St. Francis Institute of Technology, Mumbai-400 103

**Department of Information Technology**

A.Y. 2024-2025

Class: TE-ITA/B, Semester: VI

Subject: **Data Science Lab**

**Experiment – 9: To study big data analytics tools.**

**● Aim: To study Big Data Analytics tools.**

**● Objectives:** After study of this experiment, the student will be able to

● To learn streaming and batch processing tools of big data.

**● Outcomes:** After study of this experiment, the student will be able to

identify different big Data tools.

**● Prerequisite:** Fundamentals of Python Programming and Database Management System. **● Requirements:** Python Installation, Personal Computer, Windows operating system, Internet Connection, Microsoft Word.

**● Pre-Experiment Exercise:**

**Brief Theory:**

● Big data analytics

● Apache Hadoop Ecosystem

**8. Post-Experiments Exercise :**

**A. Extended Theory: (Soft Copy)**

1. **Different types of big data analytics tools**

**ANS:** Batch Processing Tools – Hadoop, Apache Spark, Flink

1. Real-Time Processing Tools – Apache Storm, Apache Kafka, Apache Flink
2. Interactive Query Tools – Apache Hive, Presto, Impala
3. Data Storage & Management – HDFS, Apache Cassandra, MongoDB
4. Machine Learning & AI Tools – Apache Mahout, TensorFlow, MLlib
5. Data Visualization Tools – Tableau, Power BI, D3.js

**2. Apache spark and spark Framework**

**ANS**:

1. Apache Spark is an open-source, distributed computing system for big data processing.
2. Supports batch processing, real-time streaming, machine learning, and graph processing.
3. Uses Resilient Distributed Datasets (RDDs) for fault tolerance and parallel computing.
4. Components include Spark Core, Spark SQL, Spark Streaming, MLlib, and GraphX.
5. Written in Scala, with APIs for Python, Java, and R.

**B. Questions:**

1. **Types of spark modes**

**ANS:   
 1.** Local Mode – Runs Spark on a single machine for testing and debugging.

1. Standalone Mode – Uses Spark’s built-in cluster manager for small to medium workloads.
2. YARN Mode – Runs Spark on Hadoop’s YARN for resource management in large clusters.
3. Kubernetes Mode – Deploys Spark applications in containerized environments.
4. Mesos Mode – Uses Apache Mesos for cluster management and resource allocation.
5. Client Mode – The driver runs on the local machine and executes Spark jobs.
6. Cluster Mode – The driver runs within the cluster for better scalability.
7. Cloud Mode – Deploys Spark on cloud platforms like AWS, Azure, or Google Cloud.
8. **What is spark context or how to create spark instance**

**ANS:**

SparkContext is the entry point for Spark applications, managing execution.  
It initializes RDDs, broadcast variables, and accumulators.  
A Spark instance is created using SparkConf, specifying configurations.  
Example in Python:from pyspark import SparkConf, SparkContext

conf = SparkConf().setAppName("MyApp").setMaster("local")

sc = SparkContext(conf=conf)

In Spark 2.0+, SparkSession is preferred over SparkContext for DataFrames.

SparkSession provides a unified entry point for working with Spark’s components.

Multiple SparkContexts cannot run in the same JVM but multiple SparkSessions can.

Closing SparkContext is necessary to free resources: sc.stop().

1. **Explain RDD in Apache spark**

**ANS:**

1. RDD (Resilient Distributed Dataset) is the core data structure in Spark.
2. It is an immutable, distributed collection of objects stored across nodes.
3. RDDs support parallel processing and fault tolerance via lineage tracking.
4. Two types of operations: Transformations (map, filter) and Actions (collect, count).
5. Transformations are lazy-evaluated, executed only when an action is triggered.
6. RDDs can be created from existing data sources like HDFS, S3, or local files.
7. RDDs can be persisted in memory or disk to optimize computations.

Example of RDD creation in Python:  
  
rdd = sc.parallelize([1, 2, 3, 4, 5])

filtered\_rdd = rdd.filter(lambda x: x % 2 == 0)

print(filtered\_rdd.collect()) # Output: [2, 4]

**C. Conclusion:**

Write the significance of the topic studied in the experiment.

**D. References:**

1. https://drive.google.com/file/d/1Mc5J1ZSOS\_-QYInDkHfyLU9nvQA2jSBa/view?usp=sharing 2. https://www.analyticsvidhya.com/blog/2021/06/real-time-data-streaming-using-apache-spark/ 3. https://www.tutorialspoint.com/apache\_spark/apache\_spark\_tutorial.pdf

**---------------------------------**