Cloud Data Management for Scientific Workflows

Peter Reimann, Tim Waizenegger,
Matthias Wieland, **Holger Schwarz**, Bernhard Mitschang

University of Stuttgart

Institute of Parallel and Distributed Systems (IPVS)

Cluster of Excellence Simulation Technology

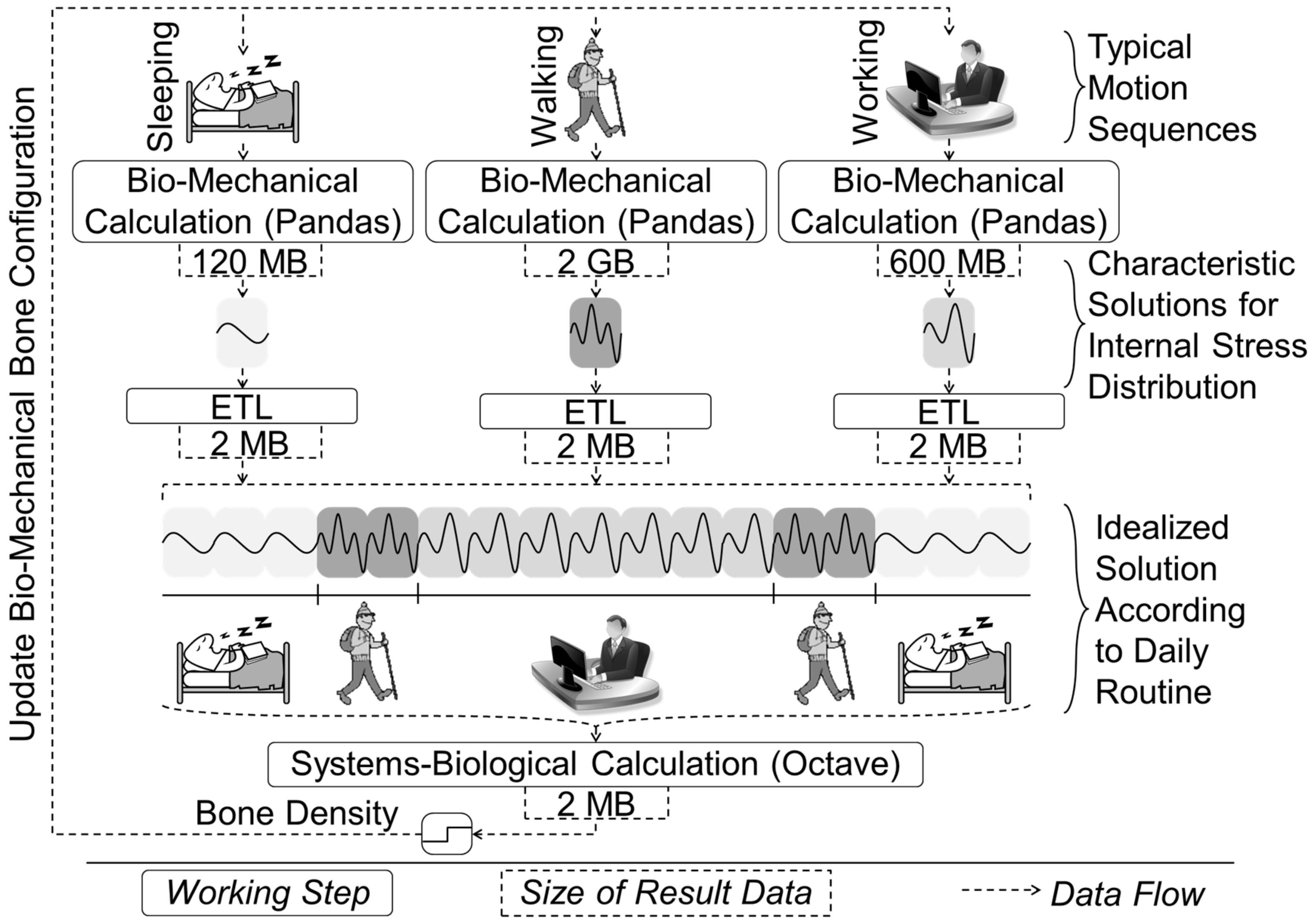
Outline

- Simulation Workflows
- Example: Simulation of Structure Changes in Bones
- Simulation software in the cloud
- Complex data provisioning in simulation workflows
- SIMPL framework
- Traceability and Reproducibility
- Conclusion

Characteristics of Simulation Workflows

- Scientific workflows
  - increasingly adopted to enable the implementation of scientific applications across various domains
  - experiments, data analyses, or computer-based simulations
- Simulation workflows
  - typically compositions of long-running numeric calculations
  - realize mathematical simulation models, e. g., based on partial differential equations
  - Coupled simulations combine various simulation tools
- Data management
  - proprietary data formats of simulation tools
  - complex data transformations
- Example:
  - Simulation of structure changes in bones

Simulation of Structure Changes in Bones


Update Bio-Mechanical Bone Configuration
Sleeping
Walking
Working
Typical Motion Sequences
Bio-Mechanical Calculation (Pandas)
Bio-Mechanical Calculation (Pandas)
Bio-Mechanical Calculation (Pandas)
120 MB
2 GB
600 MB
Characteristic Solutions for Internal Stress Distribution
ETL
ETL
ETL
2 MB
2 MB
2 MB
Idealized Solution According to Daily Routine
Systems-Biological Calculation (Octave)
2 MB
Bone Density
Working Step
Size of Result Data
Data Flow

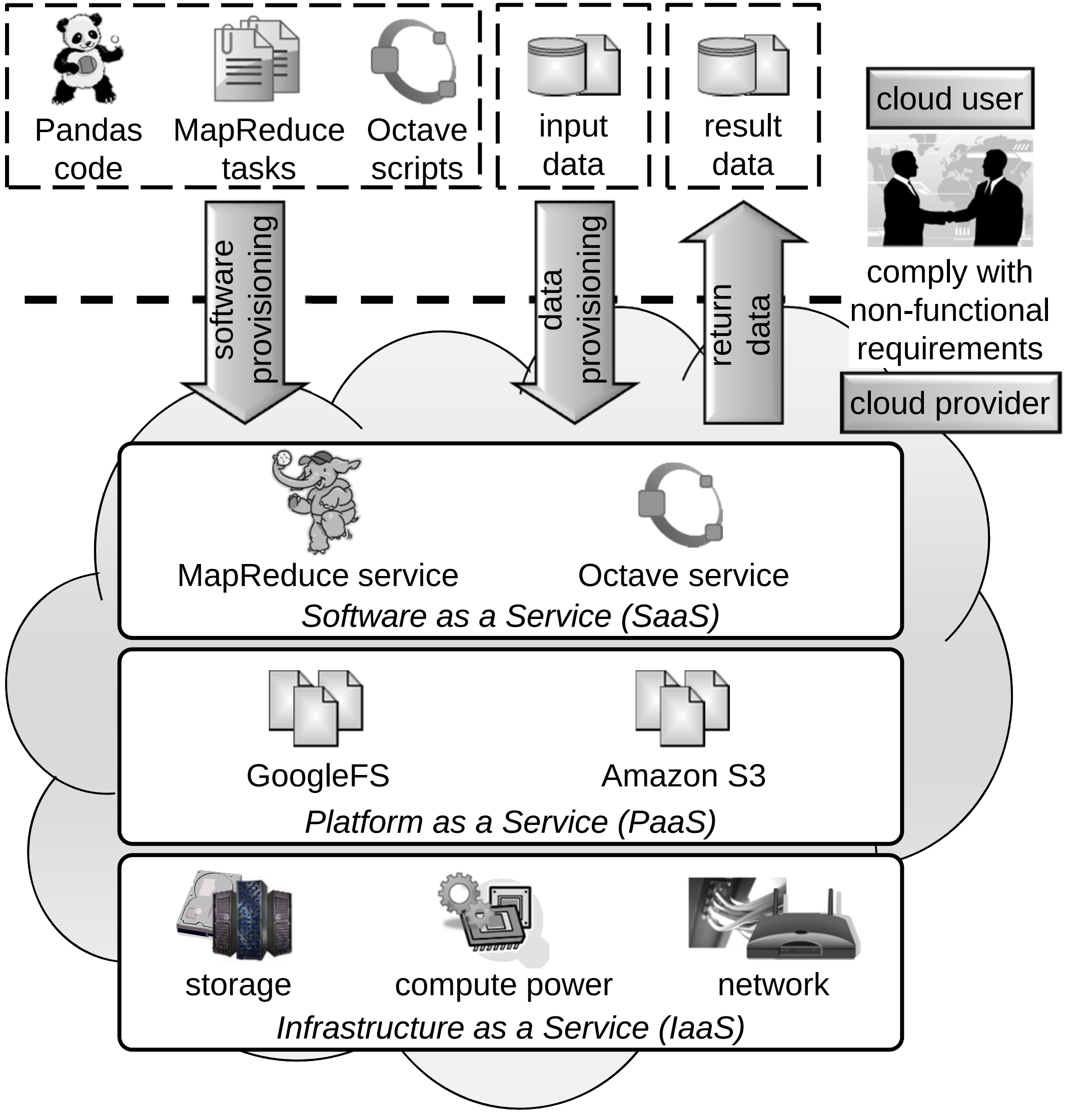
Infrastructure for Simulation Workflows

- large or specialized organizations may afford
  - high performance computing centers
  - specialized grid infrastructures
  - own or rent infrastructure
- small or medium organizations
  - run simulations more sporadically
  - high cost to own or rent infrastructure
  - high effort to provide and integrate necessary software

➡ Deploy simulation workflows in a public cloud

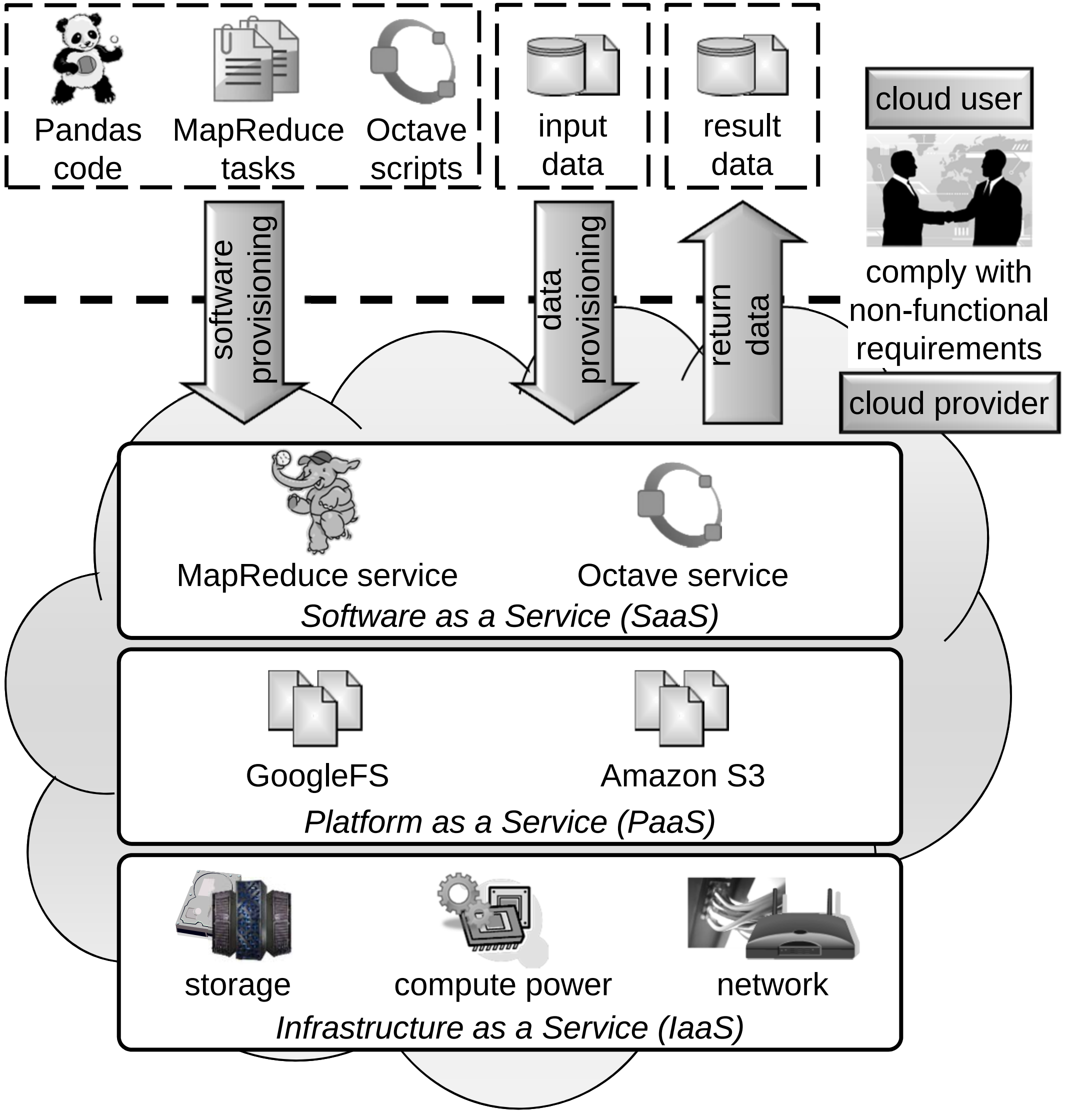
Provisioning of Simulation Software

- GNU Octave: widespread tool
- provide as SaaS
- users need to provide Octave scripts
- users provide data in Amazon S3 for example

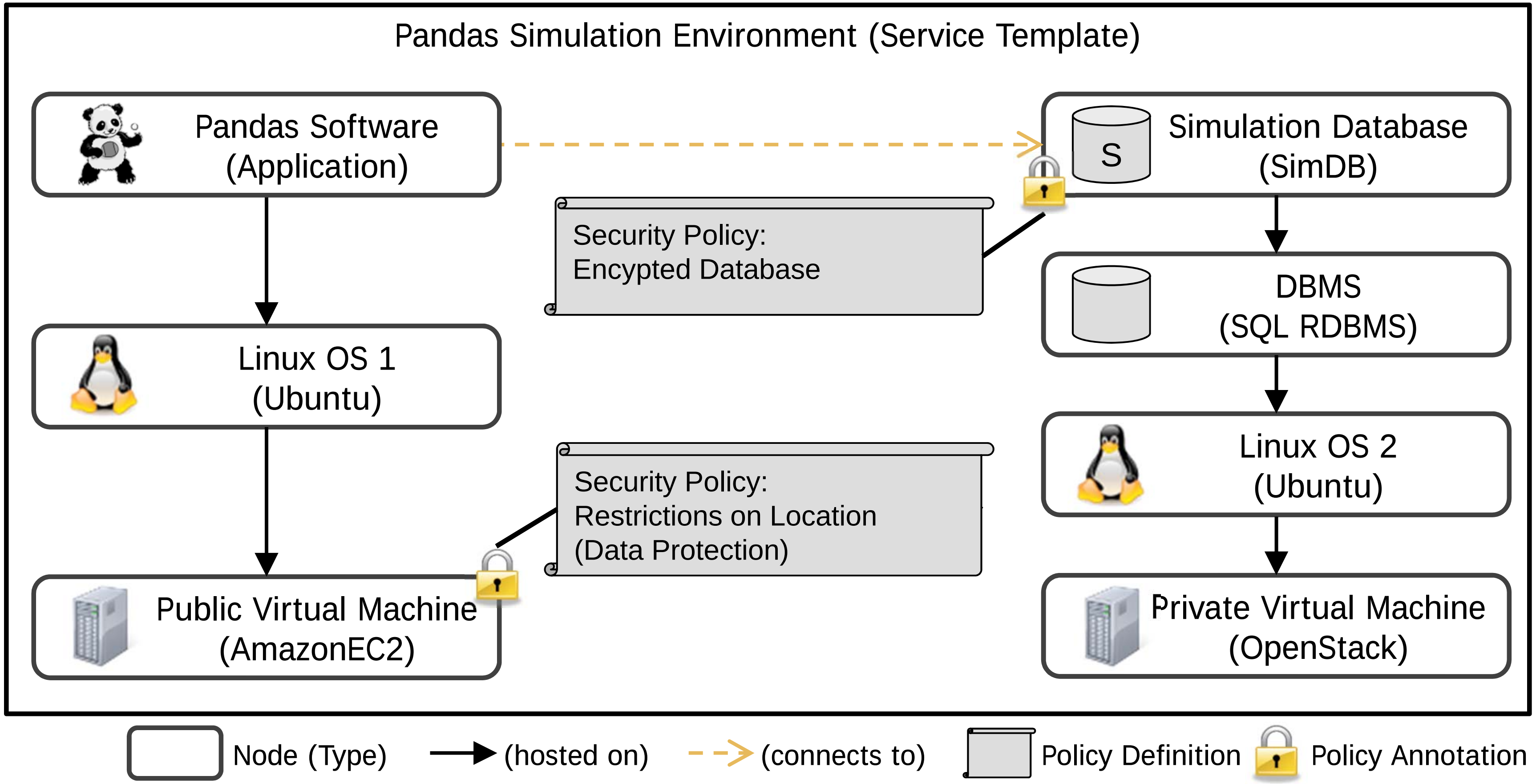


Provisioning of Simulation Software

- Pandas: Proprietary and highly specialized software
- deploy software in the cloud infrastructure



TOSCA Service Topology



Pandas Simulation Environment (Service Template)

Pandas Software (Application)

Linux OS 1 (Ubuntu)

Public Virtual Machine (AmazonEC2)

Simulation Database (SimDB)

DBMS (SQL RDBMS)

Linux OS 2 (Ubuntu)

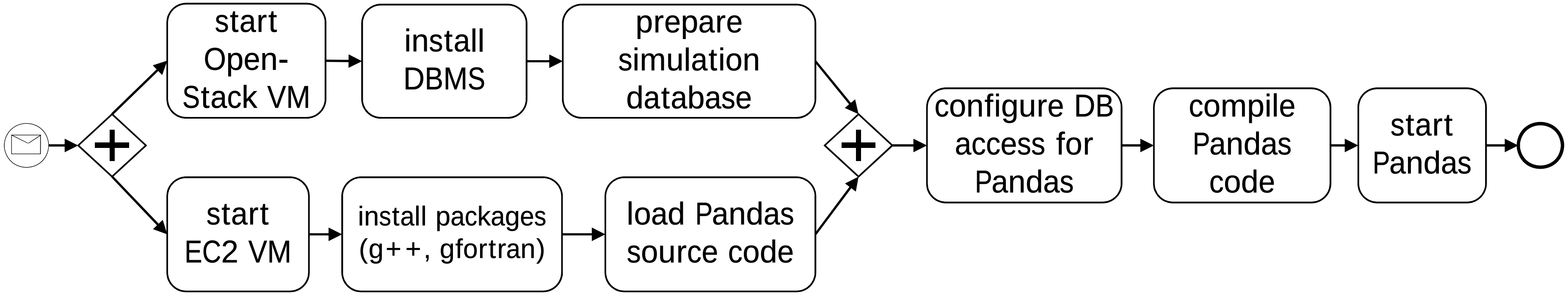
Private Virtual Machine (OpenStack)

Security Policy: Encypted Database

Security Policy: Restrictions on Location (Data Protection)

Node (Type) (hosted on) (connects to) Policy Definition Policy Annotation

TOSCA Deployment Plan

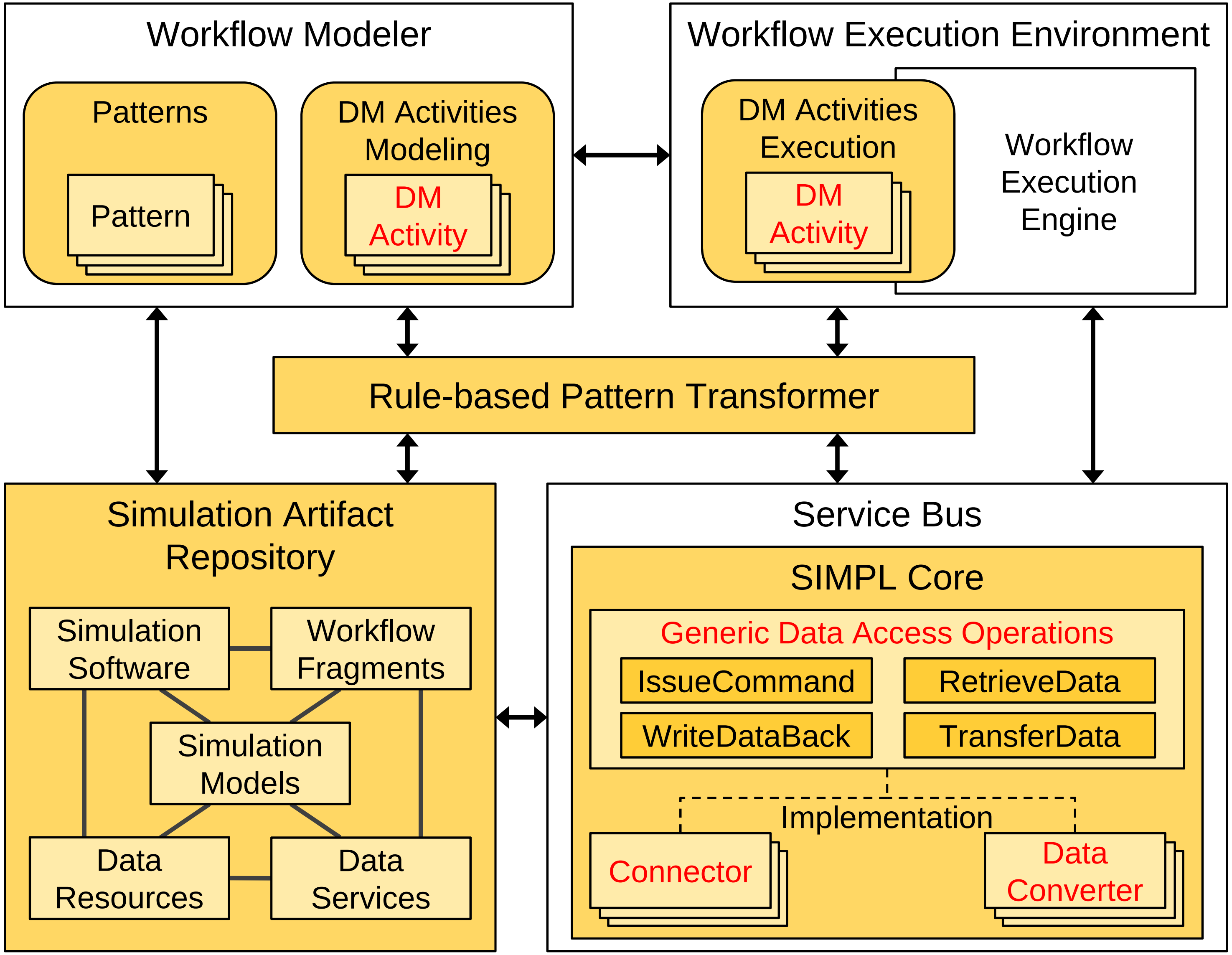


- Deployment plan: Install and configure artifacts corresponding to the topology

Data Provisioning

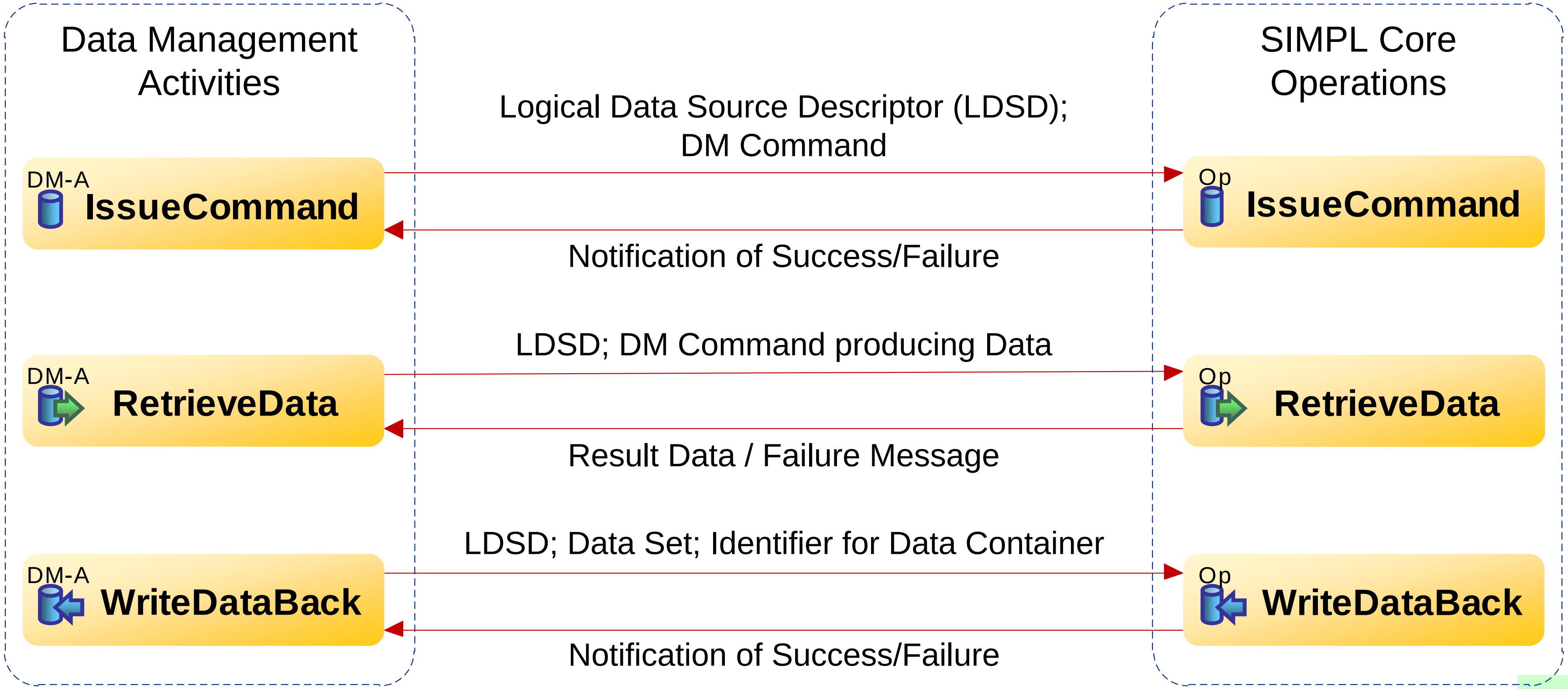
- Scientists define simulation workflows
  - limited skills in defining workflows
  - limited data management skills
  - ➡ need for abstract data management support
  - appropriate abstraction level for scientists
  - reduce need to specify technical low-level details of data management
- Coupled simulations
  - various proprietary simulation tools
  - proprietary services for handling input data and result data
  - heterogeneous data landscape
  - complex data management
  - ➡ need for generic data management
- SIMPL framework to provide abstract and generic data management for simulation workflows

SIMPL Framework



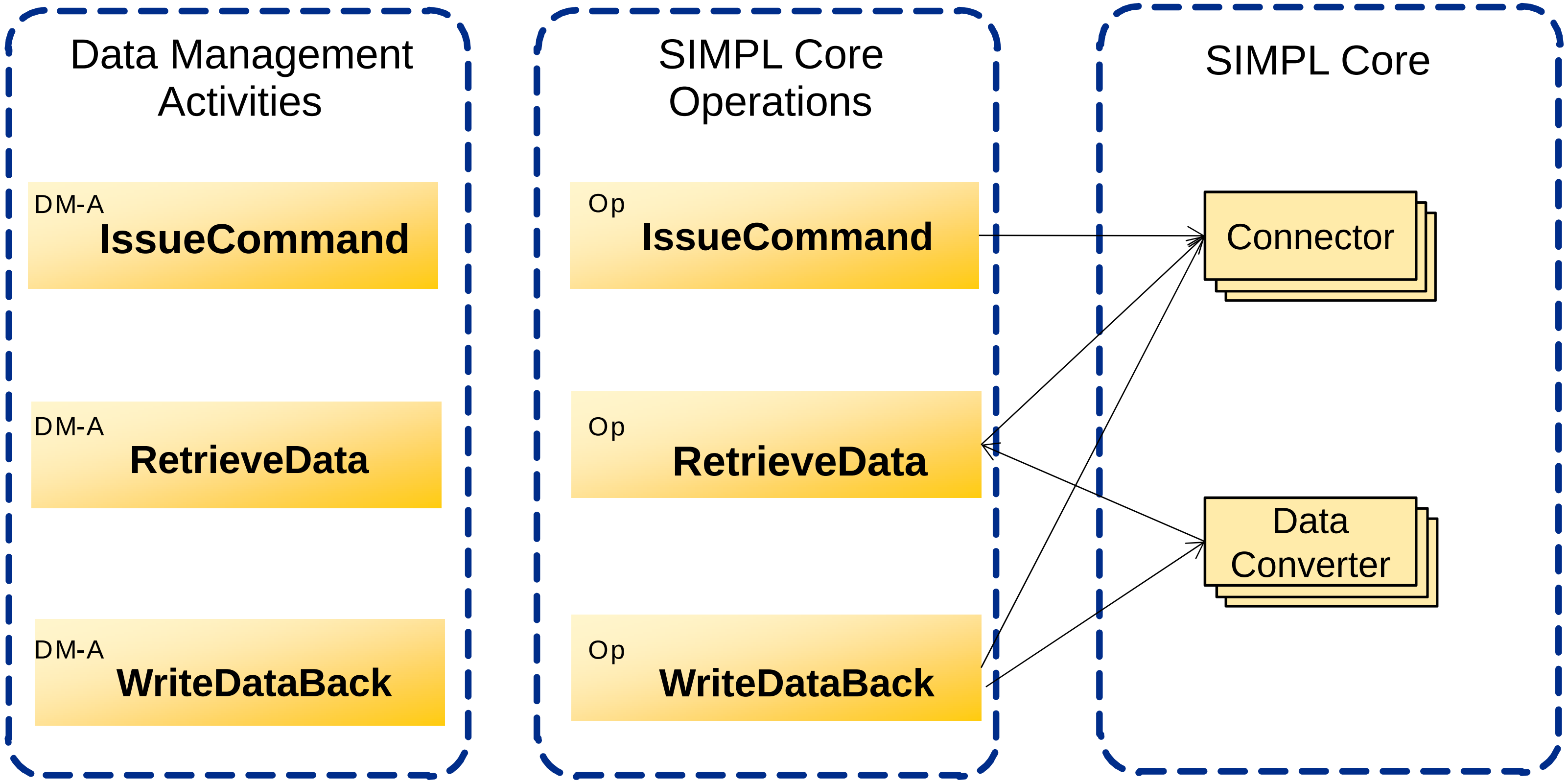
SIMPL DM Activities

- Data access abstraction by means of generic operations (BPEL extension)
  - IssueCommand, RetrieveData, WriteDataBack, TransferData
- Connectors and Data Converters of the SIMPL core implement these operations for specific data resources


Data Management Activities
SIMPL Core Operations
Logical Data Source Descriptor (LDSD); DM Command
DM-A IssueCommand
Op IssueCommand
Notification of Success/Failure
LDSD; DM Command producing Data
DM-A RetrieveData
Op RetrieveData
Result Data / Failure Message
LDSD; Data Set; Identifier for Data Container
DM-A WriteDataBack
Op WriteDataBack
Notification of Success/Failure

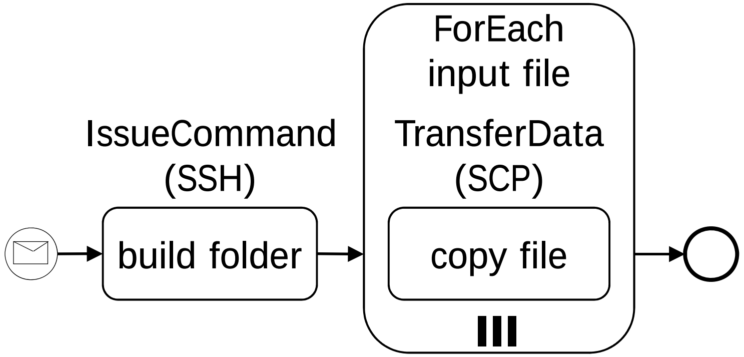
SIMPL DM Activities

- Data access abstraction by means of generic operations (BPEL extension)
  - IssueCommand, RetrieveData, WriteDataBack, TransferData
- Connectors and Data Converters of the SIMPL core implement these operations for specific data resources

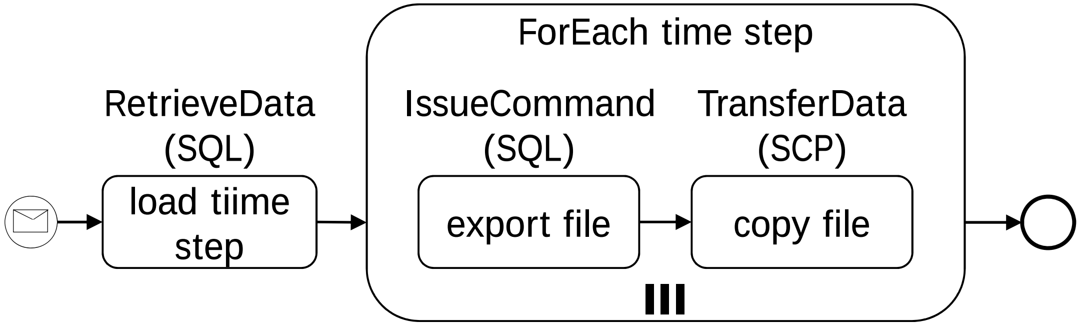


Management Plans Using SIMPL Activities

- Define data management
  as part of TOSCA management plans
- Use SIMPL DM activities
- Examples:
  Management plans for Pandas

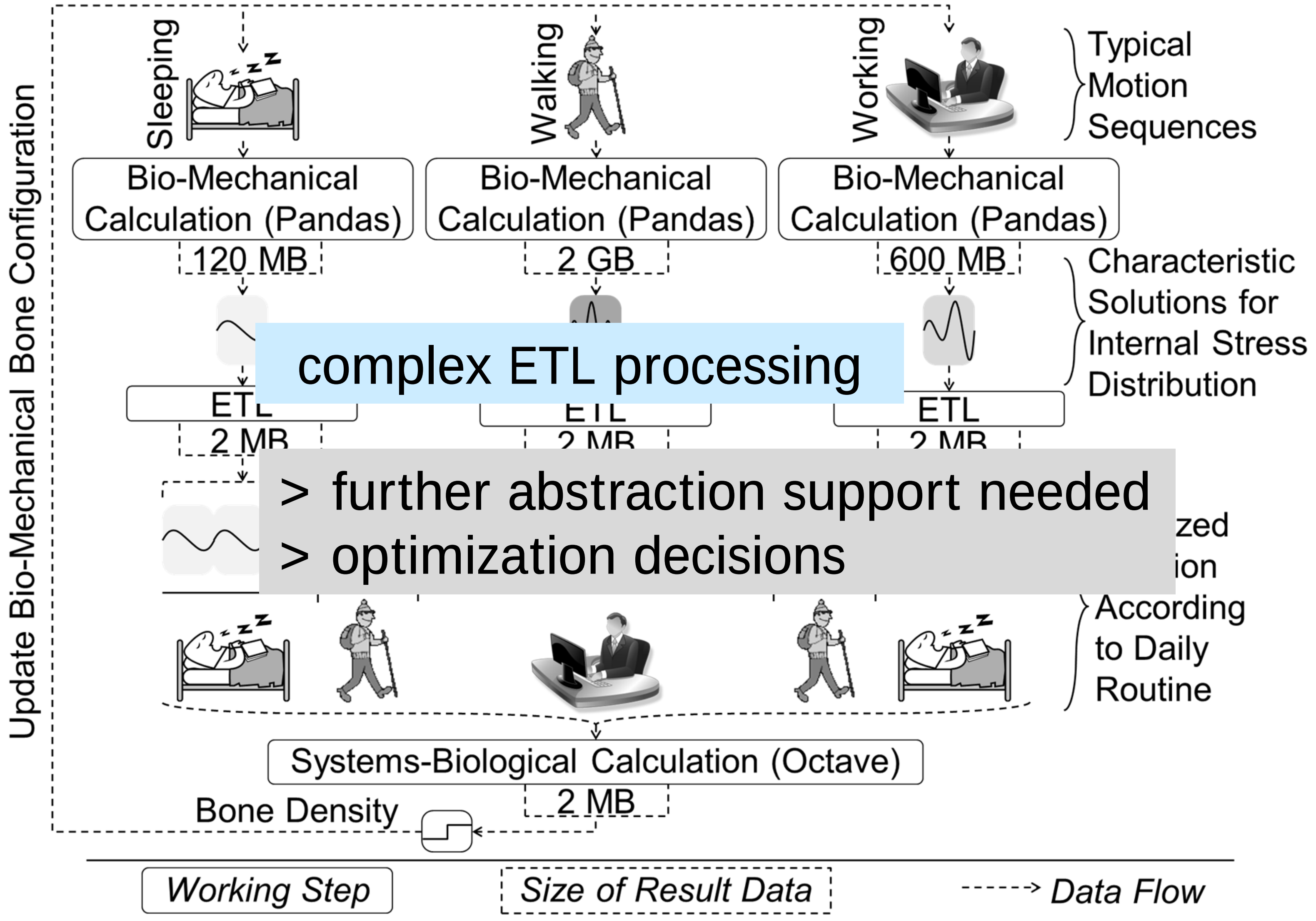


data provisioning



return result set

Simulation of Structure Changes in Bones

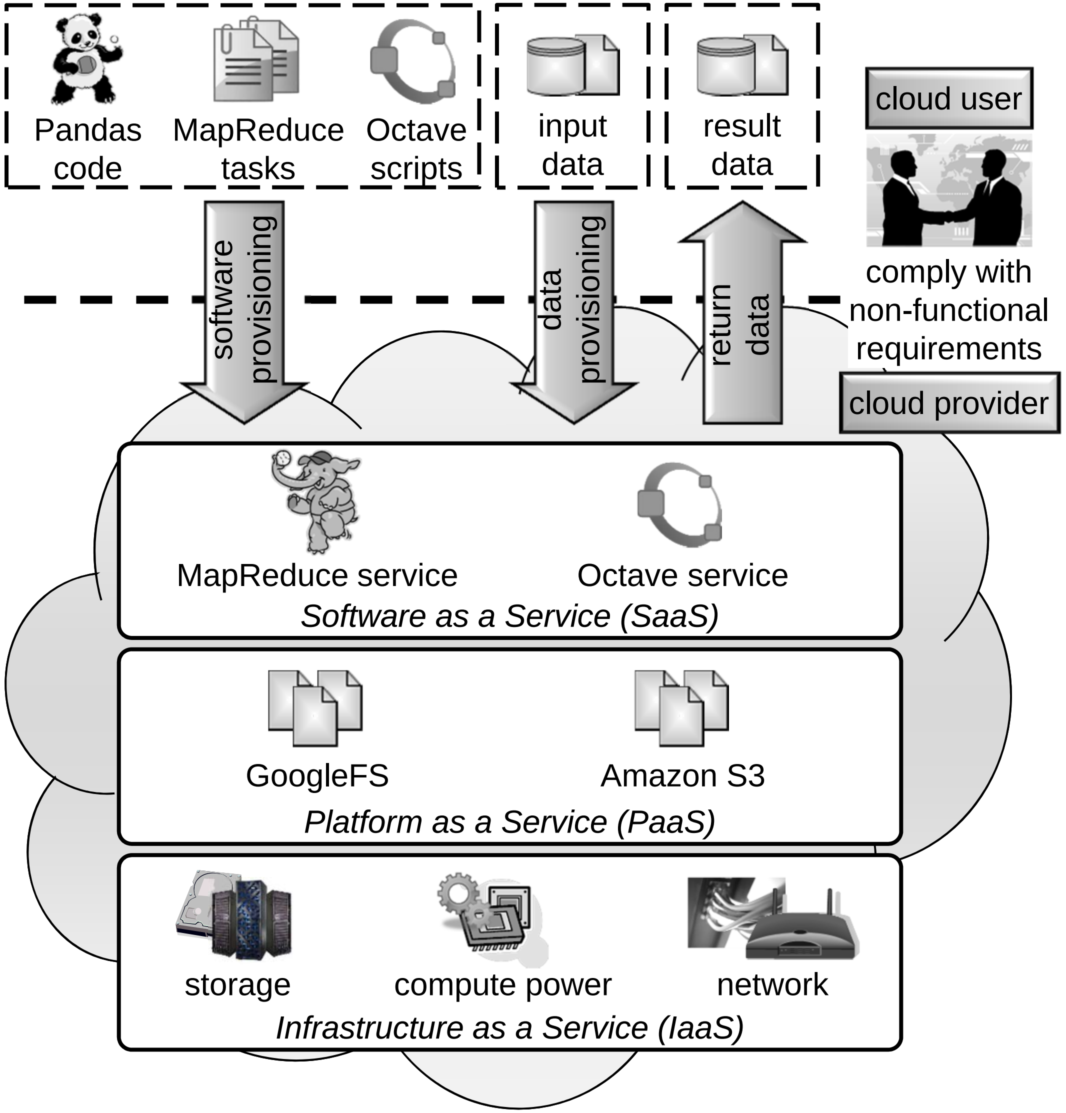

Update Bio-Mechanical Bone Configuration
Sleeping
Walking
Working
Typical Motion Sequences
Bio-Mechanical Calculation (Pandas)
Bio-Mechanical Calculation (Pandas)
Bio-Mechanical Calculation (Pandas)
120 MB
2 GB
600 MB
Characteristic Solutions for Internal Stress Distribution
ETL
ETL
ETL
2 MB
2 MB
2 MB
zed
ion
According to Daily Routine
Systems-Biological Calculation (Octave)
2 MB
Bone Density
Working Step
Size of Result Data
Data Flow

complex ETL processing

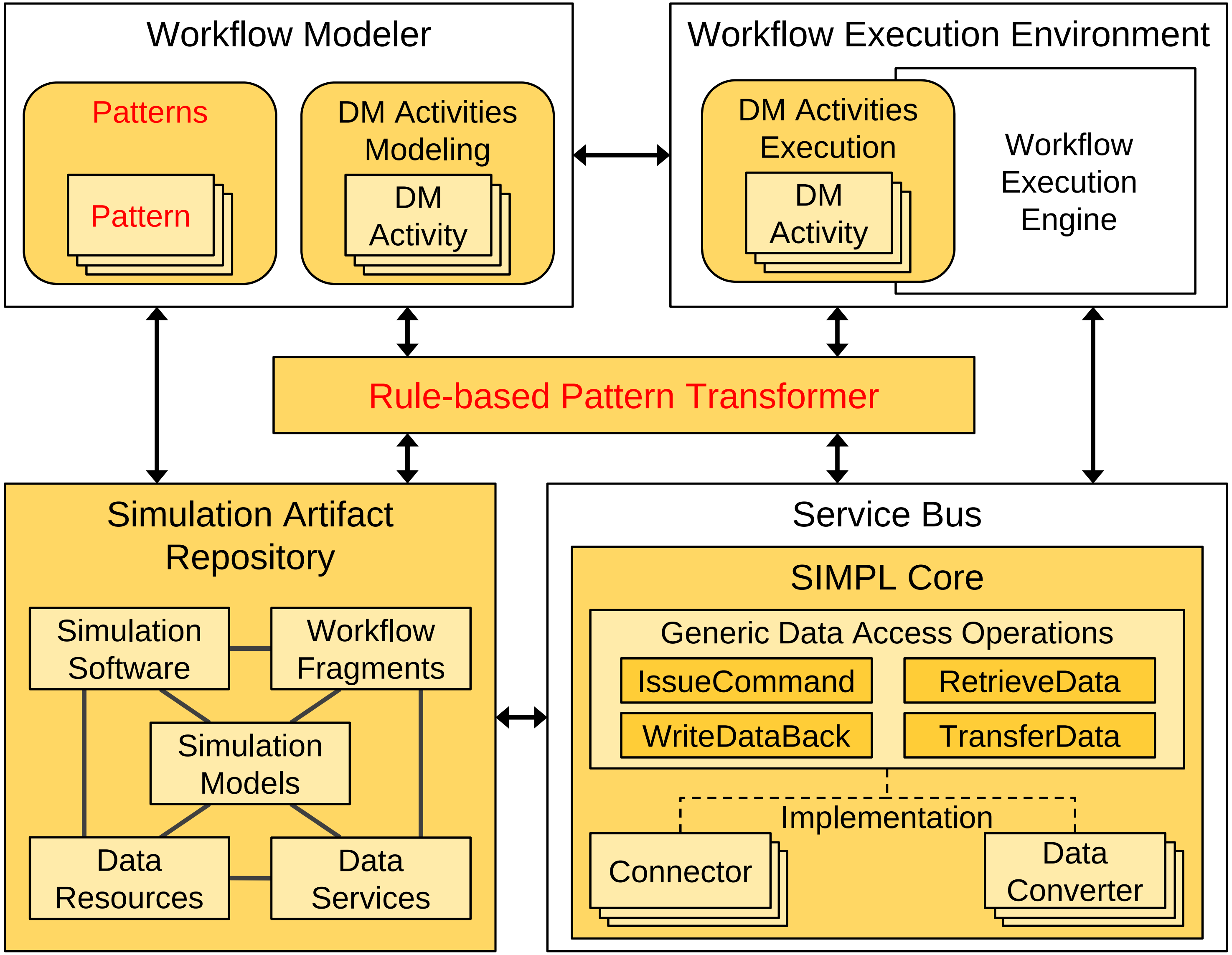
> further abstraction support needed
> optimization decisions

Data Provisioning for Simulation Software

- Complex ETL processes
- Parallel execution possible
- provide MapReduce as SaaS
- service takes MapReduce tasks as input

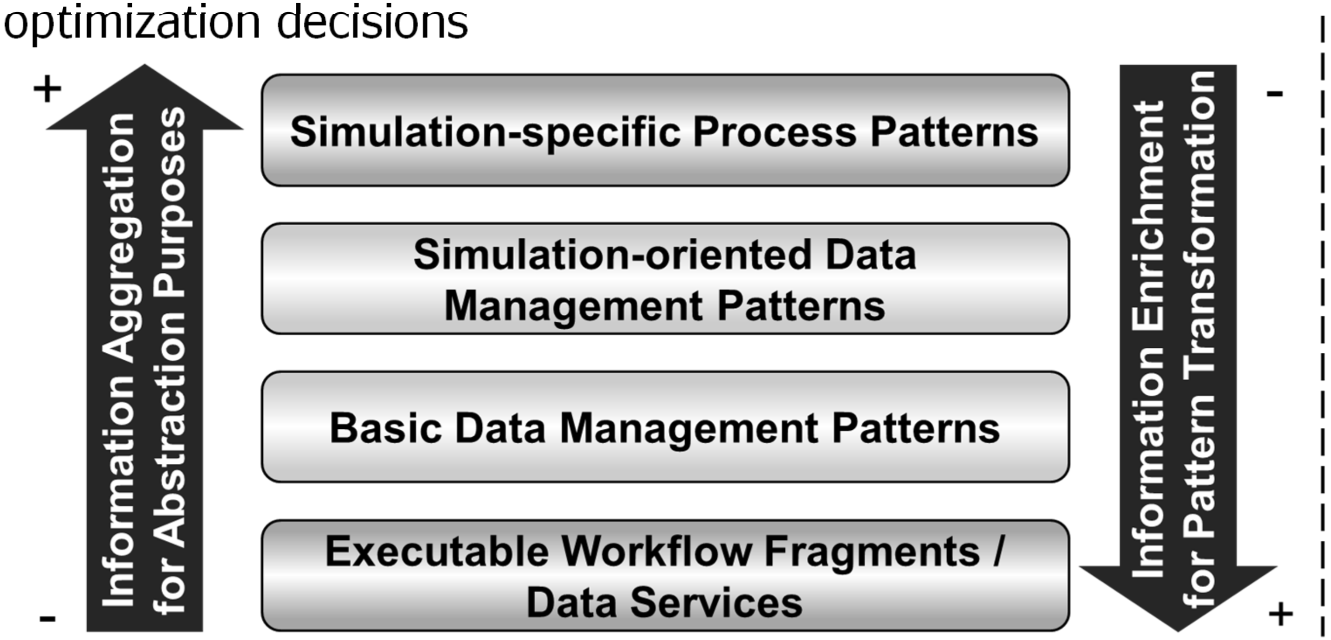


SIMPL Framework



Data Management Optimization

- Pattern-based approach allows for an abstract definition of complex data management tasks
- Rule-based pattern transformer
  - set of transformation rules
  - transfer abstract data management description into executable operations
    → pattern hierarchy
  - includes optimization decisions



+
Information Aggregation for Abstraction Purposes
-
Simulation-specific Process Patterns
Simulation-oriented Data Management Patterns
Basic Data Management Patterns
Executable Workflow Fragments / Data Services
-
Information Enrichment for Pattern Transformation
+

Traceability and Reproducibility

- Traceability
  - scientists need to analyze in detail the simulation results
  - this includes the way they were produced
  - analysis may e.g. reveal issues with data quality
  - needed information is spread over a multitude of tools and system components
  - ➡ Comprehensive provenance framework for simulations needed

- Reproducibility
  - software deployment, data provisioning and computations should be reproducible
  - OpenTOSCA CSAR archives include software artifacts and plans
  - input data have to be kept separately
  - long-term archiving necessary

Conclusion

- simulation workflows are an important means to describe simulations, in particular coupled simulations
- deploying simulation workflows in a public cloud is in particular interesting for organizations running simulations only sporadically
- abstraction support needed enabling scientists to define the necessary complex data management tasks
- SIMPL framework
- further aspects
  - traceability and reproducibility
  - data quality