1) **Explain in detail the various activities carried out in the data preparation phase of the analytical lifecycle.**

The second phase of the Data Analytics Lifecycle involves data preparation, which includes the steps to explore, preprocess, and condition data prior to modeling and analysis. In this phase, the team needs to create a robust environment in which it can explore the data that is separate from a production environment.

Usually, this is done by preparing an analytics sandbox. To get the data into the sandbox, the team needs to perform ETLT, by a combination of extracting, transforming, and loading data into the sandbox. Once the data is in the sandbox, the team needs to learn about the data and become familiar with it. Understanding the data in detail is critical to the success of the project.

The team also must decide how to condition and transform data to get it into a format to facilitate subsequent analysis. The team may perform data visualizations to help team members understand the data, including its trends, outliers, and relationships among data variables.

Each of these steps of the data preparation phase is discussed throughout this section. Data preparation tends to be the most labor-intensive step in the analytics life cycle.

ln fact, it is common for teams to spend at least 50% of a data science project's time in this critical phase. If the team cannot obtain enough data of sufficient quality, it may be unable to perform the subsequent steps in the life cycle process.

2) **Discuss the different analytical architectures to be considered for big data with different application requirements.**

From a big data analytics perspective, Hadoop is used for a number of tasks, including correlation and cluster analysis to find patterns in the unstructured data sets.

➔ **Map Reduce**

MapReduce, as discussed above, is a programming framework developed by Google that supports the underlying Hadoop platform to process the big data sets residing on distributed servers (nodes) in order to produce the aggregated results.

➔ **Pig and PigLatin**

The Pig programming language is configured to assimilate all types of data (structured/unstructured, etc.). Two key modules are comprised in it: the language itself, called PigLatin, and the runtime version in which the PigLatin code is executed

➔ **Hive**

While Pig is robust and relatively easy to use, it still has a learning curve. This means the programmer needs to become proficient.

Facebook has developed a runtime Hadoop support architecture that leverages SQL with the Hadoop platform. This architecture is called Hive; it permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements.

➔ **Jaql**

Jaql's primary role is that of a query language for JavaScript Object Notational (JSON). However, its capability goes beyond LSON. It facilitates the analysis of both structured and nontraditional data.

➔ **Zookeeper**

Zookeeper is yet another open-source Apache project that allows a centralized infrastructure with various services; this provides for synchronization across a cluster of servers.

Zookeeper maintains common objects required in large cluster situations (like a library). Examples of these typical objects include configuration information, hierarchical naming space, and others

➔ **Cassandra**

Cassandra, an Apache project, is also a distributed database system. It is designated as a top-level project modeled to handle big data distributed across many utility servers.

3) Explain the challenges of Big Data.
Refer A Q2

4) Explain the various activities carried out for Big data acquisition.
Refer A Q4

5) Elucidate the relationship between (i)Big Data and cloud computing, (ii)Big data and IoT
Refer A Q6

This disruptive technology needs new infrastructures, including software and hardware applications as well as an OS; enterprises must handle the influx of data that begins flowing in and examine it in real-time as it evolves by the minute.

That is where big data arrives into the picture; big data analytics tools have the capacity to handle large volumes of data generated from IoT devices that create a continuous stream of information.
But, in order to differentiate between them, IoT provides data from which big data analytics can extract information to generate insights required of it.

However, IoT conducts data on a completely different scale, so the analytics solution must accommodate its needs of processing and rapid ingestion followed by a fast and accurate extraction.

There are many solutions available that provide near real-time analytics on large-sized datasets, and necessarily change a full-rack database into a small server that processes up to 100 TB, so small amount of hardware is needed. The analytics database of next-generation leverages GPU technology, thus enabling even more downsizing of the hardware, i.e, 5 TB on a laptop or a big database in the car. This largely helps IoT organizations correlate the evolving number of data sets, which helps them adapt to changing trends and acquire real-time responses, solving the challenge regarding size and compromising on the performance.

6) Define Big data. Explain the types of data repositories from an analyst perspective.
Big data is an evolving term that describes a large volume of structured, semi-structured and unstructured data that has the potential to be mined for information and used in machine learning projects and other advanced analytics applications.

Big data is often characterized by the 3Vs: the extreme volume of data, the wide variety of data types and the velocity at which the data must be processed.

Data Warehouse Enterprise data warehouses (EDW) are critical for reporting and Business Intelligence (BI) tasks, although from an analyst perspective they tend to restrict the flexibility that a data analyst has for performing robust analysis or data exploration. In this model, data is managed and controlled by IT groups and DBAs, and analysts must depend on IT for access and changes to the data schemas. This tighter control and oversight also means longer lead times for analysts to get data, which generally must come from multiple sources. Another implication is that EDW rules restrict analysts from building data sets, which can cause shadow systems to emerge within organizations containing critical data for constructing analytic data sets, managed locally by power users.Analytic sandboxes enable high performance computing using in-database processing. This approach creates relationships to multiple data sources within an organization and saves the analyst time of creating these data feeds on an individual basis. In-database processing for deep analytics enables faster turnaround time for developing and executing new analytic models, while reducing (though not eliminating) the cost associated with data stored in local, “shadow” file systems. In addition, rather than the typical structured data in the EDW, analytic sandboxes can house a greater variety of data, such as webscale data, raw data, and unstructured data.

7) **Explain the key elements of Hadoop. What are the advantages provided by Hadoop for the management and analysis of big data?**

8)