Size does not matter (if your data is in a silo)

Dr. Ora Lassila
Principal Technologist
Cloud Analytics Team &
Nokia Location & Commerce

2012-11-11
Elected Member
Advisory Board
World Wide Web Consortium (W3C)

Semantic Technologies meet Recommender Systems & Big Data (SeRSy 2012)
Some speaker details

• Principal architect: “big data analytics” @ Nokia
• Elected member: W3C’s Advisory Board (1998–)
• Research (Nokia, MIT, CMU, HUT), venture capitalist, entrepreneur, software engineer
• Ph.D (D.Sc) in CS, Helsinki Univ. of Technology
• Semantic Web research and standardization work
  – co-author of a 2001 article on the Semantic Web (often characterized as science fiction)
  – co-editor of the original RDF specification
Some speaker details

• Principal architect: “big data analytics” @ Nokia
• Elected member: W3C’s Advisory Board (1998–)
• Research (Nokia, MIT, CMU, HUT), venture capitalist, entrepreneur, software engineer
• Ph.D (D.Sc) in CS, Helsinki Univ. of Technology
• Semantic Web research and standardization work:
  - co-author of a 2001 article on the Semantic Web (often characterized as science fiction)
  - co-editor of the original RDF specification

© 2012 Nokia
Things I intend to discuss

• Nokia and “big data”
• What’s difficult?
• Where do Semantic Web technologies come in?
Nokia and the third phase of mobility

A mobile-first society means “WHERE” will transform all experiences

Population shift to cities
Social and location converge
Internet of Things – gadgets and sensors

Voice 1990s
Internet 2000s
2010s

Nokia and the third phase of mobility
Nokia Maps – some of our customers

- foursquare
- inSing.com
- GARMIN
- VKонтакте
- yelp
- bing
- SingTel
- wcities
- sina.com.cn
- WRC
- tripadvisor
- YAHOO!
- lonely planet
- Windows Phone
- Tencent
- Time Out

Nokia
Connecting People
Some approximate statistics

11B “Probe points” processed per month
100M Search queries per month
2B Positioning requests per month
24M Route requests per month
80M Points of interest (POIs)
>1PB Overall data size
500K Analytics jobs per month
8B Key/value queries per month
Examples of use cases

• Correcting and improving map data
• Inferring traffic conditions (in near real time)
• Ranking POIs for recommendations
• Understanding how people move and behave
Challenges we have encountered

• Data in silos, semantics “missing”
• Multiple sources of data (overlapping, conflicting)
• Timely processing of large volumes of data
• Partial, insufficient, inaccurate, inconsistent data
• Security, privacy, etc. policies “unknown”
• Hard to share and reuse data
Challenges we have encountered

• Data in silos, semantics “missing”
• Multiple sources of data (overlapping, conflicting)
• Timely processing of large volumes of data
• Partial, insufficient, inaccurate, inconsistent data
• Security, privacy, etc. policies “unknown”
• Hard to share and reuse data
What is a “silicon”…?

• One system/app owns and controls a set of data.

• Typically, this data has “opaque” semantics:
  - proprietary data models (semantics)
  - proprietary data formats (syntax)

• As a consequence, this makes the data hard to:
  - access (from outside)
  - reuse (by other systems)
  - integrate (with data from other sources)

• Access/reuse/integration: engineering endeavors.
Perhaps this is in our future?

Whether we are talking about data, logic or presentation, locking these in an un-reusable “silo” only further fragments our information space.
More bad news…

• There is lots of excitement around new methods and systems to process big data
  – Hadoop, map/reduce
  – NoSQL, key/value databases, etc.

• BUT: focus on scalability and optimization may have made some things weaker
  – e.g., data models implicit, no explicit constraints

• As a consequence, the notion of “silos” has been emphasized and amplified
Even more bad news…

• Data format (= syntax) is an important issue, but
  – all issues wrt. formats have already been solved
  – once you decide on syntax, you should forget about it

• There is a persistent belief that as long as you understand the syntax, you have “solved the problem” (unfortunately not so)

• People are mistakenly focused on syntax
  – evidence: current public discussions on how to improve JSON focus on changing the syntax – seriously!
Things I am **NOT** interested in:

- Specific, narrow use cases (those can always be implemented one way or another)
- Converting all the world’s data to RDF ;-

Things I am **interested in**:

- Ensuring I don’t have to specify use cases in advance to build a big data platform
- Enabling sharing and *ad hoc* usage of big data
I really like the Semantic Web because…

...these technologies promote **serendipity** in

**interoperability:** is it possible to interoperate with systems and services we knew nothing about at design time?

**reuse:** when information has accessible semantics, this is easier…

**integration:** can information from various independent sources be combined?
Phase 1: Describe data thoroughly

• We need proper descriptions of
  – data models
  – operational parameters
  – policies + metadata to evaluate policies against

• Also: we need provenance
  – owner, source
  – workflow, dependencies to other datasets
Phase 1: Describe data thoroughly

• Our approach: build a “data asset catalogue” to capture enough metadata
  – initial prototype built on top of OINK [Lassila 2006]
  – eventually we settled on a more “traditional” implementation, but we need OWL for data models

• Serves as an entry point to “data discovery”, enabling automation, sharing, reuse

• Insulates SMEs from (irrelevant) physical details of our datasets
Phase 1: Describe data thoroughly

OINK-based prototype

production system
Phase 1: Describe data thoroughly

- Using OWL to describe conceptual (“semantic”) data models works fine, but production data usually has some other type of (physical) schema (SQL DDL, HiveQL DDL, even JSON Schema)

- Could we map “real world” physical schemata onto Semantic Web ontologies?
Phase 2: Ontologies as a virtual layer

• So far, only a PoC experiment...

client program

SPARQL query

D2R Server (modified)

HiveQL query via JDBC

Hive ("server mode")

Map/Reduce access to HDFS

HDFS/Hadoop

translates SPARQL queries into (potentially many) SQL/HiveQL queries, virtually constructs an RDF graph

translates relational queries into Map/Reduce programs, creating a virtual view of HDFS flat files as relational data
Some conclusions

• Data comes in many formats (and if you are lucky there might even be a physical data model associated)

• Better descriptions of data are needed (actionable metadata); goal: automation

• Semantic Web as a virtual layer; hide diversity

• These are only really the first steps, I have not said anything about how we could, for example, use reasoning…
Thank you!

• Questions, comments?

• Short rants:  @gotsemantics

• Long(er) rants:  http://www.lassila.org/blog

• Contact:  ora.lassila@nokia.com

• Thanks to:  Amy O’Connor, Yekesa Kosuru, Ian Oliver, Bob Savard, Tasneem Jodhpurwala