Management and Transfer of Large Scientific Data

Spiros Koulouzis, Adam Belloum, Marian Bubak

Informatics Institute, University of Amsterdam, The Netherlands

www.science.uva.nl/~gvlam/wsvlam

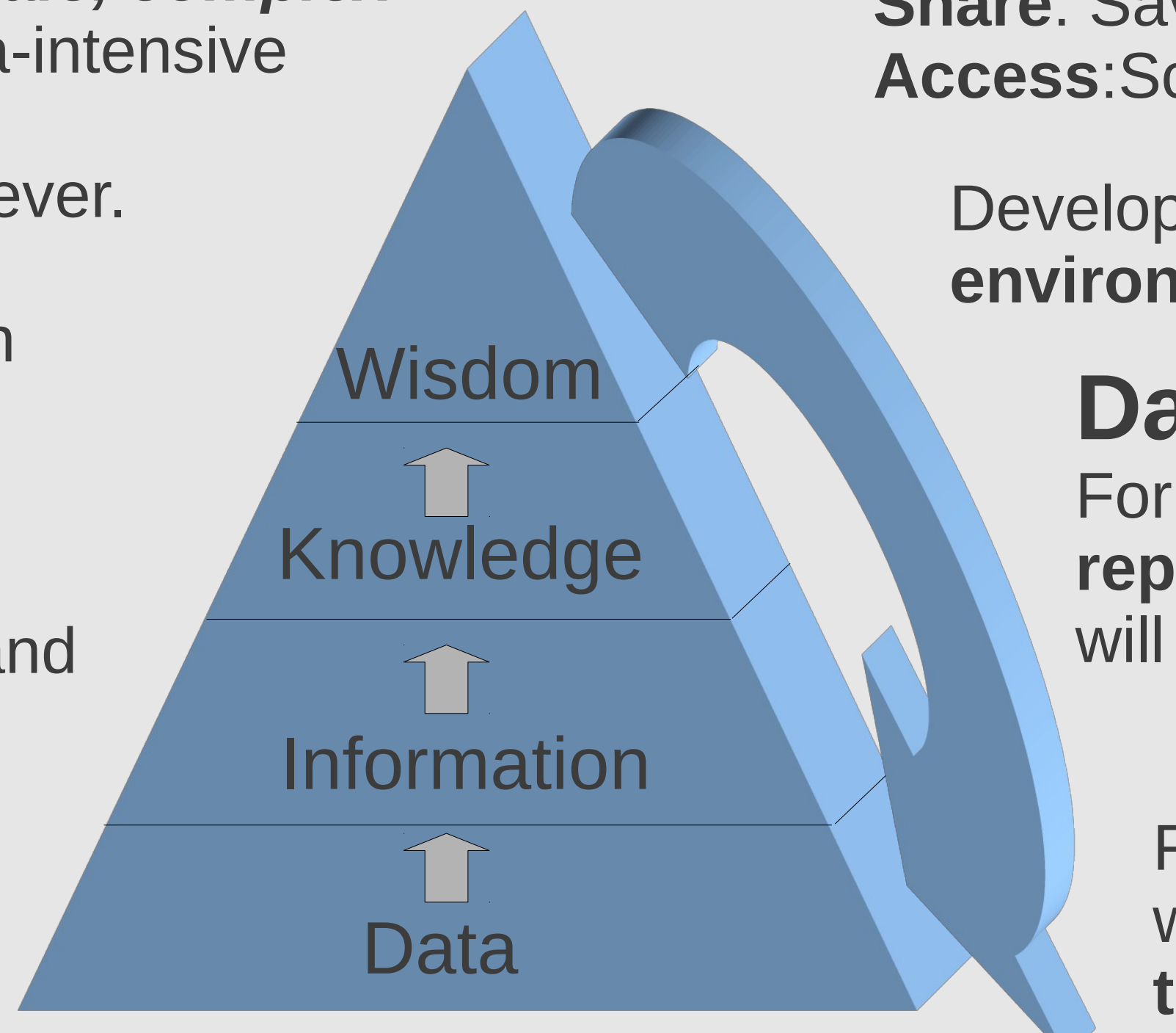
System and Network Engineering

Motivation

“*Nowadays scientists do not actually look through telescopes Instead, they are “looking” through* ***large-scale, complex data***.": [Jim Gray. The fourth paradigm: data-intensive scientific discovery].

Sciences is now producing more data then ever.

- LHC produces 15 PB/year
- LOFAR is expected to produce 1.224 GB/h



Scientific experiments generate data from instruments or simulations, process them and store the resulting information.

e-Science aims at enhancing science by enable the **sharing** of knowledge.

To achieve this eScience is promoting a **service oriented architectures** (SOA)

Challenges and Research Objectives

Manage: Who and which data to use?

Share: Save effort promote interdisciplinary research.

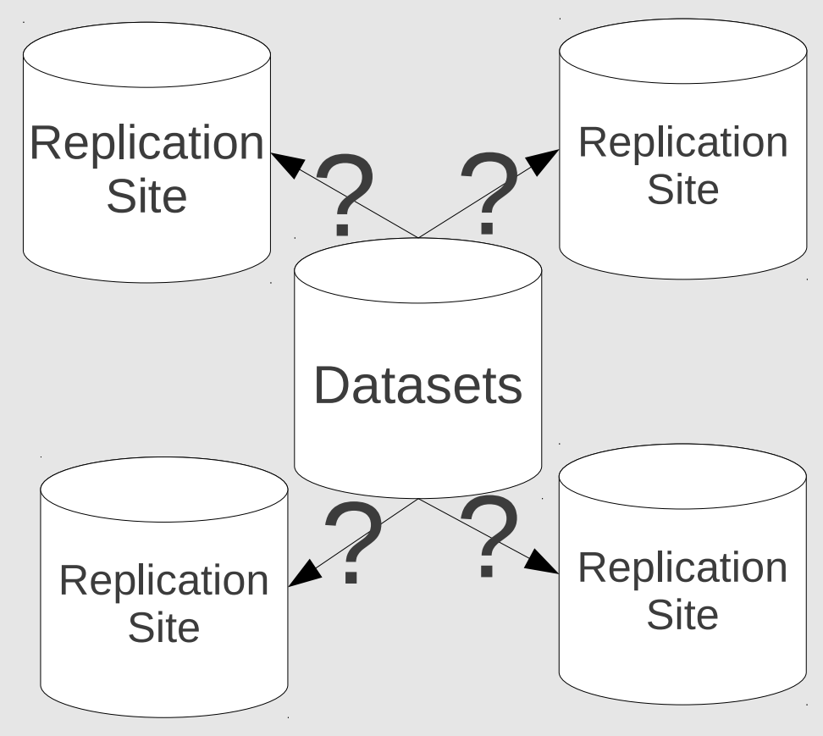
Access:Scalable access for large data

Development of a **common collaborative environment.**

Data replication

For a given dataset determine **where** and **how many replicas** will be crated.

For efficient replication we have to **move only the data that is need**, which is a small subset of a big data-set.

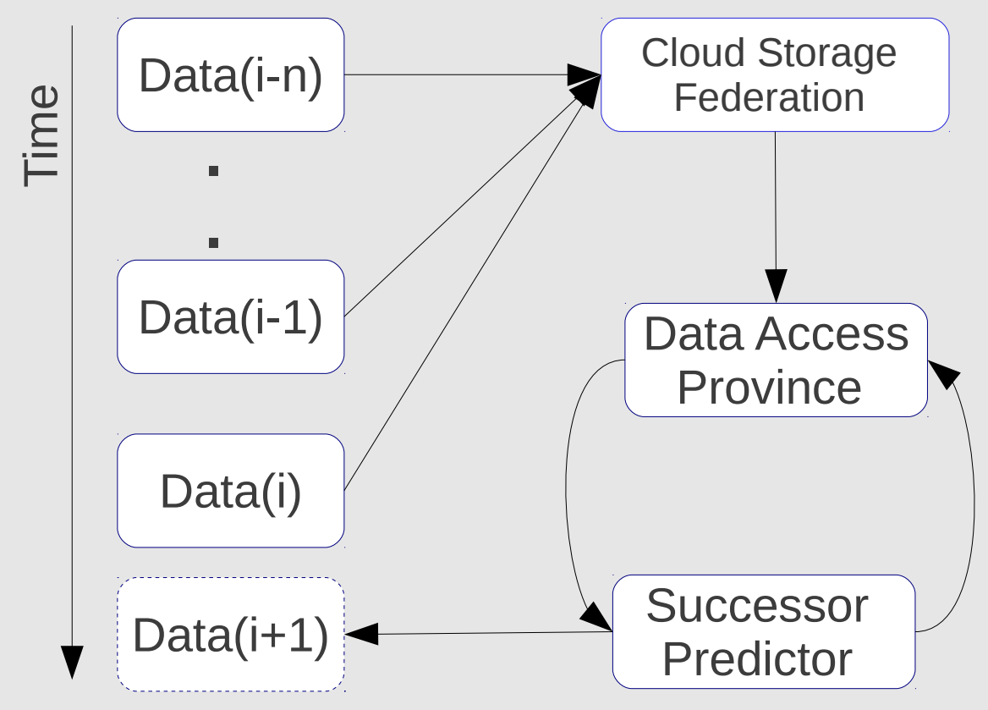


Data access prediction

What to replicate? Most scientists look at a small part of available data**.**

We need **data access**

- **prediction**
- **provenance**



prediction of the successor relationship: A mechanism for identifying inter-data relationships.

Data Streaming for SOA

Data intensive, WSs suffer from **data isolation** making the task of **moving large datasets infeasible.**

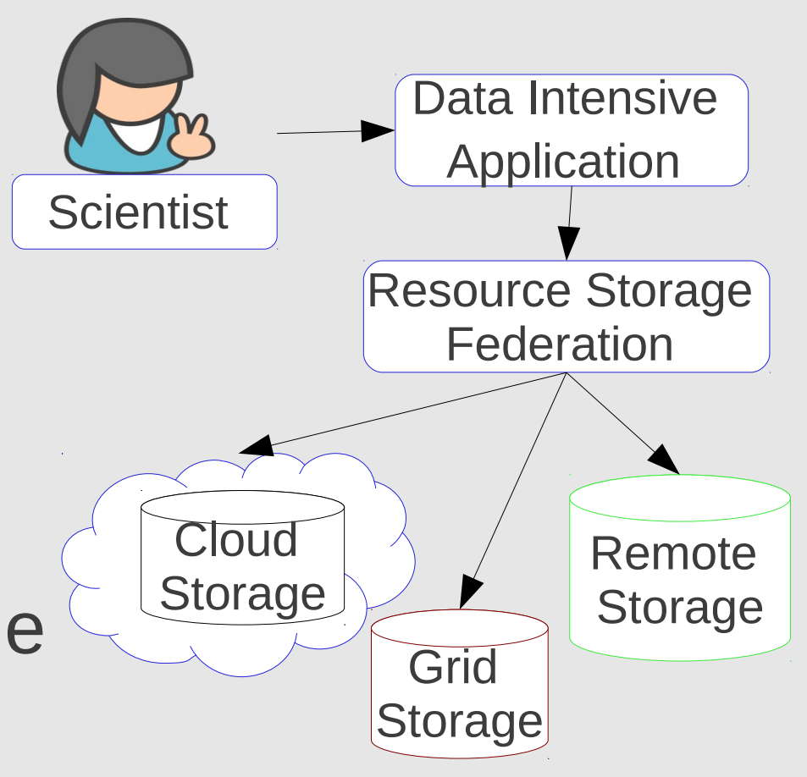
Overview of Approach

Federated Cloud Storage

Transparently integrate multiple autonomous cloud storage resources

Optimize, storage usage, speed, etc.

Efficient shearing transfer of large data-sets



Data Replication

Replicate popular data to minimum latency sites

Use 3rd party transfer

Use data striping to increase transfer speed

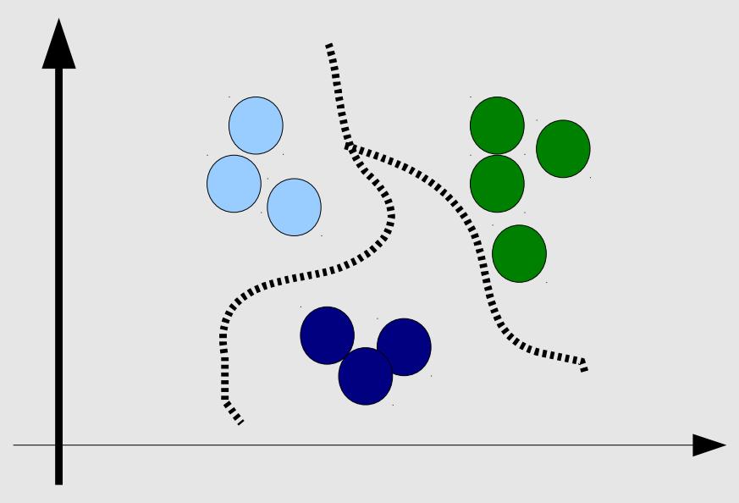
Use province to replicate sub-sets of data

Data Access prediction

Using data provenance identify and analyze data access patterns.

Get data inner-relationships.

Use metadata information to cluster data objects.



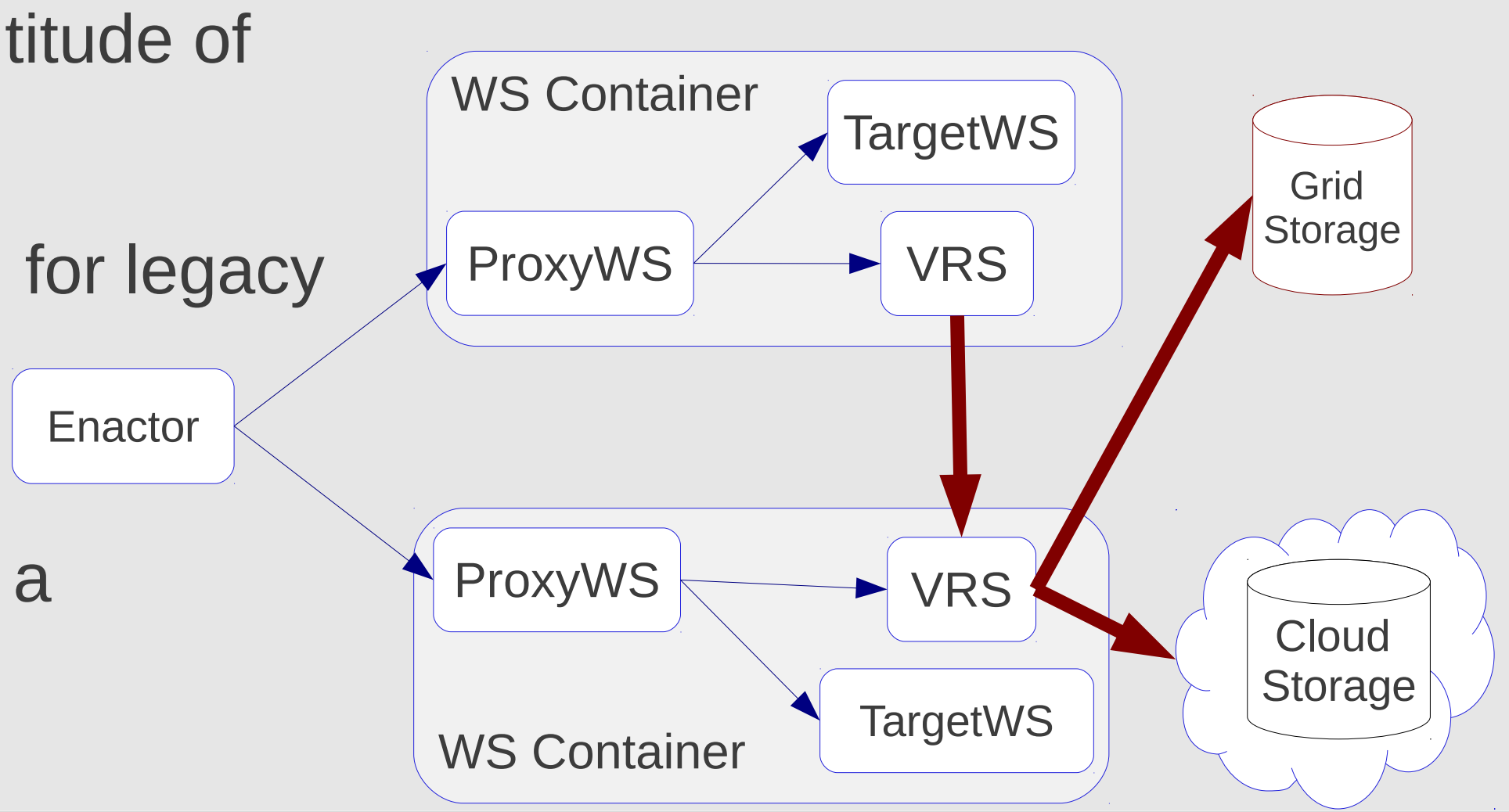
Data Streaming for Web Services

For data intensive WS we introduce streaming and data transfer proxy.

The **ProxyWS** utilizes a multitude of protocols.

It undertakes data transfers, for legacy Wss.

It can be used as an API for a direct **data streaming**.

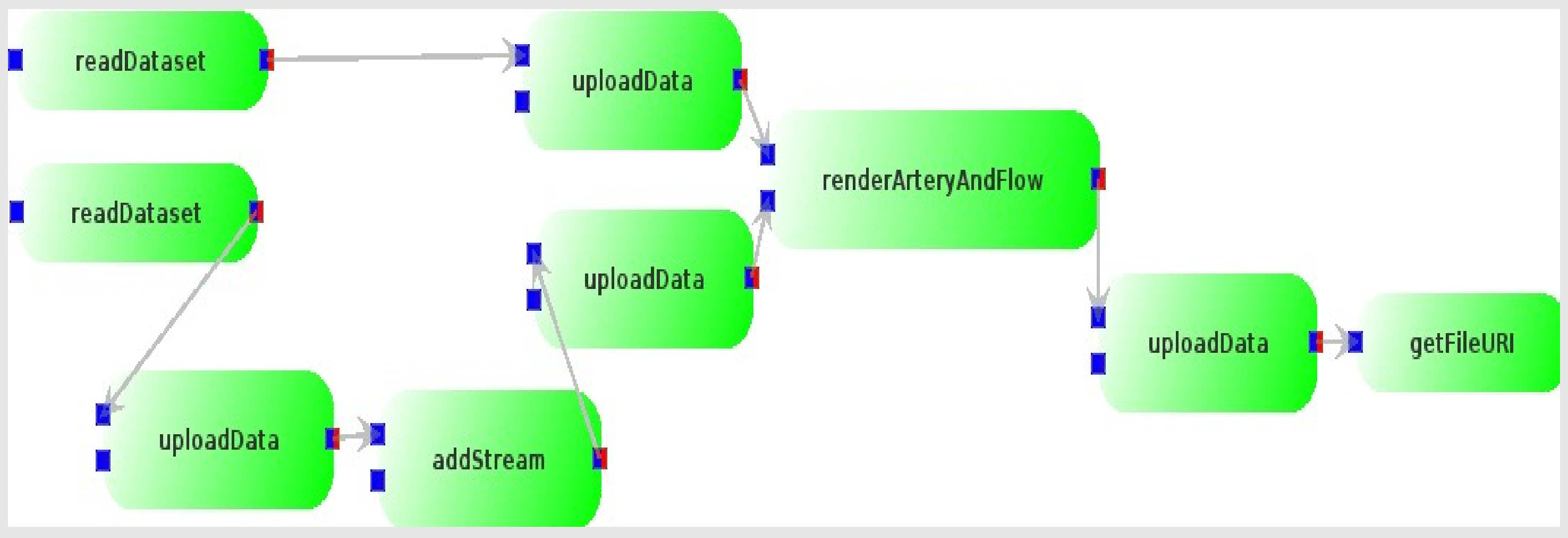


Results

Visualization Web Services for Medical Image Analysis[2]

Numerical simulation of the blood flow helps to obtain knowledge about its behavior and to develop treatments for vascular disorders.

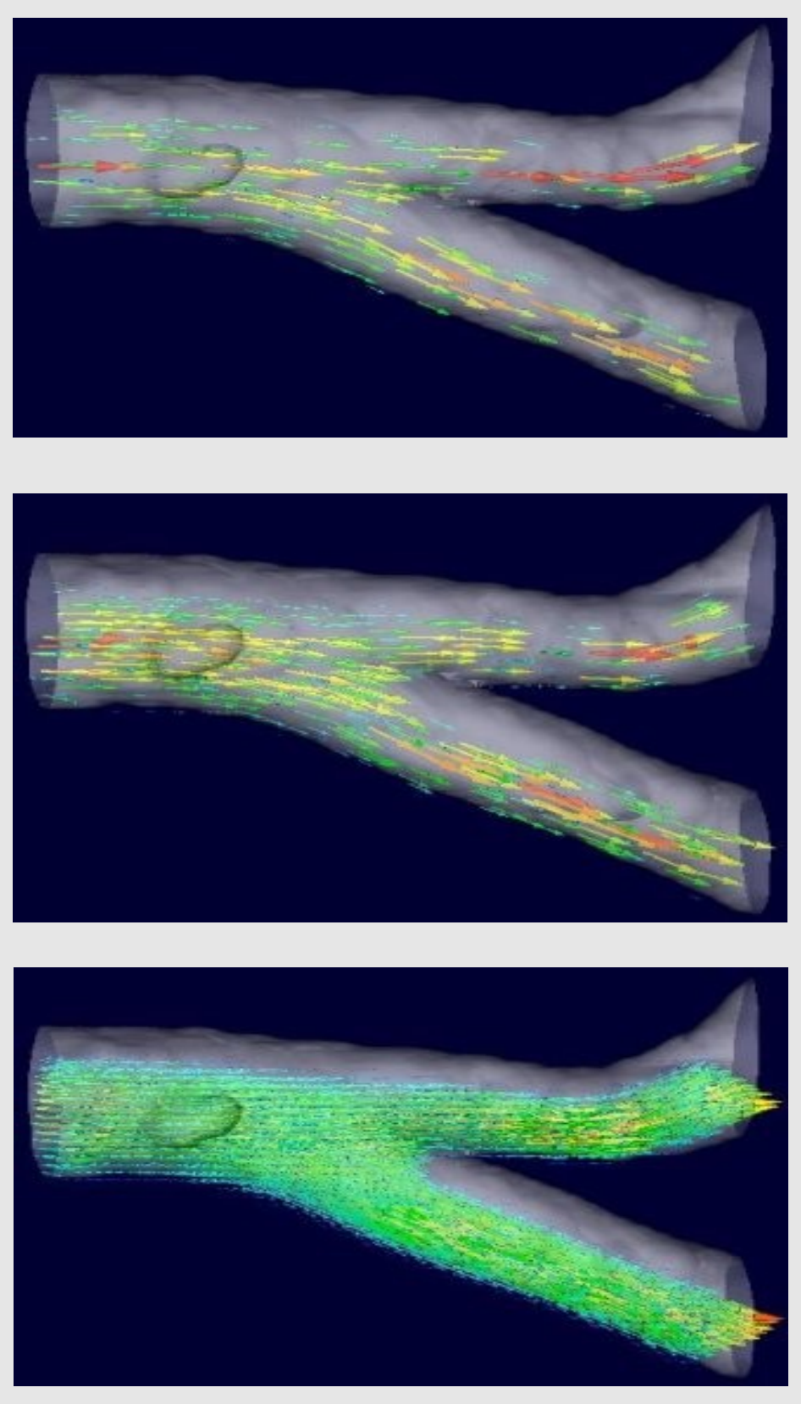
Data transfers tend to **pose serious bottlenecks in executing visualization workflows**.

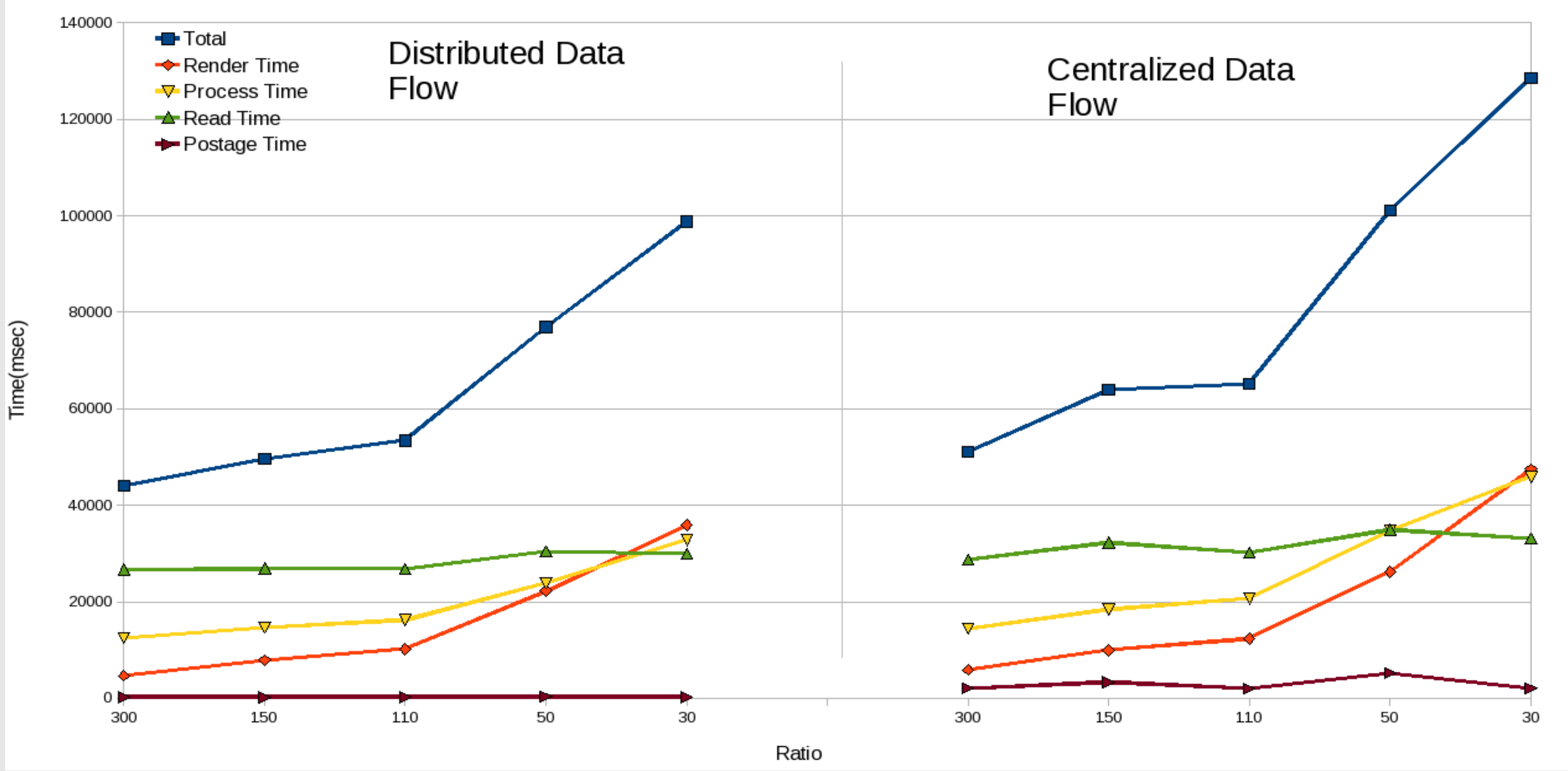


A WS Visulisation workflow

Two transport models for data-intensive medical visualizations that rely on web services:

- •direct streaming
- •loading data from a file server

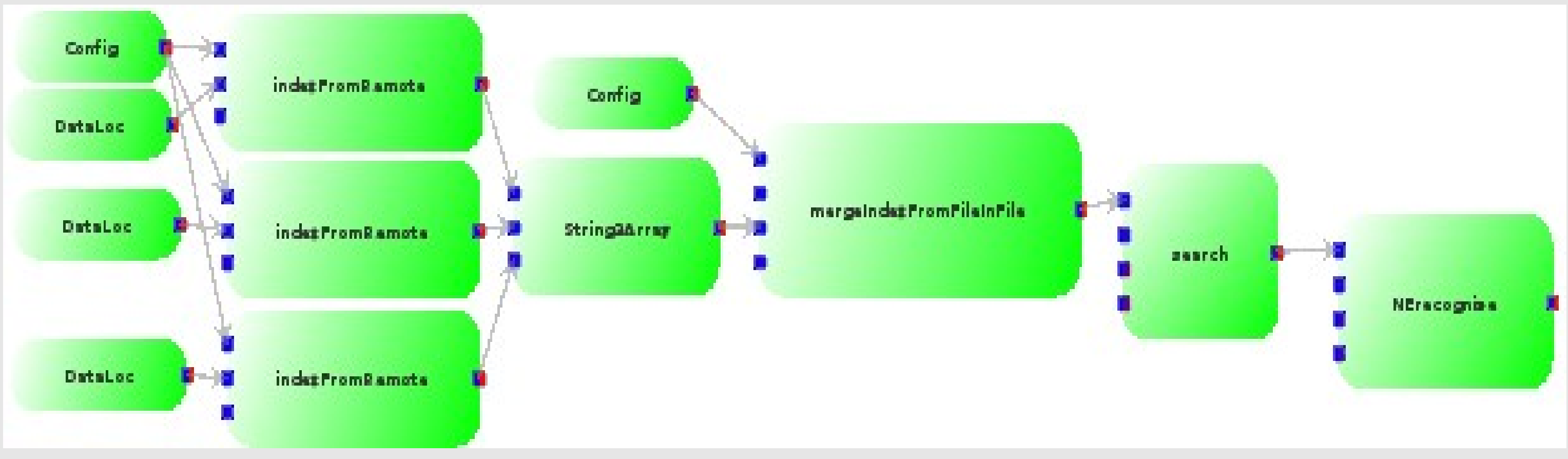




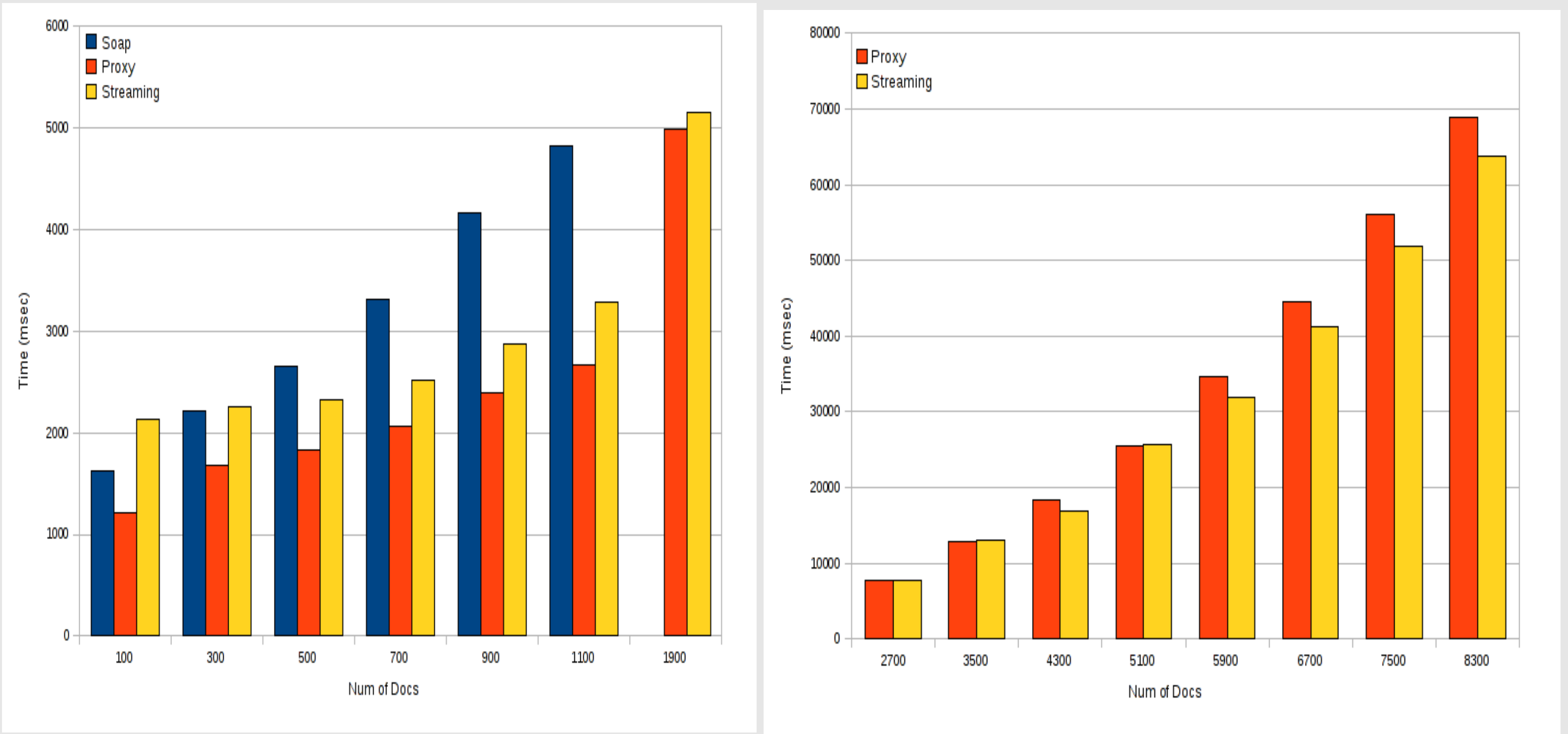
Breakdown of execution time while visualizing the 66.7 MB data set.

Indexing Web Services for Information Retrieval[1]

Indexing and Named Entity Recognition (NER) are tools that help biologists to identify and retrieve information. Indexing and recognizing units (NER) from a set of documents is a **data-incentive** procedure. Data transfers contribute to the total execution time.



Index Search & NER workflow.



Execution times for the Search and NER step.

Publications

[1] S. Koulouzis; R Cushing; K. Karasavvas; A.S.Z. Belloum; M.T.Bubak;, “Enabling web services to consume and produce large distributed data sets”, Submitted to IEEE Internet Computing, Internet-Scale Data Management

[2] S. Koulouzis; E.Z. Seinstra; A.S.Z. Belloum;, “Data transport between visualization web services for medical image analysis”, Procedia Computer Science, Volume 1, Issue 1, ICCS 2010, May 2010, Pages 1721-1730, ISSN 1877-0509

[2] Koulouzis S., Meij E J., Belloum A., "Enabling Large Data Transfers Between Web Services", 5th EGEE User Forum, April, 2010.

[3] Koulouzis, S., Meij, E.;Marshall, M.S.; Belloum, A.; , "Enabling Data Transport between Web Services through alternative protocols and Streaming" eScience, 2008. eScience '08. IEEE Fourth International Conference on , pp.400-401, 7-12Dec.2008 doi:10.1109/eScience.2008.127

Conclusions

To enable todays research, we should master the large amounts of data produced.

It can be achieved with:

- The right approaches and architectures
- Scaling complex, data-intensive applications
- Combing information from existing scientific knowledge generated by different researchers in different locations
- Identifying patterns and relationships in data usage, to make them available more efficient.

CAPACITIES

VPH-Share

gSLM

Contact:
S.Koulouzis@uva.nl