Applying Hadoop to Semantic Big-Data Processing

CIPSI and Big Data

Tomasz Wiktor Wlodarczyk

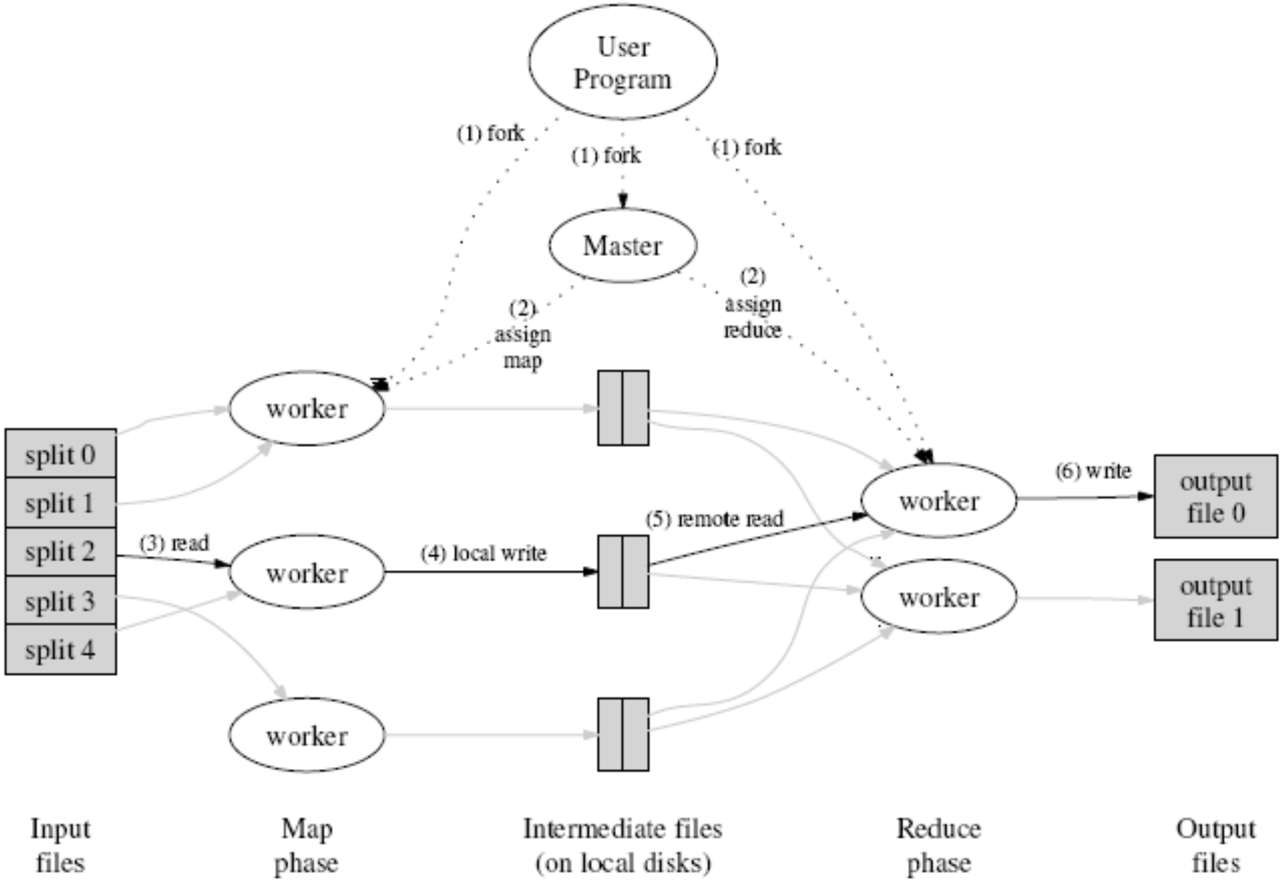
Chunming Rong

University of Stavanger, Norway

Index

- **Querying in Hadoop-based triple store**
- Scaling-out NCBO Resource Index
- CIPSI and Big Data

What is Hadoop?



(CC) Introduction to Parallel Programming and MapReduce - Google

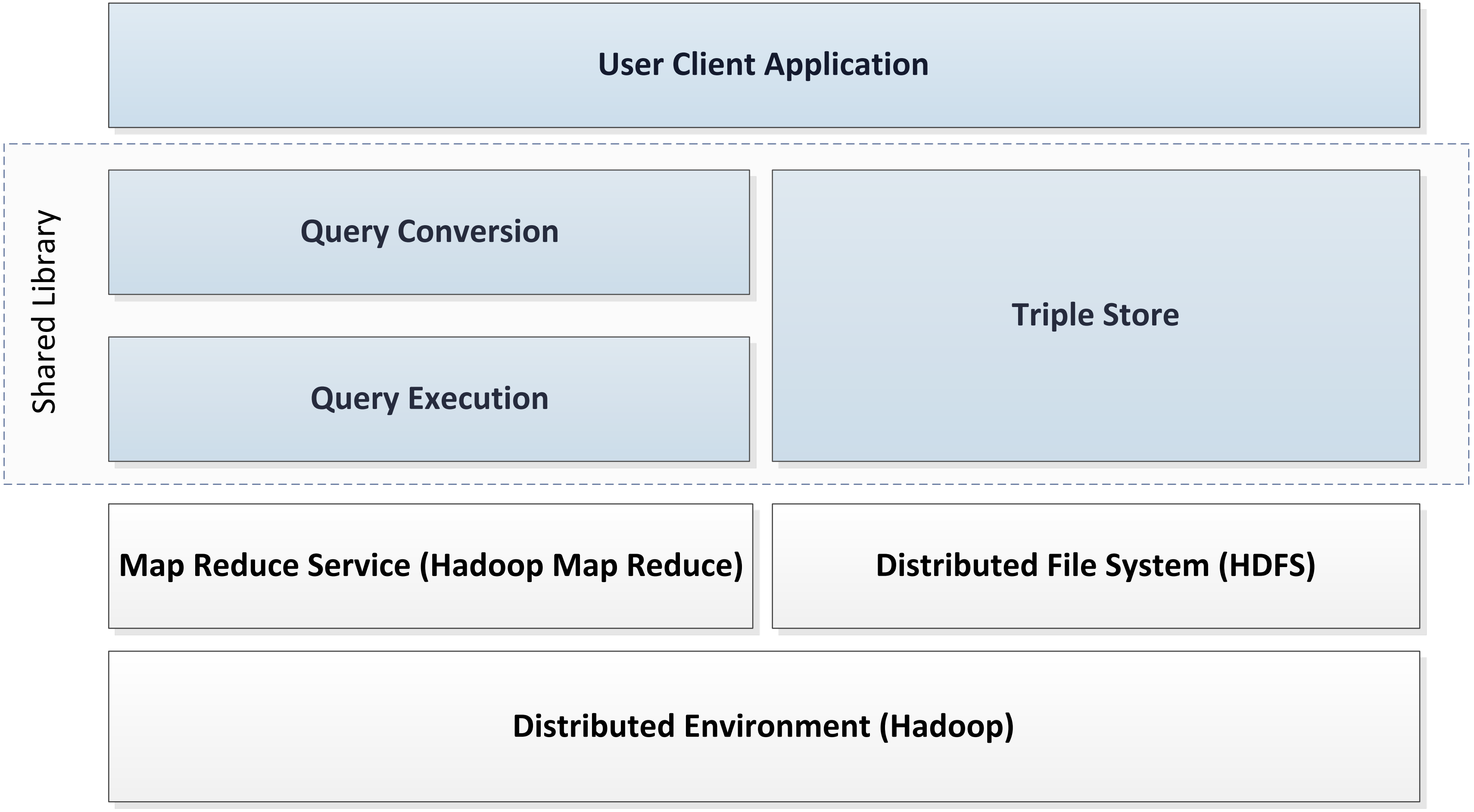
Querying in Hadoop - Motivation

- The goal was to:
  - Demonstrate querying on triple-based data without any special storage structures on a distributed system
  - Demonstrate querying for implicit data in a real-life scenario (dynamic datasets)

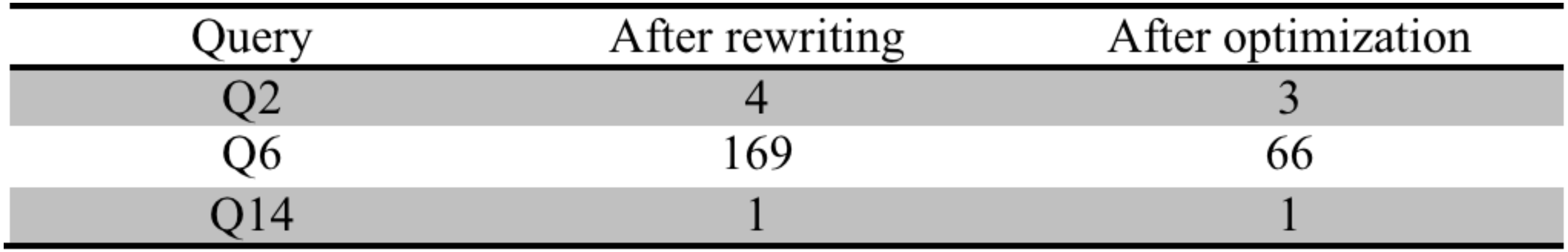
Querying in Hadoop - Data and Queries

- Lehigh University Benchmark
- 50, 100, 200, **500**, 1000 and 6000 universities
- 13.6MB for 50 U and 2.9GB for 6000 U
- Queries used were
  - 1 – high selectivity, no reasoning required
  - 2 – complex interdependence pattern, simple reasoning
  - 6 – low selectivity, complex reasoning
  - 14 – low selectivity, no reasoning required
- LUBM focuses on RDF(S) and it was a problem in the context of OWL
- Run using on Amazon EC2 2-**10**-20 small and **large** nodes using AWS in Education Research Grant

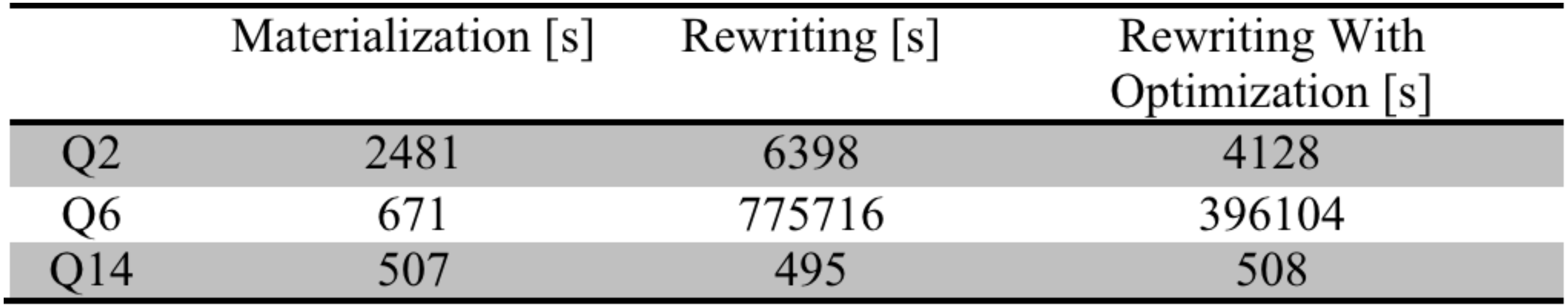
Querying in Hadoop - Architecture


User Client Application
Shared Library
Query Conversion
Query Execution
Triple Store
Map Reduce Service (Hadoop Map Reduce)
Distributed File System (HDFS)
Distributed Environment (Hadoop)

Querying in Hadoop – Results (Reasoning)

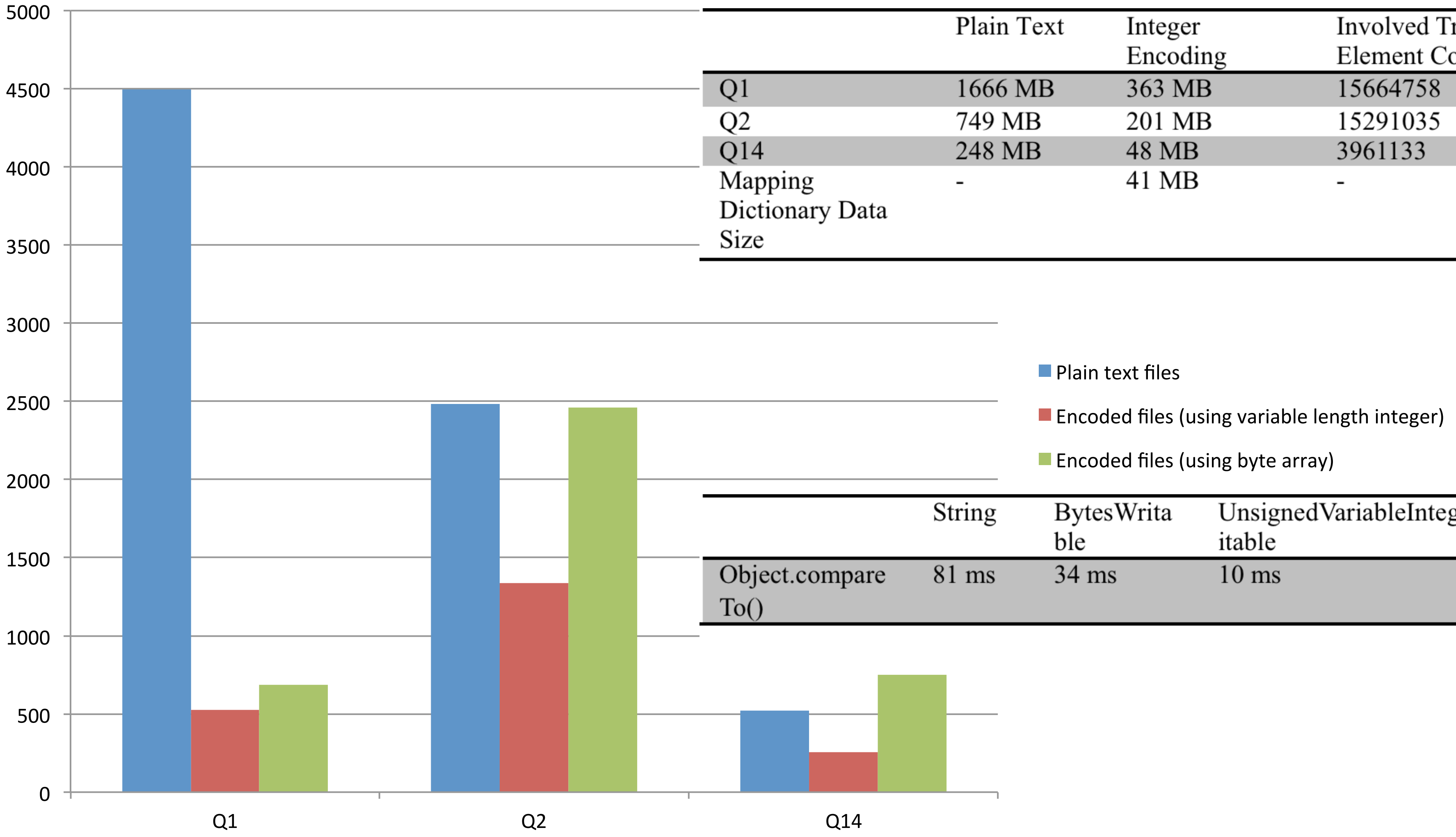


Query	After rewriting	After optimization
Q2	4	3
Q6	169	66
Q14	1	1



	Materialization [s]	Rewriting [s]	Rewriting With Optimization [s]
Q2	2481	6398	4128
Q6	671	775716	396104
Q14	507	495	508

Querying in Hadoop – Results (Encoding)



	Plain Text	Integer Encoding	Involved Triple Element Count
Q1	1666 MB	363 MB	15664758
Q2	749 MB	201 MB	15291035
Q14	248 MB	48 MB	3961133
Mapping Dictionary Data Size	-	41 MB	-

	String	BytesWritable	UnsignedVariableIntegerWritable
Object.compareTo()	81 ms	34 ms	10 ms

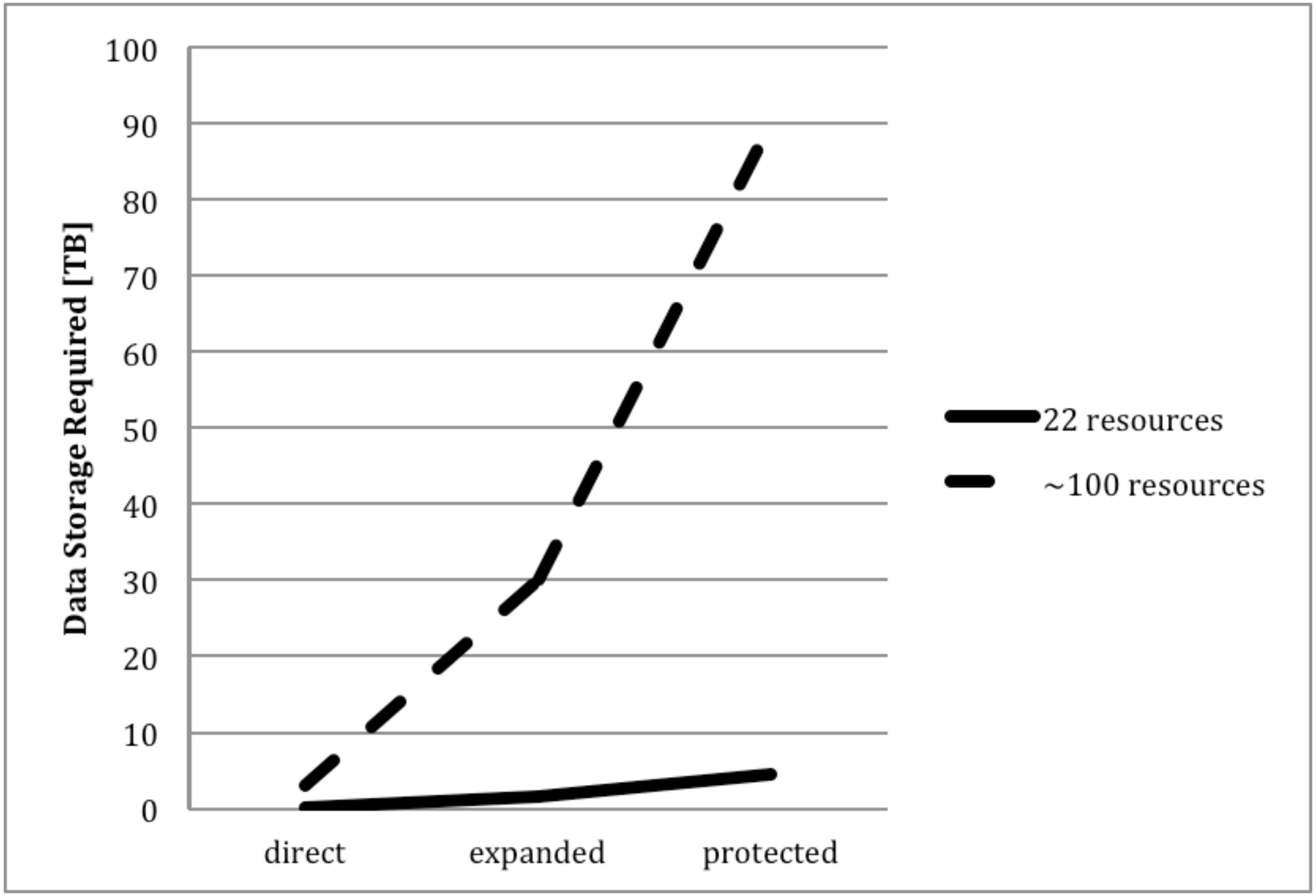
Index

- Querying in Hadoop-based triple store
- **Scaling-out NCBO Resource Index**
- CIPSI and Big Data

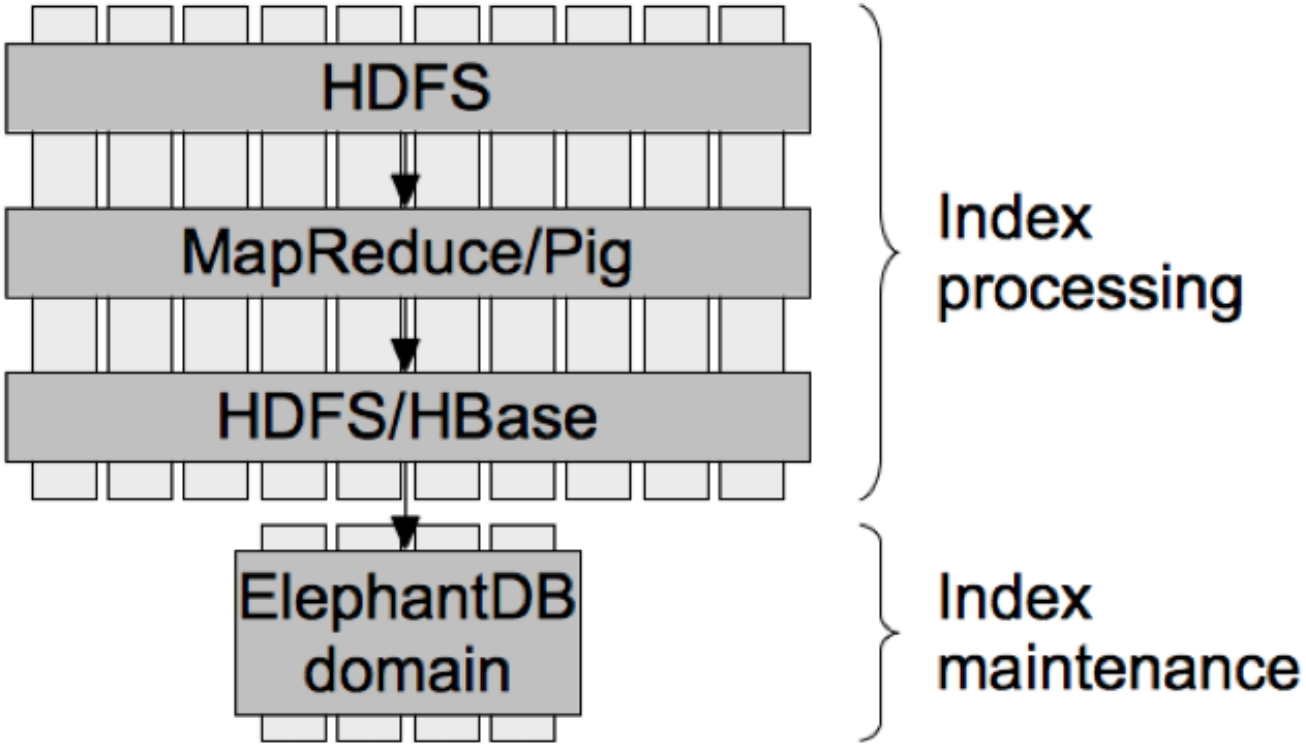
NCBO Resource Index

- System for ontology based annotation and indexing of biomedical data
- Enable users to locate biomedical data resources related to particular concepts
- Semantic expansion is used to create search index

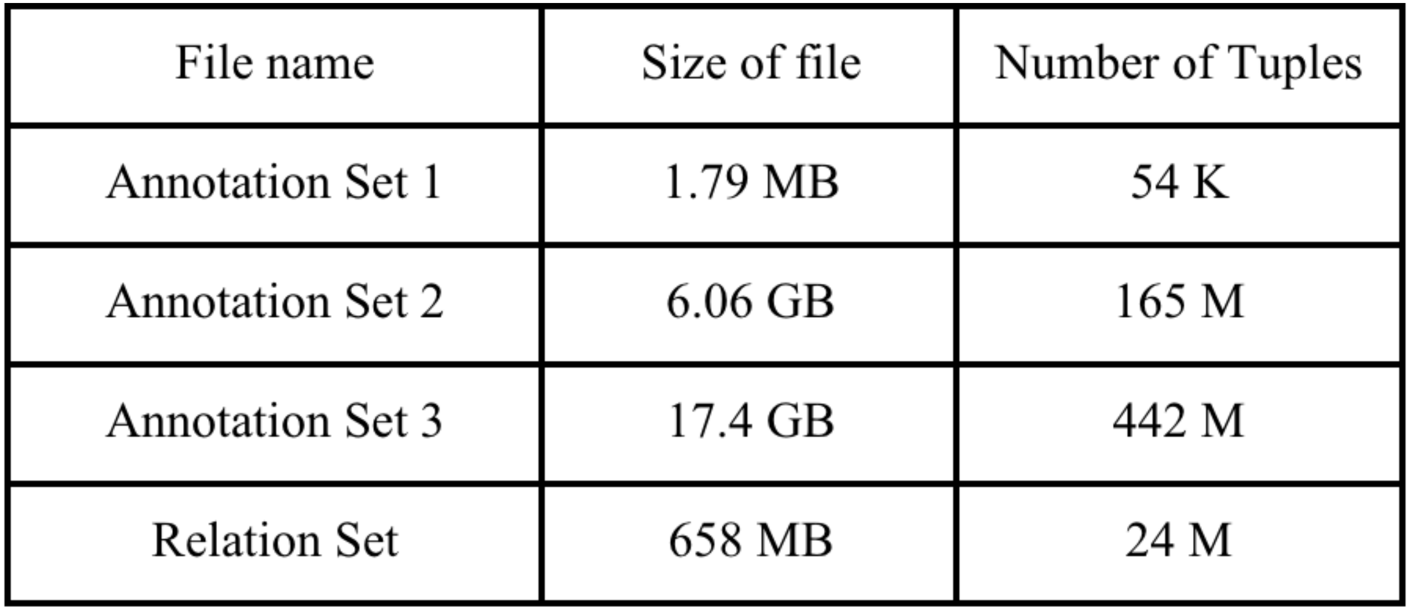
Scaling-out NCBO RI - Motivation



Scaling-out NCBO RI - Architecture

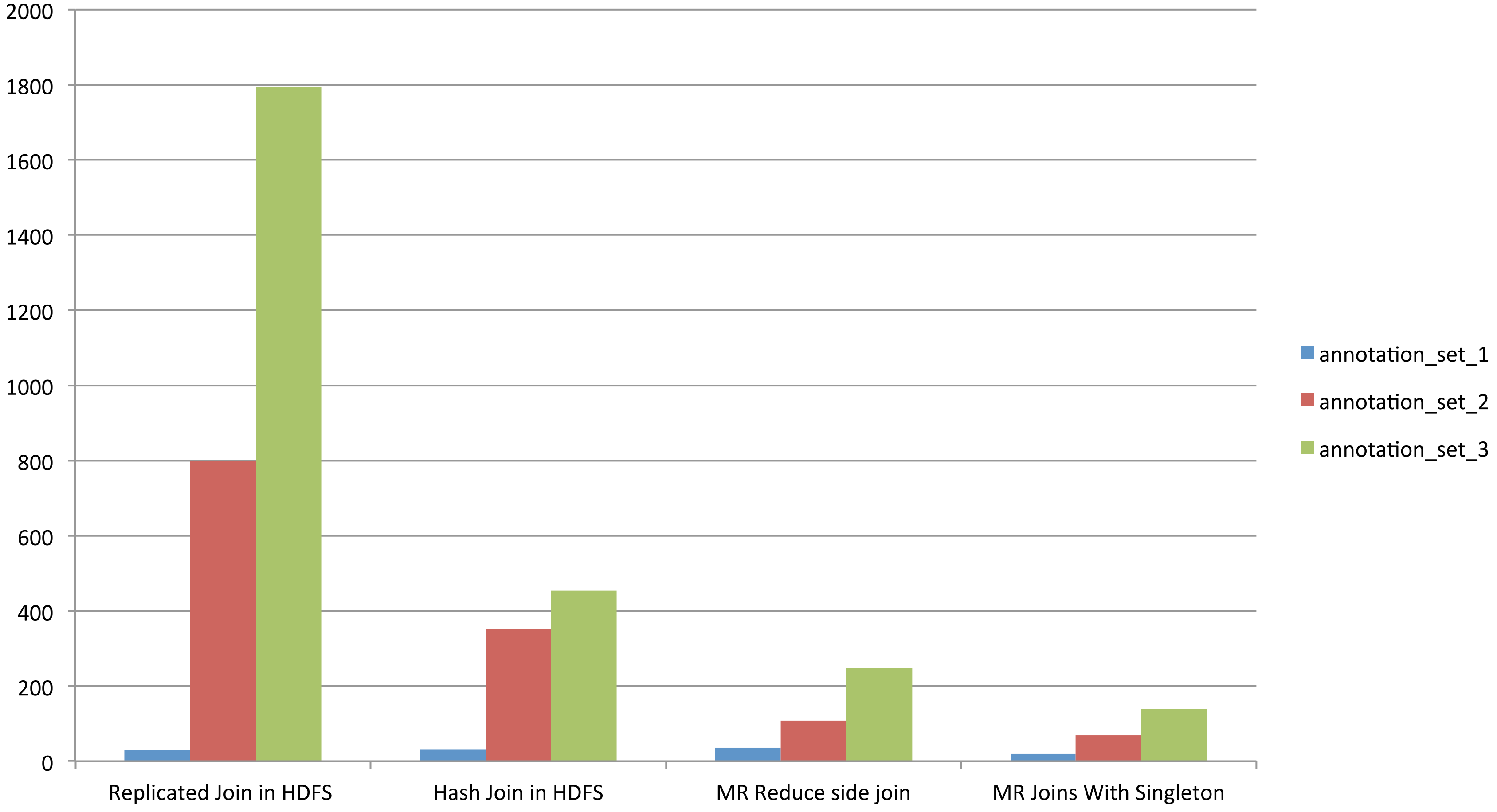


Scaling-out NCBO RI - Sets Used

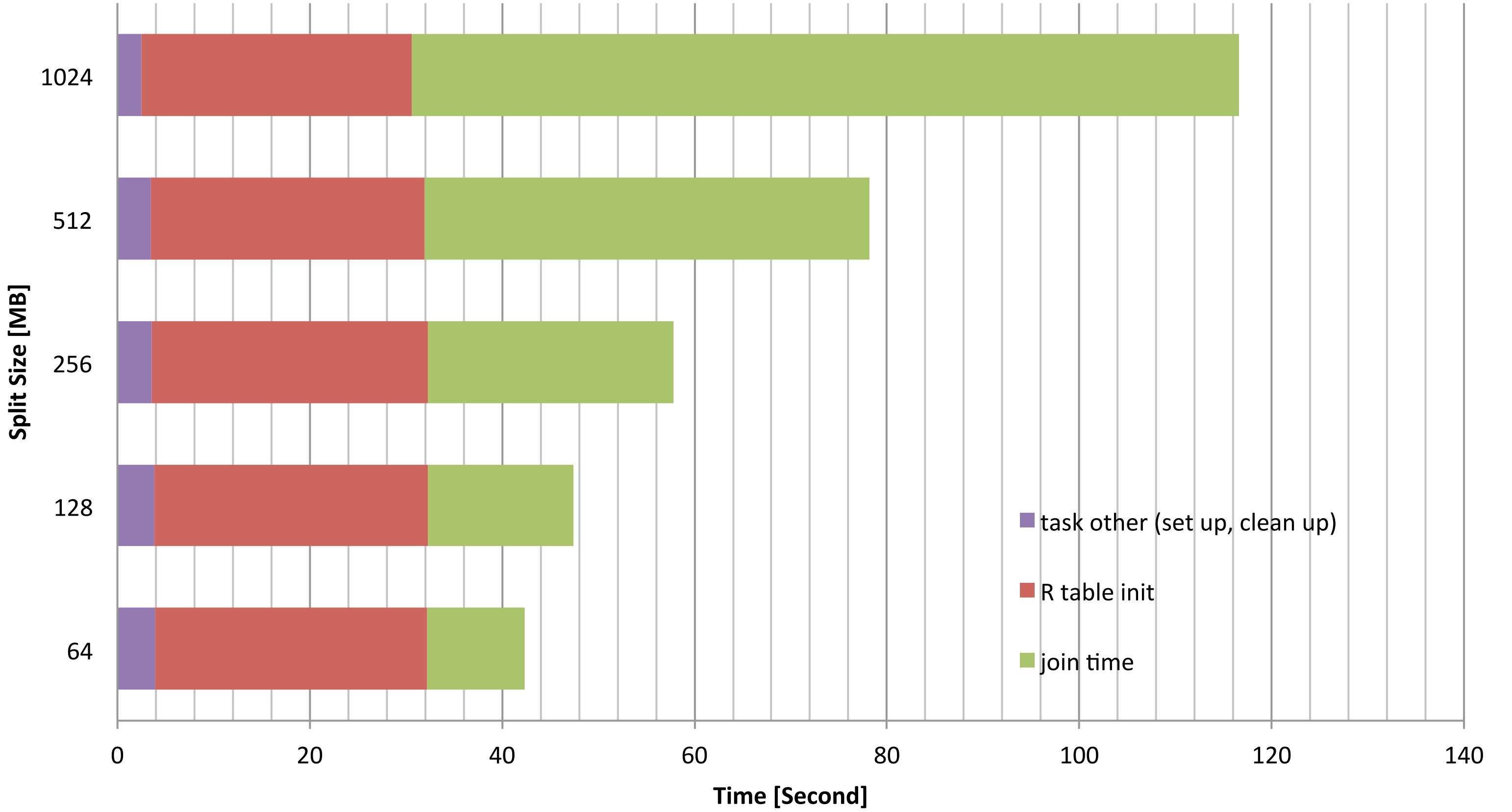


File name	Size of file	Number of Tuples
Annotation Set 1	1.79 MB	54 K
Annotation Set 2	6.06 GB	165 M
Annotation Set 3	17.4 GB	442 M
Relation Set	658 MB	24 M

Scaling-out NCBO RI - Results



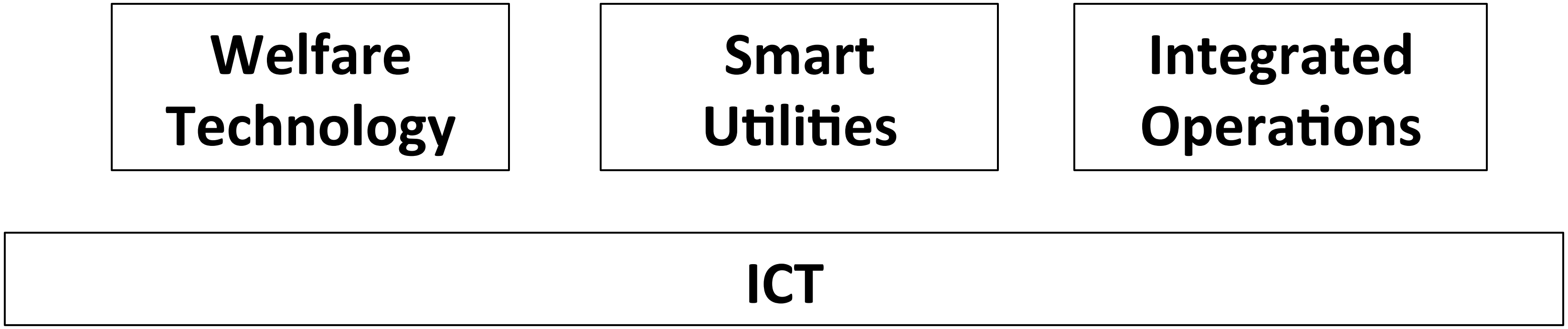
Scaling-out NCBO RI - Results



Index

- Querying in Hadoop-based triple store
- Scaling-out NCBO Resource Index
- **CIPSI and Big Data**

CIPSI:
Center for IP-based Service Innovation


Welfare Technology

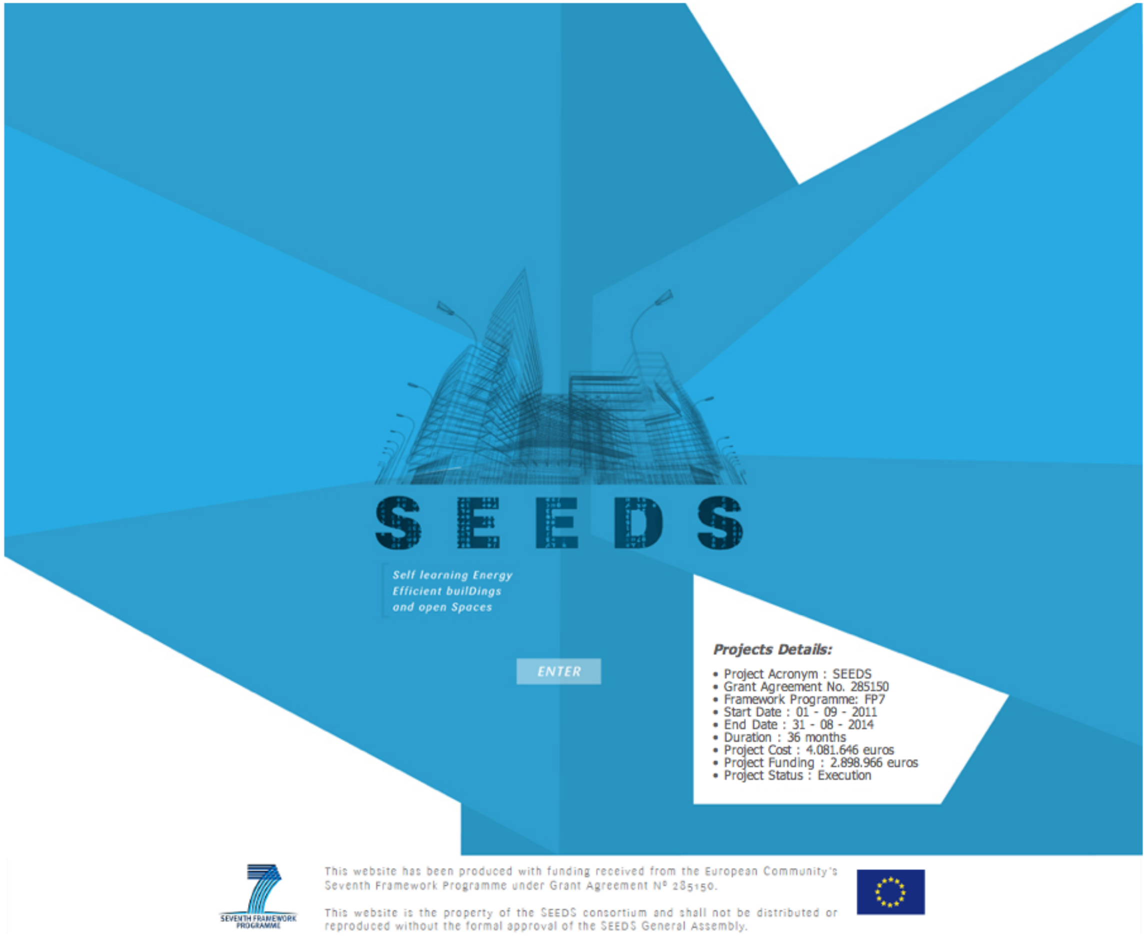
Smart Utilities

Integrated Operations

ICT

- Flere NFR og EU FP7 prosjekter ble igangsatt i tett samarbeid med Lyse/Altibox og andre lokal industriell aktørene, IRIS og SINTEF.
- Vi skal utvide vår nasjonal og internasjonal nettverk med ambisjon av SFI/NCE status og EU prosjekter i kommende årene.
- Kontaktperson: *chunming.rong@uis.no*

SEEDS: Self-learning Energy Efficient Buildings and Open Spaces (EU FP7)



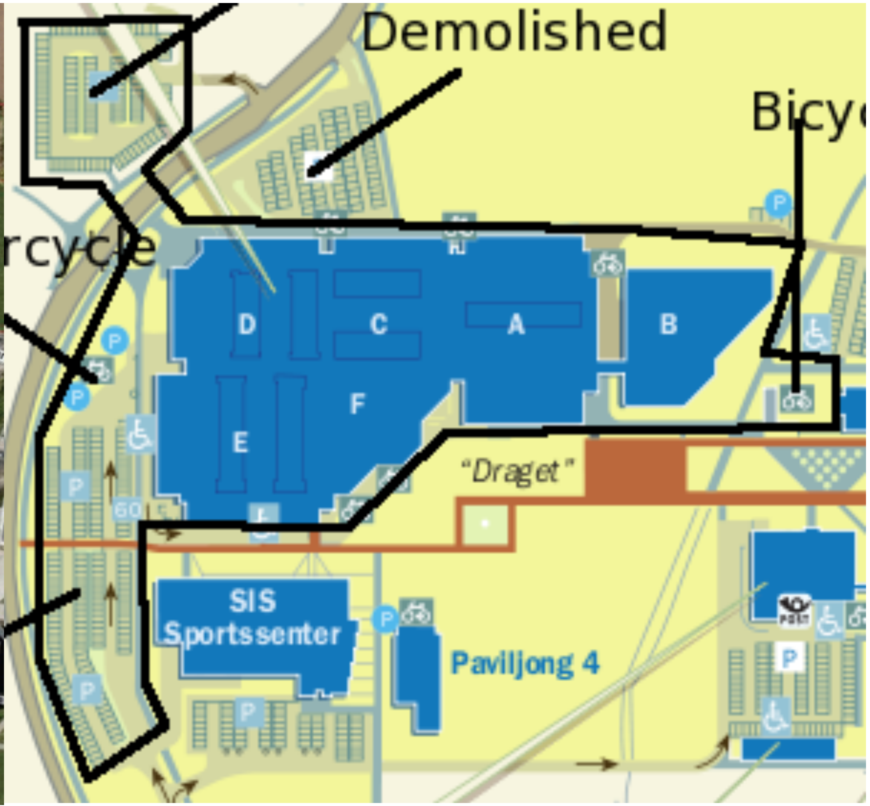
Projects Details:

- Project Acronym : SEEDS
- Grant Agreement No. 285150
- Framework Programme: FP7
- Start Date : 01 - 09 - 2011
- End Date : 31 - 08 - 2014
- Duration : 36 months
- Project Cost : 4.081.646 euros
- Project Funding : 2.898.966 euros
- Project Status : Execution

This website has been produced with funding received from the European Community's Seventh Framework Programme under Grant Agreement N° 285150.

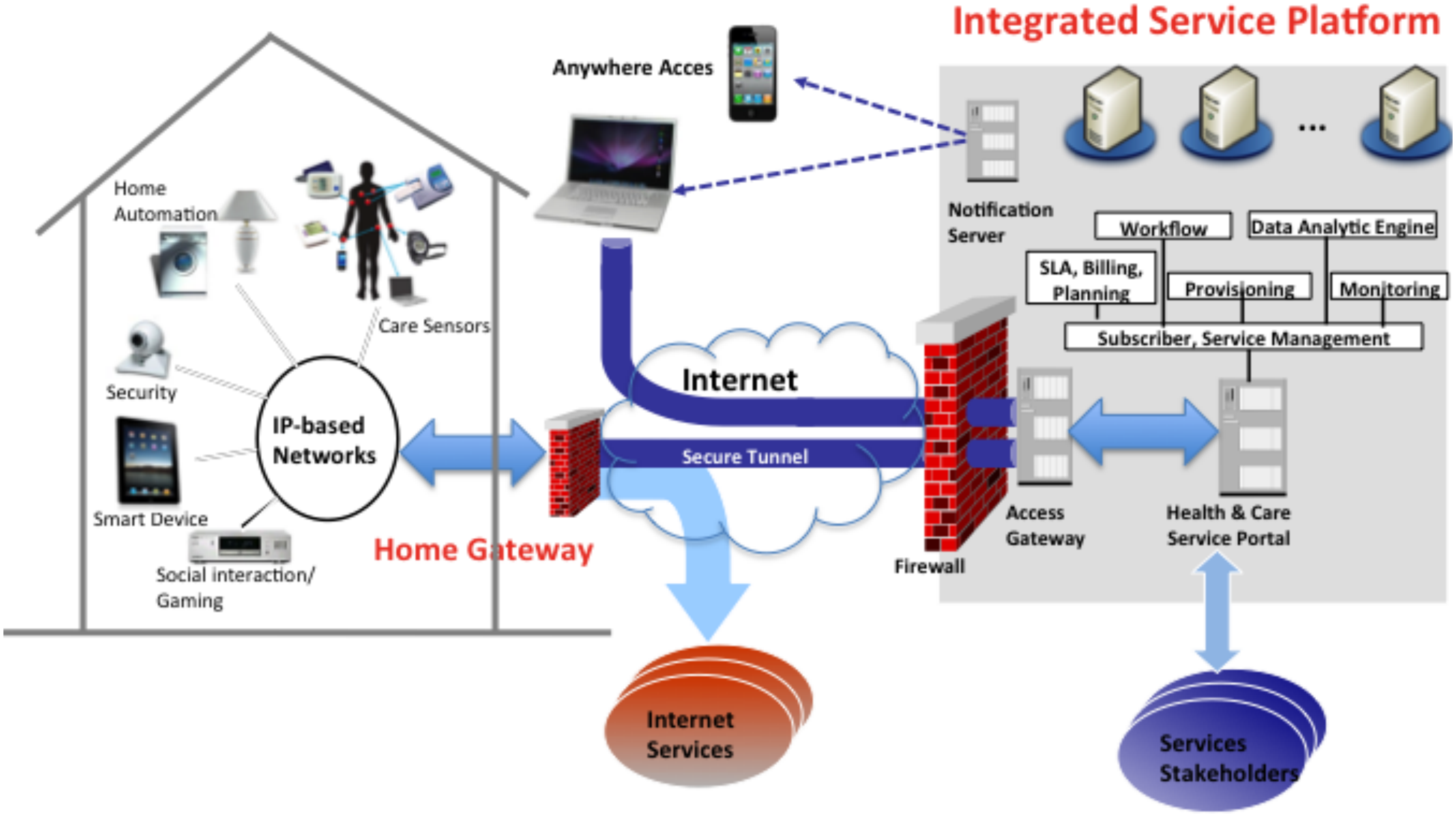
This website is the property of the SEEDS consortium and shall not be distributed or reproduced without the formal approval of the SEEDS General Assembly.

SEEDS: Self-learning Energy Efficient Buildings and Open Spaces (EU FP7)

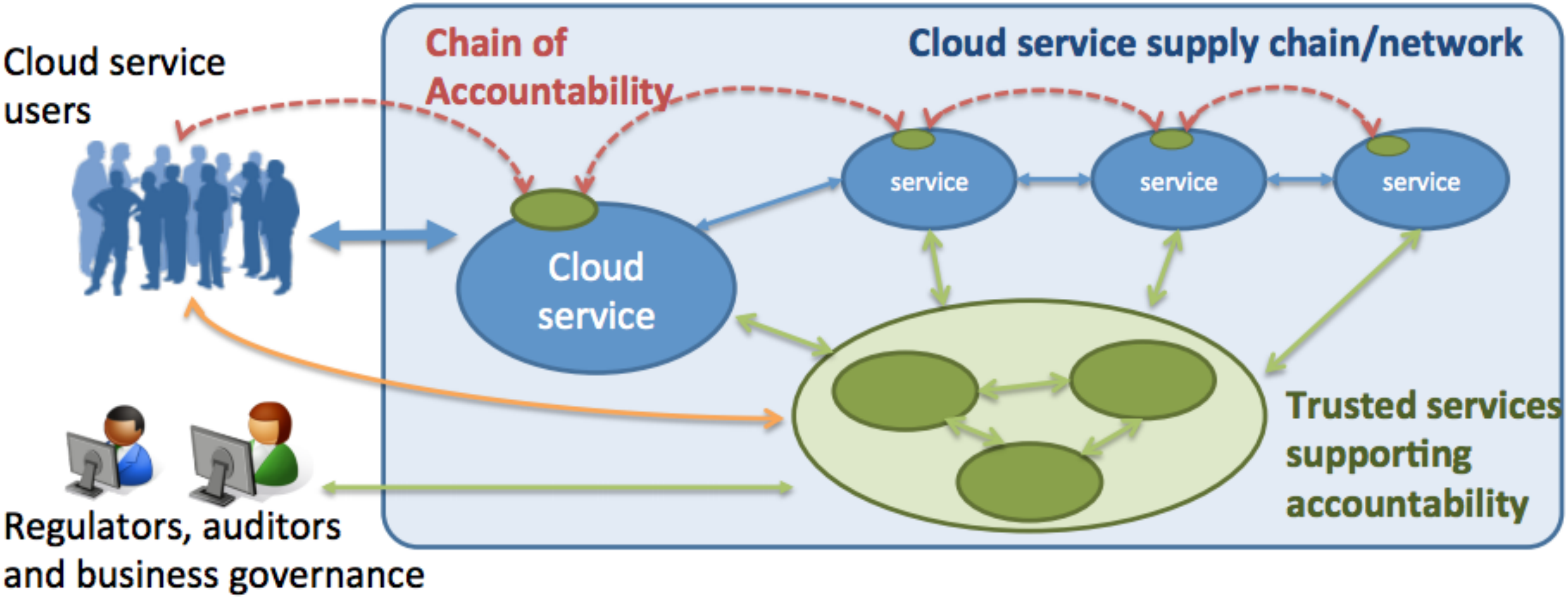


Safer@Home - Smart System to Support Safer Independent Living and Social Interaction for Elderly at Home (NFR)

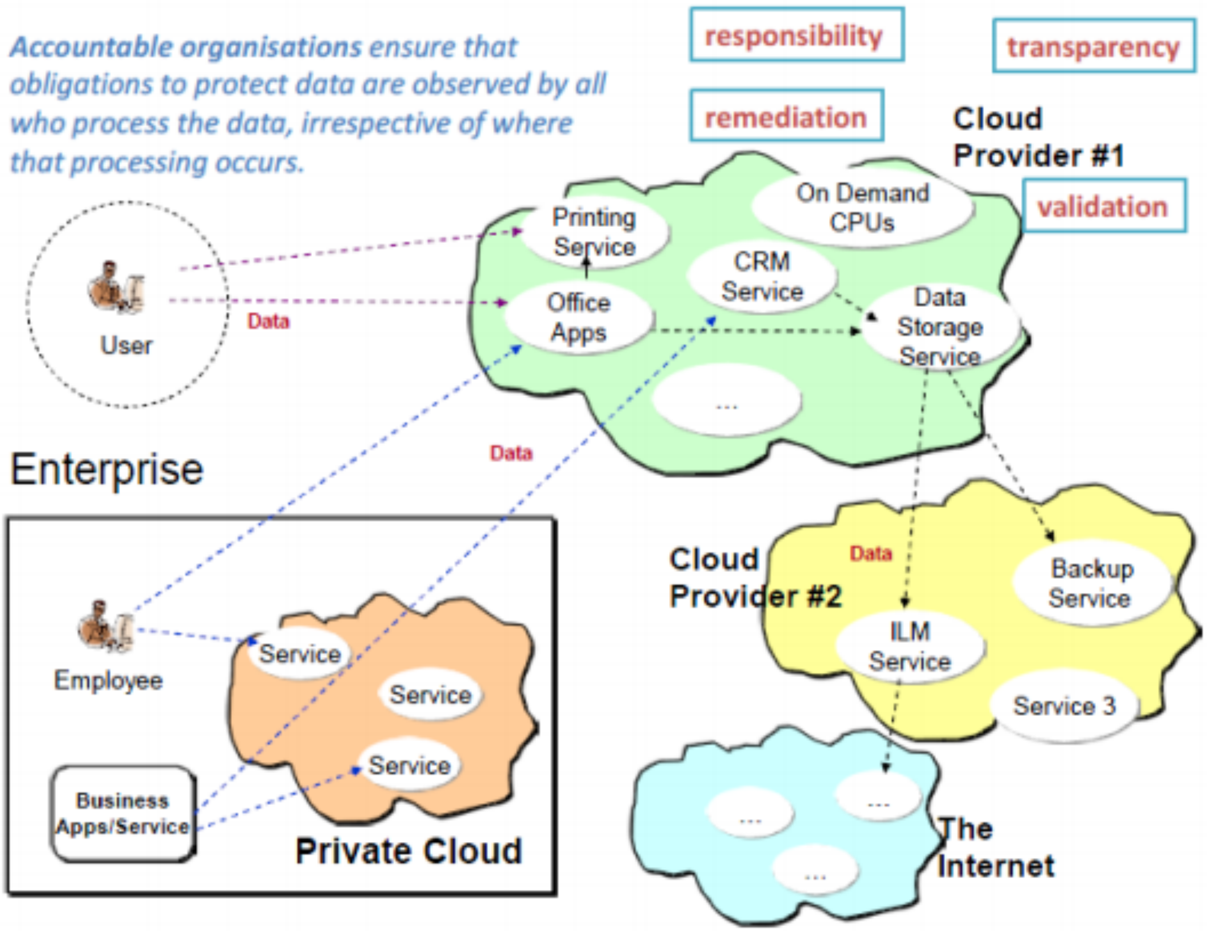
Safer@Home - Smart System to Support Safer Independent Living and Social Interaction for Elderly at Home (NFR)



A4Cloud – Accountability for Cloud (EU FP7)


Cloud service users
Chain of Accountability
Cloud service supply chain/network
service
service
service
Cloud service
Trusted services supporting accountability
Regulators, auditors and business governance

A4Cloud – Accountability for Cloud (EU FP7)



Accountable organisations ensure that obligations to protect data are observed by all who process the data, irrespective of where that processing occurs.

SCC-Computing: Strategic collaboration with China on super-computing based on Tianhe-1

- FP7-ICT-2011-7 (ICT-2011.3.4): Computing Systems
- Coordination and Support Action
- Grant Agreement
  Number 287746
- Kontaktperson: *chunming.rong@uis.no*

SEVENTH FRAMEWORK PROGRAMME

ERAC: Efficient and Robust Architecture for the Big Data Cloud (NFR)

[simula . research laboratory]

ORACLE®

UNIVERSITAS OSLOENSIS • MDCCCXI •

Universitetet i Stavanger

Lyse

Strategic Collaboration with Purdue University

- Strategic Collaboration on Big Data Analysis and Communication between Purdue University and University of Stavanger
- 4 year grant 2012-2016 from SIU
- Development of joint courses, bilateral student supervision, student exchange, guest lecturing, etc.

PURDUE UNIVERSITY

Universitetet
i Stavanger

Summary

- Querying in Hadoop-based triple store
- Scaling-out NCBO Resource Index
- CIPSI and Big Data

IEEE CloudCom 2012

- 4th IEEE International Conference on Cloud Computing Technology and Science IEEE CloudCom 2012
- The Grand Hotel, Taipei, Taiwan
- Dec 3 – 6, 2012
- 2012.cloudcom.org