التقرير 1

الطرق الثلاث

Pandas+chunksize

قراءة الملفات الكبيرة على دفعات صغيرة.

Chunksizeلتفادي استهلاك الذاكرة دفعة واحدة

Dask

مكتبة تدير المعالجة المتوازية للملفات الضخمة عبر تقسيمها تلقائيا بين الانوية و المعالجات.

Compression(gzip

يتم ضغط الملف لتقليل حجم تخزينه ثم قراءته بعد فك الضغط.

إيجابيات كل طريقة

Pandas+chunksize

- سهلة التنفيذ.
- لا تتطلب مكتبات إضافية.
- توفر تحكما جيدا في حجم الدفعة لتقليل استهلاك الذاكرة.

Dask

- اسرع من Pandas في لملفات الضخمة حيث يستخدم تعدد الانوية.
- يمكنه التعامل مع ملفات اكبر من حجم الذاكرة.
- واجهته سهلة الاستخدام .

Compression:gzip

- يقلل الحجم التخزيني بشكل كبير جدا.
- مفيد لحفظ البيانات او مشاركتها اونلاين.

سلبيات كل طريقة

Pandas+chunksize

- لا تستفيد من تعدد الانوية single threads
- بطئ مقارنة ب Daskعند الملفات الضخمة.

Dask

- يستهلك ذاكرة اكثر اثناء العمليات.
- يحتاج تنصيب مكتبة إضافية.
- لا يعطي نتائج فورية أي يحتاج compute.

Compression gzip

- ابطئ طريقة لأنها تحتاج وقت لفك الضغط.
- استهلاك عالي لذاكرة اثناء الضغط.
- غير عملي للمعالجة الفورية اثناء التحليل.

الطريقة	الوقت	الذاكرة بMB	حجم الملف ب GB
Pandas+chunksize	163.53 s=2.73min	287.64	9.61
Dask	171.49s=2.68min	963.24	9.61
Compression gzip	2975.66s=49.59min	9770.67	2.41

- من الخلال الجدول نستنتج ان الطريقة الأفضل هيPandas+chunksize
- لأنها الأقل استهلاكا للذاكرة ونتائجها ممتازة زمنيا ومناسبة في حالات التحليل السريع أما Daskفهي مناسبة للملفات العملاقة جدا اكثر من 20 جيغا بايت او المعالجة المتوازية لانها تستغل كل أنوية المعالج وتتعامل مع ملفات اكبر من RAMأما في حالة كنا نريد حفظ البيانات و أرشفتها فنستعمل compression gzipلأنها تصغر الحجم كثيرا ولكن غير مناسبة للتحليل الفوري .