CS5229 - Big Data Analytics Technologies

R.Chaumika
MSc in CS
248213J

Introduction to MapReduce

– – –

- MapReduce is a Java-based, distributed execution framework within the Apache Hadoop Ecosystem.
- MapReduce performs the processing of large data sets in distributed and parallel manner.
- MapReduce consists of two distinct tasks - Map and Reduce
- MapReduce works by breaking down a large dataset into smaller parts and processing them in parallel across multiple nodes in a cluster.
- Applications: Entertainment, E-Commerce, Data warehousing, Fraud detection.

Introduction to Apache Spark

- It is a distributed computing system designed for big data processing and analysis.
- Spark provides a unified engine for processing batch,streaming and machine-learning workloads.
- It is built on top of the HDFS and provides APIs for several programming languages including Java, Scala and Python.
- Spark also provides a wide range of built-in libraries for ML, graph processing and stream processing, making it easier for developers to build complex data pipelines.
- Applications: E-Commerce, Finance, Healthcare, Social media.

DEMO

MapReduce	Spark
It is having a very slow speed as compared to Apache Spark.	It is much faster than MapReduce.
It is unable to handle real-time processing.	It can deal with real-time processing.
It is difficult to program as you required code for every process.	It is easy to program.
It supports more security projects.	Its security is not as good as MapReduce and continuously working on its security issues.
For performing the task, It is unable to cache in memory.	It can cache the memory data for processing its task.
Its scalability is good as you can add up to n different nodes.	It is having low scalability as compared to MapReduce.
It actually needs other queries to perform the task.	It has Spark SQL as its very own query language.

THANK YOU !!!