Improving Comet Shuffle Performance

GitHub issue: https://github.com/apache/datafusion-comet/issues/1123

This document is for collaborating on ideas for speeding up Comet's shuffle performance. Once we identify ideas we want to implement, we should create GitHub issues.

Comparison with Apache Gluten Shuffle

Gluten+Velox has a single shuffle writer - ColumnarShuffleWriter.

Velox has:

- VeloxShuffleWriter (base class)
- VeloxHashShuffleWriter
- VeloxRssSortShuffleWriter (remote shuffle service?)
- VeloxSortShuffleWriter

Comet has:

- CometBypassMergeSortShuffleWriter
- CometUnsafeShuffleWriter.

Shuffle Writer

The input to ShuffleWriterExec is always a ScanExec that reads JVM batches that were produced by a native plan.

We use Arrow FFI to import the native batch into JVM and then export it back out again with no modification. This is expensive and seems like it could be avoided somehow.

Ideas:

- Can we just import the Arrow RecordBatch address as a pointer and then export it back out, without serializing the schema for each batch?

- Can we altogether avoid the transfer by putting the shuffle input native plan directly underneath the shuffle in CometShuffleExchangeExec?

Shuffle Reader