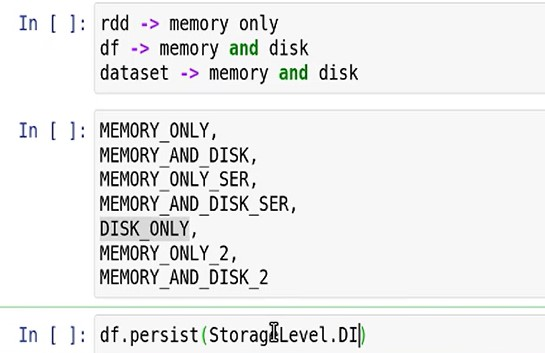
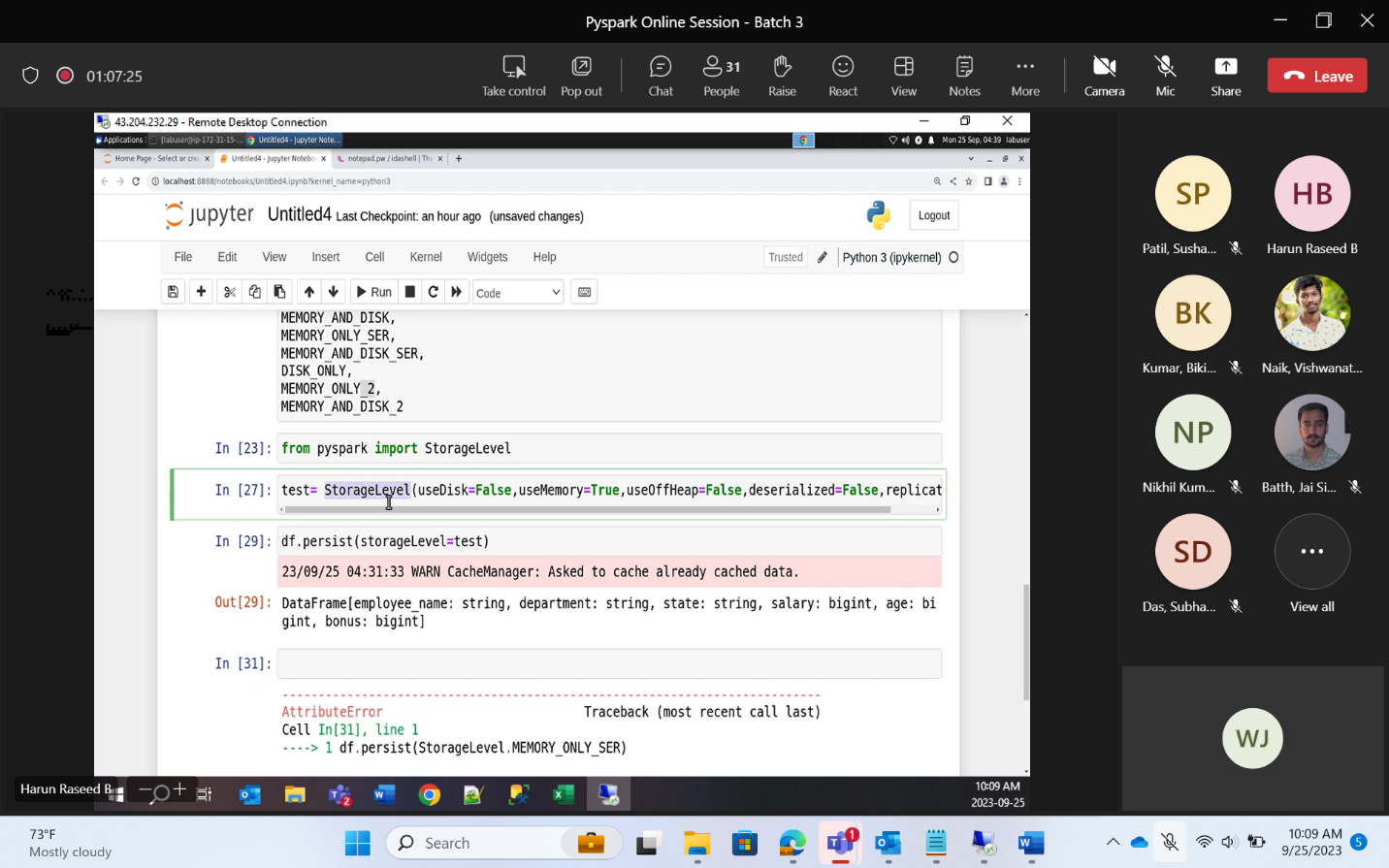
25-9-2023\_Washitwa



* 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 Operation in Spark

* If write operation happens to a table where the db is it mentioned, then it gets saved in the default database
* test = spark.sql(“describe extended emp\_tbl”)
* test.show()
* the default type 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

Managed table and external table

* On deleting the managed table, the meta data and the actual data are deleted
* On deleting the external table, the meta data is deleted but the actual data still exists as a custom path is provided to store the data