Big data in the cloud

Data velocity, volume, variety, veracity

Sam B. Siewert

July 09, 2013

Big data is an inherent feature of the cloud and provides unprecedented opportunities to use both traditional, structured database information and business analytics with social networking, sensor network data, and far less structured multimedia. Big data applications require a data-centric compute architecture, and many solutions include cloud-based APIs to interface with advanced columnar searches, machine learning algorithms, and advanced analytics such as computer vision, video analytics, and visualization tools. This article examines the use of the R language and similar tools for big data analysis and methods to scale big data services in the cloud. It provides an in-depth look at digital photo management as a simple big data service that employs key elements of search, analytics, and machine learning applied to unstructured data.

This article focuses on applications that use big data, and explains fundamental concepts behind big data analytics and how to combine that with business intelligence (BI) applications and parallel technologies like the computer vision (CV) and machine learning methods covered in Part 3 of the "Cloud scaling" series.

What distinguishes big data analytics from video analytics is the breadth of data types processed and the interactive analysis and search tools provided compared to, say, data mining or MapReduce methods used, which may be more sophisticated but take far longer to run than Google BigQuery, for example, which uses columnar search to compress and speed up interactive searches for massive amounts of unstructured data. In fact, in "An Inside Look at Google BigQuery" (see Related topics), Google explains that in tens of seconds, BigQuery can do regex text matching on a huge logging table of about 35 billion rows and 20 TB. The tool’s MapReduce functionality takes far longer to run, but does provide sophisticated data reduction.

Anyone with a Google account can use BigQuery. Or, for a quicker idea of typical big data search, download my picture and upload it to Google Images. You should get all the same pages with my image (from IBM, the University of Colorado Boulder, etc.), including at least one false positive. I have used this example mostly to make sure images I have downloaded have proper photo credits and permission granted for reuse. Along with other examples I explore here, this example gives an
idea of the spirit of big data, finding a needle in an unstructured haystack of data — in fact, more like millions of haystacks.

**Defining big data**

*Big data* is broadly defined as the capture, management, and analysis of data that goes beyond typical structured data, which can be queried by relational database management systems — often to unstructured files, digital video, images, sensor data, log files, and really any data not contained in records with distinct searchable fields. In some sense, the unstructured data is the interesting data, but it’s difficult to synthesize into BI or draw conclusions from it unless it can be correlated to structured data.

Big data also has new sources, like machine generation (e.g., log files or sensor networks), mobile devices (video, photographs, and text messaging), and machine-to-machine, where the Internet of Things reports status for purposes of maintenance planning for fleets of vehicles or aircraft or general telemetry monitoring. One way to look at this is by the characteristics of *volume*. IBM estimates that 2.5 quintillion (2,500,000,000,000,000,000) bytes of data are created now each day (see Related topics). Second, the *velocity*, where data rates are increasing because of network bandwidth — typically at gigabit rates today (gigE, 10G, 40G, 100G) compared to megabit rates. Third, the *variety*, now including more unstructured data types, like digital video streams and sensor data as well as log files. Finally, the *veracity* of data, or how much can data be trusted when key decisions need to be made on such large volumes collected at high rates. Simply knowing that data is in fact not spoofed, has not been corrupted, or comes from an expected source is difficult — it could come, for example, from one of thousands of security cameras, each producing many thousands of frames of video each hour. So, let’s outline some of the key aspects of big data, applications, and systems to better understand them.

**Where does big data come from?**

Big data has come about largely because of advances in mobile devices that now include digital video, photography, audio, and advanced email and text features. Users are collecting data in numbers that were never seen a decade ago; likewise, new applications like Google Translate provide big data server features—natural language translation for phrases spoken or typed into mobile devices. IBM sees big data as enabled by mobile first in the Global Technology Outlook for 2013 (see Related topics) and characterizes big data by *volume*, *variety*, *velocity*, and *veracity*. The data is naturally far less structured than relational database records but can be correlated to such data. This article provides detail on what constitutes big data.

Perhaps the best way to understand big data is to review its history, as *Forbes Magazine* has done (see Related topics). The scale of what has been considered big data has of course increased to the current rate of more than 2.5 exabytes per day. Interestingly, most data will never be reviewed by a human (with only 7 billion people per the US Census clock, we would each have to review more than 300MB of information each day). Given this challenge, the only logical way to use this much data is machine-to-machine automation or intelligent query of big data. Furthermore, if this much data is kept over long periods of time, how would anyone even know if some of it had been corrupted? We can of course store data digests (such as MD5, which is a form of checksum) and use redundant array of independent disks (RAID—mirrors, XOR parity, or erasure codes to...
detect and recover corrupted data), but concern is growing that some data could suffer from silent corruption (see Related topics).

The Internet Archive, a data curator, has led investigations into this concern. Overall, the veracity of big data is a challenge, but erasure codes and advanced data digesting methods show promise. Traditional methods like XOR RAID or simple mirroring — which provide only single fault protection against data loss when storage devices fail and do not handle subtle corruption scenarios caused by software bugs, data center operator errors, or media failure over time — are being replaced by RAID-6 and more advanced erasure codes. The concept of data durability for big data has become important, a topic I have researched using math models working with Intel and Amplidata. With this much data, the idea of humans reviewing it for veracity is simply not possible, and missing data might not be noticed until it is finally queried or accessed far in the future.

**Big data system design**

Architectures for data protection at scale should include protection against loss, silent corruption, malware, and malevolent modification of data by cyber-criminals or through cyber-warfare. Data is an asset and increasingly used by governments and business to make key decisions, but if the veracity of the data is unknown, the value of the data declines or may even be lost — or worse yet: bad decisions made. This topic goes beyond the scope of this article, but clearly protection against loss and undetected modification or corruption of data is necessary.

One way to better understand big data is simply to look more closely at some cloud sites that have sufficient data (petabytes, typically) along with tools to query (usually terabytes) for use by applications. Most of us use Google queries daily, but Google also provides BigQuery, which uses more sophisticated columnar storage and search (discussed in more detail as an example). Other well-known examples are Facebook (social networking), Wikipedia (general knowledge capture), the Internet Archive (digital data curators), DigitalGlobe (geographical information systems [GIS]), Microsoft® Virtual Earth (GIS), Google Earth (GIS), and numerous new big data service providers.

Companies have internal big data as well as on private cloud systems. Many big data systems are read-only for user query (with capture from machine-generated sources), but likely include strong authentication if they allow updates to databases or unstructured data, using pass phrases, requiring users to authenticate via mobile phone text message confirmation codes, with use of graphical challenges to verify human data entry and perhaps using biometric authentication more in the future.

**Big data applications**

Killer applications for video analytics are being thought of every day for CV and video analytics, some perhaps years from realization because of computing requirements or implementation cost. Nevertheless, here is a list of interesting applications:

- **Stock market sentiment analysis** using Google Trends has been shown to correlate well to historical index declines and rises, which is perhaps not surprising but interesting in terms of significance as a big data application. The article "Quantifying Trading Behavior in Financial
“Markets Using Google Trends” (see Related topics) provides evidence that use of sentiment analysis to make long and short buy-and-sell decisions for stock holdings can outperform simple buy-and-hold strategies and index fund investment. This research no doubt requires more analysis, but is compelling. An interesting consideration, though, is what will happen as these machine-based trading systems come online along with existing programmed trading.

- **Picasa photo sorting** from Google is a useful tool that allows a user to sort, query, and automatically identify faces using CV techniques combined with machine learning. This is a great way to get a feel for the value of big data services and applications. It makes it clear that big data analytics will require advanced analytics such as CV and methods like machine vision.

- **Recommendation systems** such as Pandora (music), Netflix (movies), and Amazon (books and products) use customer data and multiple agents in an approach known as **collaborative filtering**. This big data service has been the topic of much-advanced research in machine learning and data mining. Clearly, the ability to make good recommendations can increase sales and customer satisfaction.

- **Customer base analytics** can provide sentiment analysis for your customers based on social networking data (Facebook and Twitter, for example) when this textual data is correlated to BI gathered from traditional customer transactional records. Sentiment analysis allows a business to know what customers think about their products, their interest in them or competitors, what they like and dislike, etc.

- **Machine-generated data** from sources like sensor networks (for example, sensors embedded in large systems like urban transportation, traffic lights, and general infrastructure); machine-to-machine data, whereby the sensor or log data from one machine (typically in the field) is ingested by yet another machine; and log files, most often used by IT to debug problems and manage systems by exception (ignore them other than when they need human attention for recovery and continued operation).

- **Booking systems** for travel are being improved by incorporating customer preference, logistics, and prior history to make helpful suggestions for that always-arduous task of planning travel.

- **Social networking for entertainment** is replacing the social aspect of broadcast television and movie water-cooler discussion, where digital media on demand now allows anyone to watch content most anywhere and anytime, but still share the experience via social networking. Although this makes content consumption more enjoyable, it allows content creators, script writers, and artists the ability really to know their audience better than ever.

- **Medical diagnostics** have often included rule-based expert decision support systems (DSSes), but with big data, evidence exists that these systems may come out of research and become mainstream medical assistants. For example, a new DSS to assist with objective psychological evaluation of patients at risk for suicide has shown promise in research (see Related topics). Part of proving these systems is to compare them with historical data: These systems will not replace human decision-making but promise to improve it when used as a support tool.

This is by no means an exhaustive list of big data applications, but you can find more to explore in Related topics. The application of columnar query, analytics for unstructured data, MapReduce, and visualization and reasoning about big data are just getting started.
Big data in public safety and security

The integration of big data analytics with public information (or private, voluntarily provided information trusted to a custodian) can allow for rapid search of large volumes of video, voice, sensor data, and email text to improve public safety for disaster recovery, to prevent terrorist threats, and to understand public concerns. One could almost think of this as feedback compared with the one-way broadcast for emergency warning systems. Of course, concern and a potential dark side exist to big data and video/voice/email analytics if it becomes privacy invasion. Such systems require responsible use, full disclosure, and auditing of data collected in public places and networks.

Big data application privacy considerations

If companies, governments, and organizations carefully collect, analyze, and use big data, the value to the public will be apparent. If big data analytics capabilities are abused, public trust will be lost, and the value will be lost. The sentiments of users must be volunteered, and much of the value comes from knowing how people feel about what they are interacting with, where they are, or what they are reading. If a mind-reading sensor is developed, we may really have an ethical dilemma. For now, use of cameras, voice recording, or email data mining should be pursued with careful concern for privacy and in a way that maintains user trust and confidence.

As a perfect case in point, during the writing of this article, the issue of the U.S. National Security Agency phone metadata database, which can be data-mined in case of national threats, has caused significant concern (see Related topics). Obviously, much of the details will be settled in court cases, but careful consideration in the design of big data systems will no doubt save the headaches of litigation.

Example: Using R scripting

R-project Toolkit in InfoSphere Streams

InfoSphere Streams is an advanced computing platform that allows user-developed applications to ingest, analyze, and correlate information quickly as it arrives from thousands of real-time sources, handling very high data throughput rates: up to millions of events or messages per second. Version 3.1 includes an R-project Toolkit that enables you to apply complex data mining algorithms to detect patterns of interest in data streams. Learn more and give it a try.

Visual analytics is the term used to describe the visualization of big data (not to be confused with video analytics, the analysis of sequences of images to understand what they contain). Visualization has historically been a practice most often found in high-performance computing, but with the growth of unstructured data from mobile devices, social networks, machine-to-machine systems, and sensor network-generated data, the need for advanced visualization is growing for big data. Simple pie charts, the Pareto principle, X-Y plots, and bar graphs often used in business decision-making historically may not be sufficient to understand big data.

To explore this, I implemented the Lorenz equations in C and the R scripting language (a big data analysis tool). Using C and Microsoft® Excel® to understand these complex equations is limited, mostly because the modeling and analysis are not integrated and spreadsheets typically don't provide complex, multidimensional visualization. With C and Excel, I was able to produce 2-D scatter plots of the Lorenz equations, which model atmospheric convection, as shown in Figure
1. There may be a better way to visualize this data with Excel, but there was no obvious way to explore more than two dimensions.

**Figure 1. A 2-D spreadsheet plot of a Lorenz model**

![Lorenz Equation - XY Plane](image)

**Visual analytics with R**

Using R, you can import a large number of analytics and visualization packages and use them with this intuitive scripting language. For example, to better visualize the Lorenz equations, I imported the scatterplot3d package, as shown in Figure 2, allowing for a better view of the inherently 3-D Lorenz equations compared to a simple spreadsheet. You could also use scientific visualization tools such as MATLAB or even GNU plot for this type of model-based analysis, but R also includes a variety of packages that are well suited to many-dimensional analysis of data sets that aren't scientific in nature, such as BI visualizations, for which you can find many examples in [Related topics](#). The Lorenz example is a simple introduction to the power of R.
Much like MATLAB provides an interactive scientific and engineering analysis environment for model and data exploration for engineers and scientists, R provides the same for business analysts and big data analysis of all types (see Figure 3 and Listing 1). Interactive exploration of big data with tools like R and BigQuery is what distinguishes big data analysis from more batch-oriented analysis and data mining, which is often performed using MapReduce. Either way, the goal is to form new models and support decision making leveraging the volume of big data.
Figure 3. An R 3-D plot of Lorenz equations

Listing 1. Sample R script for the Lorenz equation plot

```r
code here
```

To help you understand and explore visualization, the Lorenz example for C and Excel as well as R is available for download. To explore more, you could use R to visualize data returned from Google BigQuery.

**The future of big data**

This article makes an argument for the value of big data, which has been questioned, especially when the veracity of data can't be confirmed, and provides suggestions for improvement to veracity along with concepts to deal with the volume, variety, and velocity of data. Experience to date shows that scale-out, use of advanced data durability methods, incorporation of high-rate networks for clusters, and scale-out algorithms like MapReduce and columnar search show promise for dealing effectively with big data. However, problems that were not even considered, like silent data corruption, have become new concerns because of increased volume, velocity, and variety of data, previously of less concern when bit error rates for disk drives and networks were far lower than the bytes that passed through them or were stored on them. Today's big data
architect, therefore, has to be smarter not only to protect the veracity and value of data but to
design services that make it accessible and useful now that it greatly outnumbers the human ability
to review it daily.
## Downloadable resources

<table>
<thead>
<tr>
<th>Description</th>
<th>Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scripted analytics examples</td>
<td>script-examples.zip</td>
<td>702KB</td>
</tr>
</tbody>
</table>
Related topics

- Be sure to check out the "Cloud scaling" series for more information.
- Discover the history of the three V's of big data in "3-D Data Management: Controlling Data Volume, Velocity and Variety," (Gartner, Feb 2001).
- The IBM Global Technology Outlook 2013 is a great place to start to understand how mobile systems have helped drive cloud computing and big data along with IBM's book on understanding big data. Other participants in cloud and big data that have interesting whitepapers providing insight into systems and applications include "Oracle Information Architecture: An Architect's Guide to Big Data," "Big Data Analytics: Advanced Analytics in Oracle Database, An Inside Look at Google BigQuery," "Amazon Web Services' Big Data on AWS," and the Intel big data. Of course, many other organizations are involved in big data, both large and small. And the Cloud Computing and Big Data Expo and CloudCon 2013 are good references to help you understand the full breadth of participants in this revolution in computing.
- Read An Inside Look at Google BigQuery for more information about Google's big data search tool.
- Read more about IBM's data-creation estimates.
- Silent data corruption is an interesting big data challenge that is well documented by curators of big data like the Internet Archive. You can find more in "Disk Failure Investigations at the Internet Archive," "Bit Preservation: A Solved Problem?," and "Revolutionary Methods to Handle Data Durability Challenges for Big Data."
- NPR ran a story about a suicide risk DSS titled "Analyzing the Language of Suicide Notes to Help Save Lives" (16 May 2013).
- Check out the Time magazine article "7 Things to Know About the Government's Secret Database of Telephone Data."
- The Lorenz equations used in the C and R script visualization example were formulated by Edward Lorenz to model atmospheric convection, as summarized in the Wikipedia Lorenz System wiki. The example in the R scripting language was chosen simply to show the value of R for visualization of complex data sets compared to, say, a spreadsheet, where multidimensional analysis is difficult. You can learn the art of visual analytics for big data and more by looking at the work of other analysts, such as Tom Sawyer Software's big data Visual Insights blog.
- In the developerWorks cloud developer resources, discover and share knowledge and experience of application and services developers building their projects for cloud deployment.
- Learn more about the R project and download R for Windows or R for Mac OS X or Linux®, as well. Like most languages, the best way to learn R is by example and emulating specific visualization examples like the R scatter plot example and documentation. As I did, you often want to import data into R — in my case, from my C model and Excel — which can be done in R using Data Import. You can also view a full R tutorial.
- One option for setting up cloud services and big data applications is OpenStack. IBM is a Platinum contributor to this open source cloud software. Many cloud software and solutions
companies have great products based on OpenStack—here are a few samples of newer companies to look at that specialize in big data search and OpenStack:

- Mirantis
- AMD SeaMicro
- Cloudera

- All of us no doubt have large amounts of unstructured data on home computers in the form of digital photos and video. You can download Google Picasa to a private system and use it to catalog photos with face detection and machine learning for face recognition. You can understand machine learning performance for recognition in terms of a receiver operator curve and in much more detail in "Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems."
- Download InfoSphere BigInsights Quick Start Edition, available as a native software installation or as a VMware image.
- Download InfoSphere Streams, available as a native software installation or as a VMware image.