Data Science Beyond the Laptop: Handling Data of Any Size with Arkouda

Ben McDonald, Michelle Strout, Sarah Coghlan, Oilver Alvarado Rodriguez

ChapelCon June 5, 2024
Data Scientists are modern-day explorers uncovering previously unimagined insights
Faced with the unknown, data scientists as well as explorers can suffer from the “lighthouse effect”
What if we could work with **massive-scale** data in its entirety **without** compromising interactivity?

Arkouda (αρκούδα)
Data Science Beyond the Laptop
Arkouda Performance

HPE Cray EX (May 2023)
- Slingshot-11 network (200 Gb/s)
- 8192 compute nodes
- 256 TiB of 8-byte values
- ~8500 GiB/s (~31 seconds)

• “...solving problems in a matter of seconds, rather than days...” – Tess Hayes, Bytoa
Data Science Beyond The Laptop

- Demands of data science today
- What is Arkouda?
- What is Chapel?
- The Arkouda ecosystem
- Conclusion
- Tutorial...
- The future of Arkouda
Data Science Beyond the Laptop

Data sets today

- Data scientists are drawn to Python for its interactivity
  - Code executes through the REPL (read, evaluate, print, loop)
  - Operations complete within human thought-loop
- Data science demands interactivity...
Data Science Python is only *syntactically Python*.

Image from Bill Reuss's 2020 Chapel keynote.
Data Science Beyond the Laptop

Data sets today

• Data sets have outgrown typical computers
  • Unbiased sampling is difficult
  • Unbiased sampling can eliminate rare and high-order effects

• **Data science demands scalability...**
Data Science Beyond the Laptop

Data sets today

Python must scale **beyond the laptop**, without sacrificing **interactivity**
Arkouda

**Interactivity**

Arkouda Client
(written in Python)

**Scalability**

Arkouda Server
(written in Chapel)
# Data Science Beyond the Laptop

Arkouda

*An open-source Python package providing interactive data analytics at supercomputing scale.*

Transform the way you work with big data; massive computation within the human thought loop

<table>
<thead>
<tr>
<th>EASY TO USE</th>
<th>FAST &amp; SCALABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provides an API data scientists are familiar with based on Pandas/NumPy</td>
<td>Sorting 256 TiB of data on 8,000 Nodes within seconds</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EXTENSIBLE &amp; CUSTOMIZABLE</th>
<th>POWERED BY CHAPEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly extensible ecosystem allows rapid feature development and broad project collaboration</td>
<td>Powered by a parallel distributed server written in Chapel</td>
</tr>
</tbody>
</table>
## Data Science Beyond the Laptop

The Chapel Programming Language

*From laptops to supercomputers, Chapel makes parallel programming more productive.*

<table>
<thead>
<tr>
<th>EASY TO USE</th>
<th>FAST &amp; SCALABLE</th>
<th>PORTABLE</th>
<th>GPU-READY</th>
<th>OPEN SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supports code as approachable as Python and flexible as C++</td>
<td>Scales to millions of cores with performance that rivals MPI</td>
<td>Executes on: HPE Apollo, HPE Cray EX, HPE Superdome Flex, Linux/*nix systems, Mac, NVIDIA and AMD GPUs</td>
<td>Supports high-level, vendor-neutral GPU programming without language extensions</td>
<td>Developed by HPE on GitHub in collaboration with the open-source community</td>
</tr>
</tbody>
</table>

>>> Leverage the parallel power of your hardware quickly.

>>> Scale your applications with ease.

>>> Write your code once and run it anywhere.

>>> Unlock the power of GPUs for parallel computing.

>>> Join a growing community of Chapel users and developers!
Data Science Beyond the Laptop
Arkouda

Image from Bill Reuss’s 2020 Chapel keynote
Data Science Beyond the Laptop
The Arkouda Ecosystem

<table>
<thead>
<tr>
<th>id</th>
<th>label</th>
<th>name</th>
<th>born</th>
<th>brand</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>src id</td>
<td>69</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>69</td>
<td>dst id</td>
<td>34</td>
<td>lives-with</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>89</td>
<td>relationship</td>
<td>69</td>
<td>NULL</td>
<td>NULL</td>
<td>NULL</td>
</tr>
<tr>
<td>34</td>
<td>since</td>
<td>2011</td>
<td>drives</td>
<td>89</td>
<td>NULL</td>
</tr>
<tr>
<td>69</td>
<td>bought</td>
<td>2011</td>
<td>drives</td>
<td>NULL</td>
<td>NULL</td>
</tr>
</tbody>
</table>
| 69 | owns | NULL  | NULL  | 2011  | 15
| 89 | drives | NULL  | NULL  | NULL  | NULL  |
Data science requires an intimacy with data reached through interactive exploration.

Regardless of data quantity, quality requires scalable workflows on the complete dataset.
Data Science Beyond the Laptop

Demo

Learning Objective

• How to use Arkouda
  • Launch a server
  • Connect to a server
  • Create arrays
• The Arkouda API
  • Supported file formats
  • Exploratory data analysis

Follow Along with an Arkouda Codespace

What’s in Store for Arkouda?

Ben McDonald, Michelle Strout, Sarah Coghlan, Oliver Alvarado Rodriguez

ChapelCon June 5, 2024
What’s in Store for Arkouda?

Next Steps

**Developer Experience**
- Efforts underway to make “Arkouda code” look like “Chapel code” (eliminating boilerplate)

**Resiliency**
- Plans to improve server recoverability and improve server resiliency

**User Experience**
- Considering non-blocking interaction with Arkouda via a messaging queue to improve responsiveness

**Modernization**
- Working towards a web summary interface to provide detailed server information in real time
What’s in Store for Arkouda?

Boilerplate code

- Boilerplate required to Add Arkouda functionality today
  - 85 lines of unnecessary boilerplate code
  - Repetition of code for each Arkouda supported type (int, uint, bool, etc.)

- Boilerplate required to Add Arkouda functionality with this proposal
  - 6 lines of boilerplate code, each is intuitive
  - No repetition for types, since it is handled by the server one level up
  - Looks like “regular” Chapel code, no special Arkouda-specific code other than import
Data Science Beyond the Laptop
Parquet read + GroupBy performance vs Spark

Speed Comparison, Arkouda vs. Repartitioned Spark, Data Type: int

Number of Partitions
- Arkouda
- Spark - 200.0
- Spark - 1000.0
- Spark - 2000.0
- Spark - 5000.0
- Spark - 10000.0
- Spark - 20000.0

Number of Compute Nodes
Thanks for Attending the ChapelCon 2024 Tutorial Day!

ChapelCon June 5, 2024