Aalto-yliopisto
Perustieteiden
korkeakoulu

Ontologiamuokkaimen käyttö laskentaklusterin tehokkuusanalyysissä
(valmiin työn esittely)

Santtu Klemettilä

25.03.2013

Ohjaaja: *FT Markopekka Niinimäki*
Valvoja: *Prof. Ahti Salo*

Systeemianalyysin
laboratorio

Tausta: Ontologia

- Data on yleensä hajanaista ja heterogeenistä
  - Tietokoneella ei ole älyä ymmärtää sille määrittelemätöntä dataa
- Yleisenä tavoitteena semanttinen web, jossa tarjolla on myös tiedon merkityksen kertova tieto
- Ontologia on kokonaisuus, jossa määrittyy sovellettavan aiheen käsitteet ja niiden väliset suhteet
  - Ei universaaleja ontologioita, vaan sovelluskohteille on määriteltävä tarkoituksenmukainen ontologia

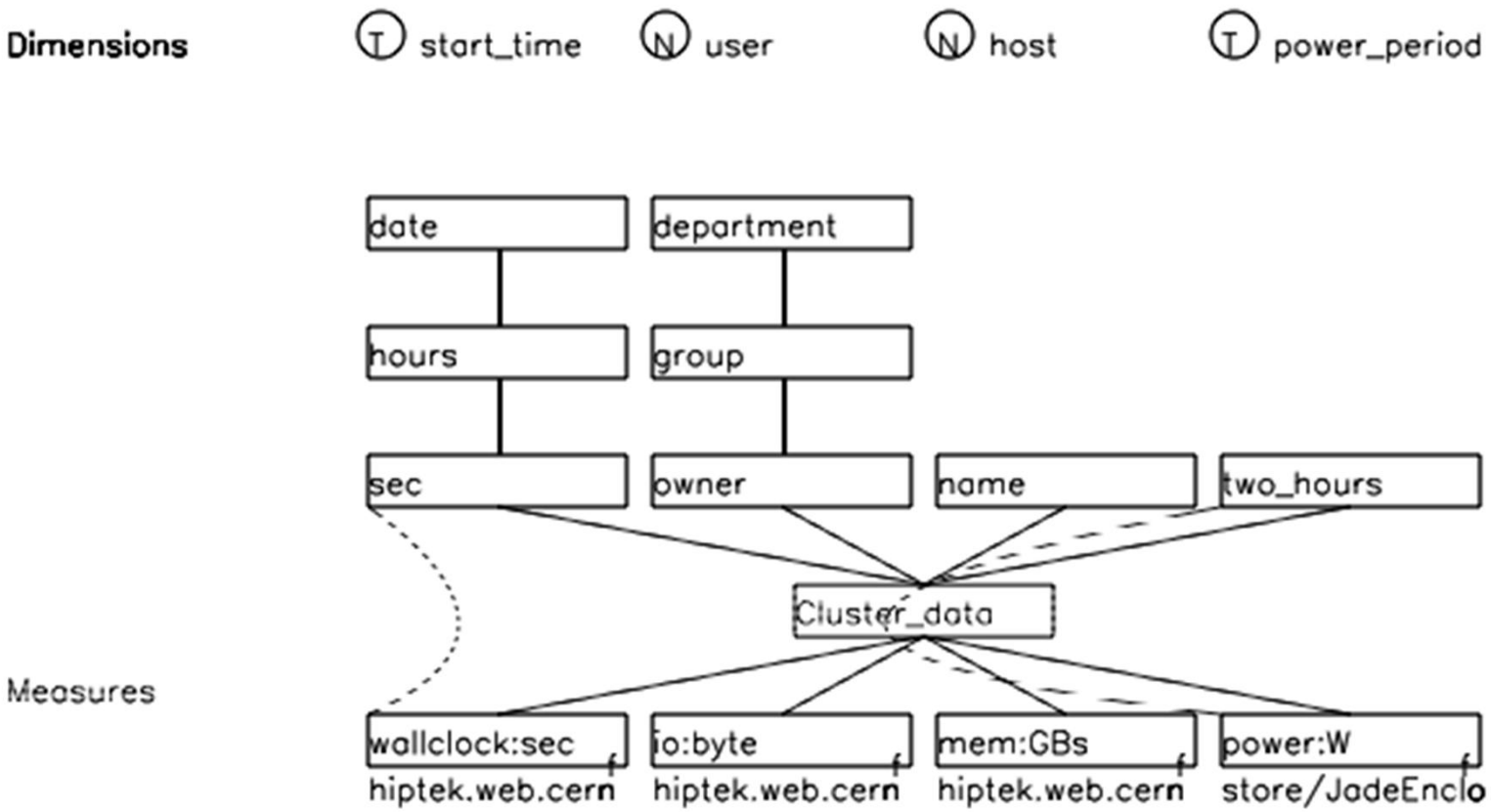
Tausta: Laskentaklusteri

- Analyysin kohteena CERNin tutkimustyössä käytettävä laskentaklusteri
  - Accounting-data koostuu laskentatöihin liitettävistä tietueista (ajankohta, muistinkäyttö, dataliikenne ym.)
  - Energiankulutusdata on osaklusterikohtaista kahden tunnin aikajaksoihin jaettua dataa
- Kahdentyyppisiä laskentatöitä
  - Monte Carlo –simulaatiot, muistinkäyttö suurta
  - Analyysityöt, dataliikenne suurta

Tavoitteet

- Ontologiamuokkaimen arviointi dataintegraatiovälineenä
  - Vaiheina laskentaklusteriontologian määritteleminen ja sen avulla datan tuonti analysoitavaksi
- Laskentaklusterin tehokkuusanalyysi
  - Tutkitaan laskentatöiden jakautumista ja laskentatöiden vaikutusta klusterin energiankulutukseen

Ontologian määrittely



Ontologiatieto (1/2)

- Data on tallennettu tiedostoihin merkein (tässä ':') ja rivinvaihdoin eroteltuna

  qname:hostname:group:owner:job_name:job_number:account:priority:submission_time:start_time:end_time:
  failed:exit_status:ru_wallclock
  arc:c26.local:ndgfops:ndgfops:arc_testjob_rls:349:sge:0:1268406240: 1268406246:1268406288:0:0:42

- Ontologiatieto tallentuu standardoituun RDF/XML-muotoon

```
<owl:Class rdf:about="#Measure"/>
    <olapcore2:Measure rdf:about="#wallclock">
      <olapcore2:hasMeasureUnit rdf:resource="sec"/>
      <olapcore2:hasMeasureSummarizabilityType rdf:resource="float"/>
      <olapcore2:hasDependencyOnDimension rdf:resource="start_time"/>
    </olapcore2:Measure>
```

Ontologiatieto (2/2)

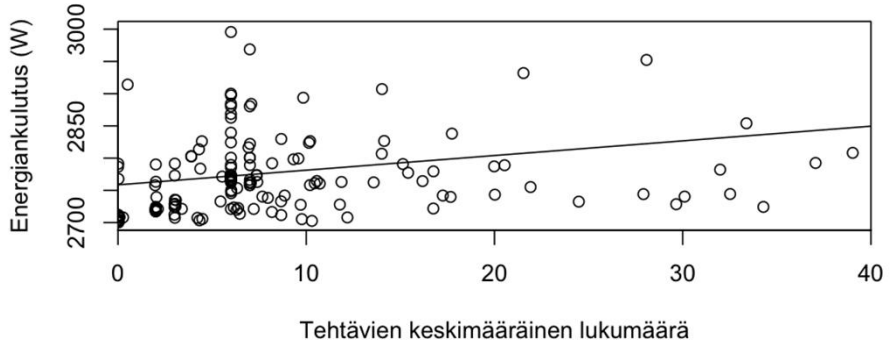
- Ontologian ulottuvuuksien ja mittojen yhteydet datalähteen nimikkeisiin tallentuu projektin omaan data_sources –muotoon

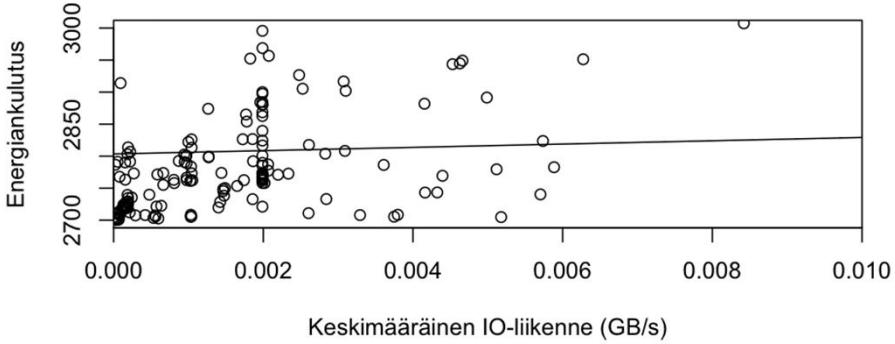
```
name="wallclock (sec)" source="http://hiptek.web.cern.ch/hiptek/ protected/accounting_feb2011_muok.txt"
    authentication="username:password" delimiter=":" variablename1="wallclock" variableget1="ru_wallclock”
```

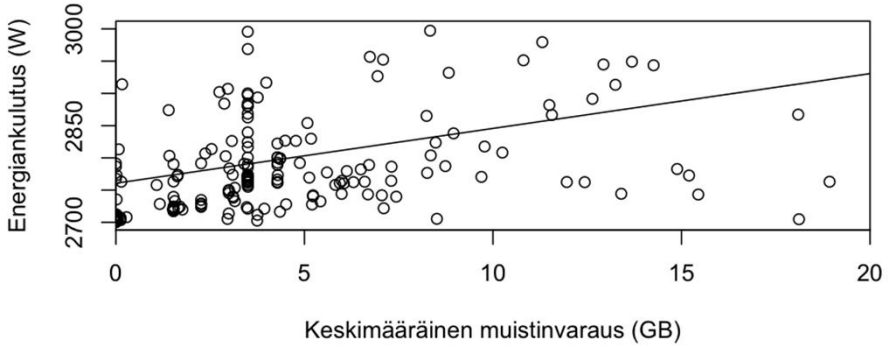
- Lopuksi muokkaimella voi tulostaa R-ohjelmistossa suoritettavan komentosarjan, jolla kartoitettu tieto saadaan helposti analysoitavaksi

```
measures=new.env(hash = TRUE, emptyenv())
require("RCurl")
data0<-
    read.table(textConnection(getURL("http://username:password@hiptek.web.cern.ch/hiptek/protected/accoun
    ting_feb2011_muok.txt")), sep=":", header=TRUE)
measures$wallclock <- data0[,c("ru_wallclock")]
```

Data-analyysi (1/2)





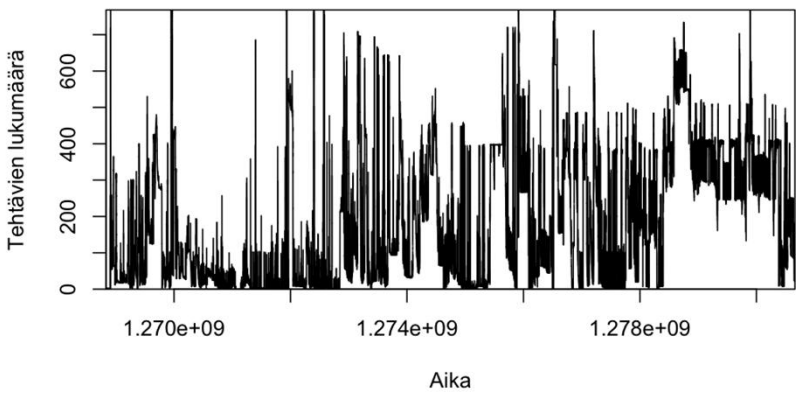


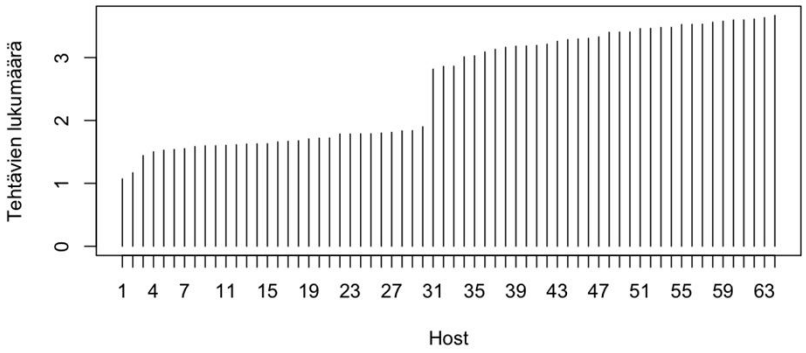
Selittävät muuttujat	R^2	Korjattu R^2	$p(\beta_{lkm})$	$p(\beta_{mem})$	$p(\beta_{io})$
lkm, mem, io	0,46	0,45	0,063	$2,6*10^{-7}$	$8,3*10^{-4}$
mem, io	0,45	0,45	-	$7,9*10^{-15}$	$6,69*10^{-4}$

Data-analyysi (2/2)

- Suurin osa energiankulutuksesta riippumatonta laskentakoneiden rasituksesta (vakiotermin perusteella 95 % energiasta kuluu koneiden päälläoloon)
- Jos laskentatyöt olisivat täydellisesti ositettavissa ja tasaisesti jakautuneita ajan suhteen, laskentakonevaatimus olisi 22 % nykyisestä määrästä eli potentiaalinen energiansäästö olisi noin 74 %.
- Tutkimalla laskentaklusterin käyttöastetta ja tarpeita, voidaan arvioida energiankulutuksen (ja muiden kustannusten) pienentämisen mahdollisuutta

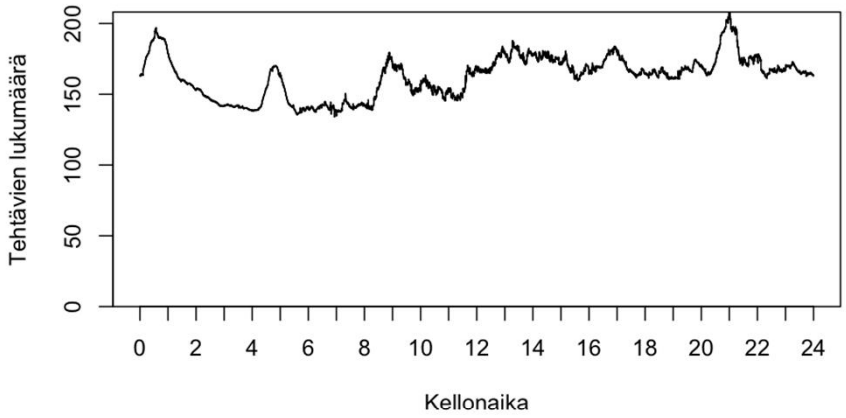
Laskentatöiden jakauma (1/2)

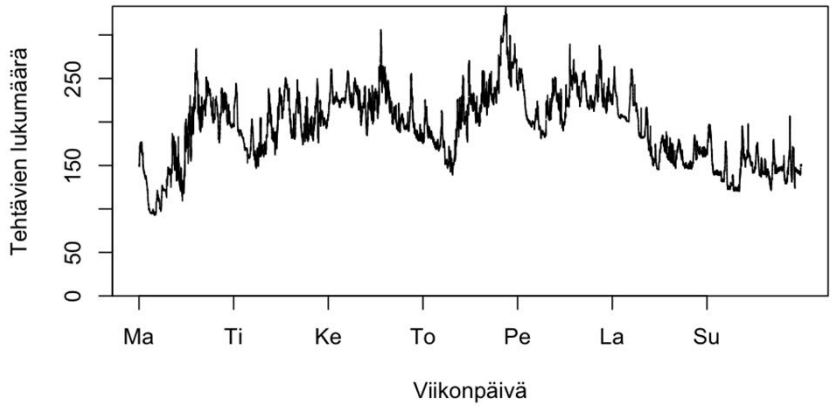




- Klusterin kapasiteetti saavutetaan vain silloin tällöin
- Käytön epätasaisuus mahdollisesti esimerkiksi työajoista johtuvaa
- Laskentakoneiden käyttö epätasaista

Laskentatöiden jakauma (2/2)





- Jakaumat kellonajan ja viikonpäivien suhteen suhteellisen tasaiset
- Varianssianalyysikään ei viittaa eroavaisuuksiin kellonaikojen tai viikonpäivien suhteen
- Kyseessä siis pidemmän aikavälin ongelma

Johtopäätökset

- Ontologiamuokkain onnistuu tavoitteessaan olla dataintegraatiota helpottava työkalu
  - Datan tuominen analysoitavaksi yksinkertaistuu
  - Erityyppisten datojen yhdistämisessä onnistutaan

- Laskentaklusterin tehokkuusanalyysistä selviää energiansäästömahdollisuus
  - Toteuttamisen järkevyys riippuu halutusta palvelutasosta
  - Käyttöasteen vaihtelevuudelle ei löydy säännöllistä tekijää

Viitteet

- Becket, Dave (toim.): RDF/XML Syntax Specification. http://www.w3. org/TR/REC-rdf-syntax/, 2004.
- Disease Ontology. http://disease-ontology.org, viitattu 31.7.2012.
- Guarino, Nicola: Formal Ontology, Conceptual Analysis and Knowled- ge Representation. International Journal of Human-Computer Studies, 43(5-6):625–640, 1995.
- Halevy, Alon, Anand Ra jaraman ja Joann Ordille: Data Integration: The Teenage Years. Teoksessa IN VLDB, sivut 9–16, 2006.
- Hendler, James, Tim Berners-Lee ja Eric Miller: Integrating Applications on the Semantic Web. Journal of the Institute of Electrical Engineers of Japan, 122:676–680, 2002.
- Resource Details for jade-cms.hip.fi. http://www.nordugrid.org/ monitor/clusdes.php?host=jade-cms.hip.fi&port=2135.
- Storage Element Performance Optimization for CMS Analysis Jobs. http://indico.cern.ch/getFile.py/access?contribId= 243&sessionId=8&resId=0&materialId=poster&confId=149557.
- Miller, Eric: An Introduction to the Resource Description Framework. Bulletin of the American Society for Information Science and Technolo- gy, 25:15–19, 1998.
- Niemi, Tapio, Santtu Toivonen, Marko Niinimäki ja Jyrki Nummenmaa: Ontologies with Semantic Web/Grid in Data Integration for OLAP. In- ternational Journal on Semantic & Web Information Systems, 3:25–49, 2007.
- Niinimäki, Marko ja Tapio Niemi: An ETL Process for OLAP Using RDF/OWL Ontologies. Journal of Data Semantics, JoDS XIII: Special Issue “Semantic Data Warehouses” , sivut 97–119, 2009.
- Niinimäki, Marko, Tapio Niemi, Stephen Martin, Jyrki Nummenmaa ja Peter Thanisch: Timely Report Production from WWW Data Sources. Teoksessa BIR 2011 Work-shops, LNBIP 106, sivut 184–195. Springer Berlin Heidelberg, 2012.
- Shoshani, Arie: Summarizability. Teoksessa Encyclopedia of Database Systems, sivut 2880–2884. Springer US, 2009.
- Ubuntu Manpage: Accounting - Grid Engine Accounting File For- mat. http://manpages.ubuntu.com/manpages/intrepid/man5/sge_ accounting.5.html, viitattu 1.8.2012.