Merging small file size splits

Background:

For hive datasource, driver is created per split, so more splits means more parallelism, this is working without any issue if we have good file size ,however , there is a case which the file size is far less than hive split size(hive.max-initial-split-size, hive.max-split-size), thus number of splits is determined by number of small files. To process a large number of small splits in parallel, it definitely introduces overhead of CPU context switch, causing performance issues when doing the data shuffling and resulting in low CPU/memory usage. Based on our testing we are seeing huge performance impact in high concurrency environment since task slot is occupied by those small split

Objective:

- 1. Small files are grouped to one split, accumulate all those file sizes it will not cross hive.maxinitial-split-size / hive.max-split-size.
- 2. Since these small files are grouped in to a split, since driver is created per split, number of usage of driver is less so overhead of CPU context switch will reduced.

High level design:

In scheduling form list of small files (HiveSplit) which will not cross total size of hive.max-split-size.

The list of small files should belong to same bucket number.

During writes, only one writer can write the same bucket file. So bucket number is obtained from split to indicate the partitioning of data to different workers or different writers.

nteraction Scheduler mer	ge splits for small files /TERE	D UNREGISTE	ERED UN	REGISTER	ED UN	REGIST	ERED	UNREG	SISTE	RE
SqlQueryScheduler	SourcePartitionedScheduler	BufferingSplitSource ConnectorAwareSplitSource		reSplitSource	HiveSplitSource					RE
THE OPTIM	JOINTEDIOTER	Lonneolori	ALD ON	L.						RE
NREGISTERS	chedule NREGISTERE	D UNREGISTE	RED UN		ED UN					RE
NREGISTERE	ED UNRE GISTERE	D UNREGISTE	RED UN		ED UN					RE
NREGISTERE	ED UNRE <u>GIS</u> TERE	D UNREGISTE	RED UN		ED UN					RE
NREG STERE		D UNREGISTE	RED UN		ED UN					RE
NREG STERE	D UNRESIS3:getNex	tBatchJNREGISTE	RED UN		ED UN					RE
NREG STERE	ED UNREGISTERE	D UNF GISTE	RED UN		ED UN					RE
NREG STERE	ED UNREGISTERE	4 : fetchNextBatch/	Async D U N		ED UN					RE
NREG STERE	ED UNREGISTERE	D UNF GISTE	RED UN		ED UN					RE
NREG STERE	ED UNREGISTERE	DUNR 5 : fetchSplits	ERED UN		ED UN					RE
NREG STERE	ED UNREGISTERE		NextBatch		ED UN	REGIST	FRED	UNREG	ISTE	RE
NREG STERE	ED UNREGISTERE		RED UN	REGISTER	ED UN	Same bucket farmed as list	number sma of HiveSplit	ll files are and place it into	STE	RE
NREGISTERE	ED UNREGISTERE	D UNREGISTE		7 : getNextBa	tch	those file size	oper, when a s it will not cr	iccumulate all oss	STE	RE
NREG STERE	ED UNREGISTERE	D UNREGISTE	RED UN		ED UN	nive.max-spii	-3126.		STE	RE
NREGISTERE	ED UNREGISTERE	D UNREGISTE			ED UN	REGIST	ERED	UNREG	ISTE	RE
NREGISTERE	ED UNREGISTERE	D UNREGISTE		REGISTER	ED UNI	REGIST	ERED	UNREG	ISTE	RE
										RE

These splits are passed to worker nodes, the driver will execute splits.

HiveSplitWrapper contains list of hiveSplit, which are grouped based on file size.

HivePageSourceProvider. createPageSource will process all the hiveSplit and result ConnectorPageSource and Page is added to list ConcatPageSource.



Code flow change in case of table scan

Driver		TableScanOperator	PageSourceManage	r ClassLoaderSafeCo	ClassLoaderSafeConnectorPageSourceProvider		ceProvider	OrcPageSourceFactory	RED
EG	ISTER	ED UNREGIS	TERED UNF	REGISTERED	UNREGISTERED	UNREG	ISTER		ERED
	1 : processir	nternal UNREGIS	TERED UNF		UNREGISTERED) UNREG		D UNREGIST	
	IL FERM	D UNREGIS	TERED UNF		UNREGISTERED) UNREG		D UNREGIST	
	ISTER		TERED UNF		UNREGISTERED) UNREG		D UNREGIST	
	ISTER:	Page 4 : createl	PageSource		UNREGISTERED) UNREG		D UNREGIST	
	ISTER	ED UNREGIS	TERE		UNREGISTERED) UNREG		D UNREGIST	
	ISTER	D UN 5 : Concat	PageSource 7 7	: createPageSource	H REGISTERED) UNREG	ISTERE	D UNREGIST	
	ISTER	ED UNREGIS	TERED	ConcatPageSource			HiveSplitWra	pper containes list of C	
	ISTER	ED UNREGIS	TERED UNF			UNRE	handled in sa ConcatPage	ame driver and create Source	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTEREL		eatePageSourc	DUNREGIST	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED	UNRE	ISTERE	D UNREGIST	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED			D UNREGIST	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED	UN12 crea	teHivePageSo	UNREGIST	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED) UNRE	ISTERE	D UNREGIST	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED) UNREG	13 : createPa	EGISTE	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED	UNREG	14 : Connector	PageSource	
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED) UNREG			RED
	ISTER	ED UNREGIS	TERED UNF		UNREGISTERED) UNREG		D UNFEGIST	ERED
	ISTER	D UNREGIS			UNREGISTERED	UNREG			

Code flow change in case of Delete/Update

