

July 2, 2019

# AWK vs Big Data

§ tech (/categories/tech/) # awk (/tags/awk/) • bigdata (/tags/bigdata/) • quotes (/tags/quotes/)

Don't sleep on the basics. Someone probably solved your problem in the 80s.

There's been a lot of talk about big data recently. Lots of people just shove data into whatever software is currently all the rage (think Hadoop some time ago, Spark, etc) and get excited with results that actually aren't that amazing. You can get very decent results by using the standard data processing toolset ( `awk/grep/sed/sort/xargs/find` ) paired with understanding of what data you process and how the software works.

One of the best pieces of software I wrote was a data mining tool working on a dataset of approximately .5TB which is not that much. The trick was that it was completing queries on that dataset in subsecond timeframe. And I did spend quite a bit of time working on performance to achieve that result.

Thus being involved with this sort of tasks, I was amused to stumble upon this tweet:

**Nick Strayer**

@NicholasStrayer · [Follow](#)



Recently I got put in charge of wrangling 25+ TB of raw genotyping data for my lab. When I started, using spark took 8 min & cost \$20 to query a SNP. After using AWK + [#rstats](#) to process, it now takes less than a 10th of a second and costs \$0.00001. My personal [#BigData](#) win.

2:34 PM · May 30, 2019



566



Reply




Copy link

[Read 18 replies](#)


The full story ([https://livefreeordichotomize.com/2019/06/04/using\\_awk\\_and\\_r\\_to\\_parse\\_25tb/](https://livefreeordichotomize.com/2019/06/04/using_awk_and_r_to_parse_25tb/)) behind this tweet is a very nice reading of how the author was re-discovering plain-text tools with some fun and insightful quotes and other tweets like:

**Nick Strayer**  
@NicholasStrayer · [Follow](#)






Me taking algorithms class in college: "Ugh, no one cares about computational complexity of all these sorting algorithms"

Me trying to sort on a column in a 20TB [#spark](#) table: "Why is this taking so long?"  
[#DataScience](#) struggles.

12:26 PM · Mar 11, 2019 

---

 56  Reply  Copy link

[Read 1 reply](#)

and, a very true one:

gnu parallel is magic and everyone should use it.

Frankly, I think that quite a number of so-called *Big Data* applications can be re-done using venerable text-processing tools and produce cheaper and faster results in the end. Which reminds me another article (<https://adamdrake.com/command-line-tools-can-be-235x-faster-than-your-hadoop-cluster.html>) where the author re-did a Hadoop task to process ~2Gb file with `awk` and ended up with 235x speed increase:

I tried this out, and for the same amount of data I was able to use my laptop to get the results in about 12 seconds (processing speed of about 270MB/sec), while the Hadoop processing took about 26 minutes (processing speed of about 1.14MB/sec).

— *'If you knew Time as well as I do,' said the Hatter, 'you wouldn't talk about wasting IT. It's HIM.'*

\$ Last updated: Feb 7, 2021 at 13:38 (EET) \$