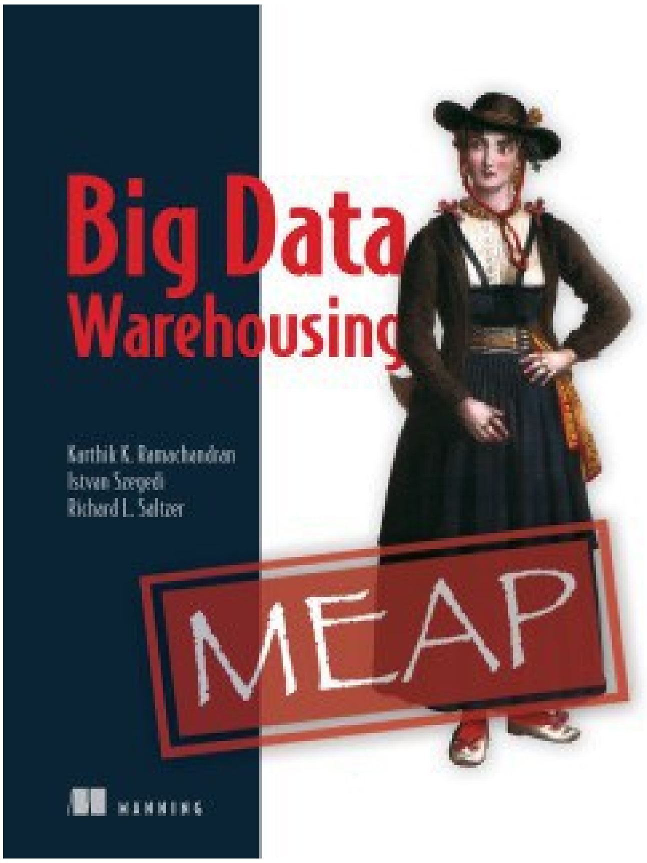
Big Data Warehousing



著者:Karthik Ramachandran

出版者:

出版时间:2016-3-30

装帧:平装

isbn:9781633430280

Big Data Warehousing teaches you new techniques for common data warehousing tasks such as data ingest, SQL queries and report generation in a big data environment. You' ll get a quick tour of using Hive and Impala to query and analyze large semi-structured datasets and learn how to build an Extract, Load, and Transform (ETL) workflow You' ll explore data extraction with Sqoop and address the practical question of schemas for modeling and transforming big data. As you progress through the book, you’ ll survey data governance with Falcon, how to build dataflows with Oozie, approaches to data processing, writing queries with SparkSQL, and data security using Apache Sentry and Knox.

作者介绍:

Karthik Ramachandran is a software engineer and Big Data expert who makes big data technologies and machine learning accessible to business users. He has extensive experience both with traditional enterprise data warehousing solutions as well as with the Hadoop ecosystem. Istvan Szegedi is a senior technical solutions architect working with enterprise data technologies and Hadoop. Richard Saltzer is a Software Engineer on Cloudera's internal data platform team where he builds scalable ingestion pipelines with Impala.

目录: PART 1: INTRODUCTION
1. HADOOP AND DATA WAREHOUSING
1.1. What’ s a Data Warehouse?
1.1.1. Operational vs. analytic systems.
1.1.2. Extract, transform and load
1.1.3. Data Requirements
1.1.4. Baseline Requirements.
1.1.5. A traditional data warehouse architecture
1.2. Defining big data - volume, velocity, variety and veracity
1.2.1. The need for distributed computing
1.3. What is the Hadoop Ecosystem?
1.3.1. What is Apache Hadoop?
1.3.2. The rest of the Hadoop Ecosystem
1.3.3. The Hadoop Ecosystem's Philosophy on Distributed Computing
1.3.4. Hadoop Distributions
1.4. Putting it all together: a Big Data warehouse architecture.
1.5. Who should read this book?
1.6. What is not covered: BI Tools.
1.7. Summary
2. INTRODUCTORY EXAMPLES
2.1. Following Along At Home
2.1.1. Installing a Preconfigured Virtual Machine
2.1.2. Understanding Local, Pseudo-distributed, and Distributed Modes.
2.1.3. Utilizing a Cloud Providers
2.1.4. Picking how you work with Hive — Hive CLI, Beeline, and Hue.
2.1.5. Impala Shell & Hue Query Editor
2.2. Analyzing data with Hive - Salary Data from Baltimore City
2.2.1. Downloading the data from opendata.gov
2.2.2. Uploading the Data into HDFS
2.2.3. Creating a table to house the raw data in Hive
2.3. Querying data with Impala - New York Social Media Stats.
2.3.1. Analyzing your first dataset with Impala.
2.4. Conclusion
PART 2: DATA INGEST & ETL
3. HDFS
3.1. What is HDFS?
3.2. Common HDFS commands.

3.2.1. Following along at home
3.2.2. Interacting with Hadoop - the fs command.
3.2.3. Creating a directory in HDFS
3.2.4. Uploading data into HDFS
3.2.5. Viewing data in HDFS
3.2.6. Copying and moving files in HDFS
3.2.7. File permissions in HDFS.
3.2.8. Deleting files and directories
3.2.9. Downloading Files and Directories
3.3. Other tools for working with HDFS
3.4. Understanding How HDFS Works
3.4.1. Blocks
3.4.2. Data replication
3.4.3. The architecture of HDFS : clients, name nodes and data nodes
3.5. Conclusion
4. DATABASES, TABLES AND VIEWS
4.1. A simple extract, load, and transform workflow
4.2. Following along at home.
4.3. How data is organized in Hive and Impala
4.4. Creating and Dropping Databases
4.5. Creating, loading, altering and deleting tables in Hive and Impala
4.5.1. Creating tables using CREATE TABLE
4.5.2. Loading data using LOAD
4.5.3. Partitioning and Bucketing Tables
4.5.4. Altering Tables
4.5.5. Deleting tables.
4.5.6. Views
4.6. Summary
5. FILE FORMATS
5.1. A simple extract, load, and transform workflow
5.2. Following along at home.
5.3. Why file formats matter.
5.3.1. Revisiting the input/output bottleneck.
5.3.2. Why file structure matters - row vs. column-oriented formats.
5.3.3. Why compression matters.
5.3.4. Converting between file formats using INSERT
5.3.5. Converting between file formats using CREATE TABLE AS SELECT
5.4. Row-oriented file formats
5.4.1. When should I use row-based storage?
5.4.2. Text Files
5.4.3. Sequence Files
5.4.4. Avro
5.5. Column -based Storage
5.5.1. RCFile
5.5.2. ORC File
5.5.3. Parquet
5.6. Summary
6. EXTRACTING DATA WITH APACHE SQOOP.
7. MODELING AND TRANSFORMING DATA
8. AUTOMATING ETL WITH OOZIE
9. DATA GOVERNANCE WITH APACHE FALCON.
PART 3: QUERY ENGINES
10. HIVE
11. IMPALA

12. SPARK SQL
PART 4: OTHER CONSIDERATIONS
13. SECURITY
・・・・・・(收起)

Big Data Warehousing_下载链接1_

标签

hadoop

bigdata

评论

Big Data Warehousing_下载链接1_

书评

Big Data Warehousing_下载链接1_