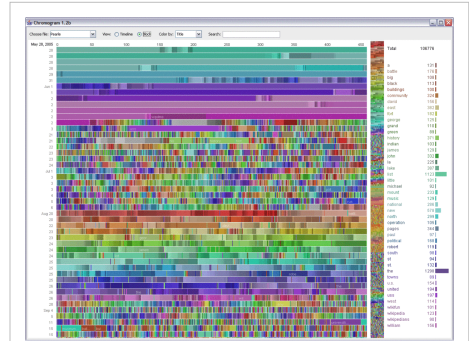


Big data

Big data^[2] are datasets that grow so large that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage,^[3] search, sharing, analytics,^[4] and visualizing. This trend continues because of the benefits of working with larger and larger datasets allowing analysts to "spot business trends, prevent diseases, combat crime."^[5] Though a moving target, current limits are on the order of terabytes, exabytes and zettabytes of data.^[6] Scientists regularly encounter this problem in meteorology, genomics^[7], connectomics, complex physics simulations^[8], biological and environmental research^[9], Internet search, finance and business informatics. Data sets also grow in size 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, wireless sensor networks and so on."^{[10] [11]} Every day, 2.5 quintillion bytes of data are created and 90% of the data in the world today was created within the past two years.^[12]



A data visualization created by IBM^[1] shows that **big data** such as Wikipedia edits by bot Pearlé are more meaningful when enhanced with colors and position.

One current feature of big data is the difficulty working with it using relational databases and desktop statistics/visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers."^[13] The size of "big data" varies depending on the capabilities of the organization managing the set. "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."^[14]

Definition

Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set.

In a 2001 research report^[15] and related conference presentations, then META Group (now Gartner) analyst, Doug Laney, defined data growth challenges (and opportunities) as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in/out), and variety (range of data types, sources). Gartner continues to use this model for describing "big data."^[16]

Examples

Examples include web logs; RFID; sensor networks; social networks; social data (due to the Social data revolution), Internet text and documents; Internet search indexing; call detail records; astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research; military surveillance; medical records; photography archives; video archives; and large-scale eCommerce.

Technologies

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. Technologies being applied to big data include massively parallel processing (MPP) databases, datamining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems.

Some but not all MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, backup, and optimize the use of the large data tables in the RDBMS. ^[17] ^[18]

Impact

The Sloan Digital Sky Survey collected more data in its first few weeks than the entire data collection in the history of astronomy back in the year 2000. Since that time, it has amassed 140 terabytes of information. The successor to this telescope, the Large Synoptic Survey Telescope, will come online in the year 2016 and will acquire that amount of data every five days. ^[19] Wal-Mart handles more than 1 million customer transactions every hour which in turn imports into databases estimated at more than 2.5 petabytes which is the equivalent of 167 times the books in America's Library of Congress. Facebook handles 40 billion photos from its user base. Decoding the human genome originally took 10 years to process when it can now be achieved in one week. ^[20]

The impact of, "Big Data," has increased the demand of information management specialists in that Oracle, IBM, Microsoft, and SAP have spent more than \$15 billion on software firms only specializing in data management and analytics. This industry on its own is worth more than \$100 billion and growing at almost 10% a year which is roughly twice as fast as the software business as a whole. ^[21]

Big Data has emerged because we are living in a society which has more of everything. There are 4.6 billion mobile-phone subscriptions worldwide and there are between 1 billion and 2 billion people accessing the internet. Basically, there are more people interacting with data or information than ever before. ^[22] Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. Cisco predicts that the amount of traffic flowing over the internet will reach 667 exabytes annually by 2013. ^[23]

Critique

Concerns have been raised about the use of big data in science neglecting principles such as choosing a representative sample by being too concerned about actually handling the huge amounts of data. ^[24] As such, the results often are biased in one way or another. Integration across heterogeneous data resources - some that might be considered "big data" and others not - presents formidable logistical as well as analytical challenges, but many researchers argue that such integrations are likely to represent the most exciting new frontiers in science ^[25].

Architecture Comparison

- Survey Distributed Databases ^[26]
- Marin Dimitrov's Comparison on PNUTS, Dynamo, Voldemort, BigTable, HBase, Cassandra and CouchDB May 2010 ^[27]
- Big Data Architecture: Comparing Aster Data, Greenplum, Gluster etc 2009 ^[28]
- HBase vs. Cassandra: NoSQL Battle! ^[29]
- Why Pick Cassandra for Real-time Transaction ^[30]
- Why Use HBase-1: from Million Mark to Billion Mark ^[31]
- Why Use HBase-2: Demystifying HBase Data integrity, Availability and Performance ^[32]
- HBase MapReduce 101 - Part I ^[33]
- HBase Architecture 101 - Write-ahead-Log ^[34]
- HBase Architecture 101 - Storage ^[35]
- Beyond Hadoop: Next-Generation Big Data Architectures ^[36] by By Bill McColl Oct. 23, 2010 about "Not Only Hadoop".
- MPI ^[37] and BSP ^[38] See wiki about Bulk Synchronous Parallel ^[39] and Apache HAMA ^[40] on Hadoop cluster.

Performance Evaluation

Existing work done by community

- 2010: Yahoo Cloud Serving Benchmark(YCSB) ^[41]
- 2010:HBase - non SQL Database, Performances Evaluation ^[42]
- 2009:HBase-0.20.0 Performance Evaluation ^[43]

References

- [1] <http://www.research.ibm.com/visual/projects/chromogram.html>
- [2] White, Tom. Hadoop: The Definitive Guide. 2009. 1st Edition. O'Reilly Media. Pg 3.
- [3] Kusnetzky, Dan. What is "Big Data?". ZDNet. <http://blogs.zdnet.com/virtualization/?p=1708>
- [4] Vance, Ashley. Start-Up Goes After Big Data With Hadoop Helper. New York Times Blog. April 22, 2010. <http://bits.blogs.nytimes.com/2010/04/22/start-up-goes-after-big-data-with-hadoop-helper/?dbk>
- [5] Cukier, K. (25 Feb 2010). Data, data everywhere. The Economist. http://www.economist.com/specialreports/displaystory.cfm?story_id=15557443
- [6] Horowitz, Mark. Visualizing Big Data: Bar Charts for Words. Wired Magazine. Vol 16 (7). June 23, 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_visualizing###ixzz0lIT2DN5j. Volu 16(7)
- [7] Community cleverness required. Nature, 455(7209), 1. 2008. <http://www.nature.com/nature/journal/v455/n7209/full/455001a.html>
- [8] Sandia sees data management challenges spiral. HPC Projects. Aug. 4, 2009. http://www.hpcprojects.com/news/news_story.php?news_id=922
- [9] Reichman, O.J., Jones, M.B., and Schildhauer, M.P. 2011. Challenges and Opportunities of Open Data in Ecology. Science 331(6018): 703-705. DOI:10.1126/science.1197962
- [10] Hellerstein, Joe. Parallel Programming in the Age of Big Data. Gigaom Blog. Nov. 9, 2008. <http://gigaom.com/2008/11/09/mapreduce-leads-the-way-for-parallel-programming/>
- [11] Segaran, Toby and Hammerbacher, Jeff. Beautiful Data. 1st Edition. O'Reilly Media. Pg 257.
- [12] <http://www-01.ibm.com/software/data/bigdata/>
- [13] Jacobs, A. (6 July 2009). The Pathologies of Big Data. ACMQueue. <http://queue.acm.org/detail.cfm?id=1563874>
- [14] Magoulas, Roger., Lorica, Ben. (Feb 2009) Introduction to Big Data. Release 2.0. Issue 11. Sebastopol, CA: O'Reilly Media. <http://radar.oreilly.com/r2/release2-0-11.html>
- [15] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (6 Feb 2001)
- [16] Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data" (<http://www.gartner.com/it/page.jsp?id=1731916>). Gartner. . Retrieved 13 July 2011.
- [17] Monash, Curt *eBay's two enormous data warehouses*, April 30, 2009 <http://www.dbms2.com/2009/04/30/ebays-two-enormous-data-warehouses/>
- [18] Monash, Curt *eBay followup — Greenplum out, Teradata > 10 petabytes, Hadoop has some value, and more*, October 6, 2010 <http://www.dbms2.com/2010/10/06/ebay-followup-greenplum-out-teradata-10-petabytes-hadoop-has-some-value-and-more/>

-
- [19] <http://www.economist.com/node/15557443>
 - [20] <http://www.economist.com/node/15557443>
 - [21] <http://www.economist.com/node/15557443>
 - [22] <http://www.economist.com/node/15557443>
 - [23] <http://www.economist.com/node/15557443>
 - [24] Danah Boyd (2010-04-29). "Privacy and Publicity in the Context of Big Data" (<http://www.danah.org/papers/talks/2010/WWW2010.html>) (html). *WWW 2010 conference*. Retrieved 2011-04-18.
 - [25] Jones MB, Schildhauer MP, Reichman OJ, and Bowers S. 2006. The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. *Annual Review of Ecology, Evolution, and Systematics* 37(1):519-544
 - [26] http://wiki.toadforcloud.com/index.php/Survey_distributed_databases
 - [27] <http://www.nosqldatabases.com/main/tag/marin-dimitrov>
 - [28] <http://www.slideshare.net/PhilippeJulio/big-data-architecture>
 - [29] <http://www.roadtofailure.com/2009/10/29/hbase-vs-cassandra-nosql-battle/>
 - [30] <http://ria101.wordpress.com/2010/02/24/hbase-vs-cassandra-why-we-moved/>
 - [31] <http://hstack.org/why-were-using-hbase-part-1/>
 - [32] <http://hstack.org/why-were-using-hbase-part-2/>
 - [33] <http://www.larsgeorge.com/2009/05/hbase-mapreduce-101-part-i.html>
 - [34] <http://www.larsgeorge.com/2010/01/hbase-architecture-101-write-ahead-log.html>
 - [35] <http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html>
 - [36] <http://gigaom.com/cloud/beyond-hadoop-next-generation-big-data-architectures/>
 - [37] <http://www.open-mpi.org/>
 - [38] <http://www.bsp-worldwide.org/>
 - [39] http://en.wikipedia.org/wiki/Bulk_synchronous_parallel
 - [40] <http://incubator.apache.org/hama/>
 - [41] <http://research.yahoo.com/files/ycsb.pdf>
 - [42] http://www.aicit.org/ijact/ppl/04_IJACT2-199028IP.pdf
 - [43] <http://www.slideshare.net/schubertzhang/hbase-0200-performance-evaluation>
-

Article Sources and Contributors

Big data *Source:* <http://en.wikipedia.org/w/index.php?oldid=464546306> *Contributors:* Arbitrarily0, Asplanchna, AtmosNews, Auntof6, Bar David, Casieg, Chire, Chris the speller, Cirt, Danielg922, Ethansdad, Fvillanustre, I42, Jac16888, Jantana, Jeremykemp, Jj1236, Jojikiba, Jordanzhang, Kuru, Melcombe, Mhiji, Od Mishehu, Ohnoitsjamie, Quibik, Ryuch, Sean Quixote, Seppemans123, SteveLoughran, Steven Walling, Tedder, Topbanana, Tothwolf, Xtzou, 17 anonymous edits

Image Sources, Licenses and Contributors

Image:Viegas-UserActivityonWikipedia.gif *Source:* <http://en.wikipedia.org/w/index.php?title=File:Viegas-UserActivityonWikipedia.gif> *License:* Creative Commons Attribution 2.0
Contributors: viegas

License

Creative Commons Attribution-Share Alike 3.0 Unported
[//creativecommons.org/licenses/by-sa/3.0/](http://creativecommons.org/licenses/by-sa/3.0/)
