CSC 460: Team 1
Modules 1 & 3

Sohail Ahmad, Priti Saha,Kenneth Romero, Joquanna Scott, Jeffrey Umanzor, Aquib Zaman

Data Science Programming Languages and Data Integration/Extraction

Agenda

1. Introduction
2. Programming Languages Overview
3. History and Evolution
4. Data Integration and Extraction Overview
5. RESTful vs. GraphQL, Web Scraping, SOAP
6. Tool Comparison
7. Industry Relevance
8. Recommendations and Conclusions
9. Q&A

Introduction

Module 1: Data Modeling/ Programing Language

Data modeling involves creating visual representations, such as graphs, to understand the connections between data points in an information system. Its goal is to illustrate the types of data used and stored, their relationships, organization methods, formats, and attributes. Common types of data models include conceptual, logical, physical, hierarchical, entity-relationship (ER), dimensional, and object-oriented models. Among these, the dimensional models; specifically the star schema and snowflake schema are most popular. The star schema organizes data into facts and dimensions in a star-like pattern, while the snowflake schema adds complexity with additional layers of dimensions.

Languages we will be discussing :

- Python: Versatile, easy to learn, ideal for data science and web development.
- R: Specialized for statistics and data visualization.
- MATLAB: Focused on numerical computing, popular in engineering and academia.

Module 3: Injection/Extraction Methods

Injection and extraction methods in databases involve techniques for manipulating data input and retrieval. Injection attacks, such as SQL injection, occur when attackers input harmful data to manipulate queries, potentially leading to unauthorized data access or modification. Conversely, data extraction methods retrieve data from databases for legitimate purposes like reporting or analysis. Different methodologies exist for injection and extraction, including REST APIs, GraphQL, SOAP, and cURL. RESTful APIs structure data around resources, using HTTP methods at endpoints for data manipulation. GraphQL offers more flexibility, allowing clients to specify exactly what data they need in one query, which helps avoid over-fetching or under-fetching. cURL is a command-line tool used for data and enabling API testing and automation of data operations. While REST and GraphQL suit structured web data, cURL is better for file-based transfers and lower-level data manipulation.

History and Evolution

Python

- Created in 1991
- Gained popularity in the early 2000’s due to its simplicity and scientific libraries
- Widely used for data integration due to extensive libraries

MATLAB

- Developed in the late 1970s
- Originally designed for numerical computing
- Still commonly used for simulations , numerical computations and visualization

R

- Created in 1993
- Initially designed with a main focus statistical computing and graphics
- Popular for its integration with data science libraries such as ggplot2 and dplyr for visualization and data manipulation

History and Evolution

RESTful APIs

- Introduced in 2000
- Gained popularity for its simplicity and scalability
- Widely used for building web services due to its resource-based architecture

GraphQL

- Created in 2012 by Facebook
- Initially developed to to streamline data fetching for Facebook’s mobile application
- Quickly growing in popularity due to efficiency and flexibility, by allowing more efficient data queries by only fetching required data fields in one request

SOAP

- Created in 1998 by Microsoft
- First designed for distributed computing and remote communication over the web using XML messaging
- Still used in enterprise environments where security and complex transaction are required

Popularity and Community Support

Python

- Very large community supported by a extensive ecosystem of libraries (NumPy, Pandas, Matplotlib)

MATLAB

- Smaller but strong community in engineering and scientific research
- Popular in fields such as aerospace, mechanical, and civil engineering where numerical simulations are needed

R

- Strong statistical community, popular in academia and research
- ggplot2 popular for data visualization, dplyr popular for data manipulation

Popularity and Community Support

RESTful APIs

- Vast community with extensive framework support (Spring, Express.js)
- Dominates modern web development and microservices architecture

GraphQL

- Growing community with strong support from companies like GitHub and Shopify
- Efficient for front-end development and single-page applications, minimizing data over-fetching

SOAP

- Strong in enterprise environments where security and transaction integrity are required
- Complex but still used in finance, government, and telecom industries

Feature Comparison of Languages

	Python	R	MATLAB
Ease of Learning	Easy, readable syntax	Moderate, steep for new user	Moderate, math focused
Use Cases	Machine Learning, data science	Statistical Modeling	Numerical Computation
Scalability	Scalable with large projects	Limited to small/medium datasets	Scalable in academia and engineering
Library Support	Extensive	Strong in statistics	Specialized toolboxes
Performance	Moderate	Slow on large data sets	High for matrices
integration	APIs, databases	Data Sources	Engineering Apps
License	Open-Source	Open-Source	Licenced, paid

Feature Comparison of Tools

	RESTful	GraphQL	SOAP
Ease of Learning	Easy	Moderate, requires understanding schemas	Complex, strict rules and protocols
Use Cases	Web services, CRUD operations	Custom data queries	Secure transactions, enterprise level data exchange
Scalability	Highly Scalable, widely used	Scalable	Scalable, used in enterprise applications
Performance	Fast	Faster (only fetches requested data)	Slower, verbose XML structure
integration	APIs, databases	Data Sources	Engineering Apps
License	Open-Source	Open-Source	Open-Source, but usually used with licensed enterprise software

Pros and Cons

Programming Languages

Python

Pros

- Versatile
- large library support
- widely used.

Cons

- Slow for heavy numerical computation.

R

Pros

- Great for statistical analysis
- Thes best for statistical Visuals

Cons

- Limited general purpose programming.

MATLAB

Pros

- Good for numerical computation.
- Engineering Applications.
- Built in Ide

Cons

- Expensive
- Smaller ecosystem

Pros and Cons

Data Integration/Extraction

RESTful

Pros

- Scalability
- Flexibility
- Compatible with multiple data formats (JSON, XML, HTML, etc)

Cons

- Susceptible to over-fetching
- Requires multiple calls for related data

GraphQL

Pros

- Queries return exact data requested
- Can get multiple related fields from a single call
- Can be combined with different architectural styles

Cons

- Learning curve
- Error handling can be difficult

cURL

Pros

- Simple method for testing APIs and endpoints
- Available on most operating systems
- Bindings developed for most languages (Python, Java, R)

Cons

- Lacks functionality compared to GUI-based alternatives (E.g Postman)

Industry and Academic Relevance

Python-Broad Prog Lang used in many fields, from data science and machine learning to web and software developments to the average joe using it to do simple tasks. Relevant companies that use python are Google,Netflix,Spotify,Nasa, and many more all for backend, frontend,data analysis.

Matlab- Used primarily for engineers and scientists and in academia, applies to machine learning and data science due to its excellence in linear algebra calculations as well as visualization.

R-Used for data analysis, Machine learning,Data Science.Primarily used by scientists and engineers in big companies like Google,Meta,Airbnb, to retrieve data and statistics.

Industry and Academic Relevance

Rest Api-Used in Mobile and Web development to access and change data.. Used for google maps,uber, twitter, facebook,weather app.

GraphQL-Used to power mobile apps, websites,and Apis, makes them faster and more efficient. Used by companies: Shopify,Netflix,Airbnb, and its founder Meta.

Soap- Used to send and receive messages/data. Used for bank transfers,flight booking, and for billing services.

Curl-Used in the cyberSecurity field to ensure secure data as well as other big tech companies like Google,Apple,Microsoft, for Api interactions and Http requests.

What should you pick?

As you saw there are many tools and paths that you guys can choose in your Data Science Journey.

Python-A free high level language that has a fast learning process, can use Pandas and numpy to manipulate and analyze data.

MatLab-Used more for data Visualization, how does your data look,like python it has a fast learning process but it has an initial cost(Free for cuny/ccny) and further cost for certain libraries.

R-An open source programming language used for data analysis and machine learning. Available for Mac,Windows,Linux.Has a weird syntax compared to many programming languages.

What should you pick?

Restful api-A web service that allows communication between client and server.Can be used to retrieve large data sets from other systems.

GraphQL-A query language where developers can fetch data from multiple data sources with a single API call. Outputs only relevant information.

Soap- Used to exchange data between applications. Used more commonly in older systems.

Curl-Used data to and from a server and to make api requests.

END