Task 1

1. Explain the meaning of virtualization and State the name of the above hypervisor.

   - **Virtualization** refers to the act of creating a virtual (rather than actual) version of something, including virtual computer hardware platforms, operating systems, storage devices, and computer network resources. **Virtualization** is the ability to run multiple operating systems on a single physical system and share the underlying hardware resources.

   - Type 2 Hypervisor **(**Hosted Hypervisor)

2. Critically evaluate the above hypervisor.

   Hypervisors can vary in terms of performance, scalability, security, management features, and compatibility with different guest operating systems. It would require a closer examination of the specific hypervisor's features, capabilities, and user experiences to evaluate its strengths and weaknesses.

3. Differentiate between type 1(bare-metal) and type 2 (hosted) hypervisors.

Criteria	Type1 hypervisor (Bare-metal or Native)	Type2 hypervisor (Hosted)
Definition	Runs directly on the system with VMs running on them	Runs on a conventional operating system
Virtualization	Hardware virtualization	OS Virtualization
Scalability	Better scalability	Not so much, because of its reliance on the underlying OS
System Independence	Has direct access to hardware along with virtual machines it hosts	Are not allowed to directly access the host hardware and its resources
Speed	Faster	Slower because of the system's dependency
Security	More secure	Less secure, as any problem in the basic operating system affects the entire system including the protected hypervisor
Examples	- VMware ESXi - Microsoft Hyper-V - Citrix XenServer (Xen) - KVM	- VMware workstation player - Microsoft virtual PC - Sun's virtual Box

4. Compare between virtual machines and containers.

Virtual Machine	Container
Heavyweight	Lightweight
Limited Performance	Native performance
Each VM runs in its own OS	All containers share the host os
Hardware-Level virtualization	OS virtualization
Startup time in minutes	Startup time in milliseconds
Allocates required memory	Requires less memory space
Fully isolated and hence more secure	Process-level isolation, possibly less secure

5. State cloud computing service model and deployment models and explain them.
   - Cloud Service Models:
     - **IaaS** — Infrastructure as a Service Cloud Service Provider provides infrastructure and resources Manufacturing organization manages OS, data and software applications
     - **PaaS** — Platform as a Service Cloud Service Provider provides infrastructure and development platform Manufacturing organization can develop its own software Applications
     - **SaaS** — Software as a Service Cloud Service Provider has a full control over cloud and software Manufacturing organization rents software applications

- **Cloud Computing Deployment Model**
  - Private: Used for a single organization
  - Community: Shared by several organization; typically externally hosted, but may be can be internally hosted by one of the organizations.
  - Public
  - Hybrid: composition of two or more clouds (private, community or public)

Task 2

6. Identify the data analytics lifecycle

- **Step 1: Business Issue Understanding**
  - Define business objectives
  - Gather required information
  - Determine appropriate analysis method
  - Clarify scope of work
  - Identify deliverables
- **Step 2: Data understanding**
  - Collect initial data
  - Identify data requirements
  - Determine data availability
  - Explore data and characteristics
- **Step 3: Data preparation**
  - Gather data from multiple sources

- Cleanse
- Format
- Blend
- Sample

❖ **Step 4: Exploratory Analysis and Modeling**

- Develop methodology
- Determine important variables
- Build model
- Assess model

❖ **Step 5: Validation**

- Evaluate results
- Review process
- Determine next steps
- Results are valid proceed to step 6
- Results are invalid revisit Steps 1 •4

❖ **Step 6: Visualization and Presentation**

- Communicate results
- Determine best method to present insights based on analysis and audience
- Craft a compelling story
- Make recommendations

7. Explain types of big data and big data Job roles

- **Business user:** Someone who benefits from the end results and can advise the project team on the value of end results and how the project results will be operationalized.

- **Project sponsor:** The project sponsor generally provides the funding and gauges the degree of value from the final outputs of the working team.

- **Project manager:** Ensures that key milestones and objectives are met on time and at the expected quality.

- **Business intelligence analyst:** Provides business-domain expertise with deep understanding of the data, KPIs, key metrics, and analytics from a reporting perspective.

- **Data engineer:** Applies deep technical skills to assist with data extraction from source systems and data ingestion on the analytic sandbox.

- **Database administrator (DBA):** Provisions and configures the database environment to support the analytical needs of the working team.

- **Data scientist:** Provides technical expertise for analytical techniques and data modeling, and applies the proper analytical techniques to given business problems to achieve the overall analytical objectives.

8. Compare between Data warehouse, Data lake, and Data mart

	Data warehouses	Data lakes	Data marts
Usage	The data analysis and reporting needs of an entire organization	The reporting needs of different kinds and difficulty, predictive analytics	The reporting needs Of a specific operational department or subject
Data stored (typically)	Larger volumes of structured data; processed	Huge volumes of structured and unstructured data; raw	A limited amount of structured data; processed
Data sources	An array Of external and internal sources, covering different areas of business	Any external or internal sources	Few sources linked to one business area
Size	Larger than 100	Larger than 100 GB	Smaller than 100 GB
Ease of creation	Difficult to set up	Difficult to set up	Easy to set up

9. Critically evaluate the Hadoop components and the meaning of map reduce and its steps with example.

Hadoop is an open-source framework that provides distributed storage and processing of large datasets across clusters of computers. It consists of several components, including:

- Hadoop Distributed File System (HDFS): A distributed file system that provides reliable and scalable storage of large datasets.
- Yet Another Resource Negotiator (YARN): A resource management and job scheduling system that allows multiple data processing engines to share a cluster.
- MapReduce: A programming model and software framework used to process and analyze large datasets in parallel.
- Hadoop Common: A set of utilities and libraries used by other Hadoop components.

MapReduce is a programming model used to process large datasets in parallel by dividing the work into smaller tasks and distributing them across a cluster of computers. It consists of two main steps:

- Map: The input data is divided into smaller chunks, and a map function is applied to each chunk in parallel. The map function takes the input data, processes it, and produces a set of key-value pairs as output.
- Reduce: The key-value pairs produced by the map function are grouped by key and passed to a reduce function. The reduce

function takes the key-value pairs, processes them, and produces a set of output data.

For example, let's say we have a large dataset of customer orders and we want to calculate the total revenue for each product. We can use MapReduce to process the dataset in parallel using the following steps:

- Map: The input dataset isdivided into smaller chunks, and the map function is applied to each chunk in parallel. The map function reads each order, extracts the product name and revenue, and outputs a set of key-value pairs where the key is the product name and the value is the revenue.
- Shuffle and Sort: The key-value pairs produced by the map function are shuffled and sorted by key. This ensures that all key-value pairs with the same key are grouped together and passed to the same reduce function.
- Reduce: The key-value pairs are grouped by key and passed to the reduce function. The reduce function takes the key-value pairs for each product, sums up the revenue, and outputs a set of key-value pairs where the key is the product name and the value is the total revenue.

10. Summarize the meaning of clustering, its types and k-mean

❖ Meaning

- Clustering is the process of dividing the datasets into groups, consisting of similar data-points
- It is unsupervised machine learning technique
- Points in the same group are as similar as possible.
- Points in different group are as dissimilar as possible.

❖ Types of clustering

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering

❖ K-Mean clustering

- It is a type of unsupervised learning used when you have unlabeled data
- It is a type of unsupervised learning used when you have unlabeled data

11. Summarize the problems that exist in big data and the advanced big data analytics techniques used to solve it

Problem to solve	Techniques to solve
I want to group items by similarity	Clustering
I want to discover relationships between actions or items	Associations rules
I want to determine the relationship between the outcome and input variables	Regression
I want to analyze my text data	Text analyze
I want to assign known labels to objects	Classification
I want to forecast the behavior of a temporal process	Time series analysis