Efficient Approximate Range Aggregation over Large-scale Spatial Data Federation (Extended Abstract)

Yexuan Shi[†], Yongxin Tong[†], Yuxiang Zeng[‡], Zimu Zhou[§], Bolin Ding, Lei Chen[‡]
[†] *State Key Laboratory of Software Development Environment, Beihang University, Beijing, China*
[‡] *Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, HongKong*
[§] *School of Information Systems, Singapore Management University, Singapore*
[†]{skyxuan,yxtong}@buaa.edu.cn, [‡]{yzengal,leichen}@cse.ust.hk, [§]zimuzhou@smu.edu.sg

2022 IEEE 38th International Conference on Data Engineering (ICDE) | 978-1-6654-0883-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDE53745.2022.00156

***Abstract*— Data federations notably increase the amount of data available for data-intensive applications such as smart mobility planning and public health emergency responses. Yet they also challenge the conventional implementation of range aggregation queries because the raw data cannot be shared within the federation and the data partition at each data silo is fixed during query processing. In this work, we propose the first-of-its-kind approximate algorithms for efficient range aggregation over spatial data federation. We devise novel single-silo sampling algorithms that process queries in parallel and design a level sampling based algorithm which reduces the time complexity of local queries at each data silo to $O(\log \frac{1}{\epsilon})$, where ϵ is the approximation ratio of the accuracy guarantee. Extensive experiments on real-world dataset validate the efficiency and effectiveness of the solutions.**

I. Introduction

Range aggregation queries over spatial data returns summarized information about the spatial objects falling within a spatial range specified as either a circle or a rectangle [1], which are fundamental queries for various big spatial data applications. There is a growing trend for the service provider of these applications to operate on a *data federation* [2], where the data from multiple data providers (*a.k.a.*, data silos) collaborate to improve the quality of services. Each data silo in the federation holds part of the entire data (*i.e.*, rows) under the same schema, and interact with the service provider without revealing its own raw data partition.

However, it is challenging to offer real-time response to frequent range aggregation queries in spatial data federation. *(i)* Traditional distributed range aggregation techniques improve query processing throughput by optimizing data partitioning. However, the data partition is fixed in the federation setting. *(ii)* Prior spatial query processing schemes fail to deliver real-time response in case of high-frequency queries. For example, real-world bike sharing applications may receive around 150 queries per second, whereas existing exact range aggregation solutions can only process 50 queries per second [3].

In this paper, we define the Federated Range Aggregation (FRA) problem and investigate efficient solutions to range aggregation queries over large-scale spatial data federation. Observing that the underlying applications demand real-time response of high-frequency queries while a small error in the result is acceptable, we focus on solutions that offer *high-throughput* and high-quality *approximate* query results. Our main contributions are as follows. *(i)* We devise a novel single-silo sampling scheme that radically reduces the communications with data silos to one round of interaction with a single silo. Such a reduction in communication cost facilitates parallel processing of FRA queries for high throughput. *(ii)* We propose a new level sampling based index called LSR-Forest for each data silo to accelerate the local range aggregation query at each data silo, which has a time complexity of $O(\log \frac{1}{\epsilon})$. *(iii)* We extensively evaluate the effectiveness and efficiency of the algorithms on real datasets.

II. Problem Statement

We focus on the scenario where multiple data silos are united as a federation for querying over a collection of spatial objects. A **spatial object** is denoted by $o = \langle l_o, a_o \rangle$, where l_o is the location of the spatial object and a_o is the corresponding measure attribute. Each **data silo** s_i contains its own spatial objects, denoted by $P_{s_i} = \{o_1, o_2, \cdots, o_{n_{s_i}}\}$. A **federation** S consists of m data silos, *i.e.*, $S = \bigcup_{i=1}^{m} \{s_i\}$. The service provider makes queries over the federation S but can only access the spatial objects in P_{s_i} via the query interface of data silo s_i.

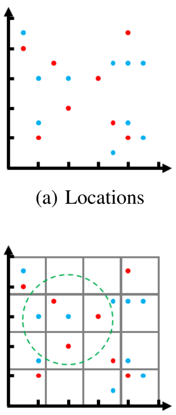
Definition 1 (FRA Query)**.** *Given a federation S possessing a collection of spatial objects P, a query range R and an aggregation function F, a federated range aggregation (FRA) query Q from the service provider aims to aggregate the measure attributes of the spatial objects within R:*

$$Q(S, R, F) = F(\{a_o \mid o \in P, o \text{ is within } R\}), \qquad (1)$$

where each data silo s_i can only access its own data partition P_{s_i}, i.e., s_i can only answer the range aggregation query of $Q(s_i, R, F) = F(\{a_o \mid o \in P_{s_i}, o \text{ is within } R\})$, and R can be either circular or rectangular.

Example 1. *Assume a federation S of two data silos (Fig. 1a). The first data silo has 10 spatial objects (marked in blue) and the second data silo holds 8 spatial objects (marked in red).*

2375-026X/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDE53745.2022.00156



(a) Locations

(b) FRA query

Data Silo #1

ID	l_o	a_o
1	(1,9)	1
2	(2,3)	4
3	(2,6)	2
4	(4,6)	1
5	(7,1)	4
6	(7,7)	2
7	(8,3)	3
8	(8,7)	2
9	(9,2)	6
10	(9,7)	2

Data Silo #2

ID	l_o	a_o
1	(1,8)	3
2	(2,2)	7
3	(3,7)	1
4	(4,4)	1
5	(6,6)	2
6	(7,3)	4
7	(8,2)	5
8	(8,9)	2

l_o: location
a_o: measure attribute

(c) Spatial objects in each data silo

Fig. 1: An example of the FRA query

The locations and measure attributes of these spatial objects are in Fig. 1c. An FRA query is shown in Fig. 1b, which asks the SUM *of the measure attributes of those spatial objects within a circular range (marked in green) centered at $(4, 6)$ with a radius of* 3.

We aim to develop high-throughput, high-quality approximate algorithms to process frequent FRA queries over large-scale data federation. The accuracy of approximate algorithms is quantified by *ϵ-approximation*.

Definition 2 (ϵ-approximation)**.** *For an FRA query with an exact result of ans, an ϵ-approximation solution should always report a result ans' such that $(1-\epsilon)ans \leq ans' \leq (1+\epsilon)ans$.*

III. Solution Overview

We optimize FRA query processing from two aspects.

- *Avoid enumerating all data silos to answer an FRA query.* A naive solution would exchange information with every data silo to answer a range aggregation query, allowing only sequential processing. Proper sampling strategies can enable parallel processing, which improves the throughput on query streams.
- *Accelerate local range aggregation queries at each data silo.* Although spatial indices such as R-trees enable $O(\log n_{s_k})$-time range aggregation queries for a data silo s_k, the time to obtain a partial aggregation answer is still a bottleneck. We argue that the local range aggregation queries can be further sped up via approximate solutions.

Fig. 2 shows an overview of our solution. Central in our solution are two techniques:

(1) **Single-Silo Sampling.** We reduce the number of silos for partial aggregation result retrieval from m to 1 to allow parallel query processing. This is achieved with a grid index to track the distribution of the spatial data partitioning and the corresponding query result estimation algorithms.

(2) **Level Sampling based Local Query.** We devise a novel level sampling based index (LSR-Forest) for fast approximate range aggregation queries at each silo. It reduces the average time cost of local range aggregation queries to $O(\log \frac{1}{\epsilon})$.

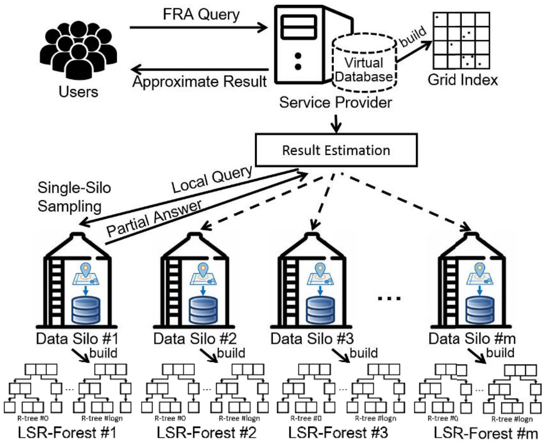


Fig. 2: An overview of our approximate solutions

IV. Experimental Study

Dataset. We conduct experiments on real-word datasets collected by three shared mobility companies in Beijing. The total size of records in this dataset is over 1TB. Each record has a collected time, the vehicle's location and affiliated company, and the number of carried passengers (as measure attribute). The proportion of the records owned by the three companies is $1 : 1 : 2$. We vary the size of data federation $|P|$ (the total number of spatial objects) from 1 million to 5 million and the number of data silos m from 3 to 15.

Compared Algorithms. We compare the exact solution *EXACT* [1], an optimal approximate histogram-based solution *OPTA* [4], our *Single-Silo Sampling* method and its extened version with *LSR-Forest*.

Summary of Experimental Results. Our approximate algorithms are significantly more efficient than *EXACT* and can achieve smaller errors than *OPTA*. Specially, our level sampling based local query notably improves the efficiency with only a small increase of error. Overall, our solutions can process over 250 queries per second, which is fit for real-time applications like federated ride-hailing services.

Acknowledgements

This work is partially supported by the National Science Foundation of China (NSFC) under Grant No. U21A20516, U1811463 and 62076017, and the Lee Kong Chian Fellowship awarded to Zimu Zhou by Singapore Management University.

References

[1] Y. Tao and D. Papadias, "Range aggregate processing in spatial databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1555–1570, 2004.
[2] J. Bater, G. Elliott, C. Eggen, S. Goel, A. Kho, and J. Rogers, "Smcql: secure querying for federated databases," *PVLDB*, vol. 10, no. 6, pp. 673–684, 2017.
[3] Y. Shi, Y. Tong, Y. Zeng, Z. Zhou, B. Ding, and L. Chen, "Efficient approximate range aggregation over large-scale spatial data federation," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
[4] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss, "Optimal and approximate computation of summary statistics for range aggregates," in *PODS*, 2001.