TRENDY TECH
UPLIFT YOUR CAREER!

Newly
Launched
Premium
Program

The Ultimate Big Data Masters Program
(Cloud Focused)

By

Sumit Mittal

CURRICULUM

Module 1 - Big Data Fundamentals

- Introduction to Big Data
- How Big Data Works
- Practice Environment
- Brief Introduction to Distributed System Architecture
- Introduction to Hadoop and its Ecosystem Tools
- Basics of Distributed Storage - HDFS Architecture
- Linux Commands
- HDFS Commands & How it Works
- Introduction to Data Lake Storage - Blob & ADLS Gen 2
- Big Data - The Big Picture with Real-Time Example

Module 2 - Distributed Processing with Pyspark

- Distributed Processing Fundamentals
- Knowing Apache Spark
- Spark Development Environments - OnPremise | OnCloud
- Understanding Spark Cluster & Cluster Modes
- Apache Spark In-Depth with Real-Time Example
- How Spark Executes Program on the Cluster
- Stages in Spark
- Understanding Spark Transformations & Actions

    Lazy Evaluation
    Narrow Vs Wide Transformations

- Accumulators & Broadcast Variables
- Repartition Vs Coalesce
- Data Caching

    Spark Storage Levels
    Cache Vs Persist

- Spark Optimization Techniques In-depth
- Internals of File Formats - Parquet | ORC | Avro
- Compression Techniques
- Introduction to Spark Data Frames

  Creating Spark Data frame

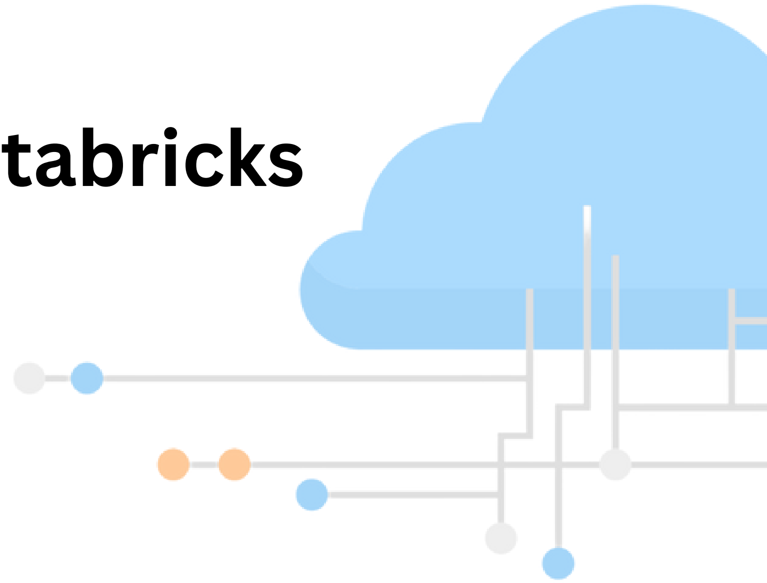
  Data frame Transformations and Actions

  Querying Spark Data frame More Data frame Transformations
- Introduction to SparkSQL
- Understanding Cluster Configurations
- How to Submit Spark Job
- Scheduling and Running Spark Jobs
- Spark Advance Optimizations - Sort Vs Hash Aggregate
- Spark Catalyst Optimizer
- Learning Hive
- Spark-Hive Integration
- Implement Your First Batch Processing Project with Pyspark

Module 3 - Azure Databricks

- Introducing Azure Databricks
- Microsoft Azure Services and Portal Overview
- What is Databricks & Why Databricks Databricks
- Pricing - Infrastructure and Software Charges
- Different Cloud Providers offering Databricks
- Databricks Features
- Databricks Community Edition
- 3 ways to Create Cluster
- All Purpose Cluster
- Job Cluster
- Cluster Pool
- When to use the Different Cluster Modes
- Databricks Benefits
- Different optimized Cluster types - Memory Optimized, Storage Optimized, Compute Optimized, General Purpose, GPU Accelerated

- Databricks File System (DBFS)
- Databricks Architecture - Control and Data Plane
- DBFS in detail
- Object Store - Blob, Datalake Gen2
- Filesystem utility- dbutils
- Data Utility & Notebook Utility & Widgets Utility
- Parameter passing from one Notebook to another
- Mount Point - How to create Mount Point
- Databricks Workspace
- Databricks CLI
- Ways to access Storage Account
- Access Key | Account Key
- SAS Key & Service Principal
- Secret Scope - Azure Key vault Backed Secret Scope & Databricks Backed Secret Scope



- Delta Lake
- Delta Table Creation
- Lakehouse Architecture
- Azure Delta Engine Optimizations
- Delta Architecture - Medallian Architecture
- Cluster Creation
- Autoloader
- Delta Live Table
- Unity Catalog

Module 4 - Azure DataFactory

- Azure Data Factory Introduction
- Data Transfer (Source to Sink)
- Data Transformation - Data Flow
- Workflow Orchestration
- Data Transfer from RDBMS to ADLS Gen2
- Azure SQL Databses
- Data Transfer from Azure SQL to ADLS Gen2
- Author, Monitor & Manage
- Data Integration Service (ADF)
- Usecases where ADF can be used
- Data Ingestion
- Data Transformation
- Data Orchestration
- Data Flow Mapping
- Data transfer from external URL to ADLS - Usecase

- **Linked Services for Source and Sink**
- **Select Transformation**
- **ADF Primary Usage**
- **Tansfer data from Blob to Datalake - Usecase**
- **Blob Connector**
- **Http Connector**
- **Datalake Instance**
- **Data Factory Instance**
- **Linked Service Creation - Blob & Datalake**
- **Dataset for Blob and Datalake**
- **Complete Pipeline setup**
- **Key Vault & Scheduled Triggers**
- **Tumbling Window Triggers**
- **Storage & Custom Events**
- **Trigger Pipeline on Custom Event - Usecase**
- **Data Ingestion from 2 Sources (Blob & Amazon S3) to ADLS Gen2**
- **Building a Complete Pipeline Using DataBricks & DataFactory**

Module 5 - Interview Readiness

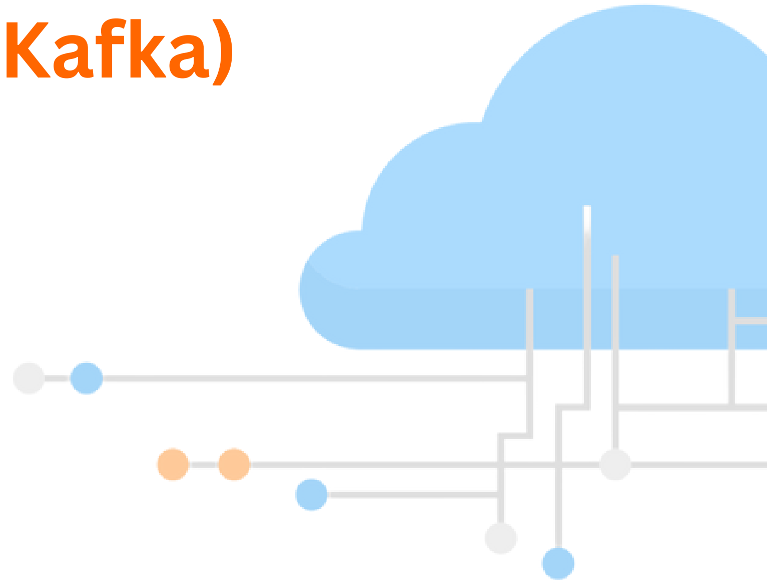
- Data Modeling

  Fact & Dimension Tables

  Data Models - Star Vs Snowflake
- System Design
- CICD - Git
- Interview Preparation Tips
- Interview Questions
- Guidance for Resume Preparation
- How to Handle Managerial Round Questions

Module 6 - Streaming

- Structured Streaming In-depth
- Benefits of Spark Structured Streaming
- Types of Data Sources
- Streaming Joins
- Streaming Dataframe
- Introduction to Kafka - Streaming Platform
- Kafka Architecture
- Installing Multi-Node Kafka Cluster
- Writing Kafka Producer and Consumer
- Scaling up the Kafka Cluster
- Integrating Kafka with Spark Structured Streaming
- Building Streaming Pipeline (Structured Streaming with Kafka)



Module 7 - More on Cloud Services

- Azure Synapse
- Azure CosmosDB
- Azure HDInsights
- Azure Logic App
- Azure Event Hub

- AWS EMR (Elastic MapReduce)
- Launch EMR Cluster Using Advanced Options
- Types of EC2 Instances
- AWS S3
- AWS Athena
- AWS Glue - Data Catalog | Crawlers
- AWS Redshift
- End-to-End Real-time Project on Cloud using other cloud services

Additional Modules

(To Crack Top Product Based Companies)

- Data Structures and Algorithms for Data Engineers - 20 hours
- Python for Data Engineers - 10 hours

TRENDY TECH
UPLIFT YOUR CAREER!

Contact
hello@trendytech.in

https://trendytech.in/