

Where would the data come from?

A map of the data a new project would read, before it starts

A new project that reads a lot of data, an AI model or a large analytics job, can get planned around the processors and the software. The data it reads sits on storage you already run, next to everything else. Map each source before kickoff: what it is, where it lives, how big it is, how fast it reads today, and how it would reach the project.

One lane per data source the project would read

1 First data source

The data What it is, and who owns it.

Where it lives today Which array, share or cloud bucket.

How big Size today, and how fast it grows.

Read rate today A quiet hour and the busiest hour, dated.

Copies or moves to get there Each copy, move or export on the way.

2 Second data source

The data What it is, and who owns it.

Where it lives today Which array, share or cloud bucket.

How big Size today, and how fast it grows.

Read rate today A quiet hour and the busiest hour, dated.

Copies or moves to get there Each copy, move or export on the way.

3 Third data source

The data What it is, and who owns it.

Where it lives today Which array, share or cloud bucket.

How big Size today, and how fast it grows.

Read rate today A quiet hour and the busiest hour, dated.

Copies or moves to get there Each copy, move or export on the way.

> ## Which copy of your data would the project actually read?

How to fill in the map, and what it does not prove

MLCommons, a consortium funded by member companies (storage makers that sell what its benchmarks measure submit results to it), describes its storage benchmark as a test of "a given storage system's ability to keep pace" with the processors it feeds. That is the question behind this map, asked about your own storage before the project starts.

The data
Ask the project team, in writing, which data it would read. Then find who owns each source in the business.

Where it lives today
The array, file share, database or cloud bucket holding the copy the project would read. More than one copy? Write each.

How big
The size shown where it lives, and how much it grew over the last year.

Read rate today
Read throughput on the array's performance view at a quiet hour and at last month's busiest hour, each dated. What the array did, not the datasheet.

Copies or moves
Every copy, export or transfer between where the data lives and where the project would run. Each one takes time, space and someone's yes.

> If the project read at your busiest hour, what else would be reading then?

What this does not prove

That the storage can carry the project. A reading is not a load test. It shows what the array did with today's work, not what it would do with the project added.

What the project needs. Only the project team can say how much it will read and when. Ask for that number in writing.

That every copy is allowed. Moving data somewhere new may need the data owner's yes. The map shows the moves, not the permissions.

Words used here

Throughput
how much data the storage delivers per second. The read rate in the map is read throughput.

Accelerator
a processor built for heavy computing, such as AI training. The "processors" in MLCommons' benchmark.

Load test
a test that adds the project's work on purpose and measures what the storage does.

Sources

- MLCommons, MLPerf Storage v2.0 results announcement (August 2025), used for the idea only.

Made by StepKnock, a small team in San Diego, paid through our partners when a customer buys. This page sells nothing.

Nothing to sign up for. No cookies, no analytics, and nothing you type here leaves this page.

stepknock.com/toolkit/checks/where-the-data-comes-from/