# Skyline Queries on Vertically Partitioned Tables

D

**José Subero**
*Universidad Simón Bolívar, Venezuela*

**Marlene Goncalves**
*Universidad Simón Bolívar, Venezuela*

## INTRODUCTION

In the last decade, new database systems have emerged to support distributed data over the Web. Systems such as Amazon DynamoDB (Sivasubramanian, 2012), Bigtable (Chang et al., 2008), HBase (Dumon, 2013) and Cassandra (Lakshman & Malik, 2010) allow to store a very large size of data commonly produced in real time. To identify relevant data in large datasets, Skyline queries have been proposed (Börzsönyi et al., 2001). Users may express their preferences using a Skyline query. To exemplify Skyline queries, suppose a tourist who wants to find hotels in the Caribbean with the highest number of stars and the lowest distance to the beach and cost per night. Also, consider the 4 hotels shown in Figure 1.

The tourist's query may be specified as a Skyline query. In this type of queries, tourist's criteria are expressed by means of functions that maximize or minimize attributes. In this example, tourist's criteria correspond to maximize the number of stars and minimize the distance and the cost. Based on these criteria, the hotels h1, h2 and h4 in Figure 1 belong to the Skyline set. They are the best hotels since no other hotel is better than them in all criteria simultaneously. It can be noted the hotel h3 is worse or equal in all criteria than the hotel h1, and therefore, h3 is dominated by h1 and it is not part of the Skyline set.

To the best of our knowledge, few works have defined evaluation algorithms for Skyline queries over Vertically Partitioned Tables (VPTs) (Balke et al., 2004; Chen et al., 2011); a VPT stores data as key-value and it may be sorted (Vasani et al., 2013). Additionaly, VPTs has received the attention of many researchers because a vertical partitioned schema achieves performance improvements that has proven to be useful in a variety and type of data processed today including data warehousing, biomedical data, and RDF data.

To illustrate the evaluation of our Skyline query example on VPTs, consider the data partition shown in Figure 2.

A naive solution to answer a Skyline query on VPTs is to join the VPTs; the results of joining the VPTs are data displayed in Figure 1. Subsequently, the Skyline may be computed on this set of joined tuples using an existing Skyline query evaluation algorithm on centralized databases (Börzsönyi et al., 2001; Xu et al., 2012; Godfrey et al., 2007). This naive solution is inefficient since it builds the set of joined tuples before executing an algorithm to return the Skyline.

A second naive solution is to compute the Skyline before joining the VPTs. A smaller amount of data will

*Figure 1. Hotel data*

| Id | Star | Distance | Cost |
|---|---|---|---|
| h1 | 3 | 3.0 | 500 |
| h2 | 4 | 5.0 | 510 |
| h3 | 3 | 4.0 | 520 |
| h4 | 3 | 2.0 | 530 |

DOI: 10.4018/978-1-4666-5888-2.ch180

Copyright © 2015, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

*Figure 2. VPTs of hotel data*

| Id | Star |
|---|---|
| h2 | 4 |
| h1 | 3 |
| h4 | 3 |
| h3 | 3 |

| Id | Distance |
|---|---|
| h4 | 2.0 |
| h1 | 3.0 |
| h3 | 4.0 |
| h2 | 5.0 |

| Id | Cost |
|---|---|
| h1 | 500 |
| h2 | 510 |
| h3 | 520 |
| h4 | 530 |

be joined because the Skyline cardinality is smaller than the dataset size (Godfrey et al., 2007). In consequence, this solution is more efficient than the first one. However, this solution may be incorrect because the query answer may be empty. Let us consider the 2 new VPTs in Figure 3 where Chain indicates the hotel chain of each hotel and Ranking corresponds to the overall score of each hotel chain. Also, the tourist would also like to maximize the chain ranking.

Notice that Star, Distance and Cost tables have the same keys while the Ranking table has a different key. Thus, the VPTs are joined by means of the Chain table. We named the Chain table as an intermediate table. An intermediate table has not preference attributes, but rather its attributes refer to keys that are defined in other VPTs of the query.

Using the second naive solution, the query answer is empty as shown in Figure 4. The first step is to find the Skyline of each VPT. The Skyline of Star, Distance, and Cost are h1, h2 and h4, respectively. The Skyline of the Ranking table corresponds to the Marriot chain. In the second step, the join operator is executed. Nevertheless, h1, h2 and h4 belong to the Eurobuilding chain while Marriott chain has the hotel h3. They do not join themselves. Consequently, the query answer using the second naive solution is empty.

*Figure 3. Skyline on VPTs of different sizes*

| Id | Star |
|---|---|
| h2 | 4 |
| h1 | 3 |
| h4 | 3 |
| h3 | 3 |

| Id | Distance |
|---|---|
| h4 | 2.0 |
| h1 | 3.0 |
| h3 | 4.0 |
| h2 | 5.0 |

| Id | Cost |
|---|---|
| h1 | 500 |
| h2 | 510 |
| h3 | 520 |
| h4 | 530 |

| Id | Chain |
|---|---|
| h3 | Marriot |
| h1 | Eurobuilding |
| h2 | Eurobuilding |
| h4 | Eurobuilding |

| Id | Ranking |
|---|---|
| Marriot | 5 |
| Eurobuilding | 4 |

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/skyline-queries-on-vertically-partitioned-tables/112592

## Related Content

Teaching Formation to Develop Computational Thinking
Klinge Orlando Villalba Condori (2018). *Global Implications of Emerging Technology Trends (pp. 59-72).*
www.irma-international.org/chapter/teaching-formation-to-develop-computational-thinking/195821

Human Resources Development in a Technology-Infused Workplace
Keri K. Stephensand Stephanie L. Dailey (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 3694-3702).*
www.irma-international.org/chapter/human-resources-development-in-a-technology-infused-workplace/112804

A Systematic Review on Author Identification Methods
Sunil Digamberrao Kaleand Rajesh Shardanand Prasad (2017). *International Journal of Rough Sets and Data Analysis (pp. 81-91).*
www.irma-international.org/article/a-systematic-review-on-author-identification-methods/178164

Digital Literacy Education for Digital Inclusion
Seunghyun Lee (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 1-9).*
www.irma-international.org/chapter/digital-literacy-education-for-digital-inclusion/112624

Scrum Software Development Methodology
Ruth Guthrie (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 7310-7318).*
www.irma-international.org/chapter/scrum-software-development-methodology/112428