Politechnika Wrocławska

„Przetwarzanie danych masowych”

Prowadzący

dr inż. Mateusz Tykierko

- Pokój: **bud. D2 pok. 101/6 (domofon guzik 3)** siedziba WCSS
- Email: mateusz.tykierko@pwr.edu.pl
- Konsultacje: wt. 11.15-13.00 bud. D2 101/6 – powiadomienia emailem
- Telefon: 71 320 20 32

Politechnika Wrocławska

Zawartość kursu

- Mechanizmy systemów operacyjnych
- Narzędzia do przetwarzania danych tekstowych
- Programowania równoległego i rozproszone
- Klastry, gridy, chmury obliczeniowe
- Języki i platformy przetwarzania danych masowych
- Algorytmy rozproszone dla przetwarzania masowych danych – macierzy, grafów, sieci, metody aproksymacji

Politechnika Wrocławska

Data mining

Eksploracja danych, drążenie danych, pozyskiwanie wiedzy, ekstrakcja danych

Data dredging

Znajdowanie „modelu” danych

Odkrywanie wiedzy z danych.

- Modele statystyczne
- Uczenie maszynowe
- Generalizacja danych
- Wydobywanie cech

Politechnika Wrocławska

Big data perspektywy

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it... - Dan Ariely

Politechnika Wrocławska

Big data perspektywy

- Torture the data, and it will confess to anything. *Ronald Coase*,
- Liczby nie kłamią, kłamcy liczą - Charles H. Grosvenor
- To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. — *Ronald Fisher*
- In God we trust. All others must bring data. — *W. Edwards Deming*,

Politechnika Wrocławska

Źródła danych

- Repozytoria danych
  - Lokalne
  - Specjalistyczne
  - Ogólne
  - Krajowej
- open data, linked open data
- Frictionless data, tabular data on the web
- 5 stars of open data
- Dostęp do danych dla celów naukowych
- Dostęp do informacji publicznej

Politechnika Wrocławska

Środowisko pracy

- System operacyjny z rodziny Linux
- Klaster obliczeniowy Bem
- System wersjonowania kodu

Politechnika Wrocławska

Klaster Bem

- Login z przedrostkiem s np. s123441 i hasło jak do poczty studenckiej
- Węzeł dostępowy: ui.wcss.pl
- Protokoły dostępu: ssh, NX, sftp
- Helpdesk: kdm@wcss.pl
- Strona informacyjna: kdm.wcss.wroc.pl

Politechnika Wrocławska

Powłoka bash

- Język programowania
- Historia
- Automatyczne uzupełnianie
- Pliki startowe
- Zmienne środowiskowe
- Aliasy
- Wejścia i wyjścia
- Potoki

Politechnika Wrocławska

Powłoka bash

- Tab – uzupełnianie składni
- ctrl-a – początek wiersza
- Ctrl – e – koniec wiersza
- Ctrl – c – zabij obecnie uruchomiony proces (sygnał SIGINT)
- Ctrl – s – zatrzymaj program (SIGSTOP)
- Ctrl – d – wznów proces
- Ctrl –d – koniec wejścia (sesji, EOF)
- Ctrl - _ - cofnij edycję
- Ctrl – u wytnij przed kursorem, ctrl-k wytnij pozostała część linii za kursorem, ctrl –t – zamień miejscami litery, ctrl –w – skasuj słowo przed, alt – w – skasuj słowo za
- Ctrl – r – wyszukiwanie w historii
- Shift page up, page down – przeglądanie bufora terminala

Politechnika Wrocławska

Metadane pliku

- Nazwa pliku – konwencja nazewnicza
- Prawa dostępu
  - User
  - Group
  - Other
  - ACL
  - r – read, w- write, x – executable, s - setuid
- Typ pliku (-,l,b,c,d)
- Daty (utworzenia, modyfikacji, ostatniego dostępu)
- Rozmiar
- Nazwa
- Polecenie chmod, stat, file

Politechnika Wrocławska

Substytucja poleceń

- Znaki *, ?, [ ], { }, [!]
- Znak $
- Wyłącznie substytucji „”, ’ ’, \
- Polecenie echo

Politechnika Wrocławska

Wejścia wyjścia

- Standardowe wejście
- Standardowe wyjście
- Standardowe wyjście błędów

- Przekierowania
  - > >>
  - 2> 2>&1
  - < <<

Politechnika Wrocławska

Zmienne środowiskowe

- Do czego służą
- env, set
- Przypisywanie wartości
- Lokalność
- Znaki ``
- Interesujące zmienne PATH, HOME, LANG, USER, TMPDIR, LD_LIBRARY_PATH

Politechnika Wrocławska

Skrypty

- Nagłówek zawiera ścieżkę bezwzględną do interpretera
  - #!/bin/bash
  - #!/usr/bin/python
  - /usr/bin/env python
  - /usr/bin/env r
- Plik jest wykonywalny i ma prawa do odczytu r,x
- Zawiera ciąg poleceń/komend zakończony znakiem końca pliku
- Wywołanie ./nazwa_skryptu, interpeter nazwa_skryptu

Politechnika Wrocławska

Potoki, sekwencje, statusy wyjścia

- |
- Sekwencje poleceń
  - &&
  - ||
  - ;

- Status wyjścia $? – wartość 0 - poprawna
- Liczba parametrów wywołania $#
- Lista parametrów wywołania $@ - narzędzie getopt
- Generowanie kodu wyjścia
  - exit 123

Politechnika Wrocławska

Kontrola procesów

- &
- bg, fg
- jobs
- kill
- ps
- trap

Politechnika Wrocławska

SSH

- Uwierzytelnianie po kluczach
- Montowanie zdalnego zasobu
- Tunelowanie portów
- Ssh agent
- Sftp – winscp,filezilla
- Montowanie dysków (sshfs, sftp netdrive)

Politechnika Wrocławska

Narzędzia - pliki

- FUSE – file system in user space
  - httpfs,ftpfs,gmailfs,minfs,wikipediafs
- find
  - czasy, uprawnienia, wielkość, typy
  - warunki, akcje
- cat,less,head,tail
- stat
- touch, ln
- tar, cpio, cp, mv
- bzip,gzip,7z
- rm, snapshot
- screen, time

Politechnika Wrocławska

Narzędzia – przetwarzanie tekstu

- Znaki niedrukowalne – tryb binarny edytora, less -r
- Kodowanie znaków diakrytycznych - iconv
- wc, join, split
- sort (-r,-n,-k), uniq (-c)
- grep (-e,-f,-v,-i,-c)
- diff,cmp,comm
- md5sum, openssl
- cut,paste
- hexdump, strings, od
- xargs, eval

Politechnika Wrocławska

Wyrażenia regularne

- . - jakikolwiek znak
- [...] - jakikolwiek wylistowany znak
- [^...] - żaden z wylistowanych znaków
- ^ - początek linii
- $ - koniec linii
- \< - początek słowa
- \> - koniec słowa
- | - alternatywa
- (...) - grupowanie
- Powtórzenia *,+,?,{m,n}
- [[:space:]],[[:digits]],
- www.regular-expressions.info

Politechnika Wrocławska

Przykłady

- sep[ea]r[ea]te - separete, seperete itp.
- ^From|Date|Subject: or ^(From|Date|Subject)
- July Jul 4th fourth 4 (July|Jul) (July?) (4th|4) (fourth|4(th)?)
- zegar 24h 0?[0-9]|1[0-9]|2[0-3] lub [01]?[0-9]|2[0-3]
- ([a-z]+) \1

Politechnika Wrocławska

Narzędzia do przetwarzania tekstu

- sed – stream editor for filtering and transforming text
- awk - pattern scanning and processing language

Politechnika Wrocławska

sed

- sed 's/^M$//'
- sed 's/\([a-z]+\) \1/\1/, - skasuj powtarzające się słowa
- sed 's/scarlet/red/g;s/ruby/red/g;s/puce/red/g, – zamień parami
- sed -e :a -e '$!N; s/\n/:/; ta, – połącz linie
- sed 10q – pokaż 10 linie
- sed '/^$/d' – skasuj puste linie

Politechnika Wrocławska

awk

- Gawk: Effective AWK Programming
- Pole, rekord
- /patten/ {action}
- BEGIN
- END
- zmienne wbudowane
- tablice asocjacyjne
- Instrukcje

Politechnika Wrocławska

Czas i utylizacja zasobów

- time
- ps
- top
- /proc

Politechnika Wrocławska

Środowisko graficzne

- X-window (X11)
  - Window managery
  - Klienci X-window
  - Serwer X-window – Xming (pod MS Windows)
  - Tunelowania X11 przez ssh (opcja w putty, -Y polecenie ssh)
  - Font serwer
  - Geometria okien, kolory, zachowanie
  - Wirtualne pulpity
  - Zmienna DISPLAY
- NoMachine ,TightVNC

Politechnika Wrocławska

Wizualizacje

- gnuplot
- chart