Custom Training

Day 18

**PySpark**

Use of persist method to serialize the data.

To manually unpersist, unpersist() method is used.

In the latest Spark versions, the unpersist happens automatically.

Least recently used algorithm.

CPU time is higher as to read files the data has to be deserialized.

Write Operations in Spark:

If write operation happens to a table where the db is mentioned, then it gets saved in the default database.

test= spark.sql(“describe extended emp\_tbl”)

test.show()

the default datatype of this table is managed table.

Managed by databricks itself.

Managed table stores metadata and the actual data. Both are present in the Spark environment itself.

In external table, custom paths are given to store the actual data.



