Upgrade ur Career

Seekho Bigdata Institute

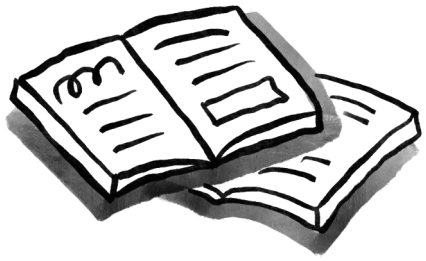
SB
Seekho Bigdata
Data is the New Oil

Big data: where information meets opportunity

1. Hadoop & HDFS
2. MAPREDUE & SQOOP
3. Shell Scripting
4. MySQL
5. Hive
6. Scala
7. Python
8. Scala Spark & Pyspark
9. DataStructures & Algorithms
10. AWS
11. Projects
12. Interview Preparation

https: //www.seekhobigdata.com
+91-9989454737

SYLLABUS HANDOUT

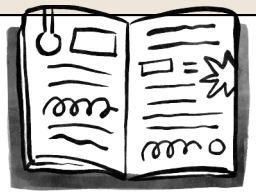


Module -1:- Introduction Bigdata

- [ ] What is Bigdata?
- [ ] What are 5 V's of Bigdata?
- [ ] What are Monolithic Systems?
- [ ] What are Distributed Systems?
- [ ] Evolution of Hadoop How Hadoop has been accepted as solution for bigdata ?

HDFS

- [ ] What is HDFS?
- [ ] What is a Block and Block Size?
- [ ] Fixed Size vs. Variable Size Block
- [ ] How to Tweak Block Size
- [ ] Using Fixed and Variable Block Size
- [ ] What is Meant by Node?
- [ ] What is a Cluster?
- [ ] What is Resource?
- [ ] Namenode
- [ ] Datanode
- [ ] HDFS Architecture
- [ ] Datanode Failure Management



- [ ] Reasons For DataNode Failure
- [ ] Heart-Beat Mechanism
- [ ] Replication Mechanism
- [ ] Properties to Check: Heartbeat Mechanism and Replication
- [ ] NameNode Failure Management
- [ ] Reasons for NameNode Failure
- [ ] What is Zookeeper?
- [ ] Secondary NameNode
- [ ] FS Image and Edit Logs
- [ ] Checkpointing
- [ ] Fault Tolerance
- [ ] What is a Split-Brain Scenario?
- [ ] What is Fencing?
- [ ] Command to Kill SNN
- [ ] Property for Checkpointing
- [ ] Checkpointing Frequency

SECURİTY FOR HDFS

SECURİTY FOR DATA STORED İN HDFS

- [ ] What is the Kerberos Authentication Mechanism in HDFS?
- [ ] What is keyTab in HDFS?
- [ ] What is the Knit Command?
- [ ] What is Pbrun?

CORRUPTED RECORDS İN HDFS
HOW CORRUPTED RECORDS ARE KNOWN

- [ ] How Data Can Be Corrupted
- [ ] How to Identify Corrupted Records in HDFS
- [ ] What is Checksum
- [ ] What is the FSCK Command
- [ ] What is the Resolution Technique for Data Corruption in HDFS

LİNUX COMMANDS

- [ ] Ls Command and its Flavours
- [ ] Cat Command and its Flavours
- [ ] mkdir command and its flavors
- [ ] rm command and its flavors
- [ ] mv command and its flavors
- [ ] cp command and its flavors
- [ ] tail command
- [ ] grep command and its flavors
- [ ] find command and its flavors

Vİ EDİTORCOMMANDS

- [ ] View a filename: vi
- [ ] Insert mode: i
- [ ] Save when updated: :w
- [ ] Quit: :q!
- [ ] Update, save, and quit: :wq!

FİLE PERMİSSİON COMMAND

- [ ] What are read permission
- [ ] What is Write Permission
- [ ] What is Execute Permission owner && group and others
- [ ] **HDFS Commands && Practicals**
- [ ] **Scenario Based Questions**

MODULE-2 :-MR AND SQOOP

- [ ] What is MapReduce?
- [ ] What are Mappers?
- [ ] What is Reducer?
- [ ] How MapReduce Works
- [ ] What is RecordReader?
- [ ] What is Combiner?
- [ ] What is Shuffling?
- [ ] What is Sorting?
- [ ] Limitations of MapReduce

SQOOP

- [ ] What is Sqoop?
- [ ] What is JDBC Driver in Sqoop?
- [ ] List-databases Argument
- [ ] List-Tables Argument
- [ ] Import Argument

- [ ] Importing Selected Columns
- [ ] Increasing Parallelism
- [ ] --num-Mappers
- [ ] Split-by Argument
- [ ] Null-Non-String
- [ ] Null-String Export Argument
- [ ] Staging in Sqoop
- [ ] Query Argument
- [ ] How to Decide Mappers
- [ ] How to Include Tables
- [ ] How to Exclude Tables
- [ ] Outliers
- [ ] Incremental Loading in Sqoop
- [ ] Append Mode
- [ ] Last Modified Mode
- [ ] Direct Argument
- [ ] Scenario-Based Questions

MODULE-3:- SHELL SCRİPTİNG

- [ ] Echo
- [ ] readonly
- [ ] arrays
- [ ] string
- [ ] Arithmetic Operations
- [ ] Loops
- [ ] For Loop
- [ ] While Loop
- [ ] Functions
- [ ] Building Scripts
- [ ] Building Sqoop Scripts
- [ ] Hive Scripts
- [ ] Cron Tab
- [ ] Orchestrating Scripts
- [ ] AWK
- [ ] SED
- [ ] Practical Lab: Building 50+ Scripts

MODULE-4:-MYSQL

- [ ] CREATE
- [ ] ALTER
- [ ] DELETE
- [ ] DROP
- [ ] TRUNCATE
- [ ] GROUP BY
- [ ] Having

- [ ] CASE and WHEN
- [ ] Window Functions
- [ ] RANK/DENSE RANK/ROW NUMBER
- [ ] Null Handling
- [ ] Date Functions
- [ ] Joins (2, 3, 4 Table Joins)
- [ ] Self Join
- [ ] MIN
- [ ] MAX
- [ ] AVG
- [ ] COUNT
- [ ] SUM
- [ ] String Functions
- [ ] CTE Expressions
- [ ] **Lab Practical: 100 Assignment Questions**

✦ MODULE -5:- HİVE

- [ ] What is a Data Warehouse?
- [ ] What is the difference between OLTP and OLAP?
- [ ] What is Hive?
- [ ] How is schema and data stored in Hive?
- [ ] Internal/Managed Tables
- [ ] External Tables
- [ ] Primary Data Types in Hive
- [ ] Complex Data Types in Hive
- [ ] Array and Its Functions
- [ ] Map and Its Functions
- [ ] Struct and Its Functions

- [ ] Different Ways of Inserting Data
- [ ] Vectorization
- [ ] MSCK Repair
- [ ] Views

OPTİMİZATİON TECHNİQUES İN HİVE

- [ ] Partitioning in Hive
- [ ] Static vs. Dynamic Partitioning
- [ ] Bucketing
- [ ] How to Calculate Number of Buckets
- [ ] Map-Side Join
- [ ] Bucket Map Join
- [ ] Sort-Merge Bucket Join
- [ ] Window Functions
- [ ] RANK()
- [ ] Dense Rank() and Row Number()
- [ ] Lead and Lag
- [ ] Problems on Lead and Lag and SCD Types
- [ ] Compression Techniques
- [ ] Miscellaneous Concepts
- [ ] Scenario-Based Questions
- [ ] Practicals on Every Topic

MODULE-6 SCALA

- [ ] Data Types in Scala
- [ ] Conditional Statements in Scala
- [ ] Loops in Scala
- [ ] Data Structures in Scala
- [ ] Arrays
- [ ] Map
- [ ] Set
- [ ] Range
- [ ] List
- [ ] Tuple

FUNCTİONAL PROGRAMMİNG

- [ ] First-Class Functions
- [ ] Higher-Order Functions
- [ ] Anonymous Functions
- [ ] Closures
- [ ] Null, Nil, Nothing, None, Unit
- [ ] 50+ Practical Problems

SCALA-OOPS

- [ ] Class and Object
- [ ] Constructor
- [ ] Polymorphism
- [ ] Encapsulation
- [ ] Abstract Class

- [ ] Access Modifiers
- [ ] Design Patterns in Scala
- [ ] Traits
- [ ] Diamond Problem
- [ ] Case Classes
- [ ] Method Overloading & Overriding
- [ ] Singleton Object
- [ ] Companion Classes
- [ ] 40-50 Practical Problems

MODULE-7 PYTHON

- [ ] Variables and Data Types
- [ ] Control Structures and Loops
- [ ] Operators
- [ ] Exception Handling
- [ ] Python Built-in Functions
- [ ] Lists
- [ ] Tuples
- [ ] Sets
- [ ] Dictionaries
- [ ] Classes
- [ ] Objects
- [ ] Inheritance
- [ ] Encapsulation
- [ ] Polymorphism
- [ ] Opening Files

- [ ] Prime Number
- [ ] Reverse a Number
- [ ] Palindrome
- [ ] Square Root of a Number
- [ ] Divisibility Rules
- [ ] Missing Number
- [ ] Reading Files
- [ ] Writing Files
- [ ] Closing Files
- [ ] Exception Handling
- [ ] NumPy
- [ ] Pandas

MODULE-8 PY-SPARK && SCALA SPARK

- [ ] What Is Apache Spark?
- [ ] What Is RDD?
- [ ] MapReduce Vs Apache Spark
- [ ] How Data Is Stored In Spark
- [ ] What Is Immutability Of RDD?
- [ ] What Is Resilient Distributed Dataset (RDD)?
- [ ] Spark Session
- [ ] Spark Context
- [ ] Parallelize()
- [ ] Read CSV, TextFile
- [ ] Lazy Evaluation
- [ ] What is DAG?
- [ ] What is Lineage Graph?
- [ ] Transformations

- [ ] Fault Tolerance
- [ ] Lineage
- [ ] map()
- [ ] filter()
- [ ] reduce vs reduceByKey
- [ ] groupByKey vs reduceByKey
- [ ] repartition & coalesce
- [ ] sortByKey()
- [ ] flatMap()
- [ ] split
- [ ] mean()
- [ ] Joins in RDD
- [ ] contains()
- [ ] parallelize
- [ ] Spark Architecture
- [ ] Broadcast Variables
- [ ] Accumulators
- [ ] Problems on RDD
- [ ] 60-70 Practical Problems

DATAFRAMES

- [ ] DataFrames
- [ ] Datasets
- [ ] DataFrame vs Dataset
- [ ] Reader API
- [ ] Read Modes
- [ ] Writer API
- [ ] Write Modes
- [ ] Infer Schema

- [ ] Explicit Schema
- [ ] Data Types in Spark
- [ ] Conditional Statements in Spark
- [ ] When and Otherwise
- [ ] Filter
- [ ] String Manipulation Functions
- [ ] Aggregations
    - Count()
    - Min()
    - Avg()
    - Sum
- [ ] GroupBy Aggregations
- [ ] Window Aggregations
- [ ] Joins
    - Different Kinds of Joins
    - Different Join Stratagies
- [ ] Log4j Mechanism
- [ ] Different Ways of Debugging
- [ ] Lead and Lag Related Problems
- [ ] Spark-SQL Date Manipulation Functions
- [ ] Practicals on Every Concept
- [ ] Benchmarking to understand Performance
- [ ] String Manipulation Functions
- [ ] Number Manipulation Functions
- [ ] Data Validation
- [ ] 400+ Wide Variety of Problems

OPTİMİZATİONS

- [ ] Serialization API Selection
- [ ] Using Broadcast Variables
- [ ] Cache and Persist
- [ ] ByKey Operation
- [ ] Predicate Pushdown
- [ ] Broadcast Join
- [ ] Partition and Bucket
- [ ] Garbage Collection Tuning
- [ ] Level of Parallelism

SPARK-İSSUES

- [ ] Out Of Memory
- [ ] Exceptions Missing
- [ ] Data Data Skewness
- [ ] Spark Job Repeatedly Fails
- [ ] Inferschema Issue
- [ ] Slow Performance Issues
- [ ] Memory Contention
- [ ] Disk Contention
- [ ] Broadcasting Large Data
- [ ] Serialization Issue
- [ ] Version Incompatibility Issue
- [ ] Cluster Instability Issues
- [ ] Small File Issue
- [ ] Result Exceeds Driver Memory
- [ ] Too Small And Large Partitions

SPARK- DEPLOYEMENT

- [ ] Build Tools
- [ ] SBT Build Tool
- [ ] Gradle Build Tool
- [ ] Maven Build Tool
- [ ] JFrog
- [ ] JIRA Tool
- [ ] Bitbucket
- [ ] GitHub
- [ ] Git Commands
- [ ] How to build a Jar
- [ ] Spark-submit
- [ ] Parameters of Spark-submit

DATA- QUALİTYCHECKS

- [ ] Check for duplicates
- [ ] Check for unique values in columns
- [ ] Check for missing values
- [ ] Find outliers
- [ ] Schema validation
- [ ] Correlations
- [ ] Cross-field validation
- [ ] Dependency check
- [ ] Text pattern analysis
- [ ] Categorical value distributions

MODULE -9 AWS

AWS S3 BASİCS

- [ ] What is Aws ?
- [ ] Aws GUI Walkthrough ?
- [ ] What is Region?
- [ ] What is edge Location ?
- [ ] What is Availability Zone and Local Zone?
- [ ] What is Multiregion Concept ?
- [ ] What are Global and Region specific services in Aws ?

AWS STORAGE

- [ ] What is S3 and how is data stored?
- [ ] The Shared Responsibility Model and Security
- [ ] Storage Tiers and Pricing
- [ ] Getting Data Into and Out of S3
- [ ] Create and Secure your AWS Account
- [ ] Upload Files to Buckets Using the AWS Console
- [ ] Move, Copy, Download, and Delete Files
- [ ] Classifying Your Buckets and Objects with Tags
- [ ] Lifecycle Management
- [ ] Retrieving Objects from Glacier
- [ ] Making Buckets or Objects Public with ACLs
- [ ] Using a Bucket Policy to Grant Public Access
- [ ] Using a Bucket Policy to Grant Access to Objects in a Bucket
- [ ] Using a Bucket Policy to Restrict Access Based on an Object Tag
- [ ] How to Enable Versioning and Encryption
- [ ] How to Set Up Cross Region Replication for Further Redundancy

AWS IAM

1. Introduction to AWS IAM

- [ ] What is AWS IAM?
- [ ] Core concepts (users, groups, roles, policies)
- [ ] Benefits of using IAM for access management

2. IAM Users

- [ ] Creating and managing IAM users
- [ ] User credentials (passwords, access keys)
- [ ] Best practices for managing user permissions

3.IAM Groups

- [ ] Creating and managing IAM groups
- [ ] Adding users to groups
- [ ] Group-level permissions and policies

4. IAM Roles

- [ ] What are IAM roles?
- [ ] Creating and managing roles
- [ ] Role-based access control (RBAC)
- [ ] Using roles with AWS services (e.g., EC2, Lambda)

5. IAM Policies

- [ ] What are IAM policies?
- [ ] AWS-managed policies vs. customer-managed policies
- [ ] Policy structure (JSON)
- [ ] Writing and attaching policies
- [ ] Policy evaluation logic

6. AWS-Managed Policies

- [ ] Overview of AWS-managed policies Common
- [ ] AWS-managed policies (e.g., AdministratorAccess, ReadOnlyAccess)
- [ ] Best practices for using AWS-managed policies
- [ ] Customizing AWS-managed policies

7. Custom Policies

- [ ] Creating custom IAM policies Policy elements (actions, resources, conditions)
- [ ] Using policy variables
- [ ] Testing and debugging custom policies
- [ ] Best practices for custom policies

AWS LAMBDA

1. Introduction to AWS Lambda

- [ ] What is AWS Lambda?
- [ ] Use cases for serverless computing
- [ ] Benefits of using AWS Lambda

2. Lambda Functions

- [ ] Creating Lambda functions
- [ ] Lambda function lifecycle
- [ ] Writing your first Lambda function
- [ ] Lambda function configuration (memory, timeout, environment variables)

3. Event Sources

- [ ] Understanding event-driven architecture Integrating
- [ ] Lambda with various AWS services (e.g., S3, DynamoDB, SNS, SQS, API Gateway)
- [ ] Custom event sources

4. AWS Lambda Triggers

- [ ] Configuring triggers for Lambda functions Invoking
- [ ] Lambda functions synchronously and asynchronously
- [ ] Handling retries and error handling

5. Permissions and Security

- [ ] AWS Identity and Access Management (IAM) roles for Lambda
- [ ] Least privilege principle
- [ ] Resource-based policies
- [ ] VPC integration and security groups

6. Lambda Layers

- [ ] What are Lambda Layers?
- [ ] Creating and using Lambda Layers
- [ ] Best practices for using Layers

7. Environment Variables and Configuration

- [ ] Using environment variables in Lambda functions
- [ ] Secrets and sensitive data management with AWS Secrets Manager and AWS Systems Manager Parameter Store

8. Logging and Monitoring

- [ ] Logging with AWS CloudWatch Logs
- [ ] Metrics and monitoring with AWS CloudWatch
- [ ] Setting up CloudWatch Alarms
- [ ] Troubleshooting and debugging

9. Error Handling and Retries

- [ ] Built-in error handling mechanisms
- [ ] Retry strategies and dead-letter queues (DLQ)
- [ ] Custom error handling

10. Deployment and Versioning

- [ ] Deploying Lambda functions using the AWS Management Console, CLI, SDKs, and CI/CD pipelines
- [ ] Versioning and aliases
- [ ] Blue/Green deployments with aliases

11. Performance Optimization

- [ ] Best practices for optimizing Lambda function performance
- [ ] Cold starts and warm starts
- [ ] Provisioned concurrency

12. Cost Management

- [ ] Understanding Lambda pricing
- [ ] Cost optimization strategies
- [ ] Monitoring and controlling Lambda usage costs

AWS EMR

- [ ] EMR Basics
- [ ] Cluster Management
- [ ] Data Ingestion
- [ ] Data Transformation
- [ ] Data Loading
- [ ] Cluster Security
- [ ] Auto Scaling
- [ ] Cluster Optimization
- [ ] Spot Instances
- [ ] EMR Best Practices
- [ ] Data Workflow Automation
- [ ] Integration with Other AWS Services

AWS REDHSİFT

- [ ] Data Warehousing Concepts
- [ ] Amazon Redshift Architecture
- [ ] Cluster Management
- [ ] Data Loading
- [ ] COPY Command
- [ ] INSERT statement
- [ ] AWS DMS
- [ ] Data Distribution: KEY, EVEN, and ALL
- [ ] Query Optimization
- [ ] Data Encryption
- [ ] Data Compression

AWS DYNAMODB

1. Introduction to Amazon DynamoDB

- [ ] What is DynamoDB?
- [ ] Use cases for DynamoDB
- [ ] Key features and benefits

2. Core Concepts

- [ ] Tables, items, and attributes
- [ ] Primary keys (partition keys and sort keys)
- [ ] Secondary indexes (Global Secondary Indexes (GSI) and Local Secondary Indexes (LSI))
- [ ] DynamoDB Streams

3. Setting Up DynamoDB

- [ ] Creating and managing tables
- [ ] Understanding capacity modes (Provisioned vs. On-Demand)
- [ ] Defining primary keys and indexes
- [ ] Setting up DynamoDB Streams

4. Data Modeling

- [ ] Designing efficient table schemas
- [ ] Understanding single-table design
- [ ] Using primary keys and secondary indexes effectively
- [ ] Best practices for data modeling in DynamoDB

5. Reading and Writing Data

- [ ] Using the DynamoDB API (PutItem, GetItem, UpdateItem, DeleteItem)
- [ ] Batch operations (BatchGetItem, BatchWriteItem)
- [ ] Querying and scanning tables
- [ ] Understanding conditional operations

6. Indexes

- [ ] Creating and managing Global Secondary Indexes (GSI)
- [ ] Creating and managing Local Secondary Indexes (LSI)
- [ ] Use cases and best practices for using indexes Querying data with indexes

AWS GLUE

- [ ] What is Glue ?
- [ ] Data Catalog
- [ ] Crawlers Data Lake and Data Warehouse Integration
- [ ] Data Brew Transformations Job Authoring & Development
- [ ] Data Source Connectors
- [ ] Target Connectors
- [ ] Serverless Execution
- [ ] Monitoring and Logging
- [ ] Security and Data Encryption
- [ ] Error Handling and Retry Mechanisms
- [ ] Data Quality and Validation
- [ ] Data Versioning

MODULE -10 DATA STRUCTURES AND ALGORİTHMS

- [ ] Datatypes
- [ ] Operators
- [ ] Conditional Statements
- [ ] Loops
- [ ] Arrays
- [ ] Strings
- [ ] Searching
- [ ] Linear Search

- [ ] Binary Search
- [ ] Sorting
- [ ] Bubble Sort
- [ ] Merge Sort
- [ ] Linked List
- [ ] Single Linked List
- [ ] Double Linked List
- [ ] Stack
- [ ] Queue
- [ ] Time Complexity
- [ ] Space Complexity

MODULE -11 PROJECTS

- [ ] Health-Care Sector Use Case
- [ ] Retail Use Case Project
- [ ] Financial Use Case Project
- [ ] LinkedIn Optimizations
- [ ] Naukri Optimization
- [ ] Resume Review Session
- [ ] Interview Preparation

MODULE-12 :INTERVİEW PREPARATİON

- [ ] Naukri Optimization
- [ ] Resume Preparation
- [ ] Linkedin Optimization
- [ ] Hive Revision with important Questions
- [ ] Spark Revision with important Questions
- [ ] Sql Revision with Important Questions
- [ ] Coding Revision with Important Questions
- [ ] 10 + coding tests
- [ ] Mcq Exams (15 exams)
- [ ] Grooming Sessions
- [ ] Physical Wriiten exams on every topic