Hadoop Didelių duomeų apdorojimas studentišku biudžetu

Maksim Norkin, ISIfm-13

Vilniaus Gedimino technikos universitetas Fundamentinių mokslų fakultetas Informacinių Sistemų katedra maksim.norkin@ieee.org

2013 m. spalio 6 d.





Susipažinimas

- Programuotojas@Ruptela
- Projektai:
 - Žmogaus eisenos atpažinimas
 - Parkinsono ligos atpažinimas
 - Akcijų biržos analitikų spėjimų analizė;

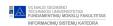




Istorija

Istorija

- Google parašė popierių apie MapReduse [1];
- Doug Cutting and Mike Cafarella 2005 išleido pirmą Hadoop versiją Java pagrindu;





Hadoop



- Kas yra Hadoop?
 - Norima atlikti kažkokius skaičiavimus su daug, daug duomenų;
 - Labai maži skaičiavimo resursai (keli seni kompiuteriai, renkantis dulkes ofiso kampe);
- Susideda:
 - HDFS
 - MapReduce





HDFS

- Reali bylų sistema, parašyta Java pagrindu;
- Hadoop Distribuded File System;
- Lanksti masto didinime (angl. scalability);
- Pateikti pavyzdį ant lentos;





${\sf MapReduce}$

- Lengvas sprendimas, norint apdoroti didelius duomenys su daug procesorių;
- Pagrindiniai tikslai, realizacijos metu:
 - Klaidų tolerancija;
 - Lanksti masto didinime (angl. scalability);
 - Automatinis lygiagretumas ir paskirstymas;





- I/O key/value rinkinys
 - \blacksquare < key, value >:
 - < userId, profile >:
 - < timestamp, log entry >;
- Programavimas vyksta naudojant dvi primityvias funkcijas
 - map (in_key, in_value) → list(out_key, intermediate_value)
 - reduce (out_key, list(intermediate_value)) → list(out_value)
- Pateikti pavyzdj ant lentos;





Privalumai

- Labai paprastas ir lengvas naudoti;
- Jokios priklausomybės nuo schemos ir duomenų tipų;
- Jokio skirtumo nuo duomenų talpinimo lygio;





Trūkumai

- MapReduce nėra aukšto lygio programavimo modelis;
- Jokios schemos ir indeksų;
- Mažas efektyvumas I/O pagrindu;





Ekosistema I

- Pig
- DataFlow



- Hive
- SQL



- Sqoop
- RDBMS





Ekosistema II

- Mahout
- Klasterizavimas, Savybių skaičiavimas







Ekosistema III

- Zookeeper
- Sinchronizacija, monitoringas, grupinis servisas







Kas naudoja?

- Adobe;
- Apple;
- Linkedin;
- Yahoo;
- Twitter;
- Beveik kiekvienas startup'as;



13 / 15

Vilnius



Ačiū

■ Klausimai?





Literatūra



Jeffrey Dean and Sanjay Ghemawat.

Mapreduce: simplified data processing on large clusters.

Commun. ACM, 51(1):107-113, January 2008.



