Pre-requisites for the Big Data Hadoop Training Course?
There will be no pre-requisites but Knowledge of Java/ Python, SQL, Linux will be beneficial, but not mandatory. Ducat provides a crash course for pre-requisites required to initiate Big Data training.

Apache Hadoop on AWS Cloud
This module will help you understand how to configure Hadoop Cluster on AWS Cloud:
- Introduction to Amazon Elastic MapReduce
- AWS EMR Cluster
- AWS EC2 Instance: Multi Node Cluster Configuration
- AWS EMR Architecture
- Web Interfaces on Amazon EMR
- Amazon S3
- Executing MapReduce Job on EC2 & EMR
- Apache Spark on AWS, EC2 & EMR
- Submitting Spark Job on AWS
- Hive on EMR
- Available Storage types: S3, RDS & DynamoDB
- Apache Pig on AWS EMR
- Processing NY Taxi Data using SPARK on Amazon EMR

Learning Big Data and Hadoop
This module will help you understand Big Data:
- Common Hadoop ecosystem components
- Hadoop Architecture
- HDFS Architecture
- Anatomy of File Write and Read
- How MapReduce Framework works
- Hadoop high level Architecture
- MR2 Architecture
- Hadoop YARN
- Hadoop 2.x core components
- Hadoop Distributions
- Hadoop Cluster Formation

Hadoop Architecture and HDFS
This module will help you to understand Hadoop & HDFS Cluster Architecture:
- Configuration files in Hadoop Cluster (FSimage & editlog file)
- Setting up of Single & Multi node Hadoop Cluster
- HDFS File permissions
- HDFS Installation & Shell Commands
- Deamons of HDFS
  - Node Manager
  - Resource Manager
  - NameNode
  - DataNode
Hadoop MapReduce Framework
This module will help you to understand Hadoop MapReduce framework:
- How MapReduce works on HDFS data sets
- MapReduce Algorithm
- MapReduce Hadoop Implementation
- Hadoop 2.x MapReduce Architecture
- MapReduce Components
- YARN Workflow
- MapReduce Combiners
- MapReduce Partitioners
- MapReduce Hadoop Administration
- MapReduce APIs
- Input Split & String Tokenizer in MapReduce
- MapReduce Use Cases on Data sets

Advanced MapReduce Concepts
This module will help you to learn:
- Job Submission & Monitoring
- Counters
- Distributed Cache
- Map & Reduce Join
- Data Compressors
- Job Configuration
- Record Reader

Pig
This module will help you to understand Pig Concepts:
- Pig Architecture
- Pig Installation
- Pig Grunt shell
- Pig Running Modes
- Pig Latin Basics
- Pig LOAD & STORE Operators
  - Diagnostic Operators
    - DESCRIBE Operator
    - EXPLAIN Operator
    - ILLUSTRATE Operator
    - DUMP Operator
  - Grouping & Joining
    - GROUP Operator
    - COGROUP Operator
    - JOIN Operator
    - CROSS Operator
  - Combining & Splitting
    - UNION Operator
    - SPLIT Operator
  - Filtering
    - FILTER Operator
    - DISTINCT Operator
    - FOREACH Operator
Sorting
  - ORDERBYFIRST
  - LIMIT Operator

Built in Functions
  - EVAL Functions
  - LOAD & STORE Functions
  - Bag & Tuple Functions
  - String Functions
  - Date-Time Functions
  - MATH Functions

Pig UDFs (User Defined Functions)

Pig Scripts in Local Mode

Pig Scripts in MapReduce Mode

Analysing XML Data using Pig

Analysing JSON data using Pig

Testing Pig Scripts

Hive
This module will build your concepts in learning:

- Hive Installation
- Hive Data types
- Hive Architecture & Components
- Hive Meta Store
- Hive Tables (Managed Tables and External Tables)
- Hive Partitioning & Bucketing
- Hive Joins & Sub Query
- Running Hive Scripts
- Hive Indexing & View
- Hive Queries (HQL); Order By, Group By, Distribute By, Cluster By, Examples
- Hive Functions: Built-in & UDF (User Defined Functions)
- Hive ETL: Loading JSON, XML, Text Data Examples
- Hive Querying Data
- Hive Tables (Managed & External Tables)
- Hive Used Cases
- Hive Optimization Techniques
  - Partitioning (Static & Dynamic Partition) & Bucketing
  - Hive Joins > Map + BucketMap + SMB (SortedBucketMap) + Skew
  - Hive FileFormats (ORC+SEQUENCE+TEXT+AVRO+PARQUET)
  - CBO
  - Vectorization
  - Indexing (Compact + BitMap)
  - Integration with TEZ & Spark
- Hive SerDer (Custom + InBuilt)
- Hive integration NoSQL (HBase + MongoDB + Cassandra)
- Thrift API (Thrift Server)
- UDF, UDTF & UDAF
- Hive Multiple Delimiters
- XML & JSON Data Loading HIVE.
- Aggregation & Windowing Functions in Hive
- Hive Connect with Tableau

Sqoop
- Sqoop Installation
- Loading Data from RDBMS using Sqoop
- Sqoop Import & Import-All-Table
- Fundamentals & Architecture of Apache Sqoop
- Sqoop Job
- Sqoop Codegen
- Sqoop Incremental Import & Incremental Export
- Sqoop Merge
- Import Data from MySQL to Hive using Sqoop
- Sqoop: Hive Import
- Sqoop Metastore
- Sqoop Use Cases
- Sqoop- HCatalog Integration
- Sqoop Script
- Sqoop Connectors

**Flume**
This module will help you to learn Flume Concepts:
- Flume Introduction
- Flume Architecture
- Flume Data Flow
- Flume Configuration
- Flume Agent Component Types
- Flume Setup
- Flume Interceptors
- Multiplexing (Fan-Out), Fan-In-Flow
- Flume Channel Selectors
- Flume Sync Processors
- Fetching of Streaming Data using Flume (Social Media Sites: YouTube, LinkedIn, Twitter)
- Flume + Kafka Integration
- Flume Use Cases

**KAFKA**
This module will help you to learn Kafka concepts:
- Kafka Fundamentals
- Kafka Cluster Architecture
- Kafka Workflow
- Kafka Producer, Consumer Architecture
- Integration with SPARK
- Kafka Topic Architecture
- Zookeeper & Kafka
- Kafka Partitions
- Kafka Consumer Groups
- KSQL (SQL Engine for Kafka)
- Kafka Connectors
- Kafka REST Proxy
- Kafka Offsets

**Oozie**
This module will help you to understand Oozie concepts:
- Oozie Introduction
- Oozie Workflow Specification
- Oozie Coordinator Functional Specification
- Oozie H-catalog Integration
- Oozie Bundle Jobs
- Oozie CLI Extensions
- Automate MapReduce, Pig, Hive, Sqoop Jobs using Oozie
- Packaging & Deploying an Oozie Workflow Application

**HBase**
This module will help you to learn HBase Architecture:
- HBase Architecture, Data Flow & Use Cases
- Apache HBase Configuration
- HBase Shell & general commands
- HBase Schema Design
- HBase Data Model
- HBase Region & Master Server
- HBase & MapReduce
DATA PROCESSING WITH APACHE SPARK

Spark executes in-memory data processing & how Spark Job runs faster than Hadoop MapReduce Job.

Course will also help you understand the Spark Ecosystem & its related APIs like Spark SQL, Spark Streaming, Spark MLib, Spark GraphX & Spark Core concepts as well.

This course will help you to understand Data Analytics & Machine Learning algorithms applying to various datasets to process & to analyze large amount of data.

- Spark RDDs.
- Spark RDDs Actions & Transformations.
- Spark SQL: Connectivity with various Relational sources & its convert it into Data Frame using Spark SQL.
- Spark Streaming.
- Understanding role of RDD.
- Spark Core concepts: Creating of RDDs: Parallel RDDs, MappedRDD, HadoopRDD, JdbcRDD.
- Spark Architecture & Components.

PROJECTS

1. Working with MapReduce, Pig, Hive & Flume
   Problem Statement: Fetch structured & unstructured data sets from various sources like Social Media Sites, Web Server & structured source like MySQL, Oracle & others and dump it into HDFS and then analyze the same datasets using PIG, HQL queries & MapReduce technologies to gain proficiency in Hadoop related stack & its ecosystem tools.
   Data Analysis Steps in:
   - Dump XML & JSON datasets into HDFS.
   - Convert semi-structured data formats(JSON & XML) into structured format using Pig, Hive & MapReduce.
   - Push the data set into PIG & Hive environment for further analysis.
   - Writing Hive queries to push the output into relational database(RDBMS) using Sqoop.
   - Renders the result in Box Plot, Bar Graph & others using R & Python integration with Hadoop.

2. Analyze Stock Market Data
   Industry: Finance
   Data: Data set contains stock information such as daily quotes, Stock highest price, Stock opening price on New York Stock Exchange.
   Problem Statement: Calculate Co-variance for stock data to solve storage & processing problems related to huge volume of data.
   - Positive Covariance, If investment instruments or stocks tend to be up or down during the same time periods, they have positive covariance.
   - Negative Co-variance, If return move inversely, If investment tends to be up while other is down, this shows Negative Co-variance.

3. Hive, Pig & MapReduce with New York City Uber Trips
   Problem Statement: What was the busiest dispatch base by trips for a particular day on entire month?
   - What day had the most active vehicles.
   - What day had the most trips sorted by most to fewest.
   - Dispatching_Base_Number is the NYC taxi & Limousine company code of that base that dispatched the UBER.
   - active_vehicles shows the number of active UBER vehicles for a particular date & company(base). Trips is the number of trips for a particular base & date.
Project #4: Analyze Tourism Data
Data: Tourism Data comprises contains : City Pair, seniors travelling, children traveling, adult traveling, car booking price & air booking price.
Problem Statement: Analyze Tourism data to find out:
- Top 20 destinations tourist frequently travel to: Based on given data we can find the most popular destinations where people travel frequently, based on the specific initial number of trips booked for a particular destination
- Top 20 high air-revenue destinations, i.e the 20 cities that generate high airline revenues for travel, so that the discount offers can be given to attract more bookings for these destinations.
- Top 20 locations from where most of the trips start based on booked trip count.

Project #5: Airport Flight Data Analysis: We will analyze Airport Information System data that gives information regarding flight delays, source & destination details diverted routes & others.
Industry: Aviation
Problem Statement: Analyze Flight Data to:
- List of Delayed flights.
- Find flights with zero stop.
- List of Active Airlines all countries.
- Source & Destination details of flights.
- Reason why flight get delayed.
- Time in different formats.

Project #6: Analyze Movie Ratings
Industry: Media
Data: Movie data from sites like rotten tomatoes, IMDB, etc. Problem Statement: Analyze the movie ratings by different users to:
- Get the user who has rated the most number of movies
- Get the user who has rated the least number of movies
- Get the count of total number of movies rated by user belonging to a specific occupation
- Get the number of underage users

Project #7: Analyze Social Media Channels:
- Facebook
- Twitter
- Instagram
- YouTube
Industry: Social Media
Data: DataSet Columns: VideoId, Uploader, Internal Day of establishment of You tube & the date of uploading of the video, Category, Length, Rating, Number of comments.
Problem Statement: Top 5 categories with maximum number of videos uploaded.
Problem Statement: Identify the top 5 categories in which the most number of videos are uploaded, the top 10 rated videos, and the top 10 most viewed videos.
Apart from these there are some twenty more use-cases to choose: Twitter Data Analysis
Market data Analysis

Partners:

NOIDA
A-43 & A-52, Sector-16, Noida - 201301, (U.P.) INDIA
+91 99-9999-3213

GHAZIABAD
1, Arund Industrial Estate, Near ITS College, Mohan Nagar, Ghaziabad (U.P.)
+91 99-9999-3213

PITAMPURA (DELHI)
Pilot No. 366, 2nd Floor, Kohat Enclave, Pitampura, (Near Kohat Metro Station) Above Allahabad Bank, New Delhi- 110034.
+91 99-9999-3213

SOUTH EXTENSION (DELHI)
D-27, South Extension-1 New Delhi-110049
+91 99-9999-3213

GURGAON
1808/2, 2nd floor old DLF, Near Honda Showroom, Sec.-14, Gurgaon (Haryana)
+91 99-9999-3213

Java

www.facebook.com/ducateducation