

Job 0 - Завантаження CSV файлу.

Job 1 - Фільтрація, вибір стовпців, групування та підрахунок.

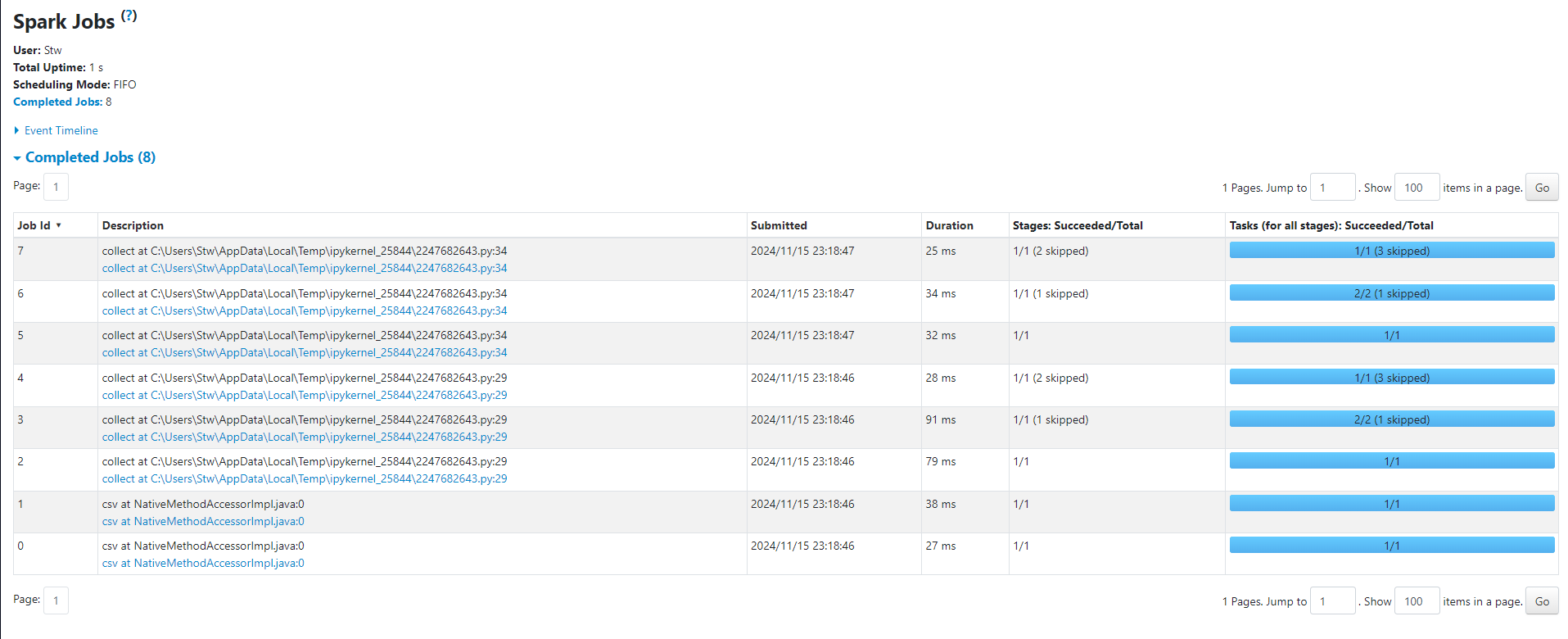
Job 2 - Перше виконання collect().

Job 3 - Додаткове фільтрування за count > 2.

Job 4 - Друге виконання collect().

Job 5 - Кешування даних.

Job 6-7 - Виконання додаткових collect() після кешування (зменшено за рахунок кешування).



Додавання проміжної дії змусило Spark виконати додаткові етапи обробки даних для підготовки результатів, що й створило три додаткові Jobs

Job 0 - Завантаження CSV файлу з диску.

Job 1 - Виконання фільтрації та групування, перше застосування collect().

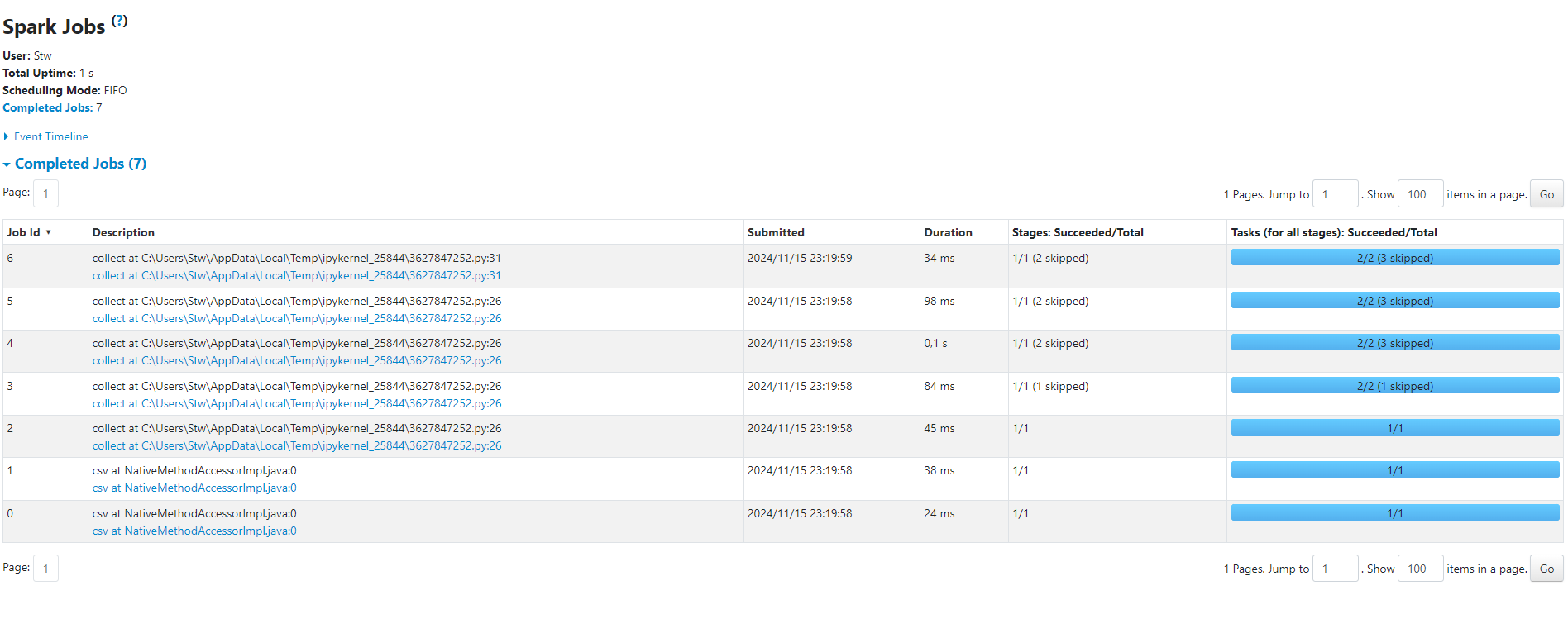
Job 2 - Збір результатів після першого collect().

Job 3 - Застосування додаткового фільтрування, друге застосування.

Job 4- Збір результатів після другого collect().

Job 5 - Використання кешування для прискорення наступних операцій, третє застосування collect(). Job 6 - Завершення обробки даних з кешуванням, четверте застосування collect().

Job 7 - Підсумкове обчислення та збір даних після кешування.



Використання cache() зменшило кількість Jobs до 7, оскільки воно зберігає проміжні результати в пам'яті, щоб уникнути повторних обчислень.

Job 0 - Завантаження CSV файлу з диску.

Job 1- Виконання першої операції collect().

Job 2 - Виконання другої операції collect().

Job 3 - Виконання третьої операції collect().

Job 4 - Виконання четвертої операції collect().

Job 5 - Виконання п'ятої операції collect().

Job 6 - Завершення операції collect() після кешування.