Ankur Dave

2703 Ridge Rd, Apt. 203   Berkeley, CA 94709
ankurdave@gmail.com   http://ankurdave.com
mobile: 650-701-7705

EDUCATION

University of California, Berkeley   August 2013 to present
Ph.D. candidate (incomplete), Computer Science   GPA: 3.87/4.0

University of California, Berkeley   August 2010 to May 2013
B.S., Electrical Engineering and Computer Science   GPA: 3.79/4.0

EXPERIENCE

UC Berkeley RISELab   *Berkeley, CA*
Graduate Research Assistant   August 2013 to present

Contributed to a variety of research systems for big data processing using Scala and C++. Advised by Prof. Ion Stoica.

- 2013–2014: Lead developer for GraphX [4], Apache Spark's distributed graph processing library. Proposed and implemented optimizations including indexed joins and incremental view maintenance (see Sec. 4 of OSDI paper [4]). Led effort to merge GraphX into Apache Spark.
- 2014–2016: Created IndexedRDD [2], a Spark package for efficient fine-grained updates to immutable in-memory datasets using a novel persistent data structure called PART. Supported the use of IndexedRDD in several AMPLab projects including G-OLA (Zeng et al., 2015) and Tegra (Iyer et al., 2017).
- 2015: Created GraphFrames [3], a library to execute Cypher graph pattern queries as relational queries in Spark SQL.
- 2016–present: Lead developer for Opaque [1], a Spark SQL package to enable in-memory encryption and access pattern hiding for relational data using Intel SGX hardware enclaves. Designed and implemented Opaque as a set of security-aware query planner rules (see Sec. 6 of NSDI paper [1]) and in-enclave relational operators, enabling drop-in security for Spark SQL. Supported the use of Opaque at IBM Research.

UC Berkeley AMPLab   *Berkeley, CA*
Undergraduate Research Assistant   September 2010 to May 2013

Contributed evaluation of user-controllable partitioning and Bagel, a Pregel-like graph processing library, to NSDI 2012 paper on Spark [6]. Created Arthur [5], a distributed replay debugger for Spark programs. Advised by Matei Zaharia and Prof. Scott Shenker.

Google   *Mountain View, CA*
Software Engineer Intern   May to August 2012

Designed an algorithm to efficiently allocate heterogeneous cluster resources to running jobs in Google's workflow execution engine. Proved NP-hardness of this optimization problem using a reduction from vertex cover. Proposed and implemented a network flow–based solution to the linear relaxation of the problem.

Facebook   *Palo Alto, CA*
Software Engineer Intern   May to August 2011

Added network usage tracking and limit enforcement to Facebook's cluster manager using the traffic control features of the Linux kernel.

Microsoft Research   *Redmond, WA*
Research Intern   June to August 2010

Implemented distributed, fault-tolerant k-means clustering on the Azure cloud and published at DataCloud 2011 workshop.

SELECTED PUBLICATIONS

[1] W. Zheng, A. Dave, J. G. Beekman, R. A. Popa, J. E. Gonzalez, I. Stoica.
Opaque: An Oblivious and Encrypted Distributed Analytics Platform. *NSDI*, March 2017.

[2] A. Dave, J. E. Gonzalez, M. J. Franklin, I. Stoica.
Persistent Adaptive Radix Trees: Efficient Fine-Grained Updates to Immutable Data. Technical report. March 2017.

[3] A. Dave, A. Jindal, L. E. Li, R. Xin, J. E. Gonzalez, M. Zaharia.
GraphFrames: An Integrated API for Mixing Graph and Relational Queries. *GRADES* workshop, June 2016.

[4] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, I. Stoica.
GraphX: Graph Processing in a Distributed Dataflow Framework. *OSDI*, October 2014.

[5] A. Dave, M. Zaharia, S. Shenker, I. Stoica.
Arthur: Rich Post-Facto Debugging for Production Analytics Applications. Technical report. January 2013.

[6] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, I. Stoica.
Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing, *NSDI*, April 2012.