GRADUATE CERTIFICATE IN BIG DATA ANALYTICS - PRACTICE MODULE

JAN 2024

LIU FAN

OVER **6,873** GRADUATE PROGRAMME ALUMNI

OFFERING OVER **150** ENTERPRISE IT, INNOVATION & LEADERSHIP PROGRAMMES

TRAINING OVER **157,000** DIGITAL LEADERS & PROFESSIONALS

- Objectives
  - To brief the participants of the graduate certificate on the requirements, conduct and assessment of the practice module
- Topics
  - Requirements of Graduate Certificate
  - Group Sizing and Effort Estimates
  - Requirements for Practice Project
  - Implementation Criteria for Group Project
  - Sample Ideas
  - Assessment of Graduate Certificate
  - Proposal for Group Project
  - Project Demonstration and Report

Requirements of Graduate Certificate

- Demonstrate **competency** in all *three* **course modules**

Module	Objective
Big Data Engineering for Analytics(BEAD)	Ingest from multiple data sources, design the right storage, and implement scalable data processing pipelines using in-memory compute frameworks such as Spark.
Recommender System (RCS)	Apply data analytics concepts and techniques to build recommender systems.
Processing Big Data for Analytics (PBDA)	Understand structuring big data projects for analytics, technical aspects of the project, model development, aggregation and monitoring.

- Demonstrate **understanding and mastery** of the skills taught in the above course modules in a **group project**
- Pass a **written examination** based on the scope of the above course modules

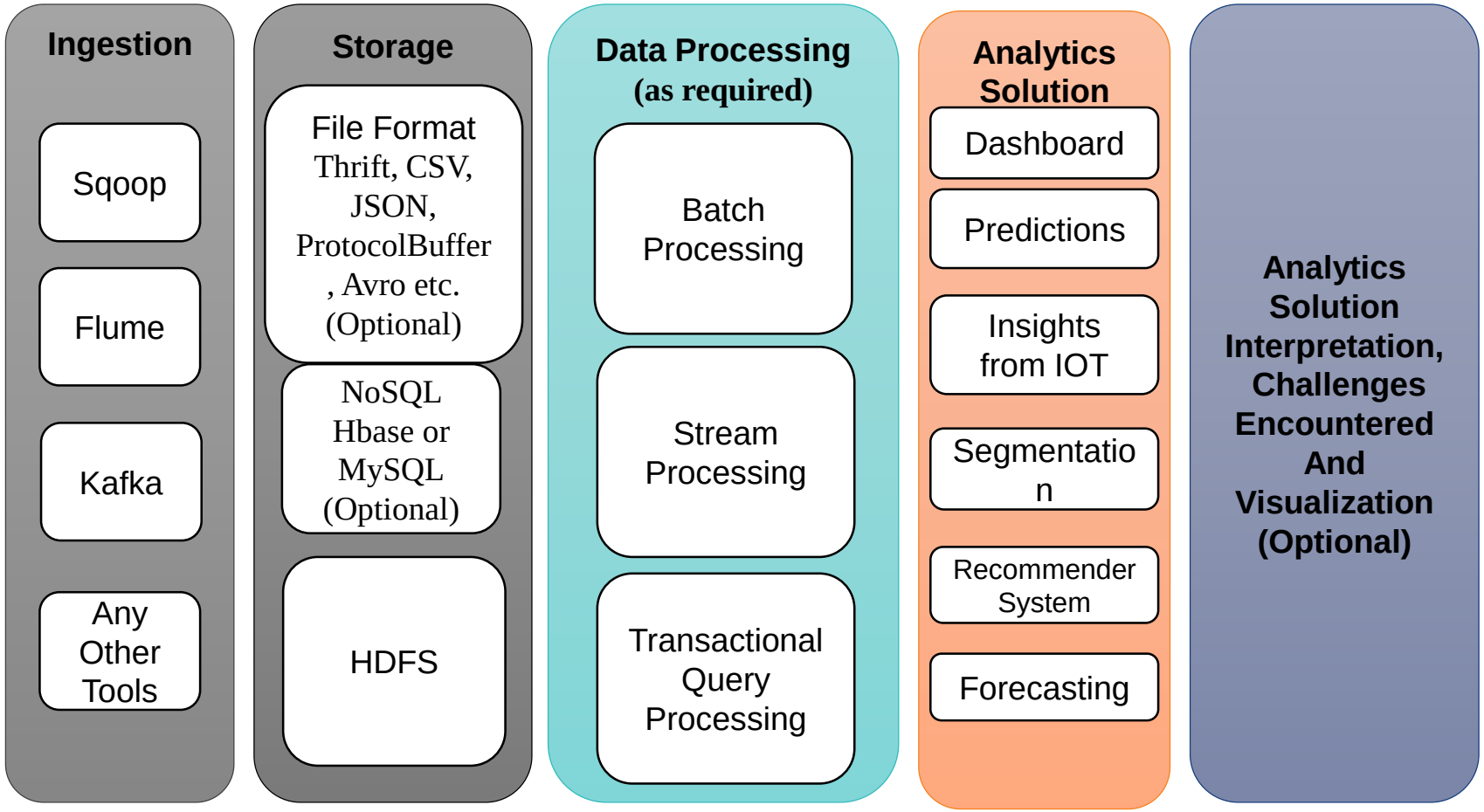
Group Sizing and Effort Estimates

- The recommended group size is **Four** or **Five**.
- Each member in the group must spend close to 10 man days of effort for sizing and complexity estimates of the project
- Groups must demonstrate their ability to apply and practice the knowledge, techniques and skills they have learnt from this certificate courses.
  - Each group would concentrate on one specific business problem that can ONLY be solved by the big data solution stack.
  - Propose and implement an existing analytics business case and implement a big data engineering solution.
  - The workshop technical environment may be taken as the basis.
  - Groups may also wish to extend the eco-system as required by the nature of assignment. Credits would be awarded for any additional efforts.

Requirements for Practice Project

Objective: The chosen project must manage processing of scalable data sets in terms of the following stages:

1. Selection of an **analytics project** involving big data
2. Select **the analytics method** (choose any one) to be used to solve the problem
3. Big Data **Ingestion**
4. Big Data **Storage**
5. Big Data **Processing** that is seen befitting the situation
6. Develop the analytics **solution** using the processed data
7. Analytics solution **Interpretation**
8. Articulation of **challenges** encountered



Implementation Criteria for Group Project

- Business Value and Creativity
  - To make the project interesting, groups must also strive to cover **a variety of methods** that allows demonstration of pipelines and stages mentioned in the previous picture
- Pragmatism
  - Should be practical and solve a realistic problem that can benefit from Big Data related technologies.
- Design of Pipeline
  - Choose some or all the stages listed in the aspects to focus as demanded by the project requirements.
- Demonstration of Analytics Processing
  - It is desirable that groups demonstrate a part of their solution using any or some of: **machine learning or deep learning or recommender system or IoT data or other data sources.** If the data is selected outside this spectrum please provide the justification.

Open Data Sets Available

- **Amazon Hosted Data Sets**
  - Amazon has public data program. https://aws.amazon.com/datasets . The data can be accessed from the respective storage free of charge and it would be easy to access from AWS hosted app. One example of app using land satellite data hosted in AWS is snapsat.org which is open source as well.
- **Singapore Government Agencies Data Sets**
  - Data.gov.sg is the primary portal for users to discover data published by the Singapore Government and its agencies. Launched in June 2011, the portal brings together over 11,000 datasets from 70 government ministries and agencies. It also offers a listing of applications developed using government data and resources for developers. http://data.gov.sg
- **Other Countries Public Data Sets**
  - Kaggle – the open data hunting ground
  - UCI – Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets.php
  - (Data.gov and http://www.nsf.gov/statistics/iris/start.cfm) Contains US Federal Government Datasets and NSF-Industrial Research and Development Information System
  - (Data.gov.uk) UK Government DataSets and UK Data Archive
  - (http://data.gov.au/) Australian Datasets
  - (Data.govt.nz/) New Zealand Datasets
  - (http://www.stat.go.jp/english/ Japan Statistics Bureau
  - (https://ec.europa.eu/eurostat/data/database) Open Data hub of the European Union
- **Google Custom Search Engine**
  - Google has a special data set search engine at: https://www.google.com/publicdata/directory

Sample Ideas

- **ClickStreams**
  - This example is from the book "Hadoop Application Architectures by Gwen Shapira et al" chapter 8. In this example the authors have the following four goals for the processing pipeline: (1) Sanitize and clean up the raw data. (2) Extract the interesting data from the raw clickstream events. (3) Transform the extracted data to create processed data set(s). In this example, create a sessionized data set. (4) Load (i.e., store) the results in a data format in Hadoop that supports high-performance querying.
  - Look for the code sample from https://github.com/hadooparchitecturebook/hadoop-arch-book/tree/master/ch08-clickstream/JavaSessionize/src/main/java/com/hadooparchitecturebook/clickstream
- **Yelp Datasets Projects**
  - This data set is a part of the Yelp Dataset Challenge conducted by crowd-sourced review platform, Yelp. Projects such as natural language processing and sentiment analysis, photo classification, and graph mining among others, are some of the projects that can be carried out using this data set containing diverse data. Inspiration: https://www.yelp.com/dataset
- **ETL (Extract Transform Load**
  - ETL is the process of reading data from one or more sources, applying some transformation on the data, and writing it to another data source. Conceptually, it consists of three steps: extract, transform and load. These steps need not be sequential; an application does not have to extract all the data before moving it to the transform step. Once it has extracted a portion of the data, it may run the three steps in parallel. Inspiration: http://www.sparkexpert.com/tag/etl/
- **Product Purchase Demographics**
  - Assume we have two datasets: User information (id, email, language, location) and Transaction information (transaction-id, product-id, user-id, purchase-amount, item-description). Given these datasets, we want to find the number of unique locations in which each product has been sold. Inspiration: https://blog.matthewrathbone.com/2016/02/09/python-tutorial.html

Sample Ideas

- **More Ideas From**
  - Official Spark Code Samples: http://spark.apache.org/examples.html and code repository https://github.com/apache/spark/tree/master/examples/src/main/scala/org/apache/spark/examples
  - http://www.insightdataengineering.com/blog/Working-With-Apache-Spark.html
  - http://www.datanami.com/2014/03/06/apache_spark_3_real-world_use_cases/
  - https://databricks.com/resources/case-studies
  - https://www.toptal.com/spark/introduction-to-apache-spark
  - http://spark.apache.org/docs/latest/mllib-feature-extraction.html
  - https://www.cloudera.com/tutorials.html
  - http://www.infoq.com/articles/apache-spark-sql

Deliverables and success criteria

- **Project Proposal** (more details next slide) + **Final Project Report** + **Demo Presentation (**Individual Performance**)**+ **Individual Reflection Report + Peer Review**
- **Demo Presentation**
  - Each group would be allocated 20 minutes to demonstrate the project. The demonstration should be a good summary of the designed features. Please take time to rehearse and test the application with different data set scenarios and computation test cases.
  - The demonstration should include:
    - Running application on pre-configured VM nodes/or other platform
    - Major scenarios and data flows of the application (the lecturers may suggest additional flows that are implied by the project)
    - Q&A could take place anytime during the presentation.
    - Detailed demonstration schedule for teams will be announced nearer to the dates. If your team has any preferences do email the lecturer 2 weeks before the presentation(isslf@nus.edu.sg).
- **Final Project Report**
  - The report produced after project demonstration serves as a journal of implementation experience. Please add installation instructions for additional tools that are being explored in the project. Creativity and innovation in the proposed work. (Promising reports may be selected for journal/conference publication)
- **Individual Reflection Report**
  - Individual feedback and reflection.

Proposal for Group Project

- Submit to CANVAS by 9th Feb 2024 for evaluation by ISS lecturers
  - Earlier submission and approval allows earlier conduct of the project
- The proposal should include the following content

Section	Content
Project Title	Title
Project Members	Names
Overview	Describe the context and the business problem solved by the system to be secured
General Architecture	Describe the general architecture of the data engineering solution proposed. This serves as the baseline for the scope of work
Scope of Work	The followings are the key aspects of analytics the project needs to focus: 1. Groups must identify the analytics problem to be solved 2. Groups must identify static or real time data available to the computing cluster using an appropriate Ingestion tool. 3. Groups must identify and implement appropriate storage strategy. 4. Describe the parts of the approach/system that are to be designed/developed to solve the problem. 5. Explain how does the scope allows you to sufficiently demonstrate the mastery of the course modules
Effort Estimates	List the rough listing tasks and their estimated efforts. This is to ensure that you have thought about the approach and the implementation effort

Assessment Schedule – TT FT

Assessment Component (Deadline)	Weightage	Deliverables
Practice Module Briefing (16th Jan 2024)	-	
Project Proposal Submission (9th Feb 2024)	10%	Submission: Project proposal submission as per template (refer to the previous slide). Approval is required. Feedback will be shared by 21st Feb 2024
Project Demo Presentation(20th Apr 2024)	15%	Submit slides one day before the presentation date 19th Apr 2024
Final Project Report & Individual Reflection Report (26th Apr 2024)	15%	Submission: (1) Final project report, solution (2) Source code
Individual Reflection Report + Peer Review (26th Apr 2024)	10%	Submission (1) Individual reflection report (2) Peer review
Written Examination (Tentatively scheduled for the 3rd week of May 2024)	50%	
Total	100%	

- The presentation and project report would be graded by a team of ISS lecturers.

Practice Module Withdrawal

Course	Registration and Payment Deadline	Withdrawal with Full Refund
Practice Module for Graduate Certificate in Big Data Analytics	30 November 2023 (Thursday)	28 January 2024 (Sunday)

Important Timeline to Note	Withdrawal Policy
By 28 January 2024* **email should reach ask-iss@nus.edu.sg latest by 28 January 2024, 11:59pm*	• Withdrawal without grade penalty • Full refund
29 January 2024 to 3 March 2024* **email should reach ask-iss@nus.edu.sg latest by 3 March 2024, 11:59pm*	• Withdrawal with/without valid documentation: **W grade** • No refund
From 4 March 2024 onwards	• Withdrawal with valid documentation: **W grade** • Withdrawal without valid documentation: **IC grade** • Participants who receive an IC grade will have 3 years to retake the practice module, else will be awarded with F grade • Participants with F grade awarded can only have one additional attempt to retake practice module, max grade capped at C+ • No refund