تقرير مقارنة طرق قراءة ملف MIMIC-IV الكبير

ملخّص لنتائج تجربة قراءة ملف `emar_detail.csv` بواسطة ثلاث طرق مختلفة: القراءة على أجزاء باستخدام (chunksize) `pandas`، استخدام `dask`، والقراءة الخطية من ملف مضغوط `.gz`.

نتائج الأداء

الطريقة	الوقت (ثانية)	الذاكرة (ميغابايت)
Chunksize	213.43	29.78
Dask	197.25	1845.58
Compression (.gz)	33.11	0.12

أفضل زمن: Compression (.gz) — 33.11s

أقل استهلاك ذاكرة: Compression (.gz) — 0.12 MB

أسرع تقنية تحليلية: Dask (مناسب للخوادم ذات RAM عالي)

التحليل

• **طريقة Chunksize**: مناسبة للبيئات ذات ذاكرة محدودة، حيث تقرأ البيانات دفعة تلو الأخرى وتسمح بتنفيذ تحليلات متتابعة دون استهلاك عالي للرام. زمن التنفيذ كان (≈213s) مع استهلاك ذاكرة منخفض (≈29.78 MB).

• **طريقة Dask**: تستفيد من المعالجة الموزّعة والمحلية، وقدّمت زمنًا أفضل بقليل (≈197s) لكنها استهلكت ذاكرة كبيرة جدًا (≈1845.58 MB) أثناء عمليات الـ compute، لذا تناسب الخوادم ذات ذاكرة كبيرة.

• **طريقة Compression (.gz)**: القراءة الخطية من الملف المضغوط كانت الأسرع والأخف (33.11s و0.12 MB) لكنها محدودة فعليًا في التحليلات المتقدمة لأنها تقرأ الخط سطراً سطراً وليست بنيويّة كـ DataFrame.

بعد الدراسة نقترح ما يلي

الحالة	الطريقة الموصى بها	السبب
أجهزة بذاكرة محدودة	Chunksize	توازن جيد بين الأداء واستهلاك الذاكرة
خوادم قوية (RAM عالي)	Dask	أداء أسرع للعمليات المعقدة والتوازي
قراءة سريعة أو عدّ الصفوف فقط	Compression (.gz)	أخف وأسرع حل لعمليات خطية