BIG DATA TECHNOLOGIES

ULTIMATE BIG DATA TECHNOLOGIES ROADMAP – THE BEST EVER!
Brought to you by Path2Proficiency – Master Big Data with Cutting-Edge Technologies!

Goal: Learn to process, analyze, and manage massive datasets using modern Big Data technologies. Develop expertise in distributed systems, batch/stream processing, and scalable data storage.

Why This Roadmap?

- **From Basics to Advanced** – Covers fundamental concepts, hands-on implementation, and real-world projects.
- **Master Distributed Data Processing Frameworks** – Spark, Hadoop, Flink, and more.
- **Learn Scalable Storage and NoSQL Databases** – HDFS, MongoDB, Cassandra, etc.
- **Work on Real-Time Stream Processing** – Kafka, Apache Storm, and others.

Phase 1: Understanding Big Data Fundamentals

Goal: Build a solid foundation in Big Data and Distributed Systems.

1. Introduction to Big Data Concepts

Topics to Cover:

- What is Big Data?
- Characteristics of Big Data (Volume, Velocity, Variety, Veracity, and Value)
- Data Generation Sources: IoT, Social Media, Sensors, etc.
- Traditional Databases vs. Big Data

References:

- Introduction to Big Data – Coursera
- Big Data Fundamentals – Udemy
- Big Data Explained – YouTube

Hands-On Task:

- Analyze basic datasets using Python (Pandas & NumPy).
- Explore JSON, CSV, and XML data formats.

2. Big Data Ecosystem Overview

Topics to Cover:

- Hadoop Ecosystem Overview

- Batch vs. Stream Processing
- Lambda and Kappa Architectures
- Role of ETL, Data Warehouses, and Data Lakes

References:

- Hadoop Ecosystem Explained – Medium
- Lambda and Kappa Architecture – YouTube

Hands-On Task:

- Create a data pipeline using batch and streaming data.
- Build a simple data lake with HDFS.

Phase 2: Distributed Storage Systems

Goal: Learn how to store and manage large datasets efficiently.

3. Hadoop Distributed File System (HDFS)

Topics to Cover:

- Basics of HDFS Architecture
- NameNode, DataNode, and Block Replication
- Read/Write Operations in HDFS
- HDFS Fault Tolerance and Scalability

References:

- HDFS Overview – Cloudera
- HDFS Crash Course – YouTube

Hands-On Task:

- Set up Hadoop on your local machine.
- Upload and retrieve data from HDFS.

4. NoSQL Databases and Key-Value Stores

Topics to Cover:

- Types of NoSQL Databases (Key-Value, Column, Document, Graph)
- MongoDB, Cassandra, Redis, and HBase
- ACID vs. BASE Properties

References:

- MongoDB Tutorial – MongoDB University
- Cassandra Basics – Datastax

Hands-On Task:

- Implement a key-value store using Redis.
- Create and query documents in MongoDB.

Phase 3: Batch Processing Frameworks

Goal: Master batch processing frameworks like Hadoop and Spark.

5. Apache Hadoop and MapReduce

Topics to Cover:

- Hadoop Cluster Setup and Configuration
- MapReduce Programming Model
- Combiner and Partitioner Concepts
- YARN and Resource Management

References:

- MapReduce Basics – Cloudera
- Hadoop Hands-On – YouTube

Hands-On Task:

- Build a MapReduce job to count word frequencies.
- Execute MapReduce tasks on a Hadoop cluster.

6. Apache Spark for Distributed Data Processing

Topics to Cover:

- Spark Architecture (RDDs, DAGs, and Executors)
- Spark vs. Hadoop – Key Differences
- PySpark, Spark SQL, and Spark Streaming
- Performance Tuning in Spark

References:

- Apache Spark Basics – Databricks
- Spark with Python – YouTube

Hands-On Task:

- Process large datasets using PySpark.
- Implement transformations and actions on RDDs.

Phase 4: Real-Time Stream Processing

Goal: Master real-time data processing using Kafka, Flink, and Storm.

7. Apache Kafka for Real-Time Messaging

Topics to Cover:

- Kafka Architecture (Broker, Producer, Consumer)
- Topic, Partition, and Offset Management
- Kafka Consumer Group and Load Balancing

References:

- Kafka Fundamentals – Confluent
- Kafka in Action – YouTube

Hands-On Task:

- Set up a Kafka cluster.
- Publish and consume messages in Kafka.

8. Apache Flink and Stream Processing

Topics to Cover:

- Flink vs. Spark Streaming
- Flink Architecture (Job Manager, Task Manager)
- Windowing and Stateful Processing in Flink

References:

- Apache Flink Crash Course – YouTube
- Flink Documentation – Apache

Hands-On Task:

- Implement stream processing using Apache Flink.
- Perform windowed operations on streaming data.

9. Apache Storm for Real-Time Computation

Topics to Cover:

- Storm Architecture (Spouts, Bolts, and Topologies)
- Fault Tolerance and Scalability in Storm
- Use Cases of Storm in Real-Time Applications

References:

- Storm Basics – Apache Storm
- Apache Storm Hands-On – YouTube

Hands-On Task:

- Build a real-time data pipeline using Apache Storm.
- Process real-time logs using Storm topologies.

Phase 5: Data Warehousing and ETL

Goal: Learn data warehousing and ETL pipeline development.

10. Data Warehousing and ETL Concepts

Topics to Cover:

- ETL (Extract, Transform, Load) Basics
- Data Warehousing Concepts (Star vs. Snowflake Schema)
- Apache Hive and Impala for Querying Big Data

References:

- ETL Fundamentals – Coursera
- Apache Hive Basics – Cloudera

Hands-On Task:

- Design and implement an ETL pipeline.
- Query large datasets using Apache Hive.

11. Apache NiFi for Data Flow Automation

Topics to Cover:

- NiFi Architecture (Processor, FlowFile, and Connection)
- Building Data Flow Pipelines with NiFi
- Integrating NiFi with Kafka and Hadoop

References:

- NiFi Documentation – Apache
- NiFi Crash Course – YouTube

Hands-On Task:

- Automate data flow between Kafka and Hadoop using NiFi.
- Create ETL workflows with NiFi processors.

Phase 6: Machine Learning and Big Data Analytics

Goal: Build machine learning models on large-scale datasets.

12. Machine Learning with Spark MLlib

Topics to Cover:

- Overview of Spark MLlib and Pipelines
- Distributed Model Training and Hyperparameter Tuning
- Feature Engineering with Spark

References:

- Spark MLlib Tutorial – Databricks
- MLlib Hands-On – YouTube

Hands-On Task:

- Build and deploy machine learning models using Spark MLlib.
- Perform feature scaling and selection in distributed datasets.

13. Big Data Analytics with Presto and ClickHouse

Topics to Cover:

- Presto and ClickHouse Architecture
- SQL-on-Hadoop Technologies
- Real-Time Analytics on Large Datasets

References:

- Presto Documentation – PrestoSQL
- ClickHouse Basics – YouTube

Hands-On Task:

- Query large datasets using Presto.
- Analyze clickstream data with ClickHouse.

Phase 7: Security and Governance in Big Data

Goal: Learn security best practices and data governance in Big Data.

14. Data Security and Privacy in Big Data

Topics to Cover:

- Data Encryption and Access Control
- Kerberos and Ranger for Hadoop Security
- Role-Based Access Control (RBAC)

References:

- Hadoop Security – Cloudera
- Kerberos Authentication Basics – YouTube

Hands-On Task:

- Configure Kerberos Authentication in Hadoop.
- Implement role-based security in Apache Ranger.

Phase 8: Capstone Projects & Portfolio Building

Goal: Build and showcase industry-level Big Data projects.

15. Real-World Capstone Project Ideas

Project Ideas:

- **Real-Time Clickstream Analysis for E-Commerce**
- **Log Data Processing and Anomaly Detection with Kafka and Flink**
- **Building a Scalable ETL Pipeline with Apache NiFi**
- **Fraud Detection in Banking using Spark and MLlib**
- **IoT Sensor Data Processing with Kafka and HDFS**

Estimated Timeline to Master Big Data:

Beginner to Intermediate: 3-4 months
Intermediate to Advanced: 5-6 months
Capstone and Deployment: 2-3 months

By Following This Roadmap, You Will:

Master Distributed Data Processing with Spark, Hadoop, and Flink.
Develop Real-Time Data Pipelines Using Kafka and NiFi.
Build and Deploy ML Models on Big Data Platforms.

Powered by Path2Proficiency – Your Guide to Excellence!