

Available Computing Resources for the SDSU Community

ExpandAI@SD

Acknowledgement: SDSU Research Foundation, Nautilus

Contents

Available Computing Resources for the SDSU Community	0
Local Resources in SDSU	1
1. VERNE: JupyterHub for Instruction	1
2. TIDE: Technology Infrastructure for Data Exploration	1
3. Computational Science Research Center (CSRC)	2
4. Discounted Commercial Cloud Services	3
5. Globus.....	3
National Resources	5
1. NSF ACCESS Program https://access-ci.org/	5
2. National AI Research Resource (NAIRR) https://nairrpilot.org/	7
3. CloudBank https://cloudbank.org	8
4. National Research Platform (NRP) https://nationalresearchplatform.org/	10
5. Others to Explore	12

Local Resources in SDSU

1. VERNE: JupyterHub for Instruction

- Available at no cost to SDSU instructors
- Managed JupyterHub instances with SDSUId authentication
- Access granted based on my.SDSU class roster
- Access to CPU and GPU resources
- Python, PyTorch, TensorFlow, R, RStudio, MATLAB, Julia, Ollama and other software images available
- Focus on but not limited to artificial intelligence and machine learning courses

Instructors can create classes and add students to them using a roster. Students can log in with their SDSUId once they are added.

Hardware:

NODE TYPE	QTY.	SPECS
GPU	8	PowerEdge R750XA (2x) Intel Xeon Gold 6338 2G CPU, 32C/64T (4x) Nvidia A100 GPU, 80 GB RAM 512 GB System RAM
CPU	4	PowerEdge R750 (2x) Intel Xeon Gold 6338 2G CPU, 32C/64T 512 GB System RAM
STORAGE	3	PowerEdge R750 (2x) Intel Xeon Gold 6336Y 2.4G, 24C/48T, 160 TB Storage 256 GB System RAM

2. TIDE: Technology Infrastructure for Data Exploration

- Available at no cost to SDSU and CSU researchers
- Managed JupyterHub instance with SDSUId authentication, like VERNE
- Python, PyTorch, TensorFlow, R, RStudio, MATLAB, Julia, Ollama and other software images available
- Access to batch jobs for longer running processes
- Access to storage, CPU, and GPU resources
- Assist with custom images and use of systems
- Focus on artificial intelligence (AI) and machine learning (ML)
- Founded by NSF

Hardware:

NODE TYPE	QTY.	SPECS
GPU	17	Dell PowerEdge R760XA (2x) Intel Xeon Silver 4410Y 2G CPU, 12C/24T (4x) Nvidia L40 GPU, 48 GB RAM 512 GB System RAM
GPU ADVANCED	1	Dell PowerEdge R750XA (2x) Intel Xeon Gold 6338 2G CPU, 32C/64T (4x) Nvidia A100 GPU, 80 GB RAM 512 GB System RAM
CPU	6	Dell PowerEdge R760 (2x) Intel Xeon Gold 6430 2.1G CPU, 32C/64T 768 GB System RAM
STORAGE	3	Dell PowerEdge R760 (2x) Intel Xeon Gold 6442Y 2.6G CPU, 24C/48T 256 GB System RAM 240 TB RAW Capacity

3. Computational Science Research Center (CSRC)

- Offers state-of-the-art computing resources for faculty and students, crafting educational programs, and facilitating industry collaborations
- CSRC has adopted a “condo” model that allows faculty to contribute resources in exchange for central support and access to a more diverse set of resources. Participants contribute resources to a shared infrastructure, much like how condominium owners share ownership and maintenance of common spaces.
- The contributing faculty have a guarantee of 60% usage of the hardware with 40% toward the shared resource pool. Usage of 100% of the contributed resources can be arranged. Usage beyond the contributed resources can also be arranged
- The Computational Sciences program at SDSU provides support for a variety of cluster and Linux SMP systems for use by its students and partner researchers. To schedule a consultation with regard to selection and use of appropriate resources for your computing needs please contact us: Email : jcastillo@sdsu.edu Web: <https://csrc.sdsu.edu/>

Hardware:

NODE TYPE	SPECS
GPU & CPU	Anthill: 40-node Dual-8 Xeon CPU, 640 total cores, 8GB RAM per core. InfiniBand. Batch scheduled.

	<p>Dugong: 7-node Quad-8 AMD CPU. 2GB RAM per core; 13-node Dual-6 Xeon 4GB RAM per core. 380 total cores. Infiniband. Batch scheduled</p> <p>Cinci: 16-node Dual-8 Xeon CPU Section 4GB RAM per core. 4-node Dual-12 Xeon Section 5.33 GB RAM per core. Dual P100 Nvidia GPU server 98 GB RAM. 368 total cores. Infiniband. Batch scheduled</p> <p>COD: 6-node Dual-8 Xeon CPU, 4GB RAM per core; 12-node Dual-10 Xeon, 12.8 GB RAM per core. Dual P100 Nvidia GPU server 128GB RAM. 352 total cores. Infiniband. Batch scheduled</p> <p>Mesxuuyan: 20-node Dual-10 Xeon CPU, 12.8GB RAM per core. 320 total cores, and 3 high-memory nodes. Omnipath. Batch scheduled</p>
CPU	<p>Notos: AI GPU cluster. 8x Tesla V100 GPU. TensorFlow, Caffe, Keras deep learning frameworks</p>

4. Discounted Commercial Cloud Services

- SDSU maintains contracts with Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform with enterprise discounts (e.g., 15% off on AWS)
- Staff certified in Azure and AWS and can assist with architecture and cost estimates
Project responsible for cloud spend. Cost recovery process in place for state and foundation funds on a quarterly basis
- Azure and AWS environments built for NIST 800-171 compliance.

5. Globus

Globus offers researchers across various institutions a sophisticated data management system. It's a cloud service used by hundreds of research institutions and High-Performance Computing facilities worldwide for secure, reliable file transfer, sharing and publication.

Globus Use Cases:

- **Transfer data**, especially useful for transferring large amounts of data. If the transfer gets interrupted, Globus automatically resumes the transfer when components come back online
- Give access to data without giving access to server
- Using Research Automation for data computation and data transfer from sources such as instrument data
- Automate instruments data collection with storage service

SDSU License:

- All data in the SDSU Globus environment is “High Assurance” and BAA compliant, which means that it has encryption, audit logging and other security features enabled that make all file transfers HIPAA compliant
- Globus S3 and Google Drive storage connectors

National Resources

1. NSF ACCESS Program <https://access-ci.org/>

ACCESS is a program established and funded by the National Science Foundation (NSF) to help researchers and educators utilize the nation’s advanced computing systems and services. Almost any computer application that requires more than a desktop or laptop could qualify as needing an advanced computing system. Examples include supercomputer applications, AI and machine learning, big data analysis and storage, and others.

ACCESS is an acronym that stands for “Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support.” NSF awarded \$52 million over five years to five lead institutions and their sub-awardees to facilitate the ACCESS program. You can find out more about how ACCESS is organized here.

These are just a few of the resources available through the ACCESS program (For a complete list see: <https://access-ci.org/resource-providers/>)

RESOURCES	RESOURCE PROVIDER	OVERVIEW	GOOD FOR
Expanse	San Diego Supercomputer Center	728x 2-socket AMD CPU nodes; 52x 4-way V100 GPU nodes; 34x 4-way H100 GPU nodes (coming)	Wide range of S&E research; modest scale computing
Bridges-2	Pittsburgh Supercomputer Center	488x 2-socket AMD CPU nodes; 24x 8-way VI100 GPU nodes	Wide range of S&E research; modest scale computing
Delta	National Center for Supercomputing Applications	132 2-socket AMD CPU nodes; 100 4-way A40 GPU nodes	Wide range of S&E research; modest scale computing
Delta AI	National Center for Supercomputing Applications	456x Grace-Hopper processors	AI and ML focused
Anvil	Purdue University	1000x 2-socket AMD nodes; 16x 4-way A100 nodes	Wide range of S&E research; modest scale computing
Stampede	Texas Advanced Computing Center	560x 2-socket Sapphire Rapids processors	Large-scale parallel applications
Jetstream-2	Indiana University	384x 2-socket AMD; 90x 4-way A100	Cloud-based, on-demand

Simplified Request & Review Framework

- **Explore ACCESS** — for getting started, evaluating resources, dissertations, small-scale Activities; Only requires an abstract, reviewed by RPs for suitability
- **Discover ACCESS** — for modest-scale work, opportunity to request courtesy review of their plans; One-page write-up, reviewed by RPs for suitability
- **Accelerate ACCESS** — for more experienced researchers with mid-scale needs; Three-page proposal, subject to panel and RP review
- **Maximize ACCESS** — for largest-scale projects, continued close scrutiny of most demanding computational work; 10-page proposal subject to panel and RP review

Explore, Discover, and Accelerate requests are made in units of “ACCESS Credits”. A Credit is essentially 1 core-hour on an AMD Rome processor. Once awarded they are exchanged for a specific system allocation. Maximize requests are made in system-specific units. (e.g., Bridges-2 GPU-hours)

Allocations Policies

- No supporting grants required
- U.S.-based investigators are eligible to lead projects
- **Graduate students can now lead projects**
- Separate projects for research, exploration, and classroom activities
- Standardized project types for flexibility. The “paperwork” required to request a project ranges: 1 paragraph; 1 page; 3 pages; 10 pages
- Universal credits that can be exchanged for any available resource
- Award duration aligns with supporting grant
- Available at no cost
- RPs are engaged in each request for their resource(s)

Step-by-Step Allocations Request

Over 95% of Requests are Approved!

- Register for an ACCESS ID
- Select the Project Type that best fits your needs. If you’re new, start with Explore and upgrade when you need more resources
- Complete the Request Form. Add co-PIs, Allocation Managers, and other Users (make sure they have an ACCESS ID)
- Exchange your allocated credits for the Available Resources
- Start your research, development, or educational (classroom) work

Link to full “Get Your First Project” guide: <https://allocations.access-ci.org/get-your-first-project>

2. National AI Research Resource (NAIRR) <https://nairrpilot.org/>

The NAIRR Pilot aims to connect U.S. researchers and educators to computational, data, and training resources needed to advance AI research and research that employs AI. It is like ACCESS, but only accepts **research/classroom projects relevant to AI**.

Resource allocation from commercial, ACCESS, and Federal Agency:

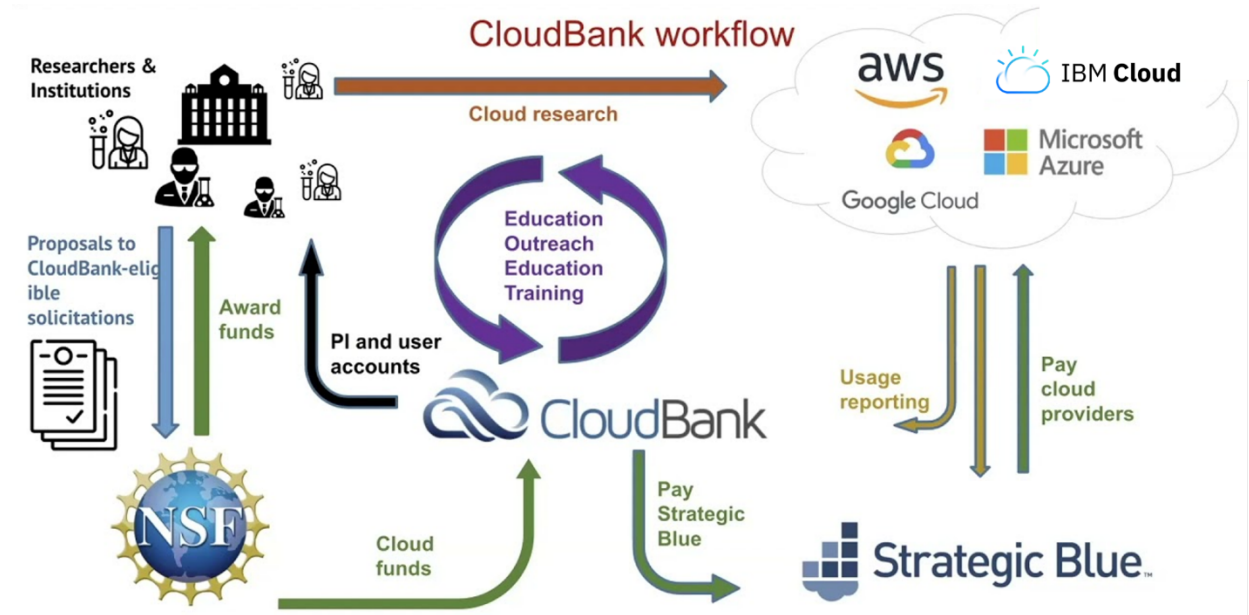
COMMERCIAL RESOURCES	COMMERCIAL RESOURCES	FEDERAL AGENCY SYSTEMS
Amazon Web Services	TACC Frontera GPU/CPU	Argonne Leadership Computing Facility
Microsoft Azure	Indiana Jetstream-2	Prototype National Research Platform
Google Cloud Platform	SDSC Expanse CPU/GPU	National Institute of Health Cloud Lab
Nvidia DGX Cloud	PSC Bridges-2	
OpenAI API	Ohio State University Supercomputer	
Vocareum AI Notebook	University of Illinois NCSA Delta	

Application, Allocation, and Support

- Proposals require an around three-page PDF; eligibility requirements are scrutinized upon submission <https://nairrpilot.org/nairr-pilot-proposal-instructions>
- Due on the 15th of every month and are reviewed by NAIRR Committee by end of the month. If rejected, a resubmission can be set up quickly
- If awarded, allocations to resources will be coordinated with resource providers (AWS, ACCESS Resources, TACC, NIH etc.)
- After being awarded, support can be found through NAIRR and/or Resource Providers
- **NAIRR Office Hours:** Every other Tuesday from 12:00 PST - 13:00 PST
https://iu.co1.qualtrics.com/jfe/form/SV_0MxSk6GjsqpbrHo

3. CloudBank <https://cloudbank.org>

CloudBank is an NSF-funded resource that gives researchers **access to commercial cloud services**.



Picture is from <https://sciencegateways.org/>






CloudBank supports

- Access to multiple public clouds
- Account management tools
- Cost monitoring
- Financial operations
- Help desk support
- Savings (no IDC)
- Education & training (e.g., classroom tools)
- Research across Artificial Intelligence, Quantum Computing, Edge Computing, Network Measurement, Cloud Bursting / Scaling, Chip Design

CloudBank Catalog

HOME / LEARN

The **CloudBank Catalog** is an at-a-glance comparison of the public clouds offered through CloudBank. This page describes broad categories of services and does not list all that the public clouds provide — click the logos below for full service listings.

Click on any category in the table below (e.g., Compute) and it will expand to show a table with types of services in the leftmost column to see a definition of the service type.






- ☰ Compute
- ☰ Storage
- ☰ Network
- ☰ Database Services
- ☰ Big Data & Analytics
- ☰ Machine Learning
- ☰ IoT
- ☰ Application Services
- ☰ Management Services
- ☰ Console & APIs

Compute

Cloud Computing Services provide information technology (IT) as a service over the Internet or dedicated network, with delivery on demand, and payment based on usage. Cloud computing services range from full applications and development platforms, to servers, storage, and virtual desktops. [Source: Dell]

	Amazon Web Services	Google Cloud Platform	Microsoft Azure	IBM Cloud	Oracle Cloud Infrastructure
IAAS	Amazon Elastic Compute Cloud	Google Compute Engine	Azure Virtual Machines	IBM Cloud Virtual Servers	Oracle Cloud Infrastructure (OCI) Compute
General Purpose	●	●	●	●	●
Compute Optimized	●	●	●	●	●
Memory Optimized	●	●	●	●	●

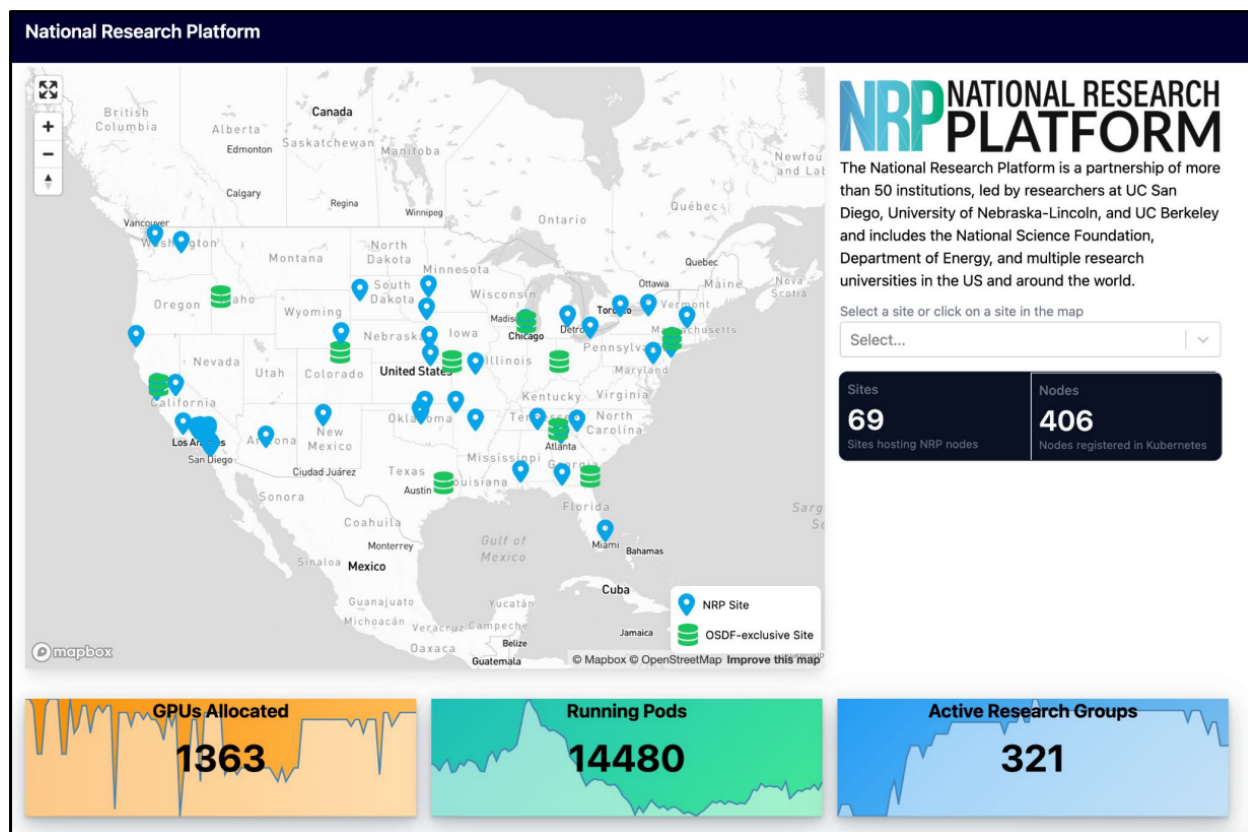
CloudBank eligibility options

Opportunity	Eligibility	More Information	Where to Apply
 NAIRR Pilot	AI-related research projects	Cloud credits are currently available to AI-related research through the National Artificial Intelligence Research Resource (NAIRR) Pilot . The cloud awards for Azure, AWS, and GCP will be managed by CloudBank. Please review our best practices guide and see an example request for help writing a successful proposal.	NAIRR Pilot Allocations Portal
 NDC-C	NDC-C awardees	NSF PIs who have received supplements from DCL 23-101 National Discovery Cloud for Climate Science (NDC-C) can request cloud computing resources thru CloudBank.	CloudBank
 Cloudbank-eligible NSF Solicitations	PIs who are submitting proposals to CloudBank-eligible NSF solicitations	PIs who are submitting proposals to NSF can request cloud resources as part of their NSF proposal if the solicitation is CloudBank-eligible (click to view list). Read more about the process here .	Research.gov
 CloudBank	Active CISE Awardees	PIs who already have an active CISE award can request modest amounts of cloud computing resources thru CloudBank. This type of request is just-in-time as opposed to requiring PIs to budget for resources at the time of proposal submission.	CloudBank
 Self-funded	Faculty with startup funds	Faculty from any US university may deposit startup funds with CloudBank to be spent in Amazon Web Services (AWS).	CloudBank

4. National Research Platform (NRP)

<https://nationalresearchplatform.org/>

- Distributed, federated Kubernetes cluster based on a “Bring Your Own Resource model”
- Focus on Big Data & AI/ML
- Access to a wide range of processors including CPUs, GPUs and FPGAs
- VERNE and TIDE are a portion of NRP
- Resource requirements out of VERNE and TIDE are available through SDSU research foundation

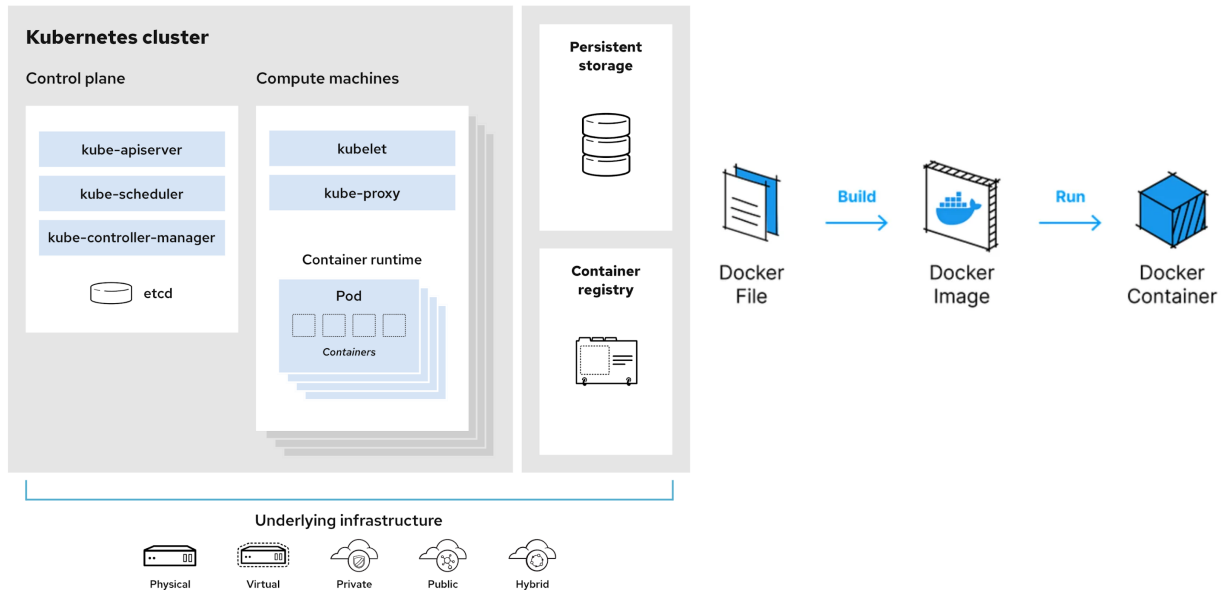


A central component of the NRP is Nautilus, a distributed computing system designed to run containerized big data applications.

The Nautilus cluster is a cluster of mainly GPUs spread across the west coast. Given the high-speed network connectivity, where your jobs actually run does not matter.

- Plenty of shared resources. Use idle GPU servers and data servers contributed by any universities.

- Isolation. Running in an isolated pod, your machine learning tasks are free from interruptions of others' tasks. You have ready-to-use dependencies.



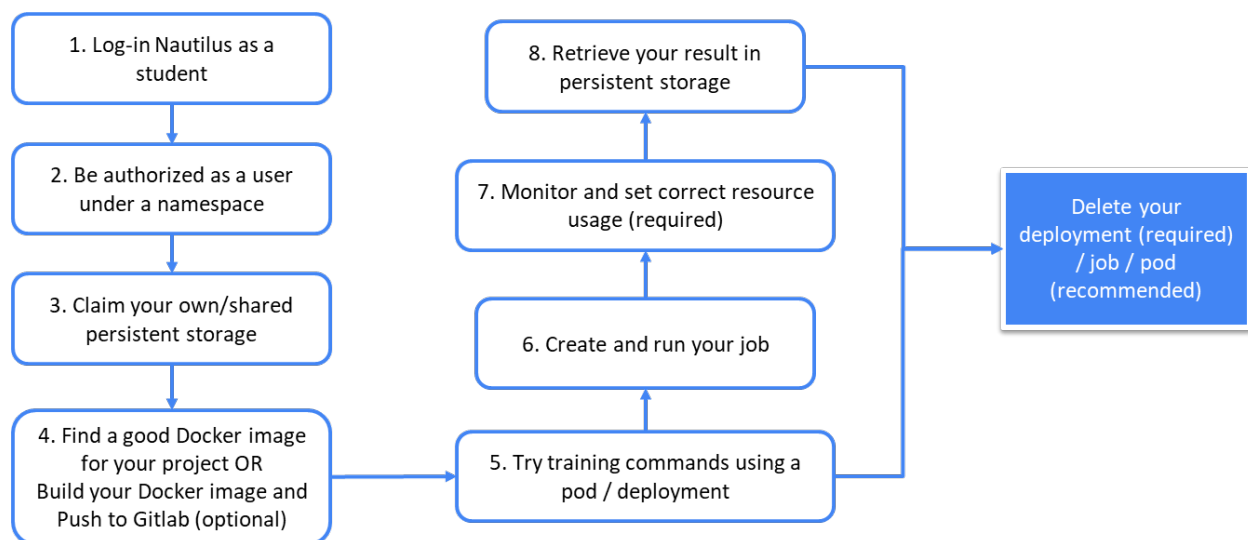
Nautilus offers versatile functions:

- Multi-GPU machine learning training & inference
- Host Jupyter pod
- Host web server
- Continuous Integration (automated testing and image building after git commit)
- Online doc collaboration
- Online code collaboration
- ...

ML tasks can be run on Nautilus in 3 ways.

- Create a pod that stays up for 6 hours. You can request at most 1 GPU. Usually for debugging.
- Create a batch-job that execute a list of commands and terminates on finish. No resource limit, but requesting much more than needed or running everrunning command violates policy.
- Create a deployment that stays up for 2 weeks. No resource limit, but requesting more than minimal violates policy. Usually not recommended for ML tasks.
- See <https://ucsd-prp.gitlab.io/userdocs/start/policies/> for more details.

The following are the standard steps for ML tasks in Nautilus:



5. Others to Explore

RESOURCES	IDEAL FOR	ADDITIONAL INFO
Leadership Class Computing Facility: Frontera, Vista, and soon Horizon	Large-scale parallel applications	https://tacc.utexas.edu/
Chameleon Cloud	Edge to cloud research	https://www.chameleoncloud.org/
CloudLab	Cloud computing research	https://www.cloudlab.us/
FABRIC	High-performance networking, cybersecurity, S&E applications	https://portal.fabric-testbed.net/
Oakridge leadership computing Facility	Research in areas of interest to DOE	https://www.olcf.ornl.gov/community/pathways-to-supercomputing/
National Energy Research Scientific Computing Center	Research in areas of interest to DOE	https://www.nersc.gov/
National center for atmospheric Research (NCAR)	Research in areas of interest to Earth system sciences	https://ncar.ucar.edu/