The Imperative of Big Data Integration in Geospatial Technologies for Epidemiological Surveillance and Population Health Risk Mapping

Dr. Lukas Weber[1]*
Prof. Amina El-Sayed[1]
[1] University of Toronto, Department of Biomedical Geoinformatics and Public Health Data Science, Toronto, Canada

ABSTRACT

Geospatial, as the name implies is an area of interest that satiates the needs of its end Users' geographic queries with a spatial touch that is, combining descriptive data about an entity along with it's a more important locational identity without which the former will be incapacitated to a great extent. This invaluable asset though still in its nascent stage due to minimal awareness among people can be given a shot in the arm when equipped with another impending boon named the Big data. The Big data concept has made its presence in the modern world so conspicuous thereby making its minuscule flaws to go unnoticed in the proceedings. This Paper attempts to identify in what ways the big data concept can be seamlessly incorporated in Geospatial domain so that both work in tandem in providing problem-oriented solutions from an ever reliable analytical perspective.

Keywords: *Big Data concept, Geospatial technology, Applications of Big Data, Data Sources, Statistical inferences*.

I. INTRODUCTION

Big Data is nothing but massive chunks of data sets that are constantly emanated or relayed by a myriad number of Sensors, wired/wireless network devices etc. Any device that outputs data is considered to be a big data contributor no matter whether it of conventional or contemporary type. Forinstance, a Datalogger which is used in Retail Stores that systematically stores the quantity and quality data about a particular product or a Weather sensor that relays periodic data regarding weather parameters. In a crux, any data set that renders a traditional data processing tool such as Relational data model useless for them to be processed is considered to be Big data. Though Size is not the only deciding factor in qualifying Big Data as since a data that proves to be big enough for one organization may not be for another due to their varying storage capacities and abilities to process, it remains to be one among many factors in the likes of data generating the frequency.

Five Concepts literally named the 5 Vs that describe the nature of Big data is:

- Volume: As discussed it represents the sheer size of the data which demands tens of thousands of Servers to run simultaneously to process and perform analytics so as to extract useful information from raw data that helps in discovering evolving trends and patterns obscured in them.
- Velocity: It describes the rate at which data are generated from a particular source and the time it takes to get processed into valuable information, every datum has its own characteristic Velocity as the frequency of generation and processing duration may not be same for a temperature sensor data and an Organization's employee data where the former tends to relay almost continually while the latter remains dormant for a while with some occasional updates.
- Variety: It describes the multidimensional aspect which clearly states that Big data is not restricted within the limits of a particular data type thereby encompassing multitudes of data in the likes of photos, videos, music files, transaction data, demographic data, inventory data that are available in different formats.
- Veracity: The larger the quantity of data, the more likely it is to contain errors. Since growing technology that has led to the umpteen deployment of cheap and affordable Internet of Things Sensors to capture data there remains a dearth of possibility for them to get misinterpreted and that is where this concept aims to stress upon. Data reliability and quality must be kept in mind if one needs to tap the benefits of Big data.
- Value: Last but not at all the least is the value we extract from a structured, semi-structured or an unstructured data through Big data analytical software such as Apache Hadoop framework that optimizes the data to suit the end user's needs.

II. METHODOLOGY

Since the very purpose of this paper is to incorporate the benefits of Big Data concept in Geospatial domain certain questions have been formulated to analyze why there is a need to gravitate towards it amidst existent data processing and analytical tools.

- What makes Big Data the need of the hour in Geospatial domain?
- How Big Data model fares among the rest?
- What is it's variants in processing?
- What the future holds?

These areas of debate were analyzed and answered in subsequent sections.

III. ANALYSIS

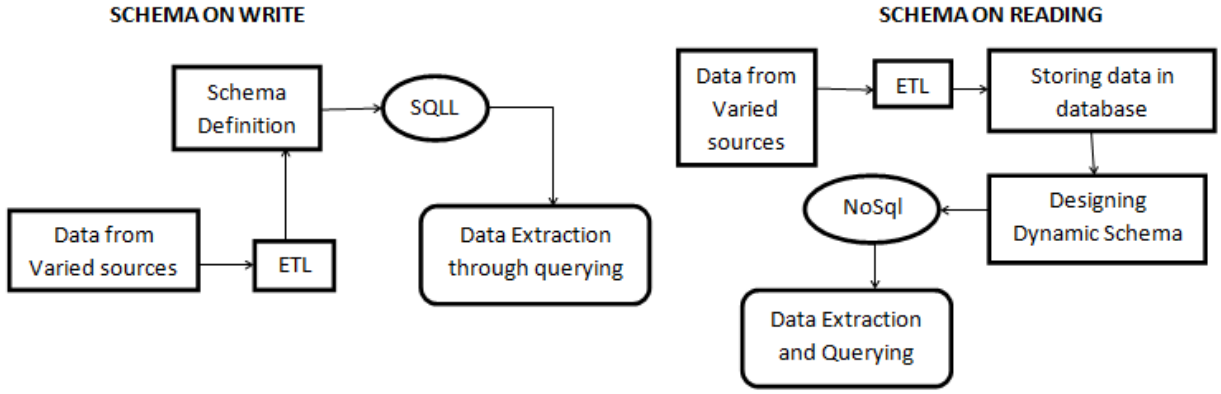
The necessity of Big Data

The reason why Big data concept is currently ruling the roost is its ability to get processed in quick time using analytical software that adopt inferential statistics technique so as to perform predictive analytics that can aid in preventing, combating or mitigating the effects of impending events, say for instance when it comes to addressing commonly spoken issues such as Ocean warming, a recent study published by NOAA Corp Scientists in Nature Journal has estimated the phenomenon's rate would be more virulent henceforth than assumed through the massive amounts of data collected from 3900 Agro Floats which are nothing but temperature sensors deployed 2000 meters below mean sea level since 2000. As far as Geospatial domain is concerned, big data concept finds a real purpose when it comes to developing forecasting applications which rely heavily on the past and if available the real-time data to estimate near accurate results using complex machine learning algorithms and the testing hypothesis of an event's occurrence using statistical inference.

Understanding The Big Data Ecosystem

As opposed to the general perception that Big Data and Business Intelligence are both one and the same, there remains a stark difference between them with the former adopting Inferential statistics while the latter adopting Descriptive statistics technique. The difference between them demands the understanding of the terms Sample and Population. Population is an en masse collection of an object of interest taking into account the varied characteristics that come along with them which might be finite or infinite in quantity whereas a Sample is considered to be a subset of Population which is definitely finite. In keeping with these definitions, Descriptive statistics tries to summarize the data that is observed in a Sample using measures of Central Tendency(Mean, Median and Mode) and Variability(Standard deviation, Variance) say for instance to evaluate the marks obtained by a group of 50 students. While Sample and Population are the same for descriptive statistics which tries to express data in a pictorial manner, it is not the case in inferential which estimates the value for a population based on a sample provided it mimics well enough the characteristics of the former, for instance estimating how the students of a University are based on a given sample of 50 Students' academic performances. Simply put, inferential statistics extrapolates the statistical parameters of a sample to estimate/predict the behavior of a population whereas Descriptive statistics summarize the characteristics of a taken Sample.

Traditional data models such as E.F Codd's path-breaking Relational model auger well for Business Intelligence where a hierarchical setup is followed due to manageable chunks of data and a relatively slower rate of change in them thereby following the motto "Design and Implement" which means a Schema for the data is designed prior to populate it into a database, popularly called the "Schema on Write" concept. This concept forms the core of Relational Database Management Systems (RDBMS) but the reason why it has not garnered much traction in Big Data workflows is it's disability to store a data prior to creating a schema for the same, since data tend to come from varied sources it is important for the Data Scientists to import and store them from which they might gain some new insights, adopting a concept that does not allow them to do so can surely hamper the value generated from it.This has led to the acceptance of a new concept named "Schema on reading" which allows one to store data in a database irrespective of whether they are in a structured or unstructured manner and an appropriate Schema can be designed for the data when it is read.



ETL is a tool that performs three database functions namely Extract, Transform and Load which describes Extracting the data from multiple sources, Transforming data into an appropriate format so as to suit the target database's specifications and finally Loading which means writing the data in the target database.As one can observe SQL databases have lost popularity in Big data environment since they always demand a fixed schema as opposed to NoSQL Databases that allow designing dynamic schema tailor-made according to the data's structure.
This ability of a Big Data ecosystem to allow data storage with least concern about its type or origin before designing a Schema for the same has given carte blanche to the Data analysts to import as per their will and that is the reason why it is high time the Geospatial domain experts start adopting them more sooner than later if they intend to keep pace with the growing needs of Customers.

Variants of Big Data Analytics

Generally, Big data ecosystems come in two variants namely Synchronous and Asynchronous analytics, it up to the concerned user's discretion on choosing which one as both have their equal share of pros and cons. Databases like NoSQL and Cassandra when coupled with necessary commodity hardware support synchronous analytics where data collected are processed then and there to get near real-time outputs, say for instance real-time monitoring of traffic flows in a metropolitan city that collects data from disparate sources like CCTV camera feeds, inputs from Police personnel, Road quality data etc. As with Asynchronous analytics, it does follow the principle of Collecting and Storing the data first before Analyzing thereby getting the name Batch processing as it tends to analyze the collected data batch by batch. As obvious as it gets. the main concern in batch processing is storage which is well covered in open source big data analytics software like Hadoop where a solid-state storage device named Hadoop Distributed File System(HDFS) is employed.

Road map ahead

Since data generating sources are increasing exponentially it is important to tap the benefits of them in enriching the end users with reliable real-time updates and valuable prognosis of impending events. In order to achieve this, incorporating big data concept in Geospatial domain is quintessential especially when it comes to Synchronous analytics, various powerful proprietary and open source GIS Servers are developed of late by keeping this in mind. ArcGIS's Geo Analytics Server is one among them which synchronously collects and processes data in real time that can help in issues like monitoring a fleet of vehicles with a configurable automated alert message being given out to a Vehicle's driver should he loiter off the route.

IV. CONCLUSION

As per the common word "No technology is foolproof ",Big data is no exception to that. Data security, data privacy and data discrimination are the three major areas of concern that can give sleepless nights to common people should they try to know whether their personal data are properly secured. Rampantcybercrimes, E-Commerce frauds are the consequences of poor data security procedures and dissemination of personal data without getting prior permissions. To contain this looming crisis. the European Union in 2018 framed the General Data Protection Regulation (GDPR) Policy that empowers its citizens to have full control of their personal data which many other Countries are

following suit. To cope with the abundance of data, protection policies are of paramount importance if we really want Big data ecosystems to change our lives.

REFERENCES

1. *A McAfee, E Brynjolfsson, TH Davenport – "Big Data:the management revolution"*
2. *S John Walker – "Big Data: "A revolution that will transform how we live, work and think"*
3. *A Labrinidis, HV Jagadish – "Challenges and opportunities with big data"*
4. *Intellipaat –Article:"7 Big Data Examples - Applications of Big Data in Real Life"*
5. *Laerd Statistics – Article: "Descriptive and Inferential Statistics"*
6. *DATAVERSITY – Technology Paper: "Data Modeling for Bg Data"*
7. *Thomas Henson – Article: "Schema on Read Vs Schema on Write Explained"*
8. *Geeks for Geeks – Article: "Difference between SQL and NoSQL"*
9. *Geeks for Geeks – Article: "Difference between Structured, Semi-Structured and Unstructured Data"*
10. *Bernard Marr, Forbes – Article:" 3 Massive Big Data Problems Everyone Should Know About"*
11. *Tech Target, Storage Magazine – "Big Data Storage and Analytics".*