**26/12/2023**

**ETL Process**

ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse. The process of ETL can be broken down into the following three stages

Extract

Transform

Load

The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area

Transform: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.

Load: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.

Real-World Applications:

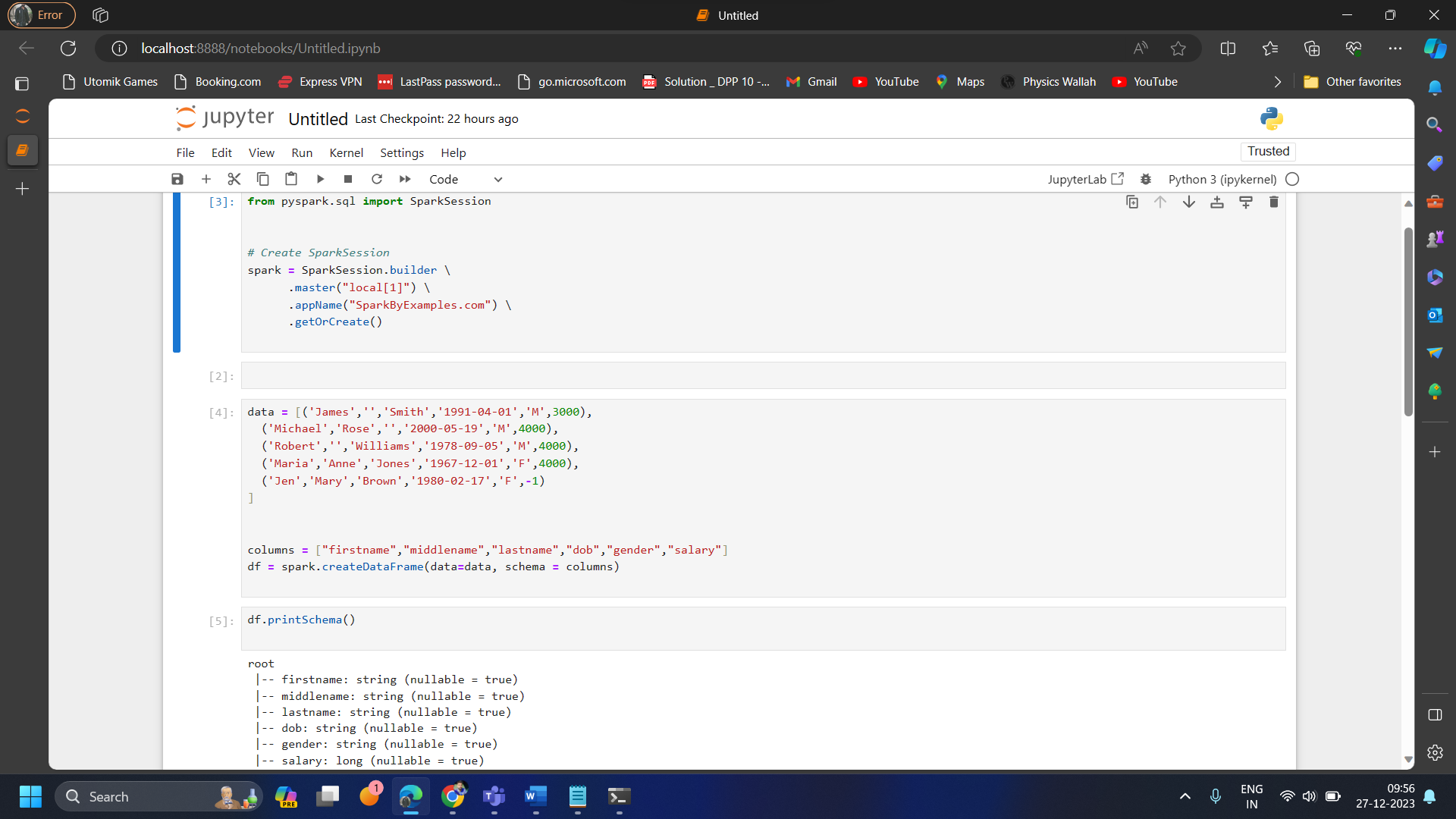
Perform large-scale data cleansing and preparation.

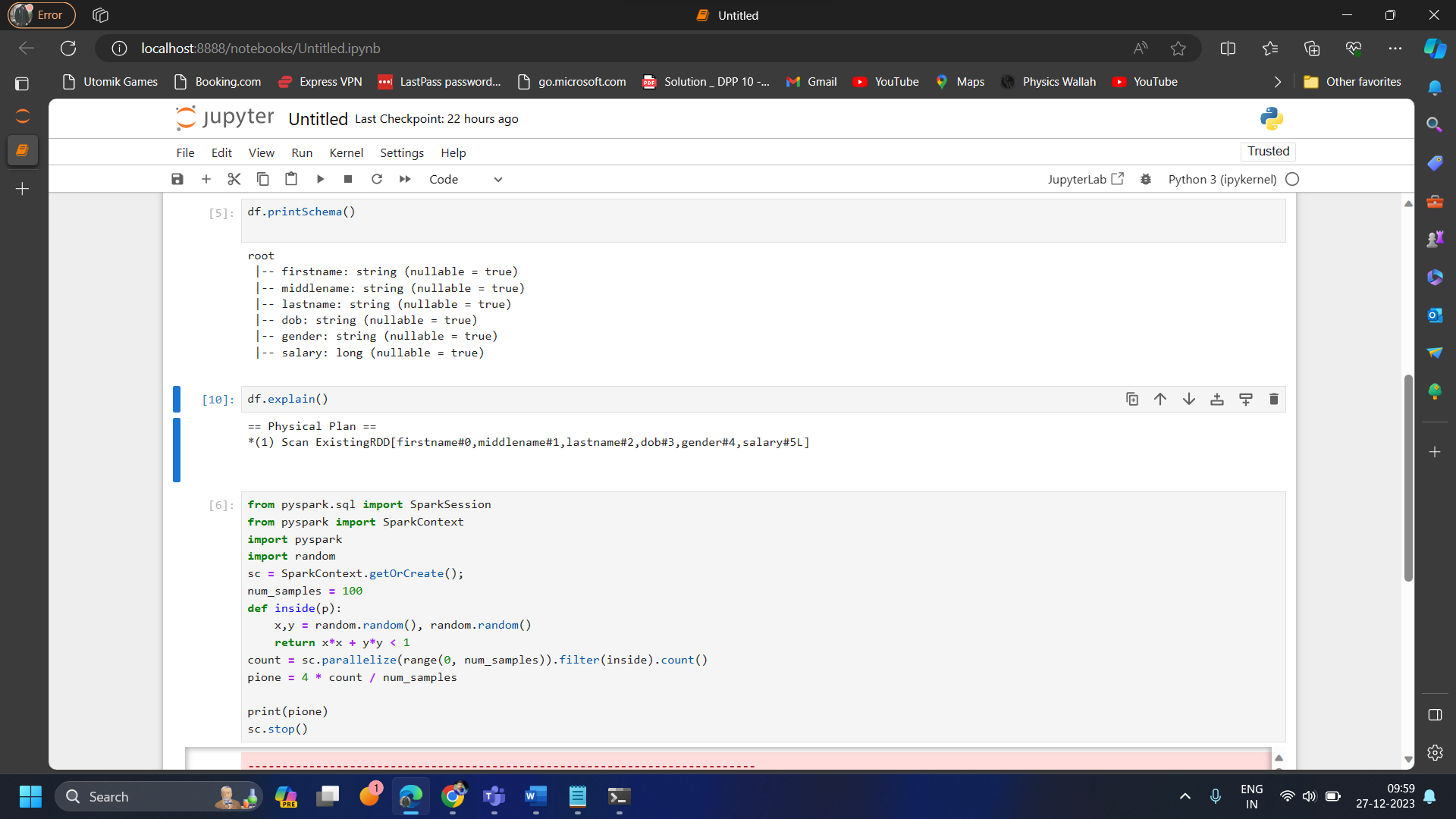
Analyse streaming data in real-time.

Create data pipelines for machine learning and AI applications.

Handle structured and unstructured data effortlessly.

**Pyspark command on jupyter notebook**

****

****