Data as a Service

A Framework for Providing Reusable Enterprise Data Services

PUSHPAK SARKAR

IEEE
IEEE PRESS

IEEE computer society

WILEY

Data as a Service

A Framework for Providing Reusable Enterprise Data Services

Pushpak Sarkar

IEEE computer society

IEEE
IEEE PRESS

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data is available.

ISBN: 978-1-119-04658-5

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Data as a Service

About IEEE Computer Society

IEEE Computer Society is the world's leading computing membership organization and the trusted information and career-development source for a global workforce of technology leaders including: professors, researchers, software engineers, IT professionals, employers, and students. The unmatched source for technology information, inspiration, and collaboration, the IEEE Computer Society is the source that computing professionals trust to provide high-quality, state-of-the-art information on an on-demand basis. The Computer Society provides a wide range of forums for top minds to come together, including technical conferences, publications, and a comprehensive digital library, unique training webinars, professional training, and the TechLeader Training Partner Program to help organizations increase their staff's technical knowledge and expertise, as well as the personalized information tool myComputer. To find out more about the community for technology leaders, visit http://www.computer.org.

IEEE/Wiley Partnership

The IEEE Computer Society and Wiley partnership allows the CS Press authored book program to produce a number of exciting new titles in areas of computer science, computing, and networking with a special focus on software engineering. IEEE Computer Society members continue to receive a 15% discount on these titles when purchased through Wiley or at wiley.com/ieeecs.

To submit questions about the program or send proposals, please contact Mary Hatcher, Editor, Wiley-IEEE Press: Email: mhatcher@wiley.com, Telephone: 201-748-6903, John Wiley & Sons, Inc., 111 River Street, MS 8-01, Hoboken, NJ 07030-5774.

Dedicated to my parents and family for making me believe that anything is possible if you dream big and work hard.

Guest Introduction

With the advent of social media and the Internet of Things (IoT), businesses are receiving a lot more data than they ever did in the past. The volume of data is increasing exponentially, the variety is increasing, and so is the velocity of its arrival. Companies who can analyze this data, derive insights and share their learnings across business lines within the company and with the ecosystem of partners externally in an effective manner to transform their businesses are invariably the ones who are going to win. This specific trend has been captured in Accenture's Technology Vision 2013 as "Data Velocity" and "Design for Analytics" and again in 2014 as "Data Supply Chain." Personally, as a Managing Director of Accenture, I have seen this trend resonate with our Fortune 500 clients across Accenture's five Operating Groups: Communications, Media & High Tech, Financial Services, Health and Public Services, Resources, and Products.

Given the need to consume data from heterogeneous sources, both internal and external to a company, hosted either in premises or in the cloud, and on the flip side, to make its own data available in exactly the same reuseable form for partners to consume, companies can no longer afford to keep data locked into silos of applications, nor can they treat it as a second class object when it comes to architecting its IT infrastructure. Data needs to be decoupled from applications so that the data generated by one application can be used effectively by a completely different set of applications, and the insights generated by analyzing the data within one business line of a company can be shared with other business lines in order to maximize the Return on Investment (RoI) on the data available to the company as a whole. I have seen this happen with a leading drugstore in the United States where sharing of data between the store's loyalty program and the sales department helped better targeting of products leading to significantly increased sales.

The most effective way of sharing the data and insights is to make data a first class object in the design of IT architecture and make it available as a service. Once exposed as a service, any application, whether internal or external to a company, can consume data in a seamless manner and use it creatively to make a tangible difference to business. In fact, there are several examples of completely new businesses created across industries from healthcare to insurance to automotive to real estate, fuelled by the sharing of data in the form of APIs by a company with its ecosystem of partners; and the huge impact created, in turn, by the ecosystem on the company's existing business due to the sharing of data, leading to mutual business benefits. For example, GM exposed their OnStar Application Programming Interface (API) to power a new business service via a start-up called RelayRides that enabled individuals to rent their personal cars, thereby disrupting the rental car business. We have seen the same

trend with Walgreens who is offering access to its data through a variety of APIs and Software Development Kits (SDKs) to fuel new businesses with its ecosystem of partners.

Similarly, there is a plethora of examples of how companies have successfully exploited the synergy across their business lines by sharing data and insights within the company, leading to higher efficiency and creation of new revenue streams. The previously cited example of the leading drug store sharing data between the customer loyalty program and the sales department fits this category. Thus, data sharing internally as well as externally has proven to be transformational for businesses across industries.

With business transformations happening across the globe based on the availability of huge amount of data and its analysis, this book on Data as a Service, providing a comprehensive view into the world of Data Engineering and its implications on business, is a must read for every IT professional and business leader.

SANJOY PAUL, PhD
Managing Director – Accenture Technology Labs

Guest Introduction

When I wrote my first book, *Data Crush*, I attempted to capture the ways in which the technical innovations of mobility, Cloud computing, and big data were leading to entirely new social and business phenomena. Several of the impacts that these new technologies have had on our world are driving the demand for Data as a Service, hence I was elated when Pushpak asked me to introduce his work, that you now hold in your hands. There are three social forces that are making Data as a Service a new business imperative, and they are quantification, appification, and cloudification. Let us look at each in turn.

Quantification is the growing trend of measuring absolutely everything, across all aspects of business. I recently met the CIO of a commercial property management company that is spending over $1 billion to quantify his business. Over a two year period, his company will connect to the Internet every lightbulb in every one of their buildings. When I asked him what data he hoped to learn from these connected bulbs his response was, "I have no idea, but what I do know is that if I don't have the data there's nothing to analyze." You will likely see this sort of pervasive data collection occurring throughout every process in every organization over the coming decade.

Appification is our growing expectation of instant gratification, at little or no cost, regardless of how irrational this expectation may be. Indeed, we are becoming so appified that we expect our needs to be met predictively. Delivering on this expectation demands that organizations not only analyze data, they must do so perpetually and rapidly. The notion that business insights only come from a Research and Development department, or from IT is outdated, because there simply is not time to push analytics to a central organization. Rather, appification means that organizations must collect, digest, and act upon data as close to the customer as possible, in both time and space.

Finally, Cloudification is the notion that the paradigm of building and owning the assets of your business has become obsolete. Cloud initially entered the world of applications with Software as a Service, and is rapidly spreading to all other aspects of business operations. More and more, companies will simply aggregate third-party services in order to meet customer needs, rather than produce those outputs themselves. Data management and analysis will follow this trend, leading to Data as a Service being the standard mode of putting data to work in organizations.

Acting upon these societal forces is challenging. Much of this mode of operating runs counter to how we have run IT for half of a century. Nonetheless, it is imperative that organizations embrace Data as a Service if they hope to remain relevant in our accelerating world. This book provides a practical, implementable approach to

reaching this goal. I trust that you will find Pushpak's guidance valuable as you work to meet the new expectations of an ever-more-competitive world.

CHRISTOPHER SURDAK
Engineer, ex-Rocket Scientist, Juris Doctor,
Technology Evangelist and author of "Data Crush,"
GetAbstract's International Book of the Year for 2014

Preface (Includes the Reader's Guide)

Typically, once every couple of decades a disruptive new technology emerges that fundamentally changes the business landscape. Innovative, high tech products that often start a trend come to the mainstream market with such rapidity that they transform the existing way of doing business. These trends also create a new market that eventually disrupts the existing market and related network, often displacing the earlier technology.

In most cases, organizations that understand underlying competitive dynamics of innovation and who adapt to these disruptive trends, win. Today such fundamental shifts take place in the world of data and analytics daily, and they are changing the global business landscape significantly.

If one closely observes the global marketplace, it is safe to say that many businesses are trying to harness an unprecedentedly large amount of data to derive new insights that support their competitive analyses. A huge amount of data that is gathered from diverse channels (e.g., social media, clickstream analysis) need to be translated by businesses to enable concrete actions. Organizations that understand the competitive dynamics at play and those that can then predictively analyze that data will win, whereas those that fail to recognize this challenge and respond to it will become extinct.

While data has always been considered an essential part of IT infrastructure across most organizations to support their business operations, today it is recognized as the key commodity upon which an enterprise runs its business and day-to-day operations. A complete paradigm shift has occurred in which data is increasingly recognized as an asset that can be commercially sold as a service, in and of itself.

Based on the author's first-hand experience and expertise, this book offers a proven framework for sharing core enterprise data using reusable data services. The book covers how organizations can generate business revenues by providing Data as a Service to their clients for fee-based subscriptions. The book goes on to explain in detail how to acquire and distribute data across heterogeneous platforms effectively using enterprise SOA principles, industry data standards, and leveraging new technologies such as data virtualization, cloud, and big data stream computing. The book also offers the following:

- Presents a comprehensive approach for introducing Data as a Service (DaaS) in any organization for the first time.
- Recommended best practices and industry standards for sharing master, reference, and big data with data consumers.

- Commercialization aspects of Data as a Service and its potential for generating revenues.
- Covers real-world applications of DaaS such as big Data as a Service.
- Real-life case studies on various innovative architecture blueprints and related patterns.

The topics covered in this book are wide ranging, starting with a presentation on the need for providing DaaS and the technical challenges involved in making that transformation. Some of the areas of the book that may particularly appeal to readers include:

- How DaaS can become a strategic enabler for sharing data with customers on company products they are interested in purchasing, browsing online, or viewing on social media.
- How the DaaS framework can help many organizations recognize monetizable intent and dependency of their customers on accessing their data while buying their company products.
- How enhanced on-demand data services can lead to potential clients by organizations that plan on mining customer, social media, and online conversations over a big data platform, using sophisticated predictive algorithms and data analytics tools.
- How to adopt best practices for successfully deploying reusable data services in your organization along with a reference architecture comprising common sets of data standards, guidelines, and processes.

Covering so much ground—from canonical modeling to data governance and XML based services—can be challenging for some readers, so the book offers a roadmap to help guide you through it.

The Reader's Guide

The Reader's Guide is provided to help readers determine who should read the book and why they need to read the book. A summary of each chapter to explain the step-by-step approach required for the successful introduction of DaaS in any organization is also provided.

The successful adoption of DaaS in any organization is based on three fundamental areas—architecture, adopting organizational processes, and ensuring the appropriate technology components are deployed. However, this should be based on real-world experiences and lessons learned from prior IT/DaaS implementations. This is one of the reasons this book includes case studies in several chapters.

The next section will guide readers on how best to use the book by sharing details of every chapter. It will also help guide readers to determine the best approach to use the DaaS framework in their current IT landscape within their organization. Figures 1 and 2 illustrate key topics in the book along with the suggested roadmap.

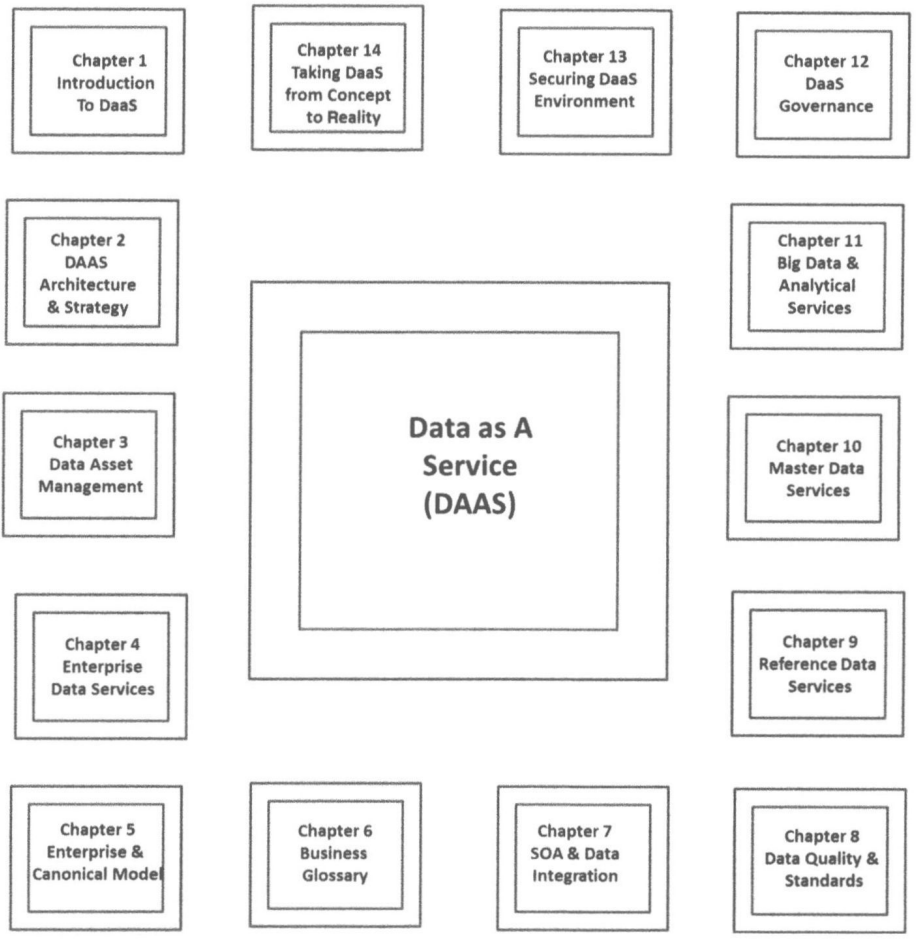

Figure 1 Key topics covered in the book by chapter

PART 1: Overview of Fundamental Concepts Includes Chapters 1 to 3

The introductory section of the book introduces you to Data as a Service (DaaS). It also provides readers with a clear overview on how an organization can deliver on the promise of providing DaaS to its business stakeholders and end customers.

Chapter 1: "Introduction to DaaS" provides a high-level overview on the core concepts of the DaaS framework. It also explores commercialization aspects of Data as a Service, its immense potential for generating revenues for most organizations, as well as some of its common limitations. It describes the details of service delivery management while suggesting necessary key steps for preparing the blueprint for enterprise data services in your organization.

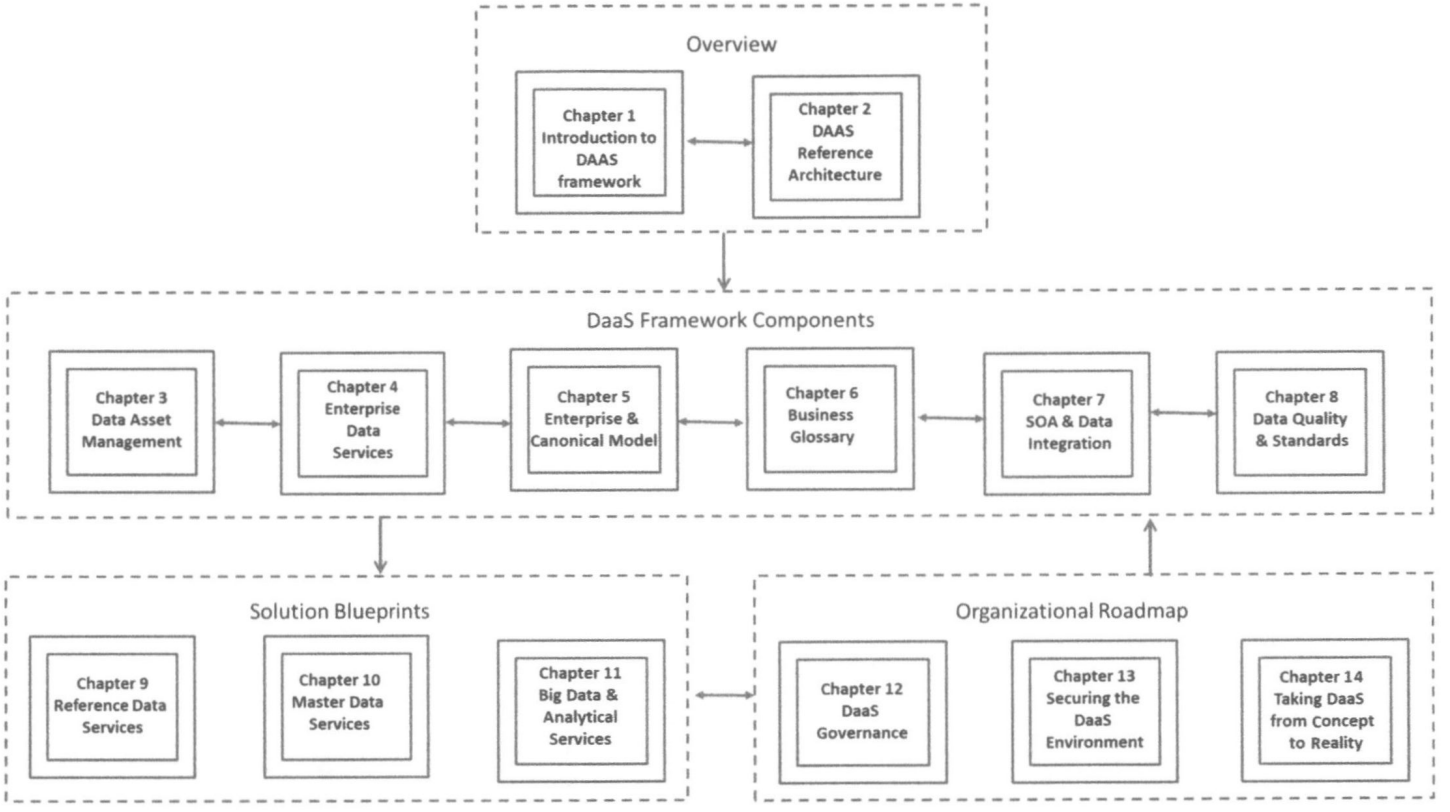

Figure 2 Roadmap the book's different chapters

Chapter 2: "DaaS Strategy and Reference Architecture" provides an overview of DaaS reference architecture along with the key components that make up the DaaS framework. It also explains the long-term significance of formally creating an enterprise data strategy in an organization that formulates a long-term roadmap to deliver Data as a Service (DaaS).

Chapter 3: "Data Asset Management" explores the significance of enterprise data and the foundational role it plays to make enterprise data services successful in any organization. It explains the underlying principles of data asset management and why companies need to treat data as a corporate asset. It also examines the various major types of enterprise data and contrasts their major features.

PART 2: DaaS Architecture Framework and Components Includes Chapters 4 to 8

This section of the book focuses on the architecture framework and components required to deploy DaaS in your organization. It also describes in detail common patterns, standards, and processes that can help shape the DaaS Reference Architecture. This section also provides readers with a high-level overview on best practices from a few related disciplines (e.g., EIM, EA, SOA, data services) to make DaaS a scalable data delivery mechanism for organizations.

Chapter 4: "Enterprise Data Services" describes the core concepts about enterprise data services as a fundamental component of the DaaS framework. It illustrates with examples how several organizations have successfully developed a set of standardized service interfaces (termed EDS) to enable data sharing with their various stakeholders (customers, vendors, regulatory agencies, government, etc.).

Chapter 5: "Enterprise and Canonical Modeling" explains the significance of enterprise and canonical modeling and its foundational role to promote consistent and reliable data exchange across disparate systems spread out over the organization. It also explains the significance of the enterprise data model (EDM) as the foundational component required for building a robust and mature set of data structures that can be reused across the entire organization.

Chapter 6: "Business Glossary for DaaS" environment provides a detailed overview of the underlying reasons why organizations need to develop a standardized business glossary for data services published for user consumption. Storing glossary terms in a shared metadata repository across the organization will improve the overall productivity of both the businesses and the external subscribers to enterprise data services (EDS).

Chapter 7: "SOA and Data Integration" provides a high-level overview on key data acquisition and integration patterns with service-oriented architecture (SOA) as the underlying foundation. It also covers a few technologies, e.g., data virtualization, stream computing for big data, data federation, which can be leveraged by the DaaS framework to publish data services with enhanced efficiency, performance, and a scalable architecture.

Chapter 8: "Data Quality and Standards" provides details on how to ensure that the quality of data published by enterprise data services is suitable and fit for public

consumption. It explains the significance of data standards for the success of any DaaS program. The chapter also discusses the role of data profiling as a foundational process for the success of any DaaS quality program. Finally, it looks at some of the major data profiling and quality measures that are critical for implementing a DaaS project in real life.

PART 3: DaaS Solution Blueprints Includes Chapters 9 to 11

This section of the book provides a number of important solution blueprints where the DaaS framework can benefit organizations across several industries. Solution blueprints of data services can be very useful for readers as they can help explain the relationship between the architecture patterns explained earlier to the specific business requirements of organizations to exchange various types of enterprise data. Solution blueprints are based on the DaaS reference architecture also explained in the earlier sections of the book. Finally, this section covers a variety of real-life case studies on how organizations have successfully utilized the DaaS framework and its architectural patterns to improve their business efficiency over the long term.

Chapter 9: "Reference Data Services" presents a detailed overview on how DaaS can be deployed successfully in organizations for disseminating shared reference data to downstream data subscribers and consumers. It also presents real-life case studies on reference data services from the financial and healthcare sectors.

Chapter 10: "Master Data Services" provides a detailed architectural pattern for designing and developing Master Data Services (MDS) that can be reused across an enterprise by using common design components and standards. It also evaluates how MDS can be utilized by organizations as an effective alternative to the existing styles of MDM implementation without physically consolidating master data in a single hub. A detailed case study on a MDS implementation at a large financial institution is presented.

Chapter 11: "Big Data and Analytical Services" explains how big data analytics users can leverage data services to access data they need for advanced analytics and take decisions in real time. This chapter includes several case studies presented from organizations that have successfully implemented big data and mobile-based analytics services, leveraging the DaaS framework. It provides a detailed solution blueprint for designing and developing big Data as a Service that can be reused across the enterprise by using the design components and standards proposed under the DaaS framework.

PART 4: Ensuring Organizational Success Includes Chapters 12 to 14

Introducing DaaS is uncharted territory for many organizations. Not all businesses are likely to face the same urgency for providing Data as a Service to their consumer, nor will they encounter the same challenges. An organizational roadmap has been