1. Record Nr. UNINA9910787905603321

Autore	Inmon W. H.
Titolo	Data architecture : a primer for the data scientist : big data, data warehouse and data vault / / W. H. Inmon, Dan Linstedt ; Steven Elliot, executive editor ; Mark Rogers, designer
Pubbl/distr/stampa	Amsterdam, Netherlands : , : Morgan Kaufmann, , 2015 ©2015
ISBN	0-12-802091-1
Edizione	[1st edition]
Descrizione fisica	1 online resource (378 p.)
Disciplina	005.745
Soggetti	Data warehousing Big data
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Note generali	Includes index.
Nota di contenuto	Cover; Title Page; Copyright; Dedication; Contents; Preface; About the authors; 1.1 - Corporate data; The Totality of Data Across the Corporation; Dividing Unstructured Data; Business Relevancy; Big Data; The Great Divide; The Continental Divide; The Complete Picture; 1.2 - The data infrastructure; Two Types of Repetitive Data; Repetitive Structured Data; Repetitive Big Data; The Two Infrastructures; What's being Optimized?; Comparing the Two Infrastructures; 1.3 - The "great divide"; Classifying Corporate Data; The "Great Divide"; Repetitive Unstructured Data; Nonrepetitive Unstructured Data Different Worlds1.4 - Demographics of corporate data; 1.5 - Corporate data analysis; 1.6 - The life cycle of data - understanding data over time; 1.7 - A brief history of data; Paper Tape and Punch Cards; Magnetic Tapes; Disk Storage; Database Management System; Coupled Processors; Online Transaction Processing; Data Warehouse; Parallel Data Management; Data Vault; Big Data; The Great Divide; 2.1 - A brief history of big data; An Analogy - Taking the High Ground; Taking the High Ground; Standardization with the 360; Online Transaction Processing Enter Teradata and Massively Parallel ProcessingThen Came Hadoop and Big Data; IBM and Hadoop; Holding the High Ground; 2.2 - What is big data?; Another Definition; Large Volumes; Inexpensive Storage; The

Roman Census Approach; Unstructured Data; Data in Big Data; Context in Repetitive Data; Nonrepetitive Data; Context in Nonrepetitive Data; 2.3 - Parallel processing; 2.4 - Unstructured data; Textual Information Everywhere; Decisions Based on Structured Data; The Business Value Proposition; Repetitive and Nonrepetitive Unstructured Information; Ease of Analysis; Contextualization
Some Approaches to ContextualizationMapReduce; Manual Analysis; 2.5 - Contextualizing repetitive unstructured data; Parsing Repetitive Unstructured Data; Recasting the Output Data; 2.6 - Textual disambiguation; From Narrative into an Analytical Database; Input into Textual Disambiguation; Mapping; Input/Output; Document Fracturing/Named Value Processing; Preprocessing a Document; Emails - A Special Case; Spreadsheets; Report Decompilation; 2.7 - Taxonomies; Data Models and Taxonomies; Applicability of Taxonomies; What is a Taxonomy?; Taxonomies in Multiple Languages Dynamics of Taxonomies and Textual DisambiguationTaxonomies and Textual Disambiguation - Separate Technologies; Different Types of Taxonomies; Taxonomies - Maintenance Over Time; 3.1 - A brief history of data warehouse; Early Applications; Online Applications; Extract Programs; 4GL Technology; Personal Computers; Spreadsheets; Integrity of Data; Spider-Web Systems; The Maintenance Backlog; The Data Warehouse; To an Architected Environment; To the CIF; DW 2.0; 3.2 - Integrated corporate data; Many Applications; Looking Across the Corporation; More Than One Analyst; ETL Technology
The Challenges of Integration

Sommario/riassunto

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 a