Wolfram|Alpha, A New Kind of Science

by
Bruce Walters

April 18, 2011

Research Paper for
Spring 2012 INFSY 556 Data Warehousing
Professor Rhoda Joseph, Ph.D.

Penn State University at Harrisburg

PENNSTATE Harrisburg

Abstract

The core mission of Wolfram|Alpha is "to take expert-level knowledge, and create a system that can apply it automatically whenever and wherever it's needed" says Stephen Wolfram, the technologies inventor (Wolfram, 2009-02). This paper examines Wolfram|Alpha in its present form.

Introduction

As the internet became available to the world mass population, British computer scientist Tim Berners-Lee provided "hypertext" as a means for its general consumption, and coined the phrase World Wide Web. The World Wide Web is often referred to simply as the Web, and Web 1.0 transformed how we communicate. Now, with Web 2.0 firmly entrenched in our being and going with us wherever we go, can 3.0 be far behind? Web 3.0, the semantic web, is a web that endeavors to understand meaning rather than syntactically precise commands (Andersen, 2010). Enter Wolfram|Alpha.

Wolfram Alpha, officially launched in May 2009, is a rapidly evolving "computational search engine," but rather than searching pre-existing documents, it actually computes the answer, every time (Andersen, 2010). Wolfram|Alpha relies on a knowledgebase of data in order to perform these computations, which despite efforts to date, is still only a fraction of world's knowledge. Scientist, author, and inventor Stephen Wolfram refers to the world's knowledge this way: "It's a sad but true fact that most data that's generated or collected, even with considerable effort, never gets any kind of serious analysis" (Wolfram, 2009-02). The knowledge is in the world, it simply needs to be captured and curated to be of use to the greater population.

Focusing on knowledge processing, rather than search, lead developer Oyvind Tafjord refers to Wolfram|Alpha as a "computational knowledge engine," comprised of three core parts. Those parts are: a computational representation of human knowledge gathered throughout history, from facts and real world data to sophisticated formulas and algorithms; a free-form input line; and the output in the form of multiple relevant interpretations to the query (Tafjord, 2009). In this way, Wolfram|Alpha seeks to make all systematic knowledge immediately computable by anyone. One needs only to enter the question or calculation, and Wolfram|Alpha uses its built-in algorithms and growing collection of data to compute the answer (Highscalability, 2009). Your author entered two challenges to Wolfram|Alpha, which surprisingly gave the correct responses within seconds (see Appendix I). Indeed impressive, and especially given the mathematics background of the system, proper responses to queries of this nature bode well for the core mission of Wolfram|Alpha: "to take expert-level knowledge, and create a system that can apply it automatically whenever and wherever it's needed" (Wolfram, 2009-02).

Wolfram|Alpha, the Special Project

Wolfram|Alpha started its life in the NKS [for New Kind of Science] "Special Projects" group. After an initial proof-of-concept with math and dates, Wolfram|Alpha moved on to units, datapaclets, and formulas, and then to specific knowledge areas such as "NFL sports" (Tafjord, 2009). Structured Data is a core component of Wolfram|Alpha. Dr. Wolfram provides us with

four components, which he calls the 4 pillars, that comprise Wolfram|Alpha. They are: 1) Curated Data, that is the data pipeline, which is organized, validated, expert judged, and converted into a form that can be combined with other data; 2) Computation, that is the science and engineering component of the system; 3) Linguistic Processing, that is the natural language input; and 4) Results presentation, that provides as clear as possible a visual "computation aesthetics" (Wolfram, 2009-05).

Wolfram|Alpha is built on Mathematica and other Wolfram Research products. Mathematica is also a repository of world knowledge and data, but with strict input syntax, well defined output, and limited to mathematics. In fact, the core computational engine of Wolfram|Alpha is essentially a gigantic Mathematica application package, which is called from webMathematica. The development environment for Wolfram|Alpha is Wolfram Workbench. Mathematica provides methods for content gathering and management, user interfaces for curating data, analysis and processing of difficult semantic interpretation tasks, such as example generation, testing interfaces, and analysis and reporting on the web site query logs which feed back into development (Tafjord, 2009). Mathematica delivers the world's largest set of algorithms, all with self-checking capabilities (Wolfram|Alpha , 2010). Following the four pillars approach, Wolfram|Alpha input is parsed and converted to Mathematica symbolic expressions. This information is passed to result "scanners" and computed with gridMathematica, which generates the results. The formatted results are packaged by webMathematica and sent to the user's browser, using AJAX for parallel CPU support (Tafjord, 2009).

Mathematica performs computation using symbolic objects. These include algebraic objects that yield exact computations as well as any expression in a formal language, as opposed to numerical objects that are typically represented by an approximation. Because of this, Mathematica, and therefore Wolfram|Alpha, can retrieve information through string processing (Roda, 2010). String processing is the heart of Natural Language Processing, and includes regular expression matching, sequence alignment, dictionary lookup, and XML processing.

Natural Language Processing

Natural Language Processing (NLP) as a science seeks to "understand" human input. The linguistic-parser develops some initial "sense" of the words, and their "use" implies that this resultant sense is able to be further understood and manipulated. Wolfram|Alpha input must be interpreted in this way in order to be mapped to the deep-web data sources (Morris, 2009). There are signs that Wolfram|Alpha works the same way that Palm Graffiti did in that people start out writing natural language queries, but pretty quickly trim it down to just the key concepts (a process known as "anti-phrasing"). And just as for Palm, we learn to write queries in a notation that Wolfram Alpha understands (Andersen, 2010). But to meet its goal, Wolfram|Alpha must understand free-form natural language, with all its variations and redundancies. The grammar for structured data is very different from this, and less forgiving. Wolfram|Alpha continues to develop and evolve this difficult process (Wolfram, 2009-02). Says Tafjord: "We want to significantly improve the quality of the input interpretation code, both through incremental and dramatic changes in the code base. This is a major ongoing R&D project" (Tafjord, 2009). NLP is the user's gateway to access the Big Data, stored within the system.

Big Data

"Big Data" is data that exceeds the processing capacity of conventional database systems. Big Data is too big, moves too fast, or doesn't fit the limits of typical database architecture. To induce value from this data, one must determine an alternative way to process it. A catch-all term, Big Data can be nebulous. Traditional input data to Big Data systems include business data, chatter from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast audio streams, banking transactions, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, and financial market data, to name a few (Dumbill, 2012). In most if not all business sectors, the promise of Big Data excites even venture capitalists (McBride, 2012).

To understand and use Big Data, we apply the three V's of volume, velocity and variety to characterize its different aspects. Volume is the benefit gained from the ability to process large amounts of information. This is the main attraction of Big Data Analytics. Velocity is the rate at which data flows into, within, or out of an organization. Variety states that the source data is diverse, and doesn't fall into neat relational structures (Dumbill, 2012). Examples of the tools normally used to harvest or manage raw Big Data include MapR's Hadoop, Fusion-IO, and Splunk. Splunk collects, indexes, and harnesses fast moving machine data from applications, servers and devices, whether they are physical, virtual, or in the cloud. Hadoop and Fusion-IO are examples of the storage and data management platform, which is referred to as the Big Data Stack, an uberdatawarehouse.

The foundational layer in the Big Data Stack is the internet cloud, which provides scalable and persistent computing power. The middle layer of the Big Data Stack is analytics, where features are extracted from data, and processed, perhaps by classification and prediction algorithms. The top layer of the Big Data Stack is the services and applications, where consumers experience a data product (Driscoll, 2011). In this regard, Wolfram|Alpha is a Big Data Stack.

But Big Data practitioners consistently report that 80% of the effort involved in dealing with data is cleaning it up in the first place (Driscoll, 2011). "I probably spend more time turning messy source data into something usable than I do on the rest of the data analysis process combined" (Dumbill, 2012). In order to be useful, the data must be carefully organized.

Curated Big Data

Wolfram|Alpha answers specific questions about objective facts, based on what is known to it. Unless specifically entered, only public information is available (Morris, 2009). But it appears that Wolfram|Alpha has its own tools to import and organize, or curate, Big Data. "Notebook" interfaces to processing and cleaning up data, flows through a database to deployed product. "Textbooks" and "handbooks" get processed into relevant formulas, grammar rules, and output code. There is dynamic interplay between adding content into existing frameworks, improving them, or adding new ones. The details of the knowledge representation, and its data structure, go to the heart of Wolfram|Alpha functionality. Similarly, Wolfram|Alpha scientists are researching deeper and more general designs for computational data representation, storage, and retrieval.

This will enable Wolfram|Alpha to handle inhomogenous, multi-source, and generally messy data (Tafjord, 2009).

A Wolfram|Alpha curated data set will reliably return the same answer freshly calculated every time. Should the answer change, it would be because the underlying data was changed, or because a developer has figured out a new and better way of doing the calculation. It would not be because someone external to Wolfram|Alpha has figured out a way to game the system (Andersen, 2010).

Wolfram|Alpha has tons of data, but for it to be accessible, a developer has to come along and enter linguistics for it. Once that is done, it is accessible by everybody. Two possible exceptions could be; premium data such as stock prices with no delay, and private instances of Wolfram|Alpha for corporate data. Those would have a subscription or be physically walled off respectively. In addition, Wolfram|Alpha security site monitors protect the system and data from attackers or unusual traffic patterns (French, 2012).

Wolfram|Alpha Technology

Wolfram|Alpha is all about processing Big Data, and there are some pretty big numbers associated with Wolfram|Alpha. At launch, on Friday, May 15, 2009, Wolfram|Alpha had two supercomputers comprising about 10,000 CPU cores, hundreds of terabytes of disk space, more than 10 trillion pieces of data. Wolfram|Alpha has 50,000 types of algorithms and consists of 5 million lines of symbolic Mathematica code. The system is able to handle about 175 million queries per day, which is perhaps 30 billion calculations, or 5 billion queries, per month. Wolfram|Alpha runs in 5 distributed colocation facilities, and in the words of some industry observers, “has what seems like enough air conditioning for the Sahara to host a ski resort.” One of the supercomputer partners, R Systems, created the world’s 44th largest supercomputer (per the June 2008 TOP500 list) for Wolfram|Alpha, called the R Smarr. R Smarr is rated at 39580 GFlops and uses Dell computers consisting of 4608 cores, 65536 GB of RAM, and Infiniband interconnect. (Highscalability, 2009).

But there are also small, say, smartphone or tablet sized elements in the Wolfram|Alpha universe. In October of 2011, Apple introduced the voice-controlled personal assistant software called Siri. Since them, according to the New York Times, nearly 25 percent of the traffic handled by Wolfram|Alpha has come from Siri (Aimonetti, 2012). Apple has a partnership with Wolfram|Alpha and has made good use of the Wolfram|Alpha Application Development Interface for Siri. For iPhone 4S users, Siri's relationship with Wolfram|Alpha is an important one (French, 2012).

In fact Siri has been driving sales of iPhones. Having the properly interfaced capability of Wolfram|Alpha on your mobile device can only continue to be helpful. But Apple is not the only company that likes Wolfram|Alpha. Microsoft also licenses the Wolfram|Alpha technology, as do several other companies that contract specialized versions for their research teams (French, 2012). Despite early criticism of Wolfram|Alpha's, the once-small company has grown to a team of over 200 and is on the cusp of releasing Wolfram Alpha Pro, which Stephen Wolfram calls "the next step of what can be done with this approach" (Aimonetti, 2012). Venture capitalists say

the Big Data wave is just starting to build, and see limitless opportunity in the mobile realm. Ever present and able to generate copious data, the mobile device is the single best data-capture device to date (McBride, 2012).

Perhaps unsurprisingly, new deployment channels for the core Wolfram|Alpha engine and API include an iPhone app, corporate data, and widgets, and the "Pro" version focuses on the corporate customer (Tafjord, 2009). Wolfram|Alpha Pro will allow corporations to add computational knowledge to a website or product, deploy Wolfram|Alpha on an enterprise platform or environment, use Wolfram|Alpha technology to curate corporate data, and create a Wolfram|Alpha interface for corporate data. In addition, Wolfram consultants may be contracted to bring technical skills, vertical industry prowess, and deep project-management expertise to satisfy customer needs (http://products.wolframalpha.com/). In a recent interview, Stephen Wolfram said that rather than sell to the search engines, "We'd rather look for things like partnerships or licensing deals for [our] APIs. I see a new field of knowledge-based computing. Imagine a spread sheet that can pull in knowledge about the entries" (Standen , 2009).

Conclusion

We have seen that Wolfram|Alpha is scalable and agile, obviously manages aggregation well, is easily deployable to the business enterprise environment with the help of consultants, and is supported by a viable company. Therefore Wolfram|Alpha meets the typical selection criteria for a data warehouse (Joseph, 2012). At $4.99/month, Wolfram|Alpha Pro version is a real bargain because it is a complete platform, available from anywhere.

We have seen that Wolfram|Alpha has a proven Application Development Interface and knowledge and computational engine. We know that Wolfram|Alpha has the tools to generate a dimensional table by accessing a web service and enriching what business users can analyze. And we know that Wolfram|Alpha has curated data, and an NLP input and output. The implications of these 4 pillars of Wolfram|Alpha become quite important when put together. Together, they can form building blocks towards a meaning-driven semantic web. The use of queries as information passing mechanisms just might get us, if not to the Web 3.0, at least some of way there (Andersen, 2010).

Gartner calls Wolfram|Alpha "Transformational," but sites the challenges of NLP. In order to be meet the stated goal, Wolfram|Alpha must discover a better means of communication with humans, one that is intuitive, effective, and context aware. The ability for technophiles and unschooled consumers alike to achieve effective responses from machines without using expertise in creating successful queries will generate new kinds of information exploitation (Andrews, 2011). Still, the question remains: Will Wolfram|Alpha fulfill its goal, to make all systematic knowledge immediately computable by anyone?

PENNSTATE Harrisburg

References

Aimonetti, Joe (7 February 2012), "Siri brings nearly 25 percent of Wolfram Alpha traffic," cnet.com, http://reviews.cnet.com/8301-19512_7-57372868-233/siri-brings-nearly-25-percent-of-wolfram-alpha-traffic/

Andersen, Espen (November 2010), "Edging Toward the Semantic Web: Protocols, Curation, and Seeds," The Norwegian School of Management, http://ubiquity.acm.org

Andrews, Whit (12 August 2011), "Natural Language Question Answering" as part of the "Hype Cycle for Business Intelligence, 2011," Gartner, http://www.gartner.com G00216086

Driscoll, Michael (09 August 2011), "Building data startups: Fast, big, and focused Low costs and cloud tools are empowering new data startups," oreilly.com, http://radar.oreilly.com/2011/08/building-data-startups.html

Dumbill, Edd (11 January 2012), "What is big data? An introduction to the big data landscape," oreilly.com, http://radar.oreilly.com/2012/01/what-is-big-data.html

French, Christopher, (March, 2012), personal interviews, Wolfram

Highscalability (15 May 2009), "Wolfram|Alpha Architecture," highscalability.com, http://highscalability.com/wolfram-alpha-architecture 2/19/2012

Joseph, Rhoda (February 2012), Class Notes from INFSY 556 Data Warehousing, Penn State University at Middletown

McBride, Sarah (21 Feb 2012), "Venture capital sees big returns in big data." Reuters, http://www.reuters.com/assets/print?aid=USTRE81G1HO20120221

Morris, Peet (June 2009), "Wolfram Alpha – how it works," Computer Weekly, http://www.computerweekly.com/feature/Wolfram-Alpha-how-it-works-part-2

Roda, Giovanna (01 June 2010), "Connecting the dots," ACM Digital Library, New York, NY, http://dl.acm.org.ezaccess.libraries.psu.edu/citation.cfm?id=1842890.1842903&coll=DL&dl=ACM&CFID=69023641&CFTOKEN=81450046

http://www.splunk.com/

Standen, James (10 April 2009), "Wolfram Alpha- Dimensional Generator?" datamartist.com, http://www.datamartist.com/wolfram-alpha-dimensional-generator

Tafjord, Oyvind (2009), "The Technology behind Wolfram|Alpha," Wolfram, http://library.wolfram.com/infocenter/Conferences/7551/

Wolfram|Alpha (2010), Wolfram, http://www.wolfram.com/mathematica/how-mathematica-made-wolframalpha-possible.html

Wolfram, Stephen (02 February 2009), Blog, Wolfram, http://blog.wolframalpha.com/2012/02/09/launching-a-democratization-of-data-science/

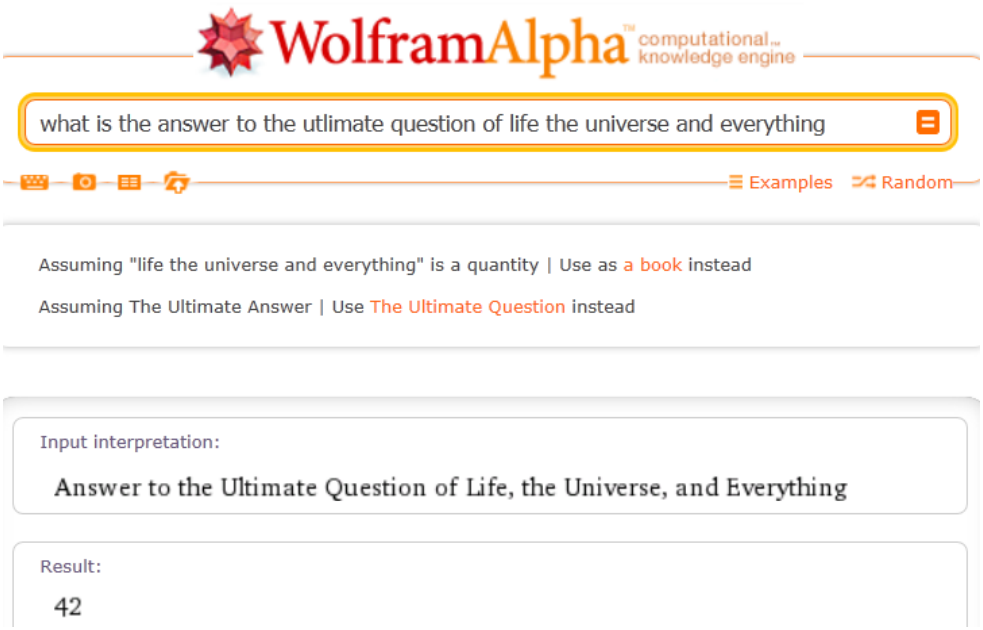
Wolfram, Stephen (21 May 2009), "Wolfram discusses the four main "pillars" or components of Wolfram|Alpha" youtube.com, http://www.youtube.com/watch?v=wirZUhqUY54&feature=related

http://products.wolframalpha.com/

PENNSTATE Harrisburg

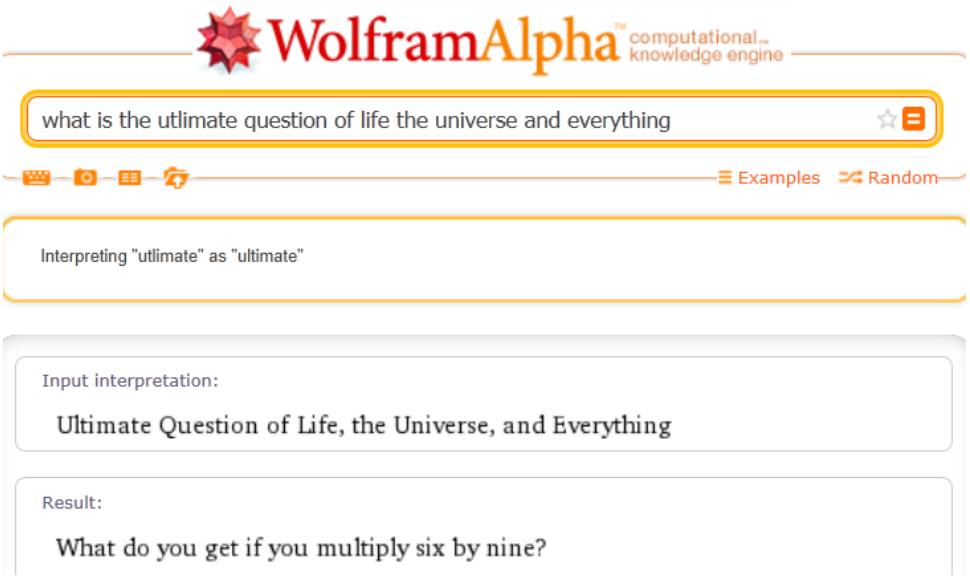
Appendix I: Standard Knowledge Test Questions

Knowledge Test Question 1



Super computer Deep Thought took millions of years to report this answer, and then suggested building an even larger computer to find the question. This larger computer resembled a rocky planet, and was called Earth. Which brings us to Knowledge Test Question 2.

Knowledge Test Question 2



Note: While 6 x 9 = 54 in base 10, it does equal 42 in base 13. Obviously the detailed answer is in progress.