Big data materials

Fortunately, there are different big data technologies that have been developed by different companies and organizations to effectively manage and analyze big data. There are different big data technologies that have been developed by different companies and organizations to effectively manage and analyze big data. Below is a list of some of the most popular big data technologies: 1. Hadoop Hadoop is an open-source big data tool developed by the Apache Foundation. It is perfect for storing and processing large amounts of data in a parallel and distributed way. Hadoop includes two main components: Hadoop Distributed File System (HDFS): This is the file system that Hadoop uses to store data. This is the file system that Hadoop uses to store data. MapReduce: This is the processing layer where data processing takes place. Hadoop is not just one tool, but a complete framework that provides different tools and components to manage big data. Below is a list of some of the most popular Hadoop components: 2. Hive Hive is a data warehousing tool that is used to store data in a structured format for easier analysis. It is built on top of Hadoop and uses the HDFS file system to store data. Hive provides a SQL-like query language known as HiveQL, which is used to query data stored in Hive tables. 3. HBase HBase is a NoSQL database that is used to store structured and semi-structured data in a columnar format. It is built on top of Hadoop and uses the HDFS file system to store data. 4. Pig Pig is a data processing tool that is used to process data stored in Hadoop. It is built on top of Hadoop and uses the MapReduce processing engine to process data. Pig provides a scripting language known as Pig Latin, which is used to write data processing scripts. 5. Sqoop Sqoop is a data transfer tool that is used to transfer data between Hadoop and relational databases. 6. Flume Flume is a data ingestion tool that is used to ingest data into Hadoop from different data sources. 7. ZooKeeper ZooKeeper is a coordination service that is used to coordinate different Hadoop services. 8. Ambari Ambari is a management tool that is used to manage Hadoop clusters. 9. Oozie Oozie is a workflow management tool that is used to manage Hadoop workflows. 10. Mahout Mahout is a machine learning tool that is used to build machine learning models on top of Hadoop.

The technology of BDA has been advancing at a rapid pace in recent years. Data management and computation have become more efficient and effective, and new platforms and services have been developed to support BDA.

Data management: Data management is a critical component of BDA. In order to effectively analyze big data, it must be properly managed. Data management includes tasks such as data collection, data storage, data processing, and data analysis.

Data collection: Data collection is the process of gathering data from various sources. Data can be collected manually or automatically. Manual data collection is often time-consuming and expensive. Automatic data collection, on the other hand, is more efficient and can be done on a large scale.

Data storage: Data storage is the process of storing data in a secure and accessible location. Data can be stored locally or remotely. Local data storage is often more convenient and faster, but it is limited in terms of capacity and security. Remote data storage is more secure and can be accessed from anywhere, but it is often slower and more expensive.

Data processing: Data processing is the process of transforming data into a format that can be analyzed. Data processing includes tasks such as data cleaning, data transformation, and data mining.

Data analysis: Data analysis is the process of extracting insights from data. Data analysis can be done using statistical methods, machine learning, or artificial intelligence. Computation: Computation is a critical component of BDA. In order to effectively analyze big data, it must be properly processed. Computation includes tasks such as data collection, data storage, data processing, and data analysis.

Examples may include data management, computation and/or general purpose

BDA service platforms such as Hadoop, Spark, Flink, Tensorflow, etc. Providers that use this service type will generally be responsible for providing at least the following components: -

# Data store(s) –

# Data ingest –

# –

# Data output –

Data security

In general, the provider will also be responsible for providing an interface for the user/tenant to configure, schedule and monitor the data processing workflows.

the technological/technical aspects of BDA. Examples may include data management, computation and/or general purpose BDA service platforms.

Data Management This is the process of collecting, cleaning, validating and storing data. These activities may require manual effort and/or application of automated data processing techniques. Data management also covers the transformation of data into formats that can be processed by various BDA algorithms.

### Computation This is the process of executing BDA algorithms as well as providing infrastructural support for running such algorithms

. ### General Purpose BDA Services This is a category of services that provide basic functionality that may be commonly used by different BDA applications. Examples include storage, data transformation, data visualization and collaboration.

### Data Analytics Process Data analytics is more than just running an algorithm. It involves a cycle of activities that include data acquisition, data preparation, modeling and

## BDA Applications BDA can be used in a wide variety of domains. This section provides examples of BDA applications in several domains.

### Health Care BDA has been used in health care to solve problems such as identifying diseases, predicting health outcomes, improving patient care and reducing health care costs.

#### Identifying Disease BDA can be used to identify diseases by analyzing data from patient medical records. For example, NLP techniques can be used to automatically extract information from medical records and to identify patterns that may be indicative of a disease.

deployment.

NameNode manages the namespace of file systemas it serves as the Master Node and keeps all meta data information about the physical file. For read operation Reduce process will collect the meta data information from file namespace available on NameNode, gathers the data from data nodes, consolidates all the data to present it to the client. NameNode performs mapping of blocks on various data nodes and also stores the metadata information and serve the client to perform various operations. In larger Hadoop cluster where there is large number of Worker nodes, NameNode functionality is divided into dedicated primary NameNode server which keeps the hierarchy of file system index and other metadata information about the files and a secondary NameNode which after a certain time interval generates the snapshot of primary NameNode's memory structure. 1 Namenode NameNode can be considered as heart of HDFS file system because Name node consist of all master or metadata information about the user data. While writing data to HDFS, to host the block replicas the HDFS client requests the N ameNode to nominate a set of three DataNode and then client writes data to the DataNodes in a pipeline fashion [14].2. 2 Datanode In Hadoop Data nodes are primarily responsible to create, replicate and delete the data of file. NameNode keeps track of all metadata information about the partitioned blocks stored on data nodes. Whenever any file system of data node is formatted, in the file system hierarchy of data node, new namespace ID is assigned to it. Data nodes keeps on sending communications pulses to NameNode, NameNode gives all type of instruction while acknowledging the DataNode.

There are many technological and technical aspects to big data analytics. The technology side includes the various tools and platforms used to collect, store, and analyze big data. The technical side involves the various algorithms and methods used to process and interpret big data.