Parquet: java.lang.IllegalArgumentException:Negative capacity

Last updated by | Veena Pachauri | Mar 8, 2023 at 11:22 PM PST

Symptom ErrorCode=ParquetJavaInvocationException,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=An error occurred when invoking java, message: java.lang.IllegalArgumentException:Negative capacity: -2146882864 total entry:21 java.nio.Buffer.<init>(Unknown Source) java.nio.ByteBuffer.<init>(Unknown Source) java.nio.ByteBuffer.<init>(Unknown Source) java.nio.MappedByteBuffer.<init>(Unknown Source) java.nio.DirectByteBuffer.<init>(Unknown Source) java.nio.ByteBuffer.allocateDirect(Unknown Source) org.apache.parquet.hadoop.codec.SnappyCompressor.setInput(SnappyCompressor.java:99) org.apache.parquet.hadoop.codec.NonBlockedCompressorStream.write(NonBlockedCompressorStream.java:48) org.apache.parquet.bytes.CapacityByteArrayOutputStream.writeToOutput(CapacityByteArrayOutputStream.java:219) org.apache.parquet.bytes.CapacityByteArrayOutputStream.writeTo(CapacityByteArrayOutputStream.java:239) org.apache.parquet.bytes.BytesInput $CapacityBAOSBytesInput.writeAllTo(BytesInput.java:392) org.apache.parquet.bytes.BytesInput SequenceBytesIn.writeAllTo(BytesInput.java:283) org.apache.parquet.hadoop.CodecFactory $HeapBytesCompressor.compress(CodecFactory.java:165) org.apache.parquet.hadoop.ColumnChunkPageWriteStore ColumnChunkPageWriter.writePage(ColumnChunkPageWriteStore.java:98) org.apache.parquet.column.impl.ColumnWriterV1.writePage(ColumnWriterV1.java:148) org.apache.parquet.column.impl.ColumnWriterV1.flush(ColumnWriterV1.java:236) org.apache.parquet.column.impl.ColumnWriteStoreV1.flush(ColumnWriteStoreV1.java:122) org.apache.parquet.hadoop.InternalParquetRecordWriter.flushRowGroupToStore(InternalParquetRecordWriter.java:169) org.apache.parquet.hadoop.InternalParquetRecordWriter.close(InternalParquetRecordWriter.java:109) org.apache.parquet.hadoop.ParquetWriter.close(ParquetWriter.java:301) com.microsoft.datatransfer.bridge.parquet.ParquetWriterBridge.close(ParquetWriterBridge.java:29) .,Source=Microsoft.DataTransfer.Richfile.ParquetTransferPlugin,''Type=Microsoft.DataTransfer.Richfile.JniExt.JavaBridgeException,Message=,Source=Microsoft.DataTransfer.Richfile.HiveOrcBridge

RootCause

The issue happens when single column size is too large, there will be a integer overflow exception throw from parquet-mr library (not owned by us).

This is the same error: https://issues.apache.org/jira/browse/PARQUET-1632 ⟹

Mitigation

1. Reducing the row count, copy by smaller batch.
2. Find the largest column, and reduce the column size (<20M), if it's a large json in a single column, try to split it to multiple columns.