Jamia Millia Islamia

Dept. Of Computer Science

Big Data Assignment-#4

Submitted By: Wasit Shafi **Submitted To**: Dr.Mansaf Alam

Roll no: 18MCA054

Q1.What is big data analytics? Differentiate between traditional data analytics and big data analytics.

Sol. Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information -- such as hidden patterns, unknown correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

Traditional data analytics and Big data analytics:

Parameters	Traditional Data	Big Data
Volume	GB	Constantly Updated(TB or PB currently)
Generated Rate	Perhour,Per day.	Morerapid(almostevery second)
Structure	Structured	Semi-structured&unstructured
Data Source	centralized	Fully distributed
DataIntegration	easy	Difficult
Data Store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real time
Update Scenarios	Repeated read and write	Write Once, Repeated Read
Data Structure	Static Schema	Dynamic Schema
Scaling Potential	Non-linear	Somewhat close to Linear

Q2.What are various sources of big data? Discus word scenario of big data.

Sol.

Top 5 sources of big data

1. **MEDIA AS A BIG DATA SOURCE**

Media is the most popular source of big data, as it provides valuable insights on consumer preferences and changing trends. Since it is self-broadcasted and crosses all physical and demographical barriers, it is the fastest way for businesses to get an in-depth overview of their target audience, draw patterns and conclusions, and enhance their decision-making. Media includes social media and interactive platforms, like Google, Facebook, Twitter, YouTube, Instagram, as well as generic media like images, videos, audios, and podcasts that provide quantitative and qualitative insights on every aspect of user interaction.

2. **CLOUD AS A BIG DATA SOURCE**

Today, companies have moved ahead of traditional data sources by shifting their data on the cloud. Cloud storage accommodates structured and unstructured data and provides business with real-time information and on-demand insights. The main attribute of cloud computing is its flexibility and scalability. As big data can be stored and sourced on public or private clouds, via networks and servers, cloud makes for an efficient and economical data source.

3. **THE WEB AS A BIG DATA SOURCE**

The public web constitutes big data that is widespread and easily accessible. Data on the Web or ‘Internet’ is commonly available to individuals and companies alike. Moreover, web services such as Wikipedia provide free and quick informational insights to everyone. The enormity of the Web ensures for its diverse usability and is especially beneficial to start-ups and SME’s, as they don’t have to wait to develop their own big data infrastructure and repositories before they can leverage big data.

4. IOT AS A BIG DATA SOURCE

Machine-generated content or data created from IoT constitute a valuable source of big data. This data is usually generated from the sensors that are connected to electronic devices. The sourcing capacity depends on the ability of the sensors to provide real-time accurate information. IoT is now gaining momentum and includes big data generated, not only from computers and smartphones, but also possibly from every device that can emit data. With IoT, data can now be sourced from medical devices, vehicular processes, video games, meters, cameras, household appliances, and the like.

5. DATABASES AS A BIG DATA SOURCE

Businesses today prefer to use an amalgamation of traditional and modern databases to acquire relevant big data. This integration paves the way for a hybrid data model and requires low investment and IT infrastructural costs. Furthermore, these databases are deployed for several business intelligence purposes as well. These databases can then provide for the extraction of insights that are used to drive business profits. Popular databases include a variety of data sources, such as MS Access, DB2, Oracle, SQL, and Amazon Simple, among others.

Big Data- Current Scenario

Big data is the big word in the current technical landscape. It has forced industries, governments, academicians and researchers to give a serious thought considering that they can mine the knowledge to increase the impact of government policies, operations of private organizations and businesses for taking better and wiser decisions for their stakeholders. The experts in Data Science consider it a game changer for every industry. The world around us and the activities that we have been doing has created a whole mesh of massive data in every area. It has shifted from hype to a real situation that must be dealt with. It has been raising many questions like how to manage big data? How useful is big data? How big data can influence the current data science scenario?

The biggest challenge faced by a new creed of professionals termed data scientists is how to quantify the value of big data? It is a huge challenge that needs funding with an assurance to the stakeholders that the investment is worth the risk involved. Several industries are investing hugely on big data. Hadoop, the most preferred open source software for distributed computing, has forcasted to grow by 58% by the year 2020. So, it is a clear indication towards the fact that the world belongs to big data now and investing in this field is definitely a sign for stepping in right direction.

Organizations can have many reasons to go ahead and jump into the sea of opportunities offered in this promising avenue. It can be better customer experience, targeted marketing, concise analysis of business, reduction in expenses, securing the business and expanding the customer base.

Q3.What are various characteristic of big data?

Sol.

The following are some Characteristics of Big Data.

- **Volume**
- **Variety**
- **Veracity**
- **Value**
- **Velocity**

1. Volume

Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot. We are currently using distributed systems, to store data in several locations and brought together by a software Framework like Hadoop.

Facebook alone can generate about billion messages, 4.5 billion times that the “like” button is recorded, and over 350 million new posts are uploaded each day. Such a huge amount of data can only be handled by Big Data Technologies

2. Variety

As Discussed before, Big Data is generated in multiple varieties. Compared to the traditional data like phone numbers and addresses, the latest trend of data is in the form of photos, videos, and audios and many more, making about 80% of the data to be completely unstructured

3.Veracity

Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments

4.Value

Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable and trustworthy data that needs to be stored, processed, analyzed to find insights.

5.Velocity

Last but never least, Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data. So, the major aspect of Big Data is to

provide data on demand and at a faster pace.

Q4.What is Hadoop? Discuss data loading technique to HDFS.

Sol.

Hadoop: Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

There are various tools and frameworks available in Hadoop Ecosystem, using which you can import and export data to HDFS & from HDFS respectively.

Few of them and their use cases:

- **Hdfs dfs -put** - simple way to insert files from local file system to HDFS.
- **HDFS Java API.**
- **Sqoop** - for bringing data to/from databases.
- **Flume** - streaming files, logs.
- **Kafka** - distributed queue, mostly for near-real time stream processing.
- **Nifi** - incubating project at Apache for moving data into HDFS without making lots of changes.