

# Niketan Pansare

Research Software Engineer,  
IBM Research Almaden.

Website: <http://ibm.biz/niketan>

3595 Granada  
Avenue, Apt 286  
Santa Clara, CA.

☎ 352-870-4782

✉ [npansar@us.ibm.com](mailto:npansar@us.ibm.com)

---

## Research Interests

My research focuses on building scalable data processing systems for large-scale machine learning. I explore the intersections of machine learning, deep learning, distributed systems, and online aggregation. I am particularly interested in investigating various challenges associated with automatic optimization of machine learning algorithms for various data and cluster characteristics.

---

## Education

- 2009–Dec 2014 **Ph.D. in Computer Science**, *Rice University*, Houston.  
Dissertation Title: *Large-Scale Online Aggregation Via Distributed Systems*  
Advisor: Prof. Christopher Jermaine
- 2007–2009 **Master of Science in Computer Engineering**, *University of Florida*, Gainesville.
- 2002–2006 **Bachelor of Engineering in Information Technology**, *VJTI*, Mumbai.
- 2003–2004 **Post Graduate Diploma in Embedded Systems**, *Electronics Corporation of India Limited*, Mumbai.

---

## Experience

- Aug 2014 – Present **Research Software Engineer**, *IBM Research*, Almaden.  
- Implemented the Spark runtime for SystemML and Spark APIs (Python, Scala and Java) for SystemML.  
- Implemented the GPU runtime for SystemML and added low-level primitives to support Deep Learning in SystemML.  
- To simplify usage, added support for Caffe APIs (for deep learning use-case) and scikit-learn as well as MLLib compatible APIs (for machine learning use-case).  
- SystemML is now a top-level Apache project and is also listed as one of Almaden Research Center's 2015 accomplishment under "Research Contributions to Spark Signature Moment".
- Summer 2011 **Research Intern**, *IBM Research*, New Delhi.  
- Worked on a novel bayesian model that performs topic modeling on the transcripts of audio files. The model incorporates the uncertainty in speech-to-text process that helps improve the accuracy of topic modeling in presence of noisy data.
- Summer 2008 **Software Development Engineer Intern**, *Microsoft*, Seattle.  
- Developed Table Analysis Tool for Cloud which is set of canned data mining tasks (such as forecasting, market basket analysis, etc) delivered as an Excel plugin.
- 2006–2007 **Software Engineer**, *MAQSoftware*, Mumbai.  
- Designed and built enterprise web application using C#, ASP.NET Ajax and XML. Also, developed data warehouse (Usage Reporting) for Microsoft using C# and SQL Server Business Intelligence Studio.

---

## Publications

- 2016 Boehm M, Dusenberry M, Eriksson D, Evfimievski A, Makari Manshadi F, **Pansare N**, Reinwald R, Reiss F, Sen P, Surve A, Tatikonda S. *SystemML: Declarative Machine Learning on Spark*. Proc. VLDB Endow., 2016.
- 2016 Boehm M, Evfimievski A, **Pansare N**, Reinwald R. *Declarative Machine Learning - A Classification of Basic Properties and Types*. ArXiv e-prints., 2016.
- 2014 **Pansare N**. *Large-Scale Online Aggregation Via Distributed Systems*. PhD Thesis, Rice University, 2014.
- 2012 **Pansare N**, Jermaine CM, Haas P, Rajput N. *Topic Models over Spoken Language*. IEEE International Conference on Data Mining (ICDM '12), December 2012.
- 2011 **Pansare N**, Borkar V, Jermaine CM, Condie T. *Online Aggregation for Large MapReduce Jobs*. Proc. VLDB Endow., August 2011.
- 2011 Sahay S, Rajput N, **Pansare N**. *Social Ranking for Spoken Web Search*. CIKM 2011.
- 2010 Arumugam S, Dobra A, Jermaine CM, **Pansare N**, Perez L. *The DataPath system: a data-centric analytic processing engine for large data warehouses*. ACM SIGMOD '10, June 2010.
- 2009 **Pansare N**. *Multi-query optimization in the Datapath system*. Master's thesis, 2009. University of Florida, Gainesville, USA.

## Projects

### Apache SystemML

- SystemML provides declarative large-scale machine learning (ML) that aims at flexible specification of ML algorithms and automatic generation of hybrid runtime plans ranging from single-node, in-memory computations, to distributed computations on Apache Hadoop and Apache Spark.

- Link: <http://systemml.apache.org/>

### Gradient Descent over MapReduce using OLA

- Extended the OLA model to multi-dimensional data and also implemented a bayesian model that improves the convergence of gradient descent (and thereby the performance of machine learning algorithms) over large data.

- Link: [http://youtu.be/JKmNI\\_eC9ss](http://youtu.be/JKmNI_eC9ss)

### Spoken Topic Model (STM)

- Implemented STM (see ICDM '12 paper) in C++ using GNU Scientific library (GSL).  
- CMU Sphinx4 speech-to-text engine was modified and data was generated by providing it with real-world audio files (TedTalks/Yale).

- The effectiveness of STM was tested by comparing it to Latent Dirichlet Allocation using off-the-shelf classifiers (SVM<sup>light</sup>, SVM<sup>multiclass</sup> and Weka).

### Online Aggregation (OLA)

- Modified Hyracks (Hadoop-like system) to provide necessary machinery for OLA.

- Dealt with Inspection paradox in a principled way to provide unbiased estimates using bayesian model implemented in C++.

- The overall system was then tested using Wikipedia traffic dataset.

- Link: <http://www.youtube.com/watch?v=GipmpNEZbWk>

### Datapath system

- Data-centric database implemented from ground-up and tested on 10TB scale TPC-H data-set.

- Developed multi-query optimizer in C++ to provide data-centric query plans.

- Link: <http://research.microsoft.com/apps/video/dl.aspx?id=133049>

- Non-research/  
Personal
- Yadmt: Tool to find best off-the-shelf classifier for your dataset using variety of statistical tests. It requires very little knowledge of machine learning literature and can also run over cluster for faster search of the classifier. Link: <http://code.google.com/p/yadmt/>.
  - Voca: Siri-like desktop app (written in Java) that is designed to run in background, with minimal user interaction/interference, and that allows users to issue voice commands. Link: <https://www.facebook.com/voca.desktop>.
  - Vigilant: Affordable and hassle-free personal safety app that connects you to friends, family and emergency services without the need to take out your iPhone. The app is geared to work globally and especially in developing countries. Link: <http://www.vigilant-app.com/>
- For detailed listing of my projects/courses, see <http://www.linkedin.com/in/niketan>.

---

## Tools

Pgm Lang Technologies **C, C++, Java, Scala, R**, Python, C#, Objective C, SQL  
**Hadoop, Apache Spark**, CUDA, Hyracks

---

## Awards/Roles

- 2016 IBM Manager's Choice Award.
- 2016 IBM Research Division Accomplishment Award.
- 2015 IBM Manager's Choice Award.
- 2015-current The Apache Software Foundation Project Management Committee (PMC) Member.
- 2013-14 IBM Ph.D. Fellowship.
- 2013-14 President of Rice Computer Science Graduate Student Association (CS GSA).

---

## Invited Talks / Presentations

- November 2016 GPU-accelerated Deep Learning in Apache SystemML - IBM Thomas J. Watson Research Center.
- November 2016 Accelerating Data Science with Apache SystemML - Future of Data - New York meetup.
- October 2016 Summary of Bay Area Deep Learning School - IBM Almaden Lunch Talk.
- October 2016 Apache SystemML - Declarative Machine Learning at Scale - University of California, Merced.
- April 2016 Declarative Machine Learning at Scale - IBM Datapalooza, Austin, TX.
- April 2016 SystemML: Declarative Machine Learning on Spark - Rice University, Houston, TX.

---

## Conference reviewer

- 2017 International Conference on Very Large Databases (VLDB).
- 2016 International Conference on Very Large Databases (VLDB) and IEEE BigData.
- 2015 IEEE BigData.