



TECHNOLOGY FEATURE *Industrial News*

Big Data and Business: How Organizations Turn Information into Competitive Advantage

Every second, 2.5 billion gigabytes of information is produced across the globe by firms. In the case of Pakistani firms, it does not lie in the availability of this information but rather their ability to convert it into meaningful actions.

By **Muhammad Zohaib Sabir** | University of Agriculture, Faisalabad | Published: September 2026

✉ 2023ag9965@uaf.edu.pk | Software Engineering

Introduction

This entire semester for this course Data Structures and Algorithms. We learn Big O notation, good sorting algorithms and how to use space efficiently. The entire lecture was theoretical. All the examples use very little data. So, I start questioning how practical is this entire theory? For example, for this Daraz website, it isn't sort few millions, it is sorting few billion.

What excites me are the differences between the theoretical idea learned in school and what it is applied for in companies; and how Big Data technology is providing the solutions to the differences. The issue I have been thinking is that what do these companies store that enormous amount of data? And how do they get meanings out of that data?

The investigation in all the above points led to finding out quickly that anyone working in the modern day technology field and claiming to be a software engineer must know what the meaning of big data is. The points of big data from technical point of view are given by this article along with the impact on the ever-growing technology sector of Pakistan.

"Big data without analysis is just noise. The real value comes from asking the right questions and building systems capable of answering them at massive scale."

What Big Data Actually Means

The term big data has been described as the four Vs and they are Volume, Variety, Velocity and Veracity I shall outline the structure of each of them.

Big Data has four characteristics, Volume refers to large data volumes, Daraz produces terabytes of data every month, and to manage the data it requires distributed compute infrastructure (Apache Spark) because it cannot be managed at scale from a single server.



Variety is unstructured data, such as comments, chat messages, and semi-structured data such as JSON. This type of data can be managed more easily with NoSQL such as MongoDB, rather than relation databases. Velocity is flowing data. For the e-commerce business, it is necessary to analyze in real-time, in order to get personalized product recommendations. The fraud detection system must distinguish fraudulent transactions within seconds; this can be done through Apache Kafka.

Veracity deals with the reliability of data. "Garbage in- garbage out," is a widely used term, and it applies perfectly here. 80% of data scientists spend their time cleaning and preparing the data.

The Technical Foundation

So how does a Big Data system function? Data is collected through Web Servers, Databases, sensors and mobile apps and Apache Kafka ingests this data reliably and quickly like a giant conveyor belt. Data is then stored in various ways. For traditional single transactional processes MySQL or a relation DB will work. But Big data needs a separate process for analytics. HDFS within Hadoop will fragment a file into several blocks across various machines for parallel processing and data is processed by Apache Spark.

Before MapReduce used to dump data into disk after every operation which made it slow. Now Apache spark has made it better by taking it in memory, now supporting streaming and Machine Learning. This is the current standard.

Lastly, the data is being used for SQL queries, statistical applications and Machine Learning for analyzing business processes, customer behavior, detect fraud transactions, building recommender engines. For the Pakistani businesses, the good news is no data centers needed, as data will be stored in the cloud services (AWS, Google cloud and Azure), millions of transactions will be processed through the resources available over the cloud.

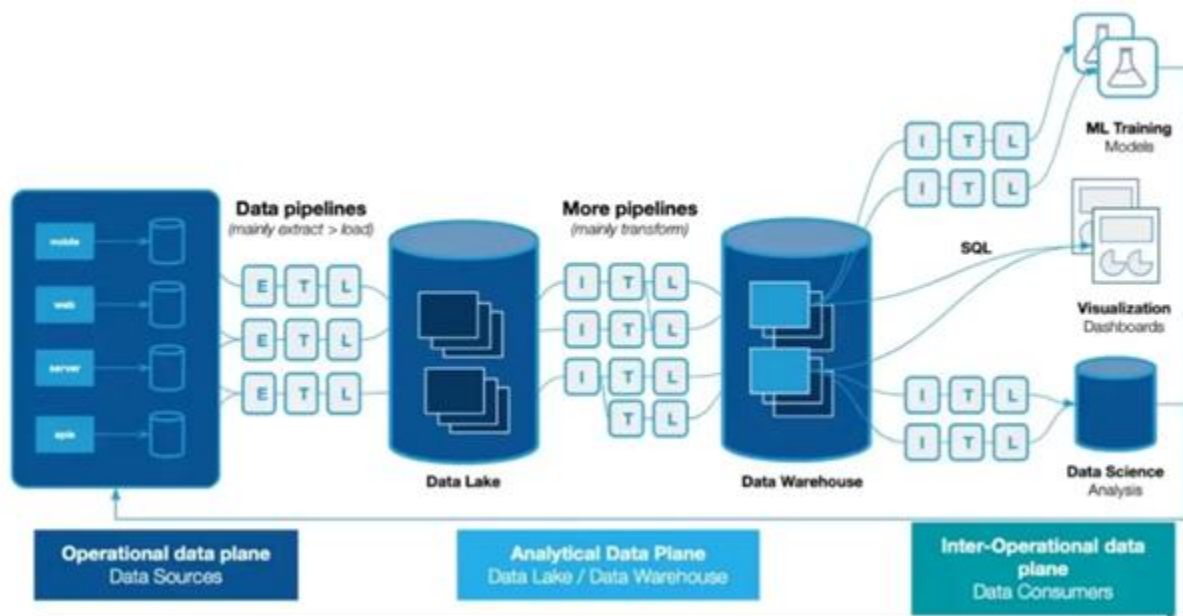


Figure 1: Typical big data pipeline showing how data flows from sources through collection, storage, processing, and finally to analytics.



Challenges Specific to Pakistan

But, the technology, though universally available, presents a lot of challenges if applied in Pakistan. The main challenge is related to people for deploying the technology; the technology is going to be operated by people capable of handling the domain of distributed systems, data engineering, and cloud computing. Pakistan is not short of able engineers, but universities are not quick enough to tackle the skill-gap.

Secondly, the proper infrastructure. We should be ensured we are not without power and high-speed internet always. We may not be able to have that in remote places instead of cities. The instability like that is what we cannot use cloud systems in Pakistan.

Third is regulation uncertainty. In countries such as EU countries there exists a regulation (General Data Protection Regulation), and Pakistan doesn't have any such data privacy law. Organizations are unclear as to what actions they can undertake concerning the customers' data and thus are hesitant towards embracing technology.

Fourth, building a private infrastructure requires huge expenditure; cloud infrastructure services are inexpensive, but then you don't have your data sovereignty; Fifth, Data silos: "In an organization there are usually many stand-alone systems scattered in the various departments of an organization e.g. HR, finance and etc, that do not "talk" to each other."

Finally, generation gap. Old type engineers are the majority in most of the companies.

The Opportunity for Pakistan

In all these hardships that Pakistan has; it has a huge opportunity of utilizing big data. There are a lot of factors that enable this to happen. Firstly, young Pakistan uses technologies quite well. With each day more people possess mobile phones and the number of digital payments is increased thus the number of data is increased also. Secondly Pakistani Universities is very known for producing excellent engineers which show that we are not far from anything when it comes to these things and we have enough capacity of using it. With the use of local resources, we will get marvellous results.

The Digital Pakistan project would be the positive sign, like Smart Cities projects too; regulation policies about data processing and AI are also being formed and the hurdle, imposed for companies to not adopt big data is being removed using cloud computing services; Pakistani startups can avail services of AWS or Google Cloud which can process a bulk amount of data at cheap rate.

It is clear to see that Big Data will be a competitive edge to place any business firm ahead of the rest when making a decision. A firm that is willing to put Big Data money into the industry in a timely manner will come out on top. There should be a curriculum put into place for data engineering, and the definition of data security need to be given as well.



About the Author

Muhammad Zohaib Sabir is a student pursuing Software Engineering in his 3rd year at the University of Agriculture, Faisalabad. He has gained a deep interest in using machine learning to solve practical problems due to courses on artificial intelligence, data structures, and systems design. His contact information is: 2023ag9965@uaf.edu.pk.

Keywords: Big Data, Distributed Systems, Data Engineering, Apache Spark, Apache Kafka, Machine Learning, Data Analytics, Business Intelligence, Pakistan, Digital Transformation

References

- [1] Dean, J. and Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113. Available at: <https://research.google/pubs/mapreduce-simplified-data-processing-on-large-clusters/>
- [2] Zaharia, M., Xin, R.S., Wendell, P., et al. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65. Available at: <https://cacm.acm.org/magazines/2016/11/209116-apache-spark>
- [3] Kreps, J., Narkhede, N., and Rao, J. (2011). Kafka: A distributed messaging system for log processing. In Proceedings of the NetDB Workshop. Available at: <https://kafka.apache.org/>
- [4] White, T. (2012). Hadoop: The Definitive Guide. Third Edition. O'Reilly Media. Available at: <https://www.oreilly.com/library/view/hadoop-the-definitive/9781491901632/>
- [5] Mayer-Schonberger, V. and Cukier, K. (2013). Big Data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt. Available at: <https://www.bigdatathebook.com/>
- [6] Labrinidis, A. and Jagadish, H.V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033. Available at: https://vldb.org/pvldb/vol5/p2032_labrinidis_vldb2012.pdf