event processing, making them suitable for applications that require low-latency data analysis.

7. **Cloud-based Data Services (e.g., AWS S3, Azure Data Lake Storage):**
    - **Hadoop:** Hadoop can be deployed on cloud infrastructure, and cloud-based storage services can be integrated with Hadoop clusters.
    - **Cloud-based Data Services:** These services provide scalable and cost-effective storage solutions, often integrated with data analytics and processing services on the cloud.

Hadoop Release

Hadoop is an open-source framework for distributed storage and processing of large datasets. The project is maintained by the Apache Software Foundation, and it has had several major releases over the years. Here are some of the notable Hadoop releases up to my knowledge cutoff date in January 2022:

Hadoop 0.1.0 (December 2005): The initial release of Hadoop, marking the beginning of the project.

Hadoop 0.20.0 (February 2010): This release introduced several enhancements, including improved security and performance, and marked a significant milestone in the project's development.

Hadoop 1.0.0 (December 2011): This release signified Hadoop's transition from the 0.x version to 1.x. It included many new features and improvements.

Hadoop 2.0.0 (October 2012): This release introduced the YARN (Yet another Resource Negotiator) resource management framework, which decoupled resource management from the MapReduce application, making Hadoop more versatile and suitable for a broader range of applications.

Hadoop 2.2.0 (October 2013): This release brought several enhancements and stability improvements to Hadoop, solidifying YARN as a key component.

Hadoop 2.7.0 (April 2015): This release included various updates and improvements, further enhancing the capabilities and performance of Hadoop.

Hadoop 2.9.0 (January 2018): The 2.9.0 release continued to refine the Hadoop framework and introduce features like improved support for erasure coding.

Hadoop 3.0.0 (December 2017): The 3.0.0 release was a significant milestone, as it marked the transition to Hadoop 3, introducing several new features and improvements. Notable additions included support for erasure coding, enhanced resource management, and improved storage.

Hadoop 3.1.0 (April 2018): This release built upon the Hadoop 3.0.0 release, adding more features and enhancements.

Hadoop 3.2.0 (December 2018): The 3.2.0 release continued to improve Hadoop's capabilities and stability.

Hadoop 3.3.0 (July 2020): This release introduced further enhancements, performance improvements, and bug fixes.

Hadoop 3.3.1 (September 2020): A maintenance release following the 3.3.0 release, focusing on bug fixes and stability.

Hadoop 3.3.2 (October 2020): Another maintenance release with additional bug fixes and improvements.

Latest is just announced 3.3.6 2023 Jun 23

Hadoop Distributions and Vendors

Hadoop Distributions: Hadoop is Apache software so it is freely available for download and use. There are following advantages of distribution:

- **Distributions provide easy to install mediums like RPMs**

  The Apache version of Hadoop is just TAR balls. Distros actually package it nicely into easy to install packages which make it easy for system administrators to manage effectively.

- **Distros package multiple components that work well together**

  The Hadoop ecosystem contains a lot of components (HBase, Pig, Hive, Zookeeper, etc.) which are being developed independently and have their own release schedules. Also, there are version dependencies among the components. For example version 0.92 of HBase needs a particular version of HDFS.

- **Tested**

  Distro makers strive to ensure good quality components.

- **Performance patches**

Sometimes, distros lead the way by including performance patches to the 'vanilla' versions.

- **Predictable upgrade path**

  Distros have predictable product release road maps. This ensures they keep up with developments and bug fixes.

- **SUPPORT**

  Lot of distros come with support, which could be very valuable for a production critical cluster.

Hadoop Vendors

1. Cloudera: Cloudera is a prominent vendor that offers the Cloudera Data Platform (CDP). They provide a comprehensive Hadoop distribution, which includes tools for data management, analytics, and machine learning. Cloudera also offers support and consulting services.

2. Hortonworks (Now Part of Cloudera): Hortonworks was a major Hadoop distribution vendor, but it has since merged with Cloudera. The combined company provides a unified platform for big data and analytics.

3. MapR (Acquired by HPE): MapR was known for its enterprise-grade Hadoop distribution and other big data solutions. Hewlett Packard Enterprise (HPE) acquired MapR's assets.

4.IBM BigInsights: IBM's Hadoop distribution, called IBM BigInsights, provides an integrated platform for data processing and analytics. IBM also offers various analytics and data science tools.

5.Amazon EMR: Amazon Elastic MapReduce (EMR) is a cloud-based Hadoop distribution provided by Amazon Web Services (AWS). It allows users to spin up Hadoop clusters on-demand for big data processing.

6.Microsoft Azure HDInsight: Azure HDInsight is Microsoft's cloud-based Hadoop distribution, available on the Azure cloud platform. It offers managed Hadoop clusters with integration into the Microsoft ecosystem.

7.Google Cloud Dataprep: Google's cloud platform provides cloud-based data processing and analytics tools, including managed Hadoop clusters using Google Cloud Dataprep.

8. Apache BigTop: While not a vendor, Apache BigTop is an open-source project aimed at creating a packaging and testing framework for Hadoop and related big data projects. It can be used to create custom Hadoop distributions.

MapReduce Online: MapReduce Online is a distribution and service provider for Hadoop, offering Hadoop-as-a-Service, which simplifies Hadoop cluster management and maintenance.

9. Pivotal HD (Discontinued): Pivotal HD was an enterprise Hadoop distribution by Pivotal, but the project was discontinued, and Pivotal was acquired by VMware.

10. Oracle Big Data: Oracle offers a Hadoop distribution as part of its big data and analytics platform, providing integration with Oracle Database and other Oracle products.

list of Top vendors offering Big Data Hadoop solutions are:

1) Amazon Elastic Map Reduce
2) Cloudera CDH Hadoop Distribution
3) Hortonworks Data Platform (HDP)
4) MapR Hadoop Distribution
5) IBM Open Platform
6) Microsoft Azure's HDInsight -Cloud based Hadoop Distribution
7) Pivotal Big Data Suite
8) Datameer Professional
9) Datastax Enterprise Analytics
10) Dell- Cloudera Apache Hadoop Solution.