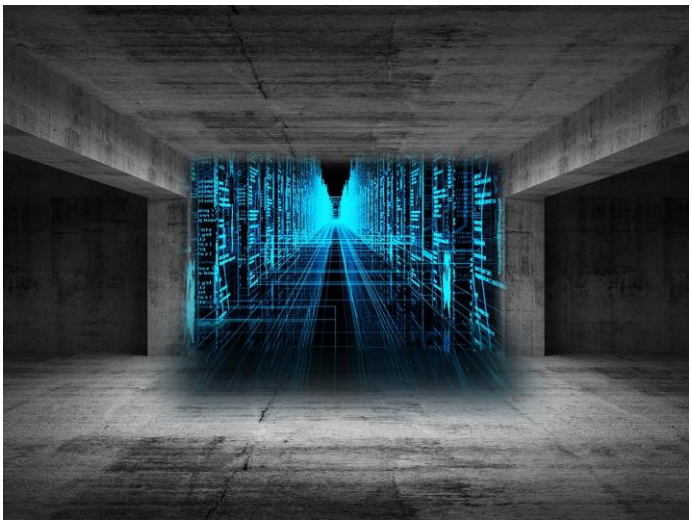

PGPM IN BIG DATA AND AI FOR BUSINESS AND MANAGEMENT

Includes a module on Designing LLM products

BY

FORE School of Management, New Delhi

Big Data Program

Table of Contents

About Program

PGPM Certificate Program in Big Data Analytics for Business and Management covers the following five distinct modules. Detailed content under each subject and module follows. We lay special and continued stress throughout the program on performing live projects on the part of students. Details about these are listed subsequently.

Module No	Big Data & AI Technologies **	Hours
1.1	**Machine Learning algorithms**	30
1.2	**Streaming Data Analytics**	30
1.3	**NoSQL and Graph Databases**	30
1.4	**Deep Learning and AI**	30
1.5	**Generative AI and Designing LLM Product**	15
	Total Hours	135**

** No of hours are indicative

Program Objectives

Applications of Big Data & AI transcend Industries. Use of predictive analytics pervades diverse disciplines such as marketing and sales, sports, molecular biology, drug-designing, waste management, finance, healthcare and the list is very long. Smart cities, for example, are the melting pot where variety of big data technologies mesh with one another to transform a city into a semi-intelligent being. In Marketing and Sales, for example, Big Data & AI are fast emerging as a potent tool to gain deeper insights into Customer behavior and thereby act as a strong driver in spurring innovation. In manufacturing, operations managers are employing advanced analytics on historical process data to identify patterns and relationships among discrete process steps and inputs, and then optimize the factors that prove to have the greatest effect on yield.

There are three principal segments to the program: One the Analytics part, second the technological part and third the contemporary Generative AI and Designing LLM products. The analytics part is about Machine Learning Algorithms and implementing them, the technological part is about learning to use

spark on Hadoop/NoSQL Databases and work on Streaming Data Analytics using Apache Kafka. As NoSQL databases are important big data storage systems, we cover them in our course. Deep Learning technology is applied in such fields as Healthcare, Personalized Marketing, Financial Fraud Detection, Facial Recognition, Recommendation Systems, Agriculture and others. Generative AI and Large Language Models (LLMs) have a wide range of applications across various industries. As for example:

a) Text Generation: Generating human-like text for content creation, storytelling, and dialogue; Summarizing long documents or conversations; Translating text between languages; Answering questions and providing informative responses.
b) Creative Applications: Generating images, music, and other media content; Assisting with ideation and brainstorming for creative projects; Designing 3D models and virtual environments.
c) Conversational AI: Powering chatbots and virtual assistants for customer service and support; Engaging in open-ended conversations and providing personalized responses; Automating repetitive tasks like scheduling and email management.
d) Analytical Applications: Analyzing large datasets to identify trends and insights; Generating reports, visualizations, and summaries from data; Providing recommendations and predictions based on data
e) Educational Applications: Tutoring and teaching by answering questions and providing explanations; Generating practice problems and grading student work; Providing personalized learning experiences
f) Research and Development: Accelerating scientific discoveries by generating hypotheses and experiments; Assisting with literature reviews and summarizing research papers; Aiding in the development of new products and services

The key difference between generative AI and LLMs is that generative AI encompasses a broader range of models that can generate diverse types of content, while LLMs are specialized in understanding and generating human-like text. The choice between the two depends on the specific application, available data, and resources.

At the end of this course, given a large dataset from any domain, a participant should:

1. Be able to clean, transform and visualize the dataset to gain deeper insights and make it ready for further analysis
2. Be able to engineer features and select a subset of appropriate machine learning or deep learning algorithms that could be applied to get the desired predictive results.
3. Make situation and context specific performance assessment as also interpret predictive models and explain them to clients
4. Apply the knowledge of image processing, image analysis, sensor-data analysis and language modeling to a wide array of disciplines such as health, process control, navigation and others.
5. Design/create knowledge-products using Large Language Models

Further:

6. Should be able to himself install, setup and configure and experiment with a complete hadoop and Kafka ecosystems (that include hive, spark, zookeeper, flume and other important layers).
7. Should be able to install, configure and be sufficiently familiar with the variety of NoSQL databases and decide for himself which one to use, when and how
8. Should be able to install fully functional various deep learning platforms including Tensorflow and PyTorch.
9. Should be able to install Web-User Interface(s) incorporating RAG for managing LLMs.

This course is project (lab) oriented: All tools, data and platforms including deep-learning platforms, hadoop-ecosystem, Kafka-streaming technologies and LLMs necessary for experimenting are either provided to the participants in advance or we help participants to install them on their laptops. There is a heavy emphasis on open-source technologies universally used almost throughout the industry. Each participant, at the beginning of the course, receives Virtual Machines (VMs) fully equipped with many software platforms, tools, packages and data to work on.

Our faculty have experience with several Industrial projects. Our students regularly execute projects on Kaggle—in fact, executing projects on Kaggle, creating a GitHub repository of projects and hosting LLM based web-interfaces on HuggingFace are a must and during the course several projects are implemented on all these platforms.

Who should attend

Data being ubiquitous, the program cuts-across job or academic profiles. The techniques taught are generic in nature. These will be valuable to anyone who wishes to analyse data to advance his/her knowledge. Specifically, the course will be useful to:

Executives
: Ambitious Executives (from Private/Public sectors) looking forward to sharpening their skills and making sense of data in order to innovate and add more value to their organization and to society.

Academicians
: Lecturers and Professors for extending the horizon of their knowledge through deepening their research skills.

Data Scientists/ Developers
: Techniques taught to them will have applications in a broad range of disciplines.

Healthcare professionals
: Healthcare professionals stand to immensely benefit from the extensive coverage of Deep Learning and LLM technologies and how these are applied in medical case studies.

Students/Research Scholars
: IInd year students currently enrolled in Engineering / MBBS / PGDM / MBA or any graduate or post-graduate program who have had an introductory course in statistics. These students can look forward to better placement opportunities with added skill set.

Eligibility

A graduate in any discipline or a student pursuing an Engineering/MBBS degree.

Module wise details

The complete course is divided into five distinct modules. Module 1.1 is about Machine Learning

Algorithms. In this module, we use variety of python based libraries. Module 1.2 is about Hadoop and Kafka eco-system: we learn to work on Hadoop and its layers; perform data extraction, build data pipelines as also push it into analytics engine. Analyzing streaming data is a major subject in its own right: in this respect we experiment with Apache Kafka and related technologies. Module 1.3 relates to NoSQL and Graph databases. The new millennium and the explosion of web content has marked a new era for database management systems. A whole generation of new specialized databases have emerged, all categorized under the name of NoSQL databases with focus on "task-oriented" database management system; selecting the right tool for the job depending upon its characteristics, nature and requirements. We cover, in depth, some often used NoSQL databases. Module 1.4 pertains to the exploding field of deep learning and AI. Deep Learning distinguishes itself from classical machine learning by its ability to work with unstructured data, like text and images, without the need for extensive pre-processing. Deep learning models, consisting of multiple layers of interconnected nodes, automatically learn and improve from data, enabling them to recognize patterns, classify phenomena, and make predictions with high accuracy. In this part we cover deep-learning technologies using very popular libraries tensorflow and PyTorch. Module 1.5 is about Generative AI. Generative AI refers to artificial intelligence systems that can generate new content such as text, images, audio, or video based on patterns learned from training data. Generative AI models learn the underlying structure and patterns in their training data, which could be large datasets of images, text, audio, or other modalities. We learn how these models work and perform projects based on Large Language Models.

Pedagogy

We strongly believe that a course in data analytics can only be practice-based rather than theory based. We also believe that a practice based course requires constant interaction with the teacher during lecture hours in real time. As it is a distance online course, the teaching pedagogy is like this: First the algorithm (or theory part) is conceptually explained without getting into mathematics and then a project is undertaken to implement the techniques. Datasets for implementation are made available in advance and so also a copy of code (or hints on it) that we need to execute. The code is numbered and copiously commented so that long after the lecture has finished, students can go back through the code/comments and refresh their knowledge. During the lecture, we execute this code (or prompt students to fill in the gaps), line-by-line and explain the steps. At his end, the student executes the required code on his laptop. Consequently, results are available at our end as also with the Students immediately. In short, both the teacher and students are working on their respective laptops simultaneously; students solve their problems and ask any questions to clarify. The whole experience is just as if everyone is sitting in a laboratory and working together.

Module 1.1: Machine Learning Algorithms Top

Machine Learning is about data cleaning, data transformation, feature engineering and predictive analytics. It is about knowledge discovery in datasets--structured or unstructured--searching through large volumes of raw data to find useful information-patterns. Data Scientists and decision makers can use this information for new sources of advantages and differentiation or for developing new business models. Broadly speaking the module objectives are three-fold:

- Generate familiarity with Big Data, Data Visualization and Data Mining algorithms: In generating this familiarity there is special emphasis on conceptual understanding of techniques rather than on mathematics. Analytics is a creative process and students are encouraged to be creative.
- Develop skills to set up predictive models with numerous types of disparate data sets. This is intended to bring home the point that predictive analytics offers a generic set of

tools that can be applied on different types of datasets, no matter what be the discipline or the Industry.

- Think differently: Expose students through projects as to novel ways of applying Big Data technologies among shifting business models.

Sr No	Subject**	Projects/ Datasets for projects***
1	Python: Data structures in Python, Pandas and Numpy.	UCI Repo: Bank Marketing Datasets
2	Data exploration, data summarization and transformation using pandas and numpy.	Practical using simple datasets Ta Feng Grocery Store dataset
3	Data Visualization: Data Visualization using Matplotlib, Seaborn and Plotly express. Developing relationships between mix of categorical and numerical features and plotting distributions	Kaggle: Insurance Dataset and also Advertising Dataset
4	Data Mining: Measures of Proximity; Cluster Analysis; Evaluation of Clusters: Cluster validation and Clustering Tendency; Curse of Dimensionality;	Clustering Census data KMeans clustering of colors in an image to reduce its size
5	Classification Analysis: Decision tree Induction: Project Objectives:) Classification using decision trees i) Using pandas and sklearn for modeling ii) Feature Engineering	Kaggle Project: Otto Group Product Classification
6	Neural Network: Project Objectives: 1. Neural Network/Deep learning models 2. Feature plotting 3. Missing values Plotting & imputation 4. Balancing unbalanced data 5. Learn to work with h2o 6. Understand and plot performance metrics (ROC/AUC) 7. Relative feature importance	Education Analytics: Dataset of schools and students in Andhra Pradesh, India
7	Random Forest and Regression Trees Project Objectives: 1. RandomForest classifier and ExtraTrees Classifier 2. Feature Engineering 3. Feature Selection 4. Feature importance from RandomForest modeling	Kaggle: airbnb-recruiting dataset
8	Techniques of Dimensionality Reduction: PCA, Random Projections and SVD (Singular Value Decomposition)	See under LightGBM and XGBoost
9	Evaluating Classification: ROC, AUC, Precision, Recall, Specificity, Sensitivity; kappa metric; Overfitting; Bias-variance trade-off; L1 & L2 regularization	Lecture
10	Gradient Boosting Technique for Machine Learning	Lecture
11	LightGBM: Light Gradient Boosting Machine: Project Objectives: 1. Learning to model with lightgbm 2. Singular value decomposition & PCA 3. Cross-validation in python 4. Parameters tuning using Bayesian optimization	Kaggle: Statoil/C-CORE Iceberg Classifier Challenge

Sr No	Subject**	Projects/ Datasets for projects***
12	eXtreme Gradient Boosting (XGBoost) Project Objectives: • Reading data from hard-disk in random chunks • Understanding & Using PCA • Pipelining with StandardScaler, PCA and Xgboost • Grid tuning of PCA and xgboost • Randomized search of parameters • Parameter search using Bayes optimization • Parameter search using Genetic algorithm • Feature importance	Kaggle Project: Discovering special particle using HiggsBoson dataset
13	Interpreting Machine Learning Models using Partial Dependence Plots and LIME Project Objectives: • Understanding individual predictions made by complex models using LIME • Learning to draw and interpret Partial Dependence plots • Plotting relative feature importance	Kaggle Project: Gender Recognition By Voice

* Extra classes, beyond normal schedule, may be held to fully make students at home with Python. These classes may be scheduled on weekdays after mutual consultation.
**Teaching sequence may alter somewhat depending upon feedback from students
***Datasets other than those mentioned here may also be introduced during classes to achieve better clarity. Datasets needed for Kaggle projects are to be downloaded from their site even though freely available; this is as per site requirements.

Module 1.2: Streaming data analytics: Hadoop, Spark and Kafka Eco Systems

Top

Hadoop based technologies are all closely interlinked. We build fairly sophisticated data pipelines as for example, streaming continuously generated invoice data into Kafka, connect Kafka to a Spark engine where Spark pulls additional static data from Hadoop, merges it with incoming stream from Kafka, performs aggregation sends the results back to Kafka as also to a local file system.

Streaming data analytics has a wide range of applications across various industries. Here are some of the most common use cases:

Log Analysis: Streaming analytics is commonly used for real-time analysis of log data from applications, servers, and IT systems. It enables quick identification of issues, anomalies, and security threats by processing logs as they are generated.

Fraud Detection: Financial institutions use streaming analytics to continuously monitor transactions and account activity. By analyzing data streams in real-time, they can quickly detect fraudulent behavior and take immediate action.

Cyber Security: Similar to fraud detection, streaming analytics helps security teams identify and respond to cyber threats by processing network traffic, security logs, and other data feeds in real-time.

Sensor Data: Organizations with large numbers of connected devices and sensors, such as in manufacturing, energy, and transportation, use streaming analytics to monitor equipment health, optimize operations, and predict maintenance needs.

Online Advertising: Advertising platforms leverage streaming analytics to track user behavior, clicks, and interests in real-time. This enables them to serve targeted ads and offers to users as they browse and interact with content.

Retail and E-commerce: Retailers use streaming analytics to analyze clickstreams, transactions, and

inventory data. This allows them to optimize pricing, promotions, and the customer experience based on up-to-the-minute insights.

S No	Subject**	Projects/Datasets for projects***
1	Introduction to Hadoop and its ecosystem Hadoop file storage formats	Lecture & Tutorial
2	Linux and Hadoop shell commands	Tutorial
3	Hadoop streaming	Using awk & sed to execute map-reduce jobs to process data in Hadoop.
4	Spark: 1. Machine Learning 2. Structured Streaming 3. Deep Learning 4. Building data pipelines with Hadoop, kafka and NoSQL databases 5. Learning Koalas 6. Spark Delta Lake 7. Using Spark NLP	Machine Learning with Spark: Spark basics—Expts with RDDs, DataFrames, Datasets and SparkSQL; Executing ML algorithms by creating end-to-end processing pipeline. Using MLlib and ML libraries Deep Learning with Spark; Image classification using Transfer Learning. Applying deep learning models at scale. Structured streaming with Spark; Streaming and aggregating both in batch mode and in continuous mode Experimenting with Koalas and Delta Lake. Building Data pipelines for streaming data Multiclass text-classification with Spark NLP
5	Apache Kafka: Building Data pipelines; transforming streaming data Apache Flink: Simple experiments with Apache Flink—Streaming analytics	Experimenting with Apache Kafka system; Streaming from files both as producer and also as consumers. Building data pipelines with log-data-generator->Kafka->Spark aggregation->Kafka; Invoice generator->Kafka->Flume->HDFS. Using Confluent and stream processing with KSQL.

**Teaching sequence may alter somewhat depending upon feedback from students

Module 1.3: NoSQL and Graph Databases

Top

NoSQL databases are a type of database management system designed to handle large volumes of unstructured and semi-structured data. Unlike traditional relational databases that use tables with pre-defined schemas, NoSQL databases use flexible data models that can adapt to changes in data structures and scale horizontally to manage growing data. NoSQL databases are classified into four main categories: document databases, key-value stores, column-family stores, and graph databases. Graph databases, specifically, are designed to handle data with well-represented relationships, making them suitable for scenarios like social relations, network topologies, and more.

Graph databases store data as nodes and edges, representing elements connected by relations. They are ideal for scenarios where complex relationships between data need to be managed efficiently. These databases are used in various applications such as social media analytics, public transport links, road maps, and network topologies.

S No	Subject	Brief topics covered
1	Introduction to NoSQL Databases and CAP theorem; Comparison with RDBMS	Lecture
2	Redis in-memory data structure store	Installation of redis. Data structures in redis; Multi key queries; Publication and Subscriptions. Use cases
3	MongoDB Document Database	Installation of mongoDB; CRUD operations; Querying & Aggregation pipelines; Fulltext search; Application Design; Replication; Sharding operations; Server Administration & Data Administration; Transactions execution Building Data pipelines with Spark and Flume. Collaborative filtering using Spark over mongoDB as backend database; Building Machine learning models using Spark with mongoDB as backend.
4	Hbase column family database on hadoop	Gaps in Hadoop and need for Google's BigTable

		Working with column family databases; Expts using hbase shell Developing SQL databases over hbase using Hive. Creating data-pipeline using flume and transferring tweets from twitter to hbase
5	TIG Stack: telegraf, InfluxDB and Grafana for collecting, storing and visualizing Time Series or IOT Data/metrics on a Dashboard	Installing & using TIG; Configuring telegraf and Influxdb; Using InfluxDB for IOT data; Visualizing metrics in grafana dashboard Developing dashboards in grafana from scratch for log data. Developing dashboards in Grafana from scratch for streaming Invoice data.
Social Network Analysis		
6	Gephi Open Graph Visualization Platform	Discovering structures in networks; Connectedness, Node importance and communities in a network; Analyzing Facebook data to discover clusters and network of friends.
7	Neo4j Graph Database	Installation of neo4j. Use cases. Graph database concepts; Creating graphs and querying graph databases; importing and modeling a relational database into neo4j

Module 1.4: Deep learning & AI**

Top

Deep learning has a wide range of applications across various industries. Here are some of the most common deep learning applications:

Computer Vision: Deep learning is widely used in computer vision tasks such as image classification, object detection, facial recognition, and image segmentation. Deep learning models like convolutional neural networks (CNNs) can analyze visual data and make predictions with high accuracy.

Natural Language Processing (NLP): Deep learning is transforming NLP by enabling machines to understand, interpret and generate human language. Applications include machine translation, text generation, sentiment analysis, and chatbots.

Healthcare: Deep learning is making significant advancements in healthcare by improving disease diagnosis, drug discovery, and personalized medicine. It can analyze medical images, electronic health records, and genomic data to identify patterns and make predictions.

Finance: Deep learning is being used in finance for fraud detection, risk management, customer relationship management, and investment modeling. It can analyze large financial datasets to identify patterns and make predictions.

Autonomous Vehicles: Deep learning is a key component of autonomous vehicles, enabling them to perceive their surroundings, make decisions, and navigate safely. Deep learning models are used for object detection, lane detection, traffic sign recognition, and sensor fusion.

Cybersecurity: Deep learning is being used in cybersecurity for malware detection, intrusion detection, and fraud prevention. It can analyze network traffic and user behavior to identify anomalies and potential threats.

Among the software we will be experimenting with are TensorFlow 2.0, Keras, TensorBoard, fast.ai, YOLO, NLTK & spaCy.

S No	Subject	Projects or brief topics covered
1	Autoencoders and anomaly detection	Recognizing similar Olivetti faces; Fraud detection
2	Deep Learning with Convolution Neural Network	Image classification: Differentiating between Rural/Urban images
3	Using very Deep Convolution networks and Data Augmentation	Image augmentation: Building Powerful Image classifiers with very less images. Experimenting with VGG16
4	Transfer Learning-I	ResNet50: Working with ResNet50: Kaggle Invasive Species Prediction using multiple images

5	Transfer Learning-II	InceptionV3: Working with InceptionV3: Kaggle project Classifying Chest X-Ray images infected with Pneumonia.
6	Natural Language Processing-I	Word2Vec transformation techniques—Experimenting with arithmetic properties of Word Embeddings: King-man equals Queen.
7	Natural Language Processing-II Recurrent Neural Networks--I	Understanding LSTM, GRUs and Bidirectional LSTM usages. Sequence classification and Sentiment analysis of tweets on Twitter
8	Natural Language Processing-III	Encoder-Decoder Network models: Speech Translation: German-English sentences translation

** Language will be python

Module 1.5: Generative AI and Designing LLM Products

Generative AI and Large Language Models (LLMs) are integral components of the AI landscape, each with distinct roles and capabilities. Generative AI encompasses a range of tools that leverage information from LLMs and other AI models to create new content through machine learning. On the other hand, LLMs are a specific type of AI model that utilizes machine learning with billions of parameters to understand and generate text.

Generative AI, including LLMs, is poised to revolutionize various industries by enabling tasks like 3D modeling, video generation, voice assistants, and more. LLMs, such as ChatGPT and Google's Bard, have gained prominence due to their text generation capabilities, with models like GPT-4 boasting over 175 billion parameters.

In practical scenarios, LLMs play a crucial role in tasks like summarizing data for case workers, creating synthetic audience personas for marketers, and analyzing trends for analysts. Generative AI, including LLMs, offers a broad spectrum of applications beyond text generation, extending to image creation, video processing, and more.

Overall, Generative AI and LLMs are interconnected, with Generative AI serving as a comprehensive category encompassing various AI models like LLMs. These technologies are reshaping industries and workflows, offering innovative solutions for content creation, data analysis, and more.

S No	Subject	Projects or brief topics covered
1	General Architecture of Transformers	HuggingFace Transformer models. HuggingFace Encoders and Decoders videos Sentiment analysis using transformers.
2	Zero-shot classification and Few-shot learning	Zero-shot image classification. Few-shot classification example
3	Streamlit for developing LLM webApps	Building powerful generative AI apps. Hosting streamlit webapp in streamlit spaces
4	Ollama and anythingLLM installation	About ollama. Students install fully functional, production oriented, totally private, secure and feature rich chatbot.
5	Embedding, vector databases and search	FAISS: library for efficient search; chroma vector database
6	Prompt Engineering	LLM prompting guide; AI Prompt Engineering isn't the Future (HBR)
7	Developing LLM applications using langchain	Getting started with Langchain; langchain and ollama; pdf chatbots with langchain and ollama

8	Biased LLMs and Ethics	Students experiment with evaluating how ethical LLM models are and how to get over any biases.

5

Top

Students Exercises/Projects

Introduction:

The ultimate beneficiary of this program are students. Our experience shows that students learn faster, if they attempt exercises/projects, make mistakes and learn from them. Students are, therefore, expected to undertake exercises and projects.

Exercises serve another purpose. We try to make students self-learn some of the important topics not possible to cover in the class. We give exercises with sufficient hints to attempt them.

There is also a third advantage. The more students perform exercises, the more a teacher can move faster and also cover easily advanced concepts.

To assist students, for each project we provide sufficient steps/code on our e-learning site. For each project, our steps/codes are quite detailed and students should be able to execute the projects on their own by following the listed steps (or at times by stealing a glance at our code).

Students will be assessed based upon their performance in Exercises and Projects.

****