"Good [morning/afternoon] everyone. I am Nirbhay Kumar, and today I am presenting my major project report on 'Technology Training on AWS, Spark, and Data Engineering.'

Slide 2: Internship Over

Currently I'm doing intership at cognizant in coimbatore, Tamil Nadu

". The training was focused on Big Data technologies, including Apache Spark, Databricks, and AWS Data Engineering. It involved a structured curriculum with self-paced learning, conceptual sessions, and doubt clarification meetings.

Slide 3: Internship Objectiv

"The main objectives of my training are

Utilizing Apache Spark and Hive for processing large-scale IoT data.

Designing an optimized cloud-based data warehouse.

Enhancing data retrieval efficiency for better performance.

Automating ETL processes for real-time data ingestion.

These objectives helped in building a solid foundation for working with big data and cloud-based solutions."

Slide 4: Technologies & Tool

"For this training, I worked with multiple technologies:

Programming: Python and SQL were used for data manipulation and queries.

Big Data: Apache Spark and Apache Hive for distributed data processing.

Cloud Platform: AWS services like S3, Redshift, Glue, and EMR.

Databases: SQL-based and NoSQL databases like MongoDB and PostgreSQL.

Slide 5: Project Progress

"My project was structured into four key phases:

SQL & Python Foundations: Implemented complex queries and stored procedures.

Data Modeling & Warehousing: Designed star and snowflake schemas, and deployed an AWS Redshift warehouse.

ETL Pipeline Development: Developed PySpark scripts for batch processing.

Spark & Hive Integration: Set up Spark jobs for distributed data processing.

Each phase contributed to building a robust data engineering pipeline."

Slide 6: Challenges Faced & Solutions

What to say:

"During the project, I encountered several challenges and implemented solutions to address them:

Slow SQL query execution: Optimized queries using indexing and optimized joins.

Managing large IoT data in Hive: Used partitioning and bucketing to improve efficiency.

Real-time processing latency: Implemented Spark Structured Streaming for real-time data processing.

These optimizations significantly improved performance and scalability."

Slide 7: Next Steps & Future Plan

"In the next phase of the project, I plan to:

Finalize IoT data processing pipelines using AWS S3, Kafka, and Spark.

Complete ETL data transformation scripts using PySpark and SQL.

Implement real-time log monitoring and alerting with Kafka and Spark Streaming.

Conduct final integration, load testing, and performance tuning.

These steps will ensure a fully optimized and scalable data pipeline."

Slide 8: Conclusion & Learnings

I gain Hands-on experience in big data engineering.

Proficiency in Spark, Hive, SQL, and AWS services.

Optimizing workflows for real-time data processing.

These learnings have helped me develop a strong foundation in cloud-based data engineering."

Slide 9: Thank You

What to say:

"That brings me to the end of my presentation. Thank you all for your time! I’d be happy to answer any questions you may have."