

Big Data Analytics & Cloud Data Pipeline Internship

(15 Days | Practical-Oriented | Industry Ready)

INTERNSHIP SYLLABUS

Internship Objectives

- Understand **Big Data** concepts & ecosystem
- Build **end-to-end** cloud data pipelines
- Work with **real datasets**
- Gain confidence to do **independent final projects**
- Industry exposure using **AWS / GCP / Azure concepts**

Tools & Technologies

- Hadoop, HDFS, MapReduce
- Apache Spark, PySpark
- Kafka (Basics)
- SQL, NoSQL (MongoDB)
- AWS (S3, EC2, Glue – concepts)
- Python for Data Engineering
- Power BI / Tableau
- Git & GitHub

DAY 1 – Introduction to Big Data

- What is Big Data?

- 5 V's of Big Data
- Traditional vs Big Data systems
- Use cases (Banking, Healthcare, E-Commerce)

Practical

Install Hadoop/Spark environment (local/cloud)

Dataset overview

Outcome

Understand why Big Data is needed

DAY 2 – Big Data Ecosystem

- Hadoop Architecture
- HDFS, YARN
- Data Lake vs Data Warehouse

Practical

Explore HDFS commands

Upload & retrieve data from HDFS

Outcome

Hands-on with distributed storage

DAY 3 – HDFS & Data Storage

- HDFS blocks & replication
- Structured vs unstructured data

Practical

Store CSV, JSON, log files in HDFS

Analyze file distribution

Outcome

Real data handling experience

DAY 4 – MapReduce Fundamentals

- Map, Shuffle, Reduce
- MapReduce workflow

Practical

Word Count program

Log file analysis using MapReduce

Outcome

Understand batch processing

DAY 5 – Introduction to Apache Spark

- Spark vs Hadoop
- Spark Architecture
- RDD, DataFrames

Practical

Spark installation

First Spark job

Outcome

Faster data processing knowledge

DAY 6 – PySpark for Data Processing

- PySpark transformations & actions
- Lazy evaluation

Practical

Data cleaning using PySpark

Filtering, aggregation

Outcome

Data preprocessing skills

DAY 7 – Big Data Analytics with Spark SQL

- Spark SQL
- DataFrames & Queries

Practical

Analyze sales dataset

Generate insights using SQL queries

Outcome

Analytical thinking using big data

DAY 8 – NoSQL & MongoDB

- SQL vs NoSQL

- MongoDB basics
- Schema-less databases

Practical

Store processed data in MongoDB

Query using MongoDB Compass

Outcome

Hybrid data storage experience

DAY 9 – Streaming Data & Kafka (Basics)

- Batch vs Streaming
- Kafka architecture
- Real-time data examples

Practical

Simulate streaming data

Process stream using Spark

Outcome

Exposure to real-time analytics

DAY 10 – Cloud Computing for Big Data

- Cloud basics
- AWS, Azure, GCP overview
- Cloud storage & compute

Practical

AWS S3 data upload (demo)

EC2 Spark environment overview

Outcome

Cloud-ready skills

DAY 11 – Cloud Data Pipeline Architecture

- ETL vs ELT

- Data pipeline stages
- Orchestration basics

Practical

Build simple ETL pipeline

Source → Process → Store → Analyze

Outcome

End-to-end pipeline clarity

DAY 12 – Data Visualization

- Why visualization matters
- KPIs & dashboards

Practical

Power BI / Tableau dashboard

Big data insights visualization

Outcome

Business-level reporting

DAY 13 – Mini Project Development

- Cloud Sales Analytics Pipeline
- IoT Sensor Data Analysis
- Log File Analytics System

Practical

Dataset selection

Pipeline design

Outcome

Project confidence building

DAY 14 – Final Project Implementation

Practical

- Complete data ingestion
- Processing & analytics
- Visualization dashboard

Outcome

Real-world project completion

DAY 15 – Project Presentation & Career Guidance

Activities

- Project demo
- Resume & interview questions
- Big Data career roadmap

Outcome

Industry readiness

