Spark shuffle and reduce phases

PeakMemoryUsage of spill = Size(AppendOnlyMap) + Size(SerializeBuffer) = 3.6 GB + 581 MB = 4.2 GB

In-memory AppendOnlyMap (3.6 GB)

stored 60 aggregated shuffled records

K	aggregated(V) (each one is about 60MB)
K1	V1 = CompactBuffer(R1,R2, Ri, Rn)
K2	V2 = CompactBuffer(R1,R2, Ri, Rn)
K3	V3 = CompactBuffer(R1,R2, Ri, Rn)
K4	V4 = CompactBuffer(R1,R2, Ri, Rn)
K5	V5 = CompactBuffer(R1,R2, Ri, Rn)
K6	V6 = CompactBuffer(R1,R2, Ri, Rn)
Ki	V7 = CompactBuffer(R1,R2, Ri, Rn)
K60	V8 = CompactBuffer(R1,R2, Ri, Rn)

Only a part of the shuffled records can be aggregated in memory

*60 * 60MB = 3.6 GB*

The size of the AppendOnlyMap has achieved the spill threshold, so the spill action is triggered

writeAnd serialize()

SerializeBuffer

Since 60 < batchSize(10,000), all the 60 records are serialized in the buffer

*stores 60 serialized records = 60 * 9.7 MB = 581 MB*

K60 V60	Ki Vi	K3 V3	K2 V2	K1 V1

flush()

Spilled records are stored on disk

*stores 60 serialized records = 60 * 9.7 MB = 581 MB*

K60 V60	Ki Vi	K3 V3	K2 V2	K1 V1

Shuffled records

K	V
K1	R1
K2	R2
Ki	Ri
Kj	Rj
Km	Rm

7,540,235 (903.4MB)

K1	R1
Ki	Ri
Kn	Rn

*60 * 27.4MB = 1.6 GB*

PeakMemoryUsage of Memory-Disk-Merge = Size(AppendOnlyMap) + Size(DeSerializeBuffer) = 1.6 GB + (2.8 GB + 1 GB) = 5.4 GB

In-memory AppendOnlyMap (1.6 GB)

stored 60 aggregated shuffled records

K	aggregated(V)
K1	V1 = CompactBuffer(R1, Ri, Rn)
K2	V2 = CompactBuffer(R1,Ri, Rn)
K3	V3 = CompactBuffer(R1,Ri, Rn)
K4	V4 = CompactBuffer(R1,Ri, Rn)
K5	V5 = CompactBuffer(R1,Ri, Rn)
K6	V6 = CompactBuffer(R1,Ri, Rn)
Ki	Vi = CompactBuffer(R1,Ri, Rn)
K48	V47 = CompactBuffer(R1, Ri, Rn)
Kj	Vj = CompactBuffer(R1, Ri, Rn)
K60	V60 = CompactBuffer(R1, Ri, Rn)

①

DeSerializeBuffer
the spilled records are read back into the buffer

The spilled record are read back and deserialized

K1	V1 = CompactBuffer(R1, Ri, Rn)
K2	V2 = CompactBuffer(R1,Ri, Rn)
K3	V3 = CompactBuffer(R1,Ri, Rn)
K4	V4 = CompactBuffer(R1,Ri, Rn)
K5	V5 = CompactBuffer(R1,Ri, Rn)
K6	V6 = CompactBuffer(R1,Ri, Rn)
Ki	Vi = CompactBuffer(R1,Ri, Rn)
K48	V48 = CompactBuffer(R1, Ri, Rn)

Deserialized records
*= 47 * 60MB*
= 2.8 GB

OOM during reading this record

Buffer[] references (1GB)

Old refs: HandleList[46,137,343] + Object[46,137,343] = 439MB
New refs: byte[92,274,687] + Object[92,274,687] = 450 MB

②

OOM

Memory-Disk-Merge
(without removing the recode in the AppendOnlyMap and deserializer)

K1	V1 = CompactBuffer(R1, Ri, Rn, R1, Ri, Rn)
Ki	Vi = CompactBuffer(R1, Ri, Rn, R1, Ri, Rn)
K48	V48 = CompactBuffer(R1, Ri, Rn, R1, Ri, Rn)

K1	V1
Ki	Vi

③

HDFS

Serialize and write each output record onto the HDFS