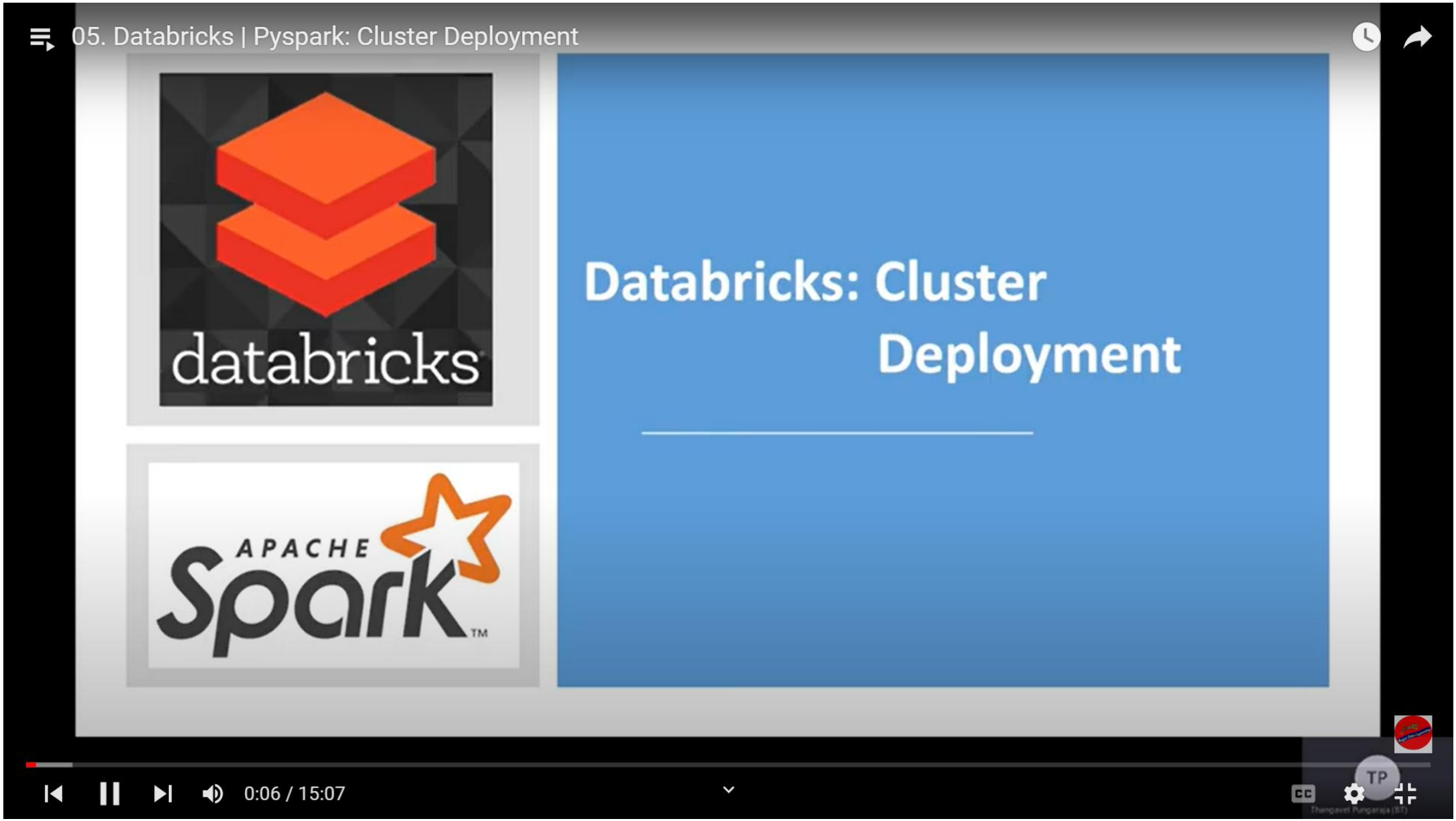
05. Databricks | Pyspark: Cluster Deployment
databricks
APACHE
Spark
Databricks: Cluster
Deployment
0:06 / 15:07

What is Cluster?

CLUSTER IS COMPUTING INFRASTRUCTURE IN DATABRICKS ENVIRONMENT.

IT IS SET OF COMPUTATION RESOURCES AND CONFIGURATIONS.

CLUSTER IS THE INFRASTRUCTURE WHICH EXECUTES DATA ENGINEERING, DATA ANALYTICS AND DATA SCIENCE WORKLOADS THAT ARE DEVELOPED IN DATABRICKS NOTEBOOK.

TP
Thangavel Punganja (BT)

Cluster Types

All Purpose Cluster:

All-purpose clusters are used to execute and analyse data collaboratively using interactive notebooks. We can manually terminate and restart an all-purpose cluster. Multiple users can share such clusters to do collaborative interactive analysis.

Job Cluster:

Job clusters are used to run fast and robust automated jobs. These are created automatically at the start of execution and terminates the cluster at the end of execution.

Pools:

Databricks pools are created to reduce the boot time and auto scaling. Clusters are attached to pools and pool is configured with set of idle and ready to use instances. So when we attach a cluster from pool to notebook, there would always be instances ready to use. When job got executed cluster will return back to pool

Cluster Modes

- **Standard:**

  This mode is suitable for single user. If no team collaboration needed, we can go for this mode

- **High Concurrency:**

  This mode is more suitable for collaboration. It provides fine-grained sharing for maximum resource utilization and minimum query latencies

- **Single Node:**

  This mode runs the job only on driver node and no worker nodes are provisioned

Cluster Runtime

Runtime is set of core components that are needed for cluster to run. Based on runtime choice, it differs in usability, performance and security.

Several options are provided by cluster. We can choose according to our need.