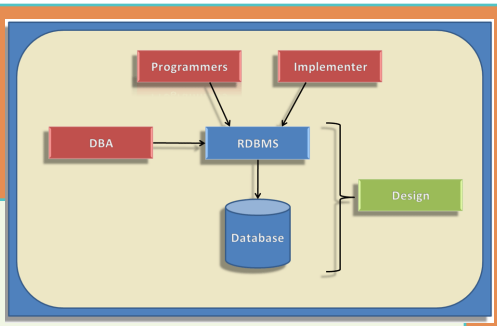
MATH MATTERS

Apply It.



The math behind... the Web of Linked Data

Technical terms used:

Resource description framework, first-order logic, query language, ontologies, machine learning

Uses and applications:

The Web of Linked Data has grown from just seven datasets in 2007 to over a thousand in 2014. It is used by several important organizations to enrich their knowledge bases by incorporating semantics. Examples include the BBC, Drug Bank, DailyMed, PubChem, CiteSeer, ACM, Craigslist, and the US and UK governments, all of which have exposed substantial Linked Data. Linked Data is also a true example of Big Data, as it exhibits all four traits of variety, volume, velocity, and veracity. The ultimate goal is to make the Web machine-readable, a concept articulated by Sir Tim Berners Lee.

How it works:

The Web of Linked Data can be visualized as a directed, labeled graph at the global scale, with nodes representing entities and edges representing the relationships between entities. This graph-based data model is denoted as Resource Description Framework (RDF) because a node or edge constitutes a resource and is typically identified using a Uniform Resource Identifier (URI), which can be fetched and dereferenced by the current World Wide Web technology stack in a manner similar to the dereferencing of Web pages.

Linked Data is often annotated using ontologies, which provide a vocabulary for the data source's metadata. Ontologies are powerful because they can be used by reasoning systems to answer intelligent questions about the data, using fragments of first-order logic. A unique aspect of Linked Data is that both data and metadata are represented as RDF graphs and can be uniformly queried, using the graph-based pattern-matching language SPARQL.

Finally, Linked Data is distributed, meaning that the graph is spread over many machines. This allows anyone in the world to publish data on the Linked Data cloud, simply by obeying a set of best practices. The most important of these practices (dubbed the fourth principle) is that data must not be published in silos but must be linked to existing Linked Data. Current research is attempting to discover these links automatically using statistical machine learning techniques. While much progress has been made on automated link discovery, the problem has not yet been satisfactorily solved.

Interesting facts:

Freebase, an important encyclopedic Linked Data source, is being used to build the Google Knowledge Graph to incorporate more semantics into Google's traditional information retrieval process.

References:

Christian Bizer, Tom Heath, and Tim Berners-Lee, Linked data-the story so far. International Journal on Semantic Web and Information Systems 5, no. 3, 1-22 (2009).

Axel-Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, and Klaus Lyko. Unsupervised link discovery through knowledge base repair. The Semantic Web: Trends and Challenges, 380-394 (2014).

Submitted by Mayank Kejriwal, University of Texas, Austin, Math Matters, Apply it! Contest, February 2015.

Society for Industrial and Applied Mathematics
www.siam.org
www.siam.org/careers/matters.php