**WEEK 3 DAY 1**

**(18-11-2024)**

**Topics covered:**

1. Basics of data engineering and why we need spark
2. Spark installation
3. Databricks platform
4. Creating first cluster
5. Creating first notebook

**Data engineering:** like our engineering course, we have multiple domains together it forms DE.

**Data file extentions:**

CSV

JSON

.txt, .pdf, .doc

**Cassandra, paraquet, arvo**

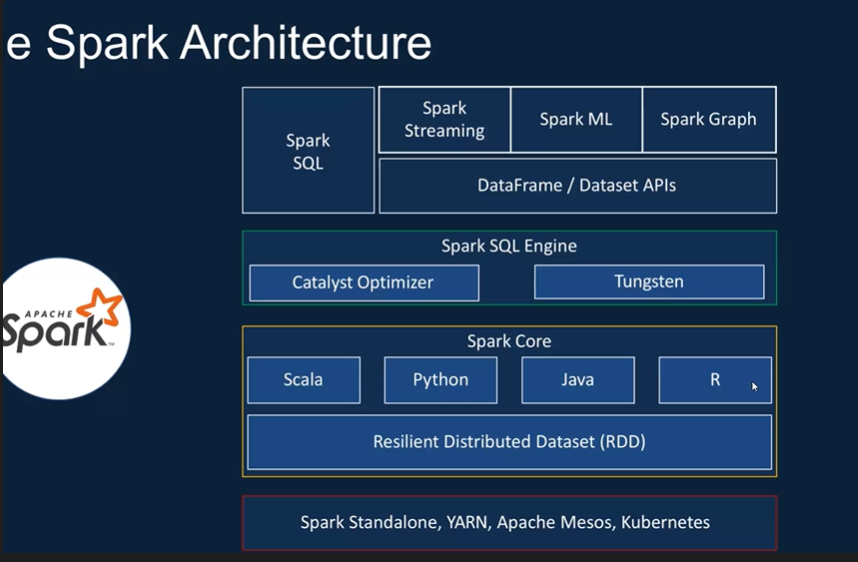
**Companies providing IT resources-we use:**

1. Microsoft:

* MS SQL, Power BI, VS Code
* MS 365, MS teams, Copilot
* Azure

1. Apache:

* [www.apache.org](http://www.apache.org).
* Hadoop, Hive, Kafka,
* Spark(SQL Spark, pyspark, Spark R, spark java)



1. Data Bricks:
   * Not azure databricks which is a data warehouse.
   * A company.

**Data Engineering Infrastructure:**

1. Data storage:
   * Therefore, we are having HDD and SDD.

* Using hardware. All data warehouse.
* Types of storage hardware:
  + **HDD (Hard Disk Drives)**: Spinning disks, cost-effective for archival and bulk storage.
  + **SSD (Solid State Drives)**: Faster access times, better for high-performance needs.
* Cloud-based storage systems:
  + Examples: **Azure Data Lake Storage**, **Amazon S3**, **Google Cloud Storage**.
  + Use in **Data Warehousing**: Cloud solutions like **Azure Synapse**, **Snowflake**, and **Google BigQuery** integrate with these storage systems.

**Learning area**: Storage tiers (hot, cold, archival), storage scalability, and cost optimization strategies.

1. Data Processing:

* CPU and MEMORY is utilized
* Storage Capacity:
  + From GB (Gigabyte) to YB (Yottabyte): 1 YB = 1024 ZB, 1 ZB = 1024 PB, 1 PB = 1024 TB, 1 TB = 1024 GB.
* Eg: Frameworks for Big Data Processing:
* Examples: **Apache Spark** for in-memory distributed computing.
* Processing large volumes of data efficiently using cluster computing.

**Learning area**: Spark architecture, resource management, and distributed data processing.

**DBMS:**

* **Relational**: MySQL, PostgreSQL, SQL Server, Oracle Database, SQLite.
* **Non-Relational (NoSQL)**: MongoDB, Redis, Cassandra, Neo4j, DynamoDB.

**Relational Databases (SQL)**

Relational databases store structured data in tables and use SQL for querying.

1. **MySQL** (Oracle)
2. **PostgreSQL** (Open Source)
3. **Microsoft SQL Server** (Microsoft)
4. **Oracle Database** (Oracle)
5. **SQLite** (Open Source)
6. **MariaDB** (MariaDB Foundation)
7. **IBM Db2** (IBM)

**Relational Database Frameworks**

1. **SQLAlchemy** (Open Source)
2. **Hibernate** (Red Hat)
3. **Entity Framework** (Microsoft)
4. **Django ORM** (Django Software Foundation)
5. **Spring Data JPA** (VMware/Spring)
6. **Knex.js** (Open Source)
7. **ActiveRecord** (Rails Core Team)

**Non-Relational Databases (NoSQL)**

NoSQL databases store unstructured or semi-structured data and are optimized for scalability and flexibility.

**Document-Based Databases:**

* Store data as documents (e.g., JSON).

1. **MongoDB** (MongoDB Inc.)
2. **Couchbase** (Couchbase Inc.)
3. **Amazon DocumentDB** (Amazon Web Services)

**Key-Value Stores:**

* Simple key-value pairs for high performance.

1. **Redis** (Redis Inc.)
2. **Amazon DynamoDB** (Amazon Web Services)
3. **Riak KV** (Basho Technologies)

**Columnar Databases:**

* Store data in columns instead of rows, suitable for analytical workloads.

1. **Apache Cassandra** (Apache Foundation)
2. **HBase** (Apache Foundation)
3. **Google Bigtable** (Google Cloud)

**Graph Databases:**

* Specialized for relationships between data points.

1. **Neo4j** (Neo4j Inc.)
2. **Amazon Neptune** (Amazon Web Services)
3. **TigerGraph** (TigerGraph Inc.)

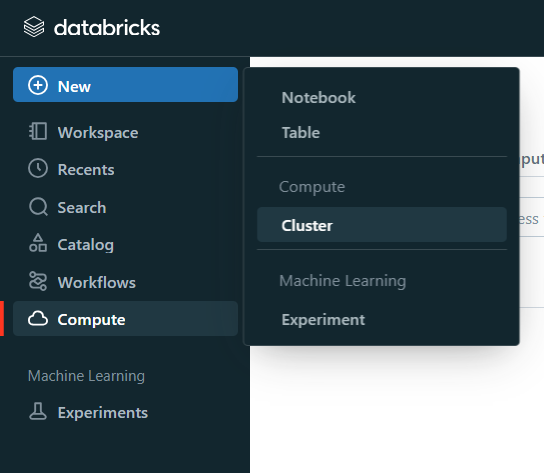
**Non-Relational Database Frameworks**

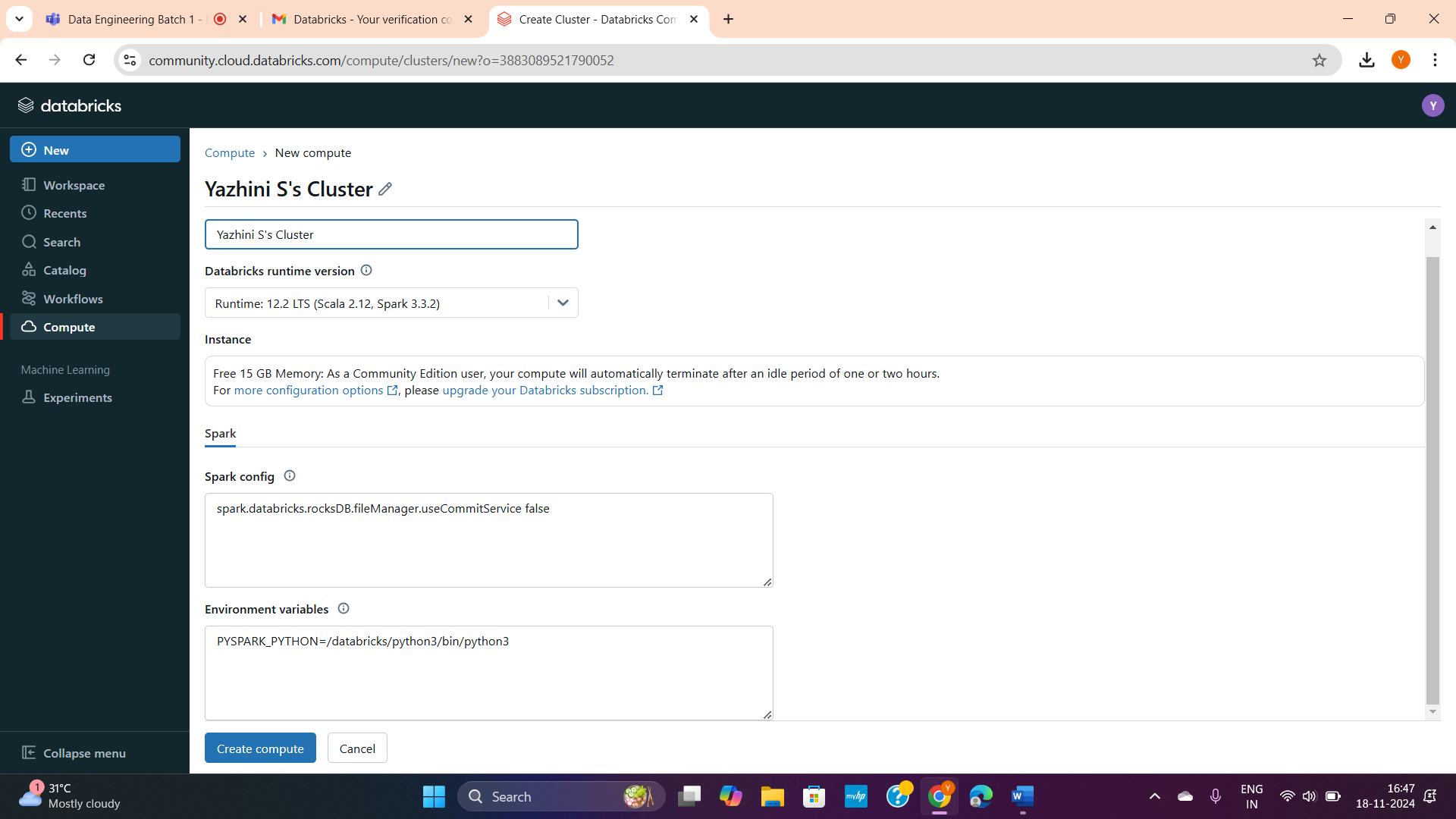
1. **Mongoose** (MongoDB Inc.)
2. **PyMongo** (MongoDB Inc.)
3. **Spring Data MongoDB** (VMware/Spring)
4. **Cassandra Driver** (DataStax/Apache)
5. **Firebase SDK** (Google)
6. **Gremlin/TinkerPop** (Apache)
7. **DynamoDB SDK** (Amazon Web Services)
8. **Hadoop** (Apache)

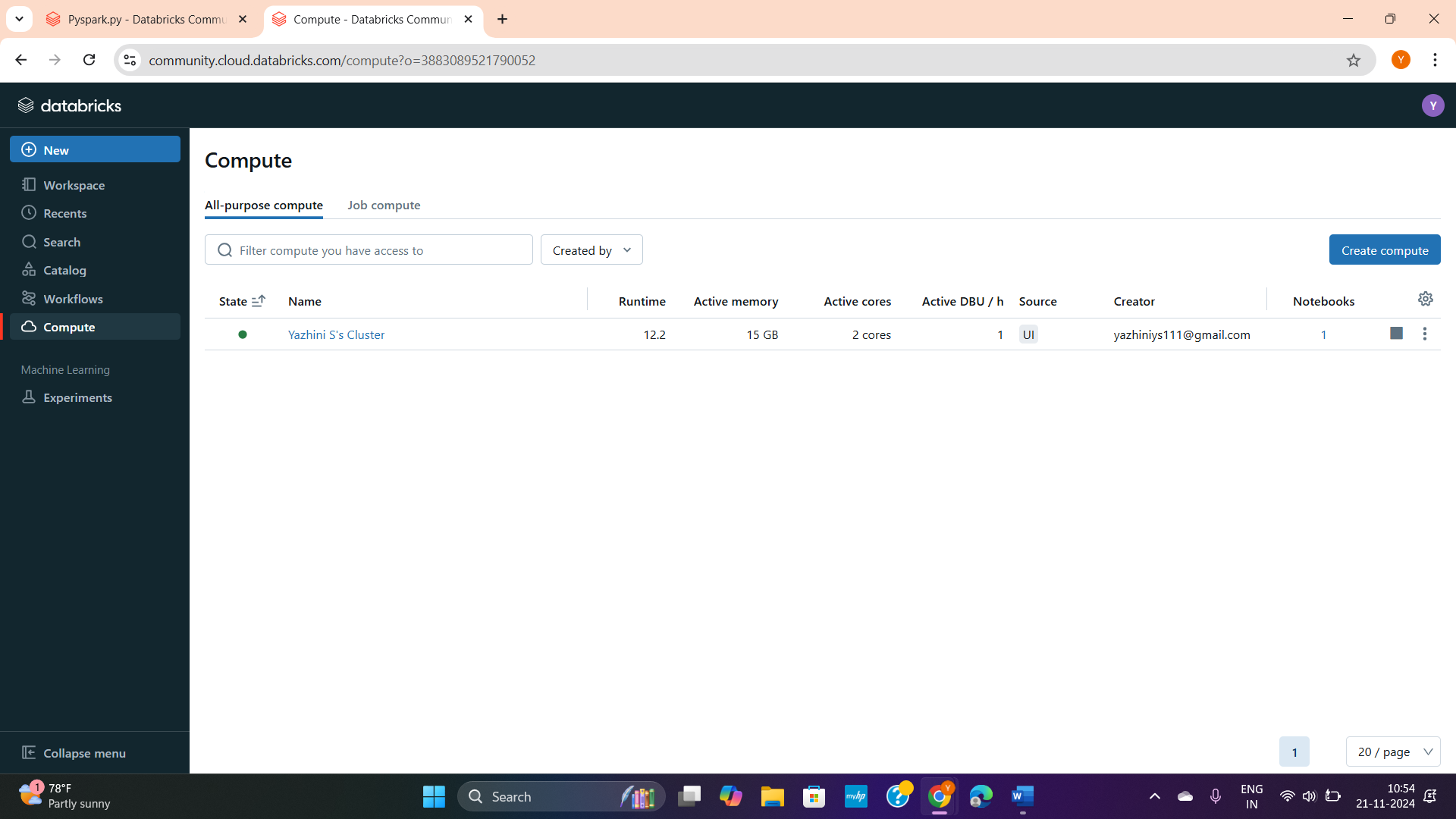
**SPARK:**

Apache Spark is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or cluster**. (refer spark ppt)**

**Data Bricks: (14 days free trial)**

****





**Notebooks created:**

