

Correction to my article posted on March 25, 2013

Dr. Bhavani Thuraisingham

January 25, 2015

My article dated March 25, 2013 (Big Data: Have we seen it before?) posted on my web site has been used by Dr. Nafeez Ahmed to promote his view that the CIA was behind Google.

<https://medium.com/@NafeezAhmed/how-the-cia-made-google-e836451a959e>

Dr. Ahmed called me from the UK on January 21, 2015 for about 5-10 minutes to find out more about the MDDS program that I was working on in the 1990s and how it impacted Google. I replied that I did not know about MDDS' impact on Google as other agencies were also funding Stanford University on massive data technologies (e.g., the digital library effort by NSF) at that time. I explained that efforts like MDDS have provided the foundations for some of the solutions being proposed to big data management today. I was given the impression that Dr. Ahmed was interested in writing about the history of Google and big data technologies. Therefore, I was disappointed and dismayed to read his article dated January 22, 2015 which has twisted my words to put a negative spin on MDDS and Google. Furthermore, Dr. Ahmed has taken pieces of information out of context from different sources and put them together to tell a story that CIA funding was behind Google.

In my article dated March 25, 2013, I made a statement that Mr. Sergey Brin (Google co-founder while at Stanford), Prof. Jeffery Ullman (from Stanford) and Dr. Chris Clifton (then at MITRE) developed the Query Flocks system. This is because at that time I believed that this was the case. However, in Dr. Ahmed's article posted on January 22, 2015, a Google Director of Communications has stated that Mr. Sergey Brin never worked on Query Flocks. Based on this information, I would like to make a change to my article dated March 25, 2013. Specifically, the second paragraph of my March 25, 2013 article should read as follows:

So was the MDDS effort successful? Some say no because the program did not solve the massive data problem. However I say yes as the program did contribute to the understanding of what was involved in handling massive amounts of data and produced solutions for some of the challenges including storage management and indexing as well as query processing. In fact Prof. Jeffrey Ullman (at Stanford) and my colleague at MITRE Dr. Chris Clifton together with some others developed the Query Flocks System, as part of MDDS, which produced solutions for mining large amounts of data stored in databases. Also, Mr. Sergey Brin, the cofounder of Google, was part of Prof. Ullman's research group at that time. I remember visiting Stanford with Dr. Rick Steinheiser from the Intelligence Community periodically and Mr. Brin would rush in on roller blades, give his presentation and rush out. During our last visit to Stanford in September 1998, Mr. Brin demonstrated to us his search engine which I believe became Google soon after.

There are also several inaccuracies in Dr. Ahmed's article (dated January 22, 2015). For example, the MDDS program was not a "sensitive" program as stated by Dr. Ahmed; it was an Unclassified program that funded universities in the US. Furthermore, Sergey Brin never reported to me or to Dr. Rick Steinheiser; he only gave presentations to us during our visits to the Department of Computer Science at Stanford during the 1990s. Also, MDDS never funded Google; it funded Stanford University.

I strongly believe that the contribution of Google to society is second to none. Google search is as important a contribution as sea, rail, or air travel. Therefore, whoever funded Google has provided the greatest service to mankind.

My original article dated March 25, 2013 is appended to this Correction.

Big Data: Have we seen it before?

Bhavani Thuraisingham

Dallas, TX

March 25, 2013

The Big Data Problem has been around as far as I can remember. Back in the 1990s the Intelligence Community started a program called Massive Digital Data Systems (MDDS) that I was managing for the Intelligence Community when I was at the MITRE Corporation. The MDDS program started in 1993 and ended in 1999. The program funded about 15 research efforts at various universities and the goal was to develop data management technologies to manage several terabytes to petabytes of data. The data management technologies included those for query processing, transaction management, metadata management, storage management, and data integration. The World Wide Web was in its very early stages back in 1994 and therefore web-based data management research was also just beginning. Furthermore, the NSF, DARPA and NASA had just started the digital library initiative. An overview of the MDDS program was presented by the MDDS Team at the Annual Intelligence Community Symposium AIPASG in 1995 and the abstract is attached to this document.

So was the MDDS effort successful? Some say no because the program did not solve the massive data problem. However I say yes as the program did contribute to the understanding of what was involved in handling massive amounts of data and produced solutions for some of the challenges including storage management and indexing as well as query processing. In fact the Google founder Mr. Sergey Brin was partly funded by this program while he was a PhD student at Stanford. He together with his advisor Prof; Jeffrey Ullman and my colleague at MITRE Dr. Chris Clifton developed the Query Flocks System which produced solutions for mining large amounts of data stored in databases. I remember visiting Stanford with Dr. Rick Steinheiser from the Intelligence Community and Mr. Brin would rush in on roller blades, give his presentation and rush out. In fact the last time we met in September 1998, Mr. Brin demonstrated to us his search engine which became Google soon after.

So now twenty years later, we have moved from Massive Data to Big Data and some of those who I worked with back in the 1990s are asking the question "What is the difference?" The challenges remain the same; we have to develop data management and analysis technologies to handle massive amounts of data. However the massiveness has changed. From several terabytes to a few petabytes of data back in the 1990s, we have now moved on to several petabytes to exabytes and even zetabytes of data. Why? Because of the World Wide Web, social media, mobile technologies and developments in the human genome project, as well progress in bioinformatics, geoinformatics, security informatics, and multimedia/video processing, among others. Massive amounts of data about entities and the relationships between them and the evolution of these relationships have to modeled, stored, managed, queried and analyzed. The world population has increased by over 25% in the past 20 years and hundreds of millions of people from all over the world are joining social media. It is expected that by 2020 50 billion devices will be online and that's only seven years away. Therefore we need to be able to securely and efficiently handle the zetabytes of data without violating the privacy of individuals. This is a daunting task.

So what do we do? First, from my experience twenty years ago, never believe that the Big Data Challenge will be solved. This is because big data will continue to get bigger. Therefore just like we developed solutions twenty years ago for massive data, we need to develop solutions for big data. I am very pleased to see that not only federal agencies such as the National Science Foundation and the Defense and Intelligence organizations are focusing on big data, practically every large company such as Google, Facebook. Microsoft, Oracle and IBM are starting initiatives. We all have to work together to tackle the big data challenge and make the world a better place to live.

TITLE: Intelligence Community Initiative in Massive Digital Data Systems

Authors: Hal Curran (NSA), Robert Kluttz (CMS), Dr. Claudia Pierce (NSA),
Dr. Rick Steinheiser (ORD) and Dr. Bhavani Thuraisingham (MITRE)

Related Application Area: Integrated and Transparent Access to Multiple Data Sources

Sponsor: Community Management Staff, Office of Research and Development,
National Security Agency

Description: The Massive Digital Data Systems (MDDS) Initiative is developing high-risk high-pay-off data management technologies for Intelligence Systems of the future. Future Intelligence Systems must effectively manage massive amounts of digital data (i.e., multi-terabytes or greater). Issues such as scalability, design, and integration need to be addressed to realize a wide spectrum of Intelligence Systems ranging from centralized terabyte and petabyte systems comprising many large objects (e.g., images) to distributed heterogeneous databases that contain many small and large objects (e.g., text). Consequently, Massive Digital Data Systems (MDDSs) are needed to store, retrieve, and manage the complex data for the Intelligence Community (IC). While several advances have been made in data management technologies, the complexity and the size of the database, as well as, the needs of the IC, such as multimedia data management, require the development of novel approaches for querying and updating the massive databases. Therefore, the Community Management Staff (CMS) of the IC has started the MDDS Initiative to identify the challenges for massive database management and to develop approaches to meet the challenges. The ultimate vision is to provide for the seamless access and fusion of massive amounts of data, information, and knowledge in a heterogeneous and real-time environment to carry out the functions of the IC with diminishing resources.

To develop massive digital data systems for the IC, several technologies have to be integrated. These include the following: Data management, Knowledge management, Information management, Information visualization, Mass storage management and Distributed processing. The technical focus of the MDDS Initiative is to develop *data management* technologies. These include developing techniques for querying, browsing, and filtering; transaction processing; accesses methods and indexing; metadata management and data modeling; and integrating heterogeneous databases; as well as developing appropriate architectures. Approaches to handle massive databases will be a major focus for this Initiative. Note that secure database management, as well as, other technologies, such as information management and visualization, are beyond the scope of this Initiative. The organizational focus of the MDDS Initiative is to provide seed money to develop data management technologies which are of high-risk and high-pay-off. Further development of the technologies produced under the MDDS Initiative for eventual use by the IC and related organizations is expected to be carried out by other agencies such as the line organizations, as well as, industrial organizations.

The MDDS initiative has awarded fifteen research contracts. These contracts are being executed through the line organizations, such as NSA and CIA. Specific topics being addressed by the research include data mining, data warehousing and heterogeneous database integration, real-time transaction processing in a heterogeneous environment, quality of service management in multimedia databases, indexing multimedia data, scalable architectures, and geographic information systems. The MDDS Team is also coordinating the research with other government agencies by conducting technology exchange meetings.

Scope: Approx: 3 - 4 million dollars per year for 3 - 4 years

Customer: DOD, IC, and other government organizations