## Scenario

For this project, you will assume the role of data engineer working for an international financial analysis company. Your company tracks stock prices, commodities, forex rates, inflation rates. Your job is to extract financial data from various sources like websites, APIs and files provided by various financial analysis firms. After you collect the data, you extract the data of interest to your company and transform it based on the requirements given to you. Once the transformation is complete you load that data into a database.

## Project Tasks

In this project you will:

* Collect data using APIs
* Collect data using webscraping.
* Download files to process.
* Read csv, xml and json file types.
* Extract data from the above file types.
* Transform data.
* Use the built in logging module.
* Save the transformed data in a ready-to-load format which data engineers can use to load the data.

1. From the lab you were asked to print out the output of the following line, and remember it as it will be a quiz question:

* ‘List of largest banks -‘
* ‘List of largest banks ‘
* ‘<!DOCTYPE html>\n<html cla’

Q2. Using the contents and beautiful soup load the data from the By market capitalization table into a pandas dataframe. The dataframe should have the country Name and Market Cap (US$ Billion) as column names. Display the first five rows using head.

Q3. True or False If the website changes your output will remain the same

* False
* True

In this digital world, everyone leaves a trace.

From our travel habits to our workouts and entertainment, the increasing number of internet

connected devices that we interact with on a daily basis record vast amounts of data

about us.

There’s even a name for it: Big Data.

Ernst and Young offers the following definition: “Big Data refers to the dynamic, large and

disparate volumes of data being created by people, tools, and machines.

It requires new, innovative, and scalable technology to collect, host, and analytically

process the vast amount of data gathered in order to derive real-time business insights

that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder

value.”

There is no one definition of Big Data, but there are certain elements that are common

across the different definitions, such as velocity, volume, variety, veracity, and value.

These are the V's of Big Data.

Velocity is the speed at which data accumulates.

Data is being generated extremely fast, in a process that never stops.

Near or real-time streaming, local, and cloud-based technologies can process information very

quickly.

Volume is the scale of the data, or the increase in the amount of data stored.

Drivers of volume are the increase in data sources, higher resolution sensors, and scalable

infrastructure.

Variety is the diversity of the data.

Structured data fits neatly into rows and columns, in relational databases while unstructured

data is not organized in a pre-defined way, like Tweets, blog posts, pictures, numbers,

and video.

Variety also reflects that data comes from different sources, machines, people, and processes,

both internal and external to organizations.

Drivers are mobile technologies, social media, wearable technologies, geo technologies, video,

and many, many more.

Veracity is the quality and origin of data, and its conformity to facts and accuracy.

Attributes include consistency, completeness, integrity, and ambiguity.

Drivers include cost and the need for traceability.

With the large amount of data available, the debate rages on about the accuracy of data

in the digital era.

Is the information real, or is it false?

Value is our ability and need to turn data into value.

Value isn't just profit.

It may have medical or social benefits, as well as customer, employee, or personal satisfaction.

The main reason that people invest time to understand Big Data is to derive value from

it.

Let's look at some examples of the V's in action.

Velocity: Every 60 seconds, hours of footage are uploaded to YouTube which is generating

data.

Think about how quickly data accumulates over hours, days, and years.

Volume: The world population is approximately seven billion people and the vast majority

are now using digital devices; mobile phones, desktop and laptop computers, wearable devices,

and so on.

These devices all generate, capture, and store data -- approximately 2.5 quintillion bytes

every day.

That's the equivalent of 10 million Blu-ray DVD's.

Variety: Let's think about the different types of data; text, pictures, film, sound, health

data from wearable devices, and many different types of data from devices connected to the

Internet of Things.

Veracity: 80% of data is considered to be unstructured and we must devise ways to produce

reliable and accurate insights.

The data must be categorized, analyzed, and visualized.

Data Scientists today derive insights from Big Data and cope with the challenges that

these massive data sets present.

The scale of the data being collected means that it’s not feasible to use conventional

data analysis tools.

However, alternative tools that leverage distributed computing power can overcome this problem.

Tools such as Apache Spark, Hadoop and its ecosystem provide ways to extract, load, analyze,

and process the data across distributed compute resources, providing new insights and knowledge.

This gives organizations more ways to connect with their customers and enrich the services

they offer.

So next time you strap on your smartwatch, unlock your smartphone, or track your workout,

remember your data is starting a journey that might take it all the way around the world,

through big data analysis, and back to you.