**Compression**

Compress techniques reduce the cost of storage and processing time.

* A large overhead in processing large data is disk and network i/o. reducing the amount of data that need to be read and write to disk can significantly decrease overall processing time.
* Compression add CPU load but for most cases it is more offset by the saving in i/o.
* All compression don’t support splitability. So, for hive we need to chose those compression which splitability.
* Some compression are optimized for storage and other for storage.

**Technique for compression:**

* Snappy
* Lzo
* Gzip
* Bzip2

**Snappy:**

* Developed by Google for high compression speeds with reasonable compression.
* Optimized for speed than storage.
* Snappy by default is not splitable but as Avro, Parquet, and ORC are container-based format and splitable, So, Snappy will also be splitable with them.

**Lzo:**

* Optimized for speed like Snappy.
* Lzo compressed files are splitable but require additional indexing step.
* As it creates splitable files. So, it is a good fit for text files.
* Lzo license prevents it being distributed in Hadoop, require a separate install.

**Gzip:**

* Very good compression but speed is slow. (provides 2.5 times better compression than Snappy)
* Gzip is also not splitable.
* Sometimes Gzip even can give a better performance bez Gzip compresses file takes few blocks, So, fewer tasks are required for processing same data. So, using smaller blocks with Gzip can lead to better performance.

**Bzip2:**

* Good compression but slower processing.
* Inherently splitable
* Compresses data 9% better than Gzip but 10 times slower than Gzip.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Snappy** | **Lzo** | **Gzip** | **Bzip2** |
| Speed | Highest | High | Average | Very slow |
| Compression | Average | Average | Good | Very Good |
| Splitable | No | Yes | No | Yes |
|  |  |  |  |  |