Spark / Scala Exercise

Create a spark application written in scala that reads in the provided signals dataset, processes the data, and stores the entire output as specified below.

For each entity_id in the signals dataset, find the item_id with the oldest and newest month_id. In some cases it may be the same item. If there are 2 different items with the same month_id then take the item with the lower item_id. Finally sum the count of signals for each entity and output as the total_signals. The correct output should contain 1 row per unique entity_id.

Requirements:

- Create a Scala SBT project
- Use the Spark **Scala API** and Dataframes/Datasets
    - Do not use Spark SQL with a sql string!
    - **If you write** `spark.sql(`"select ….") **you are doing it wrong!!**
- Output format is Parquet
- Produce a **single parquet output file**, regardless of parallelism during processing
- Use window analytics functions to compute final output in a single query

Input:

entity_id: long
item_id: integer
source: integer
month_id: integer
signal_count: integer

Output:

entity_id: long
oldest_item_id: integer
newest_item_id: integer
total_signals: integer

Example partial output:

entity_id	oldest_item_id	newest_item_id	total_signals
359781	3	3	23

152813413	1000	1000	2
224619015	0	3	12
...			

Submission Guidelines

Please submit the entire scala sbt project with all code necessary to build and run the app. You can bundle it as a zip, tarball or github link. Also include a copy of the output file that is generated from running the app.