Abstract: Real-time Log Analysis Using Hadoop and Spark

Introduction:

In the digital era, vast volumes of log data are generated continuously from servers, apps, and devices. Traditional analysis tools face challenges of latency and scalability. This project leverages Hadoop and Apache Spark to enable fast, scalable log analysis for monitoring, security, and troubleshooting.

Problem Statement:

Log data is high-volume, high-velocity, and diverse in format. Conventional tools delay analysis and fail to scale. The system addresses these issues through:

- Distributed storage via HDFS
- Real-time processing with Spark Streaming
- ML-based anomaly detection for early alerts

Tools Used:

- **Hadoop HDFS:** Scalable, fault-tolerant storage
- **Apache Spark:** Stream processing, Spark SQL, MLlib
- **Kafka/Flume:** Log ingestion
- **Python/Scala:** Implementation
- **Grafana/Kibana:** Dashboards & visualization

System Modules:

1. **Log Ingestion:** Kafka/Flume collects logs from varied sources.
2. **Storage:** Logs stored in HDFS; Spark accesses for processing.
3. **Processing:** Spark transforms and filters logs; ML detects anomalies.
4. **Analysis:** Regex/NLP extract fields, compute metrics.
5. **Alerting:** Alerts via email/Slack upon anomalies.
6. **Visualization:** Real-time dashboards track trends and errors.

Flow:

Log ⟶ Kafka/Flume ⟶ HDFS + Spark ⟶ Process ⟶ Analyze ⟶ Alert/Visualize

Conclusion:

The system ensures real-time log monitoring, faster troubleshooting, and improved security. It minimizes downtime and scales easily.

Future Scope:

- Deep learning integration
- Cloud deployment (AWS/Azure)
- NLP for auto log classification

24M11MC158
Sayyad Chandan
Aditya University