The University of Chicago Computer Science Department

PRESENTS:

"Fault-Tolerance in Apache Spark (and where it has failed)"

Imran Rashid

Developer, Cloudera

Abstract:

Apache Spark is a distributed computing platform built using Scala with fault-tolerance in mind. It has been tested rigorously and deployed in production at many companies for years. And yet, fault-tolerance issues are still surfaced. How did these faults slip through?

The talk will focus on the practical aspects for developers working with Spark: as a user of Spark, what can you reasonably expect? What should you still be wary of? And as a developer working on Spark itself, whats types of bugs have occurred? Why does it take so long to discover and fix these issues?

Time-permitting, I will also talk about scalability issues in Spark, beyond fault-tolerance.

Bio:

Imran Rashid is a Committer and PMC member of Apache Spark, and a developer at Cloudera. He has used Hadoop for the past 8 years, and been closely involved in Spark even before it was an Apache project. Imran used to write Spark programs for machine learning on terabytes of data -- now he spends his time hunting down bugs that surface when Spark is pushed to 1000s of nodes.

Wednesday, August 23, 2017
2:00pm
Ryerson 255
Host: Haryadi Gunawi