Developer Series

Join our webinar series to expand your developer skills!

Security Tuesdays
Data Science and AI Wednesdays
Cloud Native and Red Hat OpenShift Thursdays

meetup.com/IBM-Cloud-MEA

IBM **Developer**

IBM

Get started at: https://ibm.biz/BdfPQ5

Introduction to Big Data analysis & Machine Learning in Python with PySpark

Anam Mahmood
Developer Advocate, UAE

Hashim Noor
Client Technical Specialist, UAE

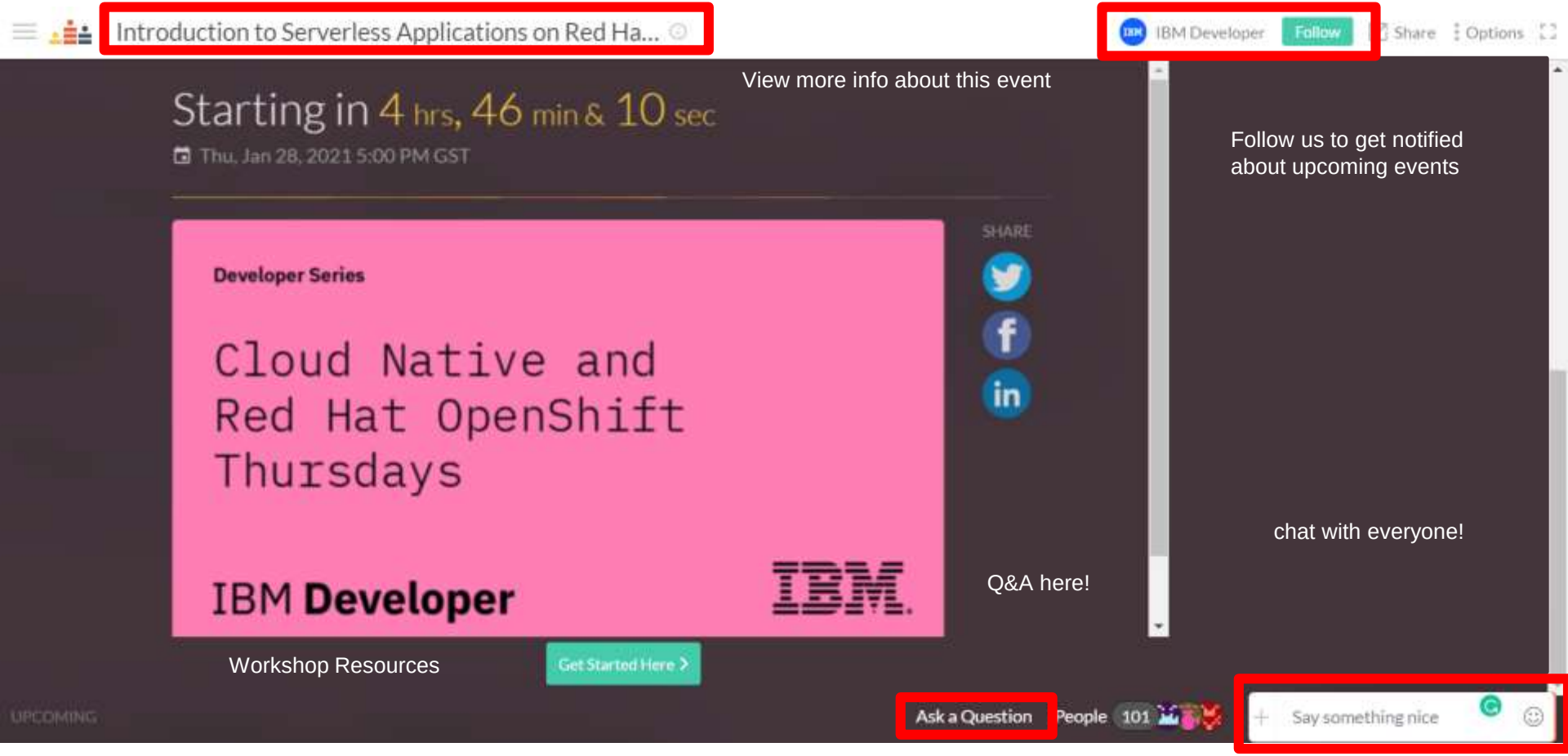
IBM Developer

IBM

Let's get started

- Sign up/Log in to your IBM Cloud Account https://ibm.biz/BdfPQ5
- Follow along for the hands-on: https://github.com/Anam-Mahmood/Introduction-to-Big-Data-analysis-Machine-Learning-in-Python-with-PySpark/blob/main/README.md

Get started at: https://ibm.biz/BdfPQ5



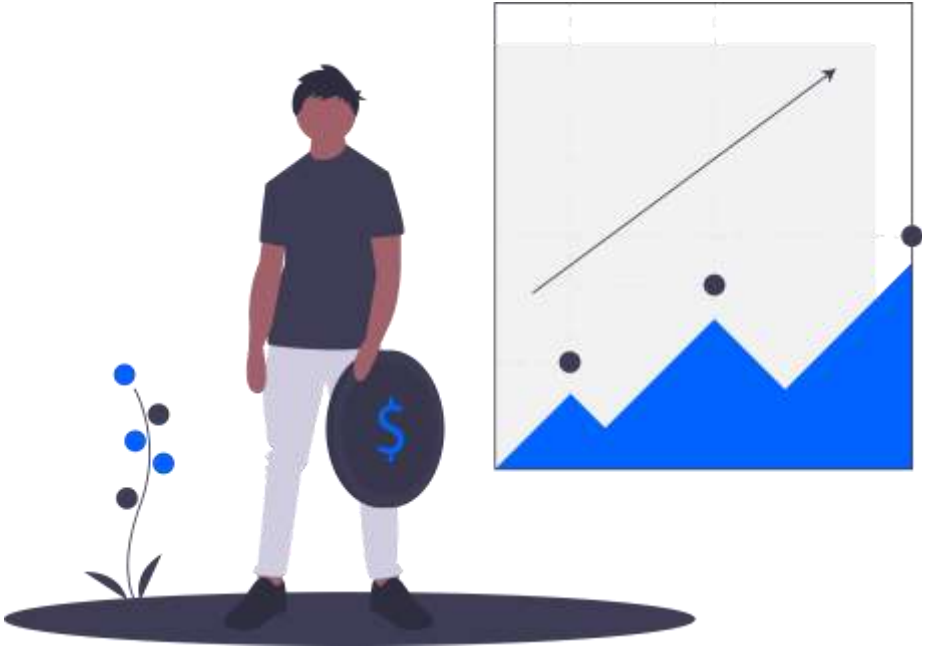
Survey

https://ibm.biz/BdfPQN

Agenda

"Data is the new oil. It's valuable, but if unrefined it cannot really be used."

-Clive Humby



Get started at: https://ibm.biz/BdfPQ5

The amount of **data** in the world was estimated **to be**

44

zettabytes

at the dawn of 2020.

https://hoteltechreport.com/news/big-data-examples

Big Data

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

https://towardsdatascience.com/what-is-big-data-lets-answer-this-question-933b94709caf

Get started at: https://ibm.biz/BdfPQ5

The 5 V's of Big Data

Velocity

Velocity is the speed at which data accumulates.

Volume

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

Variety

Variety is the diversity of the data. Variety also reflects that data comes from different sources

Veracity

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

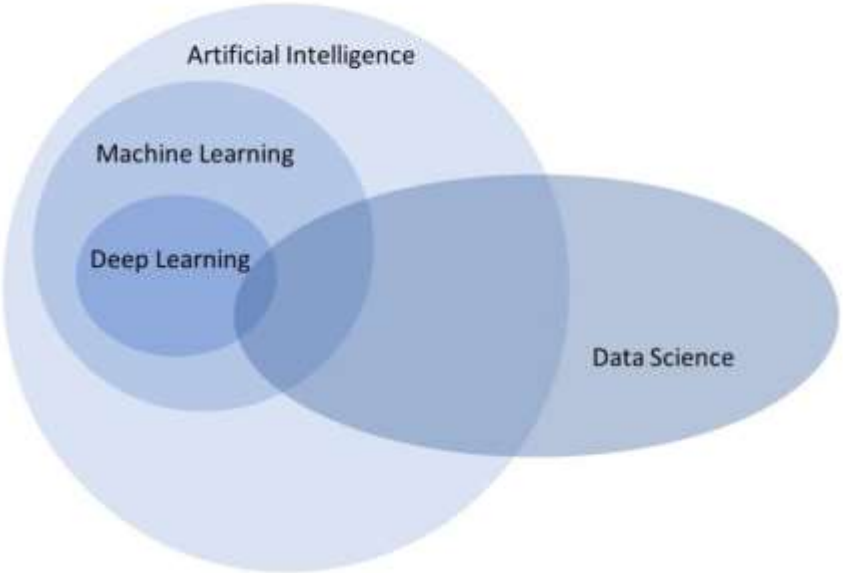
Value

Value is our ability and need to turn data into value. Value isn't just profit.

How big data is driving digital transformation?

What is Data Science?

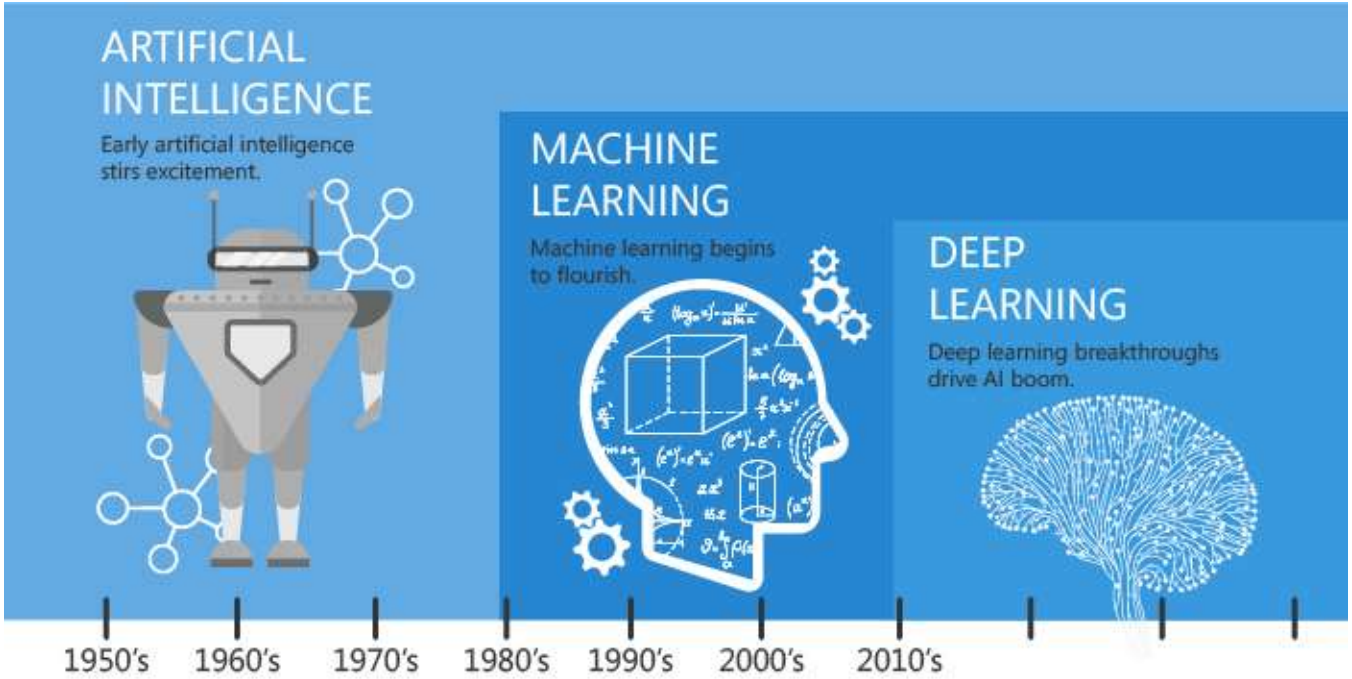
Data science is an interdisciplinary field leveraging insights from many fields to extract knowledge from data.



https://blog.finxter.com/artificial-intelligence-machine-learning-deep-learning-and-data-science-whats-the-difference/

Get started at: https://ibm.biz/BdfPQ5

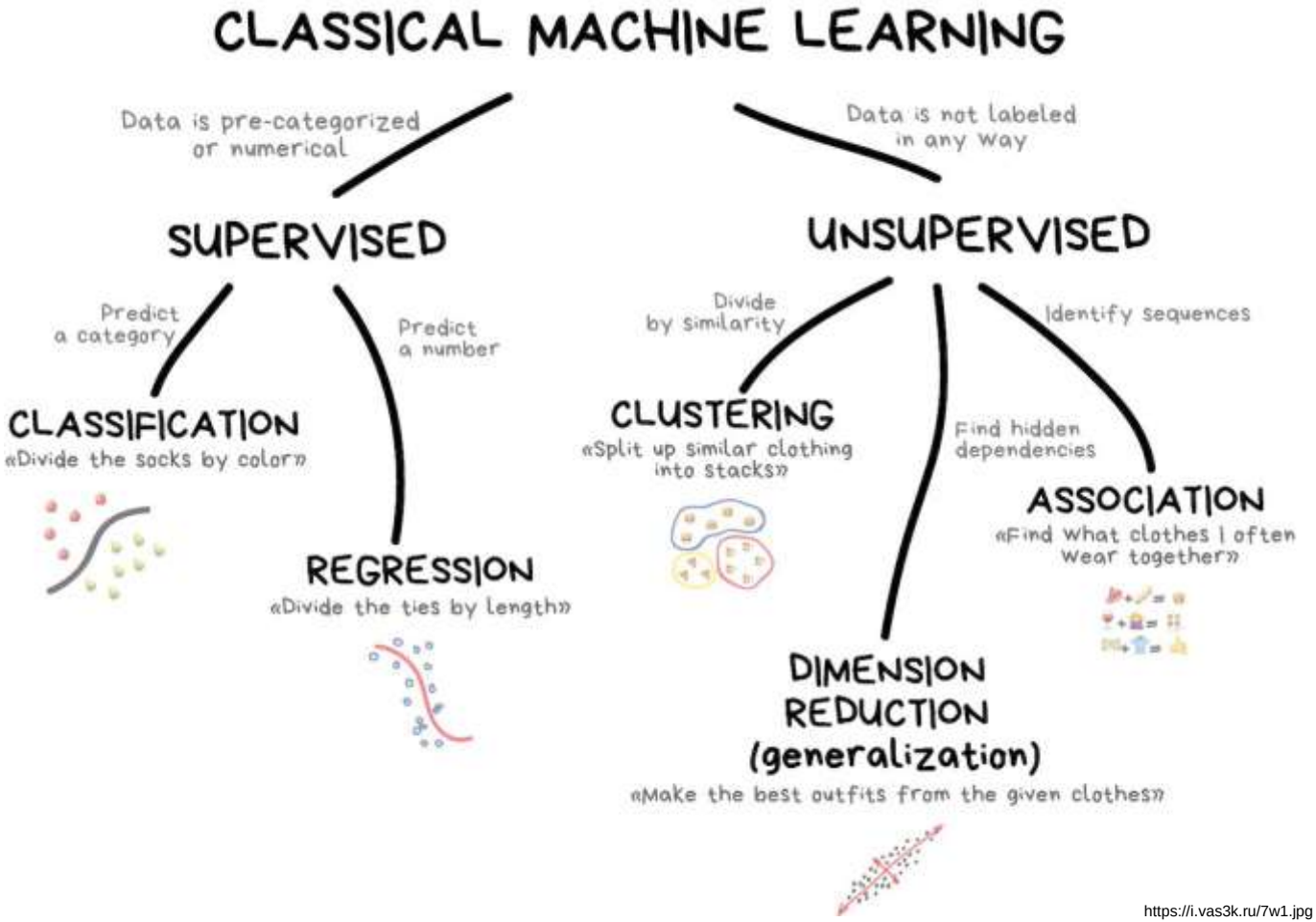
The Subsets of AI



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

https://blog.devitpl.com/learning-machine-learning/

Get started at: https://ibm.biz/BdfPQ5



https://i.vas3k.ru/7w1.jpg

Apache Spark

- Spark is an Apache project advertised as “lightning-fast cluster computing”.
- Spark provides a faster and more general data processing platform.
- Spark lets you run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop.
- Spark also makes it possible to write code more quickly as you have over 80 high-level operators at your disposal.

https://en.wikipedia.org/wiki/Apache_Spark

Get started at: https://ibm.biz/BdfPQ5

Components of Apache Spark

Apache Spark Core

It provides in-memory computing and referencing datasets in external storage systems.

Spark SQL

Spark SQL is Apache Spark's module for working with structured data.

Spark Streaming

This component allows Spark to process real-time streaming data. Data can be ingested from many sources like Kafka, etc.

MLlib

This library contains a wide array of machine learning algorithms- classification, regression, clustering, and collaborative filtering.

GraphX

Spark also comes with a library to manipulate graph databases and perform computations called GraphX.

Why Apache Spark?

- Python Pandas is intended for quick and easy data manipulation tasks.
- Pandas Dataframe help in:
  - Data Manipulation tasks such as sorting, merging data frame.
  - Modifying by updating, adding and deleting columns from a data frame.
  - Cleaning and data preparation by imputing missing data or NANs.
- **Pandas dataframe does not support parallelization**.

Features

- **Fast processing** – Big data is characterized by volume, variety, velocity, and veracity which needs to be processed at a higher speed.
- **Flexibility** – Apache Spark supports multiple languages and allows the developers to write applications in Java, Scala, R, or Python.
- **In-memory computing** – Spark stores the data in the RAM of servers which allows quick access and in turn accelerates the speed of analytics.
- **Real-time processing** – Spark can process real-time streaming data.
- **Better analytics** – Apache Spark consists of a rich set of SQL queries, machine learning algorithms, complex analytics, etc.

Use cases

E-Commerce Industry

Shopify wanted to analyse the kinds of products its customers were selling to identify eligible stores with which it can tie up - for a business partnership.

Healthcare

Many healthcare providers are using Apache Spark to analyse patient records along with past clinical data to identify which patients are likely to face health issues after being discharged from the clinic.

Media & Entertainment Industry

Apache Spark is used in the gaming industry to identify patterns from the real-time in-game events and respond.

Travel Industry

TripAdvisor, a leading travel website that helps users plan a perfect trip is using Apache Spark to speed up its personalized customer recommendations.

Use Case – Finance Industry

- Banks are using Spark to access and analyse the social media profiles, call recordings, complaint logs, emails, forum discussions and a lot more.

- **Architecture Flow**

1. User logs into Watson Studio, creates a project and initiates and instance of Cloud Object Storage and Notebook.
2. User uploads the data file in the CSV and text format to the object storage.
3. User creates a notebook from the URL provided.
4. Then enter the credentials of the dataset.
5. Run the notebook.

Hands-on

- Sign up/Log in to your IBM Cloud Account https://ibm.biz/BdfPQ5
- Follow along for the hands-on: https://developer.ibm.com/tutorials/getting-started-with-pyspark/

Summary

- Big Data is Dynamic, large and disparate volumes of data being created by people, tools, and machines.
- The 5 V's of Big Data, velocity, volume, variety, veracity and value.
- Data Science and the difference between AL, ML and Deep Learning.
- The five components of Apache spark and it's features.
- Use cases of Apache spark

Survey

https://ibm.biz/BdfPQN

Digital Developer Conference

Data & AI

Learn industry-recognized data and AI skills from IBM experts, partners, and the worldwide community

Register for free and get ready to build smart and secure data and AI solutions with a focus on experimentation, eminence, and education. The IBM Developer ecosystem of experts, enthusiasts, and client partners have created an experience to explore AI-centric solutions and platforms—dedicated to the worldwide community of developers, data scientists, and academia. You will also find regional and topical events where you can meet like-minded people—as well as a data science course with a badge.

> Conference begins
> June 8

Learn more and register for free at ibm.biz/devcon-ai.

On-demand replays available after the event

What is Call for Code?

Call for Code invites developers and problem solvers around the world to **build and contribute to sustainable Open-Source software solutions, that address social and humanitarian issues**, while ensuring solutions are deployed to make a **real difference**.

DCC DAVID CLARK CAUSE

Call for Code Creator

IBM

Call for Code Founding Partner

UNITED NATIONS HUMAN RIGHTS OFFICE OF THE HIGH COMMISSIONER

Call for Code Charitable Partner

THE LINUX FOUNDATION

Call for Code Program Affiliate

Call for Code has become the only global, **always-on** tech for good Open-Source platform to deploy & scale top projects through a host of offerings:

Why join?

- Skill Building
- Social Good
- Ideas to Action
- Community (400K+ developers, 179 nations, 15K+ applications)

Sponsors

ARROW

INGRAM

Awards

200K USD for Global Challenge winner
10K USD for University Challenge winner
5K USD for Middle East and Africa region winner

2021 Call for Code fighting CLIMATE CHANGE

Clean Water and Sanitation

Water is the natural resource that is most threatened by climate change and a prerequisite for life on earth. From intelligent solutions for small farmers to recycling showers, **technology** can make a significant impact on the **availability of water and its consumption**.

Zero Hunger

135 million people suffer from acute hunger, with climate change a major contributing factor. **Technology** can **help grow more crops in areas on the edge of drought** or quickly **distribute perishables** from small stores to local homeless shelters.

Responsible Production and Consumption

Worldwide consumption and production drives the global economy yet is inextricably linked to the environment. **Technology** can help **make recommendations on energy efficiency** to highlighting the carbon footprint of online purchases.

Resources and Events

Call for Code Main Page: ibm.biz/callforcode

FAQs: callforcode.org/faq/

Starter Kits: **Zero Hunger**, **Clean Water and Sanitation**, **Responsible production and green consumption**

Office hours **Every Monday**
from 2 PM to 3 PM GST (Dubai time)
crowdcast.io/e/mea_cfc_officehours

Join our MEA Call for Code Slack channel:
ibm.biz/mea_cfc_slack_channel

Get notified for upcoming Call for Code events by following our Meetup Page:
meetup.com/IBM-Cloud-MEA/

Go to our Crowdcast page for replays of previous events that could help you:
crowdcast.io/ibmdeveloper

Resources

IBM Developer: https://developer.ibm.com/

Meetup: https://www.meetup.com/IBM-Cloud-MEA/

Learning:

https://cognitiveclass.ai/

https://learn.ibm.com/

Big Data Fundamentals: https://cognitiveclass.ai/learn/big-data

Spark Fundamentals: https://cognitiveclass.ai/learn/spark

Thank you

Anam Mahmood
Developer Advocate, UAE
anam.mahmood@ibm.com

Hashim Noor
Client Technical Specialist, UAE
hashim.noor1@ibm.com

IBM