Open Datasets: Need of Hour

Prof. Shubhangi M. Choudhary
(Research scholar)
J. J. T. University,
Rajasthan, India
shoobhs@gmail.com

Dr.Avinash S. Jagtap
T. C. College, Baramati
Dist. Pune, India
avinash.jagtap65@gmail.com

ABSTRACT- The term open data is a type of data set which is open for access to everyone, not only open but it is available for access, reuse, modification and sharing. Governments, agencies and independent organizations have contributed on large scale and opened the floodgates of data to create and share more and more data for free and for easy access. As data skills are increasingly valuable in around every job market and in a growing professional field one just cannot simply avoid data. It is not just for big businesses but for every researchers and academicians need of data is every now and then. This research paper provides a brief introduction to the open datasets and the seven laws of universal data. Thing of relief is that every time you need not have to collect data on you own to analyse it. Tons of public data sets are available open and free to access. This paper aims to understand datasets along with its importance and provides a list of most used and popular open datasets. A comparative on open dataset and open data repository is also included in the study.

Keywords: open dataset, UCI, universal dataset, machine learning repository, open data repository

I. INTRODUCTION

In simplest word dataset is defined as a group or collection of data. Generally a dataset corresponds to the database table or statistical data matrix where every element of data matrix represents a particular value and provides information. In other words dataset can be termed as a container that holds data uploaded to Analytics. It gives a structure to maintain and manage uploaded data as well as controls how existing data and uploaded data get linked. Data set can be associated with multiple databases but association with at least one database is a must.

The term open data is a type of data set which is open for access to everyone, not only open but it is available for access, reuse, modification and sharing. Governments, agencies and independent organizations have contributed on large scale and opened the floodgates of data to create and share more and more data for free and for easy access. Open data term developed its foundation from various “open movements” such as open hardware, open source, open science, open government etc. As data skills are increasingly valuable in around every job market and in a growing professional field one just cannot simply avoid data [1]. It is not just for big businesses but for every researchers and academicians need of data is every now and then. Thing of relief is that every time you need not have to collect data on you own to analyse it. Tons of public data sets are available open and free to access. To analyse data or create data visualizations, public data sets are available. For help regarding putting, finding into form these open data sets also have write-ups on data visualization blogs along with some data visualization examples for beginners. While dealing with open data there are two terminologies one come across viz. open data set and data Repository.

Specific type of rows or records of data related to any particular entity (like bank accounts, student records etc.) is a database and repository is common term for any central storage. So Repository data is a common central point that stores files whereas a database is a set of structured files that store and organise data. Repository is a dataset where metadata for Designer objects is stored like data related to objects like modules, table, entities, and definitions etc.as part of the Repository. For maintenance of this data it contains a PL/SQL Application Programming Interface (API). Like data set it can be associated with multiple databases but association with at least one database is mandatory.

This is usually done when there is a 'higher purpose' for the data, but the data items needed to do this on different databases. In these cases a repository is necessary to bring together the discrete data items and operate on them as one. This database Repository is generally logical but sometimes physical grouping of data linked but data from distinct databases is also present. Basic difference between database and data Repository is that database is used for storing various data and Repository is for conserving and harvesting data for long term. Important thing is Repository uses data set and makes it viewable and searchable by putting lots of layers on it.

II. UNDERSTANDING DATA SETS

- **Data Set schema**

A schema is a structure that joins with the uploaded data with the existing data in hits. A simple schema comprises of a metric or an import dimension. Analytics looks for key values in hits that match with the key values in the uploaded data to import data. The additional metric values and dimensions associated with that key are added to the existing hit data when a match is found [2]. Some datasets can be accessed through multiple dimensions for the import fields.

- **Data Set types**

The particular type of data one desires to import is a data set type. Some data set types are Cost Data, User Data, Content Data, etc. as well as in which format it is required like image, audio, video, robotics etc. along with the different options for the metrics (the schema) and dimensions. While creating a Data Set, full list of metrics and dimensions is available. Some of these available dimensions and metrics can be used as import key and targets but all cannot be applied.

III. Importance of open data

Data and business have become two inseparable entities as the whole world has grown increasingly driven by data. Open data have become an integral and vital part of governance and business. The idea of data-driven governance and business is difficult to be emerged if data is not easily available and put restrictions on use and access of data [3].The governments and societies built the systems and processes and open data plays key role in reforming these. It can strengthen democracy and empower citizens with knowledge. It helps in transforming the way world engages and understands. Universal issues and global problems are well understood and as well as aware than ever before due to the spread of knowledge possible because of open data. Easy data access encouraged more and more researcher to research. Open data gives big boost to the business by opening new horizons. It is a great stimulus for machine learning. So open data is irreplaceable and has exclusive role in the development of business, governments and ultimately societies.

IV. THE SEVEN LAWS OF UNIVERSAL DATA

The term “universal data” refers to universal database which is in turn a logical thread for the organisation. This logical data thread gives access to all data and provides a freedom to use and place data as per need and requirement. The seven laws are designed by SAP in such a way that it ensures universal data solution is truly universal. By adhering to these laws value of data is maximised not only in current time but also in future.

These seven laws are,

1. Performance
2. Freedom
3. Model
4. Independence
5. Low Latency
6. Governance
7. Frictionless

A. Performance

Performance is basic and fundamental thing. An infrastructure is needed to deliver data at the speed of business demands it and as per the requirement of users. The complaint like “I can’t get access to the data when I want it, where I want it.” is still there. Elimination of the latency that

ISSN: 2394-3114
Vol-40-Issue-74-March -2020

results from segregating data into different tasks or environments from universal data solution should be achieved. The performance demanded by all processes and users must be delivered otherwise none of the rest will really matter.

B. Freedom

The word open or universal signifies freedom. So the data must be available free to use without any restrictions and constraints. In world of Industry 4.0 and digital transformation data is everywhere. It is not restricted to any enterprise. Data is accessible as well as available inside and outside four walls of enterprise, on premises and in the cloud as well. Data is generated by systems and users also data is available from a wide variety of third parties as well. So every bit of this variety of data must flow effortlessly everywhere. The capacity to innovate and to enable speedy growth depends on this freedom.

C. Model

Once the data is accessible it is important to be able to get and comprehend the data to model it as per the specific business requirements. Because getting access to right kind of data is the first step and modeling is next one important and crucial thing to achieve. Data must be assimilated to provide a 360-degree business view. So it is challenging to find and understand all of the data so that it can be modelled efficiently to meet business goals. Enterprises can operate flexibly and change direction as innovative discoveries are made if it is possible to develop sensible and precise models quickly. In this Big Data era, for data management process the most challenging aspect is data discovery reason behind is to find relationships and acumens concealed within the data.

D. Independence

Data must be independent of one’s computing limitations and vice a versa. Computing ability or power and data, each must be able to scale freely. Extra care needs to be taken to ensure that data must remain independent and elastic taking into consideration about the spread between them. This should be attained with low latency. Any ways Yesterday’s data is generally twenty four hours too late.

E. Low Latency

The time span between when data is recorded and when it is needed again is latency. Latency operations must be as less as possible. The need for high performance standards and the need for low latency operations go hand in hand. No one is willing to wait for longer not systems, users, processes and customers for data [5]. For automated processes, incorporation into transactions, and analysis, needs data to be available and ready to process as soon as it is created. It is need of an hour as the pace of current business has become too brisk.

F. Governance

Security of data from tampering and theft is of prime importance. It is major risk in case of open data or universal data and therefore of prime importance that data must be secured and in safe hands in this rapidly changing environment. Along with the security of data accessibility and usability of it must not only compromised but ensured. So some standards regarding data integrity must be there those will ensure that these standards are applicable to all relevant processes and users. Though, the security standards may vary in some use cases and modes.

G. Frictionless

Businesses nowadays have to operate in many different ways at the same and this is one of the biggest problems they face. Critical process needs to deal with mashing up structured data of one Mode with unstructured data of another Mode. And in addition to this complex new source of data can land any time. These all activities must interoperate flawlessly. Without revealing the undercurrent intricacy of the variations between different data types the enterprise data thread act as a true universal database by fetching all these diverse data sources together as the user wants them. Universal data management framework for the enterprise accepts this challenge to support all seven of these laws.

V. List of most used and popular Data Repository

- World Bank Open Data
- WHO (World Health Organization)—Open data repository
- Google Public Data Explorer
- Registry of Open Data on AWS (RODA)
- European Union Open Data Portal
- FiveThirtyEight
- U.S. Census Bureau
- Data.gov
- DBpedia
- freeCodeCamp Open Data
- Yelp Open Datasets
- UNICEF Dataset
- Kaggle
- LODUM
- UCI Machine Learning Repository

List ofseven famous public Open Data Sets available for free to access and analyse

- Google Trends (Curated by: Google)
- National Climatic Data Centre (Curated by: NOAA)
- Global Health Observatory data (Curated by: World Health Organization (WHO))
- Data.gov.sg (Curated by: Singaporean government)
- Earthdata (Curated by: NASA)
- Amazon Web Services Open Data Registry (Curated by: Amazon)
- Pew Internet (Curated by: Pew Research Centre)

VI. UCI MACHINE LEARNING REPOSITORY

It is a universal data Repository for databases, data generators and domain theories for theempirical analysis of machine learning algorithms used by the machine learning community. At present there are 463 datasets in this repository available to the machine learning community. Intelligent Systems at the University of California, Irvine and The Centre for Machine Learning maintains and hosts it. Originally it was created by David Aha as a graduate student at UC Irvine [4]. Since then researchers, educators and students from all over the world are using it as a reliable source of machine learning datasets

In this Repository each dataset has its distinct webpage that enlists all the available details along with related publications that explore it. These datasets can be downloaded as ASCII files but the more useful format is CSV. Specifications of datasets are summarized by characteristics like attribute types, number of attributes, number of instances, and year published that can be sorted and searched.

VII. CONCLUSION

Open data or universal dataset is the need of hour. It has become vital part in today’s business, research and academic zone as the world is largely and rapidly being driven by data. Because of the availability of universal dataset the understanding of the universal problems and global issues is possible. It also helps in enhancing the business and research. It is a soul of machine learning. If the access and use of data is restricted then the concept of data driven governance and business cannot be materialized.

Open dataset can aid in transforming the way one involves and comprehend with the world. Researchers, academicians and students from all over the globe make use of it as a trustworthy source of datasets for machine learning techniques. So the universal open dataset has its own exclusive place in today’s data-driven world of artificial intelligence and machine learning.

REFERENCES

[1] David Marco, Michael Jennings (2004) "Universal Metadata Models", ISBN – 0-471-08177-9, Wiley Publishing Inc.

[2] David Marco (2000) "Building and Managing the Meta Data Repository: A Full Lifecycle Guide", ISBN- 0-471-35523-2, Wiley Publishing Inc.

[3] Graham Pryor (2012) "Managing Research data", ISBN – 978-1-85604-756-2, Facet Publishing

[4] Lisa R. Johnston (2017) "Curating Research Data, Volume 1", ISBN – 978-0-83898-858-9, Association of College and Research Libraries, a division of the American Library Association.

[5] Robin Rice, John Southall (2016) "The Data Librarian's Handbook", ISBN – 978-1-78330-183-6 (e-book), Facet Publishing