IEEE GRSS HDCRS Working Group

High Performance and Disruptive Computing in Remote Sensing Summer School

29 May - 1 June, 2023, Reykjavik, Iceland

Lowering the Barrier for Modern Cloud-based Geospatial Big Data Analysis by Combined Use of Innovative and Traditional Infrastructure

Dr. Ing. Serkan Girgin[1,2]

s.girgin@utwente.nl

[1] Center of Expertise on Big Geodata Science

[2] Department of Geo-information Processing

Faculty of Geo-information Science and Earth Observation

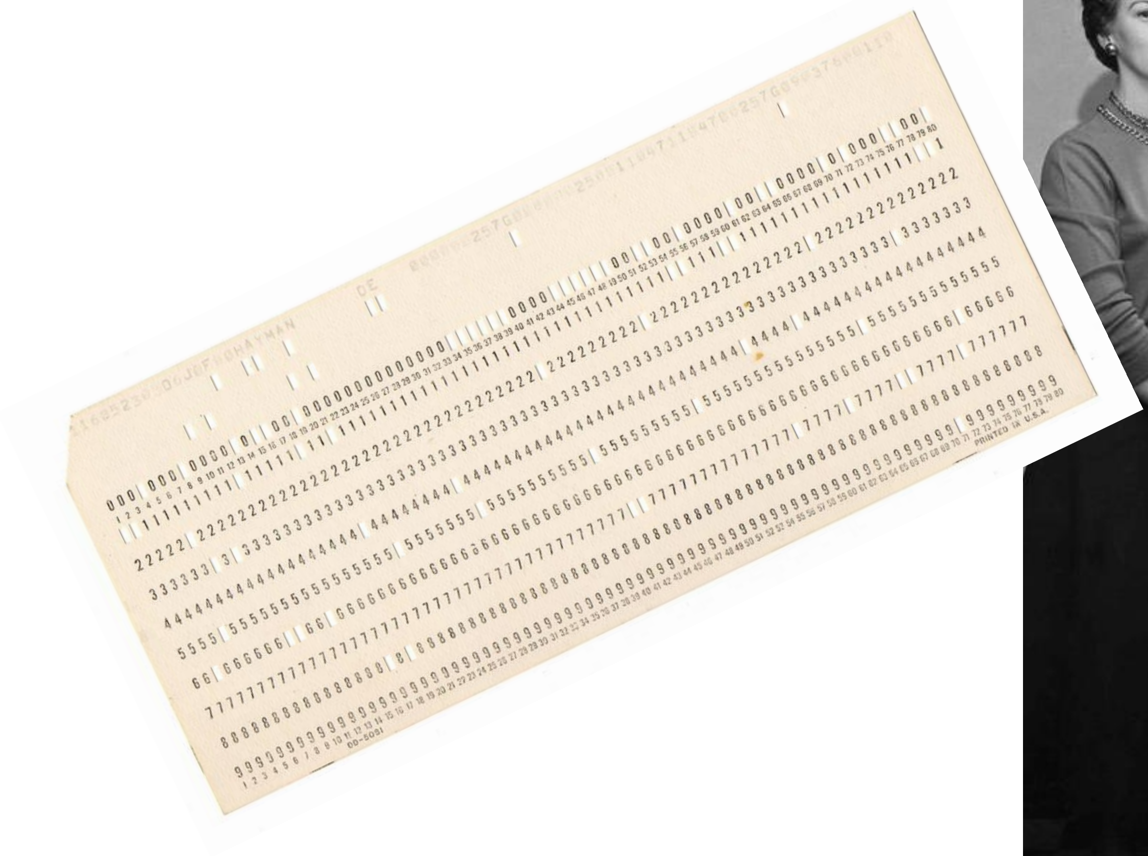
University of Twente, The Netherlands

UNIVERSITY OF TWENTE. | ITC

Big data is large and complex data that is difficult to process

(can also be defined with many V-words)

Being large, complex, or difficult are relative



ComputerHope.com

Being large, complex, or difficult are relative

Being large, complex, or difficult are relative

Amount of Data

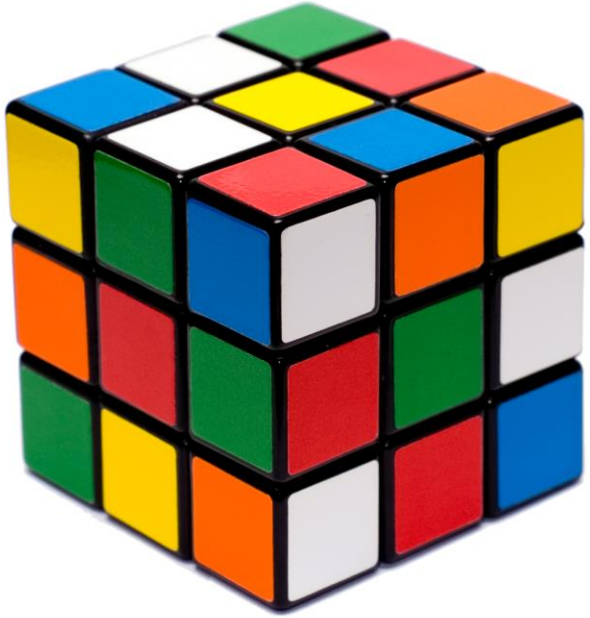
16GB

Transfer Rate

1000 Mbps

Total Transfer Time **0 Days, 0 Hours, and 2 Minutes, 17 Seconds**

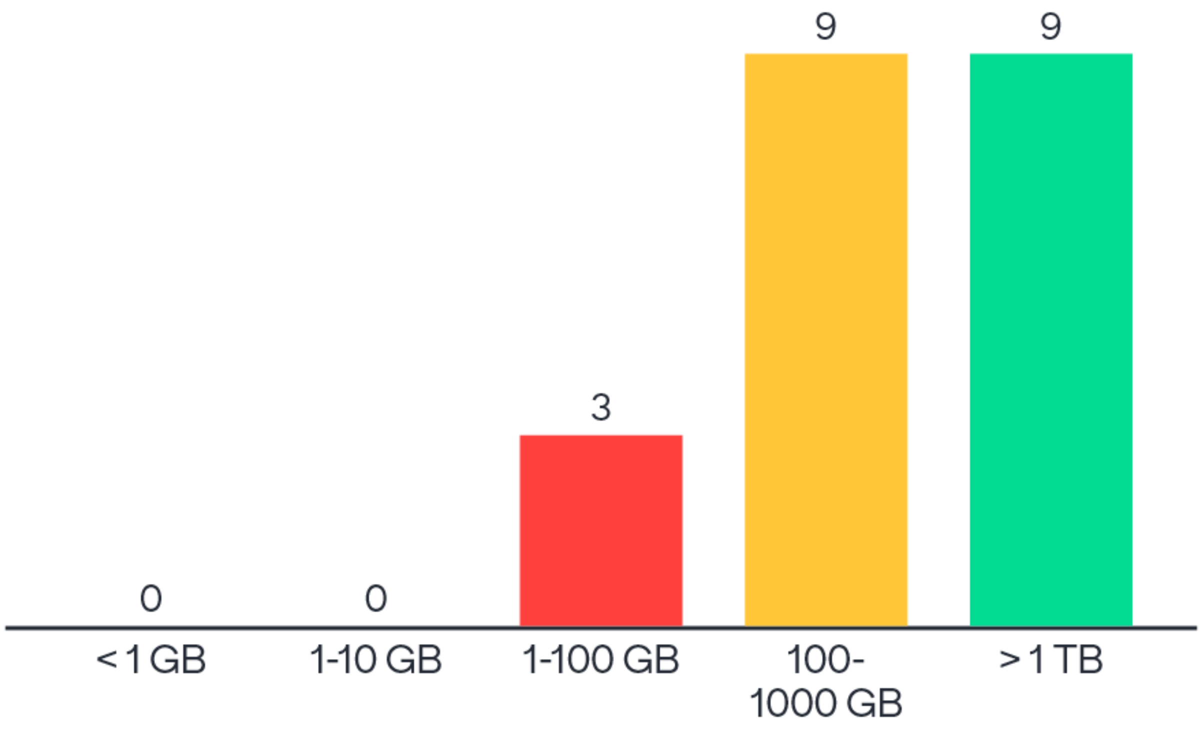
Being large, complex, or difficult are relative

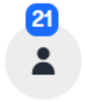


Complex & difficult
for me

Simple & easy
for Ruihang Xu

What is the size of the largest data you processed?





What are the file formats of large data you process frequently?

22

What are the complex data you need to process?

- High dimensional
- Data cubes, time series
- Large datasets in a short amount of time. Automating workflows
- Multiple tiles/time series
- Time series
- high dimensional
- SAR and Hyperspectral
- Multi-modal
- Post processing how to aggregate and get the aggregated results
- Computing power
- Geospatial
- Hyperspectral
- Heterogeneous data with different spatial and temporal resolutions, different reference coordinate systems etc
- DEM
- Time series
- Data-Cubes
- Time Series
- time-series
- Efficient access to the data and scaling
- Complex-valued raw SAR data
- Large dimensions,extent resolution
- optical data
- Multi dimensional
- SAR ...
- The data is simple, the volume is the challenge
- LiDAR point clouds, SAR data
- Time-series
- Automation to make more efficient

22

What are the challenges you encounter while processing such data?

- Time
- Computering power
- Disk space
- Optimisation and processing time
- computing power
- Heteorogeneity in plate-forme, viewing angles, storage, processing time
- Time and computational power
- Lacking ressources and time
- computational cost
- Storage and computing
- Processing speed
- Preprocessing
- Efficient access and scaling
- Unconsistant Data
- Computation and computing power
- Loading to memory
- Optimization, fast calcluation
- Quality
- storage
- Supporting operators and packages
- Tweaking the process and waiting on computation again

20

Solutions require expert know-how and infrastructure

- Local and regional studies with **medium size data**

  Analyses can be done faster by **parallel computing** on a workstation

- **Machine learning and AI** studies with medium size data

  Analyses require **special processing units** (e.g., GPU/TPU) due to computational complexity

- National, continental, and global studies with **big data**

  Analyses require **distributed computing** on a computing cluster due to computational complexity and/or large volume of data

Illustration by Storyset

Cloud computing is on-demand availability of computer system resources, especially **data storage** and **computing power**, without direct active management by the user

Cloud computing has a few distinctive features

- **On-demand self-service**: provision of computing capabilities as needed without requiring human interaction.
- **Broad network access**: availability over the Internet with standard access mechanisms for different client platforms (e.g., tablets, laptops, mobile phones).
- **Resource pooling**: dynamic assignment and reassignment of physical and virtual resources according to consumer demand.
- **Rapid elasticity**: capability to scale rapidly outward and inward proportionate to consumer demand.
- **Measured service**: accurate monitoring, control, and reporting of resource and service utilization.

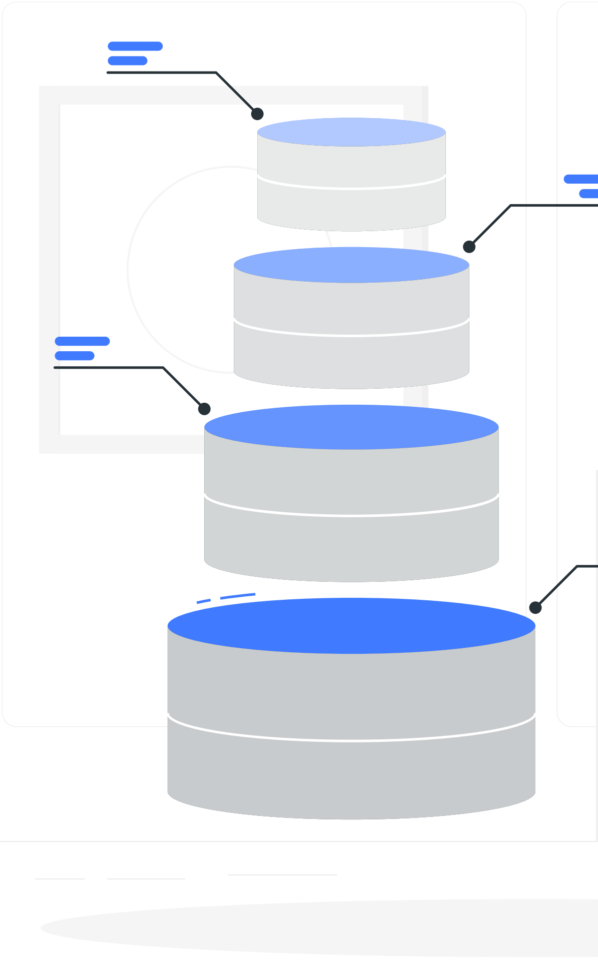


Illustration by Storyset

Computing is moving to the Cloud, so is geocomputing

- Developments in infrastructure, both hardware and software, gave a **big push** to data processing and analysis capabilities.

- **Scalable and affordable** computing is available through:
  - Open-source systems that allow computing clusters on commodity hardware
  - Proprietary cloud-based data storage and computing services
  - More accessible research ICT infrastructure and research cloud

- Using the solutions usually requires a transition in **modus operandi**.

- But challenges exists in **identifying the cases** where cloud computing can play a role and in **proper selection and efficient use** of cloud computing methods, tools, and services.

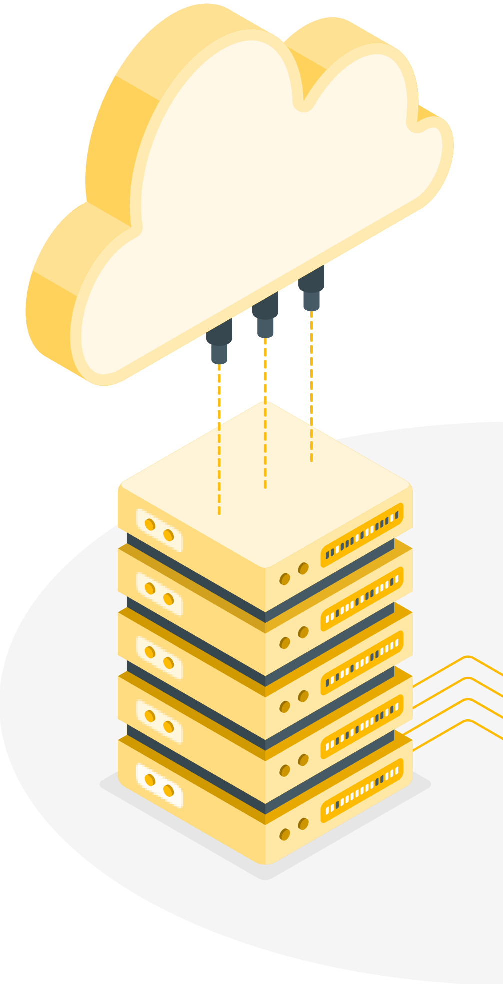
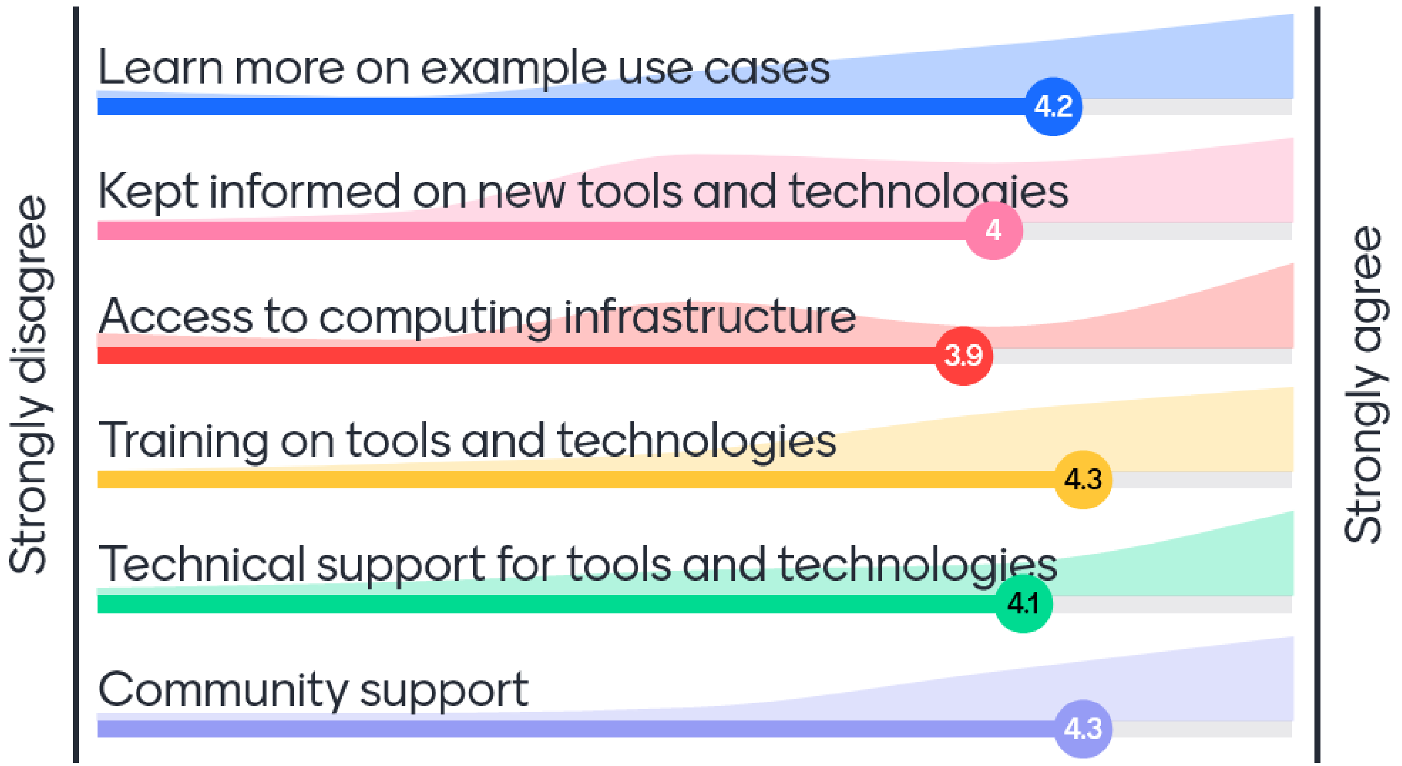


Illustration by Storyset

What are your needs for better use of big data and cloud computing technologies?

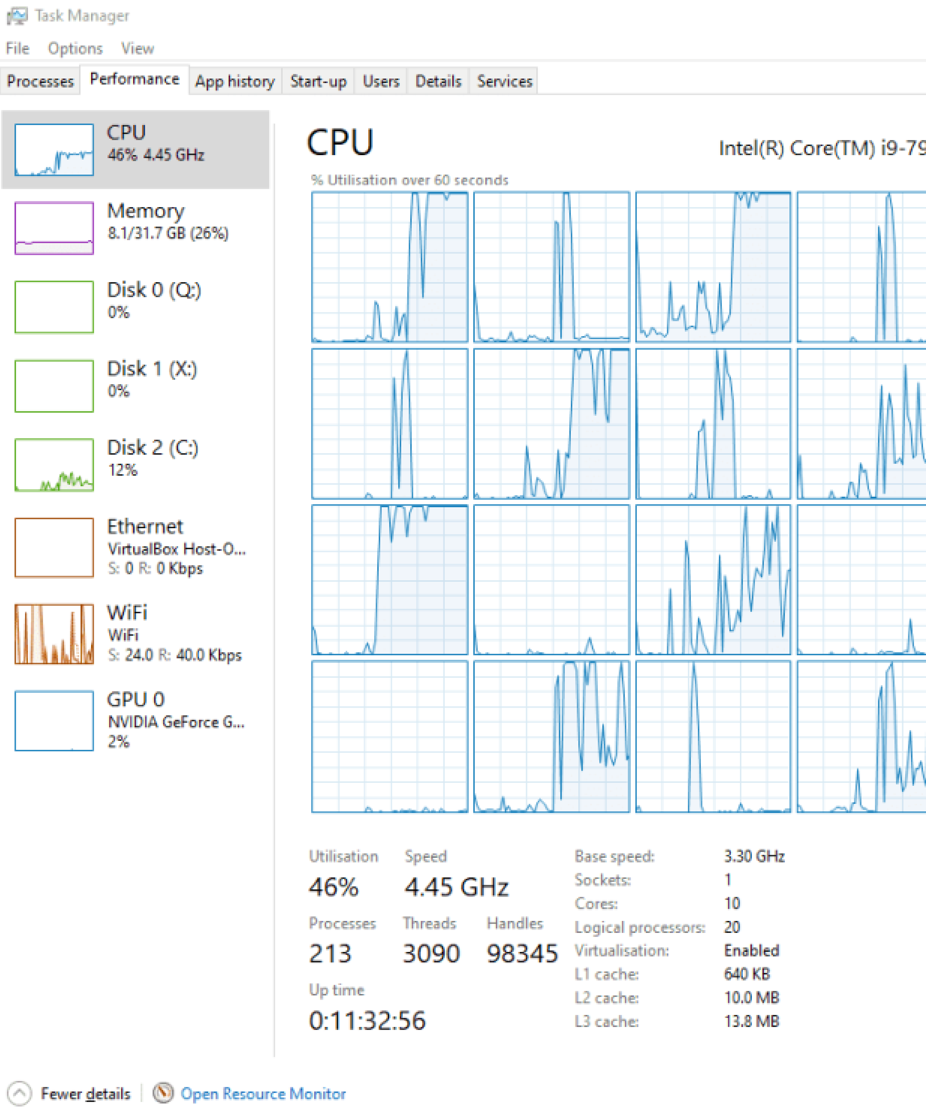


Suggestion One

Inspect what's going on!

Monitoring resource utilization helps to identify bottlenecks

- Geospatial workflows usually involve **many** software components, **including yours**.
- **Not all** software components are able to utilize available **modern computing capabilities** or utilize them efficiently.
- **Monitoring** of the resource utilization is the crucial first step to **understand** the situation.
- **Scaling without efficient use** of resources is suboptimal and costly.

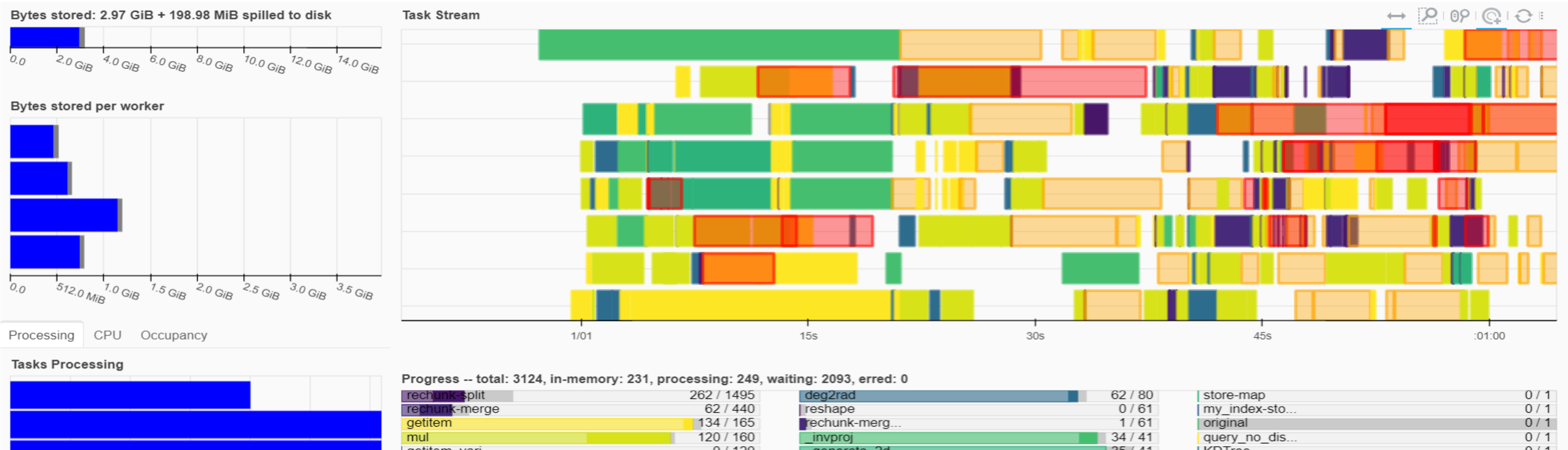


Suggestion Two

Read the documentation!

Distributed computing and the Cloud are not magical tools

- Recent tools and technologies significantly **lowered the barrier** to use cloud computing and parallel/distributed computing capabilities for geospatial workflows.
- However, **time and effort** are required to get competent in using them efficiently.
- **Scaling blindly** is suboptimal and costly.



Suggestion Three

Use the right services!

Cloud computing comes with many ..aaSs!

- Infrastructure as a Service (**IaaS**)
  - Provider supplies the infrastructure that needs to be set up by the user.
    e.g. Amazon AWS, Microsoft Azure, Google Cloud, ESA DIASs, National Research Clouds
- Platform as a Service (**PaaS**)
  - Provider supplies the platform that allows the user to deploy applications (scripts).
    e.g. Google Earth Engine, Microsoft Planetary Computer, Google Colab
- Software as a Service (**SaaS**)
- Function as a service (**FaaS**)
- Data Processing as a service (**DPaaS**)
- Data as a service (**DaaS**)
- …

Illustration by Storyset

There are also many cloud service providers!

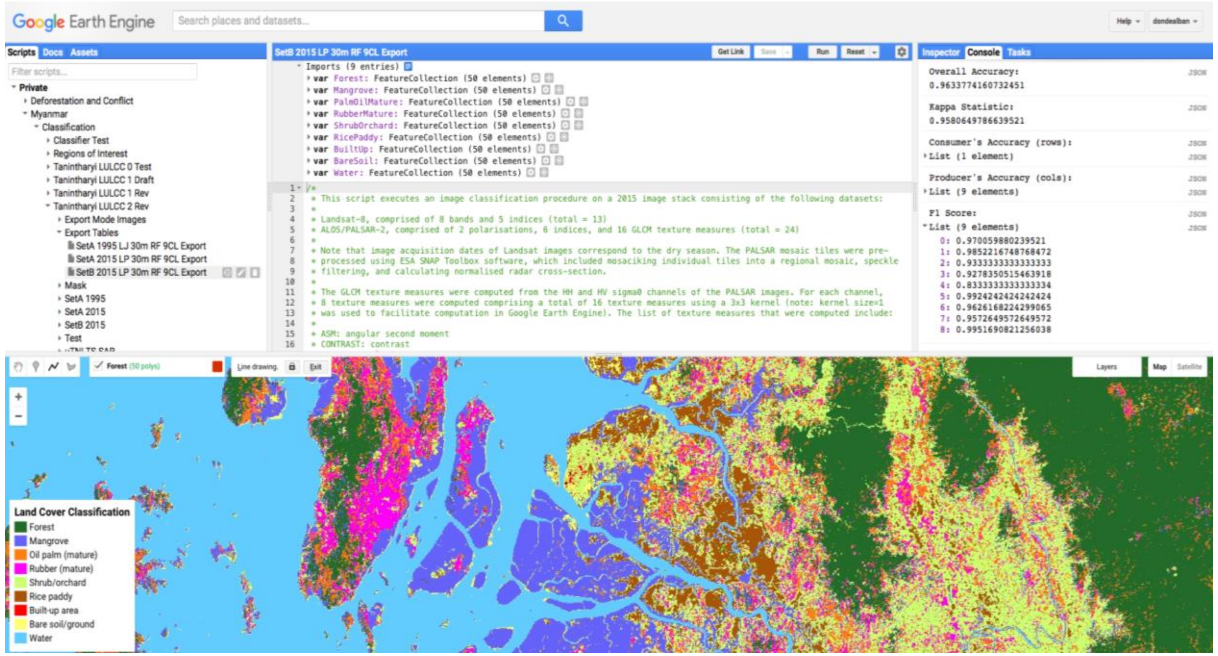
Azure   aws   Google Cloud

- Common features
  - Virtual machines
  - Cloud storage
  - Open-source software
  - Open datasets
- Different features
  - Azure Machine Learning Platform
    Cloud-based environment to train, deploy, automate, manage ML models
  - Azure Data Science Virtual Machines
    Geo AI Data Science VM with ArcGIS
  - EMR Cloud-native Big Data Platform
    EC2 + S3 clusters without provisioning, with OSS (Hadoop, Spark, etc.)
  - Google Compute Engine
    Cloud TPU (eg. ResNet-50, 90 ep.: 8 V100 GPU: 216 min, Cloud TPU V2: 7.9 min)
  - BigQuery
    BigQuery ML: create and execute ML models using standard SQL
    BigQuery GIS: analyze and visualize geodata by using standard SQL

Google Earth Engine is a gamechanger for geospatial computing

Combination of a multi-petabyte catalog of EO imagery and geospatial datasets with planetary-scale analysis capabilities available for free*.

https://earthengine.google.com



openEO develops an open API to connect R, Python, JavaScript and other clients to big Earth observation cloud back-ends in a simple and unified way.

https://openeo.org

Geocomputing on local cloud can be efficient and cost effective

- **ITC Geospatial Computing Platform** provides **GPU-enabled** general purpose (8 vCPU, 32 GB RAM) and **big data** (72 vCPU, 768 GB RAM) units with large storage, analysis ready data, ready-to-use interactive and desktop software (1500+ packages), and **shared workspaces**.
- Currently serves **550+ registered users**.
- Provided **250,000+ hours** of computing since 2021.
- Already returned **16+ times** the investment costs.
- Monthly cost is **< 200 Euro**.

https://crib.utwente.nl

Geospatial Computing Platform is designed to serve the needs of the user community*

- Designed for the primary activities:
  - Self learning
  - Exploratory research
  - Education

- Designed based on the primary criteria:

• Highly available	24/7, no queue
• Ready to use	Pre-installed software
• User friendly	Interactive user interface
• GPU enabled	GPU for each user
• Distributed-computing friendly	Computing cluster
• Low-cost	Feasible investment

Illustration by Storyset

* Girgin, S. (2020) Big Geodata at ITC: Status Quo and Roadmap

We utilized innovative solutions to develop a platform fulfilling the criteria

NVIDIA Jetson AGX Xavier Cluster

8-core CPU
NVIDIA Carmel ARMv8.2, 2.26GHz

512-core GPU
Volta architecture with 64 Tensor Cores

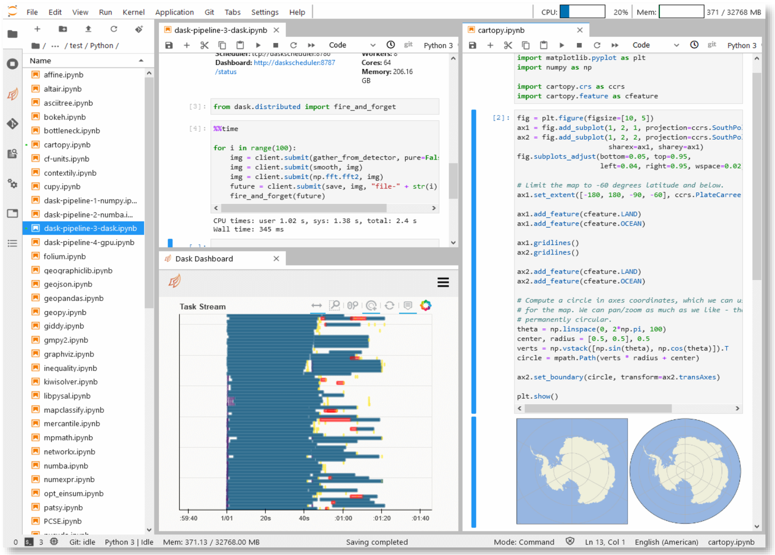
32GB memory
256-bit LPDDR4x, 2133MHz, 137GB/s

32GB internal storage

DL and CV Accelerators

Gigabit Ethernet

ⓘ https://elinux.org/Jetson_AGX_Xavier



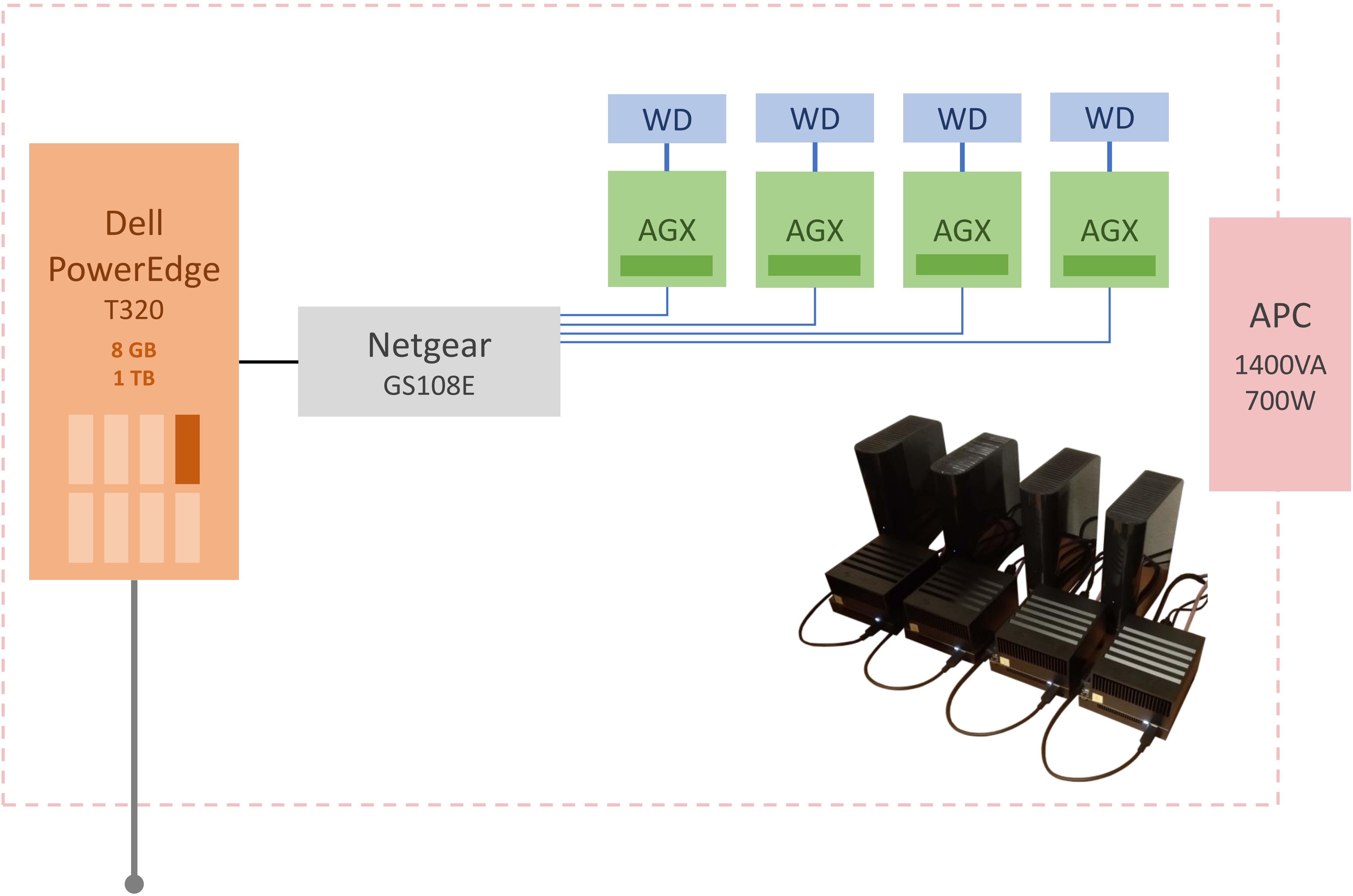
JupyterHub on Docker Swarm

Interactive
Jupyter Notebooks, Desktop Applications

Flexible
Multiple computing units, multiple language kernels

Scalable
Can be deployed with container technology

One can start small…



… but it is important to have an expansion policy

- Grow only if there is **meaningful demand**.
- **Upgrade first**, then expand.
- **Repurpose idle resources** to make them available for common use.
- Select **low-cost**, **good-performance** solutions.
- Use **refurbished** equipment if possible.

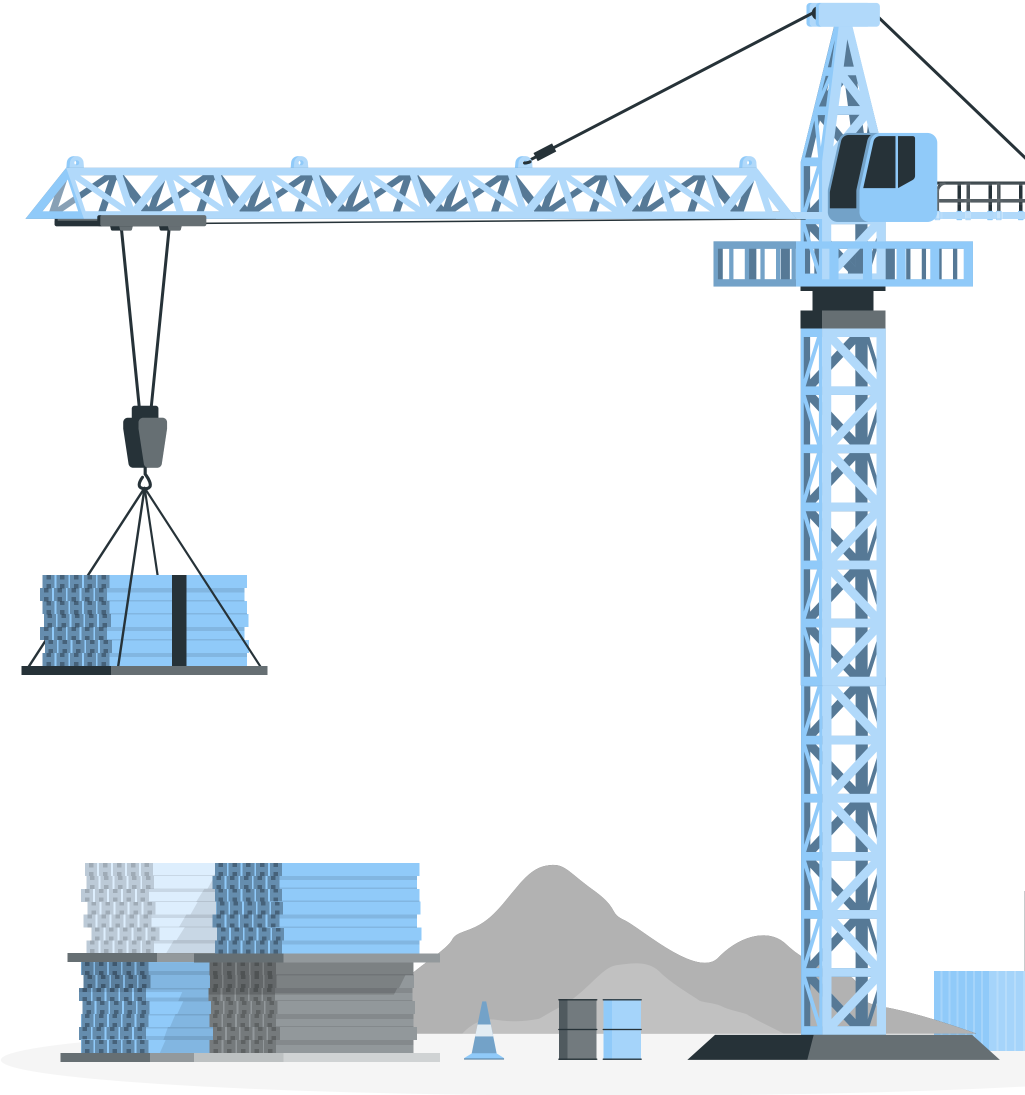
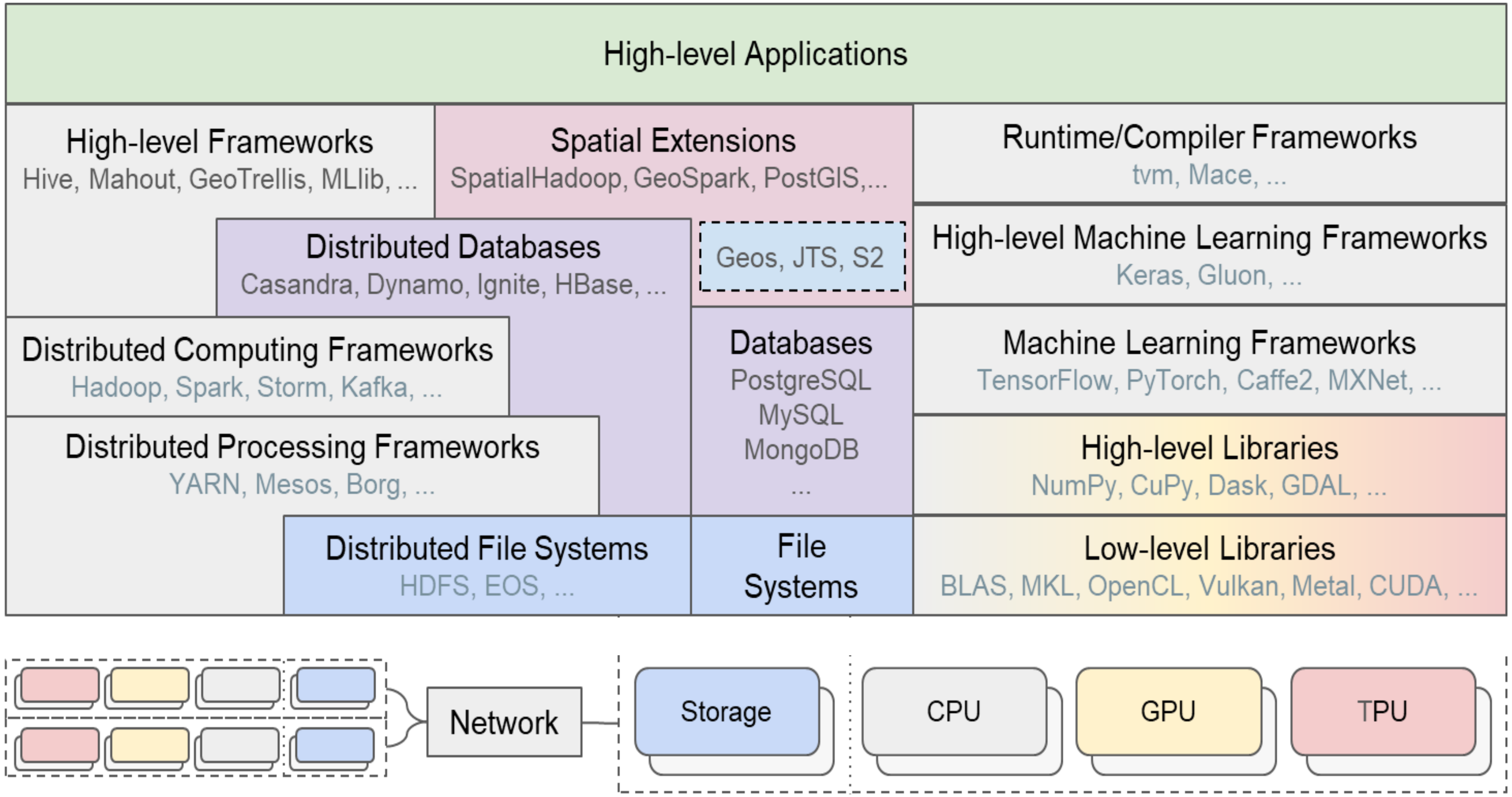


Illustration by Storyset

Today we have an efficient shared infrastructure

- **16 x** GPU-enabled General Purpose Computing Units
  8-CPU NVIDIA Carmel ARMv8.2 @ 2.26GHz, **32 GB RAM**, **512-core Volta GPU**, 64 Tensor Cores
- **6 x** General Purpose Computing Units
  8-CPU Intel Core i7-7700 @ 3.60GHz (max. 4.20GHz), **32 GB RAM**
- **3 x** GPU-enabled Big Data Computing Units
  72-CPU Intel Xeon E5-2695 v4 @ 2.10GHz (max. 3.30GHz), **768 GB RAM, NVIDIA RTX A4000 GPU**
- **1 x** GPU-enabled Big Data Computing Unit with Fast Storage
  32-CPU Intel Xeon E5-2640 v3 @ 2.60GHz (max. 3.40GHz), **768 GB RAM, NVIDIA RTX A4000 GPU, 22 TB RAID 20+2**
- **1 x** Multi-GPU Computing Unit
  32-CPU AMD Ryzen Threadripper PRO 3955WX @ 3.9GHz, 160 GB RAM, **4 x NVIDIA RTX A4000 GPU**
- **1 x** Platform Server
  72-CPU Intel Xeon E5-2695 v4 @ 2.10GHz (max. 3.30GHz), 512 GB RAM, 240 TB RAID1 (ZFS)
- **2 x** Application Servers
  12-CPU Intel Xeon E5-2420 v2 @ 2.20GHz (max. 2.7GHz), 192 GB RAM, 48 TB RAID1 (ZFS)
- **2 x** Storage Servers
  Synology, 60 TB

The platform aims to enable access to a full hardware and software stack for big geodata processing



Resource sharing is at the core of the platform

- Accessible through a **web browser** (No software installation or VPN are required)
- **No registration** is required (Login with the University credentials)
- Each user has an individual and isolated **working environment**
- Each user has access to all available* **unit resources**, including **GPU**
- Each user has access to all available* **cluster resources**
- **Replicated storage** with minimum two copies (Hardware failure protection, ZFS)
- **Distributed storage** for big data processing (HDFS)
- Automatically **balances workload** among the units

* Resource availability depends on resource usage of other active users.

It provides features to simply research activities

- **Interactive notebook**, **terminal** and **remote desktop** access are available
- Multiple **interactive languages** are supported (Python, R, Julia, Octave, Go, …)
- **Up-to-date** and **optimized** software packages are **ready to use** (No setup required)
- Users can install **additional** packages (e.g., Python, R packages)
- Distributed computing clusters are **ready to use** (Dask, Apache Spark)
- **Public** assets are shared by all users (e.g., OSM, NL 0.5m DTM, TOP10-1000, …)
- **Shared workspaces** allow assets to be shared by selected users
- Access can be granted to **external users**
- **User support** is available
- Provided and maintained at no cost (i.e., free PaaS)

Hundreds of software packages are available ready to use

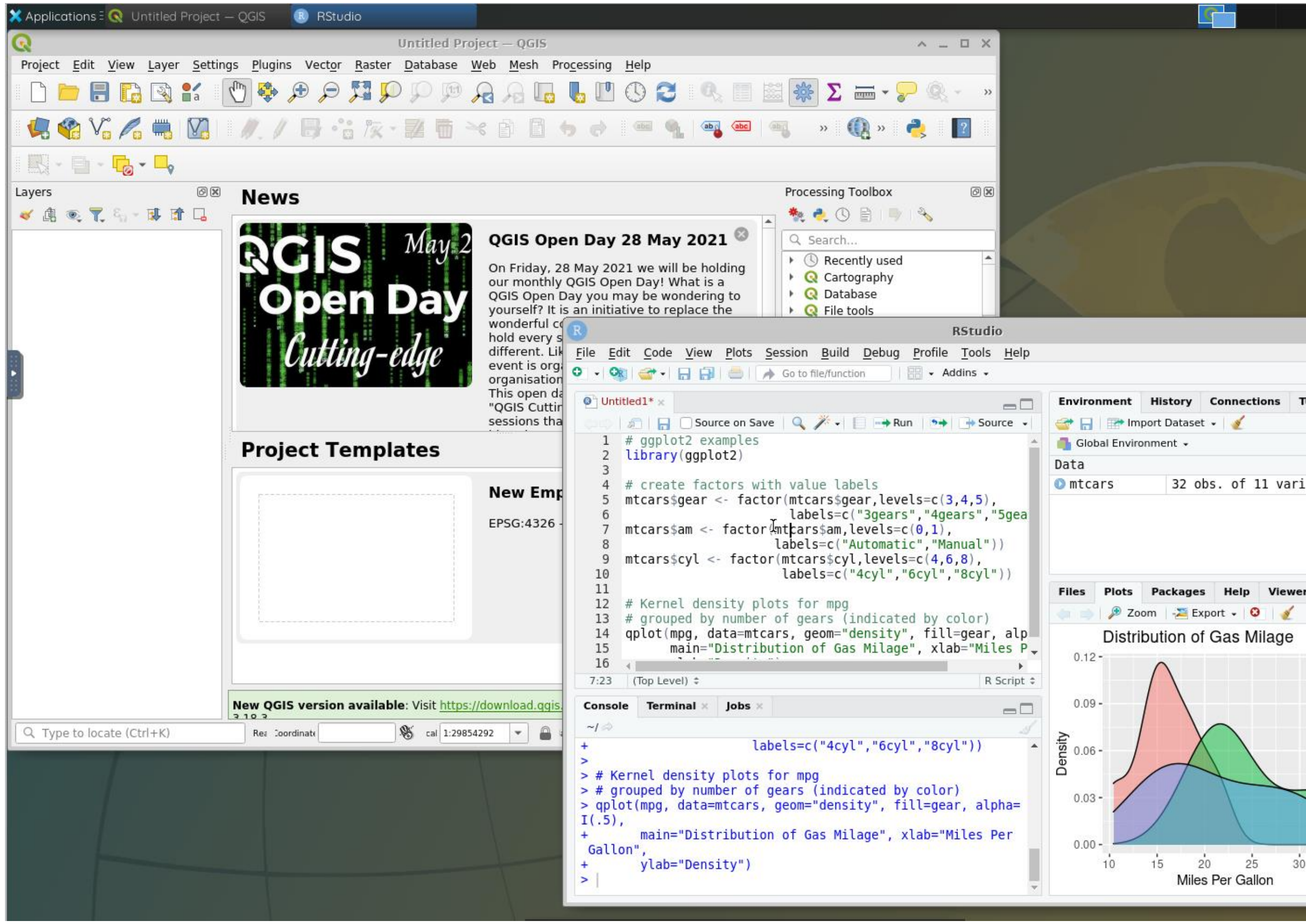
… and many more!

(**950+ Python** and **550+ R** packages)

Desktop applications are also available

- QGIS
- GRASS GIS
- SAGA GIS
- OTB
- ENVI*
- SARScape*
- Metashape*
- SNAP
- **ILWIS 3***
- **ILWIS 4***

- VS Code
- PyCharm
- R Studio
- Netlogo
- GNU Octave
- MATLAB*
- Glueviz
- Orange Data Mining



* Windows application, available only on Intel units through emulation.

* Licensed software, available only on Intel units through license server.

Prominent geospatial technologies are integrated to the platform

GeoServer
Open source server for sharing geospatial data

MapServer
Open source platform for publishing spatial data

PostgreSQL
Open source relational database

MariaDB
Open source relational database

GeoNode
Open source geospatial content management system

Dataverse
Open source research data repository software

Gitea
Open source lightweight code hosting solution

Open Data Kit
Open source platform to collect data quickly, accurately, offline, and at scale

24/7 user support is available through the Support Center

UNIVERSITY OF TWENTE. | CRIB

Support Center Home | Knowledgebase | Open a New Ticket | Check Ticket Status

Search our knowledge base [Search]

[Open a New Ticket]

[Check Ticket Status]

Welcome to the CRIB Support Center!

In order to streamline support requests and better serve you, we utilize a support ticket system. Every support request is assigned a unique ticket number which you can use to track the progress and responses online. For your reference we provide complete archives and history of all your support requests.

Quick Access

- Report a Problem
- Shared Workspace Request
- Course Workspace Registration with Canvas Integration
- External Account Request
- Account Removal Request
- Account Transfer Request
- Software Request
- Dataset Request
- Database Request

Featured Questions

How can I access to the platform?

Is it secure?

How can I use the platform?

Which programming languages are supported on the platform?

Which libraries and packages are supported by the platform?

https://support.crib.utwente.nl/

The platform is maintained and software are updated regularly

- On-demand and bi-monthly regular **rolling updates.**
- **Similar** working environment for ARM64 and Intel x86-64 units.
- Automated shared workspaces for the **departments.**
- Automated shared workspaces for **courses.**
- Automated **notification** to newcomers.
- **Daily** storage snapshots for 7 days.
- **Bi-daily** check for malicious threads .
- **Continuous** resource and performance monitoring.

The user community of the platform includes hundreds of researchers, students, and alumni

- Operational since **January 2021**
- **555*** registered users (max. ~865)
- **106*** shared workspaces for projects and courses
- **15-50*** concurrent users at a time
- **250,000+*** hours of multi-core/GPU computation
- **460+** support tickets* closed (excluding support by e-mail)
- Overall **positive feedback** from a wide-range of use cases
  4.61/5.00 according to the user survey
- Central ICT built a similar platform for **university-wide use**
  Co-developed by CRIB

Illustration by Storyset

* As of 29 May 2023

Existing and potential use cases cover a wide range of activities

- Education
  - Computation platform for **courses**
    Shared course workspaces can be created easily with Canvas integration
- Research
  - **M.Sc.** / Ph.D. thesis studies
  - Collaborative (big) data analysis and visualization
  - Strengthen **project proposals**
    Small projects can use the platform and reduce their budget needs
- Capacity Development
  - **Self-learning**
    e.g., cloud computing, distributed computing, GPU computing, ML, …
  - Computation platform for **activities**
    e.g., training workshops, hackathons, …

Lessons learned: Mixed computing resources work nicely

- Innovative = "unproven" solution
  - ✓ Serving quite well (8 core, 32 GB RAM, GPU, in cluster formation)
  - ✓ High performance/cost ratio (≈ 3 EUR/h VM, pays off in 300 h)
  - ✓ Low energy consumption (max. 30W/unit)
  - ⚠ Difficult to set up (software stack was not ready)
  - ⚠ Difficult to maintain (software need to be build)
- Support from the departments / staff
  - ✓ Donation (servers, storage servers, GPUs)
  - ✓ Donation + shared upgrade (big data unit)
  - ✗ Sharing (Jetsons)
- Low-cost solutions
  - ✓ Second-hand equipment

Lessons learned: Rolling updates do not cause major trouble

- Rolling update = State of the art
  - ✓ 950+ Python, 550+ R packages (Statistical, Spatial, EO, ML, AI)
  - ✓ QGIS, GRASS, SAGA, ILWIS, SNAP, Octave, NetLogo, MATLAB, ENVI, …
- State of the art = Difficult to maintain
  - ⚠ Dependency puzzle (multi-architecture + GPU)
  - ⚠ Stability*
- Docker virtualization / orchestration
  - 😊 Works quite well (multi-architecture)
  - 😔 Image rebuilds take a lot of time (7+ days)
- ✓ Support center (460+ closed tickets)
- 🌻 Additional services

Lessons learned: Building a community is challenging

- Computing Platform
  - ✓ In use for MSc/PhD studies, courses, projects, and trainings.
  - ✓ Overall user satisfaction is quite positive.
  - ⚠ Reaching potential users is not easy.
  - ⚠ Getting useful feedback is difficult.

ID ↑	Name	Responses
1	anonymous	STOP SENDING ME MAILS STOP SENDING ME MAILS

- CRIB
  - ⚠ Collaboration exists, but can be improved.
  - ✓ Automated e-mail to newcomers and newcomer meetings help.

Lessons learned: User stories are difficult to collect

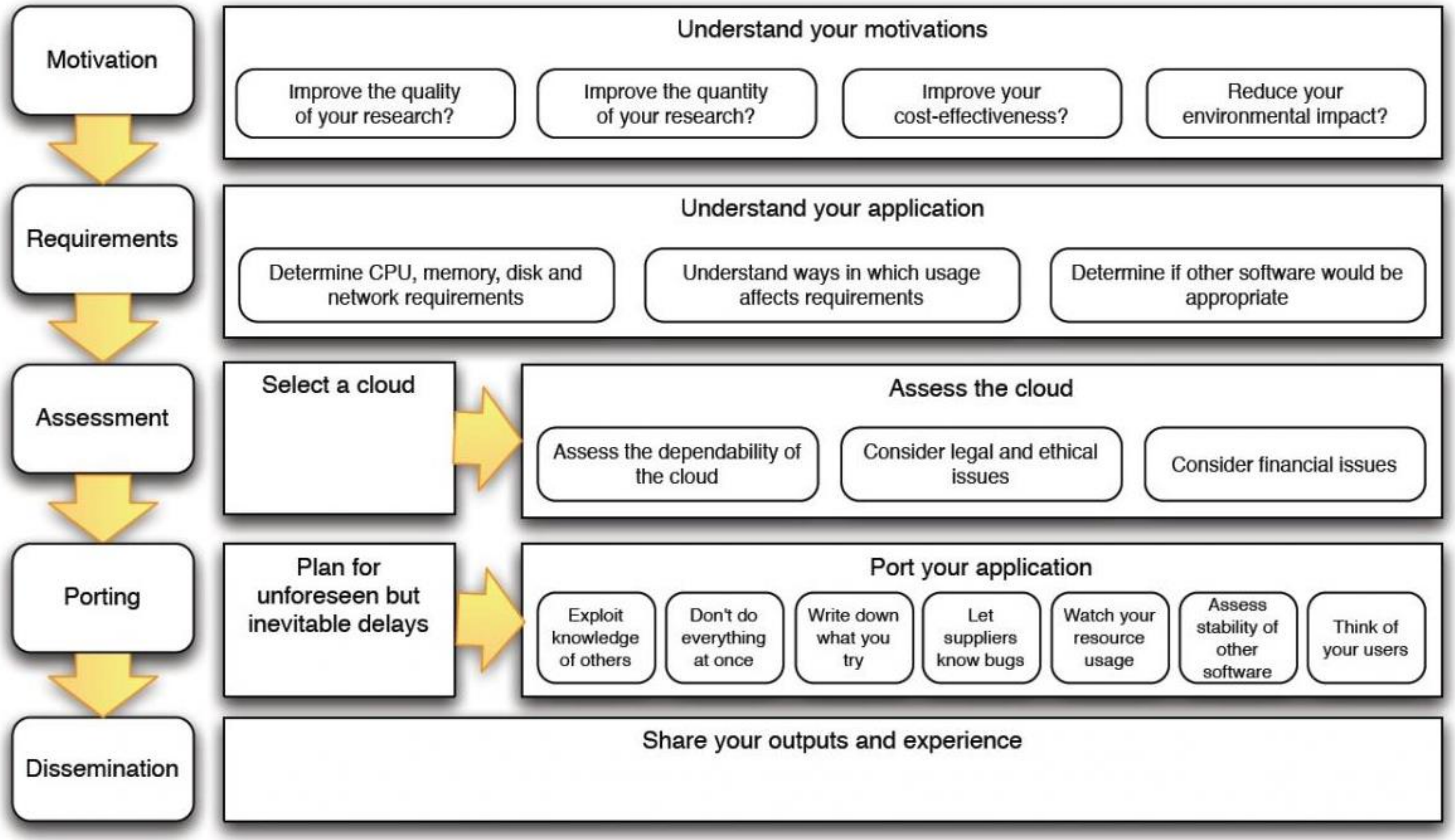
- Best way to convince people is to show examples
- Best examples are in-house from "familiar" faces
- They are not easy to collect

absquatulate (v):
to leave without saying goodbye

A few suggestions for newcomers

- **Ensure familiarity** with the cloud computing technology through short talks and lectures.
- **Improve know-how** by participating tool- and technology-specific training
- **Try and use** the infrastructure and platforms available for free or through partner organizations.
- **Follow** a hybrid approach (local + cloud) to maximize the benefits.
- **Ask for technical and scientific support** for better implementation and integration of the technology.
- **Ask for guidance** for the planning of future activities.
- **Share your knowledge** and good practices with your colleagues (e.g., for cost-effective and efficient use).

Following best practices facilitates moving to the Cloud



Source: Best practice for using cloud in research (Hong et al., 2018)

Follow us to stay informed!

https://itc.nl/big-geodata

crib-itc@utwente.nl

@BigGeodata

487 followers

CRIB YouTube Channel

240 subscribers

Big Geodata Newsletter

259 subscribers

Contact

dr.ing. Serkan Girgin MSc
Head, Center of Expertise in Big Geodata Science
Associate Prof., Dept. of Geoinformation Processing
Faculty ITC, University of Twente
s.girgin@utwente.nl
+31 53 489 55 78
https://linkedin.com/in/serkan-girgin/