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

Dr. Ora Lassila

Principal Technologist
Cloud Analytics Team
Nokia Location & Commerce

&

Elected Member
Advisory Board
World Wide Web Consortium (W3C)

2012-11-11

ISWC 2012
The 11th International Semantic Web Conference

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

NOKIA
Connecting People

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 analytic
- Elected member: W3C's Adv
- Research (Nokia, MIT
  capitalist, entre engineer
- Ph.D (D.S. Univ. of Technology
- S rch and standardization work
  d 2001 article on the Semantic Web
  haracterized as science fiction)
  – editor of the original RDF specification

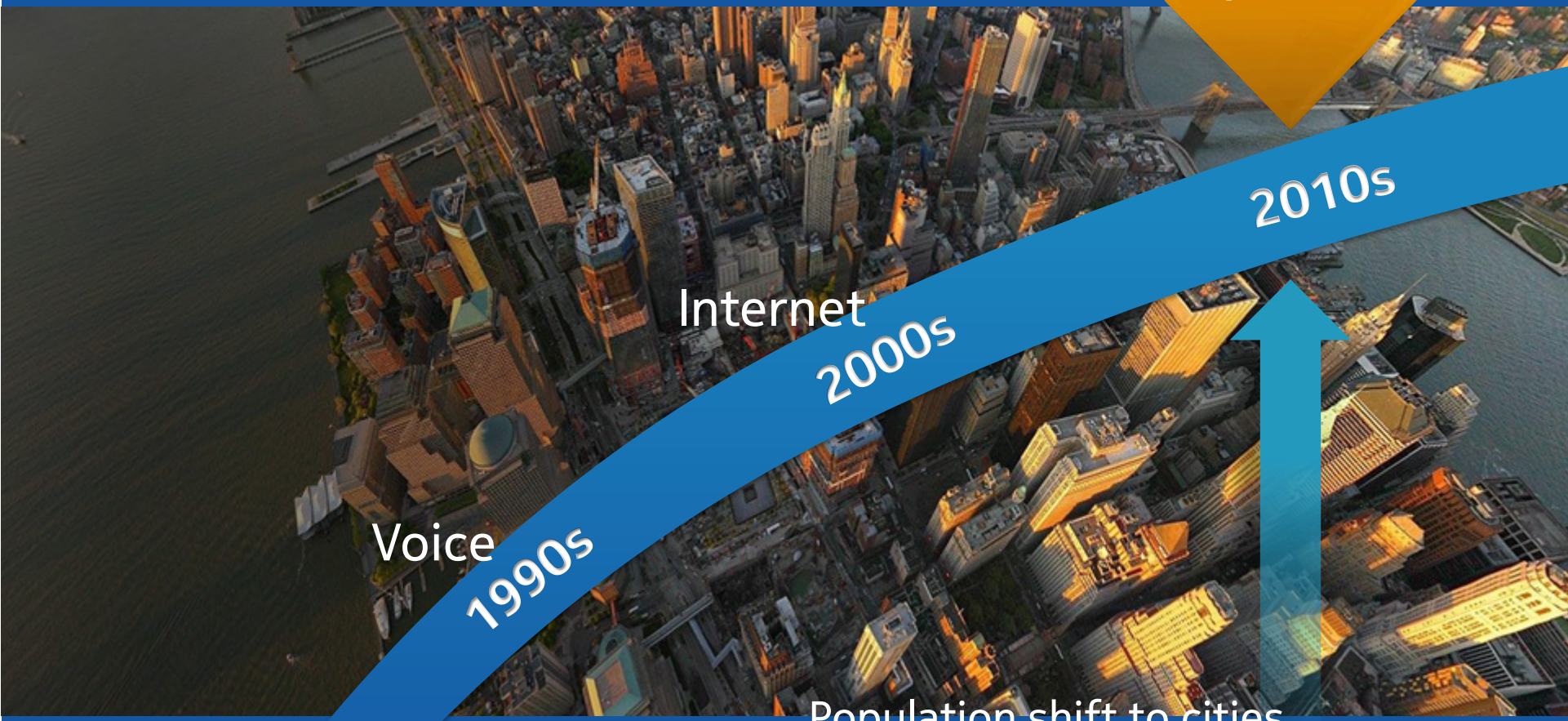
NOKIA
Connecting People

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

NOKIA
Connecting People

Nokia Maps – some of our customers

foursquare
inSing.com
GARMIN
ВКонтакте
yelp
bing
SingTel
wcities
sina.com.cn 新浪网
WRC FIA WORLD RALLY CHAMPIONSHIP
tripadvisor® get the truth. then go.™
YAHOO!
lonely planet
Windows Phone
Tencent 腾讯
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

NOKIA
Connecting People

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

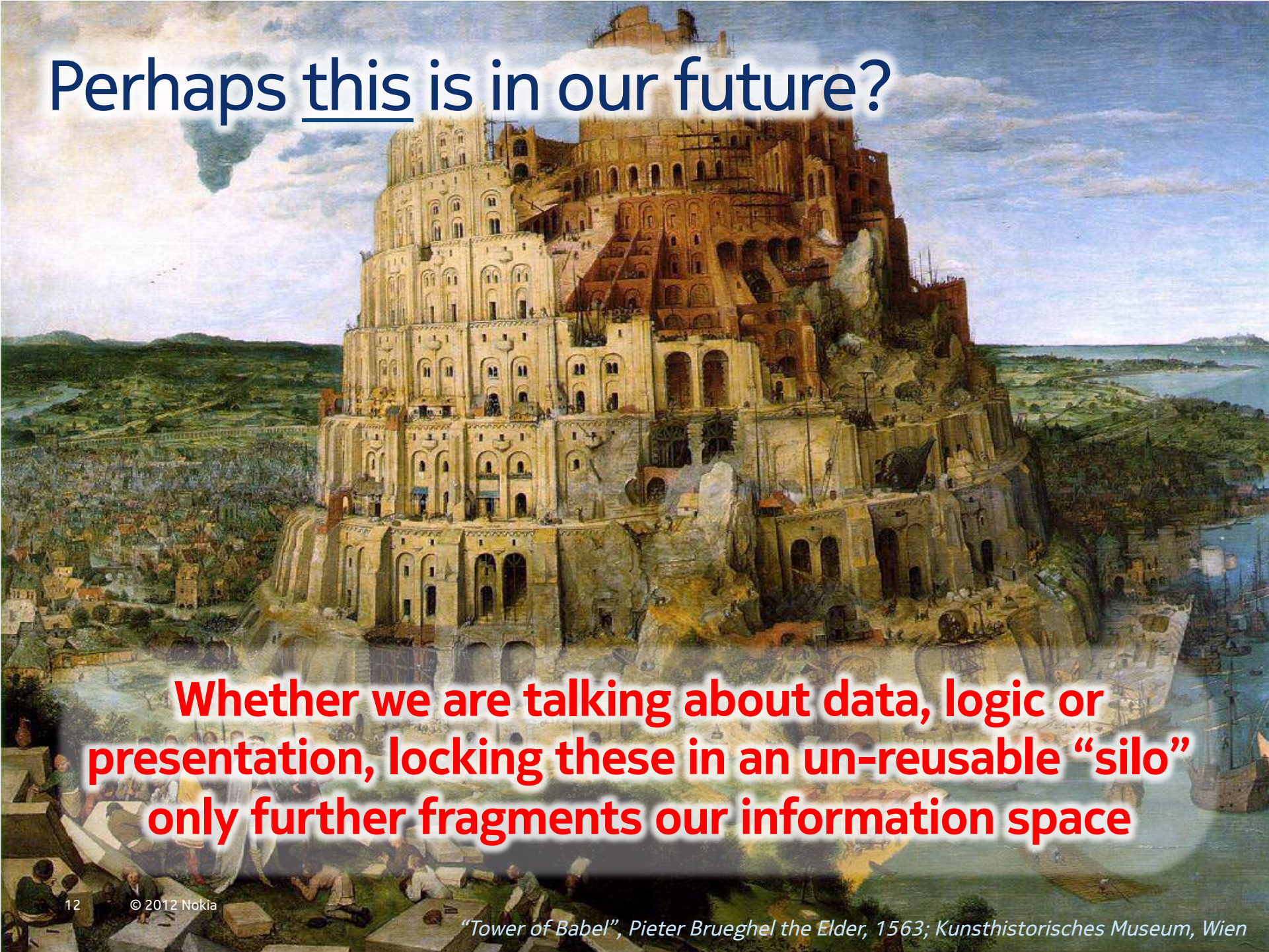
Challenges we have encountered

- 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 “silo”…?

- 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

"Tower of Babel", Pieter Brueghel the Elder, 1563; Kunsthistorisches Museum, Wien

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?

NOKIA
Connecting People

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

ACP

http://localhost:8081/node?id=1318

Bing

ACP

OINK

- home
- find

Asset | Metadata | Referrals (1)

Description: Analytical Consumer Profile
Last modified: Wednesday, 14. September, 2011 @ 12:05:00PM
Sources: SCBE Access Logs, NoA/SSO Event Logs
Dependents: (none)
Processing log: 0 entries
Logical model: (none)

ACP

http://localhost:8081/node?id=1318

Bing

ACP

OINK

- home
- find

Asset | Metadata | Referrals (1)

Also known as: catalog:3, http://dataos.nokia.com/analytics/assets/3

type	Thing, Computed Asset, Asset, 1., Catalog Object, Resource
title	*ACP*
description	*Analytical Consumer Profile*
status	*pre-production*
created on	*2011-09-14T12:00:00+00:00* ^^xsd:dateTime
modified on	*2011-09-14T12:05:00+00:00* ^^xsd:dateTime
has customer identifier	*true* ^^xsd:boolean
has device identifier	*true* ^^xsd:boolean
depends on	NoA/SSO Event Logs
depends on	SCBE Access Logs
data location	hdfs://daca2.us.msudev.noklab.net:8020/services/a
has backlog task	http://dataosd.americas.nokia.com:8080/browse/DC

Home | Search | Help

- Public Transit
- IdM
- Common Dimensions
- NAC
- Mobile Phones
- CDN
- Places Api
- Omniture
- SCBE
- Maps
- Microsoft
- Drive
  - MapSensor
    - Drive Probes
- Search

Drive Probes

Asset Name: Drive Probes
Asset ID: C55771A9-8CA0-0001-A8F0-1800505F8800
Asset Description: MapSensor Service is a service that belongs to the Map Platform/Traffic team.The MapSensor Service processes mainly the GPS probes that are collected by Nokia devices. The collected data is distributed to multiple systems, including Traffic content systems and Data OS Clusters.The probe data is collected by mobile clients and is either sent directly to server or stored on disk for later dispatch when device has WiFi connection or when device is connected to Nokia Ovi Suite

Created On: 2012-09-11 13:44:08.0
Data URL: hdfs:/services/mos/mapsensor/
Service/Application: Drive/MapSensor
Geographical Location(s):
Compression Type:
Asset Server Type:
Master of data source:
Retention duration:
Daily Amount:
Frequency: streaming
Ingestion Method: scribe
Serialization Format:
Delimiter: ,

Contains Customer Identifiable Data: false
Contains Device Identifiable Data: false
Contains Location Identifiable Data: true

Attachments:

None

Additional info:

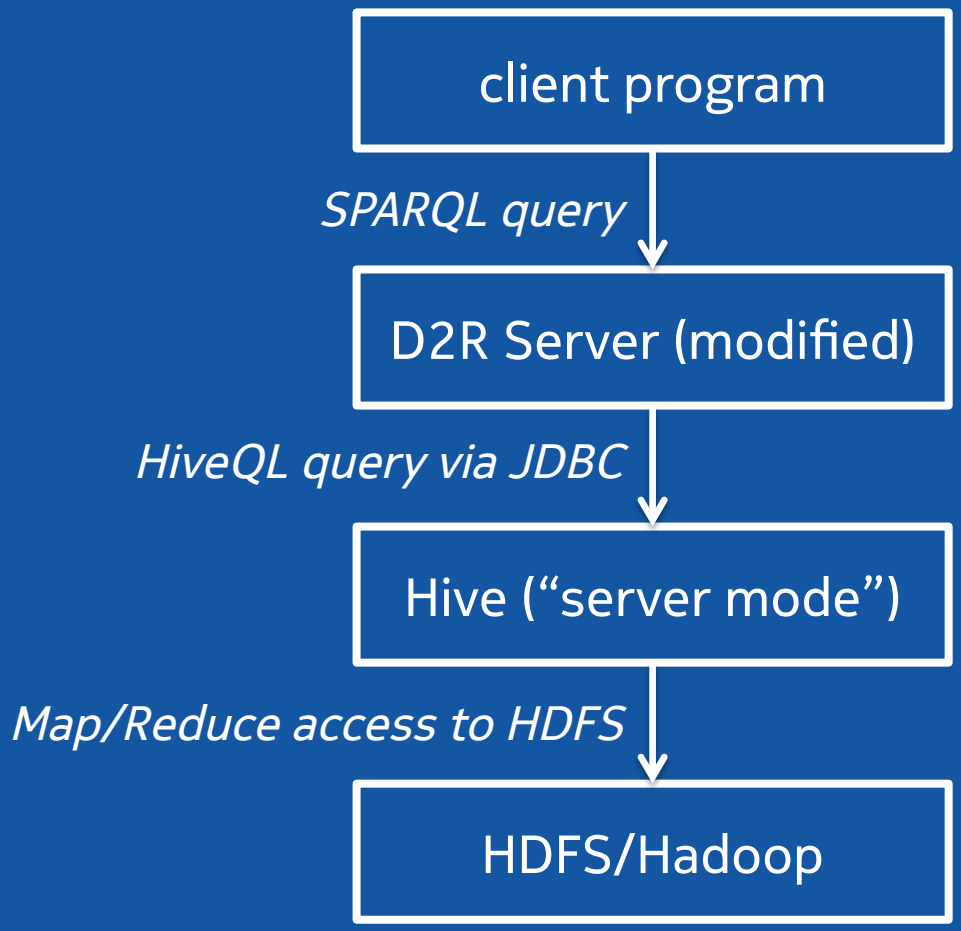
Map Sensor Wiki	Map Sensor project wiki
Privacy Policy	Privacy Policy for data usage
Confidentiality Agreement	Confidentiality Agreement required for use

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…



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

NOKIA
Connecting People

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*