Compression and Archival with Hadoop

Haddop provides ways to compress and archive files at various stages.

Compression:

Scenario 1.

The output of reduce phase can be configured to be compressed as below:

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml (3.2.1 stable as of 31st July, 2020)

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.io.*;

Configuration conf = new Configuration();

conf.addResource(new Path("/opt/hadoop/etc/hadoop/core-site.xml"));
conf.addResource(new Path("/opt/hadoop/etc/hadoop/hdfs-site.xml"));
conf.addResource(new Path("/opt/hadoop/etc/hadoop/mapred-site.xml"));
conf.addResource(new Path("/opt/hadoop/etc/hadoop/yarn-site.xml"));

/* To be used with maven install to generate a jar file to target location and wish to run MR using eclipse run configuration. Know that Hadoop components have to be up and running. */

conf.set("mapreduce.job.jar",
"/Users/JigarPandya/eclipse-workspacejun2020/MyFirstHadoopWC/target/MyFirstMavennArtifactID-0.0.1-SNAPSHOT.jar");
```

property: mapreduce.output.fileoutputformat.compress
false
Should the job outputs be compressed?

conf.setBoolean("mapreduce.output.fileoutputformat.compress",true);

property: mapreduce.output.fileoutputformat.compress.codec
org.apache.hadoop.io.compress.DefaultCodec
If the job outputs are compressed, how should they be compressed?

conf.setClass("mapreduce.output.fileoutputformat.compress.codec",org.apache.hadoop.io.compress.GzipCodec.class, org.apache.hadoop.io.compress.CompressionCodec.class);

Above shall generate a reducer output file with extension .gz accordingly.

```
Jigar-Pandyas-MacBook:output_sample JigarPandya$ tree
.
├── compressed
│   └── output
│       ├── _SUCCESS
│       └── part-r-00000.gz
└── regular
    └── output
        ├── _SUCCESS
        └── part-r-00000

4 directories, 4 files
Jigar-Pandyas-MacBook:output_sample JigarPandya$
```

You may explore more options about compression from the

CompressionCodec

Hadoop Archives

CLI

File extension .har and a scheme “har” supported to deal with archives in hadoop filesystem.

hdfs archive -archiveName arfile.har /path/to/data1 /path/to/folder2

hdfs fs -lsr har:///path/to/harfile/arlfile.har

P.s. As of date you might have observed that online storage drives let you download a folder containing subfolders and files as a zip file. It can be thought of as a real time application of compress and archival support of hadoop library. Know that it is for user convenience also and to save on network bandwidth too.

Documented By:
Jigar M. Pandya
https://www.linkedin.com/in/jigar-pandya

Document Last updated: 31st July, 2020.