From Zero to RAPIDS in 7 Days: Learning How to Use RAPIDS for Data Science on Comet

Marty Kandes
Computational & Data Science Research Specialist
High-Performance Computing User Services Group
San Diego Supercomputer Center
University of California, San Diego

NVIDIA Deep Learning Institute @ SDSC
Wednesday, August 21st 2019
1:00PM - 2:30PM PT

About Me

- High-Performance Computing Group @ SDSC
- Distributed High-Throughput Computing Group @ SDSC
- Computational Science Research Center @ SDSU
- I am most definitely **not** a data science expert

About You

Who are you?

Question 1

Are you a graduate student, post-doctoral scholar, research staff member, or professor at UCSD (or another UC campus?

Question 2

Are you a graduate student, post-doctoral scholar, research staff member, or professor at a non-UC U.S. educational institution (or another non-profit research entity?)

Question 3

Are you an industry partner?

Question 4

Do you have a (non-training) user account on Comet or TSCC?

Question 5

Does your day-to-day research work involve data science?

Question 6

What programming languages do you use day-to-day for your research?

Question 7

Do you use NVIDIA GPUs in your day-to-day research work?

Question 8

Have you ever run your own website?

Question 9

Have you seen this movie?

An Overview: From Zero to RAPIDS in 7 Days

- How to access supercomputing resources for your research
- A (very) quick, historical note on Data Science
- How to monitor your CPU/GPU resources
- How to run a Jupyter Notebook
- How to use Pandas
- More than you probably wanted to know about the Fannie Mae Single-Family Loan Performance Dataset (or How I Learned to Stop Worrying about the Performance Data and Love the Acquisition Data)
- How to accelerate your Pandas-like workflows with cuDF

A not-so long time ago in a data center not that far, far away ...

- In 2012, 99% of all computational jobs run on NSF-funded HPC resources utilized fewer than 2048 CPU-cores, while accounting for approximately 50% of the total core-hours consumed across these resources.
- Nearly 70% of all jobs actually ran on only a single compute node (16 CPU-cores) or less.

Comet: A Supercomputer Built to Serve the 99%

Comet By the Numbers

- **1944 compute nodes**: Dual-socket; 2.5 GHz Intel Xeon E5-2680v3 processors; 12 cores per processor; 128 GB DDR4 DRAM; 120 GB/s memory bandwidth; 320 GB SSD (210 GB Avail)
- **4 large-shared memory nodes**: Quad-socket; 2.2 GHz Intel Xeon E7-8860v3 processors; 16 cores per processor; 1.5 TB DDR4 DRAM; 400 GB SSD (260 GB Avail)
- **36 k80 gpu nodes**: Same as standard *compute* node, but with 2 PCIe-based NVIDIA Tesla K80 dual-GPU accelerators per node
- **36 p100 gpu nodes**: Dual-socket; 2.4 GHz Intel Xeon E5-2680v4 processors; 14 cores per processor; 128 GB DDR4 DRAM; 150 GB/s memory bandwidth; 400 GB SSD (260 GB Avail); 4 PCIe-based NVIDIA Tesla P100 GPU accelerators per node

2.76 Pflop/s

Comet By the Numbers

- **Interconnect**: Mellanox FDR (56Gbps) InfiniBand; hybrid fat-tree topology; rack-level (72 node) full bisection bandwidth; 4:1 over-subscription cross-rack bandwidth
- **Storage**: NSF-based $HOME storage (100 GB per user; weekly backups); 6.4 PB 200 GB/s Lustre-based parallel filesystem storage (intermediate-term use: at least 500 GB per group allocation in /oasis/projects; short-term use: up to 10 TB per user in /oasis/scratch; *2M inodes limit*; NO BACKUP!)
- **Applications**: More than 173 software applications and libraries maintained and deployed via Rocks (Linux) cluster distribution; accessible to users via software modules; span a wide range of scientific disciplines, including, but not limited to, bioinformatics, chemistry, data analytics, engineering, fluid dynamics, mathematics, molecular dynamics, neuroscience, and statistics
- **Scientific Impact**: 1755 PIs; 358 institutions; 1144 research allocations; 4709 direct-access users; 33000+ gateway users; 997 publications

Computing Without Boundaries

Coming September 2020

Triton Shared Compute Cluster (TSCC)

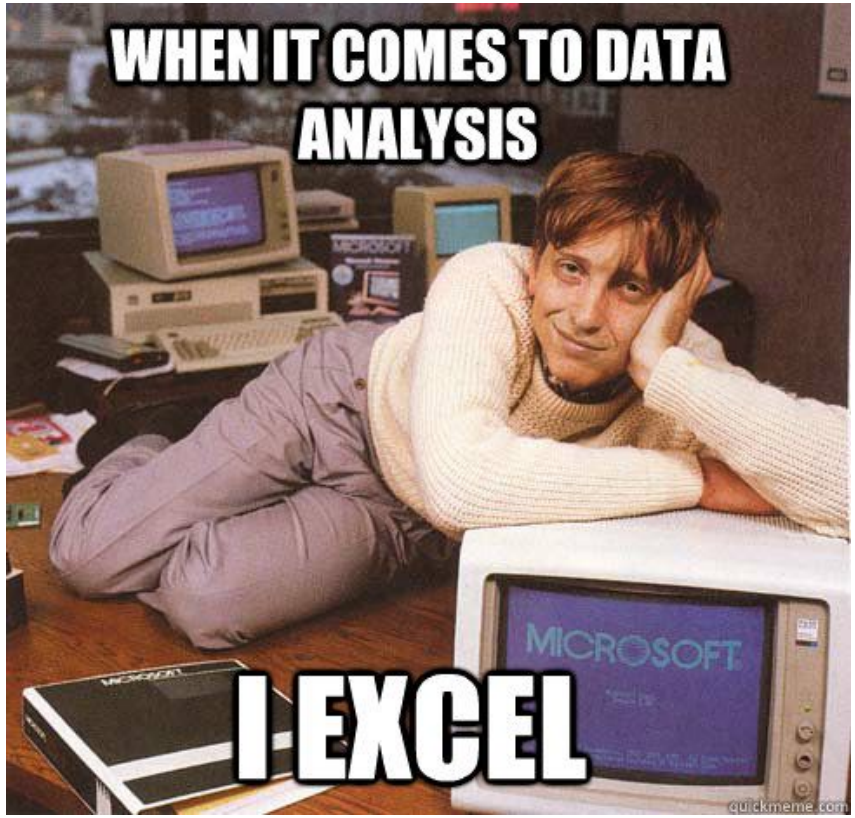
- Medium-scale research cluster (launched in 2013)
- Hybrid business model:
  1. "condo" (buy-in)
  2. "hotel" (pay-as-you-go)
- Mixed architecture: 375 CPU nodes (7k cores); 50 GPU nodes (300 GPUs); 850 TB parallel filesystem; Ethernet + Infiniband networks
- Approximately 30 participating labs/research groups

How do I get time on Comet or TSCC?

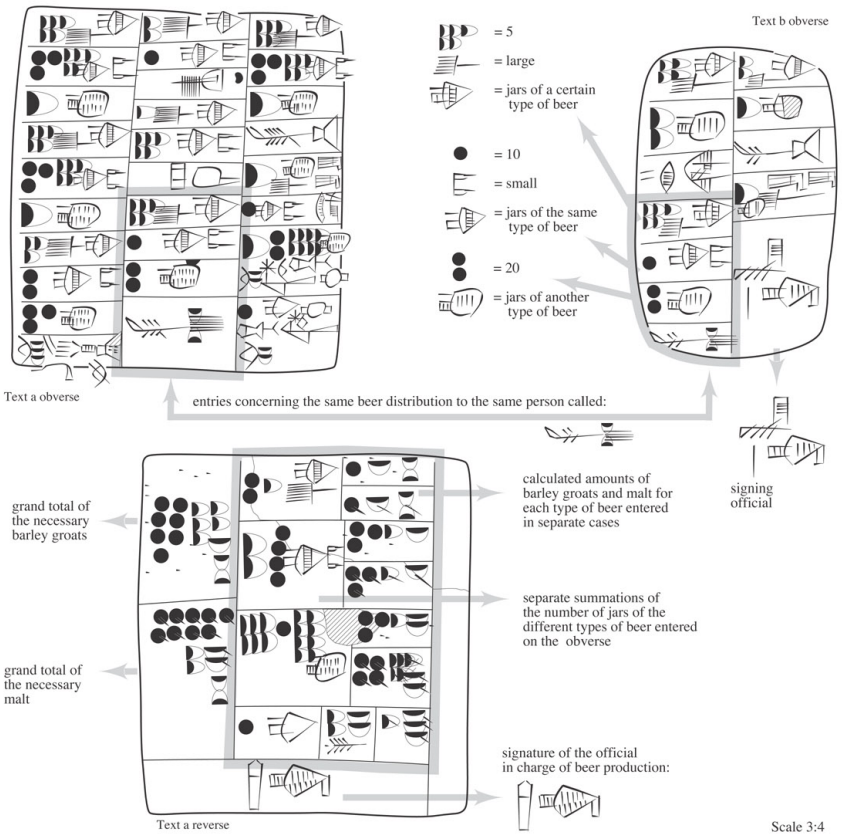
- Comet:
  1. UC: HPC @ UC Program - `https://www.sdsc.edu/collaborate/hpc_at_uc.html`
  2. UC/Non-UC/Non-Profit: XSEDE - `https://www.xsede.org/`
  3. Industry: Ron Hawkins @ SDSC, Industry Relations
- TSCC:
  1. UC/Non-UC/Non-Profit/Industry: Ron Hawkins @ SDSC, Industry Relations

In the beginning (of Data Science) ...

Mid-1980s - Today

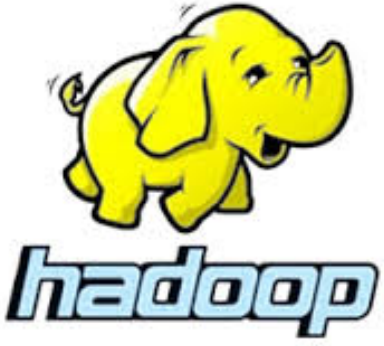


3200-3000 BCE



MSVO 3, 11 and 6
The large account to the left represents the consolidation of at least five texts, one of which is depicted above (note particularly the oblique stroke etched at the base of the sign GI [ ] in the former text, missing in the latter; it presumably indicated that the respective entry had been checked for accuracy). The counted measures of beer (jugs, probably of various sizes and/or representing beer sorts of different strengths) recorded on the obverse of MSVO 3, 11, were in the reverse of the account totaled and qualified with the amount of the grain products 'barley groats' and 'malt' required for their brewing. The entire account was signed by the responsible office 'KU SZIM'.

Today



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$