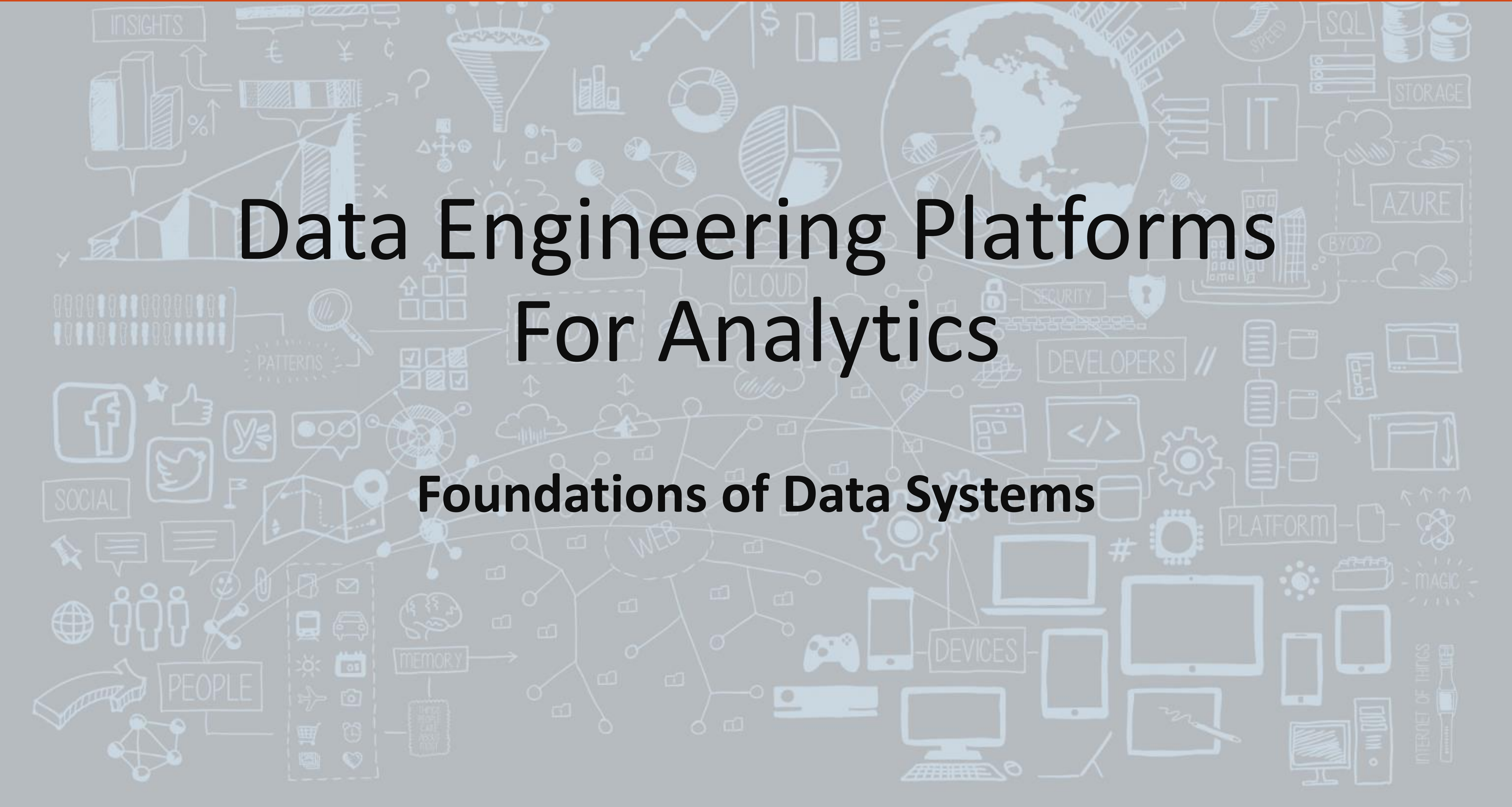


Data Engineering Platforms For Analytics

Foundations of Data Systems

Modules

- Welcome
- Introduction to data
- Data Preparation
- Database Management Systems
- Types of Data bases

Welcome

Introductions, Syllabus review

Syllabus

- Session 1 – Foundations of data systems
- Session 2 – Relational Databases
- Session 3 – Structured Query Language (SQL)
- Session 4 – Analytical Data Platforms
- Session 5 – Business Intelligence
- Session 6 – Cloud Data Pipelines
- Session 7 – NoSQL - Document and Graph Databases
- Session 8 – NoSQL - Columnar and Key Value Databases
- Session 9 – Introduction to Blockchain
- Session 10 – Team Project Presentations

Assignments & Final Project

- 4 Assignments (Individual)
  - 100 points each
  - Bi-weekly submission (Saturdays 11:59 PM CST)
- Final Project (Max 4 students)
  - Business Use Case - 10%
  - Data Analysis & Preparation - 25%
  - Data Modelling & Design - 25 %
  - Tools / Database concepts -20%
  - Dashboards and insights - 20%

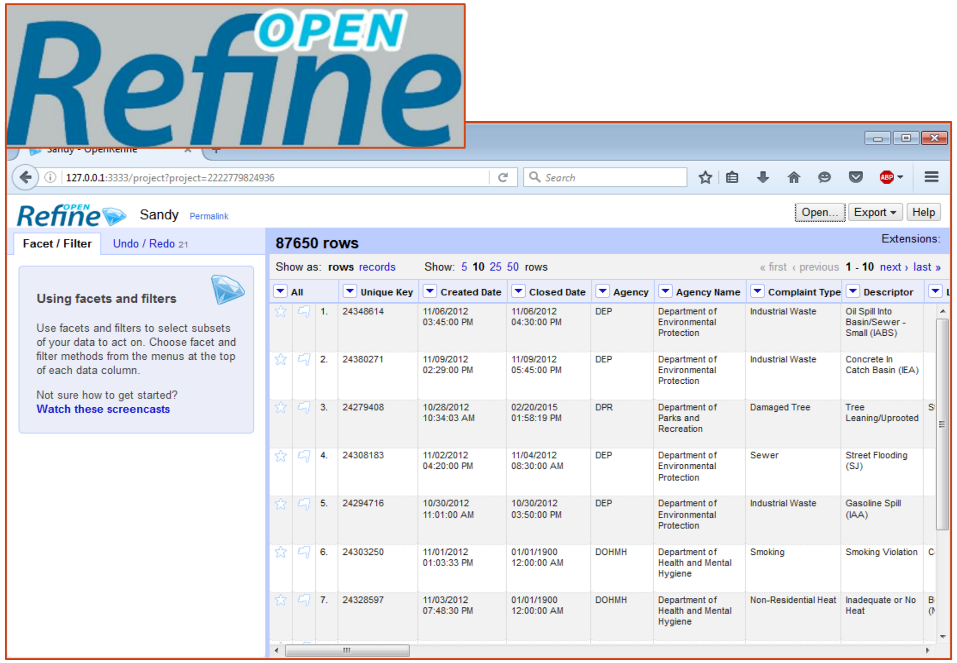
Primary Datasets

- Classic Models (In-Class exercises)
  - Sessions 2, 3, 4

- Sakila (Assignments)
  - Assignments 2,3,4

- Final Project (Sample datasets)
  - IRI Dataset
  - https://toolbox.google.com/datasetsearch/
  - https://data.cityofchicago.org/
  - https://opendata.cityofnewyork.us/
  - https://data.gov.in/catalogs/
  - https://github.com/awesomedata/awesome-public-datasets/
  - https://www.springboard.com/blog/free-public-data-sets-data-science-project/

Software Tools



Using facets and filters

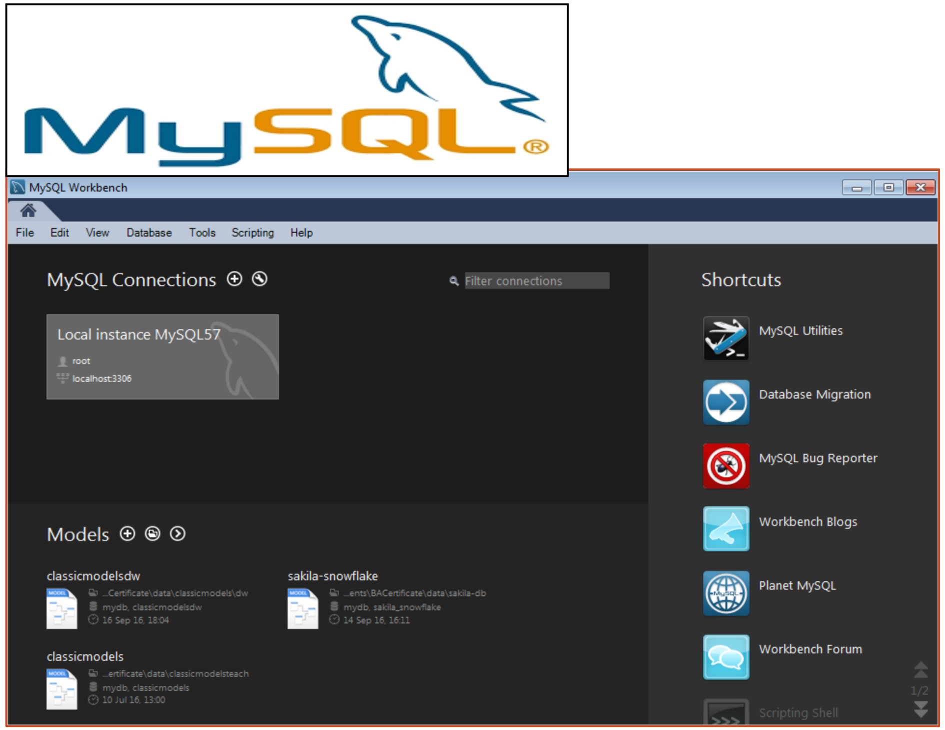
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

87650 rows

Show as: rows records   Show: 5 10 25 50 rows   « first ‹ previous 1 - 10 next › last »

All	Unique Key	Created Date	Closed Date	Agency	Agency Name	Complaint Type	Descriptor
1.	24348614	11/06/2012 03:45:00 PM	11/06/2012 04:30:00 PM	DEP	Department of Environmental Protection	Industrial Waste	Oil Spill Into Basin/Sewer - Small (IABS)
2.	24380271	11/09/2012 02:29:00 PM	11/09/2012 05:45:00 PM	DEP	Department of Environmental Protection	Industrial Waste	Concrete In Catch Basin (IEA)
3.	24279408	10/28/2012 10:34:03 AM	02/20/2015 01:58:19 PM	DPR	Department of Parks and Recreation	Damaged Tree	Tree Leaning/Uprooted
4.	24308183	11/02/2012 04:20:00 PM	11/04/2012 08:30:00 AM	DEP	Department of Environmental Protection	Sewer	Street Flooding (SJ)
5.	24294716	10/30/2012 11:01:00 AM	10/30/2012 03:50:00 PM	DEP	Department of Environmental Protection	Industrial Waste	Gasoline Spill (IAA)
6.	24303250	11/01/2012 01:03:33 PM	01/01/1900 12:00:00 AM	DOHMH	Department of Health and Mental Hygiene	Smoking	Smoking Violation
7.	24328597	11/03/2012 07:48:30 PM	01/01/1900 12:00:00 AM	DOHMH	Department of Health and Mental Hygiene	Non-Residential Heat	Inadequate or No Heat



Local instance MySQL57
root
localhost:3306

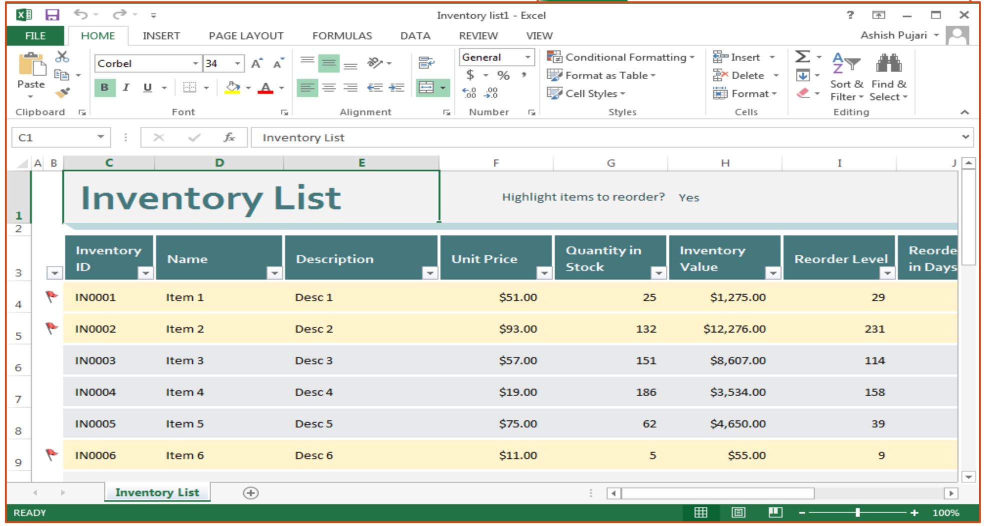
classicmodelsdw
...Certificate\data\classicmodels\dw
mydb, classicmodelsdw
16 Sep 16, 18:04

sakila-snowflake
...ents\BACertificate\data\sakila-db
mydb, sakila_snowflake
14 Sep 16, 16:11

classicmodels
...ertificate\data\classicmodelsteach
mydb, classicmodels
10 Jul 16, 13:00

Software Tools - Continued…

Excel



Font: Corbel, 34

Number format: General

C1 — Inventory List

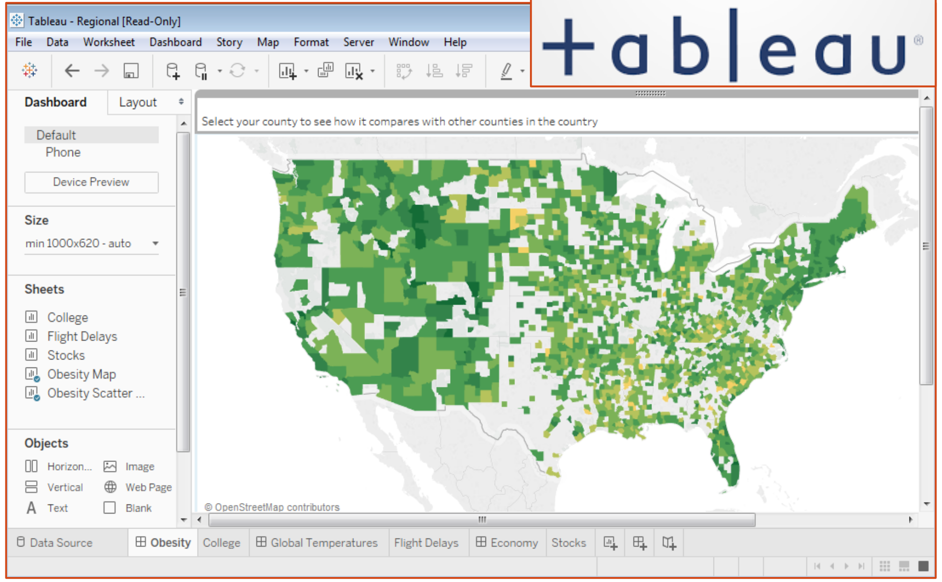
Inventory List

Highlight items to reorder? Yes

Inventory ID	Name	Description	Unit Price	Quantity in Stock	Inventory Value	Reorder Level	Reorde in Days
IN0001	Item 1	Desc 1	$51.00	25	$1,275.00	29	
IN0002	Item 2	Desc 2	$93.00	132	$12,276.00	231	
IN0003	Item 3	Desc 3	$57.00	151	$8,607.00	114	
IN0004	Item 4	Desc 4	$19.00	186	$3,534.00	158	
IN0005	Item 5	Desc 5	$75.00	62	$4,650.00	39	
IN0006	Item 6	Desc 6	$11.00	5	$55.00	9	

Inventory List

tableau



Size: min 1000x620 - auto

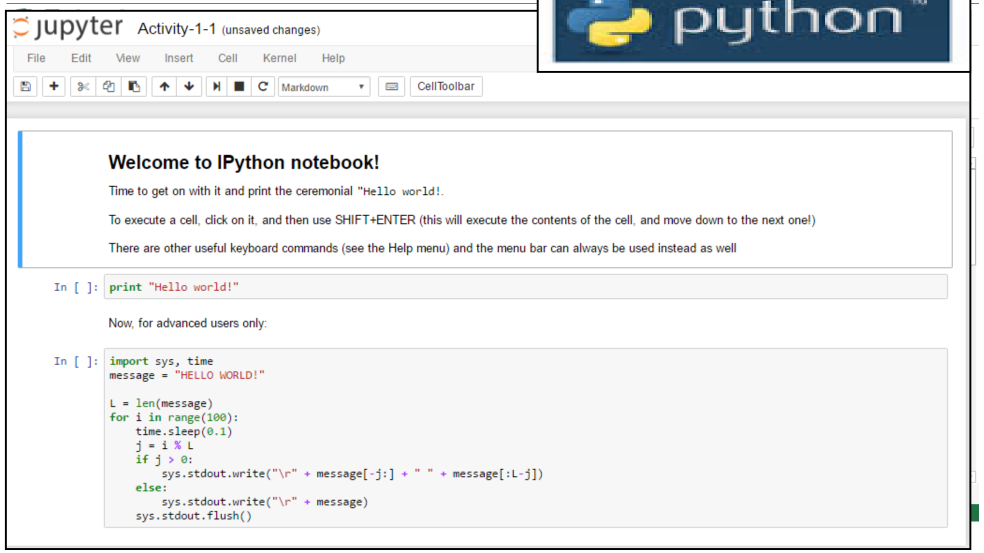
Sheets:
- College
- Flight Delays
- Stocks
- Obesity Map
- Obesity Scatter ...

Select your county to see how it compares with other counties in the country

Software Tools - Continued…

Anaconda

python™



Welcome to IPython notebook!

Time to get on with it and print the ceremonial "`Hello world!`.

To execute a cell, click on it, and then use SHIFT+ENTER (this will execute the contents of the cell, and move down to the next one!)

There are other useful keyboard commands (see the Help menu) and the menu bar can always be used instead as well

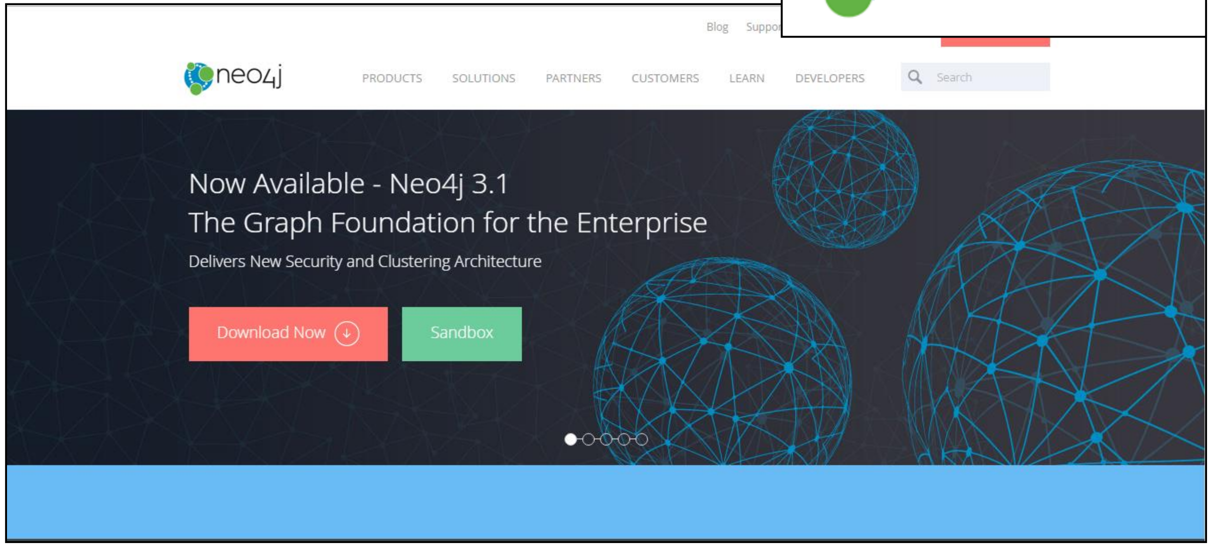
```
In [ ]: print "Hello world!"
```

Now, for advanced users only:

```
In [ ]: import sys, time
        message = "HELLO WORLD!"

        L = len(message)
        for i in range(100):
            time.sleep(0.1)
            j = i % L
            if j > 0:
                sys.stdout.write("\r" + message[-j:] + " " + message[:L-j])
            else:
                sys.stdout.write("\r" + message)
            sys.stdout.flush()
```

neo4j

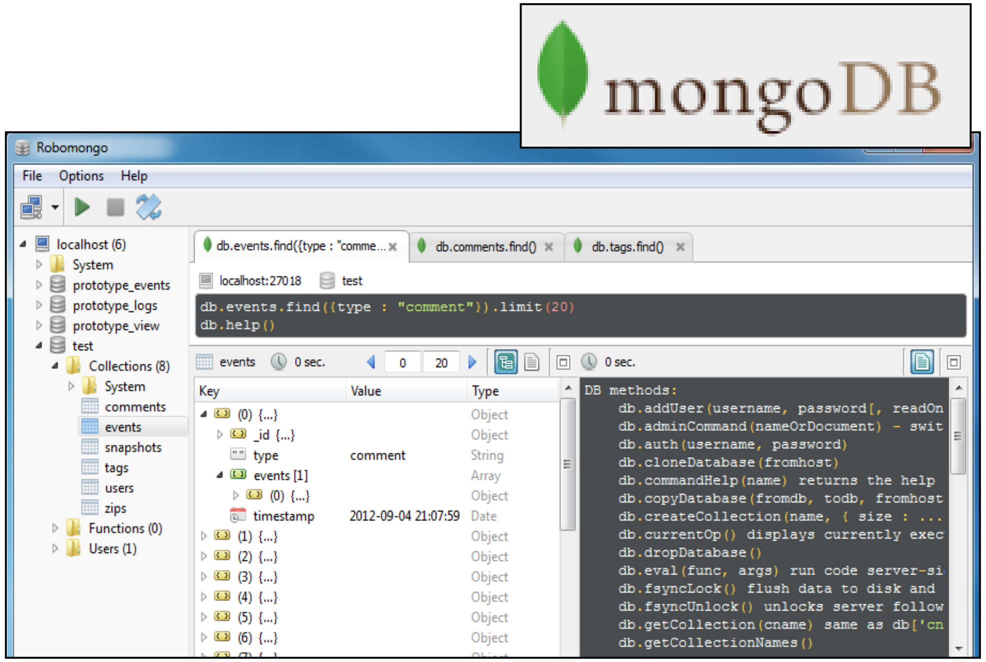


Now Available - Neo4j 3.1

The Graph Foundation for the Enterprise

Delivers New Security and Clustering Architecture

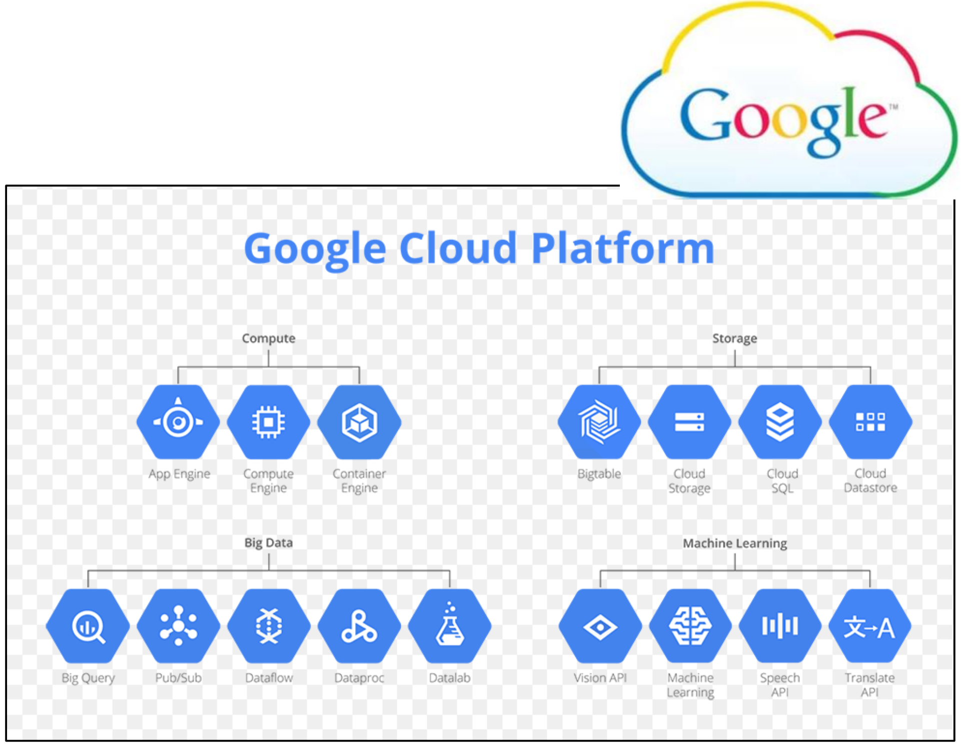
Software Tools - Continued…



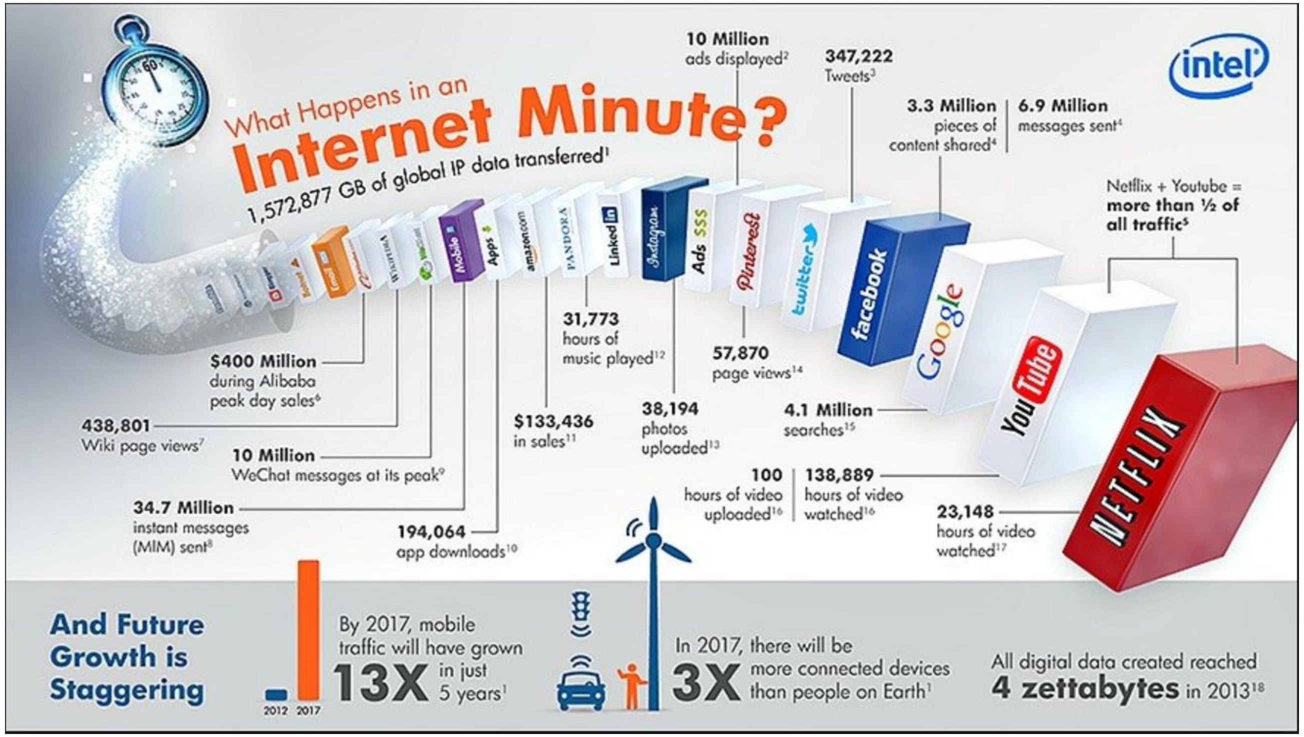
```
db.events.find({type : "comment"}).limit(20)
db.help()
```

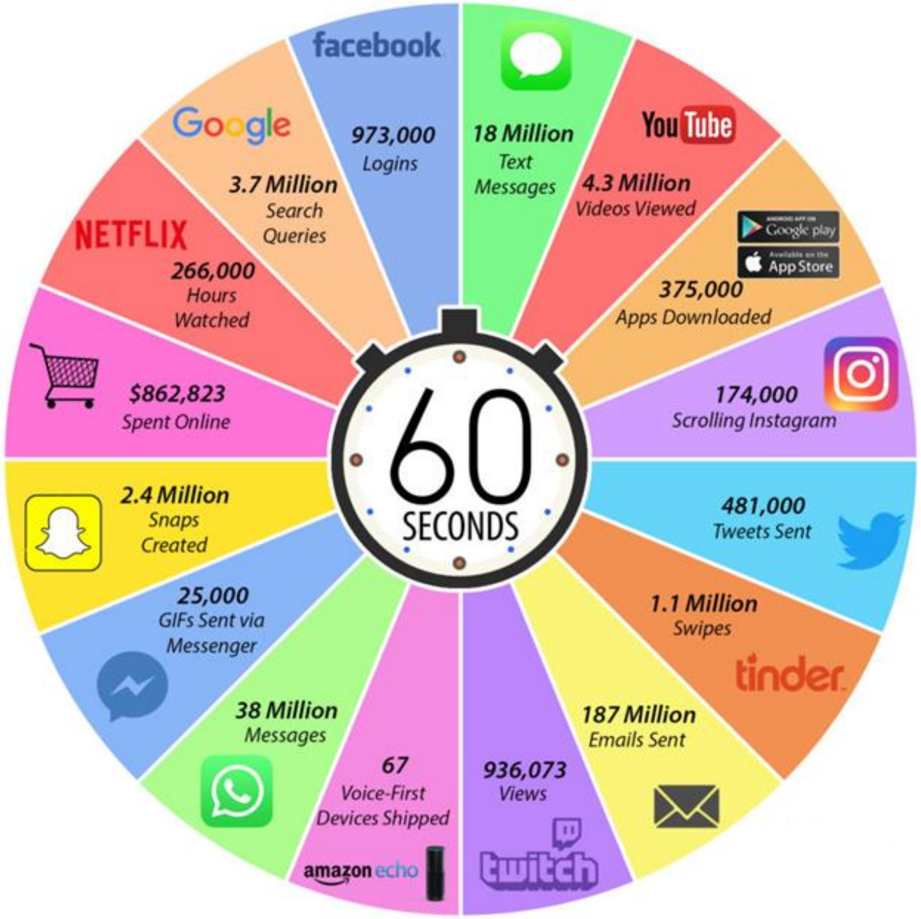
Key	Value	Type
(0) {...}		Object
_id {...}		Object
type	comment	String
events [1]		Array
(0) {...}		Object
timestamp	2012-09-04 21:07:59	Date
(1) {...}		Object
(2) {...}		Object
(3) {...}		Object
(4) {...}		Object
(5) {...}		Object
(6) {...}		Object

```
DB methods:
	db.addUser(username, password[, readOn
	db.adminCommand(nameOrDocument) - swit
	db.auth(username, password)
	db.cloneDatabase(fromhost)
	db.commandHelp(name) returns the help
	db.copyDatabase(fromdb, todb, fromhost
	db.createCollection(name, { size : ...
	db.currentOp() displays currently exec
	db.dropDatabase()
	db.eval(func, args) run code server-si
	db.fsyncLock() flush data to disk and
	db.fsyncUnlock() unlocks server follow
	db.getCollection(cname) same as db['cn
	db.getCollectionNames()
```



What Happens in an Internet Minute

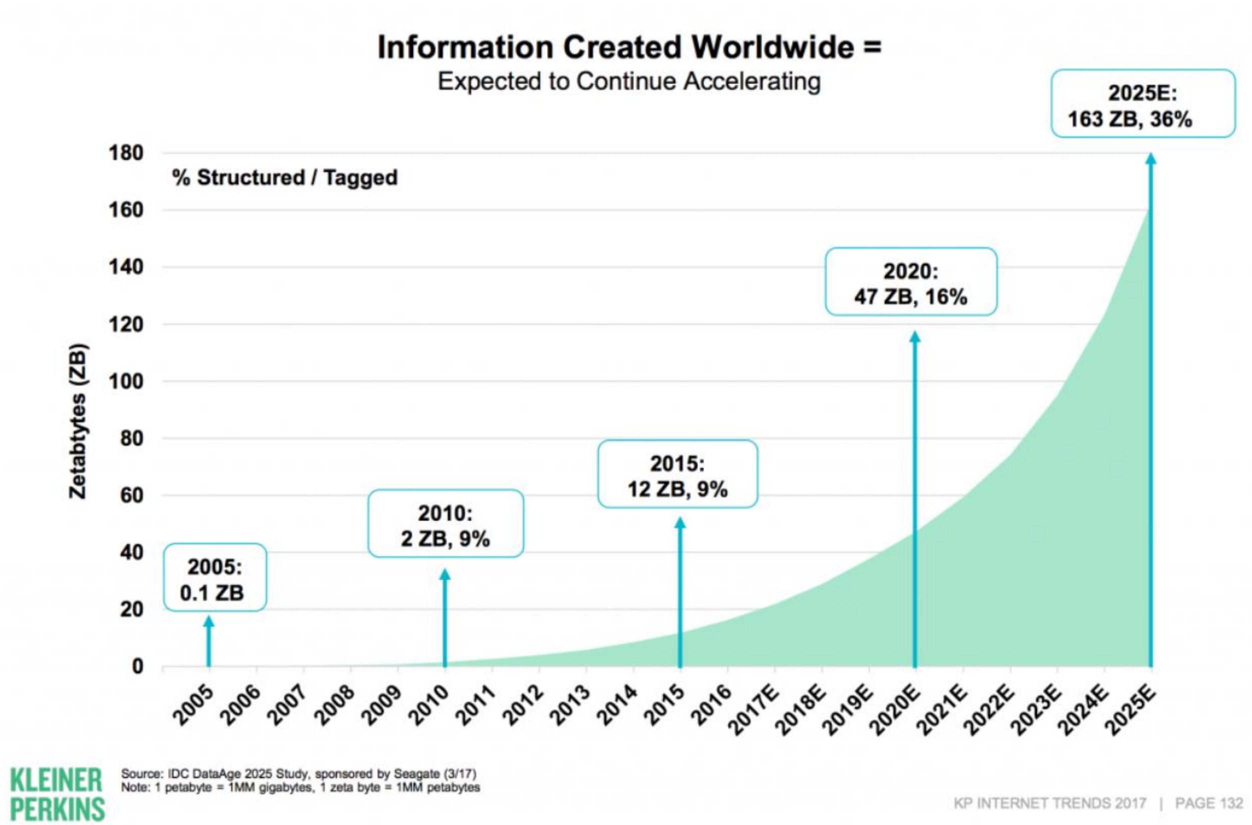




Growth of data

- The New York Stock Exchange generates about 4−5 terabytes of data per day.
- Facebook hosts more than 240 billion photos, growing at 7 petabytes per month.
- Ancestry.com, the genealogy site, stores around 10 petabytes of data.
- The Internet Archive stores around 18.5 petabytes of data.
- The Large Hadron Collider near Geneva, Switzerland, produces about 30 petabytes of data per year.

Source: Hadoop The Definitive Guide

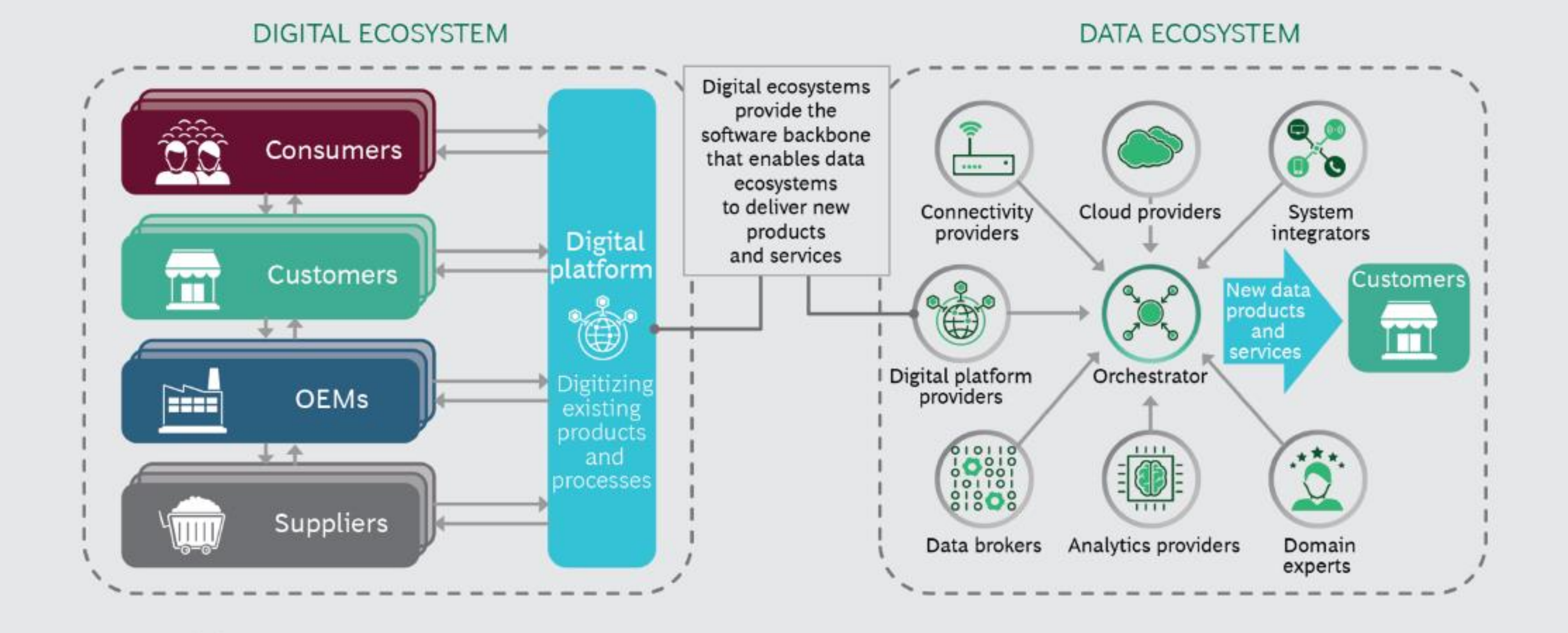


Information Created Worldwide =
Expected to Continue Accelerating

Source: IDC DataAge 2025 Study, sponsored by Seagate (3/17)
Note: 1 petabyte = 1MM gigabytes, 1 zeta byte = 1MM petabytes

KP INTERNET TRENDS 2017 | PAGE 132

Digital and Data Ecosystems

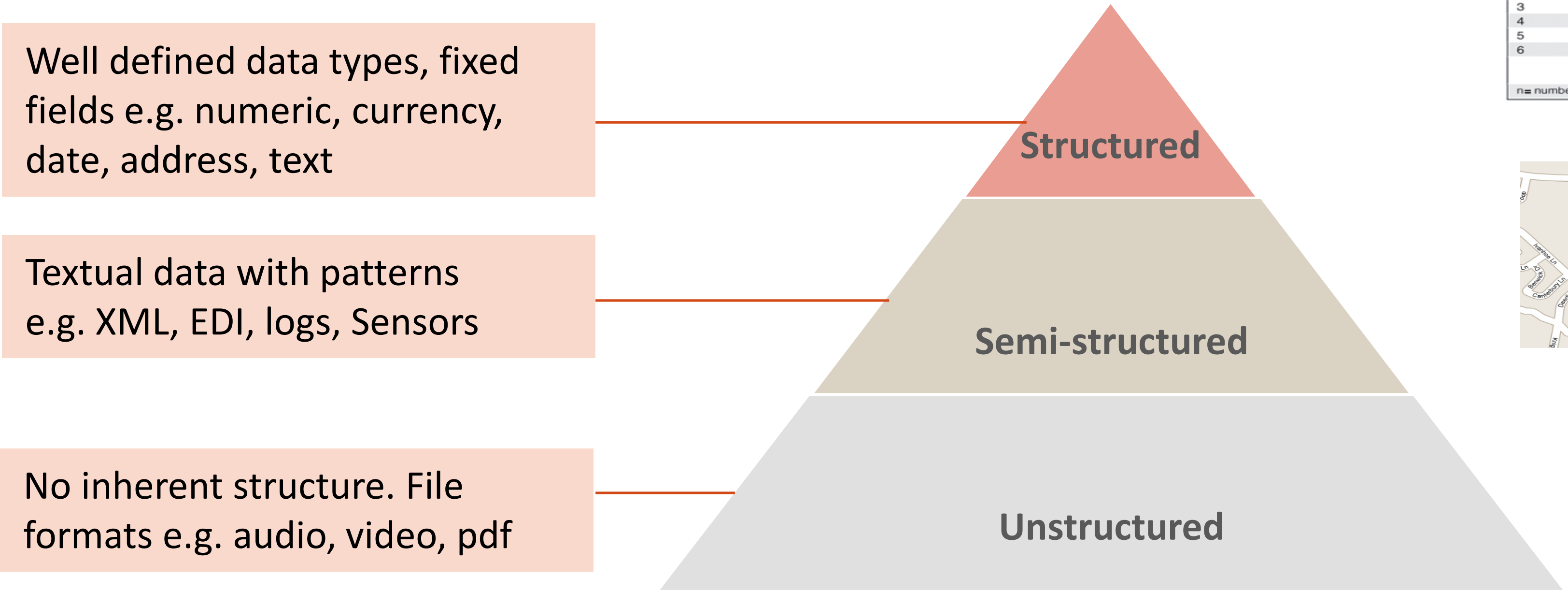

DIGITAL ECOSYSTEM
DATA ECOSYSTEM
Consumers
Customers
OEMs
Suppliers
Digital platform
Digitizing existing products and processes
Digital ecosystems provide the software backbone that enables data ecosystems to deliver new products and services
Connectivity providers
Cloud providers
System integrators
Digital platform providers
Orchestrator
New data products and services
Customers
Data brokers
Analytics providers
Domain experts

Source: BCG analysis.

Data Classification

Data Growth, Data Types, Formats

Data Formats, Classification



Well defined data types, fixed fields e.g. numeric, currency, date, address, text

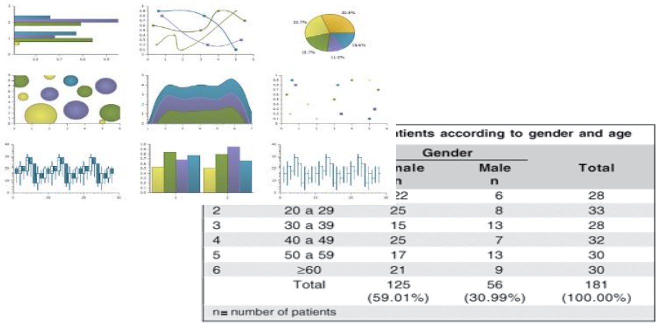
Structured

Textual data with patterns e.g. XML, EDI, logs, Sensors

Semi-structured

No inherent structure. File formats e.g. audio, video, pdf

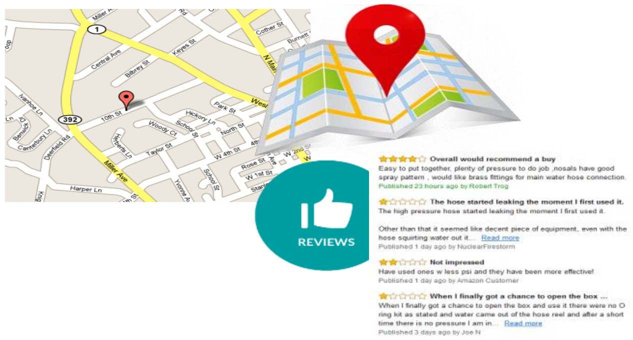
Unstructured



ıtients according to gender and age

		Gender		
		nale n	Male n	Total
		22	6	28
2	20 a 29	25	8	33
3	30 a 39	15	13	28
4	40 a 49	25	7	32
5	50 a 59	17	13	30
6	≥60	21	9	30
	Total	125 (59.01%)	56 (30.99%)	181 (100.00%)

n= number of patients



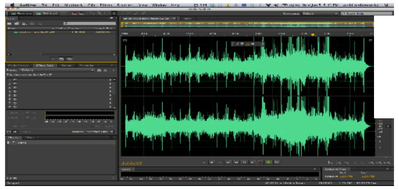
★★★★☆ **Overall would recommend a buy**
Easy to put together, plenty of pressure to do job ,nosals have good spray pattern , would like brass fittings for main water hose connection.
Published 23 hours ago by Robert Trog

★☆☆☆☆ **The hose started leaking the moment I first used it.**
The high pressure hose started leaking the moment I first used it.

Other than that it seemed like decent piece of equipment, even with the hose squirting water out it... Read more
Published 1 day ago by NuclearFirestorm

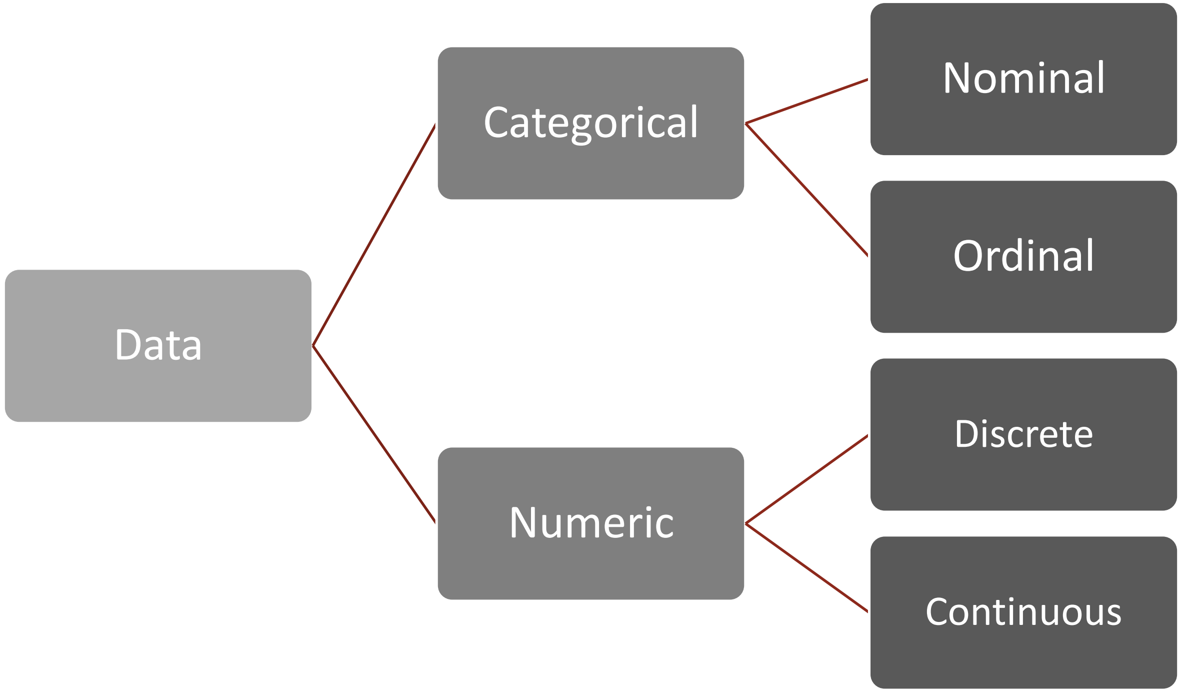
★★☆☆☆ **Not impressed**
Have used ones w less psi and they have been more effective!
Published 1 day ago by Amazon Customer

★☆☆☆☆ **When I finally got a chance to open the box ...**
When I finally got a chance to open the box and use it there were no O ring kit as stated and water came out of the hose reel and after a short time there is no pressure I am in.... Read more
Published 3 days ago by Joe N



YouTube

Types of data



Categorical

Classified or ordered

Nominal

- Descriptions or labels - No sense of sequence
- Ex: Red/blue/yellow OR M/F, etc

Ordinal

- Meaningful Order
- Ex: 1st (98.3%)/2nd (97.2%)/3rd(91.4%)

Numerical Data

Can be measured

Discrete

- Quantitative data with whole numbers
- Ex: number of students in a class

Continuous

- Can take on fractions and decimals
- Ex: height / weight

Binary Format

- Bit - smallest unit of information

0/1

- Byte - 8 bits of information

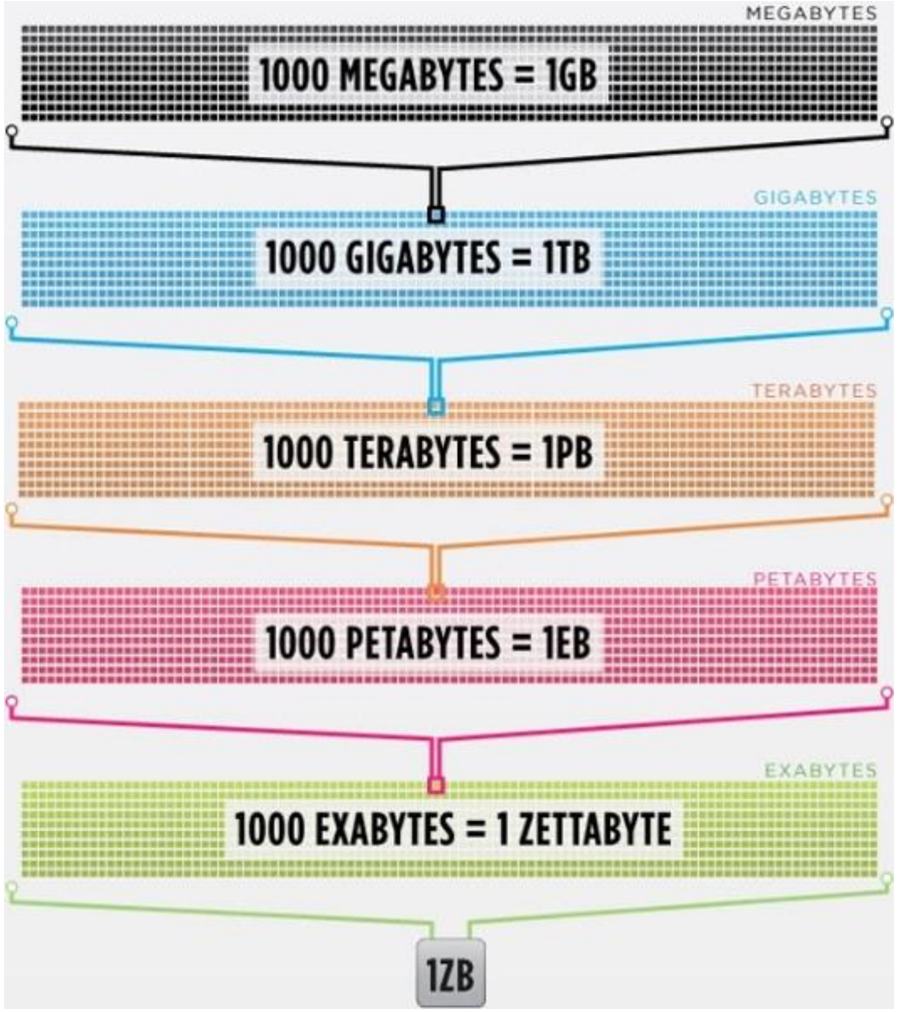
0/1 0/1 0/1 0/1 0/1 0/1 0/1 0/1

- 1 Kilo Byte = 2^{10} Bytes = 1024 bytes

- 1 Kilo Byte = 10^{3} Bytes = 1000 bytes

Data Sizes

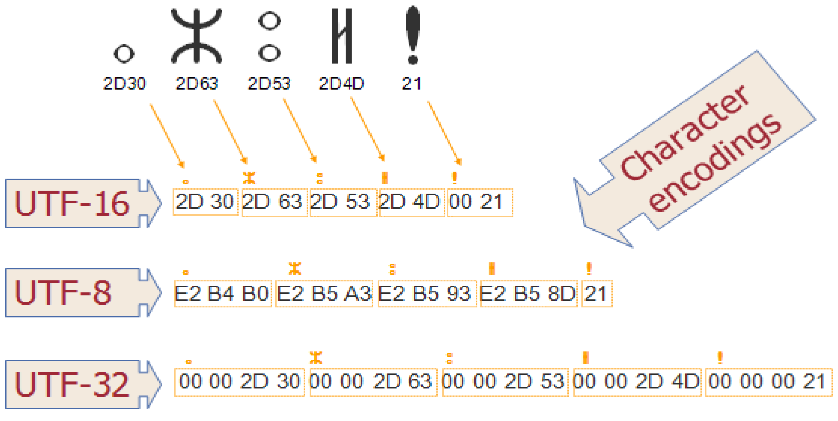
Data Size	Bytes	Examples
Kilobyte (KB)	10^3	Emails, Browser cookies
Megabyte (MB)	10^6	Music (mp3), docs, Floppy disk
Gigabyte (GB)	10^9	Movies, large docs, Flash drives, DVD
Terabyte (TB)	10^{12}	Movies collection, Hard-drive
Petabyte (PB)	10^{15}	Networking websites, Financial, Retail
Exabyte (EB)	10^{18}	Large search and video websites
Zettabyte (ZB)	10^{21}	All data on the internet



https://www.engadget.com/

Character Sets and Encodings

- Character Set is a collection of supported characters
- Encoding maps a string of characters to a string of bytes

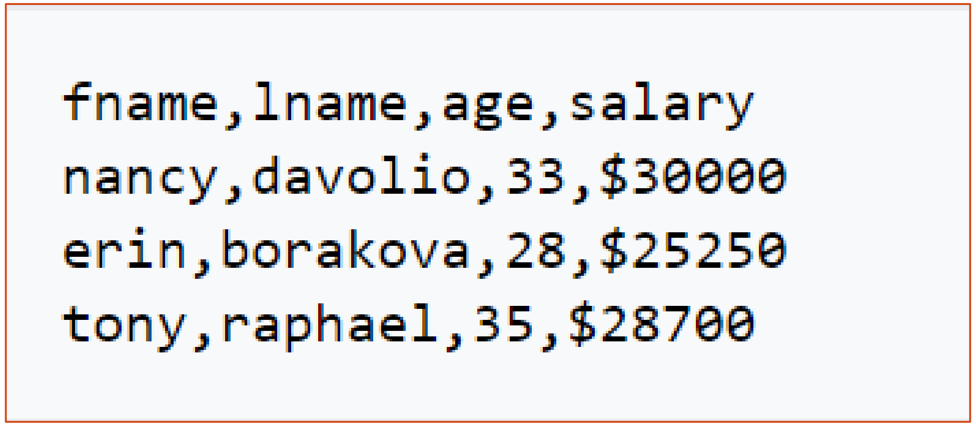

ⴰ ⵣ ⵓ ⵍ !
2D30 2D63 2D53 2D4D 21
Character encodings
UTF-16
2D 30 2D 63 2D 53 2D 4D 00 21
UTF-8
E2 B4 B0 E2 B5 A3 E2 B5 93 E2 B5 8D 21
UTF-32
00 00 2D 30 00 00 2D 63 00 00 2D 53 00 00 2D 4D 00 00 00 21

Source: w3c.org

Data	Encoding	Bits
A	ASCII /UTF-8	01000001
A	UTF-16	00000000 01000001
a	ASCII /UTF-8	01100001
7	ASCII /UTF-8	00000111
bits	UTF-8	01100010 01101001 01110100 01110011
あ	UTF-8	11100011 10000001 10000010 00001010
ऋ	UTF-8	11100000 10100100 10001011 00001010

Data Format – Text files

- Field delimiters separate data fields. Record delimiters separate groups of fields.
- Examples
  - Comma Separated Values (,)
  - Pipe Separated Values (|)
  - Tab Separated Values (tab)
  - Semicolon separated values (;)



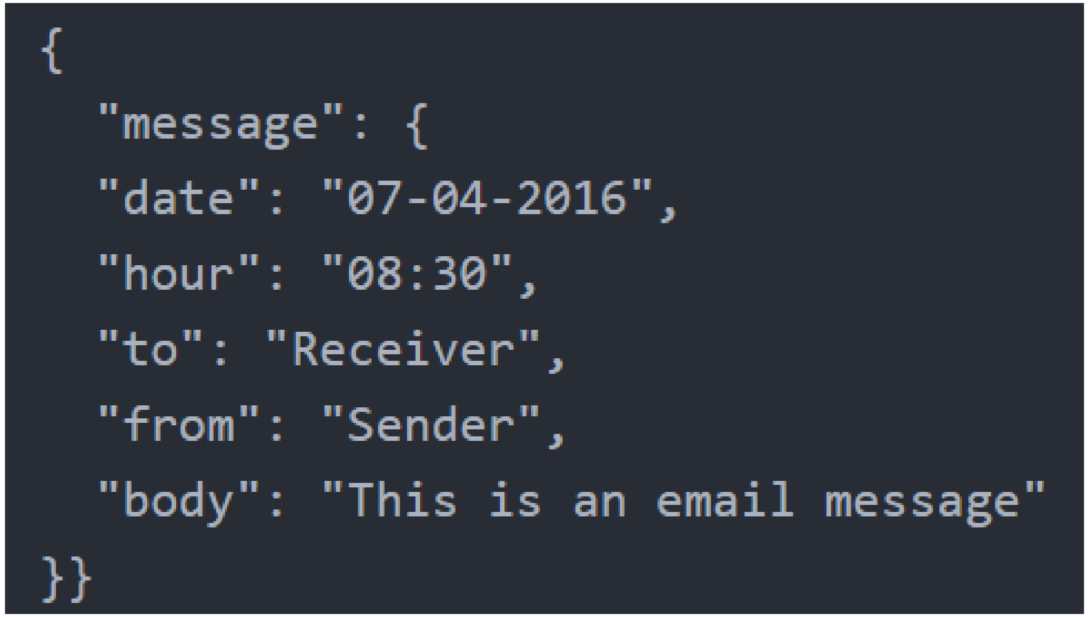
```
fname,lname,age,salary
nancy,davolio,33,$30000
erin,borakova,28,$25250
tony,raphael,35,$28700
```

CSV file

https://en.wikipedia.org/wiki/Delimiter/

Data Format - JavaScript Object Notation (JSON)

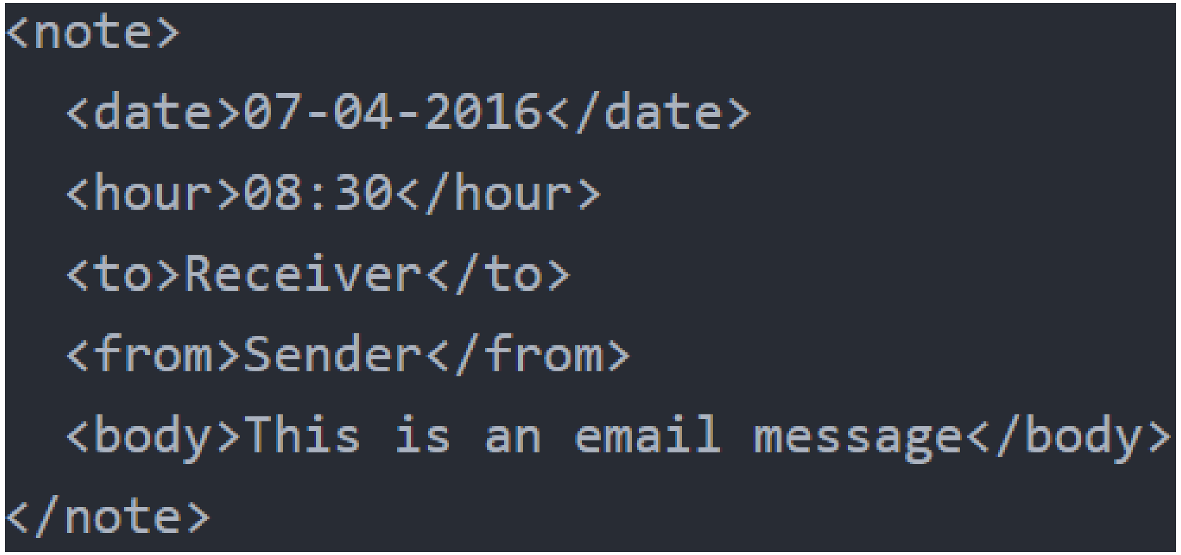
- Lightweight data-interchange format
- Platform independent
- Built on two structures :
  - Collection of name/value pairs
  - An ordered list of values



```
{
  "message": {
  "date": "07-04-2016",
  "hour": "08:30",
  "to": "Receiver",
  "from": "Sender",
  "body": "This is an email message"
}}
```

Data Format - Extensible Markup Language (XML)

- Lightweight data-interchange format
- Platform independent
- Self-descriptive
- Extensible

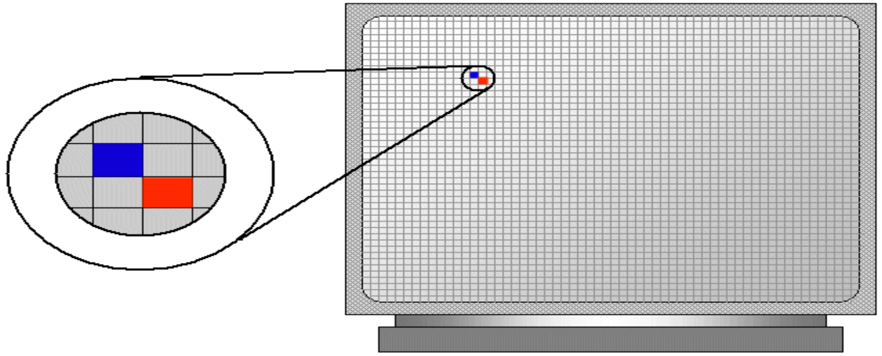


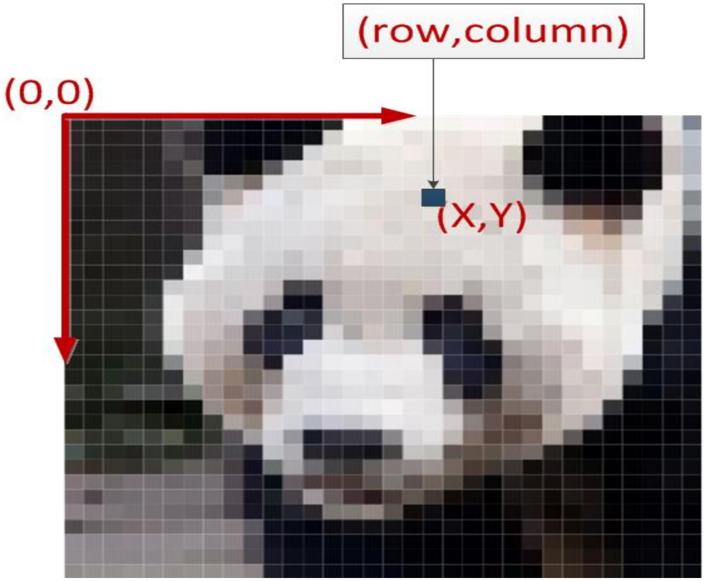
```
<note>
  <date>07-04-2016</date>
  <hour>08:30</hour>
  <to>Receiver</to>
  <from>Sender</from>
  <body>This is an email message</body>
</note>
```

Digital Images

- Raster Graphics
  - Information stored as pixels (e.g. JPEG)
- Vector Graphics
  - Information stored as vectors (e.g. SVG)

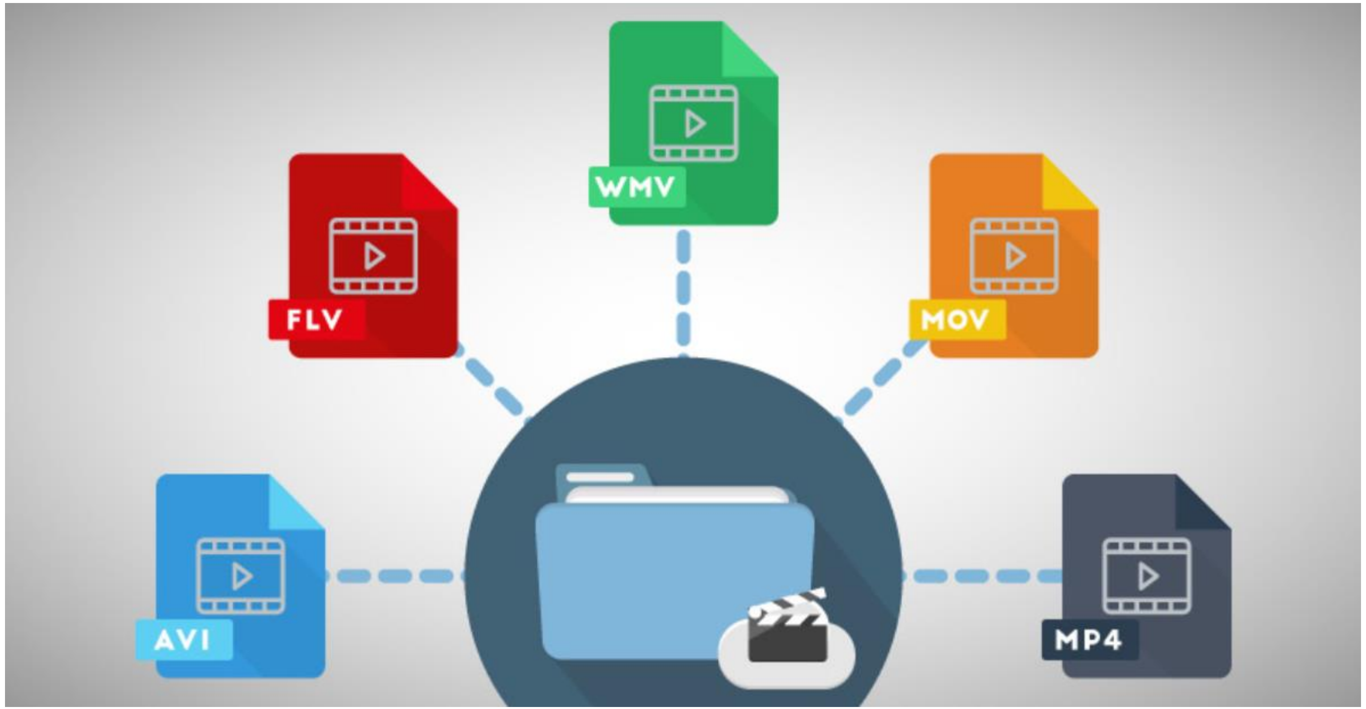
0	1	0	0	0	1	0
1	1	0	1	1	0	0
0	1	1	0	0	0	0
0	1	1	1	1	0	0
1	1	0	1	0	1	0
0	0	0	0	0	0	1





Video Formats

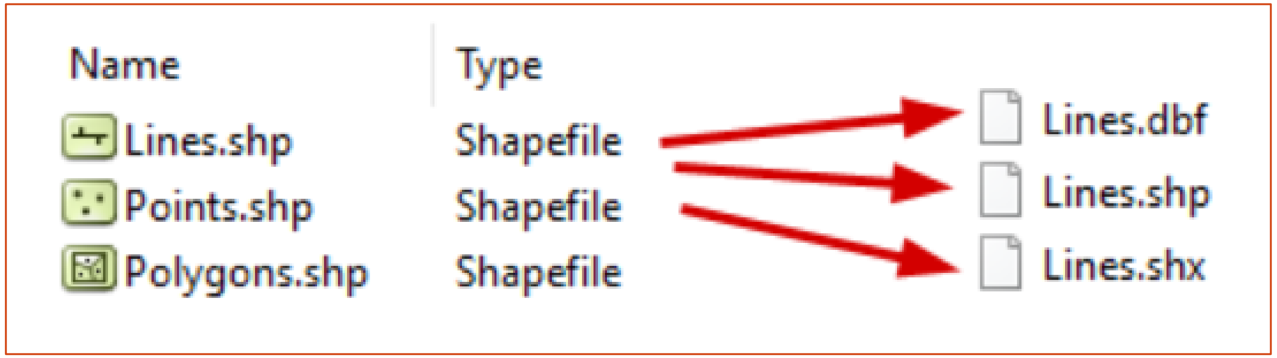
- File format for storing digital video data on a computer system. Video is almost always stored in compressed form to reduce the file size.



https://www.motionelements.com/blog/articles/what-you-need-to-know-about-the-5-most-common-video-file-formats/

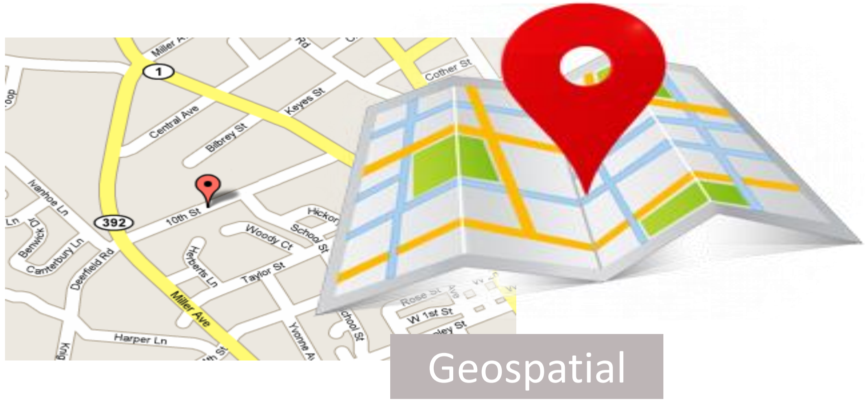
Spatial data

- Spatial data (geospatial data) represents the location, size and shape of an object such as a building, lake, mountain and includes attributes that provide more information about the entity being represented.
- The shapefile is the most common geospatial file type. The three required files are SHP is the feature geometry, SHX is the shape index position and DBF is the attribute data.



Name	Type
Lines.shp	Shapefile
Points.shp	Shapefile
Polygons.shp	Shapefile

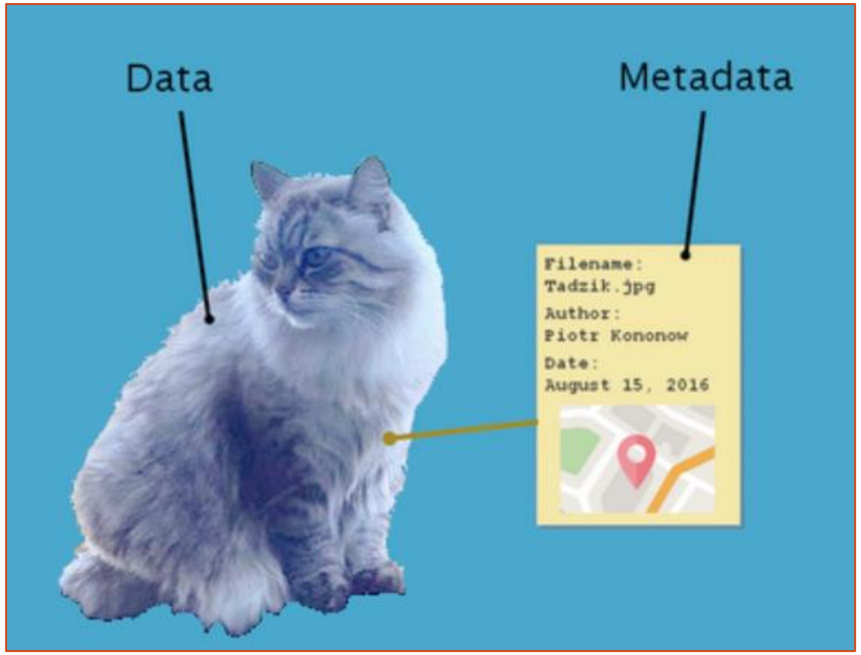
Lines.dbf
Lines.shp
Lines.shx



https://gisgeography.com/gis-formats/

Metadata

- Data that describes the structure and the properties of the data. It is essential for the proper understanding and use of the data
- Makes finding and working with particular instances of data easier

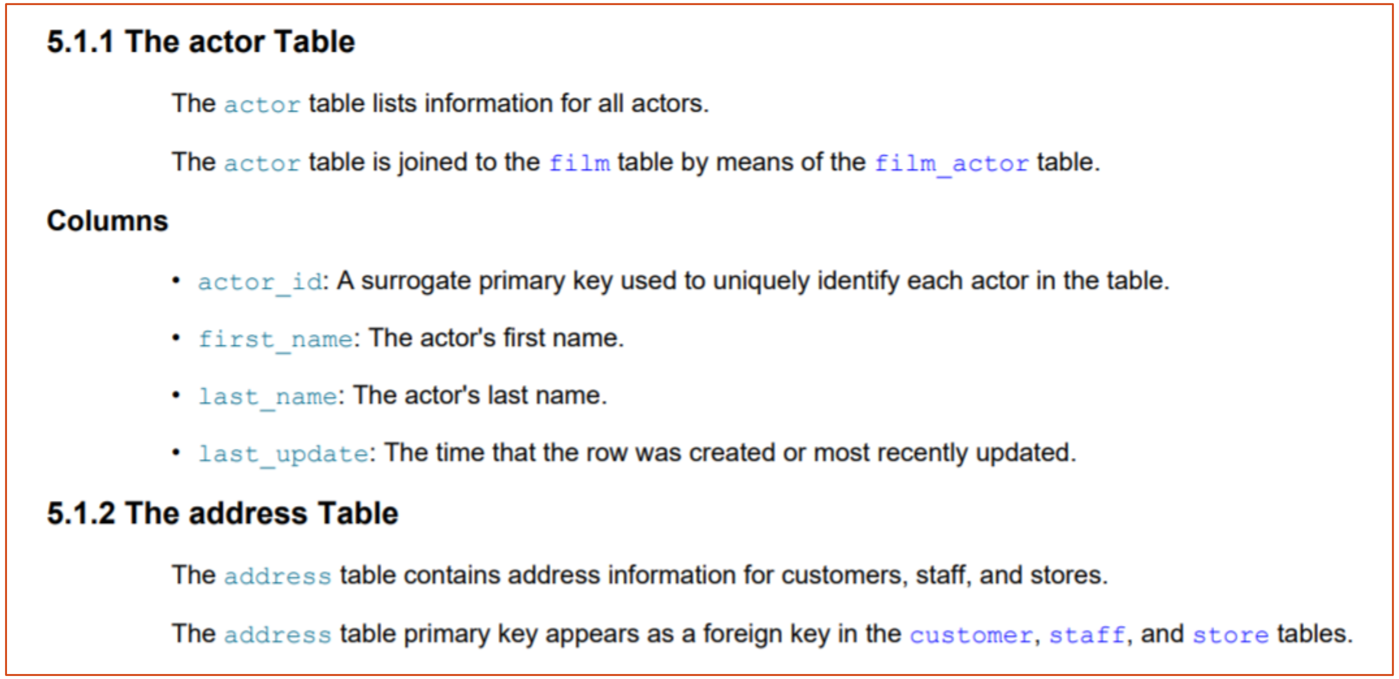


https://gisgeography.com/gis-formats/

- Data name
- Data create date
- Creator’s name
- Data owner
- Data sensitivity
- Group/user permissions
- Source of data
- Construction process of the data

Data Dictionary

- A "centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format". Also known as metadata repository.



5.1.1 The actor Table

The `actor` table lists information for all actors.

The `actor` table is joined to the `film` table by means of the `film_actor` table.

Columns

- `actor_id`: A surrogate primary key used to uniquely identify each actor in the table.
- `first_name`: The actor's first name.
- `last_name`: The actor's last name.
- `last_update`: The time that the row was created or most recently updated.

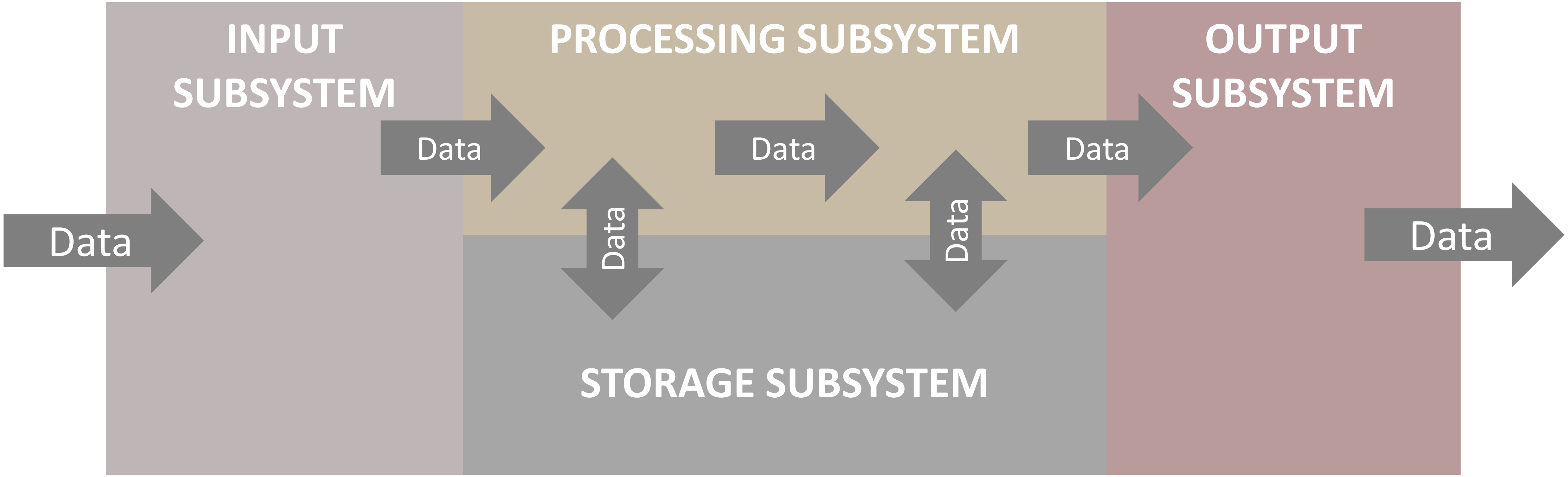
5.1.2 The address Table

The `address` table contains address information for customers, staff, and stores.

The `address` table primary key appears as a foreign key in the `customer`, `staff`, and `store` tables.

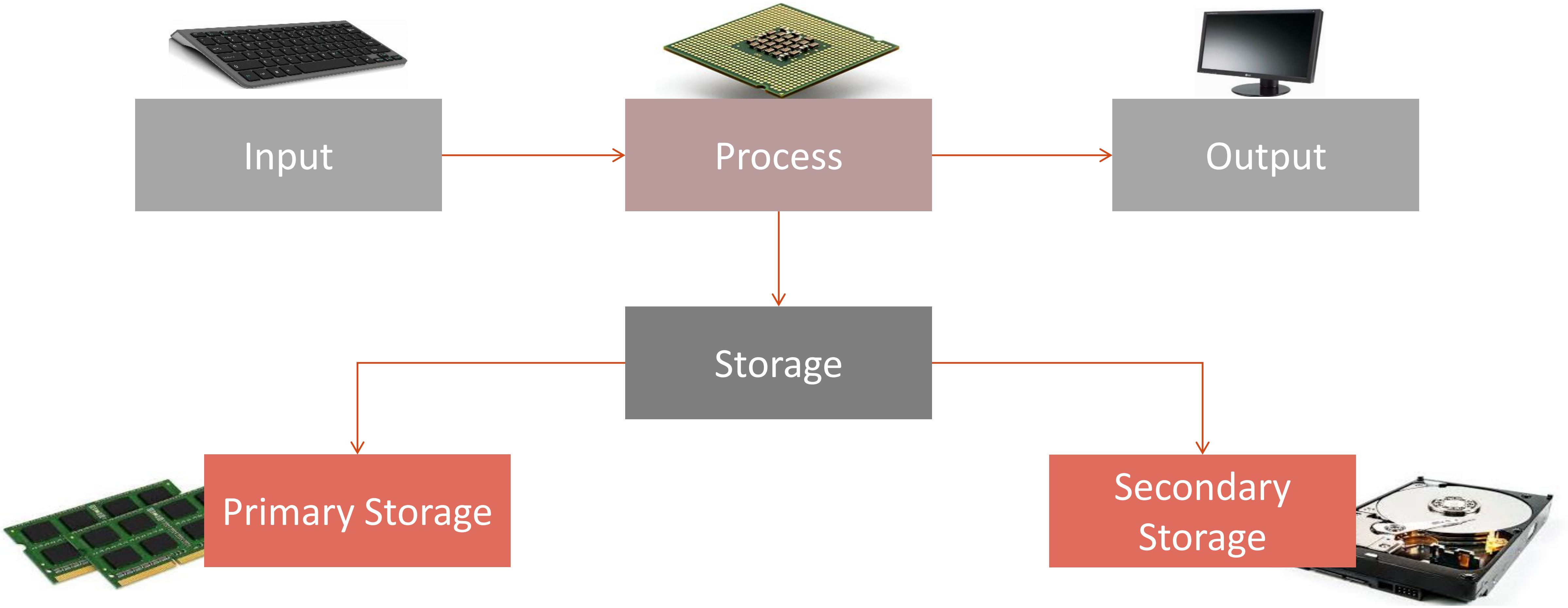
Data Preparation

Data Processing



Source: wikipedia

Data Processing



Input

Process

Output

Storage

Primary Storage

Secondary Storage

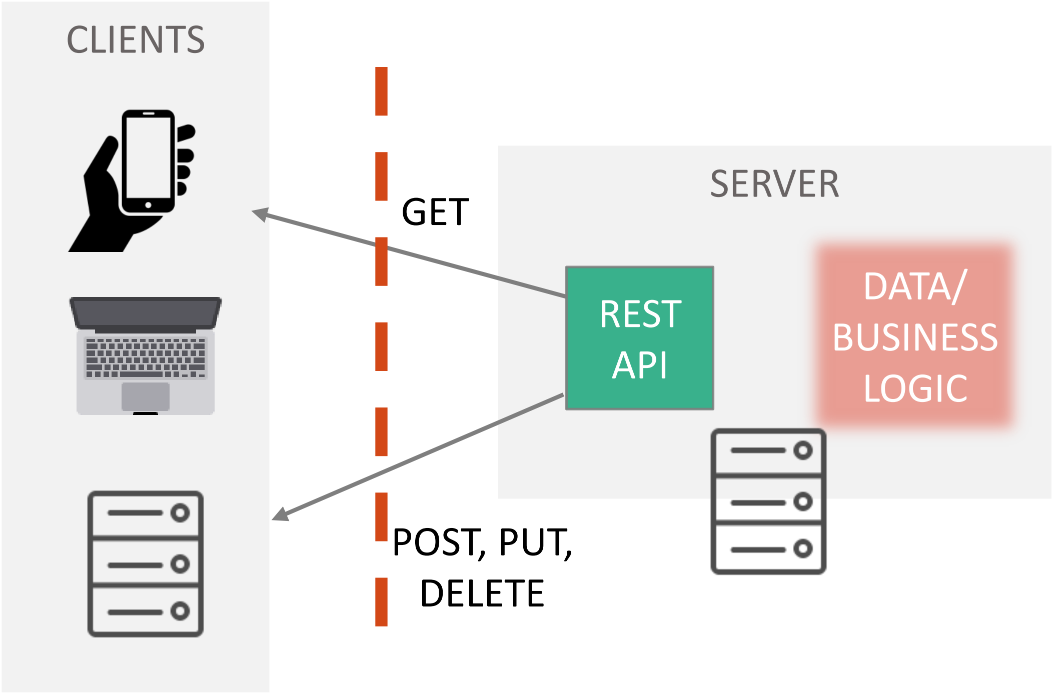
Application Programming Interfaces (APIs)

- An API is code that allows two software programs to communicate with each other
- APIs are software components or building blocks, which are used by other systems. Usually contain clearly defined methods of communication
- An API may be for a web-based system, operating system, database system, or hardware/software library

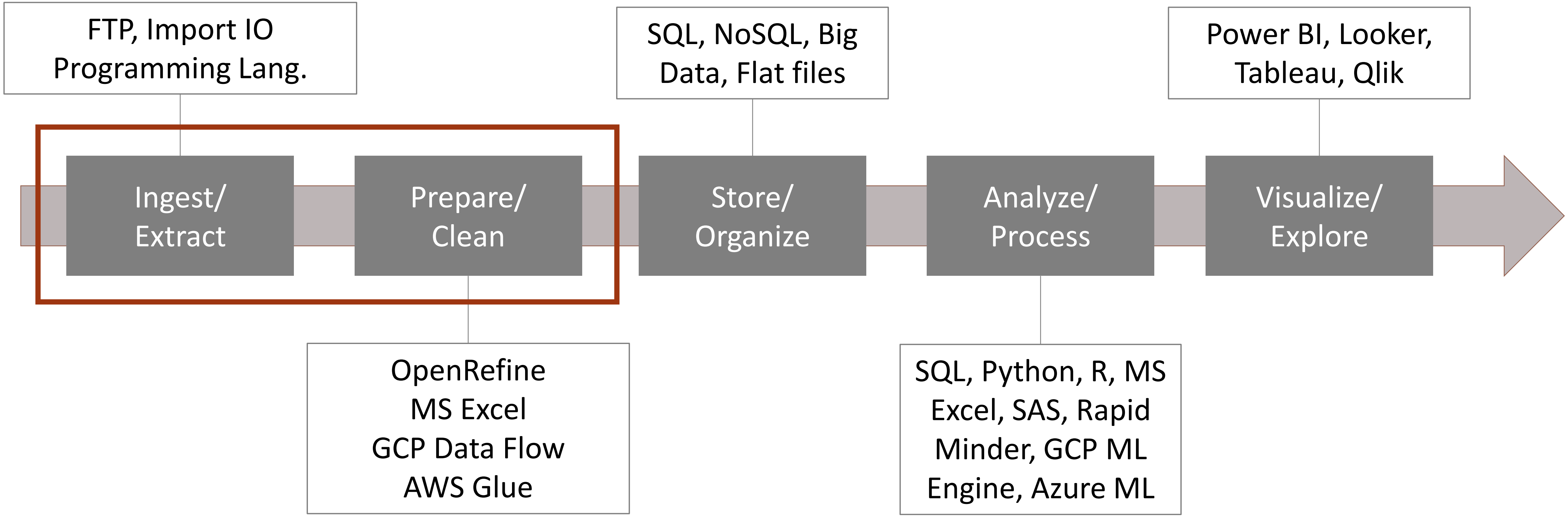
LTX Solutions

Web Services

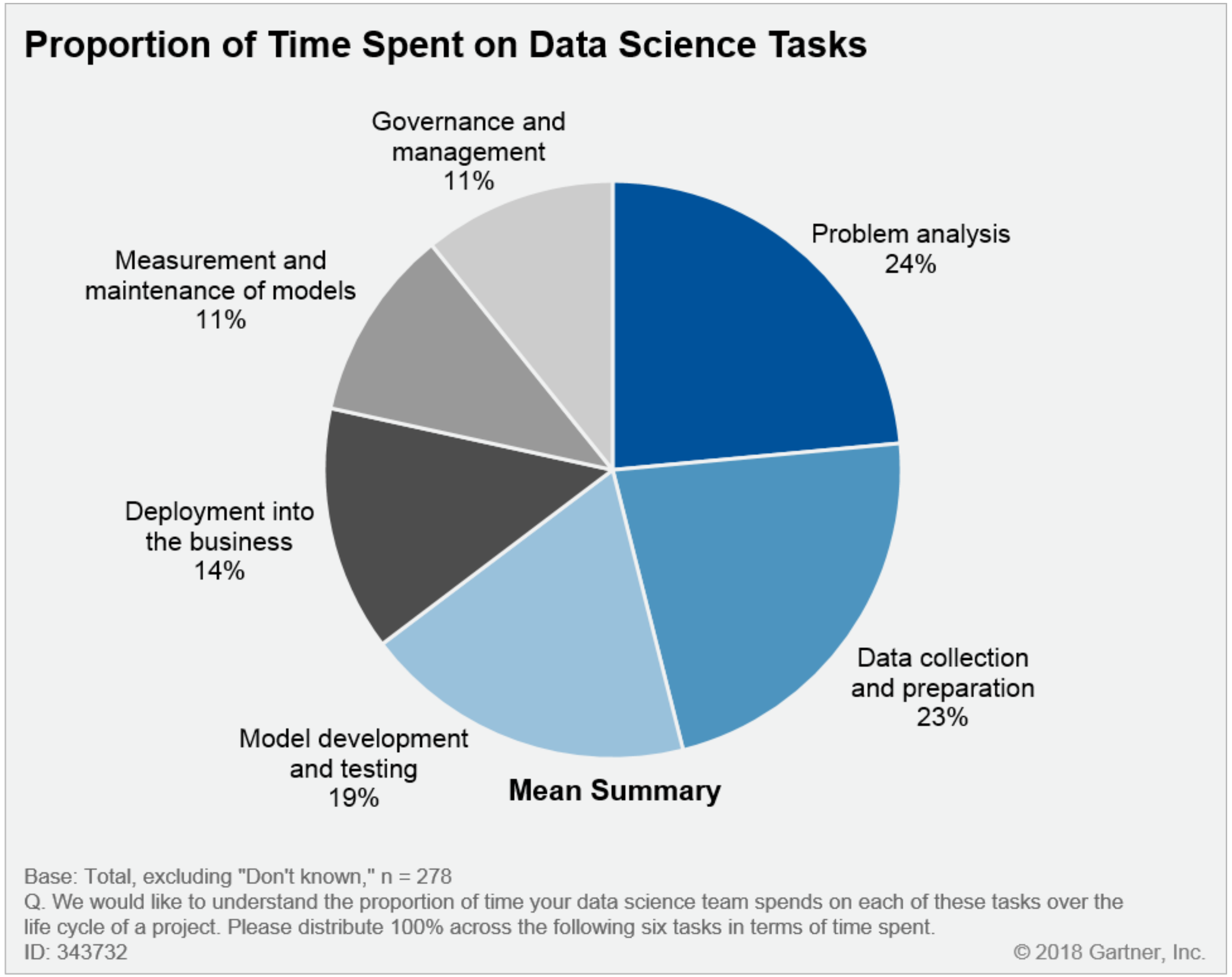
- Web Services are services or APIs offered over the internet
- REST/SOAP are widely used web services protocols
- Data exchanged between client and the server is typically in the JSON/XML format
- E.g. Weather Service, Stock Quote Service, etc.
- A curated list of web services https://www.programmableweb.com/category/all/apis



Data Analysis Pipeline – Ingest, Prepare



Data Science Tasks



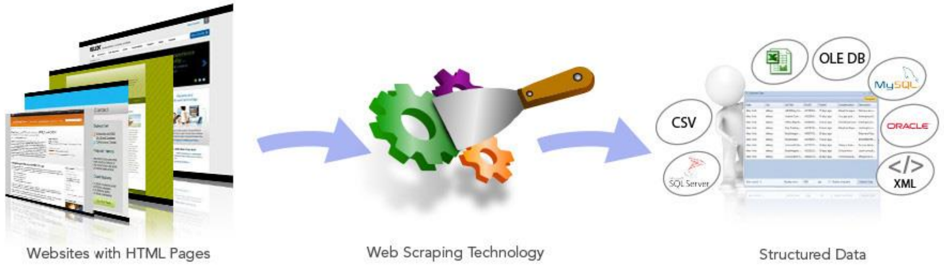
Proportion of Time Spent on Data Science Tasks

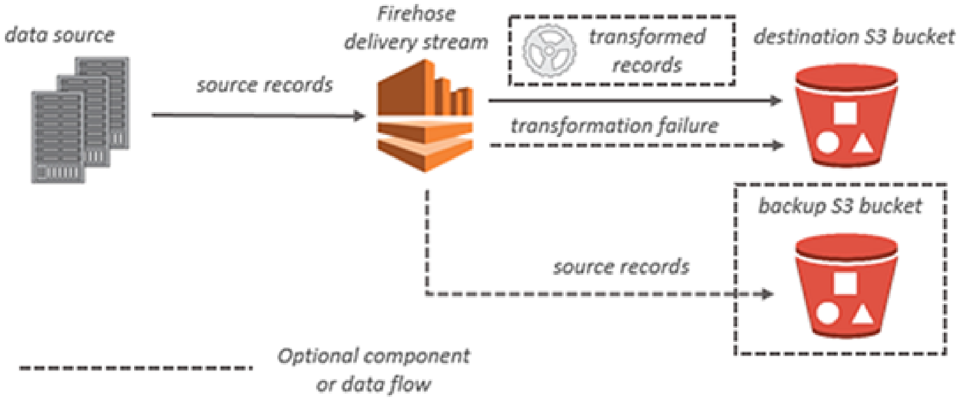
Base: Total, excluding "Don't known," n = 278
Q. We would like to understand the proportion of time your data science team spends on each of these tasks over the life cycle of a project. Please distribute 100% across the following six tasks in terms of time spent.
ID: 343732

© 2018 Gartner, Inc.

Data Ingestion

- Process of retrieving data from various sources for further processing
- Various methods of performing data ingestion:
  - File transfers (FTP, Downloads)
  - Web service (REST API)
  - Event based messaging (Publish-Subscribe, Queues, Topics)
  - Extraction Transformation Load (ETL)
  - Web Mining - scraping/crawling





Exercise – Web Scraping

Read and process data from a website

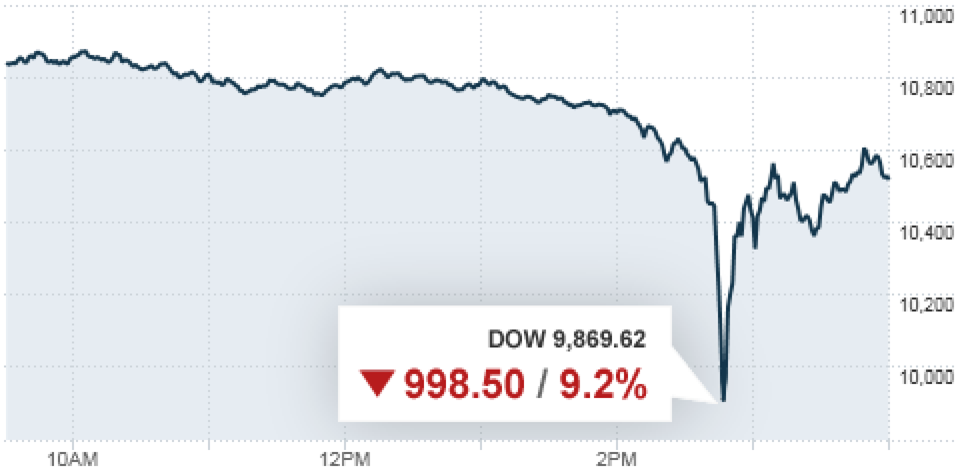
Raw Data

- Can we interpret the data ?
- Does data have errors ?
    - Typos, multiple formats, inconsistencies, missing values
- Does it meet domain standards ?
    - Currency, decimals, censored fields, etc.

```
112, ND748579, 2014-08-20, 33347.88
114, GG31455, 2016-05-20, 45864.03
114, MA765515, 2014-12-15, 82261.22
```

Sources of Errors

- Manual data entry
- Duplicate data entry
- Measurement related errors
- Absence of well defined standards
- Inconsistent data formatting
- Numeric approximations
  - software and hardware constraints



At 2:45 pm on May 6, 2010, Wall Street essentially had a heart attack. In just minutes, the stock market plunged 1000 points, for reasons traders, analysts, and business media could not explain. The "flash crash" wiped out $1.1 Trillion of investor dollars and even though most of that was quickly regained, it left the market badly shaken.

..It appears that a single keystroke error was to blame. The letter "B" was inserted in a sell order instead of the letter "M". Billion was input where Million should have been and it triggered a ripple effect through the automated financial markets.

https://ungerboeck.com

Data Quality

- Data Quality
  - Fitness of the data for its intended use in operations, decision making and planning
- Impact of data quality :
  - Consistency - all copies of the data are consistent with each other
  - Completeness – required fields are not missing or incomplete
  - Accuracy - data is correct and has been verified
  - Validity - data adheres to types, formats and business rules
  - Timeliness - data is not out of date
  - Integrity - data is appropriately referenced

Data Transformation

- Trim white spaces
- Handling missing or null values
- Formatting text, numeric and date fields
- Standardization and consistency
- Correcting domain values
    - Selecting certain attributes
    - Sorting and aggregation
    - Deriving calculated values
    - Statistical adjustments

talend

OPEN Refine

TRIFACTA
PEOPLE DATA COMPUTATION

Informatica

OpenRefine

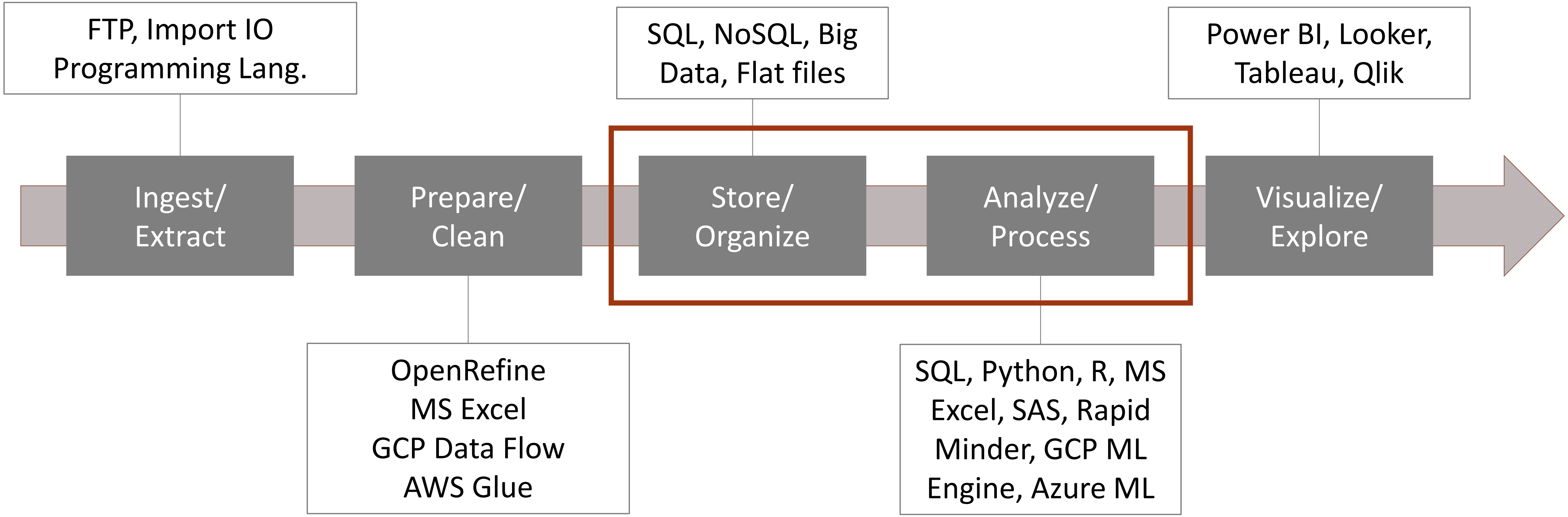
- Open source data cleaning application
- Compatible with a variety of formats
  - CSV, TSV, JSON, XML, Excel, RDF
- More powerful than spreadsheets; better visualization than programming
- Functions:
  - Import/Export, Faceting, Transforming, Clustering, Reconciling

Exercise - OpenRefine

Data preparation using OpenRefine

Database Management Systems

Data Analysis Pipeline – Store, Analyze


FTP, Import IO Programming Lang.

Ingest/ Extract

Prepare/ Clean

OpenRefine
MS Excel
GCP Data Flow
AWS Glue

SQL, NoSQL, Big Data, Flat files

Store/ Organize

Analyze/ Process

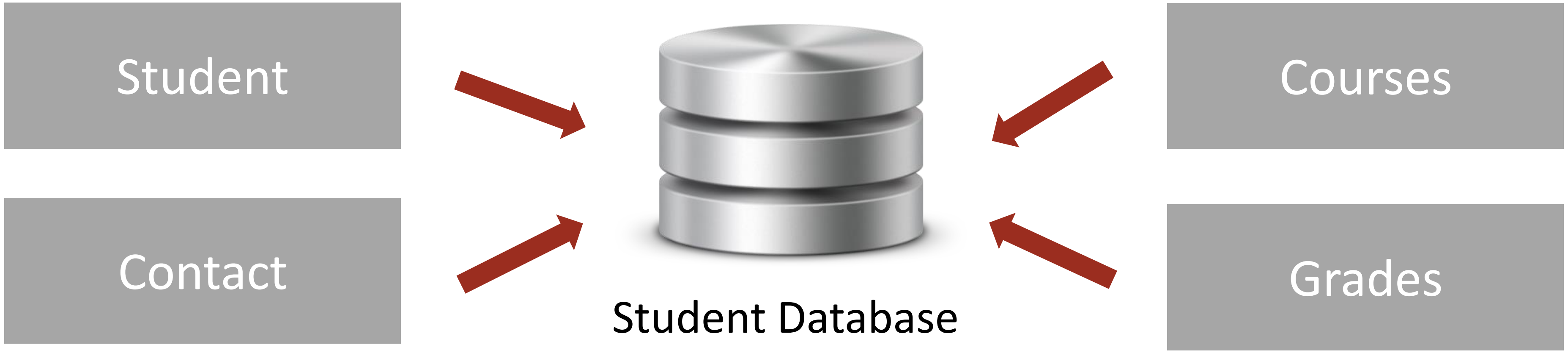
SQL, Python, R, MS Excel, SAS, Rapid Minder, GCP ML Engine, Azure ML

Power BI, Looker, Tableau, Qlik

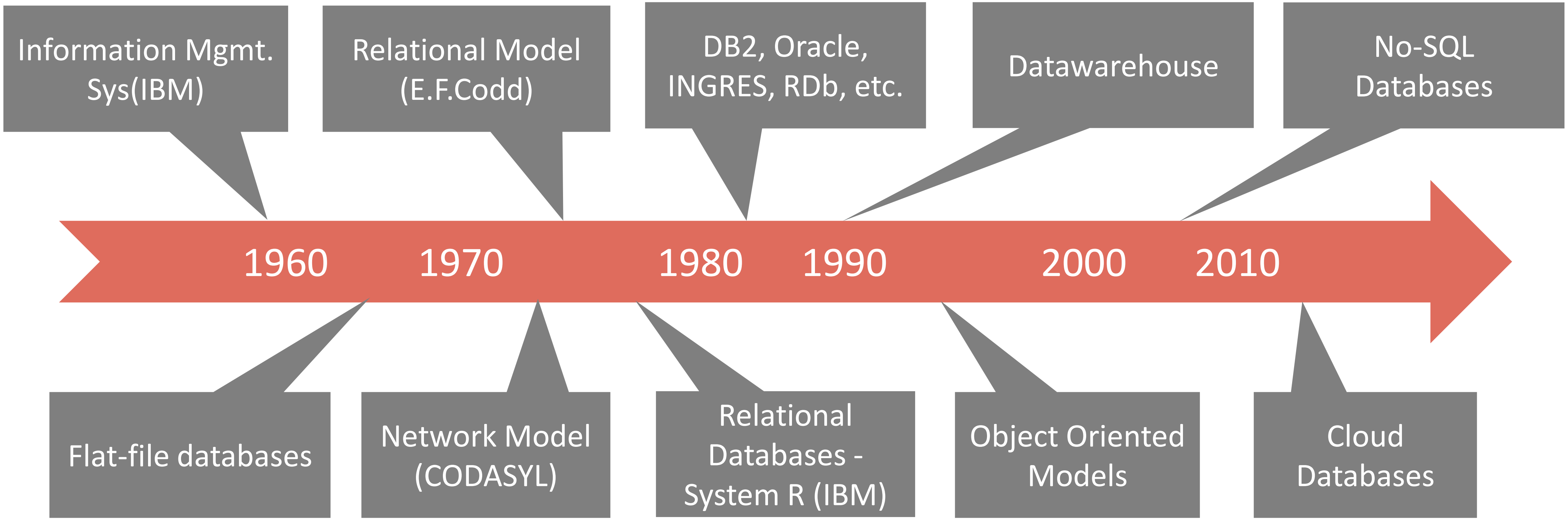
Visualize/ Explore

Database

- Shared collection of logically related data and descriptions of this data
- Modeled after real-world systems and entities



History of Databases



Principle of Data Independence

Logical Database Schema

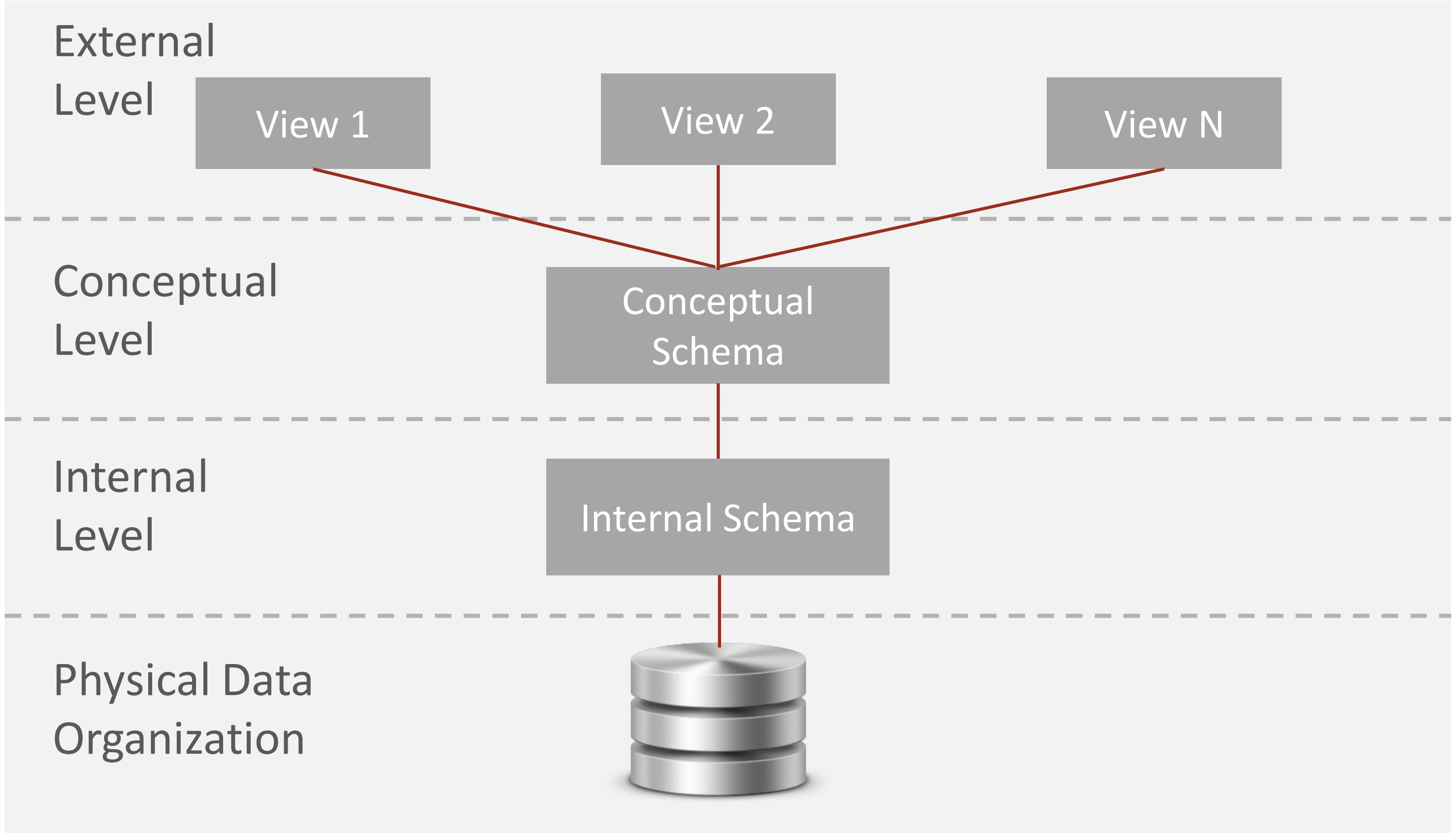
- A logical schema is a conceptual model of the data. It is primarily concerned with understanding the business entities, their attributes and their relationships
- Logical schema is design-centric in order to meet the business requirements

Physical Database Schema

- A physical model is concrete in the sense that it is implemented in the DBMS
- It contains physical objects such as data files, redo logs, control files etc., that reside on a database system

The ability to modify a schema definition in one level without affecting a scheme definition in a higher level is called **data independence.** The principle of data independence helps keep the logical model separate from the physical schema so that the database model/design can be isolated from the physical constraints

Logical Architecture (ANSI-SPARC)



ANSI-SPARC Architecture

Database Management System (DBMS)

- Aggregate of data, software, hardware and users that helps an organization manage its operational data
- Provide efficient and reliable methods of data retrieval
- Involves monitoring, administration, and maintenance of the databases

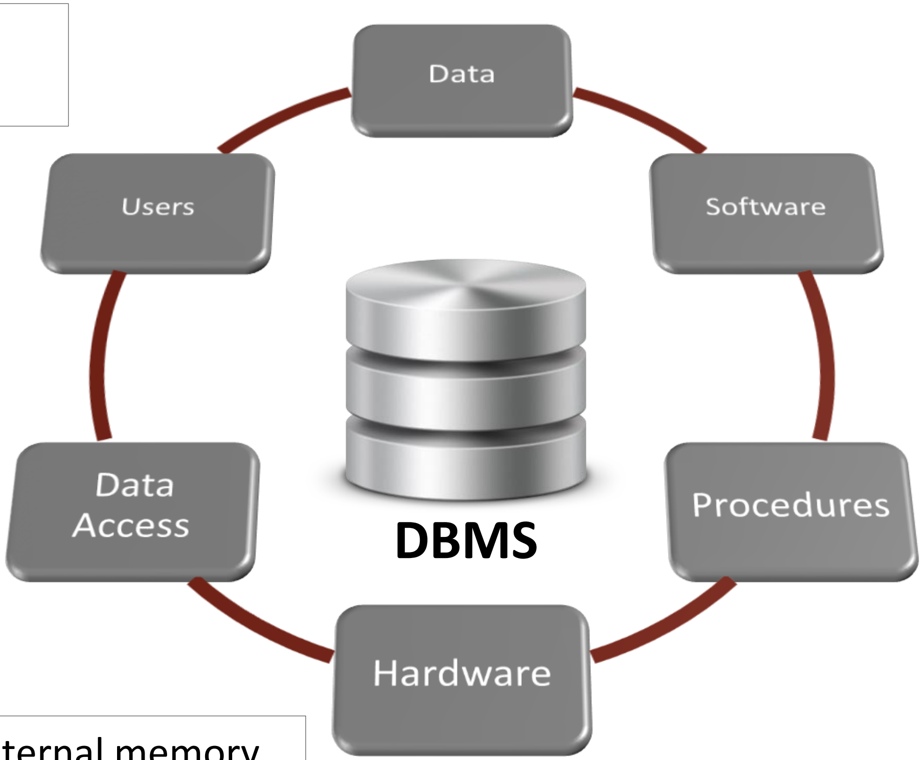
DBMS Benefits

- Program-Data Independence
- Efficiently retrieve, manipulate, store data
- Concurrent Access & Crash Recovery
- Minimize data duplication
- Data Integrity & Security
- Data Administration

DBMS Components

Administrators
Designers, Users

Query Processor
Memory Mgr.
File Systems

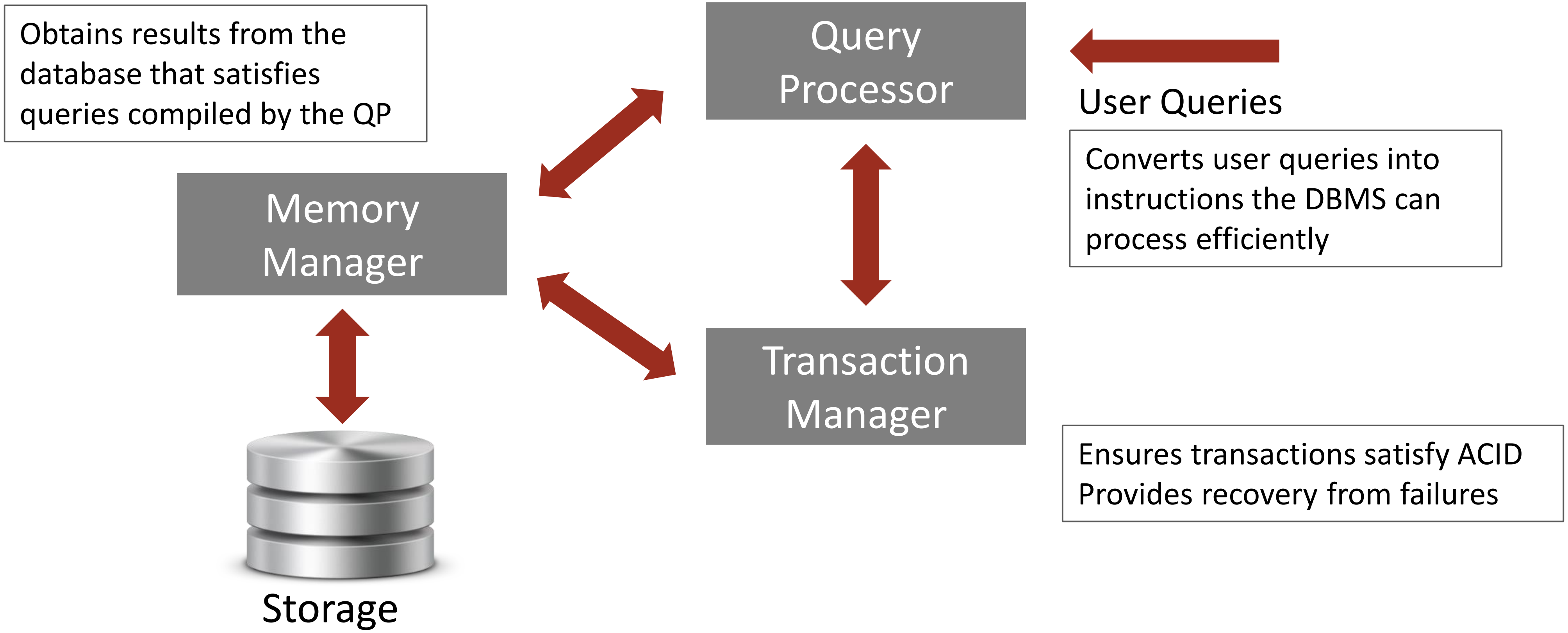


DDL, DCL, DML

Authentication,
Install, Backup

Internal memory
External memory

DBMS Software Architecture



Obtains results from the database that satisfies queries compiled by the QP

Memory Manager

Storage

Query Processor

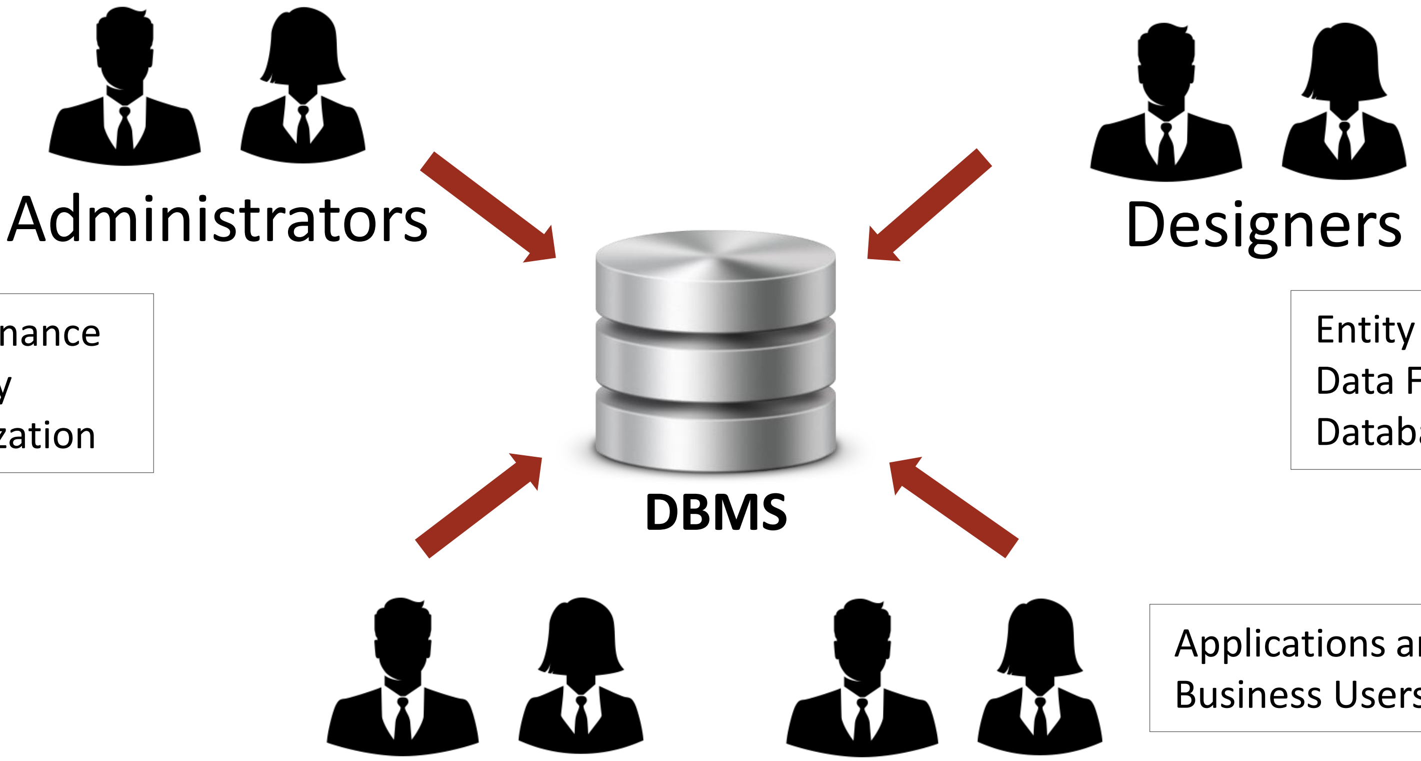
User Queries

Converts user queries into instructions the DBMS can process efficiently

Transaction Manager

Ensures transactions satisfy ACID
Provides recovery from failures

DBMS Users

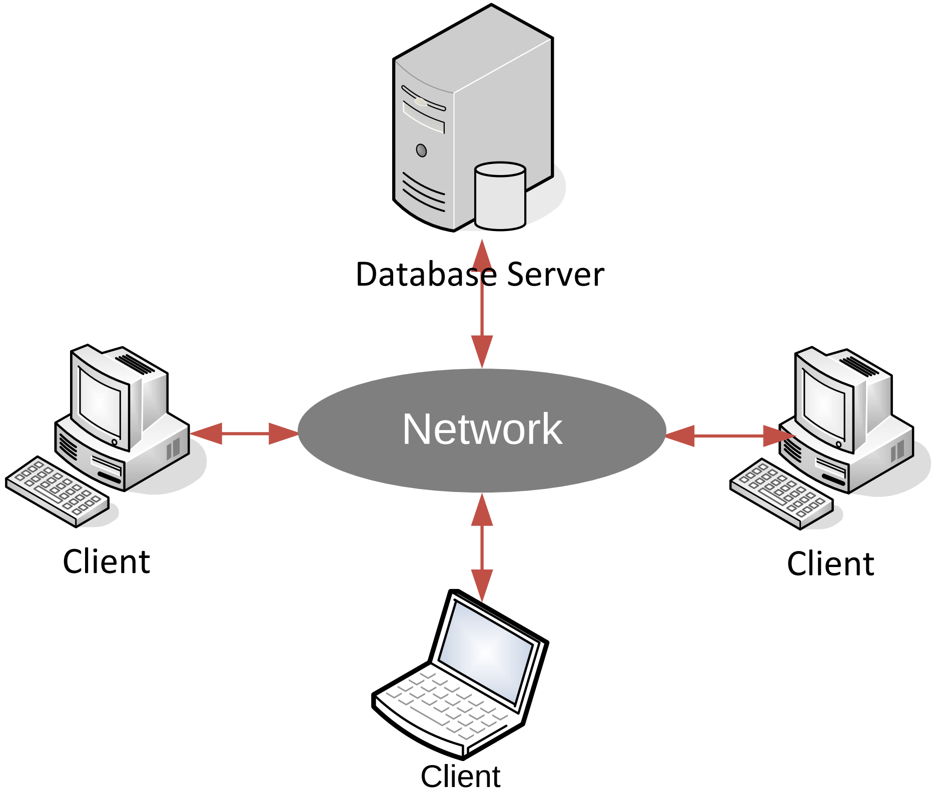


Maintenance
Security
Optimization

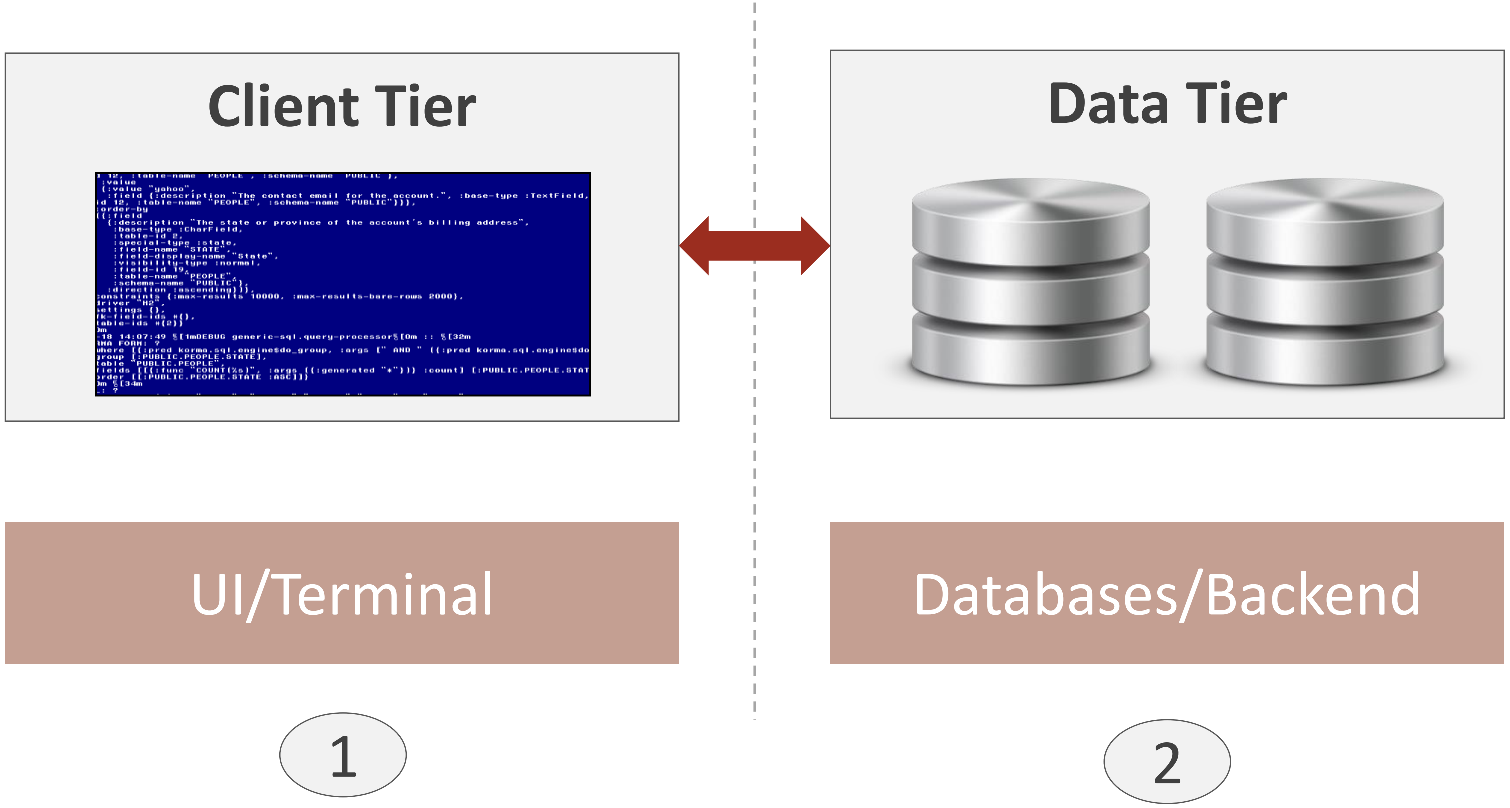
Entity Modeling
Data Formatting
Database Content

Applications and
Business Users

Client-Server Architecture



2-tier Architecture (Logical)



```
 12, :table-name "PEOPLE", :schema-name "PUBLIC"},
 :value
 {:value "yahoo",
  :field {:description "The contact email for the account.", :base-type :TextField,
id 12, :table-name "PEOPLE", :schema-name "PUBLIC"}}},
:order-by
({:field
  {:description "The state or province of the account's billing address",
   :base-type :CharField,
   :table-id 2,
   :special-type :state,
   :field-name "STATE",
   :field-display-name "State",
   :visibility-type :normal,
   :field-id 19,
   :table-name "PEOPLE",
   :schema-name "PUBLIC"},
  :direction :ascending}]},
constraints {:max-results 10000, :max-results-bare-rows 2000},
driver "H2",
settings {},
fk-field-ids #{},
table-ids #{2}}
0m
-18 14:07:49 %[1mDEBUG generic-sql.query-processor%[0m :: %[32m
RMA FORM: ?
where [{:pred korma.sql.engine$do_group, :args [" AND " ({:pred korma.sql.engine$do
group [:PUBLIC.PEOPLE.STATE],
table "PUBLIC.PEOPLE",
fields [[{:func "COUNT(%s)", :args ({:generated "*"})} :count] [:PUBLIC.PEOPLE.STAT
order [[:PUBLIC.PEOPLE.STATE :ASC]]}
0m %[34m
-: ?
```

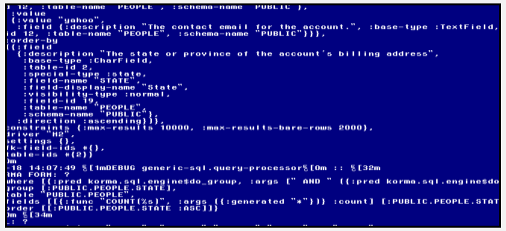
3-tier Architecture (Logical)

Presentation Tier

GUI/Frontend

1

Application Tier



```
l 12, :table-name "PEOPLE", :schema-name "PUBLIC"},
 :value
 {:value "yahoo",
  :field {:description "The contact email for the account.", :base-type :TextField,
id 12, :table-name "PEOPLE", :schema-name "PUBLIC"}}},
:order-by
[{:field
  {:description "The state or province of the account's billing address",
   :base-type :CharField,
   :table-id 2,
   :special-type :state,
   :field-name "STATE",
   :field-display-name "State",
   :visibility-type :normal,
   :field-id 19,
   :table-name "PEOPLE",
   :schema-name "PUBLIC"},
  :direction :ascending}]},
constraints {:max-results 10000, :max-results-bare-rows 2000},
driver "H2",
settings {},
fk-field-ids #{},
table-ids #{2}}
0m
-18 14:07:49 %[1mDEBUG generic-sql.query-processor%[0m :: %[32m
RMA FORM: ?
where [{:pred korma.sql.engine$do_group, :args [" AND " ({:pred korma.sql.engine$do
group [:PUBLIC.PEOPLE.STATE],
table "PUBLIC.PEOPLE",
fields [[{:func "COUNT(%s)", :args ({:generated "*"})} :count] [:PUBLIC.PEOPLE.STAT
order [[:PUBLIC.PEOPLE.STATE :ASC]]}
0m %[34m
.: ?
```

Business Logic

2

Data Tier

Databases/Backend

3

Database Transactions

- A unit of work performed against a database
- Generally results in a change of database state
- Reliable and independent of other transactions
- Either executes in entirety or is rolled back
- ACID properties

ACID Properties

- Atomicity
  - When an update occurs to a database, if part of the transaction fails, then the entire transaction fails. Atomicity requires that each transaction be an "all or nothing”.
- Consistency
  - Ensures that any transaction beings the database from one valid state to another. A consistency constraint is a predicate on data which serves as a precondition, post-condition, and transformation condition on any transaction
- Isolation
  - Ensures that concurrent execution of transactions results in a system state that would be obtained if the transactions were executed serially, i.e. one after the other
- Durability
  - Ensures that the system is able to recover committed transaction updates if either the system crashes or fails with errors

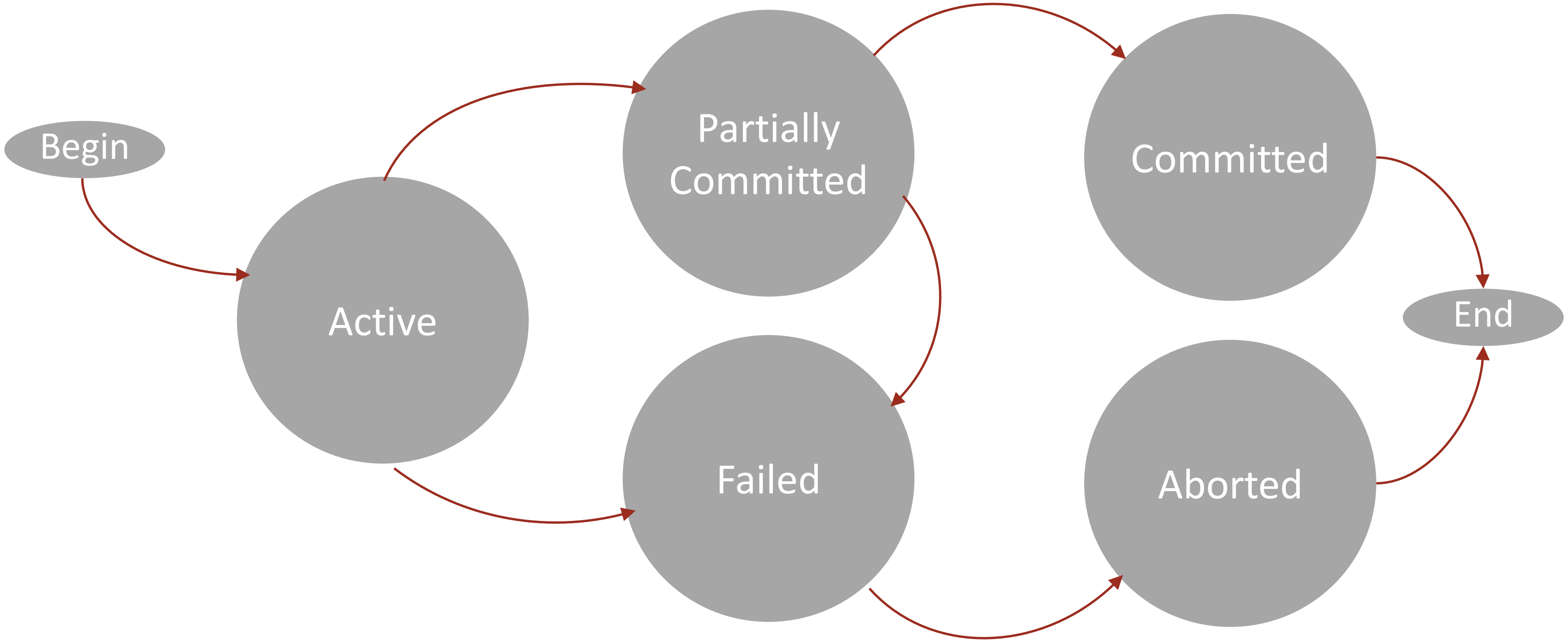
Transaction Example

ATM

Withdrawal request	Account Balance
1) Verify account details	$ 1000
2) Accept withdrawal request	$ 1000
3) Check balance	$ 1000
4) Update balance	$ 900
5) Dispense money	

Crash

Transaction Lifecycle



Exercise

MySQL DBMS Review

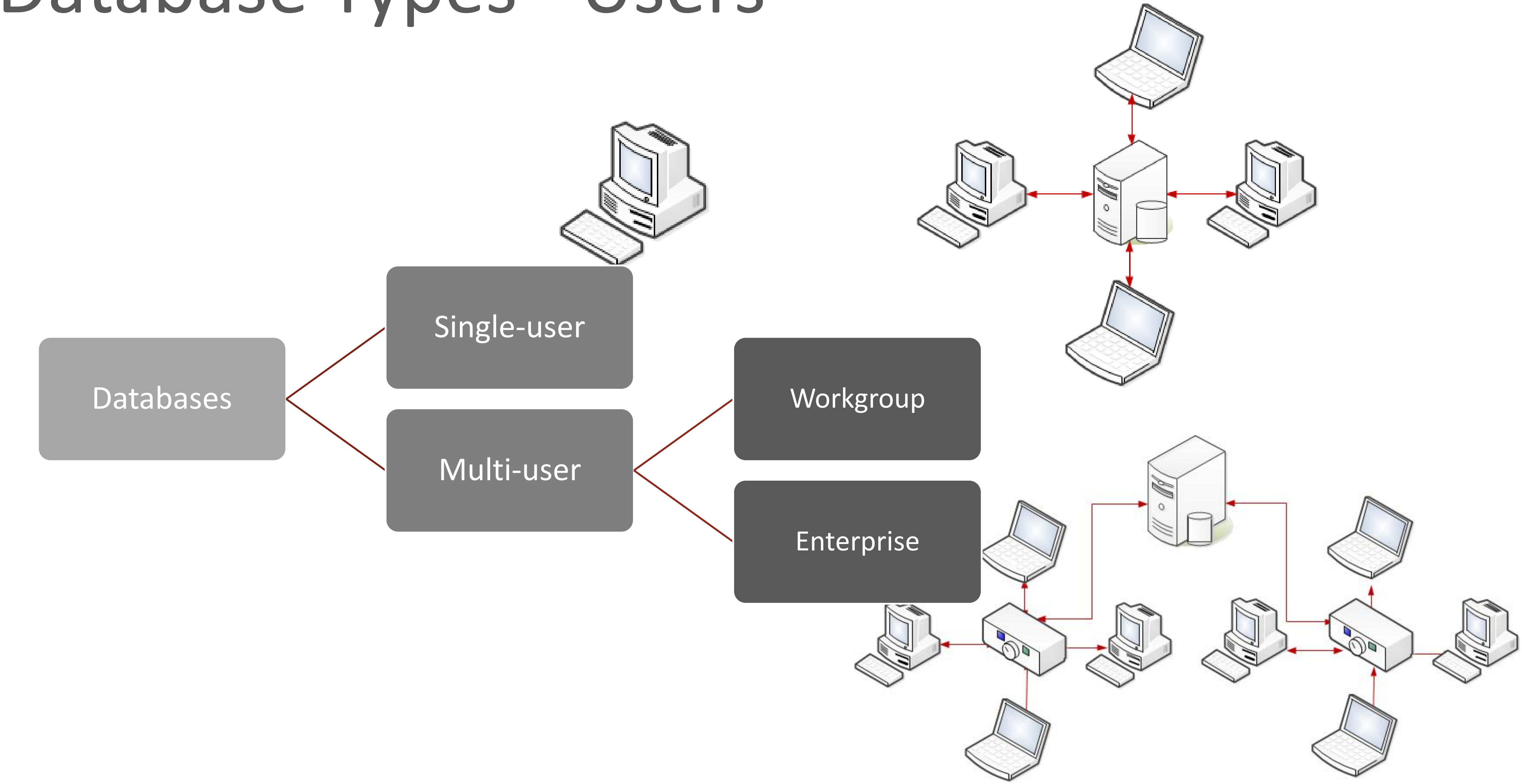
Types of Databases

Users, Location, Function, Supported Data Types

Database Classification

- Databases are usually classified based on the following attributes:
  - Number of users
    - Single user, multi-user, large scale
  - Primary function
    - Transactional vs analytical
  - Database location(s)
    - Centralized vs distributed
  - Data type and structure
    - Structured, semi-structured

Database Types - Users



Databases

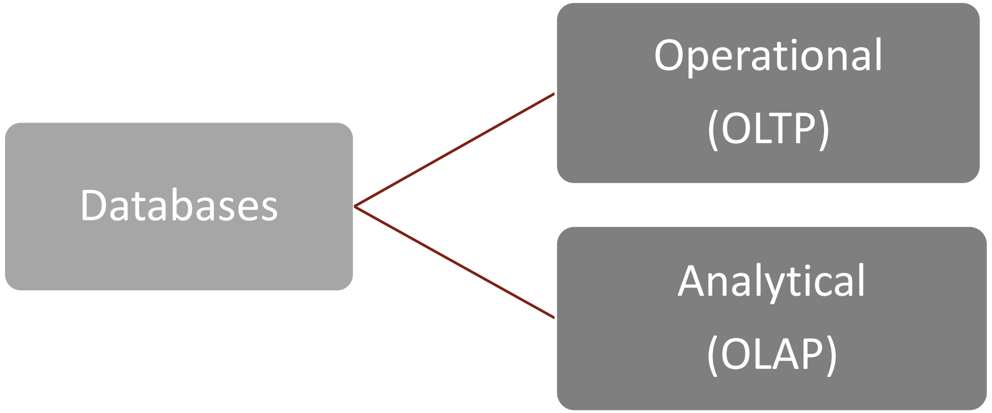
Single-user

Multi-user

Workgroup

Enterprise

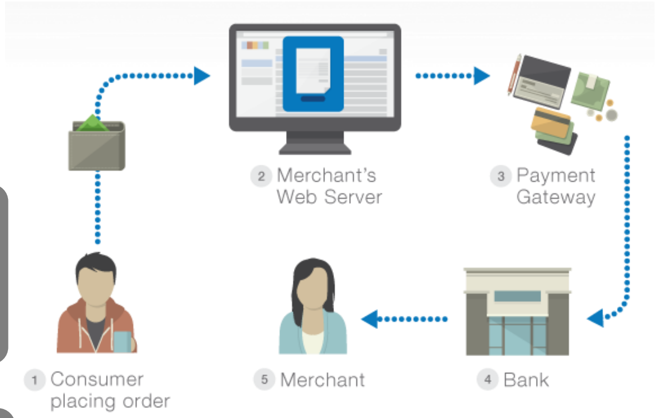
Database Types - Function

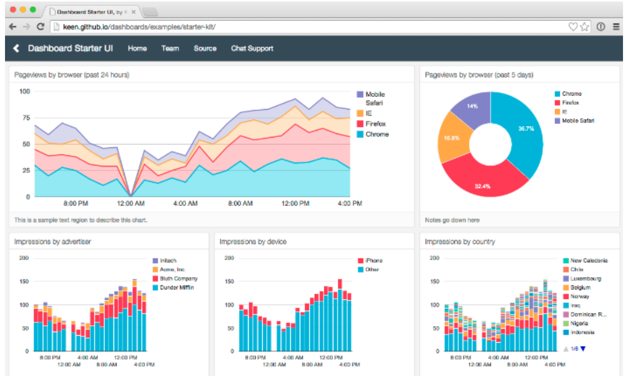


Databases

Operational (OLTP)

Analytical (OLAP)

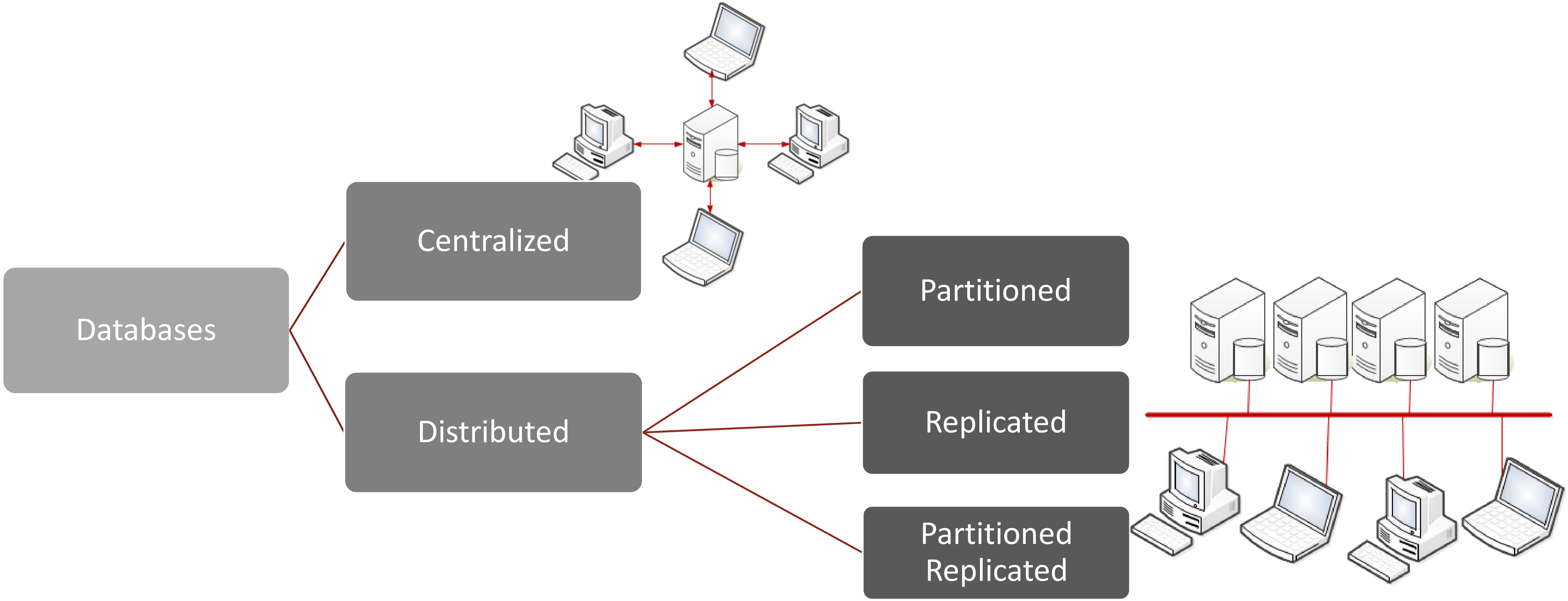



keen.github.io/dashboards/examples/starter-kit/
Dashboard Starter UI
Home
Team
Source
Chat Support
Pageviews by browser (past 24 hours)
Mobile Safari
IE
Firefox
Chrome
Pageviews by browser (past 5 days)
Impressions by advertiser
Impressions by device
Impressions by country

This is a sample text region to describe this chart.

Notes go down here

Database Types - Location(s)



Databases

Centralized

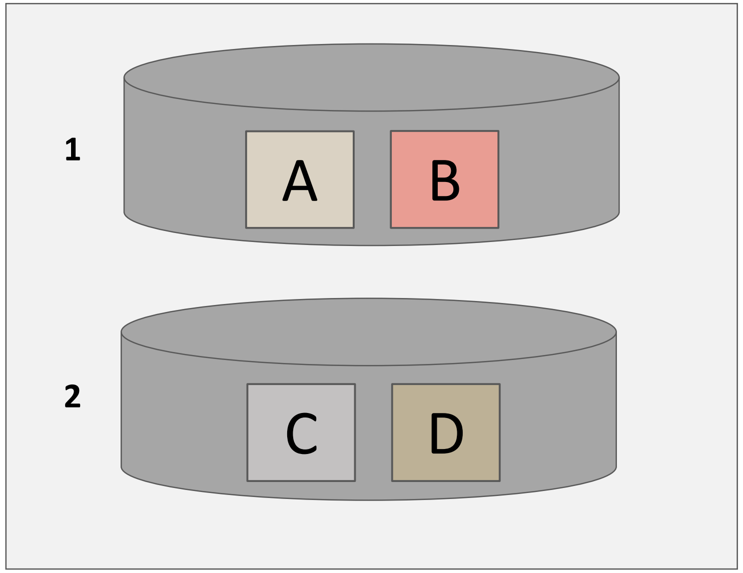
Distributed

Partitioned

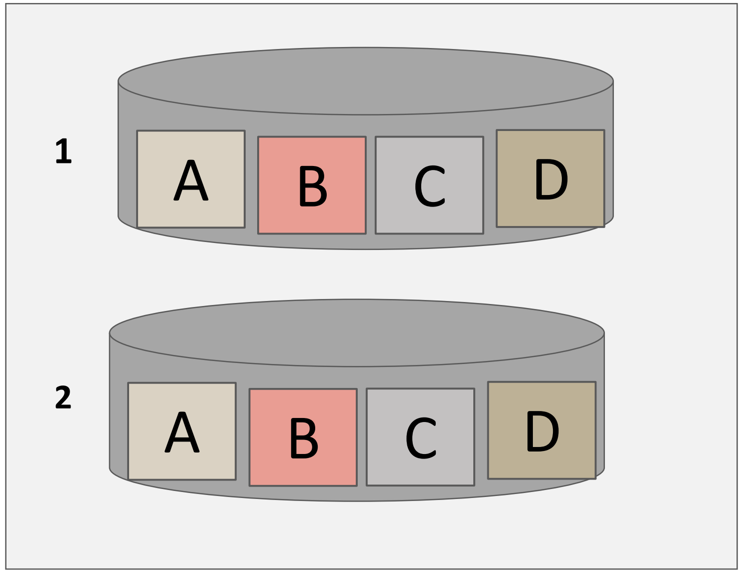
Replicated

Partitioned Replicated

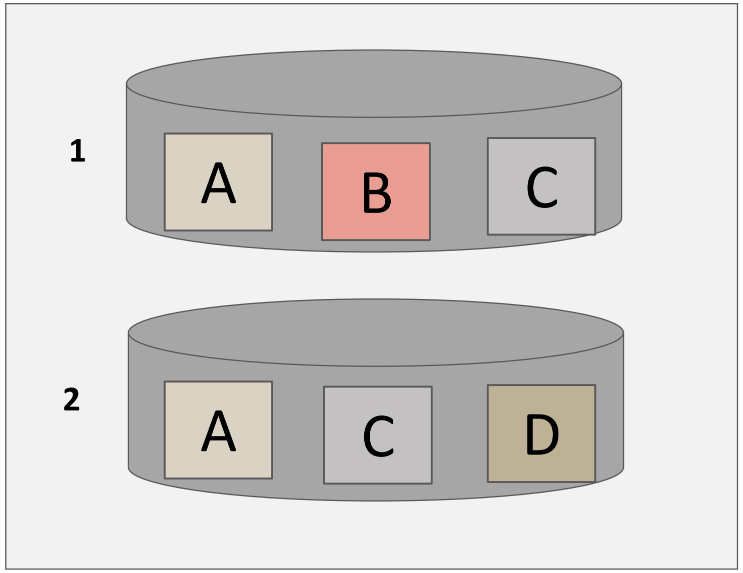
Distributed Databases



Partitioned
Non Replicated

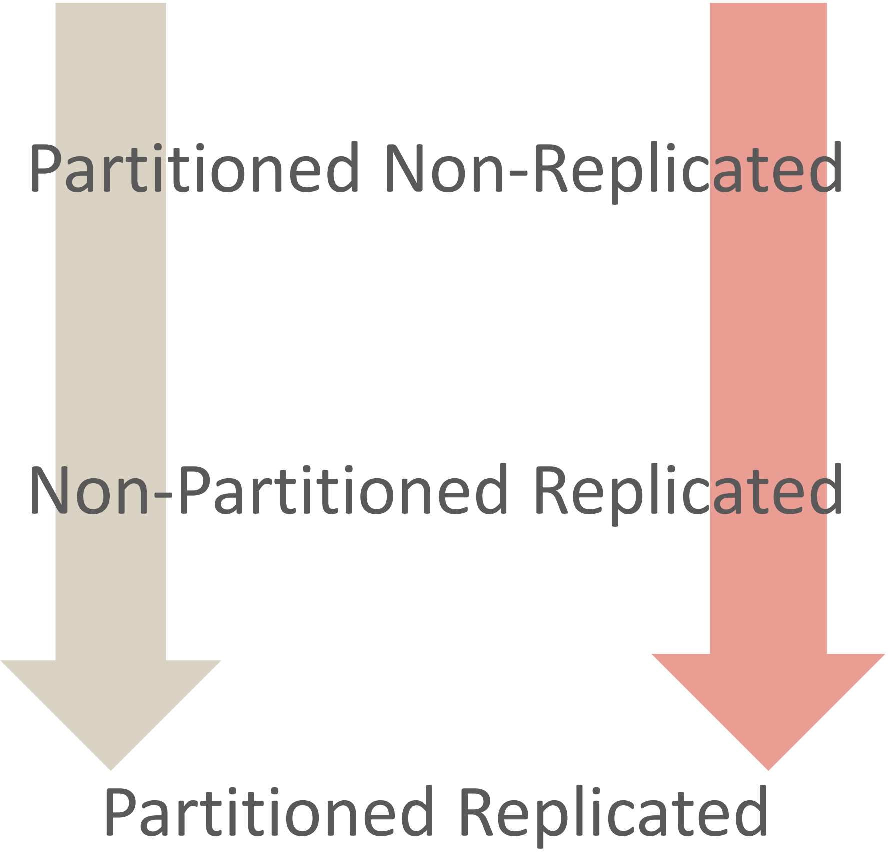


Non Partitioned
Replicated



Partitioned
Replicated

Distributed Databases



increasing flexibility, parallelism, availability

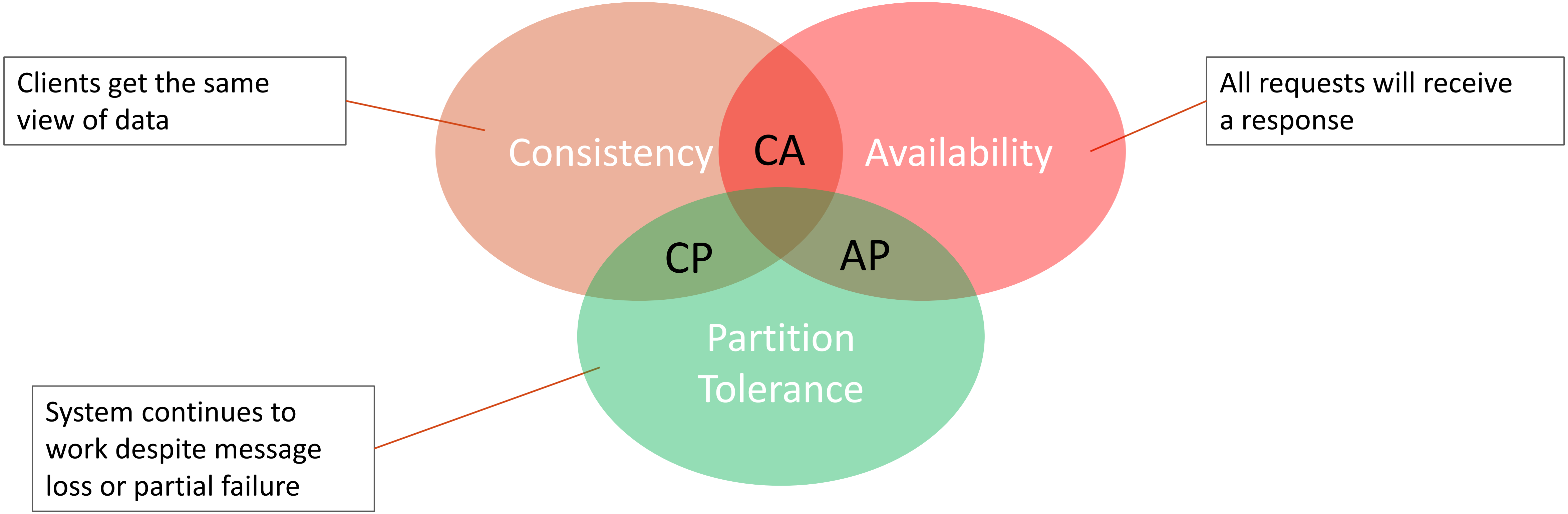
Partitioned Non-Replicated

Non-Partitioned Replicated

Partitioned Replicated

increasing cost, complexity

CAP Theorem



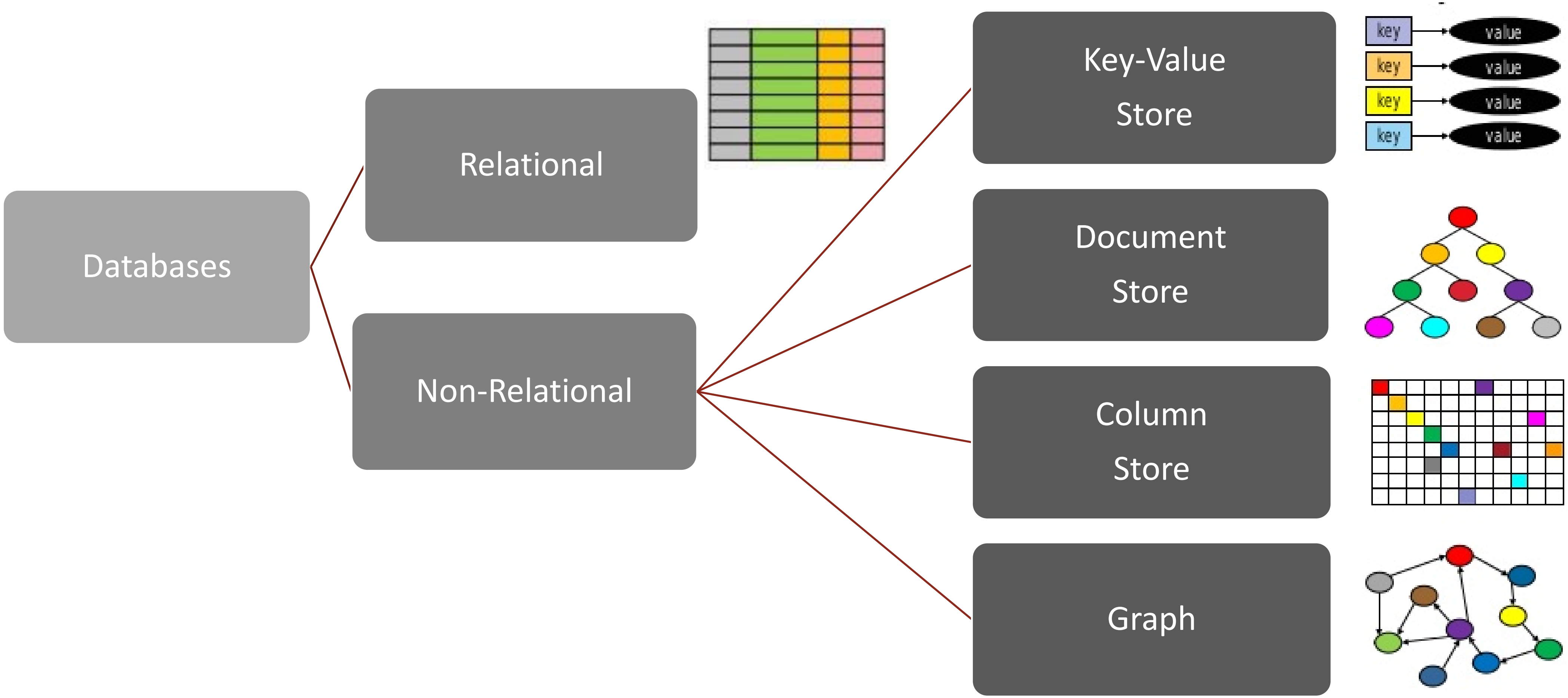
Clients get the same view of data

All requests will receive a response

System continues to work despite message loss or partial failure

It is impossible for a distributed computer system to simultaneously provide all three of the following guarantees: Consistency, Availability Partition tolerance - Eric Brewer

Database Types - Structure



Databases

Relational

Non-Relational

Key-Value Store

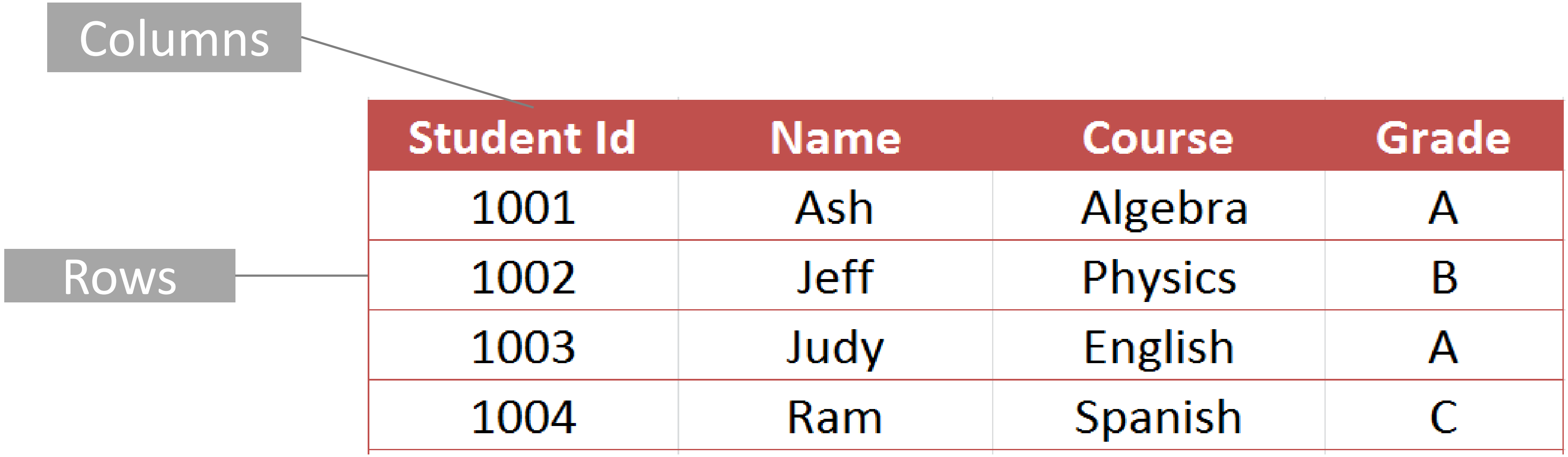
Document Store

Column Store

Graph

Relational Databases

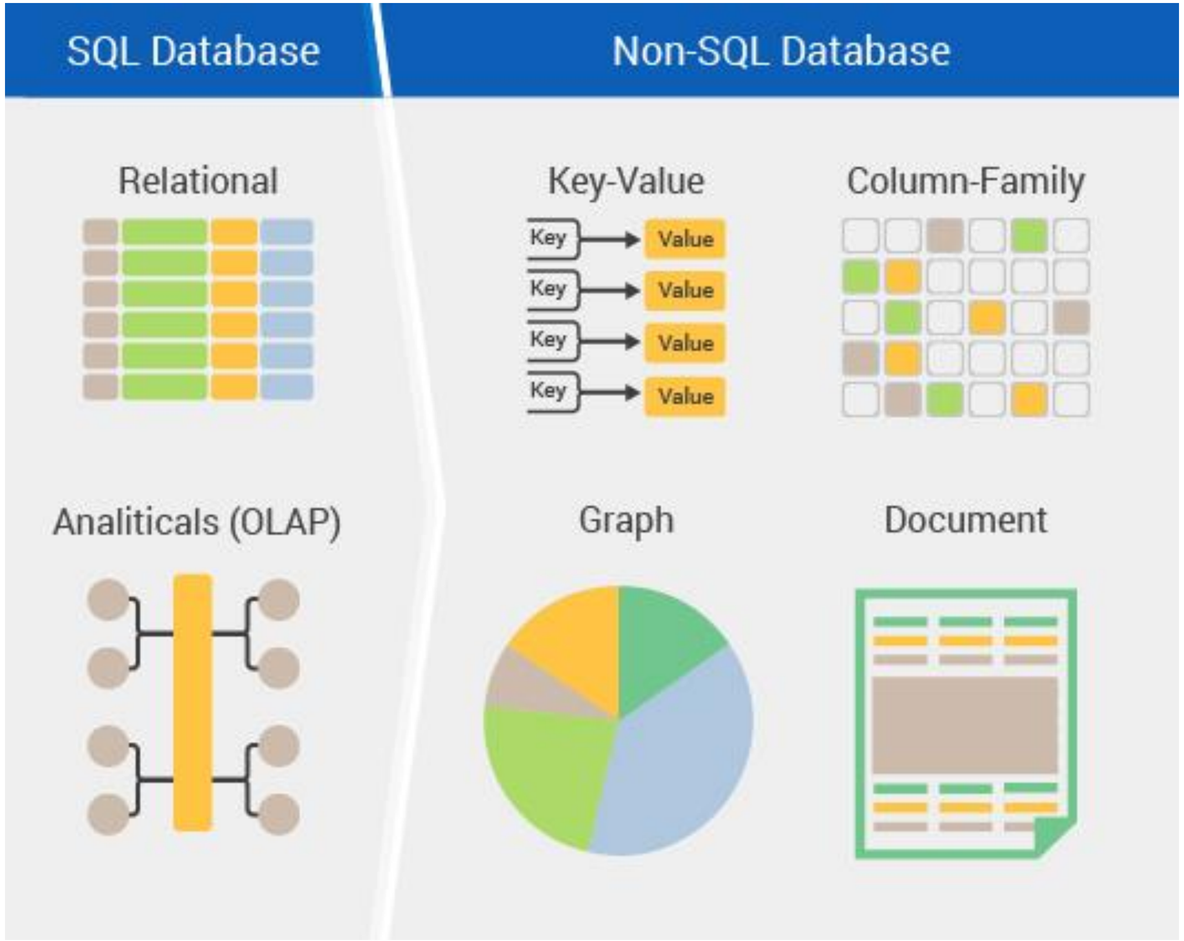
- Composed of tables with columns and rows
- Data stored across multiple related tables
- Support SQL like query languages and SQL standards



Student Id	Name	Course	Grade
1001	Ash	Algebra	A
1002	Jeff	Physics	B
1003	Judy	English	A
1004	Ram	Spanish	C

No-SQL or Non-Relational Databases

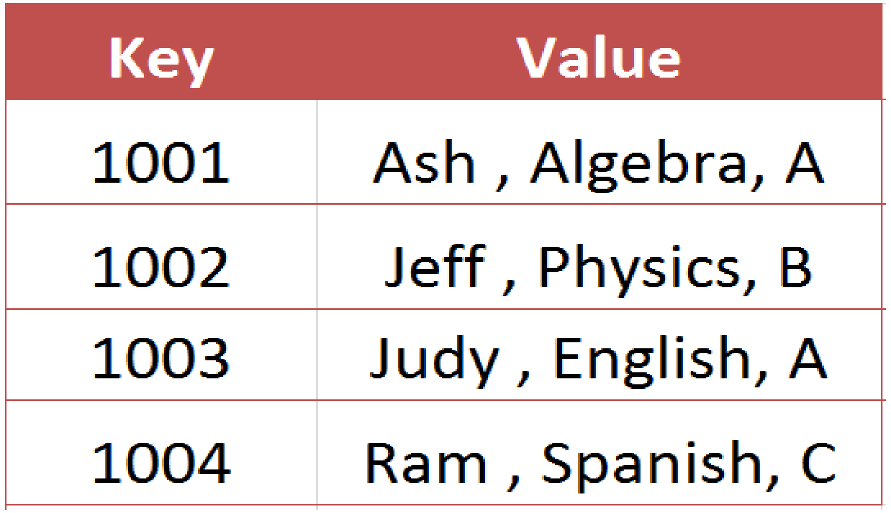
- Store semi-structured data
- Flexible schema
- Increased scalability
- Eventual consistency



https://www.netsolutions.com/

Key-Value Store

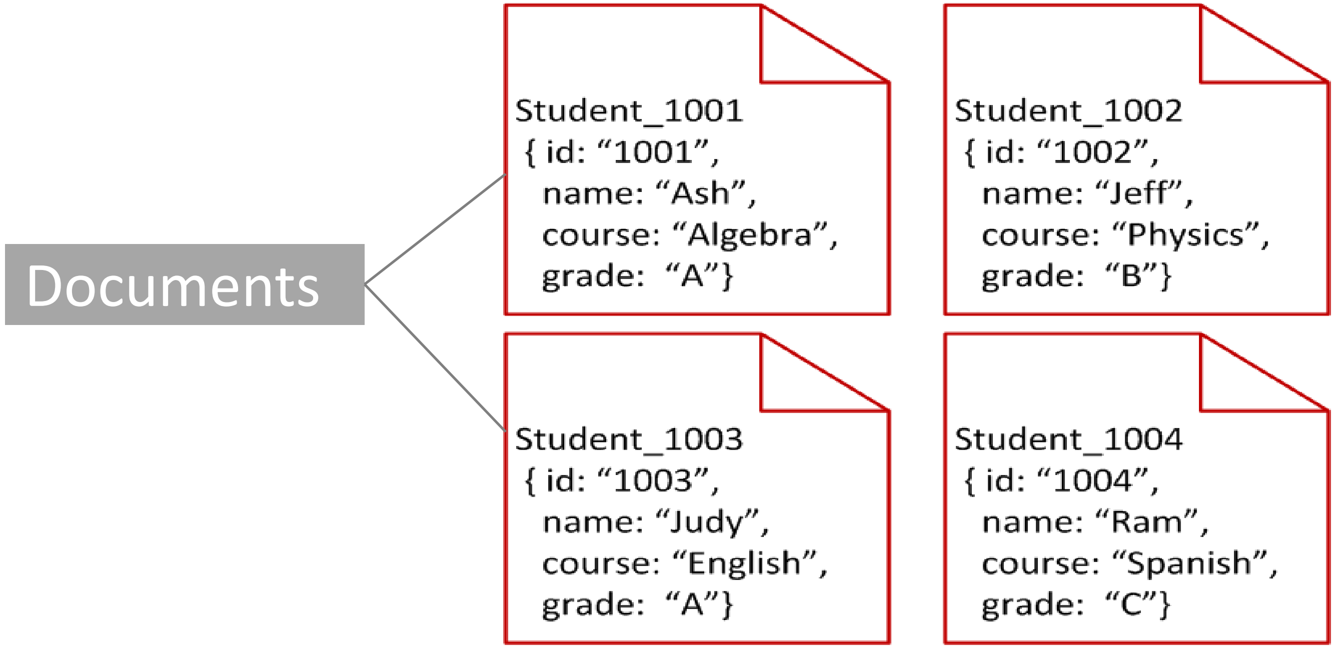
- Stores a Dictionary (or Hash) data structure
- Each value is associated with unique key



Key	Value
1001	Ash , Algebra, A
1002	Jeff , Physics, B
1003	Judy , English, A
1004	Ram , Spanish, C

Document Store

- Data stored as documents



Documents

```
Student_1001
{ id: “1001”,
  name: “Ash”,
  course: “Algebra”,
  grade:  “A”}
```

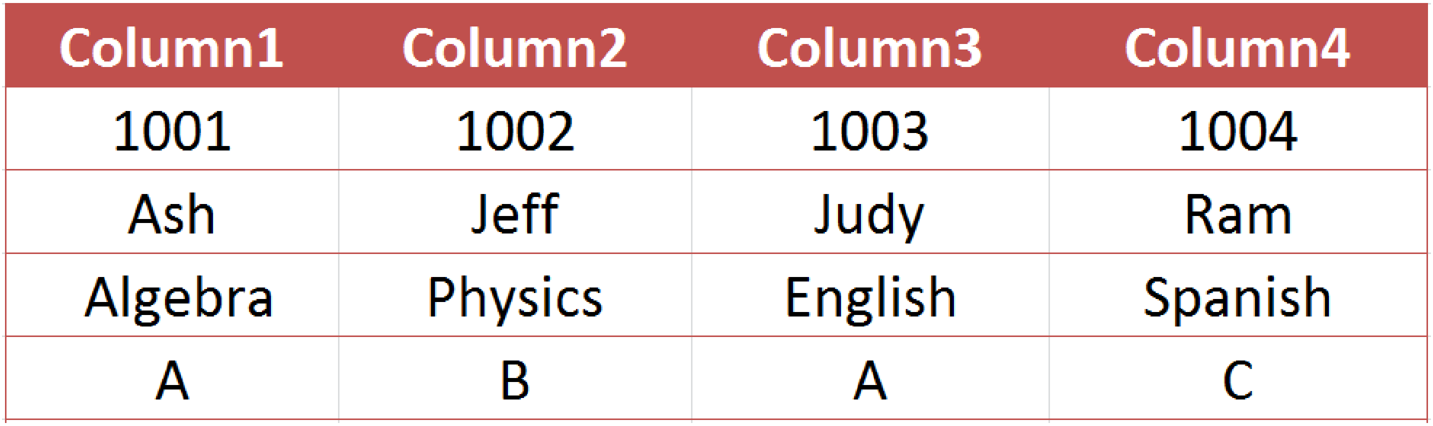
```
Student_1002
{ id: “1002”,
  name: “Jeff”,
  course: “Physics”,
  grade:  “B”}
```

```
Student_1003
{ id: “1003”,
  name: “Judy”,
  course: “English”,
  grade:  “A”}
```

```
Student_1004
{ id: “1004”,
  name: “Ram”,
  course: “Spanish”,
  grade:  “C”}
```

Column Store

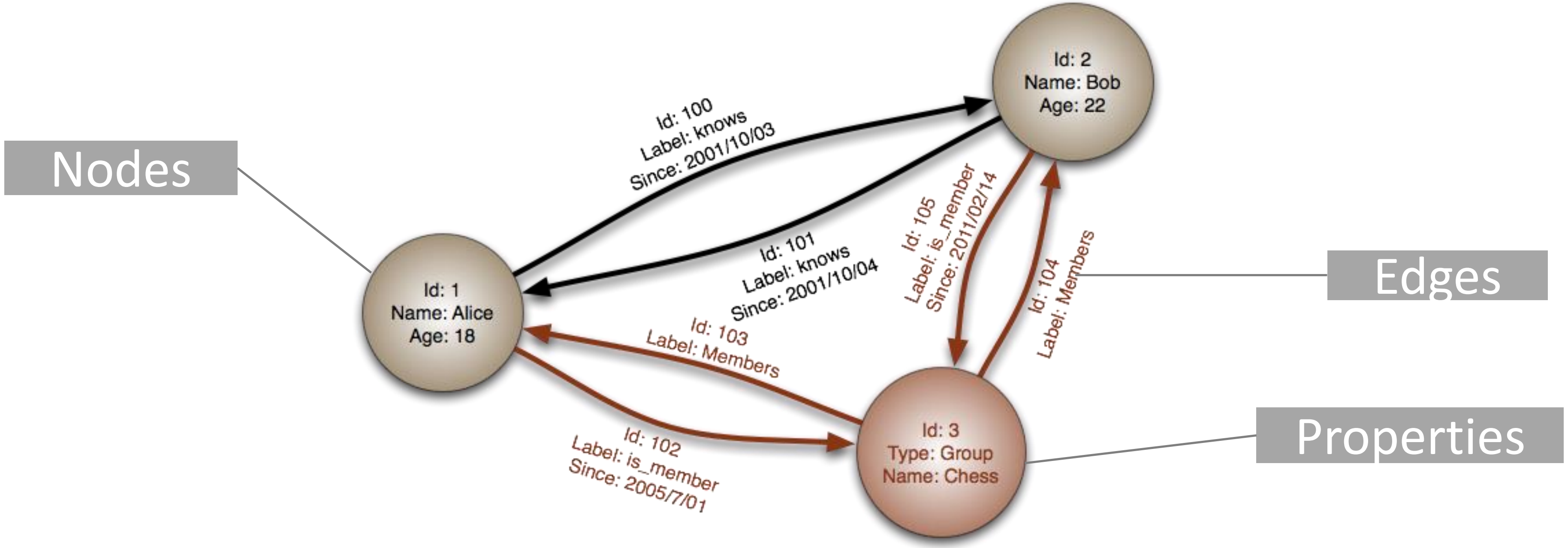
- Entities stored in columns rather than rows
- Column oriented databases



Column1	Column2	Column3	Column4
1001	1002	1003	1004
Ash	Jeff	Judy	Ram
Algebra	Physics	English	Spanish
A	B	A	C

Graph Databases

- Entities stored as graphs - Nodes, Edges, Properties



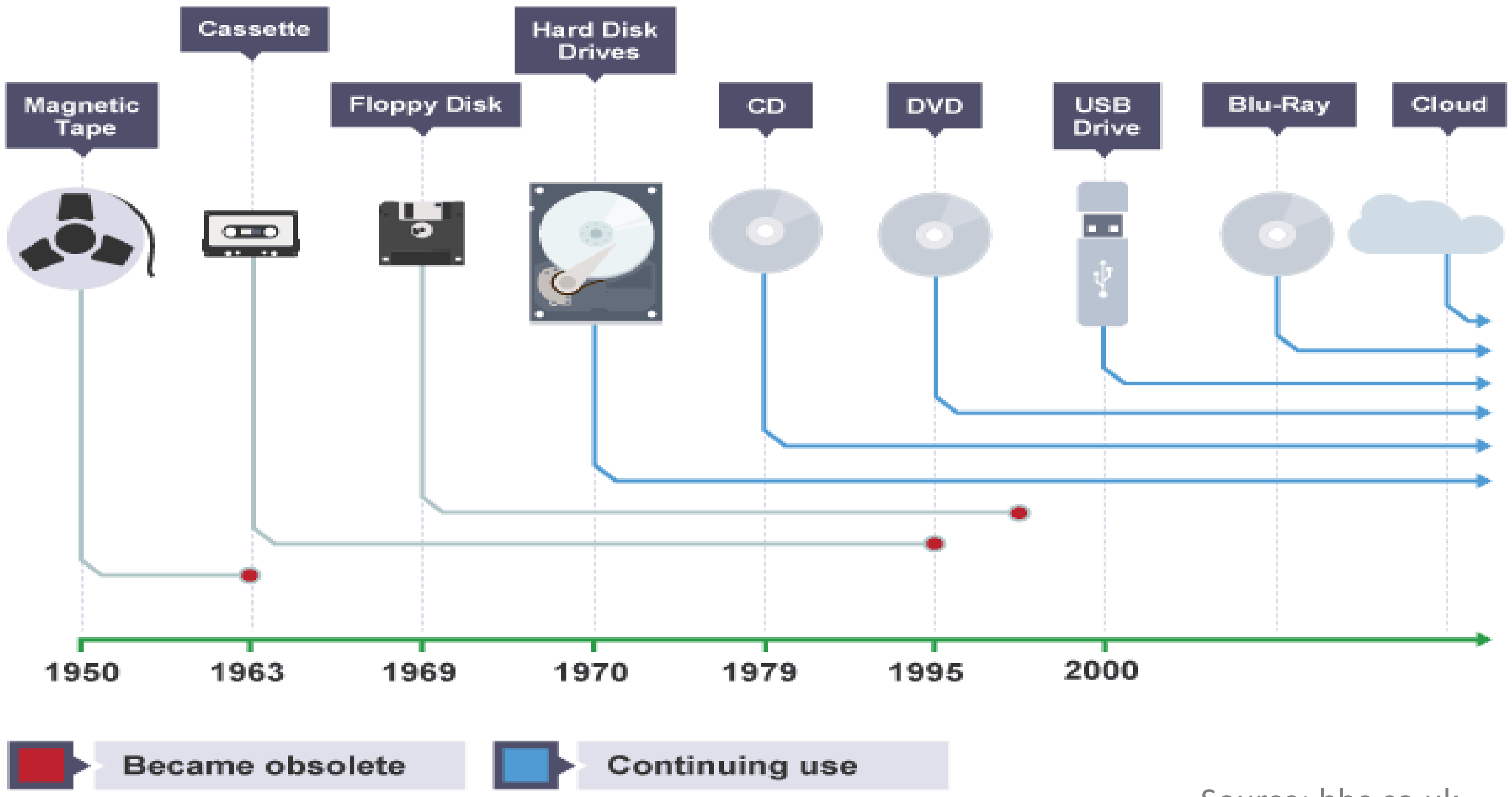
Source: Wikipedia

Popular Databases

Database Type	Database Names
Relational	MySQL ORACLE PostgreSQL IBM DB2
Key-Value	redis riak MEMCACHED
Column	APACHE HBASE Cassandra Google BigQuery
Document	mongoDB CouchDB relax amazon DynamoDB
Graph	neo4j TITAN OrientDB INFINITEGRAPH

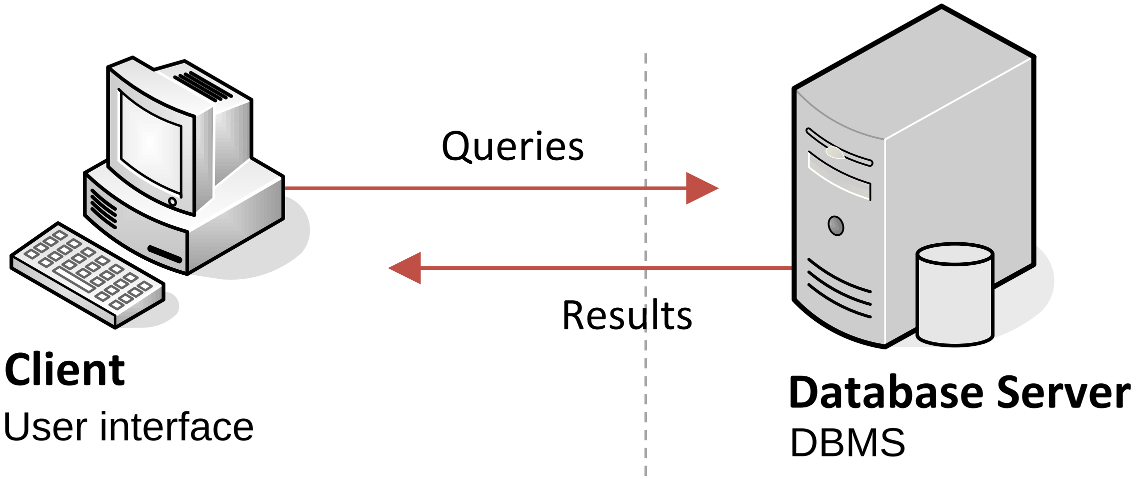
Appendix

Evolution of storage



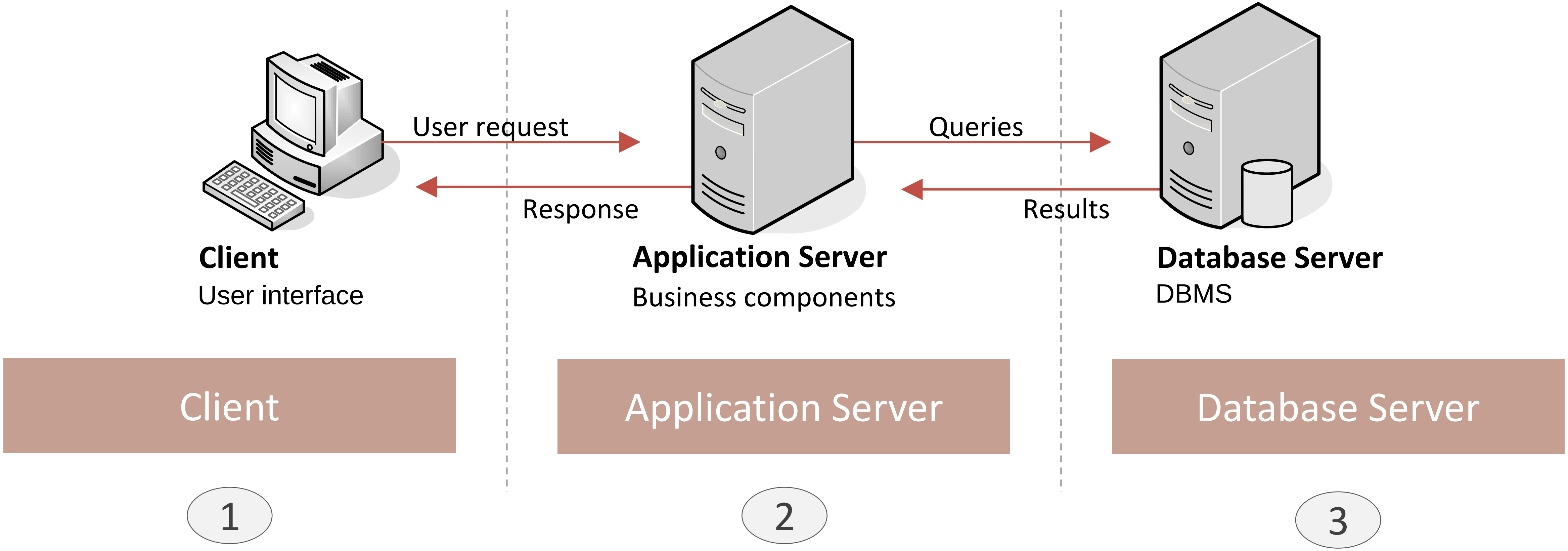
Source: bbc.co.uk

2-tier Architecture (Physical)



Client	Database Server
1	2

3-tier Architecture (Physical)


User request
Queries
Response
Results
Client
User interface
Application Server
Business components
Database Server
DBMS
Client
Application Server
Database Server
1
2
3