

Building A Scalable Data Warehouse With Data Vault 2 0

With this textbook, Vaisman and Zimányi deliver excellent coverage of data warehousing and business intelligence technologies ranging from the most basic principles to recent findings and applications. To this end, their work is structured into three parts. Part I describes “Fundamental Concepts” including conceptual and logical data warehouse design, as well as querying using MDX, DAX and SQL/OLAP. This part also covers data analytics using Power BI and Analysis Services. Part II details “Implementation and Deployment,” including physical design, ETL and data warehouse design methodologies. Part III covers “Advanced Topics” and it is almost completely new in this second edition. This part includes chapters with an in-depth coverage of temporal, spatial, and mobility data warehousing. Graph data warehouses are also covered in detail using Neo4j. The last chapter extensively studies big data management and the usage of Hadoop, Spark, distributed, in-memory, columnar, NoSQL and NewSQL database systems, and data lakes in the context of analytical data processing. As a key characteristic of the book, most of the topics are presented and illustrated using application tools. Specifically, a case study based on the well-known Northwind database illustrates how the concepts presented in the book can be implemented using Microsoft Analysis Services and Power BI. All chapters have been revised and updated to the latest versions of the software tools used. KPIs and Dashboards are now also developed using DAX and Power BI, and the chapter on ETL has been expanded with the implementation of ETL processes in PostgreSQL. Review questions and exercises complement each chapter to support comprehensive student learning. Supplemental material to assist instructors using this book as a course text is available online and includes electronic versions of the figures, solutions to all exercises, and a set of slides accompanying each chapter. Overall, students, practitioners and researchers alike will find this book the most comprehensive reference work on data warehouses, with key topics described in a clear and educational style. “I can only invite you to dive into the contents of the book, feeling certain that once you have completed its reading (or maybe, targeted parts of it), you will join me in expressing our gratitude to Alejandro and Esteban, for providing such a comprehensive textbook for the field of data warehousing in the first place, and for keeping it up to date with the recent developments, in this current second edition.” From the foreword by Panos Vassiliadis, University of Ioannina, Greece.

A cutting-edge response to Ralph Kimball's challenge to the data warehouse community that answers some tough questions about the effectiveness of the relational approach to data warehousing
Written by one of the best-known exponents of the Bill Inmon approach to data warehousing
Addresses head-on the tough issues raised by Kimball and explains how to choose the best modeling technique for solving common data warehouse design problems
Weighs the pros and cons of relational vs. dimensional modeling techniques
Focuses on tough modeling problems, including creating and maintaining keys and modeling calendars, hierarchies, transactions, and data quality

Get started with distributed computing using PySpark, a single unified framework to solve end-to-end data analytics at scale
Key Features
Discover how to convert huge amounts of raw data into meaningful and actionable insights
Use Spark's unified analytics engine for end-to-end analytics, from data preparation to predictive analytics
Perform data ingestion, cleansing, and integration for ML, data analytics, and data visualization
Book Description
Apache Spark is a unified data analytics engine designed to process huge volumes of data quickly and efficiently. PySpark is Apache Spark's Python language API, which offers Python developers an easy-to-use scalable data analytics framework. Essential PySpark for Scalable Data Analytics starts by exploring the distributed computing paradigm and provides a high-level overview of Apache Spark. You'll begin your analytics journey with the data engineering process, learning how to perform data ingestion, cleansing, and integration at scale. This book helps you build real-time analytics pipelines that help you gain insights faster. You'll then discover methods for building cloud-based data lakes, and explore Delta Lake, which brings reliability to data lakes. The book also covers Data Lakehouse, an emerging paradigm, which combines the structure and performance of a data warehouse with the scalability of cloud-based data lakes. Later, you'll perform scalable data science and machine learning tasks using PySpark, such as data preparation, feature engineering, and model training and productionization. Finally, you'll learn ways to scale out standard Python ML libraries along with a new pandas API on top of PySpark called Koalas. By the end of this PySpark book, you'll be able to harness the power of PySpark to solve business problems. What you will learn
Understand the role of distributed computing in the world of big data
Gain an appreciation for Apache Spark as the de facto go-to for big data processing
Scale out your data analytics process using Apache Spark
Build data pipelines using data lakes, and perform data visualization with PySpark and Spark SQL
Leverage the cloud to build truly scalable and real-time data analytics applications
Explore the applications of data science and scalable machine learning with PySpark
Integrate your clean and curated data with BI and SQL analysis tools
Who this book is for
This book is for practicing data engineers, data scientists, data analysts, and data enthusiasts who are already using data analytics to explore distributed and scalable data analytics. Basic to intermediate knowledge of the disciplines of data engineering, data science, and SQL analytics is expected. General proficiency in using any programming language, especially Python, and working knowledge of performing data analytics using frameworks such as pandas and SQL will help you to get the most out of this book.

Agile Data Warehouse Design is a step-by-step guide for capturing data warehousing/business intelligence (DW/BI) requirements and turning them into high performance dimensional models in the most direct way: by modelstorming (data modeling] brainstorming) with BI stakeholders. This book describes BEAM, an agile approach to dimensional modeling, for improving communication between data warehouse designers, BI stakeholders and the whole DW/BI development team. BEAM provides tools and techniques that will encourage DW/BI designers and developers to move away from their keyboards and entity relationship based tools and model interactively with their colleagues. The result is everyone thinks dimensionally from the outset! Developers understand how to efficiently implement dimensional modeling solutions. Business stakeholders feel ownership of the data warehouse they have created, and can already imagine how they will use it to answer their business questions. Within this book, you will learn: Agile dimensional modeling using Business Event Analysis & Modeling (BEAM)
Modelstorming: data modeling that is quicker, more inclusive, more productive, and frankly more fun!
Telling dimensional data stories using the 7Ws (who, what, when, where, how many, why and how)
Modeling by example not abstraction; using data story themes, not crow's feet, to describe detail
Storyboarding the data warehouse to discover conformed dimensions and plan iterative development
Visual modeling: sketching timelines, charts and grids to model complex process measurement - simply
Agile design documentation: enhancing star schemas with BEAM dimensional shorthand notation
Solving difficult DW/BI performance and usability problems with proven dimensional design patterns
Lawrence Corr is a data warehouse designer and educator. As Principal of DecisionOne Consulting, he helps clients to review and simplify their data warehouse designs, and advises vendors on visual data modeling techniques. He regularly teaches agile dimensional modeling courses worldwide and has taught dimensional DW/BI skills to thousands of students. Jim Stagnitto is a data warehouse and master data management architect specializing in the healthcare, financial services, and information service industries. He is the founder of the data warehousing and data mining consulting firm Llumino.

An Introduction to Agile Data Engineering Using Data Vault 2. 0

Super Charge Your Data Warehouse

Data Warehousing Fundamentals

Build modern data warehouses on Microsoft Azure

A Primer for the Data Scientist

Recipes for building modern data warehousing solutions

Building a Scalable Data Warehouse with Data Vault 2.0

When strange animals land in Red's yard, he and his friend Slim agree to keep it to themselves. The creatures are unlike any that they've ever seen before, and are - to them - animals that would make a fortune putting on a show at the circus. All the while, their fathers are fretting over the arrival of interstellar diplomats, on whose trade their civilisation may have to rely on to survive. Despite the urgency of the mission, however, neither hide nor hair of them have been seen...

Develop modern solutions with Snowflake's unique architecture and integration capabilities; process bulk and real-time data into a data lake; and leverage time travel, cloning, and data-sharing features to optimize data operations

Key Features

- Build and scale modern data solutions using the all-in-one Snowflake platform
- Perform advanced cloud analytics for implementing big data and data science solutions
- Make quicker and better-informed business decisions by uncovering key insights from your data

Book Description

Snowflake is a unique cloud-based data warehousing platform built from scratch to perform data management on the cloud. This book introduces you to Snowflake's unique architecture, which places it at the forefront of cloud data warehouses. You'll explore the compute model available with Snowflake, and find out how Snowflake allows extensive scaling through the virtual warehouses. You will then learn how to configure a virtual warehouse for optimizing cost and performance. Moving on, you'll get to grips with the data ecosystem and discover how Snowflake integrates with other technologies for staging and loading data. As you progress through the chapters, you will leverage Snowflake's capabilities to process a series of SQL statements using tasks to build data pipelines and find out how you can create modern data solutions and pipelines designed to provide high performance and scalability. You will also get to grips with creating role hierarchies, adding custom roles, and setting default roles for users before covering advanced topics such as data sharing, cloning, and performance optimization. By the end of this Snowflake book, you will be well-versed in Snowflake's architecture for building modern analytical solutions and understand best practices for solving commonly faced problems using practical recipes. What you will learn

Get to grips with data warehousing techniques aligned with Snowflake's cloud architecture

- Broaden your skills as a data warehouse designer to cover the Snowflake ecosystem
- Transfer skills from on-premise data warehousing to the Snowflake cloud analytics platform
- Optimize performance and costs associated with a Snowflake solution
- Stage data on object stores and load it into Snowflake
- Secure data and share it efficiently for access
- Manage transactions and extend Snowflake using stored procedures
- Extend cloud data applications using Spark Connector

Who this book is for

This book is for data warehouse developers, data analysts, database administrators, and anyone involved in designing, implementing, and optimizing a Snowflake data warehouse. Knowledge of data warehousing and database and cloud concepts will be useful. Basic familiarity with Snowflake is beneficial, but not necessary.

DW 2.0: The Architecture for the Next Generation of Data Warehousing is the first book on the new generation of data warehouse architecture, DW 2.0, by the father of the data warehouse. The book describes the future of data warehousing that is technologically possible today, at both an architectural level and technology level. The perspective of the book is from the top down: looking at the overall architecture and then delving into the issues underlying the components. This allows people who are building or using a data warehouse to see what lies ahead and determine what new technology to buy, how to plan extensions to the data warehouse, what can be salvaged from the current system, and how to justify the expense at the most practical level. This book gives experienced data warehouse professionals everything they need in order to implement the new generation DW 2.0. It is designed for professionals in the IT organization, including data architects, DBAs, systems design and development professionals, as well as data warehouse and knowledge management professionals.

- * First book on the new generation of data warehouse architecture, DW 2.0.
- * Written by the "father of the data warehouse", Bill Inmon, a columnist and newsletter editor of The Bill Inmon Channel on the Business Intelligence Network.
- * Long overdue comprehensive coverage of the implementation of technology and

tools that enable the new generation of the DW: metadata, temporal data, ETL, unstructured data, and data quality control. Leverage the power of Microsoft Azure Data Factory v2 to build hybrid data solutions Key Features Combine the power of Azure Data Factory v2 and SQL Server Integration Services Design and enhance performance and scalability of a modern ETL hybrid solution Interact with the loaded data in data warehouse and data lake using Power BI Book Description ETL is one of the essential techniques in data processing. Given data is everywhere, ETL will always be the vital process to handle data from different sources. Hands-On Data Warehousing with Azure Data Factory starts with the basic concepts of data warehousing and ETL process. You will learn how Azure Data Factory and SSIS can be used to understand the key components of an ETL solution. You will go through different services offered by Azure that can be used by ADF and SSIS, such as Azure Data Lake Analytics, Machine Learning and Databrick's Spark with the help of practical examples. You will explore how to design and implement ETL hybrid solutions using different integration services with a step-by-step approach. Once you get to grips with all this, you will use Power BI to interact with data coming from different sources in order to reveal valuable insights. By the end of this book, you will not only learn how to build your own ETL solutions but also address the key challenges that are faced while building them. What you will learn Understand the key components of an ETL solution using Azure Data Factory and Integration Services Design the architecture of a modern ETL hybrid solution Implement ETL solutions for both on-premises and Azure data Improve the performance and scalability of your ETL solution Gain thorough knowledge of new capabilities and features added to Azure Data Factory and Integration Services Who this book is for This book is for you if you are a software professional who develops and implements ETL solutions using Microsoft SQL Server or Azure cloud. It will be an added advantage if you are a software engineer, DW/ETL architect, or ETL developer, and know how to create a new ETL implementation or enhance an existing one with ADF or SSIS.

A Pragmatic Guide on Building a Data Vault

Handbook of Research on Big Data Storage and Visualization Techniques

A Guide for Solution Architects and Project Leaders

The Kimball Group Reader

DW 2.0: The Architecture for the Next Generation of Data Warehousing

The Data Warehouse Toolkit

Delivering the Promise of Big Data and Data Science

Summary Big Data teaches you to build big data systems using an architecture that takes advantage of clustered hardware along with new tools designed specifically to capture and analyze web-scale data. It describes a scalable, easy-to-understand approach to big data systems that can be built and run by a small team. Following a realistic example, this book guides readers through the theory of big data systems, how to implement them in practice, and how to deploy and operate them once they're built. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the Book Web-scale applications like social networks, real-time analytics, or e-commerce sites deal with a lot of data, whose volume and velocity exceed the limits of traditional database systems. These applications require architectures built around clusters of machines to store and process data of any size, or speed. Fortunately, scale and simplicity are not mutually exclusive. Big Data teaches you to build big data systems using an architecture designed specifically to capture and analyze web-scale data. This book presents the Lambda Architecture, a scalable, easy-to-understand approach that can be built and run by a small team. You'll explore the theory of big data systems and how to implement them in practice. In addition to discovering a general framework for processing big data, you'll learn specific technologies like Hadoop, Storm, and NoSQL databases. This book requires no previous exposure to large-scale data analysis or NoSQL tools. Familiarity with traditional databases is helpful. What's Inside Introduction to big data systems Real-time processing of web-scale data Tools like Hadoop, Cassandra, and Storm Extensions to traditional database skills About the Authors Nathan Marz is the creator of Apache Storm and the originator of the Lambda Architecture for big data systems. James Warren is an analytics architect with a background in machine learning and scientific computing. Table of Contents A new paradigm for Big Data PART 1 BATCH LAYER Data model for Big Data Data model for Big Data: Illustration Data storage on the batch layer Data storage on the batch layer: Illustration Batch layer Batch layer: Illustration An example batch layer: Architecture and algorithms An example batch layer: Implementation PART 2 SERVING LAYER Serving layer Serving layer: Illustration PART 3 SPEED LAYER Realtime views Realtime views: Illustration Queuing and stream processing Queuing and stream processing: Illustration Micro-batch stream processing Micro-batch stream processing: Illustration Lambda Architecture in depth

Data mining of massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity and national intelligence. Collections of documents, images, videos, and networks are being thought of not merely as bit strings to be stored, indexed, and retrieved, but as potential sources of discovery and knowledge, requiring sophisticated analysis techniques that go far beyond classical indexing and

keyword counting, aiming to find relational and semantic interpretations of the phenomena underlying the data. Frontiers in Massive Data Analysis examines the frontier of analyzing massive amounts of data, whether in a static database or streaming through a system. Data at that scale--terabytes and petabytes--is increasingly common in science (e.g., particle physics, remote sensing, genomics), Internet commerce, business analytics, national security, communications, and elsewhere. The tools that work to infer knowledge from data at smaller scales do not necessarily work, or work well, at such massive scale. New tools, skills, and approaches are necessary, and this report identifies many of them, plus promising research directions to explore. Frontiers in Massive Data Analysis discusses pitfalls in trying to infer knowledge from massive data, and it characterizes seven major classes of computation that are common in the analysis of massive data. Overall, this report illustrates the cross-disciplinary knowledge--from computer science, statistics, machine learning, and application disciplines--that must be brought to bear to make useful inferences from massive data. The data lake is a daring new approach for harnessing the power of big data technology and providing convenient self-service capabilities. But is it right for your company? This book is based on discussions with practitioners and executives from more than a hundred organizations, ranging from data-driven companies such as Google, LinkedIn, and Facebook, to governments and traditional corporate enterprises. You'll learn what a data lake is, why enterprises need one, and how to build one successfully with the best practices in this book. Alex Gorelik, CTO and founder of Waterline Data, explains why old systems and processes can no longer support data needs in the enterprise. Then, in a collection of essays about data lake implementation, you'll examine data lake initiatives, analytic projects, experiences, and best practices from data experts working in various industries. Get a succinct introduction to data warehousing, big data, and data science Learn various paths enterprises take to build a data lake Explore how to build a self-service model and best practices for providing analysts access to the data Use different methods for architecting your data lake Discover ways to implement a data lake from experts in different industries

The digital age has presented an exponential growth in the amount of data available to individuals looking to draw conclusions based on given or collected information across industries. Challenges associated with the analysis, security, sharing, storage, and visualization of large and complex data sets continue to plague data scientists and analysts alike as traditional data processing applications struggle to adequately manage big data. The Handbook of Research on Big Data Storage and Visualization Techniques is a critical scholarly resource that explores big data analytics and technologies and their role in developing a broad understanding of issues pertaining to the use of big data in multidisciplinary fields. Featuring coverage on a broad range of topics, such as architecture patterns, programing systems, and computational energy, this publication is geared towards professionals, researchers, and students seeking current research and application topics on the subject.

A Step-by-Step Guide to Modern Cloud Analytics

Better Data Modeling

Invaluable Data Modeling Rules to Implement Your Data Vault

Techniques for building modern cloud data warehousing solutions

The New Intimacy

Guided Steps to Data Vault Success Through Building Business-Centered Models

Collaborative Dimensional Modeling, from Whiteboard to Star Schema

You want the rigor of good data architecture at the speed of agile? Then this is the missing link - your step-by-step guide to Data Vault success. Success with a Data Vault starts with the business and ends with the business. Sure, there's some technical stuff in the middle, and it is absolutely essential - but it's not sufficient on its own. This book will help you shape the business perspective, and weave it into the more technical aspects of Data Vault modeling. You can read the foundational books and go on courses, but one massive risk still remains. Dan Linstedt, the founder of the Data Vault, very clearly directs those building a Data Vault to base its design on an "enterprise ontology". And Hans Hultgren similarly stresses the importance of the business concepts model. So it's important. We get that. But: What on earth is an enterprise ontology/business concept model, 'cause I won't know if I've got one if I don't know what I'm looking for? If I can't find one, how do I get my hands on such a thing? Even if I have one of these wonderful things, how do I apply it to get the sort of Data Vault that's recommended? It's actually not as hard as some would fear to answer all of these questions, and it's certainly worth the effort. This book just might save you a world of pain. It's a supplement to other material on Data Vault modeling, but it's the vital missing link to finding simplicity for Data Vault success.

In Designing Cloud Data Platforms, Danil Zburivsky and Lynda Partner reveal a six-layer approach that increases flexibility and reduces costs. Discover patterns for ingesting data from a variety of sources, then learn to harness pre-built services provided by cloud vendors. Summary Centralized data warehouses, the long-time defacto standard for housing data for analytics, are rapidly giving way to multi-faceted cloud data platforms. Companies that embrace modern cloud data platforms benefit from an integrated view of their business using all of their data and can take advantage of advanced analytic practices to drive predictions and as yet unimagined data services. Designing Cloud Data Platforms is a hands-on guide to envisioning and designing a modern scalable data platform that takes full advantage of the flexibility of the cloud. As you read, you'll learn the core components of a cloud data platform design, along with the role of key technologies like Spark and Kafka Streams. You'll also explore setting up processes to manage cloud-based data, keep it secure, and using advanced analytic and BI tools to analyze it. Purchase of the print book includes a free eBook in PDF, Kindle, and ePub formats from Manning Publications. About the technology Well-designed pipelines, storage systems, and APIs eliminate the complicated scaling and maintenance required with on-prem data centers. Once you learn the patterns for designing cloud data platforms, you'll maximize performance no matter which cloud vendor you use. About the book In Designing Cloud Data Platforms, Danil Zburivsky and Lynda Partner reveal a six-layer approach that increases flexibility and reduces costs. Discover patterns for ingesting data from a variety of sources, then learn to harness pre-built services provided by cloud vendors. What's inside Best practices for structured and unstructured data sets Cloud-ready machine learning tools Metadata and real-time analytics Defensive architecture, access, and security About the reader For data professionals familiar with the basics of cloud computing, and Hadoop or Spark. About the author Danil Zburivsky has over 10 years of experience designing and supporting large-scale data infrastructure for enterprises across the globe. Lynda Partner is the VP of Analytics-as-a-Service at Pythian, and has been on the business side of data for over 20 years. Table of Contents 1 Introducing the data platform 2 Why a data platform and not just a data warehouse 3 Getting bigger and leveraging the Big

3: Amazon, Microsoft Azure, and Google 4 Getting data into the platform 5 Organizing and processing data 6 Real-time data processing and analytics 7 Metadata layer architecture 8 Schema management 9 Data access and security 10 Fueling business value with data platforms

A practical guide to implementing a scalable and fast state-of-the-art analytical data estate Key FeaturesStore and analyze data with enterprise-grade security and auditingPerform batch, streaming, and interactive analytics to optimize your big data solutions with easeDevelop and run parallel data processing programs using real-world enterprise scenariosBook Description Azure Data Lake, the modern data warehouse architecture, and related data services on Azure enable organizations to build their own customized analytical platform to fit any analytical requirements in terms of volume, speed, and quality. This book is your guide to learning all the features and capabilities of Azure data services for storing, processing, and analyzing data (structured, unstructured, and semi-structured) of any size. You will explore key techniques for ingesting and storing data and perform batch, streaming, and interactive analytics. The book also shows you how to overcome various challenges and complexities relating to productivity and scaling. Next, you will be able to develop and run massive data workloads to perform different actions. Using a cloud-based big data-modern data warehouse-analytics setup, you will also be able to build secure, scalable data estates for enterprises. Finally, you will not only learn how to develop a data warehouse but also understand how to create enterprise-grade security and auditing big data programs. By the end of this Azure book, you will have learned how to develop a powerful and efficient analytical platform to meet enterprise needs. What you will learnImplement data governance with Azure servicesUse integrated monitoring in the Azure Portal and integrate Azure Data Lake Storage into the Azure MonitorExplore the serverless feature for ad-hoc data discovery, logical data warehousing, and data wranglingImplement networking with Synapse Analytics and Spark poolsCreate and run Spark jobs with Databricks clustersImplement streaming using Azure Functions, a serverless runtime environment on AzureExplore the predefined ML services in Azure and use them in your appWho this book is for This book is for data architects, ETL developers, or anyone who wants to get well-versed with Azure data services to implement an analytical data estate for their enterprise. The book will also appeal to data scientists and data analysts who want to explore all the capabilities of Azure data services, which can be used to store, process, and analyze any kind of data. A beginner-level understanding of data analysis and streaming will be required.

Join Thijs, a young Scandinavian farm boy, on his childhood adventures as he meets new friends and challenges throughout daily Norwegian life. Through hard days of harvest labor and deadly winter blizzards, Thijs journeys from boyhood pleasures to learning what it means to take his place among the men of the North.Full of child-friendly adventure and excitement, Thijs - Saga of a Norseman is a book that you and your children will want to read again and again!

The Unified Star Schema: An Agile and Resilient Approach to Data Warehouse and Analytics Design

Modeling the Agile Data Warehouse with Data Vault

Agile Data Warehouse Design

Chirp / Pollito

ETL techniques to load and transform data from various sources, both on-premises and on cloud

With Examples in SQL Server

The data vault methodology presents a unique opportunity to model the enterprise data warehouse using the same automation principles applicable in today's software delivery, continuous integration, continuous delivery and continuous deployment while still maintaining the standards expected for governing a corporation's most valuable asset: data. This book provides at first the landscape of a modern architecture and then as a thorough guide on how to deliver a data model that flexes as the enterprise flexes, the data vault. Whether the data is structured, semi-structured or even unstructured one thing is clear, there is always a model either applied early (schema-on-write) or applied late (schema-on-read). Today's focus on data governance requires that we know what we retain about our customers, the data vault provides that focus by delivering a methodology focused on all aspects about the customer and provides some of the best practices for modern day data compliance.The book will delve into every data vault modelling artefact, its automation with sample code, raw vault, business vault, testing framework, a build framework, sample data vault models, how to build automation patterns on top of a data vault and even offer an extension of data vault that provides automated timeline correction, not to mention variation of data vault designed to provide audit trails, metadata control and integration with agile delivery tools.

Building a Scalable Data Warehouse with Data Vault 2.0Morgan Kaufmann

Build a modern data warehouse on Microsoft's Azure Platform that is flexible, adaptable, and fast—fast to snap together, reconfigure, and fast at delivering results to drive good decision making in your business. Gone are the days when data warehousing projects were lumbering dinosaur-style projects that took forever, drained budgets, and produced business intelligence (BI) just in time to tell you what to do 10 years ago. This book will show you how to assemble a data warehouse solution like a jigsaw puzzle by connecting specific Azure technologies that address your own needs and bring value to your business. You will see how to implement a range of architectural patterns using batches, events, and streams for both data lake technology and SQL databases. You will discover how to manage metadata and automation to accelerate the development of your warehouse while establishing resilience at every level. And you will know how to feed downstream analytic solutions such as Power BI and Azure Analysis Services to empower data-driven decision making that drives your business forward toward a pattern of success. This book teaches you how to employ the Azure platform in a strategy to dramatically improve implementation speed and flexibility of data warehousing systems. You will know how to make correct decisions in design, architecture, and infrastructure such as choosing which type of SQL engine (from at least three options) best meets the needs of your organization. You also will learn about ETL/ELT structure and the vast number of accelerators and patterns that can be used to aid implementation and ensure resilience. Data warehouse developers and architects will find this book a tremendous resource for moving their skills into the future through cloud-based implementations. What You Will LearnChoose the appropriate Azure SQL engine for implementing a given data warehouse Develop smart, reusable ETL/ELT processes that are resilient and easily maintained Automate mundane development tasks through tools such as PowerShell Ensure consistency of data by creating and enforcing data contracts Explore streaming and event-driven architectures for data ingestionCreate advanced staging layers using Azure Data Lake Gen 2 to feed your data warehouse Who This Book Is For Data warehouse or ETL/ELT developers who wish to implement a data warehouse project in the Azure cloud, and developers currently working in on-premise environments who want to move to the cloud, and for developers with Azure experience looking to tighten up their implementation and consolidate their knowledge

Building upon his earlier book that detailed agile data warehousing programming techniques for the Scrum master, Ralph's latest work illustrates the agile interpretations of the remaining software engineering disciplines: Requirements management benefits from streamlined templates that not only define projects quickly, but ensure nothing essential is overlooked. Data engineering receives two new "hyper modeling"

techniques, yielding data warehouses that can be easily adapted when requirements change without having to invest in ruinously expensive data-conversion programs. Quality assurance advances with not only a stereoscopic top-down and bottom-up planning method, but also the incorporation of the latest in automated test engines. Use this step-by-step guide to deepen your own application development skills through self-study, show your teammates the world's fastest and most reliable techniques for creating business intelligence systems, or ensure that the IT department working for you is building your next decision support system the right way. Learn how to quickly define scope and architecture before programming starts Includes techniques of process and data engineering that enable iterative and incremental delivery Demonstrates how to plan and execute quality assurance plans and includes a guide to continuous integration and automated regression testing Presents program management strategies for coordinating multiple agile data mart projects so that over time an enterprise data warehouse emerges Use the provided 120-day road map to establish a robust, agile data warehousing program

Designing Cloud Data Platforms

Big Data, Data Warehouse and Data Vault

The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

A beginner's guide to harnessing the power and ease of PySpark 3

Youth

Data Management at Scale

Principles and Best Practices of Scalable Realtime Data Systems

The world of data warehousing is changing. Big Data & Agile are hot topics. But companies still need to collect, report, and analyze their data. Usually this requires some form of data warehousing or business intelligence system. So how do we do that in the modern IT landscape in a way that allows us to be agile and either deal directly or indirectly with unstructured and semi structured data?The Data Vault System of Business Intelligence provides a method and approach to modeling your enterprise data warehouse (EDW) that is agile, flexible, and scalable. This book will give you a short introduction to Agile Data Engineering for Data Warehousing and Data Vault 2.0. I will explain why you should be trying to become Agile, some of the history and rationale for Data Vault 2.0, and then show you the basics for how to build a data warehouse model using the Data Vault 2.0 standards.In addition, I will cover some details about the Business Data Vault (what it is) and then how to build a virtual Information Mart off your Data Vault and Business Vault using the Data Vault 2.0 architecture.So if you want to start learning about Agile Data Engineering with Data Vault 2.0, this book is for you.

Master the most agile and resilient design for building analytics applications: the Unified Star Schema (USS) approach. The USS has many benefits over traditional dimensional modeling. Witness the power of the USS as a single star schema that serves as a foundation for all present and future business requirements of your organization. Data warehouse legend Bill Inmon and business intelligence innovator, Francesco Puppini, explain step-by-step why the Unified Star Schema is the recommended approach for business intelligence designs today, and show through many examples how to build and use this new solution. This book contains two parts. Part I, Architecture, explains the benefits of data marts and data warehouses, covering how organizations progressed to their current state of analytics, and to the challenges that result from current business intelligence architectures. Chapter 1 covers the drivers behind and the characteristics of the data warehouse and data mart. Chapter 2 introduces dimensional modeling concepts, including fact tables, dimensions, star joins, and snowflakes. Chapter 3 recalls the evolution of the data mart. Chapter 4 explains Extract, Transform, and Load (ETL), and the value ETL brings to reporting. Chapter 5 explores the Integrated Data Mart Approach, and Chapter 6 explains how to monitor this environment. Chapter 7 describes the different types of metadata within the data warehouse environment. Chapter 8 progresses through the evolution to our current modern data warehouse environment. Part II, the Unified Star Schema, covers the Unified Star Schema (USS) approach and how it solves the challenges introduced in Part I. There are eight chapters within Part II: · Chapter 9, Introduction to the Unified Star Schema: Learn about its architecture and use cases, as well as how the USS approach differs from the traditional approach. · Chapter 10, Loss of Data: Learn about the loss of data and the USS Bridge. Understand that the USS approach does not create any join, and for this reason, it has no loss of data. · Chapter 11, The Fan Trap: Get introduced to the Oriented Data Model convention, and learn the dangers of a fan trap through an example. Differentiate join and association, and realize that an “in-memory association” is the preferred solution to the fan trap. · Chapter 12, The Chasm Trap: Become familiar with the Cartesian product, and then follow along with an example based on LinkedIn, which illustrates that a chasm trap produces unwanted duplicates. See that the USS Bridge is based on a union, which does not create any duplicates. · Chapter 13, Multi-Fact Queries: Distinguish between multiple facts “with direct connection” versus multiple facts “with no direct connection”. Explore how BI tools are capable of building aggregated virtual rows. · Chapter 14, Loops: Learn more about loops and five traditional techniques to solve them. Follow along with an implementation, which will illustrate the solution based on the USS approach. · Chapter 15, Non-Conformed Granularities: Learn about non-conformed granularities, and learn that the Unified Star Schema introduces a solution called “re-normalization”. · Chapter 16, Northwind Case Study. Witness how easy it is to detect the pitfalls of Northwind using the ODM convention. Follow along with an

implementation of the USS approach on the Northwind database with various BI tools.

Work with petabyte-scale datasets while building a collaborative, agile workplace in the process. This practical book is the canonical reference to Google BigQuery, the query engine that lets you conduct interactive analysis of large datasets. BigQuery enables enterprises to efficiently store, query, ingest, and learn from their data in a convenient framework. With this book, you'll examine how to analyze data at scale to derive insights from large datasets efficiently. Valliappa Lakshmanan, tech lead for Google Cloud Platform, and Jordan Tigani, engineering director for the BigQuery team, provide best practices for modern data warehousing within an autoscaled, serverless public cloud. Whether you want to explore parts of BigQuery you're not familiar with or prefer to focus on specific tasks, this reference is indispensable. Geared to IT professionals eager to get into the all-important field of data warehousing, this book explores all topics needed by those who design and implement data warehouses. Readers will learn about planning requirements, architecture, infrastructure, data preparation, information delivery, implementation, and maintenance. They'll also find a wealth of industry examples garnered from the author's 25 years of experience in designing and implementing databases and data warehouse applications for major corporations. Market: IT Professionals, Consultants.

Building with Speed and Agility on Microsoft's Cloud Platform

Data Warehouse Systems

Building a Data Warehouse

The Modern Data Warehouse in Azure

A Comprehensive Guide for IT Professionals

Building the Data Warehouse

Mastering Data Warehouse Design

The final edition of the incomparable data warehousing and business intelligence reference, updated and expanded The Kimball Group Reader, Remastered Collection is the essential reference for data warehouse and business intelligence design, packed with best practices, design tips, and valuable insight from industry pioneer Ralph Kimball and the Kimball Group. This Remastered Collection represents decades of expert advice and mentoring in data warehousing and business intelligence, and is the final work to be published by the Kimball Group. Organized for quick navigation and easy reference, this book contains nearly 20 years of experience on more than 300 topics, all fully up-to-date and expanded with 65 new articles. The discussion covers the complete data warehouse/business intelligence lifecycle, including project planning, requirements gathering, system architecture, dimensional modeling, ETL, and business intelligence analytics, with each group of articles prefaced by original commentaries explaining their role in the overall Kimball Group methodology. Data warehousing/business intelligence industry's current multi-billion dollar value is due in no small part to the contributions of Ralph Kimball and the Kimball Group. Their publications are the standards on which the industry is built, and nearly all data warehouse hardware and software vendors have adopted their methods in one form or another. This book is a compendium of Kimball Group expertise, and an essential reference for anyone in the field. Learn data warehousing and business intelligence from the field's pioneers Get up to date on best practices and essential design tips Gain valuable knowledge on every stage of the project lifecycle Dig into the Kimball Group methodology with hands-on guidance Ralph Kimball and the Kimball Group have continued to refine their methods and techniques based on thousands of hours of consulting and training. This Remastered Collection of The Kimball Group Reader represents their final body of knowledge, and is nothing less than a vital reference for anyone involved in the field.

When a little chick leaves the flock, he stumbles on to an adventure that will change him forever. This charming bilingual Spanish-English picture book is a cute read for little explorers.

Do You Know If Your Data Warehouse Flexible, Scalable, Secure and Will It Stand The Test Of Time And Avoid Being Part Of The Dreaded "Life Cycle"? The Data Vault took the Data Warehouse world by storm when it was released in 2001. Some of the world's largest and most complex data warehouse situations understood the value it gave especially with the capabilities of unlimited scaling, flexibility and security. Here is what industry leaders say about the Data Vault "The Data Vault is the optimal choice for modeling the EDW in the DW 2.0 framework" - Bill Inmon, The Father of Data Warehousing "The Data Vault is foundationally strong and an exceptionally scalable architecture" - Stephen Brobst, CTO, Teradata "The Data Vault should be considered as a

potential standard for RDBMS-based analytic data management by organizations looking to achieve a high degree of flexibility, performance and openness" - Doug Laney, Deloitte Analytics Institute "I applaud Dan's contribution to the body of Business Intelligence and Data Warehousing knowledge and recommend this book be read by both data professionals and end users" - Howard Dresner, From the Foreword - Speaker, Author, Leading Research Analyst and Advisor You have in your hands the work, experience and testing of 2 decades of building data warehouses. The Data Vault model and methodology has proven itself in hundreds (perhaps thousands) of solutions in Insurance, Crime-Fighting, Defense, Retail, Finance, Banking, Power, Energy, Education, High-Tech and many more. Learn the techniques and implement them and learn how to build your Data Warehouse faster than you have ever done before while designing it to grow and scale no matter what you throw at it. Ready to "Super Charge Your Data Warehouse"? Explore the modern market of data analytics platforms and the benefits of using Snowflake computing, the data warehouse built for the cloud. With the rise of cloud technologies, organizations prefer to deploy their analytics using cloud providers such as Amazon Web Services (AWS), Microsoft Azure, or Google Cloud Platform. Cloud vendors are offering modern data platforms for building cloud analytics solutions to collect data and consolidate into single storage solutions that provide insights for business users. The core of any analytics framework is the data warehouse, and previously customers did not have many choices of platform to use. Snowflake was built specifically for the cloud and it is a true game changer for the analytics market. This book will help onboard you to Snowflake, present best practices to deploy, and use the Snowflake data warehouse. In addition, it covers modern analytics architecture and use cases. It provides use cases of integration with leading analytics software such as Matillion ETL, Tableau, and Databricks. Finally, it covers migration scenarios for on-premise legacy data warehouses. What You Will Learn Know the key functionalities of Snowflake Set up security and access with cluster Bulk load data into Snowflake using the COPY command Migrate from a legacy data warehouse to Snowflake integrate the Snowflake data platform with modern business intelligence (BI) and data integration tools Who This Book Is For Those working with data warehouse and business intelligence (BI) technologies, and existing and potential Snowflake users Relentlessly Practical Tools for Data Warehousing and Business Intelligence Remastered Collection Snowflake Cookbook Agile Data Warehousing for the Enterprise Design and Implementation Frontiers in Massive Data Analysis The Data Vault Guru Relational and Dimensional Techniques

Data Modeling for Agile Data Warehouse using Data Vault Modeling Approach. Includes Enterprise Data Warehouse Architecture. This is a complete guide to the data vault data modeling approach. The book also includes business and program considerations for the agile data warehousing and business intelligence program. There are over 200 diagrams and figures concerning modeling, core business concepts, architecture, business alignment, semantics, and modeling comparisons with 3NF and Dimensional modeling.

Shows how to turn conflict and disappointment in a relationship into opportunities for learning, mutual growth, and intimacy

Today, the world is trying to create and educate data scientists because of the phenomenon of Big Data. And everyone is looking deeply into this technology. But no one is looking at the larger architectural picture of how Big Data needs to fit within the existing systems (data warehousing systems). Taking a look at the larger picture into which Big Data fits gives the data scientist the necessary context for how pieces of the puzzle should fit together. Most references on Big Data look at only one tiny part of a much larger whole. Until data gathered can be put into an existing framework or architecture it can't be used to its full potential. Data Architecture a Primer for the Data Scientist addresses the larger architectural picture of how Big Data fits with the existing information infrastructure, an essential topic for the data scientist. Drawing upon years of practical experience and using numerous examples and an easy to understand framework. W.H. Inmon, and Daniel Linstedt define the importance of data architecture and how it can be used effectively to harness big data within existing systems. You'll be able to: Turn textual information into a form that can be analyzed by standard tools. Make the connection between analytics and Big Data Understand how Big Data fits within an existing systems environment Conduct analytics on repetitive and non-repetitive data Discusses the value in Big Data that is often overlooked, non-repetitive data, and why there is significant business value in using it Shows how to turn textual information into a form that can be analyzed by standard tools Explains how Big Data fits within an existing systems environment Presents new opportunities that are afforded by the advent of Big Data Demystifies the murky waters of repetitive and non-repetitive data in Big Data

Data is at the center of many challenges in system design today. Difficult issues need to be figured out, such as scalability, consistency, reliability, efficiency, and maintainability. In addition, we have an overwhelming variety of tools, including relational databases, NoSQL datastores, stream or batch processors, and message brokers. What are the right choices for your application? How do you make sense of all these buzzwords? In this practical and comprehensive guide, author Martin Kleppmann helps you navigate this diverse landscape by examining the pros and cons of various technologies for processing and storing data. Software keeps

changing, but the fundamental principles remain the same. With this book, software engineers and architects will learn how to apply those ideas in practice, and how to make full use of data in modern applications. Peer under the hood of the systems you already use, and learn how to use and operate them more effectively Make informed decisions by identifying the strengths and weaknesses of different tools Navigate the trade-offs around consistency, scalability, fault tolerance, and complexity Understand the distributed systems research upon which modern databases are built Peek behind the scenes of major online services, and learn from their architectures

The Enterprise Big Data Lake

Building a Scalable Data Warehouse with Data Vault 2.0: Implementation Guide for Microsoft SQL Server 2014

Thijo - Saga of a Norseman

Data Architecture: A Primer for the Data Scientist

Big Data

Jumpstart Snowflake

Hands-On Data Warehousing with Azure Data Factory

The data warehousing bible updated for the new millennium Updated and expanded to reflect the many technological advances occurring since the previous edition, this latest edition of the data warehousing "bible" provides a comprehensive introduction to building data marts, operational data stores, the Corporate Information Factory, exploration warehouses, and Web-enabled warehouses. Written by the father of the data warehouse concept, the book also reviews the unique requirements for supporting e-business and explores various ways in which the traditional data warehouse can be integrated with new technologies to provide enhanced customer service, sales, and support—both online and offline—including near-line data storage techniques.

Here is the ideal field guide for data warehousing implementation. This book first teaches you how to build a data warehouse, including defining the architecture, understanding the methodology, gathering the requirements, designing the data models, and creating the databases. Coverage then explains how to populate the data warehouse and explores how to present data to users using reports and multidimensional databases and how to use the data in the data warehouse for business intelligence, customer relationship management, and other purposes. It also details testing and how to administer data warehouse operation.

The Data Vault was invented by Dan Linstedt at the U.S. Department of Defense, and the standard has been successfully applied to data warehousing projects at organizations of different sizes, from small to large-size corporations. Due to its simplified design, which is adapted from nature, the Data Vault 2.0 standard helps prevent typical data warehousing failures. "Building a Scalable Data Warehouse" covers everything one needs to know to create a scalable data warehouse end to end, including a presentation of the Data Vault modeling technique, which provides the foundations to create a technical data warehouse layer. The book discusses how to build the data warehouse incrementally using the agile Data Vault 2.0 methodology. In addition, readers will learn how to create the input layer (the stage layer) and the presentation layer (data mart) of the Data Vault 2.0 architecture including implementation best practices. Drawing upon years of practical experience and using numerous examples and an easy to understand framework, Dan Linstedt and Michael Olschimke discuss: How to load each layer using SQL Server Integration Services (SSIS), including automation of the Data Vault loading processes. Important data warehouse technologies and practices. Data Quality Services (DQS) and Master Data Services (MDS) in the context of the Data Vault architecture.

Provides a complete introduction to data warehousing, applications, and the business context so readers can get-up and running fast Explains theoretical concepts and provides hands-on instruction on how to build and implement a data warehouse Demystifies data vault modeling with beginning, intermediate, and advanced techniques Discusses the advantages of the data vault approach over other techniques, also including the latest updates to Data Vault 2.0 and multiple improvements to Data Vault 1.0

Over the past 5 years, the concept of big data has matured, data science has grown exponentially, and data architecture has become a standard part of organizational decision-making. Throughout all this change, the basic principles that shape the architecture of data have remained the same. There remains a need for people to take a look at the "bigger picture" and to understand where their data fit into the grand scheme of things. Data Architecture: A Primer for the Data Scientist, Second Edition addresses the larger architectural picture of how big data fits within the existing information infrastructure or data warehousing systems. This is an essential topic not only for data scientists, analysts, and managers but also for researchers and engineers who increasingly need to deal with large and complex sets of data. Until data are gathered and can be placed into an existing framework or architecture, they cannot be used to their full potential. Drawing upon years of practical experience and using numerous examples and case studies from across various industries, the authors seek to explain this larger picture into which big data fits, giving data scientists the necessary context for how pieces of the puzzle should fit together. New case studies include expanded coverage of textual management and analytics New chapters on visualization and big data Discussion of new

visualizations of the end-state architecture

Discovering the Magic at the Heart of Your Differences

Google BigQuery: The Definitive Guide

Data Warehousing, Analytics, and Machine Learning at Scale

Essential PySpark for Scalable Data Analytics

Designing Data-Intensive Applications

Cloud Scale Analytics with Azure Data Services

The Elephant in the Fridge

As data management and integration continue to evolve rapidly, storing all your data in one place, such as a data warehouse, is no longer scalable. In the very near future, data will need to be distributed and available for several technological solutions. With this practical book, you'll learn how to migrate your enterprise from a complex and tightly coupled data landscape to a more flexible architecture ready for the modern world of data consumption. Executives, data architects, analytics teams, and compliance and governance staff will learn how to build a modern scalable data landscape using the Scaled Architecture, which you can introduce incrementally without a large upfront investment. Author Piethein Strengholt provides blueprints, principles, observations, best practices, and patterns to get you up to speed. Examine data management trends, including technological developments, regulatory requirements, and privacy concerns Go deep into the Scaled Architecture and learn how the pieces fit together Explore data governance and data security, master data management, self-service data marketplaces, and the importance of metadata Discover how to build a cloud-based data warehouse at petabyte-scale that is burstable and built to scale for end-to-end analytical solutions Key Features Discover how to translate familiar data warehousing concepts into Redshift implementation Use impressive Redshift features to optimize development, productionizing, and operations processes Find out how to use advanced features such as concurrency scaling, Redshift Spectrum, and federated queries Book Description Amazon Redshift is a fully managed, petabyte-scale AWS cloud data warehousing service. It enables you to build new data warehouse workloads on AWS and migrate on-premises traditional data warehousing platforms to Redshift. This book on Amazon Redshift starts by focusing on Redshift architecture, showing you how to perform database administration tasks on Redshift. You'll then learn how to optimize your data warehouse to quickly execute complex analytic queries against very large datasets. Because of the massive amount of data involved in data warehousing, designing your database for analytical processing lets you take full advantage of Redshift's columnar architecture and managed services. As you advance, you'll discover how to deploy fully automated and highly scalable extract, transform, and load (ETL) processes, which help minimize the operational efforts that you have to invest in managing regular ETL pipelines and ensure the timely and accurate refreshing of your data warehouse. Finally, you'll gain a clear understanding of Redshift use cases, data ingestion, data management, security, and scaling so that you can build a scalable data warehouse platform. By the end of this Redshift book, you'll be able to implement a Redshift-based data analytics solution and have understood the best practice solutions to commonly faced problems. What you will learn Use Amazon Redshift to build petabyte-scale data warehouses that are agile at scale Integrate your data warehousing solution with a data lake using purpose-built features and services on AWS Build end-to-end analytical solutions from data sourcing to consumption with the help of useful recipes Leverage Redshift's comprehensive security capabilities to meet the most demanding business requirements Focus on architectural insights and rationale when using analytical recipes Discover best practices for working with big data to operate a fully managed solution Who this book is for This book is for anyone involved in architecting, implementing, and optimizing an Amazon Redshift data warehouse, such as data warehouse developers, data analysts, database administrators, data engineers, and data scientists. Basic knowledge of data warehousing, database systems, and cloud concepts and familiarity with Redshift will be beneficial.

Amazon Redshift Cookbook

The Complete Guide to Dimensional Modeling