**Assignment-1**

**Q.1) What is Big Data? What are the features of Big Data?**

* **Big Data:** Big data refers to large and complex datasets that cannot be processed, stored, or analyzed using traditional data processing tools and techniques. It encompasses a variety of data types, including structured, semi-structured, and unstructured data, and is generated at a high volume, velocity, and variety from a variety of sources, including social media, sensors, machines, and devices. Big data requires specialized tools and technologies to store, manage, process, and analyze the data effectively. These tools may include distributed computing systems, NoSQL databases, machine learning algorithms, and data visualization tools, among others. The insights derived from big data can help organizations make better business decisions, optimize operations, and develop new products and services.
* **Features of Big Data:**
* Big Data platform should be able to accommodate new platforms and tool based on the business requirement. Because business needs can change due to new technologies or due to change in business process.
* It should support linear scale-out.
* It should have capability for rapid deployment.
* It should support variety of data format.
* Platform should provide data analysis and reporting tools.
* It should provide real-time data analysis software

**Q.2) Explain Type of Digital data with Example?**

* **There are 3 types of digital data, including:**

1. **Structured Data :** Structured data is created using a fixed schema and is maintained in tabular format. The elements in structured data are addressable for effective analysis. It contains all the data which can be stored in the SQL database in a tabular format. Today, most of the data is developed and processed in the simplest way to manage information.

* **Examples –** Relational data, Geo-location, credit card numbers, addresses, etc.

1. **Unstructured Data :** It is defined as the data in which is not follow a pre-defined standard or you can say that any does not follow any organized format. This kind of data is also not fit for the relational database because in the relational database you will see a pre-defined manner or you can say organized way of data. Unstructured data is also very important for the big data domain and To manage and store Unstructured data there are many platforms to handle it like No-SQL Database.

* **Examples** – Word, PDF, text, media logs, etc.

1. **Semi-Structured Data** : Semi-structured data is information that does not reside in a relational database but that have some organizational properties that make it easier to analyze. With some process, you can store them in a relational database but is very hard for some kind of semi-structured data, but semi-structured exist to ease space.

* **Example –** XML data.

**Q.3) What is big data analysis? Explain any two Real time example or Application of Bigdata In Detail ?**

* **Big data analysis:** Big data analysis refers to the process of examining and extracting valuable insights and knowledge from large and complex datasets using advanced tools and techniques such as machine learning, artificial intelligence, and statistical analysis. The objective of big data analysis is to uncover patterns, trends, and correlations that can help organizations make informed decisions, improve their operations, and gain a competitive advantage.
* **Here are two real-time Examples**/**Application of big data analysis:**
* **Retail and E-commerce:** Big data is extensively used in the retail and e-commerce sector for customer analytics, inventory management, demand forecasting, and personalized marketing. Retailers analyze large volumes of customer data to gain insights into buying patterns, preferences, and trends, which helps them optimize pricing, promotions, and product assortment.
* **Healthcare and Life Sciences:** Big data is transforming healthcare by enabling precision medicine, disease prediction, drug discovery, and patient care optimization. Healthcare providers utilize big data analytics to process and analyze vast amounts of patient data, including electronic health records, medical imaging, genomics, and wearable devices, to improve diagnoses, treatment outcomes, and operational efficiency.
* **Finance and Banking**: Big data plays a crucial role in the finance and banking industry for fraud detection, risk assessment, customer segmentation, and algorithmic trading. Financial institutions analyze vast volumes of transactional data, social media feeds, news articles, and market data to identify fraudulent activities, evaluate creditworthiness, personalize financial services, and make data-driven investment decisions.
* **Manufacturing and Supply Chain:** Big data analytics is utilized in manufacturing and supply chain management to optimize production processes, improve product quality, and enhance supply chain visibility. By collecting and analyzing data from sensors, production lines, and logistics networks, manufacturers can identify bottlenecks, predict maintenance needs, reduce downtime, and streamline the entire supply chain.
* **Transportation and Logistics:** Big data is used in transportation and logistics for route optimization, fleet management, demand forecasting, and real-time tracking. Transportation companies analyze data from GPS devices, sensors, weather forecasts, and historical traffic patterns to optimize delivery routes, minimize fuel consumption, and improve operational efficiency.
* **Telecommunications:** Big data analytics is employed in the telecommunications industry for network optimization, customer churn prediction, and targeted marketing. Telecommunication providers analyze large volumes of network data, customer interactions, call records, and social media data to identify network issues, improve service quality, and deliver personalized offers and recommendations.
* **Energy and Utilities:** Big data is applied in the energy and utilities sector for smart grid management, energy consumption optimization, predictive maintenance, and renewable energy integration. Utility companies leverage data from smart meters, sensors, weather forecasts, and historical consumption patterns to monitor energy usage, detect anomalies, optimize distribution, and promote energy efficiency.
* **Government and Public Services:** Big data has numerous applications in government and public services, including urban planning, disaster management, crime prediction, and public health monitoring. Governments analyze diverse data sources, such as social media, satellite imagery, census data, and public records, to make informed policy decisions, enhance public safety, and improve service delivery.

**Q.4) Define Big Data and Explain the characteristics of big data?**

* **Big data** refers to extremely large and complex data sets that cannot be processed using traditional data processing tools and techniques. These data sets typically include structured, semi-structured, and unstructured data, such as text, images, videos, and sensor data. The term "big data" is often used to describe data sets that are beyond the capabilities of traditional databases, hardware, and software tools to process, store, and analyze.
* **Here are the key characteristics of big data:**

1. **Volume:** Big data sets are characterized by their size, with data sets ranging from terabytes to petabytes or more. The sheer volume of data makes it difficult to process and analyze using traditional tools.
2. **Velocity:** Big data is generated and updated at an unprecedented rate, often in real-time or near-real-time. The speed at which data is generated and processed requires advanced tools and techniques to manage and analyze it.
3. **Variety:** Big data comes in many different forms, including structured, semi-structured, and unstructured data. This variety of data types requires specialized tools and techniques to analyze and make sense of the data.
4. **Veracity:** Big data is often of uncertain quality and accuracy, making it difficult to draw reliable insights from the data. It is important to ensure that the data is accurate and reliable before analyzing it**.**
5. **Value:** The ultimate goal of big data analysis is to extract valuable insights and knowledge that can help organizations make better decisions and improve their operations.

Overall, the characteristics of big data make it challenging to manage, analyze, and extract valuable insights from. Advanced tools and techniques such as machine learning, artificial intelligence, and data visualization are necessary to process, manage and extract insights from big data**.**

**Q.5) Mention Various Distribution packages of Big Data?**

* **There are several distribution packages available for big data processing and analytics. Here are some popular ones:**
* **Apache Hadoop:** Apache Hadoop is an open-source framework that allows distributed processing of large datasets across clusters of computers using a simple programming model. It includes the Hadoop Distributed File System (HDFS) for data storage and the MapReduce processing framework.
* **Cloudera Distribution for Hadoop (CDH):** CDH is a distribution package provided by Cloudera, which includes various Apache Hadoop ecosystem components such as HDFS, MapReduce, Hive, Pig, Impala, Spark, and more. It aims to simplify the deployment and management of big data solutions.
* **Hortonworks Data Platform (HDP):** HDP is an open-source distribution package that integrates Apache Hadoop and other related projects. It provides a comprehensive platform for storing, processing, and analyzing large datasets. HDP includes components like HDFS, MapReduce, Hive, HBase, Spark, and many others.
* **MapR:** MapR is a converged data platform that combines Hadoop, Spark, and other big data technologies. It offers enterprise-grade features and performance enhancements for data storage, processing, and analytics. MapR also includes its proprietary distributed file system called MapR-FS.
* **Apache Spark:** While not strictly a distribution package, Apache Spark is a widely used big data processing engine that can work with different data storage systems. Spark provides in-memory computing capabilities and supports various programming languages, making it suitable for real-time streaming, batch processing, machine learning, and graph processing tasks.
* **Amazon EMR:** Amazon Elastic MapReduce (EMR) is a cloud-based big data platform provided by Amazon Web Services (AWS). It simplifies the deployment and management of Apache Hadoop, Spark, and other big data frameworks on AWS infrastructure. EMR offers scalability, reliability, and integration with other AWS services.
* **Google Cloud Dataproc:** Google Cloud Dataproc is a managed big data service on the Google Cloud Platform (GCP). It allows easy deployment of Apache Spark, Hadoop, and other related tools. Dataproc offers auto-scaling, high availability, and integration with other GCP services like BigQuery and Dataflow.
* **Microsoft Azure HDInsight:** Azure HDInsight is a cloud-based big data platform provided by Microsoft Azure. It supports various open-source big data technologies, including Hadoop, Spark, Hive, HBase, and more. HDInsight integrates with other Azure services like Azure Data Lake Storage and Azure Machine Learning.

**Q.6) Why Big Data is more important?**

* **Big data** has become increasingly important in today's digital landscape due to the vast amounts of data that are generated on a daily basis. The emergence of social media, IoT devices, and mobile technologies has led to an explosion of data, and the ability to collect, store, and analyze this data has become critical to the success of many organizations. Here are some reasons why big data is more important:

1. **Better Decision Making:** Big data analysis allows organizations to make better decisions based on data-driven insights. By analyzing large data sets, organizations can gain a better understanding of customer behavior, market trends, and other key factors that can influence business outcomes.
2. **Improved Efficiency:** Big data analysis can help organizations optimize their operations by identifying inefficiencies, reducing costs, and improving overall performance. For example, analyzing sensor data can help manufacturers predict maintenance needs, avoiding costly downtime and reducing maintenance costs.
3. **Competitive Advantage:** Organizations that effectively utilize big data can gain a competitive advantage over their competitors. By using data to make informed decisions, organizations can develop better products, offer more personalized services, and improve customer experiences.
4. **Innovation:** Big data analysis can also drive innovation by identifying new trends, opportunities, and areas for growth. By analyzing data, organizations can uncover new patterns and insights that can lead to new products, services, and business models.

Overall, big data has become an integral part of many industries, including healthcare, finance, retail, and manufacturing, among others. It is increasingly important for organizations to effectively manage, analyze, and extract valuable insights from big data in order to stay competitive and achieve success in today's digital landscape.

**Q.7) What is Hadoop explain history of Hadoop?**

* **Hadoop:** Hadoop is a distributed computing framework that allows for the processing of large data sets across clusters of computers.
* **History of Hadoop:** It was created by Doug Cutting and Mike Cafarella in 2005, inspired by Google's MapReduce and Google File System (GFS) papers. The name "Hadoop" comes from a toy elephant that belonged to Doug Cutting's son. The toy elephant became the project's mascot, and the name stuck. Version 0.1.0, was released in April 2006. The framework was initially developed as an open-source implementation of Google's MapReduce and GFS technologies. In 2008, Hadoop was used to power the Obama campaign's online operations, and in 2009 it became a top-level Apache project. In 2011, Hadoop 0.20.205 was released, which introduced support for security and improved HBase integration. In 2012, Hadoop 1.0 was released, which included significant performance improvements and support for Apache HBase and Apache Pig. In 2014, Hadoop 2.0 was released, which introduced support for YARN (Yet Another Resource Negotiator), a more flexible and efficient resource management system. YARN allowed Hadoop to support a broader range of workloads, including real-time data processing and interactive queries.

**Q.8) What is Big Data & Data ware house & Cloud Computing?**

* Big data, data warehouse, and cloud computing are three related concepts that are often used in the context of managing and analyzing large amounts of data. Here's a brief overview of each:
* **Big Data:** Big data refers to extremely large and complex data sets that cannot be processed using traditional data processing tools and techniques. These data sets typically include structured, semi-structured, and unstructured data, such as text, images, videos, and sensor data. The term "big data" is often used to describe data sets that are beyond the capabilities of traditional databases, hardware, and software tools to process, store, and analyze.
* **Data Warehouse:** A data warehouse is a large, centralized repository of data that is used for reporting and analysis. It is designed to support business intelligence activities, such as data mining, trend analysis, and decision making. Data warehouses are typically built using a relational database and are optimized for read-heavy workloads.
* **Cloud Computing:** Cloud computing is a model for delivering computing resources, such as servers, storage, and software applications, over the internet. Cloud computing allows organizations to access computing resources on-demand, without the need for on-premises infrastructure. Cloud computing providers, such as Amazon Web Services, Microsoft Azure, and Google Cloud Platform, offer a range of services, including compute, storage, networking, and database services.

Overall, big data, data warehouse, and cloud computing are related concepts that are all important in the context of managing and analyzing large amounts of data. Big data requires specialized tools and techniques to manage and analyze, while data warehouses are designed to support reporting and analysis. Cloud computing provides a flexible and scalable platform for managing and analyzing data, and can be used to support both big data and data warehouse workloads.

**Q.9) Difference between Structure & Unstructured Data?**

|  |  |  |
| --- | --- | --- |
|  | **Structured data** | **Unstructured data** |
| **Technology** | It is based on a relational database. | It is based on character and binary data. |
| **Flexibility** | Structured data is less flexible and schema-dependent. | There is an absence of schema, so it is more flexible. |
| **Scalability** | It is hard to scale database schema. | It is more scalable. |
| **Robustness** | It is very robust. | It is less robust. |
| **Performance** | Here, we can perform a structured query that allows complex joining, so the performance is higher. | While in unstructured data, textual queries are possible, the performance is lower than semi-structured and structured data. |
| **Nature** | Structured data is quantitative, i.e., it consists of hard numbers or things that can be counted. | It is qualitative, as it cannot be processed and analyzed using conventional tools. |
| **Format** | It has a predefined format. | It has a variety of formats, i.e., it comes in a variety of shapes and sizes. |
| **Analysis** | It is easy to search. | Searching for unstructured data is more difficult. |

**Q.10) Explain Advantage and Disadvantage of Hadoop?**

* **Disadvantages:**
* Hadoop does not provide easy tools for removing noise from the data; hence, maintaining that data is a challenge.
* It has many data security issues like encryption problems.
* Streaming jobs and batch jobs are not performed efficiently.
* MapReduce programming is inefficient for jobs involving highly analytical skills.
* It is a distributed system with low level APIs. Some APIs are not useful to developers.
* **Advantages:**
* Hadoop has many useful functions like data warehousing, fraud detection and marketing campaign analysis.
* These are helpful to get useful information from the collected data.
* Hadoop has the ability to duplicate data automatically. So multiple copies of data are used as a backup to prevent loss of data.

**Q.11) Explain Components of Hadoop with Neat & Clean diagram?**

* **Components of Hadoop:**

1. **HDFS- Hadoop Distributed File System :** It is a storage layer which allows to store data across machines and on their disks. It is open source framework ,works by rapidly transferring data between nodes.

* HDFS (Hadoop Distributed File System) is the primary storage system used by Hadoop applications. HDFS is a key component of many Hadoop systems, as it provides a means for managing big data, as well as supporting big data analytics.
* It states that the files will be broken into blocks and stored in nodes over the distributed architecture. Google published its paper GFS and on the basis of that HDFS was develop
* It's often used by companies who need to handle and store big data.

1. **Distributed File System (DFS):**

* A distributed file system (DFS) is a file system that enables clients to access file storage from multiple hosts through a computer network as if the user was accessing local storage.
* Files are spread across multiple storage servers and in multiple locations, which enables users to share data and storage resources.

1. **YARN- Yet Another Resource Negotiator:**

* It is a processing layer. It manages resources required for the data processing. It is a resource management across the cluster and then various processing frameworks such as
  + Strucutred data it can use HIVE,Impala.(query processing tools)
  + Streaming data it uses Kafka and Spark.(storing, reading and analysing streaming data) And Batch Based processing it uses Map reduce.
  + It is used for job scheduling and manage the cluster.
  + Job scheduling :It is the process of allocating system resources to many different tasks by an operating system (OS).
  + Cluster management, such as monitoring status of the cluster nodes, scheduling the compute jobs and providing a control interface between users and the cluster, are essential.

|  |
| --- |
|  |

**Q.12) List Top 10 Companies and the Cloud Services Along with Data Warehouse Storages?**

* **Top 10 Companies and the Cloud Services Along with Data Warehouse Storages:**

1. **Amazon Web Services (AWS):[** **Amazon Redshift]** Amazon Web Services (AWS) is one of the leading cloud service providers in the market. AWS offers a range of cloud-based services, including storage, computing, databases, analytics, machine learning, and more. One of the most popular services offered by AWS is Amazon Redshift, which is a fully managed, petabyte-scale data warehouse service. Amazon Redshift allows users to analyze data using SQL and business intelligence tools. It is highly scalable, and users can add or remove nodes based on their requirements.
2. **Microsoft Azure: [Azure Synapse Analytics (formerly known as SQL Data Warehouse)]**Microsoft Azure is another popular cloud service provider that offers a wide range of services, including computing, storage, databases, analytics, and more. Azure Synapse Analytics, formerly known as SQL Data Warehouse, is Microsoft Azure's data warehouse storage service. It is a cloud-based service that combines big data and data warehousing capabilities in a single offering. Azure Synapse Analytics allows users to store and analyze data using SQL and other programming languages.
3. **Google Cloud Platform (GCP):[ BigQuery]** Google Cloud Platform (GCP) is a cloud service provider that offers a range of services, including computing, storage, databases, analytics, and more. BigQuery is GCP's data warehouse storage service. It is a serverless, highly scalable, and cost-effective data warehouse that can handle petabyte-scale data. BigQuery allows users to store and analyze data using SQL and other programming languages.
4. **Snowflake:[Snowflake Data Warehouse]** Snowflake is a cloud-based data warehousing solution that offers a range of services, including data warehousing, data engineering, data science, and more. Snowflake allows users to store and analyze data using SQL. It is designed to handle large and diverse data sets and is highly scalable. Snowflake offers a unique architecture that separates storage and compute, which allows users to scale storage and compute independently.
5. **IBM Cloud:[** **IBM Db2 Warehouse on Cloud]** IBM Cloud is a cloud service provider that offers a range of services, including computing, storage, databases, analytics, and more. Db2 Warehouse on Cloud is IBM Cloud's data warehouse storage service. It is a fully-managed, cloud-based data warehouse that can handle data from a variety of sources. Db2 Warehouse on Cloud allows users to analyze data using SQL and other programming languages.
6. **Oracle Cloud:[** **Oracle Autonomous Data Warehouse]** Oracle Cloud is a cloud service provider that offers a range of services, including computing, storage, databases, analytics, and more. Oracle Autonomous Data Warehouse is Oracle Cloud's data warehouse storage service. It is a fully-managed, cloud-based data warehouse that uses machine learning to automate many of the tasks associated with data warehousing. Oracle Autonomous Data Warehouse allows users to analyze data using SQL and other programming languages.
7. **Teradata:[** **Vantage]** Teradata is a cloud-based data warehousing solution that offers a range of services, including data warehousing, data engineering, data science, and more. Vantage is Teradata's data warehouse storage service. It is a cloud-based data warehouse that allows users to store and analyze data using SQL, R, Python, and other programming languages. Vantage is designed to handle large and complex data sets.
8. **SAP:[** **SAP Data Warehouse Cloud]** SAP is a software company that offers a range of business solutions, including enterprise resource planning (ERP), customer relationship management (CRM), and more. SAP HANA is SAP's data warehouse storage service. It is an in-memory, column-oriented, relational database management system that can handle large amounts of data and is optimized for analytics. SAP HANA allows users to analyze data using SQL and other programming languages.
9. **Cloudera:[** **Cloudera Data Warehouse]** Cloudera offers Cloudera Data Warehouse as its data warehouse storage service. It is a cloud-based data warehouse that allows users to store and analyze data using SQL, Python, and other programming languages.
10. **Databricks:** Databricks offers Delta Lake as its data warehouse storage service. It is a cloud-based data lake that can handle large and diverse data sets and allows users to analyze data using SQL, Python, and other programming languages.

**Q.13) Which are the Cloud Storages available Free of Cost? Explain any three?**

* **The Cloud Storages available Free of Cost:**

1. **Google Drive:** Google Drive is a popular cloud storage service that offers 15GB of free storage that can be used for Google Photos, Google Drive, and Gmail. Users can also purchase additional storage if they need it. Google Drive is known for its easy-to-use interface and integration with other Google products.
2. **Dropbox:** Dropbox is a cloud storage service that provides 2GB of free storage, with the ability to earn more through referrals and other promotions. Dropbox is known for its file synchronization features, which make it easy to access files across different devices.
3. **OneDrive:** OneDrive is Microsoft's cloud storage service that offers 5GB of free storage, with the option to earn more through referrals and other activities. OneDrive is integrated with Microsoft Office, making it easy to create and edit documents stored in the cloud.
4. **iCloud:** iCloud is Apple's cloud storage service that offers 5GB of free storage, which can be used for photos, videos, and backups. iCloud is tightly integrated with Apple's ecosystem of products, making it easy to access files across different devices.
5. **Mega:** Mega is a cloud storage service that offers 50GB of free encrypted storage, with the ability to earn more through referrals and other activities. Mega is known for its strong encryption and privacy features.
6. **Amazon Drive:** Amazon Drive provides 5GB of free storage for Amazon Prime members, with the option to purchase additional storage. Amazon Drive is integrated with Amazon's other products, making it easy to store and access files related to Amazon purchases.
7. **Box:** Box is a cloud storage service that offers 10GB of free storage, with the ability to earn more through referrals and other promotions. Box is known for its collaboration features, which make it easy to work on files with other users.
8. **pCloud:** pCloud provides 10GB of free storage, with the option to earn more through referrals and other activities. pCloud is known for its strong encryption and security features.

**Q.14) Case Study on any one Topic?**

* **Case Study: Digital Transformation at Nike**
* Nike is a global sportswear company that has undergone a significant digital transformation in recent years. The company recognized the need to adapt to the changing retail landscape, as more consumers were turning to online channels to purchase products. In response, Nike developed a strategy to shift its focus towards digital channels and enhance the overall customer experience.
* The first step in Nike's digital transformation was the creation of the NikePlus membership program, which provides members with exclusive benefits such as early access to product launches, personalized training programs, and other rewards. This program has helped to strengthen Nike's relationship with its customers, while also providing valuable data on consumer behavior.
* Nike also invested heavily in its digital platform, with a focus on creating a seamless and personalized shopping experience for customers. The company revamped its website and mobile app, incorporating features such as product recommendations based on customer preferences, personalized content, and a streamlined checkout process.
* In addition to enhancing the online shopping experience, Nike also integrated technology into its physical retail stores. The company introduced new technologies such as the Nike Fit app, which uses augmented reality to scan a customer's feet and provide recommendations for the best shoe size and fit. Nike also implemented features such as self-checkout and in-store pickup, making it easier and more convenient for customers to shop in-store.
* The results of Nike's digital transformation have been impressive. The company has seen significant growth in its digital sales, with digital revenue increasing by 82% in Q4 of 2020. The NikePlus membership program has also been a success, with membership growing by 60% over the past year.
* Overall, Nike's digital transformation has helped the company to stay competitive in the changing retail landscape, while also enhancing the overall customer experience. By leveraging technology and data, Nike has been able to create a more personalized and seamless shopping experience for its customers, which has translated into increased sales and customer loyalty.

**(PPT Give you more understanding than PDF)** The material for the PDF has been compiled from various sources such as books, tutorials (offline and online), lecture notes, several resources available on Internet. The information contained in this PDF is for general information and education purpose only. While we endeavor to keep the information up to date and correct, **we make no representation of any kind about the completeness and accuracy of the material.** The information shared through this PDF material should be used for educational purpose only.

**Thank you…**