# Chapter 12

# Defining Big Data Analytics

### In This Chapter

- Using big data to get results
- Finding what's different with big data
- Exploring the challenges of analyzing big data
- Examining analytics tools for big data

Up until this point, we've been spending a lot of time describing the infrastructure you need to support your big data initiatives. However, because big data is most useful if you can do something with it, the question becomes, how do you analyze it?

Companies like Amazon and Google are masters at analyzing big data. And they use the resulting knowledge to gain a competitive advantage. Just think about Amazon's recommendation engine. The company takes all your buying history together with what it knows about you, your buying patterns, and the buying patterns of people like you to come up with some pretty good suggestions. It is a marketing machine, and its big data analytics capabilities have made it extremely successful.

The capability to analyze big data provides unique opportunities for your organization as well. You'll be able to expand the kind of analysis you can do. Instead of being limited to sampling large data sets, you can now utilize much more detailed and complete data to do your analysis. However, analyzing big data can also be challenging. Changing algorithms and technology, even for basic data analysis, often has to be addressed with big data.

So, in this chapter, we introduce big data analytics. We focus on the kinds of analysis you can do with big data. We also discuss some of the differences you need to think about between big data analytics and traditional analytics. In this chapter, we focus primarily on structured data analysis, although unstructured data is a very important part of the big data picture. We describe that in the next chapter.

# Using Big Data to Get Results

The first question that you need to ask yourself before you dive into big data analysis is what problem are you trying to solve? You may not even be sure of what you are looking for. You know you have lots of data that you think you can get valuable insight from. And certainly, patterns can emerge from that data before you understand why they are there.

If you think about it though, you're sure to have an idea of what you're interested in. For instance, are you interested in predicting customer behavior to prevent churn? Do you want to analyze the driving patterns of your customers for insurance premium purposes? Are you interested in looking at your system log data to ultimately predict when problems might occur? The kind of high-level problem is going to drive the analytics you decide to use. Alternately, if you're not exactly sure of the business problem you're trying to solve, maybe you need to look at areas in your business that need improvement. Even an analytics-driven strategy — targeted at the right area — can provide useful results with big data.

When it comes to analytics, you might consider a range of possible kinds, which are outlined in Table 12-1.

**Table 12-1 Big Data Analysis**

| *Analysis Type* | *Description* |
|---|---|
| Basic analytics for insight | Slicing and dicing of data, reporting, simple visualizations, basic monitoring. |
| Advanced analytics for insight | More complex analysis such as predictive modeling and other pattern-matching techniques. |
| Operationalized analytics | Analytics become part of the business process. |
| Monetized analytics | Analytics are utilized to directly drive revenue. |

REMEMBER

Before you start analyzing your data, make sure that you've dealt with all preprocessing issues. These are covered in detail in Chapter 15.

## Basic analytics

Basic analytics can be used to explore your data, if you're not sure what you have, but you think something is of value. This might include simple visualizations or simple statistics. Basic analysis is often used when you have large amounts of disparate data. Here are some examples:

- **Slicing and dicing:** *Slicing and dicing* refers to breaking down your data into smaller sets of data that are easier to explore. For example, you might have a scientific data set of water column data from many different locations that contains numerous variables captured from multiple sensors. Attributes might include temperature, pressure, transparency, dissolved oxygen, pH, salinity, and so on, collected over time. You might want some simple graphs or plots that let you explore your data across different dimensions, such as temperature versus pH or transparency versus salinity. You might want some basic statistics such as average or range for each attribute, from each height, for the time period. The point is that you might use this basic type of exploration of the variables to ask specific questions in your problem space. The difference between this kind of analysis and what happens in a basic business intelligence system is that you're dealing with huge volumes of data where you might not know how much query space you'll need to examine it and you're probably going to want to run computations in real time.
- **Basic monitoring:** You might also want to monitor large volumes of data in real time. For example, you might want to monitor the water column attributes in the preceding example every second for an extended period of time from hundreds of locations and at varying heights in the water column. This would produce a huge data set. Or, you might be interested in monitoring the buzz associated with your product every minute when you launch an ad campaign. Whereas the water column data set might produce a large amount of relatively structured time-sensitive data, the social media campaign is going to produce large amounts of disparate kinds of data from multiple sources across the Internet.
- **Anomaly identification:** You might want to identify anomalies, such as an event where the actual observation differs from what you expected, in your data because that may clue you in that something is going wrong with your organization, manufacturing process, and so on. For example, you might want to analyze the records for your manufacturing operation to determine whether one kind of machine, or one operator, has a higher incidence of a certain kind of problem. This might involve some simple statistics like moving averages triggered by an alert from the problematic machine.

## Advanced analytics

Advanced analytics provides algorithms for complex analysis of either structured or unstructured data. It includes sophisticated statistical models, machine learning, neural networks, text analytics (described in detail in Chapter 13), and other advanced data-mining techniques. (See the sidebar

"What is data mining?" later in this chapter, for more detail on data mining.) Among its many use cases, advanced analytics can be deployed to find patterns in data, prediction, forecasting, and complex event processing.

While advanced analytics has been used by statisticians and mathematicians for decades, it was not as big a part of the analytics landscape as it is today. Consider that 20 years ago, statisticians at companies were able to predict who might drop a service using advanced survival analysis or machine learning techniques. However, it was difficult to persuade other people in the organization to understand exactly what this meant and how it could be used to provide a competitive advantage. For one thing, it was difficult to obtain the computational power needed to interpret data that kept changing through time.

Today, advanced analytics is becoming more mainstream. With increases in computational power, improved data infrastructure, new algorithm development, and the need to obtain better insight from increasingly vast amounts of data, companies are pushing toward utilizing advanced analytics as part of their decision-making process. Businesses realize that better insights can provide a superior competitive position.

Here are a few examples of advanced analytics for big data:

- **Predictive modeling:** Predictive modeling is one of the most popular big data advanced analytics use cases. A predictive model is a statistical or data-mining solution consisting of algorithms and techniques that can be used on both structured and unstructured data (together or individually) to determine future outcomes. For example, a telecommunications company might use a predictive model to predict customers who might drop its service. In the big data world, you might have large numbers of predictive attributes across huge amounts of observations. Whereas in the past, it might have taken hours (or longer) to run a predictive model, with a large amount of data on your desktop, you might be able to now run it iteratively hundreds of times if you have a big data infrastructure in place.
- **Text analytics:** Unstructured data is such a big part of big data, so text analytics — the process of analyzing unstructured text, extracting relevant information, and transforming it into structured information that can then be leveraged in various ways — has become an important component of the big data ecosystem. The analysis and extraction processes used in text analytics take advantage of techniques that originated in computational linguistics, statistics, and other computer science disciplines. Text analytics is being used in all sorts of analysis, from predicting churn, to fraud, and to social media analytics. It is so important that we devote a considerable part of Chapter 13 to this issue of text analytics.
- **Other statistical and data-mining algorithms:** This may include advanced forecasting, optimization, cluster analysis for segmentation or even microsegmentation, or affinity analysis.

## What is data mining?

Data mining involves exploring and analyzing large amounts of data to find patterns in that data. The techniques came out of the fields of statistics and artificial intelligence (AI), with a bit of database management thrown into the mix. Generally, the goal of the data mining is either classification or prediction. In classification, the idea is to sort data into groups. For example, a marketer might be interested in the characteristics of those who responded versus who didn't respond to a promotion. These are two classes. In prediction, the idea is to predict the value of a continuous (that is, nondiscrete) variable. For example, a marketer might be interested in predicting those who *will* respond to a promotion.

Typical algorithms used in data mining include the following:

- **Classification trees:** A popular data-mining technique that is used to classify a dependent categorical variable based on measurements of one or more predictor variables. The result is a tree with nodes and links between the nodes that can be read to form if-then rules.
- **Logistic regression:** A statistical technique that is a variant of standard regression but extends the concept to deal with classification. It produces a formula that predicts the probability of the occurrence as a function of the independent variables.
- **Neural networks:** A software algorithm that is modeled after the parallel architecture of animal brains. The network consists of input nodes, hidden layers, and output nodes. Each of the units is assigned a weight. Data is given to the input node, and by a system of trial and error, the algorithm adjusts the weights until it meets a certain stopping criteria. Some people have likened this to a black–box (you don't necessarily know what is going on inside) approach.
- **Clustering techniques like K-nearest neighbors:** A technique that identifies groups of similar records. The K-nearest neighbor technique calculates the distances between the record and points in the historical (training) data. It then assigns this record to the class of its nearest neighbor in a data set.

Here's a classification tree example. Consider the situation where a telephone company wants to determine which residential customers are likely to disconnect their service. The telephone company has information consisting of the following attributes: how long the person has had the service, how much he spends on the service, whether he has had problems with the service, whether he has the best calling plan for his needs, where he lives, how old he is, whether he has other services bundled together with his calling plan, competitive information concerning other carriers plans, and whether he still has the service or has disconnected the service. Of course, you can find many more attributes than this. The last attribute is the outcome variable; this is what the software will use to classify the customers into one of the two groups — perhaps called stayers and flight risks.

The data set is broken into training data and a test data set. The training data consists of observations (called attributes) and an outcome variable (binary in the case of a classification model) — in this case, the stayers or the flight risks. The algorithm is run over the training data and comes up with a tree that can be read like a series of rules. For example, if the customers have been with the company for more than ten years and they are over 55 years old, they are likely to remain as loyal customers.

*(continued)*

*(continued)*

These rules are then run over the test data set to determine how good this model is on "new data." Accuracy measures are provided for the model. For example, a popular technique is the confusion matrix. This matrix is a table that provides information about how many cases were correctly versus incorrectly classified. If the model looks good, it can be deployed on other data, as it is available (that is, using it to predict new cases of flight risk). Based on the model, the company might decide, for example, to send out special offers to those customers whom it thinks are flight risks.

REMEMBER

Advanced analytics doesn't require big data. However, being able to apply advanced analytics with big data can provide some important results.

## Operationalized analytics

When you operationalize analytics, you make them part of a business process. For example, statisticians at an insurance company might build a model that predicts the likelihood of a claim being fraudulent. The model, along with some decision rules, could be included in the company's claims-processing system to flag claims with a high probability of fraud. These claims would be sent to an investigation unit for further review. In other cases, the model itself might not be as apparent to the end user. For example, a model could be built to predict customers who are good targets for upselling when they call into a call center. The call center agent, while on the phone with the customer, would receive a message on specific additional products to sell to this customer. The agent might not even know that a predictive model was working behind the scenes to make this recommendation.

## Monetizing analytics

Analytics can be used to optimize your business to create better decisions and drive bottom- and top-line revenue. However, big data analytics can also be used to derive revenue above and beyond the insights it provides just for your own department or company. You might be able to assemble a unique data set that is valuable to other companies, as well. For example, credit card providers take the data they assemble to offer value-added analytics products. Likewise, with financial institutions. Telecommunications companies are beginning to sell location-based insights to retailers. The idea is that various sources of data, such as billing data, location data, text-messaging data, or web-browsing data can be used together or separately to make inferences about customer behavior patterns that retailers would find useful. As a regulated industry, they must do so in compliance with legislation and privacy policies.

# Modifying Business Intelligence Products to Handle Big Data

Traditional business intelligence products weren't really designed to handle big data. They were designed to work with highly structured, well-understood data, often stored in a relational data repository and displayed on your desktop or laptop computer. This traditional business intelligence analysis is typically applied to snapshots of data rather than the entire amount of data available. So what's different when you start to analyze big data?

## Data

As we discuss in Chapter 2, big data consists of structured, semi-structured, and unstructured data. You often have a lot of it, and it can be quite complex. When you think about analyzing it, you need to be aware of the potential characteristics of your data:

- **It can come from untrusted sources.** Big data analysis often involves aggregating data from various sources. These may include both internal and external data sources. How trustworthy are these external sources of information? For example, how trustworthy is social media data like a tweet? The information may be coming from an unverified source. The integrity of this data needs to be considered in the analysis. We talk more about big data security and governance in Chapter 19.
- **It can be dirty.** Dirty data refers to inaccurate, incomplete, or erroneous data. This may include the misspelling of words; a sensor that is broken, not properly calibrated, or corrupted in some way; or even duplicated data. Data scientists debate about where to clean the data — either close to the source or in real time. Of course, one school of thought says that the dirty data should not be cleaned at all because it may contain interesting outliers. The cleansing strategy will probably depend on the source and type of data and the goal of your analysis. For example, if you're developing a spam filter, the goal is to detect the bad elements in the data, so you would not want to clean it.
- **The signal-to-noise ratio can be low.** In other words, the signal (usable information) may only be a tiny percent of the data; the noise is the rest. Being able to extract a tiny signal from noisy data is part of the benefit of big data analytics, but you need to be aware that the signal may indeed be small.
- **It can be real-time.** In many cases, you'll be trying to analyze real-time data streams. We cover a whole set of complexities about how to analyze this data in Chapter 16.

Big data governance is going to be an important part of the analytics equation. Underneath business analytics, enhancements will need to be made to governance solutions to ensure the veracity coming from the new data sources, especially as it is being combined with existing trusted data stored in a warehouse. Data security and privacy solutions also need to be enhanced to support managing/governing big data stored within new technologies. Governance and security and so important that we devote Chapter 19 to it.

## Analytical algorithms

When you're considering big data analytics, you need to be aware that when you expand beyond the desktop, the algorithms you use often need to be *refactored,* changing the internal code without affecting its external functioning. The beauty of a big data infrastructure is that you can run a model that used to take hours or days in minutes. This lets you iterate on the model hundreds of times over. However, if you're running a regression on a billion rows of data across a distributed environment, you need to consider the resource requirements relating to the volume of data and its location in the cluster. Your algorithms need to be data aware.

Additionally, vendors are starting to offer a new breed of analytics designed to be placed close to the big data sources to analyze data in place rather than first having to store it and then analyze it. This approach of running analytics closer to the data sources minimizes the amount of stored data by retaining only the high-value data. It is also enables you to analyze the data sooner, looking for key events, which is critical for real-time decision making. We discuss these kinds of techniques in more detail in Chapter 14.

Of course, analytics will continue to evolve. For example, you may need real-time visualization capabilities to display real-time data that is continuously changing. How do you practically plot a billion points on a graph plot? Or, how do you work with the predictive algorithms so that they perform fast enough and deep enough analysis to utilize an ever-expanding, complex data set? This is an area of active research.

## Infrastructure support

We spend a good deal of this book talking about the infrastructure needed to support big data, so we don't go into detail about that here. You might want to turn to Chapter 4 for more details on infrastructure issues. Suffice it to say that if you're looking for a platform, it needs to achieve the following:

- **Integrate technologies:** The infrastructure needs to integrate new big data technologies with traditional technologies to be able to process all kinds of big data and make it consumable by traditional analytics.
- **Store large amounts of disparate data:** An enterprise-hardened Hadoop system may be needed that can process/store/manage large amounts of data at rest, whether it is structured, semi-structured, or unstructured.
- **Process data in motion:** A stream-computing capability may be needed to process data in motion that is continuously generated by sensors, smart devices, video, audio, and logs to support real-time decision making.
- **Warehouse data:** You may need a solution optimized for operational or deep analytical workloads to store and manage the growing amounts of trusted data.

And of course, you need the capability to integrate the data you already have in place along with the results of the big data analysis.

# Studying Big Data Analytics Examples

Big data analytics has many different use cases. We mention examples throughout this book, but we now look at a few others from Internet companies and others.

## Orbitz

If you've ever looked for deals on travel, you've probably been to sites like Orbitz (`www.orbitz.com`). The company was established in 1999, and its website went live in 2001. Users of Orbitz perform over a million searches a day, and the company collects hundreds of gigabytes of raw data each day from these searches. Orbitz realized that it might have useful information in the web log files that it was collecting from its web analytics software that contained information about consumer interaction with its site.

In particular, it was interested to see whether it could identify consumer preferences to determine the best-performing hotels to display to users so that it could increase conversions (bookings). It had not been utilizing this data in the past because it was too expensive to store all of it. It implemented

Hadoop and Hive running on commodity hardware to help. Hadoop provided the distributed file system and Hive provided an SQL-type interface. It took a series of steps to put the data into Hive. After the data was in Hive, the company used *machine learning* — a data-driven (and data-mining; see the sidebar earlier in this chapter) approach to unearthing patterns in data and helping to analyze the data. For more details about Hadoop and Hive, turn to Chapters 9 and 10.

## Nokia

Nokia provides wireless communication devices and services. The company believes that its data is a strategic asset. Its big data analytics service includes a multipetabyte platform that executes over tens of thousands of jobs each day. This includes utilizing advanced analytics over terabytes of streaming data. For example, the company wants to understand how people interact with its different applications on its phones. Nokia wants to understand what features customers use, how they use a feature, and how they move from feature to feature and whether they get lost in the application as they are using it. This level of detail helps the company lay out new features for its applications and improve customer retention.

## NASA

NASA is using predictive models to analyze safety data on aircrafts. It wants to understand whether the introduction of a new technology into an aircraft will make a dramatic impact in safety. Needless to say, NASA is dealing with a massive amount of data. Each airplane each day is recording a *thousand* parameters every second for every flight. Some of this data is streaming. The company also receives text data from reports written by pilots and other crew members. NASA also throws weather data (that changes in time and space) into the mix. The data scientists there are looking to predict outcomes — for example, what pattern indicates a possible accident or incident.

# Big Data Analytics Solutions

A number of vendors on the market today support big data solutions. Here is a listing of a few solutions that you may find interesting:

- IBM (`www.ibm.com`) is taking an enterprise approach to big data and integrating across the platform including embedding/bundling its analytics. Its products include a warehouse (InfoSphere warehouse) that has its own built-in data-mining and cubing capability. Its new PureData Systems (a packaging of advanced analytics technology into an integrated systems platform) includes many packaged analytical integrations. Its InfoSphere Streams product is tightly integrated with its Statistical Package for the Social Sciences (SPSS) statistical software to support real-time predictive analytics, including the capability to dynamically update models based on real-time data. It is bundling a limited-use license of Cognos Business Intelligence with its key big data platform capabilities (enterprise-class Hadoop, stream computing, and warehouse solutions).
- SAS (`www.sas.com`) provides multiple approaches to analyze big data via its high-performance analytics infrastructure and its statistical software. SAS provides several distributed processing options. These include in-database analytics, in-memory analytics, and grid computing. Deployments can be on-site or in the cloud.
- Tableau (`www.tableausoftware.com`), a business analytics and data visualization software company, offers its visualization capabilities to run on top appliances and other infrastructure offered by a range of big data partners, including Cirro, EMC Greenplum, Karmasphere, Teradata/Aster, HP Vertica, Hortonworks, ParAccel, IBM Netezza, and a host of others.
- Oracle (`www.oracle.com`) offers a range of tools to complement its big data platform called Oracle Exadata. These include advanced analytics via the R programming language, as well as an in-memory database option with Oracle's Exalytics in-memory machine and Oracle's data warehouse. Exadata is integrated with its hardware platform.
- Pentaho (`www.pentaho.com`) provides open source business analytics via a community and enterprise edition. Pentaho supports the leading Hadoop-based distributions and supports native capabilities, such as MapR's NFS high-performance mountable file system.