**Big Data** refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered to derive real-time business insights

that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.

**The 5 important V's of big data -**

**1. Velocity -** the speed at which data accumulates.

**2. Volume -** a scale of data. drivers are an increase in data sources, Higher resolution sensors, Scalable Infrastructure.

**3. Variety -** diversity of data, structured - unstructured, different source, different format etc. drivers - wearables, mobiles, geological, video etc.

**4. Veracity -** quality in the origin of data, and its conformity to facts and accuracy Attributes - consistency, completeness, integrity, ambiguity.

**5. value -** Value is our ability and need to turn data into value. Value isn't just profit. It may have medical or social benefits, as well as customer, employee, or personal satisfaction.

**what is Hadoop?**

Larry Page and Sergy Brin from google were trying to solve their page rank algorithm which meant dealing with all the webpages in the world,

so they came up with a rather simple approach, they divided data into clusters and send these small clusters to machines,

each machine would apply the same algorithm to these clusters which are called mapping and then would return the result and the result would be

combined which is called reduce, yahoo hired Doug cutting who just replicated googles big data program and named it Hadoop.

What is Data Mining and the Steps involved in data mining?

**1. Establishing Data Mining Goals :**

The expected level of accuracy and usefulness of data is to be established as achieving more accuracy is directly proportional to the cost spent on achieving it. and after a certain level, one doesn't gain much more even after processing the data.

**2. Selection of data :**

The output of data mining is largely based upon the quality of data being used, there are times when the data is not readily available at that time one can use surveys to get that data, and selecting the right data is very important as the cost increases with more and more data you use.

**3. Pre Processing data :**

Raw data often contains irrelevant and erroneous information, so that has to be removed, and in case of randomly missing data simple solutions can be deployed but if data is missing systematically then it would lead to systematic biases in analytics so one must consider in advance if such data will be included or not in analysis.

**4. Transforming data :**

After the relevant attributes of data have been retained, this step involves reducing the number of attributes which requires to transform data, several algorithms are deployed like principal component analysis,

which can reduce the number of attributed without significant loss in information in addition to this various variables need to be transformed for example continuous variable can be transformed into a categorical variable.

**5. Storing of data :**

Efficient reading/writing from the database is required, hence data must be stored in such a manner where it is protected and the reading/writing is efficient.

**6. Mining data:**

after the above steps, the data is subjected to data mining in which data analysis methods both parametric and non-parametric, machine learning algorithm. data visualization should be performed before processing further to understand the trends in data.

**7. Evaluating mining results :**

After results have been extracted from data mining, a formal evaluation is done which tests the predictive capabilities of the model and sees how efficient and effective the model is, this is known as In sample forecast.

Note - Data mining is an iterative process where better algorithms are deployed in each iteration and stakeholders feedback is also taken into account