Data Quality Engineering

April 2022

CGI

Agenda

01

CGI at a Glance

02

Quality Engineering

03

Data Quality Engineering

04

Structured Query Language/ Python

05

Big Data and Machine Learning (Session 2)

CGI at a glance

Founded in 1976
46 years of excellence

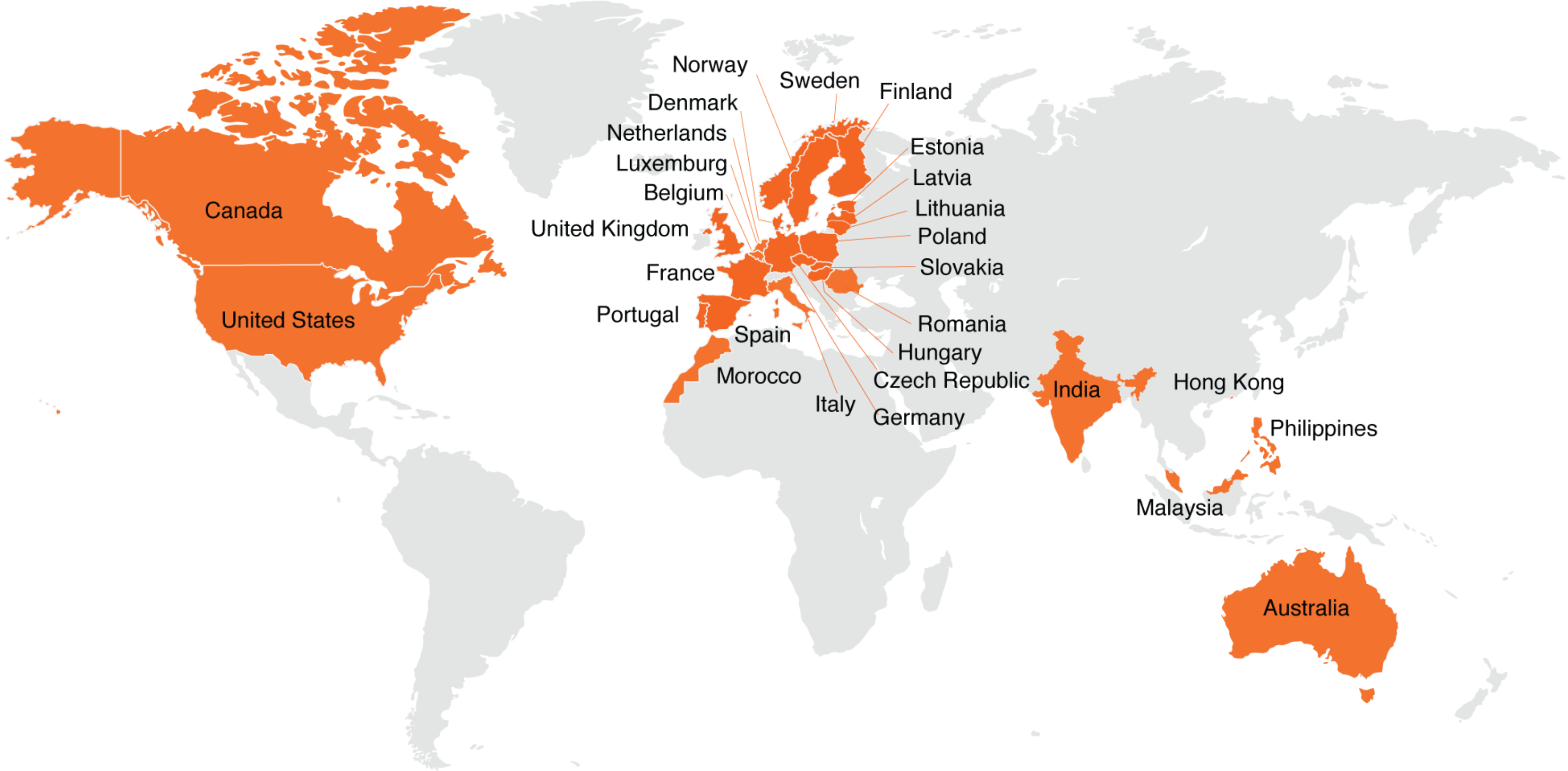
CA$12.1 billion revenue

78,000 consultants

400 locations in **40** countries

5,500 clients benefiting from end-to-end services across **10 focused industries**

170+ IP-based solutions
serving **50,000** clients



What drives us?

Our dream

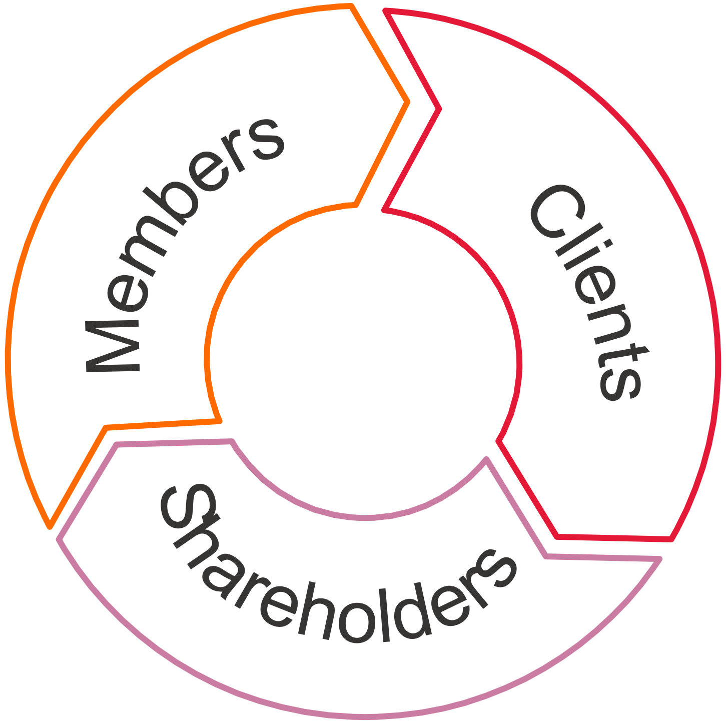
To create an environment in which we enjoy working together and, as owners, contribute to building a company we can be proud of.

Our mission

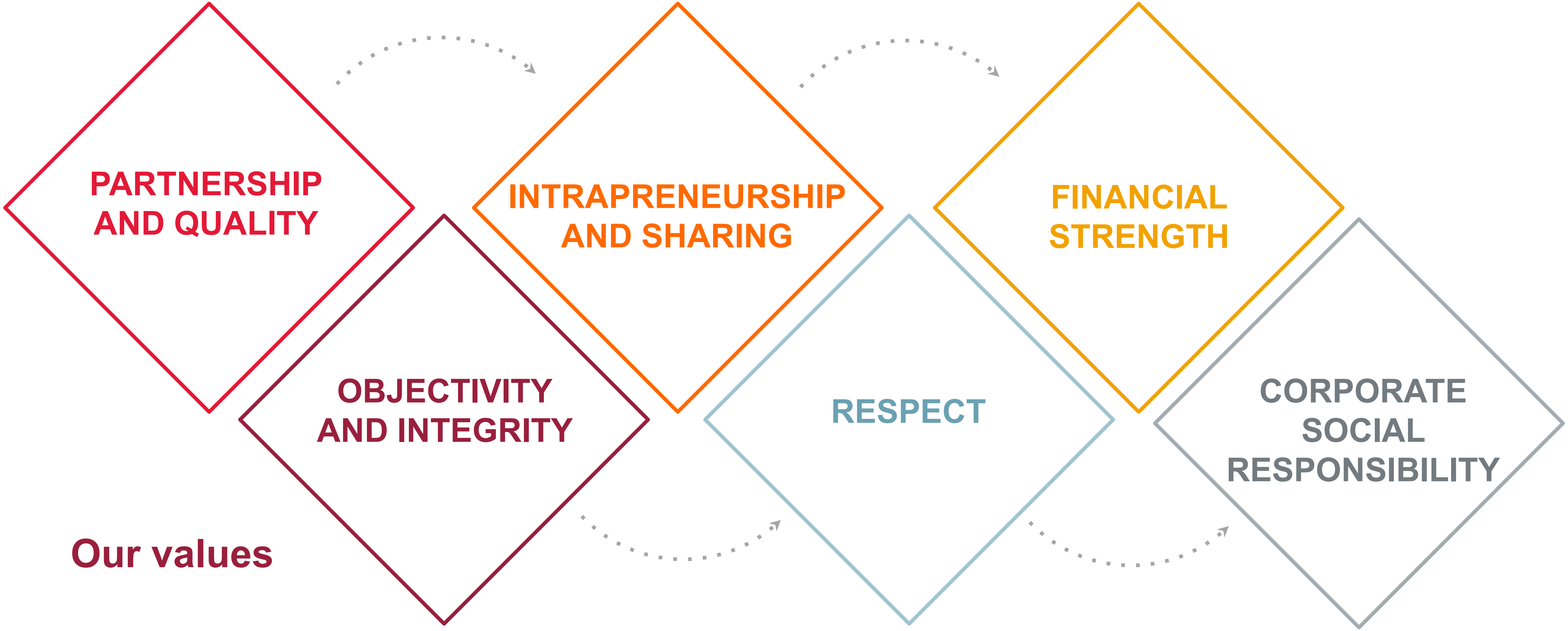
To help our clients succeed through outstanding quality, competence and objectivity, providing thought leadership and delivering the best services and solutions to fully satisfy client objectives in information technology, business processes and management. In all we do, we are guided by our Dream, living our Values to foster trusted relationships and meet our commitments now and in the future.

Our vision

To be a global world class end-to-end IT and business consulting services leader helping our clients succeed.



What guides us?



PARTNERSHIP AND QUALITY

OBJECTIVITY AND INTEGRITY

INTRAPRENEURSHIP AND SHARING

RESPECT

FINANCIAL STRENGTH

CORPORATE SOCIAL RESPONSIBILITY

Our values

A few of our clients

CGI

Financial Services

BNP PARIBAS

ING

TD

PNC

SOCIETE GENERALE

Health

CMS CENTERS for MEDICARE & MEDICAID SERVICES

DMS Defence Medical Services

HUS

NHS

Pfizer

Government

esa

HM Government

Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE

Communications

at&t

Bell

comcast.

TeliaSonera

vodafone

Utilities

eDF

Hydro Québec

NSTAR

Oil & Gas

bp

TOTAL

ExxonMobil

Manufacturing

AIRBUS

VOLVO

BOMBARDIER

EADS

MICHELIN

RioTinto

Transportation

amadeus

DB SCHENKER

KLM

NSW GOVERNMENT Transport for NSW

Post & Logistics

DHL

LIVINGSTON

LA POSTE

CANADA POST POSTES CANADA

Retail & Consumer Services

Carrefour

L'ORÉAL

LVMH MOËT HENNESSY.LOUIS VUITTON

MOLSON Coors

Auchan

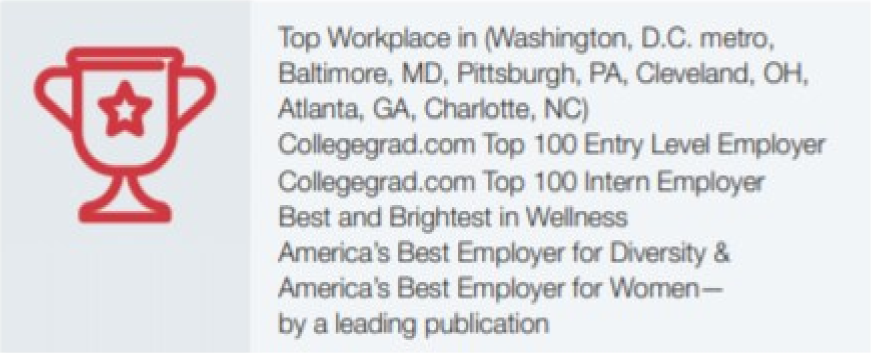
Over **5,500** commercial and government organizations worldwide

College Recruiting Overview

Early Careers at CGI

Our programs will give you the fundamentals to ease & accelerate you assimilation into CGI.

CGI’s intern program offers students real-world technical & business consulting experience



Top Workplace in (Washington, D.C. metro, Baltimore, MD, Pittsburgh, PA, Cleveland, OH, Atlanta, GA, Charlotte, NC)
Collegegrad.com Top 100 Entry Level Employer
Collegegrad.com Top 100 Intern Employer
Best and Brightest in Wellness
America’s Best Employer for Diversity &
America’s Best Employer for Women—
by a leading publication

Full Time Roles

Software Developer

- System design
- Systems and application development
- Data design
- Database administration
- Defining and maintaining data security and integrity

Programmer/Analyst

- Testing and implementation of new technology
- Software installation and configuration
- Investigate and debug errors, troubleshoot issues
- Develop and build code applications
- Write technical documents
- User Support

Business Analyst

- Business analysis
- Requirements gathering
- Direct user support and analysis
- Tracking software and
- documentation defects
- Onsite consulting and training
- System testing
- Decision analysis

Intern Roles

Business Analyst Intern

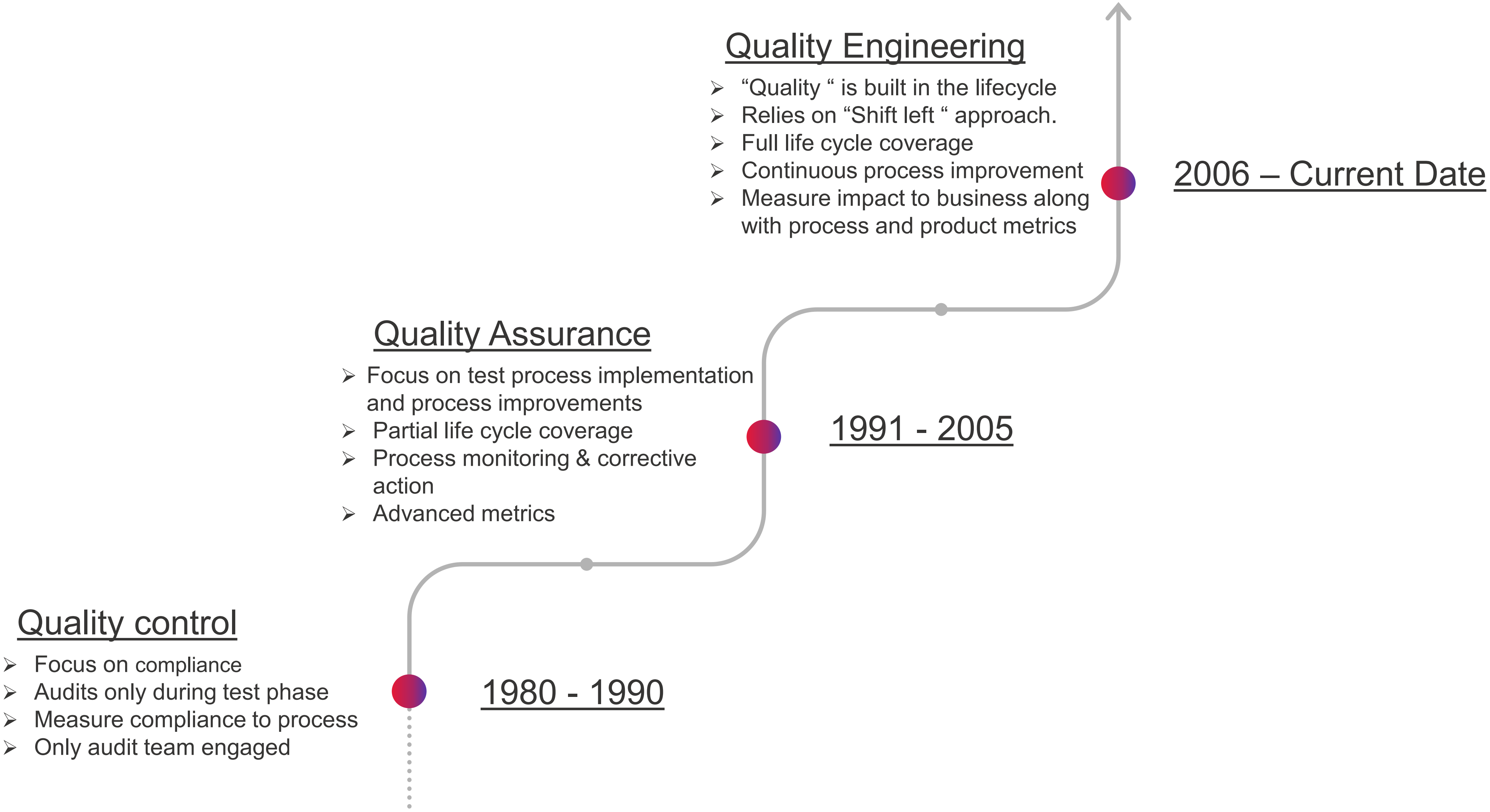
- Client requirements gathering and analysis
- Tracking testing and document of defects
- Onsite client consulting and support
- Writing program and system user manuals and/or training materials

Development/Engineering Inter

- Create systems design utilizing client requirements
- Applications development and computer programming
- Database maintenance and configuration management
- Software installs, technical testing and reporting

Quality Engineering

Quality Engineering Evolution



Quality control

- Focus on compliance
- Audits only during test phase
- Measure compliance to process
- Only audit team engaged

1980 - 1990

Quality Assurance

- Focus on test process implementation and process improvements
- Partial life cycle coverage
- Process monitoring & corrective action
- Advanced metrics

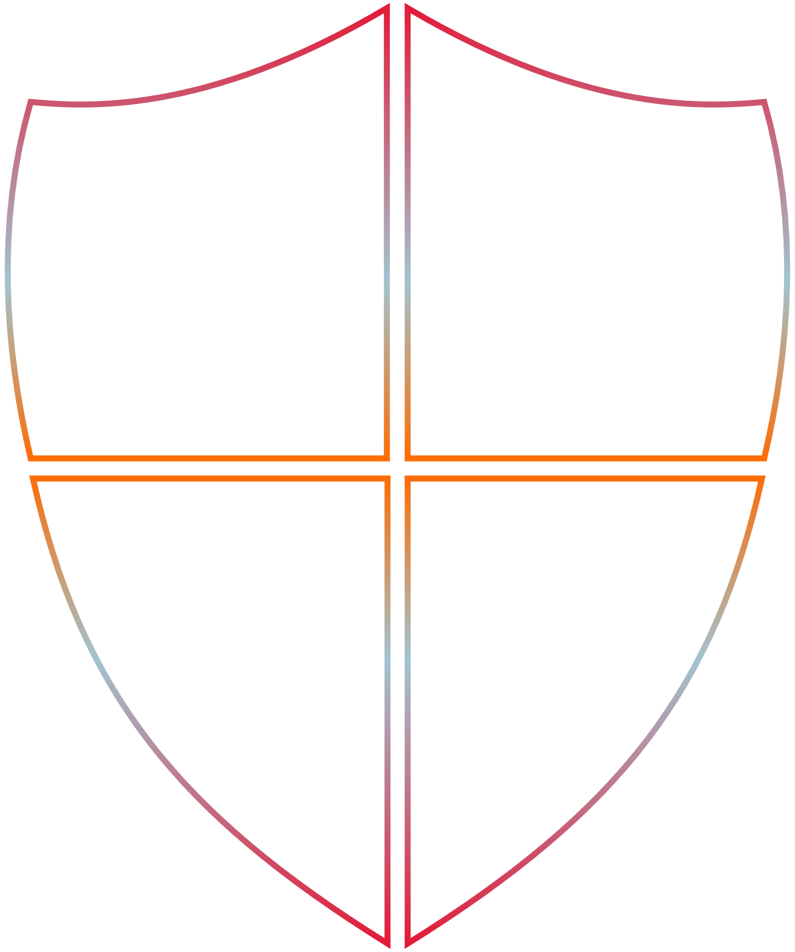
1991 - 2005

Quality Engineering

- “Quality “ is built in the lifecycle
- Relies on “Shift left “ approach.
- Full life cycle coverage
- Continuous process improvement
- Measure impact to business along with process and product metrics

2006 – Current Date

Driving Factors – Quality Engineering



Speed to market
Alignment with agile principles
.

Accountability
Everyone is accountable for software quality

Technical Debt Reduction
Focused approach to build robust applications

End User Focused
Plan and deliver value to end users

Software Testing Specializations

API Testing: API testing is a software testing practice that tests the APIs directly — from their functionality, reliability, performance, to security. Part of integration testing, API testing effectively validates the logic of the build architecture within a short amount of time.

Front End or UI Testing: The goal of Front End Testing is to test functionalities and verify that a website or app's presentation layer is bug or error-free.

Security or Vulnerability Analysis (Penetration Testing): Vulnerability assessment, one of the most important phases of penetration testing, occurs when your team maps the profile of the environment to publicly known or, in some cases, unknown vulnerabilities.

Availability Testing: Availability Testing which is also called Durability Testing is a kind of performance testing in which the application runs for a set period of time and collects failure events and repair times, and compares the availability percentage to the service level agreement.

Mobile Testing: Mobile testing is the process by which applications, software and websites designed for mobile devices are tested for functionality, usability, and consistency.

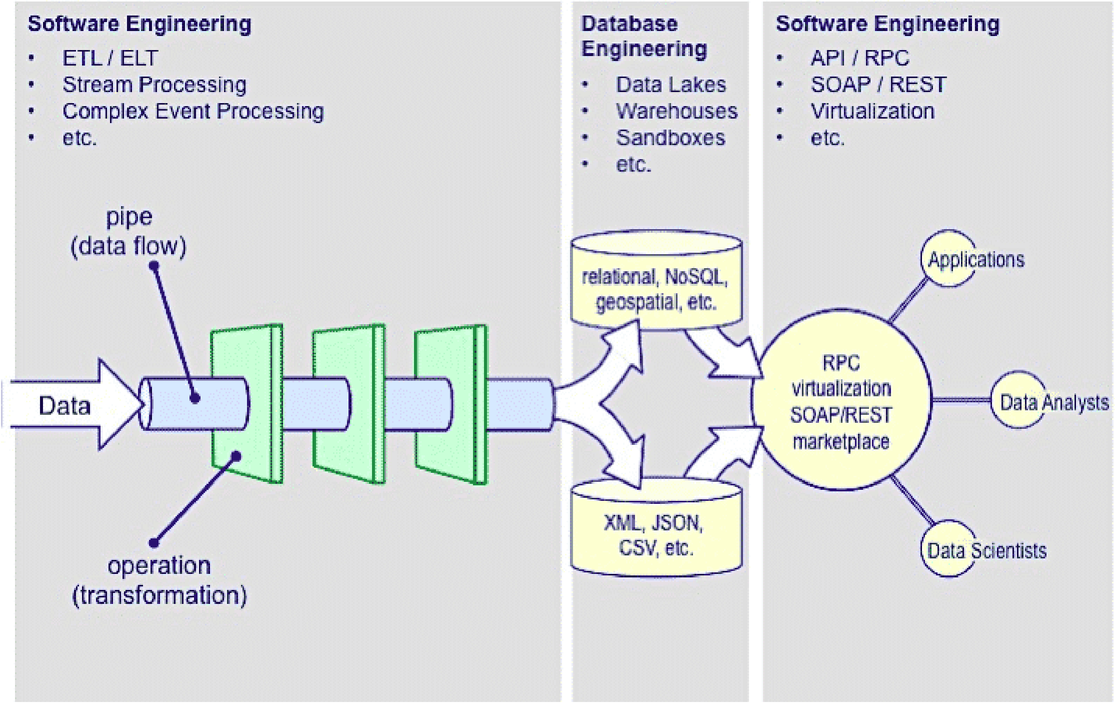
Data Quality Testing: Ensuring data is moving from one system to another, is transformed correctly, stored correctly. This needs knowledge of how different sorts of data are stored, processed and used in an application or several applications. It has below 2 common implementations:

- **Data Warehouse or ETL Testing:** It is a testing method in which the data inside a data warehouse is tested for integrity, reliability, accuracy and consistency in order to comply with the company's data framework. The main purpose of data warehouse testing is to ensure that the integrated data inside the data warehouse is reliable enough for a company to make decisions on.
- **Big Data Testing:** This is a testing process for a big data application in order to ensure that all the functionalities of the application work as expected. In Big Data testing strategy, QE members verify the successful processing of large data volumes using commodity cluster and other supportive components.

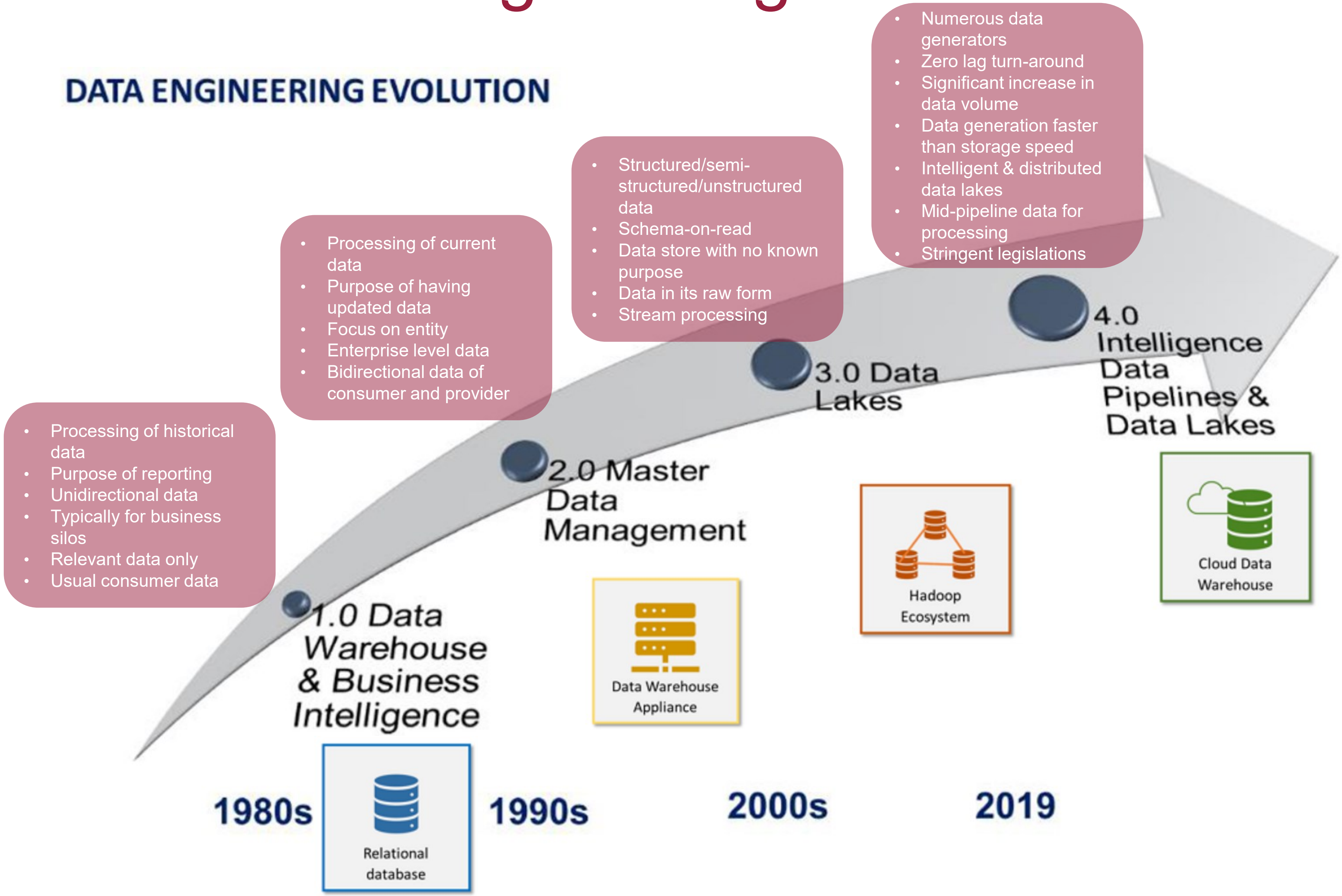
Data Engineering

What is Data Engineering?

The key to understanding what data engineering lies in the “engineering” part. Engineers design and build things. “Data” engineers design and build pipelines that transform and transport data into a format wherein, by the time it reaches the Data Scientists or other end users, it is in a highly usable state. These pipelines must take data from many disparate sources and collect them into a single warehouse that represents the data uniformly as a single source of truth.



Evolution of Data Engineering


DATA ENGINEERING EVOLUTION
1.0 Data Warehouse & Business Intelligence
2.0 Master Data Management
3.0 Data Lakes
4.0 Intelligence Data Pipelines & Data Lakes
Relational database
Data Warehouse Appliance
Hadoop Ecosystem
Cloud Data Warehouse
1980s
1990s
2000s
2019

- Processing of historical data
- Purpose of reporting
- Unidirectional data
- Typically for business silos
- Relevant data only
- Usual consumer data

- Processing of current data
- Purpose of having updated data
- Focus on entity
- Enterprise level data
- Bidirectional data of consumer and provider

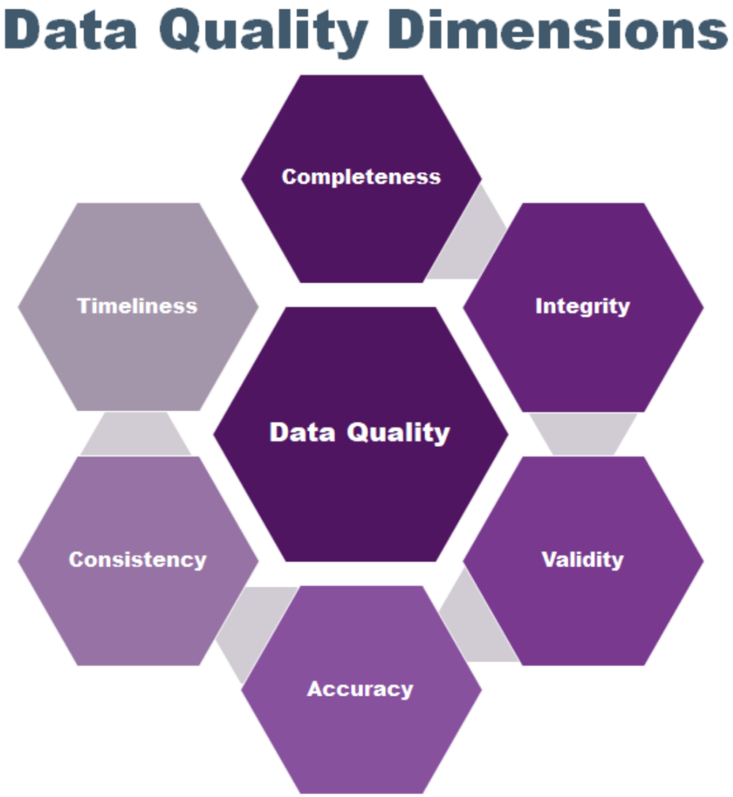
- Structured/semi-structured/unstructured data
- Schema-on-read
- Data store with no known purpose
- Data in its raw form
- Stream processing

- Numerous data generators
- Zero lag turn-around
- Significant increase in data volume
- Data generation faster than storage speed
- Intelligent & distributed data lakes
- Mid-pipeline data for processing
- Stringent legislations

Data Quality Engineering

What is Data Quality Engineering?

The Data Quality Engineering is a discipline for designing, developing, documenting and performing data quality checks across all data assets. That includes ETL jobs, reports, dashboards and data pipelines. The primary goal for this role is to ensure high quality of data delivered to internal stakeholders and customers. Validation of data in data repositories against data from source systems and validation of metrics and data in reports/dashboards against data in the repositories is a key responsibility. Principle responsibilities are to making data assets consistently accurate for users.



Types of Data Platforms

Transactional

Data generated by customers on a daily bases are persisted.

↑

ATM Transactions, Bill Payments etc.,

Business Intelligence/Analytics

Data persisted on a department for a timeline

↑

Customer buying trend across states, Patient trend across country etc.,

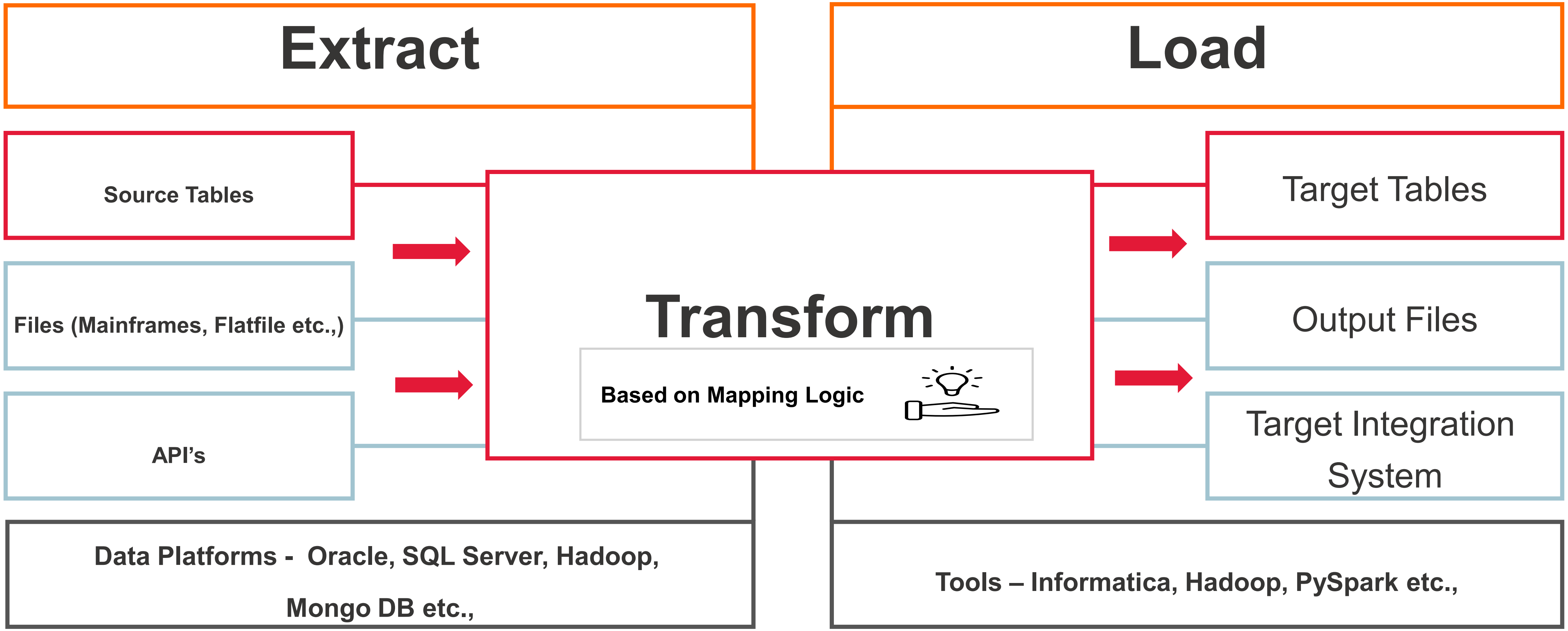
Big Data / Cloud Based Intelligent Platforms

Large variety and volume of data persisted. Velocity is also a critical factor.

↑

Customer emotions related to a product, machine learning and artificial intelligence based case studies etc.,

Data Load Process



Extract

Source Tables

Files (Mainframes, Flatfile etc.,)

API's

Transform

Based on Mapping Logic

Load

Target Tables

Output Files

Target Integration System

Data Platforms - Oracle, SQL Server, Hadoop, Mongo DB etc.,

Tools – Informatica, Hadoop, PySpark etc.,

Tools and processes to test data

SQL

- For testing RDBMS
- Validate table to table schema and data transformation
- MS SQL, Oracle, My SQL

Python

- For testing between Structured and semi-structured data or between heterogenous systems
- Validate table to file transformations
- Anaconda, NumPy , Jupyter Notebook

PySpark

- For testing big data
- Improved performance

Spark streaming

- 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

SQL syntax and query

SQL is case in-sensitive, that is keyword SELECT and select is same for SQL. Every SQL command should end with a semi-colon (;).

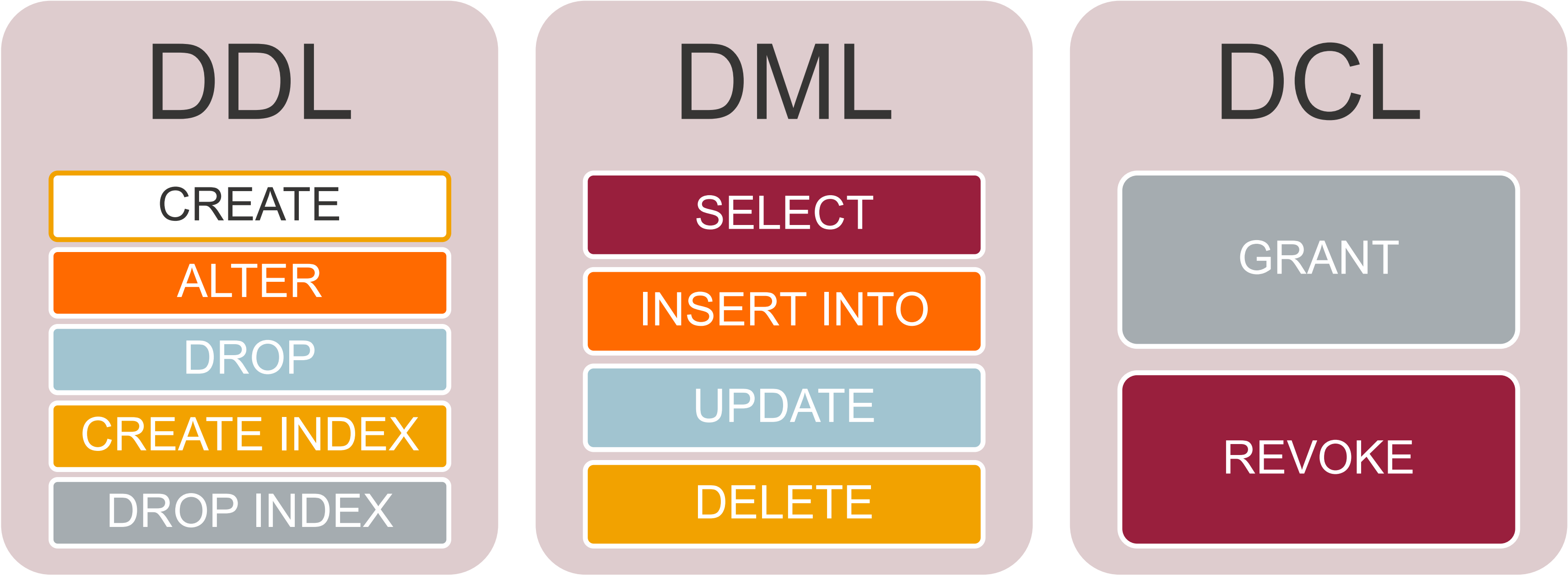
If the syntax is not proper, then executing the command would result in syntax error.

Command used to fetch data from table is called as query. A basic SQL query consists of SELECT, FROM and WHERE clause.

SQL SELECT Command example:

```
SELECT col1, col2, col3,…..
FROM table_name
        WHERE condition
<Group BY> …
<Order By> …;
```

Types of SQL commands



DDL

- CREATE
- ALTER
- DROP
- CREATE INDEX
- DROP INDEX

DML

- SELECT
- INSERT INTO
- UPDATE
- DELETE

DCL

- GRANT
- REVOKE

DDL – Data Definition Language

DML – Data Manipulation Language

DCL – Data Control Language

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

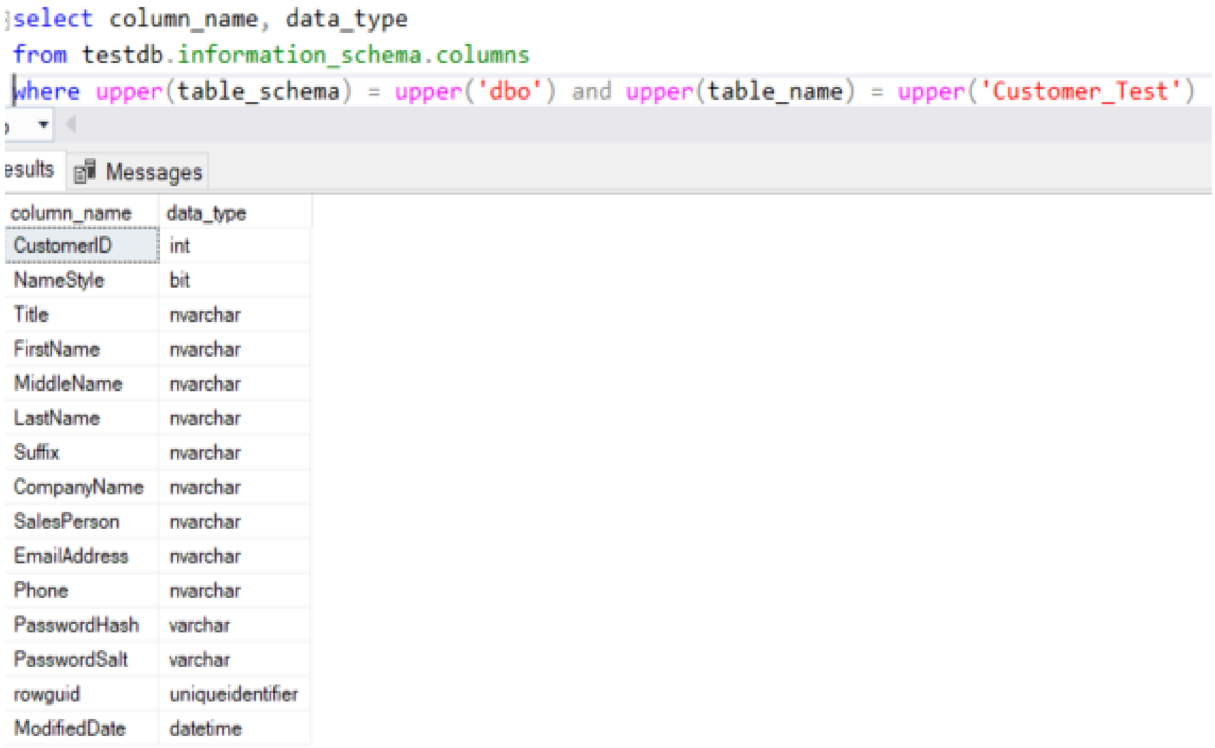
Target table – Same DB as source

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



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

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:

Microsoft Excel
ma Separated Valu

Source table – In .csv File

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

ETL Process

Target table – In MS SQL Server

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

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

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