AN ADAPTIVE EXECUTION ENGINE FOR APACHE SPARK SQL

Carson Wang (carson.wang@intel.com)

Yucai Yu (yucai.yu@intel.com)

Hao Cheng (hao.cheng@intel.com)

Agenda

- Challenges in Spark SQL* High Performance
- Adaptive Execution Background
- Adaptive Execution Architecture
- Benchmark Result

*Other names and brands may be claimed as the property of others.

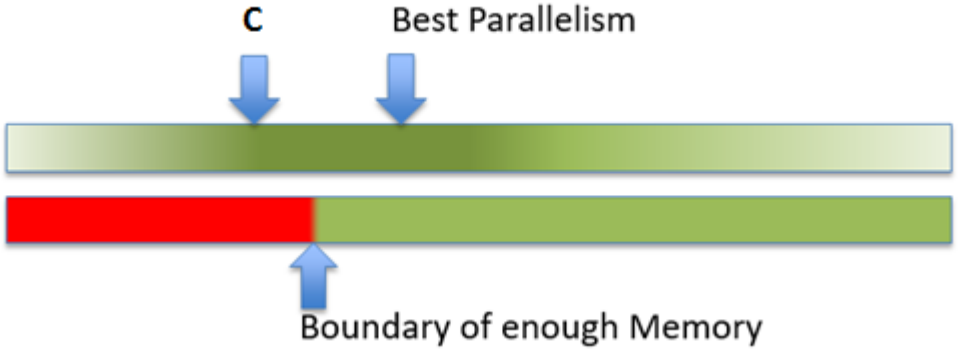
Challenges in Tuning Shuffle Partition Number

- Partition Num **P** = spark.sql.shuffle.partition (200 by default)
- Total Core Num **C** = Executor Num * Executor Core Num
- Each Reduce Stage runs the tasks in (P / C) rounds

Shuffle Partition Challenge 1

- Partition Num Too Small：Spill, OOM
- Partition Num Too Large：Scheduling overhead. More IO requests. Too many small output files
- Tuning method: Increase partition number starting from C, 2C, … until performance begin to drop

  Impractical for each query in production.

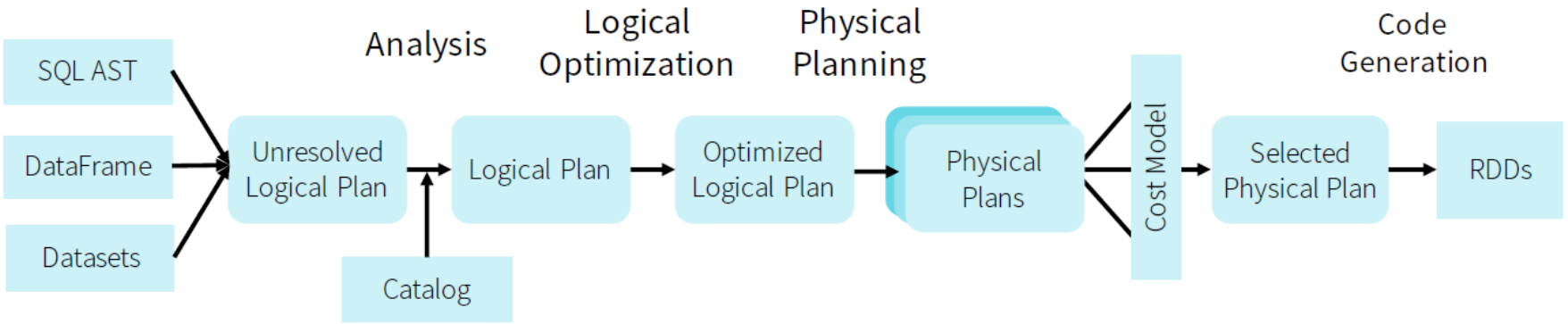


Shuffle Partition Challenge 2

- The same Shuffle Partition number doesn't fit for all Stages
- Shuffle data size usually decreases during the execution of the SQL query

Question: Can we set the shuffle partition number for each stage automatically?

Spark SQL* Execution Plan



- The execution plan is fixed after planning phase.

*Other names and brands may be claimed as the property of others.

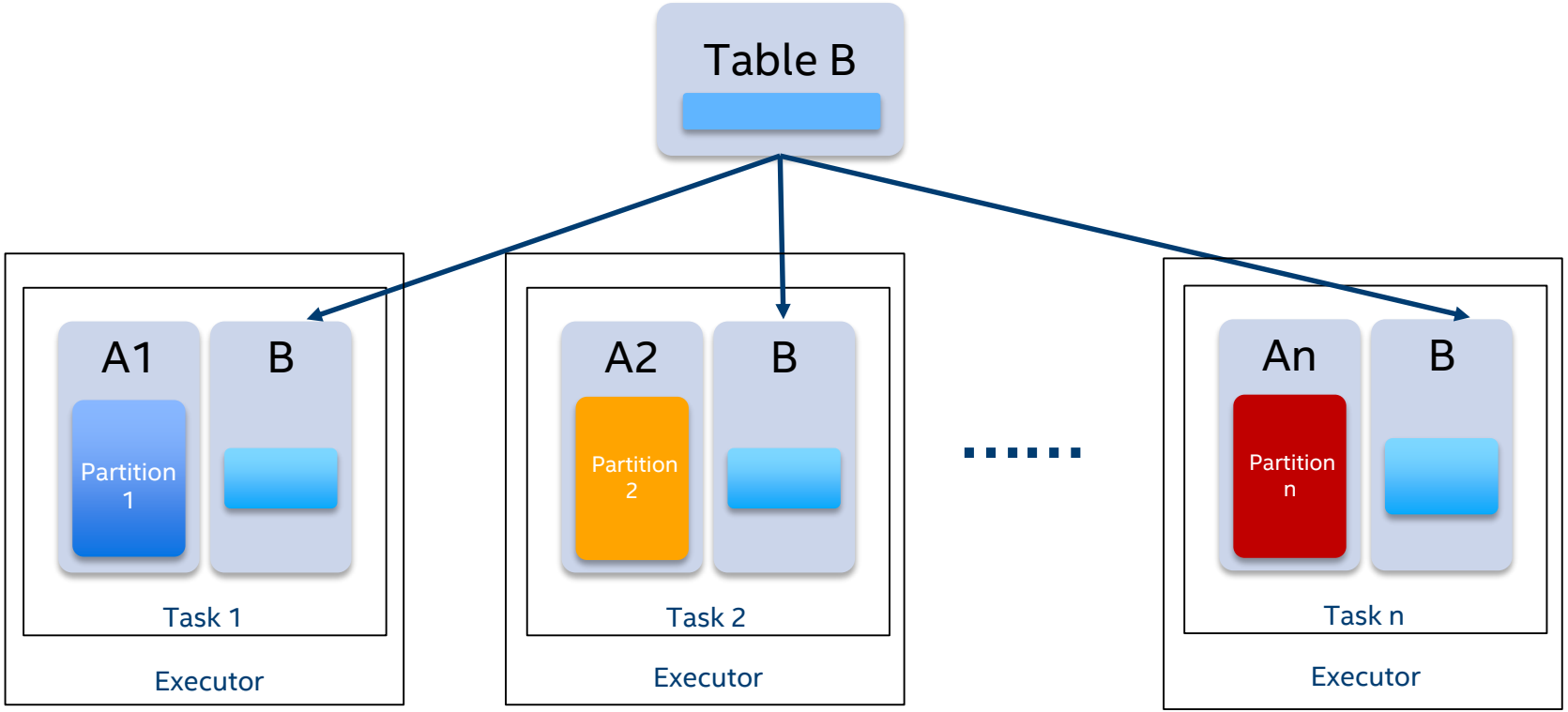
Image from: https://databricks.com/blog/2015/03/24/spark-sql-graduates-from-alpha-in-spark-1-3.html

Spark SQL* Join Selection

```
SELECT xxx

FROM A

JOIN B

ON A.Key1 = B.Key2
```

*Other names and brands may be claimed as the property of others.

Broadcast Hash Join

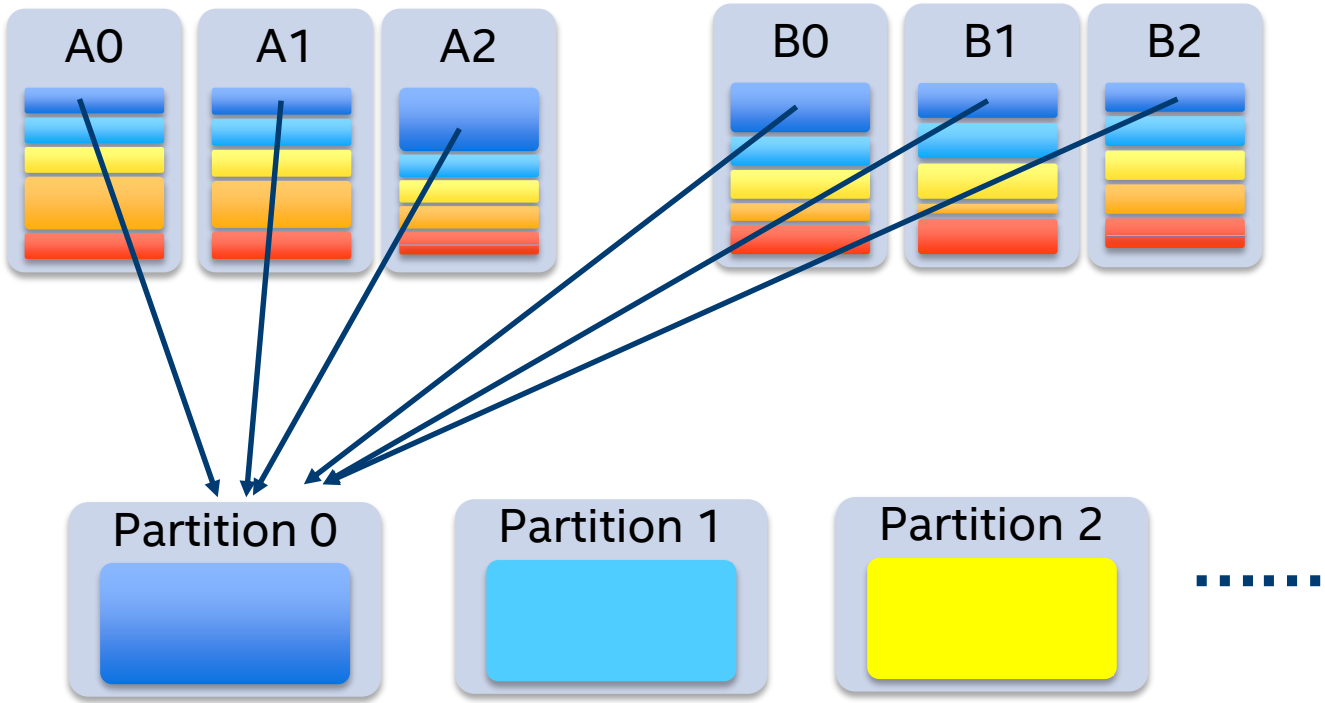


Shuffle Hash Join / Sort Merge Join

MAP

SHUFFLE

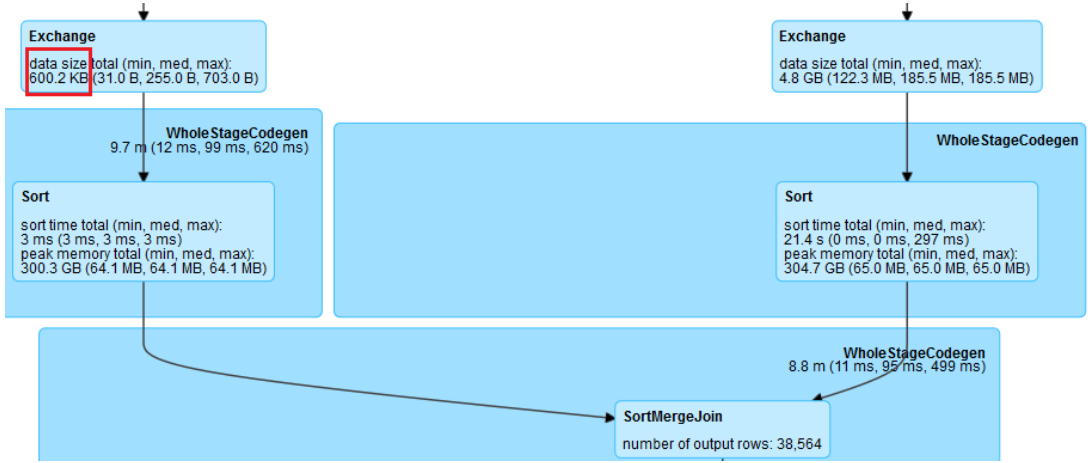
REDUCE



Spark SQL* Join Selection

- spark.sql.autoBroadcastJoinThreshold is 10 MB by default

- For complex queries, a Join may takes intermediate results as inputs. At planning phase, Spark SQL* doesn't know the exact size and plans it to SortMergeJoin.



Question: Can we optimize the execution plan at runtime based on the runtime statistics ?

Data Skew in Join

- Data in some partitions are extremely larger than other partitions.
- Data skew is a common source of slowness for Shuffle Joins.

Summary Metrics for 5600 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	15 ms	2 s	3 s	5 s	4.0 min
Scheduler Delay	0 ms	2 ms	2 ms	3 ms	1 s
Task Deserialization Time	3 ms	5 ms	6 ms	7 ms	0.4 s
GC Time	0 ms	0 ms	0.2 s	0.3 s	35 s
Result Serialization Time	0 ms	0 ms	0 ms	0 ms	3 ms
Getting Result Time	0 ms	0 ms	0 ms	0 ms	0 ms
Peak Execution Memory	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B
Shuffle Read Blocked Time	0 ms	44 ms	0.2 s	0.6 s	31 s
Shuffle Read Size / Records	0.0 B / 0	3.5 MB / 374415	4.9 MB / 747052	6.9 MB / 1783859	172.8 MB / 283732661
Shuffle Remote Reads	0.0 B	3.4 MB	4.7 MB	6.6 MB	166.6 MB
Shuffle Write Size / Records	0.0 B / 0	109.0 B / 3	132.0 B / 4	158.0 B / 6	262.0 B / 13
Shuffle spill (memory)	0.0 B	0.0 B	0.0 B	0.0 B	13.3 GB
Shuffle spill (disk)	0.0 B	0.0 B	0.0 B	0.0 B	68.2 MB

Ways to Handle Skewed Join nowadays

- Increase shuffle partition number
- Increase BroadcastJoin threashold to change Shuffle Join to Broadcast Join
- Add prefix to the skewed keys
- ……

Question 3: These involve many manual efforts and are limitted. Can we handle skewed join at runtime automatically?

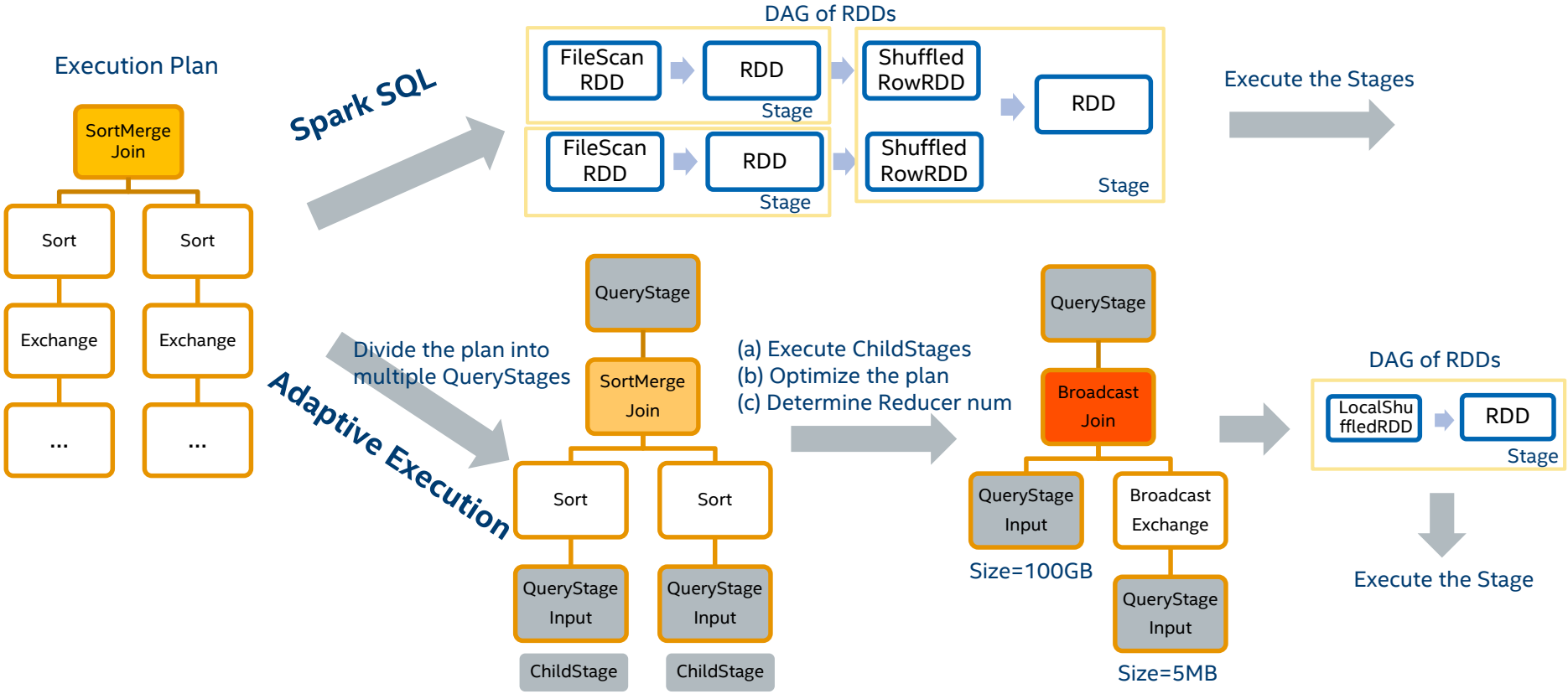
Adaptive Execution Background

- SPARK-9850: Adaptive execution in Spark*
- SPARK-9851: Support submitting map stages individually in DAGScheduler
- SPARK-9858: Introduce an ExchangeCoordinator to estimate the number of post-shuffle partitions.

A New Adaptive Execution Engine in Spark SQL*

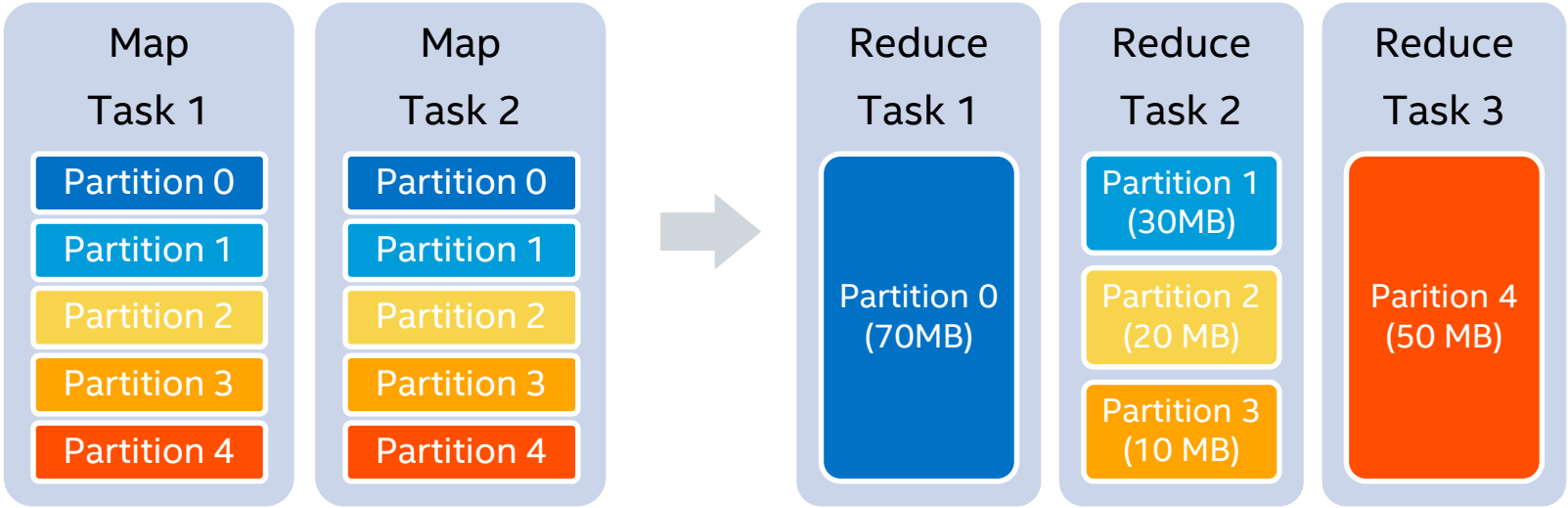
*Other names and brands may be claimed as the property of others.

Adaptive Execution Architecture



Auto Setting the Number of Reducers

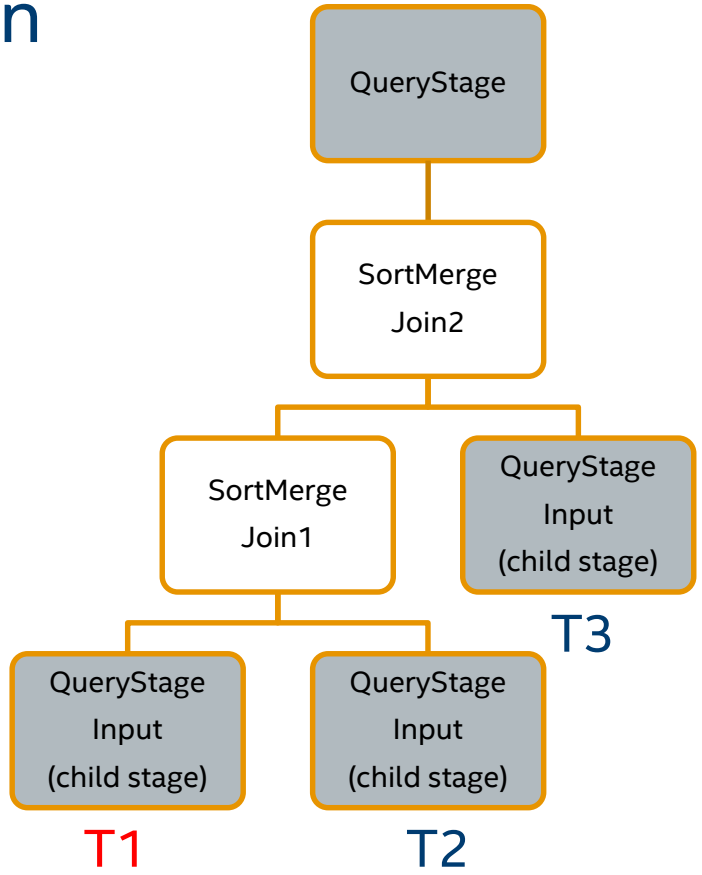
- 5 initial reducer partitions with size
  [70 MB, 30 MB, 20 MB, 10 MB, 50 MB]
- Set target size per reducer = 64 MB. At runtime, we use 3 actual reducers.
- Also support setting target row count per reducer.



Shuffle Join => Broadcast Join Example 1

- T1 < broadcast threshold
- T2 and T3 > broadcast threshold

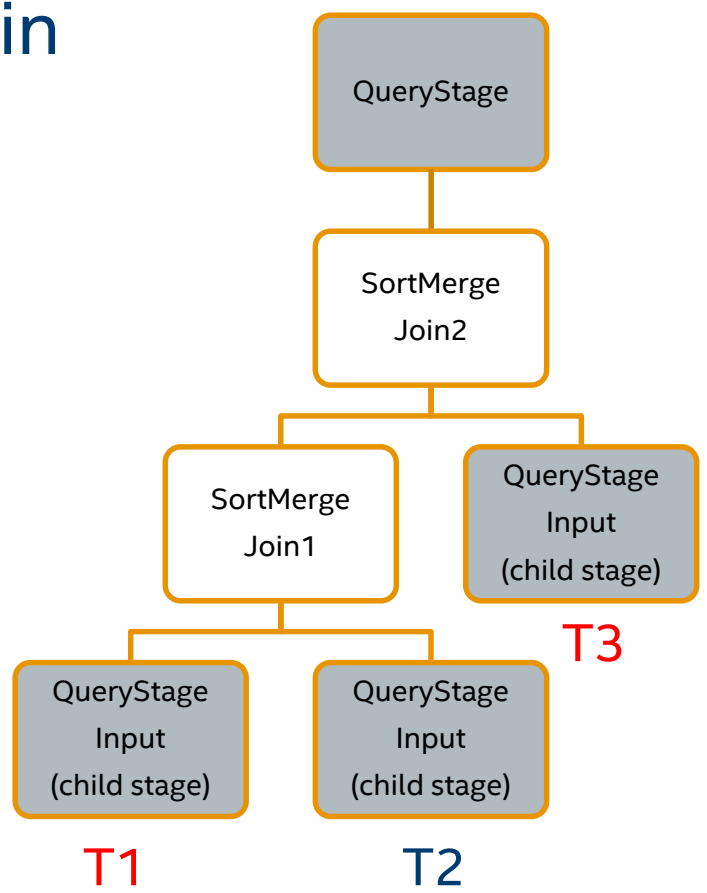
- In this case, both Join1 and Join2 are not changed to broadcast join



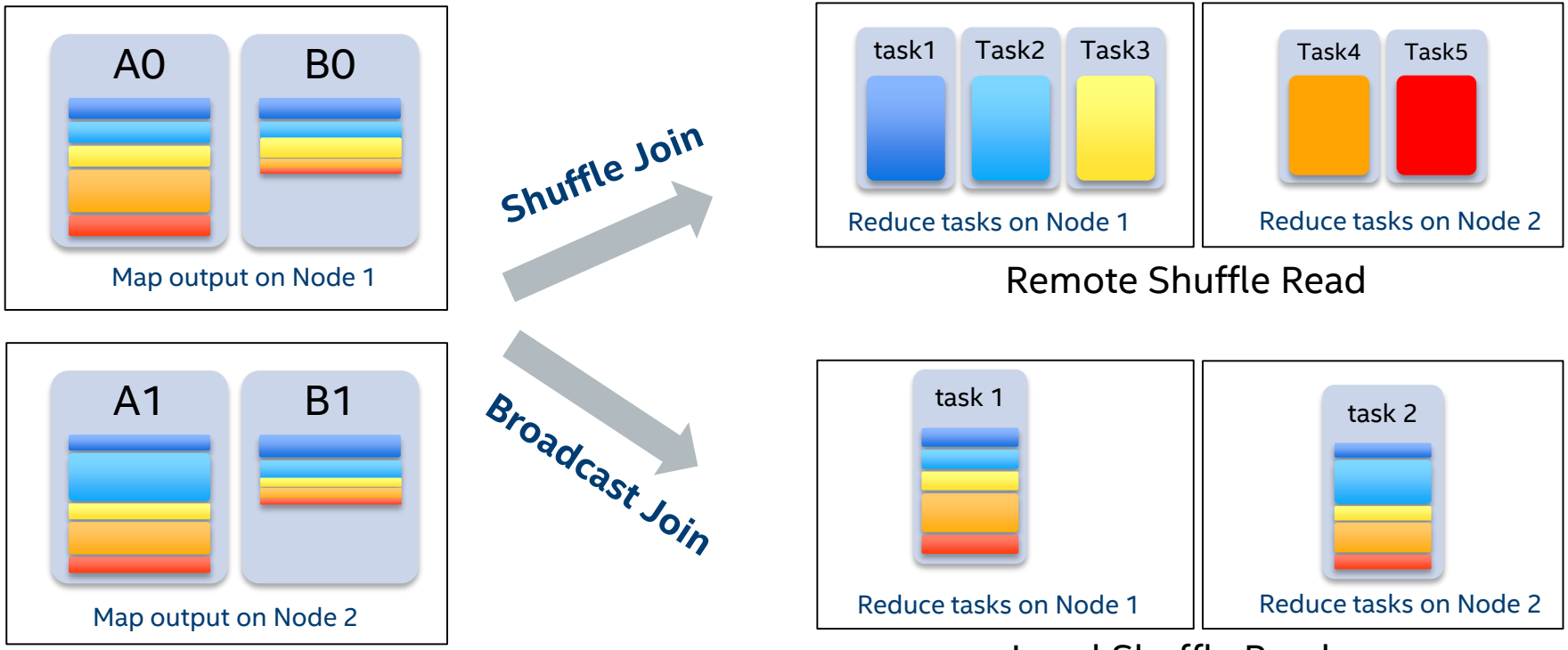
Shuffle Join => Broadcast Join Example 2

- T1 and T3 < broadcast threshold
- T2 > broadcast threshold

- In this case, both Join1 and Join2 are changed to broadcast join



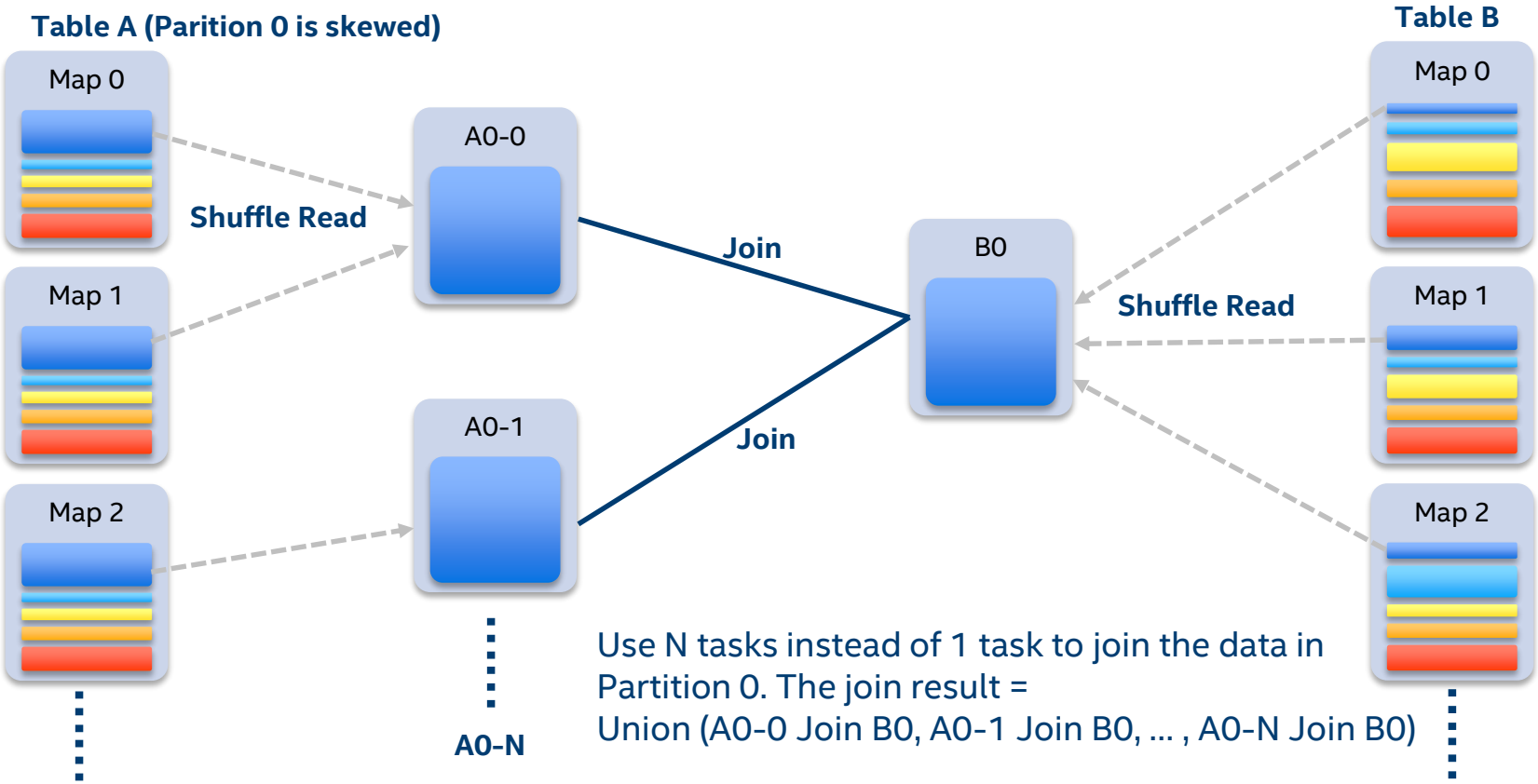
Remote Shuffle Read => Local Shuffle Read



Skewed Partition Detection at Runtime

- After executing child stages, we calculate the data size and row count of each partition from MapStaus.
- A partition is skewed if its data size or row count is N times larger than the median, and also larger than a pre-defined threshold.

Handling Skewed Join



Use N tasks instead of 1 task to join the data in Partition 0. The join result =
Union (A0-0 Join B0, A0-1 Join B0, ... , A0-N Join B0)

Benchmark Result

Cluster Setup

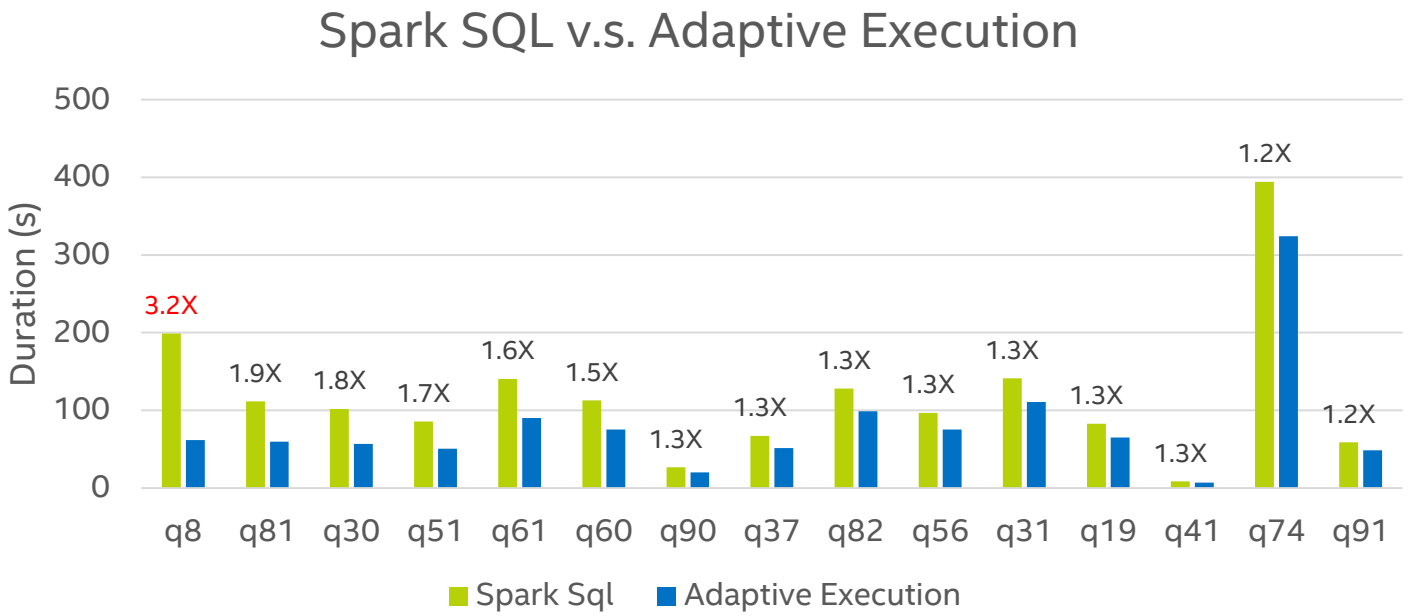
Hardware		BDW
Slave	Node#	**98**
	CPU	Intel（R） Xeon（R） CPU E5-2699 v4 @ 2.20GHz (88 cores)
	Memory	256 GB
	Disk	7× 400 GB SSD
	Network	10 Gigabit Ethernet
Master	CPU	Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz (88 cores)
	Memory	256 GB
	Disk	7× 400 GB SSD
	Network	10 Gigabit Ethernet

Software	
OS	CentOS* Linux release 6.9
Kernel	2.6.32-573.22.1.el6.x86_64
Spark*	Spark* master (2.3) / Spark* master (2.3) with adaptive execution patch
Hadoop*/HDFS*	hadoop-2.7.3
JDK	1.8.0_40 (Oracle* Corporation)

*Other names and brands may be claimed as the property of others.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

TPC-DS* 100TB Benchmark



*Other names and brands may be claimed as the property of others.

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Auto Setting the Shuffle Partition Number

- Less scheduler overhead and task startup time.
- Less disk IO requests.
- Less data are written to disk because more data are aggregatd.

Partition Number 10976 (q30)

WITH customer_total_return AS (SELECT wr_returning_customer_sk AS ctr_customer_sk, ca… run at AccessController.java:0 +details	2017/10/17 11:31:01	18 s	10976/10976			6.9 GB	
WITH customer_total_return AS (SELECT wr_returning_customer_sk AS ctr_customer_sk, ca… run at AccessController.java:0 +details	2017/10/17 11:30:52	10 s	10976/10976			17.0 GB	6.8 GB

Partition Number changed to 1084 and 1079 at runtime. (q30)

WITH customer_total_return AS (SELECT wr_returning_customer_sk AS ctr_customer_sk, ca… run at AccessController.java:0 +details	2017/10/18 04:17:22	4 s	1079/1079			5.3 GB	
WITH customer_total_return AS (SELECT wr_returning_customer_sk AS ctr_customer_sk, ca… run at Executors.java:511 +details	2017/10/18 04:17:14	7 s	1084/1084			17.7 GB	5.2 GB

*For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

SortMergeJoin -> BroadcastJoin at Runtime

- Eliminate the data skew and straggler in SortMergeJoin
- Remote shuffle read -> local shuffle read.
- Random IO read -> Sequence IO read

SortMergeJoin (q8):

SELECT s_store_name, sum(ss_net_profit) FROM store_sales, date_dim, store, (SELECT ca… run at AccessController.java:0 +details	2017/10/18 18:13:29	2.5 min	10976/10976			52.0 GB	13.2 KB
SELECT s_store_name, sum(ss_net_profit) FROM store_sales, date_dim, store, (SELECT ca… run at AccessController.java:0 +details	2017/10/18 18:12:48	37 s	12121/12121	1183.1 GB			52.0 GB
SELECT s_store_name, sum(ss_net_profit) FROM store_sales, date_dim, store, (SELECT ca… run at AccessController.java:0 +details	2017/10/18 18:13:25	2 s	10976/10976			2.5 KB	2.5 KB

BroadcastJoin (q8 Adaptive Execution):

SELECT s_store_name, sum(ss_net_profit) FROM store_sales, date_dim, store, (SELECT ca… run at Executors.java:511 +details	2017/10/18 02:48:56	10 s	12121/12121			17.4 GB	284.3 KB
SELECT s_store_name, sum(ss_net_profit) FROM store_sales, date_dim, store, (SELECT ca… run at ThreadPoolExecutor.java:1142 +details	2017/10/18 02:48:56	71 ms	69/69			2.5 KB	

*For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Scheduling Difference

- Spark SQL* has to wait for the completion of all broadcasts before scheduling the stages. Adaptive Execution can start the stages earlier as long as its dependencies are completed.

Original Spark:

SELECT i_brand_id brand_id, i_brand brand, i_manufact_id, i_manufact, sum(ss_ext_sales_p... run at AccessController.java:0 +details	2017/10/09 21:43:01	34 s	10128/10128	259.7 GB
SELECT i_brand_id brand_id, i_brand brand, i_manufact_id, i_manufact, sum(ss_ext_sales_p... run at AccessController.java:0 +details	2017/10/09 21:43:01	3 s	280/280	496.9 MB
SELECT i_brand_id brand_id, i_brand brand, i_manufact_id, i_manufact, sum(ss_ext_sales_p... run at ThreadPoolExecutor.java:1142 +details	2017/10/09 21:42:11	0.3 s	10/10	3.9 MB
SELECT i_brand_id brand_id, i_brand brand, i_manufact_id, i_manufact, sum(ss_ext_sales_p... run at ThreadPoolExecutor.java:1142 +details	2017/10/09 21:42:11	4 s	280/280	181.5 MB

50 Seconds Gap

BroadcastExchange

data size (bytes): 1,008,785,152
time to collect (ms): 19,267
time to build (ms): 24,074
time to broadcast (ms): 6,003

Adaptive Execution:

SELECT i_brand_id brand_id, i_brand brand, i_manufact_id, i_manufact, sum(ss_ext_sales_p... run at Executors.java:511 +details	2017/10/10 15:18:10	55 s	10128/10128	259.7 GB
SELECT i_brand_id brand_id, i_brand brand, i_manufact_id, i_manufact, sum(ss_ext_sales_p... run at ThreadPoolExecutor.java:1142 +details	2017/10/10 15:18:08	4 s	280/280	181.5 MB

THANK YOU

Legal Disclaimer

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others