Faysal Aziz

Industrial Big Data Analytics

with Python and Agentic AI to build production grade data and AI systems using Python LLM APIs RAG pipelines and autonomous agents - 3 DAYS

FA-0248 · 3 DAYS

Agentic & Generative AI · Software Development · Data & Analytics

---

Introduction

Modern organizations do not fail because they lack data they fail because they cannot convert overwhelming volumes of data into timely reliable decisions. The difference between average analytics teams and elite ones lies in their ability to scale computation automate reasoning and integrate intelligence directly into data workflows.

This course is designed for professionals who want to operate at that level. It focuses on how big data is actually handled in production environments using Python distributed computation frameworks and modern AI capabilities. You will not be exposed to academic abstractions or toy datasets. Instead the material reflects how experienced engineers and analysts build systems that survive real workloads real deadlines and real business pressure.

The instructor brings over 30 years of industry experience and teaches only techniques that are currently demanded in enterprise and AI driven organizations. Alongside traditional big data analysis you will learn how to integrate large language models autonomous agents and Retrieval Augmented Generation systems to create analytics platforms that reason act and explain at scale.

Learning outcomes

By the end of this course participants will be able to:

- Design and implement scalable big data pipelines using Python.
- Analyze and transform massive datasets using parallel and distributed computation.

- Apply statistical and machine learning techniques to large scale data.
- Integrate LLM APIs into analytics workflows for intelligent summarization and reasoning.
- Build Retrieval Augmented Generation systems that ground AI output in enterprise data.
- Design agentic AI systems that autonomously plan retrieve analyze and synthesize data.
- Optimize performance reliability and scalability in production analytics systems.

Prerequisites

Participants should have:

- Strong Python programming fundamentals.
- Understanding of basic statistics and data analysis concepts.
- Familiarity with APIs JSON and HTTP.
- Comfort working in a command line environment.

Training outline

1. Foundations of Big Data Analytics with Python

1. Role of Python in modern data platforms
2. Characteristics of big data volume velocity variety
3. Memory bound vs compute bound analytics
4. Data locality and execution models

2. Python Data Processing Core

1. Python execution model and performance considerations
2. NumPy internals and vectorized computation
3. Pandas architecture and scaling limitations
4. Memory management and object overhead

3. Parallel and Distributed Data Processing

1. Parallelism concepts threads processes clusters

   Dask architecture and scheduling

   1. Dask arrays dataframes delayed objects

2. Lazy execution graphs
3. Scaling Pandas workflows

PySpark architecture

1. Driver executors and cluster managers
2. RDDs vs DataFrames vs Spark SQL
3. Serialization and shuffle behavior
4. Optimization with Catalyst and Tungsten

4. Data Ingestion and Storage at Scale

Data formats and tradeoffs

1. CSV JSON Parquet ORC Avro
2. Columnar storage and compression
3. Schema evolution and enforcement
4. Reading and writing large datasets efficiently
5. Partitioning strategies
6. Batch ingestion vs streaming ingestion

5. Data Cleaning and Transformation Pipelines

1. Large scale data validation
2. Handling missing corrupt and inconsistent data
3. Transformations with distributed execution
4. Reproducible transformation pipelines

6. Exploratory Data Analysis for Massive Datasets

1. Sampling strategies for exploration
2. Approximate statistics and sketches
3. Aggregations at scale
4. Group based analysis
5. Time based and windowed analysis

7. Visualization and Insight Communication

1. Visualization limits with large datasets
2. Aggregation driven visualization strategies

3. Interactive analytics dashboards
4. Narrative driven analytics outputs

8. Machine Learning with Big Data

1. Limitations of single node machine learning
2. Distributed feature engineering
3. Model training with large datasets
4. Model evaluation at scale
5. Data leakage and bias considerations

9. Introduction to AI in Analytics Systems

1. Role of AI in modern analytics platforms
2. Large Language Models overview
3. Strengths and limitations of LLMs for data tasks
4. Cost latency and reliability considerations

10. LLM API Integration with Python

1. API based model access patterns
2. Prompt structure for analytical tasks
3. Structured data to language workflows
4. Output parsing and validation

11. Retrieval Augmented Generation RAG Systems

1. Motivation for RAG in analytics
2. Separation of knowledge retrieval and generation
3. Embedding models and vector representations
4. Vector indexing and similarity search
5. Context assembly strategies
6. Grounded generation for analytics explanations
7. Reducing hallucinations with retrieval constraints

12. Enterprise RAG Architecture

1. Data source ingestion for RAG

2. Incremental indexing strategies
3. Query routing and retrieval logic
4. Response validation and traceability
5. Security and access control in RAG systems

13. Agentic AI Concepts

1. Definition of agentic AI systems
2. Difference between single shot LLM calls and agents

   Agent components

   1. Memory
   2. Planning
   3. Tool use
   4. Feedback loops
4. Deterministic vs autonomous execution

14. Agentic AI for Data Analytics

1. Agents as analytic orchestrators
2. Decomposing analytics questions into sub tasks
3. Dynamic data retrieval and transformation
4. Multi step reasoning over datasets
5. Decision making under uncertainty

15. Agentic RAG Systems

1. Combining agent planning with retrieval pipelines
2. Autonomous query reformulation
3. Iterative retrieval and refinement
4. Cross dataset reasoning
5. Explainability and auditability of agent decisions

16. Performance Optimization and Cost Control

1. Profiling Python and distributed workloads
2. Memory optimization techniques
3. Caching and persistence strategies

4. Managing compute cost in large scale analytics
5. Optimizing LLM API usage

17. Production Deployment of Analytics and AI Systems

1. Environment management and reproducibility
2. Containerization concepts
3. Scheduling and orchestration
4. Monitoring and logging analytics pipelines
5. Failure handling and recovery

18. Capstone Integration Project

1. End to end big data analytics system design
2. Distributed processing pipeline
3. LLM powered analytics layer
4. RAG based insight generation
5. Agentic orchestration of analytics workflows
6. Technical review and optimization

© 2026 Faysal Aziz · Kuala Lumpur