04. **Utilizing Apache Spark for data representation and**

**analysis of Students Data**

**Apache Spark** is an [open-source](https://en.wikipedia.org/wiki/Open-source_software) unified analytics engine for large-scale data processing. Spark provides an [interface](https://en.wikipedia.org/wiki/Application_programming_interface) for programming clusters with implicit [data parallelism](https://en.wikipedia.org/wiki/Data_parallelism) and [fault tolerance](https://en.wikipedia.org/wiki/Fault_tolerance). Originally developed at the [University of California, Berkeley](https://en.wikipedia.org/wiki/UC_Berkeley)'s [AMPLab](https://en.wikipedia.org/wiki/AMPLab), the Spark [codebase](https://en.wikipedia.org/wiki/Codebase) was later donated to the [Apache Software Foundation](https://en.wikipedia.org/wiki/Apache_Software_Foundation), which has maintained it since.

Link-<https://github.com/kattarcoder/ctws/blob/main/unit-4/spark.ipynb>