**Big Data Assignment 3**

**1. What is Big Data, exactly? List a few sources that generate large amounts of data.**

According to Gartner, Big Data is a high-volume, high velocity, and high-variety information asset that demands cost-effective, innovative forms of information processing for enhanced insight and decision making. Big Data is a collection of data (both structured and unstructured) that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Some sources which generate a large amount of data are:

* Billions of mobile phones and tablets
* Billions of Mobile web subscribers
* Billion content pieces shared every day in TBs
* Millions of tweets generated every day
* Billions of intelligent devices
* Transaction data from many business

**2. What are the advantages of Big Data processing?**

There are several advantages of Big Data, they are:

* Better Decision Making
* Reduce costs of business processes
* Fraud Detection
* Increased productivity
* Improved customer service
* Increased agility

Apart from these advantages, in big data we scale out by increasing the number of systems involved instead of scale up. By following this mechanism big data is able use multiple commodity machines and distribute the load of storage / processing. This mechanism is economical as well as quick to implement.

**3. What causes Big Data projects to fail?**

The major reasons for big data projects failure are:

* **Improper integration:**
  + Most of the time to get the required insights, companies tend to integrate soiled data from several sources. It is not easy to build a connection to siloed, legacy systems. The cost of integration is much higher than the cost of the software. This makes simple integration one of the biggest problems to overcome.
* **Business assumptions and technical reality misalignment**
  + Most of the time, the technical capabilities don’t come up to business expectations. Organizations want the technology to be integrated to have unique functions. However, the capabilities of Artificial Intelligence and Machine Learning are limited. Being clueless about what the project is capable of doing, results in its failure. You need to be aware of the capabilities of the project before developing it.
* **Rigid project architectures**
  + Another major reason for failure is when the project architecture is inflexible and too rigid from the beginning. Moreover, certain companies wait to achieve a seamless model from the beginning rather than incrementally improving it as the project progresses.
* **Setting unachievable goals**
  + Sometimes, businesses set high expectations from the technology that’s about to be integrated into their systems. Some of these expectations are unrealistic and cannot be met. These expectations cause big data projects to fail miserably. Business leaders should set realistic goals while working on big data projects.
* **Models are taken into the production process**
  + No matter how much you invest in a big data project, it is of no use if you don’t move it into production. Machine Learning models are created by experts. However, they are left for months without anything happening. In the majority of the cases, IT companies are not equipped with the tool required to create an environment that handles an ML model. They don’t have skilled specialists with the expertise to handle these models.

**4. Give examples of the five v's of big data.**

The five v’s of big data they are:

* **Volume -** The size of the data
  + E.g., They handle more than 1 million customer transactions every hour, importing more than 2.5 petabytes of data into their database.
* **Velocity -** The speed at which the data is generated
  + E.g., Amazon captures every click of the mouse while shoppers are browsing on its website
* **Variety -** The different types of data
  + E.g., CCTV audio and video files that are generated at various locations in a city.
    - structured data (e.g., your customer ID, the timestamp of your call, your service type), and
    - unstructured data (e.g., the recording of the call, notes that the call centre operator makes during the call, the problem history related to your call).
* **Veracity:** The trustworthiness of the data in terms of accuracy
  + **E.**g., data from a medical experiment or trial.
* **Value:** Just having Big Data is of no use unless we can turn it into value
  + **E.**g., in Metlife, “We now know within a two-month period when it is highly likely that a customer will cancel his or her policy or purchase a new one.”

**5. What are the different types of resource managers in Spark.n?**

Apache Spark system supports three types of cluster managers namely-

* **Standalone Cluster Manager** 
  + Standalone mode is a simple cluster manager incorporated with Spark. It makes it easy to setup a cluster that Spark itself manages and can run on Linux, Windows, or Mac OSX. Often it is the simplest way to run Spark application in a clustered environment.
* **Hadoop YARN** 
  + YARN became the sub-project of Hadoop in the year 2012. It is also known as MapReduce 2.0.
  + YARN bifurcate the functionality of resource manager and job scheduling into different daemons.
  + The plan is to get a Global Resource Manager (RM) and per-application Application Master (AM).
  + An application is either a DAG of graph or an individual job.
* **Apache Mesos**
  + Mesos handles the workload in distributed environment by dynamic resource sharing and isolation. It is healthful for deployment and management of applications in large-scale cluster environments. Apache Mesos clubs together the existing resource of the machines/nodes in a cluster. From this, a variety of workloads may use.
  + This is node abstraction; thus, it decreases an overhead of allocating a specific machine for different workloads. It is resource management platform for Hadoop and Big Data cluster.