

Research on the architecture and working algorithms of big data storage and analysis systems

Nguyen Thi Thuy Linh

Abstract: This study investigates the architecture and operational algorithms of big data storage and analysis systems. Focusing on the core challenges of scalability, fault tolerance, and real-time processing, the research dissects the layered architecture of modern distributed frameworks, including the storage layer (e.g., HDFS), resource management (e.g., YARN), and the processing engine (e.g., Spark). A critical analysis is conducted on key algorithms, such as data partitioning strategies, replication mechanisms for fault recovery, and the MapReduce paradigm for parallel computation. The study further evaluates the performance trade-offs between batch processing and stream processing models, particularly in the context of in-memory computing. By synthesizing theoretical foundations with practical implementation patterns, this research provides a comprehensive framework for understanding how distributed systems manage petabytes of heterogeneous data. The findings offer valuable insights for optimizing system throughput and latency, contributing to the design of more efficient data pipelines for large-scale analytics.

Keywords: System architecture, algorithms, big data storage, big data analytics