Research Statement

Dippy Aggarwal
https://www.linkedin.com/in/dippyaggarwal/

My primary research interests span the areas of **database systems**, **performance benchmarking**, and **database architectures**. My Ph.D. dissertation was focused on **graph databases and relational data warehouses** and **my work at Intel** as a software development engineer focuses primarily on **performance benchmarking** of a leading database solution and machine learning using industry standard benchmarks. I have had **four collaborations in the form of my internships** during my doctoral studies, of which two were essentially research based. There is this unique merit of receiving fresh ideas, different perspectives and hands-on experience that comes from the combination of experiences coming from research and my current job at Intel.

Ph.D. Research

Graph-based solutions are receiving significant attention recently for two reasons: (1) their ability to capture relationships as first class elements in interconnected domains, and (2) their inherent resemblance to real-world scenarios such as social networks. In my work, I developed graph-based solutions coupled with provenance to offer new approaches for addressing two schema management issues: (a) impact assessment of schema evolution in a data warehouse environment, and (b) schema mapping and integration. We leverage the explicit capture of relationships to address these challenges.

Schema evolution: Schemas play a central role in structured data management. Defining a schema specifies the contents of a database using a specific format or model by defining the structure and relationships of the data stored in it, and it provides a basis for expressing queries. Data warehouses are a schema-rich, multi-layered environment consisting of many inter-related artifacts. If a user seeks to make a schema change at any level in the architecture, he or she may not be aware of the other artifacts potentially impacted by the change. We focus on defining and implementing a graph-based model for impact assessment and explanation. Impact assessment involves identification of the artifacts that depend on the evolved artifact either directly or transitively. The consequences of the change are revealed before actually propagating the change. Our work also allows changes to all schema artifacts in a multi-layered data warehouse architecture, thus addressing multiple evolution problems under one framework. The current contributions are restricted in that they do not address changes to all schema components in the warehouse architecture. Furthermore, we leverage provenance to facilitate user's understanding

for the identified impact. Along with presenting a list of artifacts that will be potentially impacted by the change, we provide a complete trace of how the evolved artifact and the impacted artifacts are related to each other.

Schema mapping and integration: In the context of the second domain of interest (schema mapping and integration) of our work, we describe a system that supports schema integration based on graph databases. Our work first looks at leveraging a graph-based solution for schema mapping. Specifically, we illustrate how schemas expressed in relational and RDF models can be transformed to a property graph to provide an information-preserving, NoSQL-compliant, standardization model for schemas expressed in heterogeneous models. Then we further extend the work by contributing a schema merging algorithm for property graphs. We consider some concrete examples from the literature to highlight how our framework supports integration over property graphs. We illustrate a modular framework that can be further extended and optimized to incorporate different schema mapping and merging algorithms. I **presented this work at a premier IEEE conference in 2015 and an extended version of the work was later also accepted as a book chapter in the Journal of Advanced Systems.**

Future applications of the work: The merit of the work comes from the fact that it is based on graphs and that makes it extensible for application to additional data models beyond the ones considered in my existing work. The methodology can be extended to any interconnected domain consisting of heterogeneous artifacts. For example, impact assessment of schema changes to the query optimizer in databases thus identifying any performance bottlenecks. The other interesting and novel part of my research comes from the idea of provenance. In addition to reporting the impact, we augment the result by offering the complete path trace that connects the evolved and impacted artifacts, thus facilitating warehouse administrator's understanding of the result.

On the work related to schema mapping and integration, my proposed approach lays a foundation for addressing the variety aspect of big data and bringing traditional data into a big data environment. Our approach offers a framework that can be further optimized and it is extensible to incorporate additional data stores. The idea of reverse engineering the graph model to obtain the schemas in the original models can also be useful to leverage tools from the native data environments.

Schema interoperability: Schema interoperability refers to the task of defining mappings between the schema elements for the purposes of facilitating conversion and exchange of metadata expressed in two different models. Recently, the advent of big data and NoSQL data stores has led to the proliferation of data models in popular usage. In the initial exploration towards my current research, I was investigating a framework for enabling mapping, exchange and integration of data represented using different formats and representation schemes. To this end, I developed a graph-based solution that allows a homogeneous, canonical representation of

all the heterogeneous models, thus bringing them to a common format. The decades of research on schema interoperability and integration and graphs can further be employed on this common graph format. In this context, I believe my work holds strong intellectual merit by not only offering a novel graph-based solution and thus advancing the field of science, but also developing an approach which can be integrated with the existing decades of scholarly research on database interoperability.

Other research collaborations

In addition to my dissertation, I have had several collaborations including both research and engineering roles and a brief description of each of them is as follows. **Two of the research areas focused on problems related to green technology.**

(1) **My current role at Intel** involves collaborating with Microsoft and we work towards performance benchmarking of leading enterprise products using industry defined database workloads. The goal is to identify performance optimizations for software so that it runs best on Intel platforms. **My current work exposes me to not only challenges in database systems but also how it relates to the underlying hardware.**

(2) One of those research collaborations came from a summer internship I did in 2011 at **Insight Centre for Data Analytics** (formerly, DERI), Ireland. My research advisor was **Dr. Edward Curry.** The project focused on the theme of **green and sustainable IT.** Our project addressed an organization's information management challenge of staying competitive in the green marketplace while remaining economically competitive. Our solution captures and visualizes information concerning the power consumption of IT devices in real time using various power-metering approaches and leveraging semantic web technologies. To this end, we developed a user interface to depict power consumption of devices in real-time and to demonstrate the potential of leveraging semantic web technologies for addressing data integration challenges. It **won the best graduate level research project award at Kentucky Celebration of Women in Computing Conference, garnered media attention, and was also presented at ACM SIGMOD workshop in 2014.**

(3) The second project in the area of green technology came from a research assistantship I had during my masters program at West Chester University, Pennsylvania (Dr. Afrand Agah). The project was on **wireless sensor networks** and involved the principles of game theory and wireless sensors. **The goal of this project was to maximize the lifetime of battery-powered wireless sensors which once deployed, are expected to run autonomously and with minimum human attendance**. In this project we investigated how various interactions in a wireless sensor network can be modeled as a

game theory framework to motivate them with a new incentive that yields equilibrium for all sensor nodes.

(4) Another project I did was related to **developing weather forecasting models for better, accurate short term forecasts**. This was a part of summer on campus work I did with a professor (Dr. Joby Hilliker) in the department of astronomy and geology. Our approach gave encouraging results for lead times of less than 6 hours. **The work was presented at the Graduate Poster Competition at West Chester University and won the best research project award.**

(5) My work at **Teradata Labs, California** focused on database systems and we looked at improving the next release of Teradata's database optimizer.

(6) Unique when compared to the projects above, I am passionate about research in computer science education. In the past, I have submitted a tutorial which was accepted for publication in Teradata University and as a result of that I earned attendance at one of the premier computer science education conferences - ACM SIGCSE in 2016. Along with my PhD advisor, I also published another education paper proposing an idea for **scaling data warehousing projects.**

Future Work

I plan to continue my research along following areas: (1) computer science education for proposing new courses/pedagogical ideas to enhance the current curriculum, (2) database benchmarking (3) investigating interoperability and information exchange between SQL and NoSQL databases, (4) query optimizations.