تقرير عن الطرق الثلاثة

Chunksize;Dask;GZIP

1*Pandas+ChunkSize

في هذه الطريقة يتم قراءة الملف على دفعات صغيرة مثل chunksize=10000 بدل تحميله دفعة واحدة.

*في الوقت أبطأ قليلا أي تأخذ وقت مقارنة ب Dask لان كل جزء يقرأ ويعالج على حدة.

*أما في الذاكرة منخفضة جدا لأن البرنامج يحتفظ فقط بجزء صغير من البايانات.

الإيجابيات:

✓ مثالية للملفات الكبيرة مثل من 10GBفما فوق.
✓ يمكنك معالجة كل جزء تدريجيا.
✓ السلبيات:
✓ لا يمكن تنفيذ بعض العمليات الا بعد تجميع النتائج يدويا.
✓ أبطأ قليلا من القراءة الكاملة.

2*Dask DATAFrame

في مكتبة Daskيتم قراءة الملفات الكبيرة بشكل متوازي عبر معالج Parallel Processingتتعامل مع الملفات كما تفعل Pandasولكن بطريقة مجزأة و موزعة تلقائيا.
*في الوقت تكون سريعة حيث أنها تستخدم Daskالمعالجة المتعددة لتسريع القراءة والمعالجة.
*في الذاكرة يكون استهلاك متوسط.

الايجابيات:

- ✓ أسرع أداء في الملفات الكبيرة.
- ✓ يستفيد من المعالجات المتعددة.

السلبيات:

- ✓ يحتاج مكتبة dask إضافية.

3*Pandas+GZIP

يتم ضغط الملف الى صيغة csv.gzلتقليل حجمه ثم قراءته.

*في الوقت يأخذ وقت كبير في قراءة الملف.

*في استهلاك الذاكرة يكون استهلاك من متوسط الى كبير نوعا ما .

الإيجابيات :

- ✓ يقلل حجم الملف بنسبة كبيرة.
- ✓ مناسبة لنقل والتخزين.
- ✓ أداء مستقر رغم الضغط.

سلبيات:

- ✓ وقت القراءة أطول قليلا.
- ✓ يحتاج وقت أولي لضغط الملف.