Do It Skills
do it. enjoy it.

MASTERS OF

DATA ENGINEERING

Master New Skills from the Comfort of Your Home with Our Online Courses!

☎ +91-989-065-2675

www.doitskills.com

About Us

DOITSKILLS is a video learning platform for online education, with certificate courses specially curated for college students and young adults. DOITSKILLS aims at making every graduate job ready, with professional skills and practical know-how for the most sought-after fields of work.

> “Our aim is simple: We strive to create high-impact, hands-on experiences that prepare learners for meaningful and productive careers.
> -DOITSKILLS”

What you will Learn:

- **Data Engineering Basics**
- **Database Management**
- **Data Visualization**
- **Data Security**
- **Real-time Processing**
- **Cloud Platforms**
- **Data Processing**
- **Pipeline Orchestration**
- **Performance Optimization**
- **Containerization**

Course Curriculum

Hadoop(HDFS)

- Architecture
- HDFS features
- Read and Write Operations in HDFS
- HDFS Developer commands, HDFS Admin commands
- HDFS Data Blocks
- Rack Awareness
- High Availability
- Fault Tolerance
- Name Node High Availability
- HDFS Federation

Hadoop(Map Reduce)

- Introduction
- Architecture
- Mapper, Shuffle, Sort, Reducer
- Key-Value Pairs
- Input format, Input split, Record reader, Output format
- Partitioner, Combiner
- Map Side Join, Reduce Side Join, Distributed Cache
- Counter
- Performance-tuning Map Reduce

Hadoop(HIVE)

- Introduction
- Architecture
- Built-In Functions
- UDFs (UDF, UDAF, UDTF)
- DDL Commands (CREATE, SHOW, DESCRIBE, USE, DROP, ALTER, TRUNCATE)
- DML Commands (LOAD, SELECT, INSERT, DELETE, UPDATE, EXPORT, IMPORT)
- Apache Hive View and Hive Index

- Hive Metastore – Different Ways to Configure Hive Metastore
- Hive Data Model – Table, Partition, Bucket
- Hive Data Types – Primitive and Complex Data Types in Hive-Complex data types: Array, Struct, Map
- Hive Operators – Relational Operators, Arithmetic Operators, Logical Operators, String Operators, Operators on Complex Types
- Hive SerDe – Custom & Built-in SerDe in Hive (e.g., JsonSerde, OpenCSVSerde, ParquetSerde, OrcSerde, XmlSerde, RegexSerde)
- Hive Partitions
- Types of Hive Partitioning with Examples
- Static Partitioning
- Dynamic Partitioning
- Bucketing in Hive – Creation of Bucketed Table in Hive
- Hive Join
- Types of Joins in Hive
- Inner Join
- Left Outer Join
- Right Outer Join
- Full Outer Join
- Self Join
- Cross Join
- Map Join
- Bucket Map Join
- Skew Join
- Sort Merge Bucket Join
- Internal vs External Table
- Configure MySQL Metastore
- HQL Select Statements
- Group By
- Having
- Grouping Sets
- Rollup and Cube
- Order By Query
- Sort By

- Clustered By
- Window Functions
- Row_number
- Rank
- Dense_rank()
- Lead()
- Lag()
- First_value()
- Last_value()
- Hive Optimization Techniques – Hive Performance
- Hive Security
- Authentication
- Authorization
- Encryption
- Hive Transaction Management

Hadoop(Sqoop)

- Sqoop Architecture
- Sqoop Features
- Sqoop Eval
- Sqoop Import
- Sqoop Import-All Tables
- Sqoop Validation
- Sqoop Export
- Sqoop Incremental Jobs
- Sqoop Jobs
- Sqoop Codegen
- Sqoop Merge
- Sqoop Metastore
- Sqoop List-Databases
- Sqoop List-Tables
- Sqoop Connectors & Drivers
- Import from Mainframe
- Hcatalog Integration
- Troubleshooting Issues in Sqoop
- Sqoop Performance Tuning

Hadoop(Hbase)

- Hbase Architecture

- Hbase Features
- Hbase Use Cases
- Hbase Operations
- Hbase Commands
- Table Management Commands in HBase
- Data Manipulation HBase Commands (Create, Truncate, Scan)
- HBase Admin API (Class Descriptor & Class HBaseAdmin)
- HBase Client API (HTable, Put, Get, Delete, Result)
- HBase MemStore (Uses, Benefits & Configuration)
- HBase Security: Kerberos Authentication & Authorization
- HBase vs RDBMS: Feature-Wise Comparison
- HBase vs Impala: Comparison
- HBase Troubleshooting (Problem, Cause & Solution)
- HBase Performance Tuning: Optimization Methods

Apache Flume

- Apache Flume Tutorial-Flume Introduction, Features & Architecture
- Apache Flume Architecture-Flume Agent, Event, Client
- Apache Flume Features-Limitations of Apache Flume
- Apache Flume Use Cases-Future Scope in Flume
- Apache Flume Source-Types of Flume Source
- Apache Flume Sink-Types of Sink in Flume
- Apache Flume Sink Processors-Types of Sink Processors
- Flume Channel Selectors-Apache Flume
- Apache Flume Channel-Types of Channels in Flume
- Flume Event Serializers-Apache Flume
- Apache Flume Interceptors-Types of Interceptors in Flume
- Flume Data Flow-Types & Failure Handling in Apache Flume
- Data Transfer from Flume to HDFS-Load Log Data Into HDFS
- Flume Troubleshooting-Flume Known Issues & Its Compatibility

Apache Spark

- Spark: (Spark Core, SQL, Streaming, MYSQL Integration, MongoDB, Cassandra, Snowflakes, ElasticSearch, SparkKafka Streaming, Hbase Integration)

Spark Core

- Spark Introduction
- Apache Spark Ecosystem – Complete Spark Component
- Features of Apache Spark – Learn the benefits of using Spark
- Apache Spark Use Cases in Real Time
- Spark Shell Commands to Interact with Spark-Scala
- Spark Shell Commands to Interact with Spark-python
- Learn SparkContext, SparkSession – Introduction and Functions
- Spark Stage, Tasks – An Introduction to Physical Execution Plan
- Spark RDD – Introduction, Features & Operations of RDD
- RDD Persistence and Caching Mechanism in Apache Spark
- Shining Features of Spark RDD You Must Know
- Introduction to Apache Spark Paired RDD
- How to Overcome the Limitations of RDD in Apache Spark?
- Spark RDD Operations – Transformation & Action with Example
- RDD lineage in Spark: ToDebugString Method
- Apache Spark Map vs FlatMap Operation
- Spark In-Memory Computing – A Beginners Guide
- Lazy Evaluation, Fault Tolerance, Directed Acyclic Graph DAG in Apache Spark
- Apache Spark Cluster Managers – YARN, Mesos & Standalone, how it works
- Spark Performance Tuning – Learn to Tune Apache Spark Job

Spark SQL

- Apache Spark SQL Tutorial – Quick Introduction Guide
- Spark SQL Features
- Spark SQL DataFrame
- Spark Dataset
- Spark SQL Optimization – Understanding the Catalyst Optimizer
- Apache Spark RDD vs DataFrame vs DataSet
- Spark MySQL Integration

- Spark Hive Integration
- Spark MongoDB Integration (including MongoDB Hands-On)
- Spark Cassandra Integration (including Cassandra Hands-On)
- Spark Hbase Integration
- Spark Elasticsearch Integration (including Elasticsearch Hands-On)
- Spark Joining Strategies
    - Spark Joins (Inner Join, Left Outer Join, Right Outer Join, Self Join, Cross Join, Full Outer Join)
    - Skew Join
    - Broadcast Join
- Spark Storage Formats (Parquet, Avro, and ORC)
- Spark DataFrame API for Window Functions (Row_number, Rank, Dense_rank, Lead, Lag, First_value, and Last_value)
- Spark SQL APITop of Form

Spark Streaming

- Spark Streaming Introduction
- Apache Spark DStream (Discretized Streams)
- Apache Spark Streaming Transformation Operations
- Spark Streaming Checkpoint in Apache Spark
- Spark Watermarking Checkpoint in Apache Spark
- Spark Kafka Integration

AWS Big Data Engineer

AWS Data Engineer: (Lambda, Glue, EMR, Kinesis, DynamoDB, RDS, EC2, S3, Redshift)

Big Data on AWS Introduction

Big Data on AWS Introduction

Cloud Computing Introduction, Advantages, and Types

- Cloud Deployment Models
- Cloud Service Categories
- AWS Cloud Platform
- AWS Cloud Architecture Design Principles – Part I
- AWS Cloud Architecture Design Principles – Part II
- Why AWS for Big Data – Reasons and Challenges
- Databases in AWS
- Data Warehousing in AWS

- Redshift, Kinesis, and EMR
- DynamoDB, Machine Learning, and Lambda
- ElasticSearch Services and EC2

Big Data on AWS - Collection

- Amazon Kinesis and Kinesis Stream
- Kinesis Data Stream Architecture and Core Components
- Data Producer
- Data Consumer
- Kinesis Stream Emitting Data to AWS Services and Kinesis Connector Library
- Kinesis Firehose
- Demo – Put and Get Records from Kinesis Data Stream
- Transferring Data Using Lambda
- Amazon SQS Lifecycle and Architecture
- IoT and Big Data
- IoT Framework
- AWS Data Pipelines and Data Nodes
- Activity, Pre-Condition, and Schedule
- Demo – Importing Data from S3 into DynamoDB Using Data Pipeline

Big Data on AWS - Storage

- Amazon Glacier and Big Data
- DynamoDB Introduction
- DynamoDB and EMR
- DynamoDB Partitions and Distributions
- DynamoDB GSI LSI
- DynamoDB Stream and Cross-Region Replication
- DynamoDB Performance and Partition Key Selection
- Snowball and AWS Big Data
- AWS DMS
- AWS Aurora in Big Data
- Demo – Amazon Athena Interactive SQL Queries for Data in Amazon S3 Part I
- Demo – Amazon Athena Interactive SQL Queries for Data in Amazon S3 Part II

Big Data on AWS - Processing

- Amazon EMR

- Demo – Analyzing Big Data with Amazon EMR
- Apache Hadoop
- EMR Architecture
- EMR Operations – Releases and Cluster
- EMR Operations – Choosing Instance and Monitoring
- Demo – Advanced EMR Setting Options
- Hive on EMR
- HBase with EMR
- Presto with EMR
- Spark with EMR
- EMR File Storage
- Demo – Analyzing Large Datasets Using Hive and Spark
- AWS Lambda

Big Data on AWS - Analysis

- Redshift Intro and Use Cases
- Redshift Architecture
- MPP and Redshift in AWS Ecosystem
- Columnar Databases
- Redshift Table Design – Part I
- Redshift Table Design – Part II
- Demo – Generating Random Dataset in EC2 and Loading it in S3
- Demo – Redshift Maintenance and Operations
- Machine Learning Introduction
- Machine Learning Algorithm
- Amazon SageMaker
- Amazon Elasticsearch
- Amazon Elasticsearch Services
- Demo – Loading Datasets into Elasticsearch
- Logstash and RStudio
- Demo – Fetching the File and Analyzing it using RStudio
- Athena
- Demo – Running Query on S3 using the Serverless Athena
- Demo – Creating a Redshift Cluster and Loading the Datasets into it from S3 – Part I
- Demo – Creating a Redshift Cluster and Loading the Datasets into it from S3 – Part II

Big Data on AWS - Visualization

- Amazon QuickSight
- Demo – Creating an Analysis with a Single Visual using Sample Data
- Demo – Creating an Analysis using Your Own Amazon S3 Data
- Big Data Visualization

Big Data on AWS - Security

- EMR Security and Security Group
- Roles and Private Subnet
- Encryption at Rest and In-Transit
- Redshift Security
- Encryption at Rest using CloudHSM
- Cloud HSM versus AWS KMS
- Limit Data Access

Azure Big Data Engineer

Azure Data Engineer: (Azure Functions, Azure Blob Storage, Azure Data factory, Azure Databricks and Azure Synapse, Azure Event Hub)

Working with Azure Blob Storage

- Introduction to Azure Blob Storage

Working with Relational Databases in Azure

- Provisioning and connecting to an Azure SQL database using PowerShell
- Provisioning and connecting to an Azure PostgreSQL database using the Azure CLI
- Provisioning and connecting to an Azure MySQL database using the Azure CLI
- Implementing active geo-replication for an Azure SQL database using PowerShell
- Implementing an auto-failover group for an Azure SQL database using PowerShell
- Implementing vertical scaling for an Azure SQL database using PowerShell
- Implementing an Azure SQL database elastic pool using PowerShell
- Monitoring an Azure SQL database using the Azure portal

Analyzing Data with Azure Synapse Analytics

- Provisioning and connecting to an Azure Synapse SQL pool using PowerShell
- Pausing or resuming a Synapse SQL pool using PowerShell
- Scaling an Azure Synapse SQL pool instance using PowerShell
- Loading data into a SQL pool using PolyBase with T-SQL
- Loading data into a SQL pool using the COPY INTO statement
- Implementing workload management in an Azure Synapse SQL pool
- Optimizing queries using materialized views in Azure Synapse Analytics

Control Flow Transformation and the Copy Data Activity in Azure Data Factory

- Implementing HDInsight Hive and Pig activities
- Implementing an Azure Functions activity
- Implementing a Data Lake Analytics U-SQL activity
- Copying data from Azure Data Lake Gen2 to an Azure Synapse SQL pool using the copy activity
- Copying data from Azure Data Lake Gen2 to Azure Cosmos DB using the copy activity

Data Flows in Azure Data Factory

- Implementing incremental data loading with a mapping data flow
- Implementing a wrangling data flow

Azure Data Factory Integration Runtime

- Configuring a self-hosted IR
- Configuring a shared self-hosted IR
- Migrating an SSIS package to Azure Data Factory
- Executing an SSIS package with an on-premises data store

Deploying Azure Data Factory Pipelines

- Configuring the development, test, and production environments
- Deploying Azure Data Factory pipelines using the Azure portal and ARM templates
- Automating Azure Data Factory pipeline deployment using Azure DevOps

Batch and Streaming Data Processing with Azure Databricks

- Configuring the Azure Databricks environment
- Transforming data using Python
- Transforming data using Scala
- Working with Delta Lake
- Processing structured streaming data with Azure Databricks

GCP Big Data Engineer

GCP Data Engineer: (GCP Data Proc, Pub Sub, Apache Beam, Composer, Gcp SQL Data Storages, Big Query and NOSQL Database)

Getting Started with Data Engineering with GCP

- Introduction to Data Engineering on GCP
- Setting Up a GCP Account and Project
- Overview of GCP Data Services

Big Data Capabilities on GCP

- Understanding what the cloud is
- Getting started with Google Cloud Platform
- A quick overview of GCP services for data engineering

Building Solutions with GCP Components

- Building Solutions with GCP Components

Building a Data Warehouse in BigQuery

- Introduction to Google Cloud Storage and BigQuery
- Introduction to the BigQuery console
- Preparing the prerequisites before developing our data warehouse
- Practicing developing a data warehouse

Building Orchestration for Batch Data Loading Using Cloud Composer

- Introduction to Cloud Composer
- Understanding the working of Airflow
- Exercise: Build data pipeline orchestration using Cloud Composer

Building a Data Lake Using Dataproc

- Introduction to Dataproc
- Exercise – Building a data lake on a Dataproc cluster
- Exercise: Creating and running jobs on a Dataproc cluster
- Understanding the concept of the ephemeral cluster
- Building an ephemeral cluster using Dataproc and Cloud Composer

Processing Streaming Data with Pub/Sub and Dataflow

- Processing streaming data
- Exercise – Publishing event streams to cloud Pub/Sub
- Exercise – Using Cloud Dataflow to stream data from Pub/Sub to GCS

Visualizing Data for Making Data-Driven Decisions with Data Studio

- Unlocking the power of your data with Data Studio
- From data to metrics in minutes with an illustrative use case
- Understanding how Data Studio can impact the cost of BigQuery
- How to create materialized views and understanding how BI Engine works

Key Strategies for Architecting Top-Notch Data Pipelines

- Key Strategies for Architecting Top-Notch Data Pipelines

User and Project Management in GCP

- Understanding IAM in GCP
- Planning a GCP project structure
- Controlling user access to our data warehouse
- Practicing the concept of IaC using Terraform

Cost Strategy in GCP

- Estimating the cost of your end-to-end data solution in GCP
- Tips for optimizing BigQuery using partitioned and clustered tables

CI/CD on Google Cloud Platform for Data Engineers

- Introduction to CI/CD
- Understanding CI/CD components with GCP services
- Exercise – implementing continuous integration using Cloud Build
- Exercise – deploying Cloud Composer jobs using Cloud Build

Snowflakes-Getting Started with Snowflake

- Creating a new Snowflake instance
- Creating a tailored multi-cluster virtual warehouse
- Using the Snowflake WebUI and executing a query
- Using SnowSQL to connect to Snowflake
- Connecting to Snowflake with JDBC
- Creating a new account admin user and understanding built-in roles

Managing the Data Life Cycle

- Managing a database
- Managing a schema
- Managing tables
- Managing external tables and stages
- Managing views in Snowflake

Loading and Extracting Data into and out of Snowflake

- Configuring Snowflake access to private S3 buckets
- Loading delimited bulk data into Snowflake from cloud storage
- Loading delimited bulk data into Snowflake from your local machine
- Loading Parquet files into Snowflake
- Making sense of JSON semi-structured data and transforming to a relational view
- Processing newline-delimited JSON (or NDJSON) into a Snowflake table
- Processing near real-time data into a Snowflake table using Snowpipe
- Extracting data from Snowflake

Building Data Pipelines in Snowflake

- Creating and scheduling a task
- Conjugating pipelines through a task tree
- Querying and viewing the task history
- Exploring the concept of streams to capture table-level changes
- Combining the concept of streams and tasks to build pipelines that process changed data on a schedule
- Converting data types and Snowflake's failure management
- Managing context using different utility functions

Data Protection and Security in Snowflake

- Setting up custom roles and completing the role hierarchy
- Configuring and assigning a default role to a user
- Delineating user management from security and role management
- Configuring custom roles for managing access to highly secure data

- Setting up development, testing, pre-production, and production database hierarchies and roles
- Safeguarding the ACCOUNTADMIN role and users in the ACCOUNTADMIN role

Performance and Cost Optimization

- Examining table schemas and deriving an optimal structure for a table
- Identifying query plans and bottlenecks
- Weeding out inefficient queries through analysis
- Identifying and reducing unnecessary Fail-safe and Time Travel storage usage
- Projections in Snowflake for performance
- Reviewing query plans to modify table clustering
- Optimizing virtual warehouse scale

Secure Data Sharing

- Sharing a table with another Snowflake account
- Sharing data through a view with another Snowflake account
- Sharing a complete database with another Snowflake account and setting up future objects to be shareable
- Creating reader accounts and configuring them for non-Snowflake sharing
- Keeping costs in check when sharing data with non-Snowflake users

Back to the Future with Time Travel

- Using Time Travel to return to the state of data at a particular time
- Using Time Travel to recover from the accidental loss of table data
- Identifying dropped databases, tables, and other objects and restoring them using Time Travel
- Using Time Travel in conjunction with cloning to improve debugging
- Using cloning to set up new environments based on the production environment rapidly

Advanced SQL Techniques

- Managing timestamp data

- Shredding date data to extract Calendar information
- Unique counts and Snowflake
- Managing transactions in Snowflake
- Ordered analytics over window frames
- Generating sequences in Snowflake

Extending Snowflake Capabilities

- Creating a Scalar user-defined function using SQL
- Creating a Table user-defined function using SQL
- Creating a Scalar user-defined function using JavaScript
- Creating a Table user-defined function using JavaScript
- Connecting Snowflake with Apache Spark
- Using Apache Spark to prepare data for storage on Snowflake

Airflow-Building Data Pipelines – Extract Transform, and Load, Building Our Data Engineering Infrastructure

- Installing and configuring Apache NiFi
- Installing and configuring Apache Airflow
- Installing and configuring Elasticsearch
- Installing and configuring Kibana
- Installing and configuring PostgreSQL
- Installing pgAdmin 4

Reading and Writing Files

- Writing and reading files in Python
- Building data pipelines in Apache Airflow
- Handling files using NiFi processors

Working with Databases

- Inserting and extracting relational data in Python
- Inserting and extracting NoSQL database data in Python
- Building data pipelines in Apache Airflow
- Handling databases with NiFi processors

Cleaning, Transforming, and Enriching Data

- Performing exploratory data analysis in Python
- Handling common data issues using pandas
- Cleaning data using Airflow

Deploying Data Pipelines in Production, Features of a Production Pipeline

- Staging and validating data
- Building idempotent data pipelines

- Building atomic data pipelines

Version Control with the NiFi Registry

- Installing and configuring the NiFi Registry
- Using the Registry in NiFi
- Versioning your data pipelines
- Using git-persistence with the NiFi Registry

Monitoring Data Pipelines

- Monitoring NiFi using the GUI
- Monitoring NiFi with processors
- Using Python with the NiFi REST API

Deploying Data Pipelines

- Finalizing your data pipelines for production
- Using the NiFi variable registry
- Deploying your data pipelines

Building a Production Data Pipeline

- Creating a test and production environment
- Building a production data pipeline
- Deploying a data pipeline in production

Beyond Batch – Building Real-Time Data Pipelines

- Beyond Batch – Building Real-Time Data Pipelines

Streaming Data with Apache Kafka

- Understanding logs
- Understanding how Kafka uses logs
- Building data pipelines with Kafka and NiFi
- Differentiating stream processing from batch processing
- Producing and consuming with Python

Data Processing with Apache Spark

- Installing and running Spark
- Installing and configuring PySpark
- Processing data with PySpark

DevOps

DevOps: (Git, Jenkins, Docker and Kubernetes)(Spark, Kafka, Airflow, Hadoop pipeline using Docker and Kubernetes, Helm chart, Terraform)

Git

- GIT Features
- 3-Tree Architecture

- GIT – Clone /Commit / Push
- GIT revert and reset
- GIT Branching strategies
- GIT Rebase & Merge
- GIT Stash, Reset, Checkout
- GIT Clone, Fetch, Pull

Jenkins

- Introduction to Jenkins
- Continuous Integration with Jenkins
- Configure Jenkins
- Jenkins Management
- Scheduling build Jobs
- POLL SCM
- Build Periodically
- Maven Build Scripts
- Support for the GIT version control System
- Different types of Jenkins Jobs
- Jenkins Build PipeLine
- Parent and Child Builds
- Sequential Builds
- Jenkins Master & Slave Node Configuration

Docker

- How to get Docker Image?
- What is Docker Image
- Docker Installation
- Working with Docker Containers
- What is Container
- Docker Engine
- Crating Containers with an Image
- Working with Images
- Docker Command Line Interphase
- Docker Compose
- Docker Hub
- Docker Trusted Registry
- Docker swarm
- Docker attach
- Docker File & Commands

- Docker containers for kafka ,spark,cassandra etl pipeline

Kubernetes

- Kubernetes Introduction
- Kubernetes Architecture
- Kubernetes Setup (Self Managed,AWS managed)
- Kubernetes Pods
- Kubernetes Services
- Kubernetes Namespaces
- Replication Controller & ReplicaSet
- Kubernetes Deployments
- Kubernetes ConfigMap
- Kubernetes Secrets
- HELM Charts
- EKS Cluster
- Monitoring