Data Quality Engineering

Session 2

April 2022

CGI

Agenda

01 Recap and Setup

02 Data QE with SQL

03 Data QE with Python

04 Big Data Testing and Pyspark

05 Machine Learning

Software Setup Steps

Steps for SQL Topics:

1. Download and Install SQL Server from: Download Microsoft® SQL Server® 2019 Express from Official Microsoft Download Center
2. Download and Install SQL Server Management Studio: Download SQL Server Management Studio (SSMS) - SQL Server Management Studio (SSMS) | Microsoft Docs
   - A Connection String will be generated towards the end of the installation process. Please copy and save it somewhere as it will be used to connect to SQL Server from python.
     Here's a sample connection string for reference -> *Server=localhost\SQLEXPRESS;Database=master;Trusted_Connection=True;*
3. Download Sample database named *AdventureWorksLT2016.bak* from AdventureWorks sample databases - SQL Server | Microsoft Docs
4. Connecting/Restoring these sample databases to SQL Server can be done by following the steps under "**Restore to SQL Server**" on the same webpage.
5. Additional files that will be needed to follow along during the session will be shared along with this deck.

For Python and Machine Learning:

- Install Anaconda Individual edition - Anaconda | Individual Edition

Tools and processes to test data

CGI

SQL	Python	PySpark	Spark streaming
• For testing RDBMS • Validate table to table schema and data transformation • MS SQL, Oracle, My SQL	• For testing between Structured and semi-structured data or between heterogenous systems • Validate table to file transformations • Anaconda, NumPy , Jupyter Notebook	• For testing big data • Improved performance	• For testing streaming data • Kafka streams

What is SQL?

SQL is the standard used to manage data in relational tables. Structured Query Language normally referred as SQL and pronounced as SEE QU EL.. ☺

SQL allows users to create databases, add data, modify and maintain data. It is governed by standards maintained by ISO(International Standards Organization).

Example of a relational table:

Employee

Emp Id	Emp Name	Age	Dept_id
1	John	40	1
2	Linda	35	1
3	Max	30	2

Department

Dept_id	Dept_name
1	Accounts
2	Production

Lets do Data QE with SQL

Overview – Validation of data flow between 2 RDBMS tables with or without transformations

What are transformations - Any kind of data correction, aggregation, summarization or manipulation done on source data to achieve the necessary values in target.

Example 1 : Data flow without transformation (Straight/Direct Move)

Source table – In a Relational DB

Emp Id	Emp_Name	Age	Dept_id
1	John	40	1
2	Linda	35	1
3	Max	30	2

ETL Process

Target table – Same DB as source

Emp Id	Emp_Nm	Age
1	John	40
2	Linda	35
3	Max	30

Example 2: Data flow example with transformation

Source table – In a Relational DB

Emp Id	Emp_Name	Age	Dept_id
1	John	40	1
2	Linda	35	1
3	Max	30	2
4	Arun	37	3

ETL Process

Target table – Same DB as source

Dept_id	Emp_Count
1	2
2	1
3	1

What to test?

Standard Data testing cases:

- Metadata validation
  - Table structure validations including column naming and order
  - Data type and data length validation for each column
- Data Profile Validation
  - Check for duplicate records and NULL values
  - Minimum, maximum and sum comparison for numeric fields
  - String length minimum and maximum comparison
  - Check for extra records in target (Ghost records)
- Data Comparison
  - Value to value comparison between source and target

```
select column_name, data_type
from testdb.information_schema.columns
where upper(table_schema) = upper('dbo') and upper(table_name) = upper('Customer_Test')
```

esults | Messages

column_name	data_type
CustomerID	int
NameStyle	bit
Title	nvarchar
FirstName	nvarchar
MiddleName	nvarchar
LastName	nvarchar
Suffix	nvarchar
CompanyName	nvarchar
SalesPerson	nvarchar
EmailAddress	nvarchar
Phone	nvarchar
PasswordHash	varchar
PasswordSalt	varchar
rowguid	uniqueidentifier
ModifiedDate	datetime

Lets do Data QE with Python

Overview – Validation of data flow between heterogenous systems

What are heterogenous systems – Heterogenous systems have different types of relational or non-relational databases which together work as a single entity to form a data warehouse or data lake

Example 1 :

Data flow between 2 different types of relational DBs:

Source table – In MS SQL Server

Emp Id	Emp_Name	Age	Dept_id
1	John	40	1
2	Linda	35	1
3	Max	30	2

ETL Process

Target table – In Oracle

Emp Id	Emp_Nm	Age
1	John	40
2	Linda	35
3	Max	30

Example 2:

Data flow between a File and a Table:

Source table – In .csv File

CustomerId	Title	FirstName	LastName
1	Mr.	Orlando	Gee
2	Mr.	Keith	Harris
3	Ms.	Donna	Carreras

Microsoft Excel
ma Separated Valu

ETL Process

Target table – In MS SQL Server

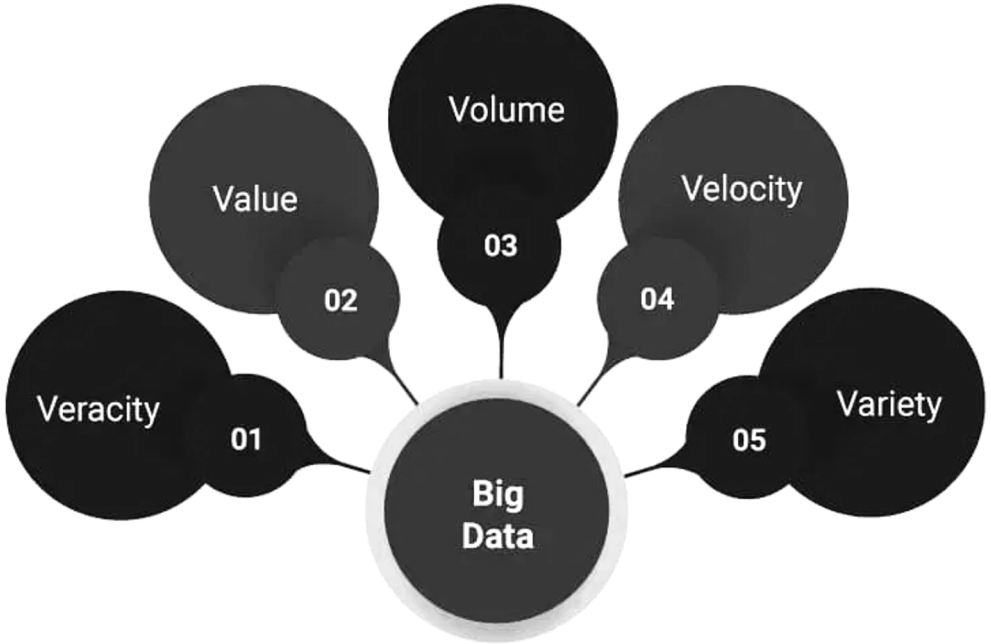
CustomerId	Title	FirstName
1	Mr.	Orlando
2	Mr.	Keith
3	Ms.	Donna

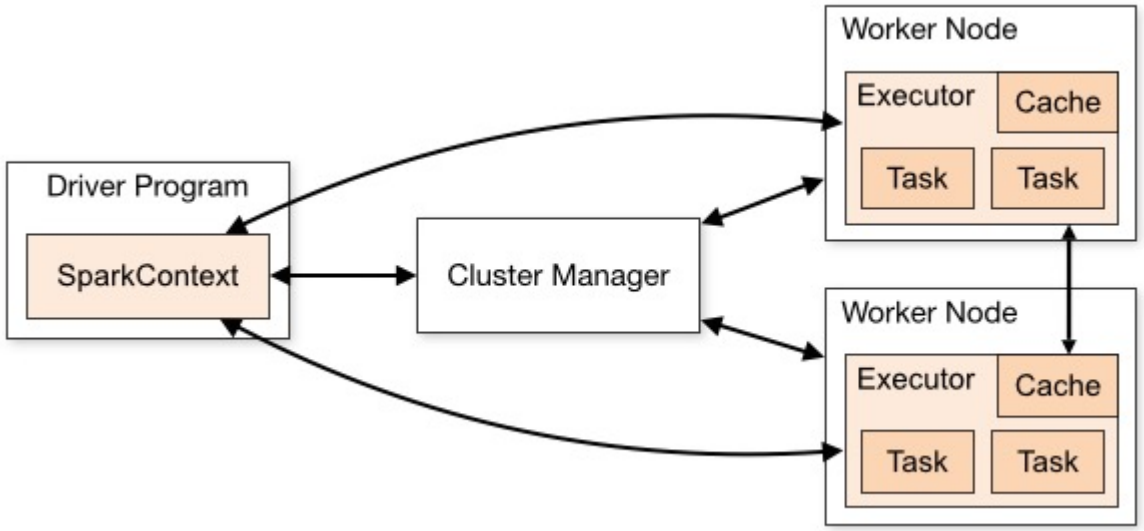
Types of Data Platforms

Transactional	Business Intelligence/Analytics	Big Data / Cloud Based Intelligent Platforms
Data generated by customers on a daily bases are persisted.	Data persisted on a department for a timeline	Large variety and volume of data persisted. Velocity is also a critical factor.
ATM Transactions, Bill Payments etc.,	Customer buying trend across states, Patient trend across country etc.,	Customer emotions related to a product, machine learning and artificial intelligence based case studies etc.,

Big Data Testing

- Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation
- Big Data Testing involves testing very large volumes of data, usually stored in a distributed system like HDFS, Hive or NoSQL databases
- Most functional test cases stay the same as any regular data testing scope but approach and technology is usually different in order to process larger data volume





What is PySpark

- Apache Spark is an open-source, cluster computing system which is used for big data solution.
- Pyspark is a python based interface to the spark execution engine. Integration to spark is achieved via Py4j library.
- It is used for processing large data sets over distributed systems which allows for parallel processing as well
- Pyspark is a very popular option for building and testing big data pipelines and is also growing in popularity for machine learning

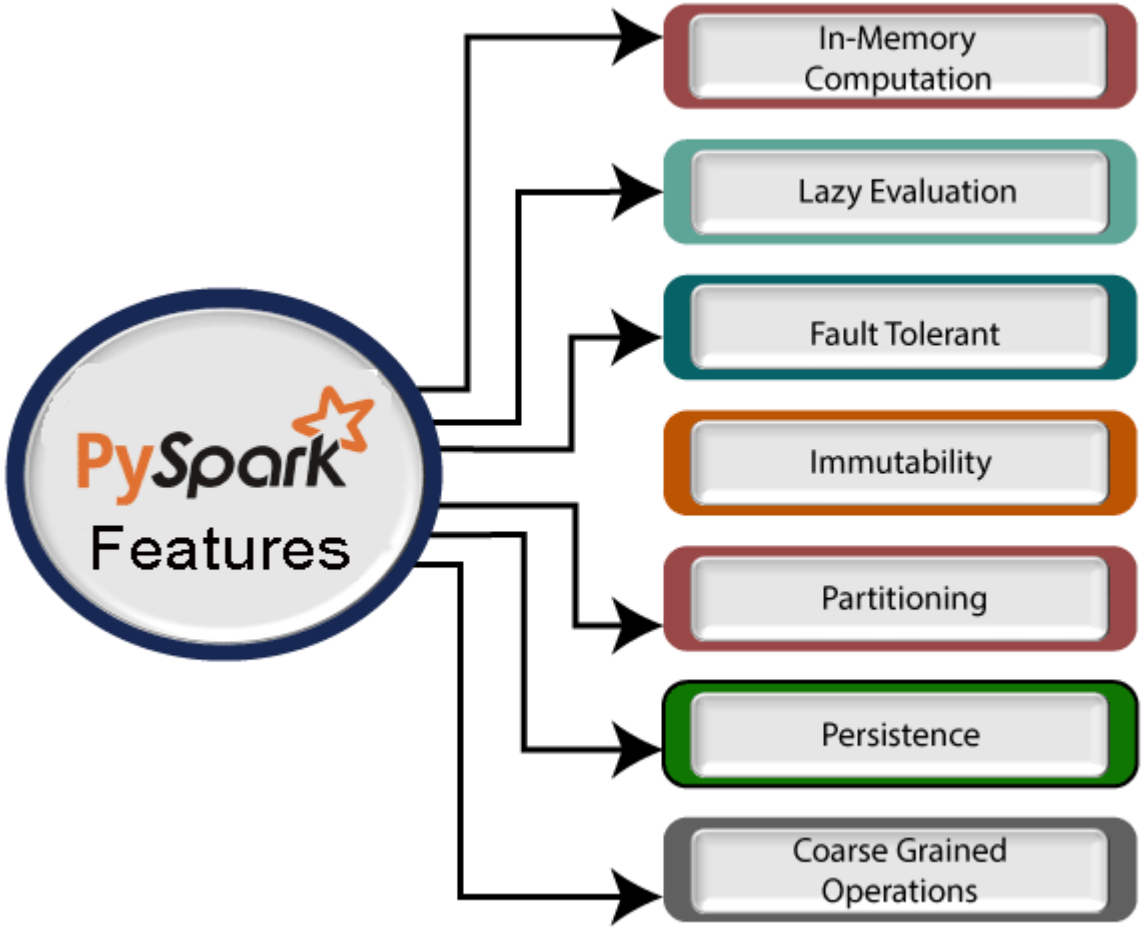
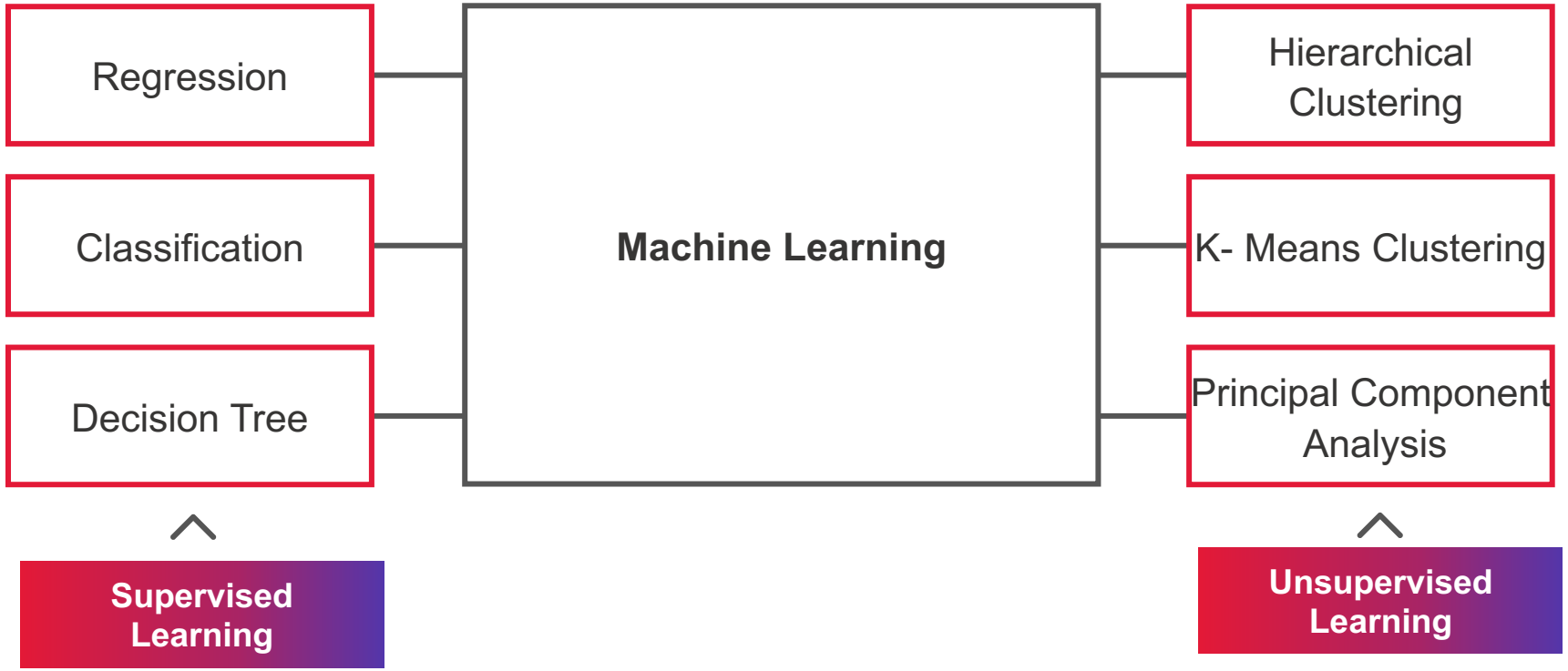


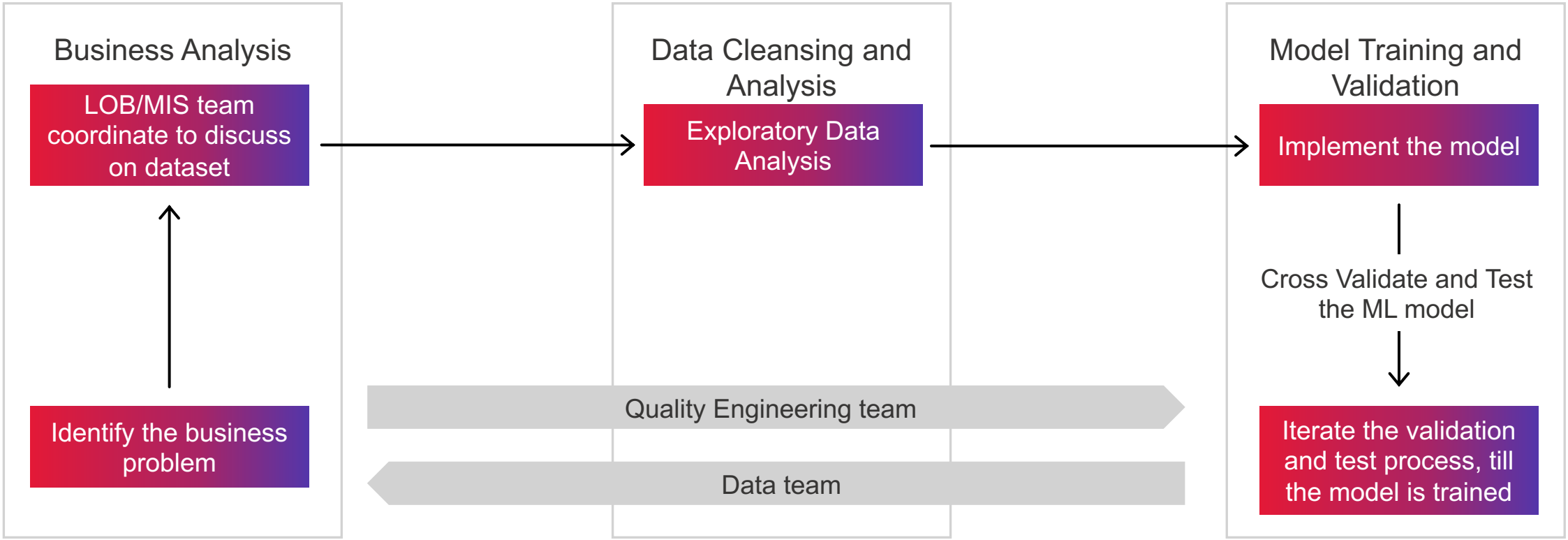
Image Reference: javatpoint.com

Why Machine Learning?

Data Monitoring	Predict Business Outcomes	Pattern recognition	Customer Segmentation	Business Strategy Planning
Machine Learning will help in monitoring PII/PHI data	Machine Learning models can be used for predicting future sales/outcomes.	Search engines use machine learning.	Banks/Financial Services utilize machine learning to focus on customer groups for lending and other financial deals	In the modern world, Machine learning is used for business strategy planning

Types of Machine Learning





Machine Learning – End to End flow

Data Team Roles – To Build Machine Learning

01 **Business Sponsor**

02 **Data Scientist**

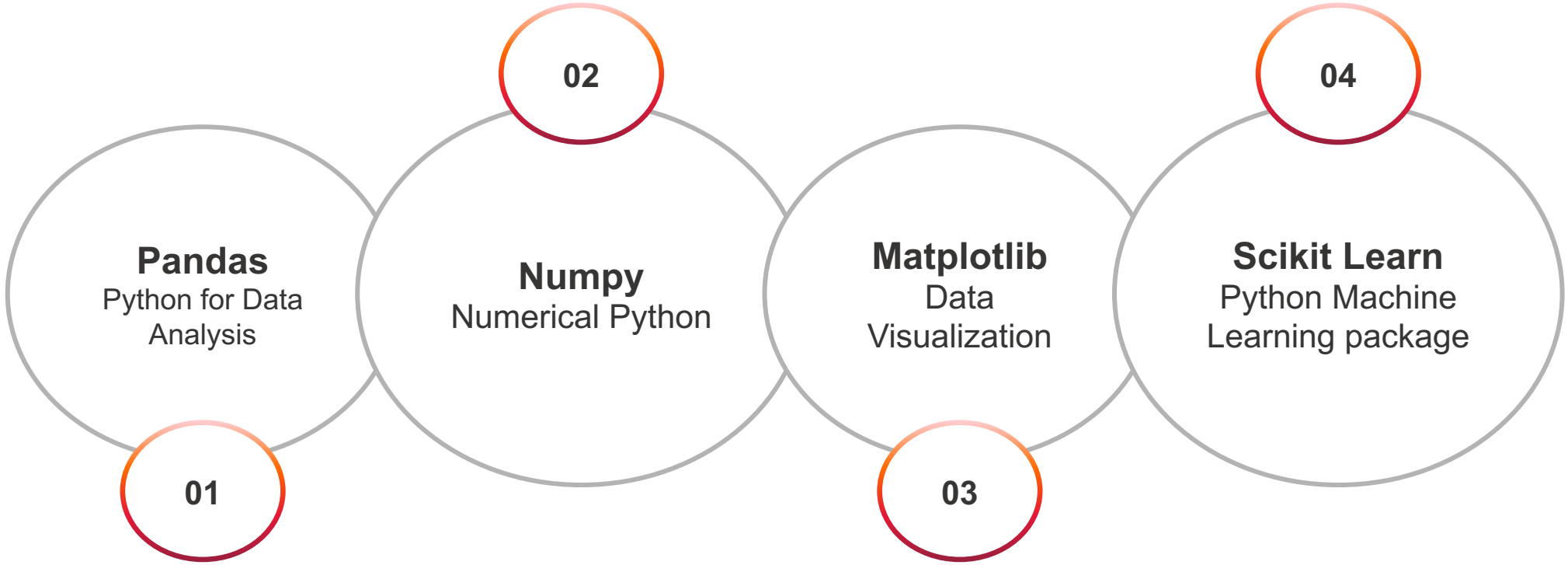
03 **Business Analyst**

04 **Business Users**

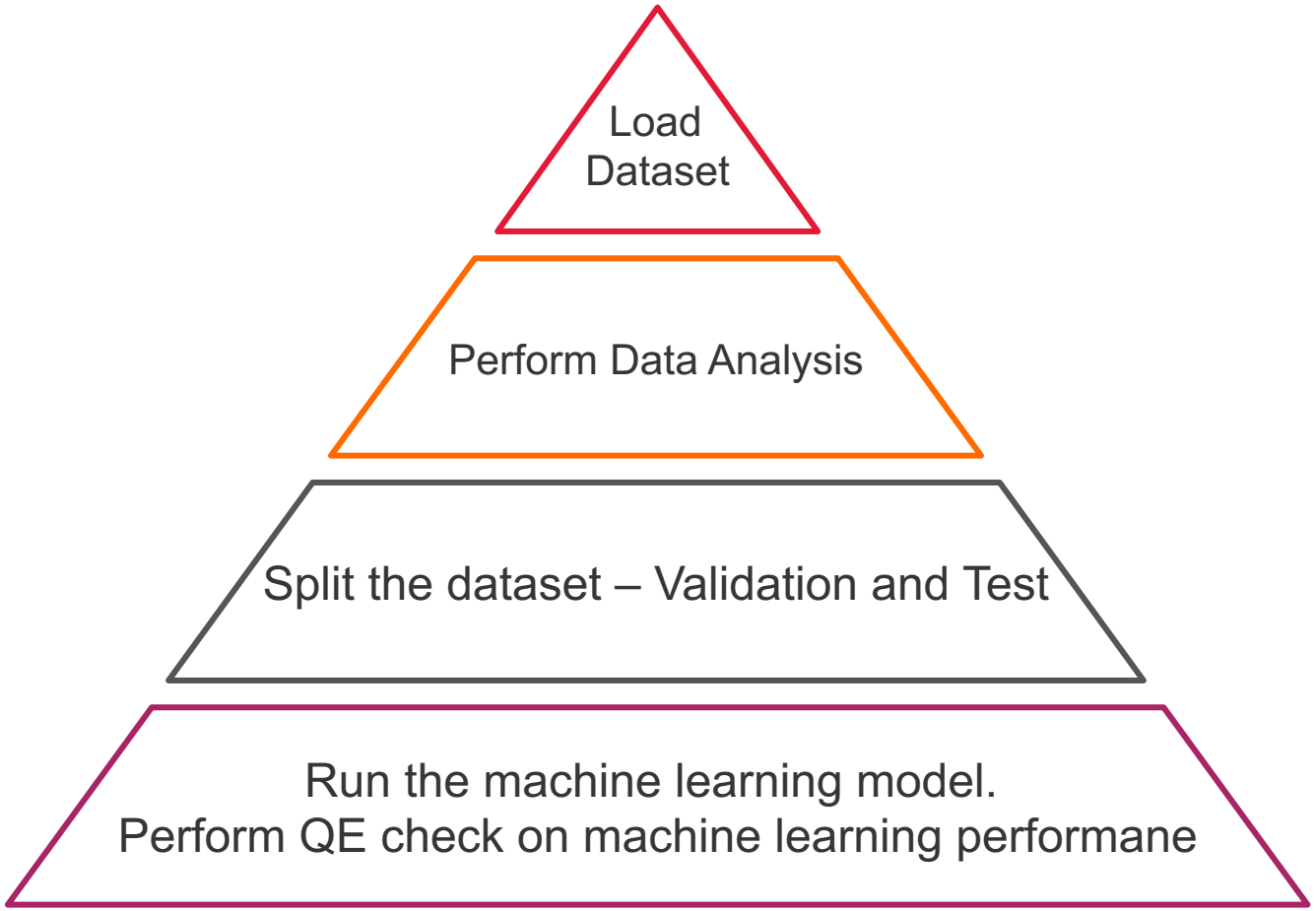
05 **Data Quality Engineers**

06 **Infrastructure Engineers**

Machine Learning Demo and Toolset



Machine Learning Demo



Questions

?

Wishing all of you great success in your career!

Please reach out to us for any questions

Heather Fusko - Heather.Fusko@cgi.com

Lakshmi Ranganathan – Lakshmi.Yeriranganathan@cgi.com

Sharath Chandran - Sharath.Chandran@cgi.com

Shobhit Sharma – Sho.sharma@cgi.com

Mrityunjay Singh – Mrityunjay.Singh@cgi.com