Proposal for a New Course

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

Course number: IS302
Course title: Fundamentals of Data Engineering – Part II (Ops)
Course prerequisites: ESC202
Course credits: [12] (2-0-6-0)
Course duration: Full semester
Course type: DC
Proposing instructors: Nisheeth Srivastava, Sruti Srinivasa Ragavan
Other faculty members interested in teaching the course:
Other departments interested in the proposed course: None

Course description:
a. *Objectives*: Creation of online-accessible, data-driven services is a cornerstone of AI/ML applications. This course will introduce students to the tools and best practices in creation, operation and maintenance of such online services at massive scales. The course will also offer hands-on experience in deploying such services on cloud platforms.
b. *Logistics*: The course will serve as a DC for IS UG students (BT, double major, digital IS minor), and as a DE for PG students (MT, MS, PhD) of the CSE and IS departments and CSE UG students (BT, double major, ML minor). The course may be offered one or more times every academic year depending on demand and availability of resources.
c. *Content*: There will be an equivalent of 24 lectures of 50 minutes each and 24 labs of 3 hours each. A weekly breakup of lecture and lab content is given below
d. *Evaluation*: Evaluation will use a combination of graded lab exercises, lab exams, take- home assignments and projects, and traditional sit-down quizzes and exams.

Weekly breakup of content: Numbers in square brackets [] against each topic indicate the number of lectures/labs for that topic.

Lecture content (24 lectures):
1. DevOps: (automate, because laziness is a virtue!) [6]
    a. Version control, CI pipelines, deployment, dev/test/prod environments
    b. Scripting, UNIX tools, APIs, client-server programming, event-driven programming
    c. Packages, package managers, dependencies, virtual environments, builds, bundling
    d. Monitoring code health: tests, code quality, logs
2. TechOps: [4]
    a. Basics of infrastructure: machines (physical / virtual), reading specs (processor speeds, CPU vs. GPU, storage). Volume, velocity, variety.
    b. Virtualization, containerization

c. Cloud, on-demand provisioning, demand forecasting, load-balancing
d. Monitoring and perf tests

3. MLOps: [6]
   a. Data formats: relational (CSV, Excel, SQL), key-value (JSON, Mongo) formats, structured data (graph, time series, point-cloud)
   b. Data storage: online services (AWS, Azure), notions of hot storage, cold storage
   c. Accessing data sources: use of popular libraries such as pandas, openpyxl, sqlite3 to access data sources.
   d. Use of popular cloud services such as S3, Sagemaker to keep up with service demands
   e. Notions of pipelines: data pipelines, ML pipelines
4. ScaleOps [6]
   a. Data storage at scale: indexing, sharding and pagination
   b. Data processing at scale: introduction to Hadoop and Spark
   c. Computing at scale: autoscaling and containerization
   d. Performance at scale: edge deployments, A/B testing, traffic splitting
   e. Monitoring at scale: MLFlow, RoboFlow
   f. Other emergent topics

Lab content (24 labs):

1. Fun with Unix, how to live without your mouse
2. Experimentation with version control, CI workflow scripting
3. Unit testing, functional testing, code quality testing
4. Building packages and installers
5. Mining logs
6. Simple web-app with API calls
7. Run CPU/GPU performance benchmarks on physical PCs and VMs with the same tech specs, Measure disk I/O, network latency, and memory speeds.
8. Set up a simple web application using Docker; Deploy it using Kubernetes; Experiment with resource allocation and auto-scaling
9. Set up a web server on AWS; Implement a load balancer and simulate traffic spikes; Use basic demand forecasting to optimize resource usage
10. Create a simple full-stack application plugging multiple sensors to a real-time dashboard
11. Work with CSV, Excel, JSON, and SQL databases; convert data between different formats; use Pandas, OpenPyXL, and SQLite3 for data manipulation.
12. Store sample datasets in hot storage (S3) and cold storage (Glacier); compare retrieval speeds and costs; automate data migration between storage types.
13. Extract data from multiple sources (APIs, databases, files); transform and clean data using Pandas; load it into a structured format (SQL, MongoDB).
14. Train an ML model using Amazon Sagemaker; automate model versioning and deployment; monitor performance and costs.
15. Use real-world datasets like stock prices or weather data; store and query time-series data using InfluxDB or TimescaleDB; train a forecasting model (ARIMA, LSTMs).

16. Implement a workflow using MLFlow or RoboFlow; automate data ingestion, preprocessing, model training, and deployment; include monitoring and failure recovery mechanisms.
17. Create a data archive for long-term cold storage, an app for intelligent pre-retrieval of relevant data into hot storage, and a dashboard consuming simple ML pipelines through S3+Sagemaker (3 labs)
18. Spin up a kubernetes cluster with multiple docker containers, with sets of containers handling data pre-processing, system monitoring, model training and prediction caching activities. We will provide artificial data in bursts. (3 labs)

Lab equipment: PCs with internet connectivity, appropriate software (browsers, Python runtime with libraries), an appropriate number of Azure, AWS, S3 licenses will be needed.

Short summary for inclusion in the Courses of Study booklet: this course will introduce tools
and techniques to build, operate and maintain scalable online data-driven, AI/ML-based services.

Textbook: There will be no textbook for this course.

Course proposer: Nisheeth Srivastava, Sruti Srinivasa Ragavan **Convener DPGC:**
Date: **Date:**

The course is approved/not approved

Chairperson, SPGC
Date: