Name: Deep Pawar (A20545137)
Professor: Joseph Rosen
Institute: Illinois Institute of Technology

CSP 554: Big Data Technologies

Fall 2024 - Research Paper Proposal

COMPARATIVE ANALYSIS OF BIG DATA TECHNOLOGIES ACROSS THE DATA LIFECYCLE: FROM ACQUISITION TO ANALYTICS

Abstract: In today's digital landscape, data has emerged as a critical asset driving innovation and insights across diverse domains, with organizations facing increasing challenges in managing their exponential growth in volume, variety, and velocity. This paper will target to present a comprehensive analysis of cutting-edge technologies throughout the data value chain, examining their roles and effectiveness in data acquisition (Like Spark, Storm, Kafka, Flink, and Samza), storage(in-memory technologies like Memcached, Hazelcast, and open-source queuing technologies like IBM MQ, Active MQ and Rdbms technologies like Oracle, IBM, PostgreSQL), Comparison of NoSQL/NewSQL storage technologies, and analysis. This paper will propose a structured model for categorizing and evaluating data processing technologies, providing a systematic framework to assess their performance, scalability, and suitability for specific big data tasks. This research will compare various technologies used at different stages of data processing, including real-time and batch processing tools, distributed storage solutions, and analytical frameworks, with particular attention to their capabilities in addressing challenges unique to large-scale data environments. This study will try to highlight key criteria for evaluating technologies, which will be supported by case studies across domains like healthcare, finance, and IoT, bridging the gap between academic advancements and industry implementations. The study will include detailed criteria for comparing these technologies, focusing on critical aspects such as high-speed data ingestion, real-time analytics, and distributed storage capabilities. This research will try to offer researchers, data engineers, and practitioners a valuable resource for navigating the complexities of the big data value chain and this research will try to help them select appropriate tools for their specific data management and analysis needs.

Keywords: Big Data, Data Acquisition, Spark, Storm, Kafka, Flink, Samza, Memcached, Hazelcast, Data Storage, IBM MQ, Active MQ, Oracle, IBM, Real-Time Processing, Batch Processing, NoSQL, NewSQL

References:

[1] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, and G. Alor-Hernández, ‘‘A general perspective of big data: Applications, tools, challenges and trends,’’ *J. Supercomput.*, vol. 72, no. 8, pp. 3073–3113, Aug. 2016, doi: 10.1007/s11227-015-1501-1.

[2] A. C. Ikegwu, H. F. Nweke, C. V. Anikwe, U. R. Alo, and O. R. Okonkwo, ‘‘Big data analytics for data-driven industry: A review of data sources, tools, challenges, solutions, and research directions,’’ *Cluster Comput.*, vol. 25, no. 5, pp. 3343–3387, 2022, doi: 10.1007/s10586-022-03568-5.

[3] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, Big Data technologies: A survey, Journal of King Saud University - Computer and Information Sciences, Volume 30, Issue 4, 2018, https://doi.org/10.1016/j.jksuci.2017.06.001.

[4] S. Sakr, "Big Data Processing Stacks," in IT Professional, vol. 19, no. 1, pp. 34-41, Jan.-Feb. 2017, doi: 10.1109/MITP.2017.6.

[5] K. Venkatram and M. A. Geetha, ‘‘Review on big data & analytics Concepts, philosophy, process, and applications,’’ Cybern. Inf. Technol., vol. 17, no. 2, pp. 3–27, Jun. 2017, doi: 10.1515/cait-2017-0013.