Big Data (DS-GA 1004) – Lecture 5 Finals Preparation Notes

Fully based on Week 5 Slides + Lecture Transcript + Expanded Details

Spring 2025

Lecture 5: NYU HPC Infrastructure and Big Data Processing

NYU HPC Research Technology Services

- Team under NYU IT responsible for research computing.
- Services provided:
  - High-Performance Computing (HPC) clusters for research and courses.
  - Big Data support, Machine Learning (ML), Deep Learning (DL), Artificial Intelligence (AI).
  - Cloud Computing support: Google Cloud Platform (GCP), Amazon Web Services (AWS).
  - JupyterHub access for courses.
  - Security Data Research Environment (SDRE) for handling sensitive data.
  - High-Speed Research Network (HSRN) for fast data transfer.
  - General research IT support (hardware advice, cloud recommendations).
- Website: `https://hpc.nyu.edu`
- Contact: `hpc@nyu.edu`

Google Dataproc Cluster (for Courses)

What is Dataproc?

- A managed Hadoop and Spark service running on Google Cloud.
- Supports HDFS, MapReduce, Spark, Hive, Trino, Pig.
- Block Size: 128 MB (optimized for large files, poor for many small files).
- Resource Management: YARN for resource allocation, job scheduling, and monitoring.
- Compared to HPC systems: YARN is simpler and more lightweight than SLURM.

Dataproc System Architecture

- Master Node: Login node (accepts SSH connections).
- Two Primary Worker Nodes: Persistent HDFS storage and compute.
- Secondary Worker Nodes (Auto-scaling): Compute-only, added/removed based on workload.

Usage for Students

- Web access: `https://dataproc.hpc.nyu.edu`
- Home Filesystem: `/home/<netid>_nyu_edu`
- HDFS Filesystem: `hdfs dfs -ls, hdfs dfs -put`
- Quota: 500 GB per user in HDFS.
- Application lifetime limit: 5 hours (Spark/YARN jobs).
- Default Spark deploy mode: `cluster` (for production).
- Debugging: `spark-shell --deploy-mode client`
- Login Node:
  - **ONLY** for job submission and debugging.
  - **NOT** for heavy computation.
  - 3 GB memory limit per user.
  - User sessions forcibly killed after 48 hours.
- Containerization: Jobs run inside containers managed by YARN.
- Logs: Output logs should go to HDFS **(not to stdout!)**

Dataproc Back-end Details

- Start with 1 TB HDFS storage, dynamically scales to 8 TB if needed.
- Ingress storage buckets available to upload very large files.
- All files deleted at semester end **(temporary cluster)**.
- Costs controlled by auto-scaling based on student usage.

Big Data on Greene HPC Cluster

Overview of Greene Cluster

- Greene: General-purpose HPC Cluster.
- Available for NYU researchers (except Langone and Abu Dhabi campuses).
- Specs:
  - 38,000 CPU cores, 220 TB memory.
  - 768 GPUs, totaling 13 TB GPU memory.
  - 12 PB parallel storage.
  - HDR 200Gb/s Infiniband network.
- Location: NYU Research Computing Center (RCDC) in Secaucus, NJ.
- Liquid-cooled NVIDIA H100 GPU servers (SD650N-V3).

Access to Greene

- Host: `greene.hpc.nyu.edu`
- 3 Load-balanced login nodes (via NYU network or VPN).
- Login methods: Terminal (Mac/Linux/Windows WSL), PuTTY, MobaXterm, VSCode Remote.
- Web access: Open OnDemand (OOD) server `https://ood.hpc.nyu.edu`

Open OnDemand (OOD)

- Browser-based access to HPC services.
- GUIs for JupyterLab, Matlab, Spark standalone cluster, Dask with Jupyter, Remote Desktop.
- Enables launching interactive jobs graphically.

Running Spark and Dask

- Spark available in both batch and standalone modes.
- Example Spark batch scripts located at:
  texttt/scratch/work/public/apps/pyspark/3.5.0/examples/spark
- Dask clusters launchable via OOD interface.

Hardware Configurations

Compute Nodes:

- Standard memory (524 nodes): 48 CPU cores, 180 GB RAM.
- Medium memory (40 nodes): 48 CPU cores, 369 GB RAM.
- Large memory (4 nodes): 96 CPU cores, 3014 GB RAM.

GPU Nodes:

- NVIDIA V100, RTX8000, A100, H100 GPUs.
- AMD MI100, MI250 GPUs.
- Multiple configurations of cores, memory, and GPU types.

Important Greene Policies

- **No compute-heavy jobs on login nodes.**
- Submit batch jobs or start interactive sessions via `srun`.
- Filesystems:
  - Home: 50 GB quota, 30,000 inode limit, backed up daily.
  - Scratch: Larger, non-backed up space for big data.
- VPN required for off-campus login (up-to-date).
- Security is strict: gateway servers exist, daily audits.

Containerization on Greene

Containers

- Singularity used instead of Docker (security reasons).
- Portable, reproducible environments for computation.
- Pack libraries, binaries, dependencies into container images.

Advantages of Containers

- Improved reproducibility (e.g., scientific papers).
- Easier compatibility across upgrades.
- Facilitates hardware/software portability.

Spark, TensorFlow, PyTorch on Greene

- Jobs can be containerized.
- Use Conda environment + Singularity + overlay filesystem.
- Manage resource requests carefully (e.g., GPUs, CPU cores, memory).
- Setup distributed training carefully (use DDP, backend specifications).

Cluster Resilience and Storage

- **Disk failures occur weekly:** Cluster tolerates with redundancy (RAID + replication).
- **Disaster recovery:** Greene storage snapshots backed up on AWS S3.
- **Redundancy:** Filesystems designed for single-point disk failure tolerance.
- **Cooling:** Liquid cooling and advanced ventilation deployed.

Wrap-up

- Big Data clusters (Dataproc) designed for course workloads **(temporary, reset each semester)**.
- Greene HPC cluster supports real research **(persistent, highly redundant)**.
- Future lectures: Submission scripts, Spark cluster configurations, GPU job handling.

Reminder: Save your work from Dataproc before semester end. All files will be erased.