**ASSIGNMENT DATE-22-12-2023**

**What are the Features of PySpark?**

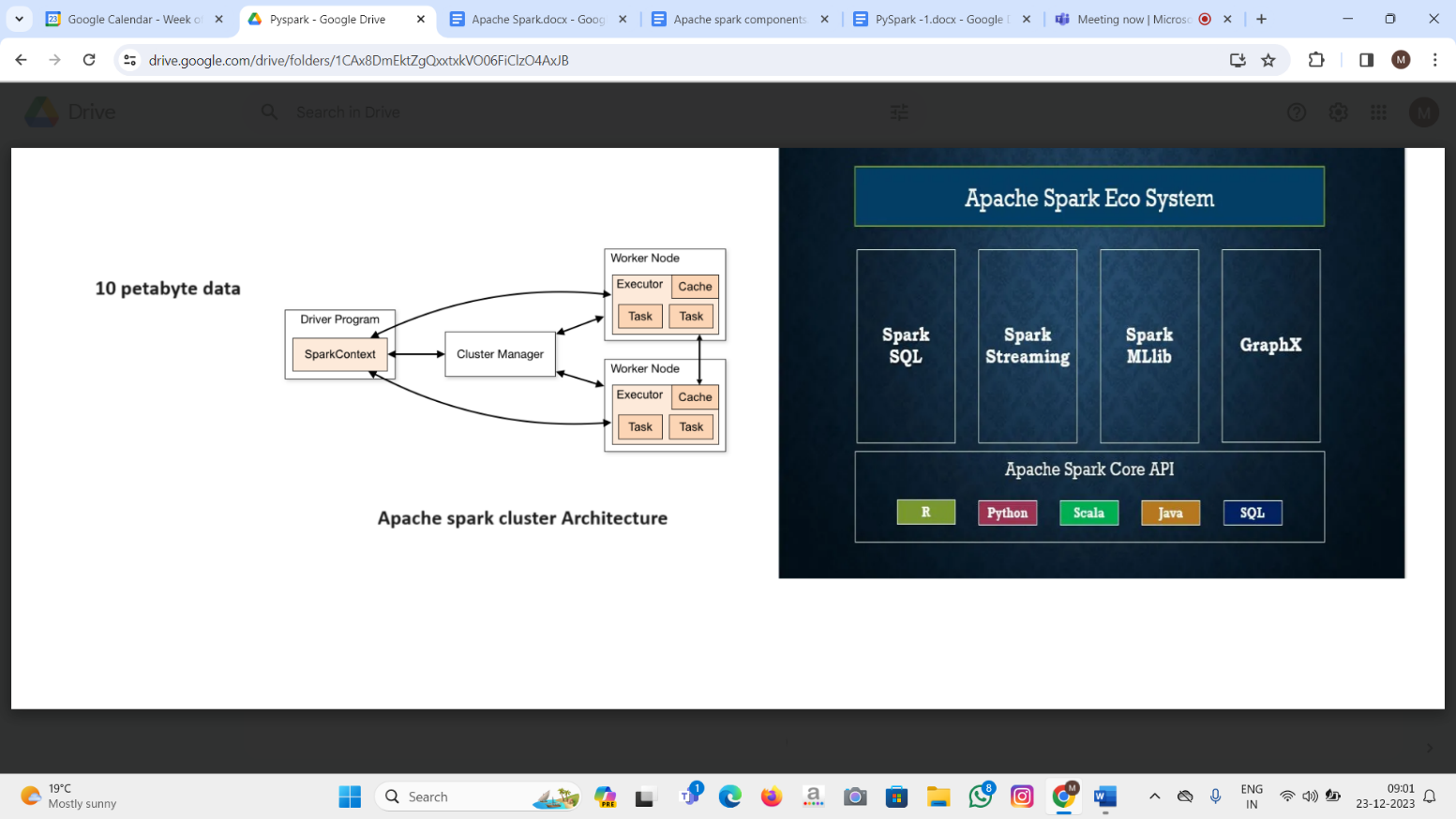
PySpark Features

* In-memory computation
* Distributed processing using parallelize
* Can be used with many cluster managers (Spark, Yarn, Mesos e.t.c)
* Fault-tolerant
* Immutable
* Lazy evaluation
* Cache & persistence
* Inbuild-optimization when using DataFrames
* Supports ANSI SQL

**Advantages of PySpark**

* PySpark is a general-purpose, in-memory, distributed processing engine that allows you to process data efficiently in a distributed fashion.
* Applications running on PySpark are 100x faster than traditional systems.
* You will get great benefits from using PySpark for data ingestion pipelines.

**PySpark Architecture**

****

**Cluster Manager Types**

As of writing this Spark with Python (PySpark) tutorial for beginners, Spark supports below cluster managers:

* Standalone – a simple cluster manager included with Spark that makes it easy to set up a cluster.
* Apache Mesos – Mesons is a Cluster manager that can also run Hadoop MapReduce and PySpark applications.
* Hadoop YARN – the resource manager in Hadoop 2. This is mostly used as a cluster manager.
* Kubernetes – an open-source system for automating deployment, scaling, and management of containerized applications.

local – which is not really a cluster manager but still I wanted to mention that we use “local” for master() in order to run Spark on your laptop/computer.