SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT I – INTRODUCTION TO DATA SCIENCE

RETRIEVING DATA

RETRIEVING DATA

- Retrieving required data is second phase of data science project.
- Sometimes Data scientists need to go into the field and design a data collection process.

1. Start working on internal data, i.e. data stored within the company

EMPATHY:

- First step of data scientists is to verify the internal data.
- Assess the relevance and quality of the data that's readily in company.
- Most companies have a program for maintaining key data, so much of the cleaning work may already be done.
- This data can be stored in official data repositories such as databases, data marts, data warehouses and data lakes maintained by a team of IT professionals.
- Data repository is also known as a data library or data archive.
- This is a general term to refer to a data set isolated to be mined for data reporting and analysis.
- The data repository is a large database infrastructure, several databases that collect, manage and store data sets for data analysis, sharing and reporting.

DEFINE:

- Data repository can be used to describe several ways to collect and store data:

a) Data warehouse is a large data repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related.

b) Data lake is a large data repository that stores unstructured data that is classified and tagged with metadata.

c) Data marts are subsets of the data repository. These data marts are more targeted to what the data user needs and easier to use.

d) Metadata repositories store data about data and databases. The metadata explains where the data source, how it was captured and what it represents.

Advantages of data repositories:

i. Data is preserved and archived.

ii. Data isolation allows for easier and faster data reporting.

iii. Database administrators have easier time tracking problems.

iv. There is value to storing and analyzing data.

Disadvantages of data repositories :

i. Growing data sets could slow down systems.

ii. A system crash could affect all the data.

iii. Unauthorized users can access all sensitive data more easily than if it was distributed across several locations.

2. Do not be afraid to shop around

IDEATE:

- If required data is not available within the company, take the help of other company, which provides such types of database.

- For example, Nielsen and GFK are provides data for retail industry. Data scientists also take help of Twitter, LinkedIn and Facebook.

- Government's organizations share their data for free with the world.

- This data can be of excellent quality; it depends on the institution that creates and manages it.

- The information they share covers a broad range of topics such as the number of accidents or amount of drug abuse in a certain region and its demographics.

3. Perform data quality checks to avoid later problem

TESTING:

- Allocate or spend some time for data correction and data cleaning. Collecting suitable, error free data is success of the data science project.

- Most of the errors encounter during the data gathering phase are easy to spot, but being too careless will make data scientists spend many hours solving data issues that could have been prevented during data import.

- Data scientists must investigate the data during the import, data preparation and exploratory phases. The difference is in the goal and the depth of the investigation.

PROTOTYPING:

- In data retrieval process, verify whether the data is right data type and data is same as in the source document.

- With data preparation process, more elaborate checks performed. Check any shortcut method is used. For example, check time and data format.

- During the exploratory phase, Data scientists focus shifts to what he/she can learn from the data.

- Now Data scientists assume the data to be clean and look at the statistical properties such as distributions, correlations and outliers.