A Seminar Report on

GeoFlink – An Efficient and Scalable Spatial Data Stream Management System

Submitted to the

Department of Computer Applications

in partial fulfilment of requirements

for the award of

Masters in Computer Applications

Degree from

APJ Abdul Kalam Technological University, Thiruvananthapuram

Under the Guidance of

Ms. Remya Anand

By

Harisankar P

(Reg no: SGI18MCA-I035)

DEPARTMENT OF COMPUTER APPLICATIONS

SNGIST GROUP OF INSTITUTIONS

NORTH PARAVUR

2018-2023

Experience Learning

SREE NARAYANA GURU
INSTITUTE OF
SCIENCE AND TECHNOLOGY

Date: 15-10-2022

CERTIFICATE

This is to certify that the seminar report titled "Geoflink-An Efficient and Scalable Spatial Data Stream Management System", submitted by Harisankar P (Reg. No: SGI18MCA-I035) in partial fulfilment of the requirement of Integrated MCA of Sree Narayana Guru Institute of Science and Technology affiliated to APJ Abdul Kalam Technological University during 2018-2023, is a record of bonafide work done by him under my supervision and guidance.

Ms. Remya Anand
(Seminar Guide)
Assistant Professor
MCA Department
SNGIST

Dr. Kavitha C.R
(Seminar Coordinator)
Professor & Head
MCA Department
SNGIST

Dr. Kavitha C.R
Professor & Head
MCA Department
SNGIST

DECLARATION

I Harisankar P, hereby declare that the seminar titled “Geoflink-An Efficient and Scalable Spatial Data Stream Management System” is a bonafide record of independent work carried out by me under the guidance of Ms. Remya Anand, Asst. Professor, Department of Computer Applications, in partial fulfilment of the requirement of Integrated MCA of APJ Abdul Kalam Technological University. I also declare that this report has not been previously submitted either in partial or in full for the award of any other degree/diploma in this institution or any other institutions/university.

Place: North Paravur
Date: 15-10-2022 **HARISANKAR P**

Abstract

Spatial data is any type of data that directly or indirectly references a specific geographical area or location. It is also called geospatial data or geographic information. It can be used to represent a physical object in a geographic coordinate system. We are currently living in a world where spatial data are increasing with the use of GPS-enabled devices. As such spatial data has a wide range of applications in business, government, and NGOs. Spatial data are generated as a high-volume continuous data stream like mobile location, vehicle GPS etc. These are huge streams of data and thus require highly scalable systems. Apache Spark, Apache Flink, and Apache samza are among one of the spatial data stream processing platforms, while they can stream spatial data they lack spatial objects, indexes query etc. Apart from these, there are other scalable spatial data stream processors like GeoSpark, and Spatial Hadoop, while they lack support for streaming workloads and handles only static and batch data, GeoFlink extends Apache flink to support spatial objects, indexes and continuous queries over spatial data streams. GeoFlink introduced a grid-based index to support efficient spatial query processing and effective data distribution across distributed cluster nodes. It is proven through a detailed study that GeoFlink achieves significantly more query throughput than existing streaming platforms.

Acknowledgement

In the name of **Almighty,** I express my sincere thanks to him keeping me fit for successful completion of the seminar.

I want to thank **Dr. Sagini Thomas Mathai, Principal,** SNGIST Group of Institutions for her kind support in all respect during our study.

I take this opportunity to express my deepest sense of gratitude and sincere thanks to everyone who helped me to complete this work successfully. I express my sincere thanks to **Dr. Kavitha C.R.**, Seminar Coordinator and Head of Department, Department of Computer Applications, N Paravur for providing me with all the necessary facilities and support.

I would like to place on record my sincere gratitude to my seminar guide **Ms. Remya Anand**, Assistant Professor, Department of Computer Applications, N Paravur for the guidance and mentorship throughout the course.

Finally, I thank my family, and friends who contributed to the successful fulfilment of this seminar work.

Harisankar P

Contents

Abstract i

Acknowledgement...ii

Chapter 1 1

Introduction...1

Chapter 2 2

Literature Review 2

2.1 SPATIAL DATA...2

2.1.1 Spatial Data 2

2.1.1.1 Geometric data 2

2.1.1.2 Geographic data...3

2.1.2 Spatial Data Analysis...4

2.1.3 Vectors and Rasters 4

2.1.4 Spatial Data Science 6

2.1.4.1 Importance of spatial data & spatial data science...................................6

2.1.5 Spatial Data Processing 7

2.1.5.1 Components of GIS 8

2.1.5.2 GIS Software 8

2.2 Apache Flink 9

2.2.1 Architecture of Flink.. 10

2.3 Geoflink 11

2.3.1 r-Neighbours..12

2.3.2 Spatial Stream Indexing...13

2.3.3 Geoflink Grid Index..13

2.3.4 Spatial Continuous Queries 14

2.3.5 Geoflink Architecture 15

Chapter 3 17

Analysis 17

Chapter 4 18

Conclusion 18

References..19

List of Figures

Figure 2.1.1.2.1: Geometric Data 3

Figure 2.1.1.2.1: Geographic Data 3

Figure 2.1.3.1: Vector data representation 5

Figure 2.1.3.2: Raster data representation 5

Figure 2.1.5.1: Spatial Data Objects 7

Figure 2.1.5.2.1: Google Earth 8

Figure 2.1.5.2.2: Esri ArcGIS 8

Figure 2.1.5.2.3: BatchGeo 8

Figure 2.2.1: Apache Flink 9

Figure 2.2.1.1: Architecture of Flink 10

Figure 2.3.1: Spatial Data Types supported by Geoflink 11

Figure 2.3.1.1: r-neighbours of a spatial object 12

Figure 2.3.3.1: Geoflink Grid Index 13

Figure 2.3.5.1: Geoflink Architecture 16

Chapter 1

Introduction

We live in an era with increasing use of GPS enabled devices, and hence spatial data is omnipresent. Many applications require real time processing of spatial data. These are done in order for the applications to perform efficient route guidance during patients tracking, disaster evacuation and so on that require real time processing of spatial data. For instance, current systems for example PostGIS and QGIS are not scalable enough to handle such huge data and throughput requirements. However scalable platform which includes Apache spark streaming, Apache flink Apache samza does not support spatial data processing as they lack spatial data objects, indexes and queries that are essential for spatial data processing. However, there are applications to handle large scale spatial data for example Hadoop GIS, Spatial Hadoop, GeoSpark, etc. In order to fill this gap Apache introduced a new extension for Apache flink called Geoflink that support real time query processing on spatial indexes, objects and queries.

Chapter 2

Literature Review

2.1 SPATIAL DATA

2.1.1 Spatial Data

Any type of data directly or indirectly referencing to a specific geographical area or location is called spatial data. These are also called geospatial or geographic data. They numerically represent a physical object in a geographic coordinate system. However, spatial data is not only a spatial component of a map, it contains location specific data which goes beyond just coordinates but also the terrain information and many other factors that makes up a geospatial location. These data that is not spatial but also relevant for processing or the data that exists alongside spatial data is called the attribute of the spatial data. Users can save spatial data in many formats as it contains more than just spatial data. Analysis of these data provides a better view of how each attributes affect the specified environment.[3]

While there are several spatial data types, the 2 key types of spatial data are geometric data and geographic data.

2.1.1.1 Geometric data

Geometric data is a spatial data type that is mapped on a two-dimensional flat surface like floor plans. Google maps is an example for application that user geometric data to provide accurate routes for travels.[3]

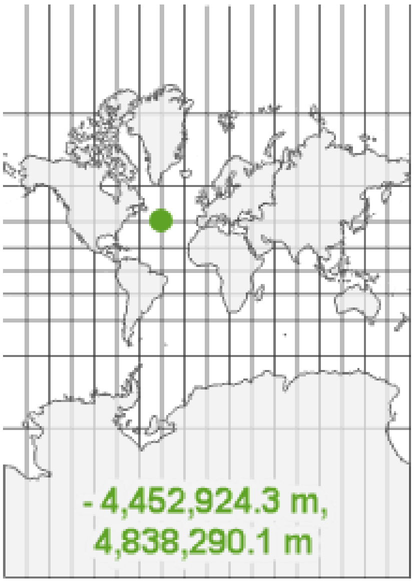

Figure 2.1.1.2.1: Geometric Data

2.1.1.2 Geographic data

Geographic data is a spatial data that is mapped around a sphere for example planet earth. Geographic data makes use of longitude and latitude relationship to map a specific object or location. A good example for geographic data is GPS (Global Positioning System).[3]

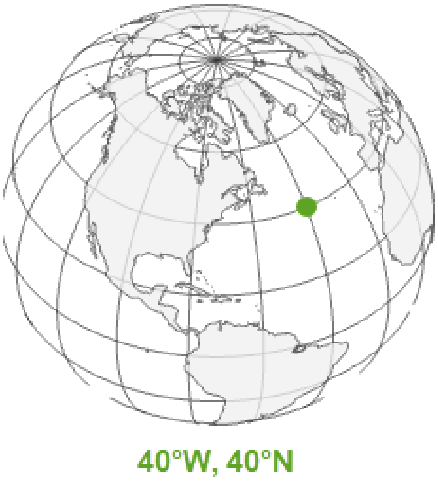

Figure 2.1.1.2.1: Geographic Data

2.1.2 Spatial Data Analysis

Geospatial analysis is the gathering, display and manipulation of images, GPS or any location data described explicitly in terms of geographic coordinates and implicitly in terms of address or postal code etc. An analysis is said to be spatial if and only if results are location based that is in other words location matters. The data that are subjected to spatial data analysis must record the locations of phenomena within some space, and very often that is the space of the Earth's surface and near-surface, in other words the geographic domain. The two important aspects of geospatial analysis are Georeferencing and geocoding.[3]

While both are used to fit data into the real-world using coordinates, however georeferencing uses vectors and rasters whereas geocoding uses address and location depicters.[3]

2.1.3 Vectors and Rasters

Vectors are graphical representation of real world. That is vector represents information in an x-y coordinate plane using points, lines and polygons. Vector data makes use of these points, lines and polygons to identify locations on the earth. Vector data is extremely useful for storing and representing data that has discrete boundaries, such as borders or building footprints, streets and other transport links, and location points.[3]

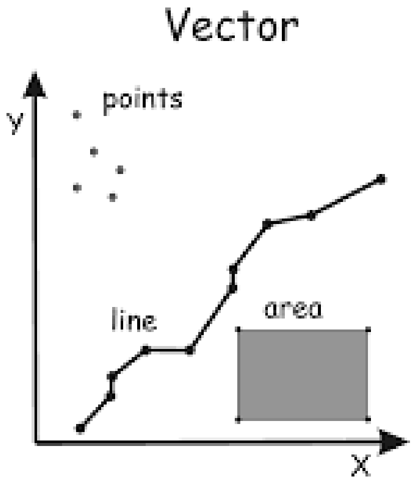

Figure 2.1.3.1: Vector data representation

Rasters represents information in a pixel grid. That is raster data provides a representation of the world as a surface divided up into a regular grid array, or cells. Raster uses a series of cells to represent locations on earth.[3]

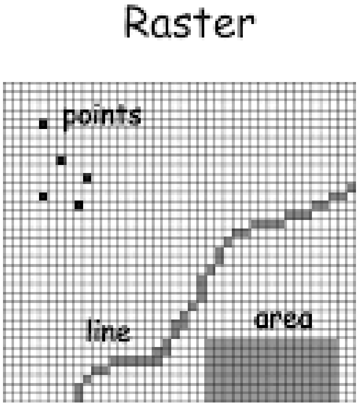

Figure 2.1.3.2: Raster data representation

2.1.4 Spatial Data Science

Spatial data science is a subset of data science that focuses on special characteristics of spatial data using modelling to know where an event happened and why that event happened there. This goes beyond just the scope of finding an event taking place in a location to the reason why that event is happening in that location. As such, spatial data science is a growing field of study and many industries are investing spatial data science as it proves to be efficient for industries to make profit by analysing the factors that would affect their products in a particular location and making changes as to increase the profit of the company.[5]

2.1.4.1 Importance of spatial data & spatial data science

Web applications use spatial data to provide contextualized results to their users, allowing for more accurate and personalized suggestions to be made. Some common examples include finding the nearest gas station, auto-selecting your nearest store, and providing news and events that are local to you. [3]

Mobile ordering apps can dynamically track you to ensure your order is started and finished at the optimal time. [3]

During the pandemic authorities used the spatial data of patients (the routes they travelled and the relationship between where they lived).[3]

2.1.5 Spatial Data Processing

The most common way that spatial data is processed and analysed is using a GIS, or, geographic information system.[5]

A system for capturing, storing, checking, integrating, manipulating, analysing and displaying data which are spatially referenced to the earth.[5]

GIS technology was developed from Digital cartography, CAD and Database Management Systems (DBMS).[5]

Data structures used by GIS are Point, LineString and Polygon.

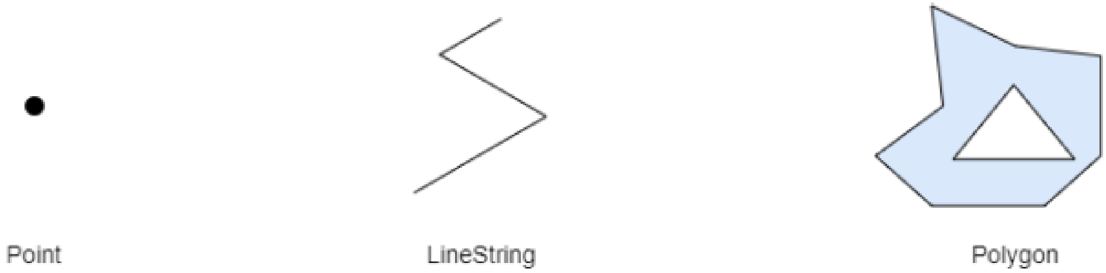

Figure 2.1.5.1: Spatial Data Objects

2.1.5.1 Components of GIS

A working GIS integrates 5 key components namely hardware, software, data, people and methods. Hardware is the computer on which a GIS operates. GIS software provides the functions and tools needed to store, analyze, and display geographic information. Geographic data and related tabular data can be collected in-house or purchased from a commercial data provider. GIS technology is of limited value without the people who manage the system and develop plans for applying it to real world problems. A successful GIS operates according to a well-designed plan and business rules, which are the models and operating practices unique to each organization. [5]

2.1.5.2 GIS Software

Figure 2.1.5.2.1: Google Earth

Figure 2.1.5.2.2: Esri ArcGIS

Figure 2.1.5.2.3: BatchGeo

These are some commonly used GIS software's. Although Esri ArcGIS is the no.1 software people find it very hard to use and most of the people prefer google maps or google earth. For navigation and other purposes. [5]

2.2 Apache Flink

Apache Flink is an open-source, unified stream-processing and batch-processing framework developed by the Apache Software Foundation. Flink is a distributed stream data flow engine capable of executing tasks in a data-parallel and pipelined manner. It can perform stateful computations over bounded and unbounded data stream. Flink has been designed to run in all common cluster environments perform computations at in-memory speed and at any scale. Flink provides a high-throughput, low-latency streaming engine as well as support for event-time processing and state management. Flink applications are fault-tolerant in the event of machine failure and support exactly-once semantics. Programs can be written in Java, Scala, Python and SQL and are automatically compiled and optimized into dataflow programs that are executed in a cluster or cloud environment. Flink does not provide its own data-storage system, but provides data-source and sink connectors to systems such as Amazon Kinesis, Apache Kafka, HDFS, Apache Cassandra, and Elasticsearch.[4]

Figure 2.2.1: Apache Flink

2.2.1 Architecture of Flink

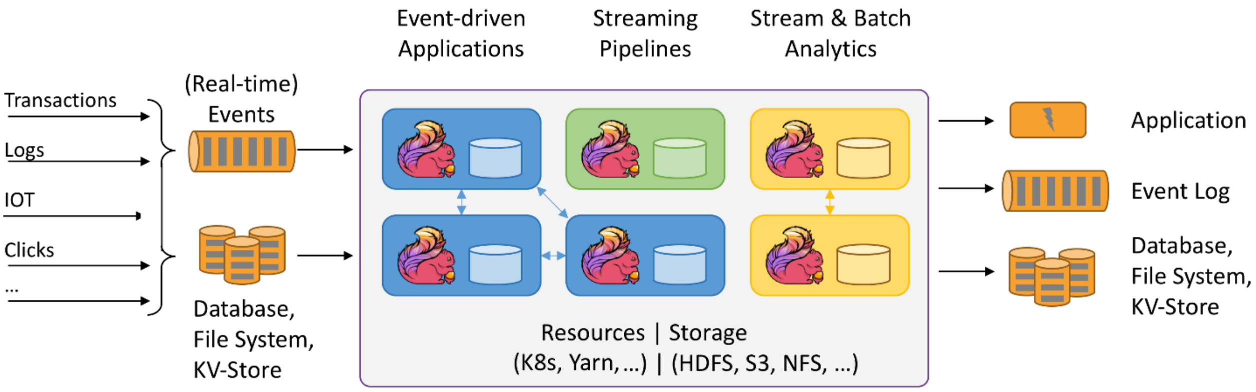

Figure 2.2.1.1: Architecture of Flink

Apache flink is a real time data processing system as such it takes both real-time events and input from database and processes it in real-time giving us a continuous stream of output of the processed input. An event-driven application is a stateful application that ingest events from one or more event streams and reacts to incoming events by triggering computations, state updates, or external actions. With a sophisticated stream processing engine, analytics can also be performed in a real-time fashion. Instead of reading finite data sets, streaming queries or applications ingest real-time event streams and continuously produce and update results as events are consumed. Data pipelines serve a similar purpose as ETL jobs. They transform and enrich data and can move it from one storage system to another. However, they operate in a continuous streaming mode instead of being periodically triggered. The processed output is then sent to database or applications using variety of sinks supported by Apache flink.[4]

2.3 Geoflink

Geoflink is an efficient and scalable spatial data stream management system developed to handle real-time spatial stream. It was introduced to overcome the problems that arise when using other streaming platforms like Apache spark streaming, Apache Kafka etc, that supports stream processing while does not support spatial objects, queries and spatial kNN which are essential for spatial data processing. It overcomes disadvantages posed by using spatial data processing software's like Hadoop GIS, Spatial Hadoop which cannot handle real-time spatial streams. To fill this gap Geoflink was introduced which extends Apache flink which can handle real-time data processing while Geoflink handles spatial data processing.[1][2]

Geoflink support spatial objects, indexes and continuous queries over spatial data streams. The 6 spatial data types supported by GeoFlink in 2D space are point, LineString, polygon, MultiPolygon, MultiLineString, MultiPolygon.[1][2]

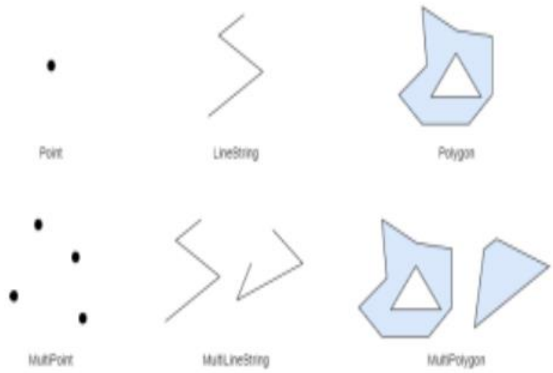

Figure 2.3.1: Spatial Data Types supported by Geoflink

2.3.1 r-Neighbours

r- neighbours computation is an important concept used in Geoflink. All of the GeoFlink's spatial queries require neighbourhood computation.in r-neighbours we define r-neighbor region and r-neighbors(ψ) given a spatial object ψ and distance r. This region which encloses the object is called r-neighbor region and all the objects that lies within or overlaps the r distance of ψ is r-neighbors(ψ). This is done in order to find what all objects are there besides the object ψ in that r region specified.[1][2]

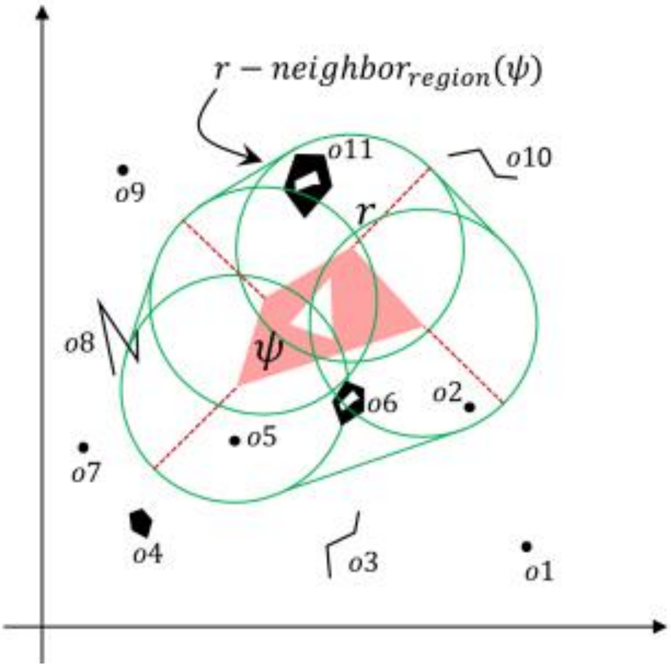

Figure 2.3.1.1: r-neighbours of a spatial object

2.3.2 Spatial Stream Indexing

Spatial data index structures can be classified into two broad categories they are tree and grid. Tree based spatial indexes can significantly speed-up the spatial query processing. However, their maintenance and restructuring cost is high especially during continuous updates (insertions and deletions). Grid based indexes enable fast updates but cannot answer queries as efficiently as tree-based indexes.[1][2]

2.3.3 Geoflink Grid Index

Geoflink is a distributed spatial data stream management system, hence an index which can work efficiently in distributed environment in the presence of heavy updates is needed. And the obvious choice is grid index which is not fast but can handle these heavy updates. The grid index is used in GeoFlink for filtering and pruning objects during spatial queries execution.[1][2]

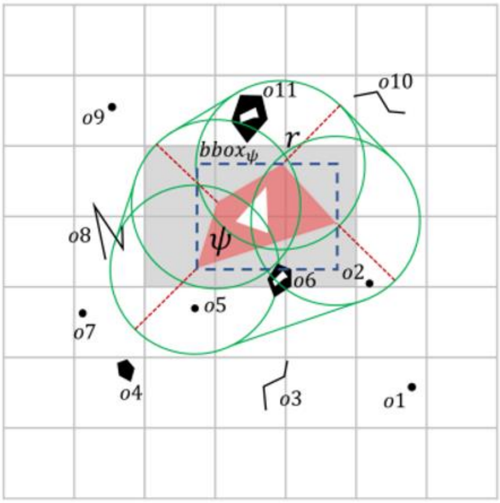

Figure 2.3.3.1: Geoflink Grid Index

2.3.4 Spatial Continuous Queries

There are 3 basic spatial continuous queries in GeoFlink:

Spatial Range- This query comes in handy when a user wants to fetch spatial objects within certain distance of query objects.[1][2]

Spatial kNN- This query is useful in fetching k nearest spatial objects of a query object.

Spatial Join- Spatial join is a useful operator where one stream is joined with another based on some query distance. Join is an expensive operator as it involves Cartesian product between the two streams.[1][2]

Each Geoflink query has 2 variants they are real-time query and window-based query.

Real-time Query - Real-time query is triggered with the arrival of new stream tuples. Precisely, as a new tuple is received by GeoFlink, it is processed by real-time query and a corresponding output is generated. [1][2]

Window-Based Query - Triggering of window-based query is based on window size and window slide step. The window-based query performs computation on all the spatial objects in the window and generates output corresponding to all the window contents. The query output is generated every slide step (Ws) in case of sliding window or every window size (Wn) in case of tumbling window.[1][2]

2.3.5 Geoflink Architecture

Geoflink architecture has 2 important layers:

Spatial Stream Layer - This layer is responsible for converting incoming data streams into spatial data stream. Apache flink treats spatial data stream as ordinary data stream which leads to insufficient processing. Geoflink converts it into spatial data stream of spatial objects.[1][2]

Real-Time spatial query processing layer - This layer provides support for a number of basic spatial operators required by most of the spatial data processing and analysis applications. Users can use Java or Scala to write the spatial queries or custom applications. This layer makes extensive use of the grid index for efficient queries' execution.[1][2]

After efficient query processing, sinks are used to provide the final processed output. Sink writes partitioned files to filesystems supported by the Flink File System.

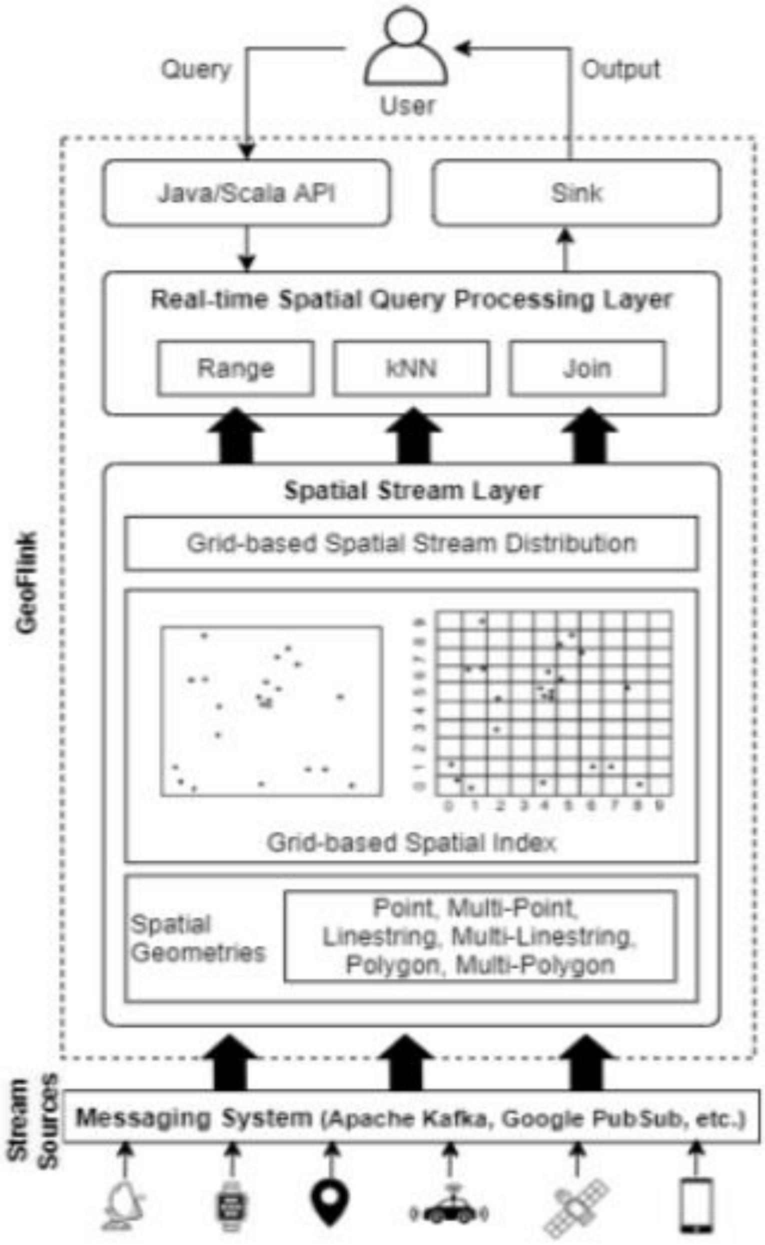
Query
User
Output
GeoFlink
Java/Scala API
Sink
Real-time Spatial Query Processing Layer
Range
kNN
Join
Spatial Stream Layer
Grid-based Spatial Stream Distribution
Grid-based Spatial Index
Spatial Geometries
Point, Multi-Point, Linestring, Multi-Linestring, Polygon, Multi-Polygon
Stream Sources
Messaging System (Apache Kafka, Google PubSub, etc.)

Figure 2.3.5.1: Geoflink Architecture

Chapter 3

Analysis

Geoflink is an efficient and scalable spatial data stream management system. It supports spatial datatypes, indexes and continuous queries. Geoflink supports spatial grid index for efficient query processing. To enable efficient processing of continuous spatial queries and for the effective data distribution among the Flink cluster nodes, a gird-based index is introduced. The grid index enables the pruning of the spatial objects which cannot be part of a spatial query result and thus guarantees efficient query processing. GeoFlink supports spatial range, spatial kNN and spatial join queries on Point, LineString, Polygon, MultiPoint, MultiLineString, and MultiPolygon spatial objects. GeoFlink supports incoming data streams in GeoJSON, CSV and WKT formats. By comparing Geoflink with other distributed spatial streaming platform Geoflink produces more desirable output along with processing of high throughput and high latency.

Chapter 4

Conclusion

Geoflink is a scalable data stream management system used for efficient processing of spatial queries which cannot be obtained by other streaming platforms. There are many cases where real-time processing of spatial queries is needed and Geoflink can achieve that feat. Geoflink proves to be an efficient tool when it comes to spatial data processing and in an era where spatial data is omnipresent and in high value this framework proves to be of great use.

References

[1] https://github.com/aistairc/SpatialFlink

[2] https://ieeexplore.ieee.org/document/9720936

[3] https://www.techtarget.com/searchdatamanagement/definition/spatial-data#:~:text=Spatial%20data%20is%20any%20type,in%20a%20geographic%20coordinate%20system.

[4] https://flink.apache.org/

[5] https://www.esri.com/en-us/what-is-gis/overview