Are knowledge graphs a good thing? (and what does it take to create one?)

Dr. Ora Lassila

Principal Graph Technologist
Amazon Neptune

2020-05-07 – KGC2020

aws

Outline

Some background

- graphs
- semantics
- the Semantic Web, Linked Data

What is a Knowledge Graph?

- why would you want one?
- some things to consider when building one

Amazon Neptune

Graphs: what and why?

As a formal concept, origins go back to the 18th century

Much of computer science is in many ways grounded in graph theory

but…

Graphs are a very natural way to think about information and especially about representations of our world

(Anecdote: “I can explain this to a 7-year old”)

aws

Semantics…? Huh…?

Much abused and overloaded term

Most folks who use the term do not know what it means

In the context of information systems, semantics

defines how data “behaves” and how machines can interpret data

and when properly applied, frees us from having to ”hard-wire” logic into software systems → “data-driven” processing

Where does semantics come from?

1. Relationships between data and definitions of data (good)
2. Relationships within data (good)
3. Hard-wired in software (not so good)

“The Semantic Web” [Berners-Lee, Hendler, Lassila 2001]

20-year old vision about making it easier to discover, interpret, combine, manage, and share information

(it is really not so much about the Web, actually…)

Predicated on

accessible data

accessible semantics – definitions of data in the form of ontologies

mechanisms of reasoning about data (with the help of ontologies)

Groundbreaking work by W3C and many member organizations to create the basic building blocks

RDF, OWL, SPARQL, etc.

rather than standardizing ”what to say”, we standardized “how to say it”

What is an ontology?

In Computer Science, an ontology is a complex logical data model

- think of them as the extreme manifestation of logical models

An ontology defines how to represent information about a domain

- concepts and their properties
- relationships between concepts
- constraints that must hold

An ontology captures the semantics of a domain to the extent that we merely need some generic software to interpret data

Linked Data

Effectively: Linked Data = Semantic Web minus reasoning

Massive undertaking to make lots of data accessible

Cynical view:

relegates RDF into a mere data structure

Optimistic view:

a necessary effort to provide data for the Semantic Web

What is a Knowledge Graph? (my view)

A means of structuring and organizing information for easier access and understanding

- “democratizes” data in an organization
- accessible data = physical access + understandable models

Graph models are typically easier to understand than relational ones

Graph modeling is very flexible and allows for the real-world diversity and heterogeneity of data

- forget “big data”… think of “wide data”

aws

Considerations when building a KG

Forces you to think of your organization's information as an asset

Because of this, you must consider how different pieces of information "go together"

- that is, how are they linked

Typically knowledge graphs are not built for a single use case

- there could be unknown future use cases → serendipity

Identity is important

You need a shared data model → "upper ontology"

What is an "upper ontology"?

Establishes basic (possibly generic) concepts

Specific domain models can leverage an upper ontology to facilitate integration with other models

Provides for best possible "partial understanding"

- something information systems traditionally have not been very good with

Public ontologies can be a good starting point, especially if you aspire to share information with external parties/organizations

Facilitates "delayed semantic commitment" → build systems without yet knowing exactly what the specific definitions of data will be

aws

Technical considerations

Graph storage

- graph database
- graph model: RDF vs. property graphs

Software patterns

- how does your software access and manipulate graph data?

Graph construction and maintenance

- what sources do you need?
- how do you keep data up-to-date?
- do you need to propagate graph changes back to original sources?

User experience

- challenge: it is hard for UI flexibility to match the new data flexibility

aws

Amazon Neptune

Fully managed cloud graph database

- scalable, secure and highly available
- full-text search, federated queries, granular change streaming

Choice of graph models and query languages

- RDF and property graphs; SPARQL and Gremlin

Many customers with KG use cases

- Siemens – master data management, production monitoring, …
- FINRA – from unstructured data to structured data (and relationships)
- Uber – managing map data and versioning
- Tableau – modeling and managing cloud infrastructure
- many Amazon internal customers – we see significant interest in graphs

aws

Thank you!

Ora Lassila

ora@amazon.com