BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES
Digital
Part A: Content Design

Course Title	STREAM PROCESSING AND ANALYTICS
Course No(s)	DSECL ZC556
Credit Units	5
Credit Model	
Content Authors	PRAVIN PAWAR

Course Description

Data is moving at very rapid space because of which necessarily of scalable systems capable of processing and analyzing this fast, streaming data has arisen. This course introduces the students with the architecture of streaming data processing systems. This course also enables students to understand the complete end-to-end solution for cost-effective analysis and visualization of streaming data with the help of various open source solutions available in this space. This course also helps students to learn the implementation and application of algorithms and data structures required for the streaming applications. Advanced streaming applications like Streaming SQL, Streaming Machine Learning will be discussed at proper length.

Course Objectives

No	
CO1	To introduce the applications of streaming data systems
CO2	To introduce the architecture of streaming data systems
CO3	To introduce the algorithmic techniques used in streaming data systems
CO4	To present survey of tools and techniques required for streaming data analytics

Text Book(s)

T1	Streaming Data: Understanding The Real-Time Pipeline, Andrew G.Psaltis, 2017, Manning Publications
T2	Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data, Byron Ellis, 2014, Wiley

Reference Book(s) & other resources

R1	Big Data – Principles and best practices of scalable real-time data systems, Nathan Marz, James Warren, 2017, Manning Publications
R2	Designing Data Intensive Applications, Martin Kleppmann, O'Reilly

Learning Outcomes:

No	Learning Outcomes
LO1	Understand the components of streaming data systems with their capabilities and characteristics
LO2	Learn the relevant architecture and best practices for processing and analysis of streaming data
LO3	Gain knowledge about the development of system for data aggregation, delivery and storage using Open source tools
LO4	Get familiarity with the advance streaming applications like Streaming SQL, Streaming machine learning

Part B: Learning Plan

Academic Term	II Semester 2019 -2020
Course Title	STREAM PROCESSING AND ANALYTICS
Course No	DSECL ZC556
Lead Instructor	Prof. Maninder Singh Bawa

Glossary of Terms

Module	**M**	Module is a standalone quantum of designed content. A typical course is delivered using a string of modules. M2 means module 2.
Contact Hour	**CH**	Contact Hour (CH) stands for a hour long live session with students conducted either in a physical classroom or enabled through technology. In this model of instruction, instructor led sessions will be for 32 CH.
Recorded Lecture	**RL**	RL stands for Recorded Lecture or Recorded Lesson. It is presented to the student through an online portal. A given RL unfolds as a sequences of video segments interleaved with exercises.
Lab Exercises	**LE**	Lab exercises associated with various modules
Self-Study	**SS**	Specific content assigned for self study
Homework	**HW**	Specific problems/design/lab exercises assigned as homework

Modular Structure

No.	Title of the Module
M1	Scalable Streaming Data Systems
M2	Streaming Data Systems Architecture
M3	Streaming Data Frameworks
M4	Streaming Analytics
M5	Advanced Streaming Applications

Detailed Lecture Plan

M1: Scalable Streaming Data Systems

Session 1 to 3 / Contact Hour 1 - 6

Time	Type	Description/Plan	Reference
Session 1	CH1	• Thinking about Data Systems • Reliable, Scalable and Maintainable Data Applications • Properties of Data	R1 Ch1 R2 Ch2
	CH2	• Scaling with the traditional databases • Big Data Systems • Desired properties of Big Data Systems	R2 Ch1
Session 2	CH3	• Data Model for Big Data • Generalized Big Data System Architecture	R2 Ch2 Class Notes
	CH4	• Real time systems • Difference between Batch processing and Stream Processing • Difference between real time and streaming systems	T1 Ch1 Class Notes
Session 3	CH5	• Streaming Data Applications • Databases and Streams • Usage patterns of Streaming Data	Class Notes R1 Ch11 Class Notes
	CH6	• Sources of Streaming Data • Complex Event Processing Systems	T2 Ch1 Class Notes
Post CH	SS	• Explore more on the non functional requirements of Data Intensive Applications ✓ Non-functional Requirements for Real World Big Data Systems ✓ IBM Big Data & Analytics RA_V1 • Explore more on the differences between the batch processing and streaming data applications ✓ Batch vs Real time data processing	

		• Identify the use cases of Complex Event Processing Systems ✓ What is stream processing? ✓ complex-event-processing

M2: Streaming Data Systems Architecture

Session 4 to 7 / Contact Hour 7 - 14

Time	Type	Description/Plan	Reference
Session 4	CH7	• Generalized Streaming Data Architecture	T1 Ch 1 T1 Ch 2
	CH8	• Lambda Architecture • Kappa Architecture	Class Notes
Session 5-6	CH9	• Streaming Data system Component • Features of Real time Architecture • A real time architecture checklist	T2 Ch2
	CH 10	• Service Configuration and Coordination Systems • Maintaining the state • Apache ZooKeeper	T2 Ch3
	CH 11	• Data Flow Manager • Managing distributed data flows	T2 Ch4
	CH 12	• Apache Kafka	T2 Ch4 Kafka Docs
Session 7-8	CH13	• Streaming Data Processor Concepts • Timing Concepts	T2 Ch 5 T1 Ch 5
	CH14	• Windowing • Joins	T1 Ch5 R1 Ch11
	CH15	• Storage for Streaming Data • NoSQL storage Systems • Choosing a Storage technology	T2 Ch6
	CH16	• Delivery of Streaming Metrics	T2 Ch7
Post CS	SS	• Explore in detail about issues with Lambda Architecture ✓ questioning-the-lambda-architecture ✓ a-brief-introduction-to-two-data-processing-architectures • Explore the Java APIs exposed by following systems ✓ Apache ZooKeeper	

		✓ Apache Kafka • Explore the data models of NoSQL data systems ✓ MongoDB ✓ Cassandra	

M3: Streaming Data Frameworks

Session 8 to 11 / Contact Hour 15 - 22

Time	Type	Description/Plan	Reference
Session 8	CH 15	• Key features of Streaming Data Frameworks • Survey of Streaming Data Systems	Class Notes
	CH 16	• Apache Spark Streaming	Spark Streaming Guide
Session 9	CH 17	• Apache Flink • Apache Samza	Flink Docs Samza Docs
	CH 18	• Apache Kafka Streaming	Kafka Streaming Guide
Session 10	CH 19	• Apache Storm Architecture	Storm Docs
	CH 20	• Apache Storm Concepts • Apache Storm Groupings	T2 Ch 5
Session 11	CH 21	• Apache Storm Running Example	Storm Docs
	CH 22	• Storm – Kafka Integration Example	Class Notes
Post CH	SS	• Compare the different streaming data platforms and identify the use cases for which they are suitable • Implement the streaming data pipeline using the Kafka Streaming library • Implement a streaming data application with Spark streaming	Kafka Streaming Guide Spark Streaming Guide

M4: Streaming Analytics

Session 12 to 13 / Contact Hour 23 - 26

Time	Type	Description/Plan	Reference
Session 12	CH 23	• Exact Aggregation of Streaming Data • Time Series Analysis	T2 Ch 8
	CH 24	• Quantization Framework • Stochastic Optimization	T2 Ch8
Session 13	CH 25	• Registers and Hash Functions • The Bloom Filter	T2 Ch 10
	CH 26	• Distinct Value Sketches • The Count-Min Sketch	T2 Ch 10
Post CH	SS	• Study illustrations for Streaming data concepts • Explore algorithms for aggregation of streaming data • Explore more about the streaming data processing algorithms for exact results	Class Notes

M5: Advanced Streaming Applications

Session 14 to 15 / Contact Hour 27 - 30

Time	Type	Description/Plan	Reference
Session 14	CH25	• Necessity of Streaming SQL • Streaming SQL : Windows • Streaming SQL : Joins • Streaming SQL : Patterns	Streaming SQL Blog
	CH26	• Apache Storm support for Streaming SQL • Apache Flink support for Streaming SQL • Streaming SQL for Apache Kafka	storm-sql flink-stream-sql Kafka Streaming SQL
Session 15	CH27	• Models for Streaming Data - Linear models • Models for Streaming Data - Logistic Regression models	T2 Ch 11
	CH 28	• Forecasting with Models - Exponential Smoothing methods • Forecasting with Models - Regression methods	T2 Ch 11
Session 15	CH 29	• Streaming ML Frameworks I	structured-streaming-ml
	CH 30	• Streaming ML Frameworks II	

Post CH	SS	• Get familiarized with Streaming SQL tools ✓ storm-sql ✓ Kafka Streaming SQL • Build and deploy machine learning models using Spark structured streaming ✓ structured-streaming-ml	

Session 16 / Contact Hour 31 - 32

Time	Type	Description/Plan	Reference
Session 16	CH31	• Review of Streaming Data Systems and Architectures	CH 1 to 16
	CH32	• Review of Streaming Data Techniques and Applications	CH 17 to 32

Evaluation Scheme:

Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

No	Name	Type	Duration	Weight	Day, Date, Session, Time
EC-1	Assignment-1	Take-home, Programming and use of platforms	-	10%	TBD
	Assignment-2		-	15%	TBD
	Quiz-1	Online	30 mins	5	TBD
EC-2	Mid-Semester Test	Closed Book	2 hours	30%	TBD
EC-3	Comprehensive Exam	Open Book	3 hours	40%	TBD

Notes:
Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 8 (contact hours 1 to 16)
Syllabus for Comprehensive Exam (Open Book): All topics

Important links and information:
Elearn portal: https://elearn.bits-pilani.ac.in
Students are expected to visit the Elearn portal on a regular basis and stay up to date with the latest announcements and deadlines.
Contact sessions: Students should attend the online lectures as per the schedule provided on the Elearn portal.
Evaluation Guidelines:

1. EC-1 consists of either two Assignments or three Quizzes. Students will attempt them through the course pages on the Elearn portal. Announcements will be made on the portal, in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted.
3. For Open Book exams: Use of books and any printed / written reference material (filed or bound) is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam which will be made available on the Elearn portal. The Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self

study schedule as given in the course handout, attend the online lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.