ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/rameshbabu-lakshmanasamy/329

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 2 & Iss: 2

Research Article

Geospatial Data Processing in Redshift and BigQuery

Rameshbabu Lakshmanasamy*

Rameshbabu Lakshmanasamy, Senior Data Engineer, Jewelers Mutual Group, USA

Citation: Lakshmanasamy R. Geospatial Data Processing in Redshift and BigQuery. *J Artif Intell Mach Learn & Data Sci 2024*, 2(2), 1452-1455. DOI: doi.org/10.51219/JAIMLD/rameshbabu-lakshmanasamy/329

Received: 03 April, 2024; **Accepted:** 28 April, 2024; **Published:** 30 April, 2024

***Corresponding author:** Rameshbabu Lakshmanasamy, Senior Data Engineer, Jewelers Mutual Group, USA

ABSTRACT

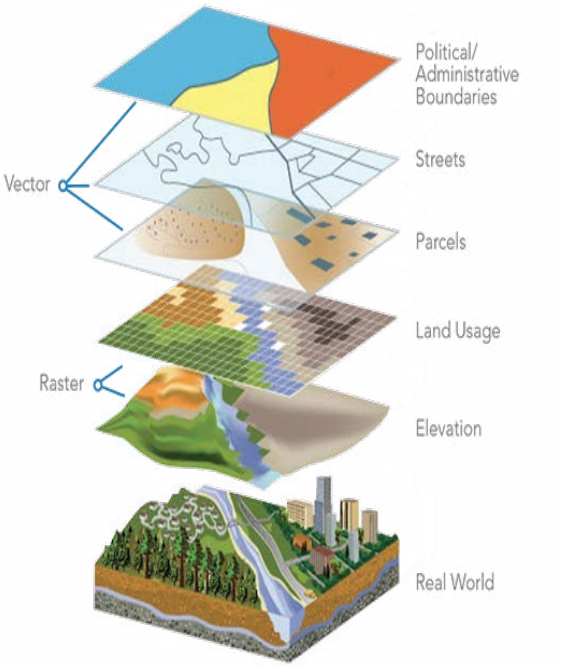
Geospatial data is information that identifies geographical locations and attributes of objects on the Earth through coordinates, addresses or areas. This comprises the natural environment and structures built by man; it can be raster data such as image satellite data or vector data types like points, lines and polygons. Geospatial data is inevitable in almost every industry, right from urban planning and development to transportation, logistics and even environmental studies (Alam et al., 2022)1. Various activities such as routing, planning of disaster situations and land use analysis are some of the functions that it fosters. Thus, by integrating other information (for example, regarding weather or economic trends), geospatial data enables such features as real-time monitoring and decision-making.

Keywords: Amazon Redshift, Google BigQuery, Geospatial Data, Performance

1. Introduction

The amount of data has grown over the years and especially IoT and satellite systems are producing even more geospatial data requiring scalable solutions. Other advanced platforms like Amazon Redshift and Google Big Query are essential when dealing with big data geospatial applications due to their capability to handle big data, scalability and provision of geospatial functions to their users. Both Redshift and BigQuery provide the functionality for spatial query, integration with PostGIS for Redshift and use of the parallel processing for the GEOGRAPHY data type for BigQuery with no infrastructure required (Shastry & Manjunatha, 2023)[2].

This research will focus on comparing Redshift and BigQuery for typical geographic operations such as distance and ranging and spatial join. It will assess their functionality, suitability for large datasets and gains so businesses can identify the optimal structure for proficiently dealing with geographic data (Armenatzoglou et al., 2022)[3].



3. Overview of Amazon Redshift and Google BigQuery for Geospatial Data

3.1. Amazon Redshift: Amazon Redshift is a cloud-based petabyte-scale data warehousing solution that allows customers to run large, complex analytical queries. Its architecture is based on a massively parallel processing (MPP) model, so queries can be processed on multiple nodes concurrently (Shastry & Manjunatha, 2023)[2]. This feature enhances query speed, making Redshift a perfect solution for organizations involved in large-scale analytics such as geographic data computation.

Available through PostGIS, an open-source extension of PostgreSQL, Redshift offers geospatial data that can be queried. PostGIS allows users to call such operations as joins, distance and intersections within SQL statements. Also, Redshift Spectrum complements these features by enabling users to scan Amazon S3 without having to load it into the Redshift cluster. Spectrum currently supports CSV, Parquet and ORC files to accommodate extensive data while utilizing only a little of the cluster's resources (Deepika et al. 2024).

Second, Redshift has more advantages in dealing with geospatial data due to the columnar storage format and data compression. These features decrease the number of times that I/O operations are accessed, which increases the speed of the query. The location of data is distributed using distribution keys, making it possible to perform interleaved querying in parallel with other nodes and speeding up the big GIS queries. Redshift's strong compatibility with other AWS services like S3, EC2 and Athena makes it a perfect environment for firms that are already using AMAZON services - an ideal platform for geospatial analytics (Yang, 2016)[4].

3.2. Google BigQuery: Google BigQuery is a cloud-based, fully managed & scalable data warehousing solution with exceptional speeds for SQL query execution on Google's platform. One of the most notable approaches BigQuery offers for geospatial data processing is the integrated GEOGRAPHY data type that enables using geographic objects (points, lines and polygons) in SQL statements. This data type is capable of both planar (comparing flat surfaces) and spherical (comparing the curves of the Earth) comparisons, thus is excellent for accurate measurement of distance as well as the joining of geographical data sets.

BigQuery also supports hundreds of geospatial functions such as ST_DISTANCE, ST_INTERSECTS, ST_WITHIN and ST_UNION, which enables users to conduct geospatial queries as part of Structured Query Language. The serverless computing feature frees users from having to manage servers and uses query scaling to efficiently handle big, fluctuating workloads without the need for capacity forecasting (Kritikos et al., 2015)[5]. Spatial joins and most other geospatial operations become feasible on BigQuery due to the distributed query execution feature, even in real-time and for extremely large-scale data.

The correlation of BigQuery with other Google Cloud Solutions like Google Map Platform, Earth Engine and Google Cloud Storage strengthens the system's geographical processing capability. It is ideal for organizations that need both geospatial analysis and geographic data from Google. Moreover, its pricing strategy is very viable since users only pay when running queries; this makes it suitable for companies with variable demand.

4. Geospatial Query Performance

Spatial queries often involve distance estimates, areas of interest, spatial joins and polygons of objects regarding Geographic Information Systems. This kind of query is ubiquitous in different fields like supply chain management, city planning and ecological system control. For instance, distance computation involves the extent of points and spatial overlay entails the comparison of geographical conditions of two sets of data, for instance, which area belongs to the flood plain. These are some of the fundamental issues on which the geospatial analytical procedures depend (Shastry & Manjunatha, 2023)[2].

In Amazon Redshift, geospatial queries are loaded from the PostGIS extension, which allows many more spatial functions to be incorporated into the SQL of Redshift. Due to the optimizations done in Redshift, such as storage of data in columnar format and data compression, precalculating values used in query maximize the throughput by minimizing the I/O operations and scanning of data to be done. Redshift also allows the creation of distribution keys for geospatial data and consequently, it improves the spatial join because less data has to be moved between nodes (Mozumder & Karthikeya, 2023)[6]. Nevertheless, experience shows that when dealing with data up to 100GB, Redshift is highly efficient. Still, it can be slow with larger sets or excessively complex polygon operations that might require an improperly tuned cluster. This is because it utilizes the cluster resources to get the benefits and, as a result, profitability depends on the degree to which the system is optimized for that load.

Google BigQuery uses GEOGRAPHY data type for geospatial analysis, which supports both planar and spherical mathematics. BigQuery is a serverless tool, which means that all the computing resources it needs will be allocated automatically in terms of dataset size and complexity of the query. This flexibility makes a massive difference in BigQuery to work with extensive geospatial data and to perform other complex operations such as a spatial join or distance query (Shastry & Manjunatha, 2023)[2]. Geospatial data processing in BigQuery is carried out using distributed query execution, which is helpful for fast processing and decision-making, for example, in the field of transportation logistic crisis management.

In terms of geospatial query, it is evident that BigQuery is always ahead of Redshift, particularly for the giant tables. Something that BigQuery can do naturally is maintain query response times as they grow in size, something that Redshift can only have issues with in terms of performance if more attention is paid to it. Also, the BigQuery billing model is different and here, it is cheaper for organizations that use processing power for short times per day, while Redshift can be costly if the organization and its size need to manage resources correctly. In general, the indicators of scalability and operations simplicity indicate that BigQuery is a better solution for those geospatial jobs that are large and potentially dynamic.

5. Supported Geospatial Operations

Introducing geospatial functionality to an Amazon Redshift database with assistance from the PostGIS extension enables the use of a myriad of sophisticated geospatial operations. Thus, PostGIS offers Redshift the means to deal with geographic objects or conduct spatial operations such as operations on spatial relationships, distance or polygons. These include `ST_INTERSECTS` used to test for the intersection between two

objects, `ST_DISTANCE` for ascertaining the distance between two objects and `ST_BUFFER` for creating buffer regions around some geographical entity. Spatial operations such as `ST_UNION' and `ST_INTERSECTION' will allow Redshift to carry out geographic boundary manipulation used in land use analysis (Mozumder & Karthikeya, 2023). Redshift also has the capability of spatial clustering via `ST_CLUSTERDBSCAN,' a function that clusters geographic points by density.

With GEOGRAPHY data type support directly in BigQuery and massive SQL support all around geospatial capabilities, BigQuery is equipped with a comprehensive geospatial platform embedded within the SQL system. Standard methods such as 'ST_DISTANCE' for distance and 'ST_INTERSECTS' for operations like intersection and touching all enable 2D spherical operations; this is perfect for geographical operations regarding the Earth. Like Redshift, the service offers mathematical operations such as `ST_UNION` and `ST_CENTROID` performed on geographic things. That being said, BigQuery outperforms other solutions when it comes to immense spatial aggregations, so users can easily group the data according to the regions.

Both platforms contain all the essential Geographic Information System (GIS) operations that include distance, join and polygon operations. Still, their distinctions are as follows: While there is PostGIS integration available at Redshift, there are more advanced features like 3D geometries and spatial clustering to be performed for the spatial analysis cases (Mozumder & Karthikeya, 2023)[5]. However, BigQuery is more about scalability and ease of use with the serverless that distributes the query nodes and increases the computing capacity in real time or dynamically to cope with large data sets that are best suited to real-time or dynamic analyses.

Yet, BigQuery could be better: it does not support 3D geometries and does not include clustering functions; for the most complex operations, the user needs to develop their algorithms from scratch. Redshift is better designed for more complicated operations. Still, in extensive data analysis, resources need to be manually managed and optimized for optimization, which, in this case, makes the system more complex compared to BigQuery's serverless structure. Finally, where post-GIS is concerned, Redshift has an edge in many complex GIS operations, whereas where real-time analytical processing is a necessity, Big Query is the better option (Yang, 2016)[4]. Depending on the need for geospatial operations and the infrastructures of the users, either of the two is chosen.

6. Use Cases for Geospatial Analytics

6.1. Amazon Redshift

Use Case 1: Logistics and Supply Chain Optimization with Route Planning.

The application of geospatial intelligence is critical in logistics, where transportation networks include delivery services and freight carriers. PostGIS functionality is integrated with Amazon Redshift and it helps companies calculate the delivery routes that consume minimal fuel and minimum time for delivery. With function like `ST_DISTANCE,` one can find the shortest distance between delivery points and serve the customers efficiently and using the `ST_WITHIN ` function to find out the points that are within the range of a particular radius from warehouses, customers can be served more effectively. In the same vein, Redshifts' `ST_CLUSTERDBSCAN` makes it possible for spatial clustering and hence, logistics companies can be able to group nearby delivery points to reduce operational costs, making it more efficient.

Use Case 2: Real Estate Analysis Involving Geographic Boundaries.

In real estate, geographic information provides firms with property values and features such as zoning regulations whereby firms can see the distances surrounding amenities or environmental hazards. ST_INTERSECTION and ST_WITHIN of Redshift's polygon operations help developers determine the extent to which given properties fall within specific zoning districts or floodplains. It can enable firms to be in a position to operate without violating legal requirements, besides helping in the right decisions on places to invest. Therefore, the information culled from Redshift enables firms in the real estate industry to make better decisions in regard to property transactions and utilize development strategies such as geographic entity acquisition based on parks, schools and any other tangible features.

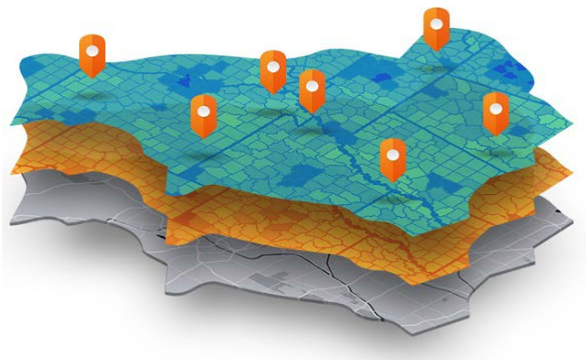
6.2. Google BigQuery

Use Case 1: Urban Planning Using City Infrastructure Data.

BigQuery fits well for tasks like analyzing geospatial data for large-scale city planning for infrastructure projects in Urban Planning. `ST_DISTANCE` can be employed by city planners to evaluate the accessibility levels of public transport since whenever they need to calculate distances between residents' places and certain stations/ stops, they can easily do so by using ST_DISTANCE. Since BigQuery can handle gigantic data sets without any additional planning, planners are able to layer different levels of geographical information, such as road maps, growth and zoning, over their analyses. It helps planners to make decisions about when to construct roads, extend railways, subways and other public transport and at which place to invest the resources of the city appropriately.

Use Case 2: Environmental Data Analysis, Tracking Deforestation Trends.

There is evidence that shows that BigQuery is efficient in monitoring the environment in general and deforestation in particular. Owners of research and environmental organizations can use satellite imagery and geospatial data to assess forest loss by year. That is, functions like `ST_INTERSECTS` can be used to compare historical and current forest boundaries and hence make extensive scale analysis of deforestation possible in BigQuery. This payment strategy of Right Media makes them useful for occasional caregiving projects, such as the effect of climate change on the forest ecosystem, thus beneficial for long-term caregiving projects.



7. Conclusion

Both Amazon Redshift and Google BigQuery excel in different aspects of geospatial analytics. Amazon Redshift is suitable for industries that require sophisticated geospatial analysis and an additional level of infrastructure configuration management, such as logistics and real estate. However, it is manual in scaling and maintaining, so it cannot be utilized with most massive data or real-time geospatial analyses.

Google BigQuery is better suited for large-scale geospatial queries as it is more cost-effective in terms of and is less suited for real-time or intermittent use as in urban planning and environmental monitoring. However, it has a relatively limited capability to handle spatial analysis and specific functions such as spatial clustering and 3D data and the frequent use of the system can also be rather costly.

References

1. https://dl.acm.org/doi/abs/10.1145/3507904
2. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003321149-7/intelligent-analytics-big-data-cloud-aditya-shastry-manjunatha
3. https://dl.acm.org/doi/abs/10.1145/3514221.3526045
4. https://scholarworks.uno.edu/td/2284/
5. https://link.springer.com/chapter/10.1007/978-3-662-46703-9_3
6. https://link.springer.com/chapter/10.1007/978-3-031-14096-9_4