Assessment 16

PySpark is an Apache Spark library written in Python to run Python applications.

PySpark is a general-purpose, in-memory, distributed processing engine that

allows you to process data efficiently.

PySpark Features:

Fault-tolerant

Immutable

PySpark Architecture:

Apache Spark works in a master-slave architecture where the master is called the Driver and

slaves are called “Workers”. When you run a Spark application, Spark Driver creates a

context that is an entry point to your application, and all operations are executed on worker

nodes, and the resources are managed by Cluster Manager.

Spark History Server:

open the command prompt and type pyspark command

first you need to set the below config on spark-defaults.conf

start the history server by starting the command.

$SPARK\_HOME/bin/spark-class.cmd