Apache Hadoop

Agenda

1. Co to jest Hadoop?
2. System plików HDFS
3. Wprowadzenie do Map-Reduce

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

Apache Hadoop

Środowisko pozwalające na:

- rozproszone przetwarzanie
- dużych zbiorów danych
- w klastrach składających się z niedrogich komputerów
- z wykorzystaniem prostego modelu programowania.

Składniki Apache Hadoop

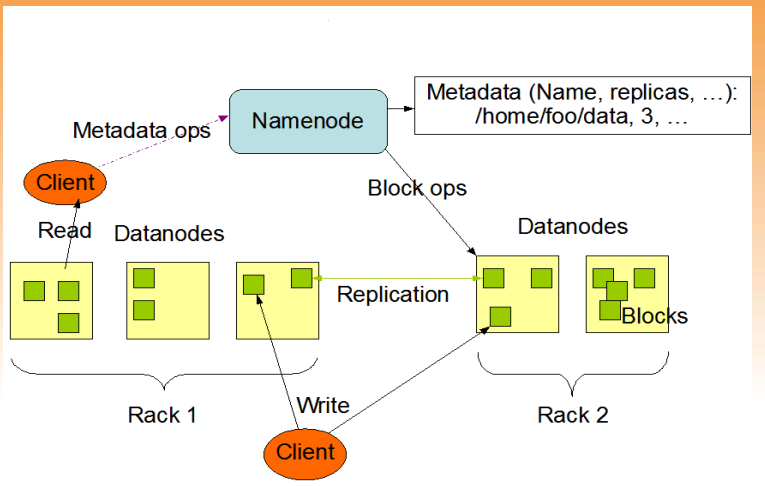
- Hadoop Common,
- Hadoop Distributed File System (HDFS™),
- Hadoop YARN,
- Hadoop MapReduce.

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

Hadoop Distributed File System (HDFS)

- odporność na awarie sprzętu,
- strumieniowy dostęp do danych,
- duże zbiory danych,
- zwarty model danych,
- „Przenoszenie obliczeń jest tańsze, niż przenoszenie danych”,
- przenośność.

Architektura HDFS



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

Cechy HDFS

- przestrzeń nazw systemu plików,
- replikacja danych,
- trwałość metadanych systemu plików,
- uszkodzenia DataNode,
- spójność danych,
- usuwanie plików.

Dostęp do HDFS

- FS shell
  - bin/hadoop fs -mkdir /foodir
  - bin/hadoop fs -cat /foodir/myfile.txt
- DFSadmin
  - bin/hadoop dfsadmin -safemode enter
- przeglądarka www

Co to jest MapReduce?

- to paradygmat programowania pozwalający na równoległe wykonywanie zadań, będący centralną częścią Hadoop,
- składa się z dwóch zadań niezależnie wykonywanych przez klaster: map i reduce,
- map – przekształca dane wejściowe w krotki,
- reduce – przekształca dane pochodzące z map.

Przykład MapReduce (1)

...
Toronto, 20
Whitby, 25
New York, 22
Rome, 32
Toronto, 4
Rome, 33
New York, 18
...

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

Przykład MapReduce (2)

- dane zwrócone przez pierwsze zadanie map

(Toronto, 20) (Whitby, 25) (New York, 22) (Rome, 33)

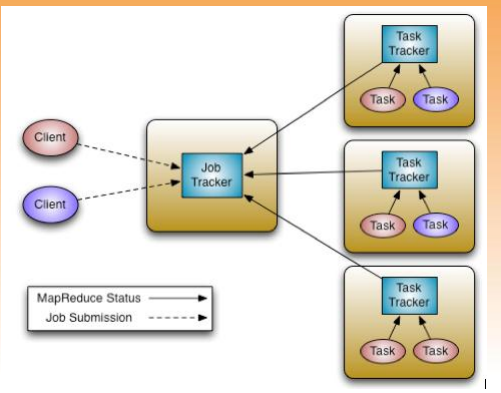
- dane zwrócone przez pozostałe zadania map

(Toronto, 18) (Whitby, 27) (New York, 32) (Rome, 37)(Toronto, 32) (Whitby, 20) (New York, 33) (Rome, 38)(Toronto, 22) (Whitby, 19) (New York, 20) (Rome, 31)(Toronto, 31) (Whitby, 22) (New York, 19) (Rome, 30)

- dane zwrócone przez zadanie reduce

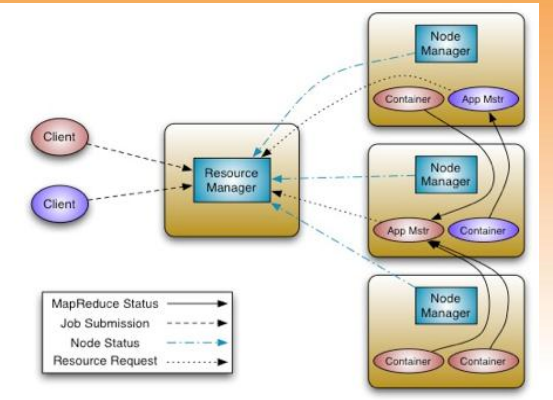
(Toronto, 32) (Whitby, 27) (New York, 33) (Rome, 38)

Architektura MapReduce1 (MR1)



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

Architektura MapReduce (MR2)



Kto korzysta z Hadoop?

- Yahoo

Pierwsze duże wdrożenie Hadoop w systemie produkcyjnym (2010) – klaster składający się z 10000 procesorów.

- Facebook

Największy (?) klaster Hadoop na świecie – 100PB w 2012, 1PB nowych danych dziennie w 2013

- 50% firm z listy Fortune 500 korzysta z Hadoop (2013)

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