INTRO TO PYSPARK

A BEGINNER'S GUIDE

BY: ANGELICA ("JELLY") SPRATLEY
AKA
LEARNING WITH JELLY

AGENDA

Overview of PySpark

Benefits of PySpark

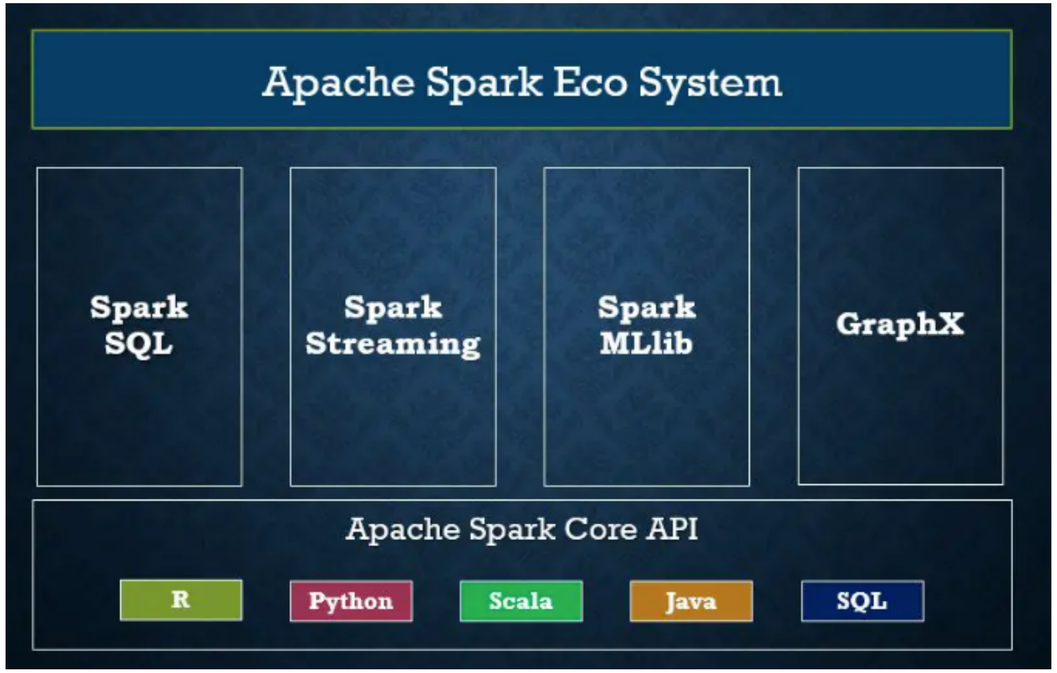
Real World Use Cases

Getting Started with PySpark

WHAT IS PYSPARK?

- PySpark is the Python Library of Apache Spark, an open-source distributed computing system

- Apache spark is a uniformed platform that executes data engineering, data science, and machine learning projects across many machines (clusters) -- HANDLES BIG DATA



COMPONENTS OF PYSPARK

- SparkContext - entry point object that connects to a Spark cluster (group of machines)

- SparkSession - entry point object that combines SparkContext, SQLContext, HiveContext to read data and execute queries

Both of these objects are created at the top of the notebook!

COMPONENTS OF PYSPARK

- Resilient Distributed Datasets (RDDs) - foundational data structure of Spark and is an immutable (can't change) collection of objects that can be processed in parallel
  - One Computer Goes Down, No Problem You Have Another Machine that can run the process (fault tolerance)
  - Tough to work with RDDs especially for data analysts! So Apache created...

COMPONENTS OF PYSPARK

- DataFrames - built on top of RDDs and organize data into rows and columns that we can change and perform tons of more tasks like data manipulation and modeling

and let's not forget...

- Spark SQL - a module (library) allows SQL code for data manipulation

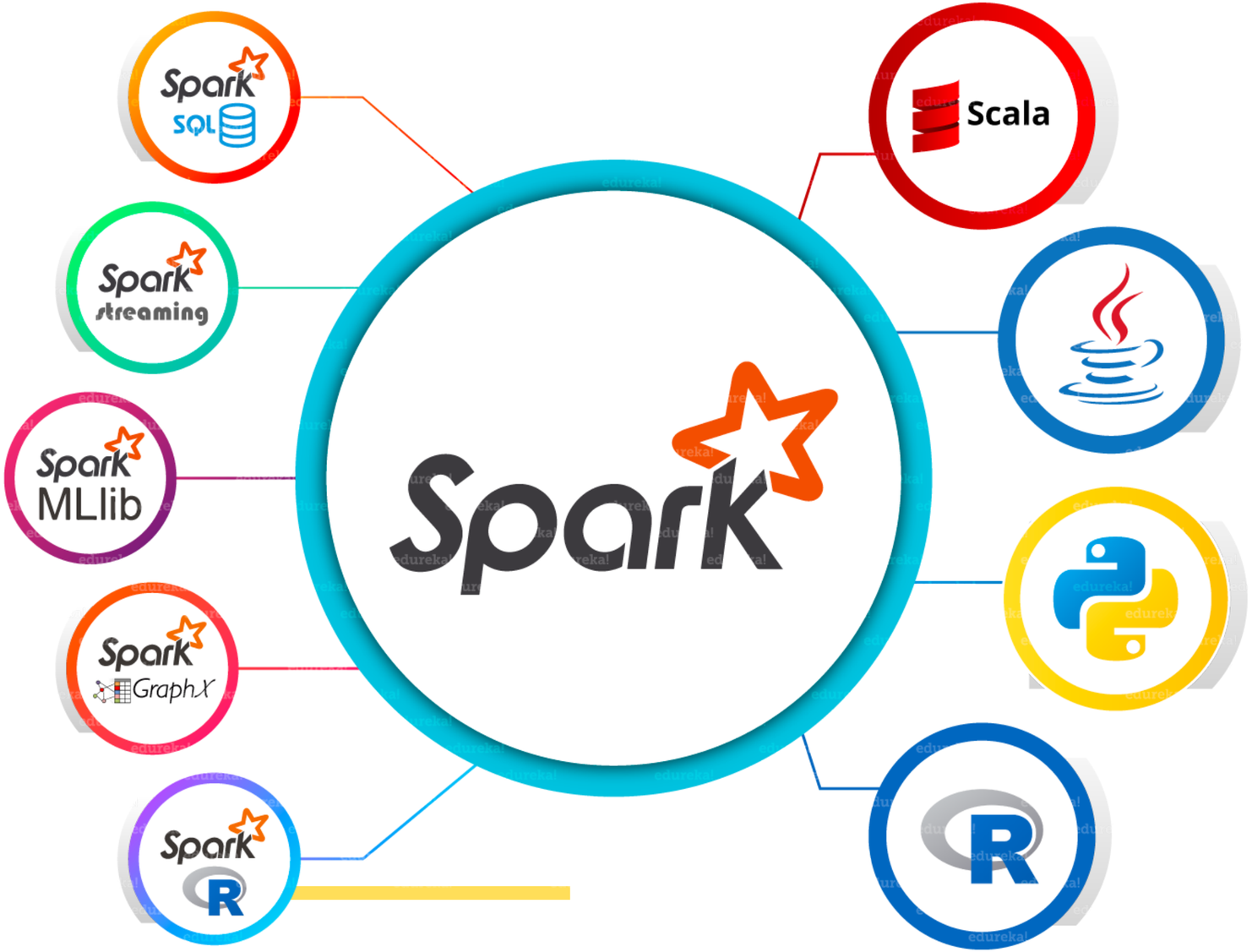
WHY USE PYSPARK?

SPEED - IN MEMORY PROCESSING

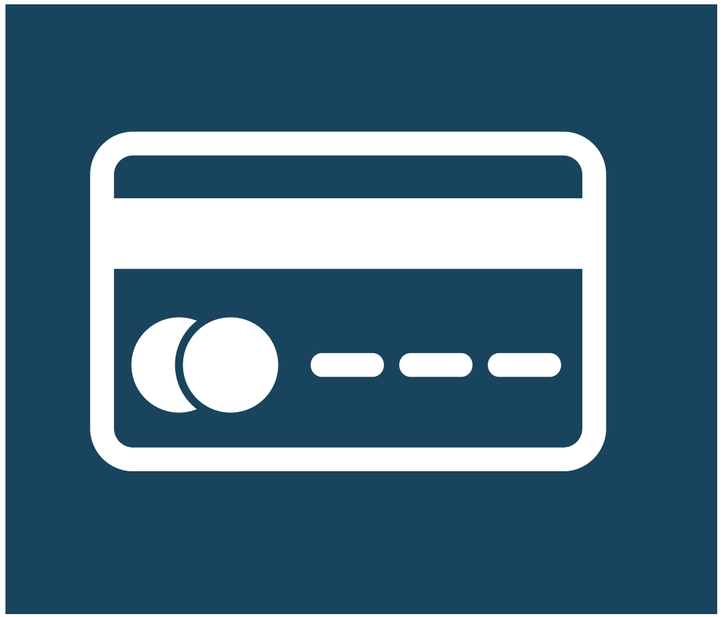
SCALABILITY - HANDLE BIG DATA

EASE OF USE - KNOW PYTHON YOU KNOW PYSPARK

PYSPARK ECOSYSTEM



REAL-WORLD USE CASES



CREDIT CARD FRAUD DETECTION

RECC SYSTEMS FOR ADS

GENOME DATA PROCESSING

GETTING STARTED

- I use GoogleColab in this repo for ease of starting a SparkContext and SparkSession
    - Start with the NYC Water Consumption EDA Notebook!

- You can also run PySpark on docker (view online documentation)

- View the original Apache Spark Site for more info!

THANK YOU