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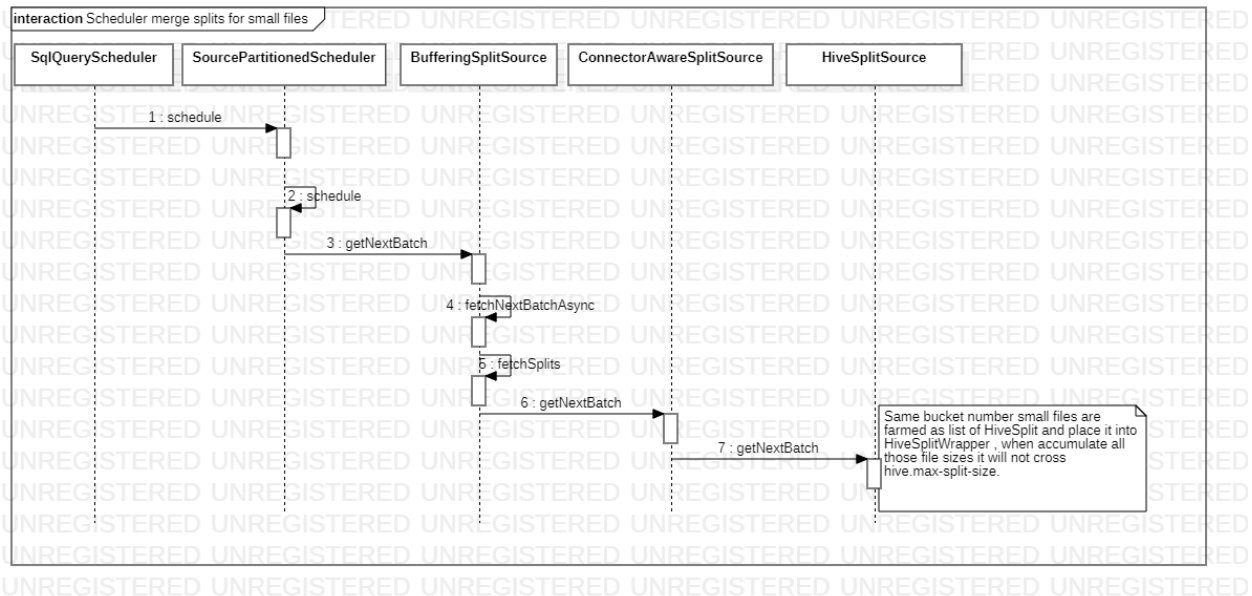
These splits are passed to worker nodes, the driver will execute splits.

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.

```
public class ConcatPageSource
        implements ConnectorPageSource
{
    private final List<ConnectorPageSource> pageSources;
    private final Iterator<Page> concatPageIterator;
    private boolean closed;
```

Code flow change in case of table scan
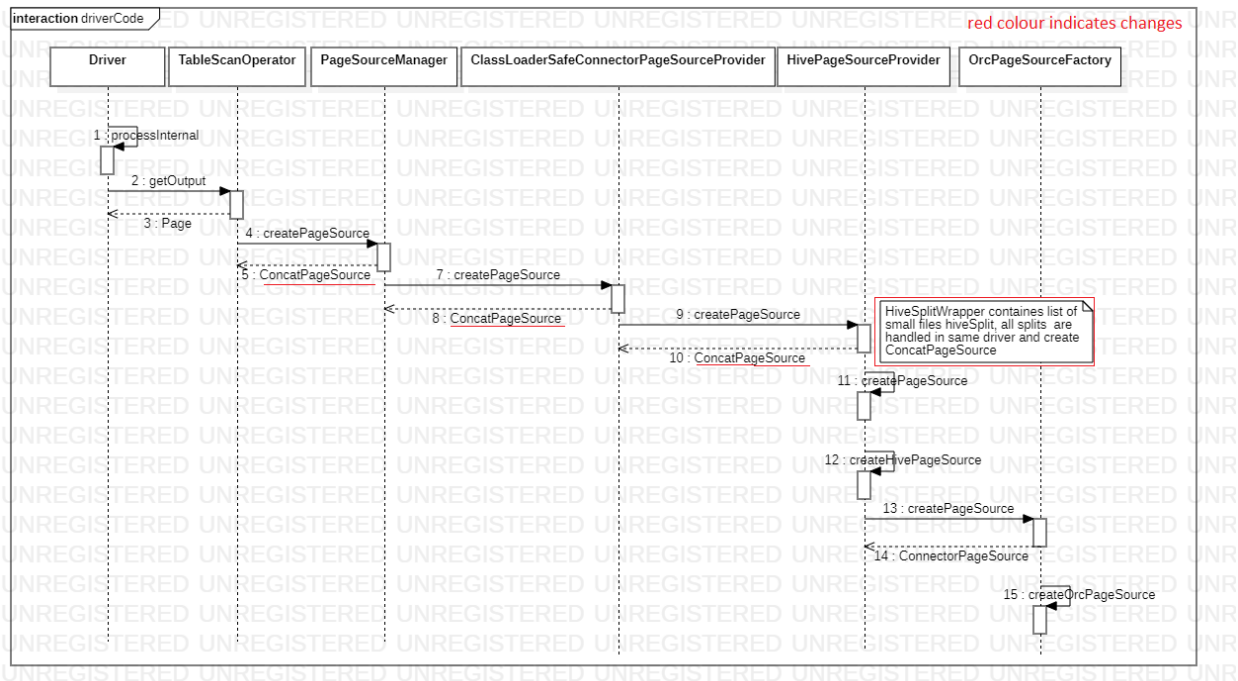
Code flow change in case of Delete/Update