**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** | SURYA PRAKASH G |

**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 | Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data, Byron Ellis, 2014, Wiley<http://www-di.inf.puc-rio.br/~endler/courses/RT-Analytics/transp/Books/Real-Time%20Analytics%20Techniques%20to%20Analyze.pdf> |
| T2 | Streaming Data: Understanding The Real-Time Pipeline, Andrew G.Psaltis, 2017, Manning Publications |

**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** |  |
| **Course Title** | STREAM PROCESSING AND ANALYTICS |
| **Course No** |  |
| **Lead Instructor** |  |

**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 | T2 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 | T1 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](https://www.researchgate.net/publication/325226328_Non-Functional_Requirements_for_Real_World_Big_Data_Systems_-_An_Investigation_of_Big_Data_Architectures_at_Facebook_Twitter_and_Netflix) * [IBM Big Data & Analytics RA\_V1](https://www.ibm.com/developerworks/community/files/form/anonymous/api/library/e747a4bd-614d-4c5d-a411-856255c9ddc4/document/bbc80340-3bf4-4e0a-8caf-a43f64a22f05/media) * Explore more on the differences between the batch processing and streaming data applications   + [Batch vs Real time data processing](https://www.datasciencecentral.com/profiles/blogs/batch-vs-real-time-data-processing) * Identify the use cases of Complex Event Processing Systems   + [What is stream processing?](https://medium.com/stream-processing/what-is-stream-processing-1eadfca11b97)   + [complex-event-processing](https://databricks.com/glossary/complex-event-processing) | |

**M2: Streaming Data Systems Architecture**

**Session 4 to 8 / Contact Hour 7 - 16**

|  |  |  |  |
| --- | --- | --- | --- |
| Time | Type | Description/Plan | Reference |
| Session 4 | CH7 | * Generalized Streaming Data Architecture * Lambda Architecture * Kappa Architecture * Streaming Data system Component * Features of Real time Architecture * A real time architecture checklist | T2 Ch 1  T2 Ch 2  Class Notes  T1 Ch2 |
| CH8 |
| Session 5-6 | CH9 | * Service Configuration and Coordination Systems * Maintaining the state * Apache ZooKeeper * Data Flow Manager * Managing distributed data flows with   Apache Kafka | T1 Ch2 |
| CH 10 | T1 Ch3  T1 Ch4  Kafka Docs |
| CH 11 | * Kafka Fundamentals Overview * Use-Cases and applications * Architecture * Kafka Topics, Producer and Consumer Using CLI * Programming Kafka * Simple Kafka Producer * Simple Kafka Consumer * Producer, Consumer Configuration * Producer, Consumer Execution * Kafka Consumer Groups | T1 Ch4 |
| CH 12 | T1 Ch4  Kafka Docs |
| Session 7-8 | CH13 | * Streaming Data Processor Concepts * Timing Concepts * Windowing * Joins | T1 Ch 5  T2 Ch 5 |
| CH14 | T2 Ch5  R1 Ch11 |
| CH15 | * Storage for Streaming Data * NoSQL storage Systems * Choosing a Storage technology * Delivery of Streaming Metrics | T1 Ch6 |
| CH16 | T1 Ch7 |
| Post CS | SS | * Explore in detail about issues with Lambda Architecture   + [questioning-the-lambda-architecture](https://www.oreilly.com/ideas/questioning-the-lambda-architecture)   + [a-brief-introduction-to-two-data-processing-architectures](https://towardsdatascience.com/a-brief-introduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data-4f35c28005bb) * Explore the Java APIs exposed by following systems   + [Apache ZooKeeper](https://zookeeper.apache.org/doc/r3.4.13/javaExample.html)   + [Apache Kafka](https://www.tutorialspoint.com/apache_kafka/apache_kafka_simple_producer_example) * Explore the data models of NoSQL data systems   + [MongoDB](https://www.mongodb.com/)   + [Cassandra](http://cassandra.apache.org/)   Self study on other frameworks |  |

**M3: Streaming Data Frameworks**

**Session 9 to 11 / Contact Hour 17 - 24**

|  |  |  |  |
| --- | --- | --- | --- |
| Time | Type | Description/Plan | Reference |
| Session 9 | CH 17 | * Key features of Streaming Data Frameworks * Survey of Streaming Data Systems | Class Notes |
| CH 18 | * Apache Spark Streaming   SELF Exploration/Assignment on the following   * Apache Flink * Apache Samza * Apache Kafka Streaming * Apache Storm | [Spark Streaming Guide](https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html)  [Flink Docs](https://ci.apache.org/projects/flink/flink-docs-stable/)  [Samza Docs](https://samza.apache.org/learn/documentation/latest/)  [Kafka Streaming Guide](https://docs.confluent.io/current/streams/index.html)  [Storm Docs](https://storm.apache.org/releases/2.0.0/index.html) |
| Session 10 | CH 19 | * Apache Spark Streaming * Spark Streaming fundamentals * Motivation * Difference between Spark Streaming API and Spark API * Architecture * Components of Spark Engine * Spark Application Architecture * Fault Tolerance * Comparison with Traditional Streaming Systems | [Spark Streaming Guide](https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html) |
| CH 20 |
| Session 11 | CH 21 | * Spark + Kafka integration | [Spark Streaming Guide](https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html) |
| CH 22 |
| Session 12 | CH 23 | * Structured Streaming * Developing application in Databricks platform | [Structured Streaming Docs](https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html) |
| CH 24 | 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](https://docs.confluent.io/current/streams/index.html)  [Spark Streaming Guide](https://spark.apache.org/docs/2.2.0/streaming-programming-guide.html) |

**M4: Streaming Analytics**

**Session 13 to 14 / Contact Hour 25 - 28**

|  |  |  |  |
| --- | --- | --- | --- |
| Time | Type | Description/Plan | Reference |
| Session 13 | CH 25 | * Exact Aggregation of Streaming Data * Time Series Analysis | T1 Ch 8 |
| CH 26 | T1 Ch8 |
| Session 14 | CH 27 | * Registers and Hash Functions * The Bloom Filter | T1 Ch 10 |
| CH 28 | * Distinct Value Sketches * The Count-Min Sketch | T1 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 15 to 16 / Contact Hour 29 - 32**

|  |  |  |  |
| --- | --- | --- | --- |
| Time | Type | Description/Plan | Reference |
| Session 15 | CH29 | * Necessity of Streaming SQL * Streaming SQL : Windows * Streaming SQL : Joins * Streaming SQL : Patterns | [Streaming SQL Blog](https://wso2.com/library/articles/2018/02/stream-processing-101-from-sql-to-streaming-sql-in-ten-minutes/) |
| CH30 | * Streaming SQL for Apache Kafka * KSQL | [Kafka Streaming SQL](https://www.confluent.io/product/ksql/) |
| Session 16 | CH 30 | * Streaming Analytics with Cloud * AWS Kinesis * Data Streams * Data Firehose * Data Analytics * AWS IoT / Streaming Analytics Service * Channels, Pipelines * Data stores & data sets * Streaming ML Frameworks | Kinesis Docs  Databricks docs  Azure Docs  Class notes |
|  | CH 31 |
| Post CH | SS | * Get familiarized with Streaming SQL tools   + [Kafka Streaming SQL](https://www.confluent.io/product/ksql/) * Build and deploy machine learning models using Spark structured streaming   + [structured-streaming-ml](https://docs.databricks.com/spark/latest/structured-streaming/index.html) |  |

**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 | 10 days | 10% | TBA |
| Assignment-2 | Take home | 15 days | 15% | TBA |
| Quiz-1 | Online | 1 day | 5% | TBA |
| EC-2 | Mid-Semester Exam | Closed Book | 2 hours | 30% | TBA |
| EC-3 | Comprehensive Exam | Open Book | 3 hours | 40% | TBA |

**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.