A Brief Introduction to Geospatial Big Data Analytics with Apache AsterixDB

Akil Sevim†, Mehnaz Tabassum Mahin†, Tin Vu†, Ian Maxon§, Ahmed Eldawy†, Michael Carey§, Vassilis Tsotras†
asevi006@ucr.edu,mmahi004@ucr.edu,tvu032@ucr.edu,imaxon@uci.edu
eldawy@ucr.edu,mjcarey@ics.uci.edu,tsotras@cs.ucr.edu
† University of California, Riverside
Riverside, California, USA
§University of California, Irvine
Irvine, California, USA

ABSTRACT

There is immense potential with spatial data, which is even more significant when combined with temporal or textual features, or both. However, it is expensive to store and analyze spatial data, and it is even more challenging with the combined features due to the additional optimization requirements. There are numerous successful solutions for big spatial data management, but they do not well support non-spatial operations. The options for the systems are even smaller for the open sources systems, and there are not a handful of options that provide good coverage of care about the spatial and non-spatial operations. This tutorial introduces Apache AsterixDB, a scalable open-source Big Data Management System, which supports standard vector spatial data types as well as non-spatial attributes, e.g., numerical, temporal, and textual. The participants will get hands-on experience on how Apache AsterixDB can efficiently process complex SQL++ queries that require multiple special handling by a team from its kitchen.

CCS CONCEPTS

• **Information systems → Parallel and distributed DBMSs**; **Information storage systems**.

KEYWORDS

geospatial big data, big data analytics, spatial analysis

ACM Reference Format:
Akil Sevim†, Mehnaz Tabassum Mahin†, Tin Vu†, Ian Maxon§, Ahmed Eldawy†, Michael Carey§, Vassilis Tsotras†. 2021. A Brief Introduction to Geospatial Big Data Analytics with Apache AsterixDB. In *3rd ACM SIGSPATIAL International Workshop on Geospatial Data Access and Processing APIs (SpatialAPI'21), November 2, 2021, Beijing, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3486189.3490018

This work is supported in part by the National Science Foundation (NSF) under grants CNS-1924694, CNS-1925610, IIS-1954644, IIS-1954962 and SES-1831615.

SpatialAPI'21, November 2, 2021, Beijing, China

ACM ISBN 978-1-4503-9103-0/21/11.
https://doi.org/10.1145/3486189.3490018

1 INTRODUCTION

The availability and the volume of the data that has geospatial features increase at a great rate as a result of the rise in GPS-enabled devices and the interest in data-driven applications. However, storing and analyzing the geospatial data requires extra care as they can get very complicated. To specify shared storage and access model for the vectoral spatial data, the Open Geospatial Consortium (OGC) and the International Organization for Standardization (ISO) define standardization. There are numerous systems as both commercial systems, such as Oracle, MS SQL Server, and open-source systems, such as PostGIS, Apache Sedona (formerly GeoSpark), and SpatialHadoop.

Apache AsterixDB[1] is an open-source, scalable Big Data Management System (BDMS) that provides a flexible data model, distributed storage and transaction, fast data ingestion, and data-parallel query execution runtime. Apache AsterixDB also supports spatial data types and operations since its foundation, and now, it implements the standards for the geospatial data. The presenting team from the University of California, Riverside, and the University of California, Irvine, continuously work on improving the geospatial infrastructure of Apache AsterixDB. This team has successfully presented several tutorials on AsterixDB, including a recent one at BOSS'21@VLDB. This tutorial will be the first to target the SIGSPATIAL community with a focus on fusing spatial and non-spatial operations. In this tutorial, the team will demonstrate how Apache AsterixDB can efficiently conduct complex data analytics tasks, including geospatial requirements, using the mechanisms it already has. As a result, the participants of this tutorial will learn how to deal with complex queries that combine, such as similarity functions, window functions, or interval joins with Apache AsterixDB. For instance, queries for finding similar reviews for businesses within a shopping mall which we define its borders with a polygon, or finding the users which reviewed businesses that are in a specific distance away from a geographical point and were "elite users" in overlapping intervals. In addition to that, Apache AsterixDB provides excellent techniques to manage both dynamic and static datasets. As a result, this tutorial will present a wider variety of queries in comparison to the GeoSpark tutorial [5] which mainly focuses on spatial analysis.

The tutorial will be a 30 minutes presentation split into three parts as follows. First, we start with an overview of Apache AsterixDB and the underlying technology and its capabilities. Second, we present data management operations include dataset and type

creation, data ingestion, external data usage, and indexing mechanisms. Lastly, **we will provide the audience dedicated Apache AsterixDB instances to have hands-on experience for running complex SQL++ queries that have both spatial and non-spatial predicates**.

2 DATASETS

We use two openly available datasets. For the hands-on experience, we will upload the files to Amazon Web Services (AWS) S3 so attendees can load them by themselves. In the following sections, we share the details of the datasets.

2.1 Yelp Dataset

The Yelp Dataset [4] consists of reviews, business, and user data for the metropolitan areas centered on Montreal, Calgary, Toronto, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland. There are approximately 8.6 million reviews for 160 thousand businesses. The dataset has five files in JSON format, one for each category consisting of businesses, reviews, users, check-ins, and tips. Figure 1 shows a selection of the business locations.

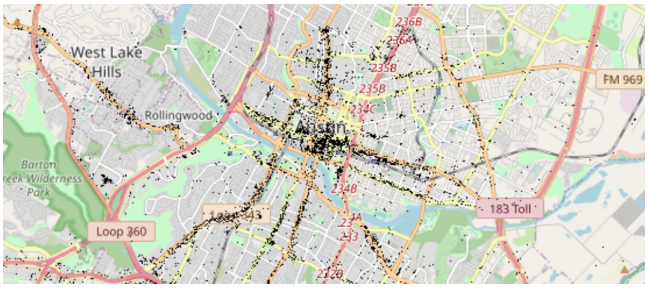

Figure 1: Bussiness Locations from Yelp Dataset

2.2 Roads Dataset

The Roads dataset [2] contains 18 million records that represent the road network of the US. Each road is represented as a Line that consists of multiple Points. Figure 2 shows a visualization for a part of the Roads datasets.

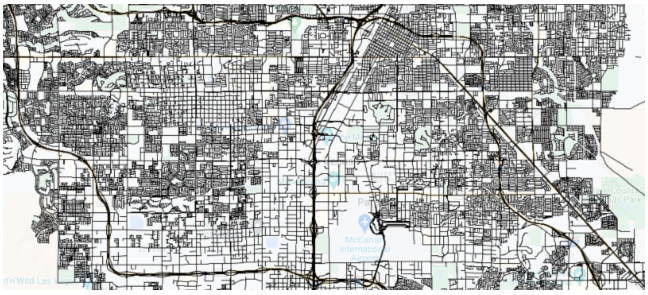

Figure 2: Roads Dataset

3 TUTORIAL OUTLINE

Our tutorial has three parts. Below, we describe the topics that we cover in each part.

Part 1 introduces Apache AsterixDB and reveals the robust architecture lying under the hood. In this part, next, we mention the recently added support for the vectoral spatial data types and spatial queries.

Part 2 first introduces UCR-Star[3], an open data repository that allows visualizing and downloading open geospatial datasets to show the audience how they can easily reach the open datasets. Then, we show the queries for creating dataverses, types, and datasets. After this point, attendees will learn how to load the datasets, and they will be able to experience Apache AsterixDB's fast ingestion mechanism. In addition to that, we mention how to use datasets externally. Lastly, we describe the usage of modification statements and the supported indexing for both spatial and non-spatial data.

Part 3 will allow the audience to have a hands-on experience. Each attendee will have a dedicated instance to replicate the operations we show, so we will first lead attendees to reach their instances. Then, we will start with queries run on single datasets. Next, the queries will be more complex, and participants will see how Apache AsterixDB can tackle complex analysis requirements, including spatial ones.

Below is the tentative outline for the tutorial.

- Overview of Apache AsterixDB
- Spatial data and query support
- Creating types and datasets
- Dataset loading
- External datasets
- Modification statements
- Indexing options
- Single dataset queries
- Multi-dataset queries

4 SUMMARY

Apache AsterixDB [https://asterixdb.apache.org/] is a scalable open-source BDMS, and with the full support for geospatial data types and operations, it increases its potential to another level. This brief tutorial will teach how Apache AsterixDB can be used effectively to overcome advanced big data analytic tasks that combine spatial, temporal and textual operations on dynamic and static datasets.

REFERENCES

[1] Apache AsterixDB. 2021. *Apache AsterixDB.* Retrieved August 27, 2021 from https://asterixdb.apache.org/
[2] US Census Bureau. 2019. All Roads. https://doi.org/10.6086/N1M906QB Retrieved from UCR-STAR https://star.cs.ucr.edu/?TIGER2018/ROADS&d.
[3] Saheli Ghosh, Tin Vu, Mehrad Amin Eskandari, and Ahmed Eldawy. 2019. UCR-STAR: The UCR Spatio-Temporal Active Repository. *SIGSPATIAL Special* 11, 2 (Dec. 2019), 34–40. https://doi.org/10.1145/3377000.3377005
[4] Yelp. 2021. *Yelp Dataset.* Retrieved August 27, 2021 from https://www.yelp.com/dataset/download
[5] Jia Yu and Mohamed Sarwat. 2019. Spatial Data Wrangling with GeoSpark: A Step by Step Tutorial. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Data Access and Processing APIs, SpatialAPI@SIGSPATIAL.* ACM.