**Task 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.

# **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

* laaS — 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

**Task 2**

**6- Identify the data analytics lifecycle**

* Problem Definition: In this stage, the business problem, research question, or hypothesis is defined, and the objectives of the analysis are established.
* Data Collection: In this stage, data is gathered from various sources, such as databases, surveys, and online sources. The data is then organized and stored in a suitable format for analysis.
* Data Preparation: In this stage, the collected data is cleaned, pre-processed, transformed, and formatted to make it ready for analysis.
* Data Analysis: In this stage, various techniques, such as exploratory data analysis, statistical analysis, machine learning, and data visualization, are applied to the data to extract insights, patterns, and trends.
* Interpretation: In this stage, the results of the analysis are interpreted and evaluated in the context of the business problem or research question.
* Communication: In this stage, the findings are presented to stakeholders through reports, dashboards, or visualizations, and recommendations are made based on the insights gained from the analysis.
* Implementation: In this final stage, action is taken based on the insights gained from the analysis to solve the business problem or address the research question. The implementation can involve changes to business processes, strategies, or products based on the recommendations made.

**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:

1. Hadoop Distributed File System (HDFS): A distributed file system that provides reliable and scalable storage of large datasets.
2. Yet Another Resource Negotiator (YARN): A resource management and job scheduling system that allows multiple data processing engines to share a cluster.
3. MapReduce: A programming model and software framework used to process and analyze large datasets in parallel.
4. 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:

1. 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.
2. 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:

1. 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.
2. 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.
3. 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.

The output of the reduce function gives us the total revenue for each product, which can be further processed or analyzed.

MapReduce has several advantages, including its ability to process large datasets in parallel, fault tolerance, and scalability. However, it also has some limitations, such as its batch processing nature and the need to write custom map and reduce functions for each data processing task.

**10- Summarize the meaning of clustering, its types and k-mean** clustering algorithm as an example of clustering algorithm

Clustering is a machine learning technique that involves grouping data points with similar characteristics or features together. The goal of clustering is to identify patterns or structures in data that may not be obvious by simply looking at it.

There are several types of clustering algorithms, including hierarchical clustering, density-based clustering, and partitioning clustering. Partitioning clustering is further divided into various algorithms, such as k-means clustering, k-medoids clustering, and fuzzy c-means clustering.

K-means clustering is a popular algorithm for partitioning clustering that involves partitioning the data into k clusters, where k is a predetermined number. The algorithm starts by randomly selecting k centroids, which are the centers of the clusters. Then, each data point is assigned to the cluster whose centroid is the closest to it. After all data points are assigned to clusters, the centroid of each cluster is updated to the mean of the data points in that cluster. This process of assigning data points to clusters and updating centroids is repeated until the centroids no longer change or a maximum number of iterations is reached.

K-means clustering has several advantages, including its efficiency, simplicity, and ability to handle large datasets. However, it also has some limitations, such as its sensitivity to the initial placement of the centroids and the assumption that all clusters have the same shape and size.

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

Big data refers to datasets that are too large and complex to be processed by traditional data processing tools and techniques. Handling big data poses several challenges, including:

* Volume: Big data is characterized by its sheer size, which makes it difficult to store, process, and analyze using traditional methods.
* Velocity: Big data is generated at a high rate and often in real-time, requiring efficient processing and analysis to extract insights in a timely manner.
* Variety: Big data comes in different formats, including structured, semi-structured, and unstructured data, making it difficult to integrate and analyze.

To address these challenges, advanced big data analytics techniques have been developed, including:

* Distributed Computing: This involves distributing the processing and storage of data across multiple computers, enabling parallel processing and faster analysis.
* NoSQL Databases: NoSQL databases are designed to handle large volumes of unstructured and semi-structured data, making them ideal for big data applications.
* Machine Learning: Machine learning algorithms can be used to automatically identify patterns and relationships in big data, enabling faster and more accurate analysis.
* Data Visualization: Data visualization tools can be used to create interactive and visually appealing dashboards that help users better understand and explore big data.
* Stream Processing: Stream processing technologies enable real-time analysis of data streams, allowing businesses to quickly respond to emerging trends and opportunities.
* Natural Language Processing: Natural language processing techniques can be used to analyze unstructured data, such astext and speech, and extract useful information from it, such as sentiment analysis and topic modeling.