**Unit 2**

**Reading Task:**

* [Streaming Data is Different](https://www.coursera.org/lecture/big-data-management/why-is-streaming-data-different-Wfb3n)
  + This video explores how streaming data is different.
* [AWS and Data Streaming](https://www.linkedin.com/learning/amazon-web-services-for-data-science/why-use-data-streaming)
  + This video takes a look at Amazon Web Services (AWS) and why it should be used for data streaming.
* [Kafka for Beginners - A Tutorial](https://www.youtube.com/watch?v=U4y2R3v9tlY)
  + This is a video tutorial for Apache's Kafka.
* [Kafka](https://kafka.apache.org/intro)
  + This Web site is a repository for information on Kafka and how it can be used.
* [Overview of Pub/Sub in Python](https://www.pubnub.com/docs/python/data-streams-publish-and-subscribe-sdk-v3%20)
  + This article provides an overview of Pub/Sub in Python.
* [Realtime Streaming Data Pipeline Using Google Cloud Platform and Bokeh](https://medium.com/datareply/realtime-streaming-data-pipeline-using-google-cloud-platform-and-bokeh-9dd0cfae647a)
  + This article discusses building a real-time streaming data pipeline using the Google Cloud Platform.
* [Spark Streaming Programming Guide](https://spark.apache.org/docs/latest/streaming-programming-guide.html)
  + This Web site provides an overview of spark streaming and related concepts.
* [Configuring Streaming Data Sources](https://sqlstream.com/docs-st/streamlab/sources/configuring-streaming-data-source/)
  + This Web site provides an overview of streaming data sources.
* [What is Streaming Data?](https://aws.amazon.com/streaming-data/)
  + This Web site provides an overview of streaming data and its benefits.
* [4 Key Components of a Streaming Data Architecture](https://www.upsolver.com/blog/streaming-data-architecture-key-components)
  + This article covers key concepts and design patterns for streaming data architecture.
* [Streaming with Pub/Sub](https://cloud.google.com/dataflow/docs/concepts/streaming-with-cloud-pubsub)
  + This Web site provides an overview of building streaming pipelines with Pub/Sub.
* [Goodbye Hadoop. Building a Streaming Data Processing Pipeline on Google Cloud](https://cloud.google.com/blog/products/data-analytics/goodbye-hadoop-building-a-streaming-data-processing-pipeline-on-google-cloud)
  + This article shares how to build a streaming data processing pipeline on Google Cloud.
* [What Is Data Streaming? A Data Architect's Guide](https://www.dataversity.net/what-is-data-streaming-a-data-architects-guide/)
  + This article examines data streaming and its applications.
* [Improvement of Kafka Streaming Using Partition and Multi-Threading in Big Data Environment](https://search-proquest-com.coloradotech.idm.oclc.org/docview/2301579574/fulltextPDF/AFBC35AC6F764DBDPQ/1?accountid=144789)
  + This paper examines applications of a Hadoop ecosystem that supports features in the manufacturing industry.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Unit 2 - Discussion Board** | | | **Task Type:** | Discussion Board | | **Deliverable Length:** | See assignment details | | **Points Possible:** | 75 | | **Description:**  **Primary Response is due by Thursday (11:59:59pm Central), and Peer Responses are due by Saturday (11:59:59pm Central).**  **Primary** **Response:** Within the Discussion Board area, write 300-500 words that respond to the following questions with your thoughts, ideas, and comments. This will be the foundation for future discussions with your classmates. Be substantive and clear, and use examples to reinforce your ideas.  For this Discussion Board, please complete the following:  For this Discussion Board, you will evaluate the characteristics of streaming data versus batch data pipelines.  *Batch data* are often processed in intervals and involve a large grouping of records, and data warehouses are often refreshed on a set interval, using batch processes. An example would be a financial warehouse refreshed nightly and that is when any new records would be added. Streaming data are quite different. They are constantly being generated and need to be processed as soon as possible. An example of streaming data would be a social media feed for a retail company or an Internet of Things (IoT) for robotic assembly lines.  For this assignment, you will evaluate the characteristics of streaming data versus batch data pipelines.   * What are the major differences between streaming data and batch data pipelines that you see? * What kind of situation would a company be inclined to use streaming data over batch data?   + Provide a real-world example of streaming data and an example of batch data.   **Responses to Other Students:** Respond to at least 2 of your fellow classmates with at least a 100-200-word reply about their Primary Task Response regarding items you found to be compelling and enlightening. To help you with your discussion, please consider the following questions:   * What did you learn from your classmate's posting? What additional questions do you have after reading the posting? * What clarification do you need regarding the posting? * What differences or similarities do you see between your posting and other classmates' postings?   **For assistance with your assignment, please use your text, Web resources, and all course materials.**  **Discussion Board Rubric**   |  |  |  |  | | --- | --- | --- | --- | | **Expectation** | **Points Possible** | **Points Earned** | **Comments** | | **Application of Learning Material Content:** Post demonstrates an understanding of Learning Material content. | 20 |  |  | | **Application of Course Knowledge:** Post contributes unique perspectives or insights gleaned from text/learning resources, or specified by assignment. | 20 |  |  | | **DB Responses:** Response substantively to two posts. Responses encourage interaction in the Discussion Board and classroom community. | 20 |  |  | | **Organization:** The post presents information logically and is clearly relevant to the discussion topic. | 8 |  |  | | **Professional Language:** Posts contain accurate grammar, spelling, and/or punctuation with few or no errors. **Any resources should be cited in APA format or style specified in the assignment.** | 7 |  |  | | **Total Points** | 75 |  |  | | **Total Points Earned** |  |  |  |   View a downloadable version of the [Discussion Board Grading Rubric](https://resources.careered.com/LCMSFileSharePreview/Resources/MSExcelWorkbook/DB_Rubric_75_300-400_Level.xlsx). | | | **Course Objectives:**  ·        Analyze the preparation, collecting, storing, securing, analyzing, interpreting, processing, and reporting of streaming data.  ·        Examine the most common challenges of managing streaming data analytics-a mix of streaming and at-rest data collection, measurement, tracking, analysis, security and reporting.  ·        Understand the differences between streaming data analytics and at-rest data analytics, as well as the foundation, techniques, and processes of streaming data analytics. | | | **Model Answer:** | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Unit 2 - Individual Project** | | | **Task Type:** | Individual Project | | **Deliverable Length:** | 3-5 pages (not including cover page and resource page) | | **Points Possible:** | 125 | | **Description:**  **Assignment Details**  Big data platforms enable many new approaches to ingesting data. Because they are distributed in nature, they allow for a very large-scale opportunity to add processing power very rapidly. Additionally, this processing power can be leveraged for the ingestion of data. Technologies such as Kafka, Nifi, and others are intended for this very purpose. These architectures are based on a publish and subscribe architecture, often referred to simply as*pub / sub*. A developer can create a scalable ingest pipeline that allows an extremely large amount of data to flow in real-time using this approach.  For this assignment, you will create a pub/sub data architecture and design for streaming data. Within this task, identify a hypothetical scenario that generates real-time data (e.g., RFIs, industrial Internet of things [IIoT], mobile apps, or game engines). First, you will produce a visual showing the data source. Second, you will show how the published component is used and how the to subscribe component is used. Third, you will discuss how the data could be stored or used to make any proposed decisions.  For the second part of the assignment, construct the pseudocode needed to create the pub/sub component of this solution. The project deliverables include the following:   * Provide a Word document that includes the following:   + A description of the data source and how the pub / sub-components will interact   + A description of how the to subscribe feature will function   + A discussion on how these data could be stored or used to make any proposed decisions   + The pseudocode solution of the pub / sub-component * Once complete, submit your assignment for grading in the Microsoft Word document.   **Individual Project Rubric**  The Individual Project (IP) Grading Rubric is a scoring tool that represents the performance expectations for the IP. This Individual Project Grading Rubric is divided into components that provide a clear description of what should be included within each component of the IP. It's the roadmap that can help you in the development of your IP.   |  |  |  |  | | --- | --- | --- | --- | | **Expectation** | **Points Possible** | **Points Earned** | **Comments** | | **Identification:** Identify the differences between streaming data analytics and at-rest data analytics, as well as the foundation, techniques, and processes of the streaming data analytics | 30 |  |  | | **Analysis:**Analyze the preparation, collecting, storing, securing, analyzing, interpreting, processing, and reporting of streaming data. | 30 |  |  | | **Visual:** Include a visual of the data being generated through a Web service or some other end point. | 30 |  |  | | **Organization:** Assignment presents information logically and is clearly relevant to discussion topic. | 25 |  |  | | **Professional Language:**Assignment contains accurate grammar, spelling, and punctuation with few or no errors. | 10 |  |  | | **Total Points** | 125 |  |  | | **Total Points Earned** |  |  |  | | | | **Course Objectives:**  ·        Analyze the preparation, collecting, storing, securing, analyzing, interpreting, processing, and reporting of streaming data.  ·        Examine the most common challenges of managing streaming data analytics-a mix of streaming and at-rest data collection, measurement, tracking, analysis, security and reporting.  ·        Understand the differences between streaming data analytics and at-rest data analytics, as well as the foundation, techniques, and processes of the streaming data analytics. | | | **Model Answer:**  The solution should include a visual of the data being generated through a Web service or some other end point. The publish routine should monitor and act on the new data as discussed and depicted in the proposed and discussed solution. | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Unit 2 - Live Chat Extra Credit Summary** | | | **Task Type:** | Live Chat | | **Deliverable Length:** | 200 words | | **Points Possible:** | 15 | | **Description:**  In this class, you have the opportunity to receive credit for attending Live Chats or reviewing the chat archive. This opportunity is available once per week. To receive these points, you can either:  1.    Attend the Live Chat session. or  2.    Review the archived chat session and submit a 200-word summary of the content. The archive review summary must meet the expectations described below.   * Summarize the content of the chat. Some ideas for what to include are a description of what information was covered and how you will apply this information to your academic or professional work. * Label the document "Chat Credit" with the date the Live Chat was presented. * A summary must be submitted within 1 week of the Live Chat to be considered for credit.   Your instructor has the discretion to determine whether to award the points. A summary must meet each of these guidelines. Points are awarded on an all-or-nothing basis.  To submit your summary, head to the assignment list and select Unit 2 - Live Chat Extra Credit Summary.  View the [Individual Project Demo](https://careered.libguides.com/ctu/success/classroom#s-lg-box-wrapper-26053866) for guidance on how to submit your chat summary as an Individual Project via the Virtual Classroom or the CTU Mobile app. | | | **Course Objectives:**  ·        Analyze the preparation, collecting, storing, securing, analyzing, interpreting, processing, and reporting of streaming data.  ·        Examine the most common challenges of managing streaming data analytics-a mix of streaming and at-rest data collection, measurement, tracking, analysis, security and reporting.  ·        Understand the differences between streaming data analytics and at-rest data analytics, as well as the foundation, techniques, and processes of the streaming data analytics. | | | **Model Answer:**  Students are required to meet all of the expectations as outlined in the assignment. | | |

**Optional Live Session Talking Points:**

* Streaming Data Pipelines Versus Batch Data Pipelines
* Streaming Data Collection
* Streaming Data Processing
* Streaming Data Storage Use
* Sources of Stream Data Generation