jamk.fi

Avoin Data

Perttu Perankoski

Opinnäytetyö
Toukokuu 2016
Tekniikan ja liikenteen ala
Insinööri (AMK), Tietotekniikka

Jyväskylän ammattikorkeakoulu
JAMK University of Applied Sciences

jamk.fi

Kuvailulehti

Tekijä(t)	Julkaisun laji	Päivämäärä
Perankoski, Perttu	Opinnäytetyö, AMK	Toukokuu 2016
	Sivumäärä 52	Julkaisun kieli Suomi
		Verkkojulkaisulupa myönnetty: x

Työn nimi
Avoin Data

Tutkinto-ohjelma
Tietotekniikka

Työn ohjaaja(t)
Mika Rantonen, Antti Häkkinen

Toimeksiantaja(t)
JYVSECTEC – Jyväskylä Security Technology

Tiivistelmä

Opinnäytetyön toimeksiantajana toimi Jyväskylä Security Technology (JYVSECTEC)-hanke, joka toimii Jyväskylän ammattikorkeakoulun (JAMK) tiloissa Jyväskylässä. JYVSECTEC kehittää ja ylläpitää kyberturvallisuuden kehitysympäristöä, jota käytetään kehitys, tutkimus ja koulutuskäyttöön.

Työn teema oli tutustua big dataan, syventyen sen jälkeen avoimeen dataan ja eri tapoihin millä sitä voidaan hyödyntää. Työ koostui big data -ilmiön ja avoimen datan tutkimisesta, avoimien tietovarantojen kartoittamisesta ja datan analysoinnista käyttäen siihen suunniteltua analytiikkatyökalua.

Työssä käydään läpi avoimen datan tietovarantoja ja rajapintoja, kuten Digitraffic ja avoindata.fi -portaali. Tutkitaan, missä eri formaateissa käytettävä data on, sekä millaisia tekijänoikeuksia creative commons –lisenssi tarjoaa avoimeen dataan. Lisäksi perehdytään eri tekniikkoihin, joilla suuria datamääriä pystytään tehokkaasti hallitsemaan ja analysoimaan.

Opinnäytetyössä keskitytiin Hadoop -pohjaisiin ohjelmistoihin kuten MapR ja Apache Drill ja näiden osiin ja siihen miten toimivat, kuten HDFS, MapReduce. Ensimmäisenä toteutuksena oli MapR Sandbox For Hadoop -järjestelmän asennus ja sen käyttäminen Jyväskylän keskilämpötilan analysointiin. Toisena toteutuksena käytettiin Apache Drill -ohjelmistoa, jolla muunneettiin CSV-tiedosto Apache Parquet -muotoon.

Avainsanat (asiasanat)
Big data, Avoin data, Creative Commons, MapR, Hadoop.

Muut tiedot

jamk.fi

Description

Author(s)	Type of publication	Date
Perankoski, Perttu	Bachelor's thesis	May 2016
		Language of publication: Finnish
	Number of pages: 52	Permission for web publication: x

Title of publication
Open Data

Degree programme
Information Technology

Supervisor(s)
Mika Rantonen, Antti Häkkinen

Assigned by
JYVSECTEC – Jyväskylä Security Technology

Abstract

This Bachelor's thesis was assigned by Jyväskylä Security Technology (JYVSECTEC) project which operates in the JAMK University of Applied Sciences (JAMK) environment. JYVSECTEC develops and maintains closed cyber identity-based infrastructure for research, development and training-services.

In this thesis the goal was to explore big data, then take a deeper look at open data and how it could be utilized. The thesis contains investigation about the big data phenomenon, different web resources for open data and an analysis of that data using analytic tools.

The thesis covers different data resources and interfaces like Digitraffic and avoindata.fi portal. The research discusses the different formats that open data uses, and what kind of copyrights creative commons license offers for open data. Various technologies that can be used for efficient data management and analysis were also examined.

This thesis focuses on Hadoop based systems, such as MapR, Apache Drill, and different parts of these systems like HDFS and MapReduce. In the first implementation, MapR Sandbox For Hadoop was installed and used to analyze average temperature in Jyväskylä. In the second implementation, Apache Drill was installed and used to convert CSV file to Apache Parquet format.

Keywords/tags (subjects)
Big data, Open data, Creative Commons, MapR, Hadoop

Miscellaneous

SISÄLTÖ

Kuviot

Taulukot

LYHENTEET

AWS	Amazon Web Services
CC	Creative Commons
CC BY	Creative Commons Nimeä
CC BY-NC	Creative Commons Nimeä-EiKaupallinen
CC BY-ND	Creative Commons Nimeä-EiMuutoksia
CC BY-SA	Creative Commons Nimeä-JaaSamoin
CC BY-NC-SA	Creative Commons Nimeä-EiKaupallinen-JaaSamoin
CC BY-NC-ND	Creative Commons Nimeä-EiKaupallinen-EiMuutoksia
CSV	Comma Separated Value
DSPL	Dataset Publishing Language
GeoRSS	Geolocation Really Simple Syndication
GNU GPL	GNU General Public License
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
HRI	Helsinki Region Infoshare
HTML	Hypertext Markup Language
IoT	Internet of Things
JSON	JavaScript Object Notation
LAM	Liikenteen Automaattiset Mittauspisteet
MapR-FS	MapR File System
PDF	Portable Document Format
RDF	Resource Description Framework
REST	Representational State Transfer

RGCE	Realistic Global Cyber Environment
RSS	Really Simple Syndication
SQL	Structured Query Language
XML	Extensible Markup Language

1 TOIMEKSIANTAJA

Tämän opinnäytetyön toimeksiantajana toimi JYVSECTEC – Jyväskylä Security Technology, joka on Jyväskylän ammattikorkeakoulun Lutakon kampuksella toimiva kyberturvallisuusteknologian kehittämishanke. JYVSECTEC:in tavoitteena on tarjota asiakkailleen kyberturvallisuuteen liittyviä testaus, - koulutus- ja asiantuntijapalveluita. Palvelujen toteutusympäristönä toimii kyberturvallisuuden kehitysympäristö RGCE (Realistic Global Cyber Enviroment). RGCE:n avulla voidaan mallintaa oikean maailman verkkoympäristöä ja siihen kuuluvia palveluita, mutta turvallisesti omassa suljetussa ympäristössä. Verkko koostuu fyysisistä ja virtuaalisista laitteista, joiden on tarkoitus mallintaa internetin toimintaa. JYVSECTEC vastaa itse RGCE:n ylläpidosta ja kehitystehtävistä. (JYVSECTEC 2016).

Toiminta sisältää mm. verkottunutta yhteistoimintaa, koulutusta ja erilaista palvelutoimintaa ja näiden yhdistämisestä yhdeksi kokonaisuudeksi. JYVSECTEC pyrkii mahdollistamaan yhteistyökumppaneiden verkostoitumisen toisten alalla toimivien tahojen kanssa. (JYVSECTEC 2016).

JYVSECTEC–projekti käynnistyi syyskuussa 2011. Projektin tavoitteena on olla Suomen johtavia kyberturvallisuuden kehittämis- ja koulutuskeskuksia ja luoda Keski-Suomeen turvallisuusalan yritysten yhteistyöverkon. Projektilla pyritään kehittämään mm. yritysten turvallisuuden tietämystä, riskien hallintaa sekä turvallisuuden ylläpitoa. Osarahoittajina ovat toimineet Keski-Suomen Liitto ja Euroopan aluekehitysrahasto. Projektin on toteuttanut Jyväskylän ammattikorkeakoulun IT-instituutti. Sen kehittäminen jatkuu vuoden 2017 loppuun asti. (JYVSECTEC 2016).

1.1 Tehtävän kuvaus ja Tavoitteet

Opinnäytetyön tavoitteena oli tutustua avoimeen dataan, jota voitaisiin käyttää erilaisilla BIG DATA analytics -tuotteilla. Avoimien datavarastojen kartoitus Suomessa ja maailmalla sekä yleisesti datan keräämisen ja jakamiseen liittyvät rajoitukset esim. tietosuoja.

Tietoperustassa tutkitaan Big Dataa yleisellä tasolla ja syventyen avoimeen dataan. Lisäksi tutkitaan muutamien toimeksiantajan valitsemien avoimen data lähteiden käyttöönottoa toimeksiantajan valitsemalla tuotteella. Tietoperustana käytetään alan kirjallisuutta avoimen datan tuottajien verkkosivuja ja ajankohtaista tietoa verkosta.

2 BIG DATA

2.1 Big Data käsitteenä

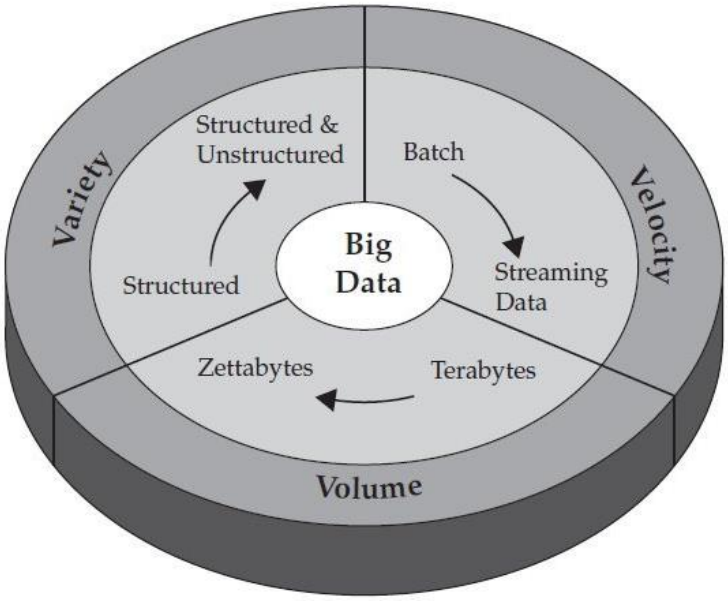
Big Datalla viitataan kahteen asiaan. Ensimmäinen on nopeasti kasvava ja monipuolistuva data, joka luo haasteita yhtiöille ja yhteiskunnalle. Toinen asia on miten tähän muuttuvaan ja kasvavaan datan haasteeseen pystytään vastaamaan. (Salo 2013,10).

Nykyteknologia on mahdollistanut uuden datan helpon ja yksinkertaisen luomisen sekä siirtämisen internetiin. Uuden datan määrä kasvaa kiihtyvällä tahdilla, koska verkkoon liitettävien laitteiden määrä myös kasvaa koko ajan. Teknologian kehittyessä nämä vaatimukset uuden datan luomiselle ja sen tallentamiselle tulevat vain kasvamaan.(Salo 2013,10–12).

Pelkästään ihmiset eivät luo uutta dataa, vaan myös erilaiset mittauspalvelut ja järjestelmät tuottavat automaattisesti suuren määrän käsittelemätöntä dataa kiihtyvällä tahdilla kuten sääasemat, autojen ajotietokoneet, älypuhelimet, reitittimet, sekä lukemattomat muut laitteet. Suurta osaa tästä datasta ei hyödynnetä millään tavalla ja suuri osa siitä johdettavasta informaatiosta jää myös tallentamatta. (Salo 2013,21).

Big datasta haetaan vastausta, miten pystytään järkevästi ja tehokkaasti siirtämään, tallentamaan, yhdistelemään sekä monipuolisesti analysoimaan kaikkea käsillä olevaa dataa. (Salo 2013,21).

Big Dataan yhdistetään usein kolmen V-kirjaimen malli (Kuvio 1). Ensimmäinen V tulee sanasta Volume eli volyymi, jolla tarkoitetaan datamäärän eksponentiaalista kasvua. Toinen V on Velocity eli tiedon vauhti, jolla dataa syötetään tietojärjestelmiin ja niistä pois. Kolmas V on Variety eli vaihtelevuus, jolla kuvataan datan rakenteen vaihtelevuutta. (Salo 2013, 21.)



Kuvio 1. Big data kolmen v-kirjaimen malli. (Eaton, C. Deroos, D. Deutsch, T. Lapis, G. Zikopoulos P. 2012)

2.2 Internet of Things

Internet of Things (IoT) -käsitteellä tarkoitetaan meneillään olevaa ilmiötä, jossa globaaliin tietoverkkoon kytkettävien laitteiden määrä kasvaa kiihtyvää vauhtia. Tämä siis tarkoittaa sitä, että mitä enemmän on verkkoon kytkettyjä laitteita, sitä enemmän virtaavaa dataa. IoT tulee siis syöttämään valtavasti uutta dataa pilvipalveluihin ja sitä kautta ruokkimaan Big Data–ilmiötä, jolla voidaan älyllistää tyhmät ja halvat laitteet kuten esimerkiksi jääkaapit. (Salo 2013, 12-13).

2.3 Strukturoitu, strukturoimaton ja semistrukturoitu data.

Data jaetaan karkeasti kahteen ryhmään; strukturoituun ja strukturoimattomaan datan. Nämä ovat vain ääripäät, sekä näiden väliin jäävää dataa kutsutaan semistrukturoiduksi dataksi. (Salo 2013,22).

2.3.1 Strukturoitu data

Termillä strukturoitu data viitataan yleensä dataan, jolla on ennalta määritelty formaatti sekä pituus. Esimerkiksi strukturoitu data sisältää numeroita, päivämääriä sekä ryhmiä koostuen sanoista ja numeroista. Näitä kutsutaan stringeiksi eli merkkijonoiksi. Stringi voi sisältää asiakkaan tietoja kuten nimen ja osoitteen. Arvellaan, että tämä kattaa noin 20 prosenttia kaikesta liikkeellä olevasta datasta. Strukturoitu data varastoidaan yleensä suuriin tietokantoihin, josta sitä voidaan hakea käyttämällä esimerkiksi SQL–kyselykieltä (Structured Query Language). Tällaisen datan alkuperä määritellään kahteen kategoriaan, tietokoneen tai laitteen luoma data, jollaista tuottavat esimerkiksi GPS-paikannustiedot (Global Positioning System), serverillä tapahtuva lokitiedostojen keräys ja pörssissä käytetty osakekauppadata. Toinen kategoria on data jota ihminen itse syöttää tietokoneeseen, kuten omien yhteystietojen täyttäminen verkkolomakkeeseen tai mainoksen klikkaaminen verkkosivulla, mikä luo uutta strukturoitua dataa. (Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013. Kappale 2).

2.3.2 Strukturoimaton data

Sturkturoimattomalla datalla tarkoitetaan dataa, joka on ennalta määrittelemätöntä. Kuten myös strukturoidussa datassa lähteenä voi olla tietokoneen tai ihmisen syöttämä data. Se voi olla kuvia, tekstiä, videota, säätietoja, asiakirjoja, paikka- ja sijaintitietoja. Esimerkkeinä tietokoneen luomasta stukturoidusta datasta voidaan ottaa Google Earth–karttapalvelu, joka sisältää säätietoja, satelliittikuvia ja maastokarttoja. Palvelut, jotka tuottavat paljon erilaista tieteellistä dataa, jota valtiot voivat käyttää

ja käyttää hyväkseen omissa palveluissaan, kuten dataa planeetan seismisistä liikkeistä, ilmakehässä tapahtuvista muutoksista ja korkeasta energiafysiikasta. Palvelu, joka tuottaa valokuvia ja videota kuten videovalvonta, keli, ja liikennekamerat. Esimerkkinä ihmisen luomasta strukturoidusta datasta voi olla yrityksen sisäisen informaation liikkuminen ja luominen, kuten sähköpostit, lokitiedostot, tutkimustulokset ja muut asiakirjat. Myös mobiilidata, joka sisältää viestit ja niiden paikannustiedot, sekä sosiaalisenmedian eri alustoilla luotu data kuuluu tähän, kuten Facebook, YouTube, Twitter, LinkedIn ja Instagram. (Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013. Kappale 2).

2.3.3 Semistrukturoitu data

Semistrukturoidulla datalla tarkoitetaan strukturoidun ja strukturoimattoman datan sekoitusta. Data sisältää molempia tyyppejä, kuten datan metatiedot, jotka on liitetty valokuviin ja videotiedostoihin. (Salo 2013, 25).

3 AVOIN DATA

3.1 Avoin Data yleisesti

Yleisesti avoimella datalla tarkoitetaan tietoa, joka on avattu kaikkien käyttöön vapaasti ja maksutta. Tällaista tietoa voi olla julkishallinnolle, organisaatioille, yrityksille tai yksityishenkilöille kertynyt data, jota voidaan hyödyntää jollain tavalla. (Opendefinition 2016).

Avoin data ja julkinen tieto eivät ole sama asia. Julkiseen tietoon on kaikilla ihmisillä pääsy. Avoin data taas on sellaista dataa, jota yksityishenkilöt ja yritykset voivat käyttää omiin tarkoituksiinsa tasavertaisesti julkisen hallinnon kanssa. (Opendefinition 2016).

Avointa dataa on mahdollisuus kerätä monista lähteistä, kuten internetsivuilta, sähköposteista ja erilaisista maksuliikenteistä. Aineisto on avointa, jos sen määritelmät

täyttävät tarvittavat ehdot. Ensimmäinen ehto on saavutettavuus, eli aineiston pitää olla kokonaisuudessaan saatavilla. Toinen ehto on, että data on käytännöllisessä ja helposti muokattavassa muodossa. Kolmantena ehtona on uudelleenjakelu, eli lisenssi ei saa rajoittaa aineiston käyttöä eikä siitä saa vaatia rojalteja, tai muita myyntiin ja jakeluun sisältyviä maksuja. (Opendefinition 2016).

3.2 Tiedostomuodot

Avoimessa datassa käytetyn tiedostomuodon pitäisi olla avoimen datan kriteerien mukaan avoin, julkisesti ja vapaasti saatavilla sekä riippumaton kaupallisista sovelluksista. Aina tähän ei kuitenkaan ole realistista mahdollisuutta. Esimerkiksi GPS-järjestelmät voivat käyttää valmistajakohtaisia formaatteja sekä usein käytetyt Excel-taulukot, jotka ovat jollain tapaa riippuvaisia Microsoftin ohjelmistosta. (Poikola 2010. 37).

Tiedostomuodon täytyy myös olla koneluettavaa, jotta sitä voidaan käyttää automatisoidusti ja ohjelmallisesti. Jos datan hyödyntäminen on liian monimutkaista voi avoimen datan hyödyt jäädä käyttämättä. (Poikola 2010. 37, 64).

Avoimeen dataan löytyy monia yleisesti käytettyjä tiedostomuotoja kuten XML, CSV, JSON ja RDF. (Poikola, A., Kola, P. & Hintikka, K. 2010. 37, 64).

3.2.1 XML

XML (Extensible Markup Language) on yleiskäyttöinen merkintäkieli. Joka on laajennettavissa uusilla merkintäelementeillä kuten RSS (Really Simple Syndication). Sitä käytetään päivittyvien verkkosisältöjen eli syötteiden välitykseen. Syötteen muoto on GeoRSS, jos siihen on lisätty paikannustietoja (Kuvio 2) (Poikola, A., Kola, P. & Hintikka, K. 2010. 64)

Kuvio 2. XML LAM-syöte

3.2.2 CSV

CSV (Comma Separated Value) on tiedosto, jossa arvot on erotettu toisistaan pilkuilla (Kuvio 3). Tiedostot voidaan avata jollain taulukkolaskentaohjelmalla. (Poikola, A., Kola, P. & Hintikka, K. 2010. 64)

	A	B
1	SUKUNIMI	YHTEENSÄ
2	Korhonen	22732
3	Virtanen	22445
4	Mäkinen	20468
5	Nieminen	20378
6	Mäkelä	19130
7	Hämäläinen	18605
8	Laine	18148
9	Heikkinen	17536
10	Koskinen	17353

Kuvio 3. CVS-tiedosto sukunimistä

3.2.3 JSON

JSON (JavaScript Object Notation) on kevyt ja ohjelmointikielestä riippumaton avoimen standardin tekstipohjainen datan siirtomuoto. Nimestään huolimatta JSON on riippumaton JavaScriptistä. JSON-tekstimuoto (Kuvio 4) noudattaa samoja merkintätapoja jotka löytyvät ohjelmointikielistä kuten C-, C++, C#-, Java, JavaScript, Perl- ja Python. (Poikola, A., Kola, P. & Hintikka, K. 2010. 64)


```
[{"trainNumber":81,"departureDate":"2016-04-
06","operatorUICCode":10,"operatorShortCode":"vr","trainType":"S","trainCategory":"Long-
distance","commuterLineID":"","runningCurrently":false,"cancelled":false,"version":39633412943,
"timeTableRows":
[{"stationShortCode":"HKI","stationUICCode":1,"countryCode":"FI","type":"DEPARTURE","trainStopp
ing":true,"commercialStop":true,"commercialTrack":"9","cancelled":false,"scheduledTime":"2016-
04-06T03:06:00.000Z","actualTime":"2016-04-06T03:05:57.000Z","differenceInMinutes":0,"causes":
[]},
{"stationShortCode":"PSL","stationUICCode":10,"countryCode":"FI","type":"ARRIVAL","trainStoppin
g":true,"commercialStop":true,"commercialTrack":"3","cancelled":false,"scheduledTime":"2016-04-
06T03:11:00.000Z","actualTime":"2016-04-06T03:10:29.000Z","differenceInMinutes":-1,"causes":
```


Kuvio 4. Jyväskylän juna-asema JSON syöte

3.2.4 RDF

RDF (Resource Description Framework) on linked data–paradigman standardi, jossa yksittäisiä tietoresursseja voidaan kuvailla niihin linkitettävien sanastojen avulla. (Poikola, A., Kola, P. & Hintikka, K. 2010. 64)

3.2.5 REST–arkkitehtuuri

REST (Representional State Trasfer) on HTTP-protokollaan perustuva arkkitehtuuri, joka soveltuu erityisesti erilaisten verkkosovellusten toteuttamiseen. HTTP-protokollan ansiosta REST on täysin riippumaton käytetystä käyttöjärjestelmästä ja ohjelmointikielestä. (National Security Agency. 2011)

REST-arkkitehtuuri on tilaton asiakas-palvelinmalli, eikä se säilytä tietoja asiakkaan tilasta. Jokaisen palvelimelle lähetetyn kutsun täytyy sisältä tarvittava tieto vastauksen tuottamiseksi. Datassa on tunnistetieto URI (Uniform Resource Identifier), jota voidaan hallita metodeilla. REST hyödyntää HTTP-metodeja palvelimen kutsuissa ja vastauksissa. Nämä metodit ovat GET-metodi, jolla voidaan hakea mitä tahansa tietoa, jossa on URI-tunniste; POST-metodi, jota käytetään uuden datan luomiseen, kuten uusien käyttäjien lisääminen; PUT-metodi, jota käytetään jo olemassa olevan datan päivittämiseen ja muuttamiseen ja DELETE-metodi, jolla voidaan poistaa dataa palvelimelta. (National Security Agency. 2011)

3.3 Creative Commons

3.3.1 Creative Commons-lisenssi yleisesti

Creative Commons (CC) on vuonna 2001 perustettu Yhdysvaltalainen voittoa tavoittelematon järjestö, joka tarjoaa erilaisia lisenssejä, joilla käyttäjä voi jakaa tietoa. Käyttäjä pystyy tietyn lisenssin avulla määräämään jaettavan aineiston tekijäoikeuksista, mitkä pitävät itsellään ja mitkä jakaa muille. (Creative Commons).

CC–käyttöluvat luotiin helpottamaan avoimen käyttöluvan myöntämistä muillekin tekijänoikeuden suojaamille aineistoille kuin tietokoneohjelmistoille. Avoimet käyttöluvat perustuvat siihen, että aineistoa saa vapaasti käyttää ja jatkojalostaa sillä edellytyksellä, että käyttöehtoja noudatetaan. Yleisin ehto on, että käyttäjäluvan antajan nimeämistiedot mainitaan. (Creative Commons).

Inspiraatiota otettiin GNU GPL:stä (GNU General Public License), joka antaa oikeudet kenelle tahansa käyttää, kopioida, muuttaa, jakaa ohjelmistoja ja niiden lähdekoodia, eikä se estä aineiston kaupallista käyttöä. Ehkä tunnetuin esimerkki GNU GPL:n käytöstä ovat Linux-käyttöjärjestelmät. (Creative Commons)

Creative Commons tarjoaa myös Creative Commons Search-metahakukoneen, jolla on helppo päästä käsiksi useisiin suuriin CC-lisensoituihin tietoaineistoihin. Search.creativecommons.org ei kuitenkaan ole hakukone, vaan se käyttää hyväkseen jo valmiita ratkaisuja ja etsii näistä CC-lisensoituja aineistoja. Tällaisia aineistoja ovat

esimerkiksi Wikimedia Commons, YouTube, SoundCloud, Flickr ja Google. Googlen hakuasetukset voidaan asettaa näyttämään vain CC-lisenssin alaisia hakutuloksia. (Creative Commons)

3.3.2 Creative Commons Nimeä (CC BY)

Creative Commons Nimeä 4.0 (Creative Commons ByAttribution 4.0) – lisenssi (Kuvio 5) sallii muiden levittää, muokata teosta sekä luoda sen pohjalta uusia teoksia. Sitä voidaan käyttää kaupallisissa tarkoituksissa, kunhan alkuperäinen tekijä mainitaan. (Creative Commons)

Kuvio 5. Creative Commons ByAttribution

3.3.3 Creative Commons Nimeä-EiKaupallinen (CC BY-NC)

Creative Commons Nimeä-EiKaupallinen (Creative Commons ByAttribution-NonCommercial) – lisenssi (Kuvio 6) toimii samalla tavalla kuin CC BY – lisenssi, mutta vain epäkaupallisessa tarkoituksessa ja alkuperäinen tekijä on mainittava. (Creative Commons)

Kuvio 6. Creative Commons ByAttribution-NonCommercial

3.3.4 Creative Commons Nimeä-EiMuutoksia (CC BY-ND)

Creative Commons Nimeä-EiMuutoksia (Creative Commons ByAttribution-NoDerivates) – lisenssi (Kuvio 7) antaa luvan kaupalliseen ja epäkaupalliseen levittämiseen,

kunhan vain aineisto jaetaan kokonaisena, muuttumattomana ja alkuperäinen tekijä mainitaan. (Creative Commons)

Kuvio 7. Creative Commons ByAttribution-NoDerivates

3.3.5 Creative Commons Nimeä-JaaSamoin (CC BY-SA)

Creative Commons Nimeä-JaaSamoin (Creative Commons ByAttribution-ShareAlike) – lisenssi (Kuvio 8) toimii samoin kuin CC BY, mutta kaikki tähän aineistoon perustuvat teokset pysyvät saman lisenssin alla, joten myös tästä johdettuja teoksia voidaan käyttää kaupallisesti. (Creative Commons)

Kuvio 8. Creative Commons ByAttribution-ShareAlike

3.3.6 CC Nimeä-EiKaupallinen-JaaSamoin (CC BY-NC-SA)

Creative Commons Nimeä-EiKaupallinen-JaaSamoin (Creative Commons ByAttributin-NonCommercial-ShareAlike) – lisenssi (Kuvio 9) toimii samoin kuin CC BY, mutta vain epäkaupallisessa tarkoituksessa. Alkuperäinen tekijä täytyy mainita ja uudet tuotokset lisätään saman lisenssin alle. (Creative Commons)

Kuvio 9. Creative Commons ByAttributin-NonCommercial-ShareAlike

3.3.7 CC Nimeä-EiKaupallinen-EiMuutoksia (CC BY-NC-ND)

Creative Commons Nimeä-EiKaupallinen-EiMuutoksia (Creative Commons ByAttribution-NonCommercial-NoDerivates) – lisenssi (Kuvio 10) sallii teoksen jakamisen sillä

ehdolla, että alkuperäinen tekijä mainitaan. Teosta ei kuitenkaan saa muokata millään tavalla, eikä sitä saa hyödyntää kaupallisesti. (Creative Commons)

Kuvio 10. Creative Commons ByAttribution-NonCommercial-NoDerivates

3.4 Avoin data Suomessa

Viime vuosina Suomessa on alettu panostamaan avoimeen dataan. Lähes kaikilla suuremmilla kaupungeilla on jo avoimen datan tietokannat ja näiden parissa toimivia hankkeita ja organisaatioita. Nykyään kaupungeilta on mahdollista saada monimuotoista avointa dataa, kun useat julkishallinnon organisaatiot ja pienemmät kunnat ovat aloittaneet tietolähteiden avaamisen.

Suurimpia suomalaisia tietoaineistoja on Helsingin kaupungin tietokeskus, joka on Helsinki Region Infosharen ylläpitämä palvelu. Tämä tietoaineisto sisältää pääasiassa data-aineistoa pääkaupunkiseudun alueesta, asukkaista ja palveluista. (Helsinki Region Infoshare).

Toinen suuri tietoaineisto on Valtionvarainministeriö, joka sisältää dokumentteja ohjeista ja standardeista. Kaikki näistä dokumenteista eivät kuitenkaan ole koneluettavassa muodossa, vaan yleisesti käytetty tiedostomuoto on HTML (Hypertext Markup Language) ja PDF (Portable Document Format).

3.4.1 Tietolähteet

Julkishallinnon tuottamien tietoaineistojen rinnalla toimii usea paikallinen tai organisaatiokohtainen sivusto, joka keskittyy tietynlaiseen dataan. Tällaisia ovat esimerkiksi kaupunkien avatut datalähteet Jyväskylässä, Tampereella, Mikkelissä ja Oulussa. Keväällä 2016 avautuu myös oikeusministeriön Finlex-tietopankki, joka sisältää suuren määrän asiakirjoja säädös- ja oikeustapaustietokannoista. Lähtökohtaisesti kaikki

palvelussa oleva data on koneluettavaa, jotta muut sovellukset ja tietojärjestelmät voivat sitä hyödyntää. (Oikeusministeriö)

Taulukkossa 1 on esitetty Suomen suurimpien kaupunkien avattujen datavarantojen osoitteet.

Taulukko 1. Tietoainestoja

LÄHDE	**url**
Jyväskylän Kaupunki	http://data.jyvaskyla.fi
Tampereen Kaupunki	http://data.tampere.fi
Mikkelin Kaupunki	https://open.mikkeli.fi/
Oulun Kaupunki	http://www.ouka.fi/oulu/oulu-tietoa/avoin-data
Turun Kaupunki	https://www.turku.fi/avoindata
Kuopion Kaupunki	https://www.kuopio.fi/web/kaupunkitietoa/avoin-data
Ilmatieteen laitos	https://ilmatieteenlaitos.fi/avoin-data
Liikennevirasto	http://www.liikennevirasto.fi/avoin-data
Verohallinto	https://www.vero.fi/fi-FI/Avoin_data

3.4.2 Avoindata.fi

Avoindata.fi on julkishallinnon avoimen datan, tietojen ja tietojärjestelmien jakamiseen tarkoitettu palvelu. Se ei ole ainoastaan julkishallinnolle suunnattu palvelu, vaan sitä voivat hyödyntää kaikki ne tahto, jotka haluavat hyödyntää julkishallinnon avointa dataa omissa palveluissaan. Palvelusta löytyy myös julkishallinnon yhteen toimivuuden edistäviä kuvauksia ja ohjeita, nämä mahdollistaa suunnittelutiedon jakamisen ja uudelleenkäytön. Palvelun tavoitteena on edistää ja helpottaa avoimen datan saatavuutta, hyödyntämistä ja käyttöä sekä edistää julkishallinnon läpinäkyvyyttä ja vähentää päällekkäistä datan keräämistä ja tuotantoa. Julkishallinnon avoimien tietoaineistojen käyttöluvaksi suositellaan Creative Commons Nimeä 4.0 – lisenssiä (Avoindata 2016)

Palvelusta on löydettävissä yli 1300 avoimen datan tietoaineistoa (Taulukko 2) ja se listaa yli 800 eri organisaatiota, vaikka suuressa osassa näistä ei ole käytettäviä tietoainestoja. Tietoaineistoja voidaan hakea palvelusta aiheen, sisältötyypin, tiedon tuottajan ja tiedostomuodon mukaan.

Taulukko 2. Avoindata.fi organisaatiot

ORGNANISAATIO	TIETOAINEISTOT	KUVAUS
Kunnat ja kunnallishallinto	707	Kaupunkien ja kuntien avoimet aineistot
Valtionhallinto	614	Valtion eri virastojen aineistot
Ulkoiset lähteet	25	Paikkatietohakemisto joka on Maanmittauslaitoksen ylläpitämä valtakunnallinen paikkatietojen metatietopalvelu
Yliopistot ja korkeakoulut	11	Eri koulujen tarjoamat avoimet aineistot
Yksityishenkilöt	7	Yksityishenkilöiden tarjoamat tietoainestot, kuten polttoaineiden hintatiedot
Yritykset ja yhteisöt	6	Yritysten tietoainestoja, kuten Yleisradio
Yhdistykset ja säätiöt	5	Teoston keräämä livemusiikkidata
Julkisen hallinnon standardit	2	Sisältää eri alojen standardeja ja niiden soveltamisohjeita
Suomi syö ja juo -hanke	1	Valokuvia suomalaisesta ruoka- ja juomakulttuurista eri aikakausilta

3.4.3 Avoimen datan rajapinnat

Suomesta löytyy jo suuri määrä palveluita, joista on mahdollisuus saada ajantasaista dataa suoraan palveluntarjoajan rajapinnasta. Tällainen on esimerkiksi Digitraffic, joka on tieliikenneviraston keräämän avoimen datan jakelukanava ja tarjoaa useita avoimia rajapintoja. Digitraffic tarjoaa ajankohtaista tietoa tie- ja rautatieliikenteestä ja tulee tulevaisuudessa laajentumaan myös meriliikenteeseen. Kaikki tästä palvelusta saatava tieto on Creative Commons Nimeä 4.0 – lisenssin alaista. (Digitraffic tieliikenne)

Liikennevirasto tarjoamaa tieliikenteeseen liittyvät avoimet rajapinnat (Taulukko 3), josta se jakaa ajankohtaisia liikennetietoja koneluettavassa XML-muodossa. Digitraffic käyttää tietolähteenä Liikenneviraston matka-aikatietopalvelua, liikenteen automaattisia mittauspisteitä (LAM), tiesääasemia, kelikamerakuvia sekä tieliikennekeskuksen häiriötiedotteita. (Digitraffic tieliikenne)

Taulukko 3. Tieliikenne Rajapinnat

RAJAPINTA	**url**
Ajantasaiset sujuvuustiedot	http://tie.digitraffic.fi/sujuvuus/ws/trafficFluency
Ajantasaiset matka-aikatiedot	http://tie.digitraffic.fi/sujuvuus/ws/journeyTime
Edellisen päivän sujuvuuden historiatiedot	http://tie.digitraffic.fi/sujuvuus/ws/dayData
Edellisen päivän 12 viikon keskimääräiset päivittäiset sujuvuustiedot	http://tie.digitraffic.fi/sujuvuus/ws/averageDayData
Ajantasaiset LAM-mittaustiedot	http://tie.digitraffic.fi/sujuvuus/ws/lamData
Ajantasaiset vapaat nopeudet	http://tie.digitraffic.fi/sujuvuus/ws/freeFlowSpeeds
Tiesääasemien ajantasaiset mittaustiedot	http://tie.digitraffic.fi/sujuvuus/ws/roadWeather
Tieasemien tilatiedot	http://tie.digitraffic.fi/sujuvuus/ws/roadStationStatuses
Kelikameroiden esiasetukset	http://tie.digitraffic.fi/sujuvuus/ws/cameraPresets
Tiejaksojen keliennusteet	http://tie.digitraffic.fi/sujuvuus/ws/roadConditions
Häiriötiedotteet	http://tie.digitraffic.fi/sujuvuus/ws/trafficDisorders

Liikenneviraston Digitraffic-palvelu tarjoaa myös rautatieverkon avoimet rajapinnat, joiden tietolähteenä toimii ratakapasiteetin hallinnan LIIKE-järjestelmä, josta tiedot poimitaan avoimeen rajapintaan käytettäväksi. Avoimesta rajapinnasta saatavat tiedot ovat reaaliaikainen junien seuranta, aikataulutiedot, historiatiedot ja junien kokoonpanotiedot. Rajapinta on REST -tyyppinen, eli http (Hypertext Transfer Protocol)– protokollaan perustuva arkkitehtuuri. Eli käyttäjä pystyy eri parametreja hyödyntämällä hakemaan vain tarvitsemiaan tietoja, tämän jälkeen rajapinta palauttaa vastauksen koneluettavassa JSON–muodossa. (Digitraffic rautatieliikenne)

Esimerkkinä voidaan hakea haluamiaan reaaliaikaisia tietoja rajapinnasta käyttäen eri parametreja. Kuten reaaliaikaiset tiedot Jyväskylän juna-asemalta rajapinnasta

http://rata.digitraffic.fi/api/v1/live-trains?station=JY

Hakuun voidaan lisätä parametreja, joilla hakua pystytään rajaamaan tarkemmin (Taulukko 4).

/live-trains?station=<station_shortcode >&arrived_trains= &arriving_trains= &departed_trains=<departed_trains> &departing_trains=<departing_trains> &version=<change_number>

Taulukko 4. Rautatieliikenne Rajapinnat

NIMI	SELITYS
station_shortcode	Aseman lyhenne, esimerkiksi JY, HKL,TPE
arrived_trains	Kuinka monta saapunutta junaa palautetaan.
arriving_trains	Kuinka monta saapuvaa junaa palautetaan.
departed_trains	Kuinka monta lähtenyttä junaa palautetaan
departing_trains	Kuinka monta lähtevää junaa palautetaan
include_nonstopping	Palautetaanko aseman ohi pysähtymättä ajavat junat
version	Palauttaa tietyn versiotyypin junat, jollei anneta arvoa, palauttaa uusimmat tiedot

3.5 Avoin Data maailmalla

3.5.1 Amazon Web Services

AWS (Amazon Web Services) on Yhdysvaltalaisen Amazon yrityksen ylläpitämä kuukausimaksullinen pilvipalvelu. Toimintansa perinteisenä verkkokauppana aloittanut yritys on kasvanut ja tällä hetkellä Amazon on yksi suurimmista pilvipalveluntarjoajista. AWS koostuu useista eri pilvipalveluista, joista tunnetuin ovat elastista pilvilaskentaa tarjoava Amazon Elastic Compute Cloud (EC2). (Amazon Web Services).

AWS tarjoaa ja ylläpitää myös suurta avoimen datan tietolähdettä, josta löytyy esimerkiksi Landsat 8–satelliitin ottamat maastokuvat, yli 3000 riisilajin perimä sekä lähes 100 miljoonaa Creative Commons–lisenssin alla olevaa kuvaa ja videota. Haluttu data voidaan ladata omaan käyttöön käyttäen Amazon EC2 – palvelua, jossa voit vuokrata virtuaalipalvelimia tai prosessoida dataa Hadoopilla käyttäen Amazon EMR–palvelua (Amazon Elastic MapReduce), joka käyttää laskennassa hyväkseen muita Amazonin palveluja. (Amazon Web Services).

3.5.2 Google Public Data Directory

Google julkaisi vuonna 2010 Google Public Data Directoryn, jonka tarkoituksena tarjota avointa dataa ja ennusteita suurelta joukolta kansainvälisiä organisaatioita kuten Maailmanpankki, Eurostat, OECD (Organisation for Economic Cooperation and Development) ja IMF (International Monetary Fund). Google Public Data Explorer tarjoaa helpon tavan tarkastella dataa, koska kaikki data on sellaisessa muodossa, jotta se pystytään esittämään erilaisissa diagrammeissa tai kartalla. Kuten esimerkiksi pystytään vertaamaan valtioiden työttömyyttä tai asuin kustannuksia. Tästä syystä kaikki data on muutettava Googlen luomaan DSPL (Dataset Publishing Language)– muotoon. DSPL on pakattu tiedostomuoto, joka rakentuu datan sisältävästä CSV-tiedostosta ja metadatan sisältävästä XML-tiedostosta. Nämä kummatkin tarvitaan, jotta sitä voidaan käyttää graafisissa esityksissä. Palvelu ei kuitenkaan vielä tue datan suoraan lataamista palvelimilta, vaan se joudutaan hakemaan alkuperäisestä lähteestä. Google tarjoaa kuitenkin linkin alkuperäiseen aineistoon. (Google Inc. 2016)

4 BIG DATAN TYÖKALUT

Suureen määrään strukturoitua ja strukturoimatonta dataa tarvitaan myös siihen soveltuvia ja suunniteltuja työkaluja. Yksi ehkä tunnetuimmista työkaluista on Hadoop. Hadoop on avoin ohjelmistokehitysprojekti, joka soveltuu hyvin suurien datamäärien käsittelyyn ja sitä pystytään käsittelemään perinteisillä analysointityökaluilla. Suuret ohjelmistotalot, jotka ovat perehtyneet Big Datan käsittelyyn ja tiedon analysointiin kuten IBM, Cloudera ja Hortonworks, käyttävät omien järjestelmiensä pohjana juuri Hadoopia.

4.1 Hadoop

Hadoop on avoimen lähdekoodin alustariippumaton ohjelmistokehitysprojekti, jonka tarkoituksena on pyrkiä helpottamaan yritysten suurien datamäärien kustannustehokasta käsittelyä. Kustannustehokkuutta lisää avoin lähdekoodi, jolloin lisenssistä ei

tarvitse maksaa. Sekä alustariippumattomuus, joka mahdollistaa eri alustojen käytön, eikä välttämättä olla kiinni kalliissa, maksullisessa alustassa. Tarkoituksena on luoda palvelinklusteri, jota voidaan käyttää suuren ja monimuotoisen tallennetun datan analysointiin nopeasti ja viiveettömästi. Hadoop myös lisää redundanttisuutta datan tallentamisessa ja analysoinnissa, kun data oletusarvoisesti tallennetaan klusteriin kolmena kopiona. Tämä tarkoittaa sitä, että data on hyvässä tallessa, jolloin laiteviat tai ohjelmistopäivitykset eivät aiheuta datan katoamista tai tilapäistä tiedoston saavuttamattomuutta. Analysointiin pätee myös hajauttaminen, analysointia ajetaan usealla palvelimella samaan aikaan rinnakkain. Tällöin vikatilanteen sattuessa, yhden palvelimen kaatuminen ei pysäytä analyysiä, eikä suuressa klusterissa edes hidasta sitä. (Salo 2013, 80–81)

Hadoop sisältää kaksi pääkomponenttia, HDFS (Hadoop Distributed File System) on edullinen ja luotettava tallennusklusteri, joka hallitsee tiedostoja verkossa. Sekä MapReduce, jota käytetään tietojen louhintaan. (Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013, kappale 9).

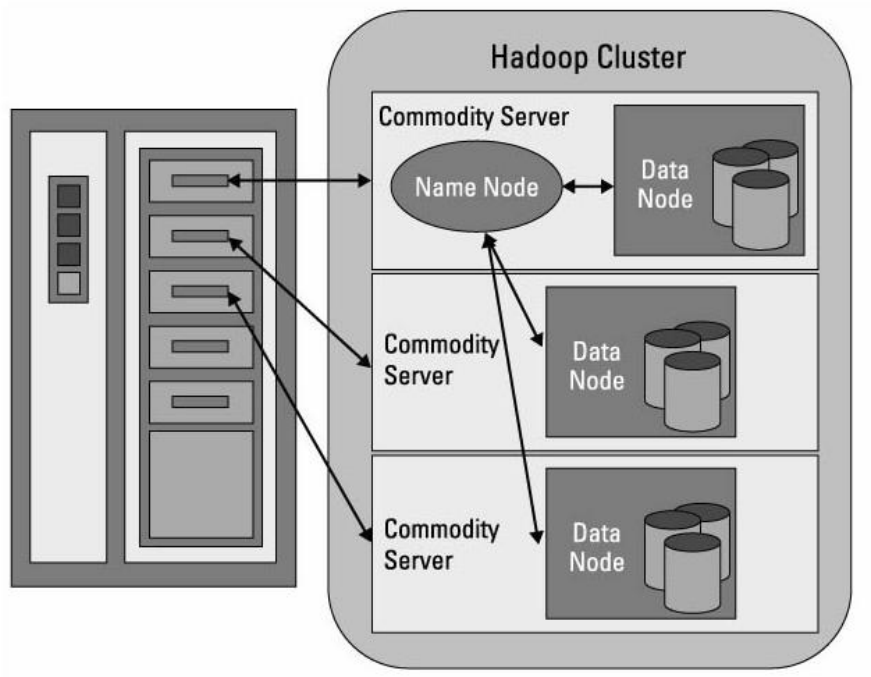
4.1.1 HDFS

HDFS (Hadoop Distributed File System) on toinen Hadoopin ydinprojekteista MapReducen lisäksi. HDFS viittaa suoraan Hadoopin tiedostojärjestelmään, joka on hajautettu useaan palvelimeen eli klusteriin (Kuvio 11). Tämä tuo edullisuutta, toimintavarmuutta ja nopeutta suurien datamäärien tallentamiseen, koska hajautettu Hadoop-klusteri mahdollistaa datan rinnakkaisen käsittelyn. Edullisuus perustuu suoraan avoimeen lähdekoodiin, jolloin lisenssin käyttöoikeuksista ei tarvitse erikseen maksaa ja mahdollisuuteen käyttää heterogeenistä laitteistoa. (Salo 2013, 82)

HDFS ei ole tarkoitus olla vain datan lopullinen tallennuspaikka, vaan toimia palveluna, jossa datan määrä ja nopeus on suurta. Koska data kirjoitetaan klusteriin vain kerran ja luetaan useasti tämän jälkeen. Tällöin ei tarvita ominaisuutta, jota muut tie-

dostojärjestelmät käyttävät. Eli jatkuvaa datan lukua ja päällekirjoitusta, tämä hidastaa palvelua. HDFS on hyvä vaihtoehto tukemaan datan analysointia. (Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013).

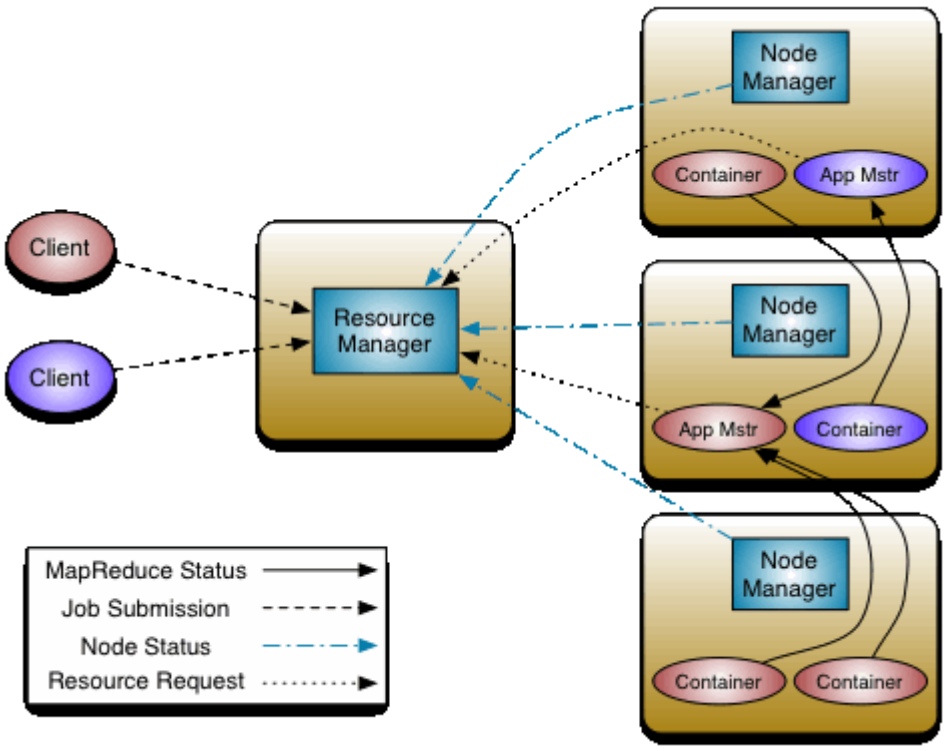
HDFS toimii hajottamalla isommat tiedostot pienempiin lohkoihin, josta jokainen kopioidaan vähintään kolme kertaa ja tallennetaan klusteriin. Jokainen lohko kahdennetaan useita kertoja, jotta yksittäinen laitevika palvelimessa ei aiheuta datan häviämistä. (Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013, kappale 9).



Kuvio 11. Hadoop Distibuted File System

4.1.2 YARN

YARN on Hadoop komponentti, joka hallitsee klusterin työkuormaa ja resursseja. YARN toimii järjestelmän resurssien ja ohjelmien valvovana elimenä. Jokainen elementti Hadoopissa keskustelee YARN:in kanssa ja sen työ on päättää hallita, miten paljon kukin työ saa klusterissa tehoa (Kuvio 12). (YARN).



Kuvio 12. YARN

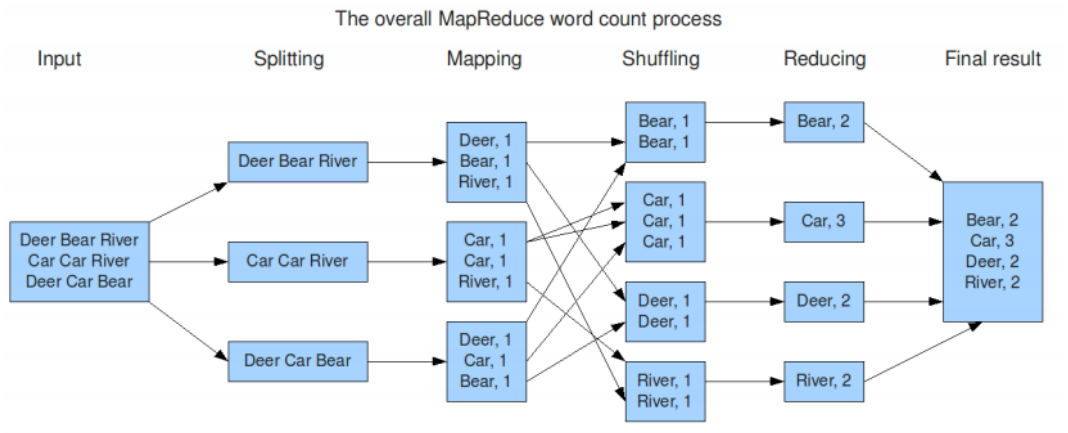
YARN on kaksiosainen komponentti. Ensimmäinen on Scheduler eli ajoittaja, joka hoitaa töiden ajoittamiseen klusterissa, eli päättää milloin mikäkin työ ajetaan. Toinen osa on ApplicationsManager eli ohjelmakontrolleri. Ajoittajan tehtävä on sijoittaa työt, sekä määrittää paljonko kyseiselle työlle annetaan resursseja käytettäväksi. Kontrolleri hallitsee itse työtä ja sen käyttämiä resursseja, sekä sen toimintaa. Näiden komponenttien toimintaa voidaan laajentaa erilaisilla lisätyökaluilla. (YARN).

4.1.3 MapReduce

MapReduce on Hadoopin YARN-pohjainen järjestelmä, jota käytetään suurten tietomassojen rinnakkaiseen käsittelyyn, jossa haut toteutetaan useissa rinnakkaisissa HDFS-nodeissa. (Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013, kappale 22).

MapReduce sisältää kolme vaihetta, jota se käyttää hajautetussa analyysissa, nämä ovat map-, shuffle- ja reduce-vaiheet (Kuvio 13). Map- ja reduce-vaiheessa ajetaan

sovelluskehittäjän omaa koodia, josta muodostuu käytetyt algoritmit ja tämä luo analytiikan älyn. Eli tässä pyritään kartoittamaan data ja muuntamaan se oikeaan muotoon, jotta sitä voidaan myöhemmässä vaiheessa yhdistellä. Shuffle-vaiheessa saadut välitulokset lähetetään map-vaiheen suorittaneelta palvelimelta reduce-vaiheen palvelimelle. Reduce-vaihe yhdistelee ja järjestelee datan luettavaan ja analysoitavaan muotoon. (Salo 2013, 83).



Kuvio 13. MapReduce

4.1.4 Hive

Apache Hive on SQL-tietokantasovellus, joka toimii Hadoop:in päällä. Se tarjoaa datan yhdistelyä, kyselyjen tekemistä tietokantaan sekä datan analysointia. Hive on suunniteltu hallitsemaan ja varastoimaan suuria datamääriä käyttäen SQL-kyselyitä. Hiven käyttää SQL:n tapaista kyselykieltä nimeltään HiveQL. Hive antaa mahdollisuuden päästä käsiksi suoraan tallennettuun dataan ja rakentaa struktuurisia tietokantoja HDFS:än päälle. Toisetkin tallennusjärjestelmät ovat tuettuja, kuten Apache HBase. Kyselyt tietokantaan voidaan suorittaa käyttäen Apache Tez, Apache Spark tai MapReducea. (Apache Hive 2016).

Hive tukee sisäänrakennettuna CSV- , Apache Parquet ja Apache ORC –tiedostomuotoja. Käyttäjät pystyvät tarpeen mukaan laajentamaan myös muihin formaatteihin. (Apache Hive 2016).

4.2 Apache Drill

Apache Drill on avoimen lähdekoodin ohjelmistokehys, joka on avoimen lähdekoodin versio Googlen Dremel ja BigQuery ohjelmistoista. Drill on suunniteltu tukemaan useita eri tietokanta järjestelmiä, kuten NoSQL-pohjaiset HBase ja MongoDB. Hadoop pohjaiset järjestelmät, kuten HDFS ja MapR-DB. Lisäksi tuettuja ovat useat pilvipalvelu ratkaisut, kuten Amazon S3, Azure Block Storage, Google Cloud Storage ja Swift. (Apache Drill 2016).

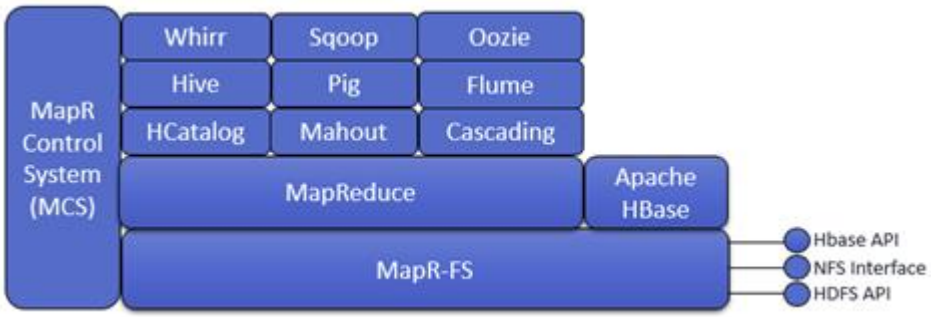
4.2.1 Apache Parquet

Apache Parquet on Apache Software Foundationin (ASF) sponsoroima projekti, jolla pyritään kehittämään tehokkaampaa tallennusmuotoa Hadoop:iin. Perinteisissä tallennusmetodeissa data tallennetaan riveissä ja on optimoitu hakemaan tietokannasta yksi tieto kerrallaan. Apache Parquet mahdollistaa sarakepohjaisen tallennuksen, jossa data tallennetaan kokonaisissa sarakkeissa. Tällöin suurissa tietoaineistoissa suoritetut haut ja datan lukeminen pystytään optimoimaan tehokkaasti. Parquet pystyy pakkaamaan sarakkeet, joka taas lisää suorituskykyä. Useimmat jo käytössä olevat Hadoop projektit voivat lukea ja kirjoittaa dataa, joka on Parquet-muodossa, kuten Hive, Drill, Pig ja MapReduce. Apache Parqueta pystytään käyttämään missä tahansa Hadoop ekosysteemissä, riippumatta ohjelmointikielestä, järjestelmän rakenteesta tai käytetystä datan muodosta. (Apache Parquet 2016)

Apache Parquet ei suinkaan ole ainut järjestelmä, joka pystyy suorittamaan tallennuksen sarakemuodossa. Hive sisältään sen oman ORC-formaatin, jolla voidaan tallentaa sarakkeissa. Se on suunniteltu lähinnä vain Hive:n lisäosaksi, eikä koko Hadoop-järjestelmän yleiseksi tallennusmuodoksi. (Apache Parquet 2016)

5 MAPR

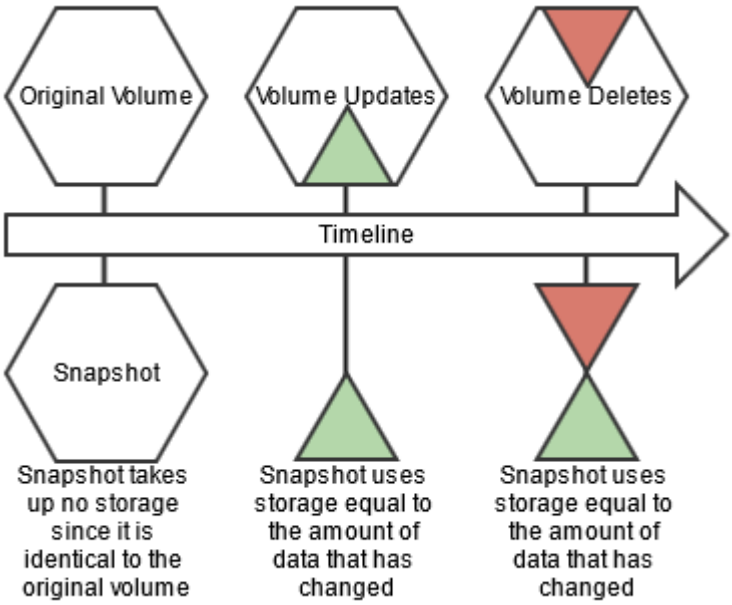
MapR on yritystason jakelu Apache Hadoop:sta, joka on suunniteltu lisäämään Hadoop:in luotettavuutta, suorituskykyä ja helppokäyttöisyyttä. MapR jakelu on täysi Hadoop palvelu, jonka ominaisuuksia ovat MapR:n oma tiedostojärjestelmä MapR-FS (MapR File System), MapReduce, täydellinen Hadoop ekosysteemi, MapR hallintajärjestelmä ja käyttöympäristö (Kuvio 14). (Minal, P. 2015).



Kuvio 14. MapR Arkkitehtuuri

MapR ominaisuudet:

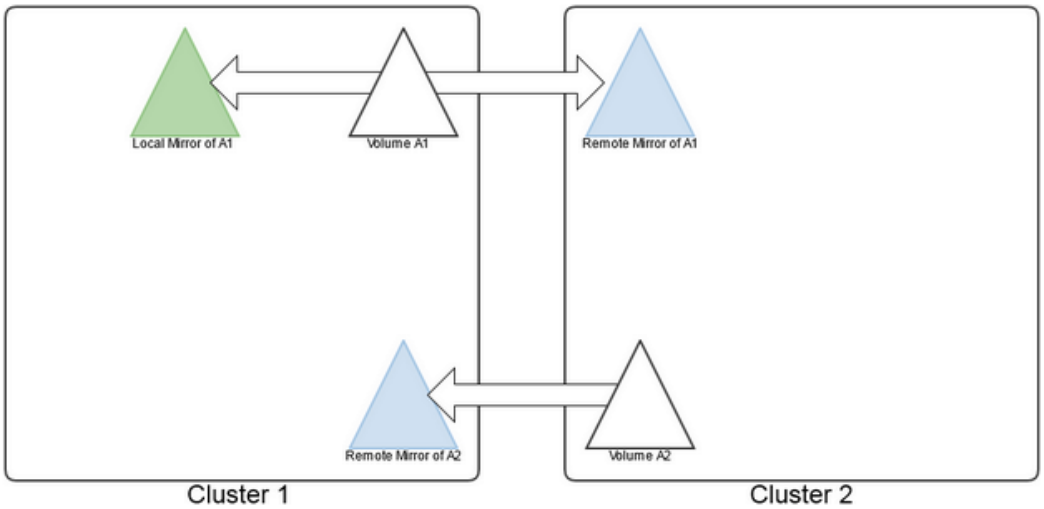
- **Datan suojaus**
  MapR Snapshot ominaisuus tallentaa kuvat tietokoneesta josta häiriön tapahtuessa se on helppo palauttaa. MapR Snapshot käyttää tehokkaasti hyväkseen saatavilla olevaa tallennustilaa ja prosessorin resursseja, tallentamalla uusissa kuvissa vain muuttuneet tiedostot (Kuvio 15).



Kuvio 15. Data Protection

- **Turvallisuus**
  Kaikki liikenne klusterin sisällä, sekä sieltä ulos ja sisäänpäin suunnattu liikenne on salattua. Jokaiselle käyttäjälle on erikseen mahdollista räätälöidä henkilökohtaiset oikeudet ja asetukset.

- **Onnettomuudesta palautuminen**
  MapR peilaa palvelimen klusterissa, joten onnettomuuden sattuessa palvelu on edelleen saatavissa (Kuvio 16).



Kuvio 16. Disaster Recovery With Mirrors

- **Integrointi**
  Klusteriin on helppo syöttää uuttaa dataa NFS jakamisen avulla. Sekä tuki muille Hadoop projekteille kuten Flume ja Sqoop.

- **Suorituskyky**
  MapR käyttää kustomoituja arkkitehtuurielementtejä klusterissaan, joka mahdollistaa lähes täyden nopeuden mitä raudasta on mahdollista saada irti.

- **Skaalautuva arkkitehtuuri**
  MapR jakelu tarjoaa suuren saatavuuden kaikille Hadoopin osille, sekä tuote toimii suoraan laatikosta, eikä vaadi suurta konfigurointia ollakseen toiminta valmis.

5.1 MapR Editions

MapR tarjoaa käyttäjilleen kaksi eri versiota järjestelmästään, joista kumpikin on räätälöity palvelemaan erilaisten yritysten ja käyttäjien eri tarpeita. Ominaisuudet nähdään taulukosta 5.

Taulukko 5. MapR Editions

	Converged Community Edition	**Converged Enterprise Edition**
	For free, unlimited production use.	For critical deployments requiring business continuity (HA/DR).
Modules		
MapR-FS	X	X
Apache Hadoop and Open Source Projects	X	X
MapR-DB	X	X
MapR Streams	X	X
Features and Capabilities		
Performance	X	X
Scalability	X	X
Standards-Based APIs and Tools	X	X
Direct Access NFS	X	X
Manageability	X	X
Integrated Security	X	X
Multi-tenancy	X	X
Advanced Multi-tenancy		X
Consistent Snapshots		X
High Availability		X
Disaster Recovery		X
Global Table Replication for MapR-DB		X
Global Replication for MapR Streams		X
Real-Time Transport for MapR-DB		X
Support Features		
Community/Forum Support	X	X
24x7 Commercial Support		X
Add-on 24x7 Commercial Support Options (Additional support subscription required)		
Apache Drill Support		X
Apache Spark Support		X
Apache HBase Support		X
Apache Solr Support		X
Impala Support		X
MapR POSIX Client	X	X

5.2 MapR Sandbox for Hadoop

MapR Sandbox for Hadoop on täysin toimiva yhden noden klusteri, jonka avulla voidaan suorittaa datan analysointia. Sandbox käyttää MapR omaa hallintajärjestelmää MCS (MapR Control System) ja Hue-käyttöliittymää. (Bevens, B. 2015).

Laitteistovaatimukset:

- VMware Player or VirtualBox is installed
- At least 20 GB free hard disk space, at least 4 physical cores, and 8 GB of RAM is available. Performance increases with more RAM and free hard disk space.
- Uses one of the following 64-bit x86 architectures:
    - A 1.3 GHz or faster AMD CPU with segment-limit support in long mode
    - A 1.3 GHz or faster Intel CPU with VT-x support

5.3 MapR Sandbox with Apache Drill

Apache Drill on avoimen lähdekoodin SQL-kyselymoduuli Hadoop:iin ja NoSQL:ään. Drill sisältää pienen viiveen jopa tuhansille yhtäaikaisille käyttäjille. Mahdollisuus suorittaa interaktiivisia SQL-kyselyitä. Eikä se ole sidottu näissä kyselyissä ennalta määritettyihin skeemoihin, vaan käyttää Schema discovery on-the-fly ominaisuutta, joka mahdollistaa uusien skeemojen lisäämisen ja käyttämisen lennosta. (MapR Drill Sandbox 2016).

Laitteistovaatimukset:

- VMware Player or VirtualBox is installed.
- At least 20 GB free hard disk space, at least 4 physical cores, and 8 GB of RAM is available. Performance increases with more RAM and free hard disk space.
- Uses one of the following 64-bit x86 architectures:
    - A 1.3 GHz or faster AMD CPU with segment-limit support in long mode
    - A 1.3 GHz or faster Intel CPU with VT-x support

6 MAPR SANDBOX FOR HADOOP TOTEUTUS

6.1 MapR Sadbox for Hadoop asennus

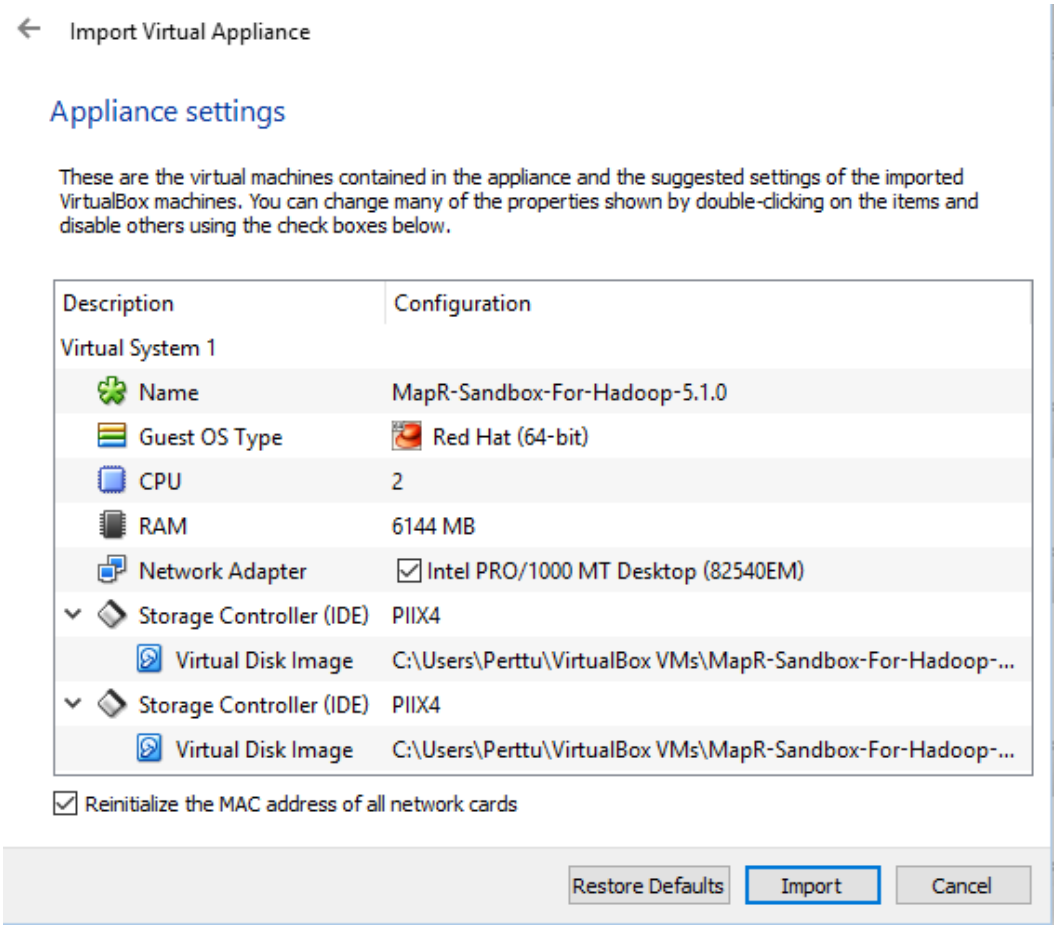
MapR tarjoaa kuvan kaikkiin Sandbox versioihin, minkä avulla minkä avulla se pystytään asentamaan VirtualBoxiin.

Osoitteesta *http://package.mapr.com/releases/* josta navigoidaan haluttuun versioon.

Tässä asennuksessa käytettiin uusinta versiota 5.1.0

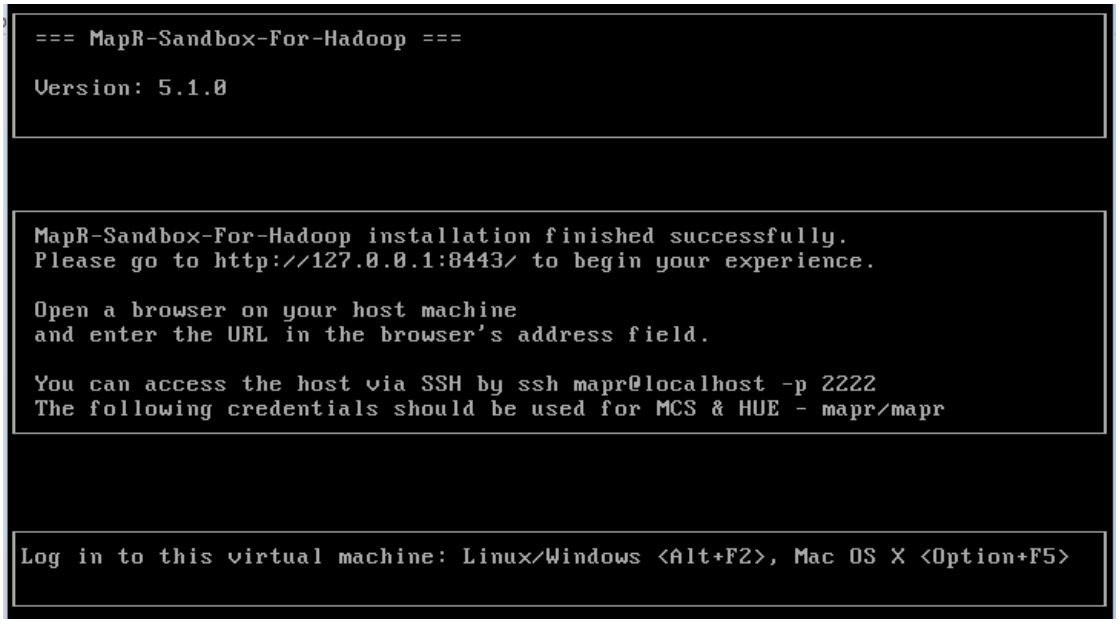
http://package.mapr.com/releases/v5.1.0/sandbox/MapR-Sandbox-For-Hadoop-5.1.0.ova

Käyttäen VirtualBoxin import toimintoa saa asennettua virtuaalikoneen ladatusta tiedostosta, samalla näemme tarvittavat järjestelmävaatimukst (Kuvio 17)

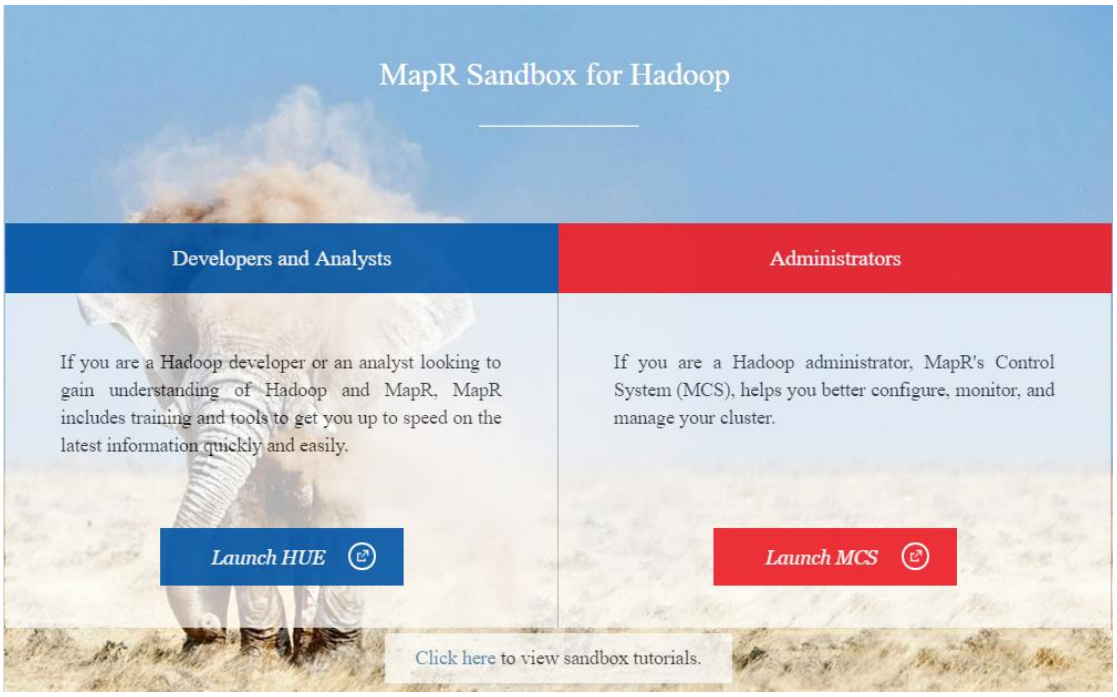


Kuvio 17. VirtualBox Hadoop import

Onnistuneen asennuksen jälkeen Sandbox serveri antaa ohjeet, jolla pääsee jatkamaan asennusta (Kuvio 18). Avataan selain ja siirrytään osoitteeseen http://127.0.0.1:8443/ , tai vaihtoehtoisesti voidaan käyttää osoitetta http://localhost:8443 (Kuvio 19).

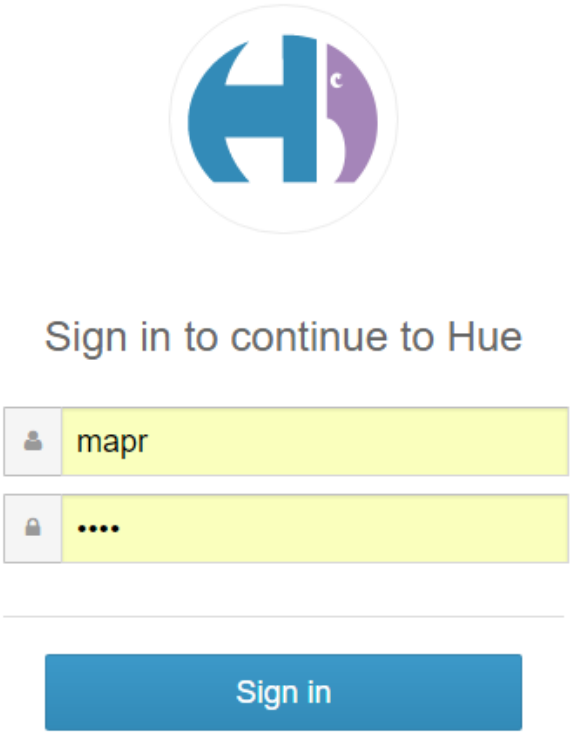


Kuvio 18. MapR Sandbox For Hadoop asennettu



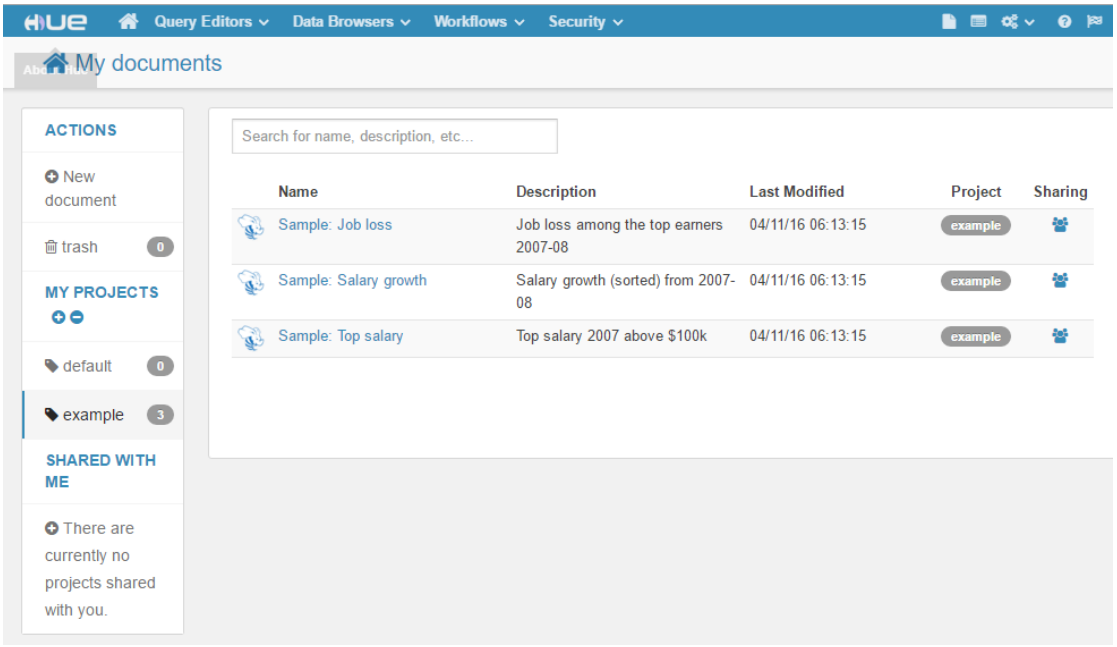
Kuvio 19. http://127.0.0.1:8443/

Kirjaudutaan Hue-käyttöliittymään tunnuksilla mapr/mapr (Kuvio 20).



Kuvio 20. Hue login

Hue–käyttöliittymästä löytyy esimerkkejä, jolla MapR Sandboxin toimintaa voidaan demota (Kuvio 21).



Kuvio 21. Hue examples

6.2 Datan lisääminen

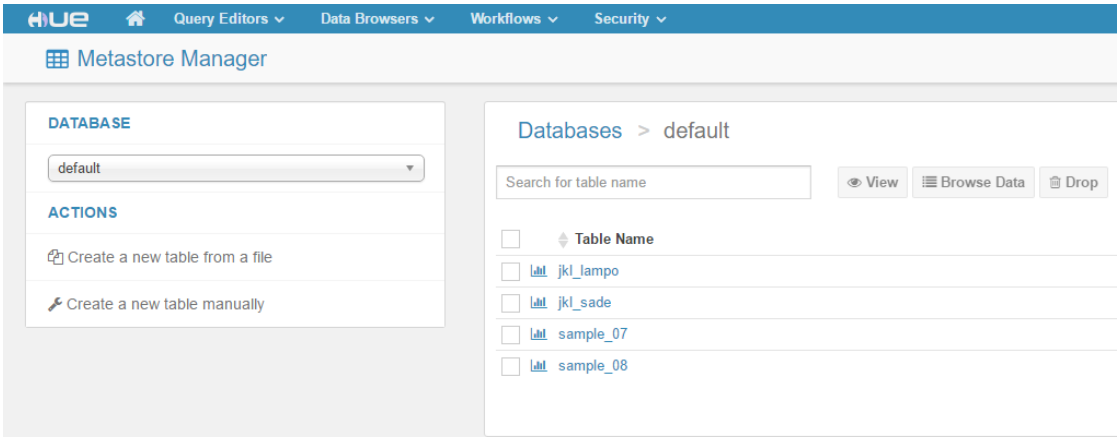
Ilmatieteen laitos tarjoaa omilla sivuillaan *http://ilmatieteenlaitos.fi/ilmasto,* mahdollisuuden tarkastella viimeisen 30 vrk:n säätietoja, lämpötila- ja sadetilastoja vuodesta 1961 sekä monia muita säähän liittyviä tilastoja. Ilmatieteen laitos antaa mahdollisuuden ladata nämä säätiedot omaan käyttöön, useassa eri formaatissa. Lataaminen on mahdollista kyseisissä formaateissa: PNG, JPEG, PDF, SVG ja CSV, jota käytetään tässä toteutuksessa (Kuvio 22). Valitsin käytettäväksi Jyväskylän heinäkuun lämpötilastot vuodesta 1961 eteenpäin.

Ladattu CSV-tiedostossa lämpötilat on merkattu käyttäen pilkkua, jota MapR ei ymmärrä ja näyttää tällöin kyseisessä kohtaa vain arvoa NULL. Tämä saatiin korjattua vaihtamalla kaikki CSV-tiedoston pilkut pisteiksi.

```
"Category";"Jakson keskilÃ¤mpÃ¶tila";"KeskilÃ¤mpÃ¶tilan
poikkeama 1981-2010 keskiarvosta"
1961;16,4;-0,3
1962;14,1;-2,6
1963;16,5;-0,2
1964;16,4;-0,3
1965;14,5;-2,2
1966;16,8;0,1
1967;16,4;-0,3
1968;15;-1,7
1969;15,8;-0,9
1970;16,2;-0,5
1971;15,7;-1
1972;19,5;2,8
1973;19,2;2,5
1974;15,6;-1,1
1975;16,5;-0,2
1976;14,4;-2,3
1977;14,5;-2,2
```

Kuvio 22. Jyväskylä lämpötila CSV

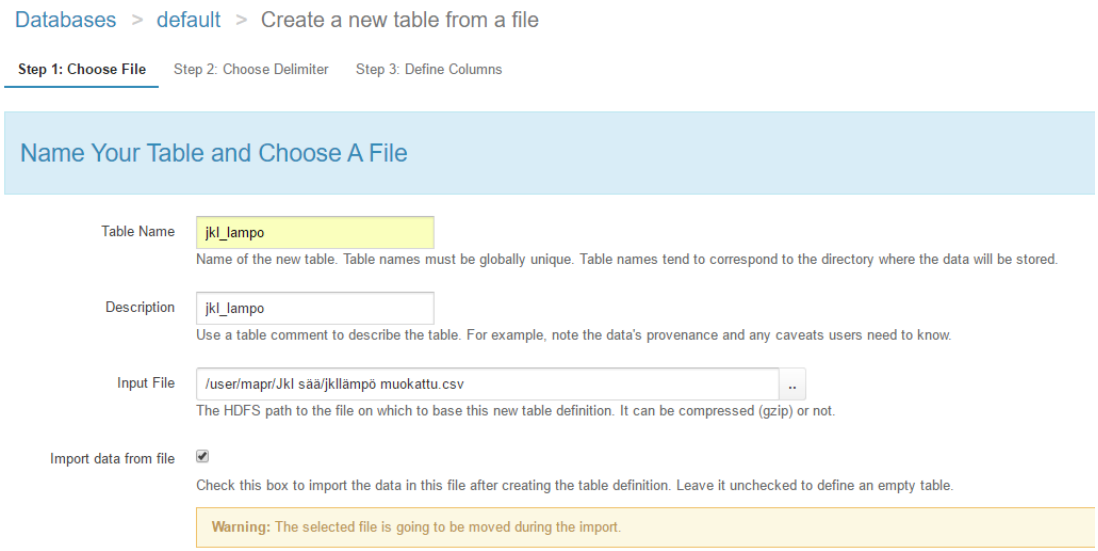
Käyttäen HUE Metastore Manageria, joka on SQL-tietokantasovellus ja se luo uuden tietokannan muokatusta CSV-tiedostosta (Kuvio 23).



Kuvio 23. Metastore Manager

Step 1:

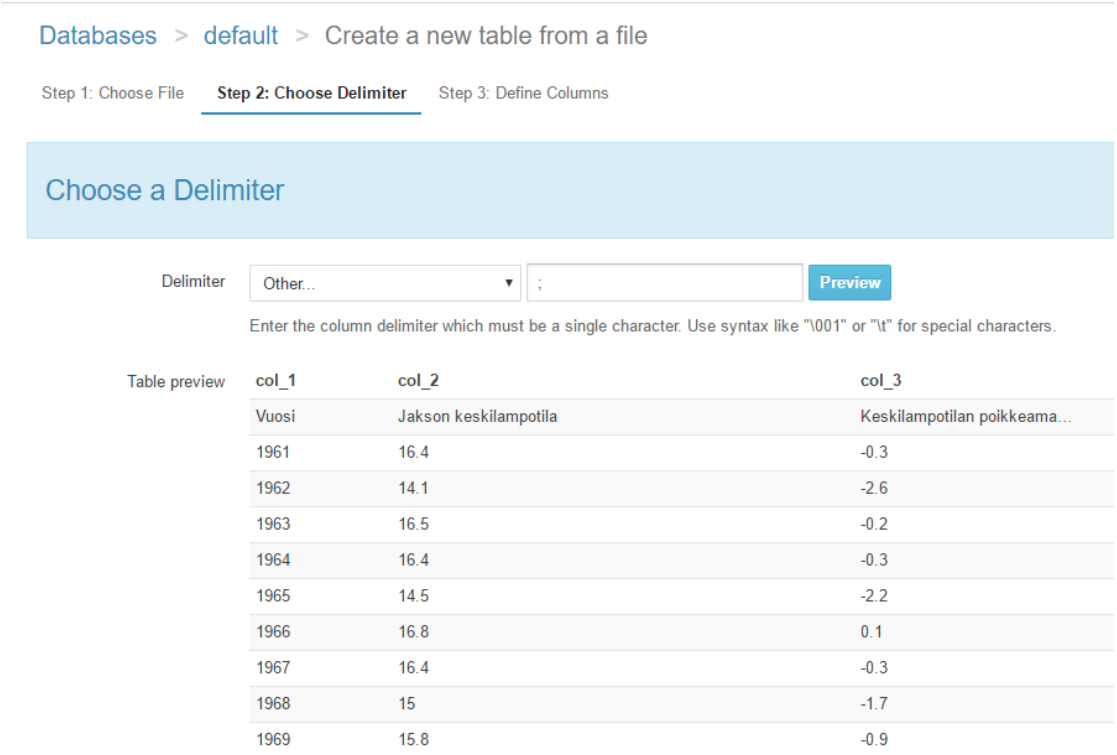
Luodaan uusi tietokanta, nimetään se ja valitaan haluttu CSV-tiedosto, joka tuodaan virtuaalikoneelle. Tiedoston lataaminen onnistuu suoraan omalta koneelta, käyttäen Hue-käyttöliittymää. Eikä sitä tarvitse siirtää erillisellä FTP-ohjelmalla. (Kuvio 24).



Kuvio 24. Uuden tietokannan luonti

Step 2:

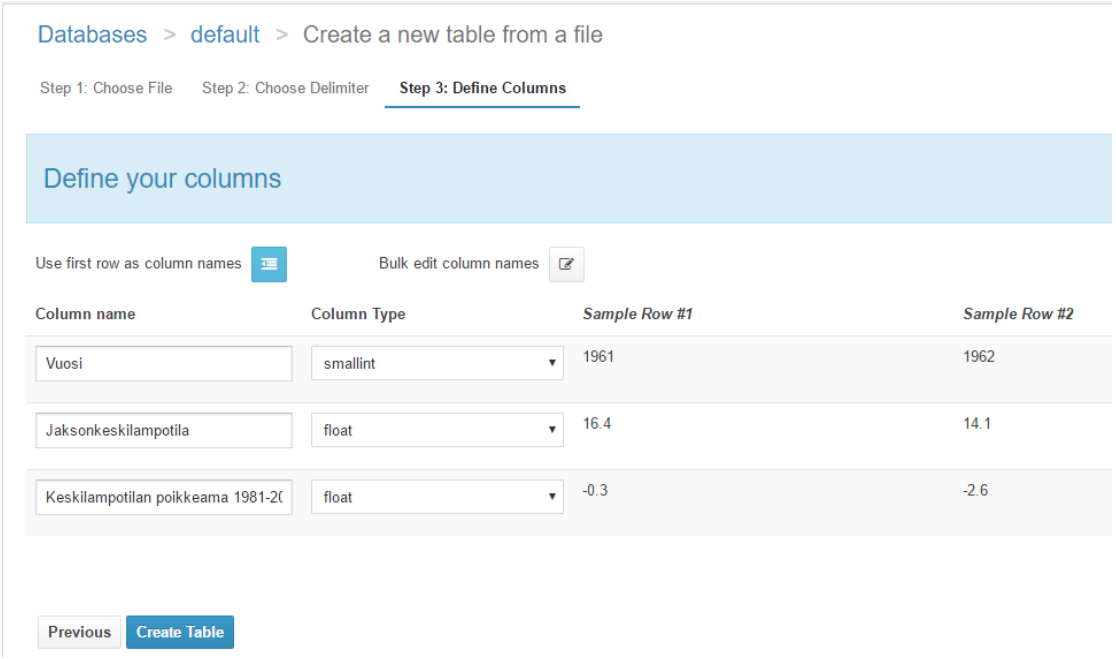
Seuraavaksi pitää valita sopiva erotin. Metastore Manager yrittää tunnistaa ladatusta tiedostosta sopivan erottimen, mutta tässä tapauksessa se ei ollut oikea. CSV-tiedoston tarkastelun jälkeen huomattiin, että kyseisessä tiedostossa erottimena käytetään puolipistettä. Joten erottimeksi täytyy valita *Other,* seuraavaan kohtaan pystytään itse syöttämään haluttu erotin, eli tässä tapauksessa puolipiste (Kuvio 25).



Kuvio 25. Erottimen valinta

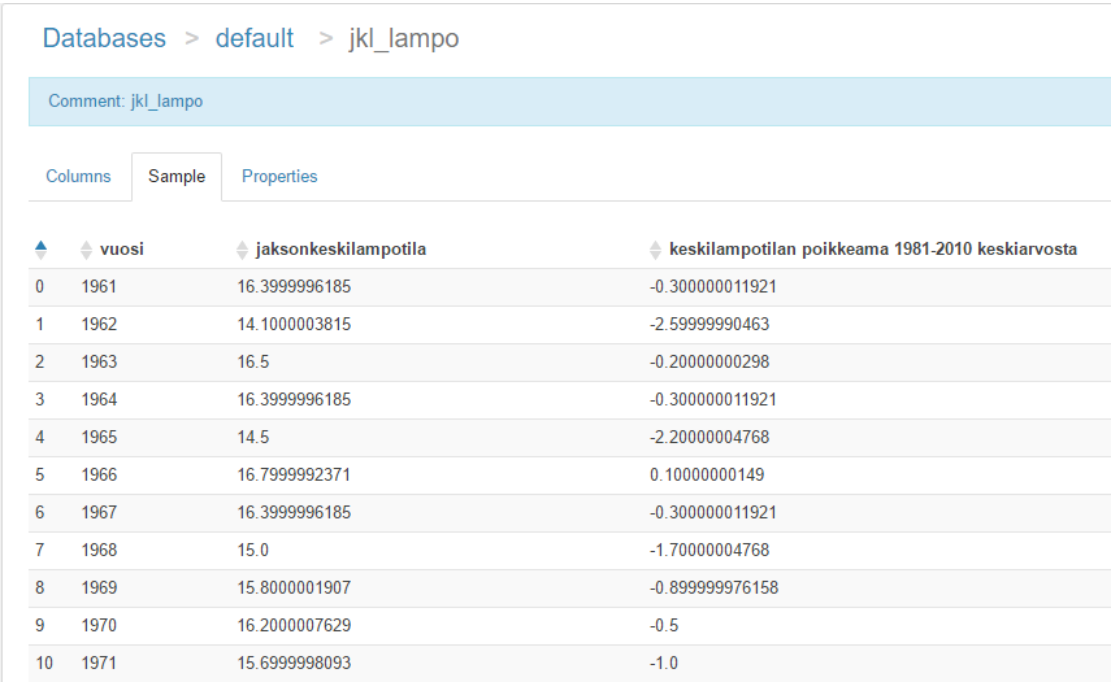
Step 3:

Tässä voidaan muokata tulevien sarakkeiden nimeä ja tyyppiä (Kuvio 26).



Kuvio 26. Sarakkeiden muoto

Ennen tietokannan luontia, ulkoasua pystytään esikatselemaan ja varmistamaan, että muotoilu on oikeanlainen (Kuvio 27).



Databases > default > jkl_lampo

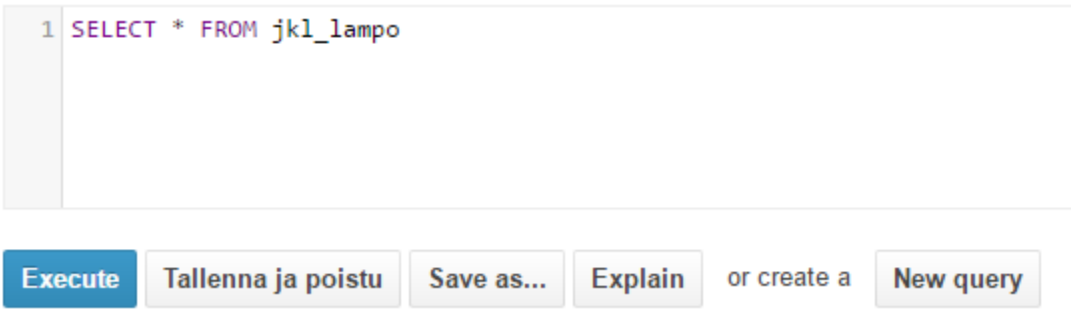
Comment: jkl_lampo

Columns | Sample | Properties

	vuosi	jaksonkeskilampotila	keskilampotilan poikkeama 1981-2010 keskiarvosta
0	1961	16.3999996185	-0.300000011921
1	1962	14.1000003815	-2.59999990463
2	1963	16.5	-0.20000000298
3	1964	16.3999996185	-0.300000011921
4	1965	14.5	-2.20000004768
5	1966	16.7999992371	0.10000000149
6	1967	16.3999996185	-0.300000011921
7	1968	15.0	-1.70000004768
8	1969	15.8000001907	-0.899999976158
9	1970	16.2000007629	-0.5
10	1971	15.6999998093	-1.0

Kuvio 27. Tietokannan esikatselu

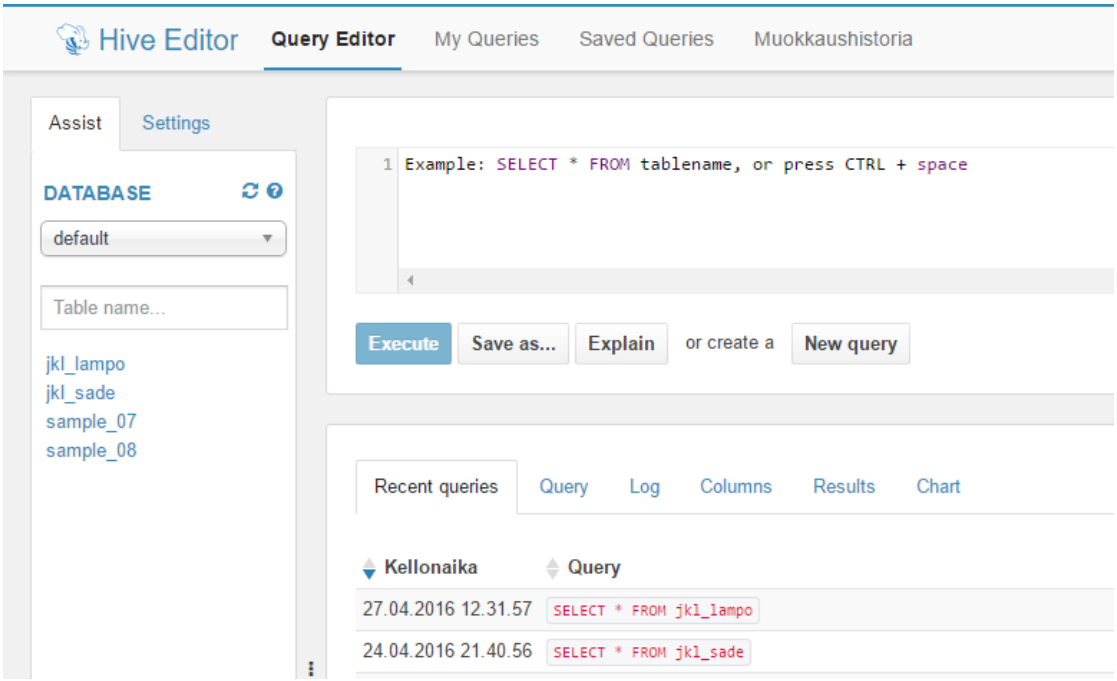
Hive Editorilla pystytään ajamaan kyselyitä luotuun tietokantaan. Komennolla *SELECT * FROM jkl_lampo* (Kuvio 28) saadaan haettua koko tietokannan tiedot (Kuvio 29).



```
1 SELECT * FROM jkl_lampo
```

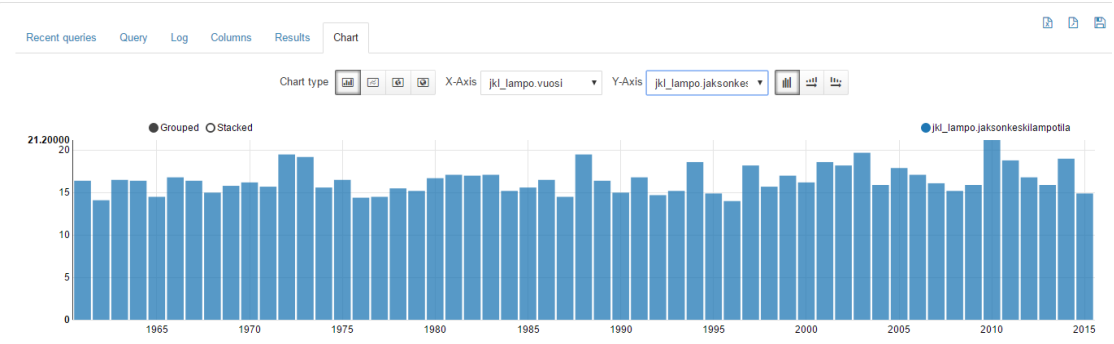
Execute | Tallenna ja poistu | Save as... | Explain | or create a | New query

Kuvio 28. Tietokantakysely

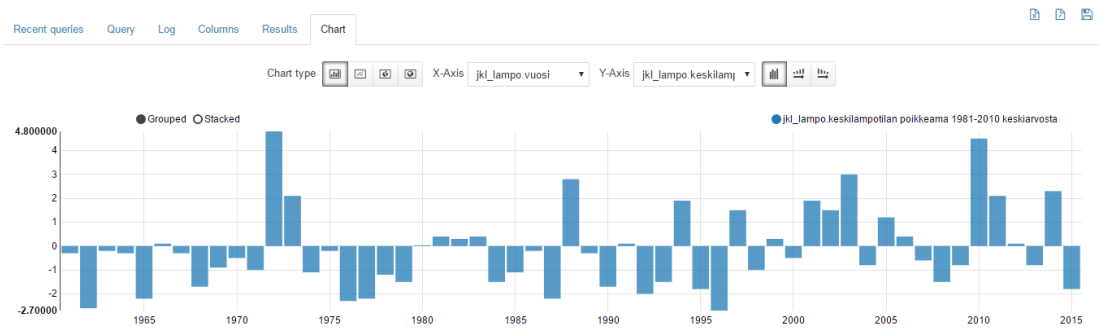


Kuvio 29. Hive Editor

Suoritetun kyselyn jälkeen, Chart välilehdestä saadaan piirrettyä diagrammia kyselyn tuloksista. X-akseliin asetetaan vuosi ja Y-akseliin lämpötila, jolloin saadaan pylväsdiagrammit Jyväskylä heinäkuun keskilämpötila 1961-2015 (Kuvio 30). Sekä heinäkuun lämpöpoikkeamat 1961-2015 (Kuvio 31).



Kuvio 30. Jyväskylä keskilämpötila heinäkuu 1961-2015



Kuvio 31. Jyväskylä lämpöpoikkeama heinäkuu 1961-2015

7 MAPR SANDBOX FOR APACHE DRILL TOTEUTUS

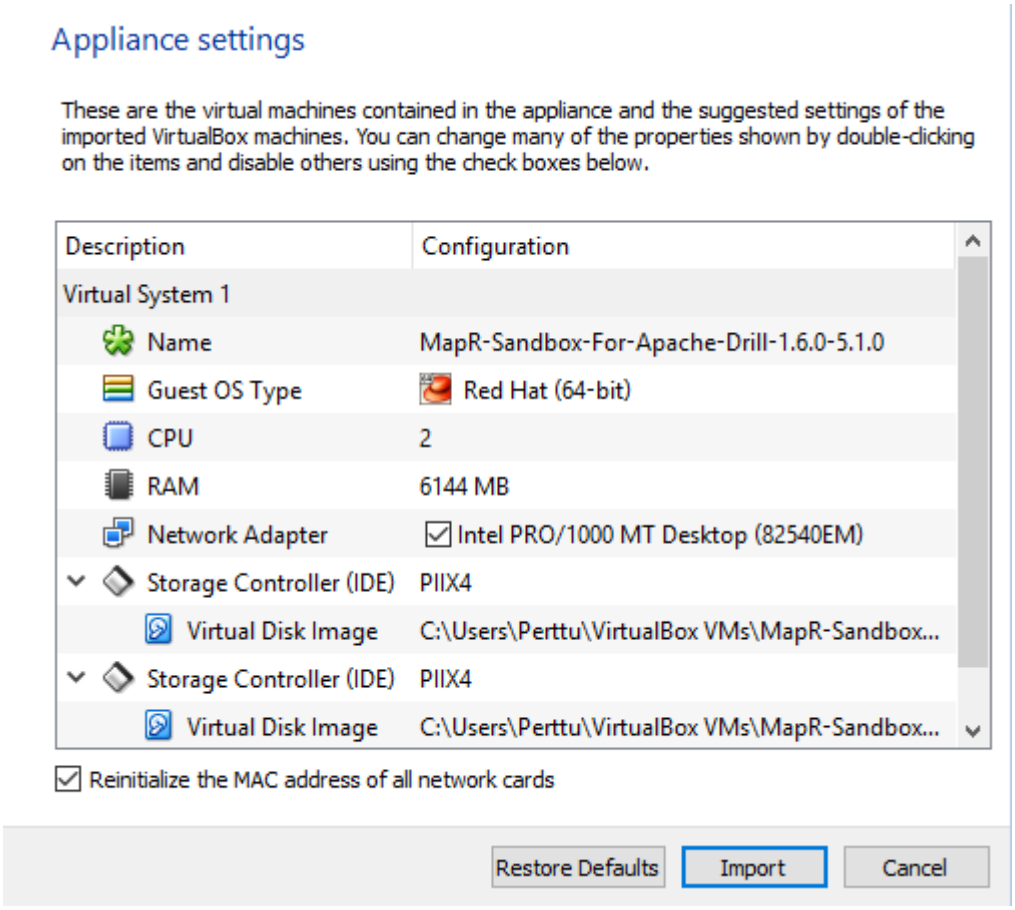
Tässä toteutuksessa muunnetaan CSV-muodossa oleva tiedosto Parquet-muotoon. Toteutuksessa käytetty CSV-tiedosto on esimerkkitiedosto lentoliikenteestä, joka löytyy osoitteesta *http://media.flysfo.com/media/sfo/media/air-traffic/Passenger_4.zip*

7.1 MapR Sandbox For Apache Drill asennus

MapR Sandbox For Apache Drill-kuvan lataaminen VirtualBox asennusta varten, vaatii yhteystietojen lähettämisen.

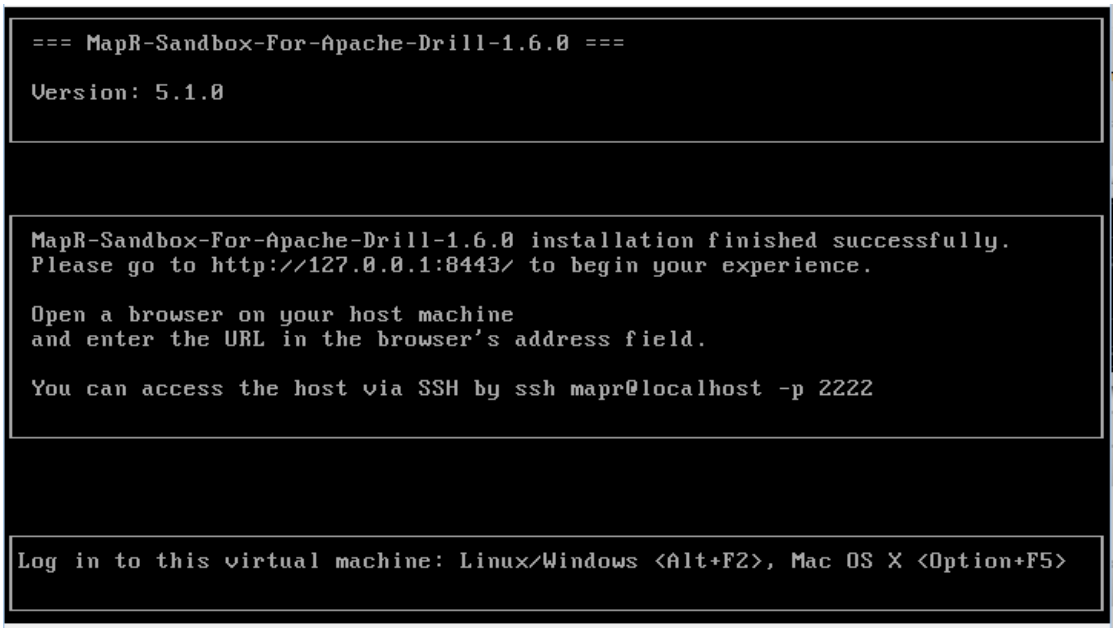
https://www.mapr.com/products/mapr-sandbox-hadoop/download-sandbox-drill

Käyttäen VirtualBoxin import toimintoa saa asennettua virtuaalikoneen ladatusta tiedostosta, samalla näemme tarvittavat järjestelmävaatimukset (Kuvio 32).



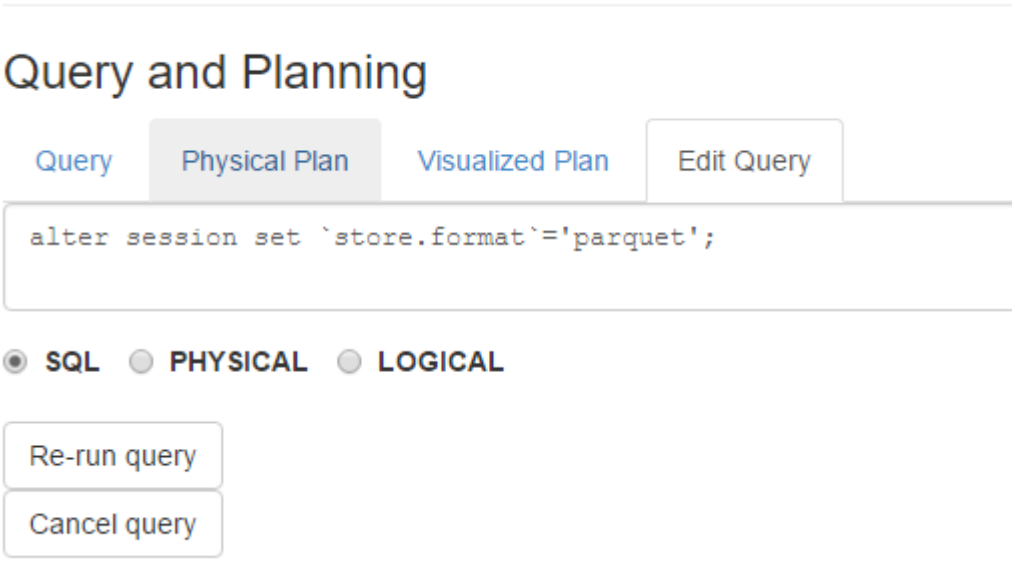
Kuvio 32. VirtualBox Drill import

Onnistuneen asennuksen jälkeen Sandbox serveri antaa ohjeet, jolla pääset jatkamaan asennusta (Kuvio 18). Avataan selain ja siirrytään osoitteeseen http://127.0.0.1:8443/, tai vaihtoehtoisesti voidaan käyttää osoitetta http://localhost:8443



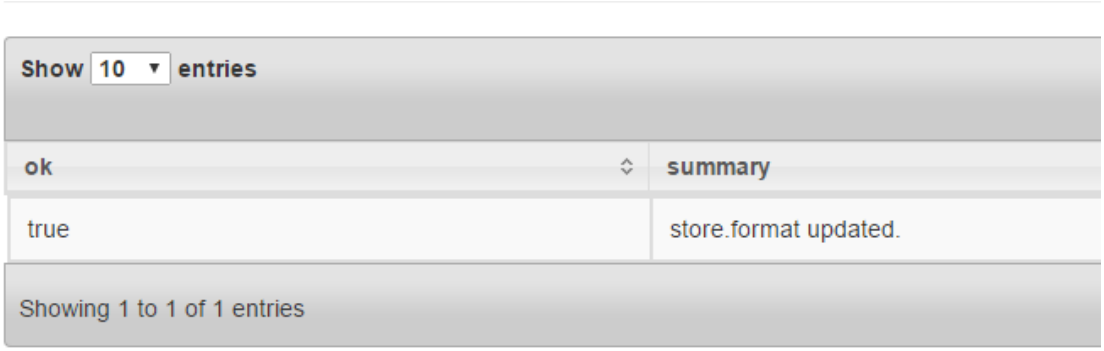
Kuvio 33. MapR Sandbox For Apache Drill asennettu

Määritetään haluttu tallennusmuoto komennolla *alter session set `store.format`='parquet';* (Kuvio 34), joka vaihtaa kyseisen istunnon aikana tallennetut taulut parquet muotoon. Tuettuja formaatteja ovat CSV, JSON ja Parquet.



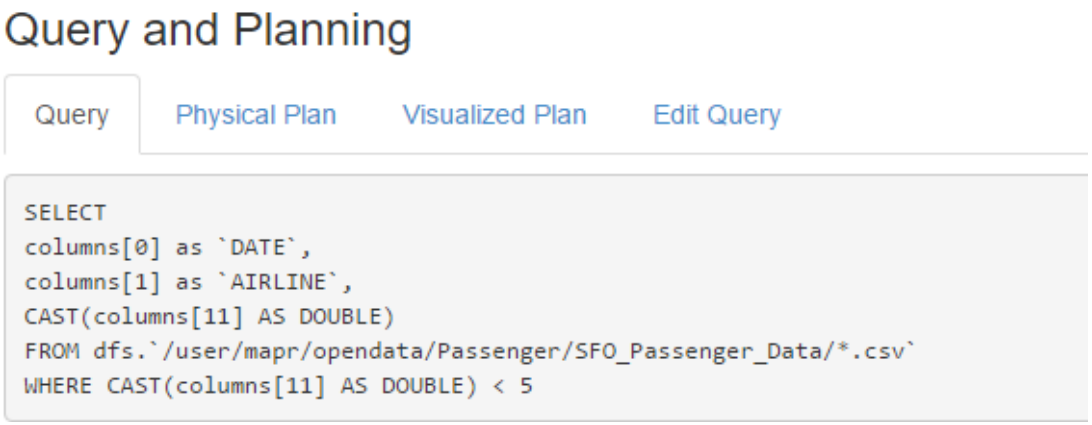
Kuvio 34. Tallennus formaatin vaihtaminen

Kun kysely suoritetaan, niin Drill kuittaa vaihdon onnistuneen. (Kuvio 35).



Kuvio 35. Kuittaus onnistuneesta vaihdosta

Käyttäen FTP-ohjelmaa, haluttu CSV-tiedosto on siirretty virtuaalikoneelle. Suoritetaan kysely, jossa todetaan järjestelmän toimivan(Kuvio 36).



Kuvio 36. Drill kysely CSV tiedostosta

Kysely palauttaa kyselyssä määritetyt sarakkeet (Kuvio 37).

Show 10 entries

DATE	AIRLINE
200610	United Airlines - Pre 07/01/2013
200611	Ameriflight
200611	Ameriflight
200611	Evergreen International Airlines
200611	Evergreen International Airlines
200612	Ameriflight
200612	Ameriflight
200710	Southwest Airlines
200711	Southwest Airlines
200712	Ameriflight

Showing 1 to 10 of 26 entries

Kuvio 37. CSV kyselyn tuloste

Seuraavaan kyselyyn lisätään toiminto, jolla luodaan uusi taulu haluttuun kansioon. Aikaisemmin määritettiin tallennusformaatti, joten uusi taulu tallentuu parque-muodossa (Kuvio 38).

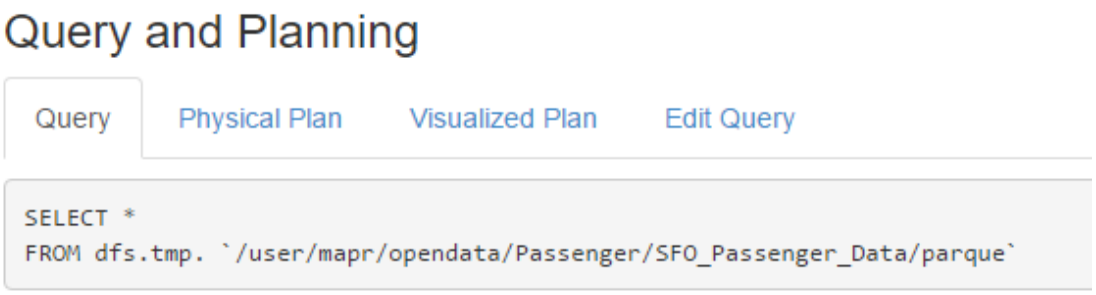
Query and Planning

Query | Physical Plan | Visualized Plan | Edit Query

```
CREATE TABLE dfs.tmp.`/user/mapr/opendata/Passenger/SFO_Passenger_Data/parque` AS
SELECT
columns[0] as `DATE`,
columns[1] as `AIRLINE`,
CAST(columns[11] AS DOUBLE)
FROM dfs.`/user/mapr/opendata/Passenger/SFO_Passenger_Data/*.csv`
WHERE CAST(columns[11] AS DOUBLE) < 5
```

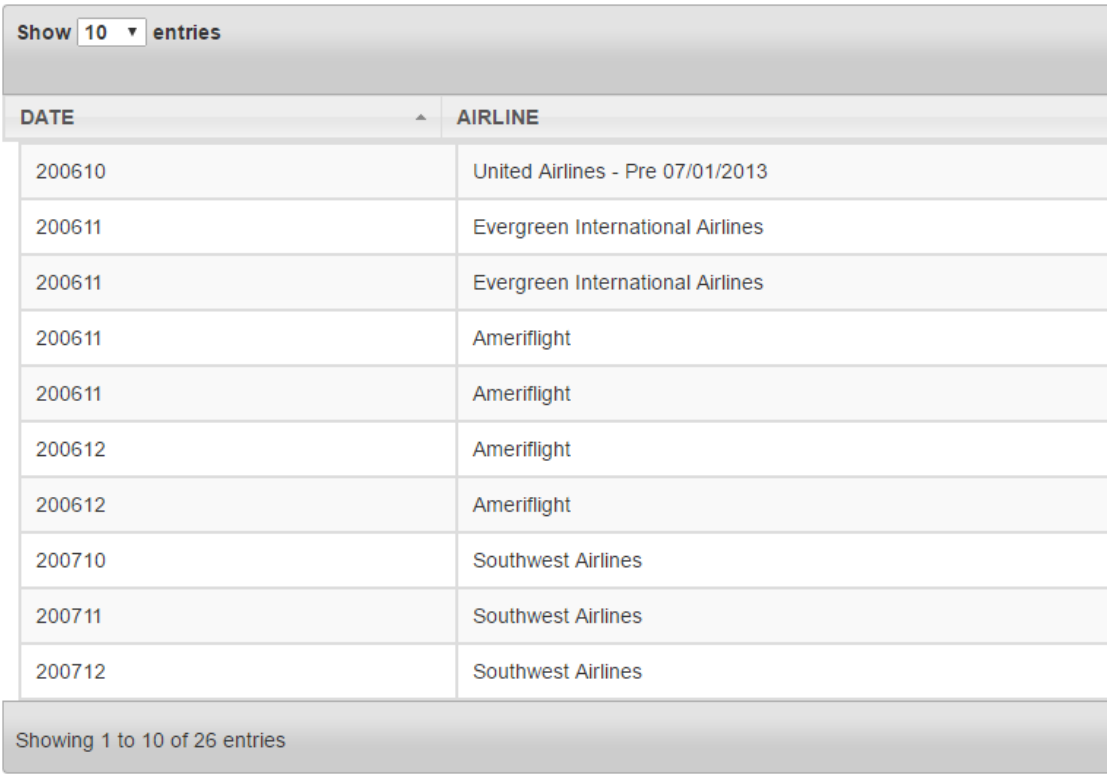
Kuvio 38. CSV-tiedoston muuntaminen Parque muotoon

Testataan onnistunut muunnos suorittamalla kysely tallennettuun parquet-tauluun (Kuvio 39).



Kuvio 39. Kysely tallennetusta Parque-tiedostosta

Kysely palauttaa Parquet-taulun (Kuvio 40).



Kuvio 40. Parque-tiedoston tuloste

8 POHDINTA

8.1 Pohdinta

Big data ja avoin data olivat käsitteenä ja käytäntönä aivan uusi, sekä ajankohtainen aihealue. Nykyään uutta dataa luodaan koko ajan kasvavalla tahdilla ja tähän tarvitaan oikeanlaisia työkaluja. Tällä hetkellä suunta on oikea kun tarjolla on koko ajan kasvava määrä eri ratkaisuja, jolla tähän ongelmaan pystytään vastaamaan.

Erilaisen avoimen datan tarjonta on laaja ja tietovarantojen kasvaa koko ajan, sitä mukaa mitä kaupungit ja organisaatiot avaavat omia tietovarantojaan julkiseen käyttöön. Avoindata.fi portaali on ehdottomasti yksi parhaista ja käyttäjäystävällisemmistä suomalaisista palveluista, julkaista ja käyttää avoimen data tietolähteitä. Liikenneministeriön ylläpitämä digitraffic palvelu ja sen avoimet rajapinnat ovat hyvin suunniteltu. Hieman suuremmalla syventymisellä, tästä palvelusta saisi aikaan useita erilaisia applikaatioita. Tämän mahdollistavat avoimet rajapinnat ja näistä saatava reaaliaikainen data. Työssä käytetty MapR Sandbox For Hadoop oli järjestelmänä kuitenkin niin suljettu, että reaaliaikaisen datan kuuntelu ei ollut mahdollista. Joten tässä täytyi tyytyä yksinkertaisempaan datan analysointiin.

Työssä mielenkiintoisena asiana tuli vastaan suuri määrä eri tiedostomuotoja ja erilaisia sovelluksia, jolla dataa pystytään tehokkaasti hallitsemaan ja analysoimaan. Tämä tarve tulee varmasti tulevaisuudessa vielä lisääntymään. Työssä käytettyjä tietoja tulee varmasti tarvitsemaan tulevaisuudessa jos siirtyy työskentelemään big datan tai datan analysoinnin pariin.

LÄHTEET

Amazon Web Services, Inc. 2016. AWS Public Data Sets. 2016. Viitattu 21.2.2016. https://aws.amazon.com/public-data-sets/

Apache Drill 2016. Schema-free SQL Query Engine for Hadoop, NoSQL and Cloud Storage. Viitattu 9.5.2016. https://drill.apache.org/

Apache Parquet 2016. Apache Software Foundation. Viitattu 9.5.2016. https://parquet.apache.org/

Apache Hive 2016. The Apache Hive data warehouse software. Viitattu 4.5.2016. https://cwiki.apache.org/confluence/display/Hive/Home

Avoindata 2016. Avoimen tiedon ja yhteen toimivuuden palvelu. 2016. Viitattu 3.3.2016. https://www.avoindata.fi/fi

Bevens, B. 2015. MapR Sandbox for Hadoop. Viitattu 4.4.2016. http://doc.mapr.com/display/MapR/MapR+Sandbox+for+Hadoop

Creative Commons. 2016. Tietoa creative commons lisensseistä. 2016. Viitattu 3.3.2016. https://creativecommons.org/licenses/

Digitraffic tieliikenne. Liikenneviraston avoin data. 2016. Viitattu 14.3.2016. http://www.liikennevirasto.fi/avoindata/palvelut/digitraffic

Digitraffic rautatieliikenne. Liikenneviraston avoin data. 2016. Viitattu 14.3.2016. http://rata.digitraffic.fi/api/v1/doc/index.html

Eaton, C. Deroos, D. Deutsch, T. Lapis, G. Zikopoulos P. 2012. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming. Viitattu 28.3.2016. http://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14296usen/IML14296USEN.PDF

Google Inc. Public Data Help. 2016. Viitattu 21.2.2016. https://support.google.com/publicdata/?hl=en#topic=1100622

Helsinki Region Infoshare.2016. Mitä on avoin data? 2016. Viitattu 14.3.2016. http://www.hri.fi/fi/mita-on-avoin-data/

JYVSECTEC 2016. JYVSECTEC-hankkeen kotisivut. Viitattu 8.2.2016. http://www.jyvsectec.fi/

Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman 2013. Big Data for Dummies. Viitattu 26.2.2016 http://www.jamk.fi/fi/Palvelut/kirjasto/Oppaat/tietotekniikka/, Books24x7

MapR Drill Sandbox. 2016. Apache Drill enables self-service data exploration on big data with a schema-free SQL query engine Viitattu 9.5.2016. https://www.mapr.com/products/apache-drill

Mapreduce 2016. A Very Brief Introduction to MapReduce. 2011. Viitattu 4.4.2016. http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf

Minal, P. 2015. MapR Overview. Viitattu 4.4.2016. http://doc.mapr.com/display/MapR/MapR+Overview

National Security Agency. Guidelines for implementation of REST. 2011. Viitattu 20.3.2016. https://www.nsa.gov/ia/_files/support/guidelines_implementation_rest.pdf

Opendefinition 2016. Avoimen tiedon määritelmä Versio: 1.1. Viitattu 29.2.2016. http://opendefinition.org/od/1.1/fi/

Poikola, A., Kola, P. & Hintikka, K. 2010. Julkinen data – johdatus tietovarantojen avaamiseen, Liikenne ja viestintäministeriö 2010. Viitattu 29.2.2016. http://www.lvm.fi/-/julkinen-data-johdatus-tietovarantojen-avaamiseen-816729

Salo, I. 2013. Big Data, Tiedon vallankumous. Viitattu 26.2.2016. Jyväskylä; Docendo Oy.

YARN. 2016. Apache Software Foundation. Apache Hadoop YARN. Viitattu 28.3.2016 http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html