Databases in Astronomy

@NOAODataLab

What is a database?

Database: organized collection of data.

Database software: application specifically designed around handling data.

Database Management System (DBMS): Also provides other things such as a privilege system, instrumentation for performance monitoring, multi-user support, an API, etc.

Relational Database: Database that organizes data into tables that "relate" to each other through keys, motivated by prior popular methods of network and hierarchical requiring rebuilding as the model changed

Physical limitations of data access

	Nanoseconds	in ms	3 Seconds
execute typical instruction	1	0.000001	3 million instruction
fetch from L1 cache memory	0.5	0.0000005	6 million fetches
branch misprediction	5	0.000005	600 million mispredictions
fetch from L2 cache memory	7	0.000007	428.5 million fetches
Mutex lock/unlock	25	0.000025	120 Million locks
fetch from main memory	100	0.0001	30 million fetches
send 2K bytes over 1Gbps network	20,000	0.02	150 thousand 2k bytes
read 1MB sequentially from memory	250,000	0.25	12000 MB
fetch from new disk location (seek)	8,000,000	8	375 seeks
read 1MB sequentially from disk	20,000,000	20	150 MB
send packet US to Europe and back	150,000,000	150	20 packets

ref: http://norvig.com/21-days.html#answers, Peter Norvig, Director of Research at Google

Given an object table with 200 columns and 4 billion rows, here are theoretical performance times only for disk in a row store database

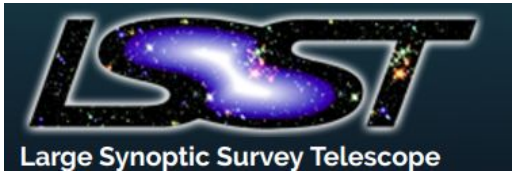
	Basic Hard Disk		Medium class SSD		Enterprise SSD	
	Metric	**Unit**	**Metric**	**Unit**	**Metric**	**Unit**
Disk Read	120	mb/s	500	mb/s	2800	mb/s
Full table scan	51657	s	12398	s	2214	s
Total Time	**861**	**min**	**207**	**min**	**37**	**min**
Rows in 3 s	**232,300**	**rows**	**967,916**	**rows**	**5,420,331**	**rows**
Rows in 10 s	**774,333**	**rows**	**3,226,388**	**rows**	**18,067,771**	**rows**
HDs needed to scan all rows in 3s	**17,219**	**drives**	**4,133**	**drives**	**738**	**drives**
HDs needed for All rows in 1 min	**861**	**drives**	**207**	**drives**	**37**	**drives**

see benchmark charts for specific drives; use the minimum benchmarks:

http://www.tomshardware.com/charts/enterprise-hdd-charts/-03-Read-Throughput-Minimum-h2benchw-3.16,3374.html

NOAO DATALAB

Examples of databases in astronomy



Large Synoptic Survey Telescope (LSST): 100s of Petabytes total. Catalog on QSERV > 30 TB, 55 billion rows as of 2011. Distributed, 150 node, SQL Server -based

Sloan Digital Sky Survey (SDSS): DR 12 catalogs were 9.1 TB and there are now 14 DRs

NOAO DATALAB

Files vs databases

Different definitions exist for a **database**:

"a usually large collection of data organized especially for rapid search and retrieval (as by a computer)" -- Merriam Webster

"a large amount of information stored in a computer system in such a way that it can be easily looked at or changed" -- Cambridge Dictionary

A comma-separated file is not well organized for rapid search and retrieval: To find a record you have to scan the whole file sequentially

Database technologies relevant for astronomy

PostgreSQL

MariaDB

MySQL

Database Management Systems (DBMS): software to administer, query, secure, backup and recover databases

Microsoft SQL Server

NOAO DATALAB

Querying a database

Structured Query Language (**SQL**)

Count objects classified morphologically as "simple" galaxies by Tractor in legacysurvey DR 4, lying within a radius of 0.1 deg of Ra and Dec. We are using the q3c style spatial indexing to support this query.

```
SELECT COUNT(*)
    FROM  ls_dr4.tractor
  WHERE type = 'SIMP'
        AND
        q3c_radial_query(ra, dec, 66.5493, 68.1717, 0.1);
```

NOAO DATALAB

What is a schema?

Schema in a database context is the definitions of the objects and their attributes. They specify the tables, indexes, views, stored procedures, column names, etc. and their *data types* such as *integer*, *real*, and *text*.

Example data definition for a table to create part of a schema:

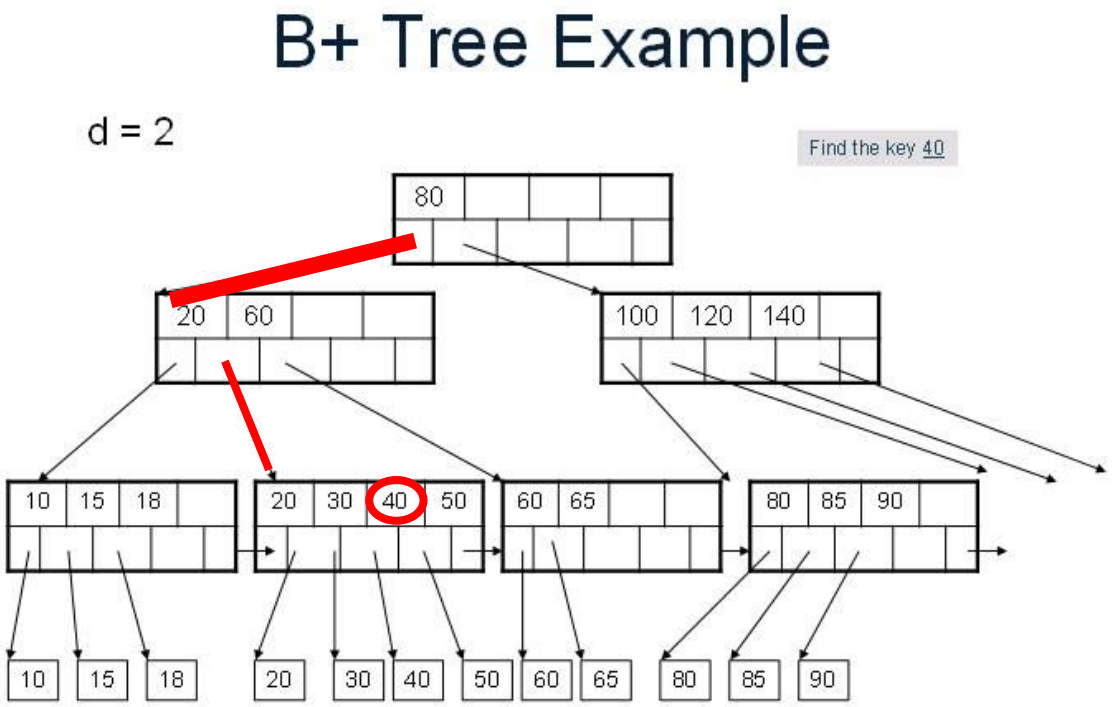
```
CREATE TABLE object (
    id bigint PRIMARY KEY,
     ra double precision,
    dec double precision,
    ...
);
```

* The ellipsis is not part of the statement

There are "schema-less" DBs and DBs that apply the schema on read. Nevertheless, a developer needs to know how things are named, what their range of values should be.

What is an index and how does it help?

An index helps the database find values in a table faster.



To find the key value 40:

Image From: CSE544:University of Washington Computer Science & Engineering Principles of Database Systems

How do you design a schema?

For a relational database, understand the **relational model.**

Identify **entities** and **attributes.**

Create an **Entity Relationship** model/diagram (E-R diagram)

Normalize the model at least to third-normal form (**3NF**)

The **physical design** varies from the **logical design** often to save space or improve performance.

For astronomical data, the physical design is more involved than the logical.

Interactive exercise

Break into 4 groups

You will be assigned a survey

Through discussion, answer some questions to help you imagine how you might design a database to serve your survey's data products

Questions

- What measurements and information will you get from your survey?
- How will you organize the information into database tables?
- What are some questions (queries) that you will ask of the data to do your science?
- Are there data products from your survey that don't fit well in a database?

A Near Earth Object Survey

Purpose: Discover fast-moving Near Earth Objects

Instrument: Dark Energy Camera

Filters: VR

Time: 30 nights

Area covered: 3000 sq. deg.

Exposure time: 40 sec (VR~23)

Number of objects: ~10^8 total, ~1000s moving objects

Cadence: 5 exposures per field separated by 5 min, repeat on two subsequent nights, 525 exposures per night

NOAO DATALAB

A Wide Area Extragalactic Survey

Purpose: Identify and measure ~200 million galaxies and stars

Instrument: Dark Energy Camera

Filters: grz

Time: 64 nights

Area covered: 6200 sq. deg.

Exposure time: ~150 sec (*grz* ~ 24.7, 23.9, 23.0)

Cadence: Three exposures per location per filter, typically separated by months

NOAO DATALAB

A Time Series Survey of the Galactic Bulge

Purpose: Identify and characterize variable stars in the Galactic Bulge

Instrument: Dark Energy Camera

Filters: ugriz

Time: 10 nights

Area covered: ~20 sq. deg.

Exposure time: ~100-300 sec (*ugriz* ~ 22.0)

Number of objects: ~30 million total

Cadence: Repeat every 1.5 hours, ~50 epochs per object per filter

NOAO DATALAB

A Wide Area Spectroscopic Survey

Purpose: Measure redshifts of ~30 million galaxies, characterize spectra of ~20 million bright galaxies and ~10 million Milky Way stars

Instrument: Dark Energy Spectroscopic Instrument

Wavelengths: 380 - 980 nm, R~5500

Time: 5 years

Area covered: 14000 sq. deg.

Exposure time: ~1000 sec