



US006343288B1

(12) **United States Patent**
Lindsay et al.

(10) **Patent No.:** **US 6,343,288 B1**
(45) **Date of Patent:** ***Jan. 29, 2002**

(54) **SINGLE PASS SPACE EFFICIENT SYSTEM AND METHOD FOR GENERATING AN APPROXIMATE QUANTILE IN A DATA SET HAVING AN UNKNOWN SIZE**

(75) Inventors: **Bruce Gilbert Lindsay**, San Jose; **Gurmeet Singh Manku**, Santa Clara; **Sridhar Rajagopalan**, San Jose, all of CA (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(*) **Notic:** Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **09/268,089**

(22) Filed: **Mar. 12, 1999**

(51) **Int. Cl.**⁷ **G06F 17/30**

(52) **U.S. Cl.** **707/7; 707/2; 707/6; 707/101**

(58) **Field of Search** **707/2, 6, 7, 101**

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,379,948 A	4/1983	Ney et al.	704/203
4,530,076 A	7/1985	Dwyer	367/135
4,817,158 A	3/1989	Picheny	704/224
4,829,427 A	5/1989	Green	707/4
5,018,088 A	5/1991	Higble	702/194
5,091,967 A	2/1992	Ohsawa	382/172
5,105,469 A	4/1992	MacDonald et al.	382/162
5,345,585 A	9/1994	Iyer et al.	707/2
5,379,419 A	1/1995	Hefferman et al.	707/4
5,664,171 A	9/1997	Agrawal et al.	502/120
5,864,841 A *	1/1999	Agrawal et al.	707/2
6,108,658 A *	8/2000	Lindsay et al.	707/101
6,195,657 B1 *	2/2001	Rucker et al.	707/5

OTHER PUBLICATIONS

Fu, et al. (IEEE publication, 2001) paper entitled "Novel algorithms for computing medians and other quantiles of dis-resident data" in Database Engineering and Application, 2001 International Symposium, pp. 145-154.*

Publication: "A One-Pass Space-Efficient Algorithm for Finding Quantiles". Agrawal et al. Proceedings of the 7th International Conference on Management of Data (COMAD-95). India. 1995.

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

(List continued on next page.)

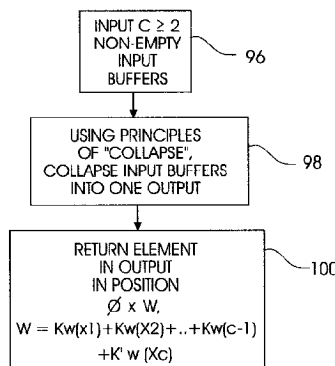
Primary Examiner—Diane D. Mizrahi

(74) *Attorney, Agent, or Firm*—John L. Rogitz

(57) **ABSTRACT**

A space-efficient system and method for generating an approximate ϕ -quantile data element of a data set in a single pass over the data set, without a priori knowledge of the size of the data set. The approximate ϕ -quantile is guaranteed to lie within a user-specified approximation error ϵ of the true quantile being sought with a probability of at least $1-\delta$, with δ being a user-defined probability of failure. B buffers, each having a capacity of k elements, initially are filled with elements from the data set, with the values of b and k depending on approximation error ϵ and the probability δ . The buffers are then collapsed into an output buffer, with the remaining buffers then being refilled with elements, collapsed (along with the previous output buffer), and so on until the entire data set has been processed and a single output remains. The element of the output corresponding to the approximate quantile is then output as the approximate quantile. In later iterations (when the height of the tree is at least equal to a predetermined height that depends on δ and ϵ), the data is sampled non-uniformly to populate the buffers to render the desired performance. Parallel processors can be used, with the final output buffers of the processors being sent to a collecting processor P_0 as input buffers to the collecting processor P_0 .

48 Claims, 9 Drawing Sheets



"OUTPUT"