International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I1P104

Volume 69 Issue 1, 18-22, January 2021

Original Article

Diagnostic, Descriptive, Predictive and Prescriptive Analytics with Geospatial Data

Prashant Tyagi

Scottsdale, Arizona, USA

Received Date: 17 November 2020
Revised Date: 02 January 2021
Accepted Date: 07 January 2021

Abstract - *This article discusses how companies can build a data lake foundation or a massively parallel processing data warehousing solution that they can leverage for addressing some of their ever-changing business climate needs through Diagnostic, Descriptive, Predictive, and Prescriptive Analytics. This article will discuss an overview of how to collate the data residing in silos and prepare the data for deeper structured analysis that will empower organizations to enhance the speed and quality of their decision-making process by converting data into some of the quick key actionable business insights. This article also discusses how certain Python Libraries on geospatial data can be leveraged to answer some of the most challenging questions faced by transportation, logistic and environmental service industries to address some of the critical issues such as Service Verification (missed customers), Route Compliance/adherence and real-time analytics for the estimated time of arrival at the customer's location.*

Keywords - *Predictive Analytics, Descriptive Analytics, Prescriptive Analytics, Diagnostic Analytics, Geospatial Data, Data Lake, Data Warehouse, Massive Parallel Processing, Hybrid Cloud, Data Access, Data Integration, Business Intelligence, Business Challenges, IoT data, Customer Service Verification*

I. INTRODUCTION

Data Analytics is pivotal to an Organization's ability to make quick decisions in an ever-changing business environment. Most of the companies that deal with Geospatial data have traditionally used route sheets generated from a green screen route editor software, static dashboards or some form of simple reports, spreadsheets to explain what is the problem statement that the company is facing (Descriptive Analytics) and the underlying cause of such issues (Diagnostic Analytics). With a solid data lake or a data warehousing kind of a storage solution having a steady flow of data coming from its many sources and tributaries, some very innovative ideas around Predictive Analytics(patterns and trends around the historical data) and Prescriptive Analytics (data-driven decisions that will help change the outcome) could also be addressed leading to greater agility in decision making by organizations and increased customer satisfaction index among its user base.

II. GEOSPATIAL DATA AND GIS

Today, most transport, logistics, courier services, waste management, utility companies, urban planning, military organizations, and the like deal with some other kinds of Geospatial Data. The word geospatial is used to indicate that data has a geographic component associated with it. This means that the records in a dataset have location information tied to them, such as geographic data in the form of coordinates like latitude and longitude of the customer's location, the address, city, or ZIP code [1]. A geographic information system, or a GIS, is an information technology system that allows for the data persistence, transformation or manipulation, analysis, and display of information with this geographic component associated with it, which is also called geospatial data. GIS allows people to connect the data they have about customers, places, or things with their corresponding physical locations and develop an even better understanding of the subject. When we can associate an object with its latitude and longitude coordinates, we can draw a relationship out of it, visualize it as to where the subject lies in the map and gain an even deeper understanding of the object than would be possible without its geographic component[2].

III. DATA INGESTION AND DATA STORAGE

Most of the transportation and logistics industries have some tablet or an IoT (Internet of Things) device which captures the GPS (Global Positioning System) coordinates of the customer locations or their business points of contact and sends those raw GPS pings/streaming data into a data lake foundation or the massive parallel processing data warehouse solution either in the hybrid cloud kind of a setup or a fully native cloud data lake foundation. This data can be stored in these data lakes set up in their raw object format in the form

of files arranged date wise, archived for future deeper structured analytics. Most organizations also have back-end relational data store on-premise, or on-cloud, or in a hybrid environment. The most common relational datastores which we come across are MS-SQL, SqlServer, Oracle, IBM-Db2, PostgreSQL, or Maria DB. The data that is residing in these databases comes from various traditional legacy systems like the CRM(Customer Resource Platforms), ERP(Enterprise Resource Management Platforms), SCM(Supply Chain Management Systems), QMS (Quality Management Systems), DMS(Dealer Management Systems), Call Center Databases and Sales and Marketing applications.

IV. DATA ACCESS AND TRANSFORMATION

Data Selection and Integration challenges abound in an era when data is becoming faster, more voluminous, more varied, and is residing across different silos. In addition, many users want greater agility and self-servicing capabilities for blending data that is coming from various disparate and coherent sources for use cases such as analyzing customer data, route data, operational data, and the like that is coming from multiple channels[3]. The amount of time taken to load new data into a data lake, data warehouse, or other target location sometimes could be very frustrating. For instance, a data scientist or data analyst may want the data to be loaded more in its raw format to clean and transform the data and prepare the data for deeper structured analysis depending on the use cases they are working on. While in other cases like for instance, the marketing managers, front-end personnel engaged with customers daily want new data loaded into a data warehouse or the database to gain a nearly real-time view of the customer activity across various channels[3]. Rather than raising an IT support ticket to get the dashboard refreshed or the new customer data loaded to the reports, the users prefer to have greater agility and self-servicing capability to be able to gain newer business insights of the data to empower them to take quick, and quality informed data-driven business decisions.

A. The emergence of Data Virtualization as a Data Integration tool

Enterprise information integration (EII) (first coined by Metamatrix), now known as Red Hat JBoss Data Virtualization, and federated database systems are terms used by some vendors to describe a core element of data virtualization: the capability to create relational Joins in a federated view[4]. The Data virtualization tool helps in combining the data coming from various disperse, dispense, coherent data sources like data warehouses, data marts, and/or data lakes and provides a single combined, holistic view and unified access of that data to the consuming applications so that the consuming applications will not have to worry about the underlying data sources in which this data is stored in. The consuming applications can access the data via the canonical views created on top of these virtualization layers. These views can then be renamed and reused as and when needed. Data Virtualization tool is not a replacement or a substitute to ETL (extract, transform, load) process but a valuable addition to the toolbox. The data virtualization platform enables developers to build and deliver data services to support various enterprise data activities and their internal and external consumers within a day, instead of one to two weeks that would take with the traditional ETL process. One must think hard to understand as to why data virtualization is in place. It provides the abstraction layer, freedom to move the logic and the accessed data to a single central point regardless of the business intelligence (BI) tool or the consumption methodology used. The availability of all the data via the virtualization layer at a central location provides a valuable capability to the professionals and citizens like Data Scientists and Data Analysts so that they can whip up some report in a very short period of time to obtain feedback or recommendation from the larger group and physicalize that later. Most virtualization tools in the marketplace have different virtual layers to optimize the performance and make the data readily available for the end-users. Layers provide a way to organize the content within a virtual database (VDB), which becomes increasingly important as more objects from varied sources are imported.
One of the most popular tools and a leader in the data virtualization space is Denedo. Denedo has different virtual database layers like the Connectivity Layer, Integration Layer, Business Entity Layer, Application Layer, and Web Services Layer, as indicated in Fig 1. Base views that mirror the source system and the clean views in which the column names are standardized, omitted, renamed intuitively need to be created in the connectivity layer. Joins, transformations, aggregations between the clean views and sources can be performed in the integration layer. The business Entity layer is the layer that has the canonical views exposed to business users. Frequently used reports, dashboards, applications, and audience-specific views could be built on top of the Business-Entity and Integration Layers sourcing data from the connectivity layer's clean views. Web Service Layer provides high concurrency real-time requirements offering both REST API and SOAP services for data access [5].

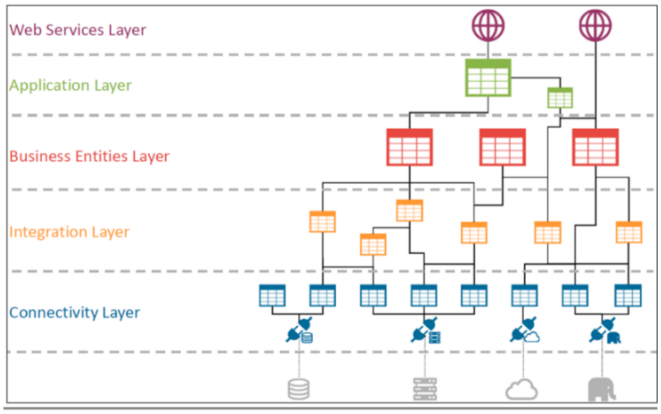

Fig. 1 Diagram illustrating the layer concept in the Denedo Data Virtualization tool [5]

V. COMMON TRANSPORTATION BASED INDUSTRY USE CASES

Most of the data virtualization software, including Denedo can get a spot in the Data Scientist/Data Analyst workflow as powerful data preparation and a governance tool for establishing a connection and importing data from this data virtualization software to Python IDE(Integrated Development Environment), such as Zeppelin, Jupyter lab, Jupyter Notebook, Spyder of Anaconda distribution. There are several Python libraries and packages that one can use to establish the connection between Python and Denedo.

A. Problem Statements or the Business Use Cases

Most logistic, transport, courier, and environmental service-based companies face the common business challenges as discussed below

a) Route Verification or Detecting Missed Customer Stops

This is one of the most common problems a courier services, waste management industry encounters in their daily operations. Not all companies have yet adopted a tablet-based IoT device installed on their trucks to give them a dynamic real-time view of their trucks' current location servicing the customers at that point in time. Most companies are yet in the phase of developing and designing a modern hybrid data architecture platform to verify the technology stack's capabilities to address some of the most common operation challenges. Pain points their business is currently facing. A missed customer stops, deliveries, and packages most often than not lead to customer dissatisfaction, leading to customer churn and revenue loss for companies. This problem is most common in transport, and logistic companies still have their drivers use traditional paper-based route sheets instead of a tablet installed in the trucks

b) Route Adherence or Route Compliance

Sometimes, due to operational challenges, the truck drivers may decide to service one customer first over another customer even though the latter one is not the next on the route sheet. Often, the route sheet which the driver carries is generated by using a network optimization software like those of the Route Editors, which considers parameters like the steep turns the trucks need to make on that route, assigned territories and boundaries, service time, sunrise/sunset times, interactive maps, road and construction work on that route, avoidance zones, turn restrictions (avoiding left turns) and the like. If the driver does not adhere to the route or there is a failure in route compliance, it can increase its operating costs, thereby decreasing its operating efficiency. For instance, UPS instructs their drivers always to take a right turn unless a left turn is unavoidable. By adopting this simple technique, the carrier saves millions of fuel each year and avoids emissions equivalent to over 20,000 passenger cars [6].

c) Estimated Time of Arrival at the Customer's Site or Location

It would help the customer get a notification message on their mobile device of the estimated time period of waiting or the number of stops remaining before they get serviced or before the truck arrives at their location. This would help prepare the customer for the arrival of the package or for them to get serviced. This kind of arrangement can increase the customer satisfaction index and reduce the company's operating costs as there will be fewer frustrating calls by the customer to the call centers.

d) Diagnostic and Descriptive Analytics catering to the above use cases

For understanding purposes, we will focus on the logistics and transportation industry who are often encountered with the above business challenges in their day to day operations. They need to gain deeper actionable business insights to make quality decisions about how many of those instances they observed in the past are out there and the reason for their occurrence. For instance, how many drivers have missed customers on their route on a day and why they are not being serviced on that day. In this case, the example of descriptive analytics would be missed pick-ups or unserved customers on that route, and the example of diagnostic analytics would be the reason or the underlying cause for missing that customer on the assigned route on that day.

e) Predictive and Prescriptive Analytics catering to the above use cases

Adding predictive analytics and AI (Artificial Intelligence) /ML (Machine Learning) capabilities that can run on bigger diverse sets have become a key priority for business analytics. Predictive analytics enables organizations to build models that capture trends and patterns around historical data and then create models to interpret data to understand future outcomes' relative likelihood [3]. Predictive analytics provides one with the raw material for making informed decisions about what is most likely to happen in the future. In contrast, prescriptive analytics provides one with data-backed decision options that can be weighed against one another.[7]. Prescriptive analytics provides recommendation actions that one needs to take to affect the outcomes differently. Prescriptive analysis often provides the finishing touch to the predictive analysis of any business [8]. Usually, we see prescriptive analysis in more data analytics matured organizations. In the business use case, which was cited above, the example of predictive analytics is the estimated arrival time at the customer's location. If, based on the route traffic conditions, we can predict the number of stops and evaluate the comparative time the driver would take to arrive at the customer's service point, that would illustrate the predictive analysis. If construction work is going on in that route or an unforeseen event like an accident, the driver can take an alternate route avoiding the traffic conditions on that route. That is the data

backed decision choice right there which the driver is making, affecting the outcome (delay in servicing the customers). This is an illustration of prescriptive analytics.

VI. DATA CONSUMPTION USING PYTHON

Data continues to grow rapidly in volume and diversity. Once ingested, most data must go through a data preparation, data curation cycle depending on the use case, including profiling, collection, integration, cleansing, transformation, and enrichment. Organizations can choose to move and centralize it in a massively parallel processing kind of a data warehouse or stream it raw into a data lake foundation kind of a solution; they can use a data virtualization platform to federate queries, sending them to the sources and viewing the results in a logical/abstract layer. Data can be managed on-premises, on various cloud and as-a-service platforms, or both. The array of options for data architecture and data management can be intimidating, not to mention the choices for how to prepare, catalog, transform and develop data pipelines for business analytics. However, having a spectrum of choices is always helpful when organizations must consider supporting many different types of users and workloads. This article will discuss one such approach as to how mathematical theorem, statistical technique, and machine learning algorithm can be applied on the business data to solve the above logistic and transportation industry use cases with geospatial data.

A. Data Curation and Integration involving Geospatial Data

Once the data has been identified for curation and integration like the employee data, customer data(latitude and longitudes of customer sites or their locations), planned route data, and the raw GPS pings (latitude and longitude coordinates) we can import this data into integrated development environment like Zeppelin Notebook or Jupyter Lab, Jupyter Notebook of Python Anaconda distribution. The individual data frames imported into Python IDE can then be cleaned, massaged, transformed, and prepared into a final data frame for deeper structured analysis. Importing the data, cleansing the data, taking care of missing values, taking care of categorical values, converting the text data into numerical data, feature scaling, standardizing, normalization and applying logarithmic scaling, and preparing the data for further analysis can be done using various Python libraries like but not limited to Pandas, Numpy, matplotlib, regular expressions, urllib, requests, to name a few.

B. Data Analysis using Haversine Distance Formula

Once the data is prepared for further deeper structured analysis, we can define the Haversine Distance formula as shown in Fig 2,

```
""" Defining the Haversine Distance Function for creating a Geo-Fence
    as the customer lat long co-ordinates as the centroid of the circle."""

def haversine(lon1, lat1, lon2, lat2):
    """
    Calculate the great circle distance between two points
    on the earth (specified in decimal degrees)
    """
    # convert decimal degrees to radians
    lon1, lat1, lon2, lat2 = map(radians, [lon1, lat1, lon2, lat2])

    # haversine formula
    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat/2)**2 + cos(lat1) * cos(lat2) * sin(dlon/2)**2
    c = 2 * asin(sqrt(a))
    r = 3956 # Radius of earth in miles. Use 6371 for kilometers
    return c * r
```

Fig. 2 Haversine Distance Formula Using Python

Haversine distance is the angular distance between two points on the surface of a sphere. The first distance of each point is assumed to be the latitude, while the second is the longitude. Both these distances will have to be in radians. In the Haversine formula, inputs are taken as GPS coordinates, and calculated distance is an approximate value [8]. We can have the customers who need to be serviced on a day (latitude and longitude coordinates) appended to a Python list. We can then also have the GPS pings (actual customer latitude and longitude coordinates) in another Python list. The customer latitude and longitude coordinates can be made the centroid of the Haversine Distance formula's imaginary circle. Each actual GPS ping could be compared to determine if the actual customer has been serviced on that day or not. The two python lists can then be compared to determine if the driver has serviced all the customers in that route or not. This will cater to the use case of 'service verification or route verification.' For the second use case of 'route adherence or route compliance,' the customer python list can be sequenced as per the planned route to be adopted by the driver. The actual coordinates of the route taken by the driver also can be ordered, and both these python lists can be compared to determine if the driver has indeed followed the same sequence of the customer route or not. For the third use of 'estimated time of arrival,' if we have a customer calling the customer care center to know their package's status, then the call center representative via the customer's address can get the coordinates, and the employee or the truck driver who is servicing that route. With this information, the call center representative can then determine which was the last customer the truck driver in that route serviced and then give an approximate estimate of the number of stops and an evaluated time range as to when this particular customer(who is on the call) would be serviced.

VII. COMMON PITFALLS TO AVOID

While determining if the customer has been serviced or not by having the customer latitude and longitude coordinates as the centroid of the largest distance circle, multiple GPS pings are recorded by the truck's GPS device if the truck, for instance, is standing at the red light. There could also be a scenario where the driver has stopped for a break near the customer's location. One needs to be cautious about such false positives as this may cause an error in interpreting the results of the data driven analytics by skewing the results and the underlying decisions that need to be taken based on this data analysis may further be impacted.

VIII. CONCLUSION

AI-driven automation evolves, it will improve self-service capabilities, so data scientists, data analysts, and professionals can do more data ingestion, preparation, and monitoring independently. Today, more organizations want to use AI/ML to get value from analytics sooner. They believe that this kind of automation could enable less technical citizens and enable data scientists and data analysts to do more with analytics and data[10]. Dashboards and Key Performance Indicators (KPI's) integrated with AI-driven insights would be more important as the organizations want to see AI make a difference in their daily operational decision-making, including individual's ability to use metrics to meet corporate business objectives [9]. As AI-driven automation becomes increasingly embedded in front-end solutions, data integration middleware, and data curation platforms, developing the correct and accurate data models is becoming critical to giving these systems the right data structure. Hence, data integrity, consistency, completeness, and quality are not issues when users work with their reports, dashboards, and analytics.

REFERENCE

[1] Caitlin Dempse|GIS Learning. What is the Difference Between GIS and Geospatial? (2014) https://www.gislounge.com/difference-gis-geospatial.

[2] The University of Arizona University Libraries GIS & Geospatial Data|https//libguides.library.arizona.edu/GIS/about-gis

[3] David Stodder Evolving from Traditional Business Intelligence to Modern Business Analytics (2020) https://tdwi.org/research/2020/09/bi-all-best-practices-report-evolving-from-traditional-bi-to-modern-business-analytics.aspx?tc=page0

[4] Data Virtualization History of Data Virtualization https://en.wikipedia.org/wiki/Data_virtualization#History

[5] Denedo Community Knowledge Base https://community.denodo.com/kb/Northbound%20Connections

[6] Jacopo Prosco, CNN| Why UPS trucks(almost) never left turn left? |https://www.cnn.com/2017/02/16/world/ups-trucks-no-left-turns/index.html#:~:text=UPS%20trucks%20almost%20never%20take,to%20over%2020%2C000%20passenger%20cars.

[7] Proponet Prescriptive Analytics Vs. Predictive Analytics: What is the difference? https://www.proponent.com/predictive-analytics-vs-prescriptive-analytics

[8] Firuze Koyunco Sept 7th, Calculating the Haversine Distance Between Two Geo-locations with Python(2020). https://codeburst.io/calculate-haversine-distance-between-two-geo-locations-with-python-439186315f1b

[9] Stratecast Analysis by Jeff Cotrupe, MBA, Mike Jude Ph.D. Comparing Total Cost of Ownership (TCO) for Business Intelligence Solutions: How to calculate Costs for Competitive Options (2018).

[10] International Journals, Engineering Research, and Technology, Science and Humanities (internationaljournalssrg.org) http://www.internationaljournalssrg.org/ssrg-journals.html.