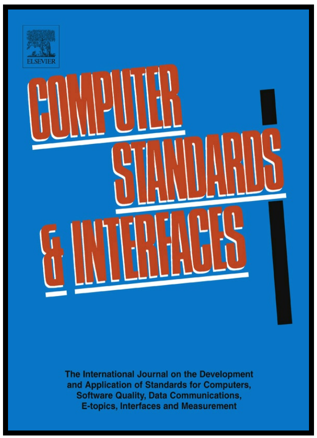
ELSEVIER
COMPUTER STANDARDS & INTERFACES
The International Journal on the Development and Application of Standards for Computers, Software Quality, Data Communications, E-topics, Interfaces and Measurement

www.elsevier.com

Big Data Normalization for Massively Parallel Processing Databases

Nikolay Golov, Lars Rönnbäck

PII: S0920-5489(17)30036-3
DOI: http://dx.doi.org/10.1016/j.csi.2017.01.009
Reference: CSI3194

To appear in: *Computer Standards & Interfaces*

Received date: 4 April 2016
Revised date: 17 January 2017
Accepted date: 20 January 2017

Cite this article as: Nikolay Golov and Lars Rönnbäck, Big Data Normalizatio for Massively Parallel Processing Databases, *Computer Standards & Interfaces* http://dx.doi.org/10.1016/j.csi.2017.01.009

Available online at www.sciencedirect.com
ScienceDirect
Computer Standards & Interfaces 00 (2017) 1–12
Computer Standards Interfaces
ELSEVIER

Big Data Normalization for Massively Parallel Processing Databases

Nikolay Golov

National Research University Higher School of Economics, Moscow, Russia

Lars Rönnbäck

Department of Computer Science, Stockholm University, Stockholm

Abstract

High performance querying and ad-hoc querying are commonly viewed as mutually exclusive goals in massively parallel processing databases. Also there is contradiction between ease of extending the data model and ease of analysis. Modern approach, called Data Lake, promises extreme ease of adding new data to a data model, while it is prone to eventually converting to Data Swamp - unstructured, ungoverned, and out of control Data Lake where due to a lack of process, standards and governance, data is hard to find, hard to use and is consumed out of context. This paper introduces a novel technique, highly normalized Big Data using Anchor modeling, that provides a very efficient way to store information and utilize resources, thereby providing ad-hoc querying with high performance for the first time in massively parallel processing databases. This technique is almost as convenient for expanding data model as a Data Lake, while it is internally protected from transforming to Data Swamp. A case study of how this approach is used for a Data Warehouse at Avito over three years time, with estimates for and results of real data experiments carried out in HP Vertica, an MPP RDBMS, are also presented. This paper is an extension of theses from The 34th International Conference on Conceptual Modeling (ER 2015) [1], it is complemented with numerical results about key operating areas of highly normalized big data warehouse, collected over several (1-3) years of commercial operation. Also, the limitations, imposed by using a single MPP database cluster, are described, and cluster fragmentation approach is proposed.

Keywords: Big Data, MPP, database, normalization, analytics, ad-hoc, querying, modeling, performance, Data Lake

1. Background

Big Data analytics is rapidly becoming a commonplace task for many companies. For example, banks, telecommunication companies, and big web companies, such as Google, Facebook, and Twitter produce large amounts of data. Nowadays business users also know how to monetize such data. For example, various predictive marketing techniques can transform data about customer behavior into great monetary worth. The main issue, however, remains to be implementations and platforms fast enough to load, store and execute ad-hoc analytical queries over Big Data [2].

Email addresses: ngolov@avito.ru (Nikolay Golov), lars.ronnback@anchormodeling.com (Lars Rönnbäck)
URL: http://www.avito.ru (Nikolay Golov), http://www.anchormodeling.com (Lars Rönnbäck)