GENERAL ASSEMBLY

DATA SCIENCE

SYD DAT 8

Week 5 - Databases and SQL
Thursday 22nd June

Stephen Samild

Manager Business & Operations Analytics at QBE Insurance

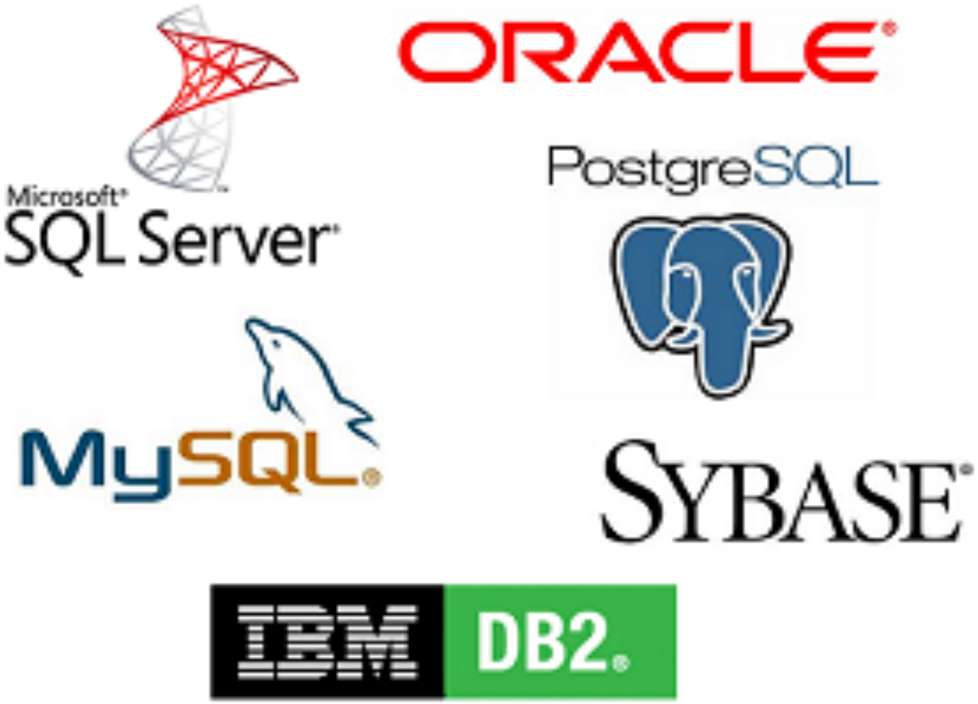
AGENDA

1. Databases
2. SQL
3. Lab: SQL
4. Discussion of Data Science Productivity Tools

DATA SCIENCE PART TIME COURSE

DATABASES

RDBMS


Microsoft SQL Server
ORACLE
PostgreSQL
MySQL
SYBASE
IBM DB2

WHAT IS A DATABASE

Databases are computer systems that manage storage and querying of data. Databases provide a way to organise data along with efficient methods to retrieve specific information.

Typically, retrieval is performed using a query language, a mini programming syntax with a few basic operators for data transformation, the most common of which is SQL.

WHAT IS A DATABASE

Databases are the standard solution for data storage and are much more robust than text, CSV or json files. Most analyses involve pulling data to and from a resource and in most settings, that means using a database.

Databases can come in many flavours, designed to serve for different use cases.

MySQL™

Microsoft® SQL Server®

mongoDB®

WHY USE DATABASES

Rules on structure make writing and retrieving data more reliable and efficient.

  - Standardised business definitions
  - Source of truth
  - Fast read

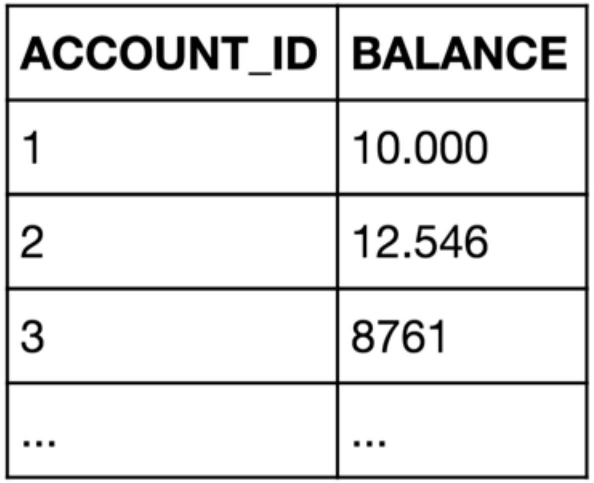
Where databases are valued:

- Operational systems
- Reporting

RELATIONAL DATABASE – TABLES

A relational database is a database based tabular data and links between data entities or concepts.

A relational database is organised into tables. Each table should correspond to one entity or concept.



ACCOUNT_ID	BALANCE
1	10.000
2	12.546
3	8761
...	...

RELATIONAL DATABASE - SCHEMA

A table is made up rows and columns, similar to a Pandas dataframe or Excel spreadsheet.

A table also has a schema which is a set of rules for what goes in each table. These specify what columns are contained in the table and what type those columns are (text, integers, floats, etc.).

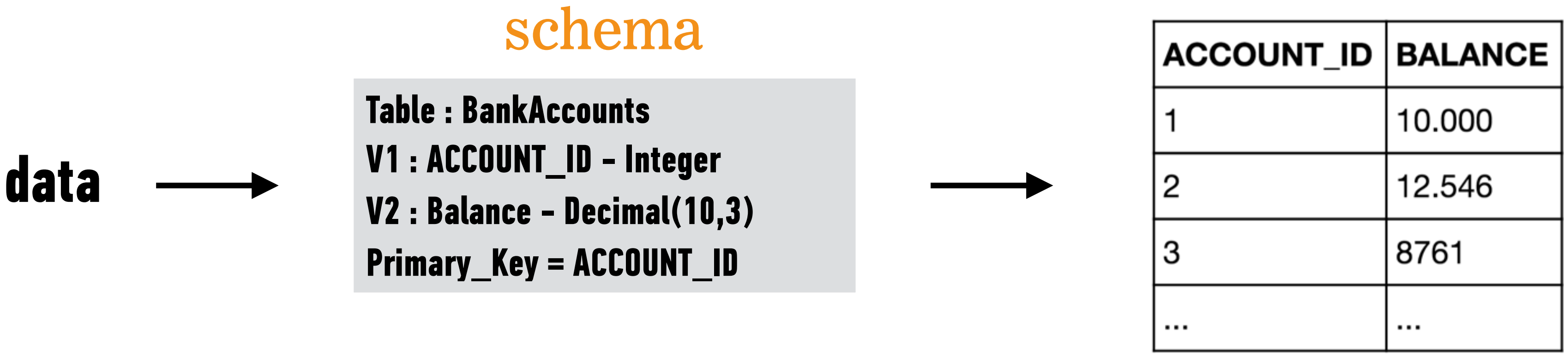


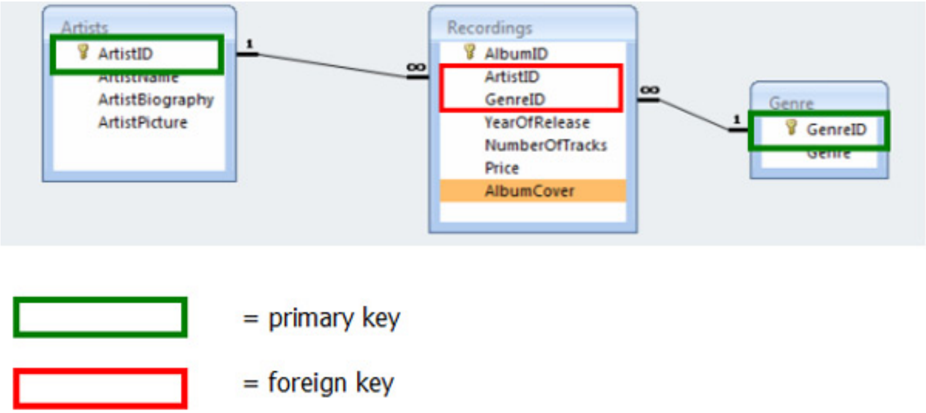
Table : BankAccounts
V1 : ACCOUNT_ID - Integer
V2 : Balance - Decimal(10,3)
Primary_Key = ACCOUNT_ID

ACCOUNT_ID	BALANCE
1	10.000
2	12.546
3	8761
...	...

RELATIONAL DATABASE – KEYS

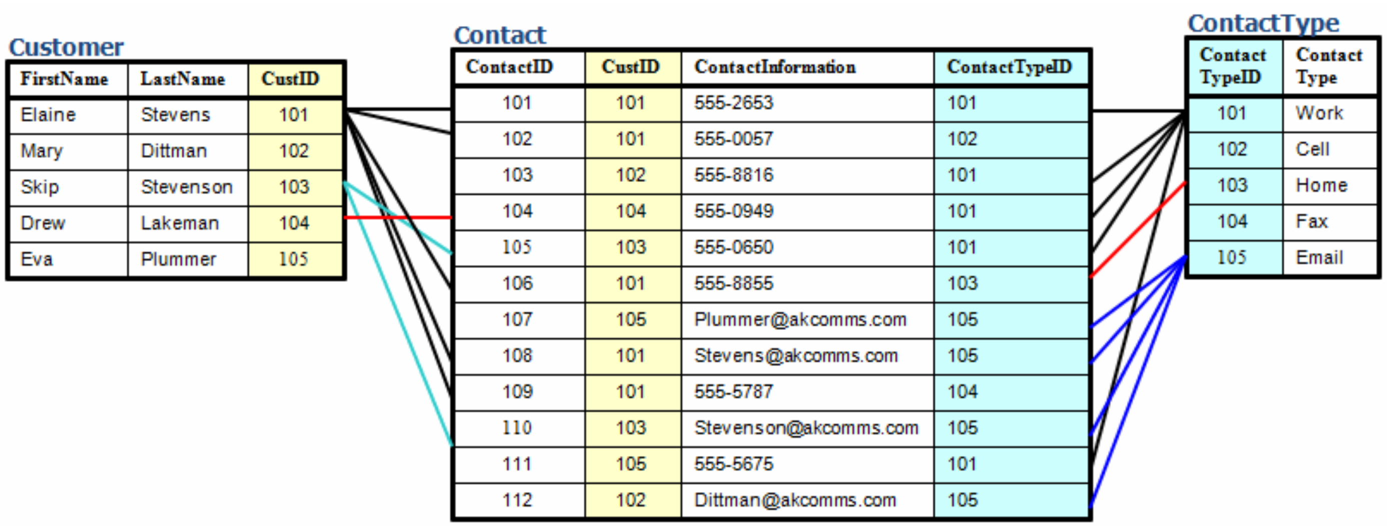
Each table typically has a primary key column. This column is a unique value per row and serves as the identifier for the row.

A table can have many foreign keys as well. A foreign key is a column that contains values to link the table to the other tables.



RELATIONAL DATABASE – DATA MODEL

Databases are ‘modelled’ to suit their intended purpose.



Customer

FirstName	LastName	CustID
Elaine	Stevens	101
Mary	Dittman	102
Skip	Stevenson	103
Drew	Lakeman	104
Eva	Plummer	105

Contact

ContactID	CustID	ContactInformation	ContactTypeID
101	101	555-2653	101
102	101	555-0057	102
103	102	555-8816	101
104	104	555-0949	101
105	103	555-0650	101
106	101	555-8855	103
107	105	Plummer@akcomms.com	105
108	101	Stevens@akcomms.com	105
109	101	555-5787	104
110	103	Stevenson@akcomms.com	105
111	105	555-5675	101
112	102	Dittman@akcomms.com	105

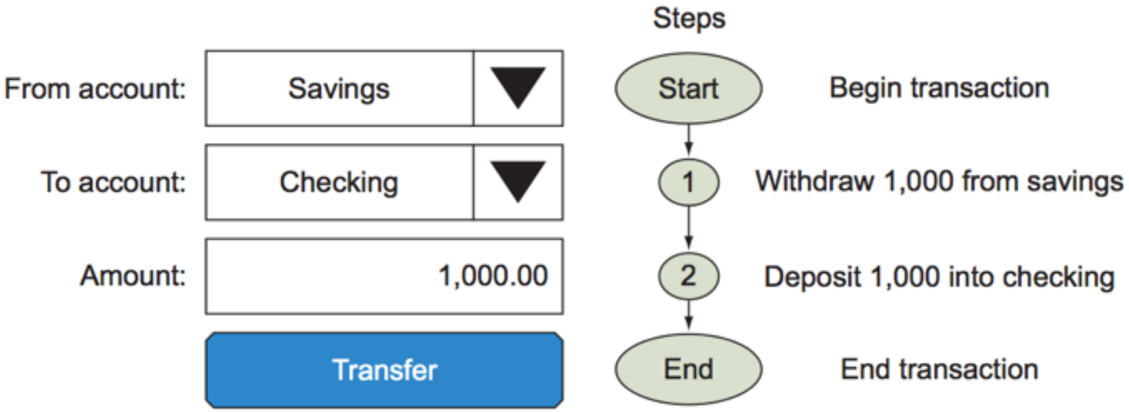
ContactType

Contact TypeID	Contact Type
101	Work
102	Cell
103	Home
104	Fax
105	Email

RELATIONAL DATABASE – TRANSACTIONAL INTEGRITY

A unit of work performed against a database is called a transaction. This term generally represents any change in database.

Example: Transfer money from an account to another.



From account: Savings

To account: Checking

Amount: 1,000.00

Steps

Start — Begin transaction

1 — Withdraw 1,000 from savings

2 — Deposit 1,000 into checking

End — End transaction

- What happens if step 1 succeeds and step 2 fails ?
- What if you request the balance between step 1 and step2 ?

RELATIONAL DATABASE - ACID

ACID is a set of properties that guarantee that database transactions are processed reliably.

Atomicity "all or nothing": if one part of the transaction fails, the entire transaction fails, and the database state is left unchanged.

Consistency ensures that any transaction will bring the database from one valid state to another.

Isolation ensures that the concurrent execution of transactions results in a system state that would be obtained if transactions were executed serially, i.e., one after the other.

Durability ensures that once a transaction has been committed, it will remain so, even in the event of power loss, crashes, or errors.

DATA SCIENCE PART TIME COURSE

NO-SQL Databases

SQL/NoSQL

SQL

- Traditional rows and columns data
- Strict structure / Primary Keys
- Entire column for each feature
- Industry standard

NoSQL

- No well defined data structure
- Works better for unstructured data
- Cheaper hardware
- Popular among Startups

SQL/NoSQL - Examples

SQL

- MySQL
- Oracle
- Postgres
- SQLite
- SQLServer
- Redshift

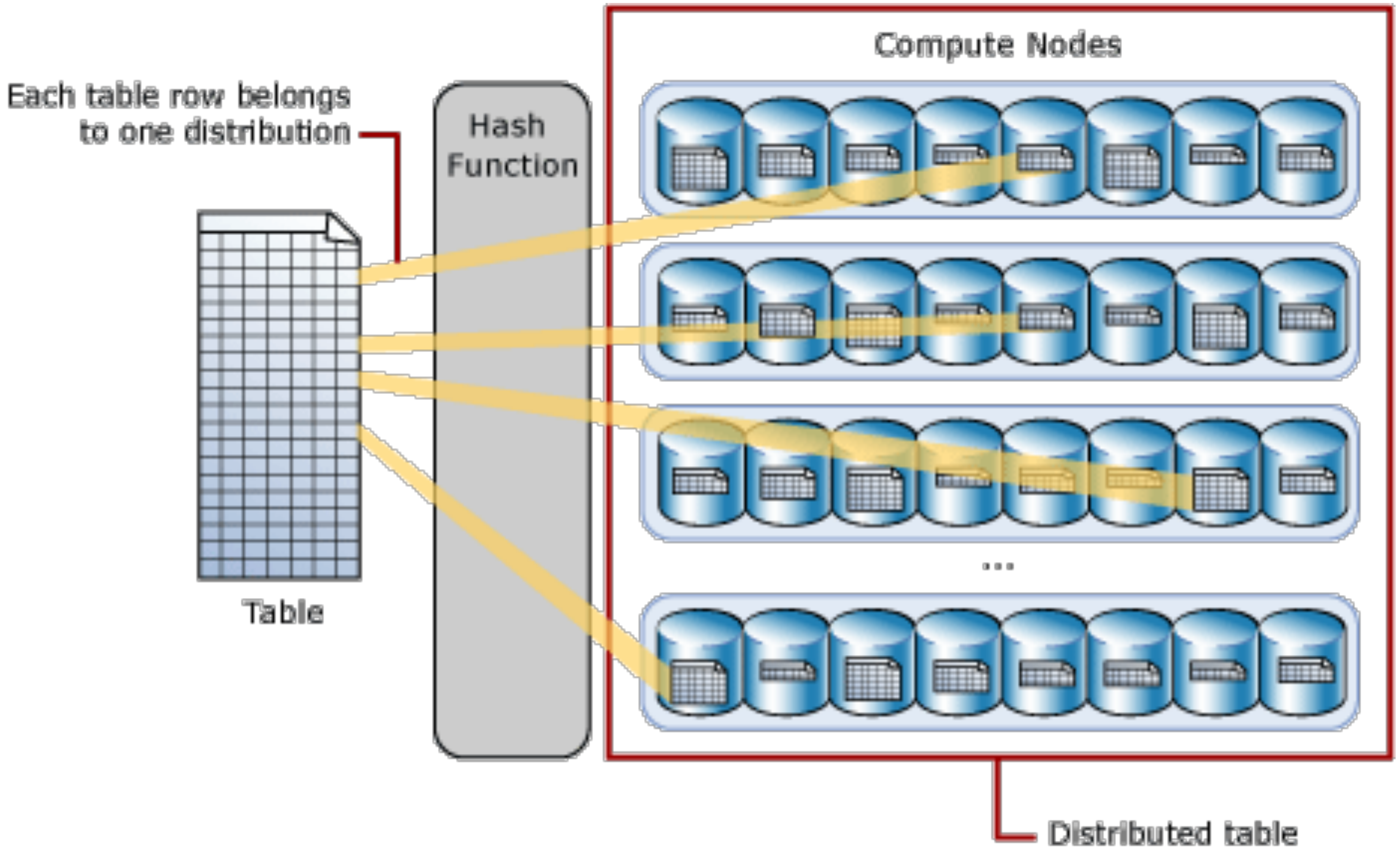
NoSQL

- MongoDB
- CouchDB
- Redis
- Cassandra
- Neo4j
- HBase

DATA SCIENCE PART TIME COURSE

BIG DATA & SQL

MPP RDBMS - Massively Parallel Processing



TERADATA

Greenplum

VERTICA
An HP Company

AMAZON REDSHIFT

BIG DATA SQL

DATA SCIENCE PART TIME COURSE

SQL

LAB

SYNCHING YOUR FORK WITH THE COURSE REPO

1. re-name your labs with lab_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD_DAT_6 local repo>
3. commit your changes ahead of sync
    - git status
    - git add .
    - git commit -m “descriptive label for the commit”
    - git status
4. download new material from official course repo (upstream) and merge it
    - git checkout master (ensures you are in the master branch)
    - git fetch upstream
    - git merge upstream/master

DATA SCIENCE PART TIME COURSE

DISCUSSION: PRODUCTIVITY TOOLS

PRODUCTIVITY TOOLS

Workflow, collaboration, operationalising ↔ Automatic Machine Learning - Auto ML

rapidminer

scienceOPS

sas

DataRobot

TPOT

Auto-Sklearn

DATA SCIENCE

HOMEWORK

Read the following

- Read Chapter 2 of http://www.redbook.io/
- https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8