The 1 Petabyte Constant Data Management Problem

Anant Jhingran
IBM Almaden Research Center, San Jose, CA
anant@almaden.ibm.com

A recent study by Hal Varian et al (http://www.sims.berkeley.edu/research/projects/how-much-info/index.html) shows that the total amount of original content is somewhere around 2 Exabytes/yr, growing at approximately 50%/yr. While they say that amount of "publicly" available information is smaller (700 TB/yr), it might be a low estimate. Let us first define "public" to mean both us as consumers, but also us as "people (or even machines) at work." Also, while the amount of "unique" content is interesting from an academic perspective, the challenge that most of us face (as consumers and as workers) is dealing with the information irrespective of its uniqueness across some sphere of "control" that defines our existence. I postulate that the amount of data in a "sphere" of influence will be in the range of 1 PB, and that is the next big data management challenge.

Let me describe four scenarios (each separated by 3 orders of magnitude on the distribution and size scale), and show why 1PB is an interesting amount, and the key point it brings out at each data point.

(1) Consider an enterprise data warehouse, that is now hitting 50 – 100 TB range, and might be comfortably in that range in a few years. However, now consider a problem where in addition to customer segmentation, we want to combine it with an "expertise" location within the enterprise, so that the right customer can be hooked up with the right employee. A good way of expertise location is through email that employees exchange among themselves, or with outside customers. Assuming 5 – 10 GB of email/employee/yr, a 10,000 employee organization is going to see 50 – 100 TB of email/yr. So the data warehouse + email could be 200 TB/yr for a medium sized company, and a 5 year history could very well be 1 PB. This example typifies achievement of a 1PB range within large centralized warehouse, but typically by *combining structured and unstructured data*.

(2) Now consider an enterprise that deals with 1000 suppliers (companies like IBM have 50,000 suppliers!), each having a 1TB data warehouse. True supply chain optimization in this environment deals with a "holistic" view of the 1PB split across 1000 databases, each with its own administrative domain and access control. The main challenges here are *"distribution of decision support" across 1000 data sources*, with strict limits on the computation that can be performed on each (earlier works on distributed databases have either dealt with ~10 databases and/or dealt with the ability to execute the full power of SQL or an equivalent query language on each of the database).

(3) A third scenario is a Napster like environment. While Napster formed a killer app in terms of music sharing (stealing?), what is technically more interesting is the distributed data management system that it (and its next generation clones) entails. *Over a million computers, each having 1GB or more of disks, cooperating together for a massive content delivery problem.* And achieving it through replication and distribution of catalogs that have the potential of giving it the quality of service and reliability that could be the envy of centralized systems.

(4) And finally, think of the deep web, which has potentially a billion nodes, and maybe 600TB – 1PB of (in this case dynamic) information. A quintessential problem here might be: "I have just launched a product X. What is the buzz about it in relation to my competitors products, and why?" This is distribution taken to its limit, with extreme variety of computation

(especially text analytics) and data management capabilities across its constituent data sources. None of the paradigms of SQL will work here.

(5) Now consider a sensor net: a trillion sensors, each communicating a 1KB of information. I am not yet convinced that once the data transmission problem is solved, it does not translate into one of the earlier data points (by collecting a 1000 sensors into one station etc.) However, from the theme of this conference, this might be an important point. Interesting research from Berkeley (the Telegraph project) is already beginning to explore some data management issues in this space,

In addition to these five points, where the (#connected databases)*(size of each database) = 1PB, there are many other interesting points, each in their limit reaching the 1PB constant. The collection of these points defines a new agenda for database research, which is currently not being addressed in a major way by the community. Here are some of the sub-problems within this grand challenge, roughly corresponding to each of the points above:

(1) What is the theory of being able to combine structured and unstructured data? We are used to formal OLAP analysis for structured data, but what is the equivalent for unstructured data? How do we drill down and roll up? Unstructured data are inherently imprecise (after all, does this document deal with databases, or a grand challenge, or bytes?) and a good theory for dealing with it in a formal way is missing.

(2) How do we deal with information integration in the presence of federation? And in particular, how does one deal with efficient computation (e.g. fraud detection) when the interfaces into various data are inherently more restrictive than full expressive power of SQL? And how does one maintain privacy as appropriate? There are some promising new results in this, primarily derivative of old statistical database research, such as Privacy Preserving Data Mining by Rakesh Agrawal et al., but much more needs to be done.

(3) Content management across a million computers breaks almost every data or content management system out there. The degree of variability in the content delivery capacity of the computers challenges our notions of SQL (where we have traditionally very carefully carved out complementary computations); in this model, redundant computation will be the norm, rather than the exception.

(4) Discovery across the deep web is inherently poorly understood problem. It challenges all our theory and practice for text analytics (try running the clustering algorithms from the literature on a trillion documents), of data gathering, of economic models, of syndication and privacy.

In all the examples I have given above, images, videos and pure scientific apps do not enter into the equation, because 1PB can get rather easily filled up taking these into account; however unless one is doing some deep analysis on these objects (and except some niche apps, commercial data analysis apps on these data types are some distance away), the data management problem is rather mundane. If I am wrong about the importance of these data types, then even 10PB might become interesting in my equation. But till then, we have our hands full challenging our conventional relational and text techniques in handling the scale of distribution and federation. This is the future of information integration, and is likely to be billions of dollars of business. So let us get on with it!

About the Author

Anant Jhingran is the Director of Computer Science at IBM’s Almaden Research Center in San Jose, CA. He manages a research team of about 150 people focused on a variety of topics, but the three top areas of research are databases, content management, and knowledge management. Some of the large scale projects going on his department are: XMLization of databases, web scale data mining, privacy, GRID computing and text analytics.

Anant obtained his PhD from the University of California, Berkeley, in Database Systems, in 1990, and his Bachelors from the Indian Institute of Technology, Delhi, in Electrical Engineering in 1985. Prior to his current position, he was the Senior Manager for Databases and Electronic Commerce at IBM’s T.J. Watson Research Center, NY.

Anant is a member of ACM and IEEE, and has been active in the database field (but not as much now as he would like to be ☺)