CERN openlab

Performance Study of Parquet Codecs

openlab Summer Students

Javier García Rubio

15/08/2019

WHICH IS THE MOTIVATION

- Large Production data systems at CERN such as NXCALS (New Generation Accelerator Logging Service) use Parquet

- We want to undestand how the different codecs (compression) affect:
  - Data Storage
  - CPU payload
  - Data Access Performance

- Oracle is developing a new Codec for Parquet and this work set up the basis for later comparisons

HOW I DID IT
WHAT DID I USE

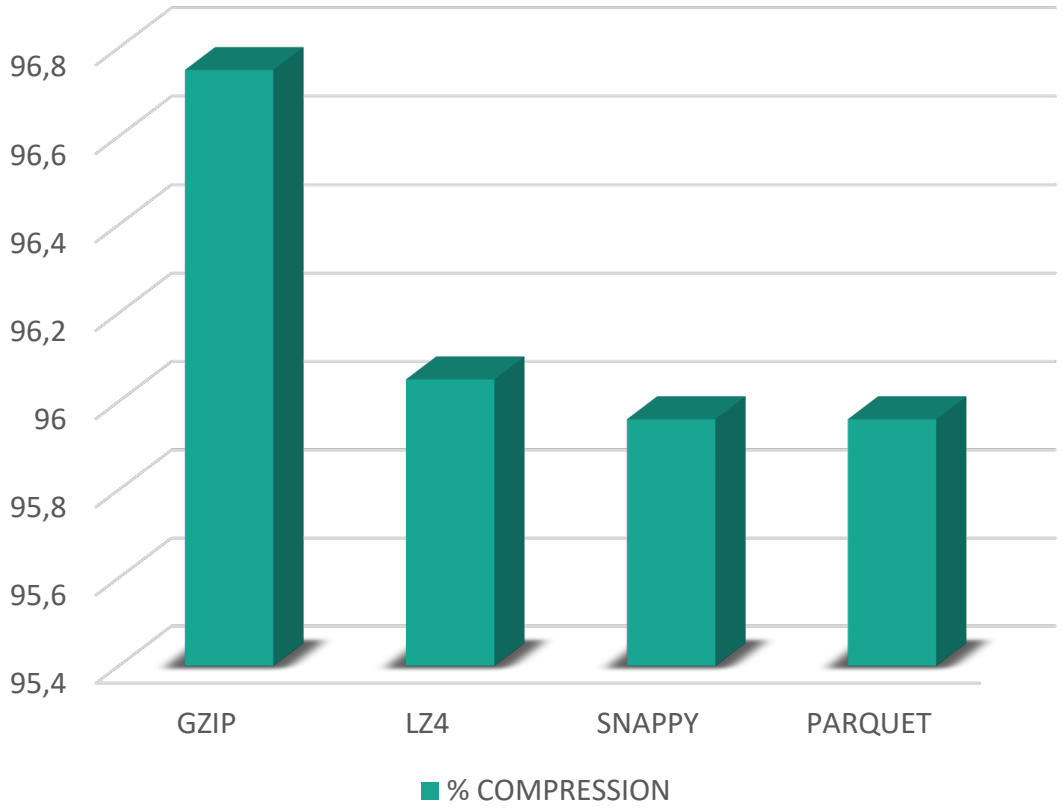
- Using real data from NXCALS:
  - Convert parquet files to json and apply the existing different codecs
    - GZIP
    - SNAPPY
    - PARQUET (NO CODEC)
    - LZ4
  - Take compression rate
  - Take CPU time
  - Evaluate the impact


APACHE
Spark
TM
Scala

WHAT ARE THE RESULTS

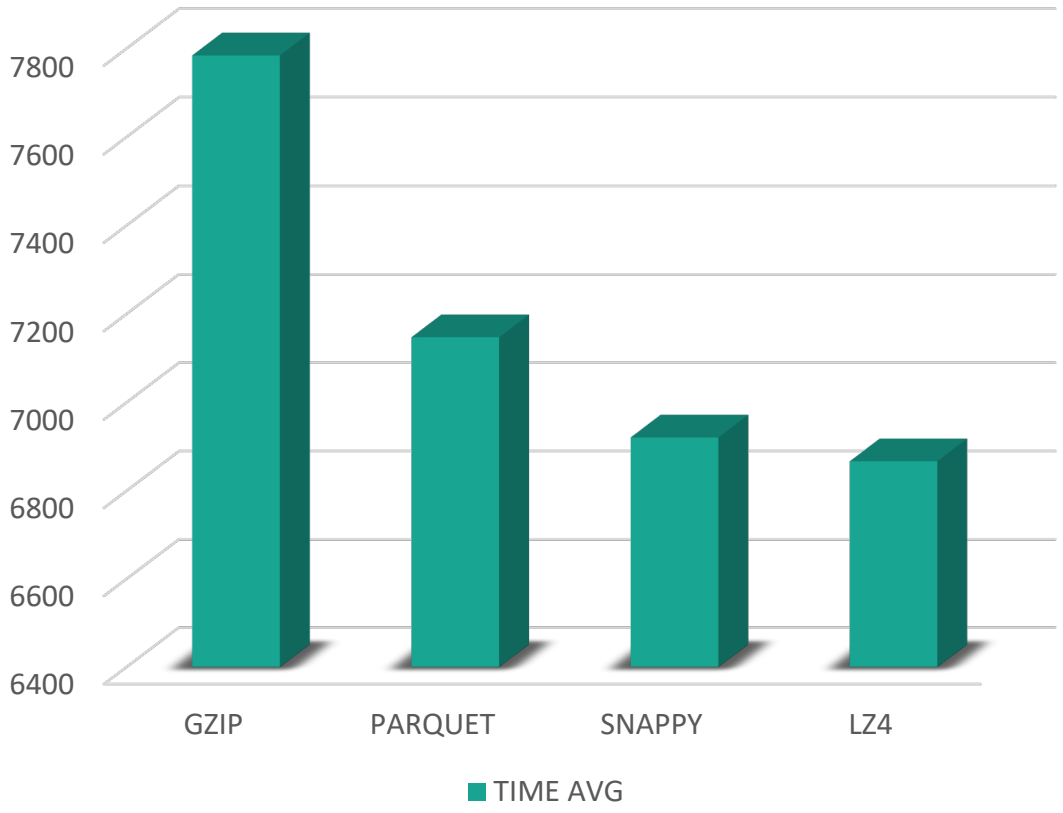
Performance Analysis (compression rate)

CODEC	% COMPRESSION
GZIP	96,8%
LZ4	96%
SNAPPY	95,9%
PARQUET	95,9%



Performance Analysis (CPU time)

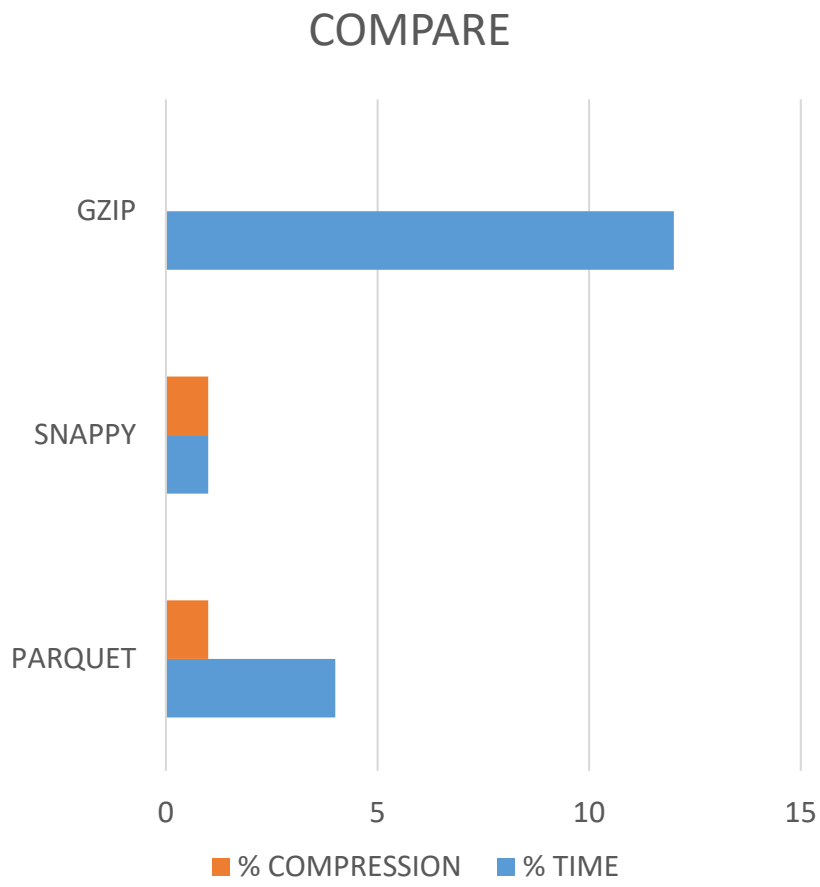
CODEC	TIME AVG (ms)
GZIP	7784,5
PARQUET	7157,5
SNAPPY	6921
LZ4	6867



WHAT ABOUT CONCLUSIONS

CONCLUSION

LZ4



WHAT NEXT

-Compare with the new Parse Aware Compression(PAC) from Oracle

- Waiting for them to make it open source

THANKS

javigr6623@gmail.com

Javier García Rubio