Old problems for Big Data + Cloud

Surajit Chaudhuri
Microsoft Research

Make Data Birth to Consumption Pipeline Friction-free

- *Ad-hoc (on-demand) search for structured data* (but not necessarily schematized)
  - New Search Model/API, Ranking
  - Fast ingestion, Indexing
- Data Profiling, Data Cleaning/Matching
  - Aided by contextual information
- Data Store to Visualization (perhaps via OLAP)
  - Make the pathway simple
- Interactive performance needed

"Auto" Cloud Big Data Platforms (aka DaaS)

- **Pure PaaS:** Define schema (optional), ingest data, query
  - Every other element of the platform should be in "auto" mode
  - Must be a win-win for provider and the users
- Questions
  - What is the right query API?
  - What are the **performance guarantees in the multi-tenant platform**?
  - How to manage hot to cold to colder data automatically?
  - How to provide **auto tuning** for cloud big data platforms by leveraging observability?