# Ata Turk - Research Statement

## Overview

In the future, all computation will move to the Cloud. Meaningful Cloud computing research is fundamentally challenging: in addition to standing on sound theoretical grounds, it has to be deployed on real systems at a realistic scale, requires significant engineering effort, and often involves large teams and industry collaboration. I find this kind of research compelling because the problems are complex, intellectually challenging, require cross-disciplinary (economics, legal, privacy, …) solutions and have very high-impact. During my time as a PhD student, startup founder, postdoc at Yahoo! research, and research scientist at the Massachusetts Open Cloud (MOC), I was fortunate to be involved in projects where I had the opportunity to experience and develop the skills necessary to succeed in this kind of research. I view my primary research direction going forward as starting new targeted research initiatives for cloud-assisted big data solutions for emerging applications such as healthcare, renewable energy, and crypto finance systems, while building on the projects I helped get started at MOC. I also plan to keep exploiting (and contributing to) fundamental computer science theory to improve performance and efficiency of Cloud applications.

My PhD initially focused on parallel information retrieval systems where I built highly efficient distributed search engines [1-4]. I also worked on combinatorial problems, developed new graph theoretical solutions for the (replicated) hypergraph partitioning problem, and then built graph partitioning tools that applied these ideas [5]. During my PhD I took a three year break to found and lead a successful Cloud-based startup that grew to twenty full time developers by the time I sold it. This experience led to a fascination with Cloud Computing and its economic and technical connections to Big Data; directly leading to my subsequent PhD research where I designed social network data storage solutions [6].

After my PhD I joined Yahoo! Research as a postdoctoral researcher. There, I had a chance to work with very large datasets and large-scale Big Data Analytics systems. I was involved in the application of machine learning solutions to large-scale problems and contributed to the improvement of personalization and suggestion systems as well as search solutions [7]. The algorithms and the software I developed was eventually productized by the Yahoo! engineering team and deployed in a system that ran daily on over thousands of servers.

In 2015, I had the opportunity to join the Massachusetts Open Cloud (MOC) project as a postdoctoral researcher in Boston University. When I joined, the MOC had just gotten its first seed funding for its vision of creating an open cloud model that enables research, while offering real service to users. At the time I joined MOC, the team only had half a PhD student and a number of core ideas with no implementation. However, this project excited me because it resonated with my focus on doing research with impact on real systems. In my time at the MOC, the research team grow to dozens of researchers with more than ten PhD students working over the fundamental research problems posed by the MOC. I am involved in most of these research projects and co-supervising many of the PhD students together with the PIs of the MOC and other involved researchers. I also played a critical role in transitioning the project from a tiny prototype to a successful production-grade service. I formed teams that worked over the set of MOC core ideas and led those teams in turning the ideas into concrete Cloud frameworks that are now being used in production and that have their own open source communities [13-16, 21, 22]. MOC currently supports many leading-edge research projects with heavy industry involvement, has attracted over 9 million dollars in industry funding, and 12 million in NSF grants, and supports hundreds of users. I was involved in all aspects of this transition. In addition to mentoring teams of graduate and undergraduate students, I led the MOC industry/academic working group in the area of Big Data, and played a major role in securing funding from the industry and funding agencies. Recently, the Cloud research we conducted has resulted in a 5 million dollar investment from Red Hat to open their first research lab at the Boston University.

In the coming sections I will describe two of the projects I currently work on in Cloud computing, my ongoing work on graph analytics and my future directions.

## 1. Cloud monitoring and operations analytics

In this section I describe my work on designing effective, and secure collection and servicing mechanisms of Cloud operational data and how I utilize this data to offer machine-learning-based optimizations for improved efficiency and security in Cloud operations.

**Cloud Monitoring Motivation:** *Public clouds are opaque.* Cloud users and Cloud system researchers have little visibility into the performance characteristics and utilization of the physical Cloud resources they use. Performance sensitive public Cloud users are forced to identify critical performance metrics by elaborate schemes such as benchmarking VMs to select the ones with desired characteristics. Cloud performance reverse engineering is turning into a research field! Optimized approaches for reverse engineering with minimal cost appear in top conferences and large companies such as Netflix open-source their automated reverse engineering libraries. Clearly the efforts by Cloud users/researchers to predict these metrics that Cloud providers can easily measure and share, but are not incentivized to release, is a waste of resources.

**Research Questions & Projects:** *What kinds of data shall we collect from public Cloud infrastructures to advance Cloud research and optimization? How shall we expose this data to researchers and end-users? How can we avoid privacy implications of sharing Cloud monitoring data?* To address these questions I am collaborating with researchers from BU and Northeastern to build Cloud monitoring and Cloud data exposure solutions. I am also collaborating with researchers from BU, IBM and Harvard to protect user privacy while exposing Cloud monitoring data.

Cloud Transparency for Enabling Cloud Research (Collaboration with researchers from Boston and Northeastern Universities): I led the monitoring team in the Massachusetts Open Cloud (MOC) project for two years and am still a part of the team. As a team our aim is to design and implement a Cloud monitoring platform that in addition to performing standard monitoring tasks such as alerting, debugging and billing, can expose users and researchers resource utilization data associated with the physical services their VMs operate on, showcase possible applications enabled by using these information sources for research and optimization (of both the Cloud service and the end-user performance), and investigate and implement mechanisms to share resource utilization data without compromising the privacy of the users of the Cloud. We designed and implemented an end-to-end monitoring framework for OpenStack-based Cloud infrastructures that monitors multiple layers (datacenter, physical, networking, middleware, ...) in the stack. All the collected data is normalized and correlated in a time-series database and is exposed thru RESTful APIs to applications. Furthermore, historical data is teased out of this database into object storage platforms (Swift) and also committed into HDFS for batch, long-term analysis using Big Data analytics platforms. We published parts of our Cloud monitoring and exposure infrastructure in Usenix CoolDC [8].

Protecting User Privacy while Exposing Monitoring Data (Collaboration with researchers from Boston and Harvard Universities): I envision a public Cloud framework where each user is provided real-time access to monitoring data from underlying Cloud layers related with his/her workloads. In such a scenario, we need to consider a stronger, *active adversarial setting*. Cloud users might act maliciously and deceptively change their workloads to obtain additional information on other users' workloads. Such active attacks have been demonstrated for co-location detection in public Clouds. I am working on mechanisms to alleviate such concerns via employing *differential privacy* to obscure the presence of user private data in the exposed data while retaining the usefulness of the data for performance optimization tasks. We proposed a *decaying*

*differential privacy* mechanism that provides higher privacy guarantees for more recent data and more accuracy guarantees for historical data and our initial analysis results appeared in ICDM PAIS [19]. I expect this project to lead to several publications and grant submissions.

**Cloud Operations Analytics Motivation:** *All computation is moving to the Cloud*. This shift necessitates development of *Cloud Analytics* solutions that securely and reliably process the vast amounts of data arriving-to and generated-in the Cloud for addressing *Cloud systems management*, *Cloud security*, and other operational requirements.

**Research Questions & Projects:** *How can we use Cloud operational data to improve security and performance of Cloud applications?* I organized and chaired the IEEE Workshop on Big Data for Cloud Operations Management [10] during the 2016 IEEE BigData conference. I am collaborating with researchers from IBM to provide Cloud Analytics as a Cloud service and we published our machine-learning-based software discovery and system analytics techniques at IBM Journal [11] and IEEE BigData [12]. We showed the potential of user profiling for detecting Cloud account compromises in IEEE BigData PSBD [20]. We demonstrated usage of cloud monitoring data for datacenter energy optimization purposes in Usenix CoolDC [8] and IGSC [9] and for detecting performance anomalies in ISC [17].

Scalable Discovery in the Cloud - Cloud Systems as Data (Collaboration with IBM): State-of-the-art cloud deployments operated by large organizations frequently scale up to thousands of cloud instances. Complications in managing such large scale deployments are exacerbated by the widespread adoption of DevOps methodologies, agile development, and Continuous Integration / Continuous Delivery practices. A main IT management challenge is tracking and understanding changes for purposes such as compliance, security, configuration drift or misconfiguration analysis. In this project I approach cloud systems as data, use techniques commonly used in information retrieval and semantic search to design cloud instance (VM, container) representations that enable efficient analysis and search of data pertinent to various operational management tasks in the cloud, offer machine learning inspired methods for rapid discovery and analysis, and build "systems knowledge bases" containing factual relations among system entities to improve the accuracy and efficiency of the proposed approaches. In collaboration with researchers from IBM T. J. Watson, we are working on building a Cloud service called DeltaSherlock that can automatically identify multiple system changes by utilizing features extracted from file systems. Our system discovery approach and DeltaSherlock service implementation appeared at IBM Journal [11] and IEEE BigData [12] and parts of the service is embedded in IBM's Bluemix service.

Cloud Security Analytics (Collaboration with researchers at Boston and Northeastern Universities): In this project, we propose analytics-based services for securing public Cloud infrastructures against prevalent classes of emerging threats such as data breaches, account and service hijacking or illegitimate use of Cloud services. The model of public Clouds creates a tension between securing the Cloud infrastructure and protecting user privacy. The main goals of this project are to provide insights into these conflicting requirements from both the Cloud provider's and users' perspective, and to offer a deeper understanding on the benefits and risks of security analytics for public Clouds. We propose methods to quantify both the privacy implications and accuracy in security analytics tasks for various data sources collected at all Cloud layers (physical, hypervisor, networking, cloud management). We also design a flexible analytics platform that supports Cloud-level and user-level analytics for addressing different threat scenarios, as well as joint analytics algorithms that respect the privacy requirements determined by users. We defined our early architecture vision in [13]. Our initial take on building end-user profiling and security analysis applications utilizing the historical monitoring data is presented in IEEE BigData PSBD [20].

Cloud Analytics for Improved Efficiency (Collaboration with researchers at Boston University and Sandia National Labs): We used the monitoring data to offer machine-learning-based performance variation detection mechanisms for HPC applications running in cloud systems in HPC-ISC [17] (Gauss Award winner). We also demonstrated usage of cloud monitoring data for datacenter energy optimization purposes in Usenix CoolDC [8] and IGSC [9]. Via these efforts we aim to expose the benefits of sharing the Cloud monitoring data with end-users and researchers and hope to incentivize public Cloud providers to share monitoring data with researchers.

## 2. Cloud-hosted Big Data Platforms

In this section, I explain my work on building Cloud-based, high-performance, end-to-end Big Data solutions for performance critical applications such as healthcare problems.

**Cloud-hosted Big Data Platforms Motivation:** *We need Cloud-based Big Data Analytics Platforms that can cater to the needs of critical applications such as Healthcare*: An increasing number of large datasets are now being made available to researchers. A good example is biomedical research datasets. Various biological and biomedical studies (e.g. anatomical, physiological, clinical, behavioral, environmental, imaging, phenotypic, genetic, molecular, etc.) are now being conducted by various research groups. Unfortunately, secure release and management of the results of these studies is a challenge. The potential impact of intelligently combining these data sources to offer holistic diagnosis approaches is not realized. Healthcare research community realizes the need for data-centric solutions that can combine different information sources from multiple possibly non-trusting entities and perform secure anonymized computation over these data sets. Furthermore, healthcare applications that have extensive computational needs and require access to GPUs, FPGAs, or other accelerators are either forced to build one-off solutions with low-utilization or forced to buy into one Cloud vendor's implementation of Big Data frameworks and workflows, facing potential high cost and vendor lock-in problems. There is need for open-source and standardized solutions that democratize the transition into cloud-hosted big-data analytics platforms.

**Research Questions & Projects:** *How can we enable critical applications such as healthcare to better utilize Big Data analytics and Cloud computing solutions?* On this topic, I am collaborating with researchers from BU, Boston Children's Hospital, Harvard, and Red Hat to develop secure and anonymized data transfer and computation orchestration mechanisms that enable hospitals to exploit Cloud resources for complex processing needs. Our initial design appeared in VLDB DMAH [16]. I am also collaborating with researchers from BU, Two Sigma, and Harvard to build Cloud-based dataset repository systems that can support the many stages datasets go through and provide secure multi-party computational capabilities on the Cloud for these datasets to enable analytics over datasets owned by non-trusting entities.

Cloud-Assisted MR Imaging (to be submitted to NIH RO1 Call - Collaboration with Boston Children's Hospital): In this project we propose a Cloud-assisted MR image processing framework that cananonymize and securely transfer MR data from/to hospital Picture Archiving and Communication Systems to the Cloud, orchestrate complex image processing workflows requiring multiple stages of computation via intelligent parallelization and containerization to achieve clinically relevant processing latencies, and provide a Cloud testbed with access to GPUs and accelerators in order to eliminate need for replicating hardware testbeds to use imaging code developed by other researchers [16]. This framework will utilize the container service, and the on-demand bare-metal Big Data provisioning frameworks [14, 15, 21, 22] we built in MOC.

Cloud Dataset Repository and Multi-Party Computation framework: (supported by the NSF CICI grant 2017-2019 – budget: one million dollar): This is a two-year infrastructure grant that I wrote together with

two other collaborators from Boston University. In this project we are building a dataset repository and computation platform that can host datasets from multiple non-trusting entities and enable computation over these datasets in a secure setting. This is a core problem NIH and NSF focus on to further the healthcare research and development. It is also of interest to many financial firms to be able to expose data in a secure and controlled setting to enable researchers conduct analysis on their datasets. We are using multi-party computation mechanisms integrated into Big Data Analytics frameworks to enable analytics that span these different entity owned data sources without exposing privacy of individual datasets.

## 3. Graph Analytics

In addition to systems research in Cloud computing, each year I spend a month in retreat to work on more fundamental computer science problems and to think about their applications. In this section I'll talk about my work on graph analytics.

**Graph Analytics Motivation:** *We need scalable Graph Analytics Algorithms to analyze and process the exponentially-growing online data and the corresponding graphs.* Human and computer generated data is growing at an unprecedented rate and the networks that capture the interactions across these entities are invaluable sources of information that can be used for bettering the society. Efficient analysis of graphs that model such interactions is becoming ever more important.

**Projects:** I worked in Graph modeling and analytics since the beginning of my PhD. I combine my theoretical knowledge in the field with my engineering experience by building tools for graph analytics and utilizing them for solving real-world problems. I built replicated (hyper)graph partitioning and declustering tools, which were published in TPDS [1] and JPDC [5] during my PhD. Then I used these tools to improve the partitioning and data placement of the distributed NoSQL solution Cassandra, which was published in TKDE [6]. I designed efficient parallel PageRank computation and random-walk crawling algorithms, which were published in TPDS [3] and SIGIR [7], respectively. Currently I am working on efficient approximation algorithms for identifying various graph properties. My first work on finding number of triangles in graphs is published in ICDM [18].

Approximating the number of graph structures and identifying local metrics in graphs (collaboration with researchers at DePaul University and The University of Buffalo): In this project I am working on extending my work on triangle counting, which was published in IEEE ICDM [18], to finding the number of other graph structures such as diamonds, butterflies, 5-vertex subgraphs etc. Another problem we are trying to tackle is identifying local metrics (e.g. local clustering coefficient) per vertex/edge in graphs.

## Future Directions

As a future direction, I would like to start new initiatives related to cloud-assisted big data solutions for healthcare, renewable energy, and crypto finance applications. After being involved in the building process of the generic open cloud solution MOC, I would like to lead a research group that builds targeted Cloud solutions for emerging fields. Furthermore, since many of the projects that I was involved in MOC and collaborated with researchers from BU, NEU, Harvard, IBM, TwoSigma, Intel, RedHat, Boston Children's Hospital, and Sandia National Labs have just started to turn into production grade systems and publications, I expect to continue my participation with the upcoming publications around these projects.

**References**

[1] A. Turk, K. Y. Oktay, C. Aykanat, "Query-Log Aware Replicated Declustering", IEEE TPDS. 24(5), pp. 987-995, 2012.

[2] T. Kucukyilmaz, A. Turk, C. Aykanat, "A Parallel Framework for In-Memory Construction of Term-Partitioned Inverted Indexes", The Computer Journal. 55(11), pp: 1317–1330, 2012.

[3] A. Cevahir, C. Aykanat, A. Turk, B. B. Cambazoglu, "Site-Based Partitioning and Repartitioning Techniques for Parallel PageRank Computation", IEEE TPDS, 22(5), pp: 786–802, 2011.

[4] B. B. Cambazoglu, E. Karaca, T. Kucukyilmaz, A. Turk, C. Aykanat, "Architecture of a Grid-Enabled Web Search Engine", Information Processing & Management 43(3), pp: 609–623, 2007.

[5] O. R. Selvitopi, A. Turk, C. Aykanat, "Replicated Partitioning for Undirected Hypergraphs", J. of Parallel and Distributed Computing. 72(4), pp: 547–563, April 2012.

[6] A. Turk, O. R. Selvitopi, C. Aykanat, H. Ferhatosmanoglu, "Temporal Workload-Aware Replicated Partitioning for Social Networks". IEEE TKDE, 2014.

[7] G. B. Tran, A. Turk, B. B. Cambazoglu, "A Random Walk Model for Optimization of Search Impact in Web Frontier Ranking", in SIGIR'15, pp 153–162, 2015.

[8] A. Turk, H. Chen, O. Tuncer, H. Li, Q. Li, O. Krieger, A. K. Coskun, "Seeing into a Public Cloud: Monitoring the Massachusetts Open Cloud", in USENIX CoolDC'16.

[9] M. Zapater, A. Turk, J. M. Moya, J. L. Ayala, A. K. Coskun, "Dynamic Workload and Cooling Management in High-Efficiency Data Centers", in IGSC'15.

[10] BDCOM'16: IEEE BigData'16: "Workshop on Big Data for Cloud Operations Management", http://bdcom16.massopencloud.org/, 2016.

[11] H. Chen, A. Turk, S. S. Duri, C. Isci, A. K. Coskun "Automated system change discovery and management in the cloud". IBM Journal of Research and Development. 60(2-3), pp. 2: 1-2: 10, 2016.

[12] A. Turk, H. Chen, A. Byrne, J. Knollmeyer, S. Duri, C. Isci, and A. K. Coskun "DeltaSherlock: Identifying Changes in the Cloud", in IEEE BigData'16, 2016.

[13] A. Oprea, A. Turk, C. Nita-Rotaru, O. Krieger, "MOSAIC: A Platform for Monitoring and Security Analytics in Public Clouds", in SecDev'16, 2016.

[14] A. Turk, R.S. Gudimetla, E. U. Kaynar, J. Hennessey, S. Tikale, P. Desnoyers, O. Krieger, "An experiment on Bare-Metal BigData Provisioning", in HotCloud'16.

[15] J. Hennessey, S. Tikale, A. Turk, E. Kaynar, C. Hill, P. Desnoyers, O. Krieger, "HIL: Designing an Exokernel for the Data Center", in SoCC'16, pp. 155–168, 2016

[16] R. Pienaar, A. Turk, J. Bernal-Rusiel, N. Rannou, D. Haehn, P. E. Grant and O. Krieger, CHIPS – A Service for Collecting, Organizing, Processing, and Sharing Medical Image Data in the Cloud, in VLDB DMAH, 2017.

[17] O. Tuncer, E. Ates, Y. Zhang, A. Turk, J. Brandt, V. J. Leung, M. Egele, and A. K. Coskun, Diagnosing Performance Variations in HPC Applications Using Machine Learning, in ISC HPC, 2017. (ISC Gauss Award Winner)

[18] D. Turkoglu, A. Turk, Edge-Based Wedge Sampling to Estimate Triangle Counts in Very Large Graphs, in IEEE ICDM, 2017. (Best paper runner up)

[19] A. Turk, M. Varia, and G. Kellaris. Revealing the Unseen: Exposing Cloud Usage While Protecting User Privacy, in IEEE ICDM PAIS, 2017.

[20] T. Tiwari, A. Turk, A. Oprea, K. Olcoz, and A. K. Coskun, User-Profile-Based Analytics for Detecting Cloud Security Breaches, in IEEE BigData PSBD, 2017.

[21] U. Kaynar, M. Abdi, M. H. Hajkazemi, A. Turk, R. Sambasivan, P. Desnoyers, O. Krieger, "D3N: A Datacenter-scale Data Delivery Network". To be submitted to Usenix ATC 2018.

[22] A. Mohan, A. Turk, R. S. Gudimetla, S. Tikale, J. Hennesey, U. Kaynar, G. Cooperman, P. Desnoyers, and O. Krieger, "M2: Malleable Metal as a Service", in IEEE IC2E 2018.