HFile Direct Read vs. HDFS SequenceFile

200,000 rows (value size: 500 – 8,000 bytes/row, 10 columns/row, no compression)
AWS EC2 m1.xlarge (4-core processor, 15 GB RAM) x 1 node

throughput
[MB/sec]

Note: HDFS SequenceFile.Reader, without compression, doesn't do block-read but record-read, resulting relatively poor performance. Enabling block-read may help.

100.00

88.79

HFile Direct V2   82.35

71.98

V1

75.00

54.53

50.00   43.77   46.69   47.05

39.79

HDFS SeqFile (io.file.buffer.size = 128KB)

30.67

28.88

25.00   20.49   22.96

16.66

5.49   13.28

Region Server (scan.cachedrows = 500)

0

500   1,000   2,000   4,000   8,000

value size [bytes/row]