# **Explain the main sources of data flood:**

There is a huge amount of information available online. And its volume is growing at lightning speed.  
1. The New York stock exchange generates about 4-5TB of data every day.

2. Facebook hosts more than 240 billion photos growing at 7PB of data every day.

3. The genealogy sites stores 10PB of data.

4. The internet archive stores around 18.5PB of data.  
 *Each minute on average:*   
5. More than 200 million emails move across the Internet (though most are spam).   
6. Twitter users post more than 300,000 new tweets.   
7. People across the globe share more than 38,000 Instagrams.   
8. YouTube users upload another 100 hours of video.  
9. Google processes more than 3.6 million web searches.   
10. 2.2 million things on Facebook get a “like” or a comment.

*It also comes from the study of volumes of raw facts, called data.*For example: - biologists collect enormous numbers of measurements on millions of cells and everything inside them.   
Astronomers fill banks of hard drives with observations of stars, galaxies and energy in deep space.   
Earth scientists assemble detailed snapshots of weather, including patterns of winds and waves throughout the world.

# **What is the difference between data and big-data?**

|  |  |
| --- | --- |
| ***DATA*** | ***BIGDATA*** |
| Data is Facts and statistics collected together for analysis but not necessarily organized in a way that give them meaning. | Big data is a term for data sets that are so large or complex that traditional data processing application soft wares are inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, and updating and information privacy. |
| It is a set of instructions which can be in any form i.e. images, photos, score cars, mobile bills, etc. It is an important part as it provides better decision by analyzing the data but its capacity is limited. | Whereas, 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. |
| Data is data in which we have sense of where it coming from and how much there will be. | Big data is data from sources in which we have no way to estimate how large it will be, how much it will grow and how much it will change |
| Data Analytics on the other hand is more of analyzing data which could be structured or unstructured. | Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. |

**What are the main reasons behind Hadoop becoming the solution for Data explosion**?

1. Hadoop being the open source, java based programming frame work thus supports huge data in distributed form.

2. Provides reliable, scalable platform for storage and analyses.

3. It is based on GFS (google file system).

4. It works in two parts i.e. HDFS and Map Reduce.  
 - With the combination of these technologies, massive amounts of data can be easily stored, processed and analyzed in a fraction of a second.

5. Hadoop supports data-intensive distributed applications that can run simultaneously on large clusters of normal, commodity, hardware.  
6. A Hadoop network is reliable and extremely scalable and it can be used to query massive data sets.   
7. Hadoop is written in the Java programming language, meaning it can run on any platform, and is used by a global community of distributors and [Bigdata](https://datafloq.com/read/?page=1&cat=31#%3Futm=internal) technology [vendors](https://datafloq.com/search?utm=internal)  
who have built layers on top of Hadoop.  
8. There are several reasons why organizations use Hadoop, being:

1. Low cost
2. Computing power
3. Scalability
4. Storage flexibility
5. Data protection