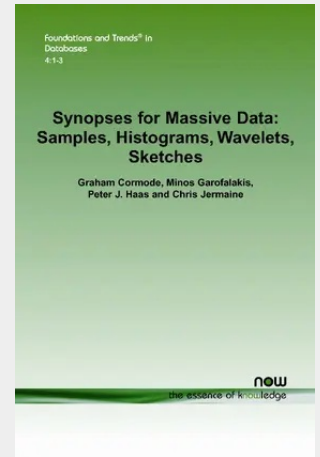


Synopses for Massive Data

Samples, Histograms, Wavelets, Sketches

Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches describes basic principles and recent developments in building approximate synopses (that is, lossy, compressed representations) of massive data. Such synopses enable approximate query processing, in which the user's query is executed against the synopsis instead of the original data. It focuses on the four main families of synopses: random samples, histograms, wavelets, and sketches. A random sample comprises a "representative" subset of the data values of interest, obtained via a stochastic mechanism. Samples can be quick to obtain, and can be used to approximately answer a wide range of queries. A histogram summarizes a data set by grouping the data values into subsets, or "buckets," and then, for each bucket, computing a small set of summary statistics that can be used to approximately reconstruct the data in the bucket. Histograms have been extensively studied and have been incorporated into the query optimizers of virtually all commercial relational DBMSs. Wavelet-based synopses were originally developed in the context of image and signal processing. The data set is viewed as a set of M elements in a vector-i.e., as a function defined on the set $\{0, 1, 2, \dots, M-1\}$ -and the wavelet transform of this function is found as a weighted sum of wavelet "basis functions." The weights, or coefficients, can then be "thresholded", e.g., by eliminating coefficients that are close to zero in magnitude. The remaining small set of coefficients serves as the synopsis. Wavelets are good at capturing features of the data set at various scales. Sketch summaries are particularly well suited to streaming data. Linear sketches, for example, view a numerical data set as a vector or matrix, and multiply the data by a fixed matrix. Such sketches are massively parallelizable. They can accommodate streams of transactions in which data is both inserted and removed. Sketches have also been used successfully to estimate the answer to COUNT DISTINCT queries, a notoriously hard problem. Synopses for Massive Data describes and compares the different synopsis methods. It also discusses the use of AQP within research systems, and discusses challenges and future directions. It is essential reading for anyone working with, or doing research on massive data.



99,00 €

92,52 € (zzgl. MwSt.)

Nicht mehr lieferbar

Artikelnummer: 9781601985163

Medium: Buch

ISBN: 978-1-60198-516-3

Verlag: Now Publishers

Erscheinungstermin: 31.12.2011

Sprache(n): Englisch

Auflage: 1. Auflage 2011

Serie: Foundations and Trends® in Databases

Produktform: Kartoniert

Seiten: 288

Format (B x H): 156 x 234 mm

