🧠 Why is reading a JSON file slower than reading a CSV file in Spark—even if both are the same size?  
  
A student recently asked me this, and I loved the curiosity behind it. Here's how I explained it:  
👉 Imagine you're asked to move 1 big rock that weighs 100 kg  
vs.  
👉 You're asked to move 100 small rocks that weigh 1 kg each.  
Both tasks involve the same total weight—but the second one is easier and faster, right?  
💡 That’s because smaller chunks can be moved in parallel.  
  
📂 Similarly, when reading files in Spark, CSV files are splittable, meaning:  
- Each line represents a complete record.  
- Spark can read different parts of the file simultaneously across multiple cores.  
🧱 On the other hand, JSON files are often non-splittable:  
- Records can span multiple lines.  
- The nested structure makes it hard to define clear split boundaries.  
- So Spark must process it more sequentially thereby slowing things down.  
  
🚀 Moral of the story?  
File format matters a lot when you’re working with distributed data processing. Choose wisely for performance! 😉

