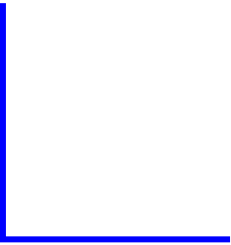




Research Topics and Review

Gregory S. DeLozier, Ph.D.
Kent State University
Nov 21, 2017



What Are Research Topics?

- * Active areas of investigation
- * Commercial opportunities
- * Areas of shortcomings in current systems
- * Capabilities that were once unfeasible

Beckman Report

Beckman Database Research Meeting

- * Periodic academic meeting
- * 8th meeting in 2013
- * Recommended topic
 - scalable big/fast infrastructure
 - diversity in data management
 - end-to-end processing of data
 - cloud services
 - managing roles of people in data life cycle
- * <https://beckman.cs.wisc.edu/>
- * <https://beckman.cs.wisc.edu/beckman-report2013.pdf>

Beckman - Big Data Trends

- * Cheap to generate data
 - Sensors
 - Social media
 - MM activities
 - IoT
- * Cheaper to process data - clouds, cores
- * Democratization of data
 - Available to many kinds of people/roles
 - Processing with many kinds of tools and contexts

Big Data Management

- * Many non-DBMS players arrived
 - e.g. Hadoop/MR
- * Movement toward DBMS principles
 - e.g. Hive
- * Volume, velocity -- known to DBMS community
- * Variety is a new challenge
- * How does the DBMS community contribute?

Scalable Fast/Big Infrastructure

- Processing size
- Parallelism
- Distribution
- Speed (capture _and_ retrieval)
- New storage paradigms
- Machine learning, sampling, and aggregation in queries

Diversity in Data Management

- * Integration and federation of diverse systems
- * No common tools or magic bullet
- * Variety of tools and architectures
- * Dataflows
 - raw (sql) -> cooked (R?)

End-to-end processing

- * From collection to extracted knowledge
 - * Few tools exist
 - * Intervention at every step is required
 - * Data 'pipeline'
-
- * Community should build effective and practical tools
 - * Exploit domain knowledge
 - * Use data knowledge bases to help understanding

Cloud Services

- * Can we offer database as a PaaS?
- * Google Big Query, Amazon Redshift, Azure SQL
- * Elasticity (change in technology as data grows?)
- * Data replication
 - Multiple copies across data centers
 - Trades with safety vs speed
- * Administration
- * Data sharing

Roles of Humans

- * Programmers
-
- * Data Scientists
- * Producers (rights? from participation to wearables)

- * Curators
- * Consumers
- * Stewards

What do all these people do, and how do they fit in?

The Red Book

The UC Berkeley "Red Book"

- * Collection of important database readings
- * It's an actual book
 - <http://www.amazon.com/Readings-Database-Systems-Joseph-Hellerstein/dp/0262693143>
- * You can get most of it on the web
 - <http://redbook.cs.berkeley.edu/bib4.html>
- * Let's go check that out...



Classic Papers



Some papers to look at...

GFS:

<http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf>

Big Table:

<http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf>

Dynamo:

<http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf>

Object Storage

Amazon S3 Object Storage

- * Amazon S3 Persistent Object Storage

- Get an account
- Create Buckets
- Put K/V pairs in Buckets

- * <http://www.smalldatajournalism.com/projects/one-offs/using-amazon-s3/>

Amazon S3 Access

- * Boto Tutorial

- * <http://boto3.readthedocs.io/en/latest/guide/index.html>

Amazon S3 As a Database

- * Not without its challenges...
- * <http://cs.brown.edu/~kraskat/pub/sigmod08-s3.pdf>