Big Data

Big Data Course

Mostafa Nabieh

وزارة الاتصالات
وتكنولوجيا المعلومات

MANNING

وزارة الاتصالات
وتقنية المعلومات
MINISTRY OF COMMUNICATIONS
AND INFORMATION TECHNOLOGY

iTi AI TALENTS HUB

UDACITY

PLURALSIGHT

Learning Space

YOUR SPACE TO LEARN
FUTURE SKILLS

Mostafa Nabieh

HP Sustainable Impact @HPSustainable · Sep 23

.@HP and the @_AfricanUnion Commission are partnering to drive #entrepreneurial skills development in Africa. This includes leveraging HP LIFE online courses, providing students with access to quality #education needed to create a better tomorrow. bit.ly/32TxO9E

HUAWEI

EPITA

ÉCOLE D'INGÉNIEURS EN INFORMATIQUE

IBM

Mostafa Nabieh

CONTENTS

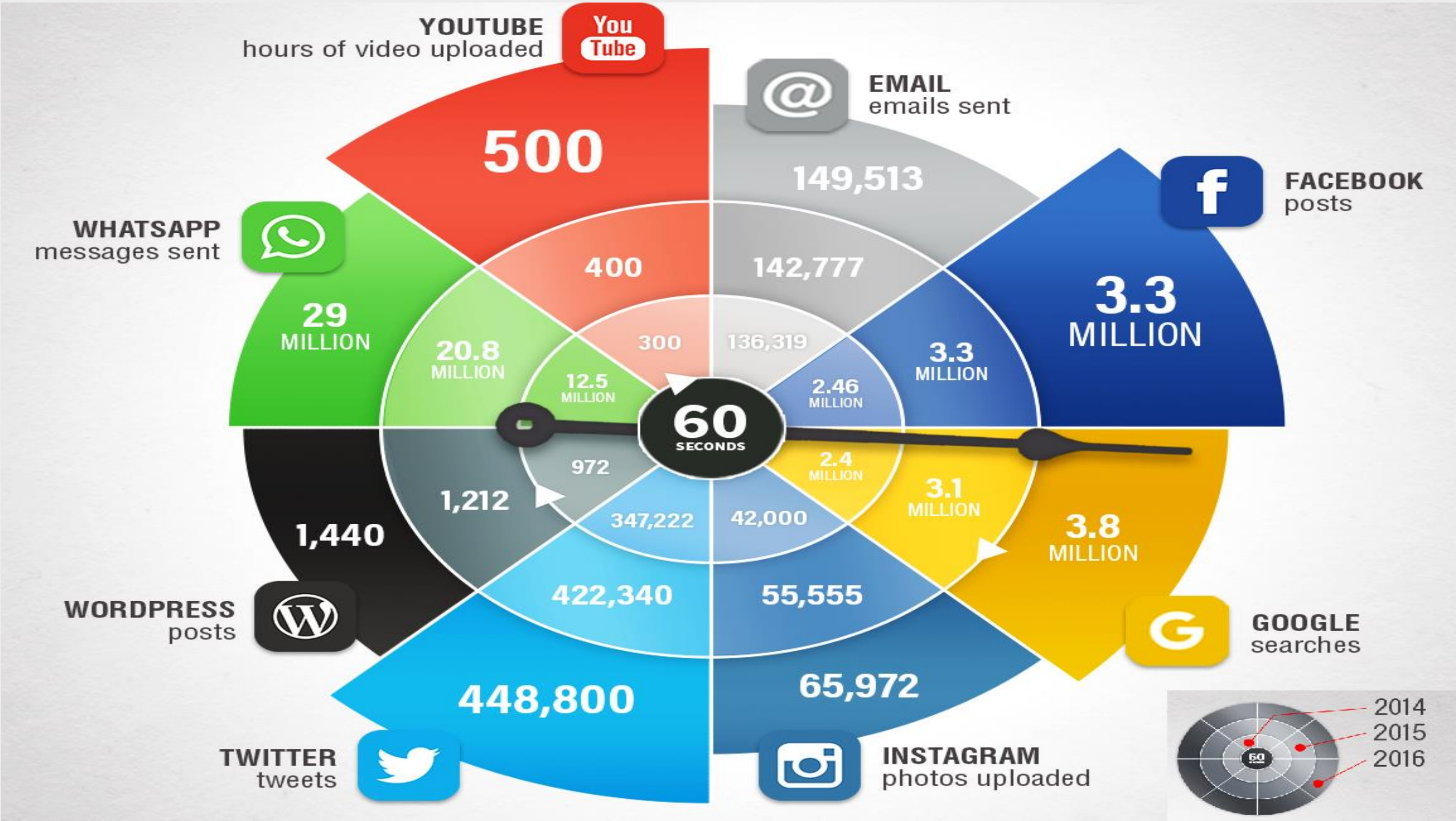
WHAT IS DATA ENGINEERING?

BIG DATA ECOSYSTEM

BIG DATA LIFECYCLE

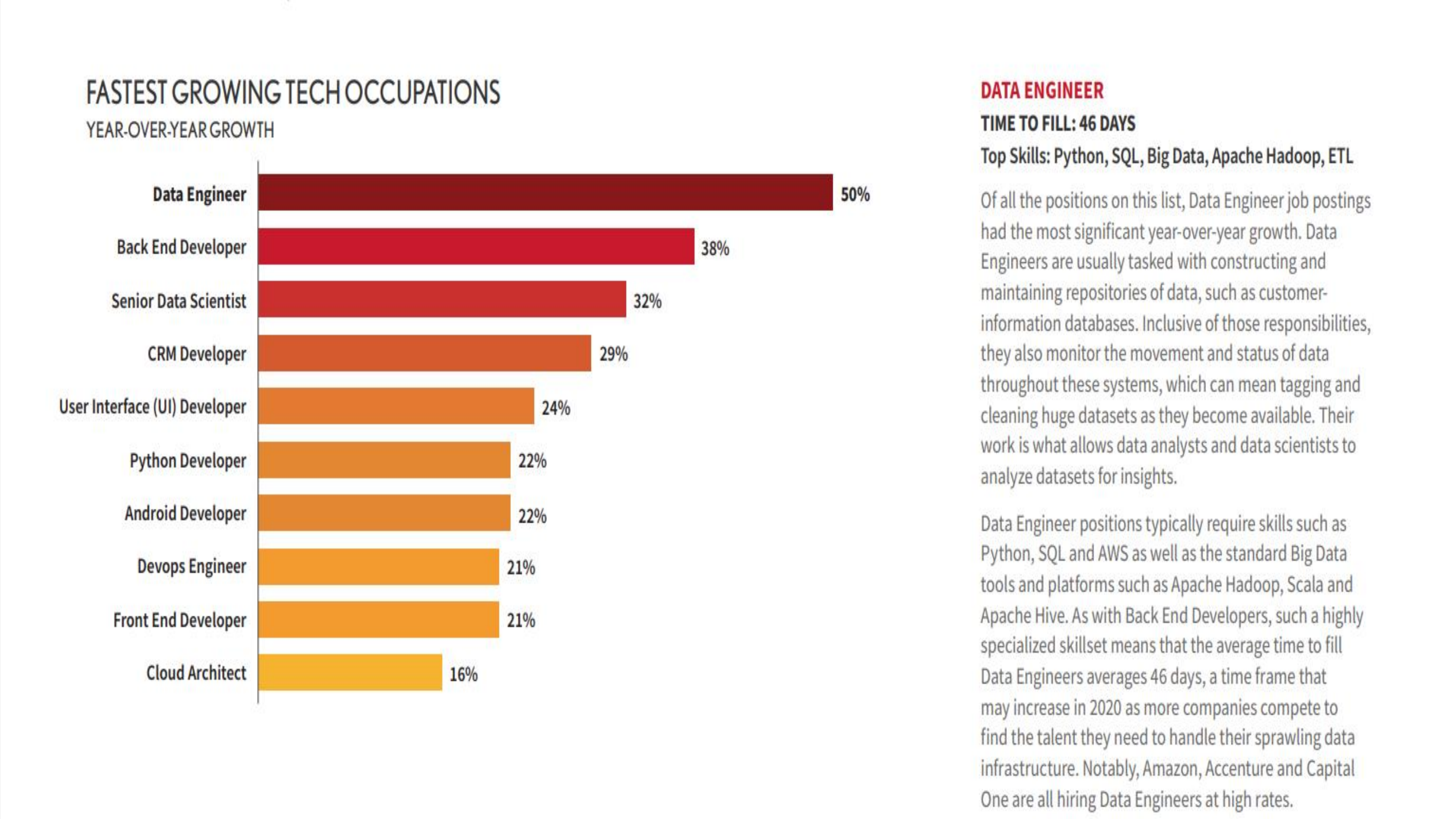
CAREER OPPORTUNITIES

What Happens in 60 Seconds



Eng.Mostafa Nabieh

Data Engineering Growing!



FASTEST GROWING TECH OCCUPATIONS

YEAR-OVER-YEAR GROWTH

DATA ENGINEER

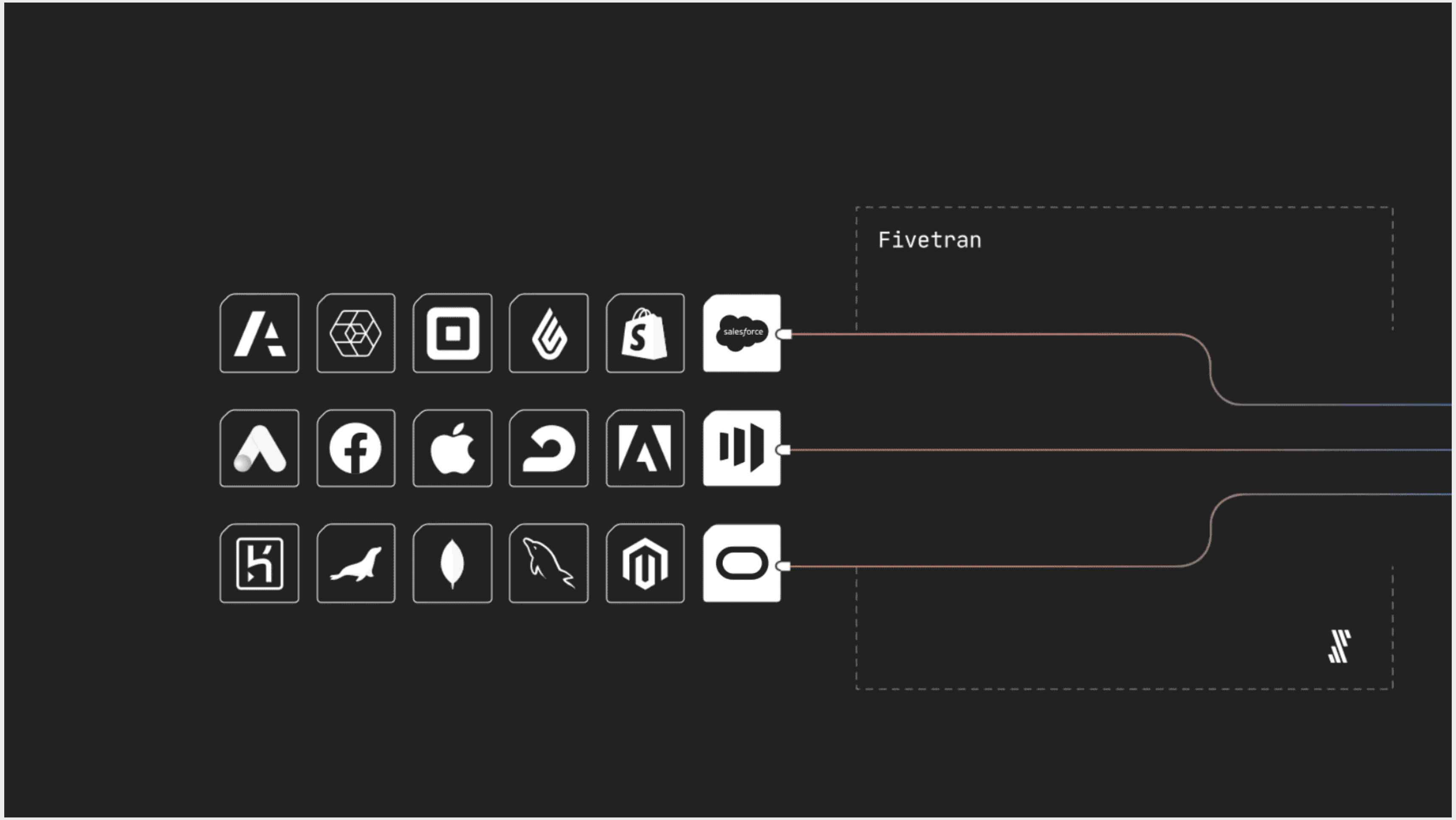
TIME TO FILL: 46 DAYS

Top Skills: Python, SQL, Big Data, Apache Hadoop, ETL

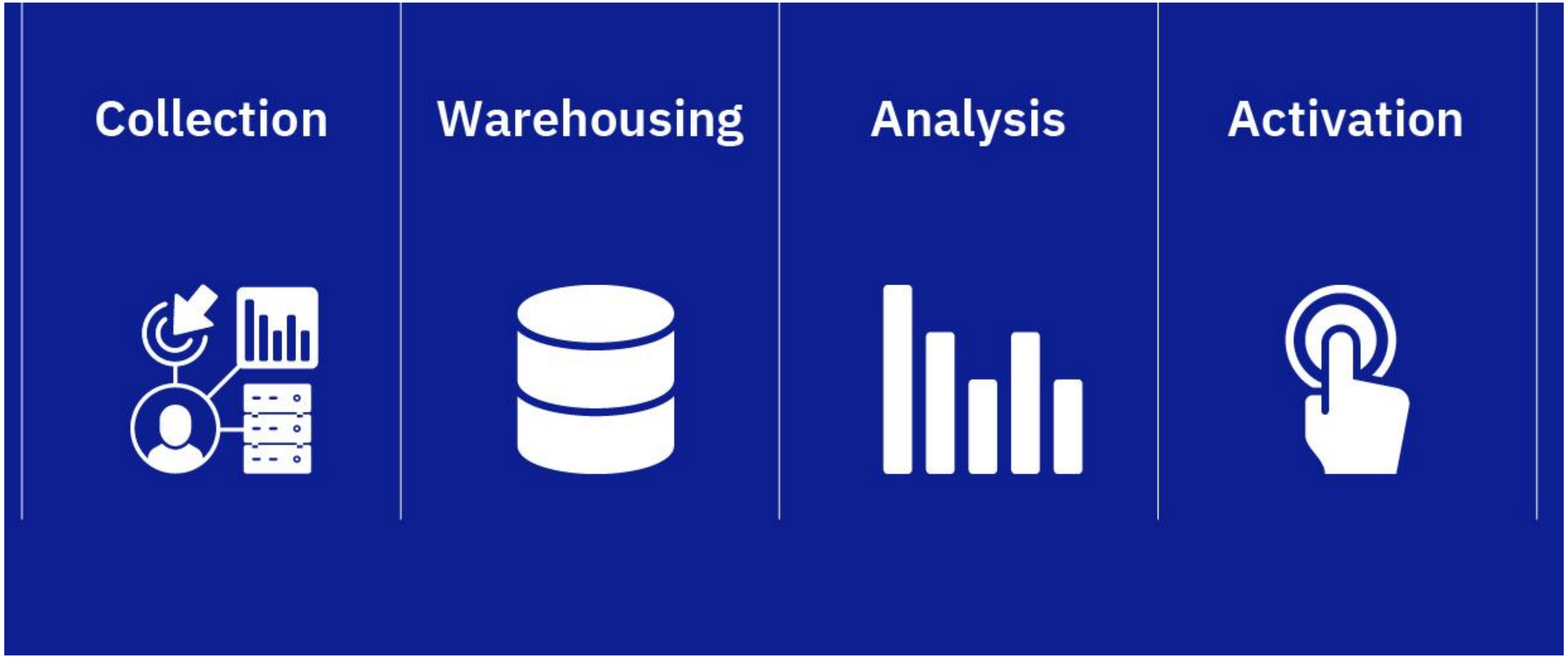
Of all the positions on this list, Data Engineer job postings had the most significant year-over-year growth. Data Engineers are usually tasked with constructing and maintaining repositories of data, such as customer-information databases. Inclusive of those responsibilities, they also monitor the movement and status of data throughout these systems, which can mean tagging and cleaning huge datasets as they become available. Their work is what allows data analysts and data scientists to analyze datasets for insights.

Data Engineer positions typically require skills such as Python, SQL and AWS as well as the standard Big Data tools and platforms such as Apache Hadoop, Scala and Apache Hive. As with Back End Developers, such a highly specialized skillset means that the average time to fill Data Engineers averages 46 days, a time frame that may increase in 2020 as more companies compete to find the talent they need to handle their sprawling data infrastructure. Notably, Amazon, Accenture and Capital One are all hiring Data Engineers at high rates.

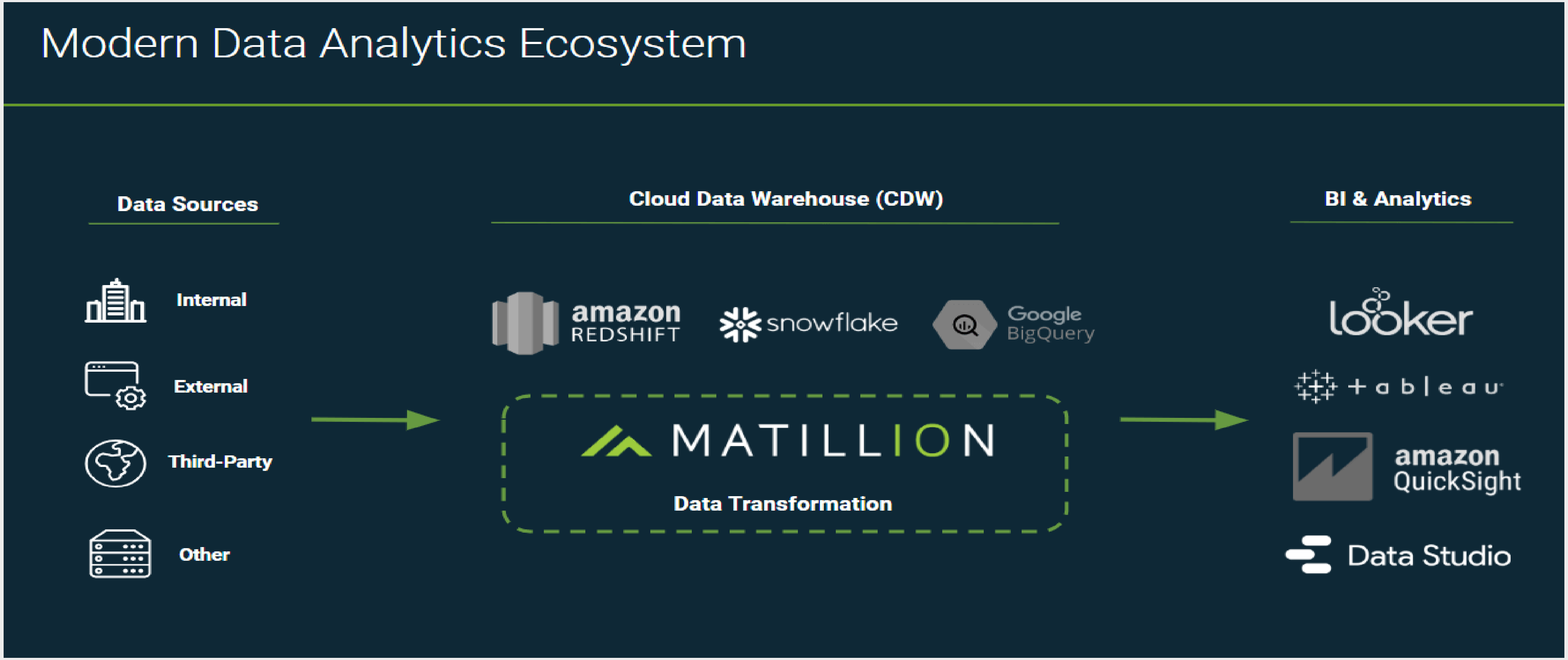
Modern Data Ecosystem



Modern Data Ecosystem



Modern Data Ecosystem



Modern Data Analytics Ecosystem

Data Sources

- Internal
- External
- Third-Party
- Other

Cloud Data Warehouse (CDW)

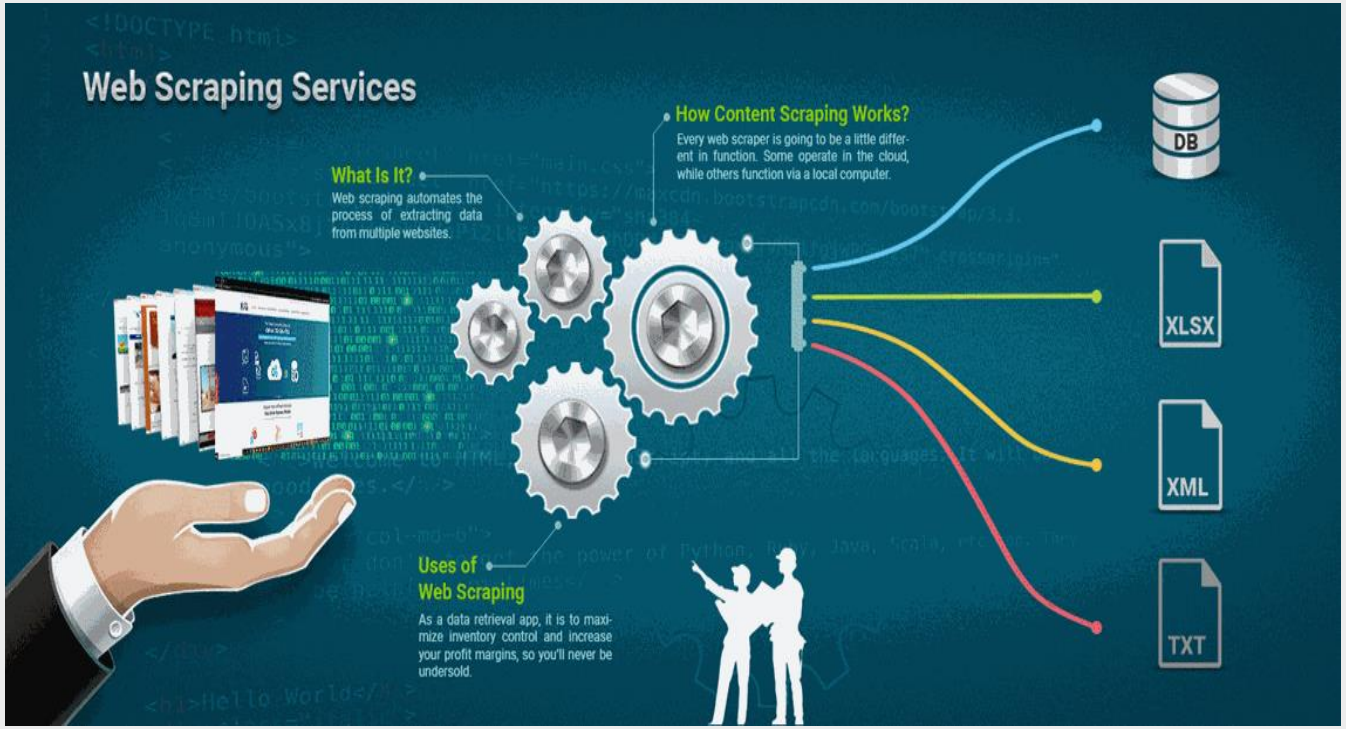
Data Transformation

BI & Analytics

Data Sources

- *Text*
- *Images*
- *Videos*
- *Clickstreams*
- *User conversations*
- *Social media platform*
- *IOT*
- *legacy database*

Web scraping or API



Web Scraping Services

What Is It?

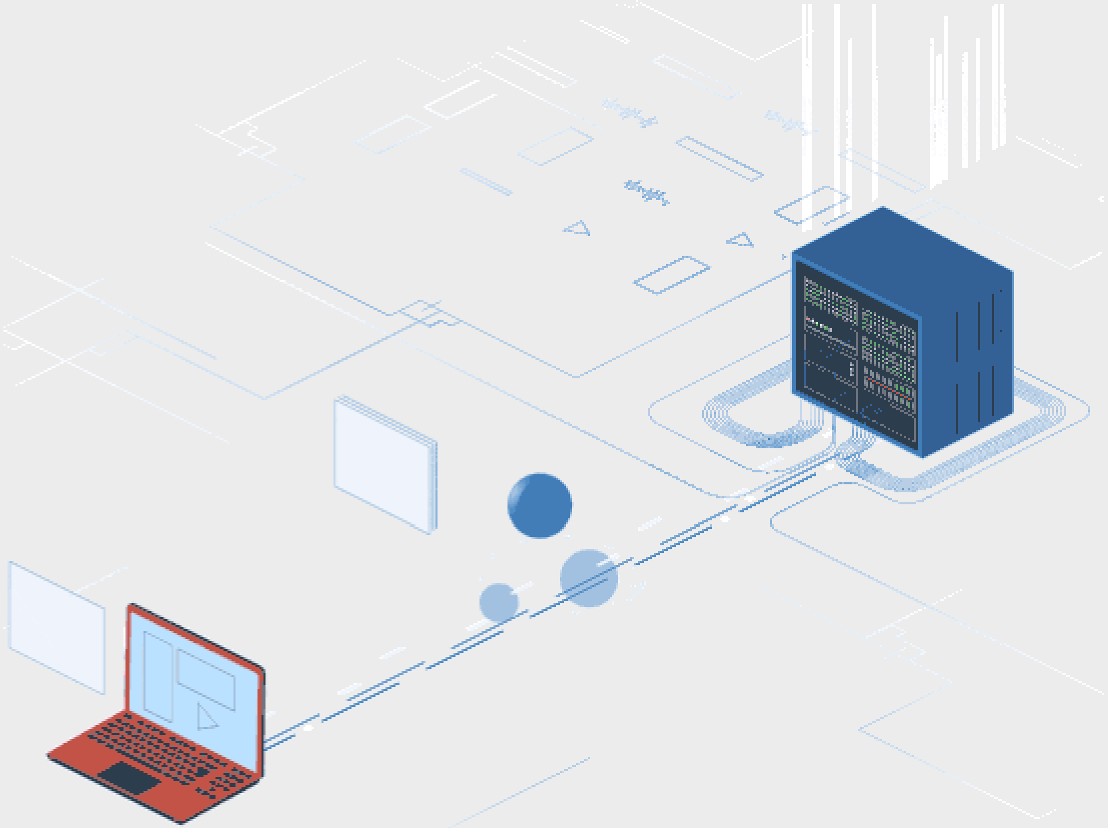
Web scraping automates the process of extracting data from multiple websites.

How Content Scraping Works?

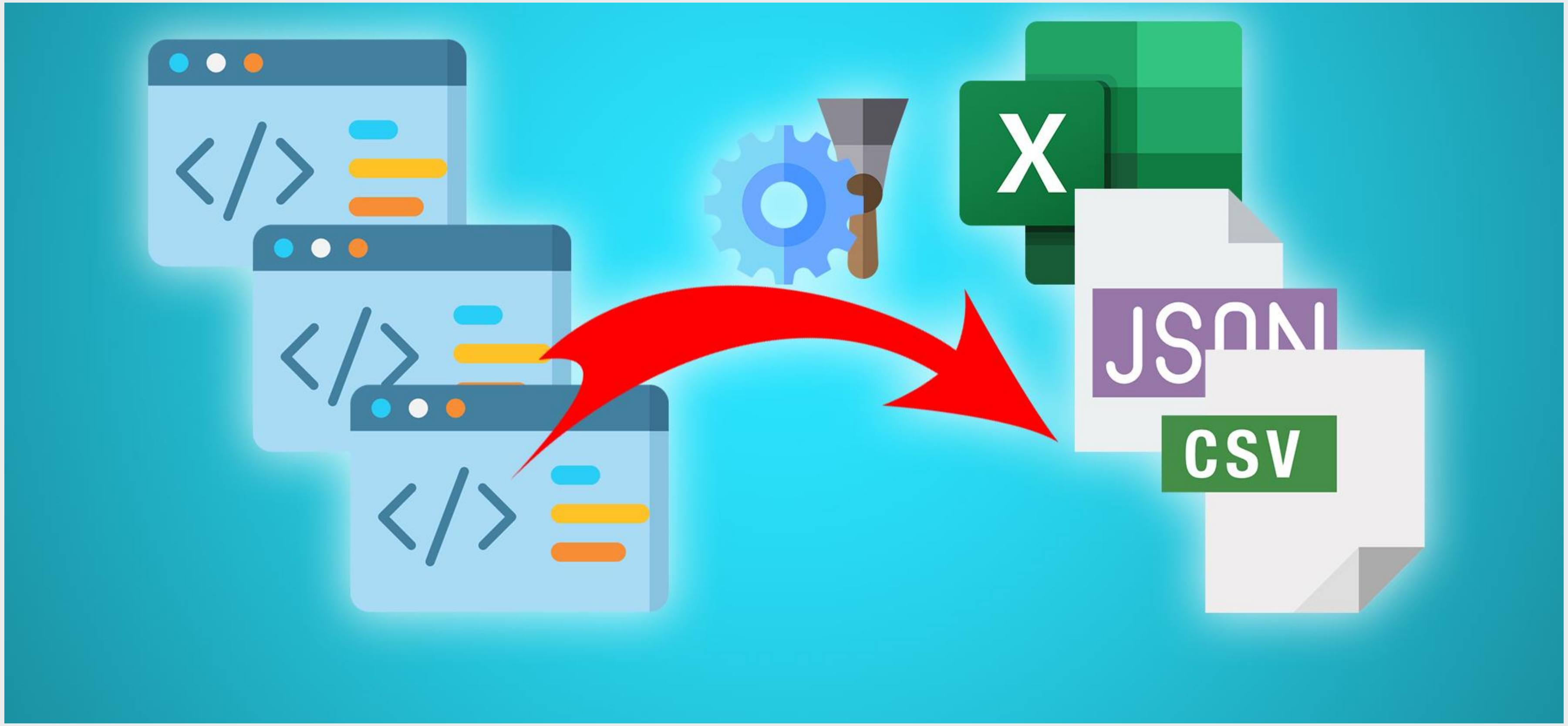
Every web scraper is going to be a little different in function. Some operate in the cloud, while others function via a local computer.

Uses of Web Scraping

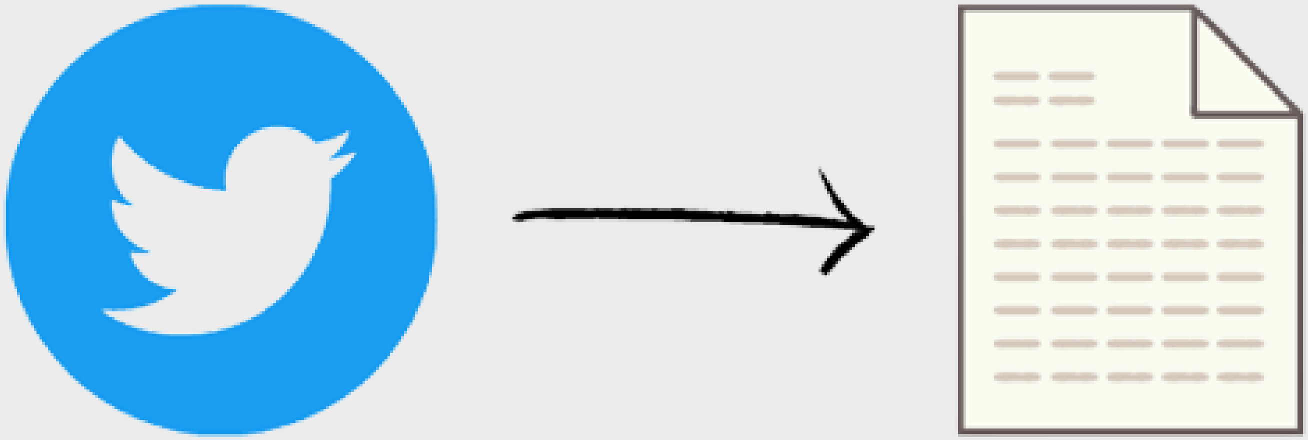
As a data retrieval app, it is to maximize inventory control and increase your profit margins, so you'll never be undersold.



Web Scraping



API



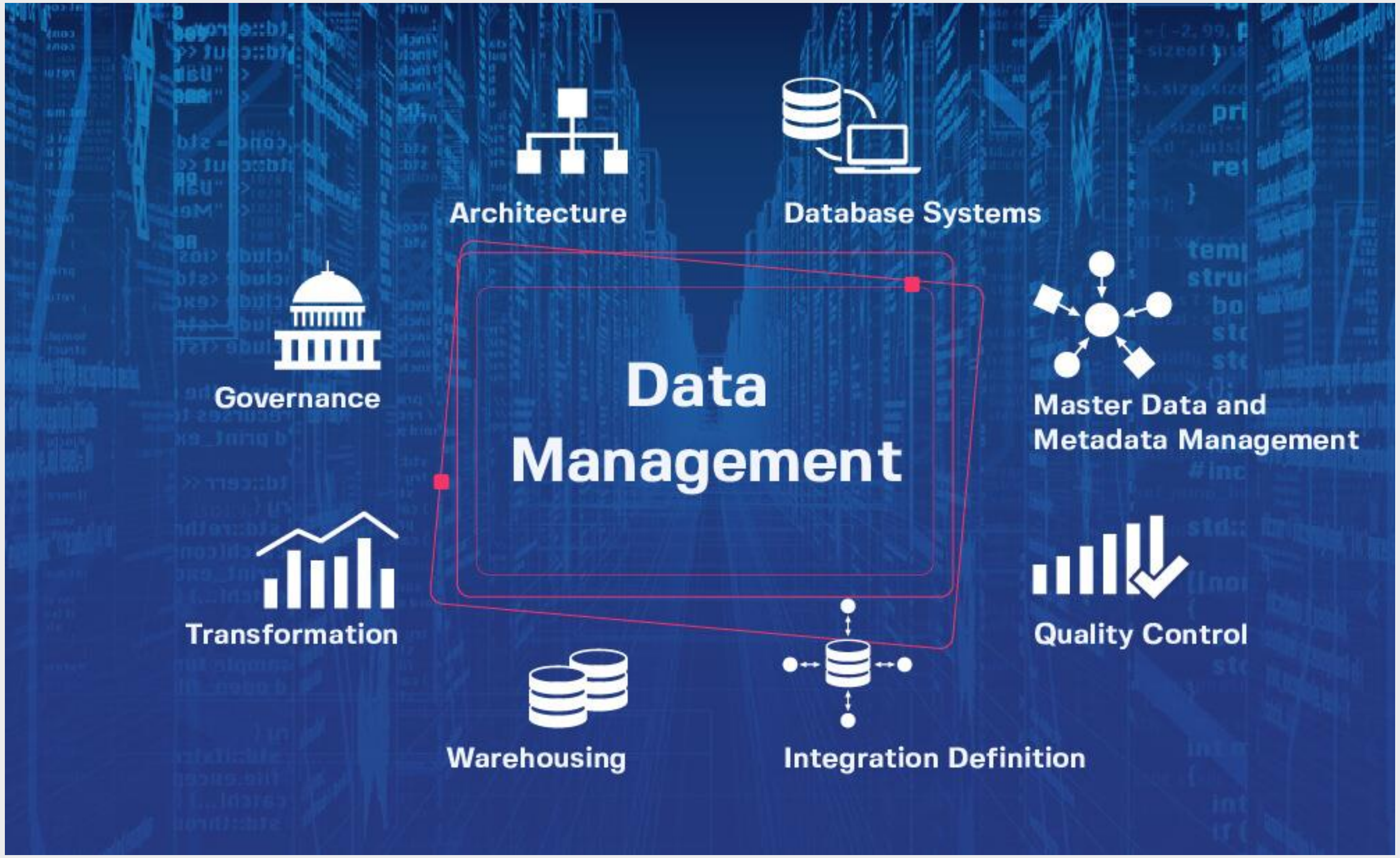
Data From Twitter API

Data Management

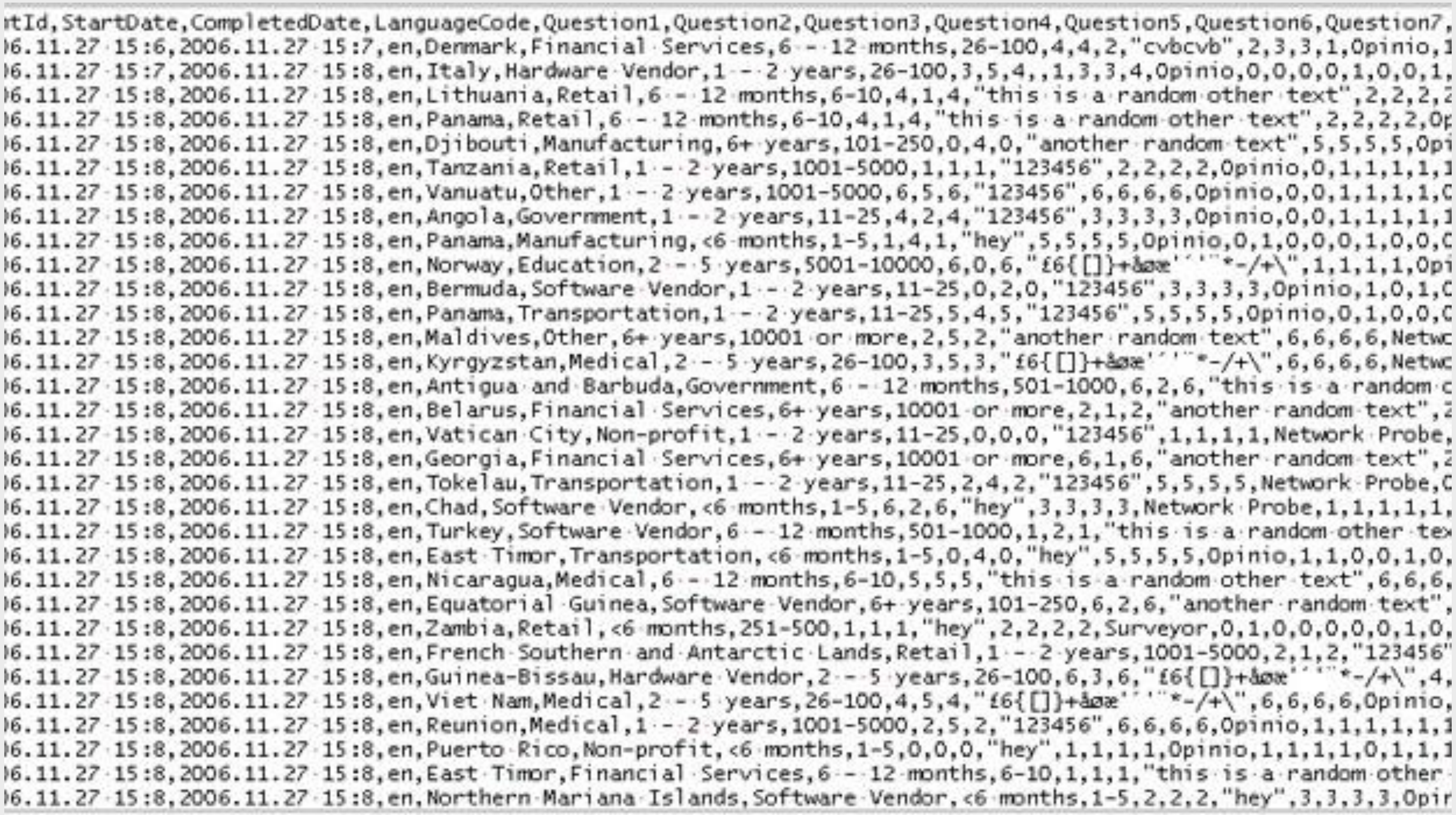
Raw data is in a common place, it needs to get organized, cleaned up, and optimized for access by end-users.

The data will also need to conform to compliances and standards enforced in the organization.

Data Management

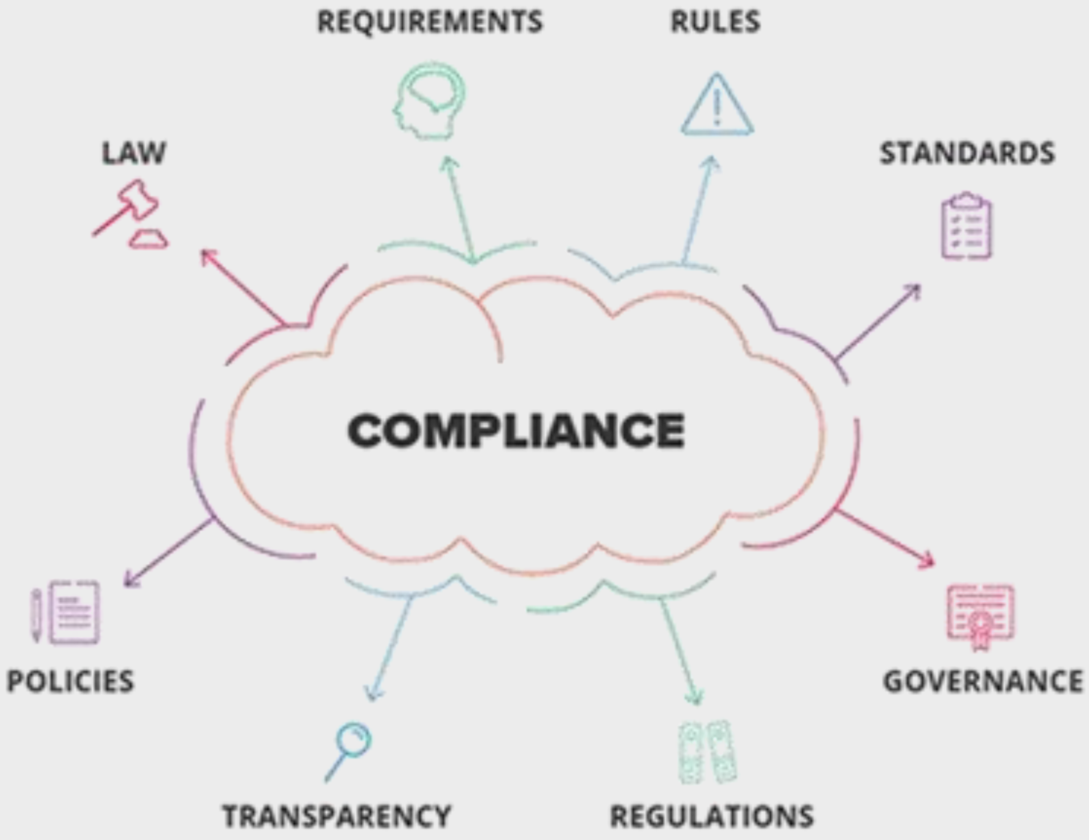


Raw Data

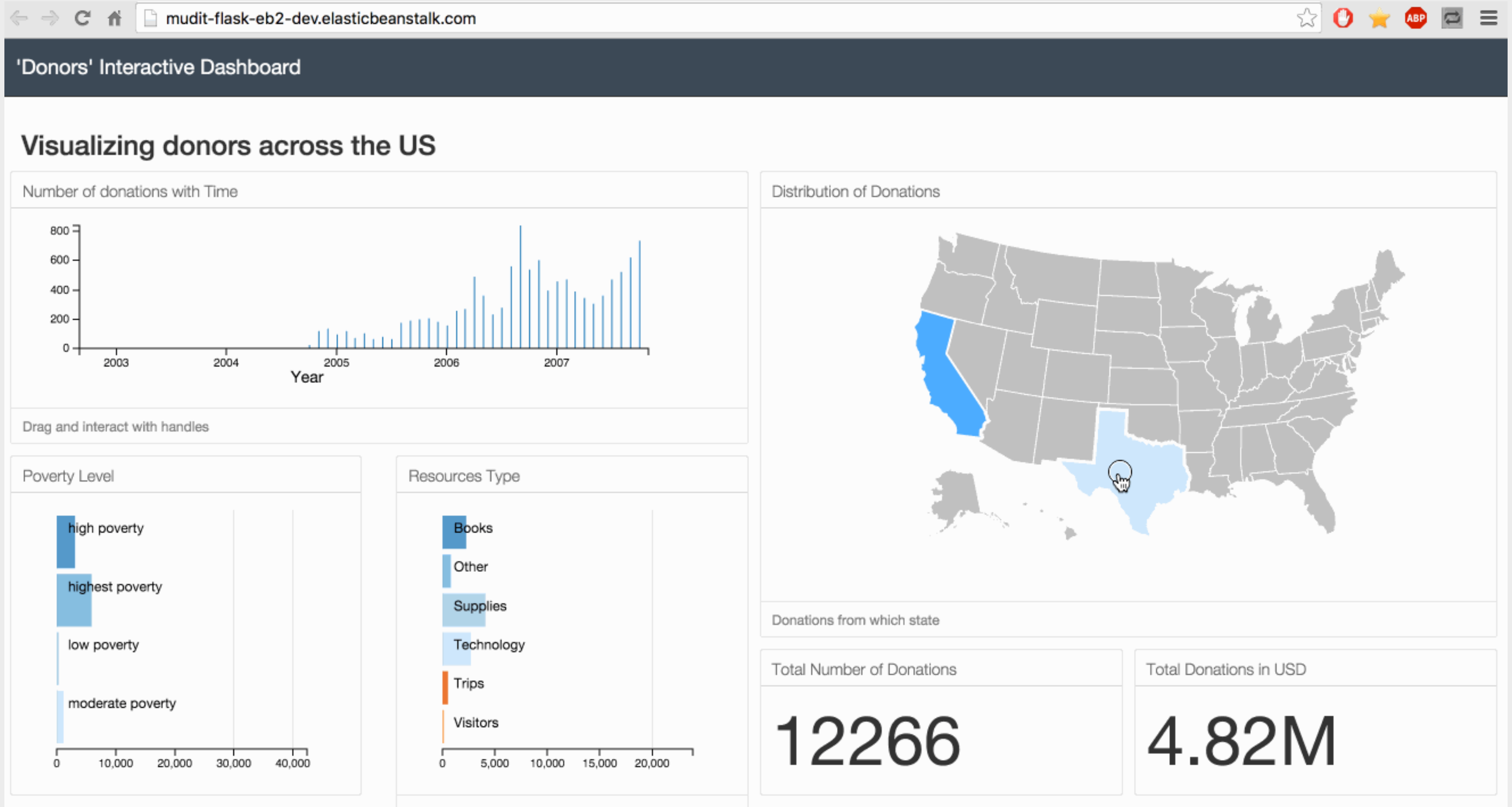


```
tId,StartDate,CompletedDate,LanguageCode,Question1,Question2,Question3,Question4,Question5,Question6,Question7,
6.11.27 15:6,2006.11.27 15:7,en,Denmark,Financial Services,6 - 12 months,26-100,4,4,2,"cvbcvb",2,3,3,1,Opinio,1
6.11.27 15:7,2006.11.27 15:8,en,Italy,Hardware Vendor,1 - 2 years,26-100,3,5,4,,1,3,3,4,Opinio,0,0,0,0,1,0,0,1,
6.11.27 15:8,2006.11.27 15:8,en,Lithuania,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2
6.11.27 15:8,2006.11.27 15:8,en,Panama,Retail,6 - 12 months,6-10,4,1,4,"this is a random other text",2,2,2,2,Op
6.11.27 15:8,2006.11.27 15:8,en,Djibouti,Manufacturing,6+ years,101-250,0,4,0,"another random text",5,5,5,5,Opi
6.11.27 15:8,2006.11.27 15:8,en,Tanzania,Retail,1 - 2 years,1001-5000,1,1,1,"123456",2,2,2,2,Opinio,0,1,1,1,1,1
6.11.27 15:8,2006.11.27 15:8,en,Vanuatu,Other,1 - 2 years,1001-5000,6,5,6,"123456",6,6,6,6,Opinio,0,0,1,1,1,1,C
6.11.27 15:8,2006.11.27 15:8,en,Angola,Government,1 - 2 years,11-25,4,2,4,"123456",3,3,3,3,Opinio,0,0,1,1,1,1,1
6.11.27 15:8,2006.11.27 15:8,en,Panama,Manufacturing,<6 months,1-5,1,4,1,"hey",5,5,5,5,Opinio,0,1,0,0,0,1,0,0,C
6.11.27 15:8,2006.11.27 15:8,en,Norway,Education,2 - 5 years,5001-10000,6,0,6,"£6{[]}+åøæ'´'¨*-/+\",1,1,1,1,Opi
6.11.27 15:8,2006.11.27 15:8,en,Bermuda,Software Vendor,1 - 2 years,11-25,0,2,0,"123456",3,3,3,3,Opinio,1,0,1,C
6.11.27 15:8,2006.11.27 15:8,en,Panama,Transportation,1 - 2 years,11-25,5,4,5,"123456",5,5,5,5,Opinio,0,1,0,0,C
6.11.27 15:8,2006.11.27 15:8,en,Maldives,Other,6+ years,10001 or more,2,5,2,"another random text",6,6,6,6,Netwo
6.11.27 15:8,2006.11.27 15:8,en,Kyrgyzstan,Medical,2 - 5 years,26-100,3,5,3,"£6{[]}+åøæ'´'¨*-/+\",6,6,6,6,Netwo
6.11.27 15:8,2006.11.27 15:8,en,Antigua and Barbuda,Government,6 - 12 months,501-1000,6,2,6,"this is a random c
6.11.27 15:8,2006.11.27 15:8,en,Belarus,Financial Services,6+ years,10001 or more,2,1,2,"another random text",2
6.11.27 15:8,2006.11.27 15:8,en,Vatican City,Non-profit,1 - 2 years,11-25,0,0,0,"123456",1,1,1,1,Network Probe,
6.11.27 15:8,2006.11.27 15:8,en,Georgia,Financial Services,6+ years,10001 or more,6,1,6,"another random text",2
6.11.27 15:8,2006.11.27 15:8,en,Tokelau,Transportation,1 - 2 years,11-25,2,4,2,"123456",5,5,5,5,Network Probe,C
6.11.27 15:8,2006.11.27 15:8,en,Chad,Software Vendor,<6 months,1-5,6,2,6,"hey",3,3,3,3,Network Probe,1,1,1,1,1,
6.11.27 15:8,2006.11.27 15:8,en,Turkey,Software Vendor,6 - 12 months,501-1000,1,2,1,"this is a random other tex
6.11.27 15:8,2006.11.27 15:8,en,East Timor,Transportation,<6 months,1-5,0,4,0,"hey",5,5,5,5,Opinio,1,1,0,0,1,0,
6.11.27 15:8,2006.11.27 15:8,en,Nicaragua,Medical,6 - 12 months,6-10,5,5,5,"this is a random other text",6,6,6,
6.11.27 15:8,2006.11.27 15:8,en,Equatorial Guinea,Software Vendor,6+ years,101-250,6,2,6,"another random text",
6.11.27 15:8,2006.11.27 15:8,en,Zambia,Retail,<6 months,251-500,1,1,1,"hey",2,2,2,2,Surveyor,0,1,0,0,0,0,0,1,0,
6.11.27 15:8,2006.11.27 15:8,en,French Southern and Antarctic Lands,Retail,1 - 2 years,1001-5000,2,1,2,"123456"
6.11.27 15:8,2006.11.27 15:8,en,Guinea-Bissau,Hardware Vendor,2 - 5 years,26-100,6,3,6,"£6{[]}+åøæ'´'¨*-/+\",4,
6.11.27 15:8,2006.11.27 15:8,en,Viet Nam,Medical,2 - 5 years,26-100,4,5,4,"£6{[]}+åøæ'´'¨*-/+\",6,6,6,6,Opinio,
6.11.27 15:8,2006.11.27 15:8,en,Reunion,Medical,1 - 2 years,1001-5000,2,5,2,"123456",6,6,6,6,Opinio,1,1,1,1,1,1
6.11.27 15:8,2006.11.27 15:8,en,Puerto Rico,Non-profit,<6 months,1-5,0,0,0,"hey",1,1,1,1,Opinio,1,1,1,1,0,1,1,1
6.11.27 15:8,2006.11.27 15:8,en,East Timor,Financial Services,6 - 12 months,6-10,1,1,1,"this is a random other
6.11.27 15:8,2006.11.27 15:8,en,Northern Mariana Islands,Software Vendor,<6 months,1-5,2,2,2,"hey",3,3,3,3,Opir
```

Data Compliance



END-USER



'Donors' Interactive Dashboard

Visualizing donors across the US

Number of donations with Time

Drag and interact with handles

Distribution of Donations

Donations from which state

Poverty Level

Resources Type

Total Number of Donations

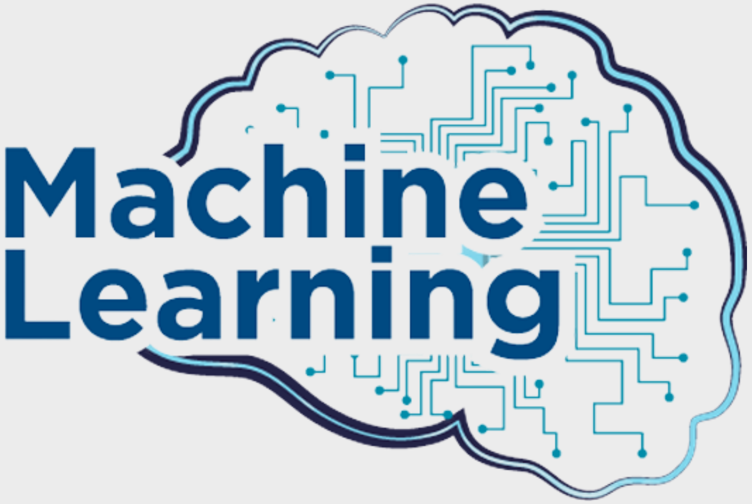
12266

Total Donations in USD

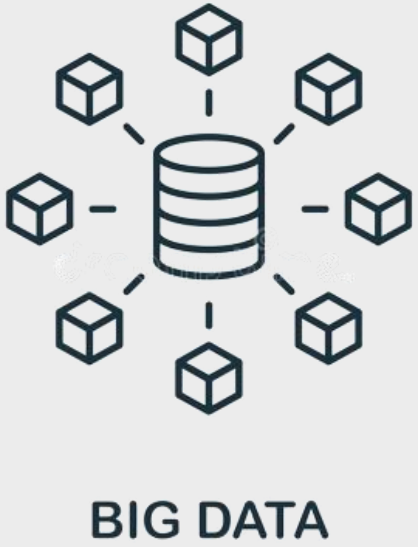
4.82M

EMERGING TECHNOLOGIES SHAPING THE MODERN DATA ECOSYSTEM

Cloud Technologies



Machine Learning



Big Data

EMERGING TECHNOLOGIES SHAPING THE MODERN DATA ECOSYSTEM

Every enterprise today has access to limitless storage, high-performance computing, open-source technologies, machine learning technologies, and the latest tools and libraries.

Data Scientists are creating predictive models by training machine learning algorithms on past data.

Big Data is paving the way for new tools and techniques and also new knowledge and insights.

Data Professional Team

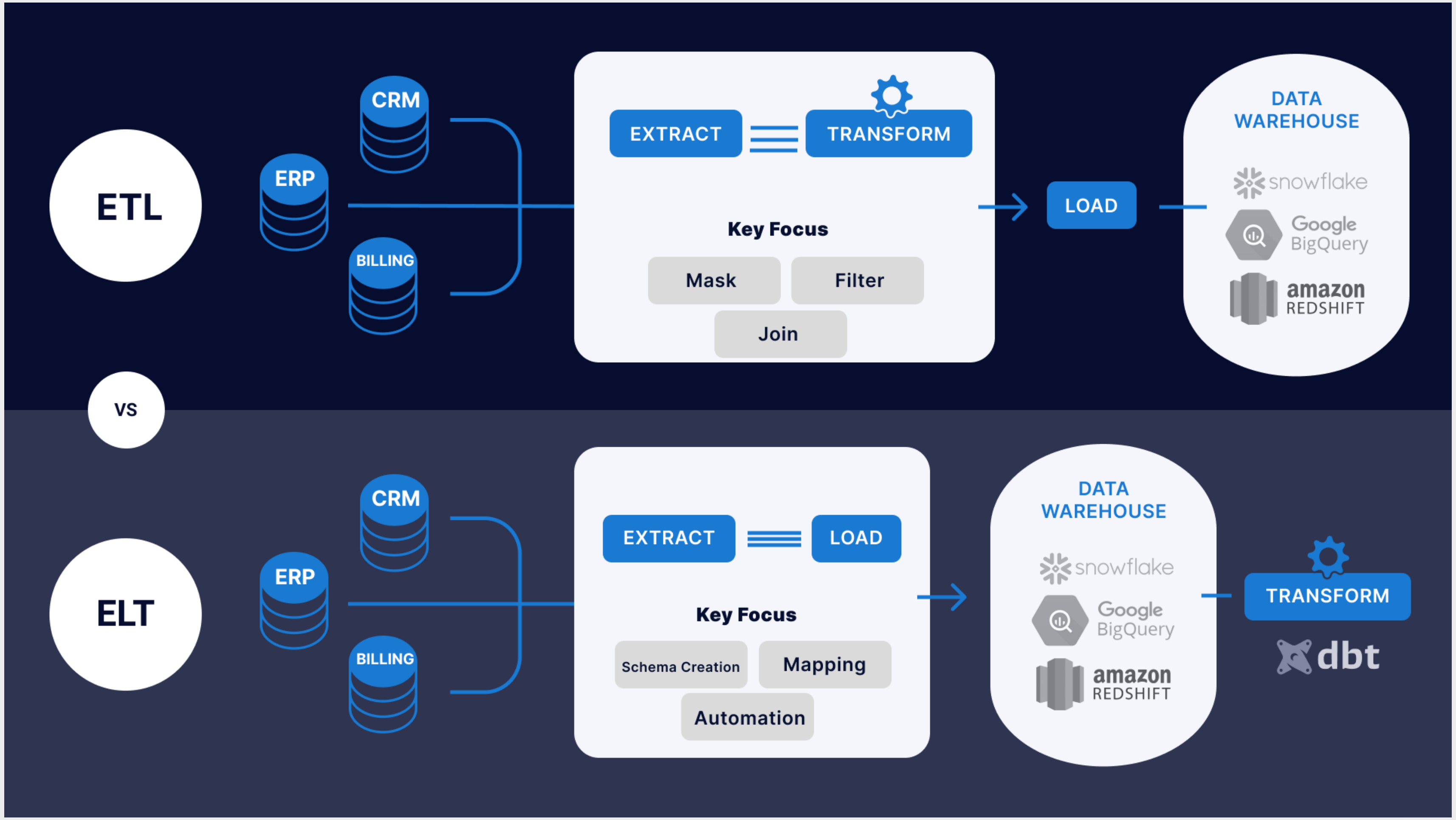
DATA ENGINEERS

DATA ANALYSTS

DATA SCIENTISTS

BUSINESS INTELLIGENCE

DATA ENGINEERS WORK WITHIN THE DATA ECOSYSTEM TO:



Eng.Mostafa Nabieh

Data Engineers Skills

- *Good knowledge of programming*
- *Sound knowledge of systems and technology architectures*
- *In-depth understanding of relational databases and non-relational data stores*

Data Analysts

- *Inspect and clean data for deriving insights*
- *Identify correlations, find patterns, and apply statistical methods to analyze and mine data*
- *Visualize data to interpret and present the findings of data analysis*

Data Analysts Skills

- *Good knowledge of spreadsheets, writing queries, and using statistical tools to create charts and dashboards*
- *Programming skills*
- *Strong analytical and story-telling skills*

Data Scientists

Analyze data for actionable insights

Create predictive models using Machine Learning and Deep Learning

Data Scientists Skills

KNOWLEDGE OF MATHEMATICS AND STATISTICS

UNDERSTANDING OF PROGRAMMING LANGUAGES, DATABASES, AND BUILDING DATA MODELS

DOMAIN KNOWLEDGE

Business Intelligence

- *Business Analysts leverage the work of Data Analysts and Data Scientists to look at possible implications for their business and the actions they need to take or recommend.*

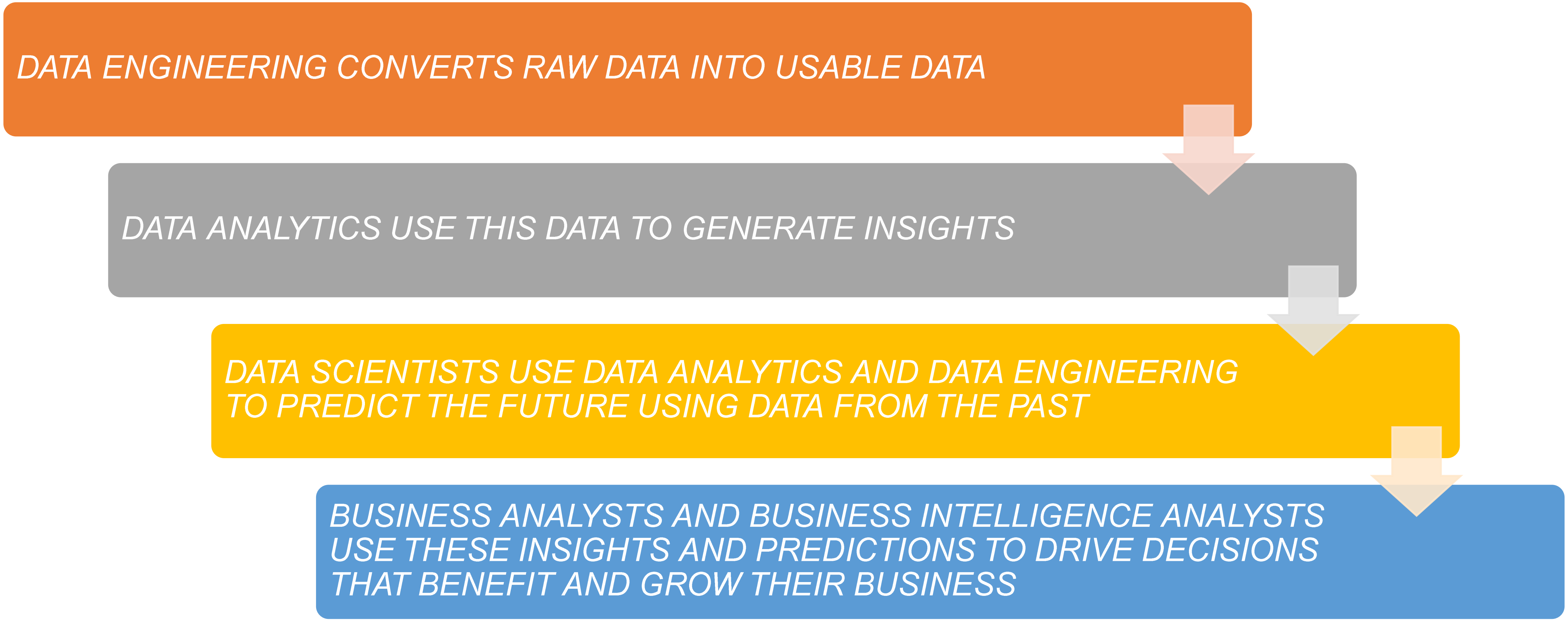
BI Analysts

FOCUS ON MARKET FORCES AND EXTERNAL INFLUENCES THAT SHAPE THEIR BUSINESS

ORGANIZE AND MONITOR DATA ON DIFFERENT BUSINESS FUNCTIONS

EXPLORE DATA TO EXTRACT INSIGHTS AND ACTIONABLES THAT IMPROVE BUSINESS PERFORMANCE

To Summarize



DATA ENGINEERING CONVERTS RAW DATA INTO USABLE DATA

DATA ANALYTICS USE THIS DATA TO GENERATE INSIGHTS

DATA SCIENTISTS USE DATA ANALYTICS AND DATA ENGINEERING TO PREDICT THE FUTURE USING DATA FROM THE PAST

BUSINESS ANALYSTS AND BUSINESS INTELLIGENCE ANALYSTS USE THESE INSIGHTS AND PREDICTIONS TO DRIVE DECISIONS THAT BENEFIT AND GROW THEIR BUSINESS

Languages

Query languages
SQL for relational databases and SQL-like query languages for NoSQL databases

Programming languages
Python, R, Java

Shell and Scripting languages
Unix/Linux Shell and PowerShell

Databases and Data Warehouses

RDBMS
MySQL, Oracle Database, PostgreSQL

NoSQL
Redis, MongoDB, Cassandra, Neo4J

Data Warehouses
Oracle Exadata, Amazon RedShift

Operating Systems

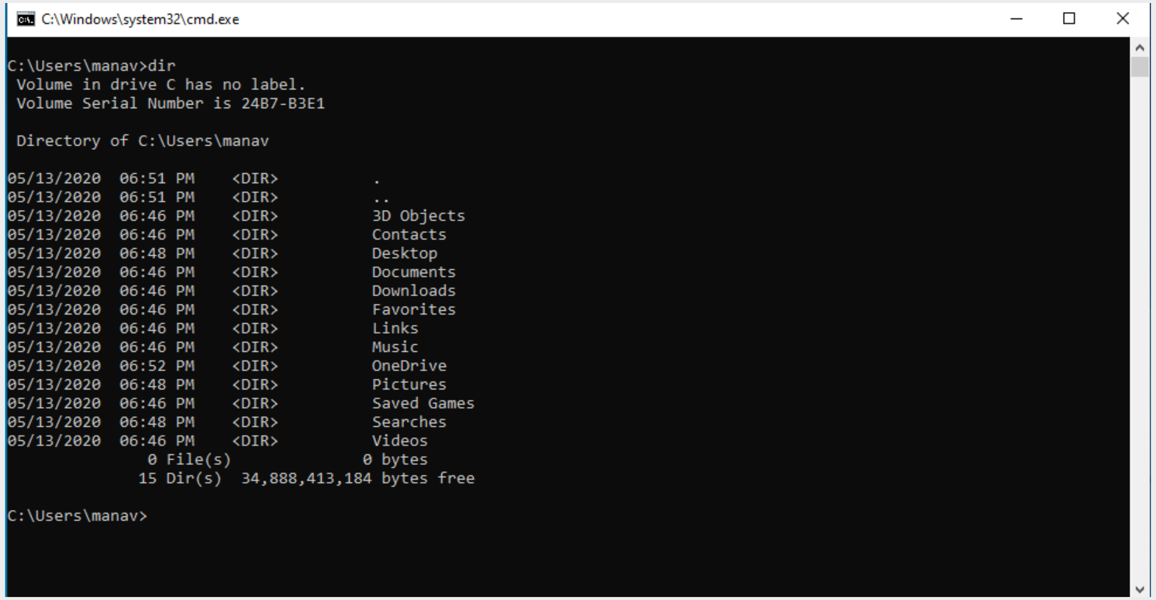
UNIX

LINUX

WINDOWS ADMINISTRATIVE TOOLS

SYSTEM UTILITIES & COMMANDS

Linux vs Windows

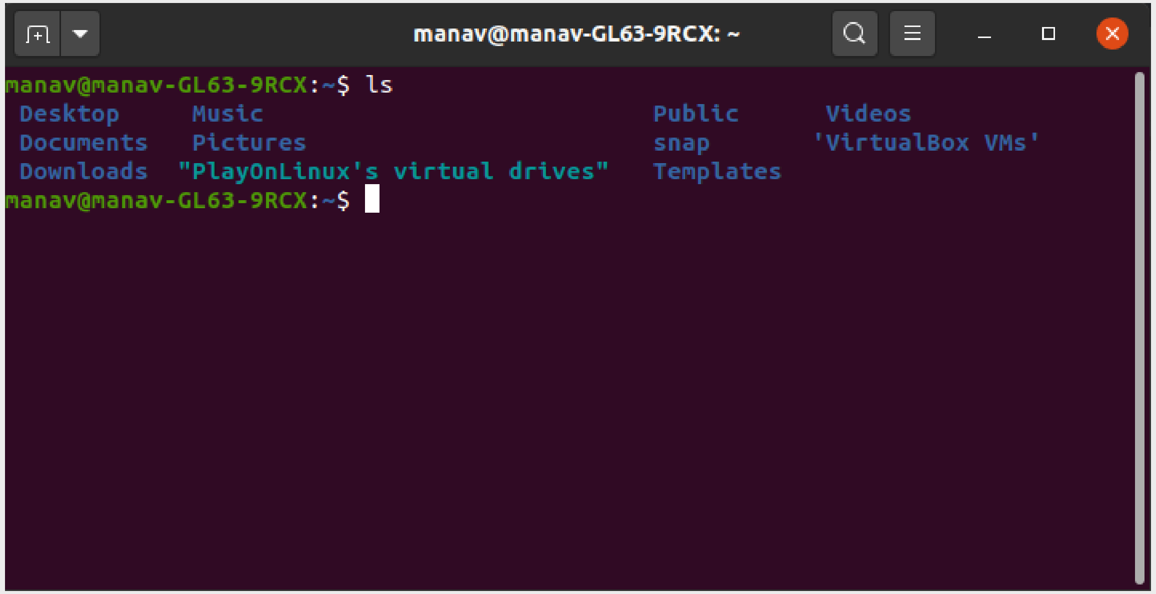


```
C:\Users\manav>dir
 Volume in drive C has no label.
 Volume Serial Number is 24B7-B3E1

 Directory of C:\Users\manav

05/13/2020  06:51 PM    <DIR>          .
05/13/2020  06:51 PM    <DIR>          ..
05/13/2020  06:46 PM    <DIR>          3D Objects
05/13/2020  06:46 PM    <DIR>          Contacts
05/13/2020  06:48 PM    <DIR>          Desktop
05/13/2020  06:46 PM    <DIR>          Documents
05/13/2020  06:46 PM    <DIR>          Downloads
05/13/2020  06:46 PM    <DIR>          Favorites
05/13/2020  06:46 PM    <DIR>          Links
05/13/2020  06:46 PM    <DIR>          Music
05/13/2020  06:52 PM    <DIR>          OneDrive
05/13/2020  06:48 PM    <DIR>          Pictures
05/13/2020  06:46 PM    <DIR>          Saved Games
05/13/2020  06:48 PM    <DIR>          Searches
05/13/2020  06:46 PM    <DIR>          Videos
               0 File(s)              0 bytes
              15 Dir(s)  34,888,413,184 bytes free

C:\Users\manav>
```



```
manav@manav-GL63-9RCX:~$ ls
 Desktop     Music                            Public     Videos
 Documents   Pictures                         snap      'VirtualBox VMs'
 Downloads  "PlayOnLinux's virtual drives"    Templates
manav@manav-GL63-9RCX:~$ 
```

Big Data Processing Tools

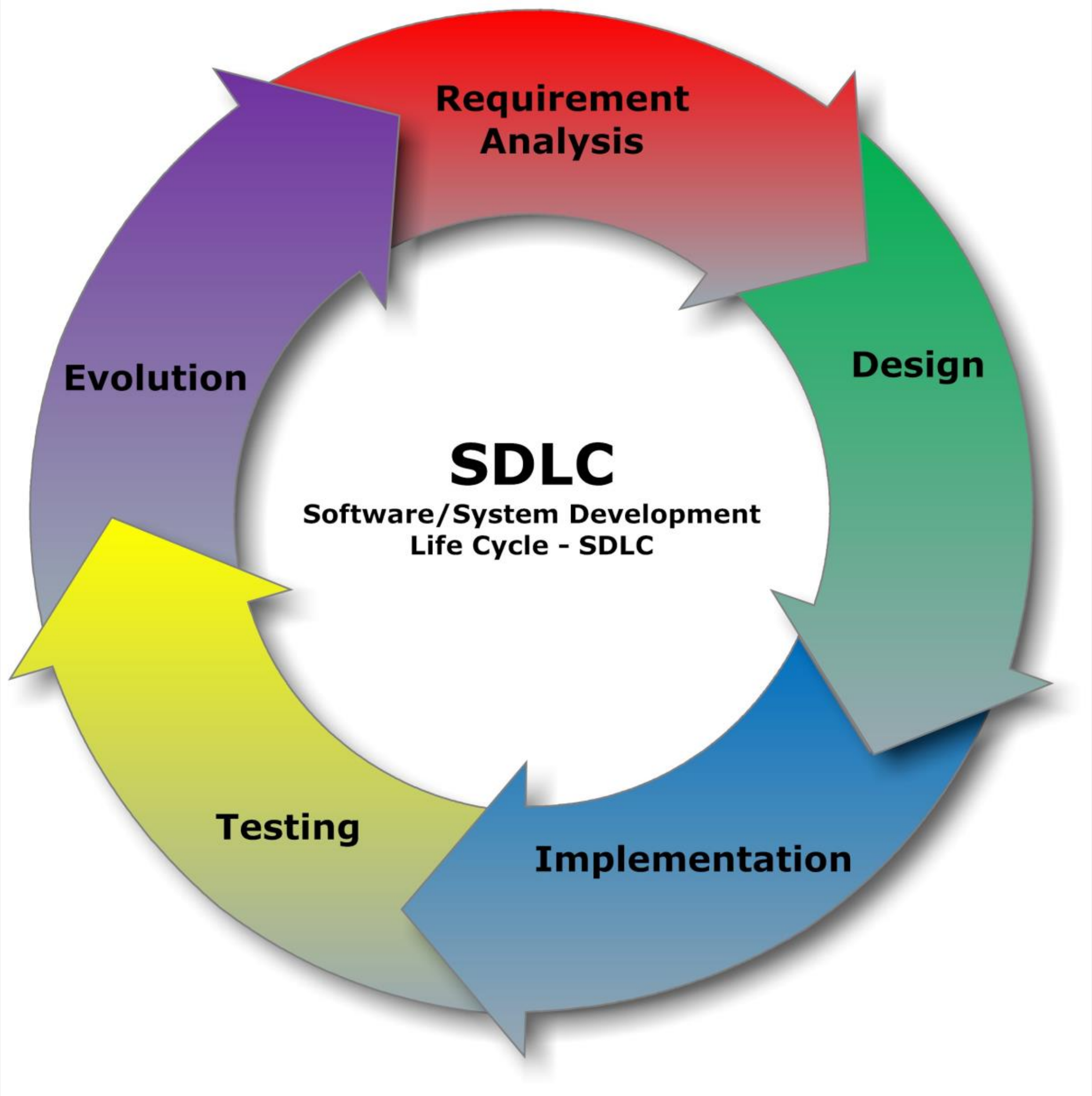
Data Pipelines

Functional Skills

CONVERT BUSINESS REQUIREMENTS INTO TECHNICAL SPECIFICATIONS

WORK WITH THE COMPLETE SOFTWARE DEVELOPMENT LIFECYCLE IDEATION -> ARCHITECTURE -> DESIGN -> PROTOTYPING -> TESTING -> DEPLOYMENT -> MONITORING

SDLC


Requirement Analysis
Design
Implementation
Testing
Evolution
SDLC
Software/System Development Life Cycle - SDLC

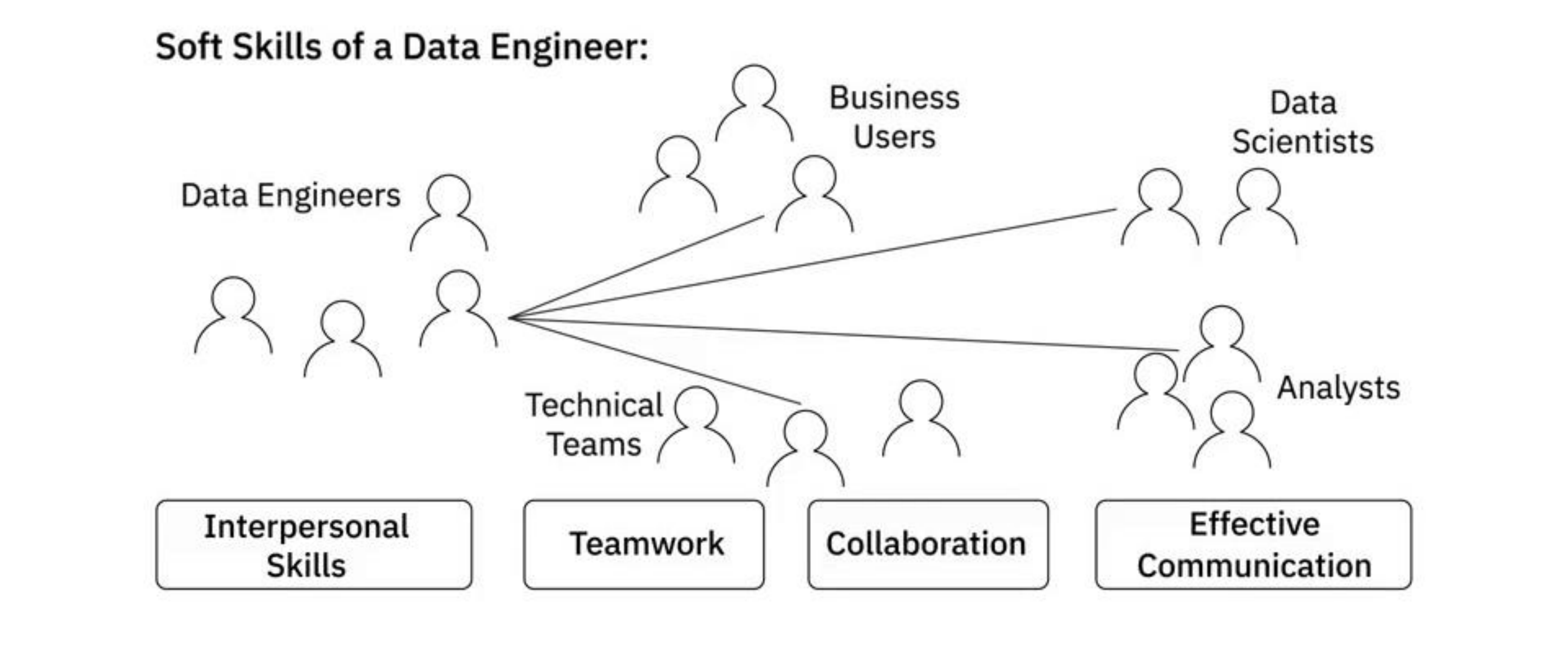
Cloud

Functional Skills of a Data Engineer:

UNDERSTAND DATA'S POTENTIAL APPLICATION IN BUSINESS

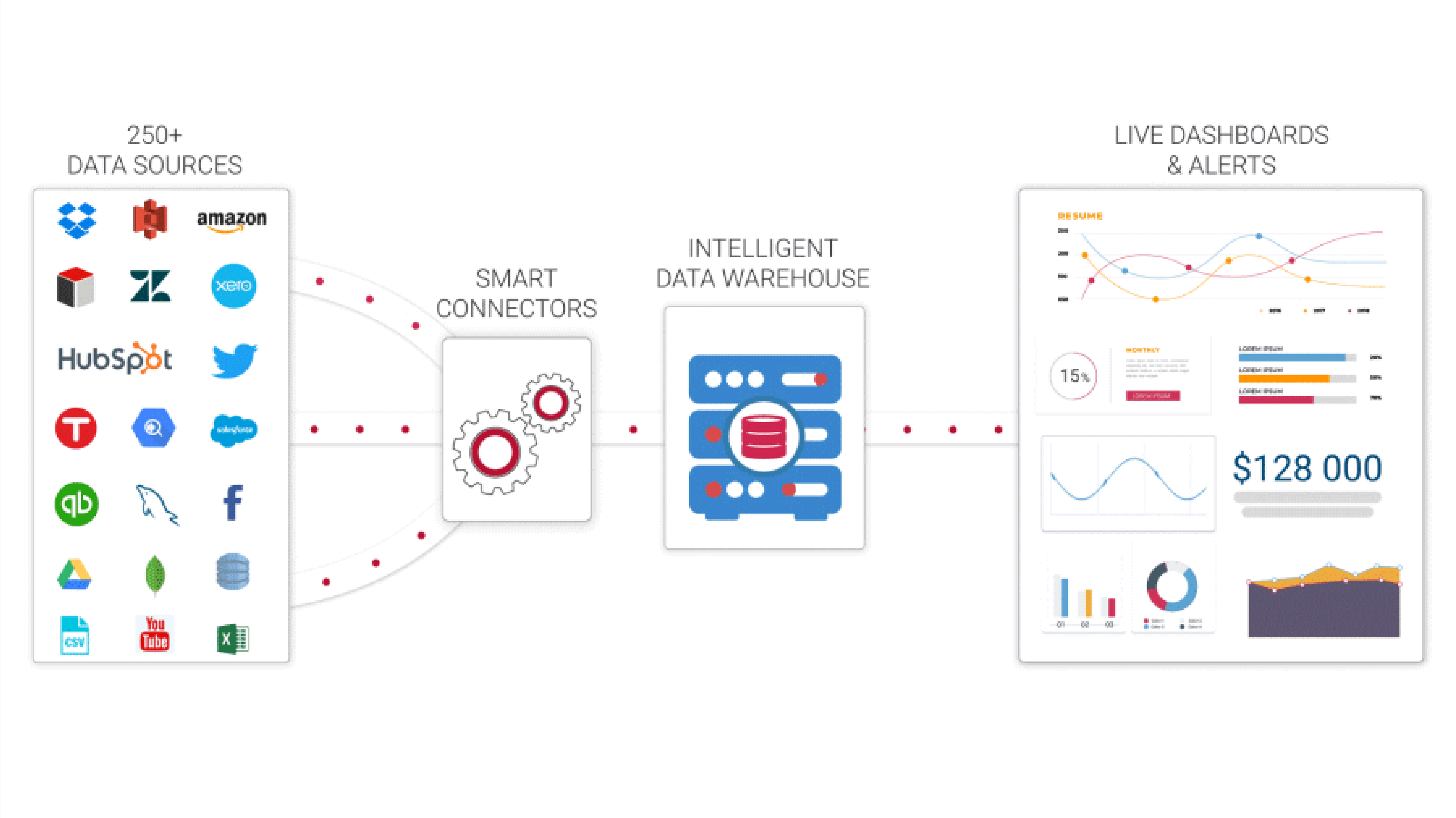
UNDERSTAND RISKS OF POOR DATA MANAGEMENT DATA QUALITY I DATA PRIVACY I SECURITY I COMPLIANCE

Soft Skills


Soft Skills of a Data Engineer:
Business Users
Data Scientists
Data Engineers
Technical Teams
Analysts
Interpersonal Skills
Teamwork
Collaboration
Effective Communication

Pipelines

Pipelines


250+
DATA SOURCES
SMART
CONNECTORS
INTELLIGENT
DATA WAREHOUSE
LIVE DASHBOARDS
& ALERTS
amazon
HubSpot
RESUME
15%
$128 000

Thank you