



US006108658A

# United States Patent [19]

[11] Patent Number: **6,108,658**

Lindsay et al.

[45] Date of Patent: **Aug. 22, 2000**

[54] **SINGLE PASS SPACE EFFICIENT SYSTEM AND METHOD FOR GENERATING APPROXIMATE QUANTILES SATISFYING AN APRIORI USER-DEFINED APPROXIMATION ERROR**

[75] Inventors: **Bruce Gilbert Lindsay**, San Jose; **Gurmeet Singh Manku**, Santa Clara; **Sirdhar Rajogopalan**, San Jose, all of Calif.

[73] Assignee: **International Business Machines Corporation**, Armonk, N.Y.

[21] Appl. No.: **09/050,434**

[22] Filed: **Mar. 30, 1998**

[51] Int. Cl.<sup>7</sup> ..... **G06F 17/30**

[52] U.S. Cl. .... **707/101; 707/6; 707/7; 707/2**

[58] Field of Search ..... **707/2, 6, 7, 101**

[56] **References Cited**

**U.S. PATENT DOCUMENTS**

4,379,948	4/1983	Ney et al. ....	179/1 SC
4,530,076	7/1985	Dwyer .....	367/135
4,817,158	3/1989	Picheny .....	381/47
4,829,427	5/1989	Green .....	364/300
5,018,088	5/1991	Higbie .....	364/574
5,091,967	2/1992	Ohsawa .....	382/22
5,105,469	4/1992	MacDonald et al. ....	382/17
5,345,585	9/1994	Iyer et al. ....	395/600
5,379,419	1/1995	Heffernan et al. ....	395/600
5,664,171	9/1997	Agrawal et al. ....	395/602
5,864,841	1/1999	Agrawal et al. ....	707/2

**OTHER PUBLICATIONS**

Sridhar Rajagopalan et al, "Approximate Order Statistics in One Pass and Little Memory: Theory and Database Applications", IBM Almaden Research Center, Nov. 3, 1997, pp. 1-18.

Viswanath Poosala et al, "Improved Histograms for Selectivity Estimation of Range Predicates", SIGMOD, Jun. 1996, pp. 294-305.

Raj Jain & Imrich Chlamtac, "The P<sup>2</sup> Algorithm for Dynamic Calculation of Quantiles and Histograms Without Storing Observations", Communications of the ACM, vol. 28, No. 10, Oct. 1995, pp. 1076-1085.

Khaled Alsabti et al, "A One-Pass Algorithm for Accurately Estimating Quantiles for Disk-Resident Data", Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997, pp. 346-355.

Richard J. Lipton et al, "Efficient sampling strategies for relational database operations", Elsevier Science Publishers A.V., 1993, pp. 195-226.

Rakesh Agrawal & Arun Swami, "A One-Pass Space-Efficient Algorithm for Finding Quantiles", 7th Int'l Conf. Management of Data (COMAD-95), Prune, India 1995, pp. 1-12.

L.G. Valiant, "A Theory of the Learnable", Communications of the ACM, vol. 27, No. 11, Nov. 1984, pp. 1134-1142.

Ira Pohl, "A Minimum Storage Algorithm for Computing The Median", IBM Thomas J. Watson Research Center, Nov. 17, 1969, pp. 1-6.

(List continued on next page.)

*Primary Examiner*—Wayne Amsbury

*Assistant Examiner*—Thuy Pardo

*Attorney, Agent, or Firm*—Gray Cary Ware Freidenrich

[57]

**ABSTRACT**

A system and method for finding an  $\epsilon$ -approximate  $\phi$ -quantile data element of a data set with N data elements in a single pass over the data set. The  $\epsilon$ -approximate  $\phi$ -quantile data element is guaranteed to lie within a user-specified approximation error  $\epsilon$  of a true  $\phi$ -quantile data element being sought. B buffers, each having a capacity of k elements, initially are filled with sorted data elements from the data set, with the values of b and k depending on  $\epsilon$  and N. The buffers are then collapsed into an output buffer, with the remaining buffers then being refilled with data elements, collapsed (along with the previous output buffer), and so on until the entire data set has been processed and a single output buffer remains. A data element of the output buffer corresponding to the  $\epsilon$ -approximate  $\phi$ -quantile is then output as the approximate  $\phi$ -quantile data element. If desired, the system and method can be practiced with sampling to even further reduce the amount of space required to find a desired  $\epsilon$ -approximate  $\phi$ -quantile data element.

**35 Claims, 6 Drawing Sheets**

