Big Data Assignment 3

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

Big Data is a collection of data 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.

The bulk of big data generated comes from three primary sources: *social data, machine data and transactional data.*

2. What are the advantages of Big Data processing?

1. Using big data cuts your costs

A recent Tech Cocktail article looks at how Twiddy & Company Realtors cut their costs by 15%. The company compared maintenance charges for contractors against the average of its other vendors. Through this process, the company identified and eliminated invoice-processing errors and automated service schedules.

2. Using big data increases your efficiency

Using digital technology tools boosts your business's efficiency. From using tools such as Google Maps, Google Earth, and social media, you can do many tasks right at your desk without having travel expenses. These tools save a great amount of time, too.

3. Using big data improves your pricing

Use a business intelligence tool to evaluate your finances, which can give you a clearer picture of where your business stands.

4. You can compete with big businesses

Using the same tools that big businesses do allows you to be on the same playing field. Your business becomes more sophisticated by taking advantage of tools that are available for your use.

5. Allows you to focus on local preferences

Small businesses should focus on the local environment they cater to. Big Data allows you to zoom in on your local client's likes/dislikes and preferences even more. When your business gets to know your customer's preferences combined with a personal touch, you'll have an advantage over your competition.

6. Using big data helps you increase sales and loyalty

The digital footprints that we leave behind reveal a great deal of insight into our shopping preferences, beliefs, etc. This data allows businesses to tailor their products and services to exactly what the customer wants. A digital footprint is left behind when your customers are browsing online and posting to social media channels.

7. Using big data ensures you hire the right employees

Recruiting companies can scan candidate's resumes and LinkedIn profiles for keywords that would match the job description. The hiring process is no longer based on what the candidate looks like on paper and how they are perceived in person.

3. What causes Big Data projects to fail?

5 Reasons why big data projects fail

1. Improper integration

Various technological problems cause big data projects to fail. One of the most important of these problems is 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.

Nothing magical will happen if you link every data source. The outcomes will be zero. One of the biggest parts of the problem is the siloed data itself. When you put data into a common environment, it is hard to figure out what the values mean. Knowledge graph layers are needed to enable machines to interpret the data mapped underneath. Without this information, you only have a data swamp that is of no use to you. Since you would have to invest in security to stop any potential data breaches, improper integration means big data will only be a burden on your company's finances.

2. 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.

3. Rigid project architectures

From resources to skills, talent to infrastructure, most companies have it all. Yet they fail to implement a successful big data project. Why does that happen? This happens 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.

Even if the project isn't complete and you haven't achieved the perfect model, it is still possible to acquire considerable business value. Even if you have a subset of data to work with, you can implement machine learning to reduce the associated risks.

4. 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.

5. Models are taken into the production process

This is one of the biggest reasons why most big data projects fail. 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.

Big data is a collection of data from many different sources and is often describe by five characteristics: *volume, value, variety, velocity, and veracity.*

- **Volume:** the size and amounts of big data that companies manage and analyze
- **Value:** the most important “V” from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- **Veracity:** the “truth” or accuracy of data and information assets, which often determines executive-level confidence

The additional characteristic of variability can also be considered:

- **Variability:** the changing nature of the data companies seek to capture, manage and analyze – e.g., in sentiment or text analytics, changes in the meaning of key words or phrases

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

There are three Spark cluster manager, Standalone cluster manager, Hadoop YARN and Apache Mesos. Apache Spark supports these three type of cluster manager.

Cluster manager is a platform (cluster mode) where we can run Spark. Simply put, cluster manager provides resources to all worker nodes as per need, it operates all nodes accordingly.

We can say there are a master node and worker nodes available in a cluster. That master nodes provide an efficient working environment to worker nodes.

There are three types of Spark cluster manager. Spark supports these cluster manager:

Standalone cluster manager

Hadoop Yarn

Apache Mesos

Apache Spark also supports pluggable cluster management. The main task of cluster manager is to provide resources to all applications. We can say it is an external service for acquiring required resources on the cluster.

1. Standalone Cluster Manager

It is a part of spark distribution and available as a simple cluster manager to us. Standalone cluster manager is resilient in nature, it can handle work failures. It has capabilities to manage resources according to the requirement of applications.

We can easily run it on Linux, Windows, or Mac. It can also access HDFS (Hadoop Distributed File System) data. This is the easiest way to run Apache spark on this cluster. It also has high availability for a master.

2. Hadoop Yarn

This cluster manager works as a distributed computing framework. It also maintains job scheduling as well as resource management. In this cluster, masters and slaves are highly available for us. We are also available with executors and pluggable scheduler.

We can also run it on Linux and even on windows. Hadoop yarn is also known as MapReduce 2.0. It also bifurcates the functionality of resource manager as well as job scheduling.

3. Apache Mesos

It is a distributed cluster manager. As like yarn, it is also highly available for master and slaves. It can also manage resource per application. We can run spark jobs, Hadoop

MapReduce or any other service applications easily.

Apache has API’s for Java, Python as well as c++. We can run Mesos on Linux or Mac OSX also.