



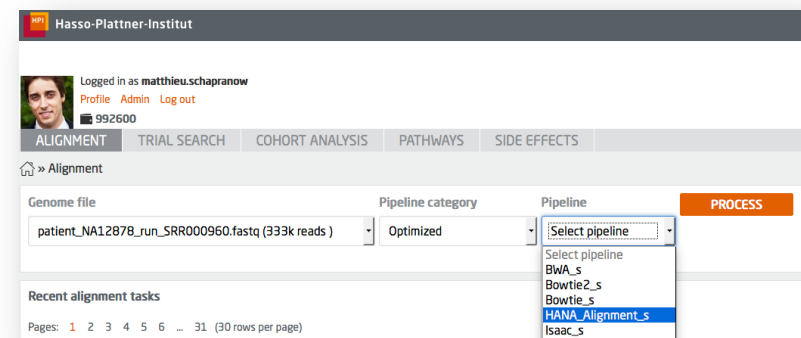
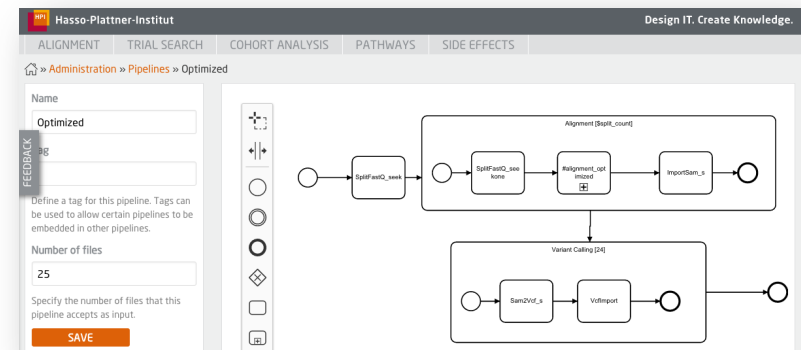
Reproducible Research: Pipeline Execution

Dr. Matthieu-P. Schapranow
Code of Life

November 2016

From Model to Execution

1. Design time (researcher, process expert)
 - Definition of parameterized process model
 - Uses graphical editor and jobs from repository
2. Configuration time (researcher, lab assistant)
 - Select model and specify parameters, e.g. aln opts
 - Results in model instance stored in repository
3. Execution time (researcher)
 - Select model instance
 - Specify execution parameters, e.g. input files



Scheduling and Execution of Pipelines

1. Trigger task execution

Webservice

Tasks

ID	Pipeline	Params
12	BWA	xyz.fastq
13	Stanford	A_1.fastq
14	Bowtie	xyz.fastq

2. Schedule subtasks

Scheduler

In-Memory Database

Subtasks

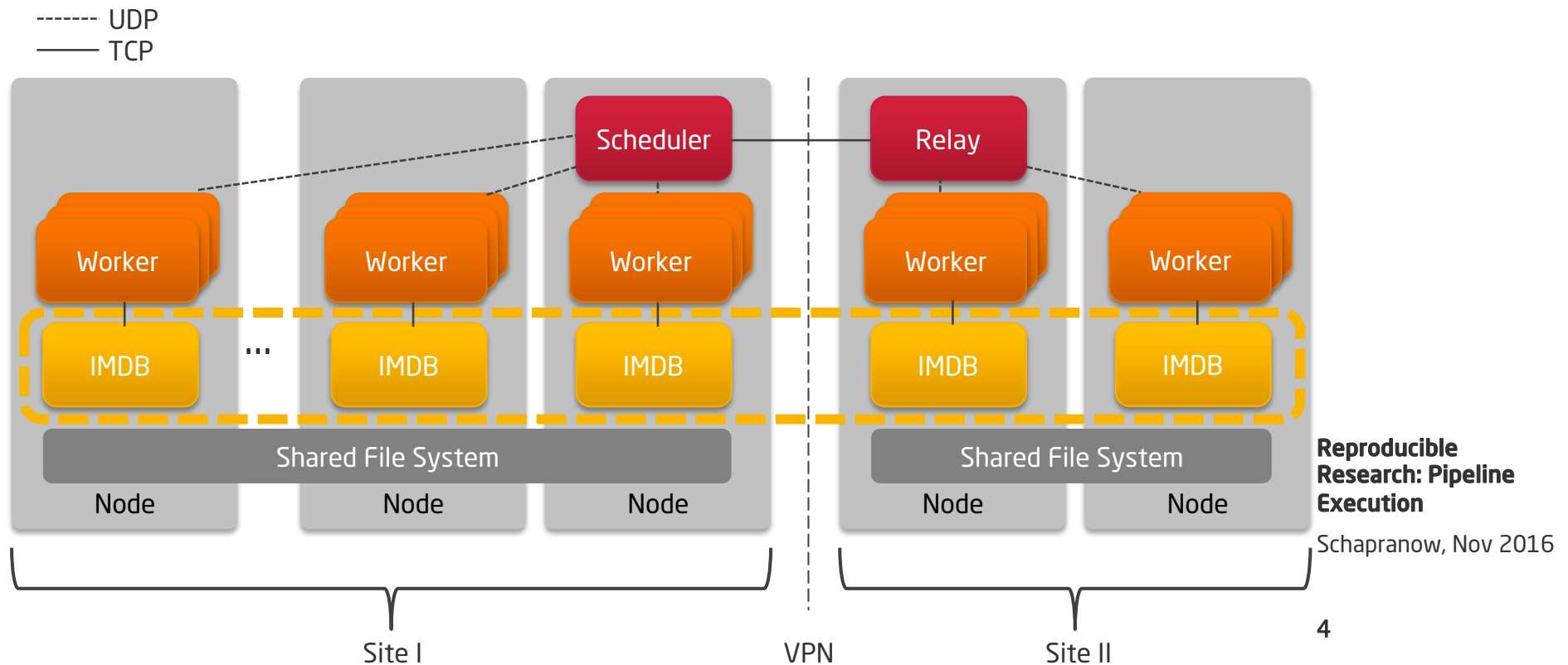
Task	ID	Job	Status	Params
12	97	Split	done	xyz.fastq
12	98	Import	todo	abc.vcf
12	98	Import	done	abc.vcf

3. Execute subtasks

Worker

Worker

Software Components and Communication



IMDB Structure TASKS



Table Name:

TASKS

Columns Indexes Further Properties Runtime Information

	Name	SQL Data Type	Dimens	Column Store Data Type	Key	Not Null	Default
1	ID	INTEGER		INT		X	
2	STATUS	BIGINT		FIXED			
3	PIPELINE_ID	INTEGER		INT		X	
4	PARAMETERS	VARCHAR	5000	STRING			
5	FASTQ_READCOUNT	BIGINT		FIXED			0
6	CREATED_AT	TIMESTAMP		LONGDATE			
7	USER	INTEGER		INT			-1

WORKER

- JOBSTATISTICS
- NODE_GROUPS
- PIPELINES
- SESSIONS
- SUBTASKS
- TASKS

Reproducible

Research: Pipeline

Execution

Schapranow, Nov 2016

1,925	0	11	{ "file": { "name": "smallexample.fastq__2894", "pretty_name": "smallexample.fastq", "type": "use...	0	Oct 27, 2015 5:19:03.769 PM
1,924	2	73	{ "file": { "name": "patient_NA12878_run_SRR000960.fastq__333970", "type": "user", "user_id": "...	333,970	Oct 27, 2015 3:24:10.865 PM
1,921	2	68	{ "file": { "name": "patient_NA12878_run_SRR000960.fastq__333970", "type": "user", "user_id": "...	333,970	Oct 27, 2015 8:12:58.572 AM
1,920	2	69	{ "file": { "name": "patient_NA12878_run_SRR000960.fastq__333970", "type": "user", "user_id": "...	333,970	Oct 27, 2015 8:12:48.055 AM
1,899	2	69	{ "file": { "name": "ERR005584.filt.fastq__65316", "type": "user", "user_id": 3 }, "read_count": 65316 }	65,316	Oct 23, 2015 10:18:36.615 PM
1,895	2	69	{ "file": { "name": "417kb.fastq__2263", "type": "user", "user_id": 1 }, "read_count": 2263 }	2,263	Oct 23, 2015 10:01:09.444 PM
1,894	3	68	{ "file": { "name": "417kb.fastq__2263", "type": "user", "user_id": 1 }, "read_count": 2263 }	2,263	Oct 23, 2015 9:51:25.764 PM

5

IMDB Structure

SUBTASKS

Table Name:

SUBTASKS

Columns	Indexes	Further Properties	Runtime Information				
	Name	SQL Data Type	Dimens	Column Store Data Type	Key	Not Null	Default
1	SUBTASK	INTEGER		INT			
2	TASK	INTEGER		INT			
3	STATUS	INTEGER		INT			
4	JOB	VARCHAR	50	STRING			
5	PARAMETER	VARCHAR	5000	STRING			
6	WORKER	INTEGER		INT			
7	UPDATED_AT	TIMESTAMP		LONGDATE			

SUBTASK	TASK	STATUS	JOB	PARAMETER	WORKER	UPDATED_AT
108,916	1,983	1	BWA_s	{"index": 3, "number_of_instances": 25, "reference": "hg19", "filename": "ugac3tudx080uk8...	2,000	Nov 11, 2015 4:00:29.534 PM
108,915	1,983	1	BWA_s	{"index": 11, "number_of_instances": 25, "reference": "hg19", "filename": "2cqngkjrxo6efb...	2,002	Nov 11, 2015 4:00:29.207 PM
108,916	1,983	0	BWA_s	{"index": 3, "number_of_instances": 25, "reference": "hg19", "filename": "ugac3tudx080uk8...	2	Nov 11, 2015 4:00:29.199 PM
108,891	1,983	2	SplitFa...	{"filename": "63zynd0cj9wm6n49.fastq"}	2,000	Nov 11, 2015 4:00:29.169 PM
108,915	1,983	0	BWA_s	{"index": 11, "number_of_instances": 25, "reference": "hg19", "filename": "2cqngkjrxo6efb...	2	Nov 11, 2015 4:00:28.706 PM
108,889	1,983	2	SplitFa...	{"filename": "fyed8cl9p4mkut1m.fastq"}	2,002	Nov 11, 2015 4:00:28.682 PM
108,914	1,983	1	BWA_s	{"index": 17, "number_of_instances": 25, "reference": "hg19", "filename": "qa61bwu2yvk40...	2,001	Nov 11, 2015 3:59:51.547 PM
108,914	1,983	0	BWA_s	{"index": 17, "number_of_instances": 25, "reference": "hg19", "filename": "qa61bwu2yvk40...	2	Nov 11, 2015 3:59:50.285 PM
108,886	1,983	2	SplitFa...	{"filename": "pq45v2h820v59ckm.fastq"}	2,001	Nov 11, 2015 3:59:50.262 PM
108,913	1,983	1	BWA_s	{"index": 18, "number_of_instances": 25, "reference": "hg19", "filename": "7nb56bhunugcc...	1,000	Nov 11, 2015 3:59:40.927 PM
108,913	1,983	0	BWA_s	{"index": 18, "number_of_instances": 25, "reference": "hg19", "filename": "7nb56bhunugcc...	1	Nov 11, 2015 3:59:40.345 PM
108,910	1,983	2	SplitFa...	{"filename": "7ewxddmppvj1vnpn.fastq"}	1,000	Nov 11, 2015 3:59:40.308 PM

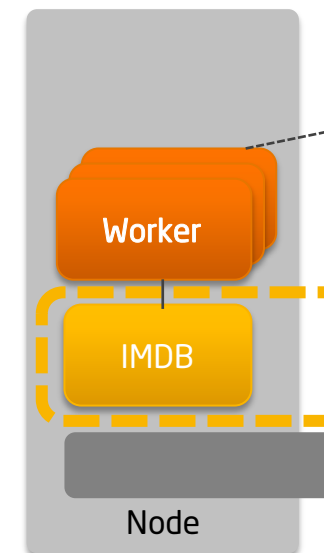
- ▼ WORKER
 - JOBSTATISTICS
 - NODE_GROUPS
 - PIPELINES
 - SESSIONS
 - SUBTASKS
 - TASKS

**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

Runtime Layer Worker

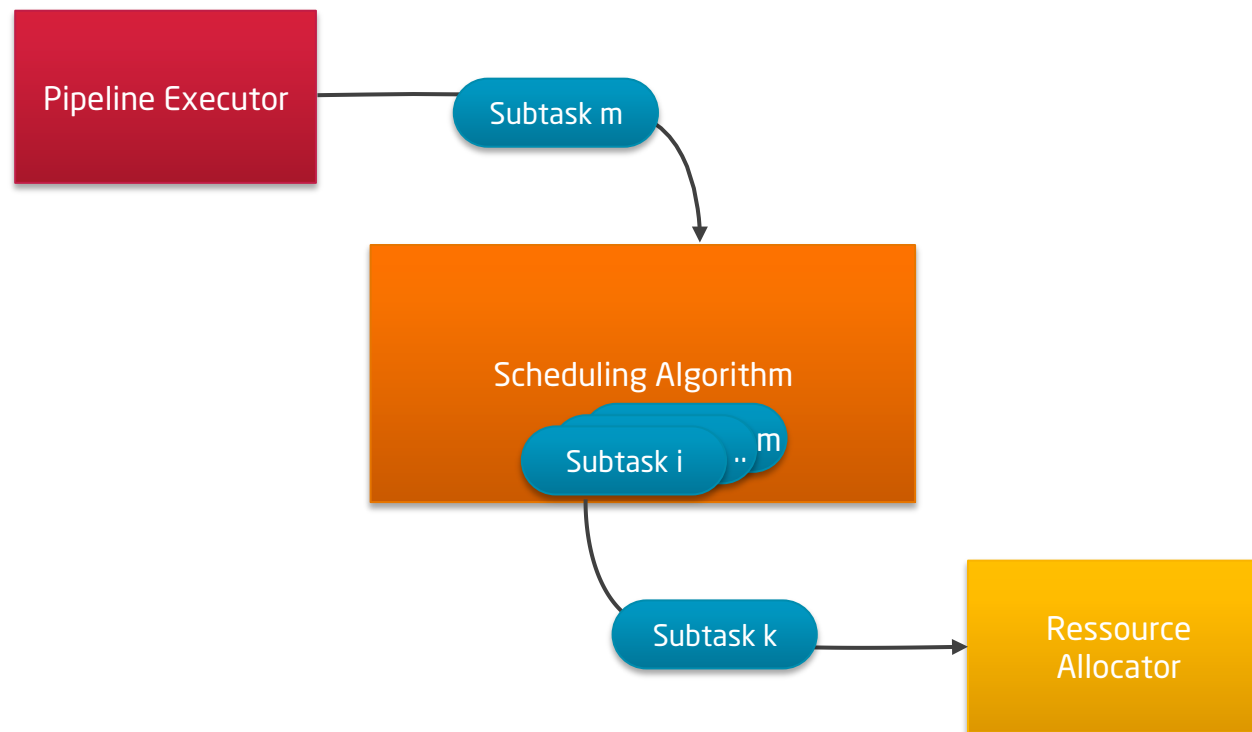
- Workers execute jobs one by one
- Subtask execution status in IMDB:
 - Ready (0),
 - In Progress (1),
 - Done (2), or
 - Erroneous (3).
- Jobs implemented as Python modules/classes
 - Can contain arbitrary code
 - Have access to IMDB
 - Can read/write to shared working directory



**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

Coordination Layer Scheduler

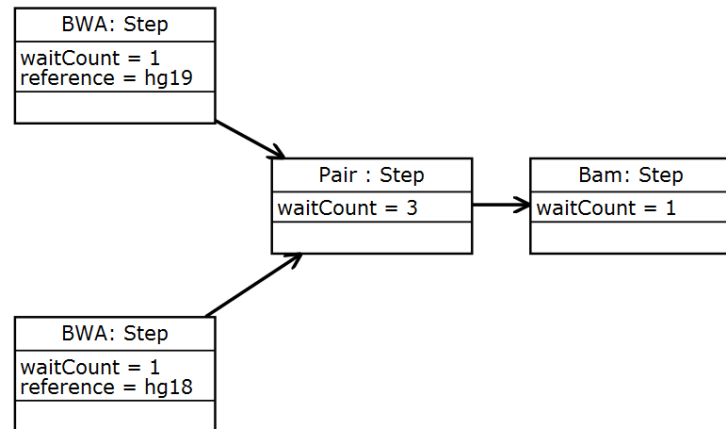


**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

Scheduling Pipeline Executer

```
<xpdl:Activity CompletionQuantity="1" Id="newpkg1_wp1_act2" Name="BWA">
  <xpdl:Implementation>
    <xpdl:No/>
  </xpdl:Implementation>
  <xpdl:Performers>
    <xpdl:Performer>newpkg1_wp1_par1</xpdl:Performer>
  </xpdl:Performers>
  <xpdl:NodeGraphicsInfos>
    <xpdl:NodeGraphicsInfo BorderColor="#000000" FillColor="#99FF99">
      <xpdl:Coordinates XCoordinate="239.0" YCoordinate="219.0"/>
    </xpdl:NodeGraphicsInfo>
  </xpdl:NodeGraphicsInfos>
</xpdl:Activity>
```

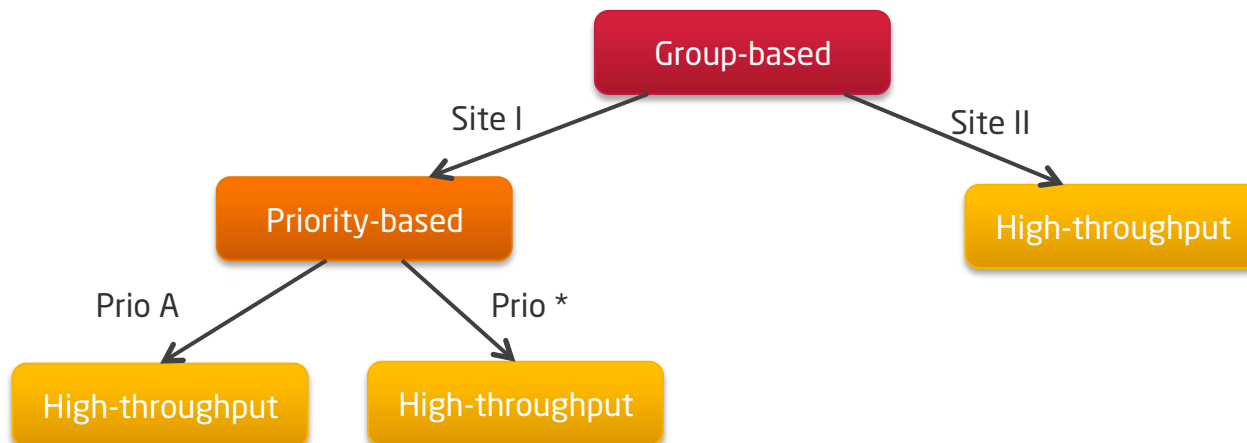


**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

Scheduling Algorithms

- Scheduling algorithms are plug-in software modules
 - "User-/Group-based" to let users execute their tasks on their local site only
 - "Priority First" to prefer important users
 - "High Throughput", i.e. "shortest task first" to deal with high load
- Scheduling algorithms can also be composed hierarchically

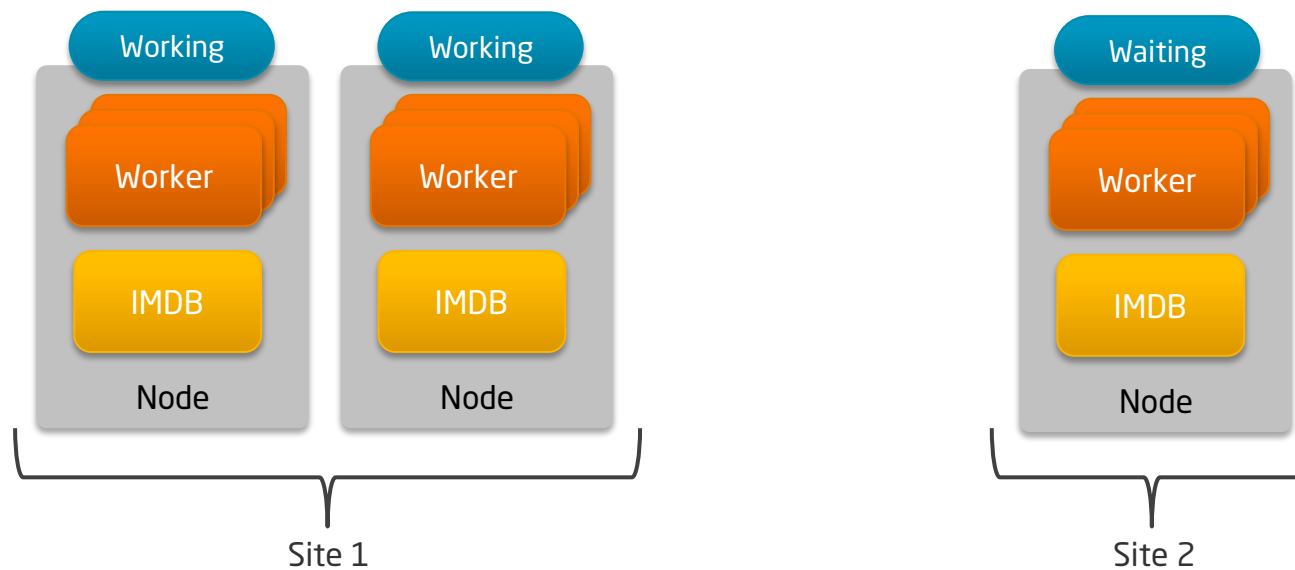


**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

Scheduling Resource Allocator

- Maintains lists of running and idle nodes
- Idle worker requests new sub task for its assigned groups
- If there is no matching sub task, it sleeps until a new sub task gets ready

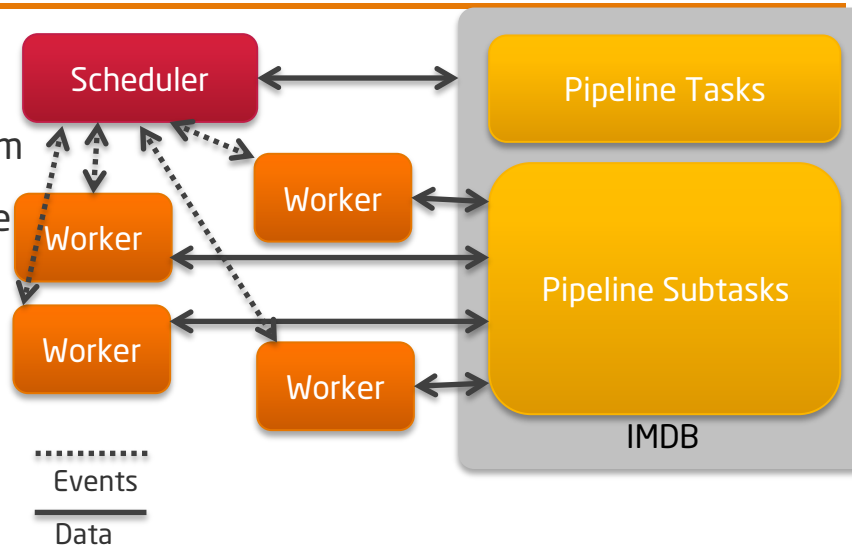


**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

Recoverability

- All execution data is stored in IMDB
- Temporary files on a shared file system
- In case of any failure, the system-wide state can be restored



TypeError: NoneType object is not subscriptable

```
2015-11-04 18:01:30 INFO [ContinuingCoordinator] will start task with ID 1969
2015-11-04 18:01:30 INFO [ContinuingCoordinator] Will continue old but unfinished task 1969 with 52 already done subtasks.
2015-11-04 18:01:31 ERROR [ContinuingCoordinator] Taskbook (most recent call last):
```

**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

What is Next?

- How to bring all puzzle pieces together?
- How to enable real-time data analysis?



Human genome/biological data
600GB per full genome
15PB+ in databases of leading institutes



Human proteome
160M data points (2.4GB) per sample
>3TB raw proteome data in ProteomicsDB



Hospital information systems
Often more than 50GB



PubMed database
>23M articles



Cancer patient records
>160k records at NCT



Medical sensor data
Scan of a single organ in 1s
creates 10GB of raw data



Prescription data
1.5B records from 10,000 doctors
and 10M Patients (100 GB)



Clinical trials
Currently more than 30k
recruiting on ClinicalTrials.gov

**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016

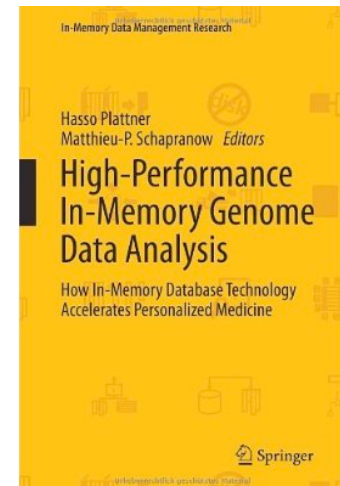
Keep in contact with us!



Dr. Matthieu-P. Schapranow
Program Manager E-Health & Life Sciences

Hasso Plattner Institute
August-Bebel-Str. 88
14482 Potsdam, Germany

schapranow@hpi.de
<http://we.analyzegenomes.com/>



**Reproducible
Research: Pipeline
Execution**

Schapranow, Nov 2016