SQL Introduction

http://dilbert.com/strip/1995-11-17

Today's lecture

1. SQL and schema definitions
2. Single-table queries
3. Multi-table queries

DISCLAIMER

SQL was initially developed in the early 1970s

- Do not reinvent the wheel, this course is based on information from:
  - CS145 (2016), Stanford
  - IIS (2009), EPFL
  - INF725, SD202 (2016), Télécom ParisTech
  - SoSe (2005), Freie Universität Berlin
  - Database Management Systems (3rd Edition), Ramakrishnan and Gehrke.
  - Database Systems Concepts (6th Edition), Abraham Silberschatz, Henry F. Korth, and S. Sudarshan.

- Database Systems: A Practical Approach to Design, Implementation and Management (6th Edition), Thomas M. Connolly and Carolyn E. Begg.

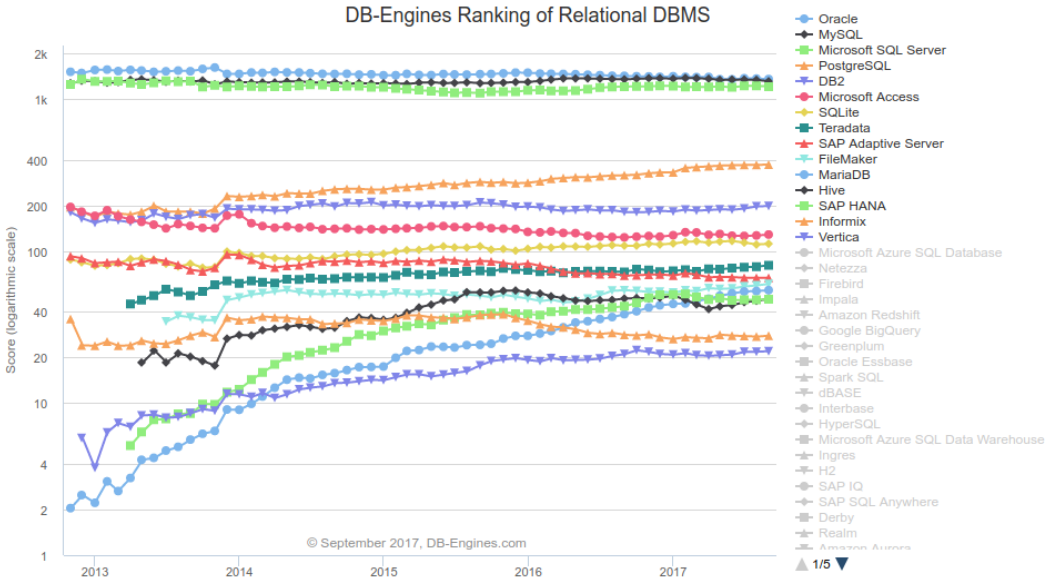
What is SQL?

- **S**tructured **Q**uery **L**anguage
- A standard language for querying and manipulating data
  - Not a programming language!
  - **Very high-level** <-- Highly optimized
- Originally based upon *relational algebra* and *tuple relational calculus*
- Employed as query language for most **R**elational **D**ata**B**ase **M**anagement **S**ystems (RDBMS)

How to pronounce it?

https://www.reddit.com/r/ProgrammerHumor/comments/7z0eoj/how_to_pronounce_sql/

- Many standards and implementations ... *but*
  - Implementations are incompatible between vendors and do not necessarily completely follow standards



SQL consists of

Data *Definition* Language (DDL)

- Define relational schemata
- Create/alter/delete tables and their attributes

Data *Manipulation* Language (DML)

- Insert/delete/modify tuples in tables
- Query one or more tables

Data *Control* Language (DCL)

- Control access to data stored in a database (Authorization)

Data Types in SQL

Atomic types

- Characters: CHAR[(length)], VARCHAR[(length)]
- Numbers: INT, BIGINT, SMALLINT, FLOAT
- Others: MONEY, DATETIME

Tuple or **row**

- A single entry having the attributes specified by the schema

Attribute or **column**

- A typed data entry present in each tuple in the relation

Table (relation)

- Tuples ensemble

Database Schema

The organization of data as a **blueprint** of how the database is constructed

- Divided into database tables in the case of relational databases

Example

Flights Database example

Table Schema

The **schema** of a table is the *table name*, its *attributes*, and their *types*:

```
Planes(ID: INT, Model: CHAR, Built_date: DATE, Number_of_seats: INT)
```

SQL Constraints

Used to specify rules for data in a table.

Commonly used constraints in SQL:

- **NOT NULL** - Ensures that a column cannot have a NULL value
- **UNIQUE** - Ensures that all values in a column are different
- **PRIMARY KEY** - A combination of a NOT NULL and UNIQUE. Uniquely identifies each row in a table
- **FOREIGN KEY** - Uniquely identifies a row/record in another table
- **CHECK** - Ensures that all values in a column satisfies a specific condition
- **DEFAULT** - Sets a default value for a column when no value is specified

- **INDEX** - Use to create and retrieve data from the database very quickly

Keys

A **key** is a **minimal subset of attributes** that acts as unique identifier for tuples in the relation

It is an implicit constraint on tuples, if two tuples agree on the value(s) of the key, then they **must** be the same tuple.

```
Students(sid: INT, name: CHAR, gpa: FLOAT)
```

1. Which attribute would you select as a key?
2. Is a key always guaranteed to exist?
3. Can we have more than one key?

Foreign Keys

Suppose we have two tables

```
Students(sid: INT, name: CHAR, gpa: FLOAT)
Enrolled(student_id: INT, course_id: CHAR, grade: CHAR)
```

And we want to impose the following constraint:

- *"Only registered students can enroll in courses"*, in other words, a student must exist in the *Students* table to enroll in class.

Students

sid	name	gpa
101	Anne	3.2
123	Mary	3.8

Enrolled

student_id	course_id	grade
123	564	A
123	537	A+

Foreign Keys

What if we try to insert a tuple into *Enrolled*, but there is no such student in *Students*?

- INSERT is rejected -> Foreign keys are **constraints**

What if we delete a student from *Students*?

- Depending on the configuration of the database there are tree options:
  1. An error occurs and no tuples are deleted. OR
  2. The delete operation is propagated and all courses are removed for that student. OR
  3. Each course for that student is set to NULL.

NULL

Whenever we don't have a value

Can mean many things:

- Value does not exists
- Value exists but is unknown
- Value not applicable
- etc.

Example: In the following table, we can add a student 'Jim' who just enrolled on his first class

```
Students(sid: INT, name: CHAR, gpa: FLOAT)
```

sid	name	gpa	
123	Bob	3.9	
143	Jim	NULL	<--

We can constrain a column to be NOT NULL, e.g., "name"

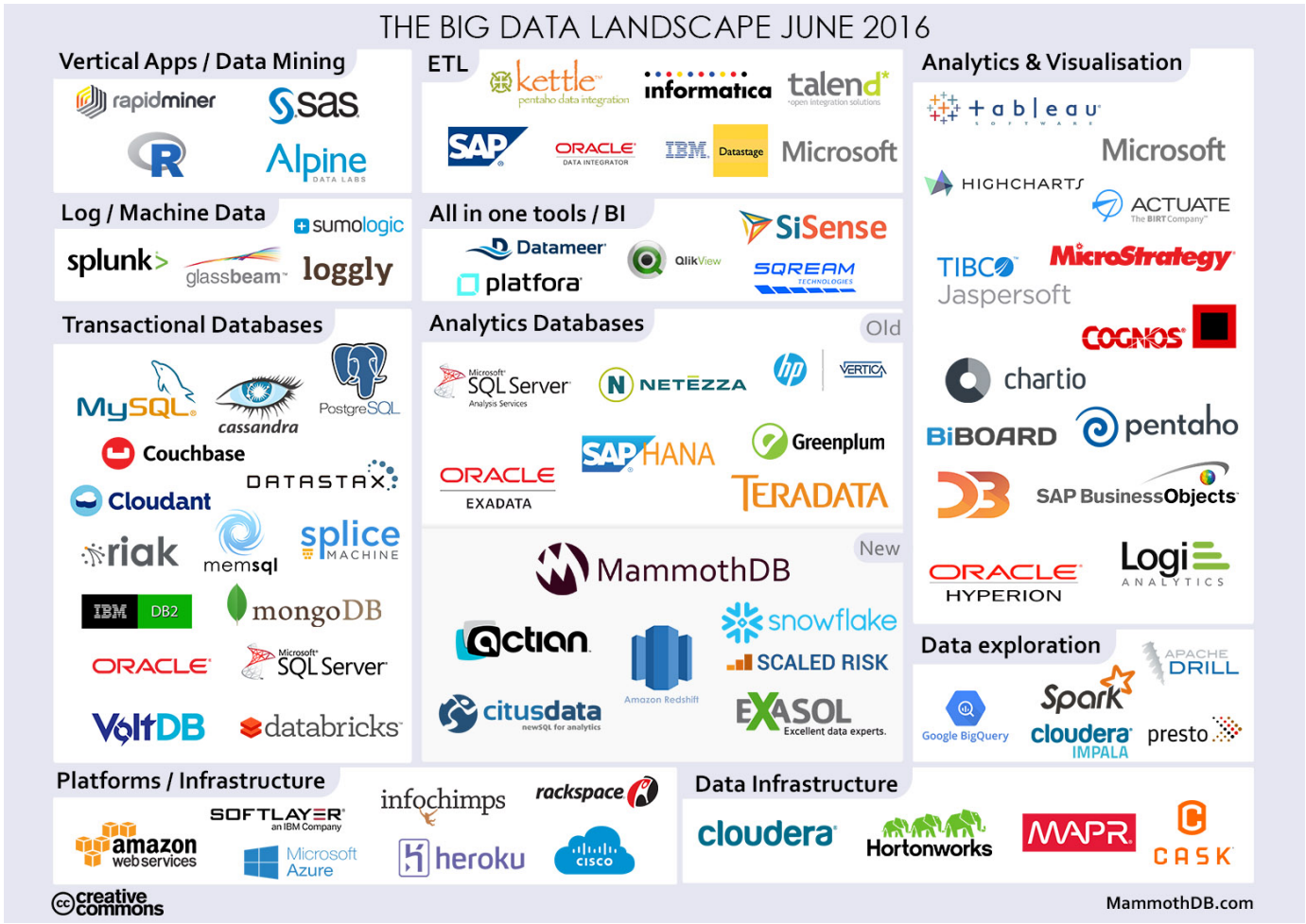
So far

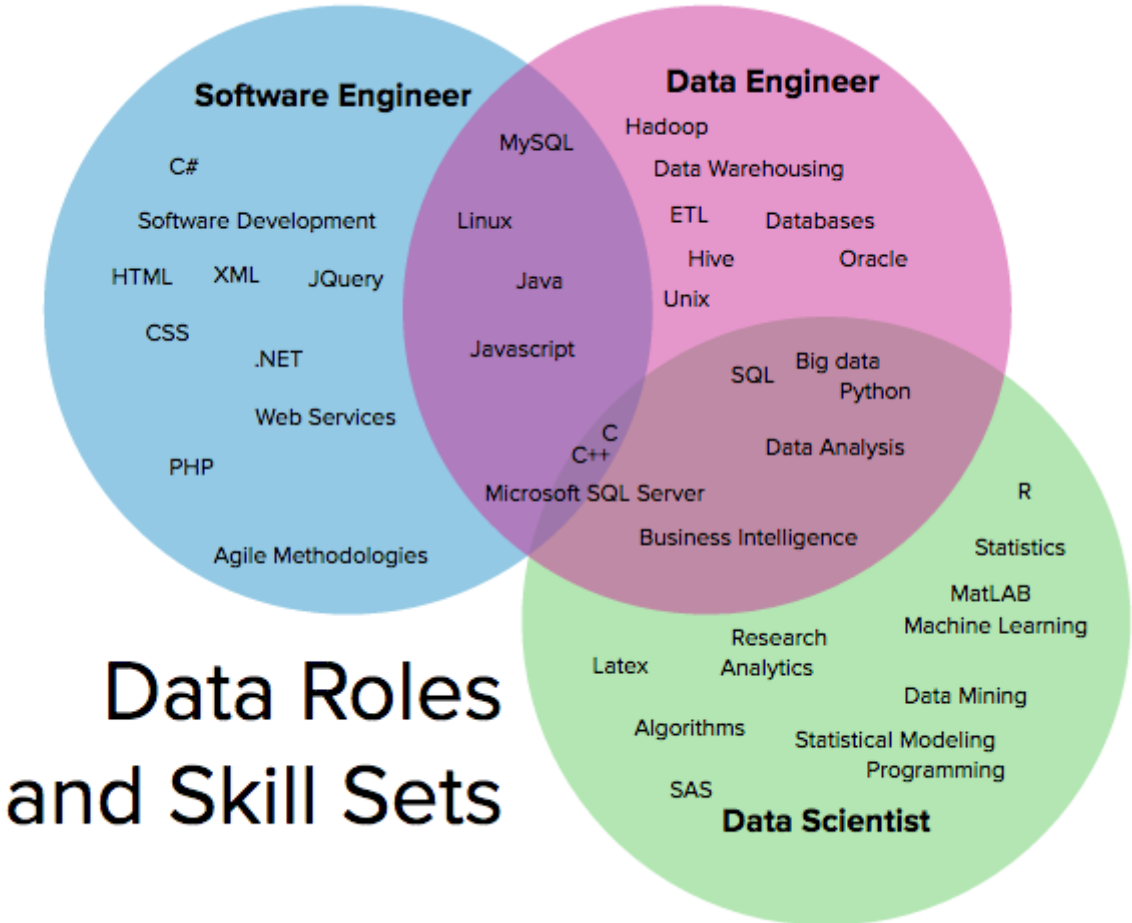
- Schema and Constraints are how databases understand the semantics (meaning) of the data
- They are useful for optimization
- SQL supports general constraints:
  - Keys and foreign keys are the most important

Does it still matter?

Does it still matter?

YES!


THE BIG DATA LANDSCAPE JUNE 2016
Vertical Apps / Data Mining
rapidminer
sas
R
Alpine DATA LABS
ETL
kettle pentaho data integration
informatica
talend open integration solutions
SAP
ORACLE DATA INTEGRATOR
IBM Datastage
Microsoft
Analytics & Visualisation
tableau SOFTWARE
Microsoft
HIGHCHARTS
ACTUATE The BIRT Company
TIBCO Jaspersoft
MicroStrategy
COGNOS
chartio
BIBOARD
pentaho
SAP BusinessObjects
ORACLE HYPERION
Logi ANALYTICS
Log / Machine Data
sumologic
splunk>
glassbeam
loggly
All in one tools / BI
SiSense
Datameer
platfora
QlikView
SQREAM TECHNOLOGIES
Transactional Databases
MySQL
cassandra
PostgreSQL
Couchbase
DATASTAX
Cloudant
riak
memsql
splice MACHINE
IBM DB2
mongoDB
ORACLE
Microsoft SQL Server
VoltDB
databricks
Analytics Databases
Old
Microsoft SQL Server Analysis Services
NETEZZA
hp
VERTICA
SAP HANA
Greenplum
ORACLE EXADATA
TERADATA
New
MammothDB
snowflake
actian
SCALED RISK
Amazon Redshift
citusdata newSQL for analytics
EXASOL Excellent data experts.
Data exploration
APACHE DRILL
Google BigQuery
Spark
cloudera IMPALA
presto
Platforms / Infrastructure
amazon web services
SOFTLAYER an IBM Company
infochimps
rackspace
Microsoft Azure
heroku
cisco
Data Infrastructure
cloudera
Hortonworks
MAPR
CASK
creative commons
MammothDB.com


Software Engineer
C#
Software Development
HTML
XML
JQuery
CSS
.NET
Web Services
PHP
Agile Methodologies
Data Engineer
MySQL
Hadoop
Data Warehousing
Linux
ETL
Databases
Hive
Oracle
Java
Unix
Javascript
SQL
Big data
Python
C
C++
Data Analysis
Microsoft SQL Server
R
Business Intelligence
Statistics
MatLAB
Machine Learning
Research
Latex
Analytics
Data Mining
Algorithms
Statistical Modeling
Programming
SAS
Data Scientist
Data Roles and Skill Sets

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY