WEST UNIVERSITY OF TIMIȘOARA
FACULTY OF MATHEMATICS AND INFORMATICS

Big Data

Report

Geospatial Big Data

Advisor(s): Lect. Dr. Adrian Spataru
Student(s): Daniel Ciucur

TIMISOARA, JULY 2023

Contents

1 Introduction

1.1 What is Geospatial data?

Geospatial data is information that describes objects, events or other features with a location on or near the surface of the earth.

Geospatial data typically combines **location information** (usually coordinates on the earth) and **attribute information** (the characteristics of the object, event or phenomena concerned) with **temporal information** (the time or life span at which the location and attributes exist). The location provided may be static in the short term (for example, the location of a piece of equipment, an earthquake event, children living in poverty) or dynamic (for example, a moving vehicle or pedestrian, the spread of an infectious disease).

Geospatial data typically involves large sets of spatial data gleaned from many diverse sources in varying formats and can include information such as census data, satellite imagery, weather data, cell phone data, drawn images and social media data. Geospatial data is most useful when it can be discovered, shared, analyzed and used in combination with traditional business data.

Geospatial analytics is used to add timing and location to traditional types of data and to build data visualizations. These visualizations can include maps, graphs, statistics and cartograms that show historical changes and current shifts. This additional context allows for a more complete picture of events. Insights that might be overlooked in a massive spreadsheet are revealed in easy-to-recognize visual patterns and images. This can make predictions faster, easier and more accurate.

Geospatial information systems (GIS) relate specifically to the physical mapping of data within a visual representation. For example, when a hurricane map (which shows location and time) is overlaid with another layer showing potential areas for lightning strikes, you're seeing GIS in action.

1.2 Types and examples of Geospatial data

Geospatial data is information recorded in conjunction with a geographic indicator of some type. There are three primary forms of geospatial data:

- raster data

- vector data (point, line, polygon)

- graph data

Vector data is data in which points, lines and polygons represent features such as properties, cities, roads, mountains and bodies of water. For example, a visual representation using vector data might include houses represented by points, roads represented by lines and entire towns represented by polygons.

Raster data is pixelated or gridded cells which are identified according to row and column. Raster data creates imagery that's substantially more complex, such as photographs and satellite images.

Graph data mainly appears in the form of road networks. Here, an edge represents a road segment, and a node represents an intersection or a landmark. The trajectories of vehicles on the road network are represented by sequences of road segments (edges).

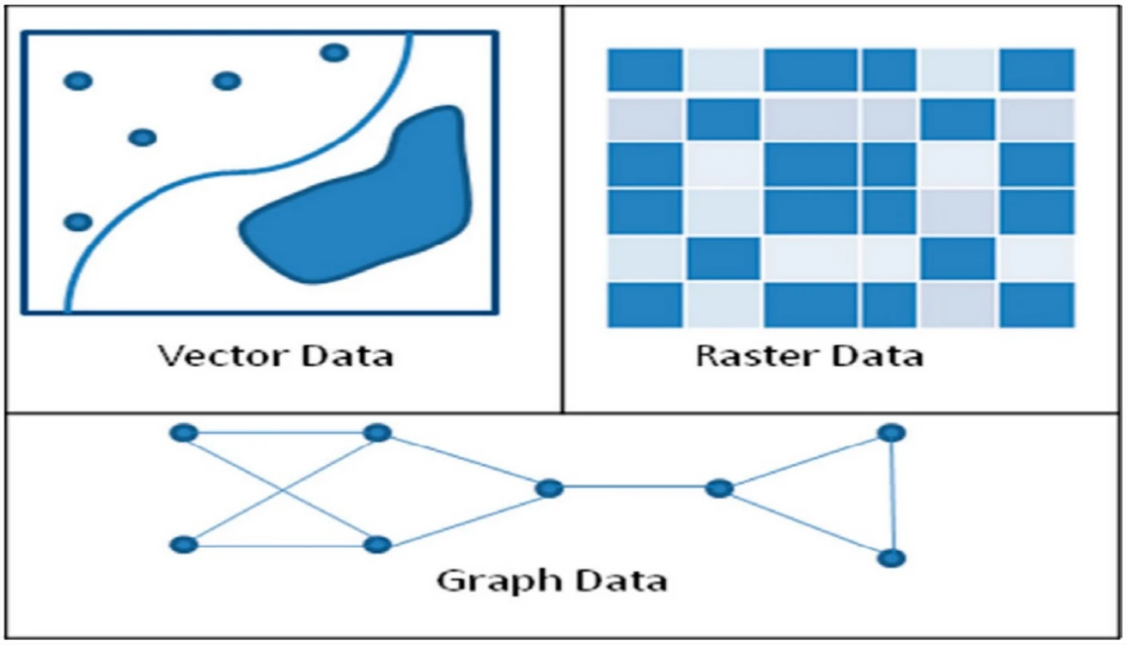


Figure 1.1: Types of Spatial Data

1.3 Sources of Geospatial Big Data

We have many sources of geospatial big data, I will summarize them bellow:

- Earth observations

  Earth observation systems generate massive volumes of disparate, dynamic, and geographically distributed geospatial data with in-situ and remote sensors. Remote sensing, with its increasingly higher spatial, temporal, and spectral resolutions, is one primary approach for collecting Earth observation data on a global scale. The Landsat archive, for example, exceeded one petabyte and contained over 5.5 million images several years ago (Wulder et al., 2016; Camara et al., 2016). As of 2014, NASA's Earth Observing System Data and Information System (EOSDIS) was managing more than nine petabytes of data, and it is adding about 6.4 terabytes to its archives every day (Blumenfeld, 2019). In recent years, the wide use of drone-based remote sensing has opened another channel for big Earth observation data collection (Athanasis et al., 2018).

- Geoscience model simulations

  The rapid advancement of computing power allows us to model and simulate Earth phenomena with increasingly higher spatiotemporal resolution and greater spatiotemporal coverage, producing huge amounts of simulated geospatial data. A typical example is the climate model simulations conducted by the Intergovernmental Panel on Climate Change (IPCC). The IPCC Fifth Assessment Report (AR5) alone produced ten petabytes of simulated climate data, and the next IPCC report is estimated to produce hundreds of petabytes (Yang et al., 2017; Schnase et al., 2017). Beside simulations, the process of calibrating the geoscience models also produces large amounts of geospatial data, since a model often must be run many times to sweep different parameters (Murphy et al., 2014). When calibrating ModelE (a climate model from NASA), for example, three terabytes of climate data were generated from 300 model-runs in just one experiment (Li et al., 2015).

- Internet of Things

  The term Internet of Things (IoT) was first coined by Kevin Ashton in 1999 in the context of using radio frequency identification (RFID) for supply chain management (Ashton, 2009). Simply speaking, the IoT connects "things" to the internet and allows them to communicate and interact with one another, forming a vast network of connected things. The things include devices and objects such as sensors, cellphones, vehicles, appliances, and medical devices, to name a few. These things, coupled with now-ubiquitous location-based sensors, are generating massive amounts of geospatial data. In contrast to Earth observations and model simulations that produce structured multi-dimensional geospatial data, IoT continuously generates unstructured or semi-structured geospatial data streams across the globe, which are more dynamic, heterogeneous, and noisy.

- Volunteered geographic information

Volunteered geographic information (VGI) refers to the creation and dissemination of geographic information from the public, a process in which citizens are regarded as sensors moving “freely” over the surface of the Earth (Goodchild, 2017). Enabled by the internet, Web 2.0, GPS, and smartphone technologies, massive amounts of location-based data are being generated and disseminated by billions of citizen sensors inhabiting the world. Through geotagging (location sharing), for example, social media platforms such as Twitter, Facebook, Instagram, and Flickr provide environments for digital interactions among millions of people in the virtual space while leaving “digital footprints” in the physical space. For example, about 500 million tweets are sent per day according to Internet Live Stats (2019); assuming the estimated 1% geotagging rate (Marciniec, 2017), five million tweets are geotagged daily.

1.4 Geospatial big data challenges

Dealing with large geospatial data sets presents many challenges. For this reason, many organizations struggle to take full advantage of geospatial data.

First, there is the sheer volume of geospatial data. For example, it is estimated that 100 TB of weather-related data is generated daily. This alone presents considerable storage and access problems for most organizations. Geospatial data is also stored across many different files, which makes it difficult to find the files that contain the data needed to solve your specific problem.

In addition, geospatial data is stored in many different formats and calibrated by different standards. Any effort to compare, combine or map data first requires a significant amount of data scrubbing and reformatting.

Finally, working with raw geospatial data requires specialized knowledge and the application of advanced mathematics to conduct necessary tasks, such as geospatial alignment of data layers. Unless analysts are proficient and experienced at this work, they will not get value from the data or make progress toward their organization’s business goals.

1.5 Geospatial analytics market and industry

The geospatial analytics market is presently experiencing considerable and steady growth; in fact, the market is expected to grow in value to USD 96.3 billion by 2025, achieving a 12.9% annual sales growth during the 5-year period under review.

Many industries make use of geospatial analytics:

- Governments can take insights about health, disease and weather and use them to better advise the public when a natural disaster strikes, or an emergency health event occurs.
- Electric utilities providers

  can use data to help predict possible service disruptions and optimize maintenance and crew schedules.
- Insurers

  can do a more accurate job of projecting risks and warning policy holders about potential issues they may soon be facing.
- Farm and agricultural lenders

  can improve the methodology they use to assess credit risk scores and reduce bad loan placements.

2 Frameworks for Spatial Data

2.1 Spatial Hadoop

SpatialHadoop is a full-fledged MapReduce framework with native support for spatial data. SpatialHadoop is a comprehensive extension to Hadoop that pushes spatial data inside the core functionality of Hadoop. SpatialHadoop runs existing Hadoop programs as is, yet, it achieves order(s) of magnitude better performance than Hadoop when dealing with spatial data.

2.1.1 Spatial Hadoop Arhitecture

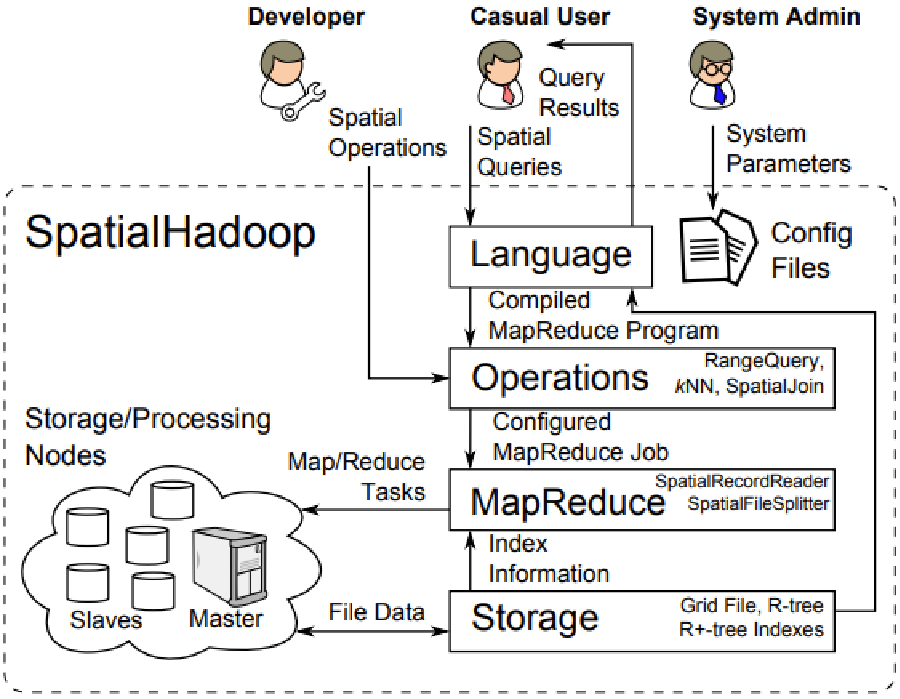


Figure 2.1: Spatial Hadoop Arhitecture

A SpatialHadoop cluster contains one master node that accepts a user query, breaks it into smaller tasks, and carries out the tasks on multiple slave nodes. There are three types of users who interact with SpatialHadoop, casual users, developers and administrators.

Casual users are non-technical users who access SpatialHadoop through the provided language to process their datasets. Developers are more advanced users who have deeper understanding of the system and can implement new spatial operations, which could be specific to some applications. Administrators are able to tune up the system through adjusting system parameters in the configuration files provided with SpatialHadoop installation.

SpatialHadoop adopts a layered design of four main layers, namely, language, storage, MapReduce, and operations layers.

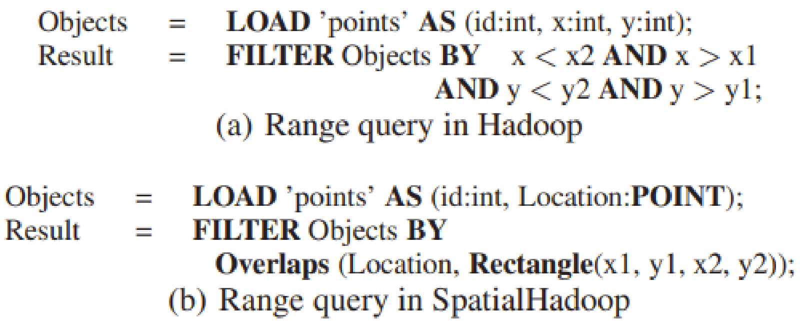
1) The language layer provides a simple high level SQL-like language that supports spatial data types and operations. Spatial hadoop does not have its own language. Pigeon extension is added to Pig language in SpatialHadoop. As map-reduce-like paradigms require huge coding efforts, a set of declarative SQL-like languages have been proposed, e.g., HiveQL, Pig Latin, and SCOPE. Spatial Hadoop does not provide a completely new language. Instead, it provides, Pigeon, an extension to Pig Latin language by adding spatial data types, functions, and operations that conform to the Open Geospatial Consortium (OGC) standard. Pigeon adds support for OGC-compliant **spatial data types** including, *Point*, *LineString*, and*Polygon*. Since Pig Latin does not allow defining new data types, Pigeon overrides the *bytearray* data type to define spatial data types. Conversion between bytearray and geometry is done automatically on the fly which makes it transparent to end users. Pigeon adds basic **spatial functions** which are used to extract useful information from a single shape; e.g., *Area* calculates the area of a polygonal shape. Pigeon supports OGC standard **spatial predicates** which return a Boolean value based on a test on the input polygon(s). For example, *IsClosed* tests if a linestring is closed while Touches checks if two geometries touch each other. **Spatial analysis functions** perform some spatial transformations on input objects such as calculating the *Centroid* or *Intersection*. These functions are usually used to performs a series of transformations on input records to produce final answer. **Spatial aggregate functions** take a set of spatial objects and return a single value which summarizes all input objects; e.g., the *ConvexHull* returns one polygon that represents the minimal convex polygon that contains all input objects. In addition to the functions in Pigeon, we do the following changes to the language: KNN to perform k-nearest neighbor query, FILTER for range query and JOIN for spatial joins.

2) The storage layer employs a two-level index structure of global and local spatial index structures. The global index partitions data across computation nodes while the local index organizes data inside each node.

3) The MapReduce layer has two new components, namely, SpatialFileSplitter and SpatialRecordReader that exploits the global and local indexes, respectively, to prune data that do not contribute to the query answer.

4) The operations layer encapsulates the implementation of various spatial operations that take advantage of the spatial indexes and the new components in the MapReduce layer. SpatialHadoop is initially equipped with an efficient implementation of three basic spatial operations, namely, range query, kNN, and spatial join. Other spatial operations can be added to the operations layer using a similar approach of the implementation of basic spatial operations.

2.1.2 Comparison with Hadoop



```
Objects   =   LOAD 'points' AS (id:int, x:int, y:int);
Result    =   FILTER Objects BY   x < x2 AND x > x1
                                  AND y < y2 AND y > y1;
```

(a) Range query in Hadoop

```
Objects   =   LOAD 'points' AS (id:int, Location:POINT);
Result    =   FILTER Objects BY
                Overlaps (Location, Rectangle(x1, y1, x2, y2));
```

(b) Range query in SpatialHadoop

Figure 2.2: Range query in Hadoop vs Spatial Hadoop

Spatial Hadoop is a comprehensive extension to Hadoop that pushes spatial constructs and the awareness of spatial data inside Hadoop code base. As a result, SpatialHadoop works in a similar way to Hadoop where programs are written in terms of map and reduce functions, and hence existing Hadoop programs can run as is on SpatialHadoop. Yet, if the programs deal with spatial data, SpatialHadoop will have order(s) of magnitude better performance than Hadoop.

According to one of the pappers I studied for this this report ("A Demonstration of SpatialHadoop: An Efficient MapReduce Framework for Spatial Data") the query in Figure 2.2 was run for 70M spatial objects on a 20 node cluster. For Hadoop the execution of the query took 200 seconds, while Spatial Hadoop took 2 seconds for the same query.

SpatialHadoop is distributed as an open source, which allows contributors in the research community to further extend its functionality. The basic components shipped in the core of Spatial Hadoop help in implementing more spatial operations in different applications efficiently. As case studies, SpatialHadoop already has three spatial operations, range queries, k-nearest-neighbor queries, and spatial join.

2.2 Beast

Beast is based on well-established research and has been released to assist the research community with analyzing big spatio-temporal data. Beast provides a set of extensible components that naturally integrate with Spark to build exploratory data science pipelines. Beast can install in less than a minute on an existing Spark cluster and provides a wide array of features including loading vector and raster data represented in standard file formats, synthetic data generation for benchmarking, load-balanced spatial partitioning, data summarization, interactive visualization, and more. Beast builds on several research projects; its goal is to make all this research widely available to researchers in one integrative and coherent system.

2.2.1 Beast Arhitecture

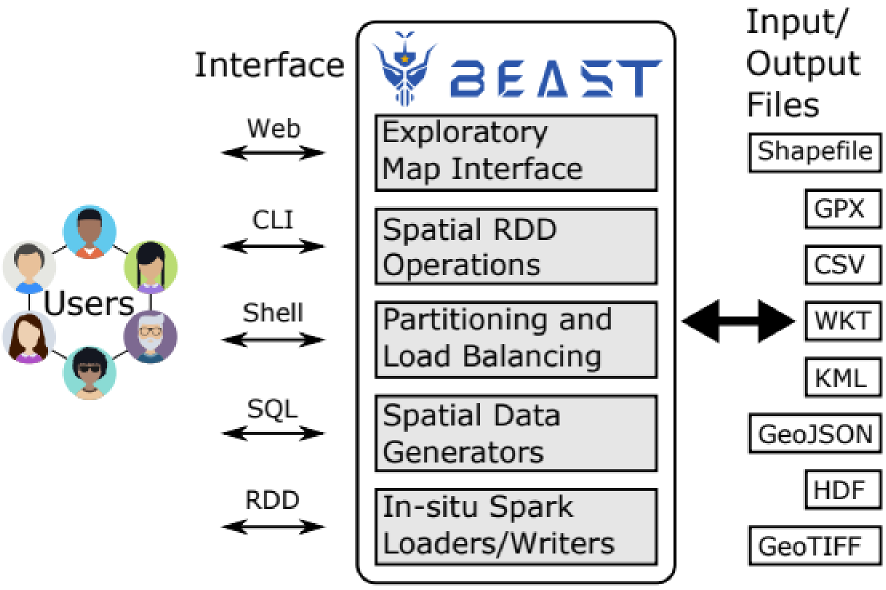


Figure 2.3: Beast Arhitecture

Beast which consists of five main components. First, to support in-situ data processing, Beast provides a set of parallel loaders and writers for popular file formats such as Shapefile, CSV, GeoJSON, and GeoTIFF. This component also provides scalable spatial data generators for stress testing and benchmarking.

Second, the spatial partitioner and load balancer component provides a set of spatial partitioning techniques which can group spatially relevant records into partitions while balancing the load across the executor nodes. The partitioned data can be written to disk in any of the standard file formats to be reused by Beast or any other system.

Third, to support interactivity, the interactive query processor offers a set of data synopses to facilitate approximate query processing, e.g., sample, point histogram, Euler histogram, and Bloom filter. It also uses these synopses to build some approximate algorithms such as clustering and selectivity estimation.

Fourth, the scalable join framework is crucial for big variety data since it allows users to integrate multiple datasets together. Beast provides a set of distributed join algorithm with various optimizations to handle big spatial data efficiently and uses a rule-based optimizer to choose the most appropriate algorithm.

Finally, the exploratory map interface helps users in visually exploring the input data or the query results on an interactive map interface.

Users can interact with all components of Beast through various interfaces. The web interface provides a graphical interface for some features in Beast such as the map visualization, data retrieval, and conversion.

UCR-Star is a an example of a web application built using this web interface. The command-line interface (CLI) gives quick access to some common features in Beast such as data conversion, indexing, and visualization. The interactive shell extends the Spark Scala shell with all features of Beast. It allows developers to try out the features of Beast or write short code snippets.

3 Real-life projects related to Geospatial BigData

3.1 Movebank

> "Movebank is a free, online database of animal tracking data hosted by the Max Planck Institute of Animal Behavior. We help animal tracking researchers to manage, share, protect, analyze and archive their data." - www.movebank.org

Movebank is presented as an ecosystem of tools used by thousands of researchers to collect, manage, share, visualize, analyse and archive their animal tracking and other animal-borne sensor data. Users add sensor data through file uploads or live data streams and further organize and complete quality control within the Movebank system. All data are harmonized to a data model and vocabulary. The public can discover, view and download data for which they have been given access to through the website, the Animal Tracker mobile app or by API. Advanced analysis tools are available through the EnvDATA System, the MoveApps platform and a variety of user-developed applications. Data owners can share studies with select users or the public, with options for embargos, licenses and formal archiving in a data repository.

Movebank can be considered a "platform" as it consists of more services. Of course at the core there is a PostgreSQL database hosted on servers at the Max Planck Computing and Data Facility in Garching, Germany. But besides that we also have a web application (movebank.org) where data owners can manage projects and data, and through which the public can browse data, connect with owners and access data for which they have appropriate permissions. Movebank's HTTP- and JSON-based RESTful APIs support transfer of data into and out of Movebank, thus allowing automated interactions with tag manufacturers, analysis software, other websites and mobile apps and advanced users. These allow development of third-party applications and provide scaling to transfer large volumes using automated procedures. All data transfer by API is subject to the same data model and access permissions as the database and webapp.

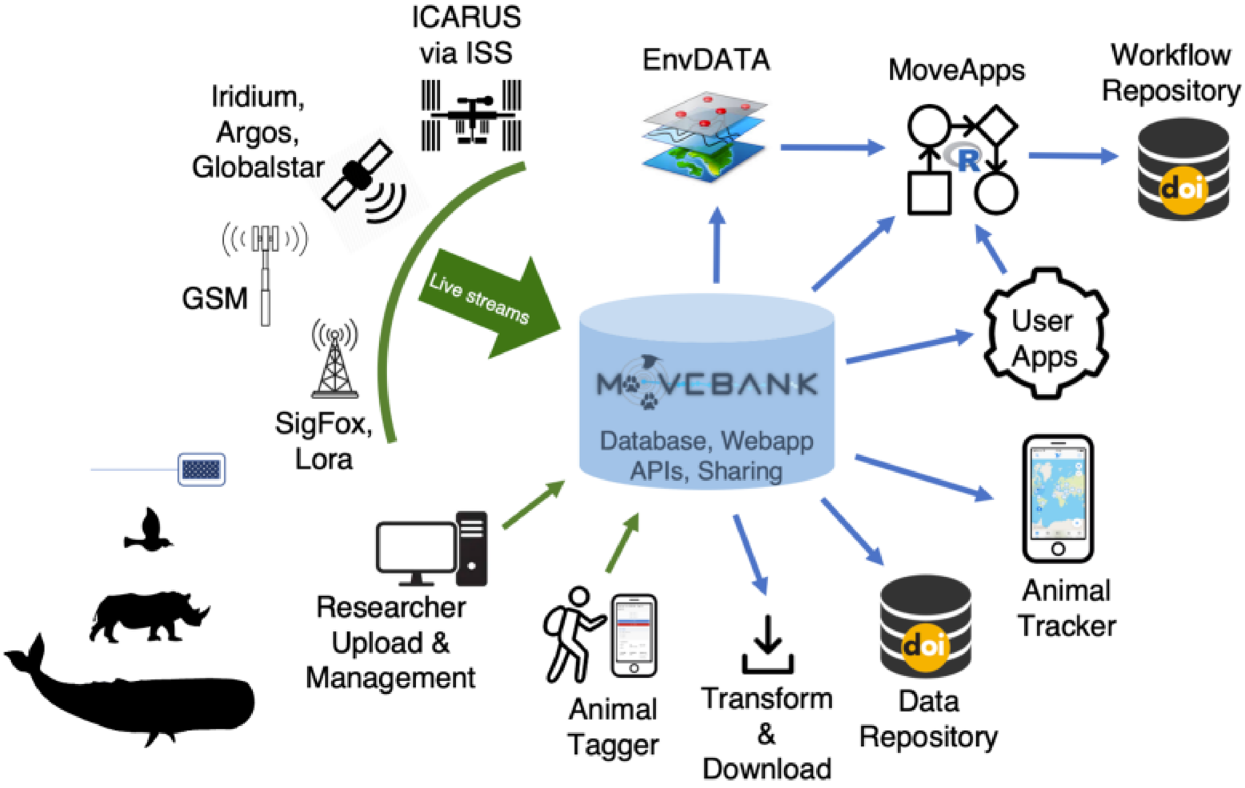
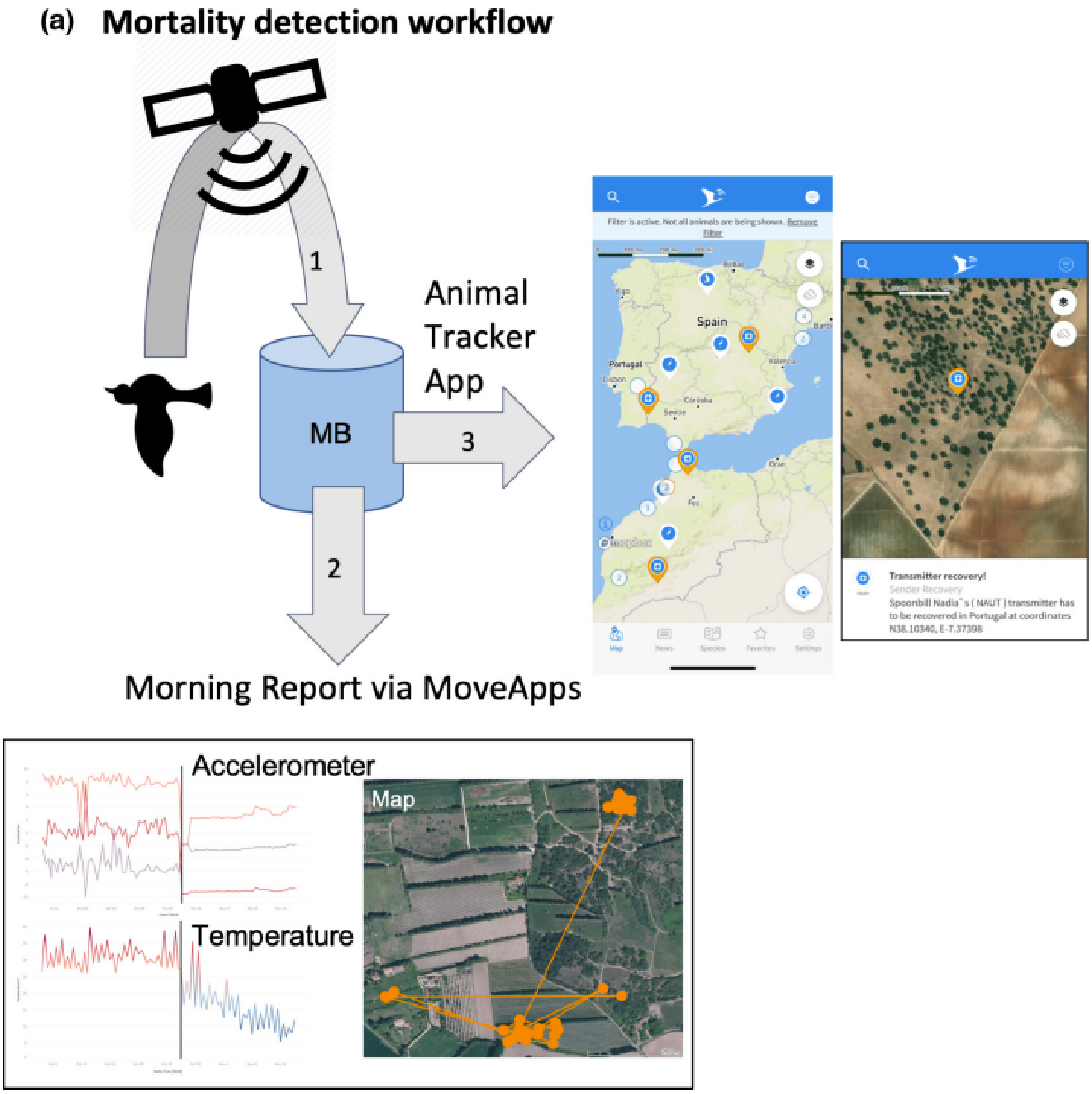


Figure 3.1: The Movebank cyberinfrastructure ecosystem of tools to acquire, manage and analyse animal tracking data

3.1.1 Case study: Cause of mortality for 171 white storks

In the papper "The Movebank system for studying global animal movement and demography" I also managed to find details regarding a case study that was done using Movebank which had to do with finding the mortality cause of 171 white storks.



Filter is active. Not all animals are being shown. Remove filter

Transmitter recovery!
Sender Recovery
Spoonbill Nadia`s (NAUT) transmitter has to be recovered in Portugal at coordinates N38.10340, E-7.37398

Figure 3.2: Mortality detection workflow

The workflow is consisting of (1) near real-time automated data transfer through wireless networks to the Movebank database where (2) an automated daily ‘Morning Report’ through the MoveApps platform that alerts researchers to the possible death of an animal through sensor streams (examples from a Eurasian blackbird Turdus merula with vertical black line indicating time of death). This location is then accessible to staff, collaborators or citizen scientists through the (3) Animal Tracker app so they can conduct a forensic investigation at the site.

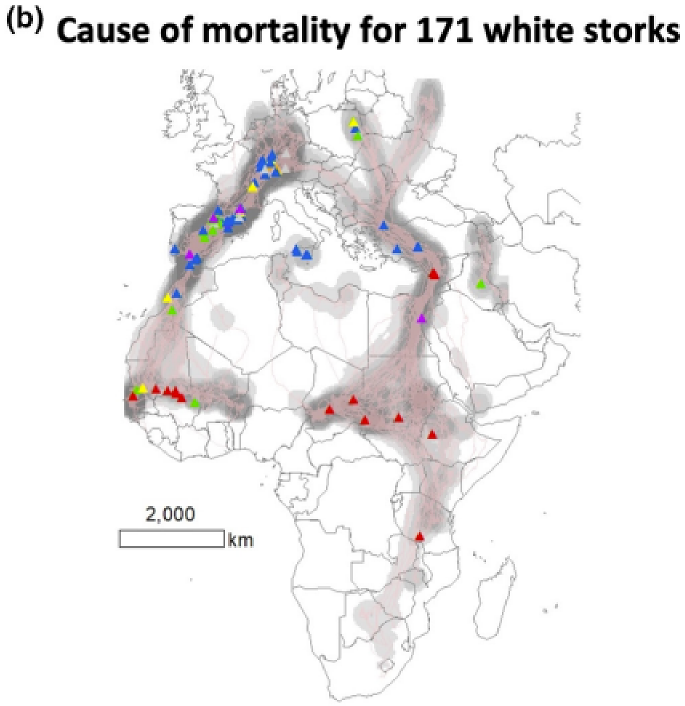


Figure 3.3: Storks mortality map

Results of this approach for 171 storks tracked across Europe and Africa. Shading on the map shows density of tracking locations, pink lines show individual bird tracks and coloured triangles show the location and cause of mortality events.

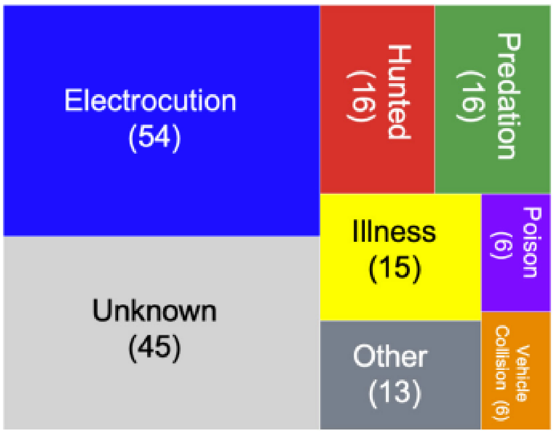


Figure 3.4: Causes of mortality of the storks in the study

4 Conclusion

In my report I managed study the topic of **Geospatial data**. I was interested to know what types of geospatial data we have, how we collect this data and what challenges we have. After that I studied tools/technologies which were created to help us deal with geospatial data, for this I analyzed **SpatialHadoop** and **Beast**. Lastly, I researched an open-source project which consists of an ecosystem of tools used by thousands of researchers to collect, manage, share, visualize, analyse and archive their animal tracking and other animal-borne sensor data.

References

[1] Harshawardhan S Bhosale and Devendra P Gadekar. A review paper on big data and hadoop. *International Journal of Scientific and Research Publications*, 4(10), 2014.

[2] Ahmed Eldawy, Vagelis Hristidis, Saheli Ghosh, Majid Saeedan, Akil Sevim, AB Siddique, Samriddhi Singla, Ganesh Sivaram, Tin Vu, and Yaming Zhang. Beast: Scalable exploratory analytics on spatio-temporal data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.

[3] Ahmed Eldawy and Mohamed F Mokbel. A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. *Proceedings of the VLDB Endowment*, 6(12), 2013.

[4] Ahmed Eldawy and Mohamed F Mokbel. The era of big spatial data. In *2015 31st IEEE International Conference on Data Engineering Workshops*. IEEE, 2015.

[5] Antonin Guttman. R-trees: A dynamic index structure for spatial searching. In *Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, 1984.

[6] Roland Kays, Sarah C Davidson, Matthias Berger, Gil Bohrer, Wolfgang Fiedler, Andrea Flack, Julian Hirt, Clemens Hahn, Dominik Gauggel, Benedict Russell, et al. The movebank system for studying global animal movement and demography. *Methods in Ecology and Evolution*, 13(2), 2022.

[7] Jae-Gil Lee and Kang Minseo. Geospatial big data: Challenges and opportunities. *Big Data Research*, 2, 02 2015.

[8] Zhenlong Li. Geospatial big data handling with high performance computing: Current approaches and future directions. *High Performance Computing for Geospatial Applications*, 2020.

[9] Yassine Loukili, Younes Lakhrissi, and Safae Elhaj Ben Ali. Geospatial big data platforms: A comprehensive review. *KN-Journal of Cartography and Geographic Information*, 72(4), 2022.