B B M R I • N L

Biobanking and Biomolecular Resources Research Infrastructure

Using the Dutch Life Science GRID for RNA-seq analysis in the BBMRI Biobank-based Integrative Omics Study

LUMC: Michiel van Galen, Wibowo Arindrarto, Matthijs Moed, Leon Mei

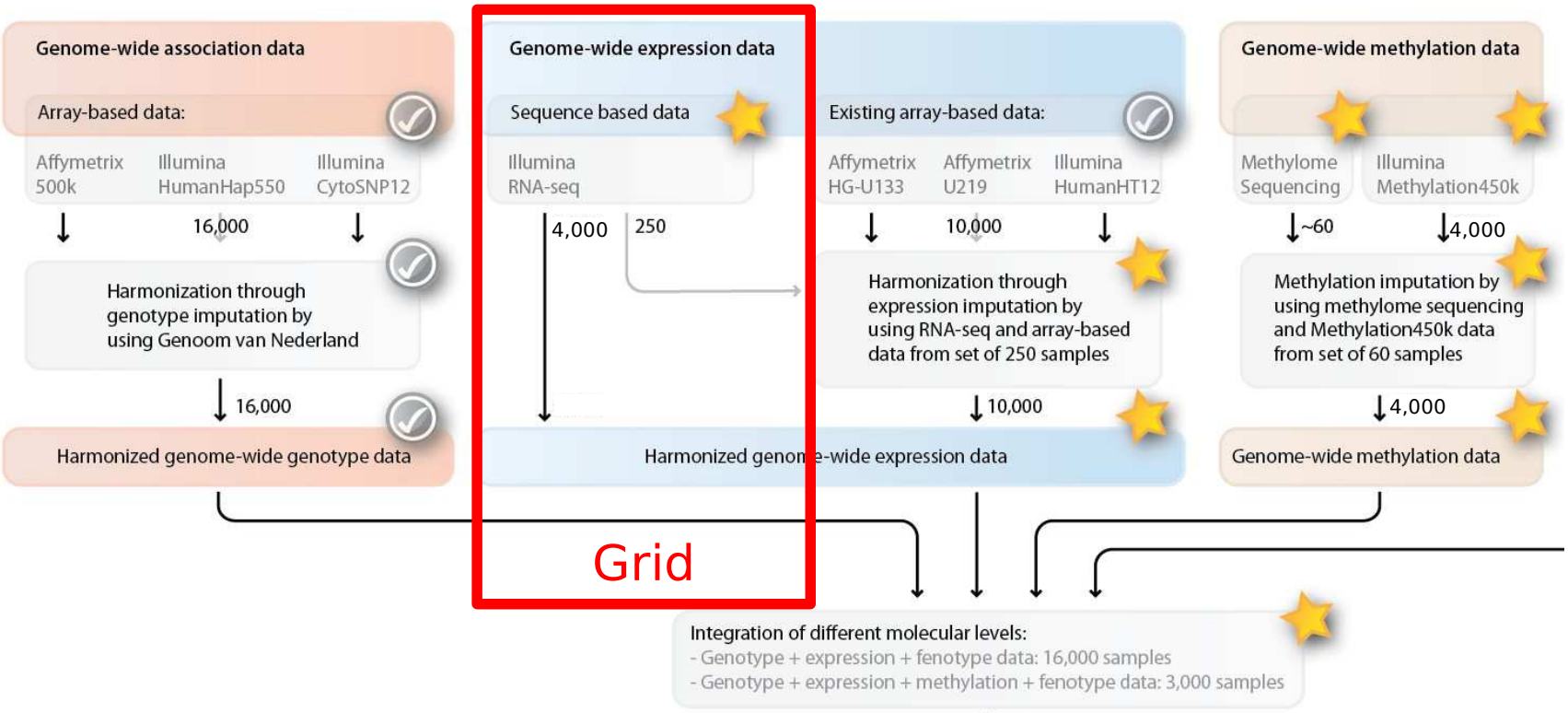
ErasmusMC: Jeroen van Rooij, Marijn Verkerk

UMCG: Freerk van Dijk SURFsara: Jan Bot

Content

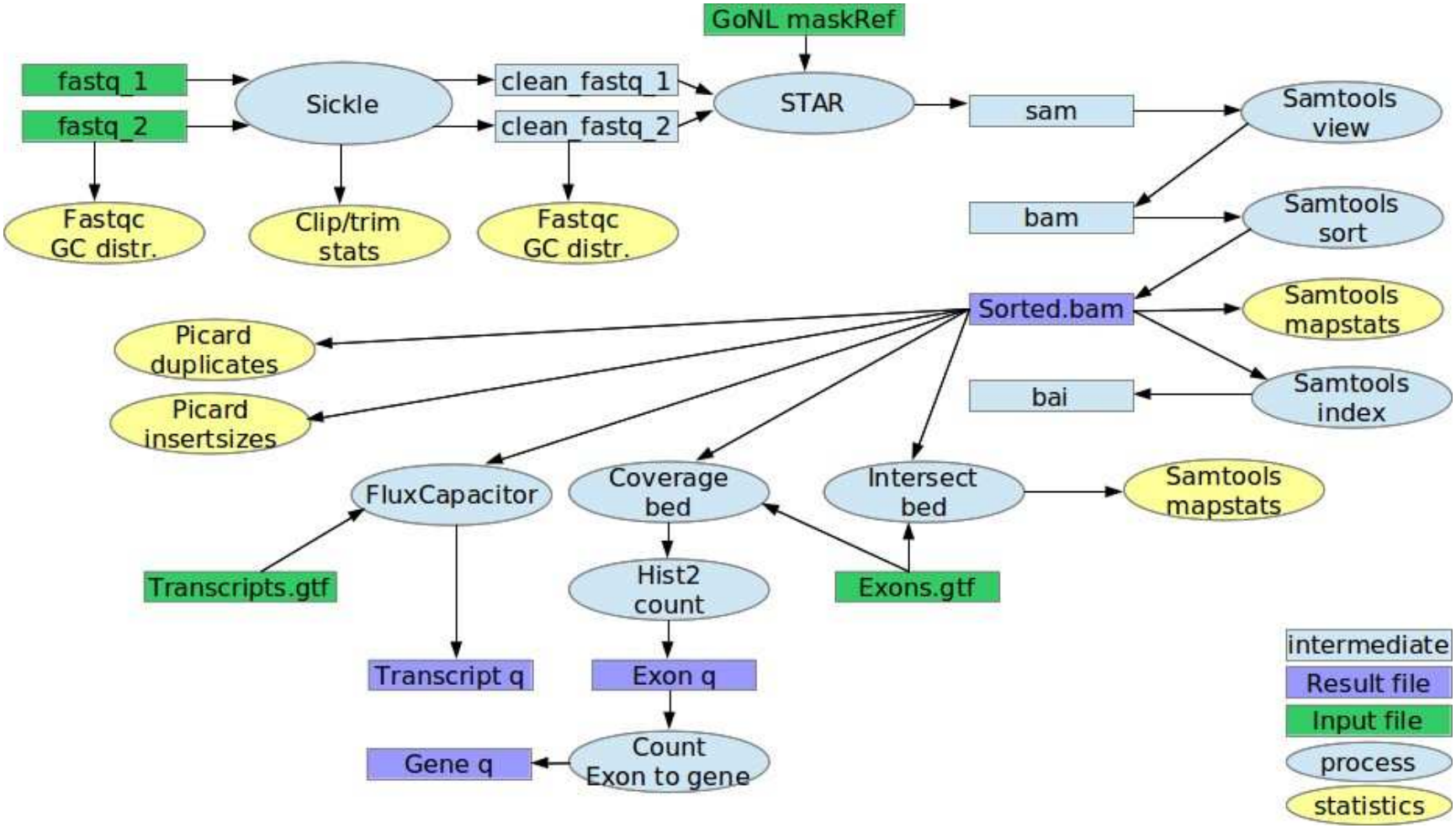
- Background of the study
- GRID structure
- Job management
- Best practices

BBMRI-NL Functional Genomics Project (RP3)



Note: number of samples in this slide have been adjusted in the current project

Analysis pipeline



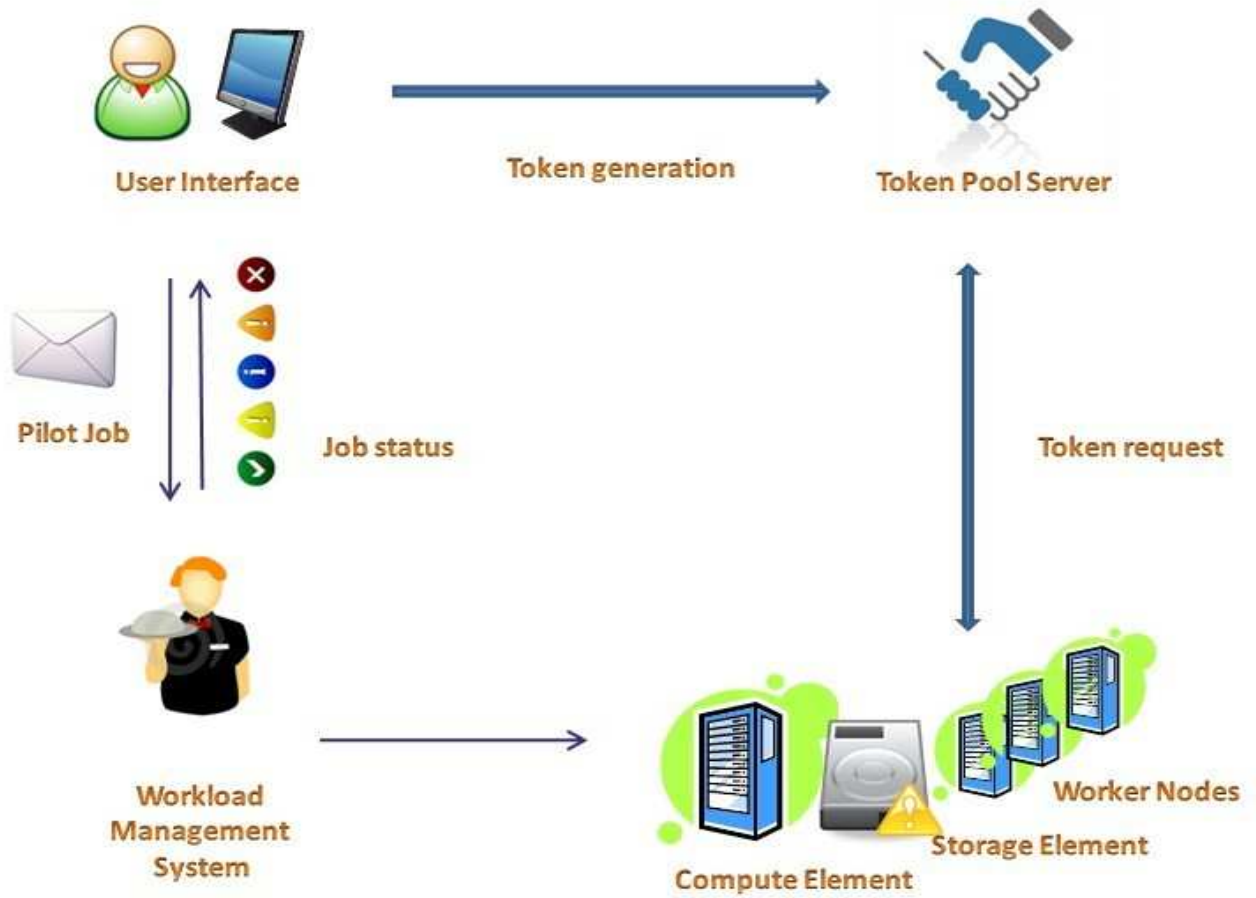
Dutch Life Science Grid

11 dedicated clusters in WUR, NKI, AMC, VU, LUMC, RU Nijmegen, Hubrecht, TU Delft, ErasmusMC , Keygene, RUG

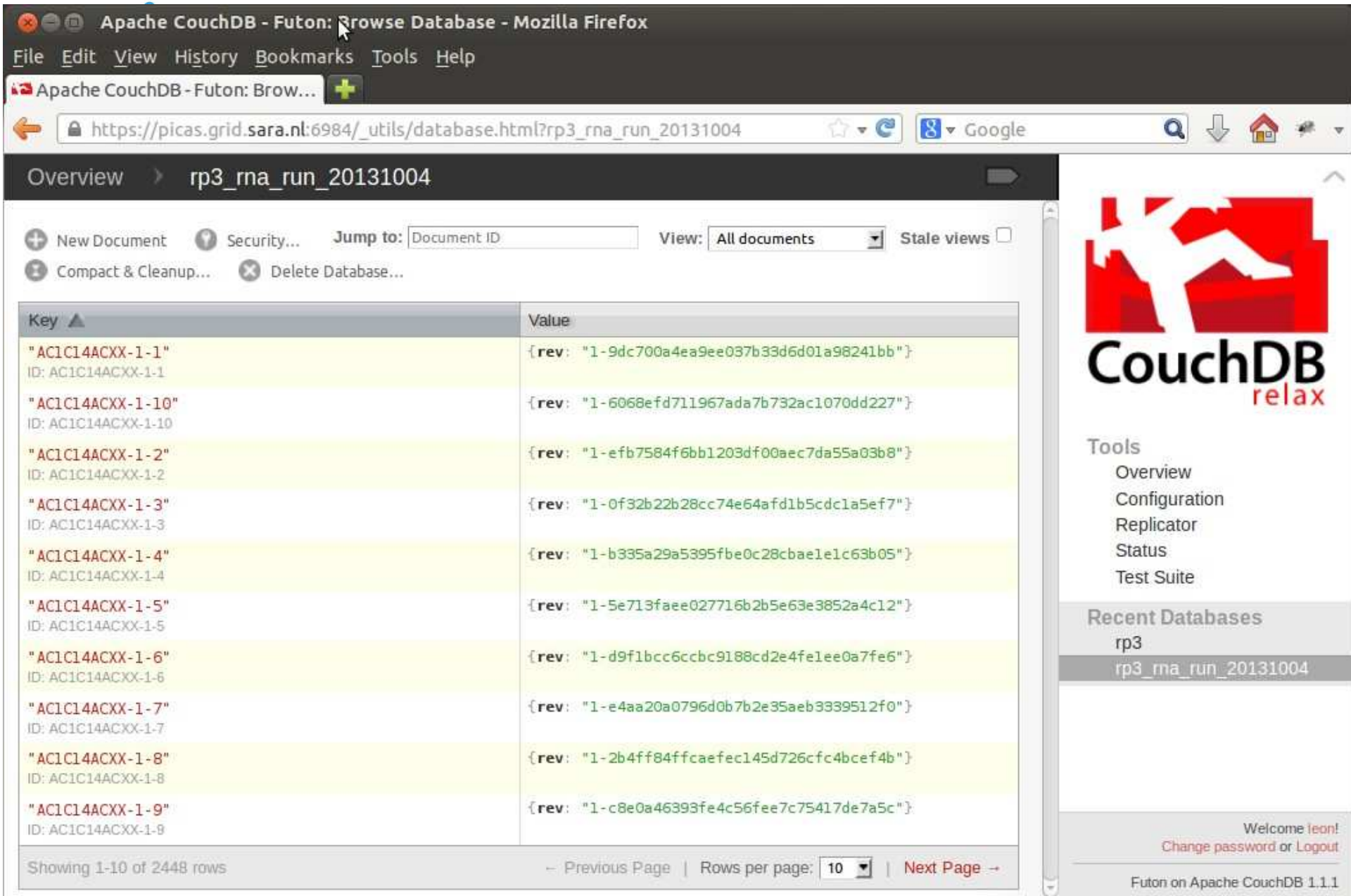
Each cluster has 128 CPUs, 512G memory, 48TB storage.

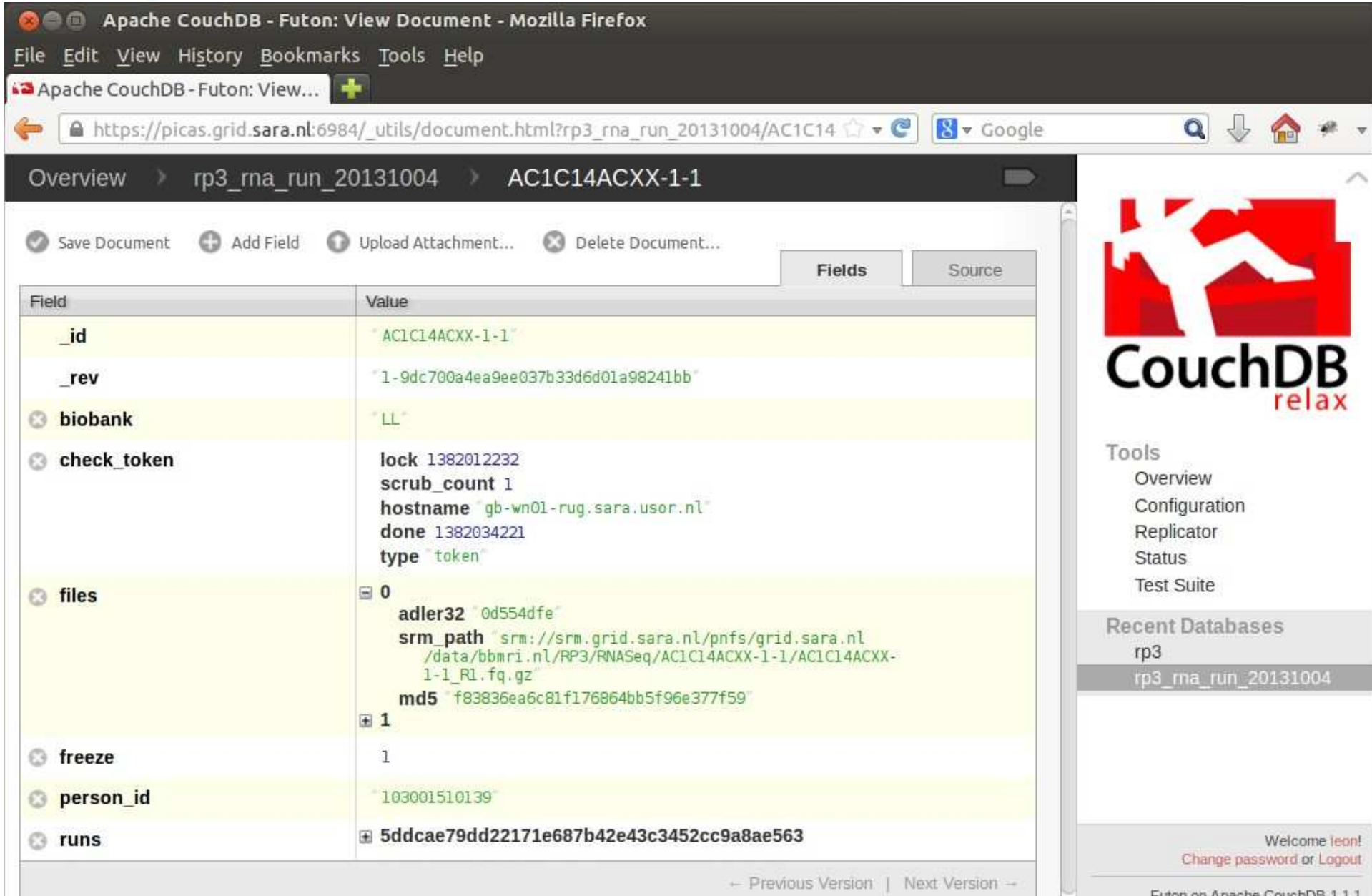
Additional resources can be added to LSG at NIKHEF, GINA, RUG sites.

PiCaS – use CouchDB as token pool server

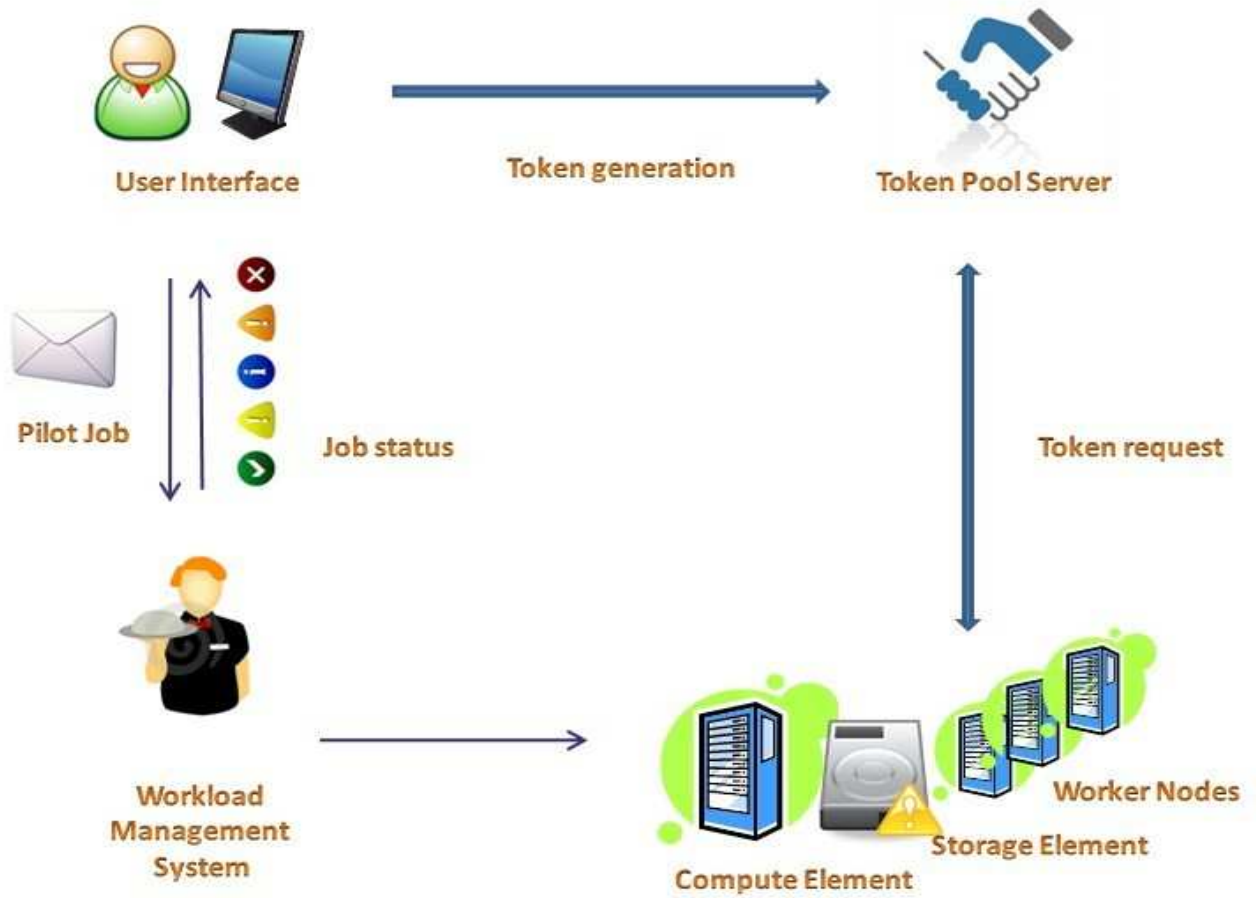


Picture taken from SARA wiki


Apache CouchDB - Futon: Browse Database - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Apache CouchDB - Futon: Brow...
https://picas.grid.sara.nl:6984/_utils/database.html?rp3_rna_run_20131004
Google
Overview
rp3_rna_run_20131004
New Document
Security...
Jump to: Document ID
View: All documents
Stale views
Compact & Cleanup...
Delete Database...
Key
Value
"AC1C14ACXX-1-1" ID: AC1C14ACXX-1-1 {rev: "1-9dc700a4ea9ee037b33d6d01a98241bb"}
"AC1C14ACXX-1-10" ID: AC1C14ACXX-1-10 {rev: "1-6068efd711967ada7b732ac1070dd227"}
"AC1C14ACXX-1-2" ID: AC1C14ACXX-1-2 {rev: "1-efb7584f6bb1203df00aec7da55a03b8"}
"AC1C14ACXX-1-3" ID: AC1C14ACXX-1-3 {rev: "1-0f32b22b28cc74e64afd1b5cdc1a5ef7"}
"AC1C14ACXX-1-4" ID: AC1C14ACXX-1-4 {rev: "1-b335a29a5395fbe0c28cbae1e1c63b05"}
"AC1C14ACXX-1-5" ID: AC1C14ACXX-1-5 {rev: "1-5e713faee027716b2b5e63e3852a4c12"}
"AC1C14ACXX-1-6" ID: AC1C14ACXX-1-6 {rev: "1-d9f1bcc6ccbc9188cd2e4fe1ee0a7fe6"}
"AC1C14ACXX-1-7" ID: AC1C14ACXX-1-7 {rev: "1-e4aa20a0796d0b7b2e35aeb3339512f0"}
"AC1C14ACXX-1-8" ID: AC1C14ACXX-1-8 {rev: "1-2b4ff84ffcaefec145d726cfc4bcef4b"}
"AC1C14ACXX-1-9" ID: AC1C14ACXX-1-9 {rev: "1-c8e0a46393fe4c56fee7c75417de7a5c"}
Showing 1-10 of 2448 rows
← Previous Page | Rows per page: 10 | Next Page →
CouchDB relax
Tools
Overview
Configuration
Replicator
Status
Test Suite
Recent Databases
rp3
rp3_rna_run_20131004
Welcome leon!
Change password or Logout
Futon on Apache CouchDB 1.1.1

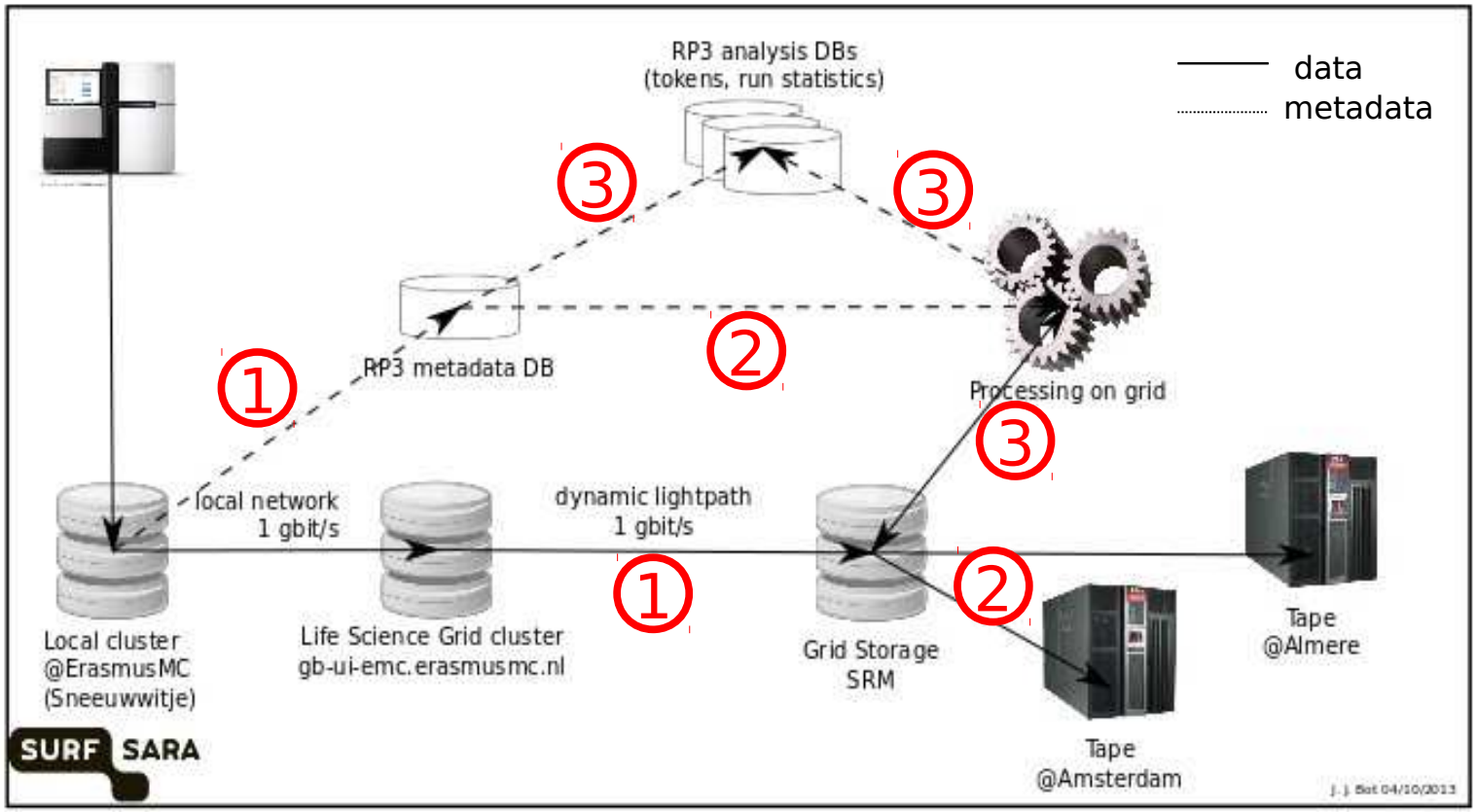

Apache CouchDB - Futon: View Document - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Apache CouchDB - Futon: View...
https://picas.grid.sara.nl:6984/_utils/document.html?rp3_rna_run_20131004/AC1C14
Google
Overview
rp3_rna_run_20131004
AC1C14ACXX-1-1
Save Document
Add Field
Upload Attachment...
Delete Document...
Fields
Source
Field
Value
_id
"AC1C14ACXX-1-1"
_rev
"1-9dc700a4ea9ee037b33d6d01a98241bb"
biobank
"LL"
check_token
lock 1382012232
scrub_count 1
hostname "gb-wn01-rug.sara.usor.nl"
done 1382034221
type "token"
files
0
adler32 "0d554dfe"
srm_path "srm://srm.grid.sara.nl/pnfs/grid.sara.nl/data/bbmri.nl/RP3/RNASeq/AC1C14ACXX-1-1/AC1C14ACXX-1-1_R1.fq.gz"
md5 "f83836ea6c81f176864bb5f96e377f59"
1
freeze
1
person_id
"103001510139"
runs
5ddcae79dd22171e687b42e43c3452cc9a8ae563
← Previous Version | Next Version →
CouchDB
relax
Tools
Overview
Configuration
Replicator
Status
Test Suite
Recent Databases
rp3
rp3_rna_run_20131004
Welcome leon!
Change password or Logout

PiCaS – use CouchDB as token pool server



Picture taken from SARA wiki

Data management workflow



#1 data upload #2 verification, backup #3 process runs

Scale up the computation

- Getting it to work is not easy, but worthwhile!
- Production run summary (2300 RNA-seq samples)
  - More than 300 pilot jobs submitted in parallel, across the whole country
  - Every pilot job takes 10 cores and 40Gb RAM for 36 hours
  - Processing one sample at a time
  - Analysis finished within 4 days, using over 150 000 core hours!

Grid best practice #1

- Grid is great on running a production pipeline
  - Same setting, many many samples
- Programming for Grid is not easy
  - Close collaboration with Grid experts (Jan Bot)
  - Team or pair programming
- Debug on Grid is not easy
  - Create sufficient logging, progress flags, environment variables, library paths, etc.
  - Grid monitoring, local settings may trick you

Grid best practice #2

- Start testing on Grid as early as possible in your project
  - Running successfully on UI machine can not guarantee things will run on Grid clusters
- Data staging, e.g., use a local copy of frequently accessed files (reference genomes).
- Move files on SRM is not trivial, so think carefully on your file naming and directory structure and do that together with your power users!

Grid best practice #3

- Think before you start, data management
- Data administration
  - Sample and file naming, accommodate sample swaps and redoes
  - Database schema, reliably linking all various data and outputs
- Data validation
  - Integrity checks (md5, adler32)
  - Backups (database, data)

Where to go from here

- GRID – Bulk analysis
  - Alignments and quality control
- HPC Cloud – Downstream analysis
  - Less computationally intensive tasks
  - Access to the GRID generated output
  - Familiar interface
- VM example configuration:
  - 8 cores
  - 64Gb RAM

Recap

- Now you have seen :
  - How the Dutch Life Science GRID was used in the perspective and experience of a BBMRI RNA-seq integrative omics study
  - How the GRID is structured
  - How jobs are managed and executed
  - Some of challenges we faced and what we have learned from them
- Hopefully this showcase gave you an idea of how this project made use of the GRID

Acknowledgement

- LUMC: Bas Heijmans, Peter-Bram 't Hoen, Rene Luijk
- UMCG: Lude Franke, Dasha Zhernakova, Patrick Deelen, Pieter Neerincx
- AMC: Rick Jansen
- ErasmusMC: Aaron Isaacs, Joyce van Meurs
- SURFsara Grid support team