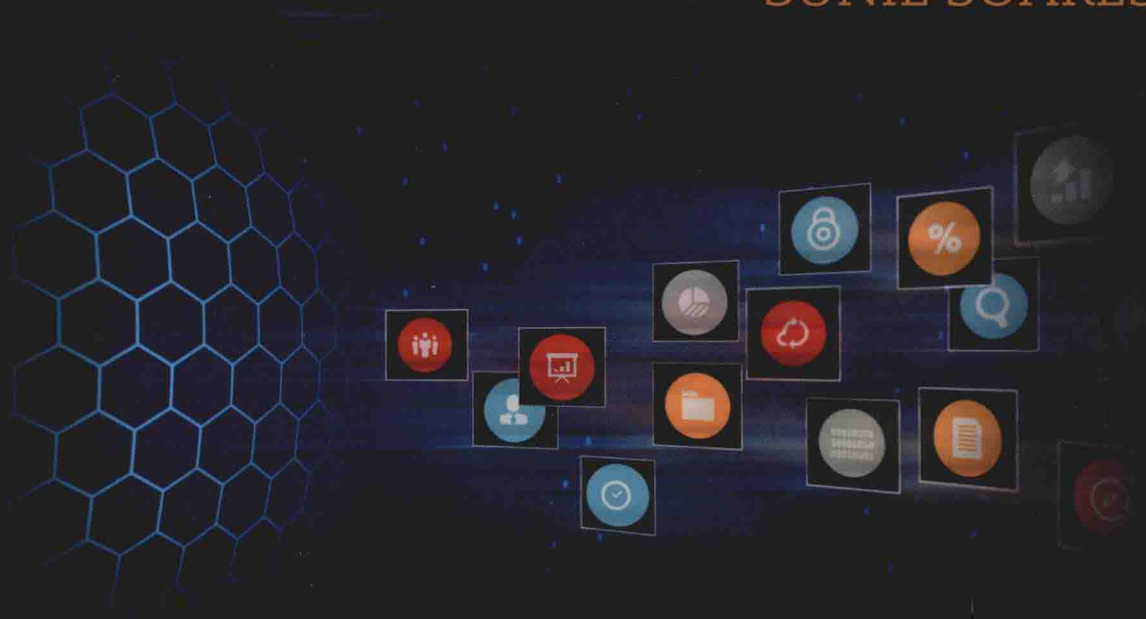
"... a comprehensive, detailed guide ..."

JOHN R. TALBURT, PHD, UNIVERSITY OF ARKANSAS AT LITTLE ROCK

# Data Governance Tools

## *Evaluation Criteria, Big Data Governance, and Alignment with Enterprise Data Management*

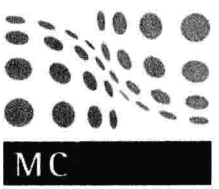
SUNIL SOARES



# Data Governance Tools

Sunil Soares






MC Press Online, LLC
Boise, ID 83703 USA

**Data Governance Tools: Evaluation Criteria, Big Data Governance, and Alignment with Enterprise Data Management**
*Sunil Soares*

First Edition


**MC Press Online, LLC** 3695 W. Quail Heights Court, Boise, ID 83703-3861 USA • (208) 629-7275
service@mcpressonline.com • www.mcpressonline.com • www.mc-store.com

ISBN: 978-1-58347-844-8 WB201410

Dedicated to my beautiful daughters, Maya and Lizzie.

Many thanks to my wife Helena, who came up with the idea for this book.

A big thanks to my parents Cecilia and Hubert for their prayers and guidance.

I also want to acknowledge the Information Asset team, including Jatin Bhoir, Michelle D'Sa, Royson Mendonca, Yanxin Shi, and Dorothy Xavier. The Enterprise Data Management lab is a critical success factor in our client engagements and in the development of this book.

# About the Author

**Sunil Soares** is the founder and managing partner of Information Asset, a consulting firm that specializes in data governance. Prior to this role, Sunil was director of information governance at IBM, where he worked with clients across six continents and multiple industries. Before joining IBM, Sunil consulted with major financial institutions at the Financial Services Strategy Consulting Practice of Booz Allen & Hamilton in New York.

Sunil's first book, *The IBM Data Governance Unified Process* (MC Press, 2010), details the almost 100 steps to implement a data governance program. This book has been used by several organizations as the blueprint for their data governance programs and has been translated into Chinese. Sunil's second book, *Selling Information Governance to the Business: Best Practices by Industry and Job Function* (MC Press, 2011), reviews the best practices to approach information governance by industry and function. His third book, *Big Data Governance (MC Press, 2012)*, addresses the specific issues associated with the governance of big data.

Sunil lives in New Jersey and holds an MBA in Finance and Marketing from the University of Chicago Booth School of Business.

# FOREWORD

*by Aditya Kongara*

Enterprise Data Management (EDM) over the past few years has quickly become an important discipline as organizations look to establish governance over their information assets. Effective data management needs the three pillars of people, process, and technology to be mature and well-functioning.

I have spent the majority of my career in large financial services organizations and working with Big Four consulting firms setting up data management and governance programs. In my opinion, the technology pillar of EDM is as important as the other two pillars.

Assume you are the data governance lead at a large bank that has to pass a data audit from the regulators. The bank's systems consist of hundreds of thousands of data elements spread over hundreds of databases and schemas. How do you demonstrate data lineage to the regulators without a metadata tool? Are you able to convince the Chief Information Security Officer that all instances of sensitive data have been discovered? Can you do that without a data discovery tool? Are your SQL queries robust and automated enough to produce data quality scorecards on a regular basis? For these reasons and others listed in the book, I feel that companies will increasingly have to rely on data management tools to automate various manual tasks.

I have known Sunil Soares for many years in a variety of job roles. I am excited by his knowledge and passion for data governance and for his thought leadership around tools. This book is a great read for any practitioner who wants to be successful in the data management and governance field.

*Aditya Kongara*
*Head of Enterprise Data Management*
*American Family Mutual Insurance Company*

# FOREWORD

*by John R. Talburt*

This book on data governance tools could not have come at a better time for the field of information quality. I say this having been in the most fortunate position to observe the explosive growth and evolution of information and data quality over the past three decades, from both a practitioner and academic perspective. Given this perspective, let me start by giving a bit of background that I think explains why this book is so timely.

Deeply rooted in practice, the emerging field of information quality had its genesis in the seemingly endless data cleaning efforts that were necessary to launch the data warehousing movement of the 1980s. From cleaning and correcting data, it started to mature, first embracing root cause analysis, then later fully adopting and incorporating the principles of TQM (Total Quality Management). Having embraced the concept of managing information as product, it continued to develop and mature. In its current incarnation, information quality goes far beyond just repairing things gone wrong, to having a seat at the table for information architecture planning and design, and now is an integral part of information policy and strategy in the role of data governance.

Like data warehousing, data governance is one of those new ideas that in retrospect seems so obvious. Why wouldn't any enterprise want to have a clear policy around and a shared understanding of its information assets? But like data warehousing, it has taken some time to "iron out the wrinkles" and make data governance really work. Now that we know that it does work, the competitive advantage imparted by a well-defined data governance program has elevated it to an essential part of corporate strategy.

Accepting data governance as essential is one thing, but making it work is another. In the early years of information quality, everyone had to develop their own tools to try and get the job done. It was not long before the demand for easier tools with more functionality created a market demand that was addressed by the many data quality tool

vendors we see today. Now we see a repeat of this cycle with data governance. Many vendors now offer various tools and suites of tools to help organizations implement data governance programs. However, one difference is that data governance programs are more diverse because the reasons for adopting them and their goals are often quite different.

This comes to the point of why this book is so timely and important. In one source, the reader can have an overview of the various categories of data governance tools and their key components. This book also gives a clear description of how and where these tools integrate into the data management strategy of the enterprise. Moreover, it is written by someone with extensive experience in data governance implementation, someone who has been there and knows how it works. This experience is reflected in the large amount of detail and concrete examples given in the book.

One really invaluable section of this book is the survey of data governance tools offered by the leading vendors. The overview will be a tremendous help to those still on the sidelines and getting ready to start a data governance program, as well as those who have started on their own, but now see the potential value in adopting a third-party system.

Another very helpful section is on big data governance tools. It contains a great discussion on the use of Hadoop MapReduce and NoSQL tools to gain insights into data. There are also sections explaining approaches to streaming computing and text analytics.

All in all, *Data Governance Tools* is a comprehensive, detailed guide to the landscape of data governance tools that will be valuable to everyone involved with enterprise data management, both from business and IT. I hope that everyone will take advantage of the wealth of information that it provides.

*John R. Talburt, PhD, IQCP*
*Director of the Information Quality Graduate Program*
*University of Arkansas at Little Rock*

# FOREWORD

*by Aaron Zornes*

While Sunil's prior books represented a Rosetta Stone for IT professionals to map their traditional IT experiences (MDM, RDM, data governance, etc.) to big data, at last we now have a "Domesday Book" to categorize and better understand the vast menagerie of solutions that comprise the data governance software market. There is quite a lot more beyond Microsoft Excel and SharePoint, and Sunil's "reference architecture" provides the foundational touchstone.

Given the synergy and codependence between MDM and data governance, Sunil's latest book is a must read for any MDM practitioner who is charged with establishing or upgrading the data governance processes inherently necessary for enterprise MDM or RDM programs. Among other benefits, it provides a much appreciated reference architecture and set of evaluation criteria, as well as examples illustrating the practical application of these tools.

In my consultancy practice and experience, MDM and RDM mandate the application of data governance (not just people and processes, but also software tools) to be effective and sustainable. Clearly, data governance for MDM is moving beyond simple stewardship to convergence of task management, workflow, policy management, and enforcement. Moreover, it is now time for MDM vendors to instantiate their data governance marketing claims and finally move from "passive-aggressive" mode to "proactive" data governance mode. The evaluation criteria provided in this book is proof that MDM vendors have recently begun to deliver (especially IBM, Informatica, Orchestra Networks, and SAP).

*Data Governance Tools* is the plenary source that can successfully tutor and guide you into becoming a "data governance professional." Moreover, it is a key asset that I'll be sharing with the 3,000+ annual attendees of my MDM & Data Governance Summit series.

*Aaron Zornes*
*Chief Research Officer, The MDM Institute*
*Conference Chairman, The MDM & Data Governance Summit*
*(London, New York City, San Francisco, Shanghai, Singapore, Sydney, Tokyo, Toronto)*

# PREFACE

Data governance is the formulation of policy to optimize, secure, and leverage information as an enterprise asset by aligning the objectives of multiple functions. Data governance programs have traditionally focused on people and process. Cost has historically been a key consideration because data governance programs have often started from scratch, with little to no funding. As a result, Microsoft Excel and SharePoint have been the tools of choice to document and share data governance artifacts. While the marginal cost of these tools is zero, they are often missing critical functionality. Meanwhile, vendors have matured their data governance offerings to the extent that organizations need to consider tools as a critical component of their data governance programs.

It is not always clear, however, what "data governance tools" really mean. In this book, I review a reference architecture for data governance software tools. I seek to define the category called "data governance," as well as lay out evaluation criteria for software tools, the vendor landscape, and the alignment with big data.

This book consists of the following sections:

1. *Introduction*

   The chapters in this section provide an introduction to data governance and the Enterprise Data Management (EDM) reference architecture.

2. *Categories of Data Governance Tools*

   These chapters discuss key data governance tasks that can be automated by tools for business glossaries, metadata management, data profiling, data quality management, master data management, reference data management, and information policy management.

3. *The Integration Between Enterprise Data Management and Data Governance Tools*

   This section is an overview of the integration points between EDM tools and data governance. EDM tools relate to data modeling, data integration, analytics and reporting, business process management, data security and privacy, and information lifecycle management.

4. *Big Data Governance Tools*

   The chapters in this section provide an overview of how data governance tools interact with big data technologies, including Hadoop, NoSQL, stream computing, and text analytics.

5. *Evaluation Criteria and the Vendor Landscape*

   This section is a review of the overall evaluation criteria for data governance tools. This section also provides an overview of key vendor platforms, including ASG, Collibra, Global IDs, IBM, Informatica, Orchestra Networks, SAP, and Talend.

This book is geared toward business users and is relatively nontechnical. Sample roles who might be interested in this book include the following:

- Chief Information Officer
- Chief Data Officer
- Data Governance Lead
- Business Intelligence Lead
- Data Warehousing Lead
- Enterprise Data Management Lead
- Chief Information Security Officer
- Chief Privacy Officer
- Chief Medical Information Officer

All the best, and happy reading.

# Contents