Apache DataFusion Comet 0.8.0 Release

The Apache DataFusion PMC is pleased to announce version 0.8.0 of the Comet subproject.

Comet is an accelerator for Apache Spark that translates Spark physical plans to DataFusion physical plans for improved performance and efficiency without requiring any code changes.

Comet aims to provide 100% compatibility with Apache Spark. Any operators or expressions that are not fully compatible will fall back to Spark unless explicitly enabled by the user. Refer to the compatibility guide for more information.

This release covers approximately TBD weeks of development work and is the result of merging TBD PRs from TBD contributors. The complete change log is available here (TBD).

Release Highlights

Native Scan Improvements

There are notable improvements to the experimental native_datafusion scan:

- INT96 support
- Support for complex and nested types, including Structs, Arrays, and Maps

There are some known limitations

- There are schema coercion bugs for nested types containing INT96 columns, which can cause incorrect results.
- There are compatibility issues when reading integer values that are larger than their type annotation, such as the value 1024 being stored in a field annotated as int(8).
- A small number of Spark SQL tests remain unsupported (#1545)

Native Shuffle Improvements

Significant enhancements to the native shuffle mechanism include:

- **Lower memory usage** through `interleave_record_batches`
- **Support for complex types** in shuffle data (note: hash partition expressions still require primitive types)
- **Reclaimable shuffle files**, reducing disk pressure
- **Respects `spark.local.dir`** for temporary storage
- **Per-task shuffle metrics**, providing better visibility into execution behavior

Performance Improvements

- Up to **4x speedup** in jobs using `dropDuplicates`, thanks to optimizations in the `first_value` and `last_value` aggregate functions
- Improved performance when **`native_datafusion`** is enabled
- Introduction of a **global Tokio runtime**, which resolves potential deadlocks in certain multi-task scenarios

🔧 Spark Compatibility

- **Added support** for **Spark 3.5.5**
- **Dropped support** for **Spark 3.3.x**
-

Acknowledgements

We would like to thank everyone who has helped with these releases through their helpful conversations, code review, issue descriptions, and code authoring. We would especially like to thank the following authors of PRs who made this release possible, listed in alphabetical order by username: TBD

Getting Involved

The Comet project welcomes new contributors. We use the same Slack and Discord channels as the main DataFusion project and have a weekly DataFusion video call.

The easiest way to get involved is to test Comet with your current Spark jobs and file issues for any bugs or performance regressions that you find.

There are also many good first issues waiting for contributions. See the contributors guide to get started with contributing to the project.