Task 4: Structured Streaming Processing with HDFS Logs

References:

1. Stream Processing with Apache Spark Mastering Structured Streaming and Spark Streaming By Gerard Maas, Francois Garillot, 2019. `https://search.library.qmul.ac.uk/iii/encore/record/C__Rb2531938__SStream%20Processing%20with%20Apache%20Spark%20Mastering%20Structured%20Streaming%20and%20Spark%20Streaming__Orightresult__U__X2?lang=eng&suite=def`
2. Week 8 Lecture, Lab and related resources.

Task 4: Analysis of HDFS Logs dataset (25 points)

In this task, you will analyze HDFS log data using Structured Streaming Processing. Please refer to the provided HDFS log dataset and the streaming process. Note that the example tables/pictures provided in this task are based on sample results. Your results may differ depending on how you process and display the data. When a screenshot is required, include your running timestamp in your screenshot. The example tables/pictures illustrate the expected format for your screenshot.

Fields Descriptions

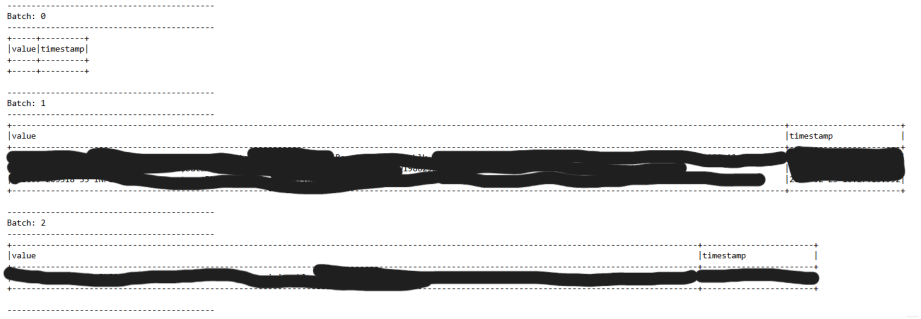
Field	**Description**
Timestamp	The exact time the log was recorded.
Log Level	Indicates the severity of the log message (e.g., INFO, WARN, ERROR).
Component	The component of the system that is recording the log (e.g., dfs.DataNode$DataXceiver, dfs.FSNamesystem).
Block ID	The ID of the block involved in the log entry (e.g., blk_1608999687919862906).
Source IP:Port	The IP address and port of the source node involved in the log operation (e.g., /10.250.19.102:54106).
Destination IP:Port	The IP address and port of the destination node (e.g., /10.250.19.102:50010).
Block Size	The size of the block being transferred (e.g., size 91178).
Message	A description of the action performed in the log entry (e.g., ”Receiving block”, ”PacketResponder terminating”).

1. Loading the Dataset (2 points)

- Explicitly specify the host value as STREAMING_SERVER_HDFS (default: `'default_host'`) and the port as STREAMING_SERVER_HDFS_PORT (default: `'default_port'`) to load the dataset.

- Create a query to display the dataset in the console using append mode and ensure that fields are not truncated.

- Provide a screenshot of the output for batch 0, 1, and 2 in your report. For example:



```
-------------------------------------------
Batch: 0
-------------------------------------------
+-----+---------+
|value|timestamp|
+-----+---------+
+-----+---------+

-------------------------------------------
Batch: 1
-------------------------------------------
+------------------------------------------------------------------------------------------------+---------------+
|value                                                                                           |timestamp      |
+------------------------------------------------------------------------------------------------+---------------+
[redacted]                                                                                        [redacted]
+------------------------------------------------------------------------------------------------+---------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+-------------------------------------------------------------------------------------+---------------+
|value                                                                                |timestamp      |
+-------------------------------------------------------------------------------------+---------------+
[redacted]
+-------------------------------------------------------------------------------------+---------------+

-------------------------------------------
```

2. Defining a Watermark (2 points)

- Define a watermark on the timestamp column with a delay of 5 seconds. (`https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.DataFrame.withWatermark.html`) .

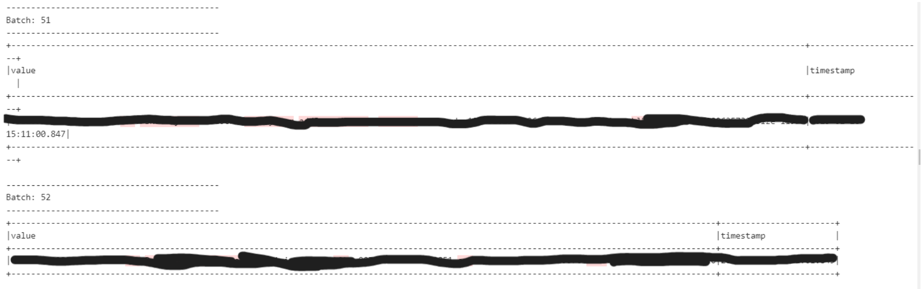
- Explain the purpose of defining a watermark in your report.

3. Analyzing Data Patterns (3 points)

- Create a query to display the dataset in the console using append mode and ensure that fields are not truncated.

- Identify and highlight any patterns or anomalies in the data for batch 51 and batch 52.

- Provide a screenshot of the output for batch 51 and batch 52 in your report, along with a brief analysis of the observed patterns or anomalies. For example:



```
-------------------------------------------
Batch: 51
-------------------------------------------
+------------------------------------------------------------------------------------------------------------------------------+----------------------
--+
|value                                                                                                                         |timestamp
  |
+------------------------------------------------------------------------------------------------------------------------------+----------------------
--+
[redacted]
15:11:00.847|
+------------------------------------------------------------------------------------------------------------------------------+----------------------
--+

-------------------------------------------
Batch: 52
-------------------------------------------
+----------------------------------------------------------------------------------------------------------------+-------------------+
|value                                                                                                           |timestamp          |
+----------------------------------------------------------------------------------------------------------------+-------------------+
[redacted]
+----------------------------------------------------------------------------------------------------------------+-------------------+
```

4. Analyzing DataNode Activity (6 points)

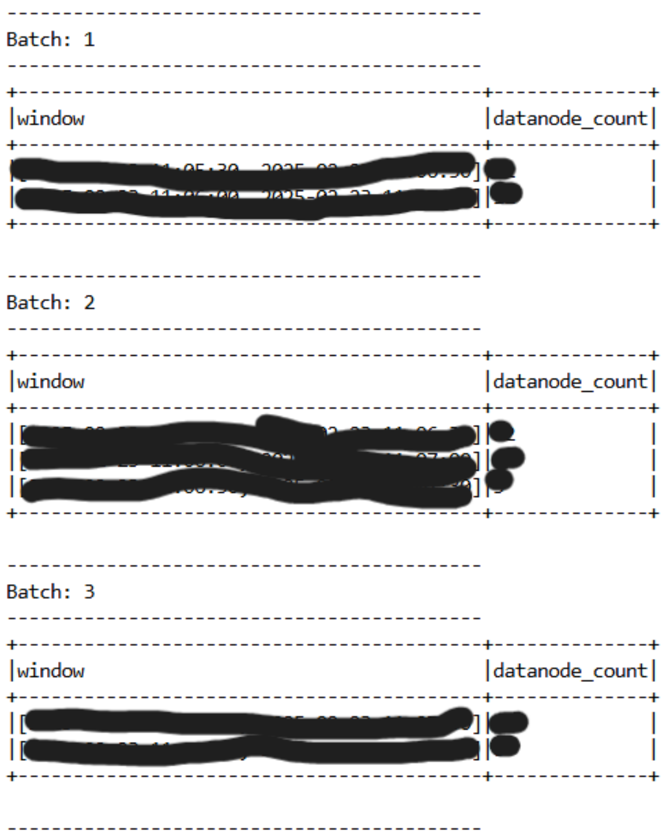
- Create a query to count the number of log entries in the dataset where the ”Component” field contains the string `"DataNode"` in each log entry, within each time window.

- Name the result column as `datanode_count` and write the output to the console in update mode, ensuring that fields are not truncated.

- Set the `windowDuration` to ’60 seconds’ and the `slideDuration` to ’30 seconds’.

- Analyze the trends in DataNode activity over time based on the counts. Identify any significant peaks or patterns.

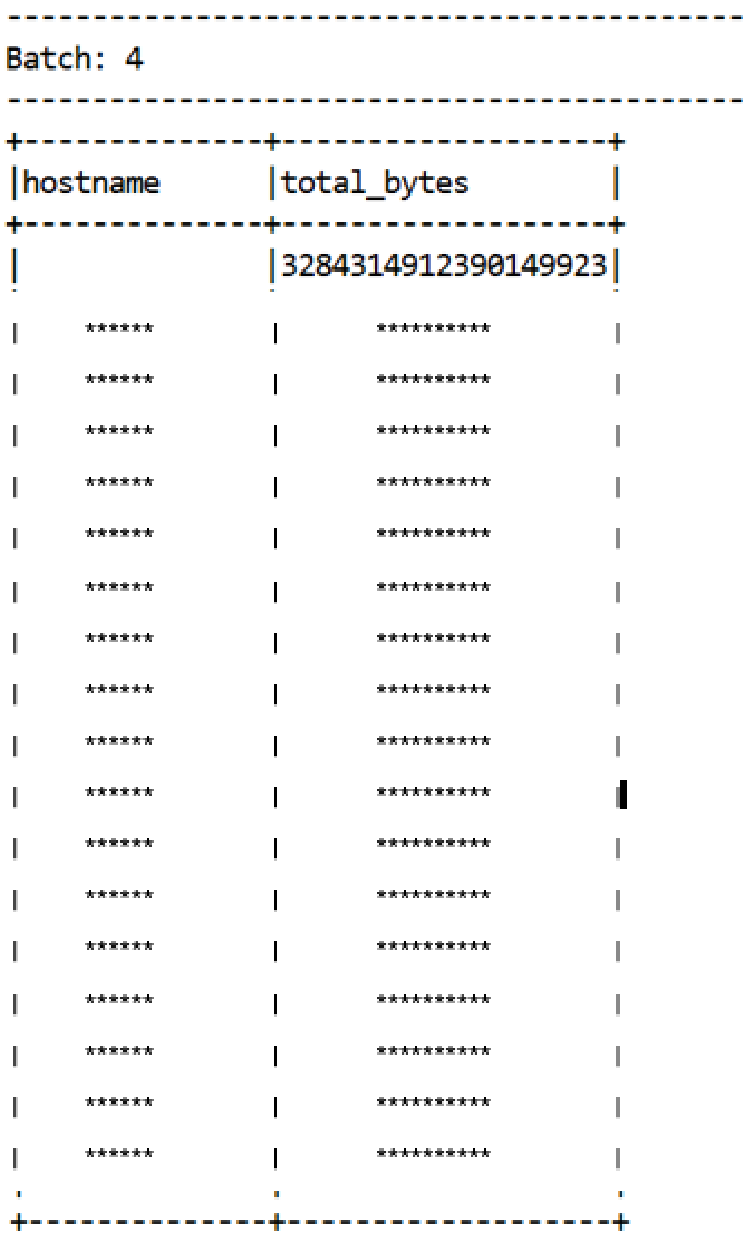
- Include a screenshot of the output for batches 1, 2, and 3 in your report, along with a brief analysis of the observed trends. For example:



```
-------------------------------------------
Batch: 1
-------------------------------------------
+------------------------------------------+--------------+
|window                                    |datanode_count|
+------------------------------------------+--------------+
|[redacted]                                |[redacted]    |
|[redacted]                                |[redacted]    |
+------------------------------------------+--------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+------------------------------------------+--------------+
|window                                    |datanode_count|
+------------------------------------------+--------------+
|[redacted]                                |[redacted]    |
|[redacted]                                |[redacted]    |
|[redacted]                                |[redacted]    |
+------------------------------------------+--------------+

-------------------------------------------
Batch: 3
-------------------------------------------
+------------------------------------------+--------------+
|window                                    |datanode_count|
+------------------------------------------+--------------+
|[redacted]                                |[redacted]    |
|[redacted]                                |[redacted]    |
+------------------------------------------+--------------+

-------------------------------------------
```

5. Aggregating Data (6 points)

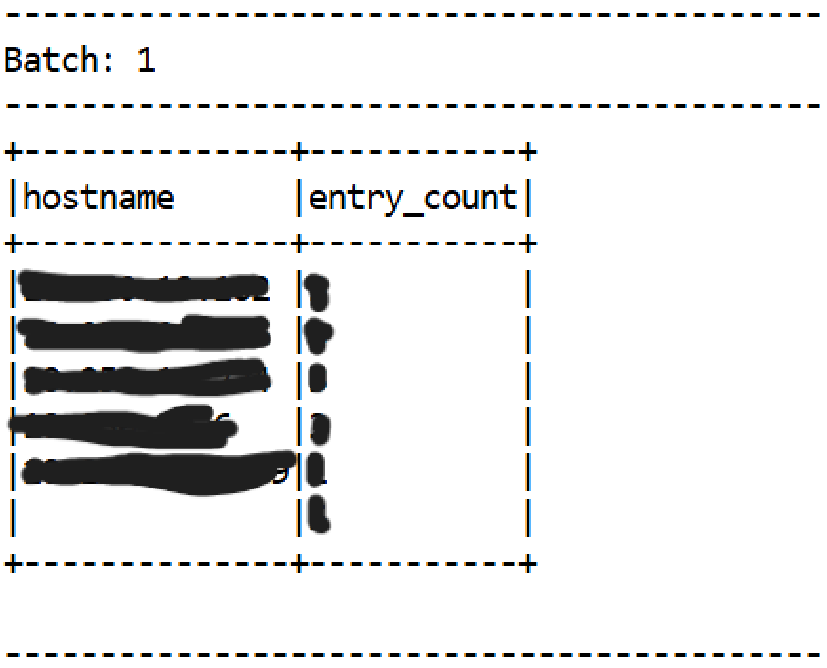
- Group the dataset by `hostname` and aggregate the bytes transferred for each host.
- Name the new column as `total_bytes` and sort the results by `total_bytes` in descending order.
- Write the output to the console in complete mode, ensuring that fields are not truncated.
- Include a screenshot of the output for batch 4 in your report. For example:



```
-------------------------------------------
Batch: 4
-------------------------------------------
+--------------+-------------------+
|hostname      |total_bytes        |
+--------------+-------------------+
|              |3284314912390149923|
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
|    ******    |    ***********    |
+--------------+-------------------+
```

6. Filtering and Triggering (6 points)

- Create a query that filters the incoming log data to include only `INFO` entries with a specific block ID pattern (e.g., `blk_`).
- Group the data by `hostname` and count the number of such entries for each host.
- The query should trigger every 15 seconds, outputting the results to the console in complete mode, ensuring that fields are not truncated.
- Include a screenshot of the output for batch 5 in your report. For example for the batch 1:



```
-------------------------------------------
Batch: 1
-------------------------------------------
+--------------+-----------+
|hostname      |entry_count|
+--------------+-----------+
|[redacted]    |[redacted] |
|[redacted]    |[redacted] |
|[redacted]    |[redacted] |
|[redacted]    |[redacted] |
|[redacted]    |[redacted] |
|              |[redacted] |
+--------------+-----------+

-------------------------------------------
```