Test 2 – Big Data Fundamentals

Date – 30-Nov-2021 (Tue)

Student Name – Aadarsha Chapagain

Student ID – C0825975

Marks 10

Q1. What are biggest privacy issues with Big Data? Talk in detail about a government legislation which impacts privacy.

Some of the privacy issues with Big data are

Discrimination or bias

Data related to the individual and organization are collected from various sources such as social media and internet. These platform contains the information specifics to the user. When analysis is performed the data might be biased and certain insight or result may point towards a specific community or group which create discrimination. For example, an google AI used is image processing classified an African man as an ape. Such results or scenario clearly creates discrimination.

Breaches

There are always great deal of chances of data breaches which leaks private data of the users. Data breaches are not new things since the rise of IOT and social media and are happening quite often which snitch the rights to privacy of users.

Accuracy

Sometimes the accuracy of the big data is questionable and cannot be relied upon. These may happen to errors or inaccuracy in the collected data or simply because of bad choice of algorithm. This might result in bad prediction or poor decision making in the organization since every decision is based on insights obtained from the data these days. When the decisions are made which included individuals they might have to suffer because of the inaccuracy of data which is serious in case if medicines, diagnosis and services.

Anonymity can be useless.

Expressing opinion and views about different topics and subject matter without revealing the identity is considered as anonymous behaviour in internet. But this is near to impossible with such computational power and amount of data collected. For example, Recently Netflix releases its users rating for the movies without revealing the identities of the users but some scientist were

able to co-relate that data to imdb rating of the users. Although Netflix released rating with specifying users information it could be related to use different means.

The seven foundation principles of privacy by design is the concept developed to ensure the privacy of the user. The privacy policy needs to be adapted while designing or developing any kind of information or communication technologies or large scale networked system.

Proactive not reactive: preventative not remedial
This policy expects that there is always the risk of privacy invasion so it suggest that rather then waiting for invasion to happen, efforts should be dedicated not to stop it or to take preventive measure before that happen.

Privacy as the default setting
Privacy should not be achieved after configuration or managing different parameters in every system the privacy should be the default setting or builtin setting. Individual should not be expected to work on setting privacy.

Privacy Embeded in design
While designing every system or technology privacy should be considered. Without reducing any functionality privacy should be integral part of the system.

Full functionality: positive-sum, not zero-sum
All the legitimate objectives should be there in the system which creates positive sum or win-win situation. There should be no trade off between usability and privacy in any system.

End to End encryption
There should be the provision of privacy from the first steps of information collected to finish line and throughout the life cycle of the data involved.

Visibility and transparency: keep it open
Individual should always know how there information is collected and used at different steps or process in the system and should be used only after they agree about it.

Respect for user privacy: keep it user-centric
The user should be given the choice to allow and refuse the use of their data. Strong privacy by default and notice are some of measures that can be used to do so.

Test 2 – Big Data Fundamentals

Date – 30-Nov-2021 (Tue)

Q2. What is Data Warehouse and Data Lake, list similarities and differences.

Data warehouse is repository of the data collected from different source and medium to support the analytical solution in the organization. The analytical solution may be reporting, decision support system, different kinds of queries. Data are cleaned, integrated, and stored in the form in which analysis can be performed in Data warehouse.

A data lake is repository which collect the structured and instructed data from the different source. Usually, data lake stores the data in its own or native format without any cleaning and integration.

Both the data warehouse and data lake are repository of the data collecting the data from different sources but the difference lies between their objectives. Data ware has predefined set of objectives which is not the case in data lake.

Some of the differences in the data warehouse and data lakes are in terms of points discussed below.

Storing of data

While developing data warehouse there are certain analytical and decision-making question related to the data which needs to be answered and according to that questions the required data are stored. The unnecessary data which does not help in achieving the objective are removed from the data warehouse. But the scenario is different in case of the data lake, data lake stores all the data that has been collected in the hope that it might be used some day.

Data types

In data warehouse data are collected and transformed in the format or types that can be used easily. After collection the data from the heterogenous source the data are cleaned and transformed in the certain format in the data cleaning and integration phase.

In the data lakes there is no cleaning and transforming phase, data are mostly stored in the form in which they were collected.

Processing of the data

There is major difference between data lake and warehouse in processing of data. Data lake use Extarct, load and transform (ELT) to process data where as Data warehouse uses Extract, Transform and load(ETL) so that only transformed and filtered data are loaded in the warehouse.

Users

Data lake is ideal for all kind of users. A data scientist or any user can go to the data lake and obtained the type of the data he need to achieve his goal but data warehouse may be ideal only for some kind of users.

Test 2 – Big Data Fundamentals

Date – 30-Nov-2021 (Tue)

Since most of the organization use operational or transactional data they will go to data warehouse to perform analysis and reporting so for those users data warehouse is ideal. For users in organization performing day to day work data lake does not make any sense.