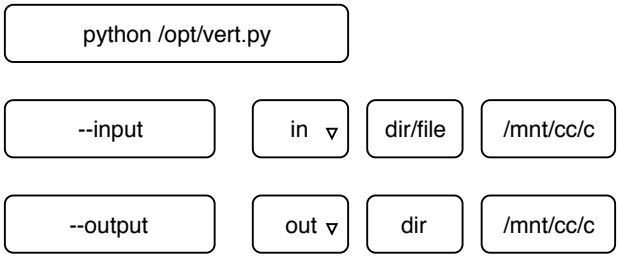
Zápis z konzultácie
Aplikácia pre paralelné spracovanie dat
KNOT

Katarína Galanská

05.06.2019

Parametre

procházení starých jobů



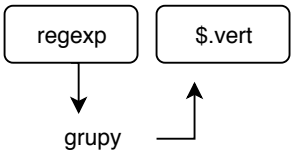
- obecné (nie viazané na konkrétne kroky)
- Je potreba vedieť kontrolovať, či ku všetkým vstupným súborom existujú súbory vo výstupnej složke. Možnosť vypnúť alebo zapnúť kontrolu.
- Vedieť predhadzovať vstup v podobe
  - súboru
  - složky
  - všekých složek ve složce
- Kontolola vstupných a výstupných súborov bude spustená ako ďalší krok
- musí byť možný výber kontrol, kontroly sú typu:
  - kontrola na logy (veľkosť logu)
  - kontrola výstupných súborov
- typy parametrov:
  - conf, in, out, log, err

GUI Prehľad

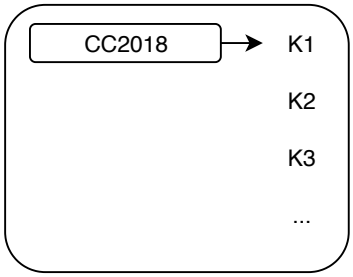
- zobrazovať zoznam n serverov
- zobrazovať úlohy na strojoch a vidieť jednotlivé kroky, ktoré prešli

Paralelizmus

- spustiť n-tý krok, až keď všetky servery budú mať hotový n-1-tý krok
- krok spustenia hashov je prerekvizita
- v kroku definovať v koľkých beží procesorech
- 2-3-5 procesov na krok
- riešiť timeouty - 1 proces beží výrazne dlhšie, po rozumnej dobe proces vypnúť
- keď dojde k timeoutu, vypne sa proces, celý krok končí chybou, do ďalšieho kroku nepostúpim
- 2 možnosti správania, keď neexistje výstupný súbor
  - rerun
  - skus to znova
- špatný súbor zmazať a spustiť iba nad súbormi, ktoré nemajú výstupné súbory
- spustiť s tým, že len dospracujem súbpry, ktoré nemajú výstupné súbory
- výstupný súbor nemusí byť složka, ale aj súbor
- Regulárnym výrazom vyberiem názov súboru, rozseknem ho na grupy a tým nahradím meno súboru



- konfigurácia uložená v databázi
- prvé naklikávanie je pracné, potom načítam konfiguráciu



Štatistiky

- ako dlho trvá vertikalizace na súbor?
- koľko RAM?
- doba behu?
- konkrétne joby, krok č.1 spracovával 1 súbor priemerne 10 minút, celkovo trval x hodin
- štatistiku spracovania na 1 súbor je potrebná pre ladenie nástrojov, 95% trvá dva dni a zbytok deň
- čaš na súbor , na složku, RAM (priemerný čas/okamžitý čas, vzorkovať a priemerovať, nie je potrebná 100% presnosť)
- graf s počtom súborov spracovaný za daný čas
- čas uvedený ako časový interval alebo len dĺžka intervalu
- stav serverov (farebne rozlíšiteľné)
- vyťaženie serverov (netreba zaťaženie RAM)
- prehľad jobov, beží tento job? beží tento krok? koľko súborov je spracovaných z tohto kroku?