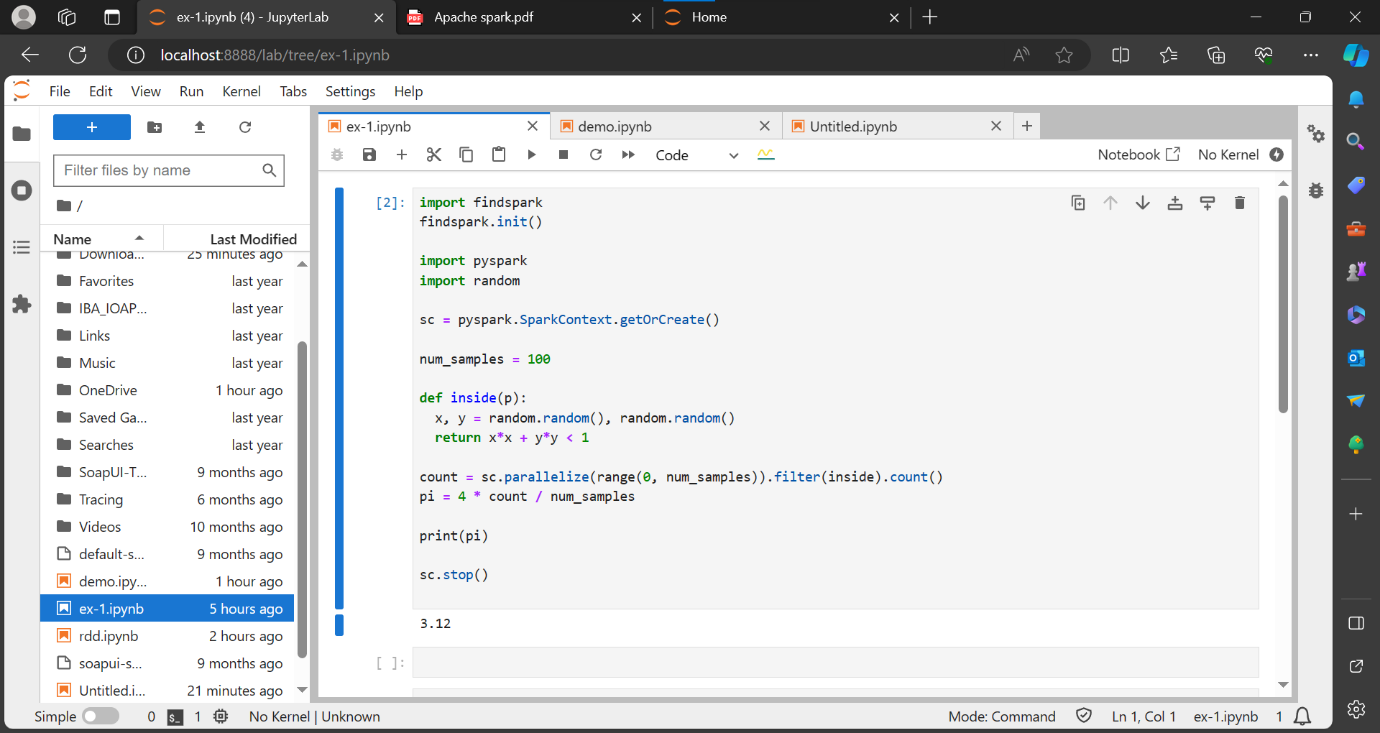
**ASSESSMENT-19 Vinutha S**

**26/12/2023**

**Installed jupyter using the command “py -m pip install jupyter”.**

**Launched by using command “jupyter lab”.**

****

Why we Choose PySpark for ETL?

* Performance: PySpark leverages in-memory computing, making ETL processes faster than ever.
* Ease of Use: Python developers can seamlessly transition to PySpark due to its Pythonic syntax.
* Scalability: Handle massive datasets with ease, thanks to Spark’s distributed processing.
* Rich Ecosystem: PySpark integrates with popular tools and libraries, making it versatile for various data tasks.

The PySpark ETL Workflow:

* Extract: Retrieve data from various sources like databases, files, or APIs.
* Transform: Clean, aggregate, and manipulate data to fit your analysis needs.
* Load: Store the transformed data into a database or data warehouse for analysis.