# Apache Pig



## *Agenda*

1. Apache Pig
2. Struktury danych w Pig

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
Człowiek – najlepsza inwestycja
Projekt nr POKL.04.03.00-00-131/12 „Inżynieria wiedzy dla inteligentnego rozwoju"
współfinansowany przez Unię Europejską w ramach środków Europejskiego Funduszu Społecznego
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
MATERIAŁY DYDAKTYCZNE I SZKOLENIOWE NA STUDIA PODYPŁOMOWE ORAZ NA SZKOLENIA DYSTRYBUOWANE SĄ BEZPŁATNIE

# *Apache Pig*

- język wysokiego poziomu,
- kod jest kompilowany do sekwencji zadań MapReduce,
- może być łatwo rozszerzany.



# *Metody uruchamiania Pig*

- tryb interaktywny działający lokalnie
  pig/bin>pig -x local
- tryb interaktywny działający jako MapReduce
  pig/bin>pig -x
- tryb uruchamiania skryptów
  pig/bin>pig -x local myscript.pig
  pig/bin>pig -x muscript.pig

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
Człowiek – najlepsza inwestycja
Projekt nr POKL.04.03.00-00-131/12 „Inżynieria wiedzy dla inteligentnego rozwoju"
współfinansowany przez Unię Europejską w ramach środków Europejskiego Funduszu Społecznego
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
MATERIAŁY DYDAKTYCZNE I SZKOLENIOWE NA STUDIA PODYPŁOMOWE ORAZ NA SZKOLENIA DYSTRYBUOWANE SĄ BEZPŁATNIE

## *Typy danych w Pig*

- typy skalarne
  int 24,
  chararray tewnty-four etc.
- krotki
  <15, Bob, 10.06>
- bag – grupy krotek
  <15, Bob, 10.06>, <John, 2>
- mapy
  [ 'apache' : <'search', 'news'> ; 'cnn' : 'news' ]



## *Wyrażenia w Pig*

- wyrażenia produkują relacje (zazwyczaj)
  - relacja to bag, który posiada nazwę,
- wyrażenia pobierają jako dane wejściowe relację (zazwyczaj),

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
Człowiek – najlepsza inwestycja
Projekt nr POKL.04.03.00-00-131/12 „Inżynieria wiedzy dla inteligentnego rozwoju" współfinansowany przez Unię Europejską w ramach środków Europejskiego Funduszu Społecznego
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
MATERIAŁY DYDAKTYCZNE I SZKOLENIOWE NA STUDIA PODYPLOMOWE ORAZ NA SZKOLENIA DYSTRYBUOWANE SĄ BEZPŁATNIE

## *Wczytywanie danych*

Operator LOAD. Może wczytywać różne typy danych:

- BinStorage()
- PigStorage()
- TextLoader()
- JsonLoader()

A = LOAD '/datadir/datafile' using PigStorage()

A = LOAD '/datadir/datafile' using PigStorage() as (f1:int, f2:chararray)



## *Operatory relacyjne*

- FILTER
  b = filter data by pubyear = 2003;
- SORT
  c = sort data by author ASC;
- GROUP
  grupuje krotki wg określonego klucza
- FOREACH
  tworzy nową na bazie istniejącej
- FLATTEN
  usuwa zagnieżdżenia z krotki

# *Demo*

WordCount w Pig