Data Lakes Workshop

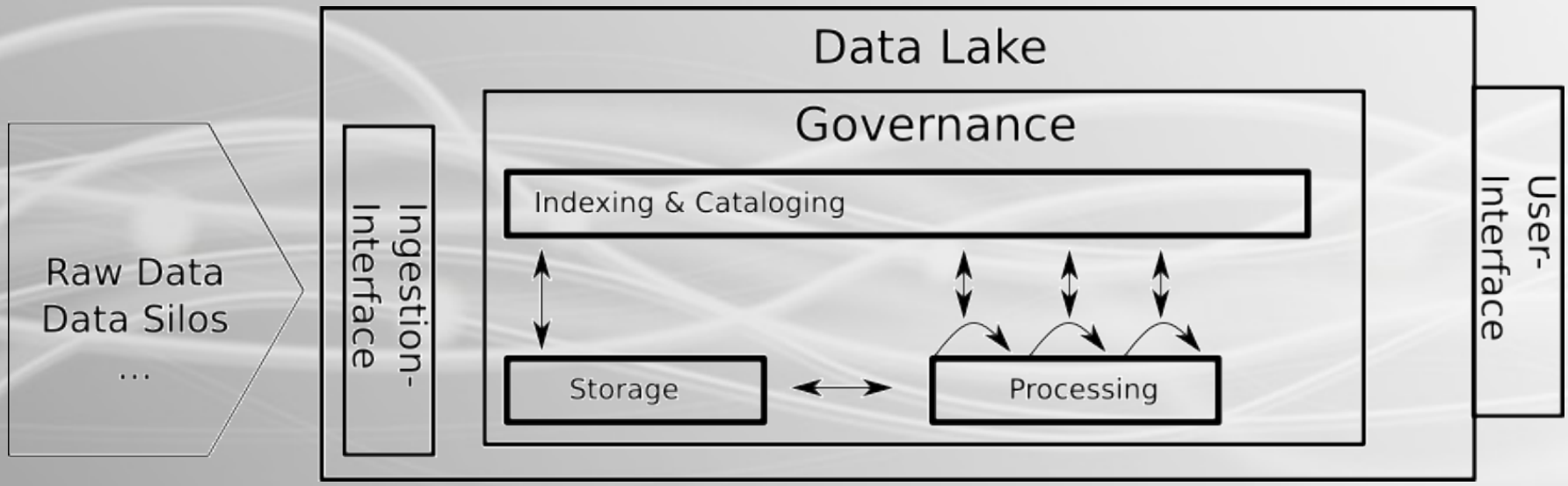
Hendrik Nolte, Julian Kunkel, Piotr Kasprzak

12.11.2021

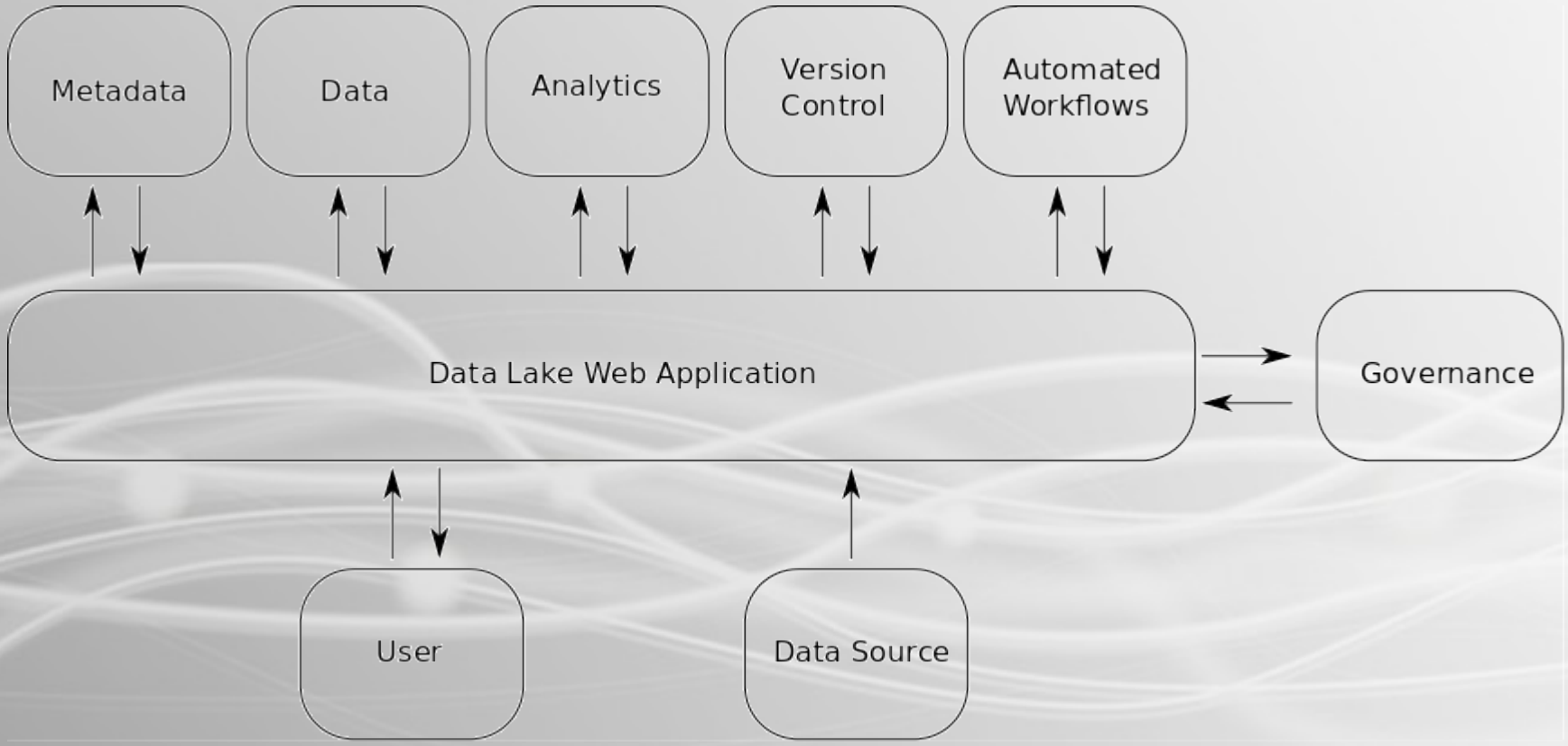
Content

- Quick Overview over these two sub-projects:
  - Development of a data lake solution
    - Combines various tools under one management and control
    - Additional software layers, e.g., for governance
  - Secure data processing for highest privacy levels
    - Hardened workflow

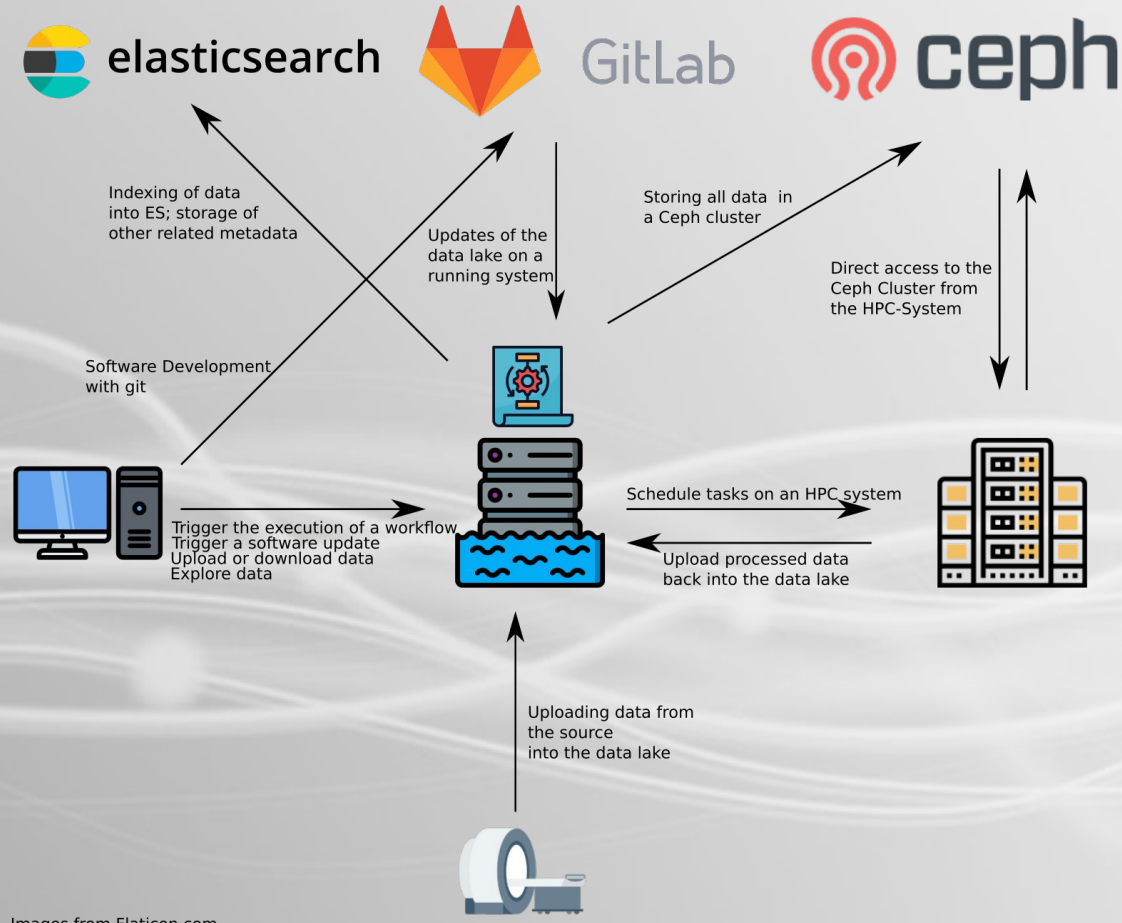
Data Lake Basic Concept



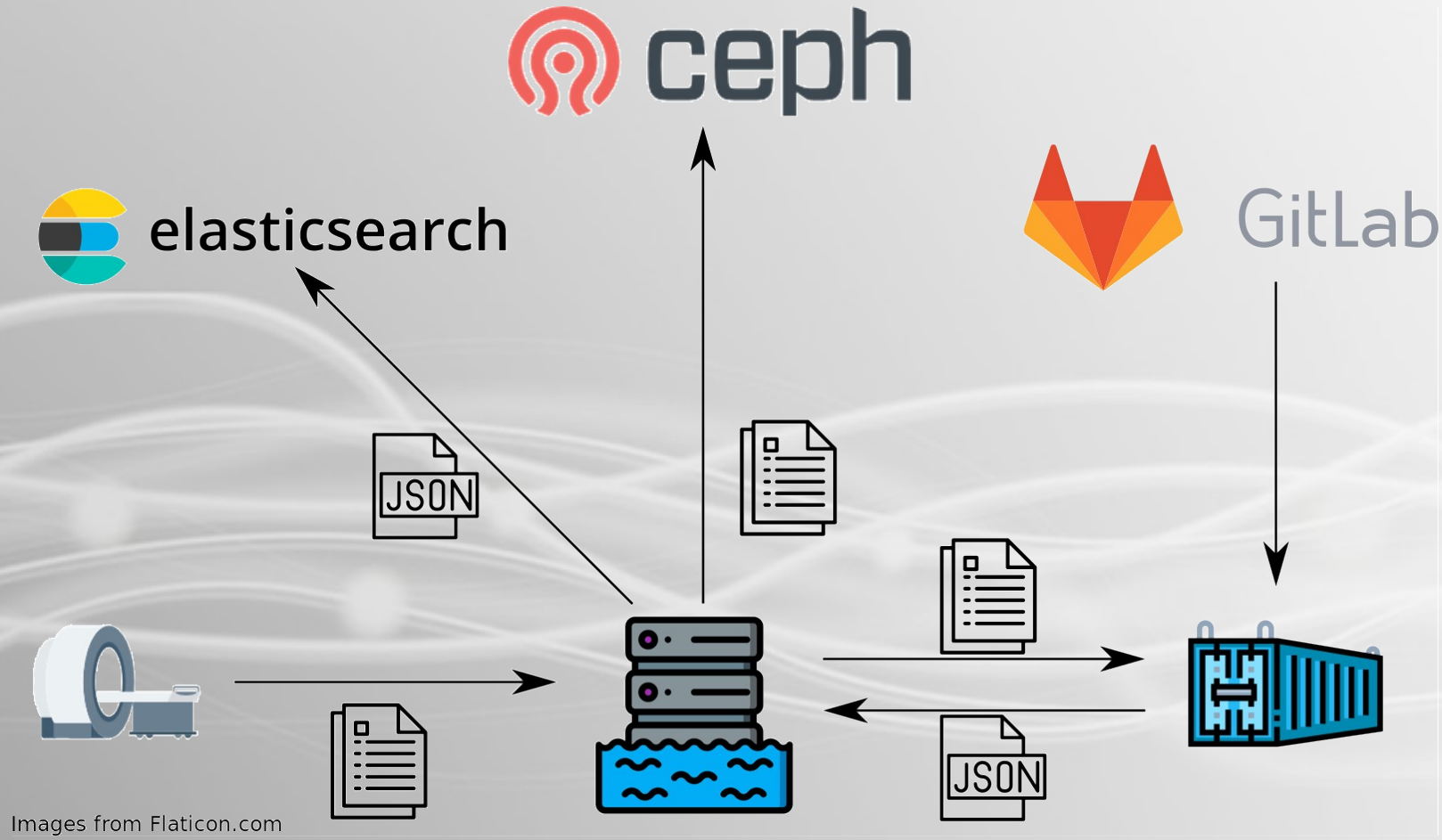
High-Level Architecture



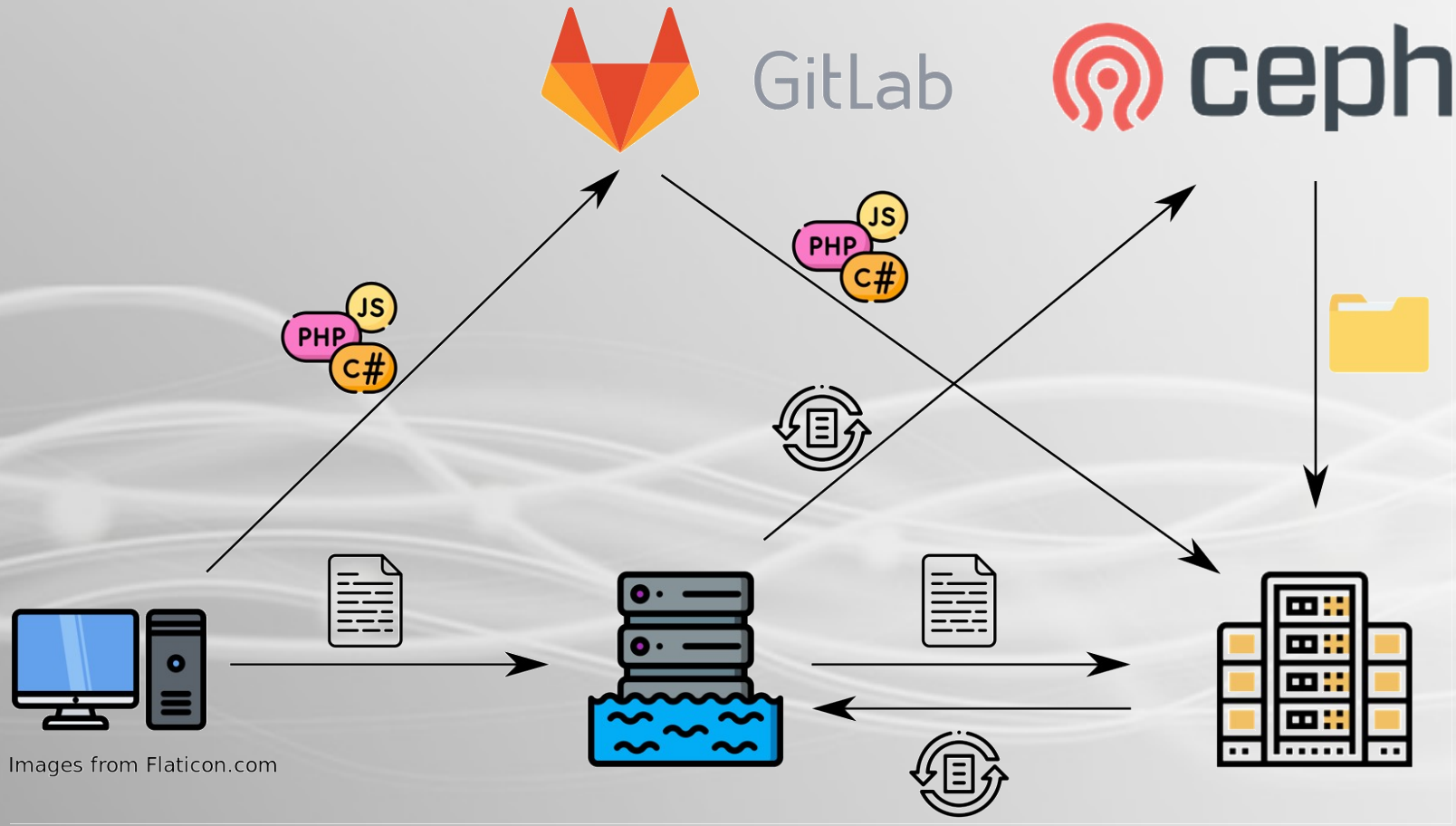
Our Implementation



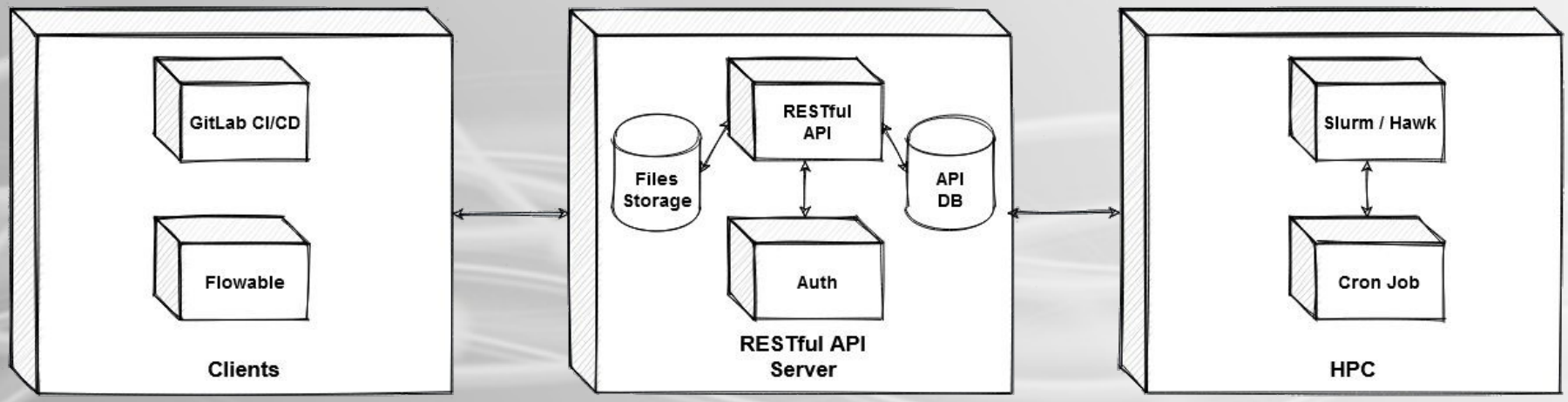
Ingestion


ceph
elasticsearch
GitLab
JSON
JSON
Images from Flaticon.com

Analytics



HPCSerA


GitLab CI/CD
Flowable
Clients
RESTful API
Files Storage
API DB
Auth
RESTful API Server
Slurm / Hawk
Cron Job
HPC

Job Manifest

In [15]:
```
import data_lake

conn = data_lake.Connection(username='',password='', url='data-lake.gwdg.de')
```

In [8]:
```
job = data_lake.Job()
job.comment = 'Beispiel Job'
job.container_name = 'bart-DeepDeepLearning'
job.compute.append('cd /program/hpc-statustagung-rep/')
job.compute.append('python3 example_python_script.py')
```

In [9]:
```
job.git = [
    {
        "uri" : "git@gitlab.gwdg.de:hnolte1/hpc-statustagung-rep.git",
        "type" : "build",
        "bash" : "mkdir /program/output/"
    }
]
```

In [10]:
```
job.job_name = "Output-Test"
job.data_category = "kspace"
job.env_var = {
    "OUTPUT_FOLDER" : "/program/output/",
    "NUM_REPETITIONS" : "2"
}
```

In [11]:
```
job.output_dir='/program/output'
```

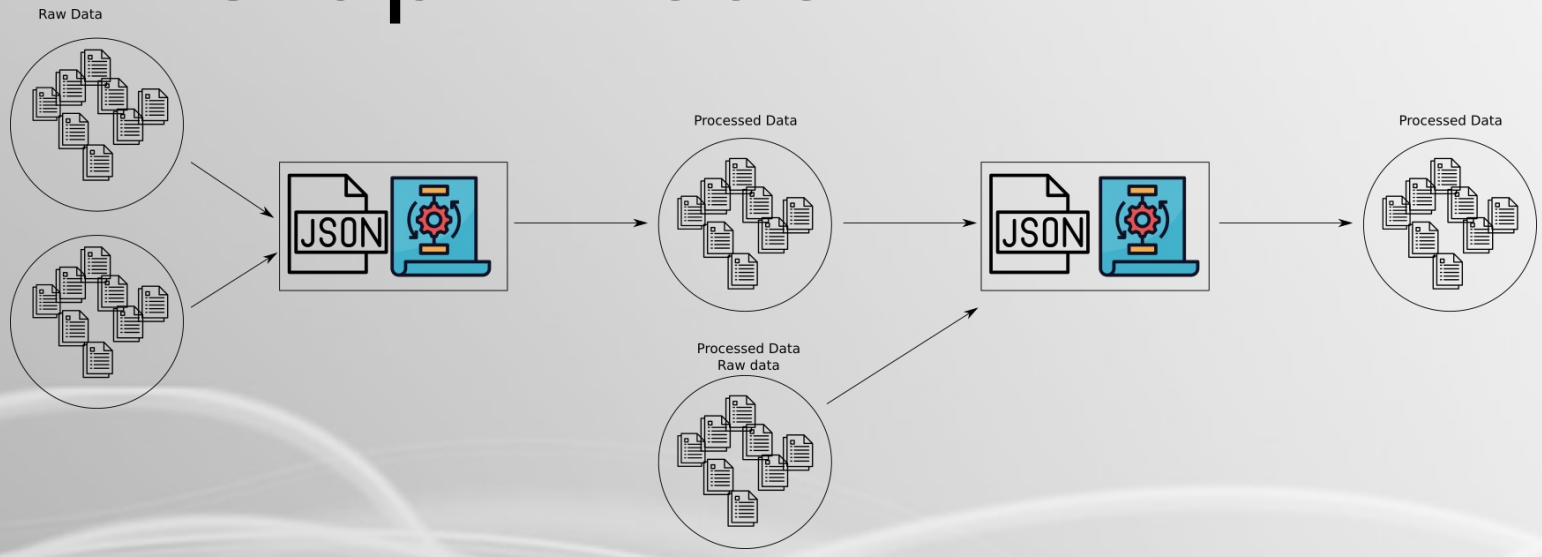
In [ ]:
```
res = job.submit(conn)
```

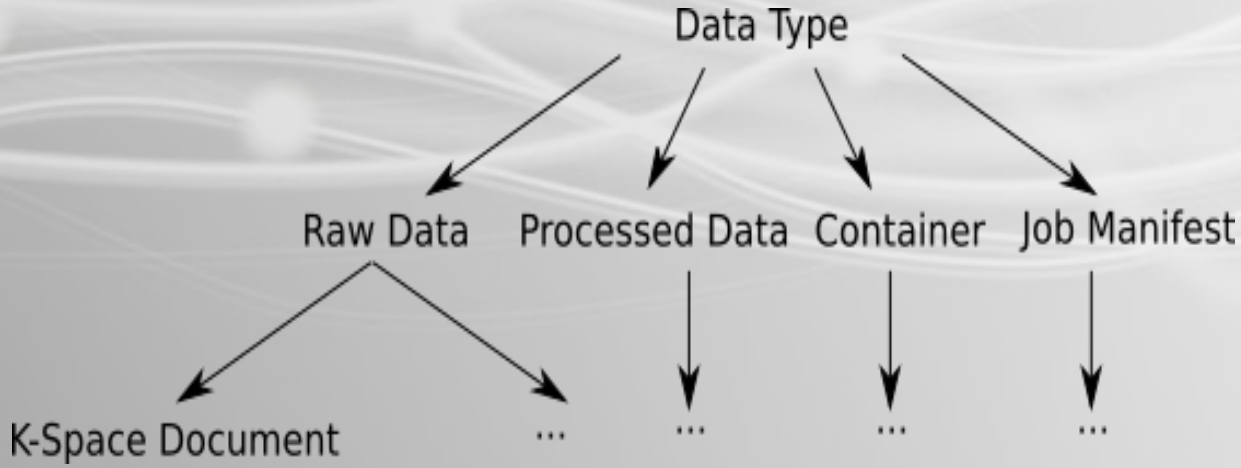
In [ ]:
```
res = data_lake.show_jobs(conn, 'DataOutput','output_file_1.txt',format='GitRepositories,GitCommits,ContainerName,DataOutput,Comment,NUM_REPETITIONS,CreateDate')
```

In [17]:
```
print(res.content.decode('utf-8'))
```

```
+----------------------+------------------------------------------+--------------------------+------------------+---------------+----------------+----------------------------+
|    GitRepositories   |                GitCommits                |       ContainerName      |     DataOutput   |    Comment    | NUM_REPETITIONS |         CreateDate         |
+----------------------+------------------------------------------+--------------------------+------------------+---------------+----------------+----------------------------+
| hpc-statustagung-rep, | a1077ca42de7e8c4bc2f6d9265ddd07627588115, | bart-DeepDeepLearning.sif | output_file_1.txt |  Beispiel Job |        2       | 2021-04-16T15:09:24.563806 |
|                      |                                          |                          | output_file_0.txt |               |                |                            |
|                      |                                          |                          |                  |               |                |                            |
| hpc-statustagung-rep, | a1077ca42de7e8c4bc2f6d9265ddd07627588115, |        bart-ddl.sif       | output_file_2.txt |  Beispiel Job |        3       | 2021-04-06T20:36:43.716102 |
|                      |                                          |                          | output_file_1.txt |               |                |                            |
|                      |                                          |                          | output_file_0.txt |               |                |                            |
|                      |                                          |                          |                  |               |                |                            |
| hpc-statustagung-rep, | a1077ca42de7e8c4bc2f6d9265ddd07627588115, |        bart-ddl.sif       | output_file_1.txt | Beispiel Job 3 |        2       | 2021-04-07T13:11:48.496271 |
|                      |                                          |                          | output_file_0.txt |               |                |                            |
```

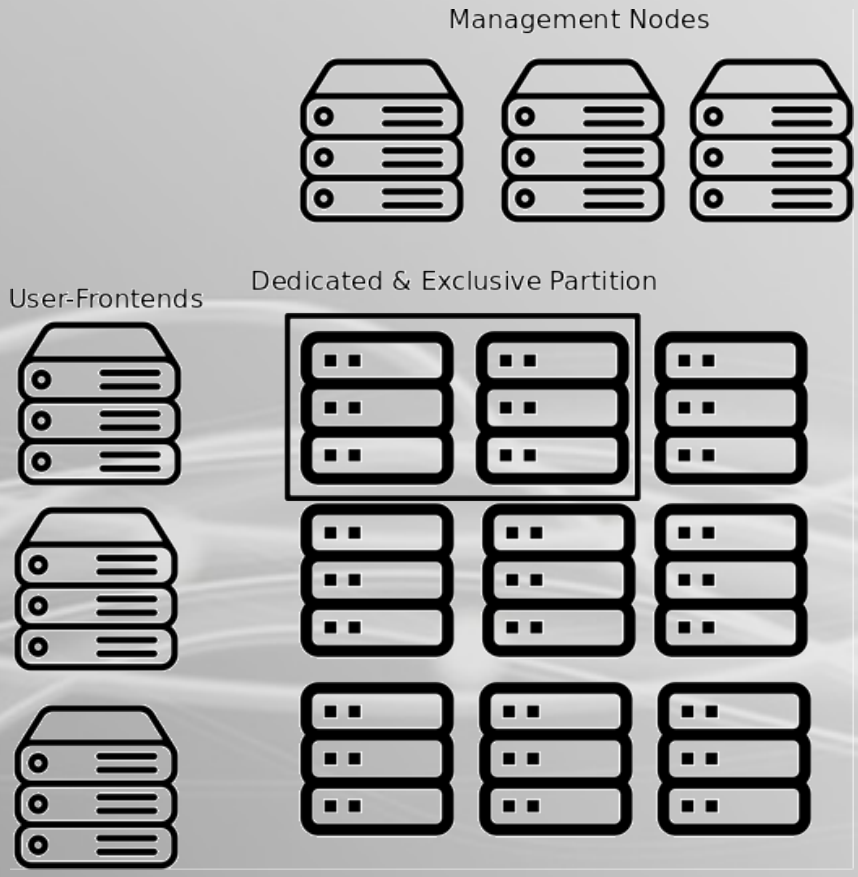
Provenance-Centred Graph Model

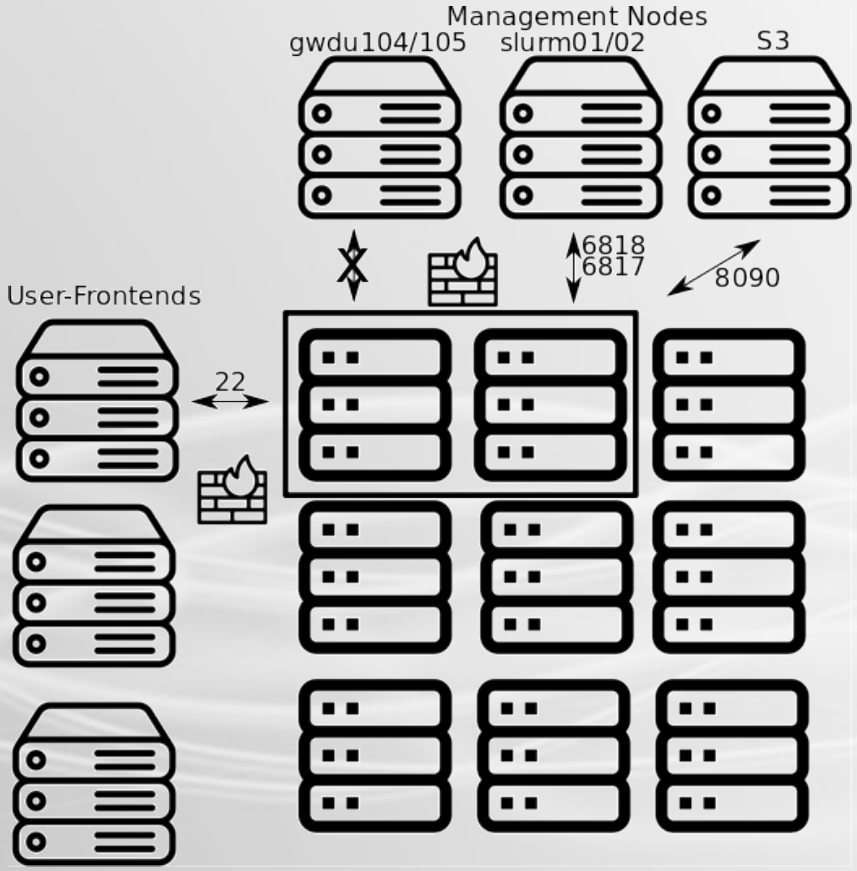




Secure Workflow

Secure Nodes





Secure Workflow