Podcast Name: *ACM ByteCast*
Episode: Matei Zaharia - Episode 32

Welcome to the *ACM ByteCast* podcast, a series from the Association for Computing Machinery! The podcast features conversations with researchers, practitioners, and innovators at the intersection of computing research and practice about their experiences, lessons learned, and visions for the future of computing. In this episode, host Bruke Kifle is joined by Dr. Matei Zaharia, Chief Technologist and Co-Founder of Databricks and Assistant Professor of Computer Science at Stanford. His research has been recognized through the 2014 ACM Doctoral DIssertation Award, NSF Career Award and the U.S. Presidential Early Career Award for Science and Engineers.

Matei was born in Romania, grew up mostly in Canada and got into computer science during university. His biggest draw into the field was his interest in computing and the speed computers allowed for trying out the latest techniques. At the University of Waterloo, he was fortunate enough to work with a networking professor who got him interested in research. After applying to PhD programs, he ended up at UC Berkeley and became interested in large-scale data centers and Cloud computing. His development of the multi-language engine Apache Spark was a key inflection point in his journey. Spark, he explains, is a framework for writing programs and processing data. At its core is the API which allows you to write single machine code to create large scale programs. It is also a useful framework for combining high level libraries into a larger application. It was the first to open up large scale systems beyond software engineers.

Matei began working on Spark in 2009 as a grad student. At the time, many tech-centric companies expressed interest in the program. He realized that there was enough interest to justify creating a company in the field and helped cement their ability to contribute to the Open Source project. It is important for stakeholders to have a common presentation of data. Thus, the data model in both Spark and Matei's current company Databrick is the same for all users. However, there is the option to write a specialized function other people can use. They believe there is an opportunity to simplify things within one engine, one data load.

Then, Matei highlights the different factors influencing the need for this new paradigm shift towards lakes compared to more traditional data warehouses. Data warehouse systems were typically designed to be deployed on their own servers. When you have everything in the cloud, however, it becomes a problem to have data locked into something which only one system can read. Basically, data lake is low-cost storage where you can store files in any format. It is the best solution for cheaply storing large amounts of data without loading it into limited proprietary systems. Lakehouse is an emerging trend which combines data warehouse performance and management features with low-cost storage and open format.

Next, the conversation shifts to discussing Matei's other project, MLflow, an open source platform for managing the end-to-end machine learning lifecycle. He then highlights exciting recent announcements in the field of AI and machine learning. The benefits of open source include improving access, adoption and ease of extension with other tools of your choice. When

Matei isn't shaping the future of Databricks, he is actively involved in the future of computer science as an Assistant Professor at Stanford. Listen as he shared about his exciting research and endeavors in the realm of academics and research. The biggest gaps he has noticed in computing education, he reveals, is the recent shift towards software as a service. He would love to see a class in which students employ a service on day one and keep it operating throughout the semester. As the conversation wraps up, he touches on juggling his industry work with his role in academia. Finally, he looks forward to the future of data management, machine learning and computing at large.

Links:
Learn more about Matei Zaharia and Databricks.
Learn more about the Association for Computing Machinery (ACM) at acm.org.
Learn more about the ACM ByteCast podcast at acm.org/bytecast.

Key Takeaways:
0:29 - Introduction to this episode of *ACM ByteCast.*
1:44 - What led Matei to the field of computing?
3:34 - What is Apache Spark?
12:18 - The key stakeholders of Databricks.
17:00 - Compelling examples of the one engine, one data load solution.
20:03 - What is influencing the need for this new paradigm?
24:09 - How MLFlow addresses challenges in the ML lifecycle.
29:28 - Exciting announcements in the field of AI.
33:47 - Motivations contributing to open source software.
36:27 - Matei shares about his work at Stanford University.
46:15 - The biggest gaps Matei observes in computing education.
49:37 - How Matei juggles his industry work with his role in academia.
52:21 - The future of the field of data management, machine learning and computing.

Tags:
ACM bytecast, computing, machine learning, computer science, technology, research, data centers, UC Berkeley, Apache, Apache Spark, python, java, coding, code writing, programming, data processing, software development, open source, software engineers, data engineer, Databricks, data warehouse systems, cloud, cloud migration, applications, open format, lakehouse, machine learning, AI, digitization, MLflow, data analytics, Stanford University