Hadoop
Didelių duomenų 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.

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

Susipažinimas

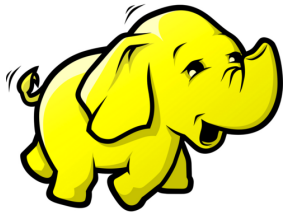
- Programuotojas@Ruptela
- Projektai:
  - Žmogaus eisenos atpažinimas
  - Parkinsono ligos atpažinimas
  - Akcijų biržos analitikų spėjimų analizė;

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

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

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

HDFS

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

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

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;

Programavimo modelis

- I/O – key/value rinkinys
  - $< key,\ value >$;
  - $< userId,\ profile >$;
  - $< timestamp,\ log\ entry >$;
- Programavimas vyksta naudojant dvi primityvias funkcijas
  - map (in_key, in_value) $\rightarrow$ list(out_key, intermediate_value)
  - reduce (out_key, list(intermediate_value)) $\rightarrow$ list(out_value)
- Pateikti pavyzdį 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

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
FUNDAMENTINIŲ MOKSLŲ FAKULTETAS
INFORMACINIŲ SISTEMŲ KATEDRA

Ekosistema II

- Mahout
- Klasterizavimas, Savybių skaičiavimas

Ekosistema III

- Zookeeper
- Sinchronizacija, monitoringas, grupinis servisas

Kas naudoja?

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

Ačiū

- Klausimai?

Literatūra

Jeffrey Dean and Sanjay Ghemawat.
Mapreduce: simplified data processing on large clusters.
Commun. ACM, 51(1):107–113, January 2008.