ASSIGNMENT # 1

**1.** Explain **using your own words**, in 4 or 5 lines, the difference between *Big Data*, *Data Analytics* and *Distributed Systems*.

- Big Data refer to a very large set of data that no conventional database management or information management tool can really work with. Indeed, we procreate trillions of bytes of data every day. It's the information from everywhere: messages we send to each other, videos we post.

- Data analytics is the process of exploring, transforming and analyzing data to identify trends and patterns that reveal relevant actionable insights and improve efficiency.

- A distributed system is a collection of computer programs that use computing resources on multiple separate computing nodes to achieve a common, shared goal. Distributed systems aim to eliminate bottlenecks or central points of failure in a system.

**2. Go to:**

<https://www.datamation.com/big-data/big-data-pros-and-cons.html>

Make the reading and answer the following questions:

1. What are the nine pros of Big Data? Explain each one of them in one or two lines.

1/ Better decision-making: it helps us to make better decisions in storage to give and be better organized and have all the data possible recorded in our data  
  
2/ Increased productivity: the latest data tools help us work faster and save more time helping workers in this field to be more productive  
  
3/ Reduce costs: reducing costs is the most important element for companies and it costs them less, which is why it is advantageous for them  
  
4/ Improving customer service: it is for the improvement of customer service and the latter responded successfully  
  
5/ Fraud detection: this helps to detects any fruad activity or illegal scam acitivites  
  
6/ Increase in income: the majority of companies use lr big data for better results and better revenues  
  
7/ Increased agility: It is better agility in all organisations. Because support faster and more frequent changes to their business strategies and tactics.   
  
8/ More innovation: there is a vast oppurtunites of innovating using big data. Almost all moder technologies are running through bigdata.  
  
9/ Faster time to market: Big data is very fast and update in market easily

1. What are the nine cons of Big Data? Explain each one of them in one or two lines.

1/ Need for talent: big data nowadays really helps us but to change all our systems takes us more time .

2/ Data quality: for data quality need high technical support

3/ Need for cultural change: need to change time to time

4/ Compliance: Failing to comply with these regulations can lead to severe legal consequences, including hefty fines, legal action, and reputational damage.

5/ Cybersecurity risks: there is high chance of cyber risk

6/ rapid evolution the technology improves more and more so that's why you have to be up to date.

7/ Hardware needs: it is a big problem for organizations to record huge size data it is not always easy.

8/ Costs: big data tools are very advantageous but are very expensive for companies.

9/ Difficulty integrating legacy systems: the old ones find it difficult to integrate because their data is stored differently

**3.** Is Hadoop and Big Data the same? (Give your personal answer in 5 or maximum 6 lines).

No, Hadoop and Big Data are not the same thing. Big Data refers to the large volume, variety, and velocity of data that is being generated and collected today. Hadoop is a software framework designed to store and process Big Data across distributed computer systems. Hadoop provides a way to store and process large volumes of data across a cluster of commodity hardware, while also offering fault tolerance and scalability. Other Big Data technologies databases can also be used for processing and analyzing Big Data.

**4.** Scenario:

1. You have 100 terabytes of music: 44 terabytes of Indian music, 33 terabytes of rock music and 23 of electronic and hip-hop music. Then, create the partitioning and the replication in different nodes.

5. Practice: Intro to *Hadoop*.

A) Go to:

1. The virtual machine in the computer.  
     
   B) Send me your *gmail* account. If you do not have one, you have to create it, because I will give you access to two *Cloudera* files.  
     
   C) Go to your *gmail* account, download and import the Cloudera files to the virtual machine.  
     
   D) Run Cloudera