The State of the Art in Supporting "Big Data"

by

Michael Stonebraker

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

What is “Big Data”

- Too much Volume (I have too much data)
- Too much Velocity (Its coming at me too fast)
- Too much Variety (Its coming at me from too many places in too many formats)

Too Much Data -- The Data Warehouse World (structured data)

- Mature (and large) commercial market with several well-regarded vendors
- I know of a couple dozen of these in production use on petabytes of data
  - E.g. Zynga, E-Bay
  - That is about 20 Mbytes for every person in the US!
- No reason why this technology won't scale as customers want larger installations
  - Expect data warehouses to get larger by 1-2 orders of magnitude over the rest of this decade.

Too Much Data -- The Hadoop/Hive World (semi-structured data)

- I know over another 20 or so petascale Hadoop installations
  - E.g. Facebook
- No reason this technology won’t continue to scale
- And probably converge with the data warehouse world

Too Much Data -- The Data Scientist World

- Predictive Modelling, data mining, data clustering, recommendation engines, ….
- Complex analytics – not in SQL
- Not well understood
  - World of research, start-ups, …
- My prediction:
  - As the world moves from simple analytics to complex analytics, the server side technology will mature to meet the need

Too Fast

- Often a legacy problem
  - Rise in stock market volume breaking the legacy real-time infrastructure of investment banks
- Usually solvable by throwing money/hardware at the problem
- Usually amenable to aggregation in the sensor network to knock down the velocity
  - E.g. car insurance sensors

Too Fast

- Some problems yet to be solved (query languages, integration of storage with “on-the-wire processing)
  - But I see no showstoppers here
- Technology is capable of handling “the firehose” that will result from “the internet of things”

Too Many Places

- Mature technology for integrating 20 data sources
  - Extract-Transform and Load (ETL) vendors
- But how to integrate 10,000?
  - Novartis has 10,000 bench chemists and biologists, each with an (independently constructed) data set of experimental results
  - Company wants to integrate these 10,000 data sources
  - And add additional ones from the public web

Too Many Places

- Research problem!
  - Killing most CIO’s that I know
- Very active area of investigation
- Startups in this space

- If there is any achilles heel in big data, this is it!

DBMS Security

- Works well
  - i.e. I have never heard of the DBMS screwing up in this area.

Encryption

- Can be entrusted to the DBMS
  - Appropriate when there are many clients sharing data
  - Don’t want the encryption key to be on 500 desktops
- Can be entrusted to the client
  - Appropriate when there is personal (single user) data
  - See Nickolai’s talk this afternoon

Leaks

- Usually insiders (think Edward Snowdon)
- Or unguarded desktops (my password on a post-it note on my PC)
- No possible way for the DBMS to prevent this

However

- DBMS can write a “command log” (everything everybody did)
  - Enables after-the-fact auditing
  - Sniff the log for suspicious behavior (unusual activity)
  - Would be a nice DBMS add-on
- But it is a human management problem to actually use it!!!