Course Title: Big Data Fundamentals & Concepts

- **Duration:** 5 Days (approximately 35-40 instructional hours)
- **Course Goal:** To provide participants with a comprehensive, foundational understanding of the big data ecosystem, including its core concepts, key technologies, common architectures, and business applications. This course will demystify the terminology and technologies of big data, enabling professionals to participate in strategic discussions, understand the potential, and contribute to data-driven initiatives in their organizations.
- **Core Focus:** Understanding the "what, why, and how" of big data from a conceptual and strategic perspective, rather than deep, hands-on coding.

Learning Objectives:

Upon successful completion of this workshop, participants will be able to:

- **Define** Big Data and articulate the "5 Vs" (Volume, Velocity, Variety, Veracity, Value).
- **Differentiate** between traditional data processing and big data paradigms.
- **Understand** the core components of the Hadoop ecosystem, including HDFS and MapReduce (and its successor, Spark).
- **Explain** the principles of distributed storage and distributed processing.
- **Describe** the role of modern data processing engines like Apache Spark.
- **Differentiate** between various types of NoSQL databases and their use cases.
- **Understand** the process of a typical big data pipeline, from ingestion to analysis.
- **Identify** common business use cases and applications of big data across industries.
- **Recognize** the key roles in a big data team (e.g., Data Engineer, Data Scientist).
- **Discuss** the strategic implications of cloud computing, data governance, and ethics in the context of big data.
- **Develop** a high-level architectural concept for a big data solution to a business problem.

Target Audience:

- Business Managers, Strategists, and Decision-Makers.
- IT Leaders, Project Managers, and Program Managers.
- Business Analysts, Data Analysts, and BI Professionals transitioning to big data.
- Marketing, Finance, and Operations professionals who work with large datasets.
- Anyone seeking a foundational understanding of the big data landscape.

Prerequisites:

- Basic computer literacy and a foundational understanding of data concepts (e.g., what a database is).
- No programming or deep technical expertise is required.

Teaching Methodology:

This is an interactive and conceptual workshop focused on building understanding. The methodology includes:

- **Interactive Lectures & Analogies:** Using clear, simple language and relatable analogies to explain complex technical concepts.
- **Case Study Analysis:** Deconstructing how real-world companies use big data to solve problems and create value.
- **Live Demonstrations (High-Level):** The instructor will showcase the interfaces and outputs of various big data tools (without requiring participants to code) to make them tangible.
- **Group Discussions & Brainstorming:** Exploring use cases, challenges, and strategic implications.
- **Architectural Design Workshops:** Participants work in groups to sketch out high-level solutions to business problems.

ADEPT TRAINING CENTER

Materials Provided:

- Comprehensive digital or printed workbook with key concepts, architectural diagrams, and case study summaries.
- A "Big Data Glossary" of essential terminology.
- Templates for a high-level big data solution design.
- A curated list of resources for continued learning (articles, videos, blogs).
- Certificate of Completion.

Detailed Day-by-Day Curriculum

Day 1: The "Why" – The Big Data Revolution

- **Module 1: What is Big Data and Why Now? (Approx. 2 hours)**
    - Welcome, introductions, and course objectives.
    - The Data Deluge: The exponential growth of data.
    - Defining Big Data: The **3 Vs (Volume, Velocity, Variety)**.
    - Expanding the definition: The **5 Vs (adding Veracity and Value)**.
    - The limitations of traditional databases and data warehouses in the face of big data.
    - **Activity:** "Big Data in My Industry" – Participants brainstorm examples of Volume, Velocity, and Variety of data in their respective fields.
- **Module 2: The Big Data Paradigm Shift (Approx. 2.5 hours)**
    - **Scaling Up (Vertical Scaling):** The traditional approach of buying bigger servers.
    - **Scaling Out (Horizontal Scaling):** The big data approach of using many commodity computers.
    - The core concepts of distributed, fault-tolerant systems.
    - Shifting from "Schema-on-Write" (traditional) to "Schema-on-Read" (big data).
- **Module 3: The Big Data Ecosystem: An Overview (Approx. 3 hours)**
    - Introducing the major categories of tools in the big data landscape.
    - A high-level map of a typical Big Data Pipeline:
        1. **Ingestion:** Getting data in.
        2. **Storage:** Storing it reliably.
        3. **Processing:** Transforming and analyzing it.
        4. **Analysis & Visualization:** Making sense of it.
    - Introduction to the Apache Hadoop project as the foundation.
    - **End of Day 1 Recap & Q&A.**

Day 2: The Foundation – Hadoop & Distributed Storage

- **Module 4: Introduction to the Hadoop Ecosystem (Approx. 2.5 hours)**
    - The history and significance of Hadoop.
    - Hadoop's two core components: **HDFS** (Storage) and **MapReduce** (original Processing).
    - The role of **YARN** (Yet Another Resource Negotiator) as the "operating system" of the Hadoop cluster.
- **Module 5: Deep Dive into HDFS (Hadoop Distributed File System) (Approx. 3 hours)**
    - The principles of distributed storage.
    - How HDFS works:
        - Breaking large files into **Blocks**.
        - **Replication** for fault tolerance.
        - The role of the NameNode (the "librarian") and DataNodes (the "bookshelves").

ADEPT TRAINING CENTER

- Understanding how HDFS enables massive, low-cost data storage.
- **Analogy Workshop:** Using a "giant library" analogy to explain HDFS concepts.

- **Module 6: Understanding MapReduce (The Original Processing Engine) (Approx. 2 hours)**
  - The concept of "bringing the compute to the data."
  - The two phases:
    - **Map Phase:** Parallel processing on each data node.
    - **Reduce Phase:** Aggregating the results from the map phase.
  - A simple, conceptual walkthrough of a "word count" example.
  - Understanding the limitations of MapReduce (batch-oriented, slow).
  - **End of Day 2 Recap & Q&A.**

Day 3: Modern Processing & Data Storage

- **Module 7: The Modern Engine: Apache Spark (Approx. 3 hours)**
  - Why Spark has largely replaced MapReduce.
  - Key advantages: In-memory processing, speed, and versatility.
  - Core concepts (explained simply): RDDs, DataFrames, and lazy evaluation.
  - Spark's ecosystem: Spark SQL, MLlib (Machine Learning), Streaming, GraphX.
  - **Demonstration:** Showing the speed difference between a MapReduce and a Spark job (via a pre-recorded video or a live demo).
- **Module 8: SQL on Big Data: Apache Hive (Approx. 2 hours)**
  - The challenge: Not everyone is a programmer.
  - How Hive allows users to write standard SQL queries that are translated into processing jobs (MapReduce or Spark).
  - The concept of an external table and schema-on-read.
  - The role of Hive in data warehousing and business intelligence on big data.
- **Module 9: NoSQL Databases: Beyond Relational (Approx. 2.5 hours)**
  - The limitations of rigid schemas in relational databases for big data.
  - Introduction to NoSQL.
  - The four main types of NoSQL databases and their use cases:
    1. **Key-Value Stores** (e.g., Redis) - for caching, session management.
    2. **Document Databases** (e.g., MongoDB) - for product catalogs, user profiles.
    3. **Column-Family Stores** (e.g., Cassandra, HBase) - for time-series, IoT data.
    4. **Graph Databases** (e.g., Neo4j) - for social networks, recommendation engines.
  - **Workshop:** "Choose the Right Database" – Participants are given several scenarios and must choose the most appropriate NoSQL (or relational) database.
  - **End of Day 3 Recap & Q&A.**

Day 4: The Data Pipeline, Use Cases & Roles

- **Module 10: Data Ingestion: Getting Data into the System (Approx. 2.5 hours)**
  - **Batch Ingestion:** Moving large, bounded datasets. Tools like **Apache Sqoop** (for relational databases).
  - **Streaming Ingestion:** Moving continuous, real-time data. Tools like **Apache Flume** and **Apache Kafka** (for logs, events, IoT data).
  - The concept of a data pipeline or ETL/ELT process in a big data context.
- **Module 11: Real-World Big Data Use Cases & Applications (Approx. 3 hours)**
  - A deep dive into several industry case studies:
    - **Retail/E-commerce:** Recommendation engines, customer 360, market basket analysis.

ADEPT TRAINING CENTER

        - **Finance:** Fraud detection, risk modeling.
        - **Healthcare:** Genomic analysis, predictive health.
        - **Telecommunications:** Network optimization, customer churn prediction.
        - **Manufacturing/IoT:** Predictive maintenance.
    - **Case Study Workshop:** In groups, participants choose an industry and develop a high-level big data use case.
- **Module 12: Roles & Skills in the Big Data World (Approx. 2 hours)**
    - Defining the key roles:
        - **Data Engineer:** Builds and maintains the data pipelines and infrastructure.
        - **Data Analyst:** Uses tools like Hive/SQL to query data and create reports/dashboards.
        - **Data Scientist:** Applies statistical analysis and machine learning to build predictive models.
    - The skills required for each role and how they collaborate.
    - **End of Day 4 Recap & Q&A.**

Day 5: The Cloud, Governance & Your Big Data Strategy

- **Module 13: Big Data in the Cloud (AWS, Azure, GCP) (Approx. 2.5 hours)**
    - Why the cloud is the dominant platform for big data today.
    - Benefits: Scalability, elasticity, pay-as-you-go, managed services.
    - Overview of key big data services on a major cloud provider (e.g., AWS):
        - **Storage:** S3 (Data Lake).
        - **Processing:** EMR (Managed Hadoop/Spark).
        - **Data Warehousing:** Redshift.
        - **Streaming:** Kinesis.
- **Module 14: Big Data Governance, Security & Ethics (Approx. 2 hours)**
    - The challenge of ensuring data quality (Veracity) at scale.
    - Data security in a distributed environment.
    - Data privacy and regulatory compliance (e.g., GDPR).
    - The ethical implications of big data and AI.
- **Module 15: Capstone Workshop: Designing a Big Data Solution (Approx. 2.5 hours)**
    - **Final Workshop:** Participants are given a new, detailed business problem (e.g., "A retail company wants to analyze customer website clickstream data to personalize marketing").
    - In teams, they must sketch out a high-level architectural solution, identifying:

1. Key data sources.
2. The right ingestion tools (batch vs. streaming).
3. The storage solution (e.g., HDFS/S3).
4. The processing engine (e.g., Spark).
5. The analysis tools (e.g., Hive, BI tools).

    - Teams present their architectural diagrams for peer and instructor feedback.

- **Module 16: Course Review & The Future of Data (Approx. 1 hour)**
    - Review of the key concepts of the big data ecosystem.
    - The future: The convergence of big data, AI, and real-time analytics.
    - Open forum for final questions and how to continue learning.
    - Course evaluation and closing.

ADEPT TRAINING CENTER