PRENTICE HALL

UNDERSTANDING BIG DATA SCALABILITY

BIG DATA SCALABILITY SERIES, **PART I**

CORY ISAACSON

PRAISE FOR *UNDERSTANDING BIG DATA SCALABILITY*

"This book is useful to anyone who works with data and wants to learn more about scaling. Cory helps you understand what causes databases to slow down as data volumes grow over time. He then reviews a number of strategies that you have at your disposal to manage the growth, including software and database tuning, hardware upgrades, read-replication, and ultimately horizontal partitioning of data."

—Dan Lynn, cofounder of FullContact

"*Understanding Big Data Scalability* presents the fundamentals of scaling databases from a single node to large clusters. It provides a practical explanation of what Big Data systems are, and the fundamental issues to consider when optimizing for performance and scalability. Cory draws on his many years of database experience to explain the issues involved in working with data sets that can no longer be handled with single monolithic relational databases.

"When transitioning from a traditional relational database deployment, it is tempting to ignore traditional database discipline regarding data modeling and data integrity. Much of this has been motivated by the proliferation of schema-less NoSQL databases. In spite of this trend, Cory shows why it is still important to carefully structure your data to maintain data integrity and allow sharding in such a way as to avoid costly distributed scan/shuffle operations. He discusses a practical approach to this called *relational sharding*. This is a commonsense method that avoids the pitfalls of black-box sharding. Cory's approach is particularly relevant now that relational data models are making a comeback via SQL interfaces to popular NoSQL databases and Hadoop distributions.

"*Understanding Big Data Scalability* addresses practical problems in Big Data processing systems using real-life examples. This book should be especially useful to database practitioners new to the process of scaling a database beyond a traditional single-node deployment."

—Brian O'Krafka, software architect

3 WHAT IS BIG DATA?

Given that the subject of this book is Big Data, it will be useful to establish what Big Data is and how it applies to your database environment. In this chapter I will cover not only the definition of Big Data but where Big Data comes from, the types of Big Data, and how to know when you have a Big Data problem.

WHAT IS BIG DATA ANYHOW?

You hear incessant talk these days about Big Data; it seems as though everyone is jumping on the Big Data bandwagon. As is always the case with any new technological advance, it's extremely important to cut through the marketing hype and get down to the basics of the subject so you can accomplish your objectives based on your own understanding.

NOTE

With any major advance in database technology, vendors in the market will scramble to incorporate their offerings with the goal of capitalizing on the momentum. This is very common; in fact, often you will see a technology vendor or service provider "bending" the definition to fit their particular offering. This is not particularly bad; after all, it is how we get great new database products and services to market in order for data architects to take advantage of them. My only advice is to push through the hype and gain your own understanding of Big Data, so you can properly evaluate the offerings available to you based on your specific needs and requirements.

A Formal Definition for Big Data

In searching for a formal definition of Big Data, I found this excerpt from Wikipedia to be very useful[1]:

> Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."
>
> As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data. . . . [In addition to scientific applications, the] limitations also affect Internet search, finance, and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×1018) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.
>
> Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers." What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

While the definition is highly technical (even after some edits), it really does a thorough job of describing the scope of Big Data. Let's walk through some of the key points:

1. Source: http://en.wikipedia.org/wiki/Big_data

- *"Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools . . ."*

  Clearly the important thing about Big Data is the size of the data set, and the key message is that it's tough (i.e., often impossible) to manage Big Data with traditional database management tools.

- *"The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization."*

  Big Data is hard to deal with: You have to collect it, manage it (i.e., "curation"), store it, and move it around. And, of course, if you cannot search, analyze, or visualize Big Data, it's of no use to anyone.

- *"The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data . . ."*

  This is a fascinating point: Big Data most often concerns one large set of related data, rather than smaller sets that are searched individually. More and more we see the need to crawl these gigantic data sets in order to find meaningful answers to business or analytical questions.

- *". . . allowing correlations to be found to 'spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.'"*

  Here we see that Big Data is used for all kinds of purposes; in fact, this list represents only a tiny fraction of the unlimited use cases imaginable. Big Data can be utilized for any purpose an individual or organization can dream up, with the purpose of explaining, predicting, or identifying information with the end goal of improving the useful, actionable knowledge within the organization or individual.

- *"As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data . . ."*

  There are, of course, limits on how much Big Data you can collect, manage, and process. While the Wikipedia definition suggests the limit is on the order of exabytes, for most of us the practical limits today are nowhere near that large. In my experience, tens to hundreds of terabytes (and sometimes a few petabytes) better describes what most organizations will deal with for their use cases.

- *"Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices . . ."*

  Big Data is coming from an incredible number of high-volume sources, and I will review some of the more common sources later in this chapter.

- *". . . as of 2012, every day 2.5 exabytes (2.5×1018) of data were created."*

  It's incredible that the world is generating 2.5 *exabytes* of data each and every day. That is a *lot* of data, and our job is to tap into the portion of it that is useful and meaningful to the organizations and purposes we serve.

- *"Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead 'massively parallel software running on tens, hundreds, or even thousands of servers.'"*

  This is the clincher: To successfully utilize Big Data requires *parallel software* (i.e., a true database cluster). Thus, the central concept of this book: how to scale and manage a database cluster for your Big Data endeavors.

You can see that the formal Wikipedia definition is focused on Big Data analytics use cases. From direct experience I believe Big Data concepts apply to a much broader sphere with many different database scenarios, including OLTP (online transaction processing), traditional data warehouse applications, and NoSQL engines.

A Practical Big Data Definition

Most discussions of Big Data assume that Big Data is relegated to the realm of advanced analytics, often of unstructured data. This is a very limited definition, and the scope of Big Data will intrude on virtually any advanced application with a high growth need. It's a fact that databases only get larger with time; this fact alone will place more and more applications into the Big Data realm over time.

Here is a more practical definition of Big Data: A monolithic database meets the criteria for Big Data when *you* have a scalability and performance problem with *your* database.

Simply put, when your database runs out of steam and can no longer handle the load and volume your application is generating, you have hit the Big Data threshold. This is shown in Figure 3.1. The database starts to run into trouble (or outright fails) when too much load is placed on the engine. It doesn't matter what technology you use, what type of DBMS engine, or really how big your database is (the Wikipedia definition indicates this as well—it just depends on the application requirement). When you run into performance barriers as your database grows, this qualifies as a genuine Big Data problem—one that you should solve with scalable database technologies to overcome the limitation. In other words, it's important to think in terms of a parallel Big Data cluster for your database tier.

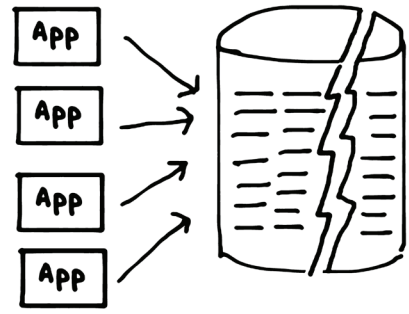

Figure 3.1 A Big Data problem indicated by a database that cannot handle the load

NOTE

It is true that in many cases initial optimizations, such as adding an index, optimizing a query, or upgrading your hardware, will tide you over (that is, until the next performance barrier is hit). But if you are consistently running into database slowdown and performance issues, it's time to adopt a more comprehensive Big Data strategy.

SOURCES OF BIG DATA

Big Data has become a hot topic due to the plethora of data sources literally flooding into today's databases as online applications proliferate. This section discusses just a few of today's drivers and sources of Big Data, as highlighted in Figure 3.2. These include search engines, social networks, mobile devices, system data, and the Internet of Things (IoT).

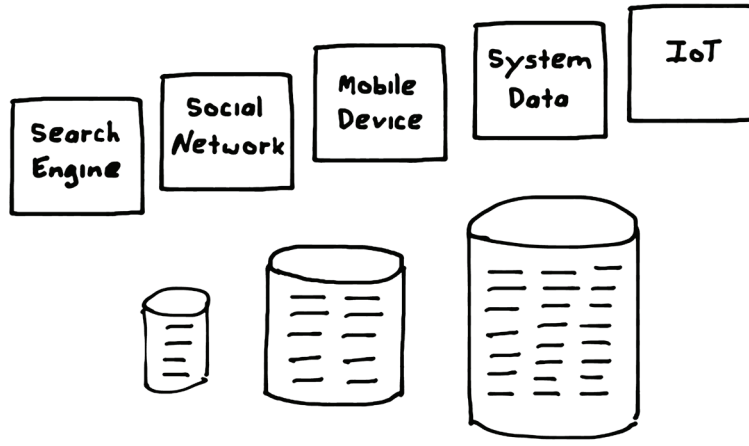

Figure 3.2 Sources of Big Data

THE ADVENT OF THE SEARCH ENGINE

Modern search engines, with Google being the most notable example, have actually defined the Big Data paradigm. These search engines literally transformed the world, integrating the Internet and World Wide Web into society in ways unimaginable before this technology. The concept of crawling Web sites to enable instantaneous searches for virtually any conceivable topic was a huge game changer. While everyone is familiar with search engines (you probably use them many times a day without even thinking about it), what is important for our purposes is the massive influence of the search engine on Big Data techniques and approaches.

NOTE

Google invented and popularized one of the first truly parallel frameworks, called Bigtable, to support its search engine capabilities. The Bigtable paper, released in 2006, is a seminal work on Big Data, and recommended reading for any data architect with a serious interest in understanding the internals of a fully scalable data architecture.[2]

THE RISE OF SOCIAL NETWORKS

Social networks have achieved unprecedented user volumes, a true historical first in the information age. Along with all these users comes the resultant huge data volume.

Facebook, the giant of all social networks, has exceeded one billion users in 2013, with over 600 million active users each and every day.[3] The amount of data generated from this volume of users is of gigantic proportions, and Facebook has excelled in developing and adopting Big Data approaches to effectively deal with the virtual flood of traffic. In fact, Facebook released figures in mid-2012 that at that time the social network was processing 2.5 billion pieces of content, totaling more than 500 terabytes per day.[4]

Twitter is another mega-player in the industry, and while not strictly a social network, its paradigm of tweets and followers certainly fits the bill. With over 600 million users and 135 million active users daily, generating 58 million tweets per day, this also represents a mammoth amount of data.[5]

2. You can access the Bigtable paper at http://research.google.com/archive/bigtable.html.
3. Source: http://news.yahoo.com/number-active-users-facebook-over-230449748.html
4. Source: http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/
5. Source: http://www.statisticbrain.com/twitter-statistics/