## Syllabus Part 1

### Course Description

This course provides a comprehensive introduction to the fundamental components of big data processing and analytics. Students will gain practical knowledge and hands-on experience with popular big data technologies, including Hadoop, MapReduce, Hive, Spark, HBase, Kafka, Solr, and Nifi. The course covers the core concepts, architectures, and tools used in big data environments, preparing students to tackle large-scale data processing, storage, and analysis challenges. Through a combination of lectures, demonstrations, and hands-on exercises, students will learn how to harness the power of these components to effectively manage and derive insights from massive volumes of data.

### Course Prerequisites

None

### Course Objectives

Students who successfully complete this course will be able to:

1. Understand the fundamentals of big data and its challenges.
2. Gain proficiency in deploying and configuring Big Data clusters.
3. Develop MapReduce programs for distributed data processing.
4. Utilize Hive for data warehousing and SQL-like querying on Hadoop.
5. Master the concepts and techniques of Apache Spark for advanced data processing and analytics.
6. Learn how to store and retrieve data using HBase and other NoSQL databases.
7. Explore real-time data streaming and processing with Kafka.
8. Build search applications and perform search analytics using Solr.
9. Manage data flow and create pipelines with Nifi for efficient data processing.
10. Evaluate and select appropriate big data architectures for specific use cases.
11. Apply the acquired knowledge and skills to complete a real-world big data project.
12. By the end of this course, students will have a strong foundation in big data technologies and be equipped with the necessary skills to tackle big data challenges in various domains, such as business intelligence, data analytics, and machine learning. They will also have completed a comprehensive big data project that showcases their ability to design and implement scalable solutions for processing and analyzing large volumes of data.

### Grading Scale

* 93 – 100% = A
* 90 – 92% = A-
* 87 – 89% = B+
* 83 – 86% = B
* 80 – 82% = B-
* 77 – 79% = C+
* 73 – 76% = C
* 70 – 72% = C-
* 67 – 69% = D+
* 63 – 66% = D
* 60 – 62% = D-
* 0 – 59% = F

### Topic Outline

**Week 1: Introduction to Big Data and Infrastructure**

* Overview of big data concepts and challenges
* Introduction to Hadoop ecosystem and its components
* Setting up a Big Data cluster
* Understanding distributed storage and processing

**Week 2: Hadoop Basics and MapReduce**

* Understanding Hadoop Distributed File System (HDFS)
* MapReduce fundamentals and its role in big data processing
* Writing and executing MapReduce programs
* Hands-on exercises with Hadoop and MapReduce

**Week 3: Hive and Data Warehousing**

* Introduction to Hive and its role in big data analytics
* HiveQL: SQL-like querying on Hadoop
* Data warehousing concepts in big data
* Creating tables, loading data, and running queries in Hive

**Week 4: Spark Fundamentals**

* Introduction to Apache Spark and its advantages over MapReduce
* Spark architecture and components
* RDD (Resilient Distributed Datasets) and transformations
* Hands-on exercises with Spark

**Week 5: Advanced Spark Programming**

* Working with Spark DataFrames and Datasets
* Spark SQL for querying structured data
* Spark Streaming for real-time data processing
* Implementing machine learning algorithms with Spark MLlib

**Week 6: HBase and NoSQL Databases**

* Introduction to HBase and its role in big data storage
* Understanding NoSQL databases and their characteristics
* Creating tables, inserting and retrieving data in HBase
* HBase schema design and best practices

**Week 7: Kafka and Real-time Data Streaming**

* Introduction to Apache Kafka and its role in data streaming
* Kafka architecture and components
* Producing and consuming messages with Kafka
* Building real-time data processing pipelines with Kafka

**Week 8: Solr and Search Analytics**

* Introduction to Apache Solr and its role in search analytics
* Indexing and querying data with Solr
* Faceted search and filtering in Solr
* Building search applications using Solr

**Week 9: Nifi and Data Flow Management**

* Introduction to Apache Nifi and its role in data flow management
* Creating data pipelines with Nifi
* Data ingestion, transformation, and routing with Nifi
* Monitoring and managing data flows in Nifi

**Week 10: Big Data Architectures**

* Understanding different big data architectures
* Batch processing vs. real-time processing
* Lambda architecture and Kappa architecture
* Evaluating and selecting appropriate architectures for specific use cases

**Week 11-12: Big Data Project**

* Students will work on a hands-on project applying the concepts learned throughout the course
* Project planning, data acquisition, processing, and analysis
* Implementing a scalable and efficient big data solution
* Presenting the project results and lessons learned