TRENDY TECH

UPLIFT YOUR CAREER!

Compression Codecs

IMPORTANT

Copyright Infringement and Illegal Content Sharing Notice

Compression techniques helps us to reduce storage costs and processing time.

A major overhead in processing large amounts of data is disk and network I/O, reducing the amount of data that needs to be read and written to disk can significantly decrease overall processing time. This includes compression of source data, but also the intermediate data generated as part of data processing. Although compression adds CPU load, for most cases this is more than offset by the savings in I/O.

Although compression can greatly optimize processing performance, not all compression formats supported on Hadoop are splittable. Because the MapReduce framework splits data for input to multiple tasks, having a non splittable compression format is an impediment to efficient processing. If files cannot be split, that means the entire file needs to be passed to a single MapReduce task, eliminating the advantages of parallelism and data locality that Hadoop provides. For this reason, splitability is a major consideration in choosing a compression format as well as file format.

Different compression techniques:

1. Snappy
2. Lzo
3. Gzip
4. Bzip2

Some of them are optimized for speed and others optimized for storage.

Snappy

Snappy is a compression codec developed at Google for high compression speeds with reasonable compression. Although Snappy doesn't offer the best compression sizes, it does provide a good trade-off between speed and size. Processing performance with Snappy can be significantly better than other compression formats. It's important to note that Snappy is intended to be used with a container format like Avro, Orc, Parquet since it's not inherently splittable.

LZO

LZO is similar to Snappy in that it's optimized for speed as opposed to size. Unlike Snappy, LZO compressed files are splittable, but this requires an additional indexing step. This makes LZO a good choice for things like plain-text files that are not being stored as part of a container format. It should also be noted that LZO's license prevents it from being distributed with Hadoop and requires a separate install, unlike Snappy, which can be distributed with Hadoop.

Gzip

Gzip provides very good compression performance (on average, about 2.5 times the compression that’d be offered by Snappy). But in terms of processing speed its slow. Gzip is also not splittable, so it should be used with a container format. Note that one reason Gzip is sometimes slower than Snappy for processing is that Gzip compressed files take up fewer blocks, so fewer tasks are required for processing the same data. For this reason, using smaller blocks with Gzip can lead to better performance.

Bzip2

bzip2 provides excellent compression performance, but can be significantly slower than other compression codecs such as Snappy in terms of processing performance. Unlike Snappy and Gzip, bzip2 is inherently splittable. In the examples we have seen, bzip2 will normally compress around 9% better than GZip, in terms of storage space. However, this extra compression comes with a significant read/write performance cost. This performance difference will vary with different machines, but in general bzip2 is about 10 times slower than GZip. For this reason, it's not an ideal codec for Hadoop storage, unless your primary need is reducing the storage footprint. One example of such a use case would be using Hadoop mainly for active archival purposes.

We have learnt compression Codecs in Hadoop

Happy Learning!!!

Google REVIEWS

5 Star Google Rated Big Data Course

LEARN FROM THE EXPERT

9108179578

Call for more details

TRENDYTECH
UPLIFT YOUR CAREER!

Follow US

Trainer	Mr. Sumit Mittal
Phone	9108179578
Email	trendytech.sumit@gmail.com
Website	https://trendytech.in/courses/big-data-online-training/
LinkedIn	https://www.linkedin.com/in/bigdatabysumit/
Twitter	@BigdataBySumit
Instagram	bigdatabysumit
Facebook	https://www.facebook.com/trendytech.in/
Youtube	https://www.youtube.com/channel/UCbTggJVf0NDTfWX-C_gUGSg