Avro Files

Apache Avro is an efficient data serialization framework. Avro defines a file format that uses optimized binary encoding to store data efficiently. (For an explanation of binary encoding, see “What Is Binary Encoding” from wisegeek.com.) The Avro format is widely supported by big data tools, and it’s also designed to work across different programming languages and tools outside the typical big data system. Avro files are suitable for long-term data storage.

An Avro data file includes an embedded schema definition, which makes the file *self-describing* —the file itself provides information about what's in the file. Avro is also built to handle *schema evolution*. This means that it’s possible to add, remove, or modify columns in a Hive or Impala table without needing to make changes to the existing data stored within Avro files. The Avro framework will accommodate these schema changes, so the table and the existing data files will continue to be compatible, even though their schemas do not perfectly match. (For more about schema evolution, see “How Schema Evolution Works” within the linked page. Note that most of the details in that article pertain to Avro schemas in general, but some details are specific to the use of Avro within the Oracle NoSQL Database and so are not applicable to the big data systems presented in this course.)

Overall, Avro offers excellent interoperability and excellent performance, making it a popular choice for general-purpose big data storage.