

Making Everything Easier!™

Cisco Systems Special Edition

Big Data & Analytics

FOR
DUMMIES®
A Wiley Brand

Brought to you by



**Dan Sullivan
Bob Eve
Raghunath Nambiar
Robert Novak
Renee Yao**



About Cisco

Cisco brings 30 years of networking and infrastructure leadership to the table. We help customers create a connected infrastructure from the data center to the edge. Cisco and its partners connect all of a company's data and hyper-distributed environments to unlock data from a hardware or software standpoint.

The Unified Computing System (UCS) Integrated Infrastructure for Big Data provides a secure and scalable infrastructure. Cisco leads the market by integrating technology with innovative and disruptive software companies. While the industry is bringing data to compute, Cisco is bringing the computing and analytics to the data to take advantage of the valuable insight that it holds. With new sources of data coming in, Cisco offers a way to capture, organize, prepare, and handle the data, while providing the speed, consistency, and repeatability necessary for deploying and managing a successful Big Data and Analytics service.

Big Data & Analytics

FOR
DUMMIES[®]
A Wiley Brand

Cisco Systems Special Edition

Big Data & Analytics

FOR
DUMMIES[®]
A Wiley Brand

Cisco Systems Special Edition

**by Dan Sullivan, Bob Eve,
Raghunath Nambiar,
Robert Novak, and Renee Yao**

FOR
DUMMIES[®]
A Wiley Brand

Big Data & Analytics For Dummies®, Cisco Systems Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

Copyright © 2016 by John Wiley & Sons, Inc., Hoboken, New Jersey

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, For Dummies, the Dummies Man logo, The Dummies Way, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc., and/or its affiliates in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc., is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

ISBN 978-1-119-23616-0 (pbk); ISBN 978-1-119-23617-7 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Development Editor: Elizabeth Kuball

Copy Editor: Elizabeth Kuball

Acquisitions Editor: Amy Fandrei

Editorial Manager: Rev Mengle

Business Development Representative:
Karen Hattan

Production Editor: Siddique Shaik

Special Help: Rex Backman,

Marcus Phipps, J.D. Stanley,

Gary Serda, Sean McKeown,

Ron Graham, Dave Kloempken,

Bob Eve, Jim McHugh, Brian Sak,

Danny Dunn

Table of Contents

Introduction	1
About This Book	1
Foolish Assumptions	1
Icons Used in This Book.....	2
Beyond the Book.....	2
Chapter 1: Introducing Big Data and Analytics	3
Defining Big Data and Analytics	3
Looking at the emergence of big data and analytics	4
Describing and predicting with statistics.....	5
Finding patterns with machine learning	5
Working with new tools for big data and analytics	6
Deploying Big Data and Analytics in the Enterprise.....	7
Chapter 2: The Business Case for Big Data and Analytics	9
Looking at Some Popular Use Cases.....	9
Cutting costs with predictive analysis.....	10
Personalizing recommendations for customers.....	10
Enhancing patient care with analytics.....	11
Additional popular use cases.....	12
Supporting Processes and Procedures for Big Data and Analytics	13
Chapter 3: The First Steps to Analytics: Data Collection, Preparation, and Virtualization	15
Big Data Basics	16
Data precision and tolerance for noise.....	16
Collecting Data from Across the Enterprise and Beyond.....	17
Batch processing	17
Stream processing	18
Preparing Big Data for Analysis	19
Preparing data.....	19
Using Cisco Data Preparation	20
Integrating data beyond the big data systems.....	20



Chapter 4: Infrastructure for Big Data and Analytics 23

Key Requirements of Big Data Infrastructure 23
Scalability 23
Cisco Big Data and Analytics Infrastructure 24
 Cisco Unified Computing System..... 25
 Cisco UCS Integrated Infrastructure for Big Data..... 25
 Cisco Application Centric Infrastructure..... 26

Chapter 5: Three Must-Haves to Ensure Success 27

Understanding Your Data and Applications..... 27
Planning for Scalable Infrastructure..... 28
Building Relations with Vendors
 and Business Partners 29

Appendix: Cisco Big Data and Analytics Partners 31

Cloudera..... 31
Hortonworks..... 32
IBM..... 32
MapR..... 32
SAS 33
SAP 34
Splunk..... 34
Platfora..... 34

Introduction

Big data and analytics is a rapidly expanding field of information technology. Big data incorporates technologies and practices designed to support the collection, storage, and management of a wide variety of data types that are produced at ever-increasing rates. Analytics combine statistics, machine learning, and data preprocessing in order to extract valuable information and insights from big data.

About This Book

Big Data & Analytics For Dummies, Cisco Systems Special Edition, is a guide to the rapidly evolving fields of big data management and data science. Big data has moved from a problem faced by a handful of large, data-intensive organizations to a common business problem. Big data is different from related domains, requiring new technologies needed to manage and exploit it. Statistics, machine learning, and data mining are important sets of technologies that enable a new field of data science, delivering analytic services to the enterprise. You don't need a history of related domains to make sense of big data and analytics.

Foolish Assumptions

In preparing this book, I assumed a few things about you:

- ✔ You may not work in information technology, but you work with information.
- ✔ You're challenged with managing growing volumes of data while struggling to get value from it.
- ✔ You're more interested in understanding the full life cycle of big data and analytics than in drilling down into the details of particular techniques or algorithms.
- ✔ You're a pragmatist who wants to build systems and solve problems.

Icons Used in This Book

You find several icons in the margins of this book.



The Tip icon points out important factors in decision making, or sometimes just helpful pointers for getting something done a little faster.



The Remember icon reminds you of facts or trends that businesses often overlook.



Pay attention to the Warning icon and you'll avoid costly pitfalls.

Beyond the Book

You can find additional information beyond what I cover in this book by visiting the following websites, which provide some resources and next steps:

- ✓ www.cisco.com/go/bigdata
- ✓ www.cisco.com/go/bigdata_design
- ✓ www.cisco.com/go/ucs

Chapter 1

Introducing Big Data and Analytics

.....

In This Chapter

- ▶ Defining big data and analytics
 - ▶ Gaining insight into the practice of analytics
 - ▶ Learning about established vendors in the big data market
-

Big data is an increasingly valuable asset to business, but its potential will only be realized through effective analysis. This chapter defines big data and analytics, and highlights key vendors in this space.

Defining Big Data and Analytics

Sometimes a quantitative change leads to qualitative change. With more data, you need more efficient data processing. This phenomenon is apparent in information management where the increasing volume and variety of types of data require fundamental change in the way we manage data. In addition to traditional data sources, such as financial management and customer relationship management tools, we now have data available from:

- ✔ Online browsing activities and click stream analysis
- ✔ System logs capturing information about events across networks, devices, and applications running on IT infrastructure
- ✔ Social media sentiment and activity

- ✓ A constant stream of product feedback in the form of online reviews
- ✓ Mobile device information, including location data and activity of billions of cellphone users
- ✓ Third-party vendors who are building more complex and rich sets of data about individuals and organizations

Traditional methods of collecting, filtering, integrating, and analyzing data are insufficient for the scale and complexity of the data available to today's enterprises. Big data demands new kind of infrastructure.



Big data and analytics entails not just a change in volume of data, but also a change in the way we work with the data.

Looking at the emergence of big data and analytics

It is reasonable to ask why traditional business intelligence (BI) type reporting is not sufficient to extract maximum value from big data. After all, BI is a well-established domain with best practices and a 20-plus-year track record of providing value across a range of industries.

BI reporting systems do a number of things well. They are responsive to ad hoc queries against structured data. For example, if a sales analyst needs to know the current quarter sales volumes of a product line compared to last year's sales for the same quarter, she could easily get the data from a data warehouse or data mart reporting tool.

Similarly, if a regional manager wanted to drill down into performance data to identify underperforming stores, he could work with an online analytic processing (OLAP) cube or dimensional data warehouse report to find that data. Each of these scenarios requires isolating a subset of data, such as sales by quarter or performance by store, and then aggregating that subset of data, such as summing sales over time or averaging margins across stores.

Certainly, you could run reports such as these over big data sets. However, you would not uncover correlation and connection within those data sets. To extract more value, you

have to deploy two additional types of analysis: statistics and machine learning.



You can use existing business intelligence tools with big data, but you won't realize the full potential of big data without tools that exploit the unique properties of big data.

Describing and predicting with statistics

Statistics is a branch of mathematics that focuses on describing populations of data and using data to make predictions about future events. Descriptive statistics are widely used in traditional BI reporting. Any time you run a report that shows minimums, maximums, means, or standard deviations, you're working with descriptive statistics. Some examples include

- ✓ Personalized recommendations in retail, insurance, and other industries, offering an array of options from which customers can choose
- ✓ Predictive maintenance, such as estimating the likely time to failure of a critical component of an industrial machine

The other type of statistics is predictive statistics. These methods use data to build models, which are mathematical formulas for making predictions based on set of data.



Descriptive statistics help understand the characteristics of data sets, while predictive statistics use data sets to make inferences about new instances of similar populations. Describing the size of customer segments and their average marginal revenue, which is a type of descriptive statistics. Predicting the impact of a different sales offers on changes in product sales is an example of predictive statistics.

Finding patterns with machine learning

Machine learning is a field of computer science dedicated to developing algorithms that can identify a wide range of

patterns in data. These patterns come in many forms and include the following:

- ✓ Detecting products frequently bought together
- ✓ Classifying fraudulent transactions
- ✓ Making recommendations to a customer based on the choices of similar customers
- ✓ Discerning subgroups, or clusters, of similar customers, transactions, events, or other entities within a large population

The emerging field of big data and analytics employs additional techniques to those found in business intelligence platforms, and this means new tools are needed.

Working with new tools for big data and analytics

In later chapters, we drill down into detail about different big data and analytics tools, but for now it's sufficient to highlight different types of big data tools and their uses. Here are several you'll be hearing more about throughout this guide:

- ✓ **Hadoop:** A big data storage and processing platform that is almost synonymous with big data. The term *Hadoop* is also used to describe the broad ecosystem of data management, processing, and visualization projects that work with the core Hadoop platform.
- ✓ **NoSQL:** A NoSQL (originally referring to “non-SQL” or “nonrelational”) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases.



The big data and analytics field is rapidly changing, and new tools are being released frequently. Watch for specialized tools that can help solve your particular problems.

Deploying Big Data and Analytics in the Enterprise

Enterprises have a number of options for building their big data and analytics platform, including both open-source and commercially developed tools. Many organizations will have multiple business use cases that could benefit from big data and analytics. Before deploying big data and analytics in the enterprise, it's important to consider the full range of platform options and the business use cases that will initially benefit from the new capabilities.

The market for big data platforms and analytics tools is maturing as indicated by the growing number of established vendors with product offerings in this area. In addition, open-source projects continue to create well-designed and efficient applications and support tools for big data and analytics.

Many companies began developing big data and analytics tools to meet their specific internal requirements and then contributed them to the open-source community. That applies to everything from Hadoop itself to NoSQL platforms (such as MongoDB and Cassandra) to analysis and visualization tools.

Infrastructure is a crucial element to the success of big data and analytics. Enterprises should consider platforms that enable highly scalable and highly available infrastructure for big data and analytics applications.

Cisco big data and analytics solutions offerings

Cisco and its partner ecosystem can offer comprehensive solutions for end-to-end big data and analytics.

Cisco UCS Integrated Infrastructure for Big Data integrates industry-leading compute, storage, connectivity, and management capabilities into a unified, fabric-based architecture optimized for big data and analytics workloads. Cisco UCS Integrated Infrastructure for Big Data is a highly efficient, scalable, high-performance solution that can help your organization grow quickly and cost-effectively. Use it to deliver insights faster, and reduce total cost of ownership (TCO). The Cisco

innovations allow you to unlock the intelligence in your data to help you create a sustainable, competitive business advantage.

As enterprise big data workloads increase in size and complexity, the network will play a crucial role in ensuring workloads are completed and insights are delivered in a timely fashion. Cisco offers an ideal solution to solve network constraints caused by increasing network traffic: Cisco Application Centric Infrastructure (ACI) can apply policies and dynamically load balance across the application infrastructure for optimal performance.

Chapter 2

The Business Case for Big Data and Analytics

In This Chapter

- ▶ Identifying the common use cases of big data and analytics in your enterprise
 - ▶ Integrating big data and analytics into your application ecosystem
-

Studies show that big data and analytics can significantly improve business process to customer retention. To get the maximum benefits of big data and analytics, you need to understand the business, understand the customers, and have a long-term plan to monetize ever-growing data.

In this chapter, we offer sample use cases that can enhance customer-focused decision-making. We also describe key aspects of successful implementations and deployments.

Looking at Some Popular Use Cases

From forming sales predictions to offering purchase suggestions to customers, big data can inform a multitude of business concerns. Here are a few use cases to give you an idea of what big data can do for business.

Cutting costs with predictive analysis

Businesses love saving money and making money. One of the best ways to do that with big data is by using predictive analysis. By analyzing large business data sets with predictive analysis, companies can predict, with increasing accuracy, how well certain products will sell, when they will need more of a certain product, and so on. This can cut down on waste, and reduce unsold products sitting in a warehouse for years at a time.

You can take this one step further, however. What would happen if you let your customers know what you were predicting?

Personalizing recommendations for customers

Businesses can use individual customer data, such as purchasing history or location data, to personalize certain marketing practices. For instance, when a customer adds an item to an online shopping cart, the vendor may want to immediately recommend a cross-sell product.

Similarly, when a mobile device user passes a restaurant, the sales generation service may push a coupon to increase the chances of the person stopping in for lunch.

This sort of marketing can also evolve into stopping *customer churn* (customers stopping their patronage of your business). When you can offer personalized product suggestions or coupons, a customer may be more likely to stick around and purchase from you again.

Figure 2-1 shows an example of predictive analytics through data points that correspond to particular values along the x-axis and y-axis. These could represent the price of a commodity over time, the average sales margin over time, or some other measure that changes over time and is of interest to a business. A method known as *linear regression* can find a line that best fits the historical pattern of data.

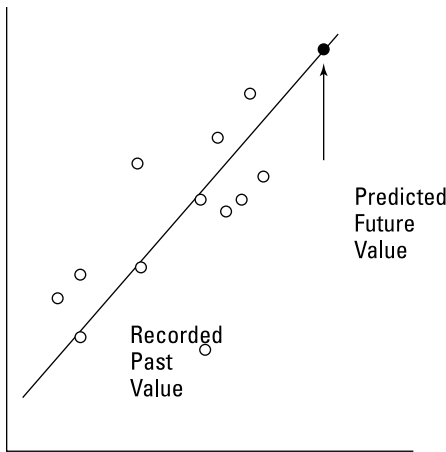


Figure 2-1: Predictive statistics is applied to big data to help make projections about future events and values.

However, it's important not to try to find too close a fit to historical data. The goal is to find a pattern that is useful for predicting future values, not have a model that exactly describes the past. The model that best fits the past may not give enough room to account for measurement errors or variations in your customer data.



Finding an optimal data model is challenging and not always obvious. Be sure to understand your algorithms and their limitations before deploying models to production.

Enhancing patient care with analytics

A major hospital system and research organization started using big data and analytics to improve healthcare delivery. Initial efforts were successful, but the big data and analytics team soon found itself outgrowing its homegrown cluster. To continue to enable leading-edge clinical research, the BI leadership team turned to a Cisco solution that included UCS servers and a Hadoop vendor.

As a result of upgrading their homegrown cluster, clinicians were able to receive information that helped improve patient

outcomes, while IT was able to store and process more clinical data without incurring additional costs.

Big data and analytics can revolutionize the healthcare industry. For example, it can

- ✔ Improve quality, safety, efficiency, and reduce health disparities
- ✔ Engage patients and family
- ✔ Improve care coordination and public health
- ✔ Maintain the privacy and security of patient health information
- ✔ Enhance clinical outcomes
- ✔ Improve population health outcomes
- ✔ Increase transparency and efficiency
- ✔ Empower individuals
- ✔ Generate more robust research data on health systems

Additional popular use cases

Here are some more use cases for big data and analytics:

- ✔ Financial services firms can assess risk and achieve compliance by combing data from different sources.
- ✔ Pharmaceutical firms can combine data from multiple places to get product to market faster.
- ✔ Manufacturing firms can manage inter-train forces to better manage acceleration and braking.
- ✔ Food and beverage firms can evaluate online sales data across thousands of stores to gain business and marketing insights.

Supporting Processes and Procedures for Big Data and Analytics

In addition to the obvious resources needed for big data and analytics, such as compute and storage services, organizations need to be prepared to support analysts, application developers, and product managers with regards to their use of big data.



Big data requires a full life cycle management program that includes attention to data, workflows, and model generation.

In particular, consider the following:

- ✓ **How to design and deploy big data infrastructure that is scalable, easy to use, and readily managed.**
- ✓ **How to govern access to data.** Systems of record, such as financial reporting systems, require strict governance. Big data and analytics is more tolerant of minor discrepancies and places more emphasis on flexibility and adaptability. It should be governed accordingly.
- ✓ **How to protect intellectual property developed to meet the enterprise's requirements.** Procedures should be in place to preserve this intellectual property and make it accessible to others in the organization.
- ✓ **How to manage a typical life cycle that begins with hypothesizing about a business need, collecting and exploring data, building and evaluating new models, and deploying these models to production.** Mature users of big data and analytics will have multiple projects at varying stage of the life cycle continually underway.

As you can see from the use cases in this chapter, big data and analytics is built on a wide range of potential application areas.

Chapter 3

The First Steps to Analytics: Data Collection, Preparation, and Virtualization

In This Chapter

- ▶ Understanding the need for a variety of data from across the enterprise
 - ▶ Transforming data into analytics-friendly forms
 - ▶ Integrating big data with other data
 - ▶ Reviewing Cisco solutions for data preparation and data virtualization
-

Big data and analytics begins with the collection of source data, and the ingestion of that data into scalable storage systems. Data preparation steps follow. Yet some analytic data may continue to reside outside this big data environment, so additional data integration may be required.

We begin this chapter with a primer on big data basics, particularly the factors that drive the need for diverse sources of data. We describe the key differences between two modes of big data ingestion: batch and streaming. Finally, we identify key data preparation and integration activities along with useful tools to perform them.

Big Data Basics

Data is valuable. Once collected, data also enables for many forms of analysis, spanning what has happened in the past to what might happen in the future. Big data and analysis then drives valuable insights to give you an edge over your competitors.

Big data volume, complex processing flows, and advanced analytics provide an opportunity and challenge. Get it right, and the insights derived can lead to excellent business outcomes. Get it wrong and the result can be poor decisions and bad outcomes and ultimately lead to costly mistakes.

Data precision and tolerance for noise

Some analysis requires precise and accurate data. Financial reporting is a prime example where a penny difference in EPS can result in large stock price changes. Other analysis can be more exploratory and, thus, less precise. For example, you may need to evaluate trends or respond quickly to an outside event such as a competitor's marketing campaign.

Big data and analytics often must be more tolerant of errors and missing data than other analysis. There are several reasons why data may not be completely accurate or complete in big data systems:

- ✔ There may be errors in the source system.
- ✔ Data may be missing in source systems or lost in transit.
- ✔ Transformations may filter data before it lands in the big data store.

Tolerance for data errors is largely determined by the business use case. If the data is used by a data science team building predictive models, the statistical techniques used may accommodate potential data quality problems (to a point). However, when data is used for compliance reporting or some other process that requires accurate and consistent information, tolerance for error is low.

Collecting Data from Across the Enterprise and Beyond

The goal of big data and analytics is to uncover and validate important business insights. As such, more data is better, because more data allows deeper insights, and more types of data allow broader insights.

Enterprises have no shortage of data sources. Some are well established and well understood. Financial data about company overall performance; detailed data about customer transactions down to the individual customer, store, SKU, and more; and detailed logs of systems activities are all examples. Some data resides within enterprise IT systems; other data resides in the cloud. Some data comes from third-party providers; other data comes from sensors and devices. Data, data everywhere!

Big data changes the data collection and analysis landscape because of the volume of data, the different types of data that can be analyzed, and the pace at which data is generated.

Data for analysis is typically aggregated from source systems into big data systems using either batch processing or stream processing. Batching data into blocks of work is the traditional ingestion mode. With the proliferation of sensors and other data collection devices, stream processing, in which data is continually processed, is becoming more important as well. Figure 3-1 is a visual representation of the difference between the two.

Batch processing

Data is continuously generated and collected in files or databases. With batch processing, data for analysis moves into a big data repository for periodic analysis. Data is aggregated in logical chunks — for example, point-of-sale transactions from a store in the last 60 minutes or all accounting transactions during a day.

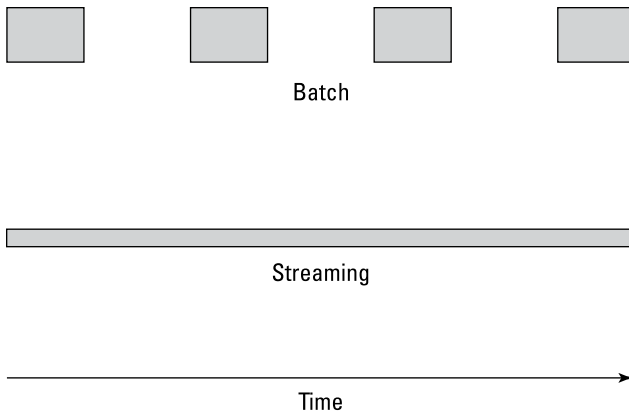


Figure 3-1: Batch processing versus stream processing.

Definition, execution, and management of batch processing are typically performed using extract, transform, and load (ETL) and/or extract, load, and transform (ELT) tools.

Streaming is the process of collecting, integrating, and processing data as it is generated. It's mostly used when batch processing is too slow.

Stream processing

Stream processing combines collection and preliminary analysis of data as it is generated. It's often used when batch processing could delay needed insight. For example, when a customer adds an item to an online shopping cart, a cross-sell recommendation analysis immediately determines and displays an offer for a related product.

Sometimes stream processing is used to break up the analytic workload with some analysis occurring at collection and other analysis done later within the big data system. For example, in information security and fraud detection applications, immediate analysis greatly reduces breach and fraud risk in real time, whereas subsequent analysis can be used to track historical risk and fraud trends.



Sometimes stream processing is implemented using mini-batches. That is, a number of input records are collected and then processed instead of processing each record as it arrives. When records arrive in quick succession, mini-batch processing may be a more efficient alternative to immediate processing of the stream.

Stream processing is executed using tools such as Cisco Connected Streaming Analytics. Cisco ParStream is an example of a fast-ingestion, high-compression database purpose-built for capturing high-volume data streams.

Preparing Big Data for Analysis

After you've collected data from your data sources and ingested it into a scalable big data storage system, the next step is to prepare that data for analysis.

This preparation is needed because data from different sources may not share the exact same definitions, formats, and so on, thus making it hard to combine and analyze. For example, two source systems might store dates in different formats. Analysts use data preparation tools to transform these dates into a consistent form, thereby enabling consistent analysis.

Preparing data

Data preparation is the set of activities required to reformat and transform data from the way it was collected and subsequently stored in the big data storage system into a form that is suitable for analysis. This stage sounds relatively simple, but it can account for a majority of effort and time in any data analysis project, with some estimates as high as 80 percent.

Amongst business analysts today, Microsoft Excel has been the *de facto* standard for data preparation. Although Excel is an ideal tool for many use cases, big data and analytics requires analytic tools that scale far beyond what you can expect from a desktop application. This doesn't mean, however, that tool designers can ignore the ease-of-use features that contribute to Excel's popularity.

Data preparation tools that scale to big data loads and match Excel's ease of use are proving to be the new way to go to reduce data preparation time and enable end-users to prepare data themselves.

Using Cisco Data Preparation

Cisco Data Preparation is a self-service tool that allows analysts and other nonprogrammers to prepare big data for analysis. The tool brings the ease of use of Microsoft Excel to big data and analytics. That, combined with the rich set of data preparation functionality, creates an effective and efficient tool for business analysts. This leaves analysts with more time to analyze the data and gain important insights.

Ease of use, however, does not imply lack of power. Users can employ an exploration-based process to iteratively test hypotheses and evaluate feedback. The tool also incorporates features to guide analysts and automate time-consuming tasks. Cisco Data Preparation scales well beyond what desktop tools, like spreadsheets, are capable of handling.

Another advantage of Cisco Data Preparation is that it helps improve governance by reducing the need for ad hoc data preparation steps. There is no need for an analyst to download a data set, perform some operations in an Excel spreadsheet, and then load the results back into a sanctioned process flow. These kinds of ad hoc fixes often solve a short-term problem but lead to procedures that are difficult to maintain, reproduce, and govern.



Be sure to review business analysts' data preparation usage trends. It can provide valuable insights into where IT should focus future data and analytic investments.

Integrating data beyond the big data systems

Despite best efforts, important analytic data will always exist outside the big data storage systems. For example, data managed by cloud applications such as Salesforce.com and NetSuite live in the cloud. And in some cases, data sovereignty regulations prevent physical consolidation of source data into big data systems.

Businesses must integrate these other sources with their big data systems to gather the full set of insights required to successfully compete. For example, customer data that includes only purchase and payment history is of limited use. When combined with data about the customer's behavior on a website — such as what products the customer searched for, which reviews were read, and how these activities change over time — the opportunity for better customer engagement and cross-selling is greatly enhanced.

Data virtualization

Data virtualization is a flexible data integration technique that spans big data storage systems and other information sources to provide a complete view of all data, regardless of where the data resides.

It allows your business to pull data directly from multiple sources on demand, transform it, and deliver it to your analytic tool of choice as needed. Data from external sources can remain in their original system of records, such as data warehouse or cloud repository.



Data virtualization does not eliminate the technical challenges of data integration, but it does simplify and accelerate the integration process.

Cisco Data Virtualization

Cisco Data Virtualization is a proven data integration platform used by hundreds of enterprises to integrate big data, traditional enterprise data sources, cloud sources, Internet of Things (IoT) device sources, and more.

Cisco Data Virtualization lets IT and end-users access data from multiple sources without having to delve into implementation details for each data source.

When compared to traditional ETL/ELT-based approaches to this type of integration, Cisco Data Preparation is faster to build and deploy, often enabling enterprise-grade integration of diverse data sources in days, not months.

It is also more economical, with typical hardware and software costs up to 90 percent lower than standard ETL implementation.

And it solves practical problems such as integrating data sources that cannot — or should not — reside in a big data system:

- ✔ Governance or compliance rules on source systems prevent copying the data.
- ✔ Source data is easy to access and analyze without replication.
- ✔ Source data volumes are too large to move.

Chapter 4

Infrastructure for Big Data and Analytics

In This Chapter

- ▶ Identifying key requirements of big data infrastructure
 - ▶ Reviewing the capabilities and advantages of Cisco Unified Computing System
-

Big data requires a more scalable and extensible infrastructure to be designed and deployed. Traditional business intelligence (BI) platforms are often appliance type design or more regimented in configuration. For this reason, scalability and manageability are critical for a big data infrastructure.

Key Requirements of Big Data Infrastructure

When enterprises extend their IT infrastructure to support big data and analytics, they're committing to maintaining and managing a new set of services.

Scalability

Big data is characterized in part by the volume of data captured and analyzed by businesses. The data by itself is like a raw natural resource that must be extracted and processed to transform it into something of value.

In addition, the volume of data can easily grow exponentially, so the infrastructure must be able to scale quickly and consistently to meet the needs of growing businesses.

Scalable big data infrastructure includes three core components:

- ✔ **Compute:** The rapid growth of data requires an environment that can scale processing power to keep up with the data. This usually means adding more identical servers or like servers to the existing environment since most big data platforms scale horizontally. It's also critical for a large server deployment to be configured consistently to ensure sustainable performance.
- ✔ **Storage:** A big data and analytics environment often becomes the repository of valuable data for the business. Data grows at an increasing rate and often has substantial retention requirement. A sustainable big data environment must support rapid expansion and changing performance concerns. As data ages, the platform should adjust to changing performance requirement versus capacity concerns.
- ✔ **Network:** Big data means lots of data movement that requires high-performance networking to keep up with processing and analysis. High latencies, insufficient bandwidth, and other network performance issues will cripple an otherwise capable big data infrastructure, leading to degraded analytics capability and lower business value.

Big data offers the potential for valuable insights about business operations and market opportunities, but it requires a scalable infrastructure that meets the compute, storage, and networking demands.

Cisco Big Data and Analytics Infrastructure

Cisco has developed an enterprise-grade infrastructure offering built around Cisco Unified Computing System (UCS) and Cisco Application Centric Infrastructure (ACI).

Cisco Unified Computing System

The Cisco UCS is a hardware platform designed to support the needs of big data. When compared to other hardware vendor offerings, UCS offers the following advantages:

- ✓ **Scalability:** Consistent and repeatable configuration whether deploying 1 or 1,000 servers
- ✓ **Manageability:** Policy-driven configuration and centralized management to reduce operational overhead
- ✓ **Performance:** Optimized component, configuration, and network infrastructure

UCS is a complete hardware platform that includes servers, network, and storage with integrated management. Of course, Cisco Services is available to support initial installation and ongoing operations.

Cisco UCS Integrated Infrastructure for Big Data

Cisco UCS Integrated Infrastructure for Big Data provides specific configuration designed for big data workloads. Use it to deliver insights faster, and reduce total cost of ownership (TCO).

The core components include:

- ✓ **Cisco UCS 6200 and 6300 platform fabric interconnects:** One Cisco UCS Manager instance can manage two Cisco UCS 6300 or 6200 Series Fabric Interconnects, up to 20 Cisco UCS 5100 Series Chassis, up to 40 total Cisco UCS 2200 or 2100 Series Fabric Extenders, and 160 Cisco UCS B-Series Blade Servers or Cisco UCS C-Series Rack Servers.
- ✓ **Cisco UCS Manager:** Cisco UCS Manager provides policy-driven configuration and a central point of management and monitoring for the entire UCS system.

It helps significantly reduce management and administration expenses by automating routine tasks to increase operational agility while reducing risk.

UCS Manager manages all the system components as a single logical entity. It can be accessed through an intuitive graphical user interface (GUI), a command-line interface (CLI), or an XML application-programming interface (API). Cisco UCS Manager uses service profiles to define the personality, configuration, and connectivity of all resources within Cisco UCS, radically simplifying provisioning of resources so that the process takes minutes instead of days. This simplification allows IT departments to shift their focus from constant maintenance to strategic business initiatives.

- ✔ **Cisco UCS C-Series Rack Servers:** Cisco UCS C-Series Rack Servers deliver unified computing in an industry-standard form factor with flexible compute and storage options to reduce total cost of ownership and increase agility. Cisco UCS Virtual Interface Card (VIC) offers flexible network options configured on demand through Cisco UCS Manager.
- ✔ **Cisco UCS Director Express for Big Data:** Cisco UCS Director Express for Big Data provides a single-touch solution to deploy Hadoop on Cisco UCS servers. It also provides a single management frame across both physical infrastructure and Hadoop software.

Cisco Application Centric Infrastructure

Cisco ACI delivers policy-driven network configuration across Cisco Nexus 9000 switches to reduce TCO, automate IT tasks, and accelerate data center application deployments. Cisco ACI offers a scalable Software Defined Networking (SDN) platform to automatically monitor and adapt to changing network conditions and topology.

Chapter 5

Three Must-Haves to Ensure Success

.....

In This Chapter

- ▶ Understanding your data and applications
 - ▶ Planning for scalable infrastructure
 - ▶ Building relations with vendors and partners
-

Big data and analytics is a challenging undertaking, but a valuable one. Act on these three factors to expedite deployment, mitigate risk, and ensure a successful big data and analytics outcome.

Understanding Your Data and Applications

Data and applications are interwoven components of big data and analytics.

Batch processing will always have a role, but near real-time operations are becoming essential to today's business. Mobile devices, sensors, and other Internet of Things (IoT) technologies will alter the balance of batch and streaming data.

Big data is dynamic in terms of volume, types, and use cases for data. Remember that as business environments and markets change, the demands placed on your big data and analytics infrastructure will change as well.

Don't think of a big data project as a destination but as a journey. You'll reach a useful state that you'll build on for the next phase as business and data requirements change.

Planning for Scalable Infrastructure

Scalability is an essential characteristic of big data and analytics. Data volumes will grow over time and require additional storage.

It's important to select an infrastructure platform — compute, storage, and network — that can scale as your workload grows.

Use an infrastructure platform designed and tested for big data and analytics application, rather than reinventing the wheel with an ad hoc infrastructure. Cisco and its big data ecosystem partners provide integrated solutions that accelerate time to value.

UCS Integrated Infrastructure Platform described in this book provides an end-to-end hardware and software solution that adapts.

Plan to manage infrastructure to the dynamic needs of enterprises. Leverage the management capabilities of the Cisco UCS platform, including

- ✔ Cisco UCS
- ✔ Cisco UCS Fabric Interconnects
- ✔ Cisco UCS Manager
- ✔ Cisco UCS Director Express
- ✔ Cisco ACI

Apply Cisco Validated Design (CVD) to achieve predictable and tested performance and scalability between Cisco UCS and its partners' joint solutions. The most recent version of the CVD can be found at www.cisco.com/go/bigdata_design.

Building Relations with Vendors and Business Partners

Data professionals can spend years studying system designs and working in complex production environments, but your business demands faster results. Don't be afraid to enlist outside resources and take advantage of their experience in big data and analytics.

Cisco is a long-term partner to enterprises deploying big data and analytics infrastructure, as well as an established user of these technologies. It can bring its own experience and the expertise of its third-party vendor ecosystem to supplement its customers' in-house resources. Through its extensive joint products and engineering interaction, Cisco and its partners deliver accelerated solutions to various customer requirements.

Appendix

Cisco Big Data and Analytics Partners

.....

Many vendors are staking a claim to some part of the big data market. Here are examples of some vendors that offer products that scale to the meet the data collection, storage, and analysis demands of big data and analytics.

Cloudera

Cloudera delivers a modern data management and analytics platform built on Apache Hadoop and the latest open-source technologies. Cloudera helps some of the world's leading organizations help solve challenging business problems with Cloudera Enterprise, a fast, easy, and secure data platform that helps customers efficiently capture, store, process, and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly, and at lower cost than has been possible. Cloudera also offers comprehensive support, training, and professional services.

Cloudera Enterprise runs seamlessly on and works intelligently with the Cisco UCS Integrated Infrastructure for Big Data. Together they enable clients to store, protect, and access data when needed in an industry-compliant environment. As clients grow, Cloudera and Cisco can scale to meet needs without slowing business activity down. Organizations benefit from proactive and predictive support, including alerts of potential issues to minimize downtime. In addition, data is well protected with authentication protocols and data encryption. Learn more at www.cloudera.com.

Hortonworks

Hortonworks is a leading commercial vendor of Apache Hadoop, the open-source platform for storing, managing, and analyzing big data. Hortonworks Data Platform, its distribution of Apache Hadoop, provides an open and stable foundation for enterprises and a growing ecosystem to build and deploy big data solutions.

Hortonworks is a trusted source for information on Hadoop, and together with the Apache community, Hortonworks is making Hadoop an enterprise data platform. Hortonworks provides technical support, training, and certification programs for enterprises, systems integrators, and technology vendors.

IBM

Cisco and IBM have an almost two-decade history of working together, deep technical expertise, global resources, and world-class support, demonstrated by more than 25,000 shared customers.

In 2014, Cisco joined with IBM to create the VersaStack Solution, the world's first integrated infrastructure based on the IBM Storwize family of virtualized storage technologies and the Cisco Unified Computing System (Cisco UCS). Further extending this partnership, the two companies created a comprehensive set of solutions for big data and analytics, combining innovations from Cisco UCS such as programmable infrastructure with open-source software with enterprise-grade capabilities in IBM BigInsights for Apache Hadoop. These solutions are designed and optimized for common use cases, pre-tested, pre-validated, and fully documented by Cisco and IBM engineers to ensure dependable deployments that can scale from small to very large as workload demands.

MapR

MapR provides a converged data platform that integrates the power of Hadoop and Spark with global event streaming, real-time database capabilities, and enterprise storage, enabling customers to harness the enormous power of their data.

Organizations with the demanding production needs, including sub-second response for fraud prevention, secure and highly available data-driven insights for better healthcare, petabyte analysis for threat detection, and integrated operational and analytic processing for improved customer experiences, run on MapR.

MapR is used across financial services, retail, media, healthcare, manufacturing, telecommunications, and government organizations, as well as by leading Fortune 100 and Web 2.0 companies. Amazon, Cisco, and Google are part of MapR's broad partner ecosystem.

MapR has partnered with Cisco to create a validated and tested infrastructure design that can be rapidly deployed, managed, and scaled with high reliability for companies now living in a data-driven world. In addition to this validated, integrated, architecture for big data, Cisco and MapR currently offer joint services and software through Cisco's Advanced Services team. Together, they're looking at future technology integrations.

SAS

SAS, a privately held company based out of Cary, North Carolina, is a market leader in software analytics. SAS software is installed in more than 148 countries on more than 80,000 sites worldwide. Ninety-one of the top 100 companies on the 2015 Fortune Global 500 use SAS to make business decisions. SAS generates more than \$3 billion annually in revenue and reinvests 25 percent back into research and development (R&D). SAS sells into a broad set of verticals, including a very strong presence in financial services and health and life sciences. SAS solutions cross the spectrum of the analytics space including big data and analytics and visualization. SAS professionals work closely with business stakeholders to bridge the gaps between data, discovery, and deployment on an enterprise scale.

SAS offers visual analytics and high-performance analytics on Cisco UCS. SAS has also integrated its intelligent contact center software for analytics with Cisco call center software, providing keen insights to call center managers. Learn more at www.sas.com.

SAP

A market leader in enterprise application software, SAP applications and services enable customers to operate profitably, adapt continuously, and grow sustainably.

Together, Cisco and SAP offer differentiated, scalable, highly secure end-to-end solutions. With SAP Applications on Cisco UCS, you can reduce deployment risks, complexity, and total cost of ownership (TCO). This can transform the way people connect, communicate, and collaborate.

Splunk

Splunk, Inc., is a market-leading platform that powers Operational Intelligence. Splunk pioneers solutions that make machine data accessible, usable, and valuable to everyone. More than 10,000 customers in over 100 countries use Splunk software and cloud services. You can try Splunk solutions for free: www.splunk.com/free-trials.

Cisco and Splunk have built more than 20 apps for Cisco infrastructure integration and management, including apps for UCS and ACI. Splunk doesn't have equivalent apps for other compute platforms for any other vendor. Together, Cisco and Splunk have future plans for innovations in the IT operations and security analytics areas.

Platfora

Platfora is the number-one big data discovery platform built natively on Apache Hadoop and Spark. Platfora enables business users and data scientists to visually interact with petabyte-scale data in seconds, allowing them to work with even the rawest forms of transaction, customer interaction, and machine data.

Learn more about Platfora at www.platfora.com, read its blog at www.platfora.com/blog, or follow @platfora on Twitter.

Notes



Handwriting practice lines consisting of 20 solid horizontal lines.

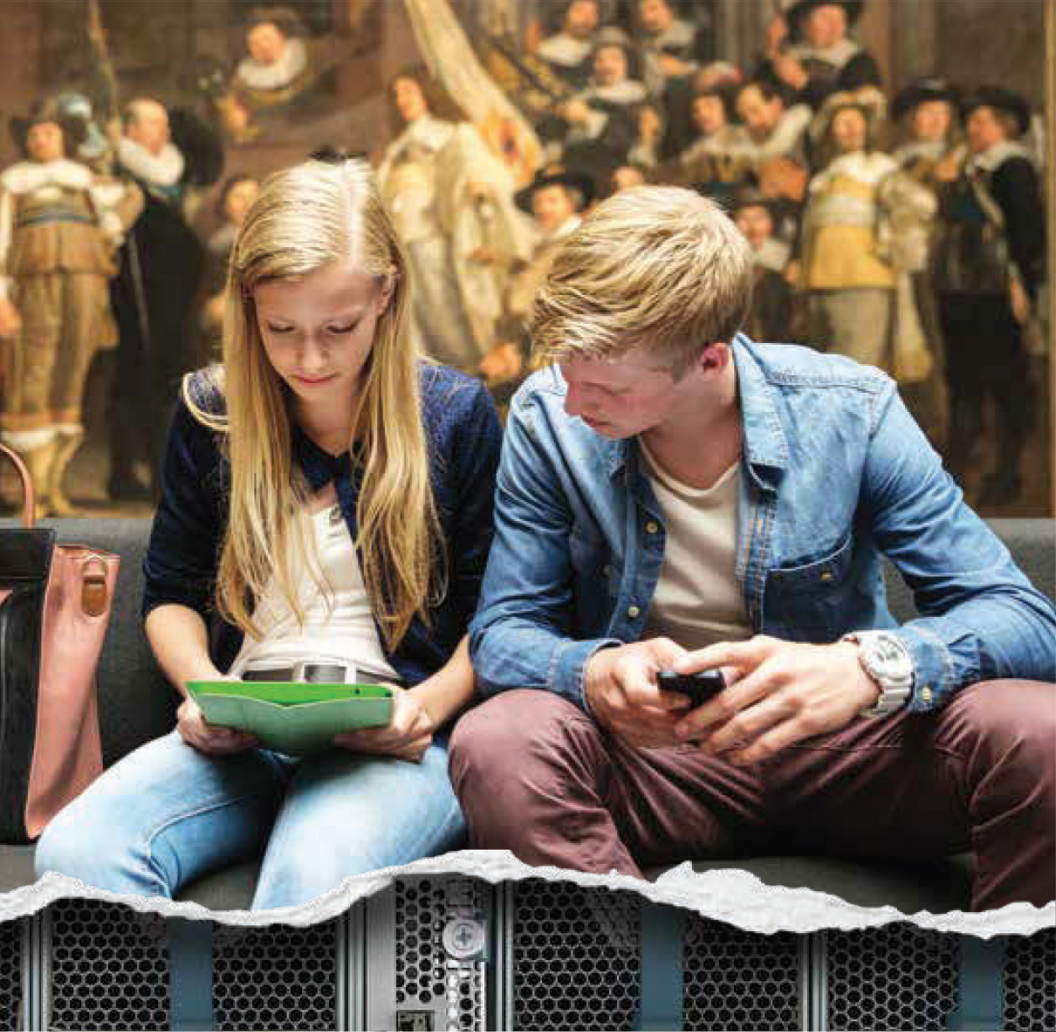
Notes



Notes



A series of 20 horizontal solid lines for writing, arranged in a column below the dotted line.



The center of endless possibilities

Insight is the key to disruption—changing the status quo.

Cisco solutions for data centers and real-time analytics, built on Cisco UCS servers, help you turn your big data challenges into impactful business decisions.

cisco.com/go/disrupt



Intel, the Intel logo, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation in the U.S. and/or other countries.

Cisco UCS®
with Intel® Xeon®
processors



These materials are © 2016 John Wiley & Sons, Inc. Any dissemination, distribution, or unauthorized use is strictly prohibited.

Learn about common use cases, data preparation, data virtualization, and infrastructure aspects of successful big data deployments

This book offers an overview of what is needed to get started in big data and analytics.

- *Learn about common deployment scenarios — Understand tool and application options for working with big data*
- *Learn major phases of big data and analysis — Learn how to collect, prepare, analyze, and visualize big data*
- *Selecting the optimal infrastructure — Delve into the capacity, performance, and scalability of your infrastructure to meet your application demand*

Dan Sullivan, MSc, is an author, systems architect, and consultant with more than 20 years of IT experience. **Bob Eve** has held executive-level engineering, marketing, and business development roles at Oracle, PeopleSoft, Mercury Interactive, Informatica, and Cisco. **Raghunath Nambiar** is Chief Architect of Emerging Technology Solutions and a Cisco Distinguished Engineer. **Robert Novak** is a big data whisperer for Cisco's partner organization; he walks softly, but carries a big shell script. **Renee Yao**, Cisco Product and Solutions Marketing, focuses on Big Data and Analytics for Cisco UCS.




Open the book and find:

- An overview of big data and analytics
- The business case for big data and analytics
- Tips on big data collection and integration
- Data and analytics tools and techniques

Go to Dummies.com
for more!

FOR
DUMMIES[®]
A Wiley Brand

 Also available
as an e-book

ISBN: 978-1-119-23616-0
Not for resale

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.