Datastat Research
Building Capacity in Training and Research

DATASTAT TRAINING IN

OFFICIAL COURSE PROSPECTUS

Center for Advanced Research & Professional Development

Nairobi Training Center • Global Virtual Academy • Corporate On-Site

DATA SCIENCE

Big Data Storage and Distributed Computing Training Course

Comprehensive Course Syllabus, Practical Modules & Upcoming Schedules

PROGRAM SPECIFICATIONS & HIGHLIGHTS

DURATION
5 Days Practical Hands-on

DELIVERY MODES
In-Person Class & Live Virtual

CERTIFICATION
Accredited Certificate of Completion

INTAKES / DATES
Multiple Cohort Schedules Included

PREPARED FOR:

Valued Participant

Email: training@datastatresearch.org | Issued: November 25, 2026

Academic & Professional Training Year: 2026

DATASTAT TRAINING INSTITUTE
P.O. Box 102802-00101, Nairobi, Kenya
Training Centers: Nairobi • Mombasa • Kigali • Dubai • Live Virtual Academy
Tel: +254 724 527 104 | +254 734 969 612
Email: training@datastatresearch.org
www.datastatresearch.org

Big Data Storage and Distributed Computing Training Course

Data Science • 5 Days Intensive • Accredited Certification

Course Overview

The Big Data Storage and Distributed Computing Training Course is an advanced professional programme designed to equip participants with practical and strategic skills for managing, storing, processing, and analysing massive datasets across distributed computing environments.

Key Skills & Software Tools Covered

Apache Hadoop • Apache Spark • Apache Kafka • MongoDB • PostgreSQL

Upcoming Training Schedules

Confirmed cohort intakes and venues. Both physical in-person classroom and live virtual online seats are available.

TRAINING DATES	LOCATION / VENUE	FORMAT / DURATION	INVESTMENT (FEE)
30 Nov 2026 – 4 Dec 2026	Nairobi	Classroom / 5d	**$1,500 USD**
7 Dec 2026 – 11 Dec 2026	Nairobi	Classroom / 5d	**$1,500 USD**
14 Dec 2026 – 18 Dec 2026	Nairobi	Classroom / 5d	**$1,500 USD**
21 Dec 2026 – 25 Dec 2026	Nairobi	Classroom / 5d	**$1,500 USD**

* Corporate discounts and customized cohort scheduling are available for teams of 3 or more participants.

Course Curriculum & Modules

Big Data Storage and Distributed Computing Training Course

Course Introduction

The Big Data Storage and Distributed Computing Training Course is an advanced professional programme designed to equip participants with practical and strategic skills for managing, storing, processing, and analysing massive datasets across distributed computing environments. The course provides comprehensive knowledge of Big Data storage architectures, distributed systems, cloud storage, data lakes, data warehouses, Hadoop Distributed File System (HDFS), distributed databases, parallel processing, cluster computing, and scalable data infrastructure. Participants will learn how modern organizations design highly available, fault-tolerant, secure, and scalable storage and computing environments capable of handling the volume, velocity, and variety of contemporary enterprise data.

The training combines theoretical concepts with practical enterprise applications and real-world Big Data storage and distributed computing case studies. Participants will explore data partitioning, replication, distributed file systems, resource management, parallel computing, batch and stream processing, cloud-native storage, performance optimization, disaster recovery, and distributed database management. Particular emphasis is placed on building cost-effective and resilient Big Data infrastructure that supports advanced analytics, artificial intelligence, machine learning, Internet of Things (IoT), business intelligence, and digital transformation. By completing the course, participants will be able to evaluate storage and computing requirements and design appropriate distributed architectures aligned with organizational performance, scalability, security, and business objectives.

Learning Objectives

By the end of the Big Data Storage and Distributed Computing Training Course, participants will be able to:

- **Explain the fundamental concepts and principles of Big Data storage and distributed computing.**
- **Design scalable storage architectures for large volumes of structured, semi-structured, and unstructured data.**
- **Configure and manage distributed file systems such as HDFS for enterprise Big Data environments.**
- **Apply data partitioning, replication, sharding, and redundancy techniques to improve system reliability.**
- **Design distributed computing architectures using parallel and cluster-based processing.**
- **Compare relational, NoSQL, distributed database, data lake, and data warehouse technologies.**
- **Implement efficient batch and real-time data processing architectures.**
- **Evaluate cloud-based storage and distributed computing solutions for enterprise applications.**
- **Apply performance optimization, fault tolerance, backup, recovery, and disaster recovery techniques.**
- **Design an integrated Big Data storage and distributed computing architecture that supports analytics, AI, machine learning, and digital transformation.**

Target Audience

- **Big Data engineers and data engineers.**
- **Cloud computing and cloud infrastructure professionals.**
- **System and network administrators.**
- **Database administrators.**
- **Data architects and enterprise architects.**
- **Software developers and DevOps professionals.**
- **Data scientists and machine learning engineers.**
- **IT managers and digital transformation professionals.**
- **Business intelligence and analytics professionals.**
- **Researchers, consultants, and technical professionals involved in Big Data infrastructure, distributed systems, data management, and enterprise technology.**

Course Modules

Module 1: Fundamentals of Big Data Storage and Distributed Computing

This module introduces the principles underlying modern Big Data storage systems and distributed computing architectures. Participants examine why traditional centralized storage and computing approaches often struggle with massive datasets and learn how distributed architectures provide scalability, availability, reliability, and processing efficiency.

Key topics include:

- Big Data characteristics and infrastructure requirements
- Distributed computing principles
- Centralized versus distributed architectures
- Scalability and elasticity
- Horizontal versus vertical scaling
- Distributed storage concepts
- Fault tolerance and high availability
- CAP theorem and distributed systems
- Enterprise Big Data infrastructure

Case Study: Global E-Commerce Data Platform — examining how a large online retailer can distribute customer, transaction, product, and website activity data across multiple computing and storage nodes to support millions of users.

Module 2: Distributed File Systems and Hadoop Distributed File System

This module provides an in-depth examination of distributed file systems, with particular emphasis on the Hadoop Distributed File System (HDFS). Participants learn how large datasets are divided, replicated, stored, and accessed across clusters of commodity servers.

Key topics include:

- Distributed file system architecture
- HDFS architecture
- NameNode and DataNode
- Data blocks and block placement
- Data replication
- Rack awareness
- HDFS permissions and quotas
- HDFS snapshots
- Data balancing and cluster management
- HDFS performance optimization

Case Study: Media and Entertainment Big Data Storage — designing an HDFS-based storage architecture for an organization managing millions of large video, audio, image, and multimedia files.

Module 3: Data Partitioning, Replication, Sharding and Fault Tolerance

This module explores techniques for distributing data across multiple servers while maintaining performance, reliability, and availability. Participants examine partitioning, sharding, replication, redundancy, failover, and recovery mechanisms used in distributed data environments.

Key topics include:

- Data partitioning strategies
- Horizontal and vertical partitioning
- Database sharding
- Consistent hashing
- Replication strategies
- Synchronous and asynchronous replication
- Data redundancy
- Fault tolerance
- Automatic failover
- Distributed recovery mechanisms

Case Study: Global Banking Transaction System — developing a distributed storage strategy that enables a financial institution to maintain reliable access to transaction data despite server and network failures.

Module 4: Distributed Databases and NoSQL Storage Technologies

This module examines modern distributed databases and NoSQL technologies designed for high-volume and high-velocity data. Participants compare document, key-value, column-family, and graph-based databases and determine when different technologies are appropriate.

Key topics include:

- Distributed database architecture
- Relational versus NoSQL databases
- Key-value databases
- Document databases
- Column-family databases
- Graph databases
- Data consistency models
- Distributed transactions
- Database replication
- NoSQL scalability and performance

Case Study: Real-Time Social Media Platform — designing a distributed NoSQL architecture capable of storing user profiles, social interactions, posts, comments, and high-volume engagement data.

Module 5: Distributed Computing, Parallel Processing and Cluster Management

This module focuses on how large computational workloads can be divided across multiple machines. Participants explore parallel processing, cluster computing, distributed execution, resource management, and workload optimization.

Key topics include:

- Parallel computing concepts
- Distributed processing
- Cluster architecture
- Batch processing
- Resource allocation
- Job scheduling
- MapReduce processing
- Apache Spark architecture
- Cluster performance management
- Distributed workload optimization

Case Study: Telecommunications Network Analytics — using distributed computing to process millions of call records and network events to identify service problems, usage patterns, and network optimization opportunities.

Module 6: Data Lakes, Data Warehouses and Cloud-Based Big Data Storage

This module examines modern enterprise approaches to storing analytical data, including data lakes, data warehouses, data lakehouses, and cloud storage platforms. Participants learn how to select storage architectures based on data types, analytical workloads, performance requirements, and cost considerations.

Key topics include:

- Enterprise data warehouse architecture
- Data lake architecture
- Data lakehouse architecture
- Object storage
- Cloud-based Big Data storage
- Data lifecycle management
- Storage tiering
- Data compression
- Data formats and partitioning
- Storage cost optimization

Case Study: Enterprise Data Lake Implementation — developing a centralized data lake that integrates customer, operational, financial, IoT, and external datasets to support enterprise analytics and machine learning.

Module 7: Real-Time Distributed Data Processing and Streaming

This module introduces real-time data processing and distributed streaming architectures for organizations that require immediate insights from continuously generated data. Participants explore streaming pipelines and technologies used to process high-velocity data.

Key topics include:

- Real-time data processing
- Batch versus stream processing
- Event-driven architectures
- Distributed messaging systems
- Apache Kafka fundamentals
- Stream processing
- Data ingestion pipelines
- Event processing and monitoring
- Real-time analytics
- Streaming performance optimization

Case Study: Real-Time Fraud Detection — designing a distributed streaming architecture that processes financial transactions as they occur and identifies suspicious transaction patterns with minimal latency.

Module 8: Performance Optimization, Security, High Availability and Disaster Recovery

The final module integrates the technical concepts covered throughout the programme and focuses on creating secure, resilient, high-performance distributed data environments. Participants learn how to monitor infrastructure, identify bottlenecks, optimize workloads, and develop disaster recovery strategies.

Key topics include:

- Distributed system performance monitoring
- CPU, memory, disk, and network optimization
- Storage performance optimization
- Data security and access control
- Encryption and authentication
- High-availability architectures
- Backup and recovery
- Disaster recovery planning
- Capacity planning
- Business continuity
- Distributed computing architecture optimization

Case Study: Mission-Critical Enterprise Data Platform — designing a highly available distributed Big Data infrastructure for an organization whose operations depend on continuous access to large-scale analytical and transactional datasets.

Required Software

Apache Hadoop, Apache Spark, Apache Kafka, MongoDB, PostgreSQL

Training Methodology

- Interactive instructor-led sessions
- Hands-on AI tool demonstrations
- Group-based leadership simulations
- Real-life case study discussions
- Personalized leadership development plans
- Post-training mentoring and AI coaching sessions

Register as a group from 3 participants for a Discount

Send us an email: info@datastatresearch.org or call +254724527104

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to DATASTAT CONSULTANCY LTD account, as indicated in the invoice so as to enable us prepare better for you.

Certification & Learning Methodology

- Hands-on practical delivery with real-world datasets, case studies, and lab exercises.
- Post-training mentorship, access to class resources, and customized follow-up support.
- Participants who successfully complete the course receive an accredited Datastat Certificate of Completion.

How to Enroll & Reserve Your Seat

1. Online Registration: Visit https://datastatresearch.org/courses/big-data-storage-and-distributed-computing-training-course to register online.
2. Email Booking: Send participant details to training@datastatresearch.org for official invoice & reservation.
3. Support & Assistance: Inquire via phone +254 724 527 104 | +254 734 969 612.