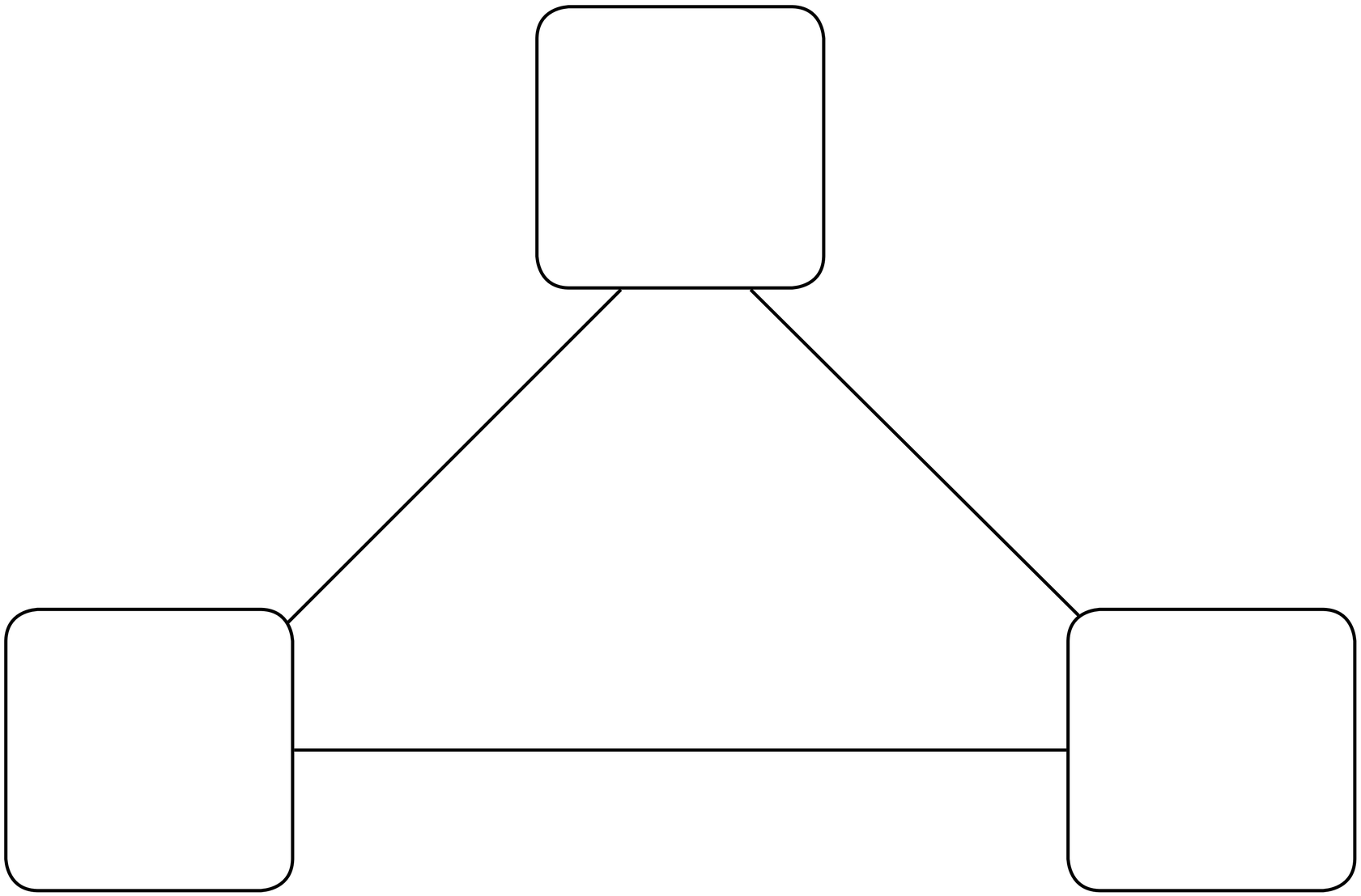
Yandex

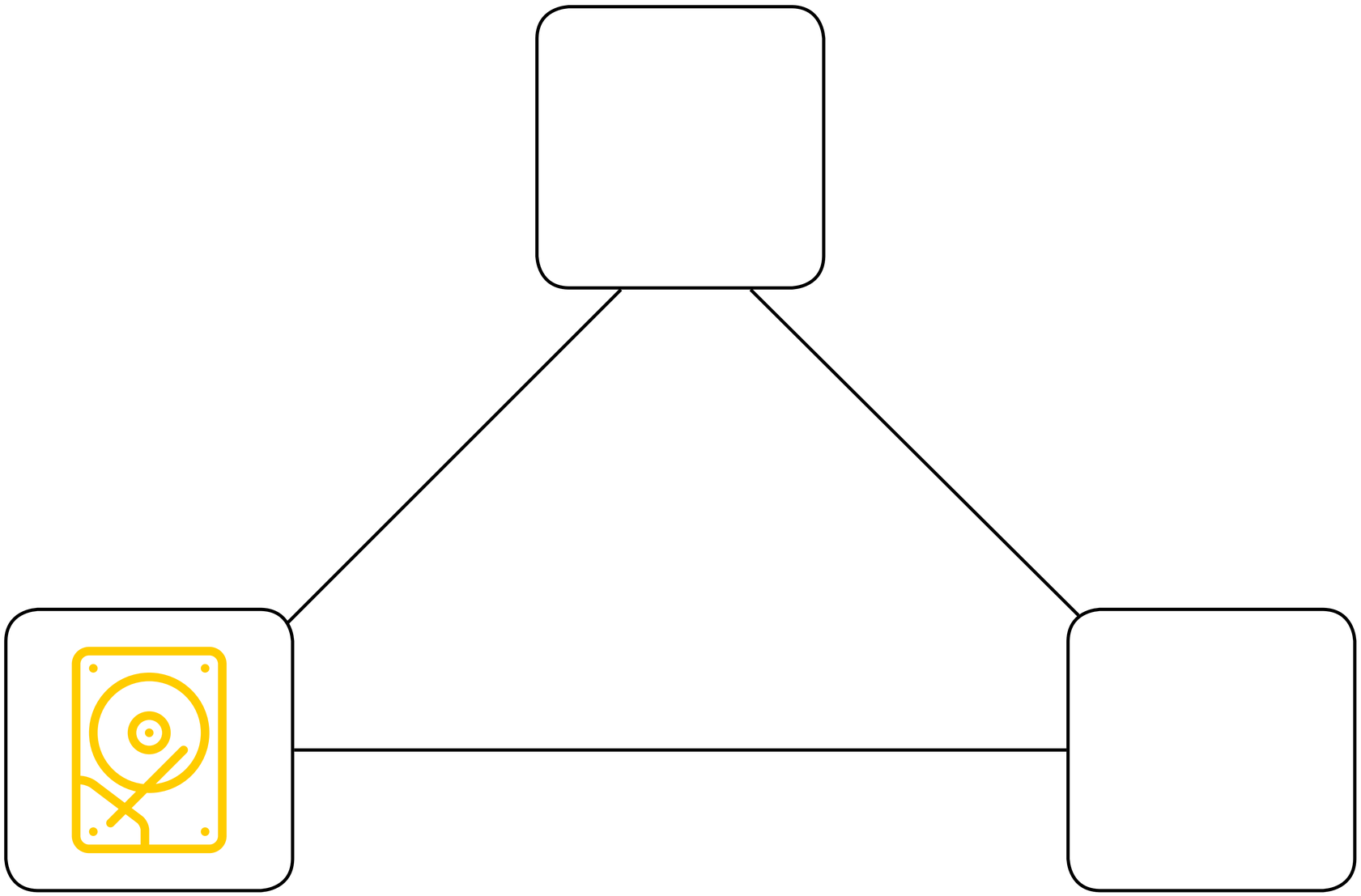
MapReduce

Compression

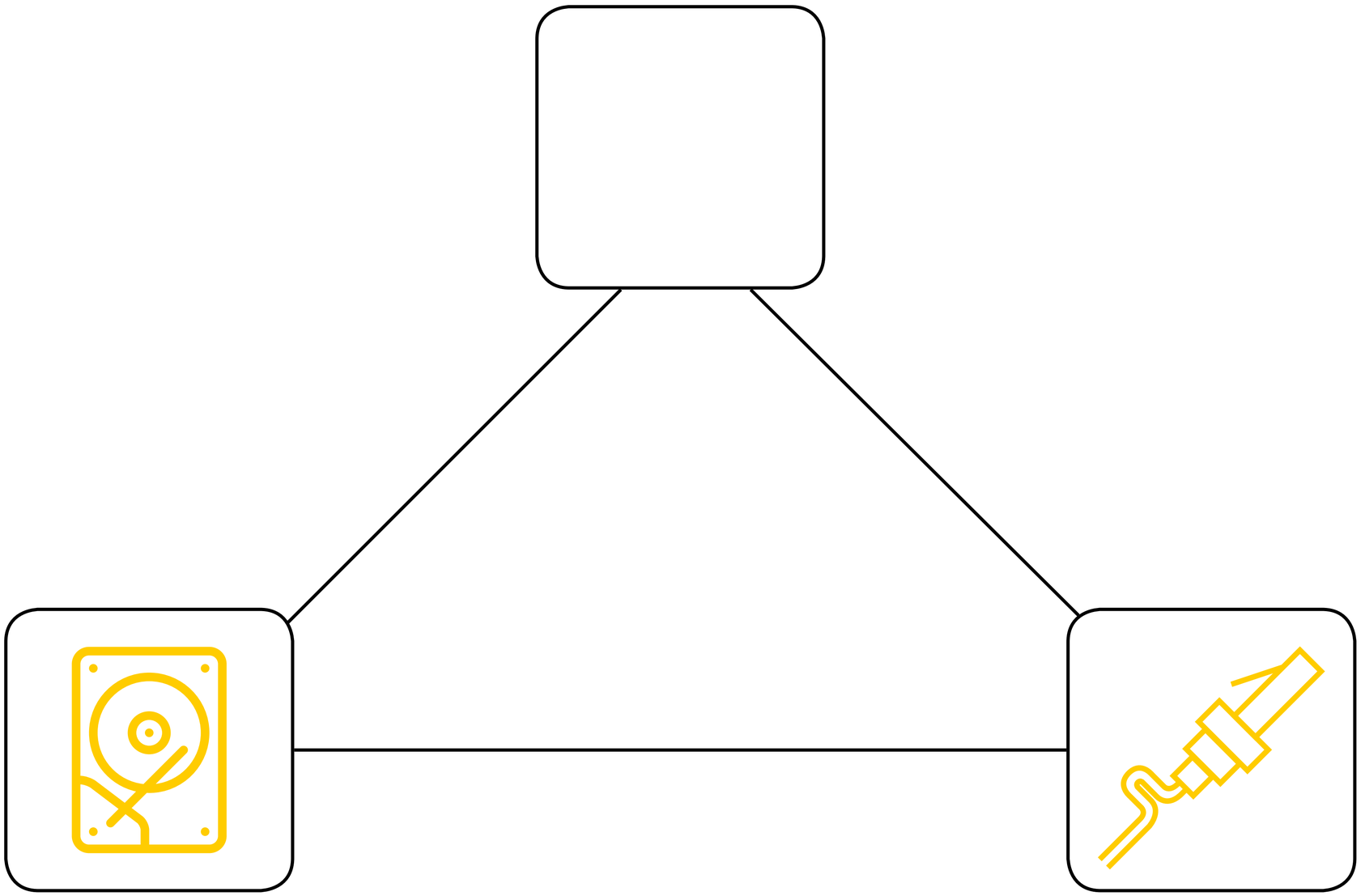
Compression Tradeoff



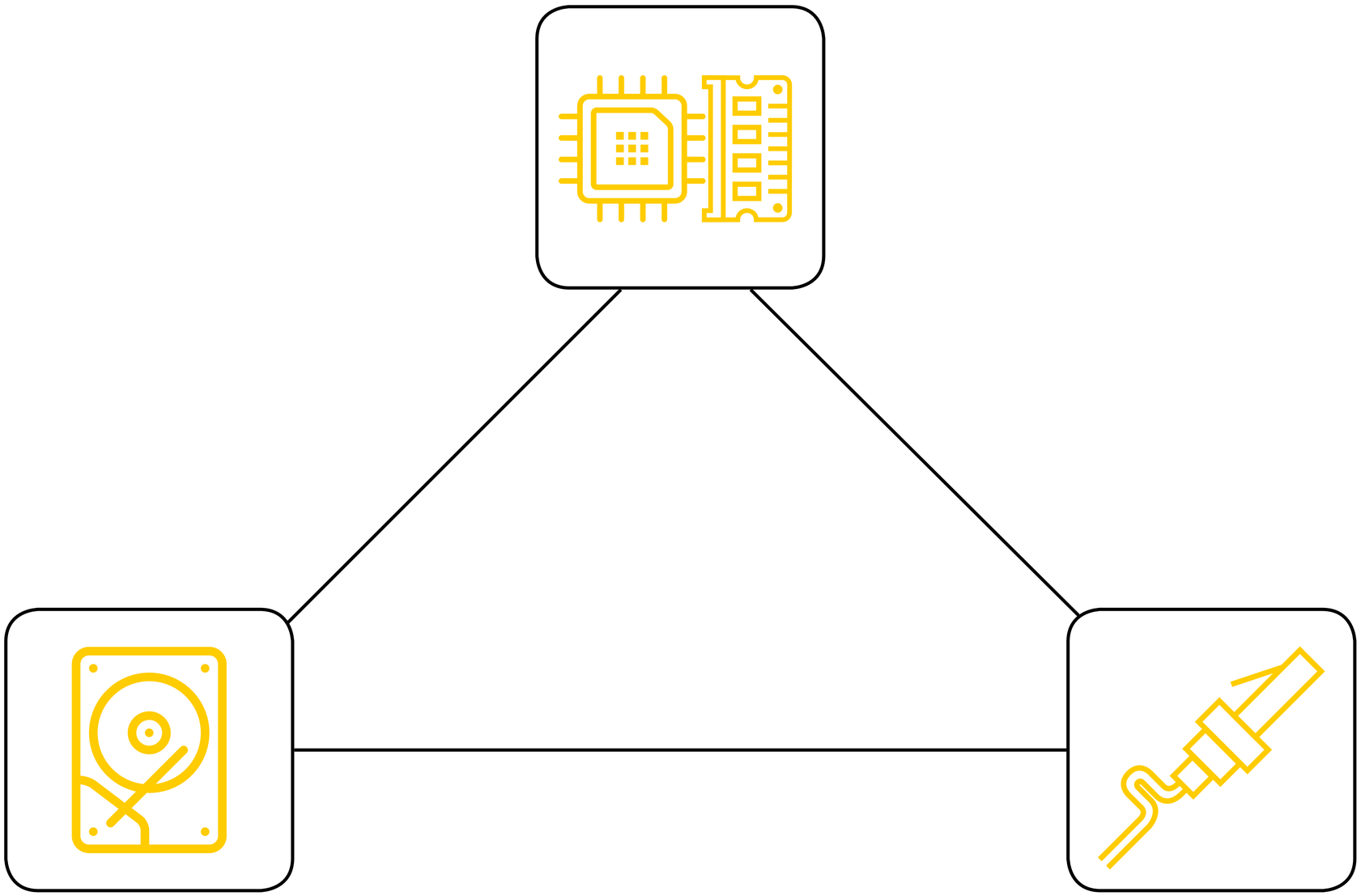
Compression Tradeoff



Compression Tradeoff



Compression Tradeoff



Compression Format	Splittable	Comments

Compression Format	Splittable	Comments
.deflate .gz (gzip)	NO	Uses DEFLATE algorithm

Compression Format	Splittable	Comments
.deflate .gz (gzip)	NO	Uses DEFLATE algorithm
.bz2 (bzip)	YES	more effective than gzip, but slower compression

Compression Format	Splittable	Comments
.deflate .gz (gzip)	NO	Uses DEFLATE algorithm
.bz2 (bzip)	YES	more effective than gzip, but slower compression
.lzo	YES*	decompres way faster than gzip, compres less efficient

Compression Format	Splittable	Comments
.deflate .gz (gzip)	NO	Uses DEFLATE algorithm
.bz2 (bzip)	YES	more effective than gzip, but slower compression
.lzo	YES*	decompres way faster than gzip, compres less efficient
.snappy	NO	faster than LZO for decompression

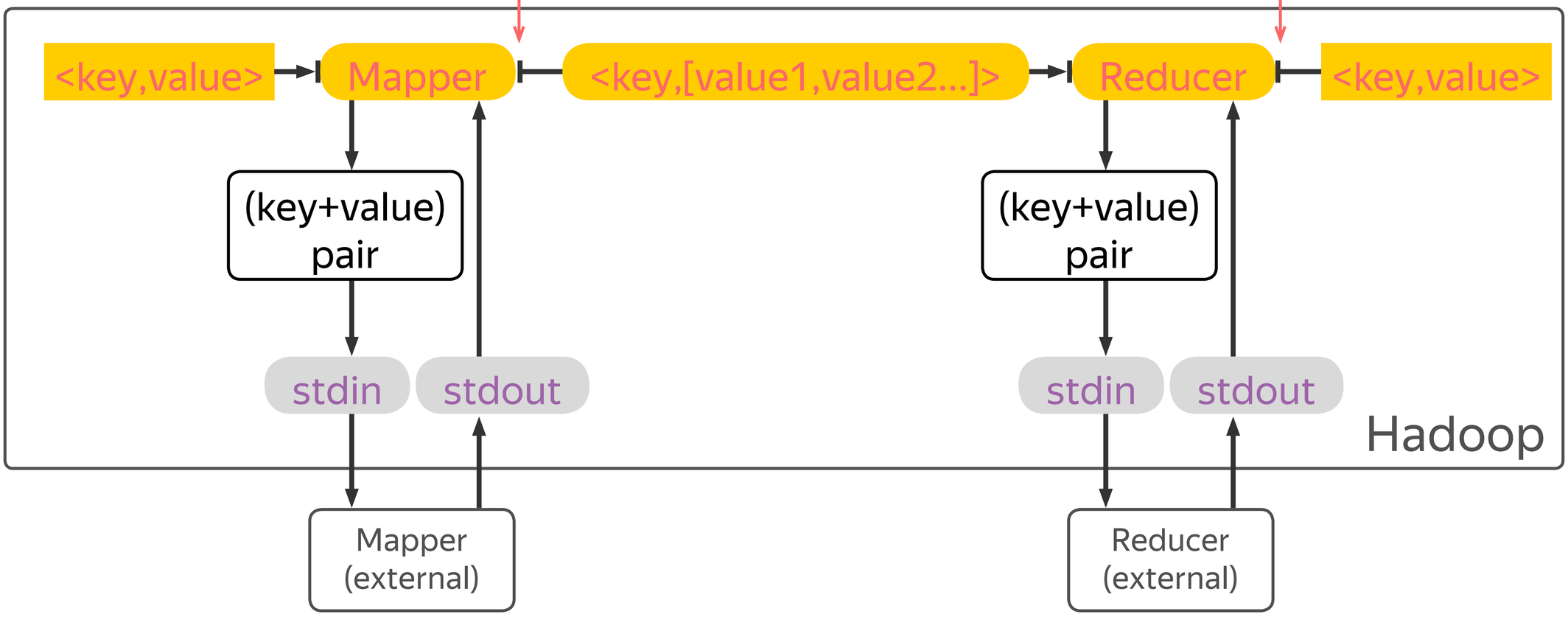
Compression Format	Splittable	Comments
.deflate .gz (gzip)	NO	Uses DEFLATE algorithm
.bz2 (bzip)	YES	more effective than gzip, but slower compression
.lzo	YES*	decompres way faster than gzip, compres less efficient
.snappy	NO	faster than LZO for decompression

flags: -1(optimised for speed)...-9(optimised for space)

Comptession format	Hadoop CompressionCodec
DEFLATE	org.apache.hadoop.io.compress.DefaultCodec
gzip	org.apache.hadoop.io.compress.GzipCodec
bzip2	org.apache.hadoop.io.compress.BZip2Codec
LZO	com.hadoop.compression.lzo.LzopCodec
LZ4	org.apache.hadoop.io.compress.Lz4Codec
Snappy	org.apache.hadoop.io.compress.SnappyCodec

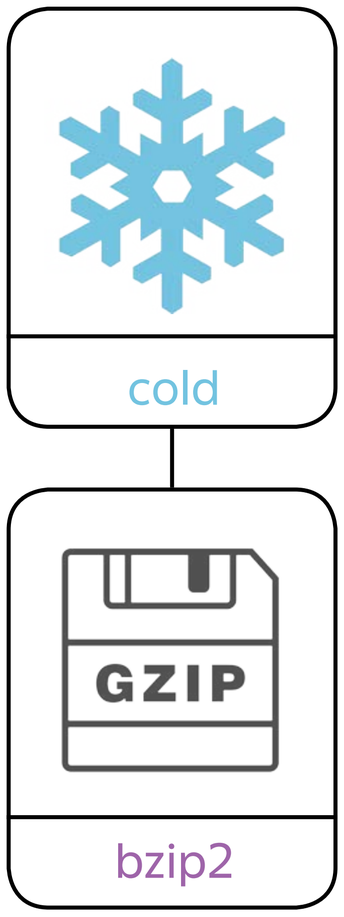
-D mapreduce.compress.map.output=true
-D mapreduce.map.output.compression.codec=...

-D mapreduce.output.compress=true
-D mapreduce.output.compression.codec=...

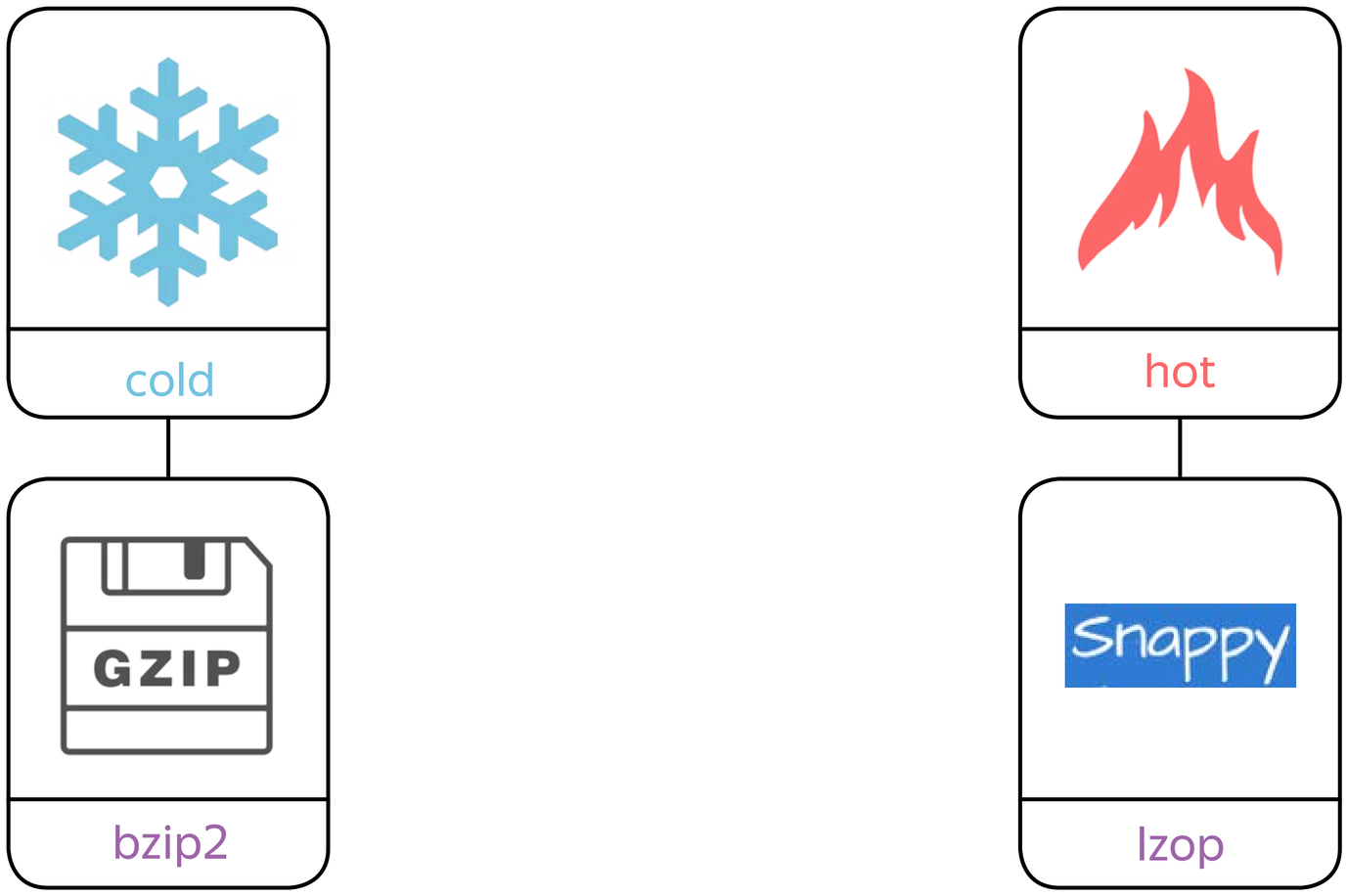


Summary

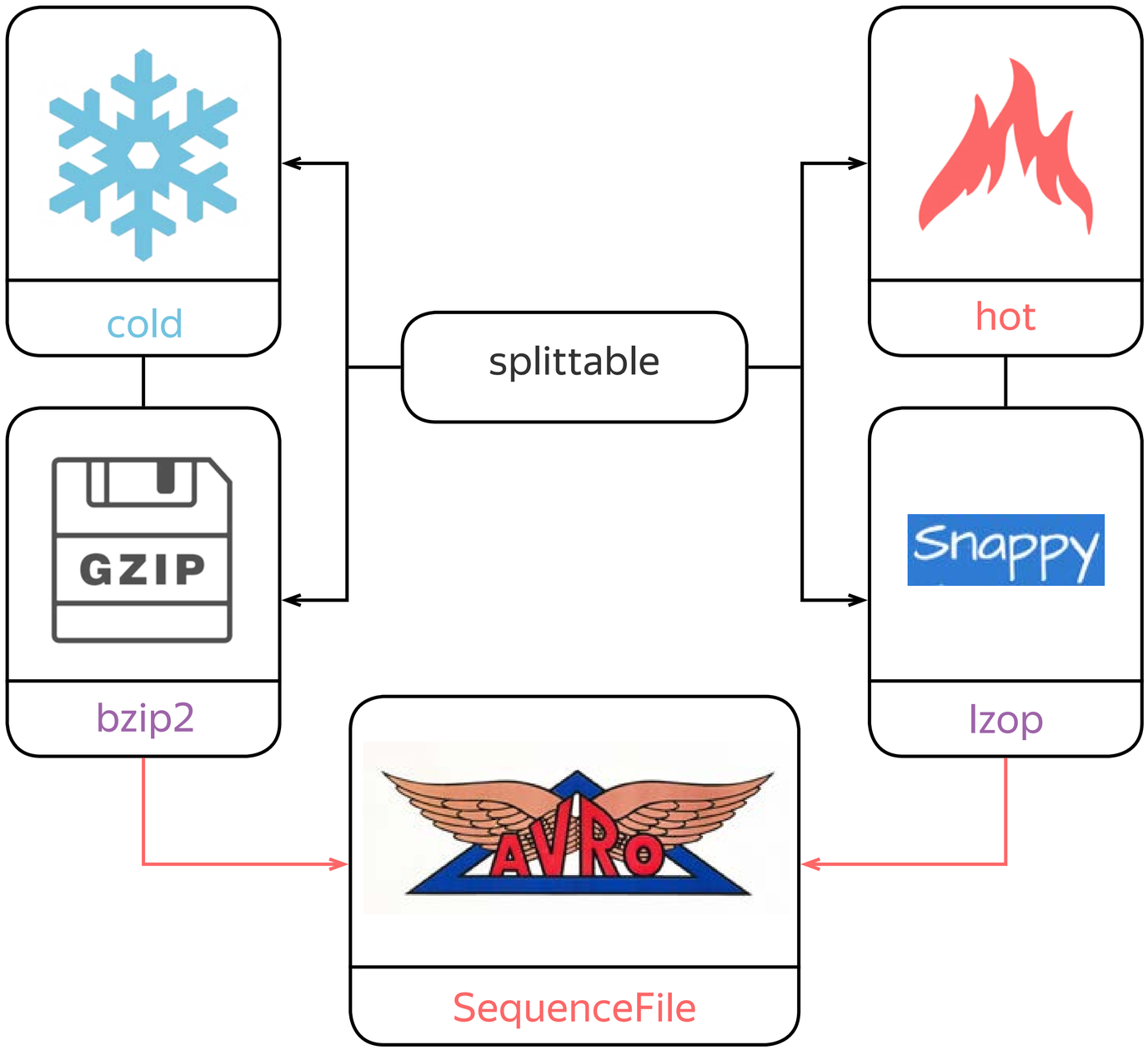
Summary



Summary

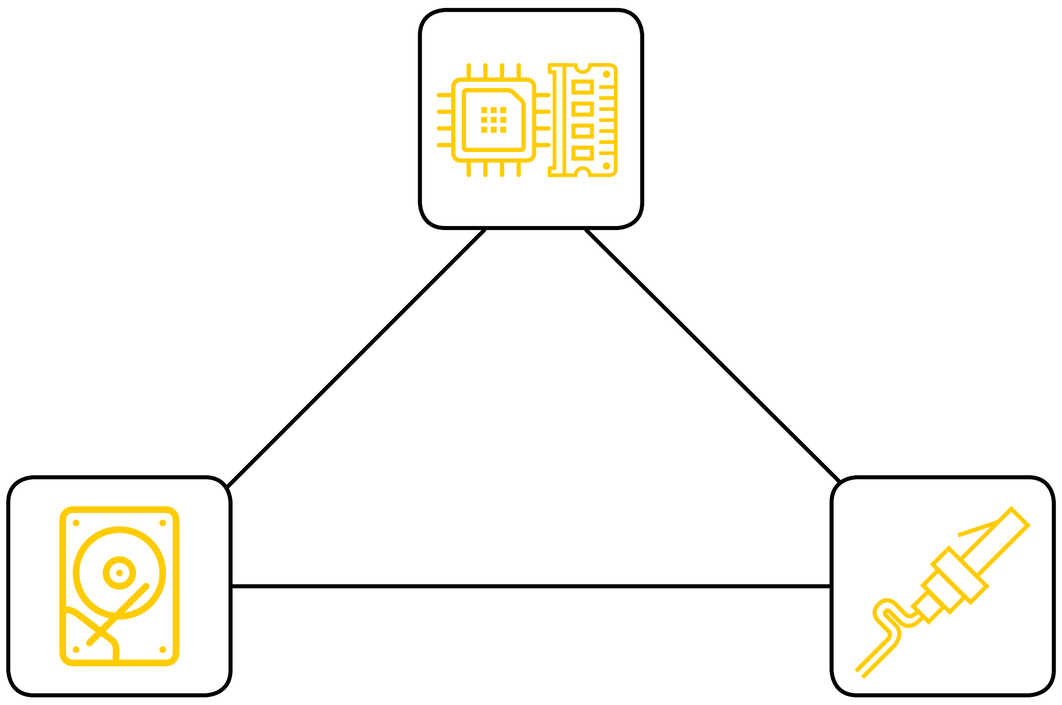


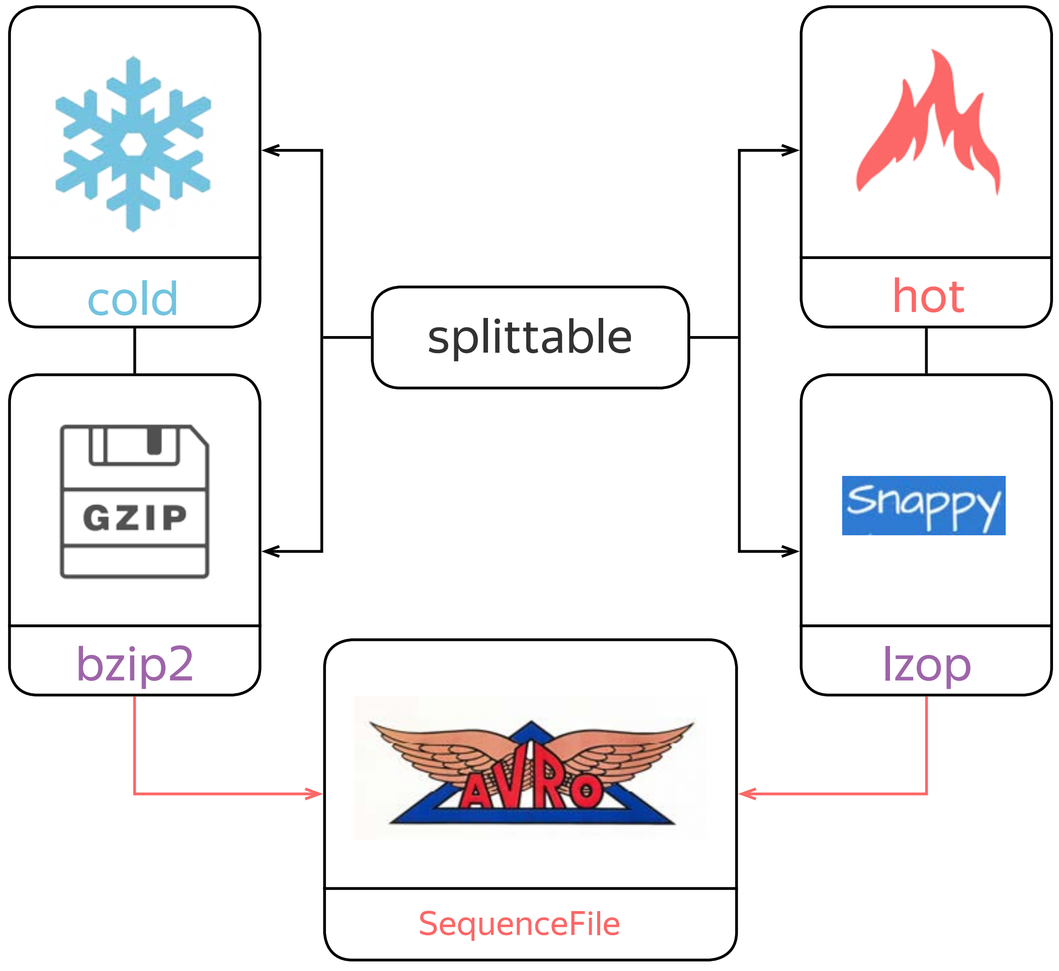
Summary


cold
hot
splittable
GZIP
Snappy
bzip2
lzop
AVRO
SequenceFile

Summary

› you know where and how to use compression to
  › optimise MapReduce Application





BigDATAteam