Hands-On utilization of OWLer

#OWI onboarding

Who are we?

Michael Dinzinger

Saber Zerhoudi

Jelena Mitrović

Michael Granitzer

Open WebSearch .eu

OWler Open Web Crawler

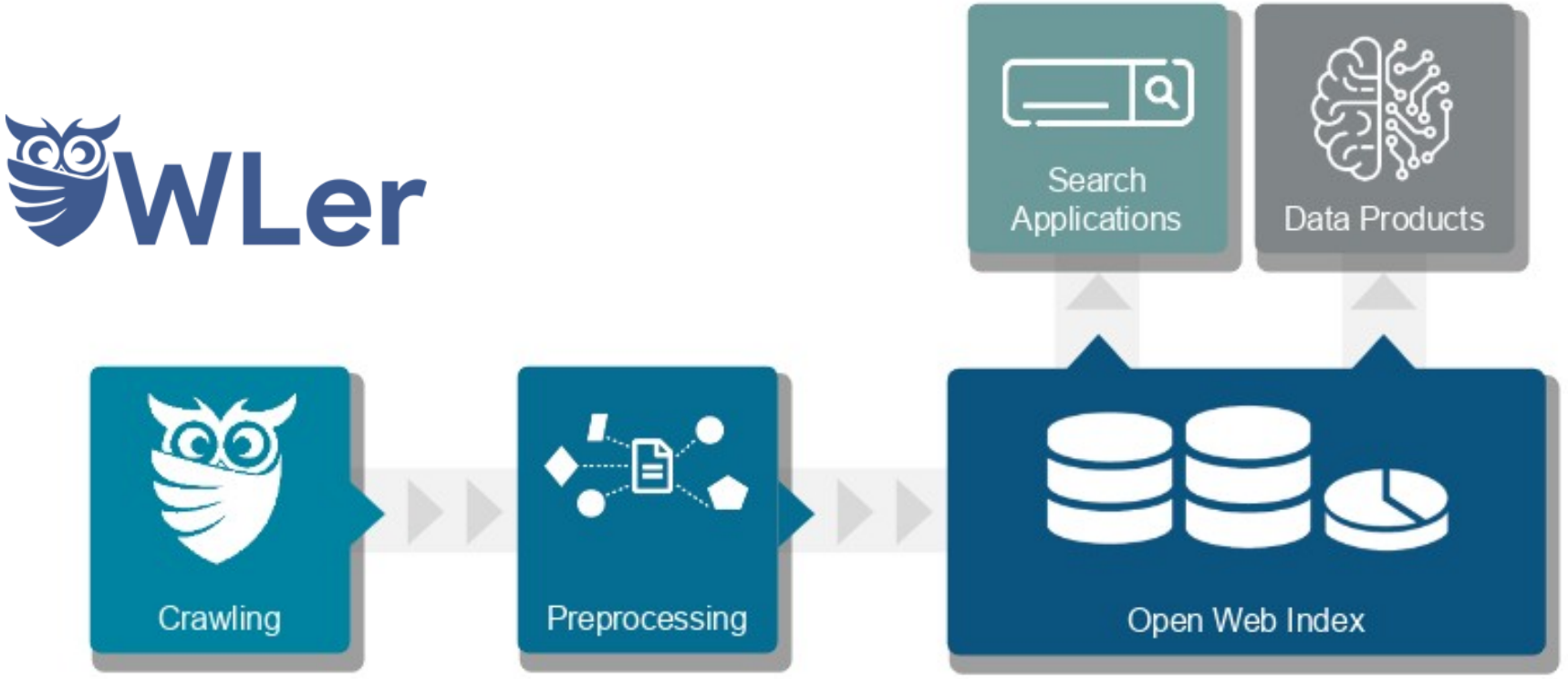
UNIVERSITÄT PASSAU

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

CSC ICT Solutions for Brilliant Minds

lrz CERN DLR

Technical pipeline of OpenWebSearch.EU



Insights on the Ongoing Crawling

18 VMs + 2 Servers
VM w/ 6-10 vCPU cores & 16-45 GB RAM

Up to 100 Million daily visits
w/ 89.7% successful fetches

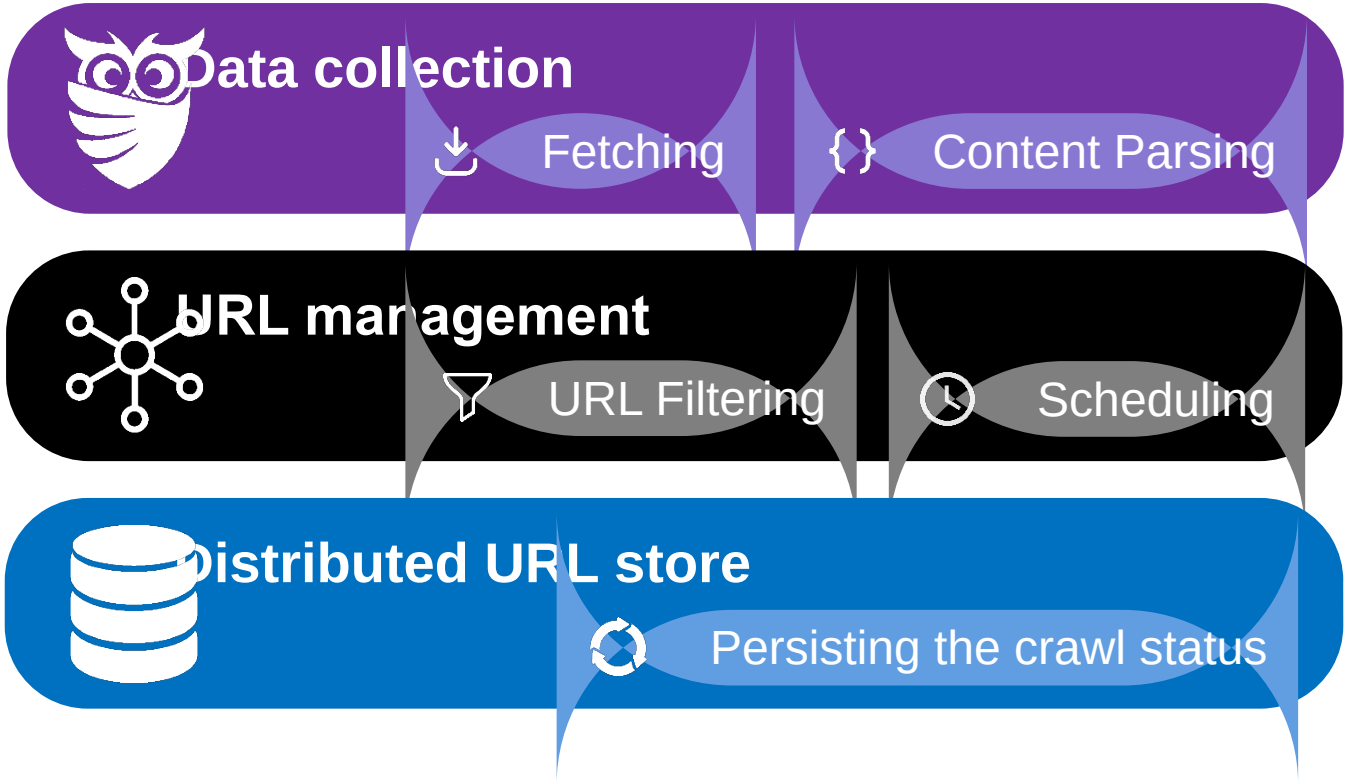
Up to 3.5 TiB WARC data per day
uploaded to three S3 buckets

On avg. 5 links are added to crawl
randomly sampled

Crawl space grows to 1.2 Billion URLs
in the first 5 days of the crawl

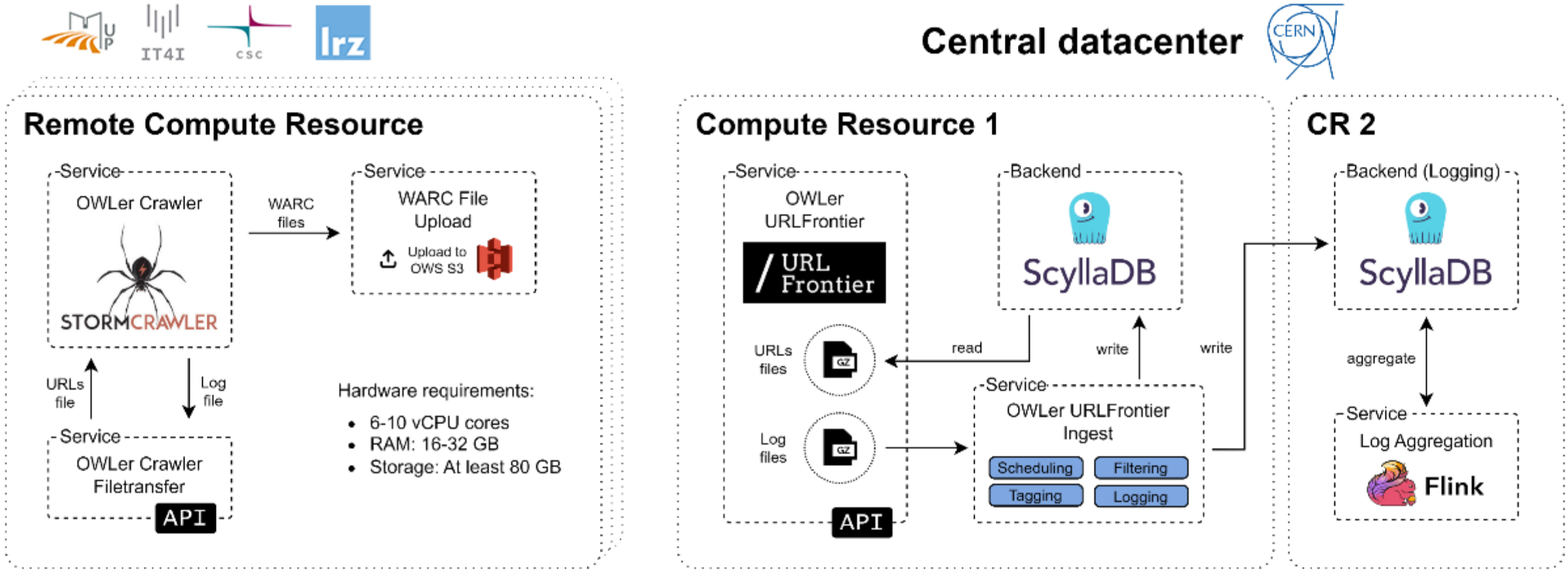
The Open Web Crawler

System for general-purpose crawling:



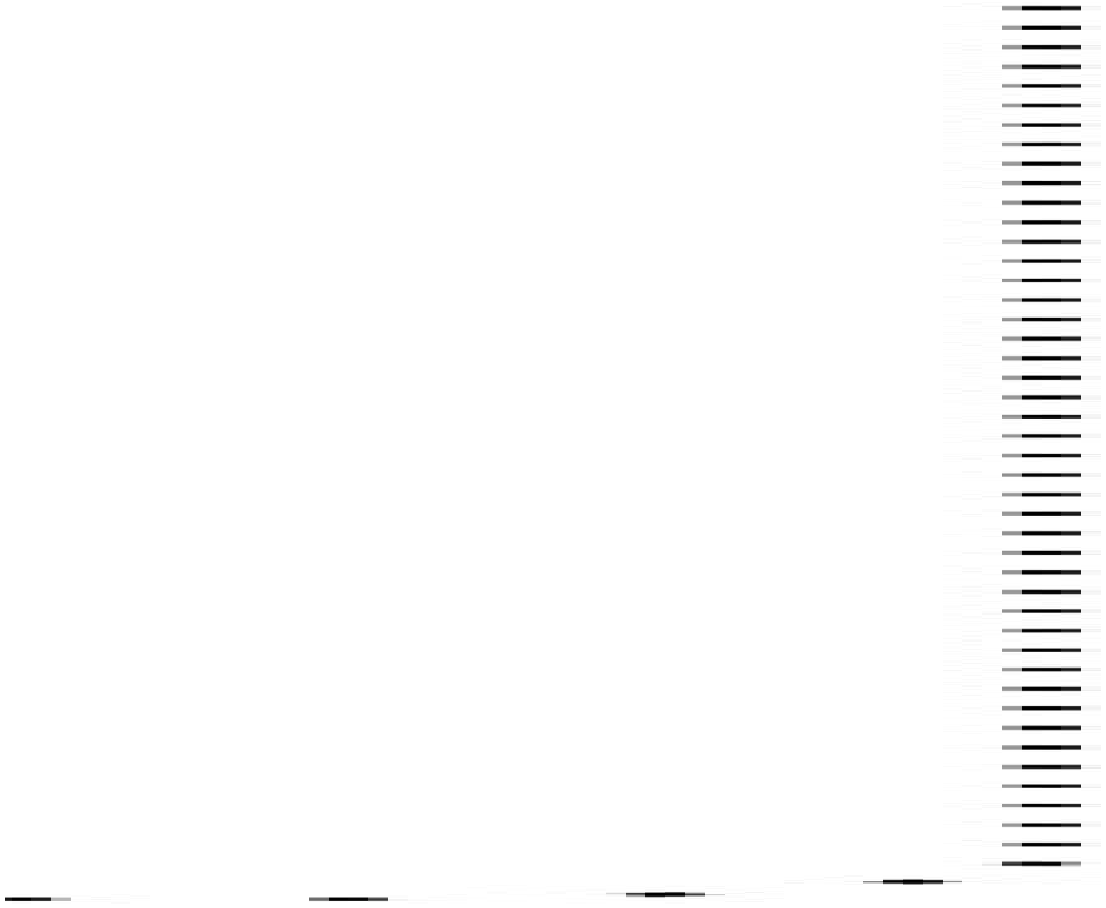
Scraping of specific web platforms:

System Overview



System Overview

System Overview



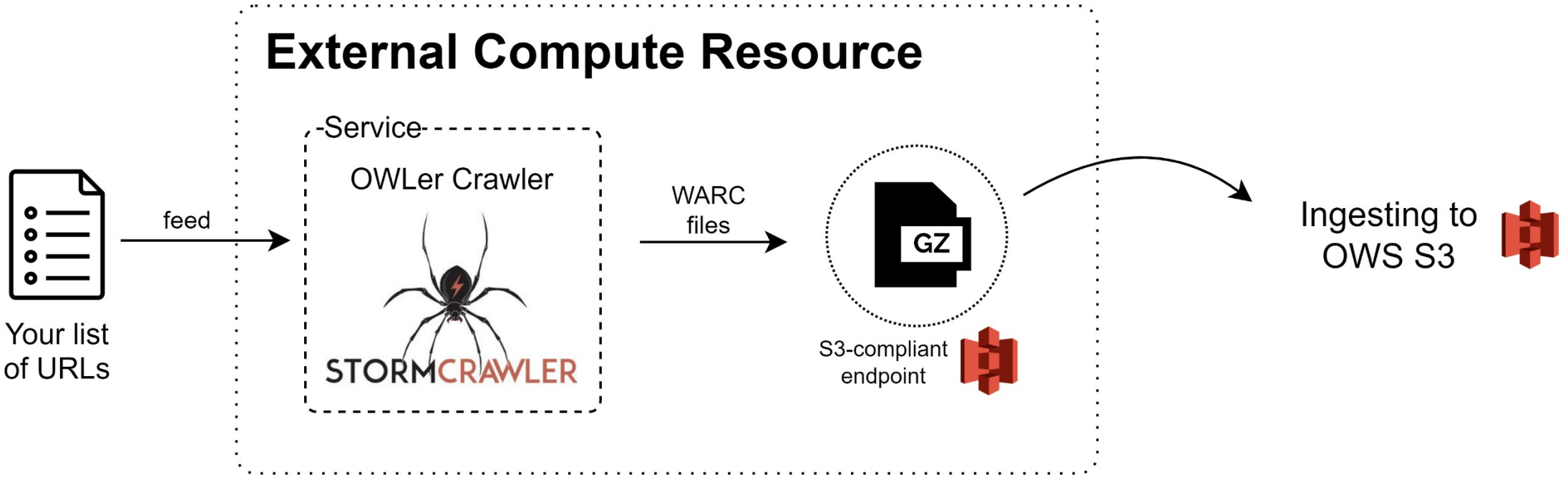
On-Boarding:
You can join the OWS crawling efforts!

Open WebSearch .eu

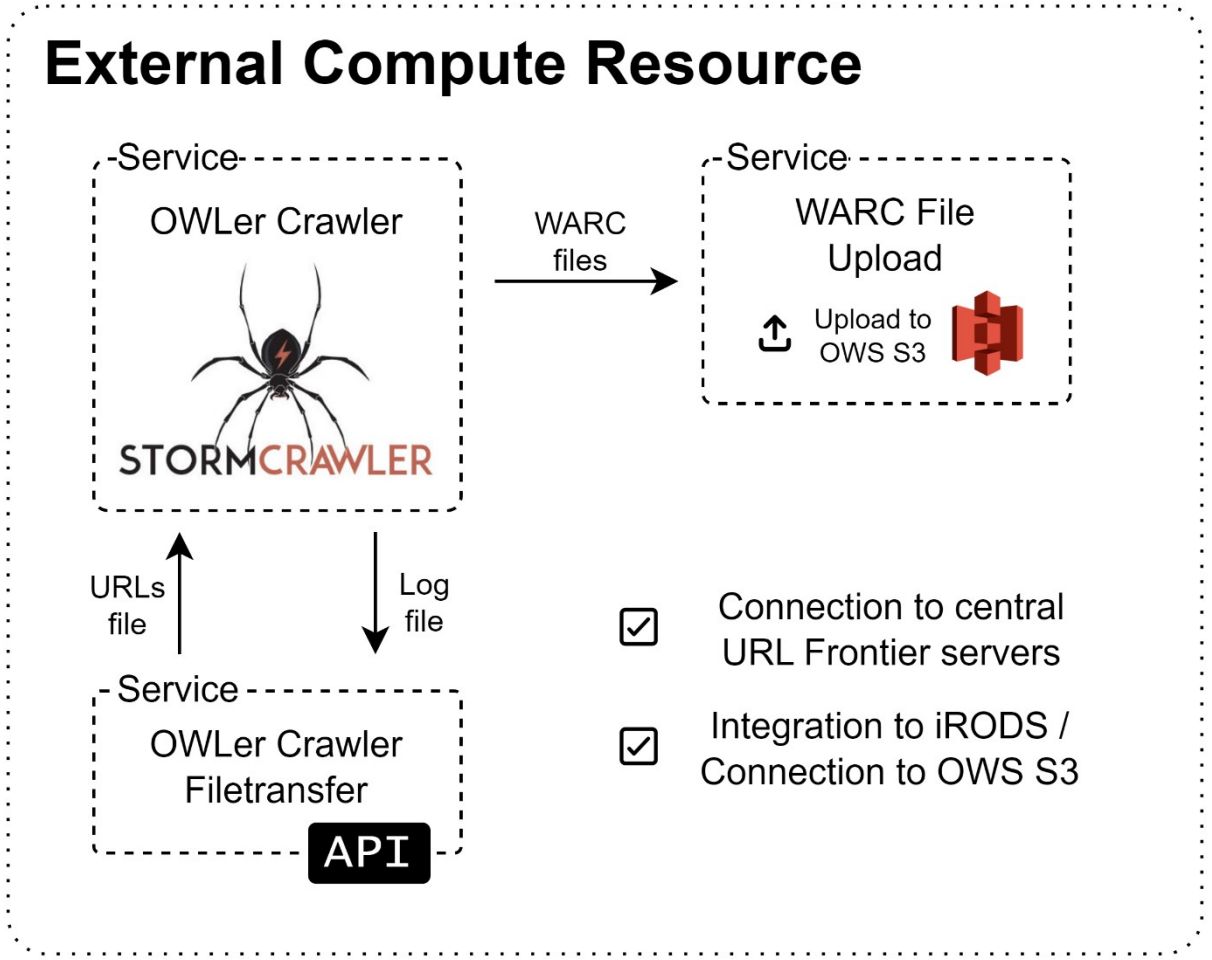
Use case 1:
„Sponsoring the WARC files of your crawl“

Use case 2:
„Providing computing nodes for crawling“

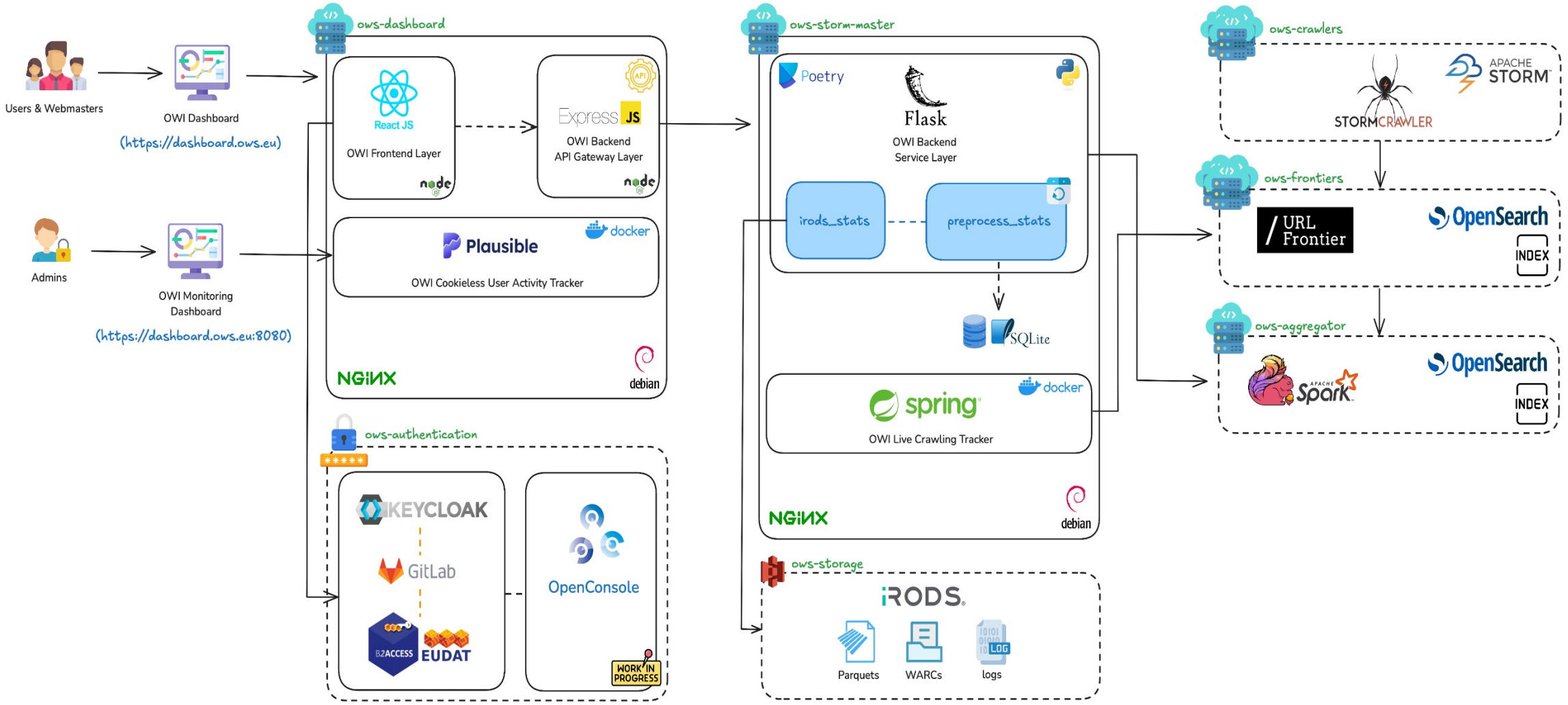
Use case 1:
„Sponsoring the WARC files of your crawl“



Use case 2:
„Providing computing nodes for crawling“



OWLer Dashboard



Thank you for listening!

Questions?

Website: ows.eu
OWLer Dashboard: dashboard.ows.eu
Code: opencode.it4i.eu/openwebsearcheu-public

StormCrawler: stormcrawler.apache.org
URLFrontier: urlfrontier.net