DATA Engineering Assignment

**Big data:** Is a process of examining large data set to uncover hidden patterns, unknown correlations and collection of huge information which is hard to store, process or transfer using traditional method. To solve above problems bigdata tools where boomed.

**Some of the examples of bigdata:** Stock market data, social media data, organization data.

**Source of big data:** Human generated data, machine (system generated data), Transactional data, social media.

**5V in bigdata**

**Volume:** Defines the size of data and it’s a key parameter which can help us to understand if the give volume of data falls under bigdata or not.

**Variety**: Variety defines different forms of data which is produced by source like

Unstructured: Any data with unknown form or the structure is classified as unstructured data.

Example: Audio files, video files, Gif.

Semi-structed: Any data with both form unstructured or structure is classified as semi-structured data.

Example: JSON file, XML, Parque.

Structured: Any data with known form or the structure is classified as structured data.

Example: CSV, Excel etc.

**Velocity:** Determines the speed of data flow from source if the flow of data is massive and continuous then it’s considered as bigdata ex: social media.

**Variability:** This refers to inconsistency of data at times, if it is continuously changing, then it can have an impact on the quality of data.

**Value:** If the data extracted and processed brings some meaningful or profitable insights then that can be considered as a value of data. It’s nothing but Junk to treasure.

**Hadoop**

Hadoop provides a scalable solution to store and process huge data sets in parallel and distributed fashion.

The main components of Hadoop are

1. YARN: Yet another resource negotiator
2. HDFS: Hadoop distributed file system
3. MapReduce