Scalable Workflow

Adam Barker
University of St Andrews
www.adambarker.org

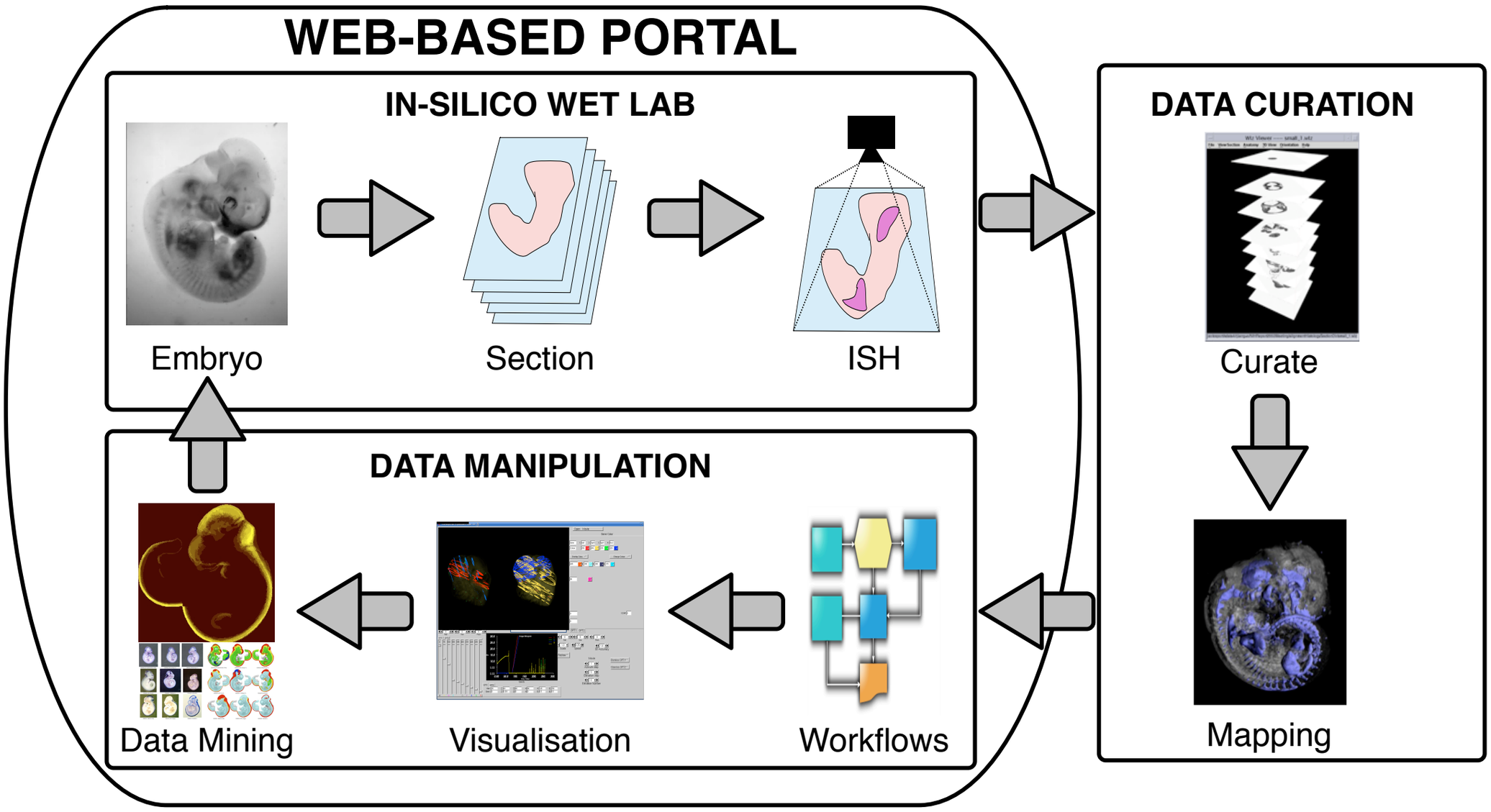
Overview

- Big Data Lab
- Motivation
- Service Oriented Architectures
- Workflow orchestration using Taverna
- WS proxy architecture
- Performance analysis
- Future work and collaboration

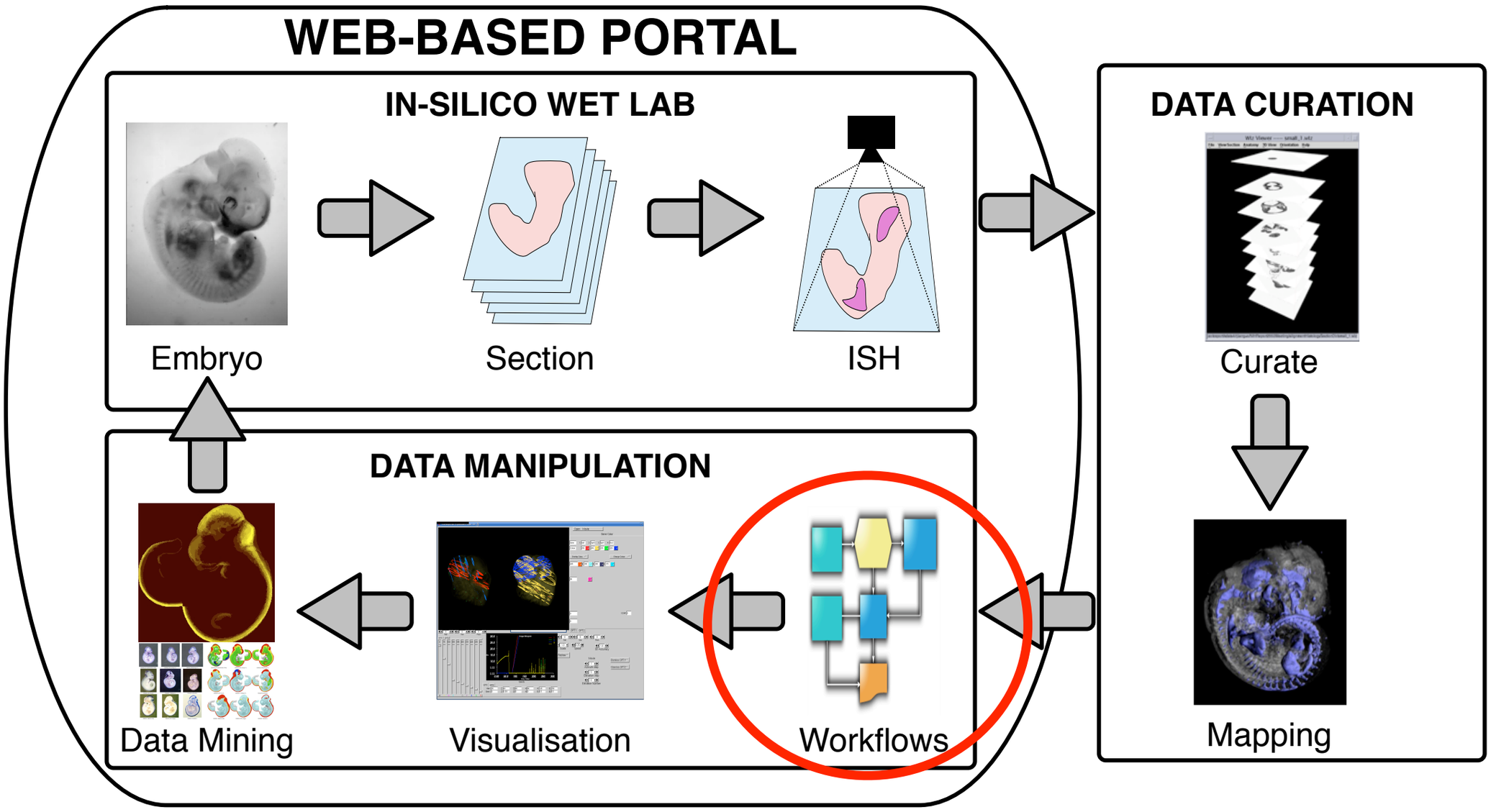
Big Data Lab

choreography
systems
cloud-computing
mobile-code
workflow
semantic
St-Andrews
scalable
analytics
real-time
agile
big-data
on-demand
autonomic
standups
SOA
orchestration
multi-disciplinary
algorithmic-trading
peer-to-peer
data-intensive
distributed
large-scale
monitoring
collaborate
economic
disrupt
decentralised
noSQL
innovate

DGEMap: Gene Expression patterns in early human development



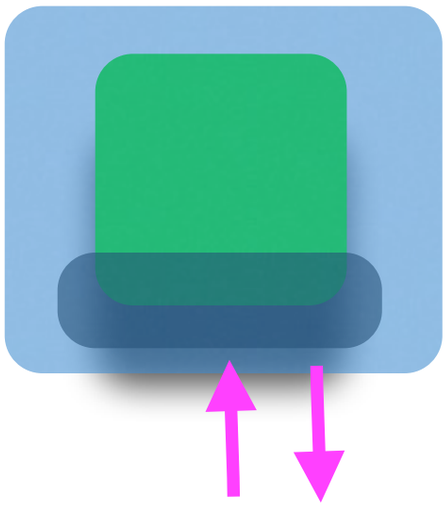
DGEMap: Gene Expression patterns in early human development



Large image data

Web Services

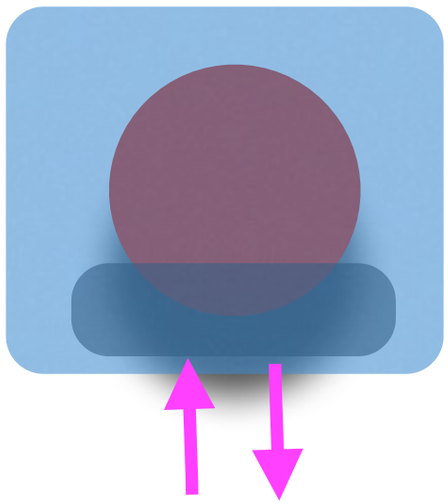
WSDL
Interface



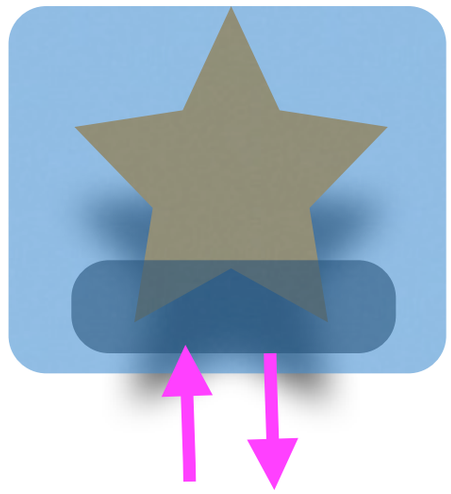
SOAP
Message
Processor

SOAP/XML

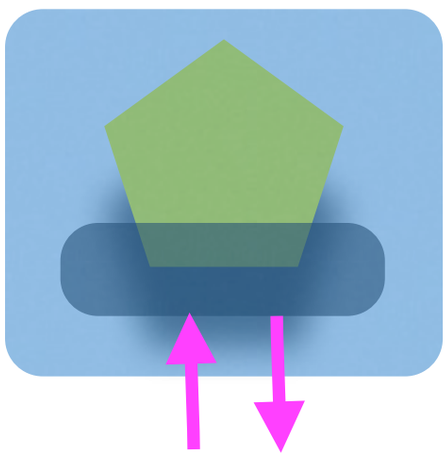
Application
Logic



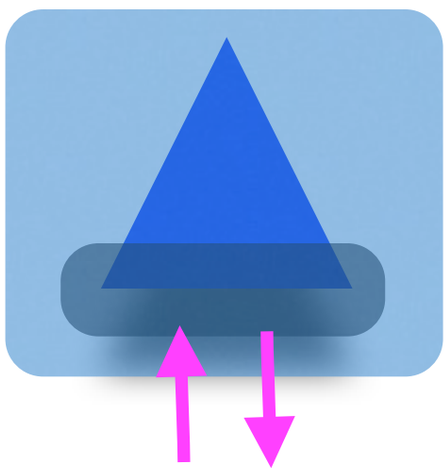
SOAP/XML



SOAP/XML

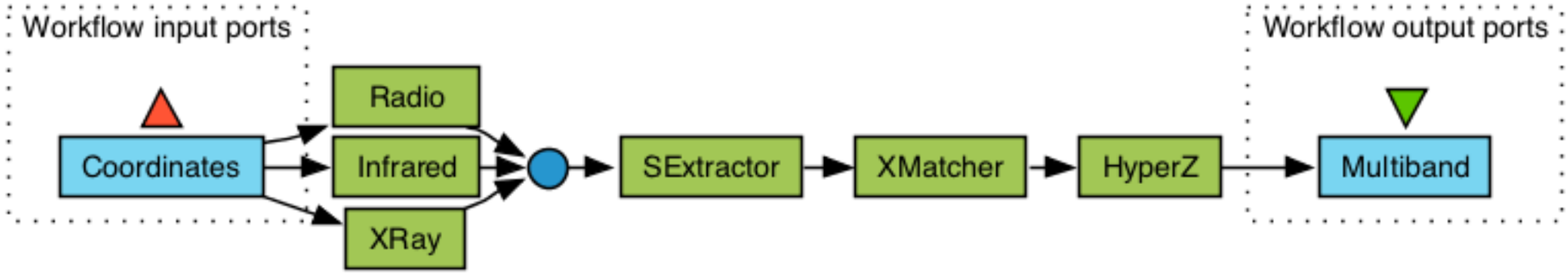


SOAP/XML

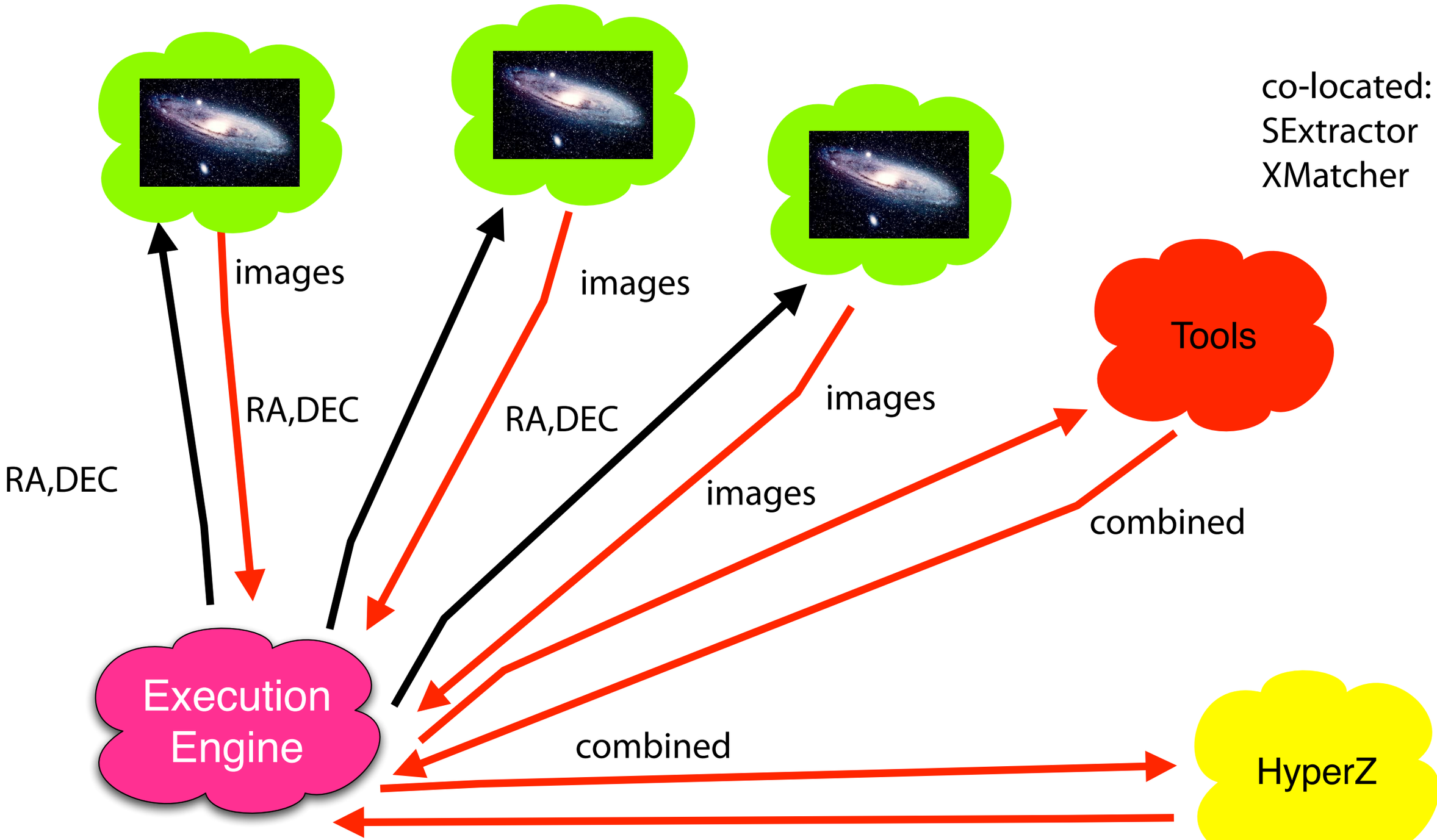


SOAP/XML

Here's a Taverna Workflow



Orchestration



UNDERGROUND STATION
VICTORIA
Underground

Bottleneck

- Although service-oriented workflows can be composed as DAGs using tools such as Taverna
  - In reality they are usually orchestrated from a single workflow engine
- Intermediate data (e.g., large images) are routed through a single centralised engine
  - Routing intermediate data through a single engine creates a bottleneck
  - Decreases the performance of a workflow

Hybrid Architecture

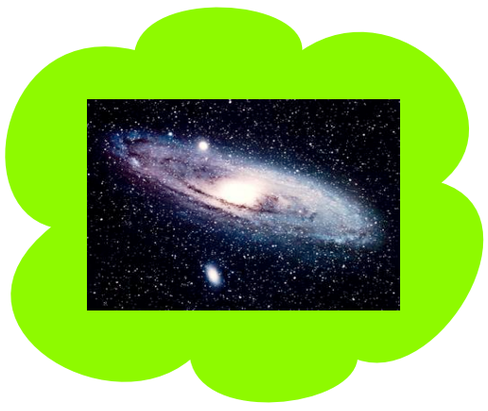
- Semi-decentralised execution of service-oriented workflows
  - Maintains the simplicity of centralised orchestration
  - Benefits from distributed data flow
- Reduces intermediate data transfer
- Reduces expensive workflow engine to services link
- Speedup the execution time of a workflow

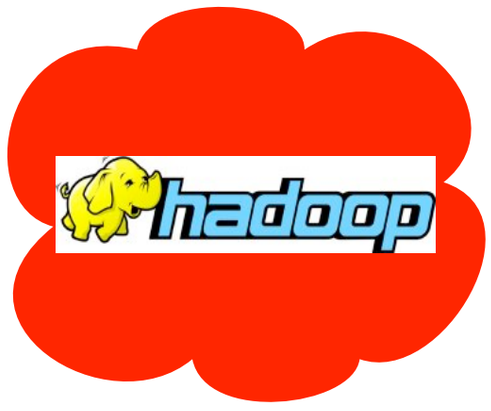
Assumption I

Assumption 2

Assumption 3

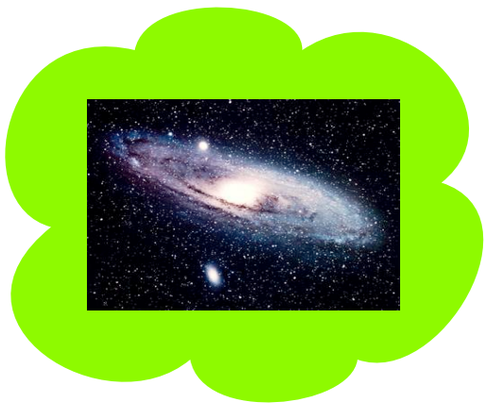
WS-Proxy

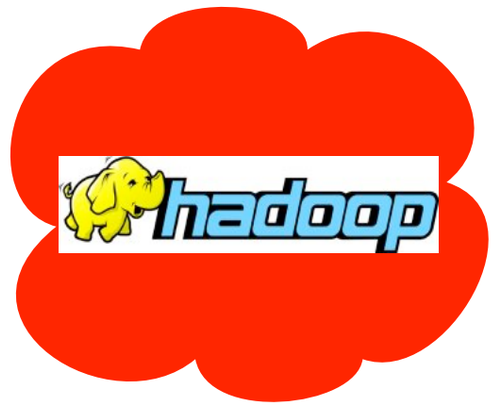




Execution Engine

WS-Proxy

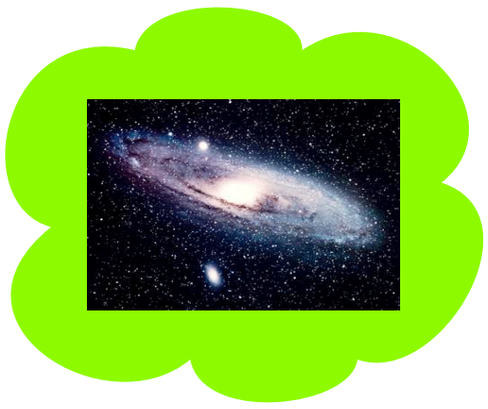


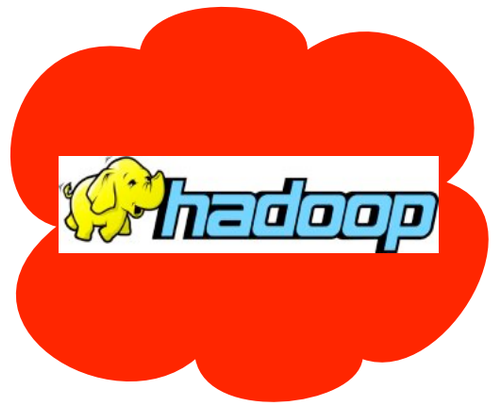


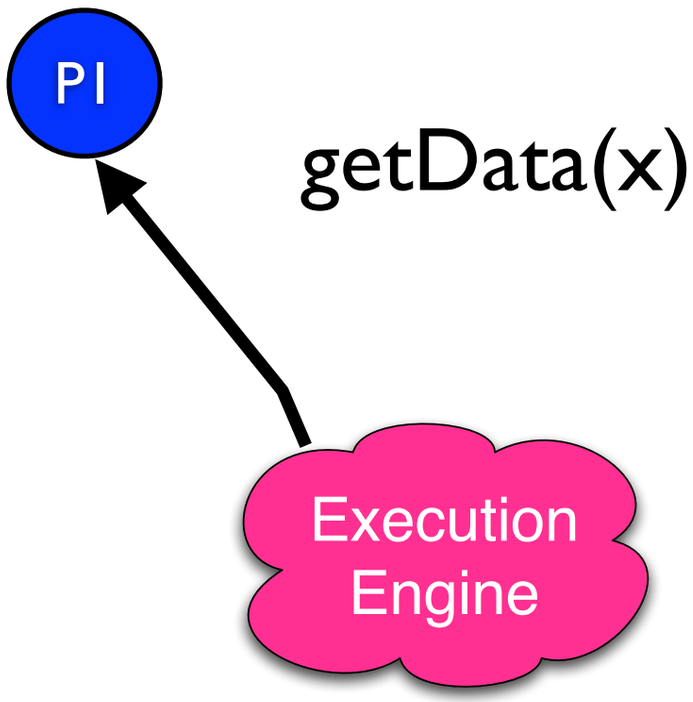
PI

Execution Engine

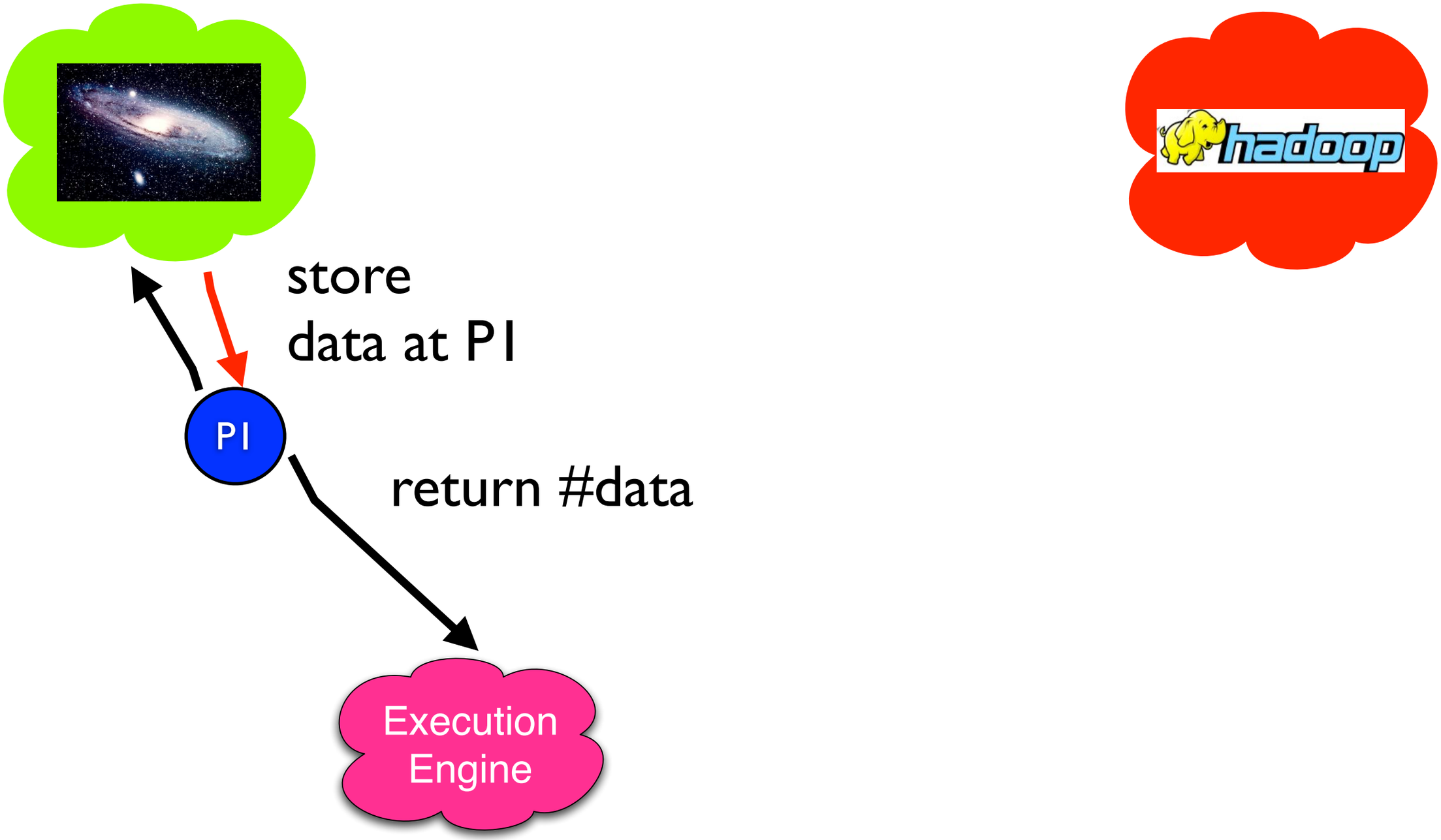
WS-Proxy



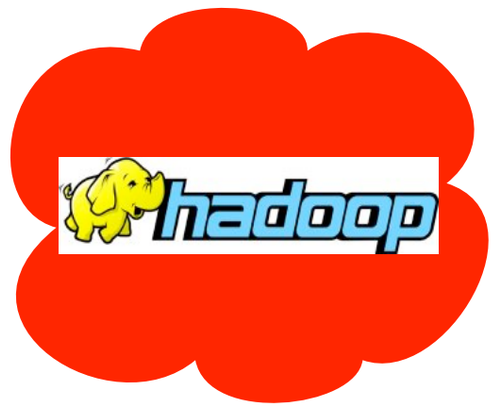


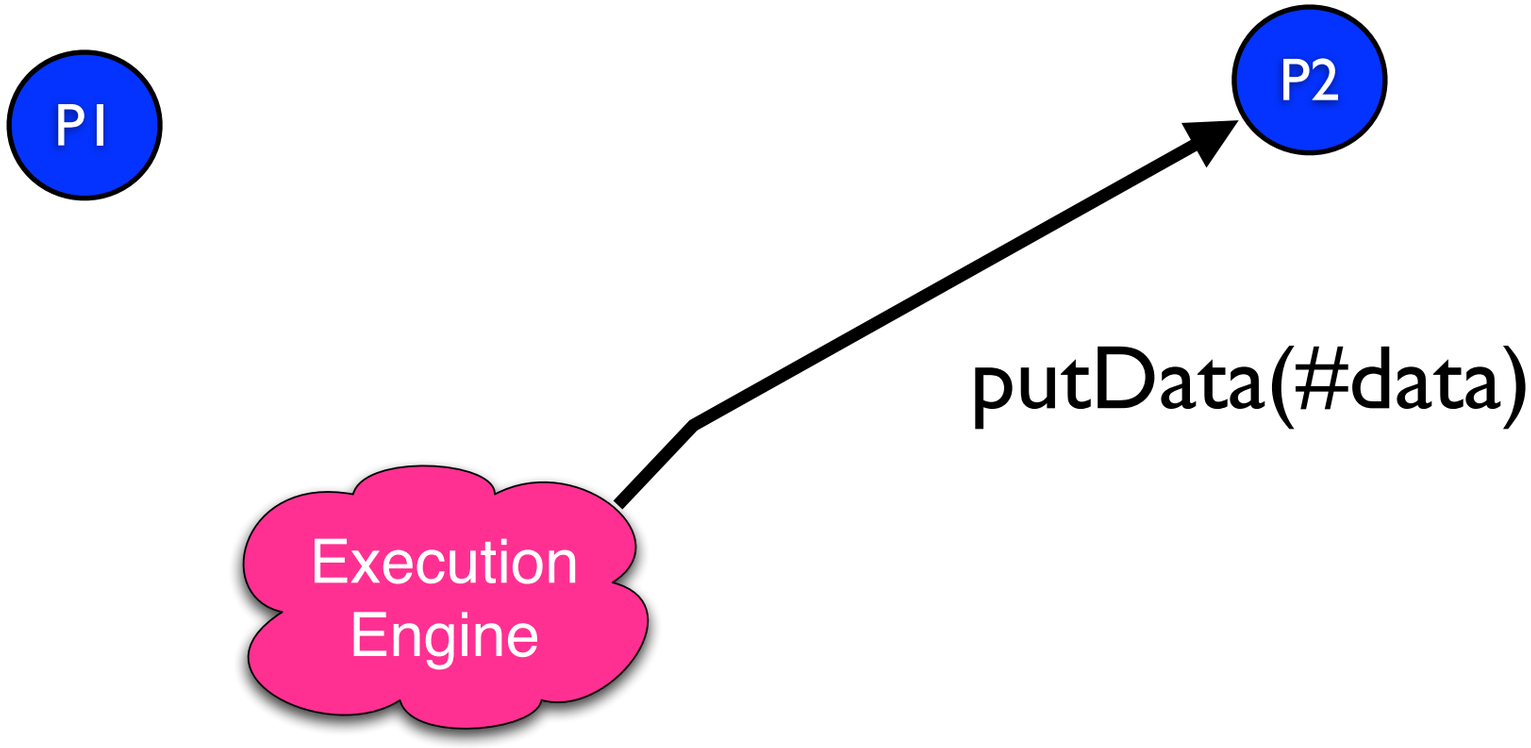


WS-Proxy

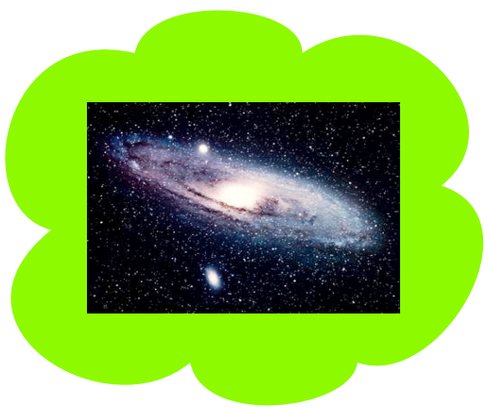


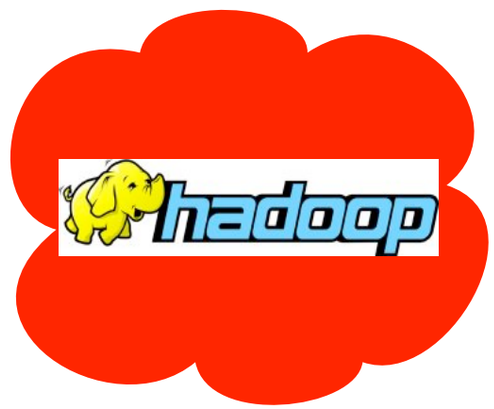
WS-Proxy

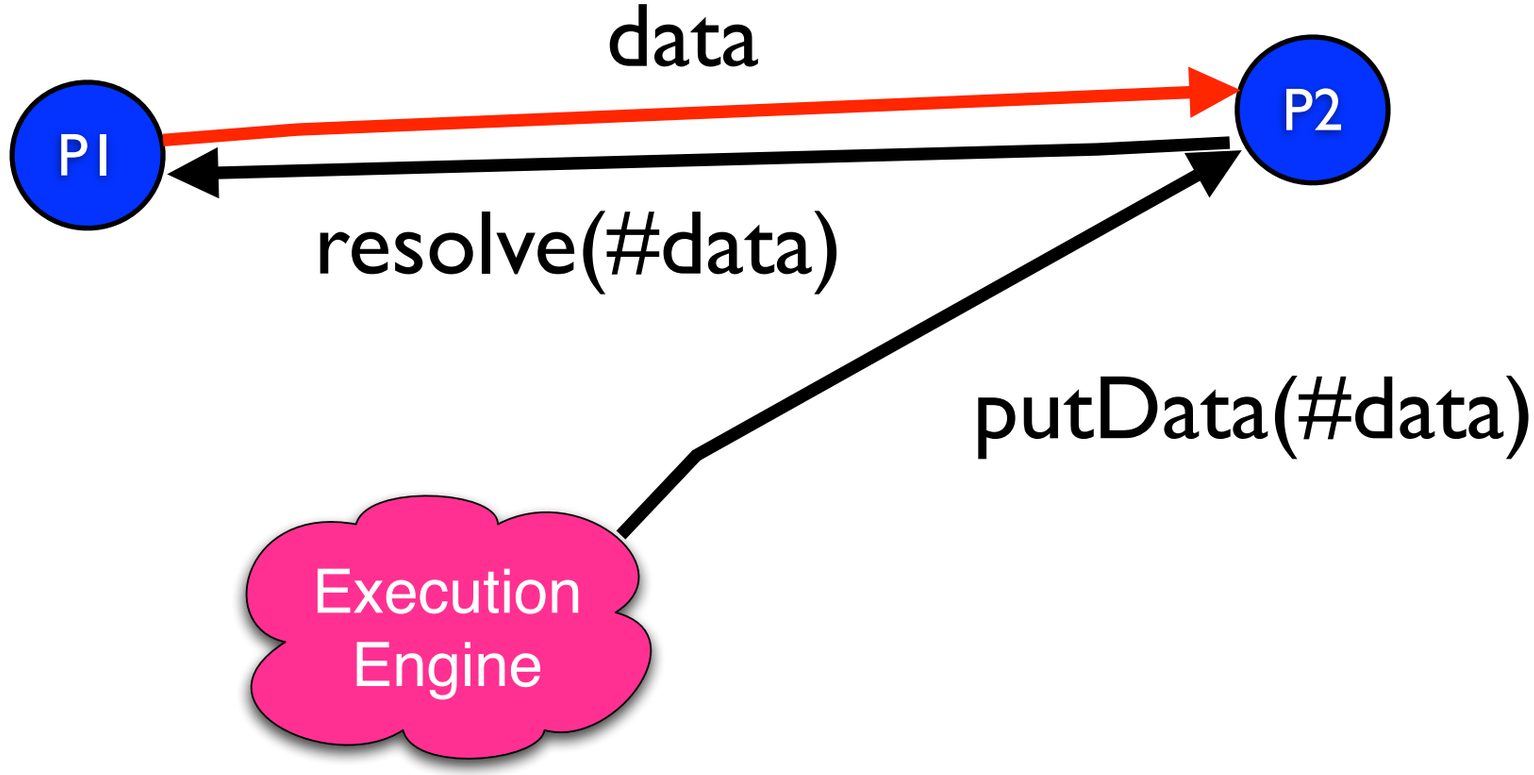




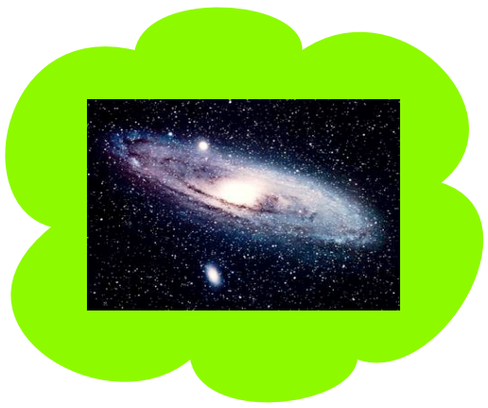
WS-Proxy

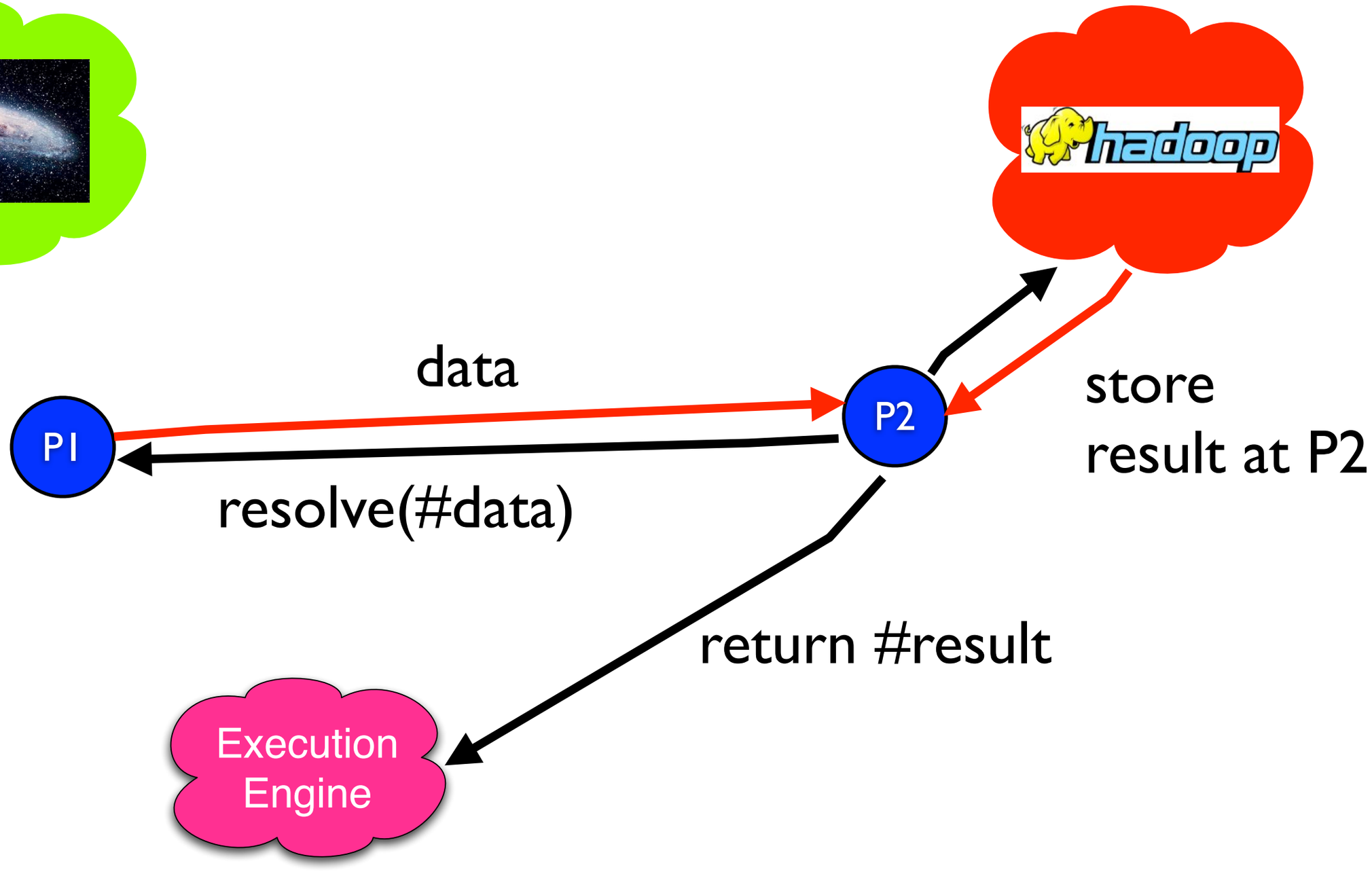



hadoop


data
P1
P2
resolve(#data)
putData(#data)
Execution
Engine

WS-Proxy



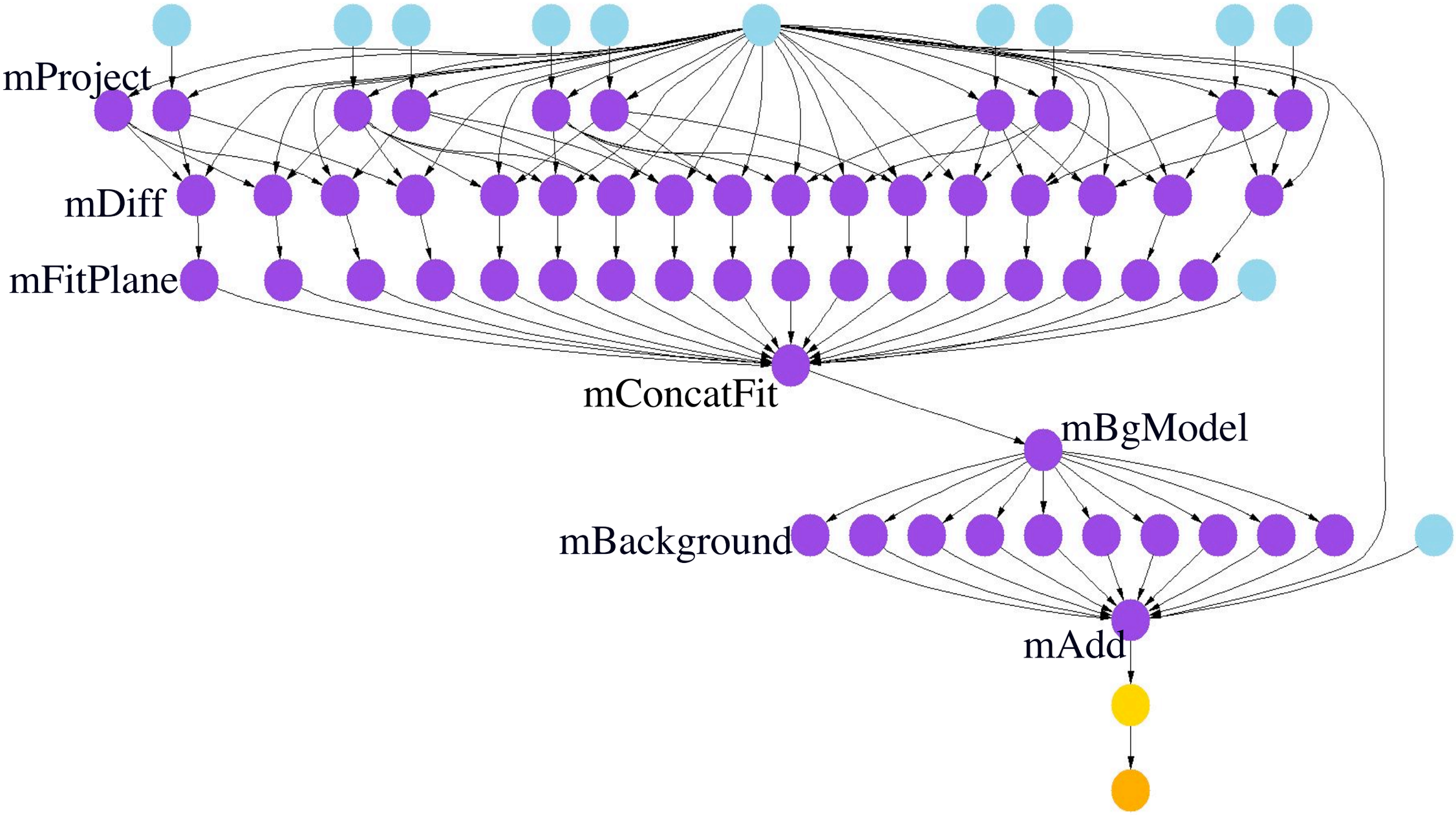

hadoop
data
P1
P2
store
result at P2
resolve(#data)
return #result
Execution
Engine

WS-Proxy

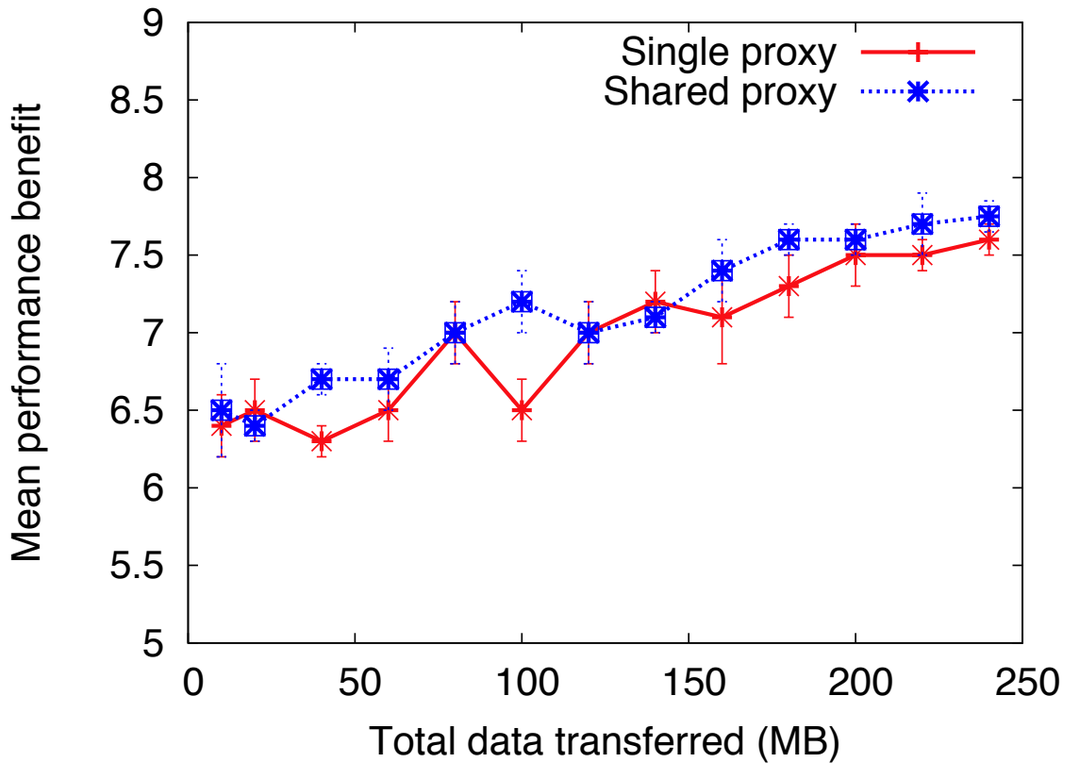
- Mirrors the same interface as the WS it is managing
- Workflow engine sees no difference
  - End points change (proxy), reference type (input, output)
- Proxies store intermediate data and talk to one another
  - Data are globally identifiable
- Proxy responsible for 1...N services
- Assume that proxy can be deployed as closely (n/w distance) as possible to back-end services
- Simple to install and configure: WAR file in /webapps

amazon
web services™

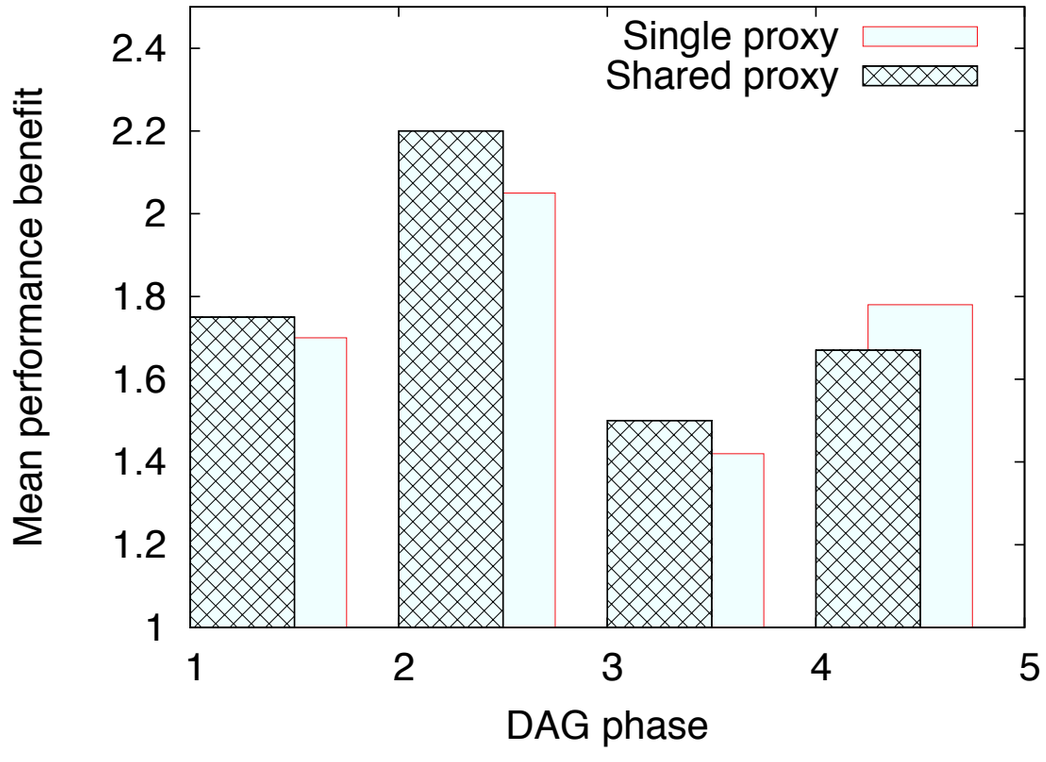
Performance Analysis



Performance Analysis



(a) Montage end-to-end performance



(b) Montage DAG phase performance

Single = 1 service per proxy
Shared = 4 services per proxy

PLANETLAB

amazon web services™

Questions...?

- Adam Barker, Jon B. Weissman and Jano I. van Hemert. Reducing Data Transfer in Service-Oriented Architectures: The Circulate Approach. To appear in the IEEE Transactions on Services Computing, 2013.

- Adam Barker, Christopher D. Walton and David Robertson. Choreographing Web Services. IEEE Transactions on Services Computing, volume 2, number 2, pages 152-166, IEEE Computer Society, April-June 2009.

- Adam Barker, Jon B. Weissman and Jano van Hemert. Eliminating the Middle Man: Peer-to-Peer Dataflow. In HPDC'08: Proceedings of the 17th International Symposium on High Performance Distributed Computing, pages 55-64. ACM, June 2008.