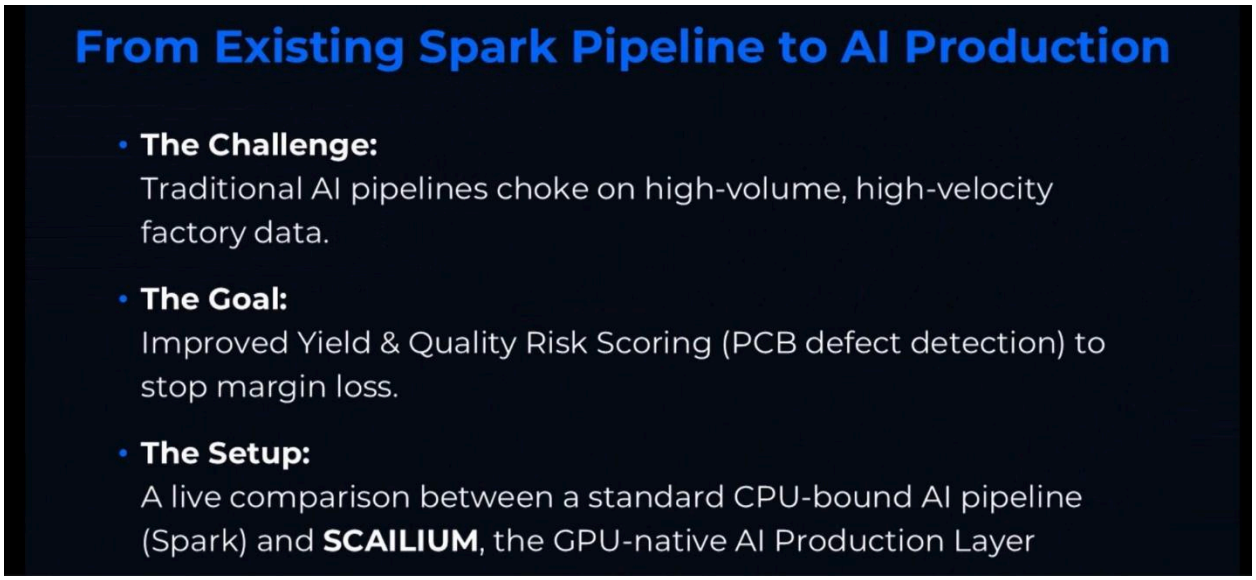
Demo Guide

Preparation:

- Access Demo Video: www.scailium.com/demos
- Password: SC26
- When demoing, enlarge the video (full screen)
- Pause it as needed for further explanation

On Screen:



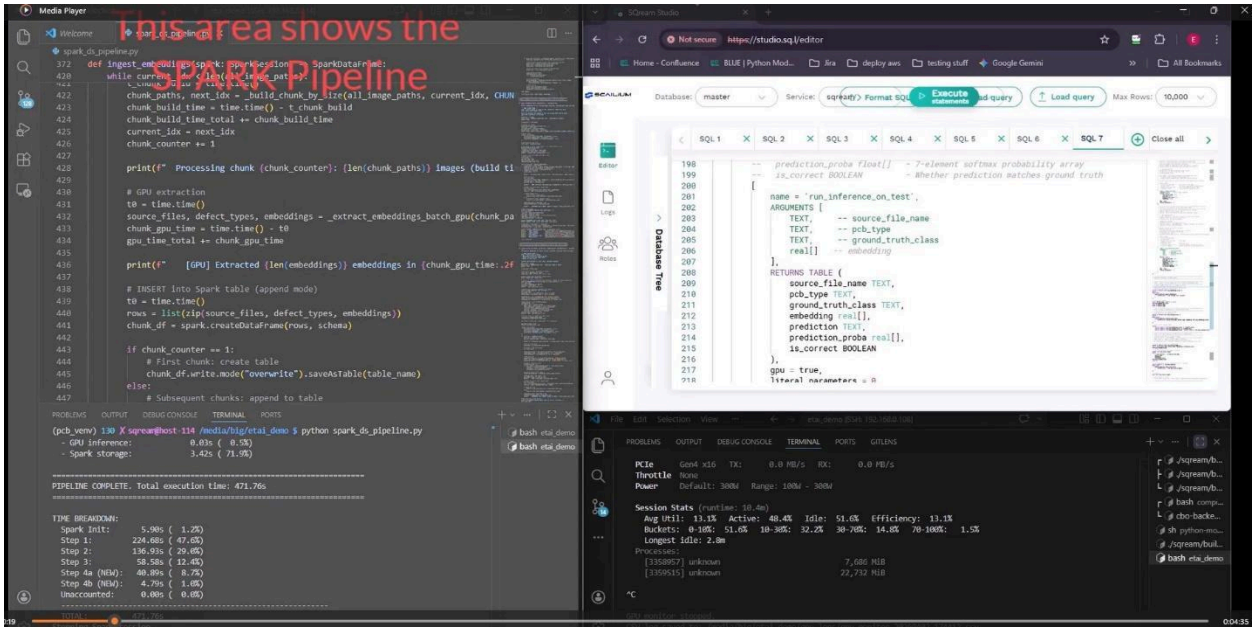
What to say:

“Let me walk you through a simple side-by-side comparison.

This demo starts with an existing Spark-based machine learning workflow and shows what happens when that same workflow runs through SCAILIUM instead. The point is not to introduce a different model or a different use case. The point is to show what changes when the production path changes.

In this example, the business goal is manufacturing quality. We want to find defects sooner, improve yield, and move from raw factory data to action faster and more accurately”

On Screen:



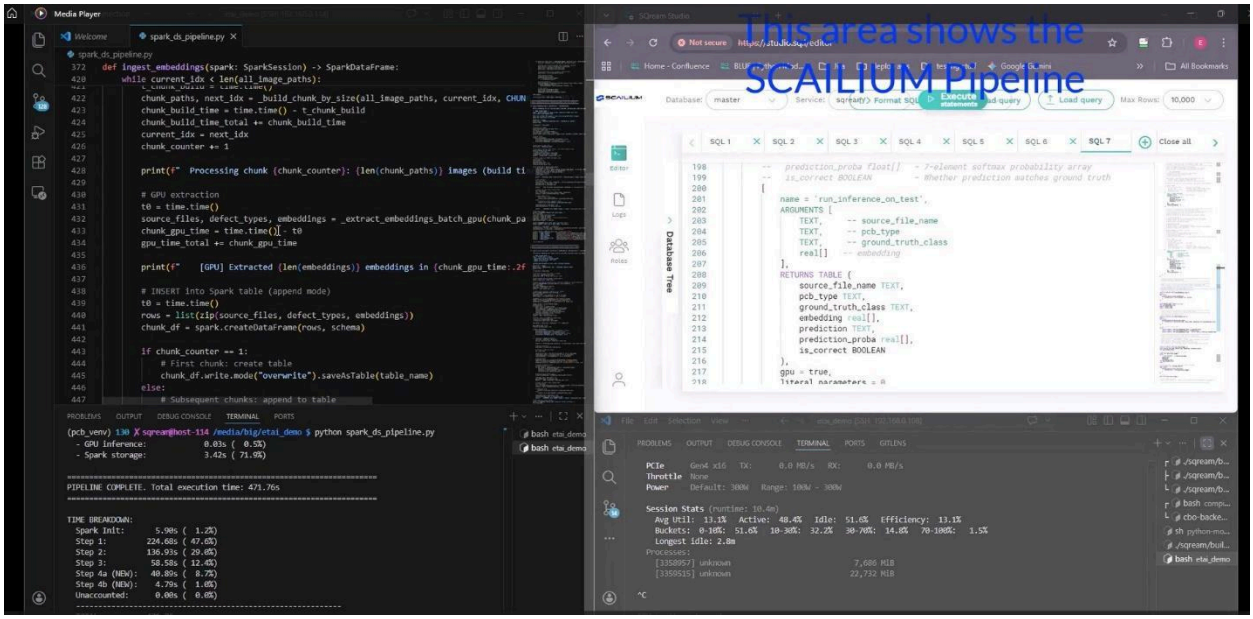
What to say:

“On the left, you are looking at the original Spark pipeline.

This is the baseline many teams already know. It takes the data in, prepares it, trains the model, and then runs prediction. There is nothing unusual here. This is the familiar way teams build and run machine learning workflows today.

That is important, because we are not comparing two different business processes. We are starting from the same practical workflow a customer may already have.”

On Screen:



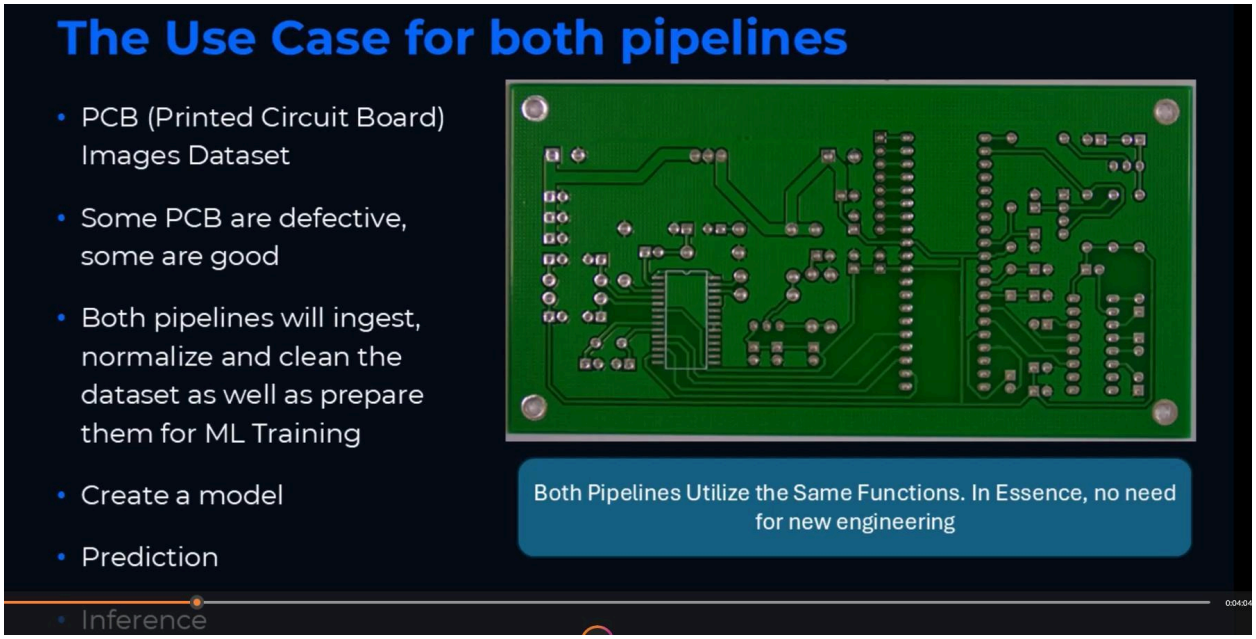
What to say:

“On the right, you see that same workflow running through SCAILIUM.

This is one of the most important points in the demo. We did not rebuild the application from scratch. We reused the workflow logic and ran it through a better production path. That is why this matters in the real world. Nobody wants a science project. People want a faster, cleaner way to operationalize what they already have.

So from this point on, think of this as the same business workflow with a different engine under it. No rip and replace”

On Screen:

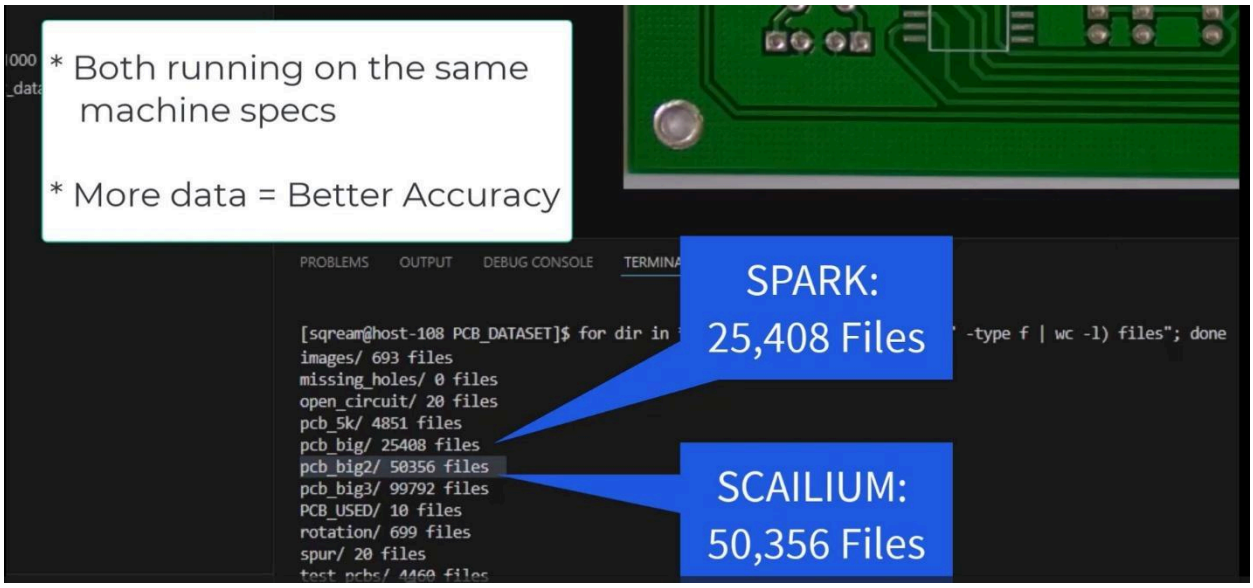


What to say:

“This is the use case both pipelines are working on.

We are using PCB images. Some boards are good, and some are defective. Both sides take those images, prepare the data, train the model, and then run prediction. In simple business terms, this is about catching quality problems earlier, reducing scrap and rework, and protecting yield.”

On Screen:



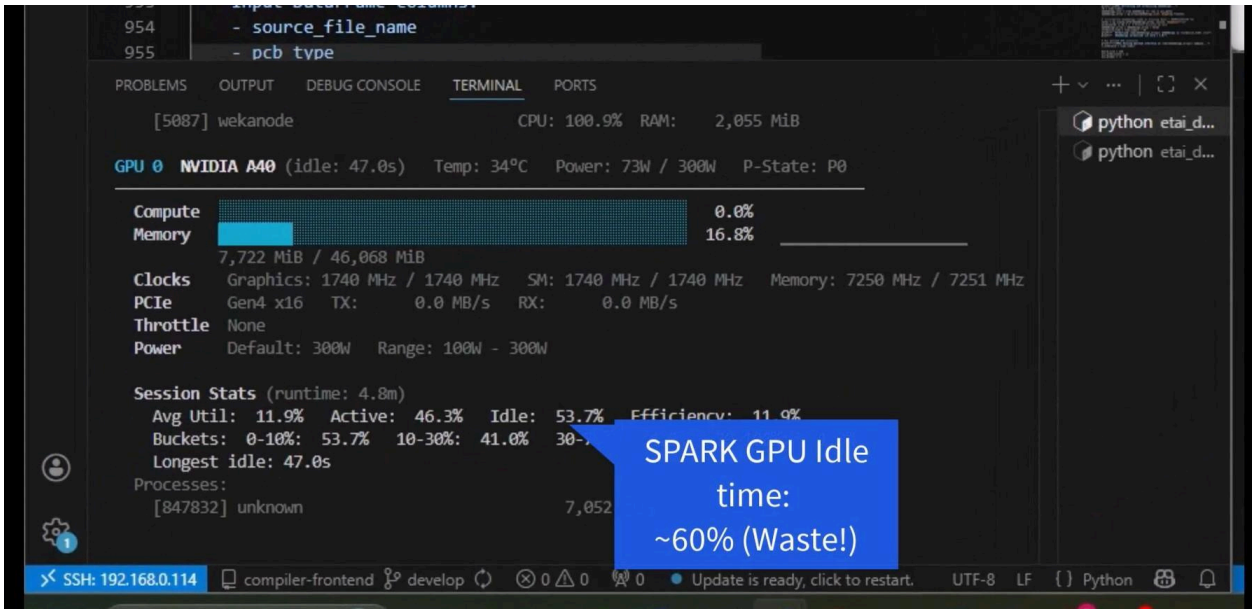
What to say:

“Here we show the scale of the comparison.

Both sides are running on the same machine specs, but SCAILIUM is handling a larger data load. That matters because production AI does not fail on small, tidy samples. It fails when live data volume shows up and the system has to keep moving.

So SCAILIUM is not getting the easier case here. It is doing more work. In a few seconds you will see that it’s even faster and more efficient (even under heavier load)”

On Screen:



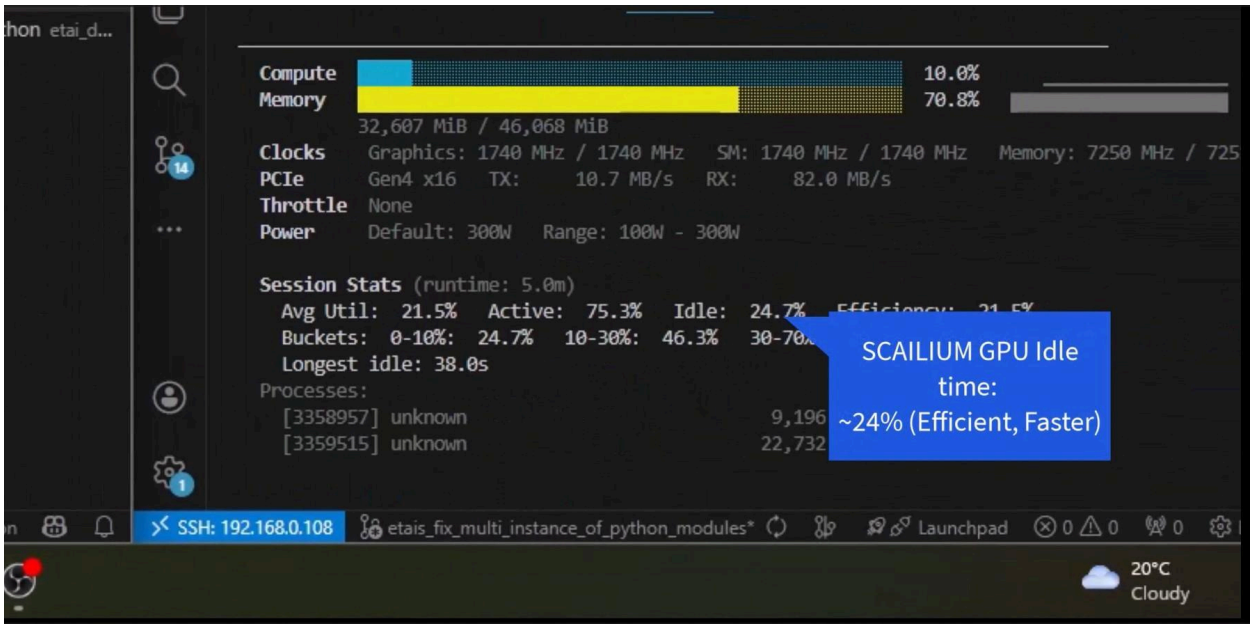
What to say:

“Now we get to the real bottleneck.

On the Spark side, the workflow keeps pushing work back through the CPU path. That leaves the GPU waiting too often (FYI – while waiting, the GPU still uses energy). The hardware is there, the spend is there, but too much of that compute sits idle instead of producing useful output.

This is where the business problem starts to show up. Slower pipelines mean longer time to result, lower infrastructure efficiency, and more waste inside the same hardware footprint.”

On Screen:



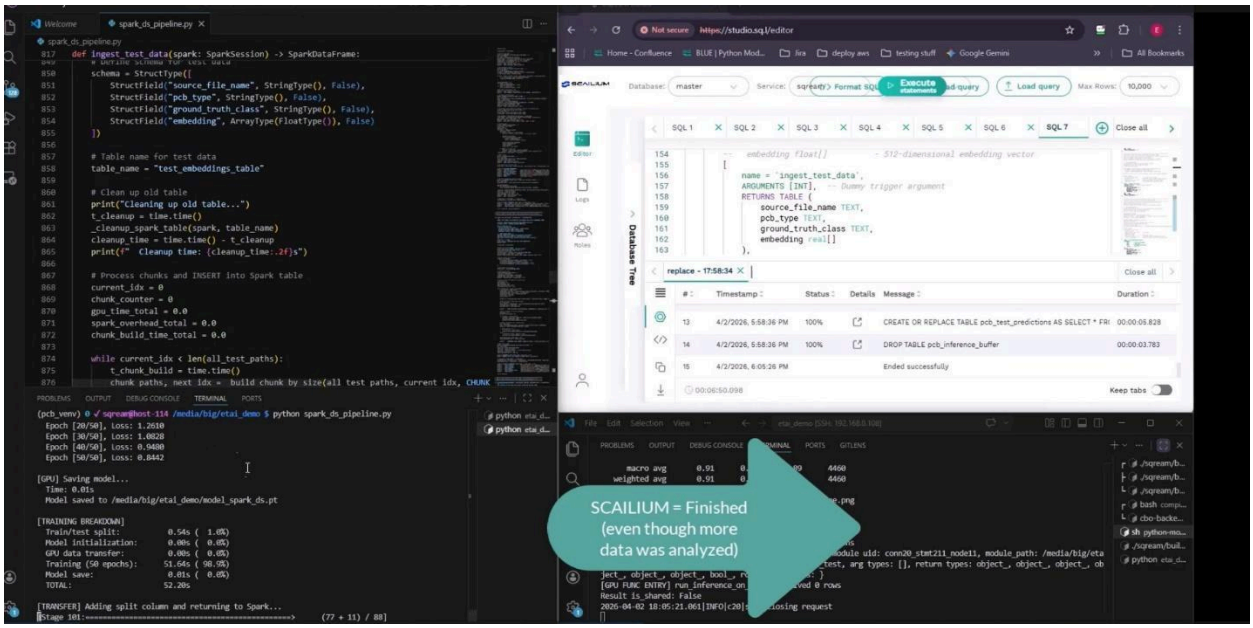
What to say:

“Now compare that with the SCAILIUM side.

Because more of the production path stays on the GPU, idle time drops sharply. That means the system spends more time doing useful work and less time waiting on the wrong part of the stack. This is the heart of SCAILIUM.

We are not trying to replace storage. We are not trying to replace the model. We are fixing the path in the middle so the compute stays fed and productive.”

On Screen:



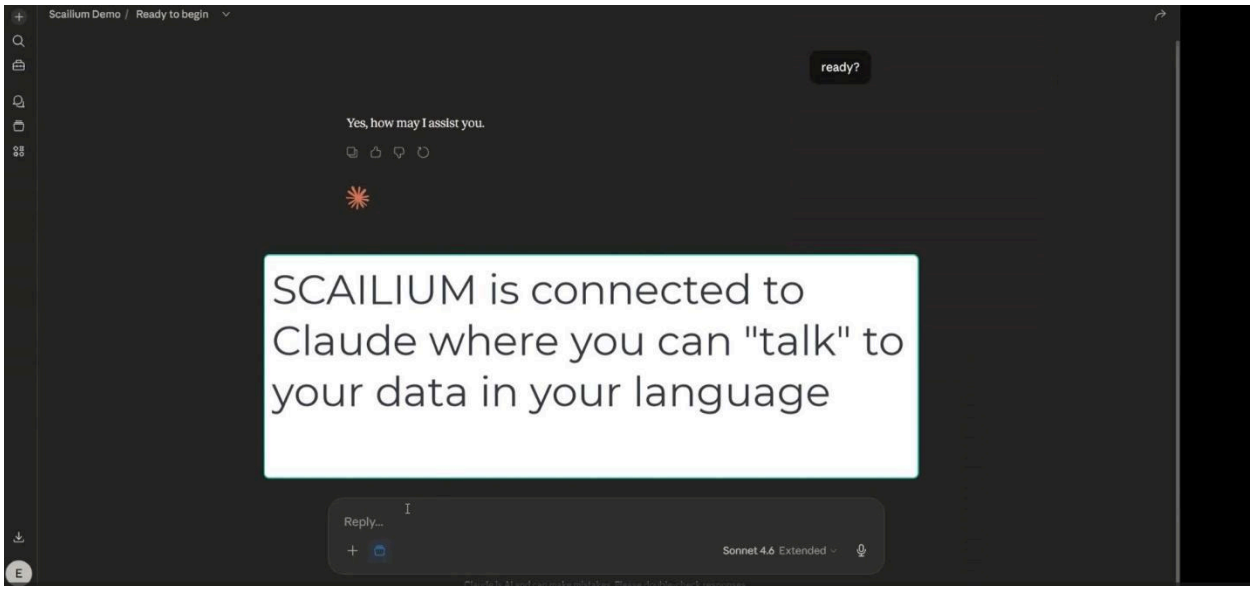
What to say:

“SCAILIUM finishes first, even though it processed more data. That is the outcome you want the prospect to remember. Same business workflow. Minimal change. More data handled. Less idle compute. Faster completion.

That is what a stronger production path looks like. It is not theory. It is not a future promise. It is a visible operational result. The business value is clear: First, higher throughput. The pipeline finishes sooner, which means faster time to inference and faster time to action. Second, better AI economics. Higher utilization means less idle waste and better output from the same hardware and power budget. Third, it fits the existing stack. There is no demand here for a storage replacement, a model replacement, or a major rebuild.

That combination is what makes SCAILIUM practical. It improves the economics and the speed of AI production without turning into a giant transformation project.”

On Screen:



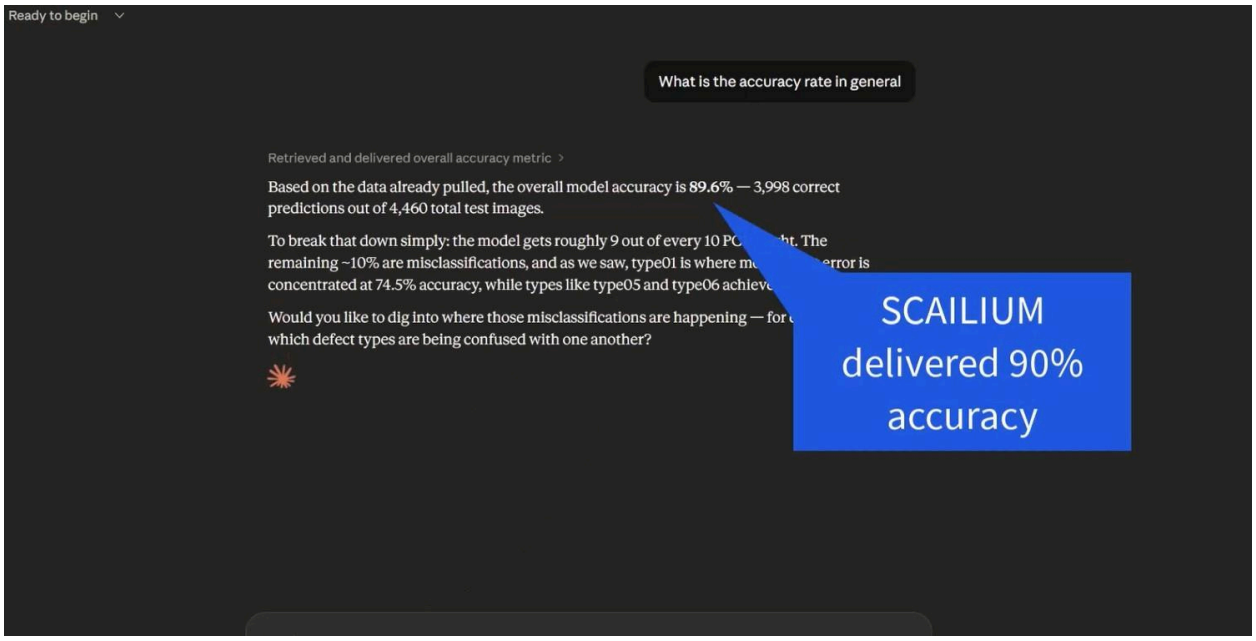
What to say:

“Now the story moves from pipeline performance to usability.

Because SCAILIUM turns raw production data into production-ready context, that output becomes much easier to use. A business user does not need to understand the raw schema or the full data model before asking a useful question. An AI agent gets the same benefit. It works from current, governed, production-ready context.

This matters because business value does not come from data sitting in a system. It comes from turning that data into decisions and actions. In this example, Claude is connected to SCAILIUM where business users can “talk” to their data.”

On Screen:

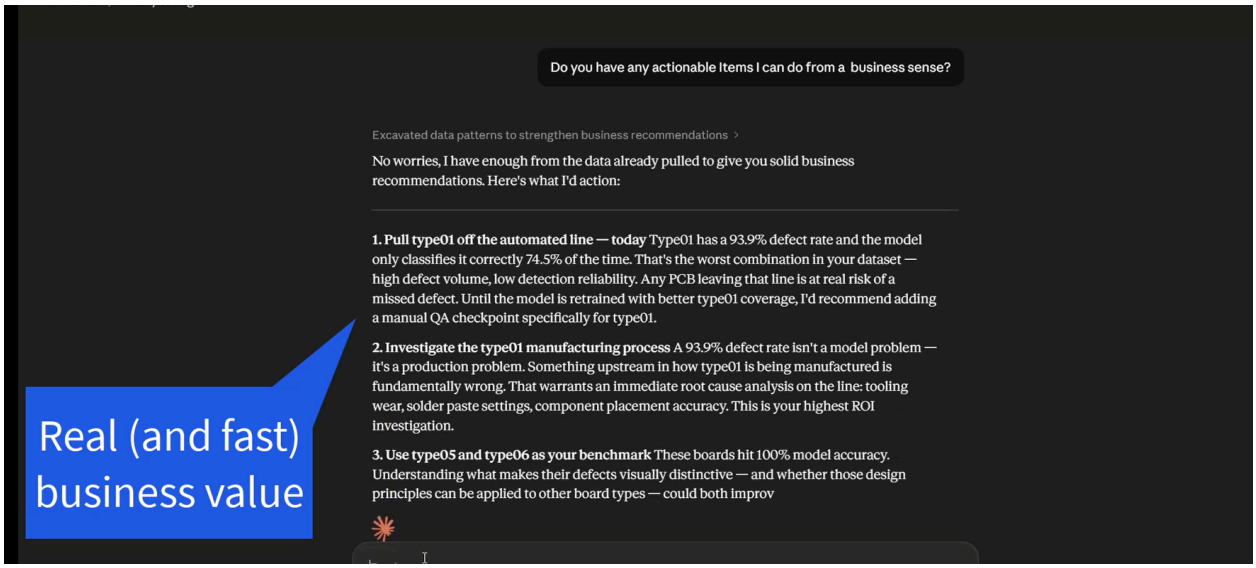


What to say:

“The assistant is working directly from the production-ready context created by SCAILIUM. Someone asks a simple question in plain language, and the system returns the result directly. In this example, it reports about 90 percent overall accuracy. The SPARK accuracy was only around ~80%.

The bigger point is not only the number. The bigger point is that the data is now usable. It is current, understandable, and ready for action by both people and software. That is the shift from dashboards and analysis toward business action.”

On Screen:

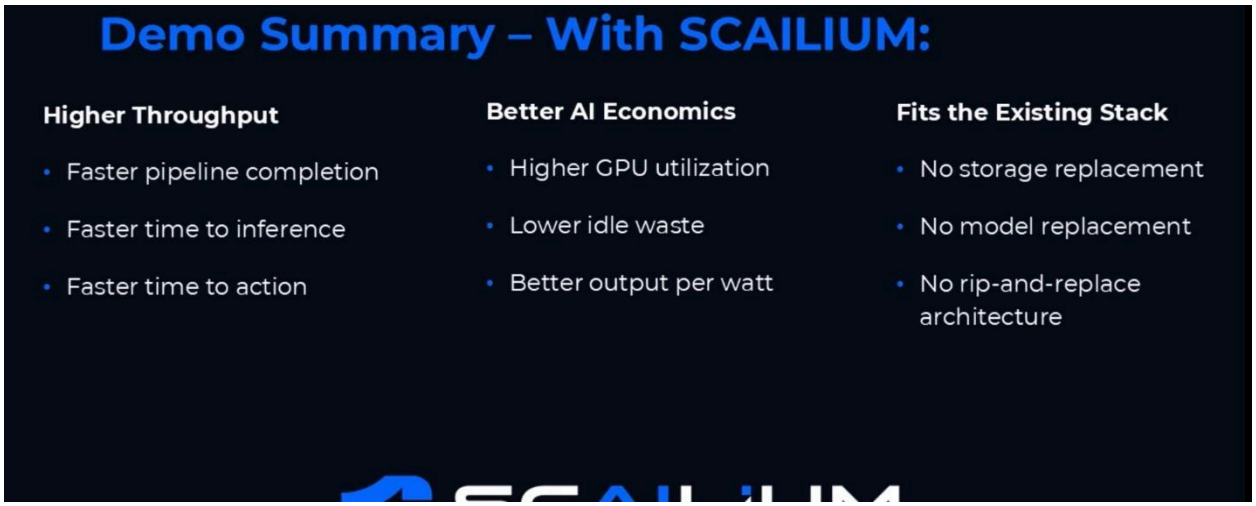


What to say:

“Here is another example where we simply ask “Do you have any business related action items?”

SCAILIUM, via Claude provides real recommendations.”

On Screen:



What to say:

“What this demo shows is simple. We took an existing Spark workflow, ran it on SCAILIUM with minimal change, reduced GPU idle time, processed more data, finished faster, and made the output easier to use for both people and AI agents. That is what we mean by the AI Production Layer.”