DATA+AI SUMMIT 2022

Spark AQE SkewedJoin Optimization and Practice in ByteDance

Lietong Liu
Software Engineer, ByteDance

ORGANIZED BY databricks

Who am I?

- Lietong Liu

Software Engineer in Data Engine team , Bytedance

- LakeHouse Analytics

LAS, which is short for LakeHouse Analytics. Spark is one of the most important engine for LAS. The improvements demonstrated in this session are launched on LAS and you can try LAS on volcengine, which is a Chinese public cloud platform

https://www.volcengine.com/product/las

Agenda

- Motivation
- Enhancements
- Practice in ByteDance
- User guidance
- Summary

Motivation

Spark AQE SkewedJoin

Table A inner joins Table B, and the 0th partition (A0) in Table A is a skewed partition. Under normal circumstances, A0 will join the 0th partition (B0) in Table B. Because A0 is skewed at this time, Task 0 will become a long-tail task.

Spark AQE SkewedJoin

Spark AQE will split the data of A0 into N copies, and deal with the partition with N tasks. Each task will only read the shuffle output files of several MapTasks, as shown in the figure below, while A0-0 will only read the data that belongs to A0 in Stage0#MapTask0. Then, these N tasks will read the data of Partition 0 in Table B and join.

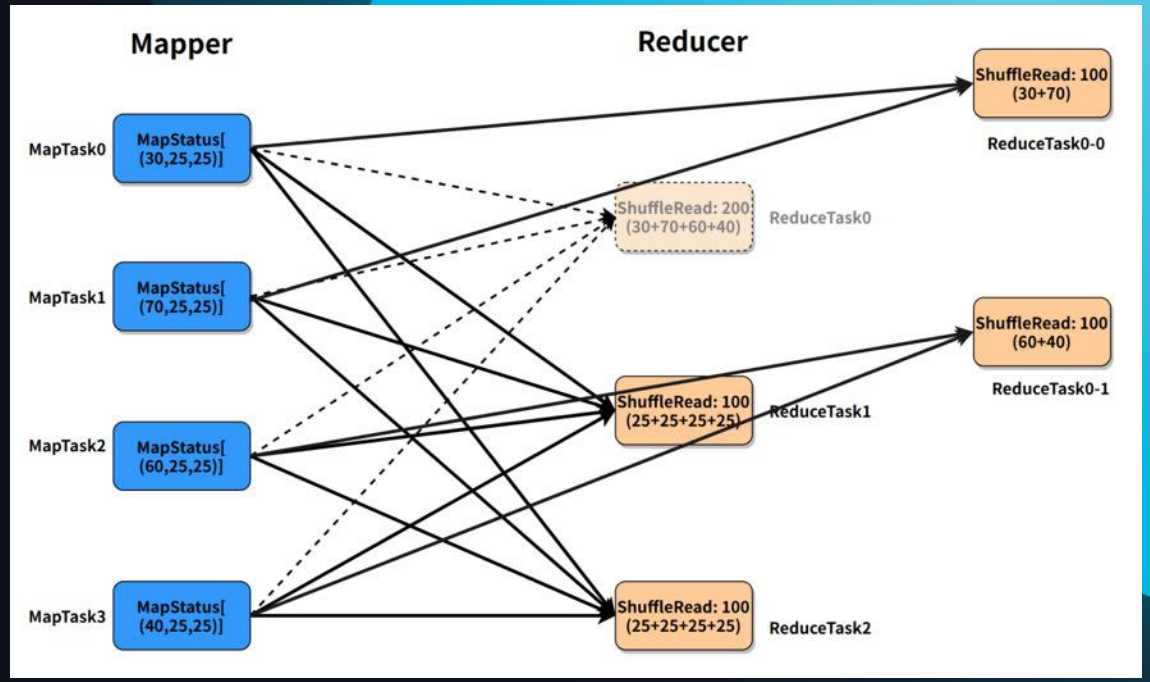
Motivation

- Inaccurate statistics would cause failures in identifying data skew
- Uneven segmentation would leads to unsatisfactory processing
- Do not support complex scenarios such as continuous joins in the same stage

Enhancements

Improve the ability to identify data skew

- ReduceTask0 can be divided into two parts; ReduceTask0-0 reads the data of MapTask0 and MapTask1, and ReduceTask0-1 reads the data of MapTask2 and MapTask3. After splitting, the ShuffleRead of the two tasks is 100.



Improve the ability to identify data skew

Default parameters related to MapStatus

Configuration Parameters	Default Value	Meaning
spark.shuffle.minNumPartitionsToHighlyCompress	2000	Number of partitions to determine if MapStatus should use HighlyCompressedMapStatus
spark.shuffle.accurateBlockThreshold	100M	Threshold in bytes above which the size of shuffle blocks in HighlyCompressedMapStatus is accurately recorded. This helps to prevent OOM by avoiding underestimating shuffle block size when fetch shuffle blocks.

Improve the ability to identify data skew

Stage summary

Summary Metrics for 20000 Completed Tasks

Metric(percentile)	Min	25th	Median	75th	90th	95th	99th	Max
Duration	0 ms	6.6 min	7.8 min	9.1 min	11 min	12 min	15 min	47 min
GC Time	0 ms	5 s	10 s	16 s	23 s	29 s	40 s	1.4 min
Shuffle Serial Read Time	0 ms	15.5 h	20.0 h	25.6 h	28.5 h	29.9 h	32.3 h	125.0 h
Shuffle HDFS Read Time	0 ms	23 s	30 s	37 s	44 s	48 s	57 s	7.9 min
Shuffle Read Size / Records	0.0 B / 0	179.1 MB / 5358074	209.9 MB / 6185166	262.6 MB / 7654176	367.4 MB / 10712763	472.4 MB / 13775954	870.9 MB / 25787683	4.5 GB / 135657737
Shuffle Write Time	0 ms	0.9 s	2 s	7 s	9 s	11 s	20 s	2.4 min
Shuffle Write Size / Records	0.0 B / 0	161.9 MB / 4090795	191.6 MB / 4918037	241.4 MB / 6386699	339.9 MB / 9444531	439.1 MB / 12509414	823.9 MB / 24519253	4.3 GB / 134389837
Shuffle HDFS Write Time	0 ms	2 s	3 s	5 s	9 s	13 s	31 s	4.5 min
Shuffle Spill Time	0 ms	8 s	26 s	38 s	57 s	1.3 min	2.4 min	13 min
Shuffle spill (memory)	0.0 B	1920.0 MB	3.9 GB	4.7 GB	7.5 GB	9.1 GB	20.1 GB	115.9 GB
Shuffle spill (disk)	0.0 B	268.3 MB	461.5 MB	558.6 MB	850.1 MB	1089.2 MB	2.3 GB	12.7 GB

AQE Statistics

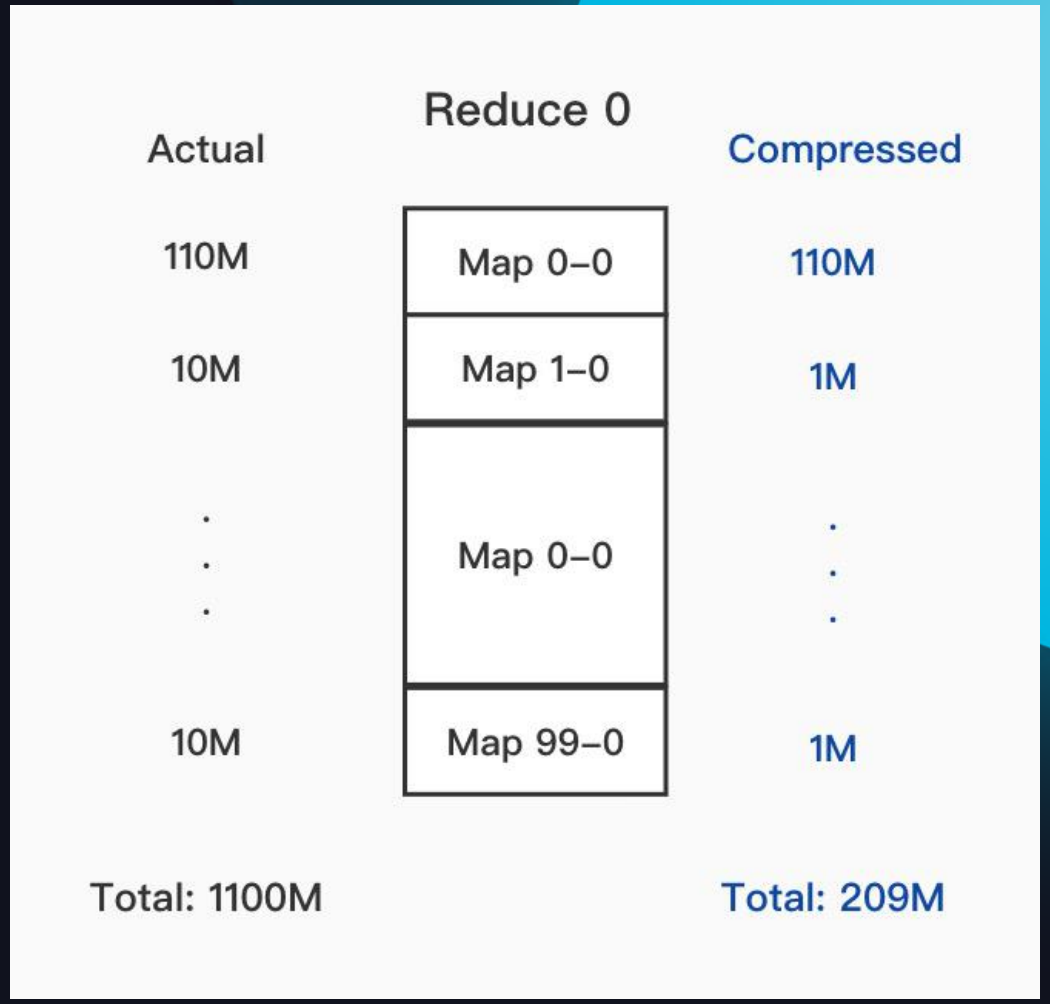
```
INFO HandleSkewedJoin: HandlingSkewedJoin left medSize/rowCounts: (245884546, 6220012) right medSize/rowCounts (20906271, 1266431)
INFO HandleSkewedJoin: left bytes Max : 245884546
INFO HandleSkewedJoin: left row counts Max : 6220012
INFO HandleSkewedJoin: right bytes Max : 20906271
INFO HandleSkewedJoin: right row counts Max : 1266431
INFO HandleSkewedJoin: skewed partition number is 0
```

Improve the ability to identify data skew

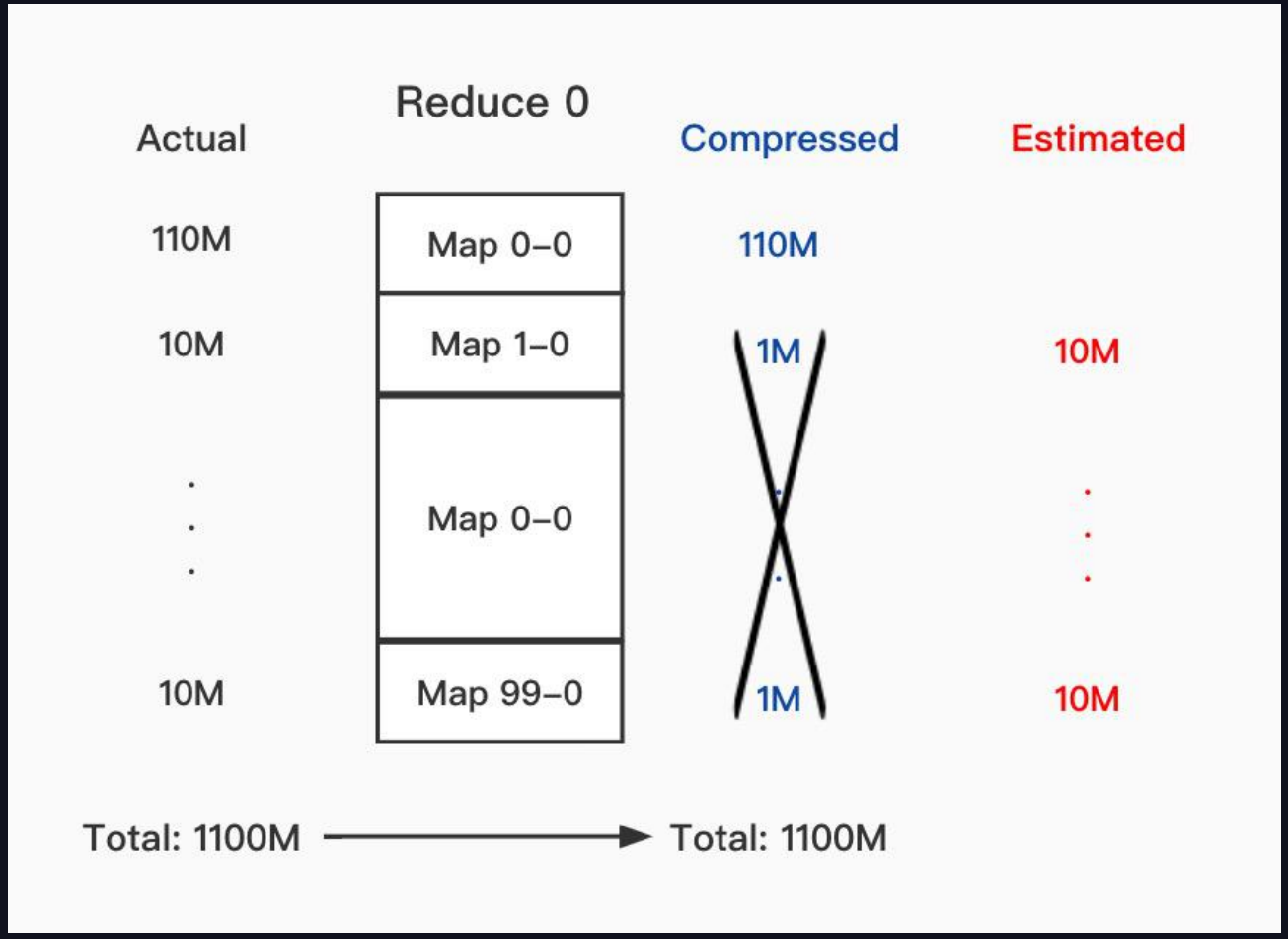
- After receiving the detailed MapStatus, the Driver first uses the data to update the accumulative input data of each ReduceTask, and then compresses the MapStatus
- Caches are used to ensure that each MapStatus will only be decompressed once when the Driver end consumes MapStatus, which greatly reduces the overhead brought by optimization.

Improve the uniformity of segmentation of skewed data

The total data and distribution of each ReduceTask calculated through HighlyCmpressdMapStatus would be quite different from the actual data.



Improve the uniformity of segmentation of skewed data



Using the accurate data size of ReduceTask to deduce the corresponding data size of each MapperTask and obtain the more accurate data distribution.

Improve the uniformity of segmentation of skewed data

Before

Summary Metrics for 11955 Completed Tasks

Metric(percentile)	Min	25th	Median	75th	90th	95th	99th	Max
Duration	0 ms	15 ms	3 s	51 s	2.7 min	4.1 min	6.8 min	34 min
GC Time	0 ms	0 ms	77 ms	1 s	6 s	12 s	39 s	3.2 min
Shuffle Serial Read Time	0 ms	0 ms	1 s	54 s	4.5 min	8.4 min	21 min	1.8 h
Shuffle HDFS Read Time	0 ms	0 ms	4 ms	15 ms	0.1 s	0.4 s	2 s	43 s
Shuffle Read Size / Records	0.0 B / 0	124.0 B / 1	4.4 MB / 118656	121.3 MB / 2966151	354.9 MB / 8510468	585.7 MB / 12670019	2.3 GB / 42337410	9.9 GB / 340550179
Shuffle Write Time	0 ms	0 ms	0 ms	8 ms	0.4 s	0.8 s	3 s	4.6 min
Shuffle Write Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 0	0.0 B / 0	80.6 MB / 1912348	272.8 MB / 6049066	1288.4 MB / 23522099	9.7 GB / 325889050
Shuffle HDFS Write Time	0 ms	0 ms	0 ms	0 ms	2 s	5 s	31 s	3.6 min
Shuffle Spill Time	0 ms	0 ms	0 ms	1 s	6 s	11 s	42 s	4.6 min
Shuffle spill (memory)	0.0 B	0.0 B	0.0 B	512.0 MB	1984.0 MB	3.5 GB	12.5 GB	120.1 GB
Shuffle spill (disk)	0.0 B	0.0 B	0.0 B	95.8 MB	393.6 MB	806.3 MB	3.2 GB	18.8 GB

Mediam ShuffleReadSize: 4MB

Max ShuffleReadSize: 9.9GB

Duration: 2h

After

Summary Metrics for 8924 Completed Tasks

Metric(percentile)	Min	25th	Median	75th	90th	95th	99th	Max
Duration	0 ms	3 s	8 s	21 s	45 s	1.2 min	4.3 min	6.0 min
GC Time	0 ms	0.4 s	2 s	5 s	10 s	15 s	23 s	37 s
Shuffle Serial Read Time	0 ms	6 ms	0.7 s	3 s	11 s	25 s	5.0 min	34 min
Shuffle HDFS Read Time	0 ms	0 ms	0 ms	0 ms	0.2 s	0.5 s	2 s	27 s
Shuffle Read Size / Records	0.0 B / 0	46.3 MB / 1209350	149.1 MB / 3458582	281.0 MB / 6376723	496.3 MB / 9960970	571.9 MB / 11523793	682.1 MB / 13449822	1427.8 MB / 29539823
Shuffle Write Time	0 ms	0 ms	0 ms	2 s	6 s	8 s	13 s	22 s
Shuffle Write Size / Records	0.0 B / 0	0.0 B / 0	0.0 B / 0	76.4 MB / 1758261	276.0 MB / 5747392	450.1 MB / 9343103	618.7 MB / 12229367	1445.4 MB / 26502591
Shuffle HDFS Write Time	0 ms	0 ms	0 ms	2 s	7 s	11 s	21 s	1.0 min
Shuffle Spill Time	0 ms	0 ms	0 ms	0 ms	0 ms	0 ms	10 s	44 s
Shuffle spill (memory)	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B	2048.0 MB	2.6 GB	8.6 GB
Shuffle spill (disk)	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B	198.3 MB	538.5 MB	2.2 GB

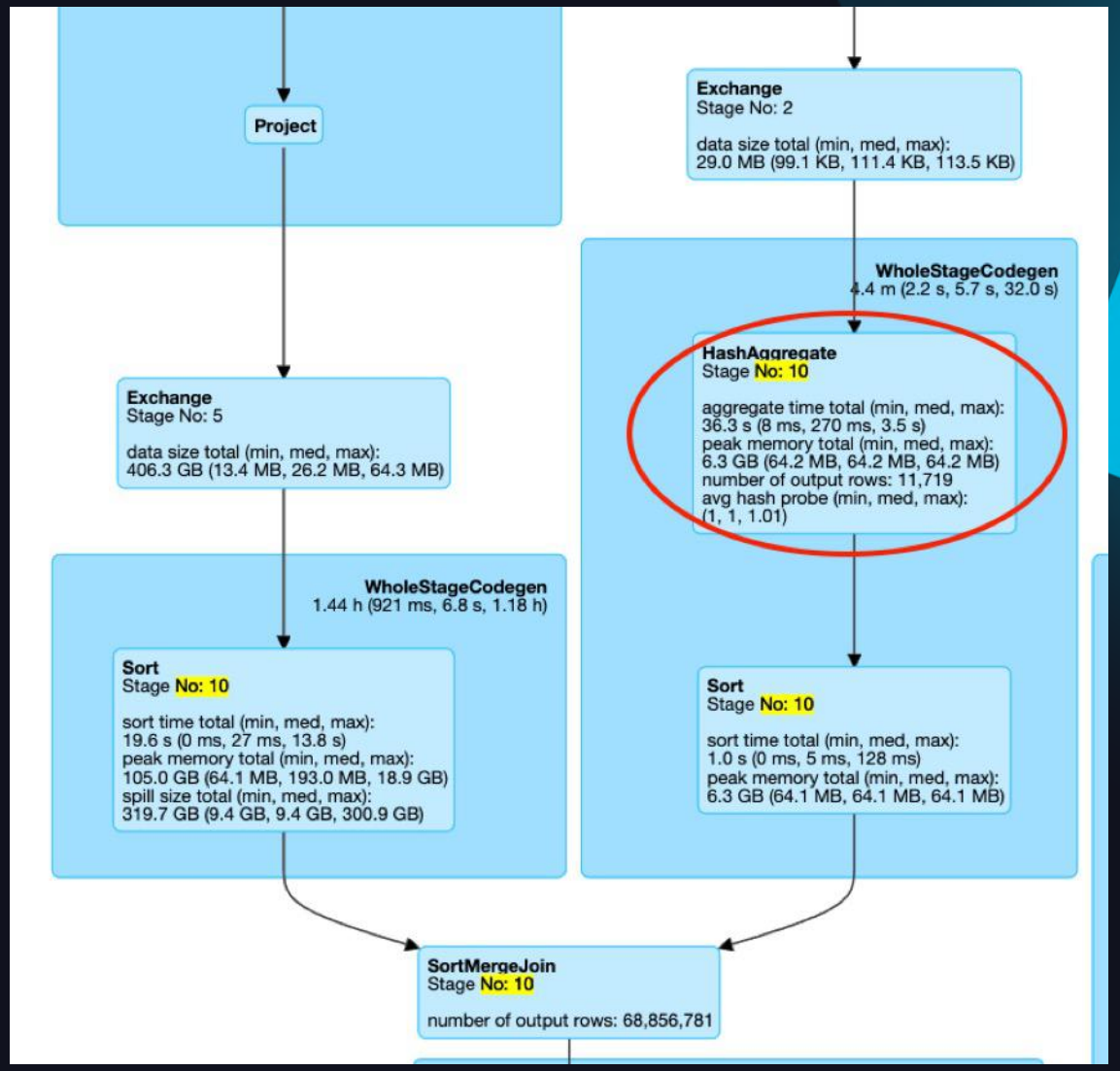
Mediam ShuffleReadSize: 149MB

Max ShuffleReadSize: 1427 MB

Duration: 20min

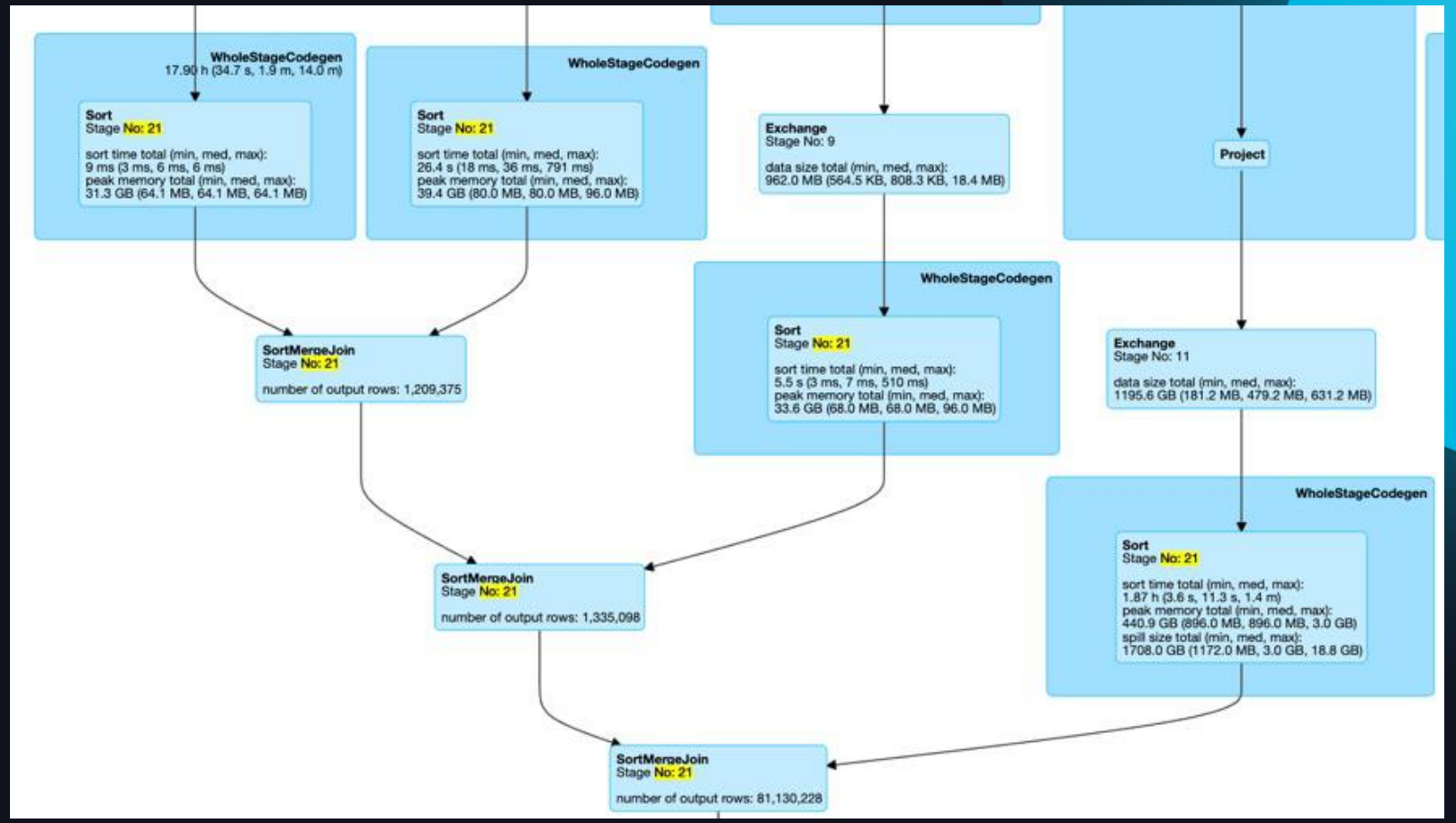
Support more scenarios

JoinWithAggOrWin



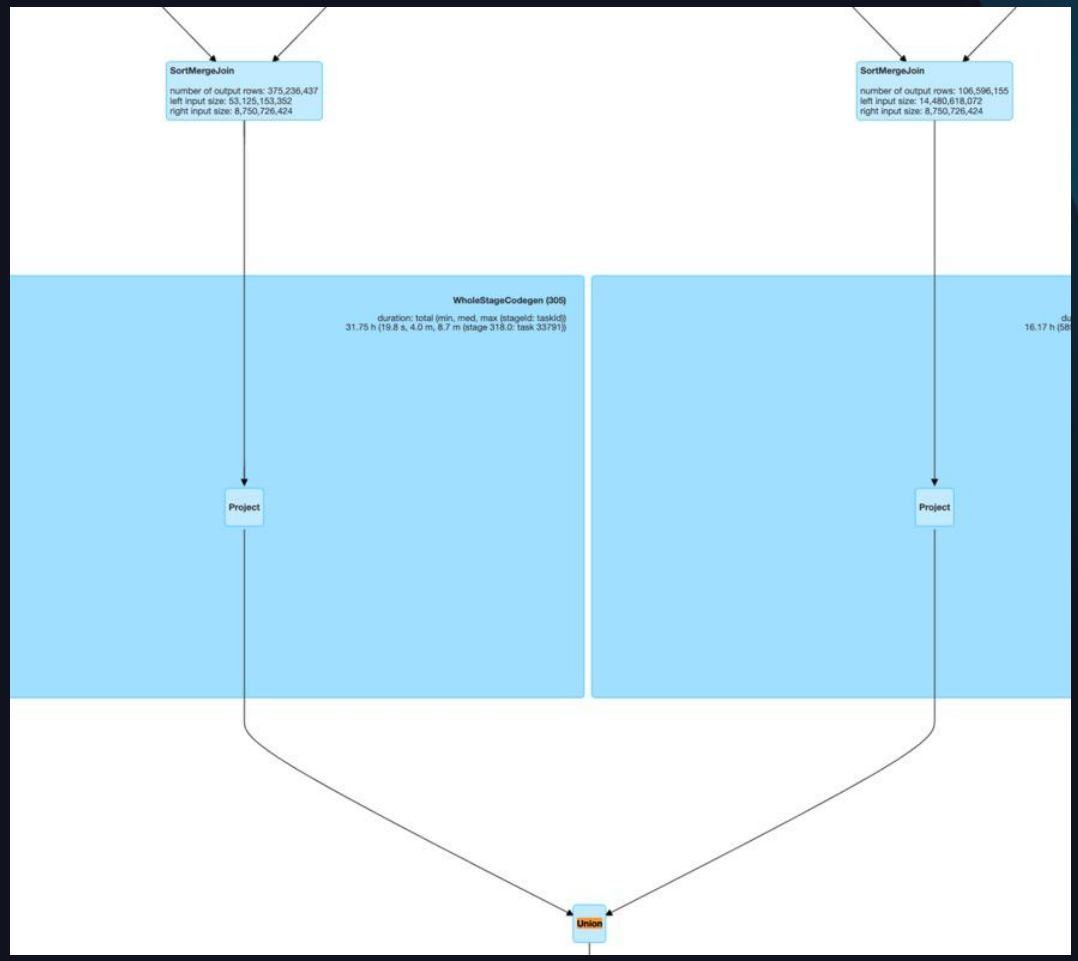
Support more scenarios

MultipleJoin



Support more scenarios

MultipleJoinWithUnion



Support more scenarios

BucketJoin

Support bucketJoin with optimizing data skew in unbucket side.

ShuffledHashJoin

Suppor ShuffledHashJoin with data skew.

MultipleJoinWithAggOrWin

Similar to JoinWithAggOrWin, support MultipleJoinWithAggOrWin.

Practice in ByteDance

Practice in ByteDance

12000+ Spark apps has been covered by the optimization

The performance of the covered apps improved by **35%** by average.

Self-develop optimization contributes about **30%** of covered apps.

User guidance

Where AQE SkewedJoin cannot help

- Can’t help if most data in skewed partition came one mapper
- Can’t help if skewed side of join contains operator which has specified required distribution such as WindowExec or AggregateExec
- Can’t help skewed BroadcastHashJoin

When AQE SkewedJoin does not perform well

- Increase spark.shuffle.minNumPartitionsToHighlyCompress, make sure it equal to numShufflePartitions.
- Reduce spark.shuffle.accurateBlockThreshold, which wil increase memory usage of Driver.
- Reduce spark.sql.adaptive.skewJoin.skewedPartitionFactor

Summary

Summary

- In Bytedance, we
  - Improve the ability to identify data skew
  - Improve the uniformity of segmentation of skewed data
  - Support more complex scenarios
- Over 12000 spark apps benefited from the optimization per day, and the performance of the covered apps improved by 35% by average.
- Introduce the scenarios that SkewedJoin does not take effect and how to modify related configs when SkewedJoin does not perform well.

DATA+AI SUMMIT 2022

Thank you

Lietong Liu

Software Engineer, ByteDance