Data Analytics

Themis Palpanas
Paris Descartes University

LIPADE
Laboratoire d'Informatique Paris Descartes

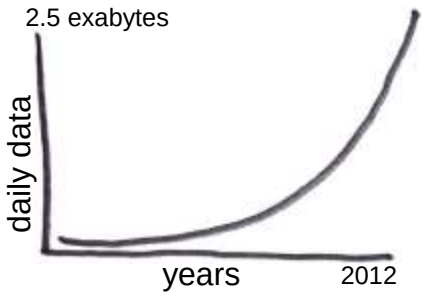
big data

2

Big Data

- large scale data
- streaming data
- heterogeneous data
- private data
- uncertain data

confluence of all the above!



[IBMbigdata]

Every two days we create as much data as much we did from dawn of humanity to 2003

[Eric Schmidt, Google]

big data V's
(it is not about size only)

volume | velocity | variety | veracity

5

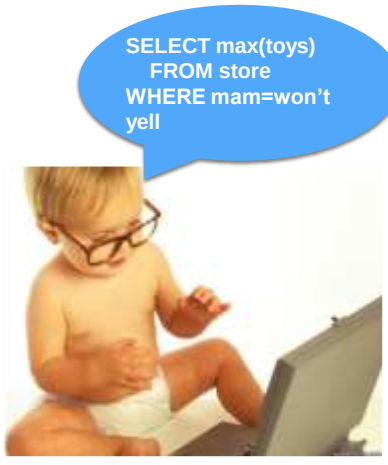
there are good chances we already have the data for the next big breakthroughs in say biology, medicine, etc. but we simply cannot extract the knowledge

6

today

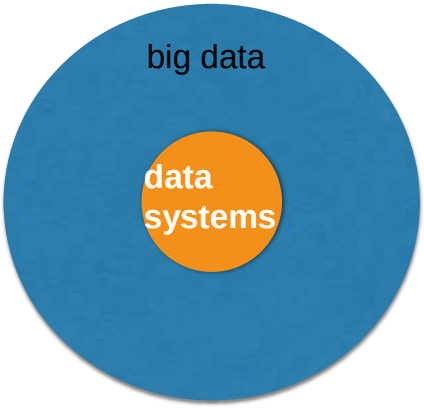
tomorrow

soon everyone will need to be a “data scientist”

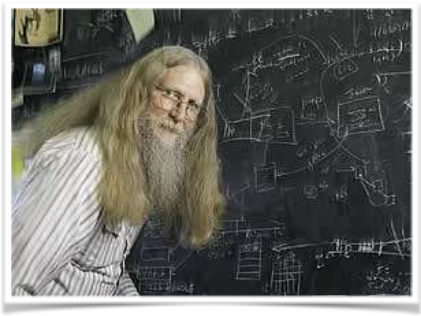


big data **systems**

data systems are in the middle of all this



9



"relational databases are the foundation of western civilization"

Bruce Lindsay, IBM
ACM SIGMOD Edgar F. Codd Inovations award 2012

10

dbs are everywhere…

facebook

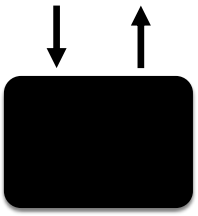
Google

1
1

5 decades of research

IBM, Microsoft, Oracle, Teradata,etc.
and a gazillion start-ups today

declarative interface
ask ''what'' you want

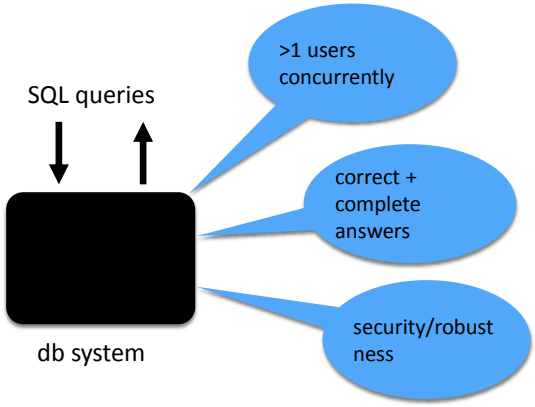


db system

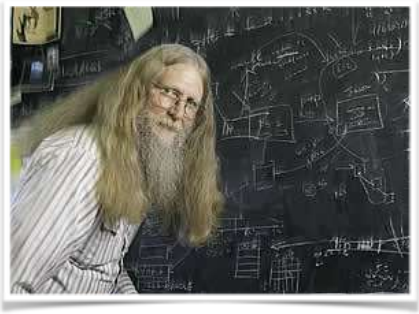
the system decides "how" to best store and access data

why is this good

1
2



13

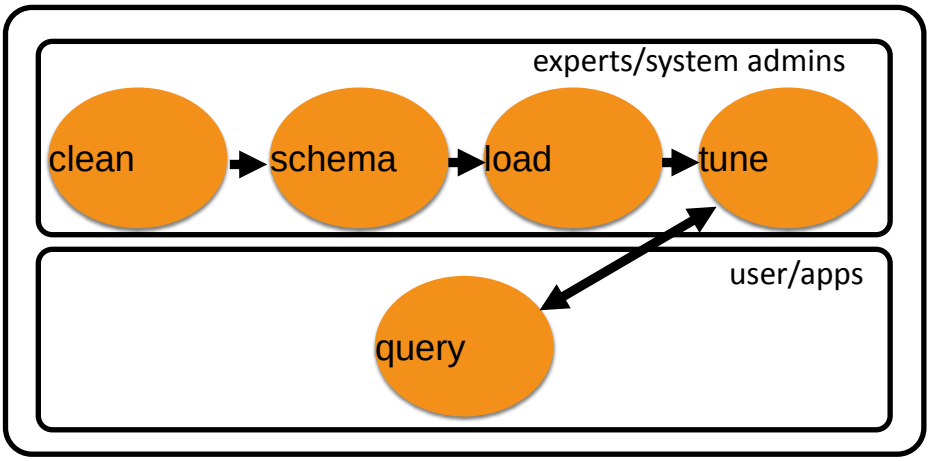


"Three things are important in the database world: ***performance***, ***performance***, and ***performance***"

Bruce Lindsay, IBM
ACM SIGMOD Edgar F. Codd Inovations award 2012

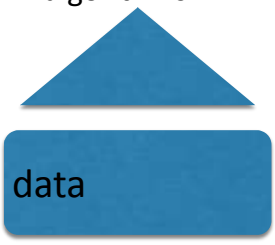
14

essential steps in using a database system



15

data systems architectures



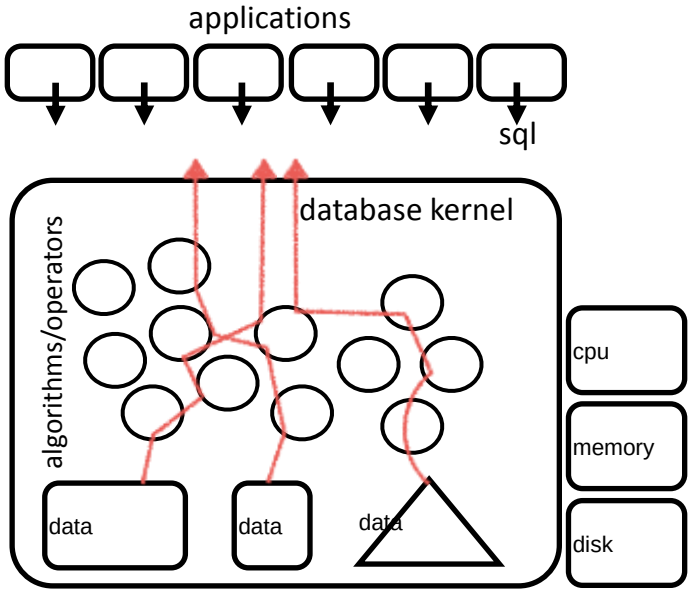
some problems:
how to **store** data
how to **access** data

how to best answer a **complex** query
(e.g., which data to access first and how)

how to answer millions of queries **concurrently**

how to guarantee **correctness** and **availability**

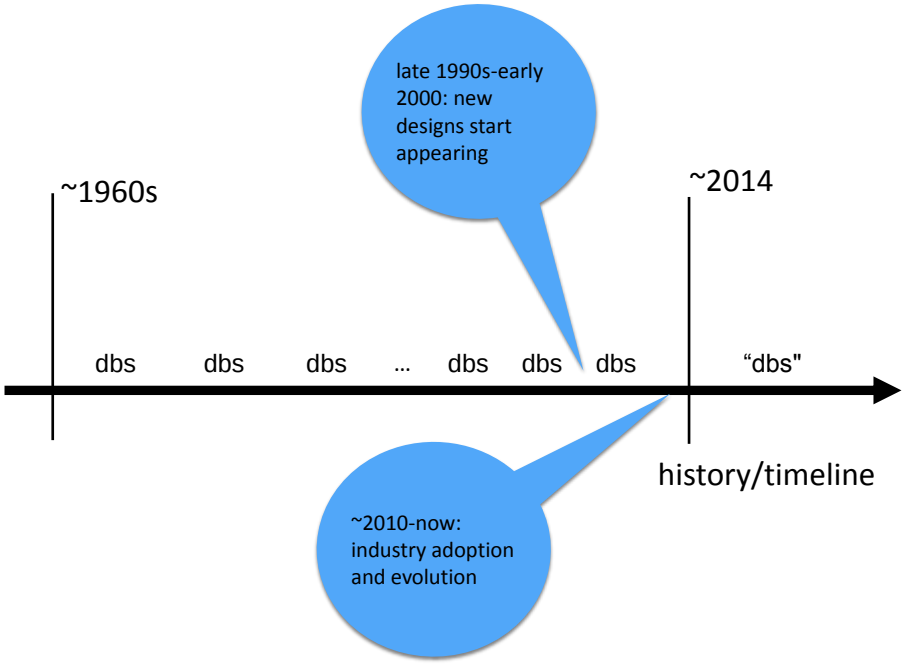
16



17

scale up vs scale out

18



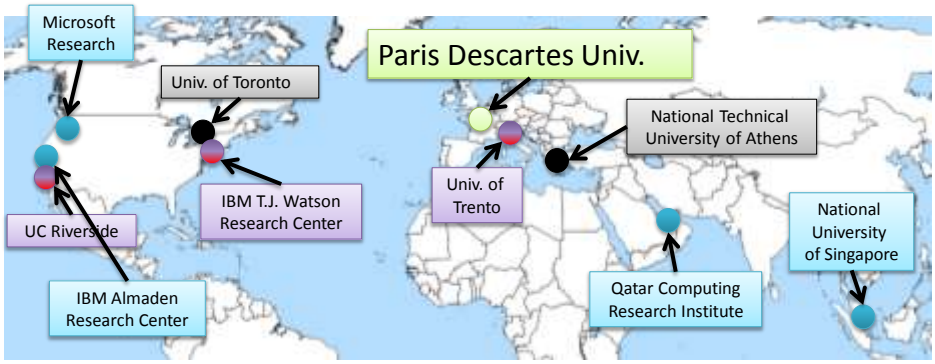
19

data systems design (and research) is kind of an art

20

who is who

Background and Experience



- founder and director of database group of University of Trento
- director of database group of Paris Descartes University
- awards: IBM SUR Award, 3 Best Paper Awards
- 8 US patents:3 implemented in commercial products (IBM DB2, IBM MDBT)
- **expertise in data bases, data mining, data sequences**

Our Work

- large scale data
- streaming data
- heterogeneous data
- private data
- uncertain data

Our Work

- large scale data
  - Managing and Analyzing Very Large Scientific Data
    - work with Harvard University (USA)
- streaming data
- heterogeneous data
- private data
- uncertain data

Our Work

- large scale data
- streaming data
  - Identifying Important Events in Fast Data Streams
    - work with AT&T Research Labs (USA)
  - Streaming Analytics for Green Manufacturing
    - work with Intel (Ireland), Volvo (Sweden), SAP (Germany)
  - Real-Time Monitoring of Human Behavior Patterns
    - work with IBM Research (Ireland), Telecom Italia (Italy)
- heterogeneous data
- private data
- uncertain data

Our Work

- large scale data
- streaming data
- heterogeneous data
  - Sentiment Analytics on Large Text Collections
    - work with Hewlett-Packard Labs (USA)
  - Entity Resolution in Web Data
    - work with L3S Institute (Germany)
  - Fraudulent Activities Detection
    - work with Legitscript (USA), Vodafone (Italy)
- private data
- uncertain data

Our Work

- large scale data
- streaming data
- heterogeneous data
- private data
- uncertain data
  - Processing and Mining Uncertain Data
    - work with IBM T.J. Watson Research Center (USA)