CPS 516: Data-intensive Computing Systems

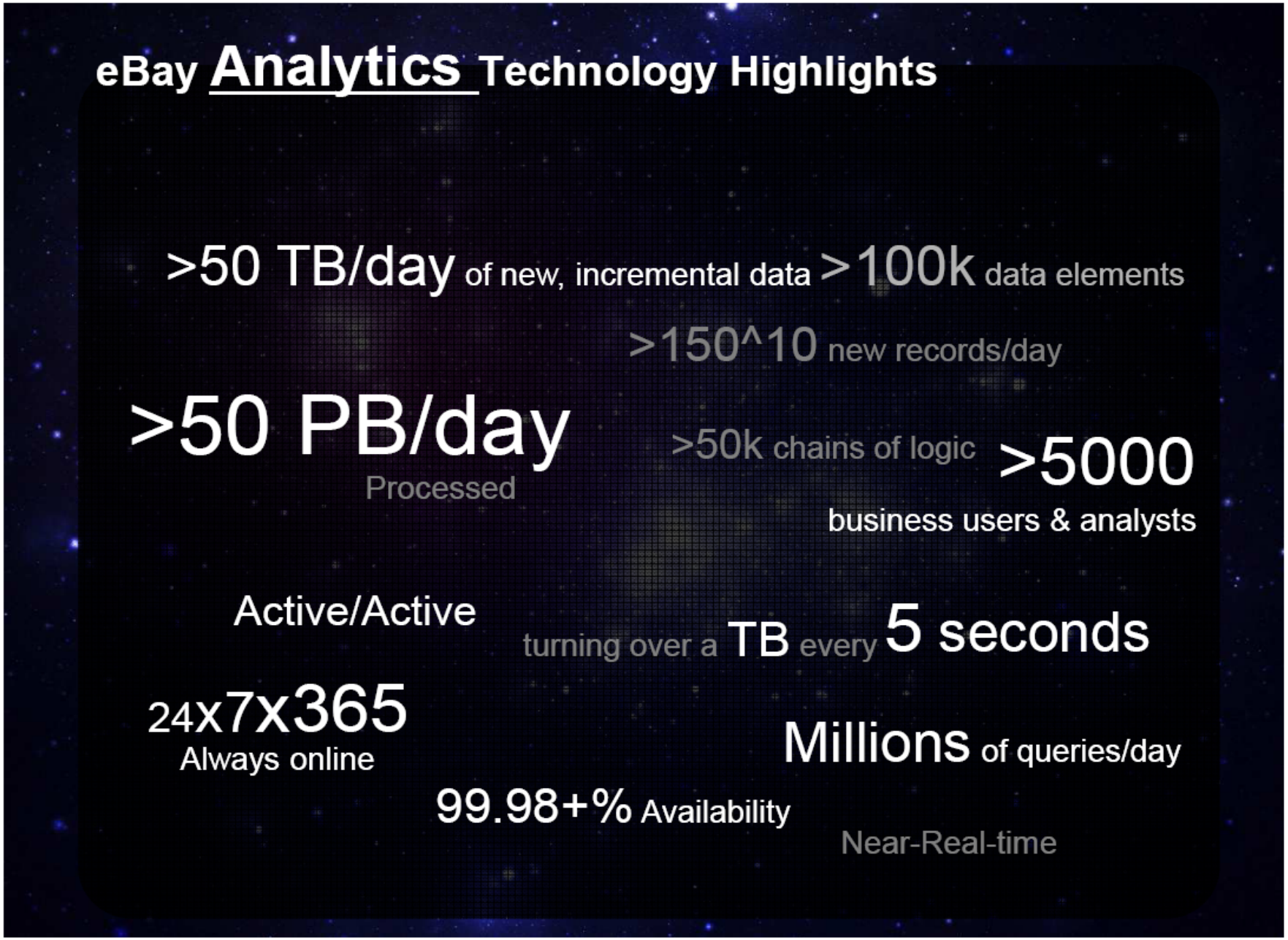
Instructor: Shivnath Babu

TA: Zilong (Eric) Tan

The World of Big Data

- eBay had 6.5 PB of user data + 50 TB/day in 2009

From http://www.umiacs.umd.edu/~jimmylin/



From: http://www.cs.duke.edu/smdb10/

The World of Big Data

- eBay had 6.5 PB of user data + 50 TB/day in 2009

How much do they have now?

See http://en.wikipedia.org/wiki/Big_data

Also see: http://wikibon.org/blog/big-data-statistics/

From http://www.umiacs.umd.edu/~jimmylin/

FOX AUDIENCE NETWORK

- Greenplum parallel DB
  - 42 Sun X4500s ("Thumper") *each* with:
    - 48 500GB drives
    - 16GB RAM
    - 2 dual-core Opterons
- Big and growing
  - 200 TB data (mirrored)
  - Fact table of 1.5 trillion rows
  - Growing 5TB per day
    - 4-7 Billion rows per day

- Also extensive use of R and Hadoop

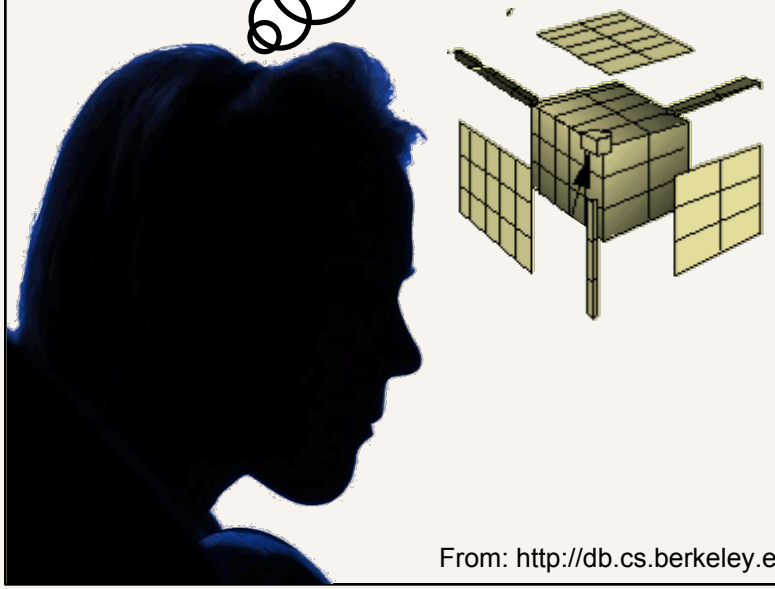
Yahoo! runs a 4000 node Hadoop cluster (probably the largest). Overall, there are 38,000 nodes running Hadoop at Yahoo!

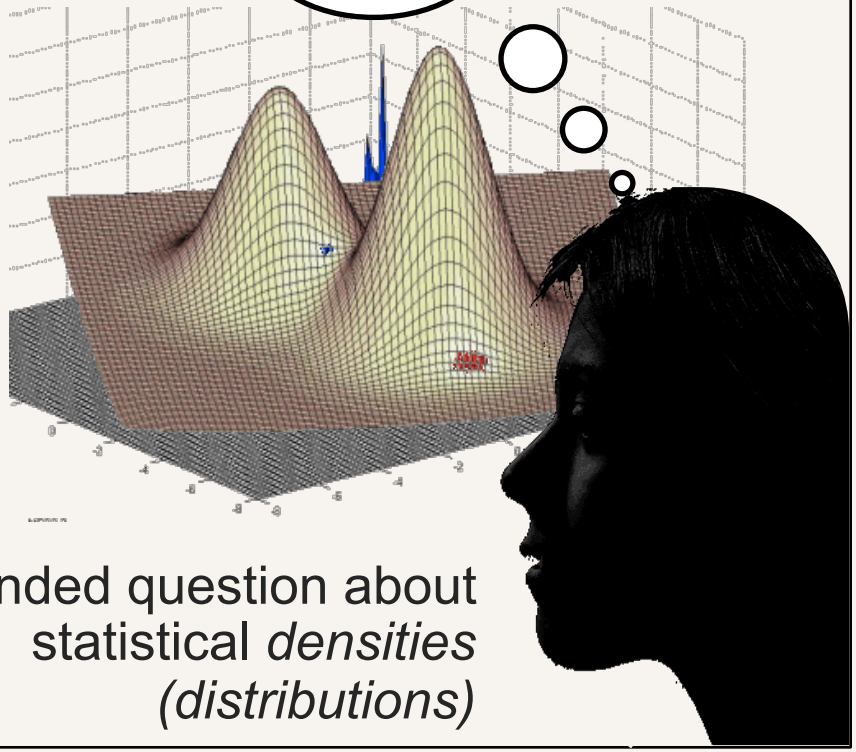
From: http://db.cs.berkeley.edu/jmh/

As reported by FAN, Feb, 2009

How many female WWF fans under the age of 30 visited the Toyota community over the last 4 days and saw a Class A ad?

How are these people similar to those that visited Nissan?

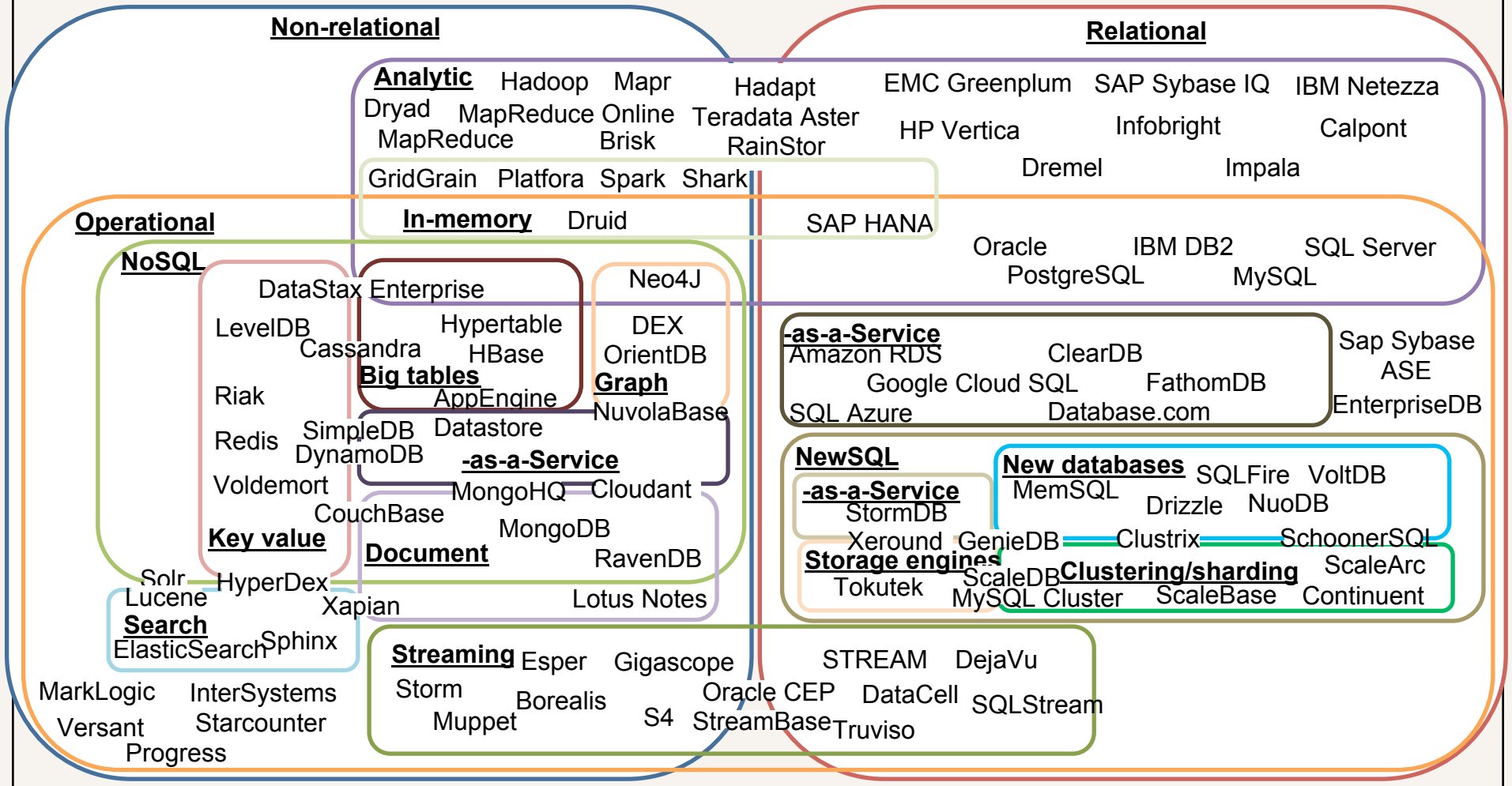




Open-ended question about statistical *densities (distributions)*

From: http://db.cs.berkeley.edu/jmh/

“No One Size Fits All” Philosophy



An extension of the figure given in http://blogs.the451group.com/information_management/2012/11/02/updated-database-landscape-graphic

What we will cover in class

- Scalable data processing
  - Parallel query plans and operators
  - Systems based on MapReduce
  - Scalable key-value stores
  - Processing rapid, high-speed data streams
- Principles of query processing
  - Indexes
  - Query execution plans and operators
  - Query optimization
- Data storage
  - Databases Vs. Filesystems (Google/Hadoop Distributed FileSystem)
  - Data layouts (row-stores, column-stores, partitioning, compression)
- Concurrency control and fault tolerance/recovery
  - Consistency models for data (ACID, BASE, Serializability)
  - Write-ahead logging

Course Logistics

- Web pages: Course home page will be at Duke, and everything else will be on github
- Grading:
  - Three exams: 10 (Feb) + 15 (March) + 25 (April) = 50%
  - Project: 10 (Jan 21) + 10 (Feb 1 – Feb 21) + 10 (Feb 22 – March 10) + 20 (March 11 – April 15) = 50%
- Books:
  - No one single book
  - *Hadoop: The Definitive Guide*, by Tom White
  - *Database Systems: The Complete Book*, by H. Garcia-Molina, J. D. Ullman, and J. Widom

Project Part 0: Due in 2 Weeks

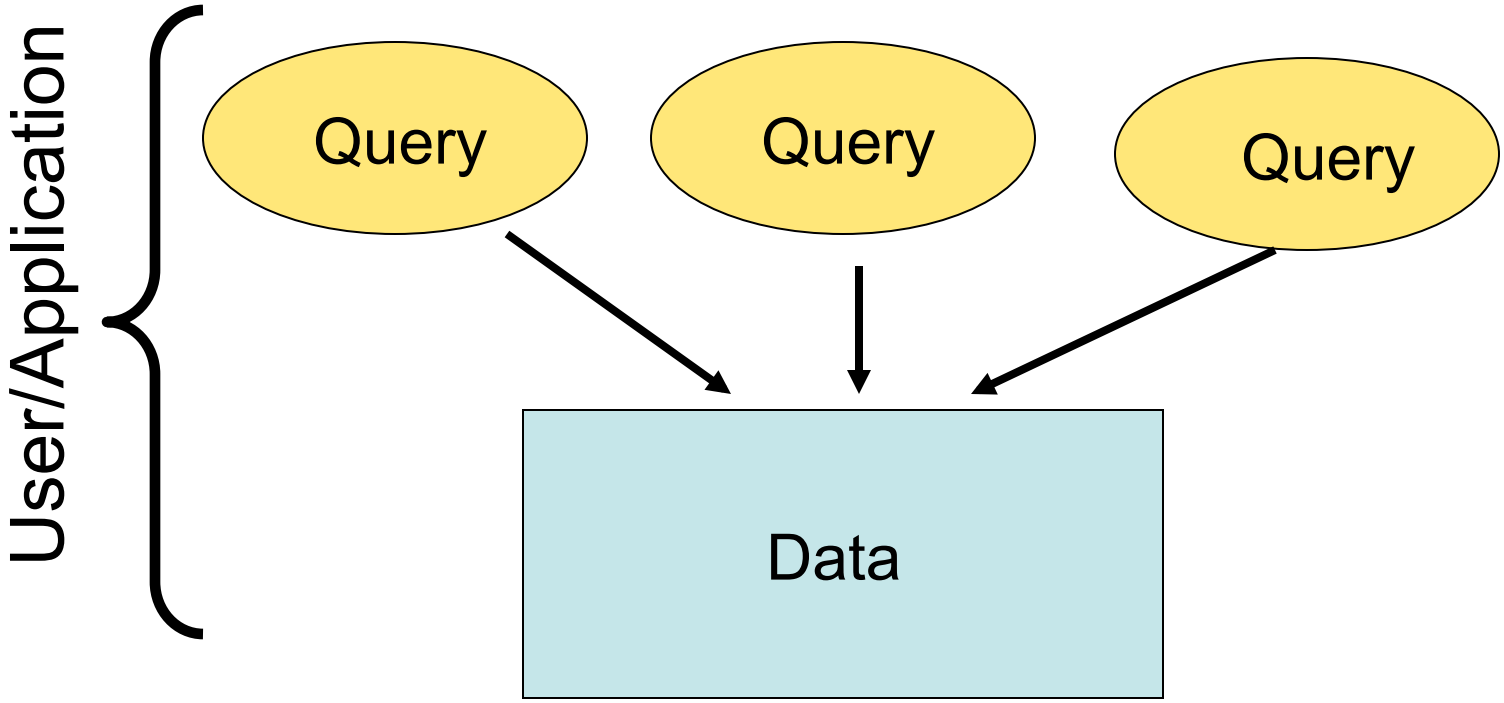
- For every single system listed in the “Data Platforms Map”, give as a list of succinct points:
  - Strengths (with numbered references)
  - Weaknesses (with numbered references)
  - References (can be articles, blog posts, research papers, white papers, your own assessment, …)
- Your own thoughts only. Don’t plagiarize. List every source of help. We will enforce honor code strictly.
- Submit on github (md format) into repository given by Zilong
- Outcomes: (a) Score out of 10; (b) Project leader selection.

Project Parts 1, 2, 3

- Shivnath/Zilong will work with project leaders to assign one system per project. Will also try to have one mentor per project
- Each student will join one project. Project starts Feb 1
- Part 1: Feb 1 – Feb 21
  - Install system
  - Develop an application workload to exercise the system
  - Run workload and give demo and report
- Part 2: Feb 22 – March 15
  - Identify system logs/metrics and other data that will help you understand deeply how the system is running the workload
  - Collect and send the data to a Kafka/MySQL/ElasticSearch routing and storage system set up by Shivnath/Zilong. Give demo and report
- Part 3: March 16 to April 15
  - Analyze and visualize the data to bring out some nontrivial aspects of the system related to what we learn in class. Give demo and report

Primer on DBMS and SQL

Data Management



DataBase Management System (DBMS)

Example: At a Company

Query 1: Is there an employee named “Nemo”?
Query 2: What is “Nemo’s” salary?
Query 3: How many departments are there in the company?
Query 4: What is the name of “Nemo’s” department?
Query 5: How many employees are there in the “Accounts” department?

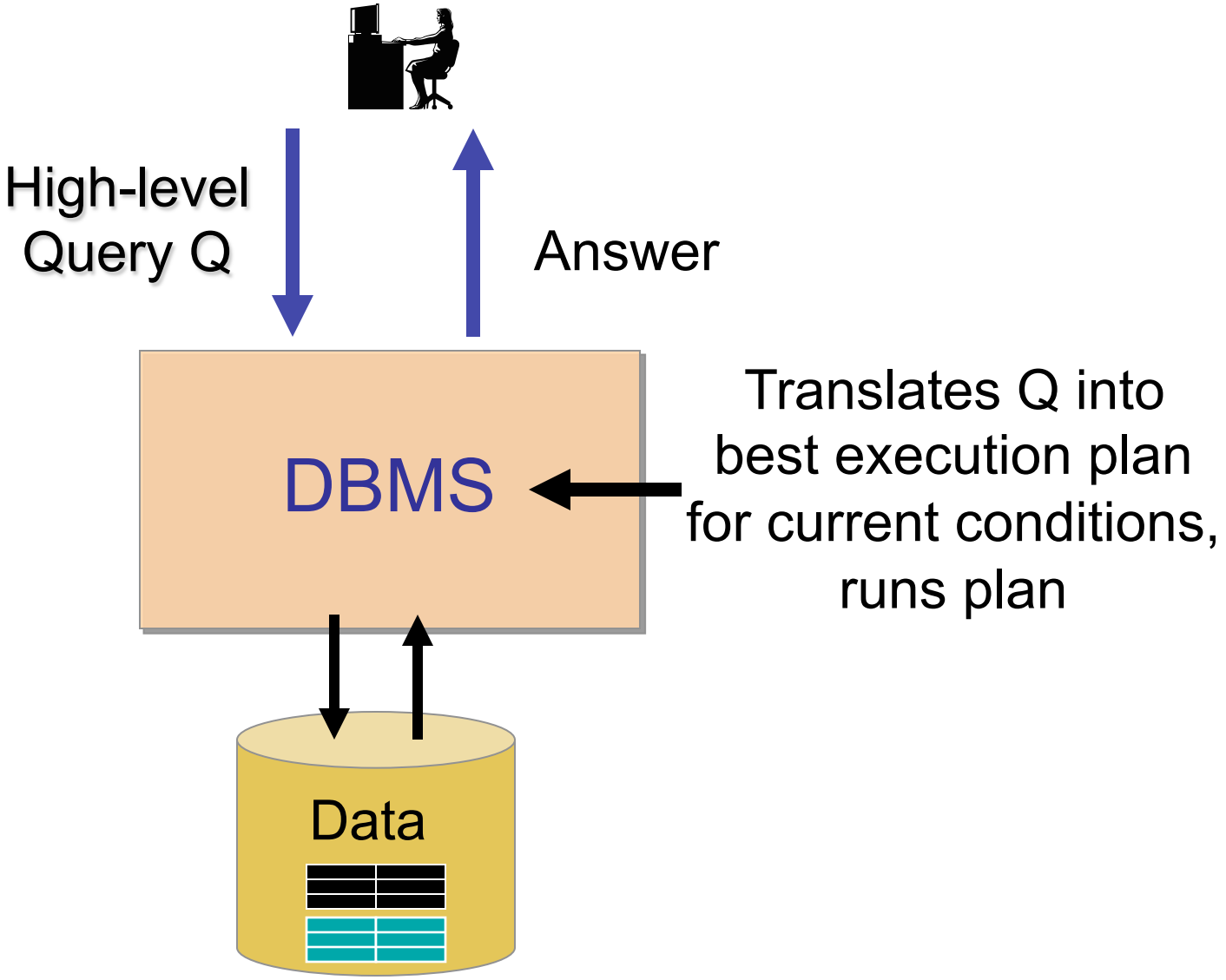
Employee

ID	Name	DeptID	Salary	…
10	Nemo	12	**120K**	…
20	Dory	156	**79K**	…
40	Gill	89	**76K**	…
52	Ray	34	**85K**	…
…	…	…	…	…

Department

ID	Name	…
12	IT	…
34	Accounts	…
89	HR	…
156	Marketing	…
…	…	…

DataBase Management System (DBMS)



Example: Store that Sells Cars

Owners of Honda Accords who are <= 23 years old

Make	Model	OwnerID	ID	Name	Age
Honda	Accord	12	12	Nemo	**22**
Honda	Accord	156	156	Dory	**21**

Join (Cars.OwnerID = Owners.ID)

Filter (Make = Honda and Model = Accord)

Filter (Age <= 23)

Cars

Make	Model	OwnerID
Honda	Accord	12
Toyota	Camry	34
Mini	Cooper	89
Honda	Accord	156
…	…	…

Owners

ID	Name	Age
12	Nemo	**22**
34	Ray	**42**
89	Gill	**36**
156	Dory	**21**
…	…	…

DataBase Management System (DBMS)

High-level Query Q

Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

Data