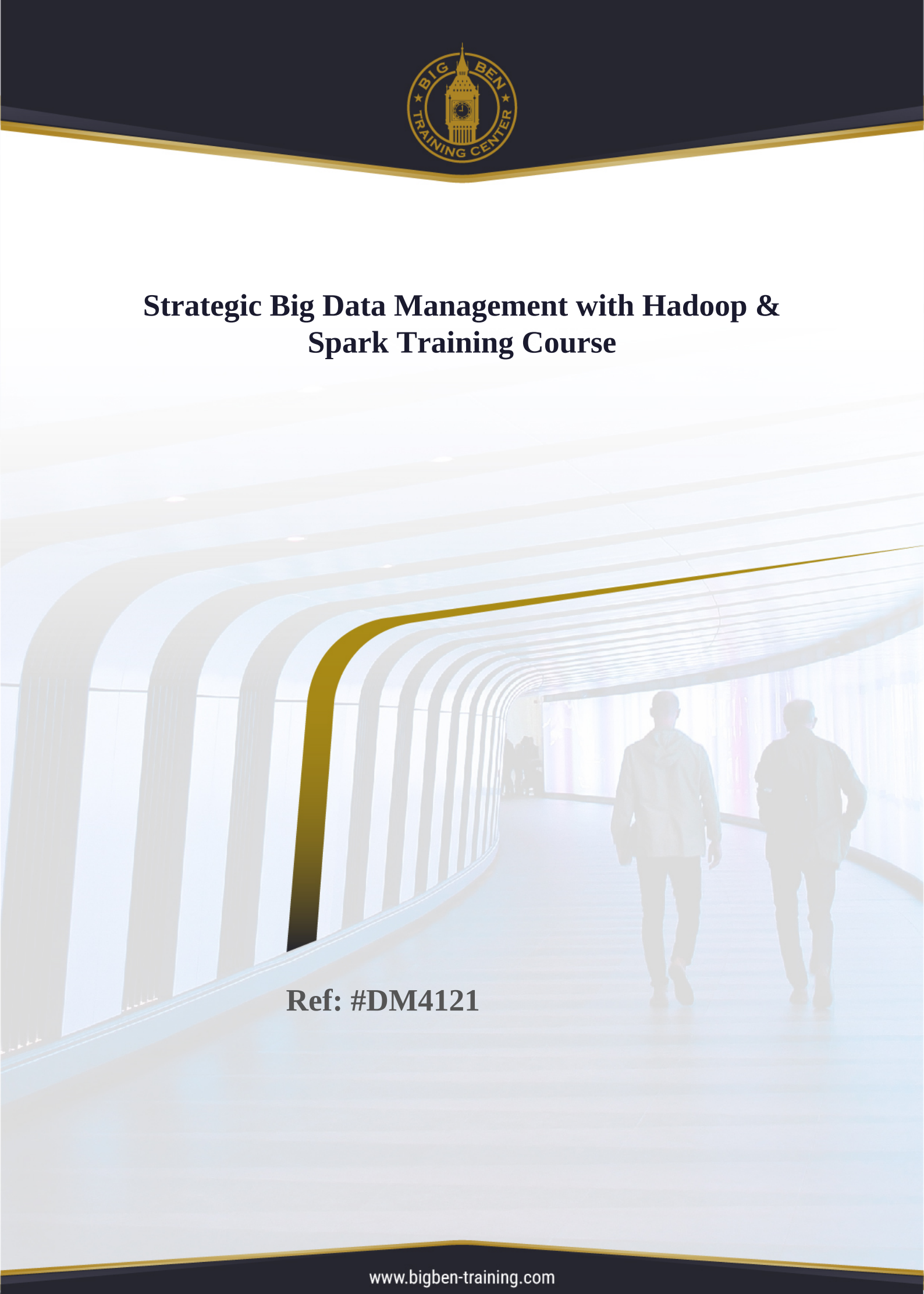
Strategic Big Data Management with Hadoop & Spark Training Course



Ref: #DM4121

Course Introduction / Overview:

The era of big data has created immense opportunities for businesses to gain new insights and a competitive edge. However, managing vast volumes of diverse data requires a new set of tools and skills. This training course is designed to give participants a deep and practical understanding of how to manage and analyze big data using the two most important frameworks in the ecosystem, Hadoop and Spark. We will explore how these technologies work together to store, process, and analyze massive datasets efficiently. The course goes beyond theory, focusing on hands-on application and real-world use cases. Participants will learn how to build big data pipelines, perform advanced analytics, and integrate these technologies into their existing data architecture. In their book, "Big Data: A Revolution That Will Transform How We Live, Work, and Think," Viktor Mayer-Schönberger and Kenneth Cukier describe how the ability to process big data is fundamentally changing every industry. At BIG BEN Training Center, we understand that mastering these tools is essential for any professional looking to leverage the power of big data. This training course will give you the skills to handle big data challenges and turn them into strategic business opportunities.

Target Audience / This training course is suitable for:

- Data engineers and data architects.
- Business intelligence professionals.
- Data scientists and analysts.
- IT professionals.
- Solutions architects.
- Database administrators.
- Anyone involved in big data initiatives.

Target Sectors and Industries:

- Technology and software.
- Financial services.
- Telecommunications.
- Retail and e-commerce.
- Healthcare.
- Government agencies and the public sector.
- Manufacturing.

Target Organizations Departments:

- Data and Analytics.
- IT Infrastructure.
- Business Intelligence.
- Data Science.
- Research and Development.
- Operations.
- Marketing.

Course Offerings:

By the end of this course, the participant will have mastered the following skills:

- Understand the core components of Hadoop and Spark.
- Store and manage big data with the Hadoop Distributed File System (HDFS).
- Process and analyze large datasets using Spark.
- Build big data pipelines for real-time and batch processing.
- Integrate Hadoop and Spark with other data tools.
- Optimize performance for big data workloads.
- Understand the architectural differences between Hadoop and Spark.
- Develop a strategic approach to big data management.

Course Methodology:

This training course is highly practical and hands-on. We use a project-based learning approach where participants will work on a series of real-world scenarios, from data ingestion to advanced analytics. The curriculum is built around live coding sessions, interactive labs, and collaborative projects that give participants direct experience with Hadoop and Spark. Our expert trainers will guide you through each step, helping you to troubleshoot problems and understand best practices. We believe that the best way to learn these complex technologies is by building something tangible. This methodology makes sure that participants leave not just with theoretical knowledge but with the practical skills and confidence to build a big data solution in their own organization.

Course Agenda (Course Units):

Unit One: ## Introduction to Big Data.

- What is big data and why it matters?
- The evolution of big data technologies.
- Hadoop: its ecosystem and core components.
- Spark: an overview and its key advantages.
- The difference between batch and real-time processing.
- The role of big data in modern business.
- Case study: a big data use case.

Unit Two: ## Hadoop Fundamentals.

- The Hadoop Distributed File System (HDFS).
- Writing data to and reading data from HDFS.
- Introduction to MapReduce.
- Hadoop's role in the modern data stack.
- Managing a Hadoop cluster.
- Securing data in Hadoop.
- Hadoop ecosystem tools like Hive and Pig.

Unit Three: ## Apache Spark for Big Data Processing.

- Introduction to Apache Spark.
- The Spark architecture and its components.
- Resilient Distributed Datasets (RDDs).
- Data Frames and Spark SQL.
- Performing data transformations and actions.
- Using Spark for data cleaning and preparation.
- Running a Spark job.

Unit Four: ## Building Data Pipelines.

- Introduction to data pipelines.
- Designing an ingestion pipeline.
- ETL/ELT processes with Spark.
- Using Spark Streaming for real-time data.
- Integrating Spark with other systems.
- Best practices for building a production pipeline.
- Case study: building a complete data pipeline.

Unit Five: ## Big Data Strategy and Advanced Topics.

- Comparing Hadoop and Spark.
- Optimizing big data workloads.
- Managing big data security and governance.
- Introduction to machine learning with Spark.
- Future trends in the big data ecosystem.
- Final project: designing a big data solution.

FAQ:

Qualifications required for registering to this course?

There are no requirements.

How long is each daily session, and what is the total number of training hours for the course?

This training course spans five days, with daily sessions ranging between 4 to 5 hours, including breaks and interactive activities, bringing the total duration to 20 - 25 training hours.

Something to think about:

As big data continues to grow in velocity and volume, how can organizations move beyond simply storing information to actively use platforms like Hadoop and Spark to uncover predictive insights and create a competitive advantage?

What unique qualities does this course offer compared to other courses?

This training course is unique because it provides a comprehensive, comparative look at both Hadoop and Spark. While many other programs focus on just one of these technologies, this course gives participants a clear understanding of how they work together, when to use each one, and how to build a complete big data ecosystem. The curriculum is highly practical, with a strong focus on hands-on labs and a final project that simulates a real-world big data challenge. This combination of a holistic view and practical application makes this program an excellent investment for anyone who needs to not only understand big data but also needs to be able to work with it effectively.