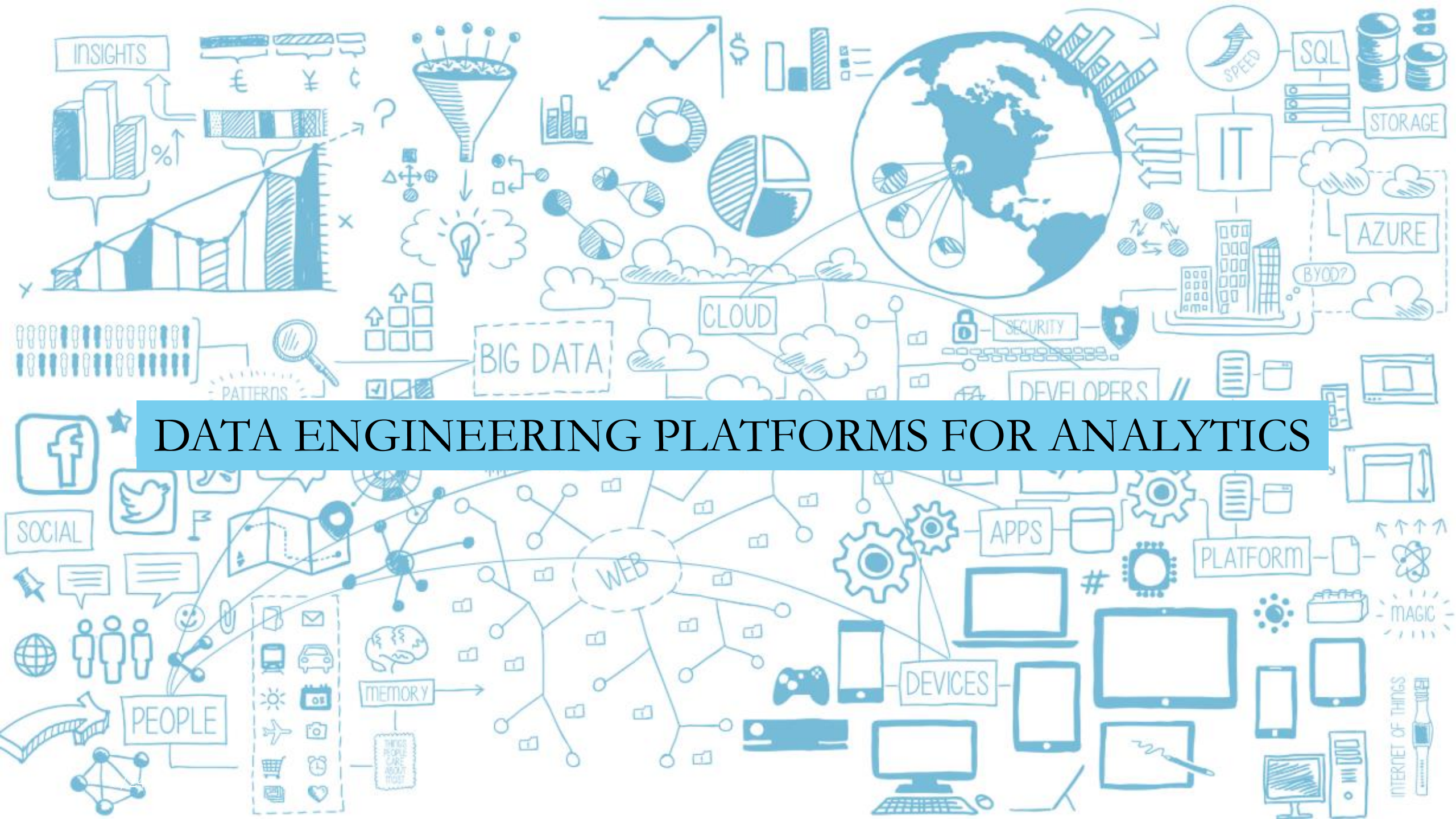
DATA ENGINEERING PLATFORMS FOR ANALYTICS



Welcome

Introductions, Syllabus Review

Introductions

The Team

- Instructor
  - Shreenidhi Bharadwaj (sbharadwaj@uchicago.edu)
  - https://www.linkedin.com/in/ssbharadwaj/

- TA & Grader
  - Jon Chan (jchan530@uchicago.edu)
  - https://www.linkedin.com/in/jchan530/

Syllabus

- Session 1 – Foundations of data systems
- Session 2 – Relational Databases
- Session 3 – Structured Query Language(SQL)
- Session 4 – Advanced SQL
- Session 5 – Analytical data platforms

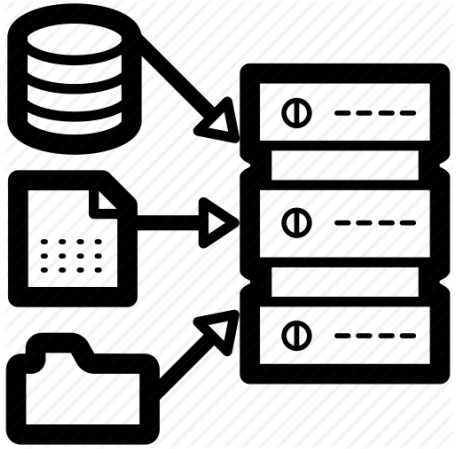
Syllabus

- Session 6 – Business Intelligence
- Session 7 – Data Pipelines (GCP)
- Session 8 – Document & Graph databases
- Session 9 – Columnar, Key Value DB & Blockchain
- Session 10 – Final Exam & Project Presentations

Assignments, Final Project & Final Exam

- 4 Assignments (Individual)
  - 100 points each
  - Bi-Weekly submission (Saturdays 11:59 PM CST)
- Final Project (Max 4 students)
  - Business Use Case - 10%
  - Data Ingestion, Analysis & Preparation - 25%
  - Data Modelling & End to End Platform Design – 25 %
  - Tools / Data Store Evaluation / Data Transformation – 20%
  - Design of Dashboards/reports for insights - 20%
- Final Exam : 100 Points
  - 50 Multiple Choice Questions, 10 SQL & 4 MongoDB questions

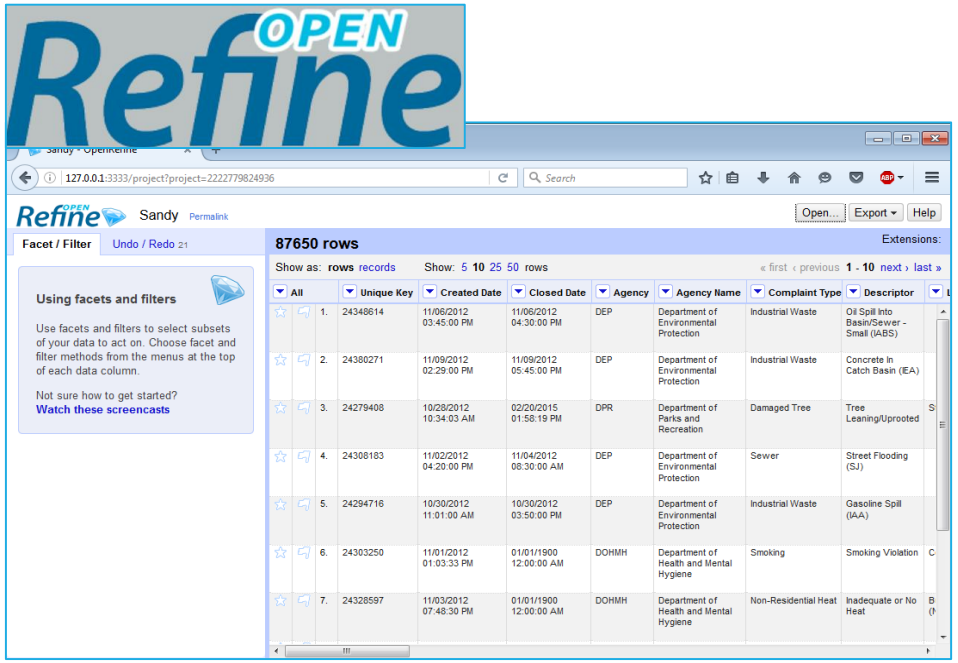
Primary Datasets



- Classic Models (In-Class exercises)
  - Sessions 2, 3, 4
- Sakila (Assignments)
  - Assignements 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/

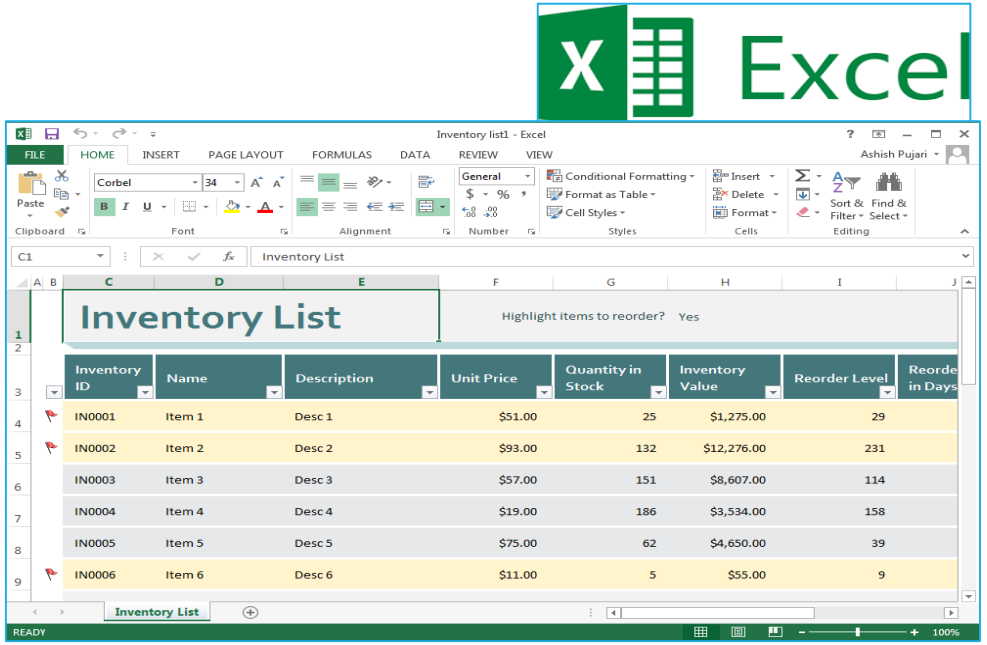
SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

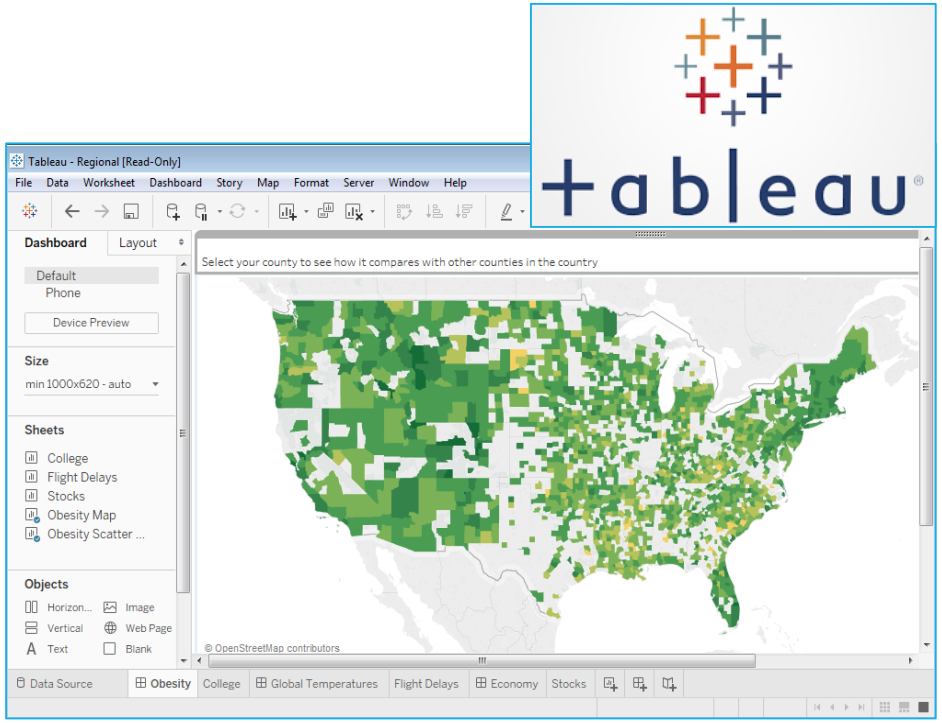
Tools (Software)



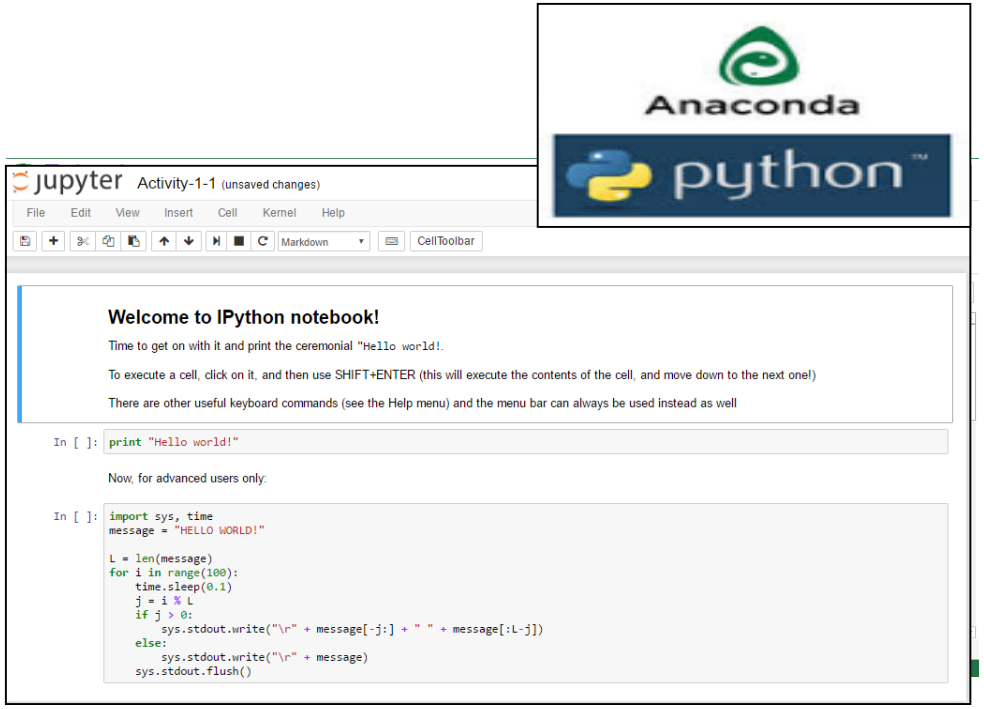
SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

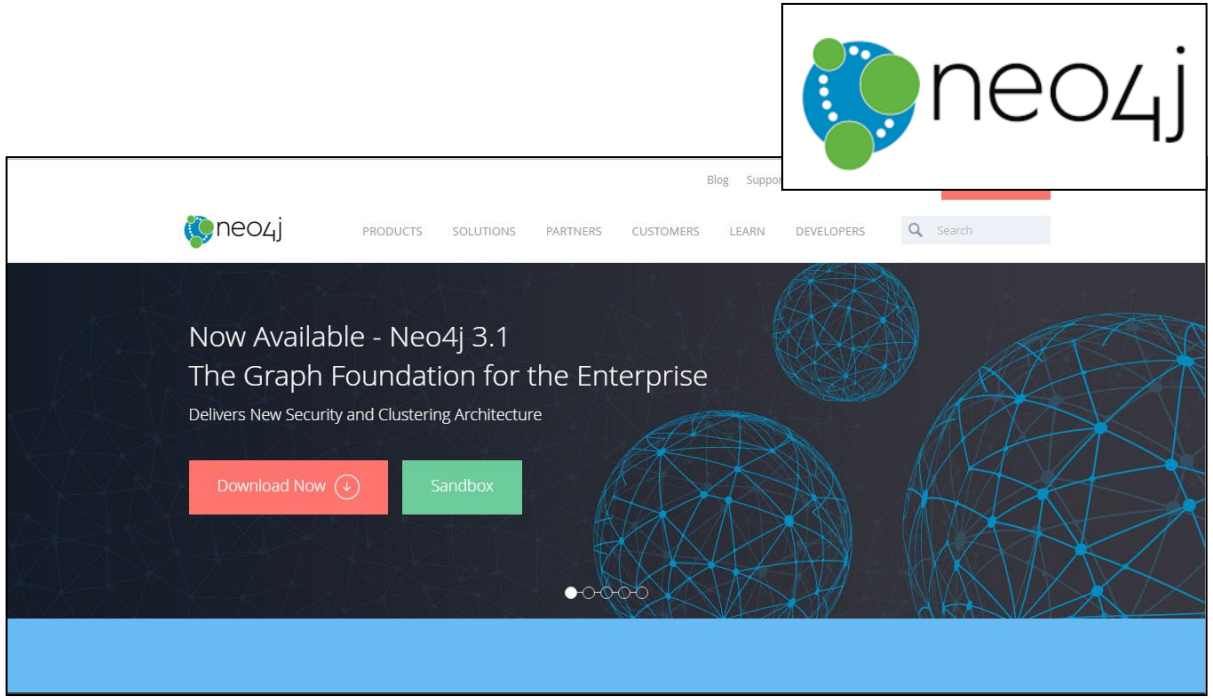
Tools (Software) - Continued…



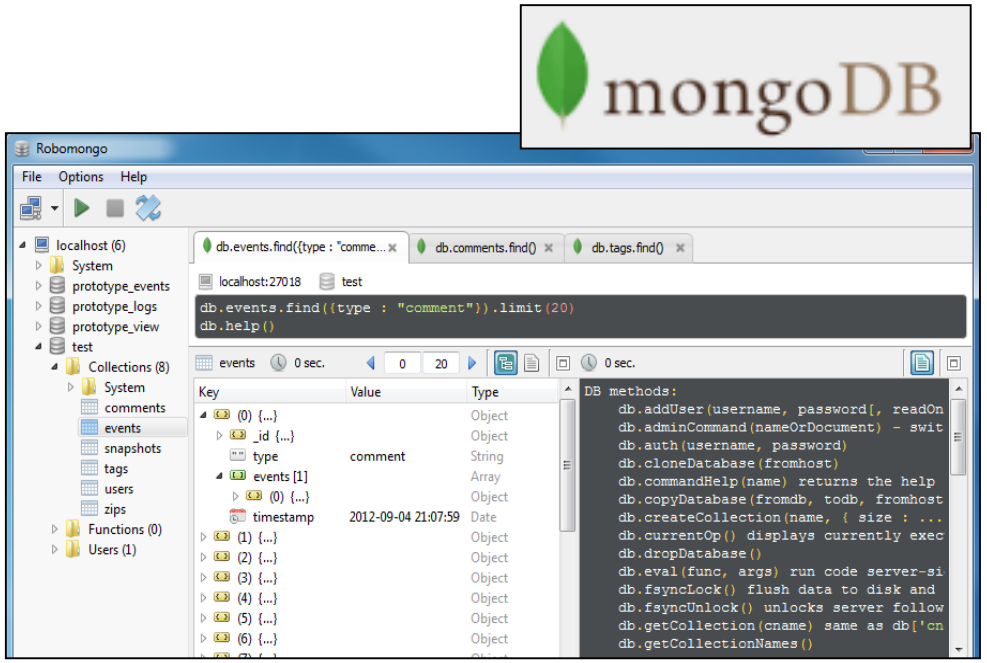


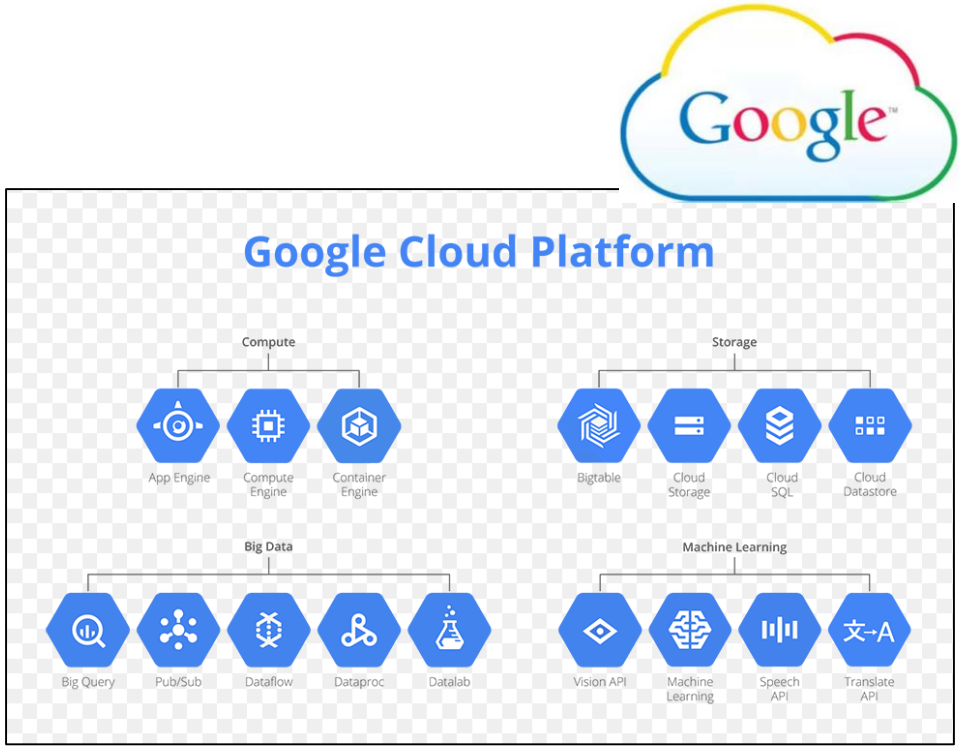
Tools (Software) - Continued…





Tools (Software) - Continued…



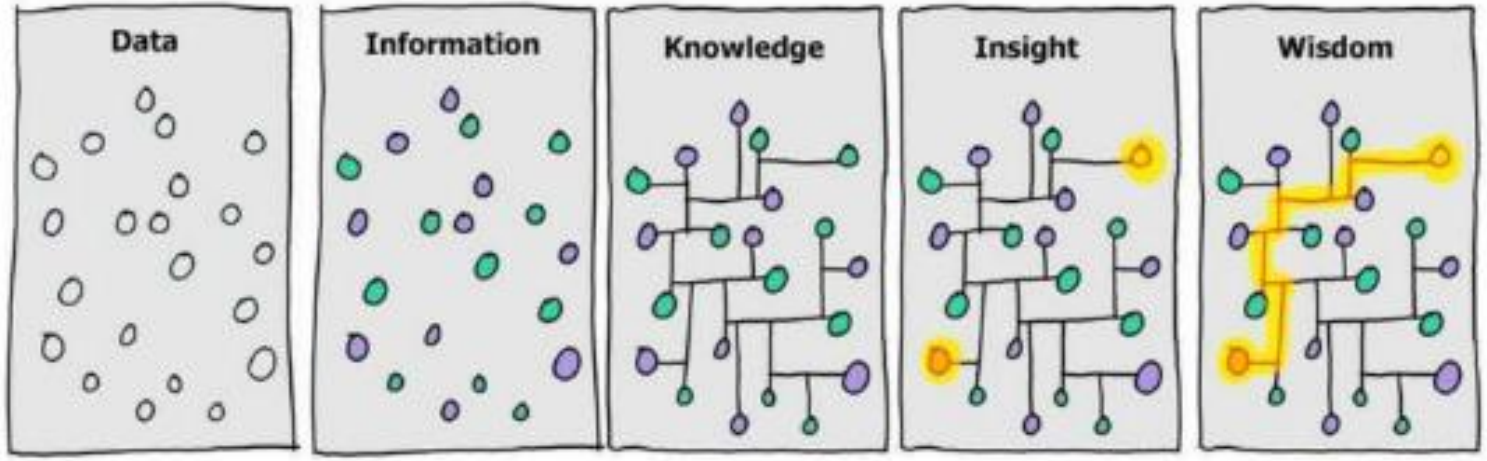


SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Data Strategy

Data To Wisdom Via Information, Knowledge & Insight.

- **'Data'** is represented by a series of random dots that could mean something – or nothing.
- **'Information'**, is where meaning or relationship is applied to the raw data.
- **'Knowledge'** is gained when we are able to draw connections between different pieces of information
- **'Insight'** is the ability to synthesize knowledge in order obtain a deep understanding of a problem.
- **'Wisdom'** is the ability to use insight to facilitate informed decision making.



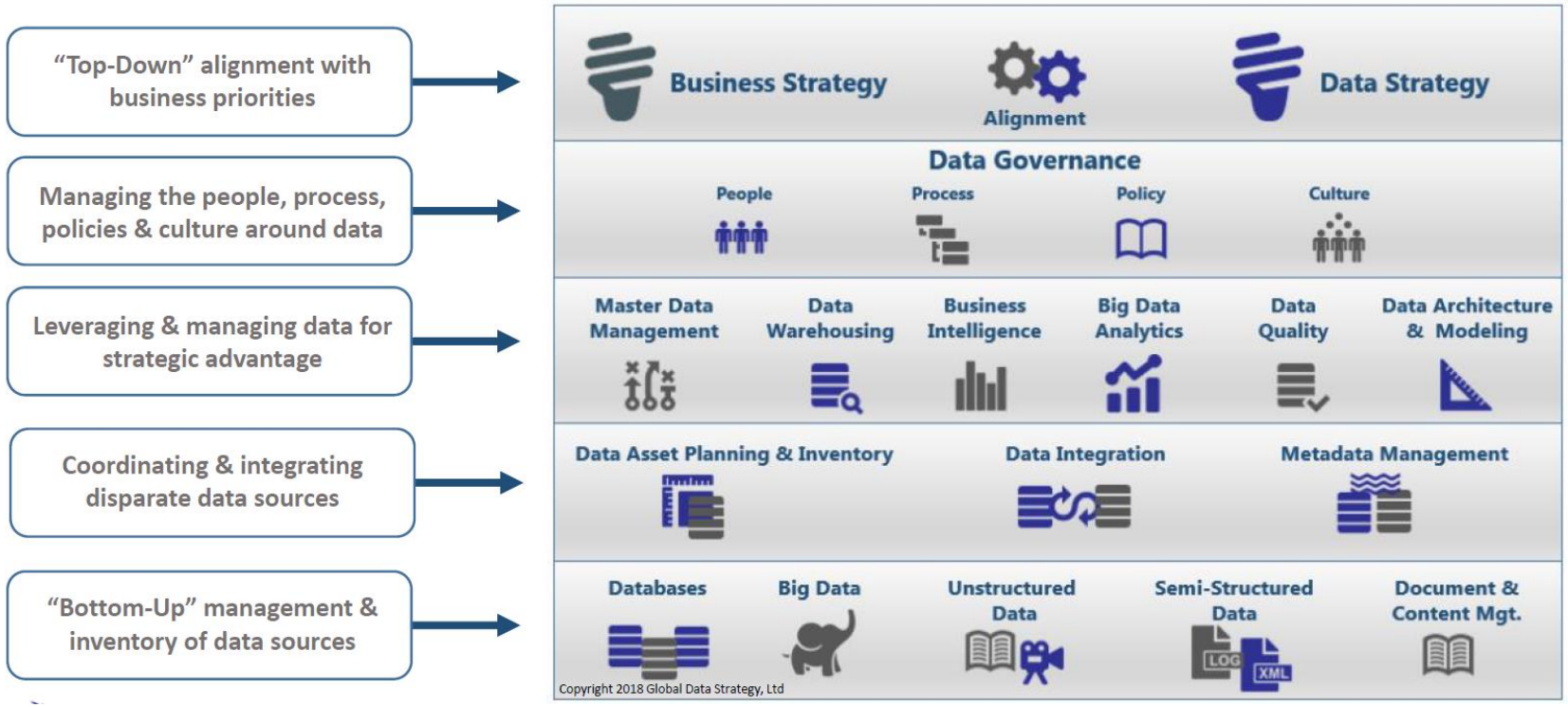
https://www.theifactory.com/news/gaining-wisdom-from-data/

What is data strategy ?

- A roadmap to enable data driven decision-making and build applications that helps an enterprise achieve its strategic imperatives.
- A Successful Data Strategy links Business Goals with Technology Solutions

Aligning Business Strategy and Data Strategy

- A Successful Data Strategy links Business Goals with Technology Solutions



Bridge the gap

Analysts need:
Collaboration

- A way to share their work
- A way to ensure their work is based on trusted data
- A means to showcase their contribution of models and analyses
- A way for other analysts and teams to contribute back to their creations

IT needs:
Centralization

- A solution that empowers all users to easily work with data
- Visibility and controls for governance and compliance
- Reduced burden - domain experts manage domain data
- Insight into which data is most valuable so they can focus their efforts

The organization needs:
Socialized data access

- Awareness of what data and assets are being created and by whom
- All users working and collaborating in the same place
- Agility and governance in one solution
- Analysis is conducted based on trusted and certified information
- Consistency in models across groups, teams, enterprise

https://globaldatastrategy.com/

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

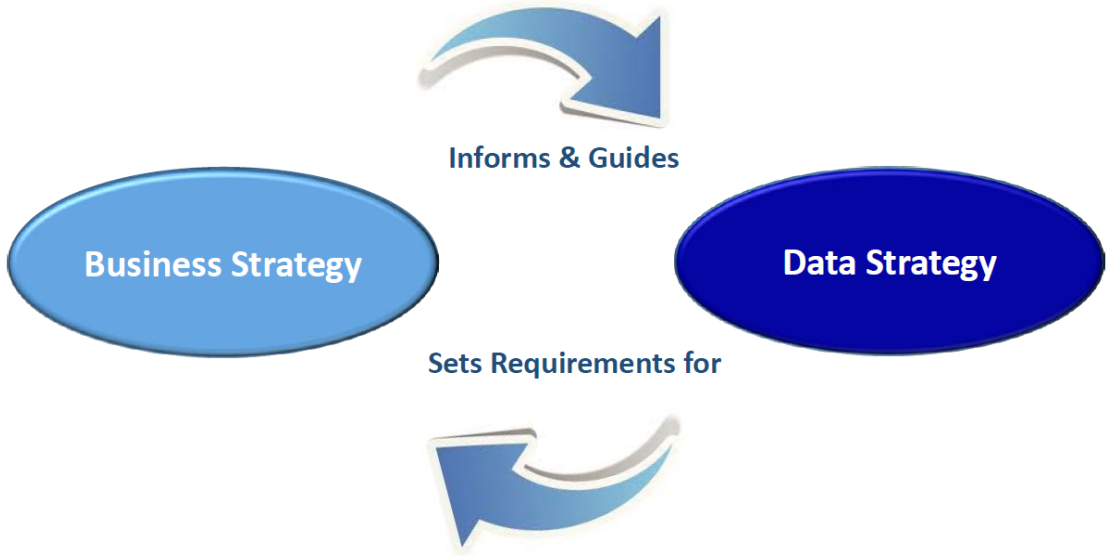
Data culture matters

- Decision making
  - The fundamental objective in collecting, analyzing and deploying data is to make better decisions.
- C-suite imperatives, and the board
  - There must be an ongoing, informed conversation with top decision makers and those who lead data initiatives throughout the organization
- Democratization of data
  - Data has to flow across the organization seamlessly
- Risk Management
  - Companies must manage risk around regulations and compliance

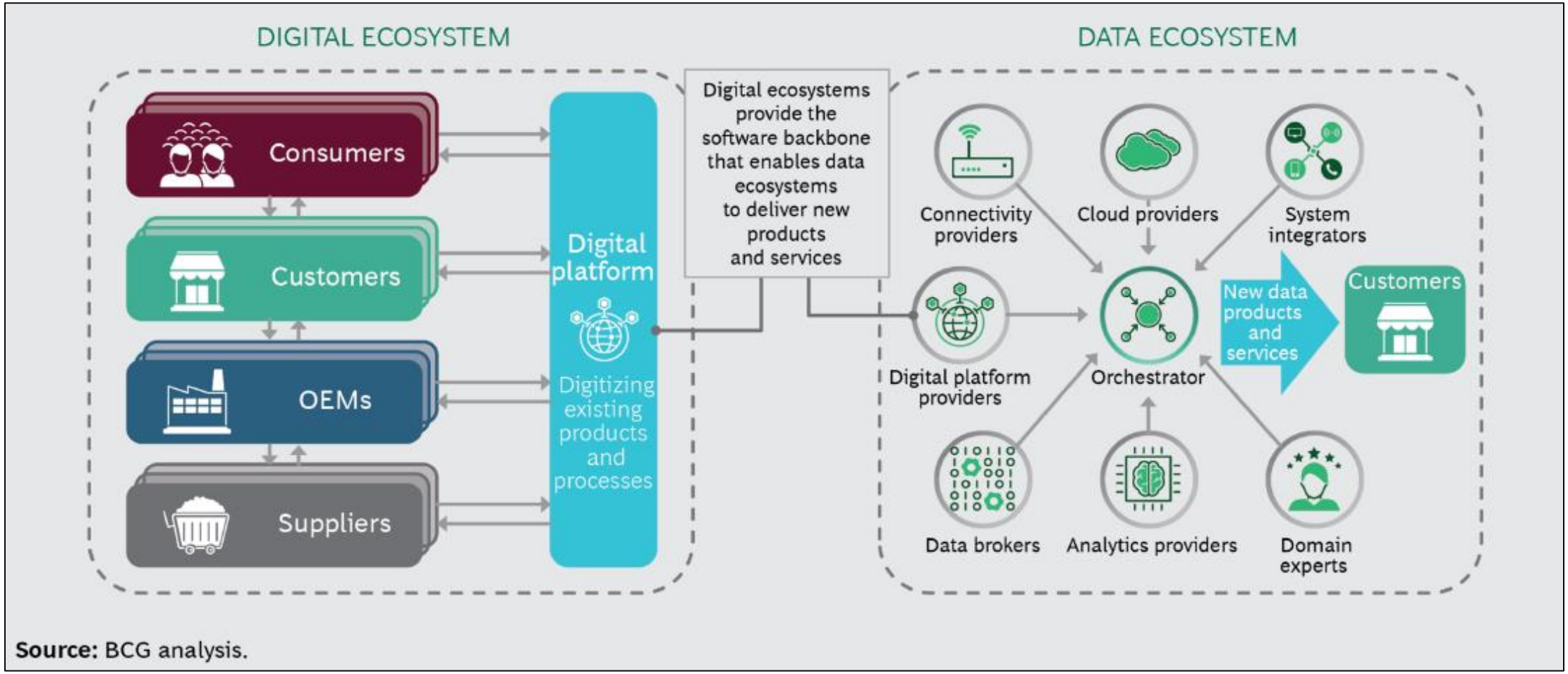
https://www.mckinsey.com/

Data Strategy focus areas

- Data Ingestion Strategy
- Data Storage Strategy
- Data Delivery Strategy
- Data Analytics Strategy
- Data Governance Strategy

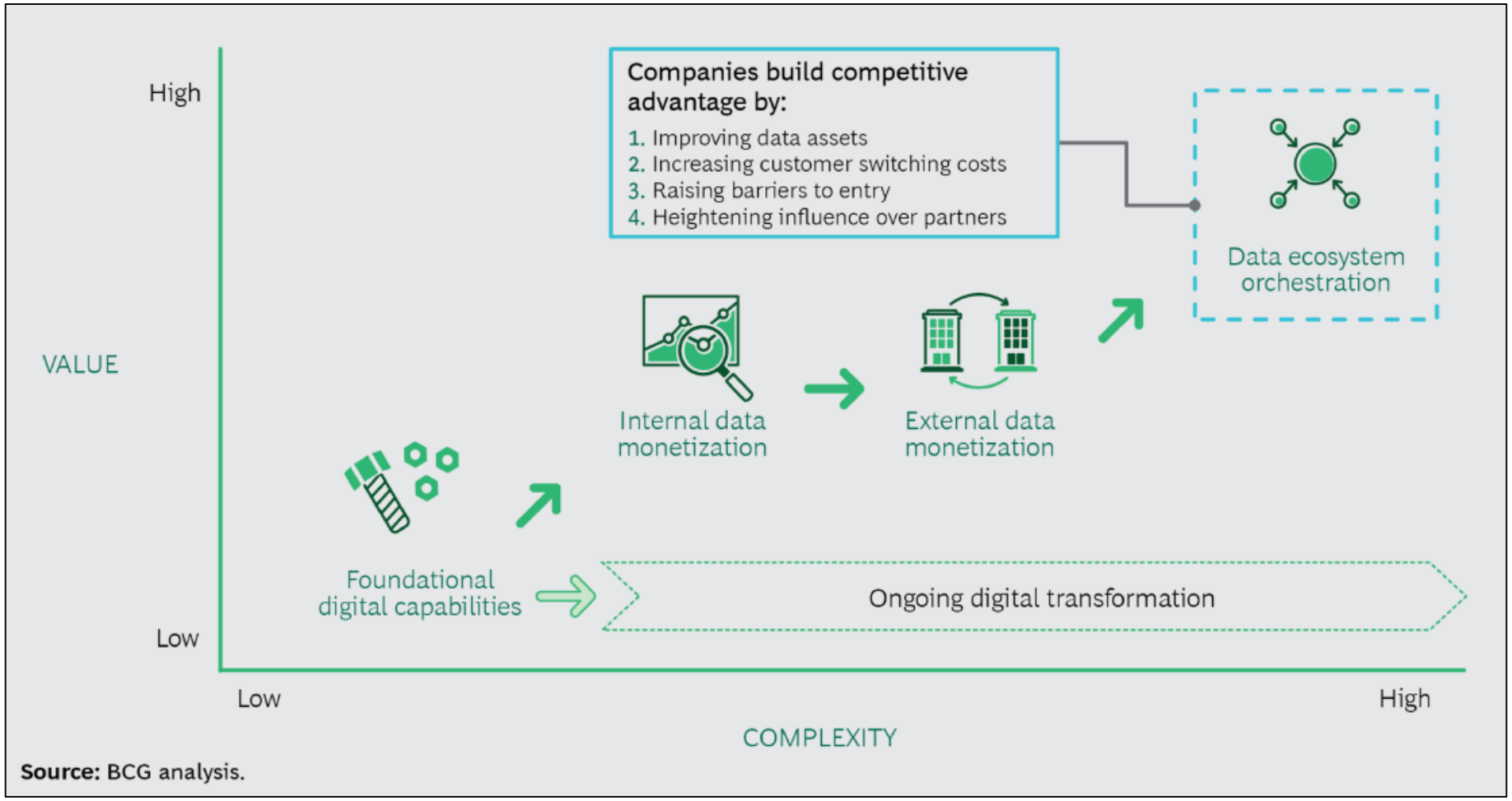


Data and digital ecosystems

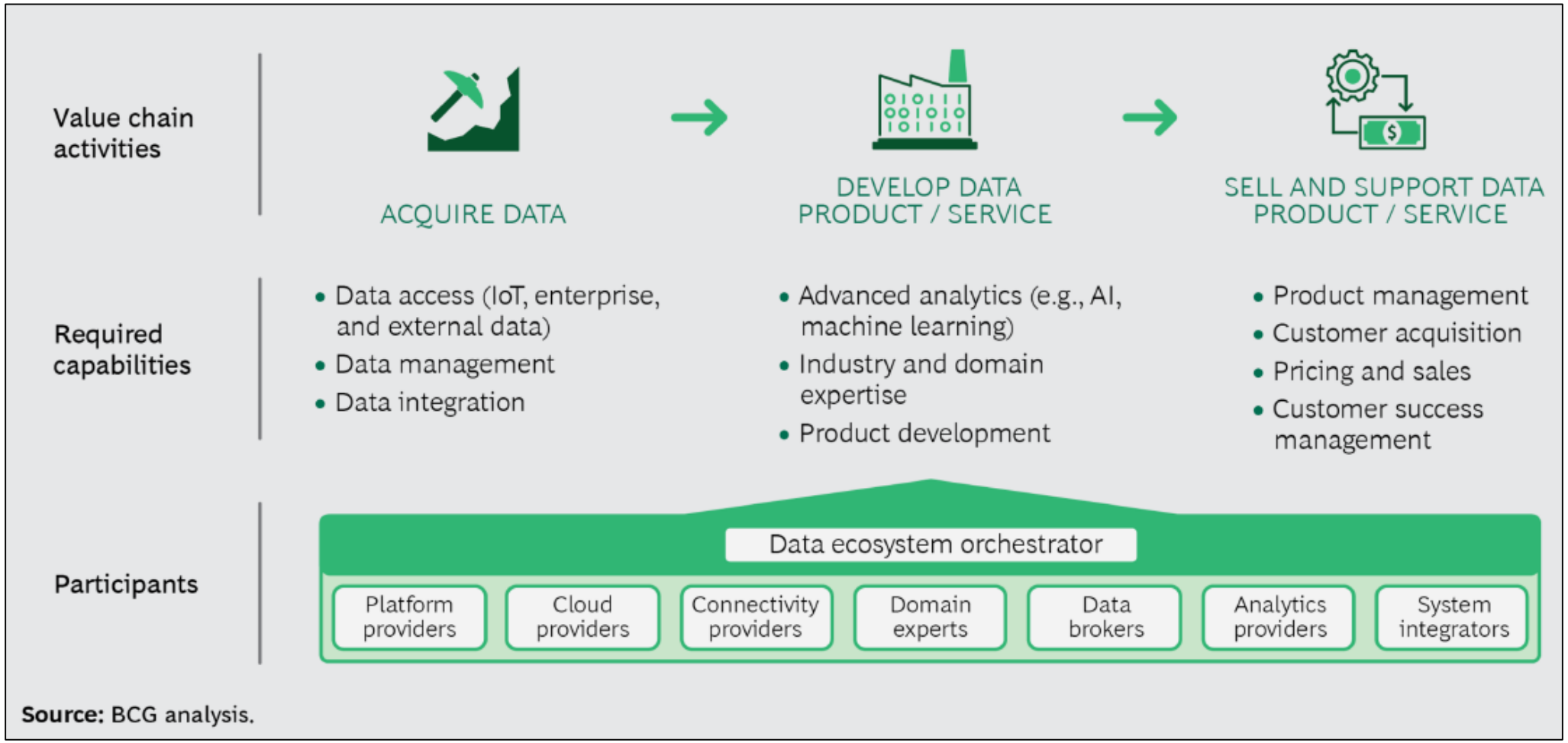


Source: BCG analysis.

Data monetization journey

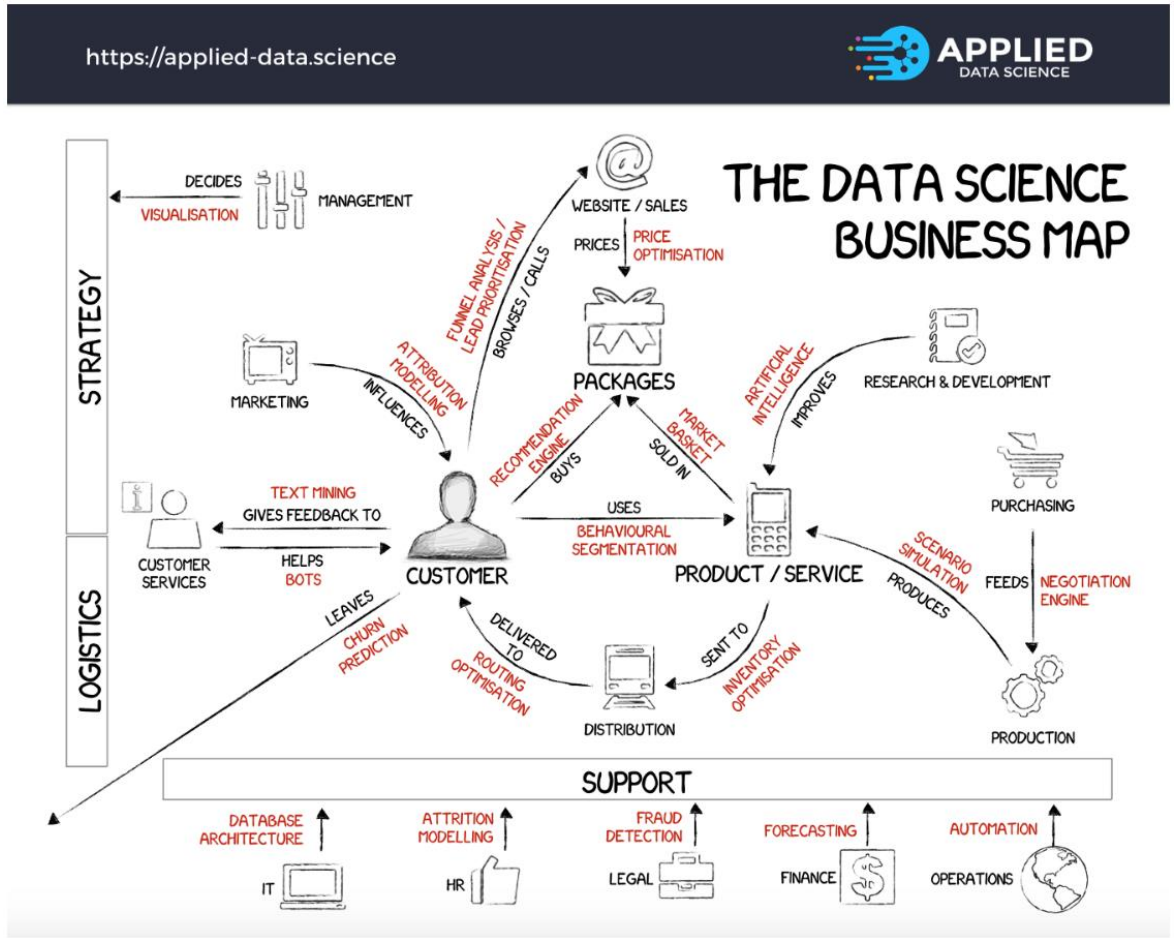


Data value chain



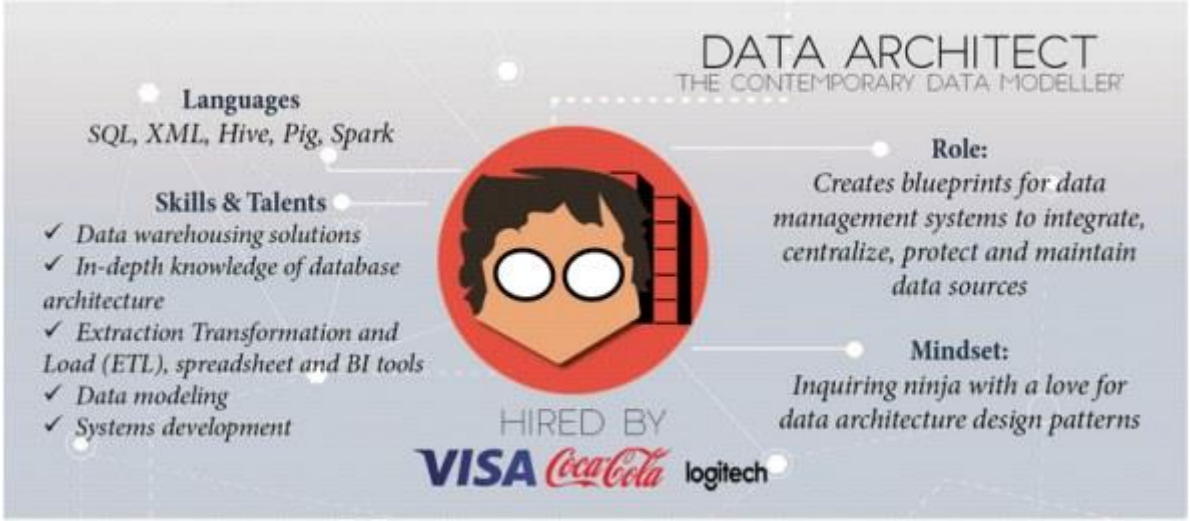
The Data Science Business Map

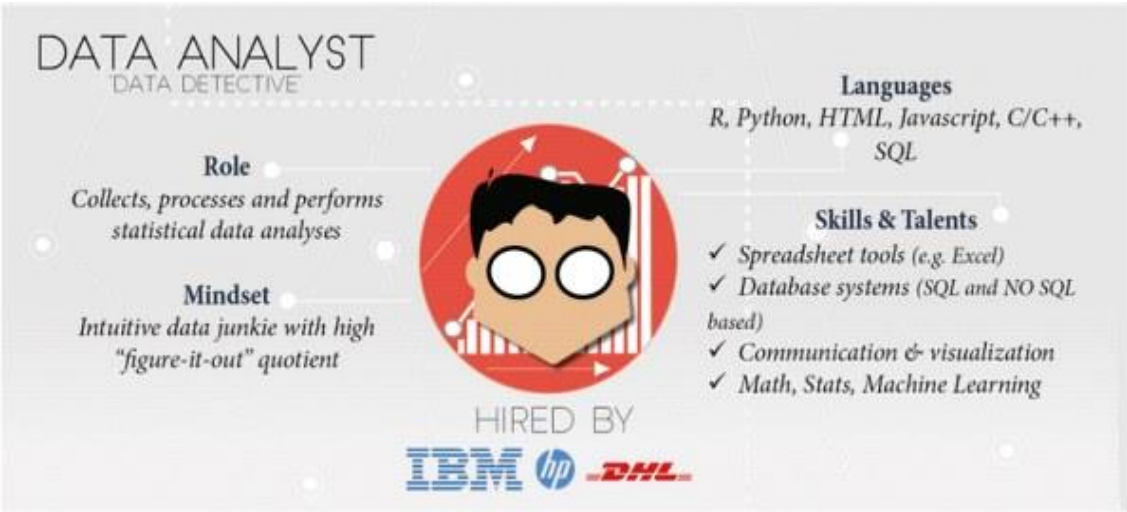
- 3 key elements
  - The customer
  - The product / service
  - The package that links the customer to the product / service



https://medium.com/applied-data-science/every-arrow-on-this-diagram-is-a-data-science-project-775339e51828/

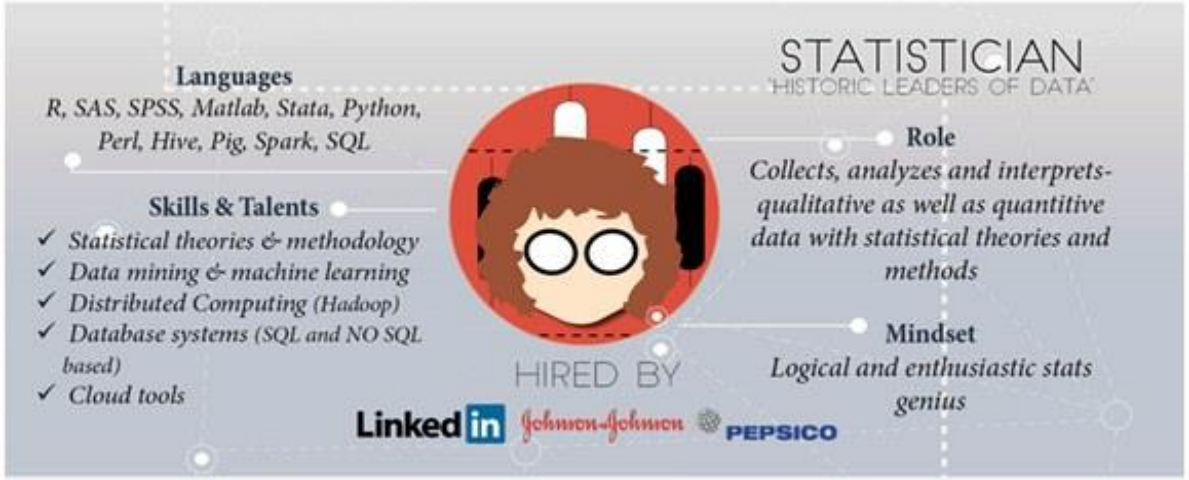
Roles

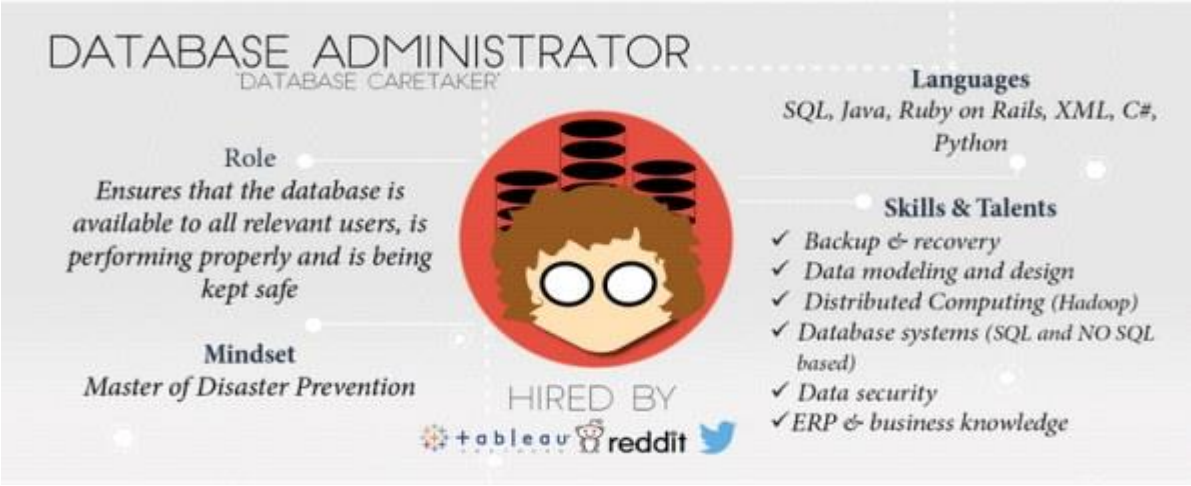


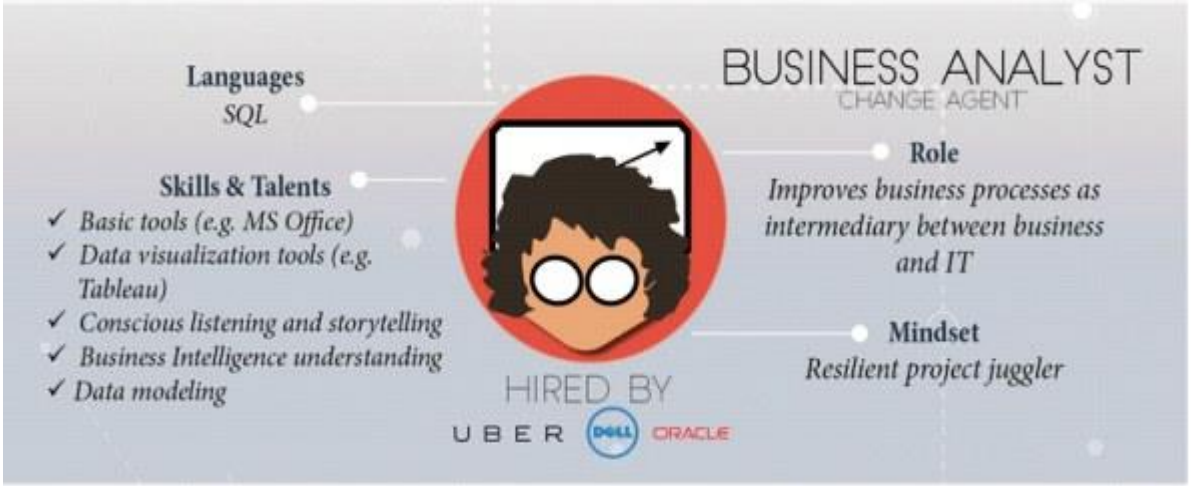


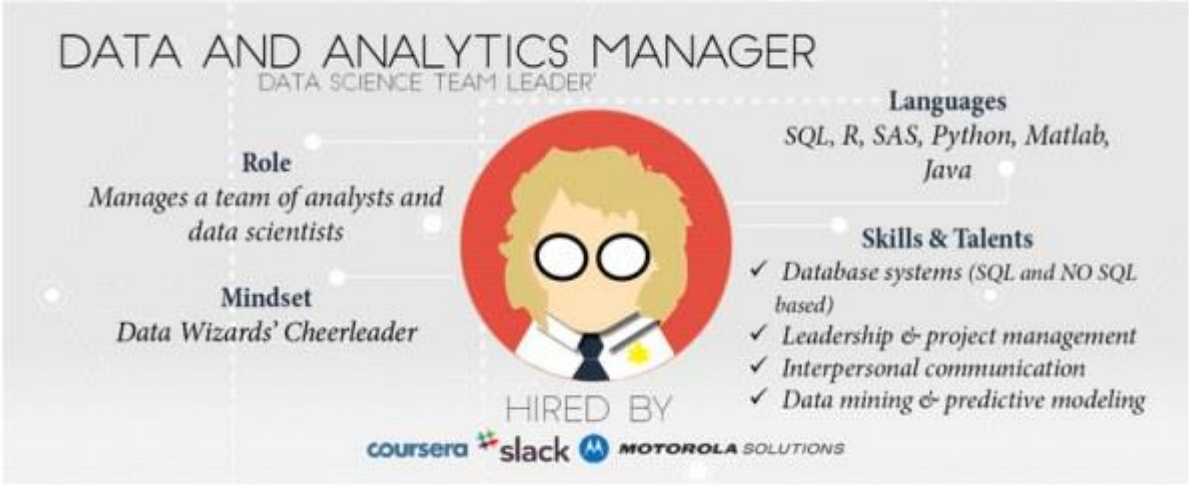
https://www.kdnuggets.com/

Roles









https://www.kdnuggets.com/

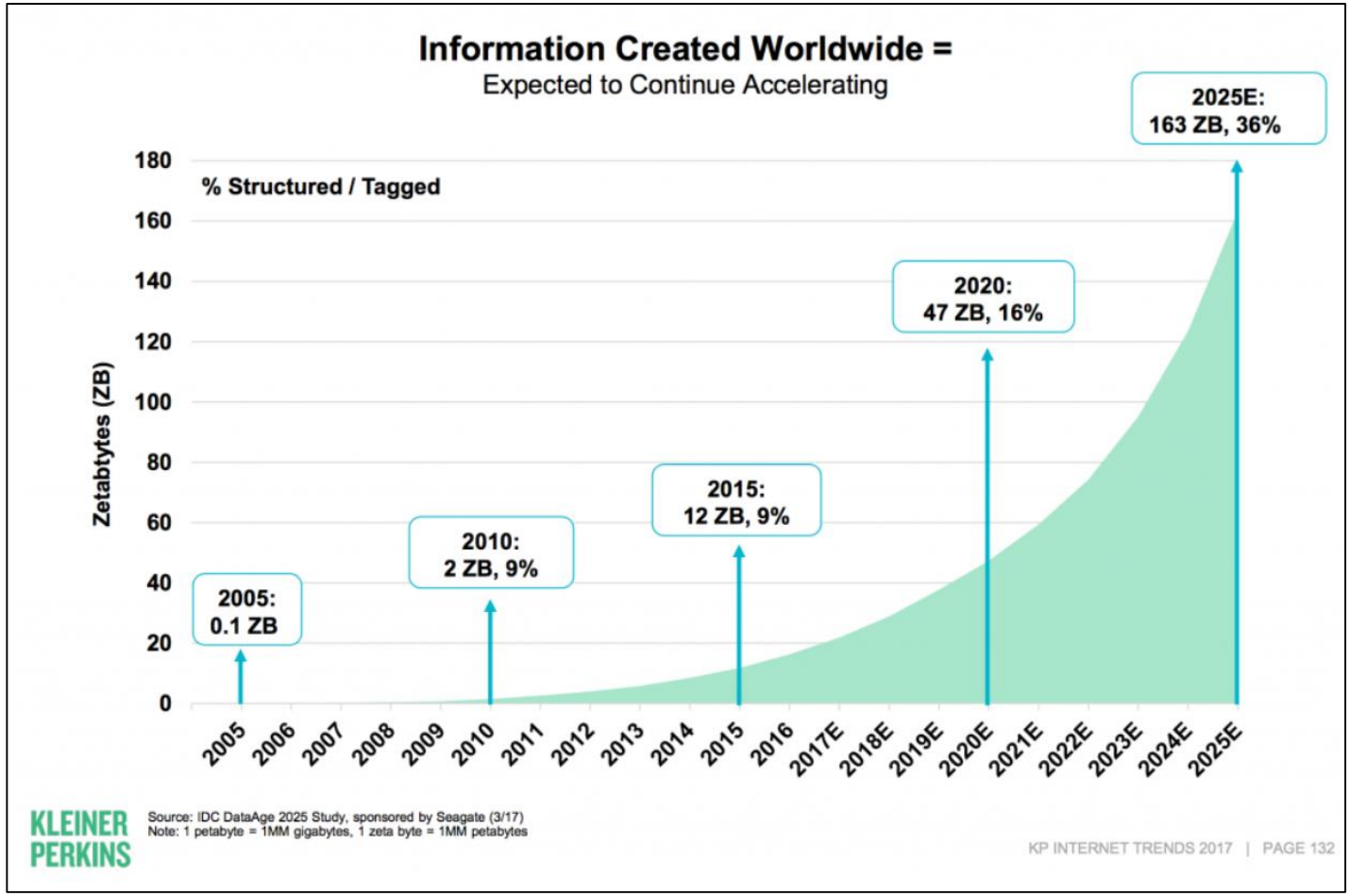
SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Data

Data Growth, Classification, Formats

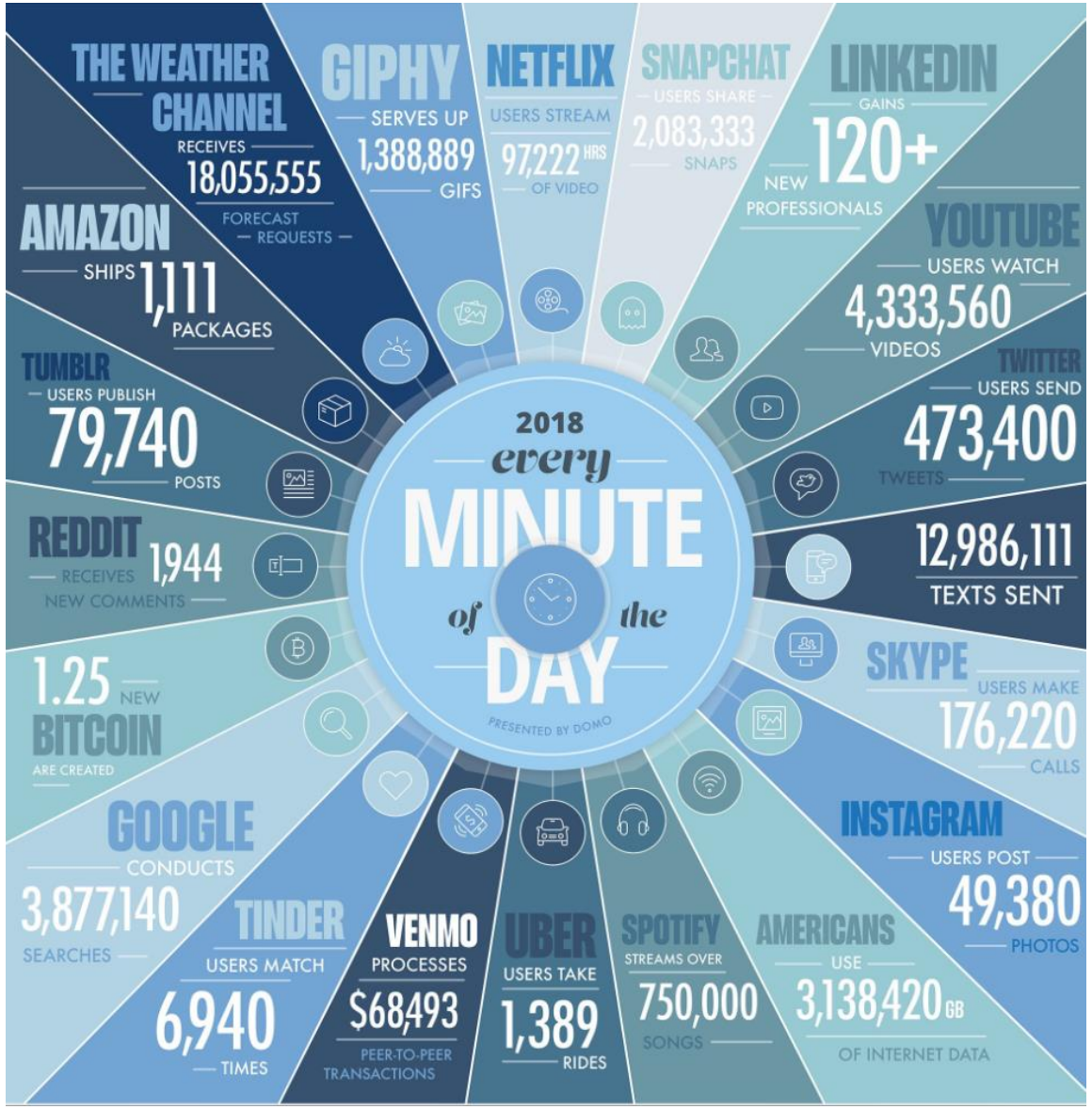
Data Revolution

- The data revolution and the data driven innovation using technology…
- https://www.youtube.com/watch?v=4ycC0DJqrpc



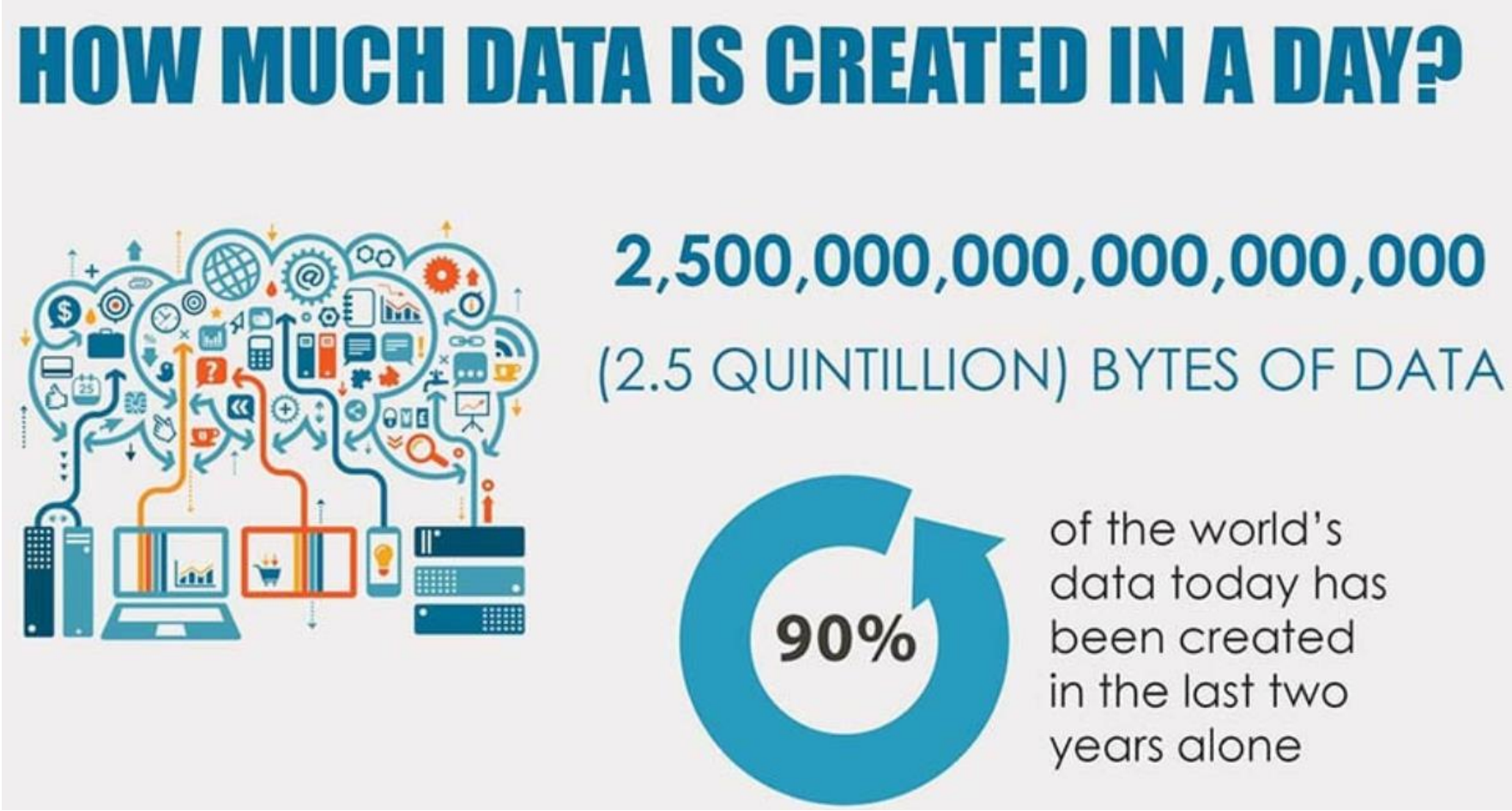
Data in 1 minute

- Weather Channel
- Giphy
- YouTube
- Texts
- Google



https://www.entrepreneur.com/article/314672/

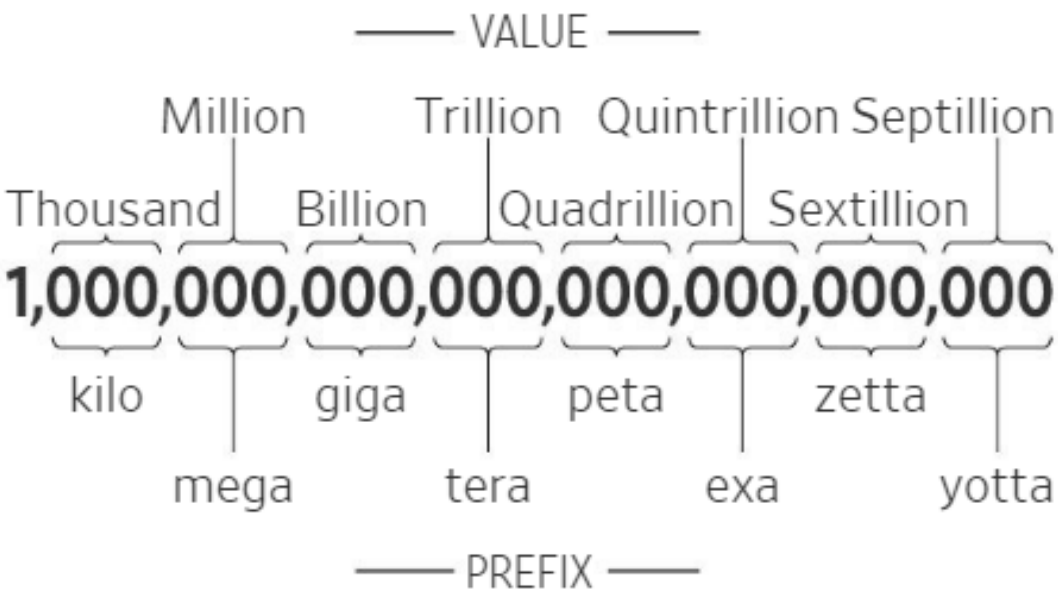
Data in a day



https://samikshyagautam.com/simple-introduction-of-data-science/

Counting Up

- The largest named prefix for a number is yotta, or 1 followed by 24 zeros, also known as septillion. A British metrologist proposes adding prefixes for even larger numbers.

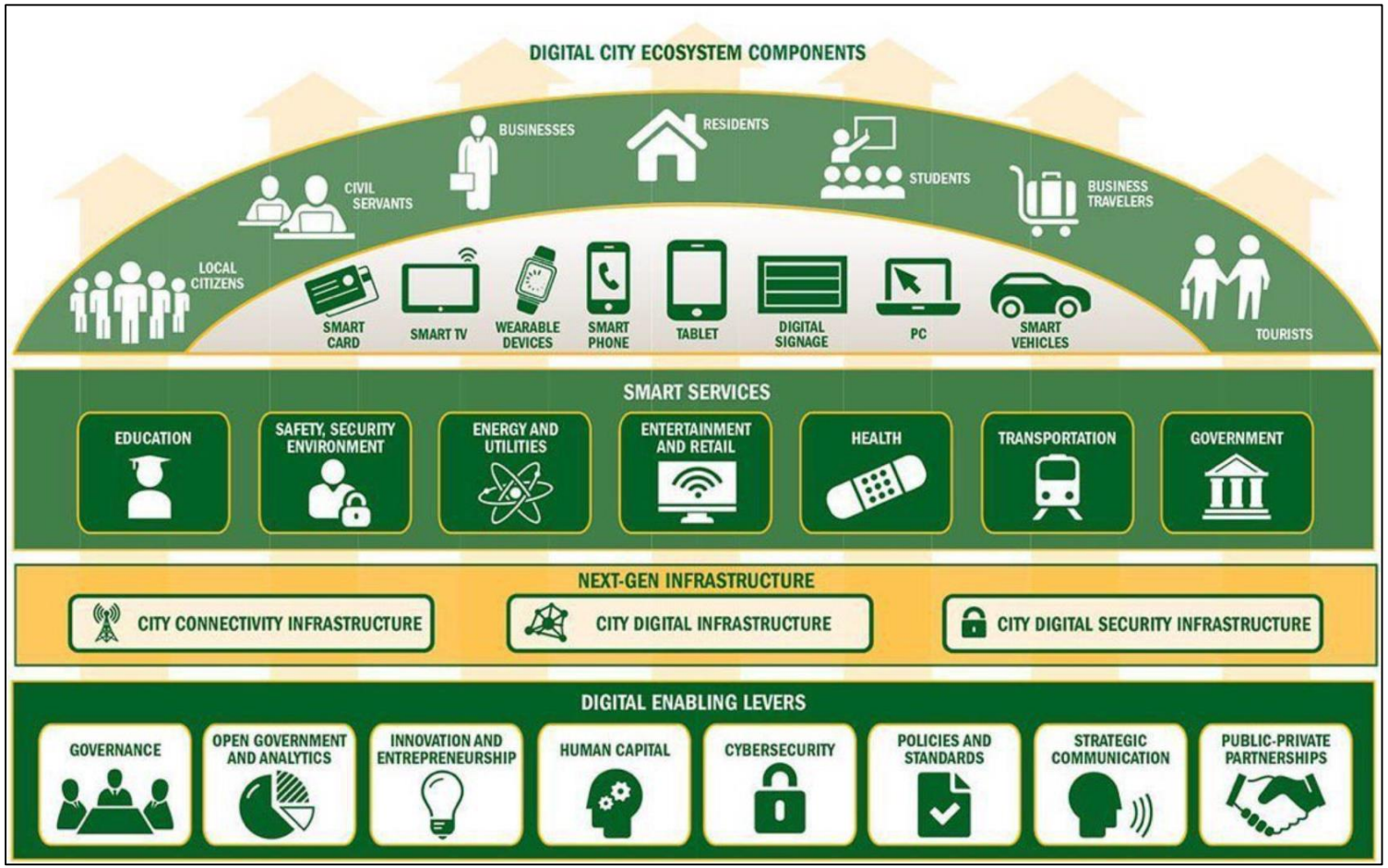


Source: National Institute of Standards and Technology

https://www.wsj.com/articles/to-keep-track-of-worlds-data-youll-need-more-than-a-yottabyte-11552048200/

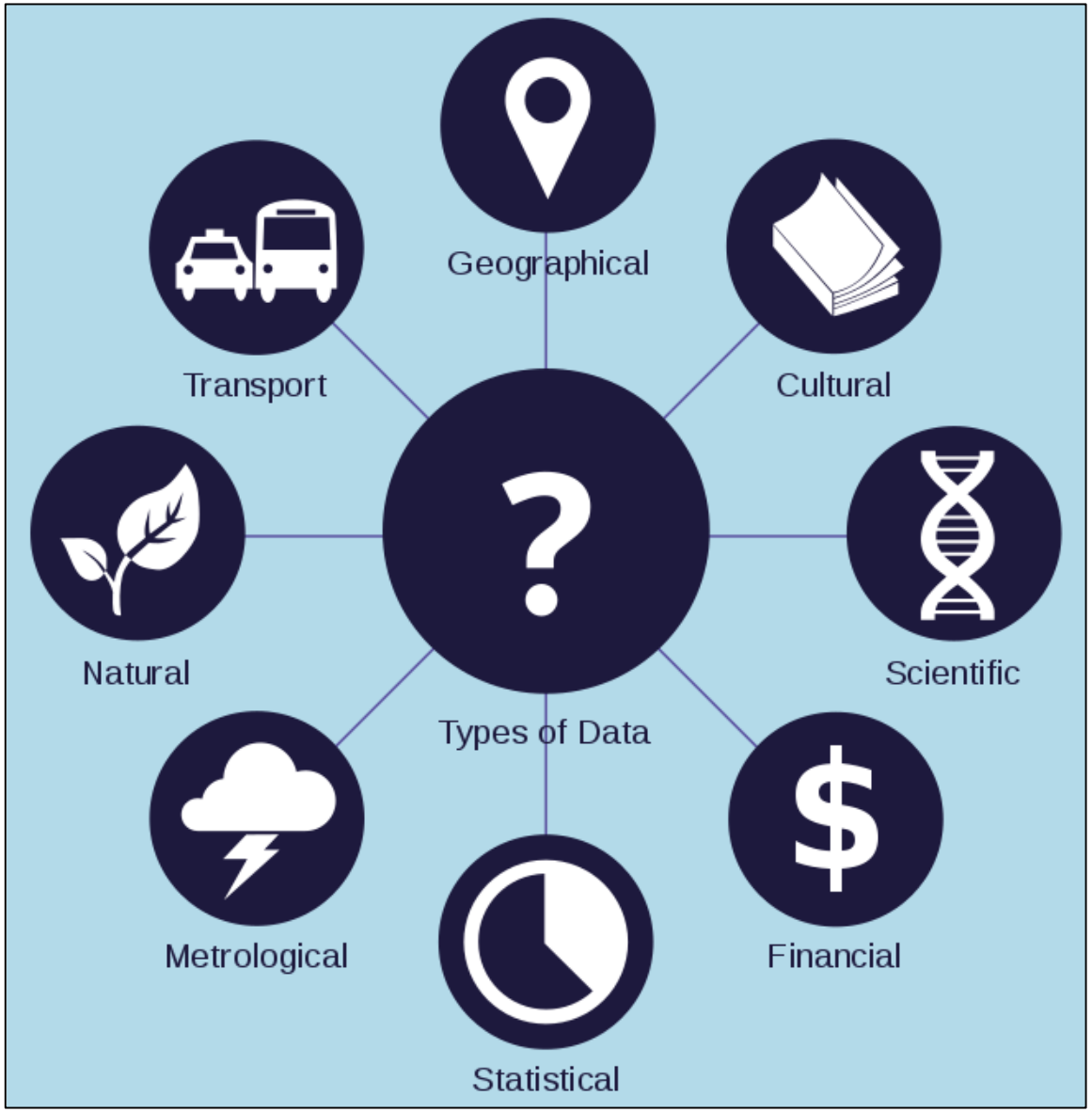
Ecosystem of digital smart city

- A **smart city** is an urban area that uses different types of electronic data collection sensors to supply information which is used to manage assets and resources efficiently.



Factors related to growth of data

- Sources of data
  - Users, Social Media, Internet, Artificial Systems, IoT, Natural Systems
- Business Analytics
  - Business Intelligence, Data Science
- Facilitated by technology
  - Hardware, Software, Networking



https://en.wikipedia.org/

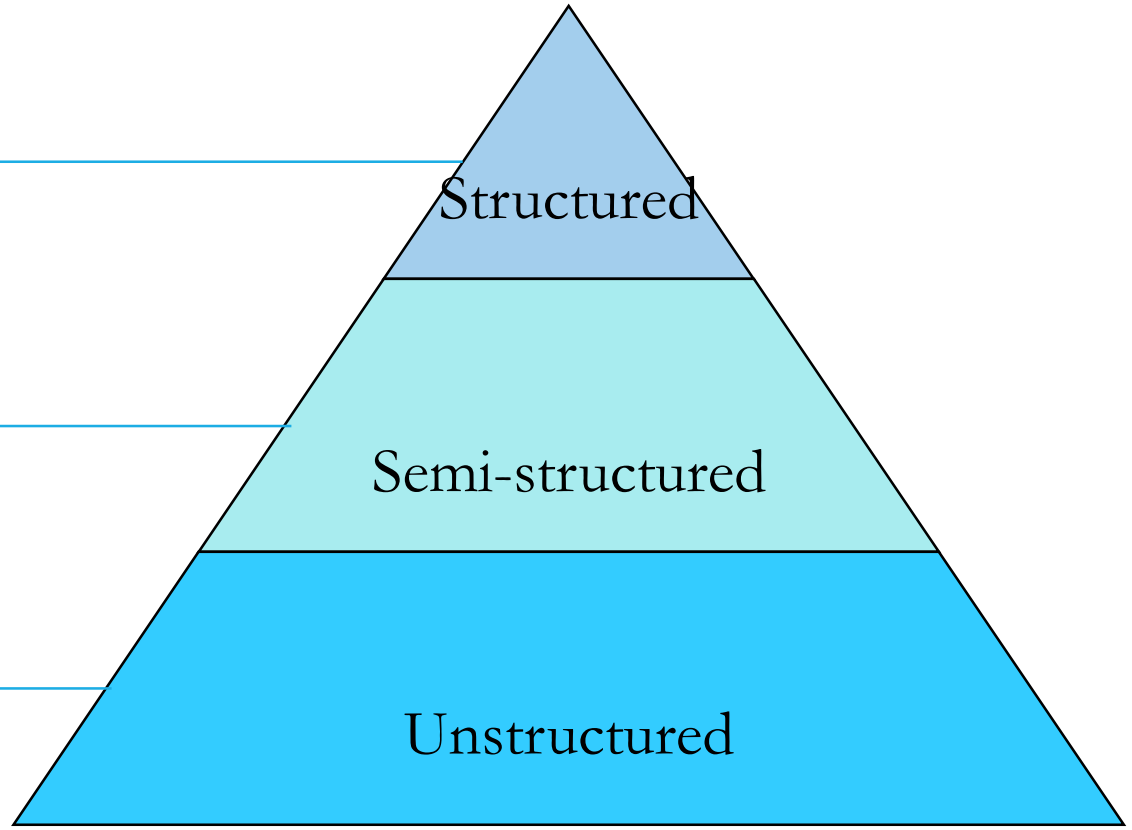
SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Data Classification

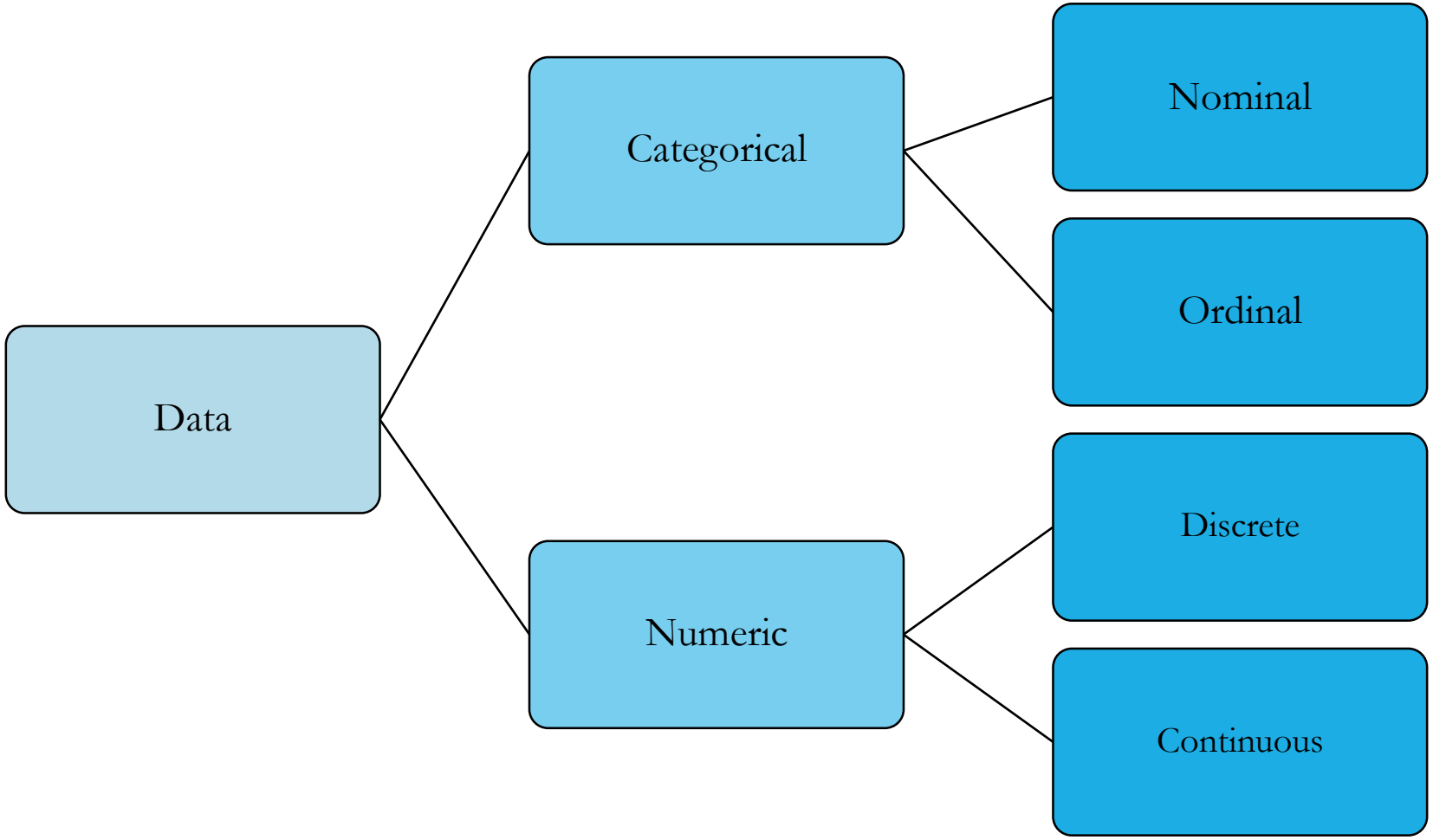
Well defined data types, fixed fields e.g. numeric, currency, date, address, text

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

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



Types of data



Categorical Data

- Classified or ordered
  - Nominal
    - Descriptions or labels - No sense of sequence
    - Ex: Red/blue/yellow OR M/F OR Chicago/NY/SFO
  - Ordinal
    - Meaningful Order
    - Ex: 1st (98.3%)/2nd (97.2%)/3rd(91.4%)

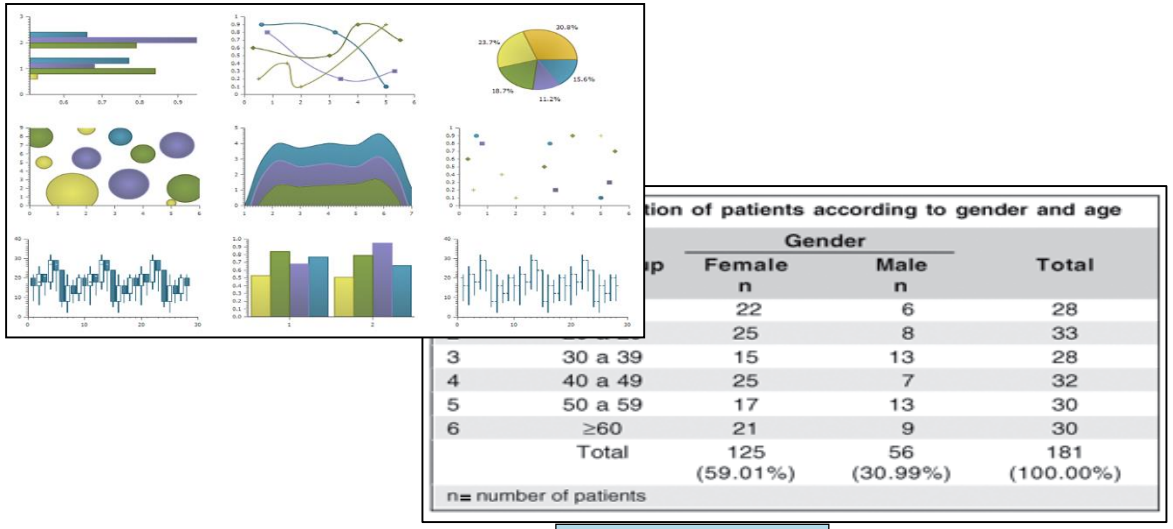
 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Numeric 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

Data Formats

Text

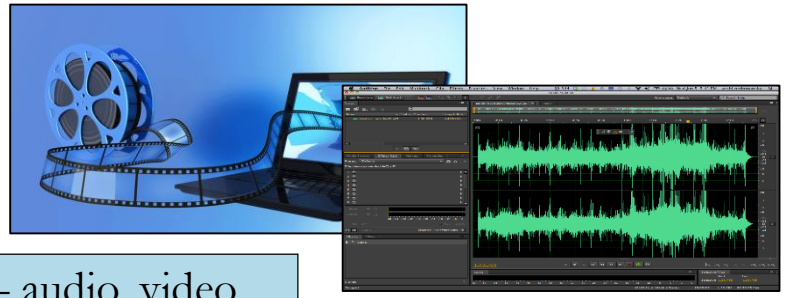


tion of patients according to gender and age

		Gender		
		Female n	Male n	Total
		22	6	28
		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

Numerical



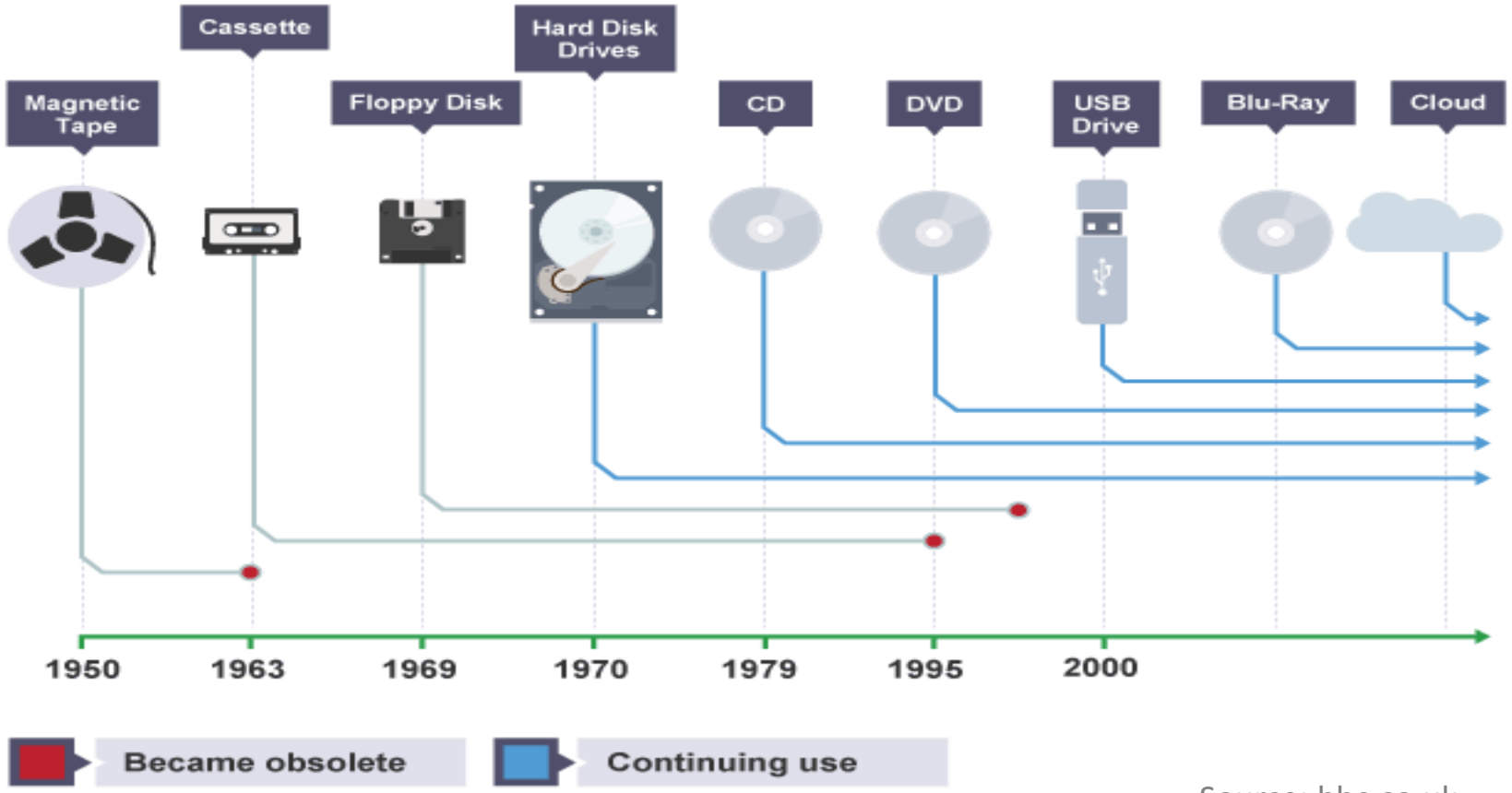
Media – audio, video

Geospatial

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

Evolution of storage



Source: bbc.co.uk

Binary Format

- Bit - smallest unit of information

0/1

- Byte - 8 bits of information

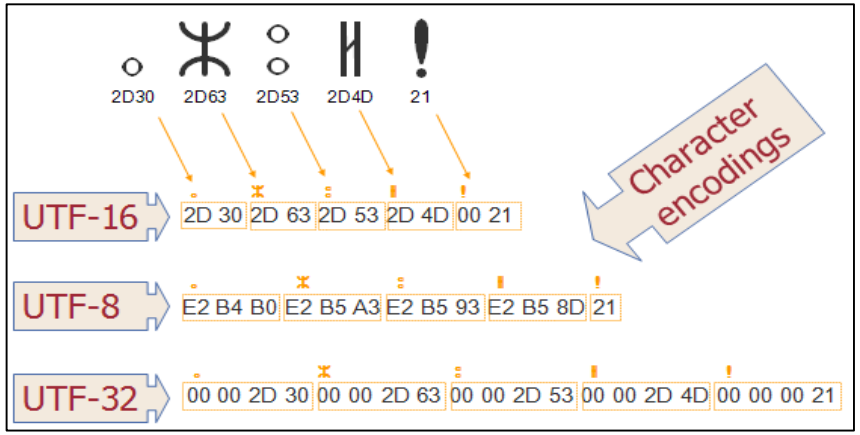
0/1 0/1 0/1 0/1 0/1 0/1 0/1 0/1

- 1 Kilo Byte $= 2^{10}$ Bytes $= 1024$ bytes
- 1 Kilo Byte $= 10^{3}$ Bytes $= 1000$ bytes

SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Character Sets and Encodings

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



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 formats – 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/

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

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>
```

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

JavaScript Object Notation (JSON)

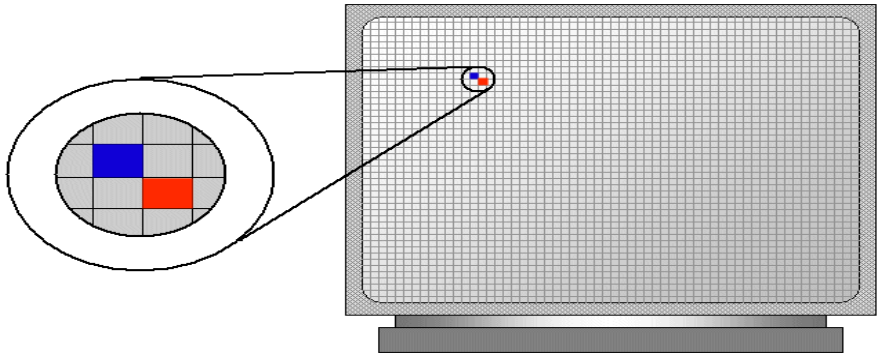
- Lightweight data-interchange format
- Platform independent
- Built on two structures :
  - Collection of name/value pairs
  - An ordered list of values

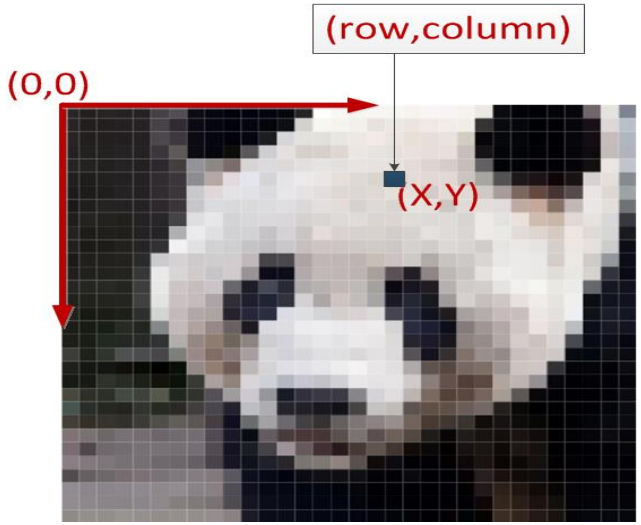
```
{
  "message": {
  "date": "07-04-2016",
  "hour": "08:30",
  "to": "Receiver",
  "from": "Sender",
  "body": "This is an email message"
}}
```

Digital Images

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

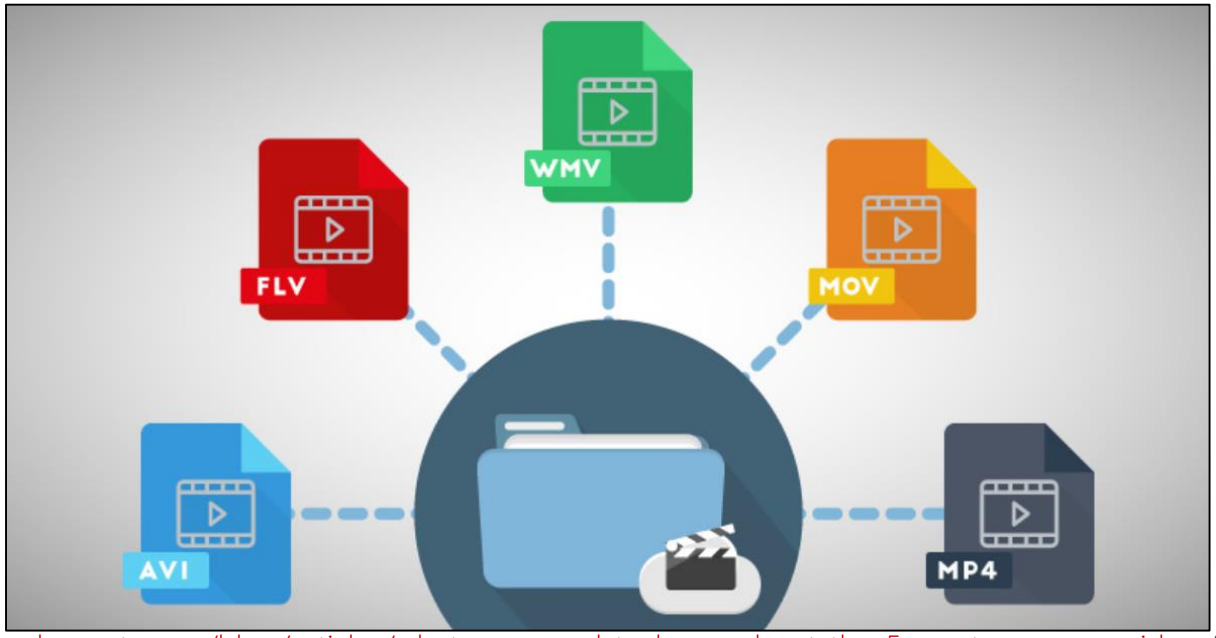
0	1	0	0	0	1	0
1	1	0	1	1	0	0
0	1	1	0	0	0	0
0	1	1	1	1	0	0
1	1	0	1	0	1	0
0	0	0	0	0	0	1





Video Formats

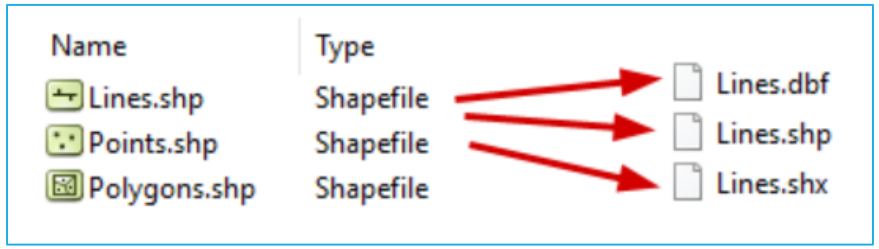
- File format for storing digital video data on a computer system. Video is almost always stored in compressed form to reduce the file size.



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

Spatial data

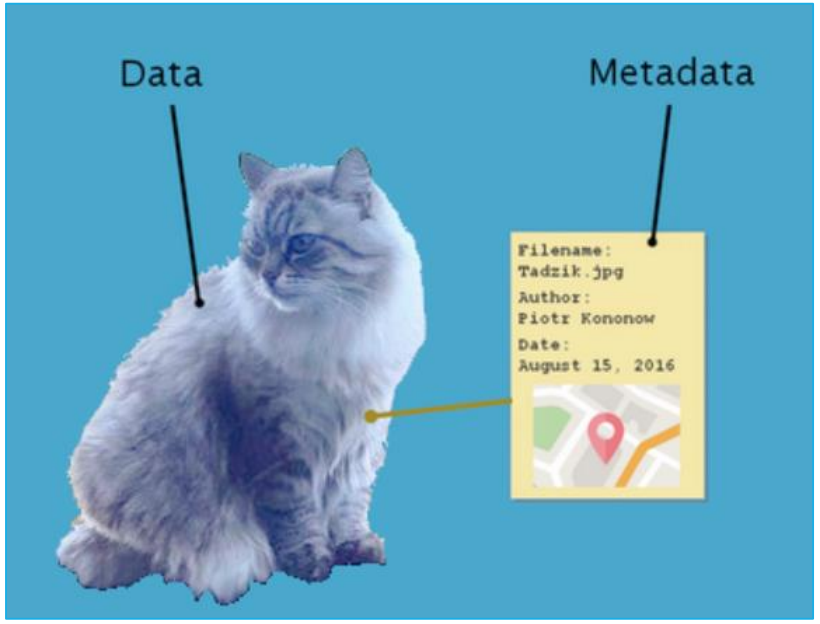
- Spatial data (geospatial data) represents location, size and shape of an object such as a building, Lake, mountain and includes attributes that provide more information about the entity that is being represented.
    - The shapefile is the most common geospatial file type. The three required files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



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



https://dataedo.com/

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.

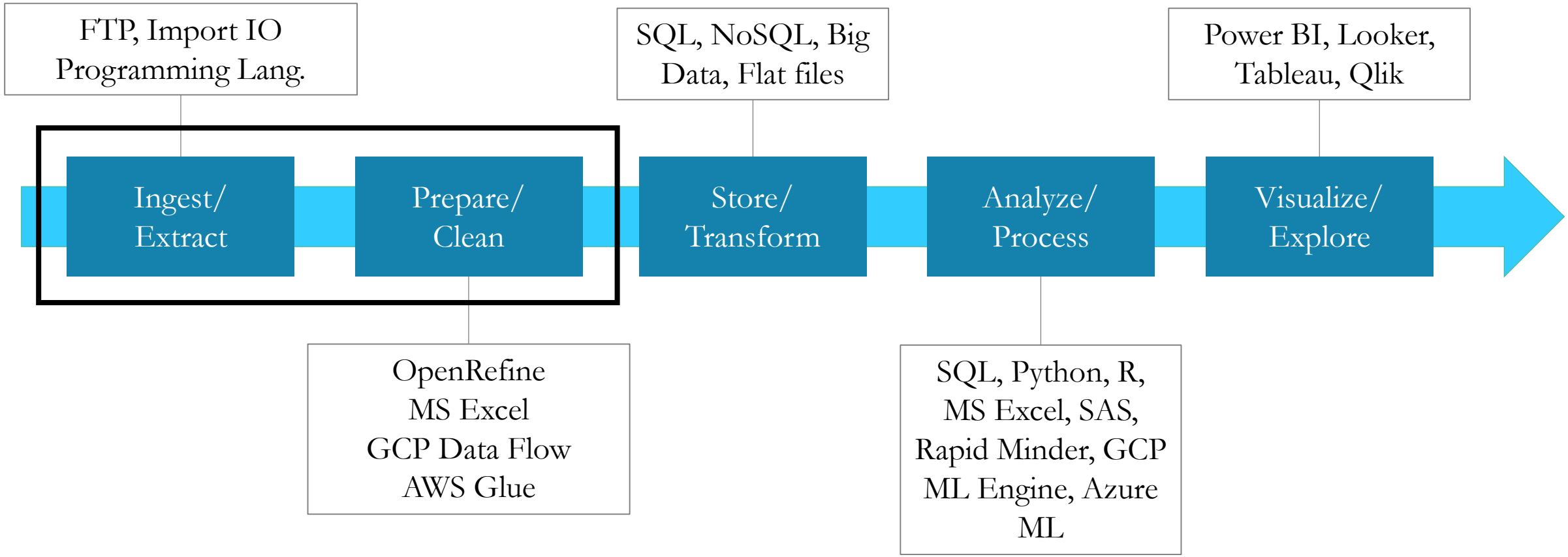
5.1.2 The address Table

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

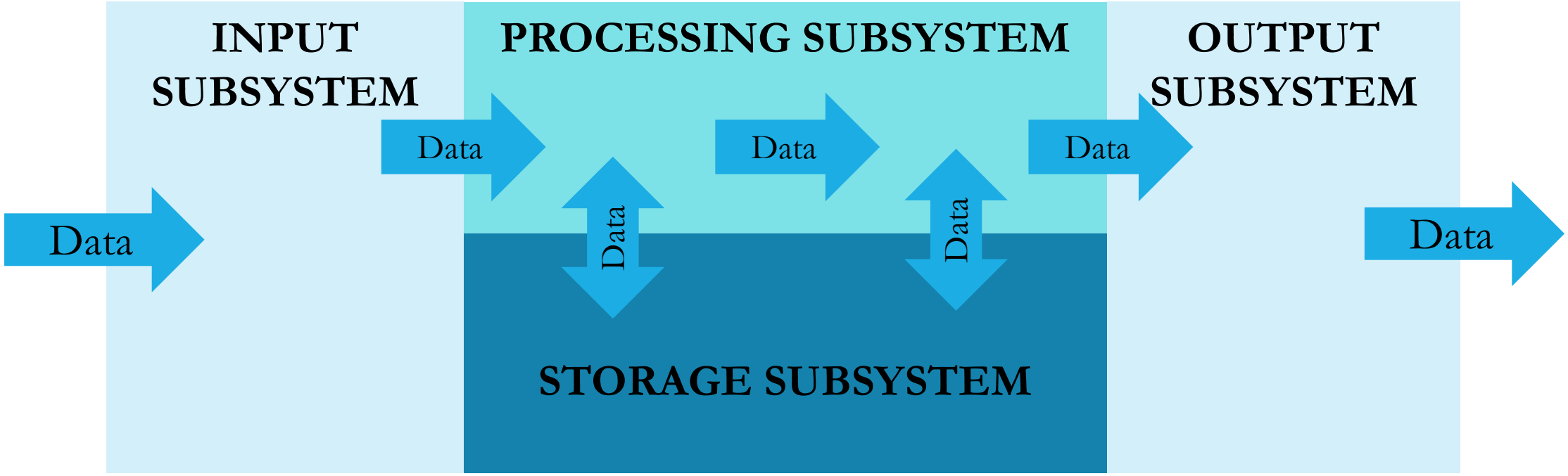
The `address` table primary key appears as a foreign key in the `customer`, `staff`, and `store` tables.

Data Ingestion & Preparation

Data Analysis Pipeline – Ingest, Prepare

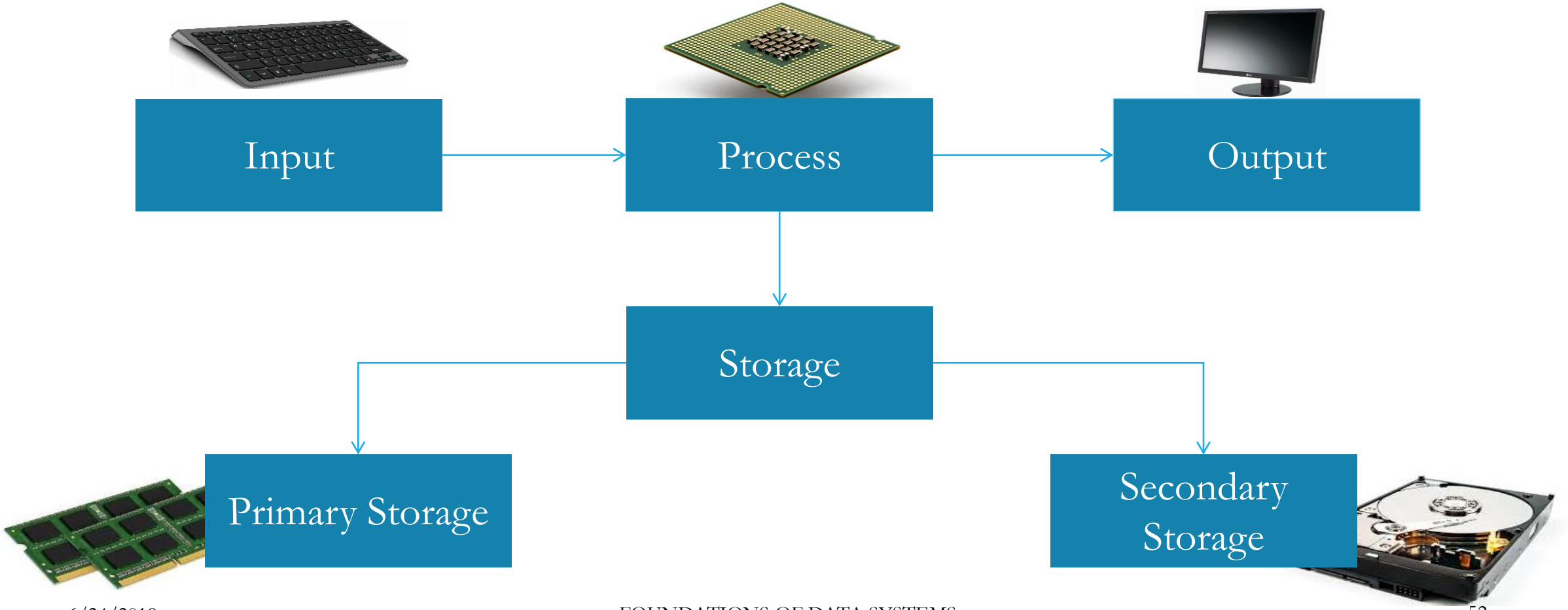


Data Processing



Source: wikipedia

Data Processing



Memory versus storage

- Memory (DRAM) is a component in your computer that allows for short-term data access

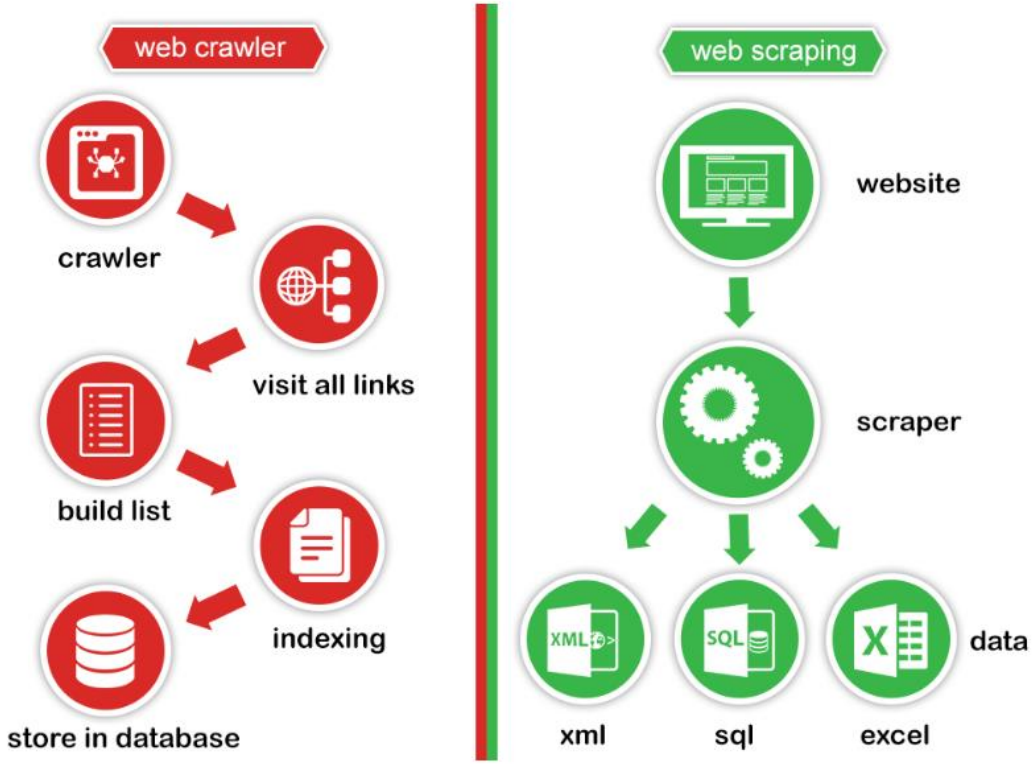
- Storage (whether in the form of a hard drive or SSD) is the component in your computer that allows for long-term data access

Data Ingestion

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

Web Scraping versus Web Crawling

- Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites.
  - Specific data gathered and copied from the web into a central local database or spreadsheet, for later retrieval or analysis.
- Web Crawling is the process of locating information on World Wide Web(WWW), indexing all the words in a document, adding them to a database, then following all hyper links and indexes and adds that information also to the database.
  - Collects and prepares data



http://prowebscraping.com/web-scraping-vs-web-crawling/

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Data ingestion requirements

- Define the incoming data from a business perspective.
- Document the context, lineage, and frequency of incoming data.
- Classify the security level (public, internal, sensitive, restricted) of the incoming data.
- Document the creation, usage, privacy, regulatory, and encryption business rules that apply to the incoming data.

Data ingestion requirements

- Identify the data owner (sponsor) or the source of the ingested data.
- Identify the data steward(s) charged with monitoring the health of the data.
- Continuously measure data quality as it resides within the enterprise

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
```

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Sources of Errors

- Manual entry
- Measurement errors
- Duplicate data entry
- Lack of well defined standards
- Numeric approximations
  - software and hardware constraints

Data Quality

- Data Quality - Fitness of the data for its intended use in operations, decision making and planning
  - 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

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

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

File naming best practices

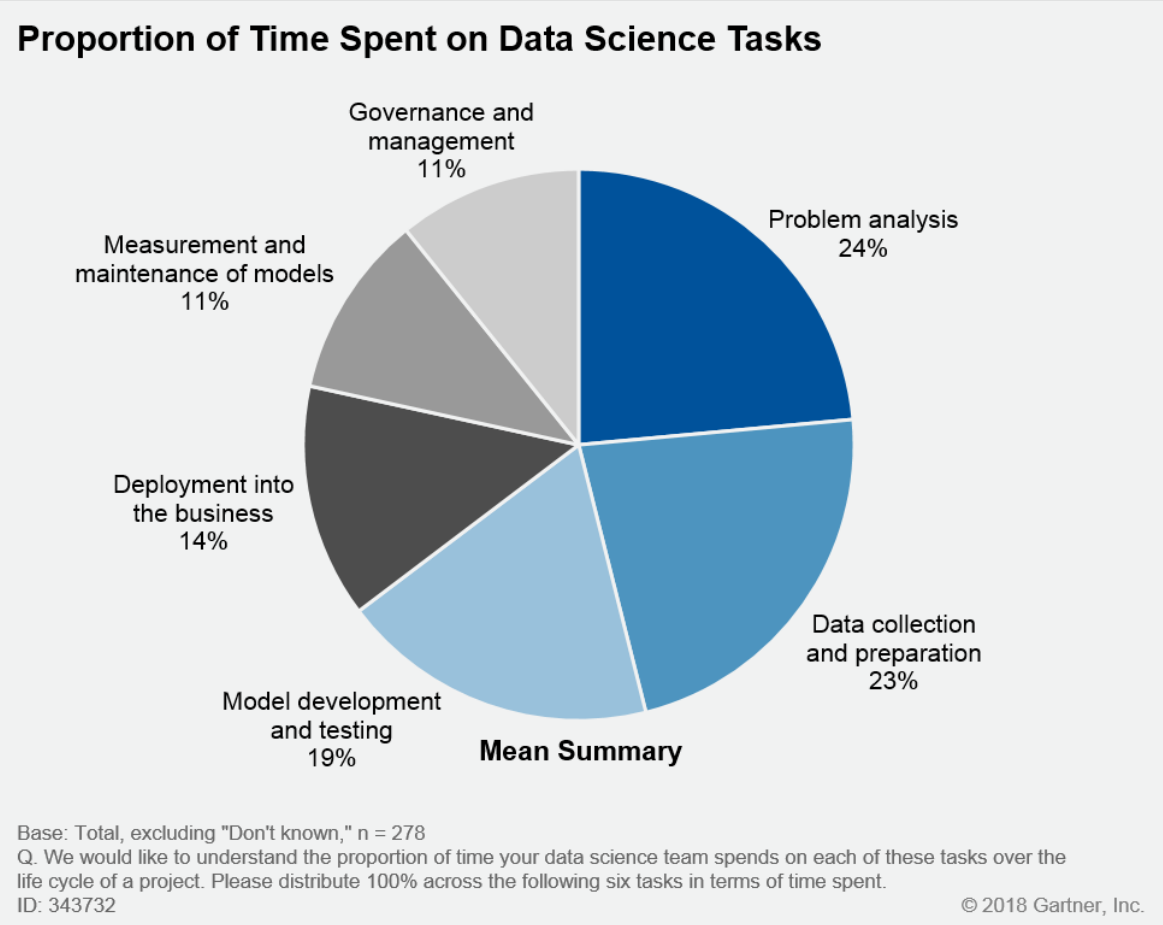
- Information to consider including in file names includes whatever information will allow you to distinguish your files from each other and clearly indicate to you what is in them.
- Consider including a README.txt file in your directory that explains your naming convention along with any abbreviations or codes you have used.
  - For additional details please visit the URL below http://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

Additional tips for file naming

- A good format for dates is YYYYMMDD (or YYMMDD). This makes sure all your files stay in chronological order, even over many years.
- Don't make file names too long; longer names do not work well with all types of software.
- Special characters should be avoided: ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ' "
- For sequential numbering, use leading zeros to ensure files sort properly. For example, use "0001, 0002…1001, etc" instead of "1, 2…1001, etc."
- Do not use spaces, because they are not recognized by some software. Instead use underscores (file_name), dashes (file--name), no separation (filename), or camel case (FileName).

Time spent in data science

- Problem analysis
- Data collection and preparation
- Data governance and management



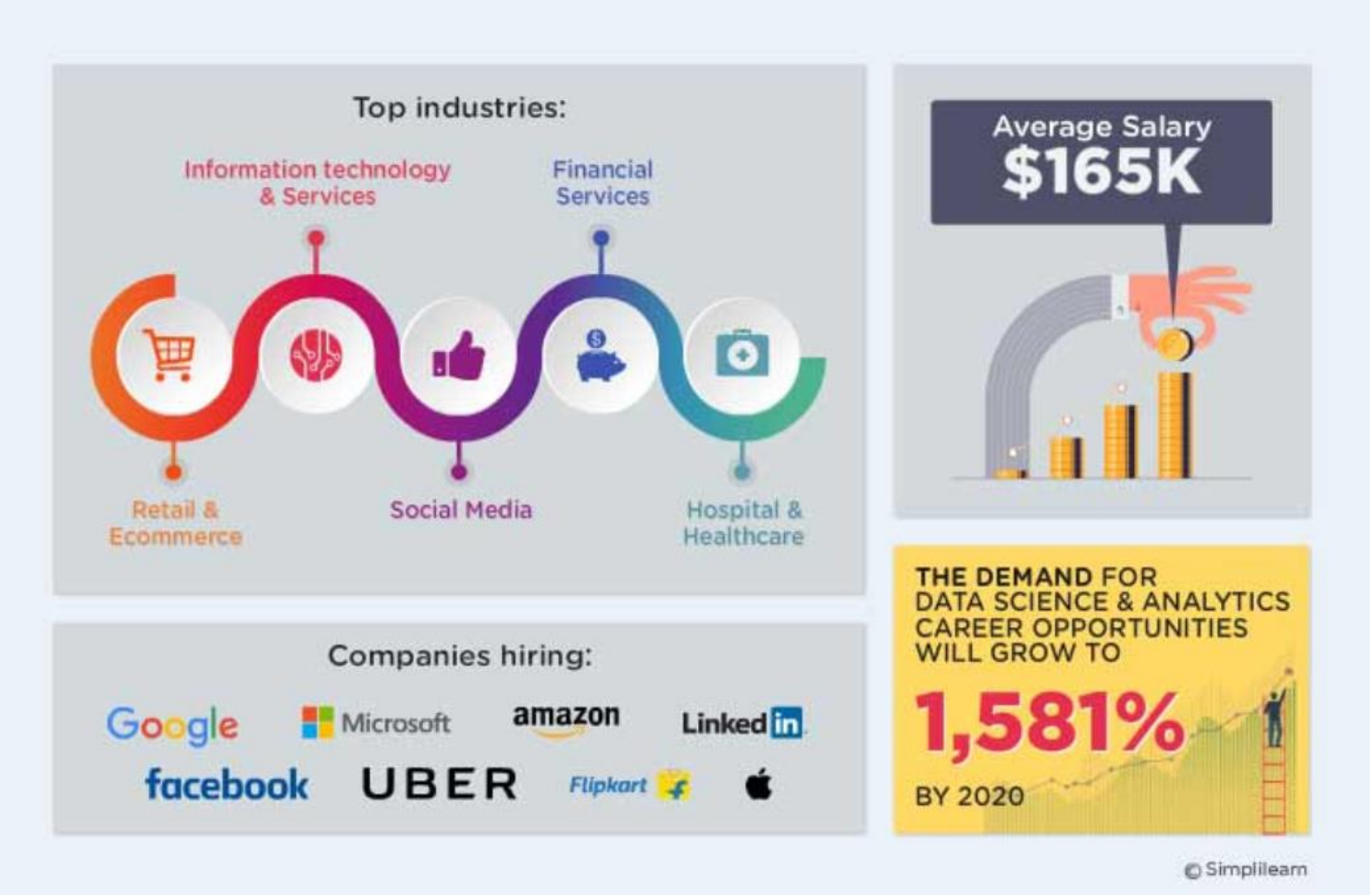
 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

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

Python for Data Engineering & Data Science

- Python is a general purpose programming language which is easy to use, for analytical and quantitative computing.



Exercise 1 – Data Ingestion

Data Ingestion using web scraping

Additional References:

- https://www.dataquest.io/blog/web-scraping-beautifulsoup/
- https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf
- https://hackernoon.com/web-scraping-tutorial-with-python-tips-and-tricks-db070e70e071

Exercise 2 – Open Refine

Data preparation using Open Refine

SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Exercise 3 – Pandas

Introduction & data wrangling with pandas

SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

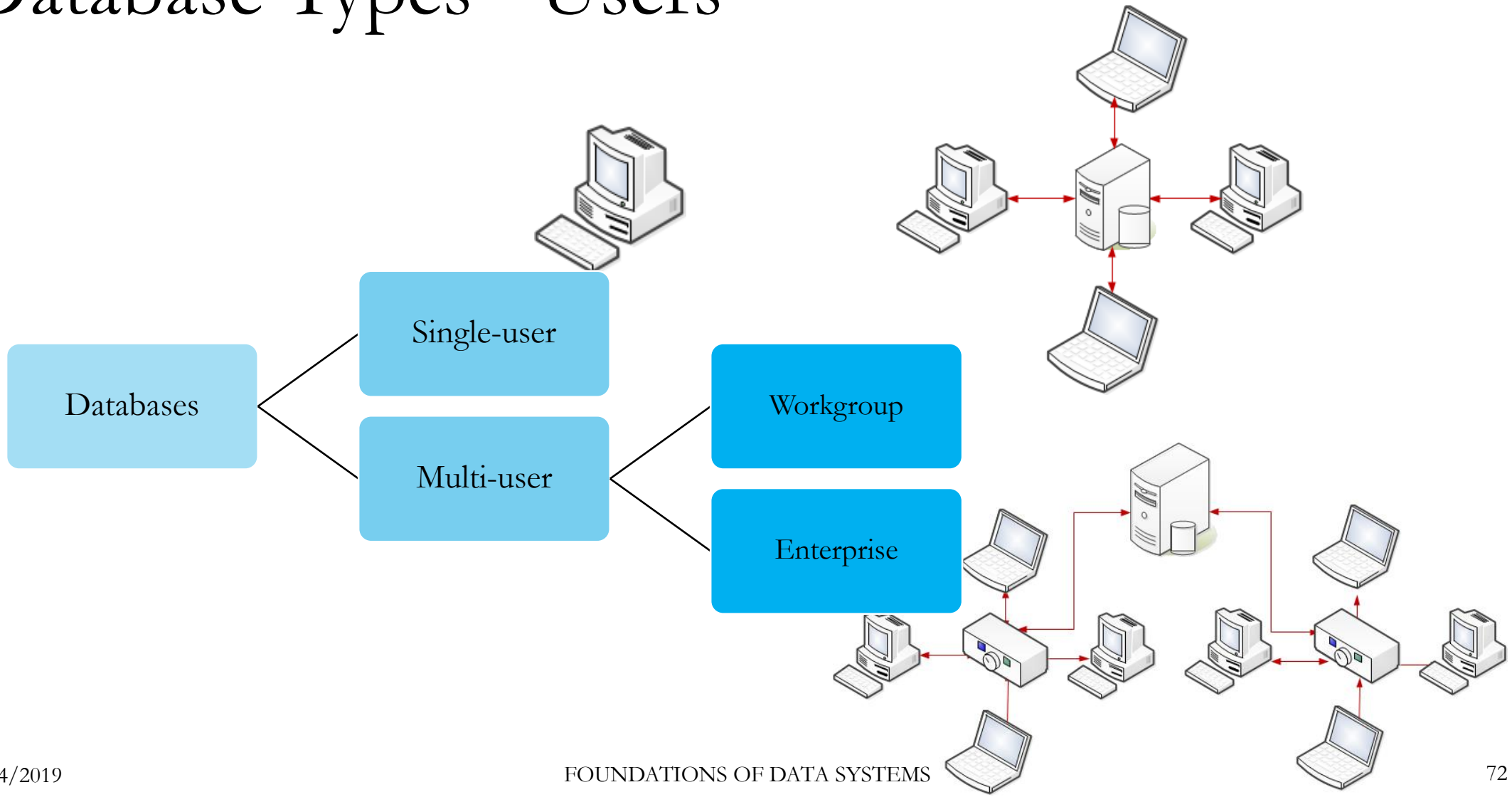
Types of Databases

Database Classification

- Number of users
- Primary Function
- Database location(s)
- Data type and structure

SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

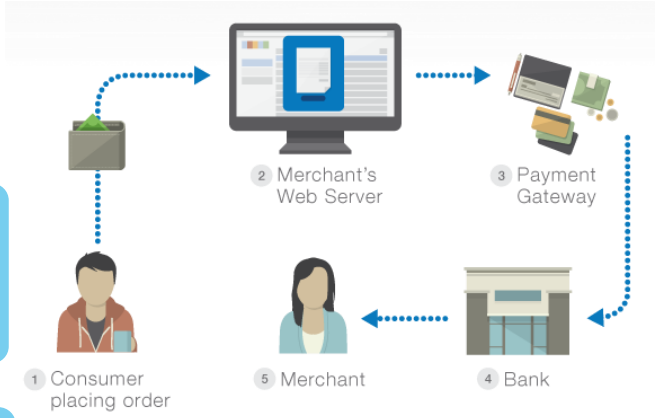
Database Types - Users

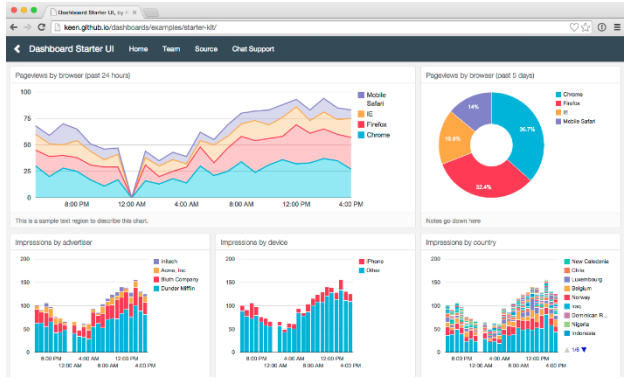


Database Types - Function

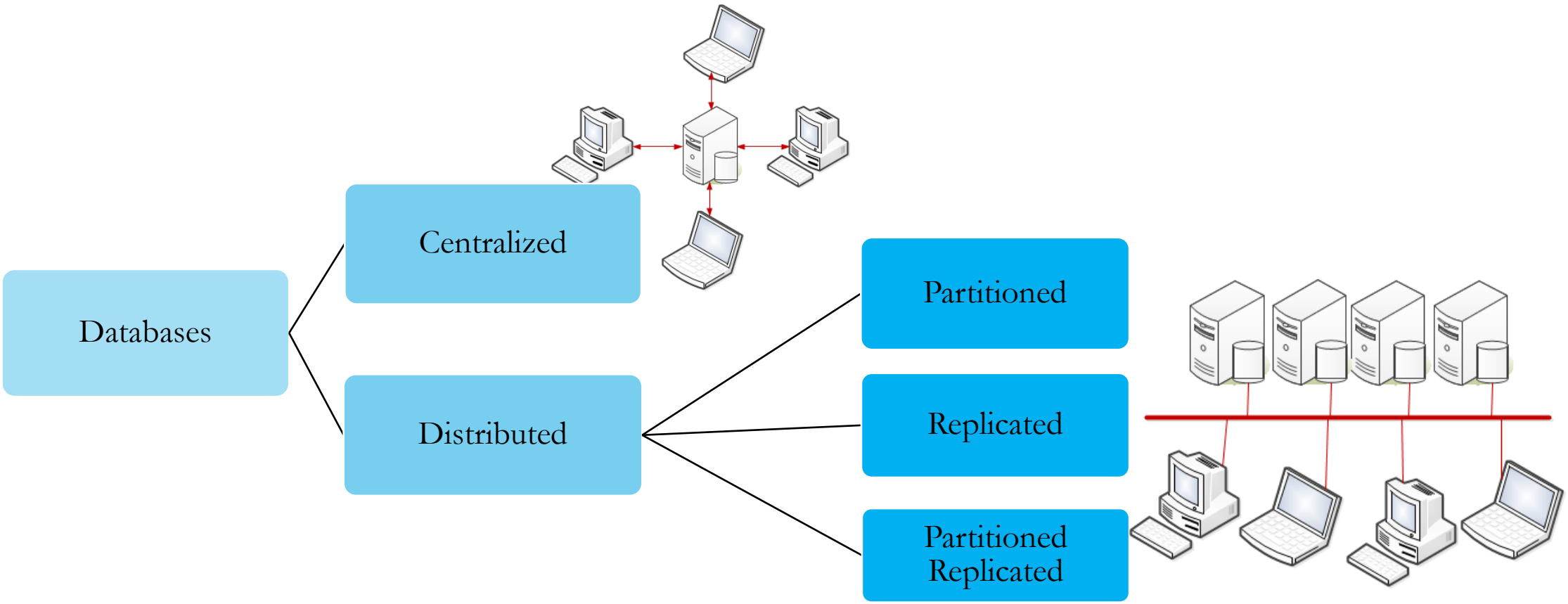
Databases

- Operational (OLTP)
- Analytical (OLAP)

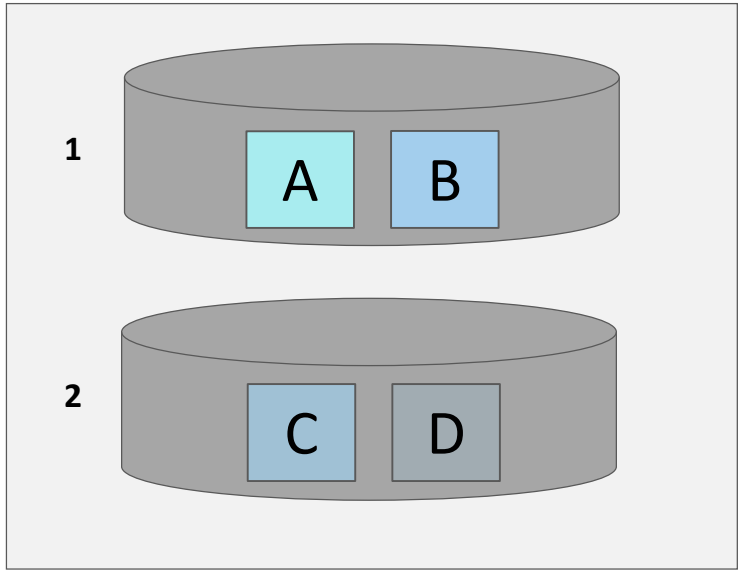




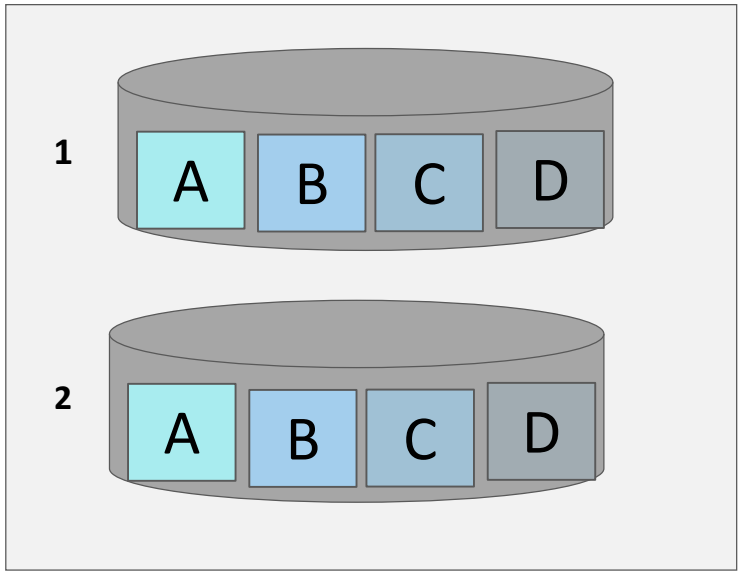
Database Types - Location(s)



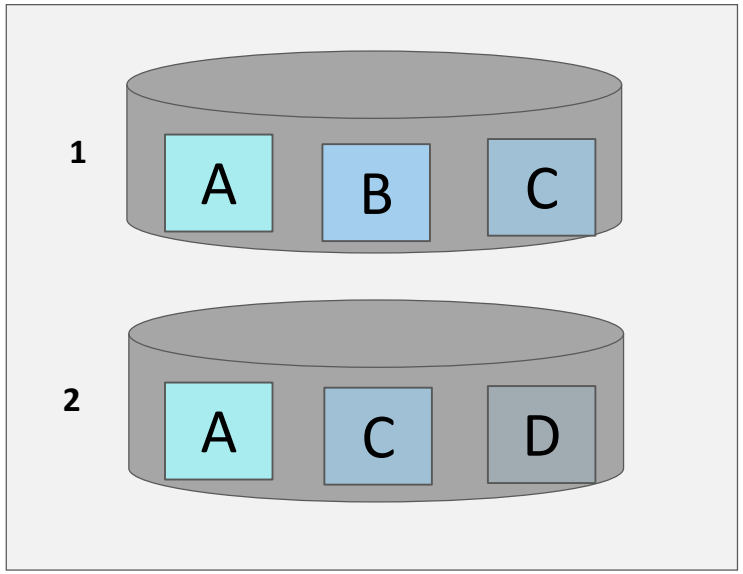
Distributed Databases



Partitioned
Non Replicated



Non Partitioned
Replicated



Partitioned
Replicated

Distributed Databases

Increasing flexibility, parallelism, availability

Increasing cost, complexity

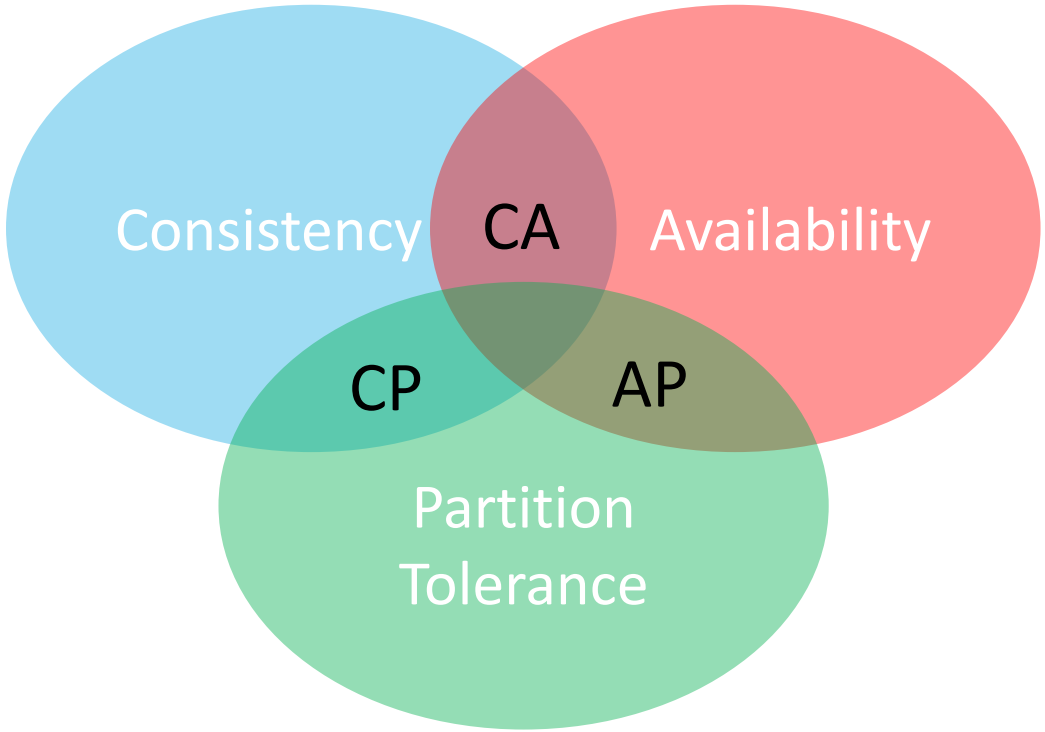
Partitioned Non-Replicated

Non-Partitioned Replicated

Partitioned Replicated

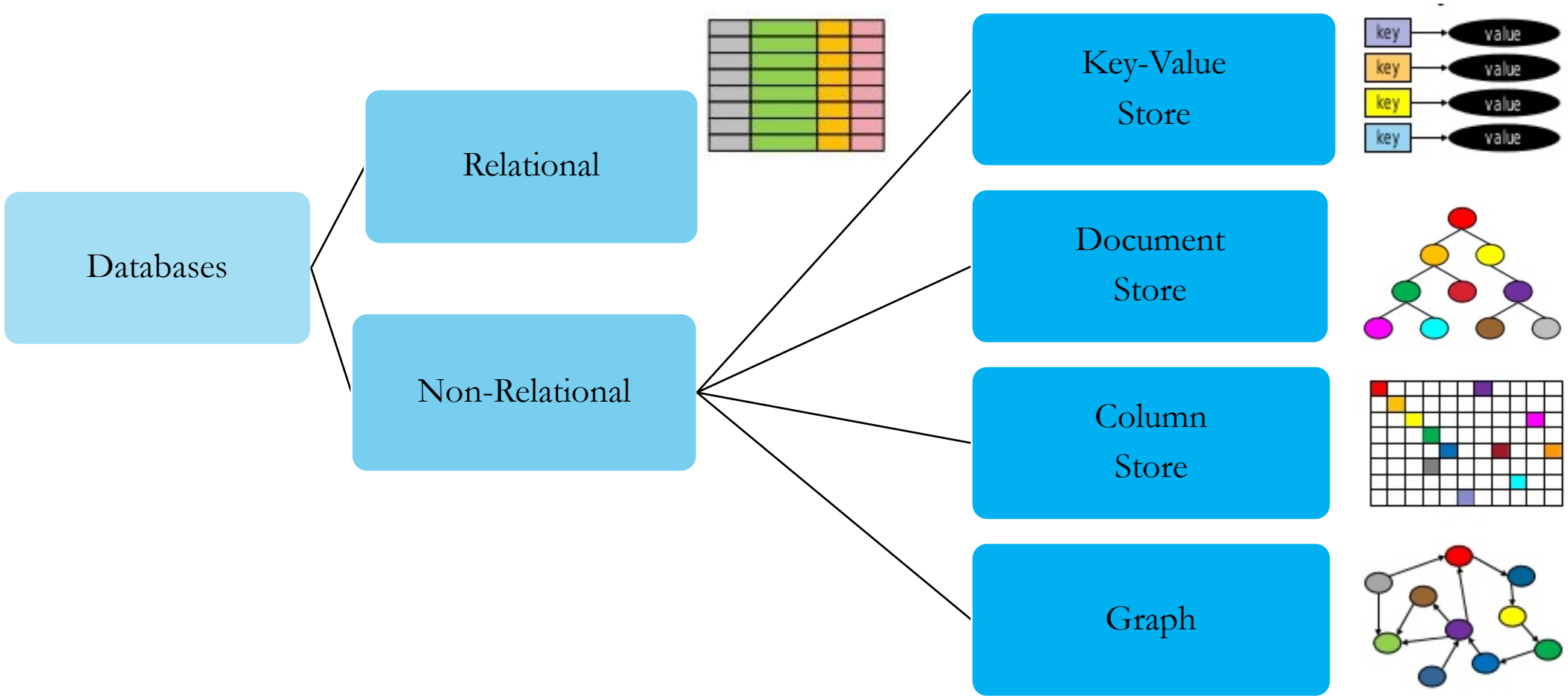
CAP Theorem

- Impossible for a distributed computer system to provide all three guarantees simultaneously
  - Consistency
    - Same view of data
  - Availability
    - All requests will receive a response
  - Partition tolerance
    - System continues to work despite message loss or partial failure



(Eric Brewer)

Database Types - Structure



Relational Databases

- Composed of tables with columns and rows
- Data stored across multiple related tables

Columns

Rows

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

SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Non-relational Databases

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

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

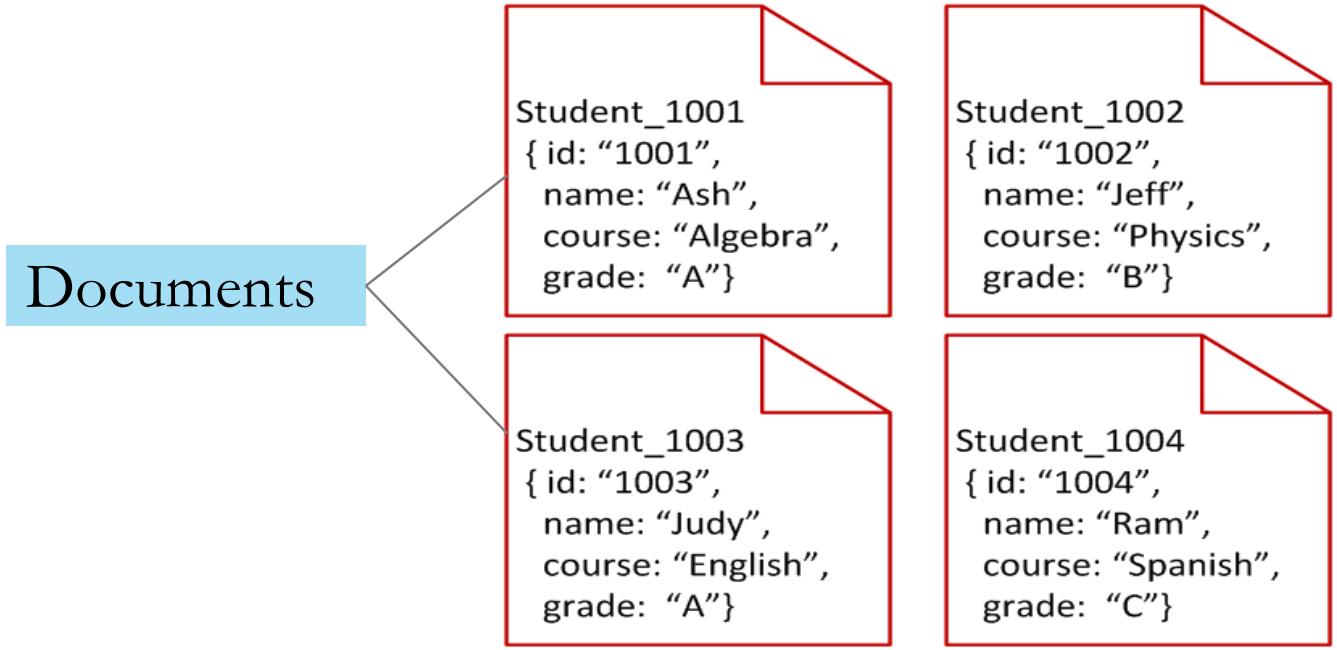
Key-Value Store

- 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



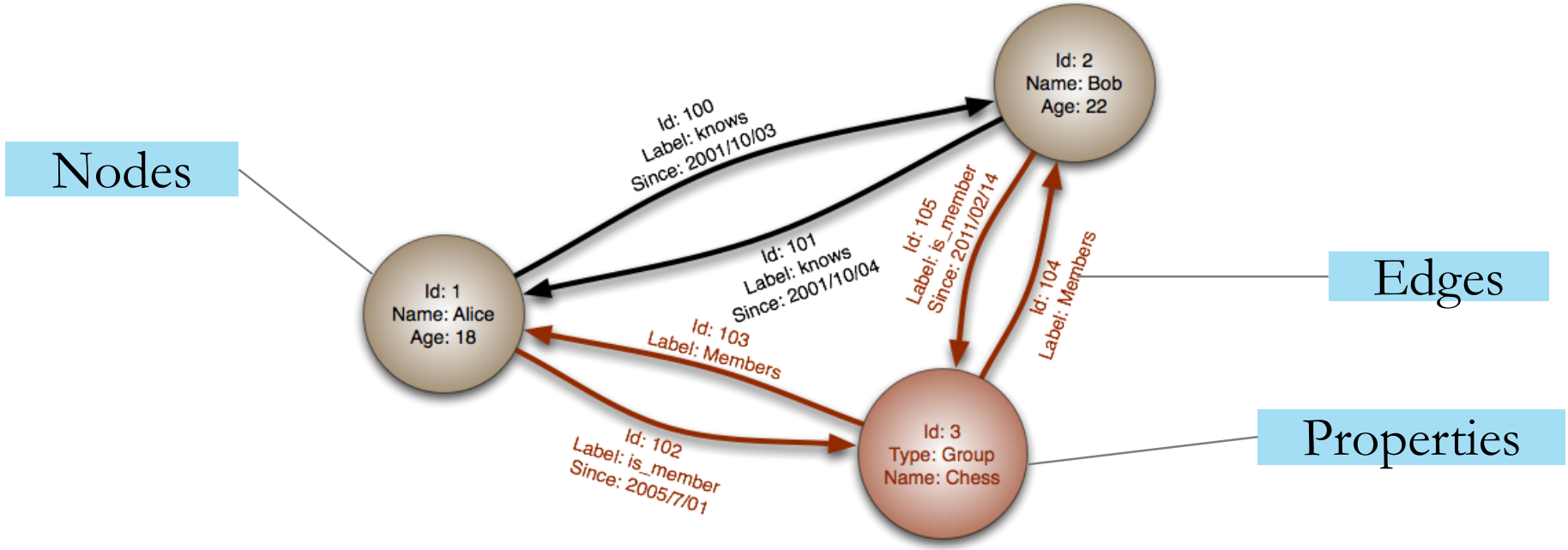
Column Store

- 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



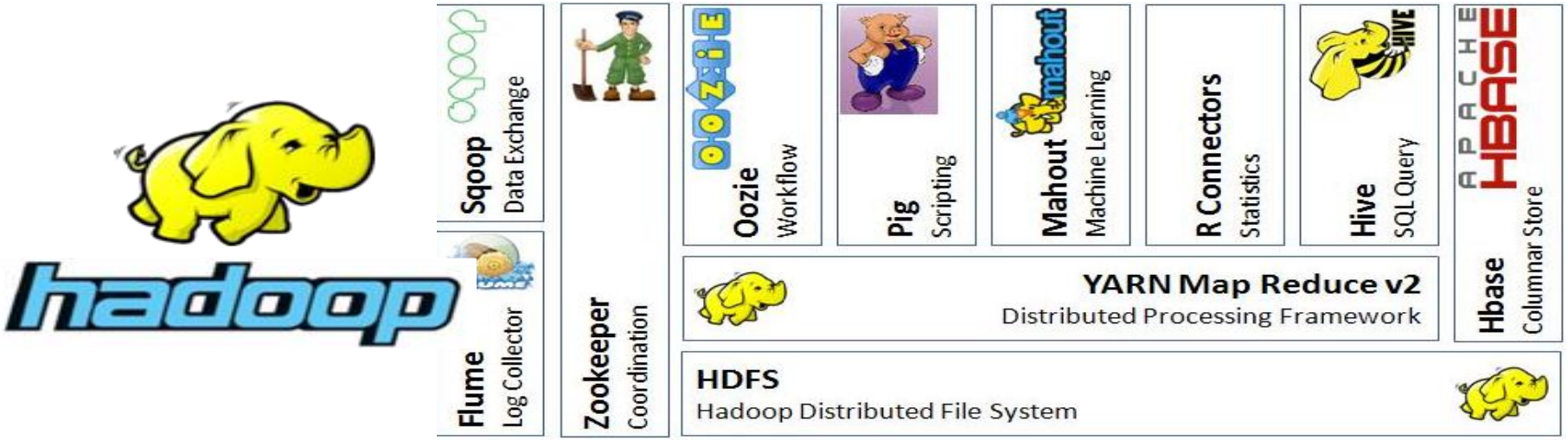
Source: Wikipedia

Popular Databases

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

Big Data

- Systems designed to handle very large datasets
- 3 V's – Volume, Velocity, Variety



Appendix

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Open Refine - Clustering

- Finding groups of different values that might be alternative representations of the same thing.
  - For example, the two strings "New York" and "new york" are very likely to refer to the same concept
- Key Collision Methods
  - Create an alternative representation of a value (a "key") that contains only the most valuable or meaningful part of the string
- Nearest Neighbor Methods
  - Provide a parameter (the radius, or k) which represents a distance threshold: any pair of strings that is closer than a certain value will be binned together.

https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth/

 SHREENIDHI BHARADWAJ | ALL RIGHTS RESERVED| DO NOT DISTRIBUTE

Open Refine - Key Collision Methods

- Fingerprinting (default)
  - Method is fast and simple & works relatively well in a variety of contexts and least likely to produce false positives.
- N-Gram Fingerprint
  - Similar to the fingerprint method described above but instead of using whitespace separated tokens, it uses n-grams, where the n (or the size in chars of the token) can be specified by the user.
- Phonetic Fingerprint
  - Uses a phonetic fingerprinting (specifically, Metaphone3 method for English and the Cologne phonetic keyer for German), which is a way to transform tokens into the way they are pronounced

https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth/

Open Refine - Nearest Neighbor Methods

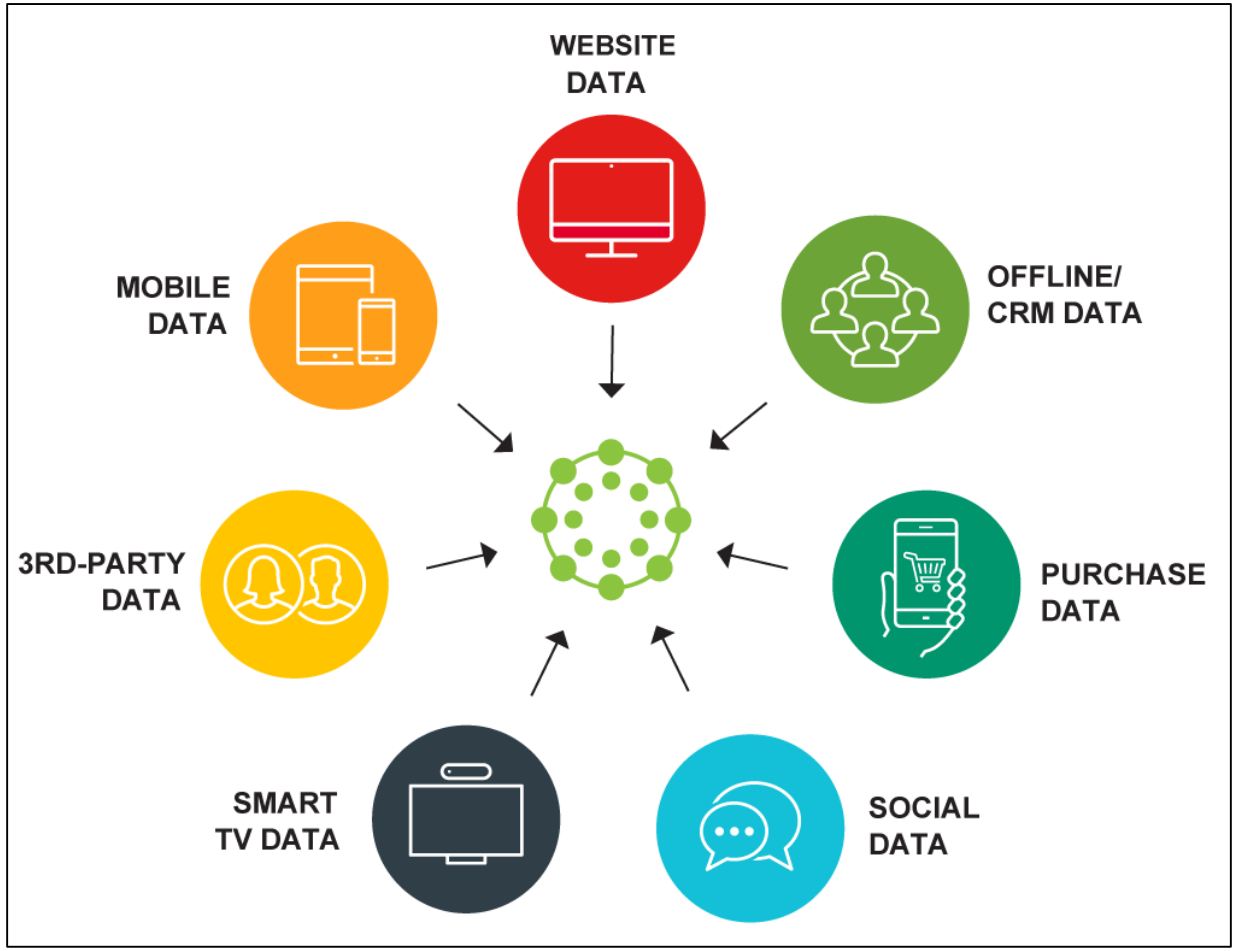
- Levenshtein Distance (also known as "edit distance")
  - Measures the minimal number of 'edit operations' that are required to change one string into the other.
- PPM (Kolmogorov complexity)
  - Estimate 'similarity' between strings and has been widely applied to the comparison of strings originating from DNA sequencing.

```
d(A,B) = comp(A+B) + comp(B+A) / (comp(A+A) + comp(B+B));
```

https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth/

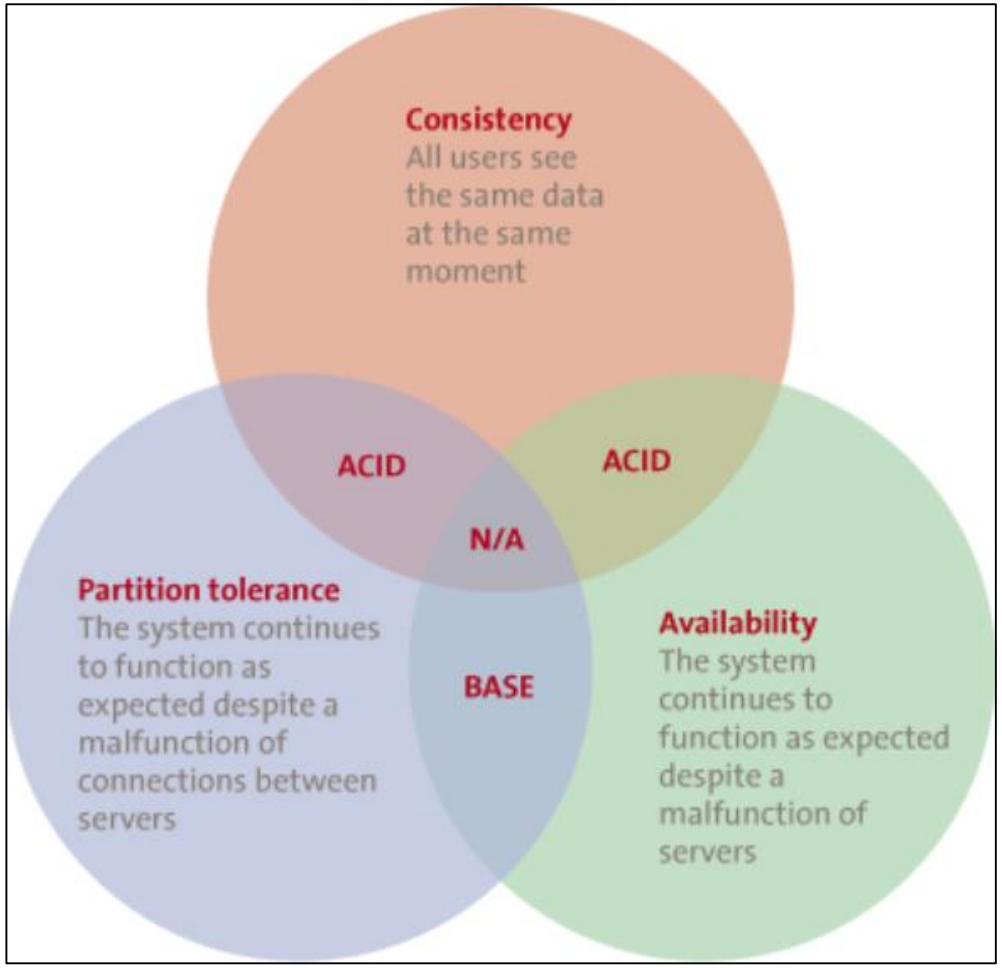
Data Management Platform (DMP)

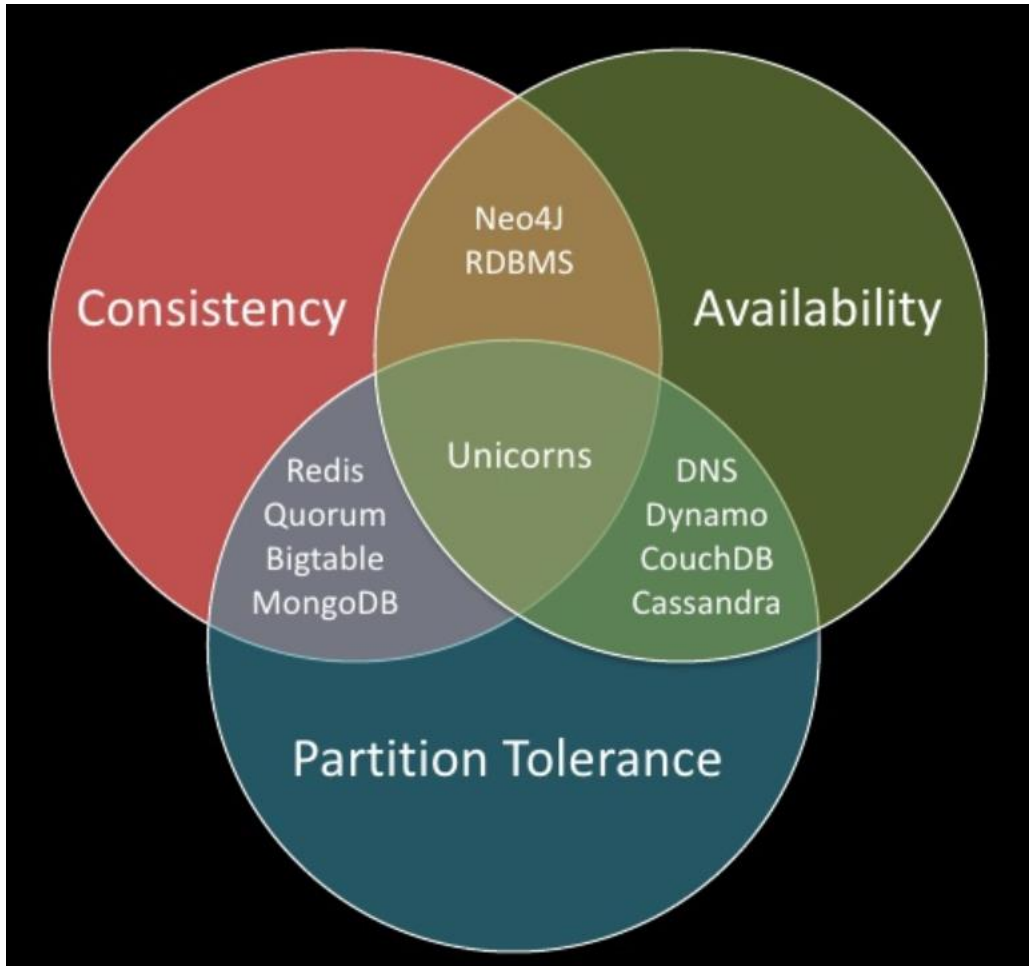
- Platform used for collecting and managing data, mainly for digital marketing purposes
- Generate audience segments, used to target specific users in online advertising campaign



http://xcademy.in/

CAP Theorem





http://www.anishsneh.com/ ; http://refactor.ir/

Readings

- http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/
- https://www.csoonline.com/article/2130877/data-breach/the-biggest-data-breaches-of-the-21st-century.html
- https://www.cnbc.com/2018/03/27/facebook-data-scandal-should-be-a-wake-up-call-about-online-footprint.html
- https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them
- https://www2.deloitte.com/insights/us/en/industry/public-sector/chief-data-officer-government-playbook/data-as-an-asset.html

References

- https://www.iconfinder.com/
- https://en.wikipedia.org/