Webcrawler

...

“Scraping the web on the budget”

Jarosław Kosowski

Robert Kowalski

Mateusz Szynka

Charakterystyka i wymagania

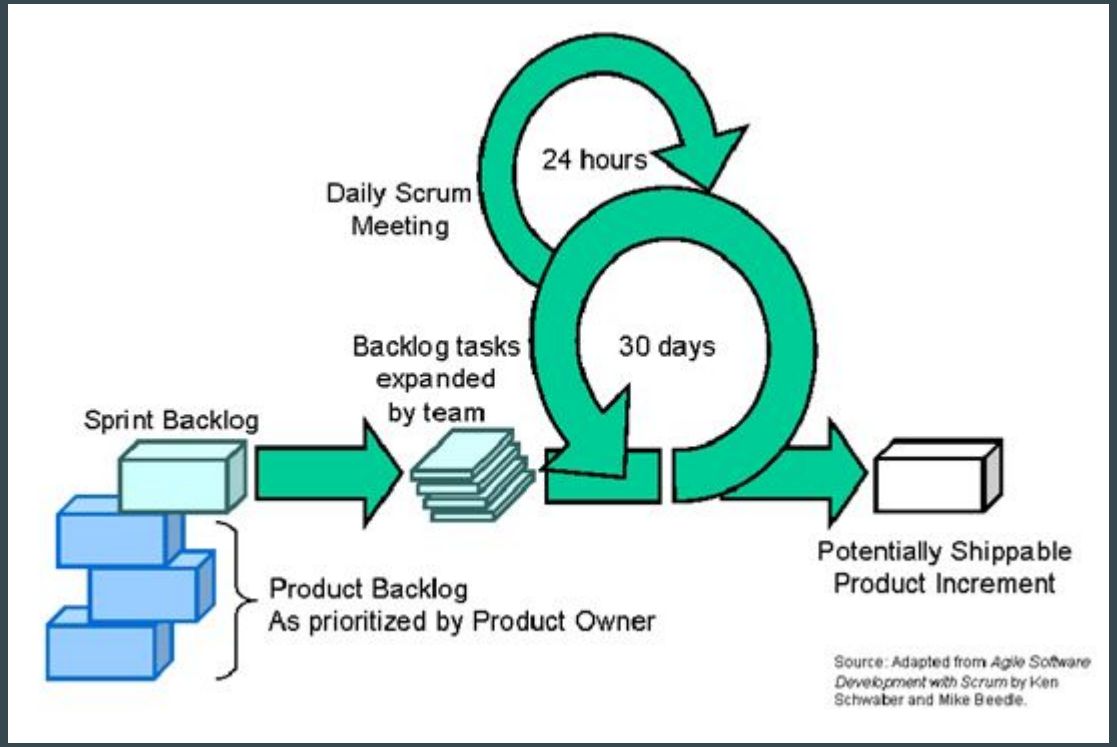
- ➔ automatyczne crawlowanie po serwisie od zadanego linku
- ➔ jednolita struktura wyników
- ➔ rozszerzalność źródeł danych (szablony)

Metodologia pracy

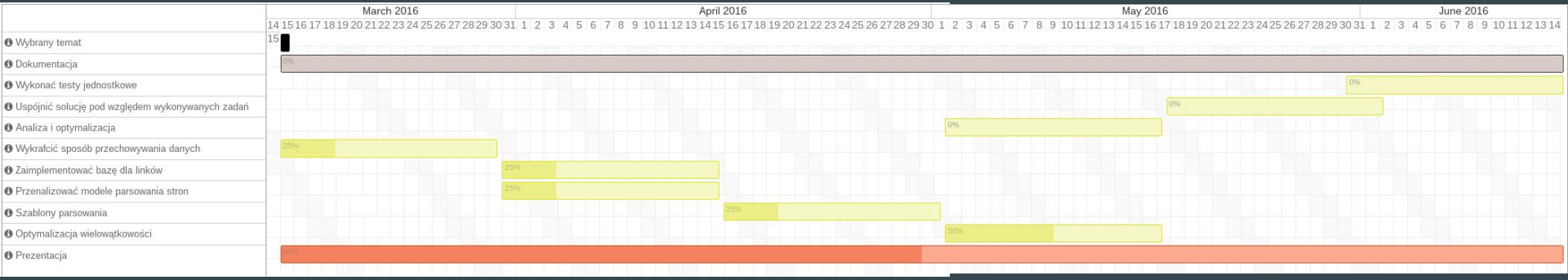
Scrum

Sprinty 2-tyg

Dynamiczny podział zadań

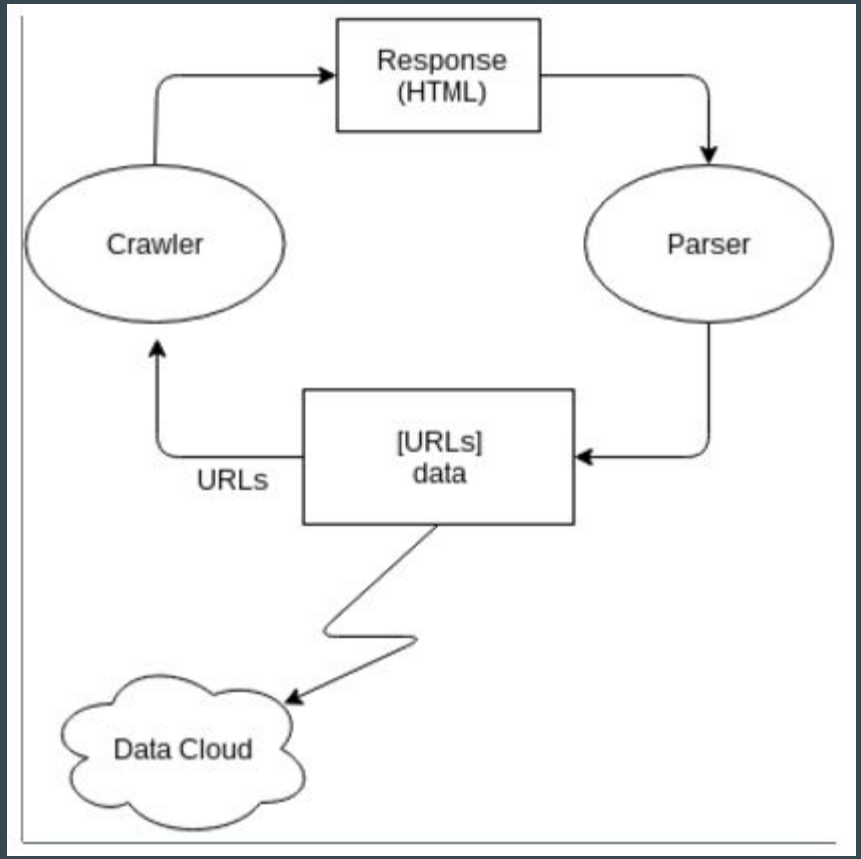


Harmonogram działań



Przeszukiwanie zasobów

1. Crawl
2. Parse
3. Commit



Szablony danych

Rozszerzanie przeszukiwania

Format XML/JSON

Możliwość generacji szablonów na podstawie stron

KEEP
CALM
AND
HOPE FOR
THE BEST