Aalto University
School of Science
Degree Programme of Computer Science and Engineering

Hussnain Ahmed

Using Big Data Analytics for Measuring Energy Consumption Patterns:

Applying big data for energy efficiency

Master's Thesis
Espoo, June, 2014

DRAFT! — June 18, 2014 — DRAFT!

Supervisors: Professor Professor Matti Vartiainen, Aalto University
Professor Jukka Nurminen, Aalto University
Instructor: Sanja Scepanovic M.Sc. (Tech.)

Aalto University
School of Science

Aalto University
School of Science
Degree Programme of Computer Science and Engineering

ABSTRACT OF
MASTER'S THESIS

Author:	Hussnain Ahmed		
Title:			
Using Big Data Analytics for Measuring Energy Consumption Patterns: Applying big data for energy efficiency			
Date:	June, 2014	**Pages:**	40
Professorship:	Data Communication Software	**Code:**	T-110
Supervisors:	Professor Matti Vartiainen Professor Jukka Nurminen		
Instructor:	Sanja Scepanovic M.Sc. (Tech.)		

A dissertation or thesis is a document submitted in support of candidature for a degree or professional qualification presenting the author's research and findings. In some countries/universities, the word thesis or a cognate is used as part of a bachelor's or master's course, while dissertation is normally applied to a doctorate, whilst, in others, the reverse is true.

!Fixme **Abstract text goes here (and this is an example how to use fixme).** Fixme! Fixme is a command that helps you identify parts of your thesis that still require some work. When compiled in the custom `mydraft` mode, text parts tagged with fixmes are shown in bold and with fixme tags around them. When compiled in normal mode, the fixme-tagged text is shown normally (without special formatting). The draft mode also causes the "Draft" text to appear on the front page, alongside with the document compilation date. The custom `mydraft` mode is selected by the `mydraft` option given for the package `aalto-thesis`, near the top of the `thesis-example.tex` file.

The thesis example file (`thesis-example.tex`), all the chapter content files (`1introduction.tex` and so on), and the Aalto style file (`aalto-thesis.sty`) are commented with explanations on how the Aalto thesis works. The files also contain some examples on how to customize various details of the thesis layout, and of course the example text works as an example in itself. Please read the comments and the example text; that should get you well on your way!

Keywords:	big data, energy, smart grid,energy efficiency, hadoop, analytics,machine learning, classification, CIVIS
Language:	English

Acknowledgements

I wish to thank all students who use LaTeX for formatting their theses, because theses formatted with LaTeX are just so nice.

Thank you, and keep up the good work!

Espoo, June, 2014

Hussnain Ahmed

Abbreviations and Acronyms

2k/4k/8k mode	COFDM operation modes
3GPP	3rd Generation Partnership Project
ESP	Encapsulating Security Payload; An IPsec security protocol
FLUTE	The File Delivery over Unidirectional Transport protocol
e.g.	for example (do not list here this kind of common acronymbs or abbreviations, but only those that are essential for understanding the content of your thesis.
note	Note also, that this list is not compulsory, and should be omitted if you have only few abbreviations

Contents

Chapter 1

Introduction

In modern era we have seen phenomenal increase in human dependency on information and communication technology (ICT) enabled products and services. It has transformed the way of life on the planet. From fulfilling very basic physiological needs like health and food to the human needs of communicating with others and being part of wider social groups, we need and depend on ICT. There are many research areas and opportunities that are emerging as bi-products of this continuous transformation. One of them is the availability of digital traces of human activities. With every instance of use of these services we produce a digital trace that can be recorded and analysed. Big Data is a term that is being widely used to refer to these digital traces of human activity. Ubiquity of computing resources, fast and highly mobile connectivity and advent of social media usage has caused a great surge in volumes of data. Realizing the true potentials of data, businesses are not only utilizing it as source of decision making but new revenue lines and opportunities are emerging that are reshaping the business models of many companies around the globe.

To support this transfiguration, we have seen a rapid development in distributed parallel computing, data communication software and machine learning. Industry giants like Google and Yahoo has opened technologies and tools like MapReduce and Hadoop to facilitate these advancement and open source software communities like Apache Software foundation has further developed the tools to provide a complete ecosystem for handling big data and generate insights. The new specialized big data companies like Cloudera and Hortonworks has emerged that has acted as catalyst for this data revolution. In this research we try to formulate a model for end to end big data analytics platform based on these technologies that can ingest data from heterogeneous sources, process it in an efficient way, mine the data to generate the insights based on business logic and then present the information using interactive visualizations. This thesis includes the development as well as implementation of the mentioned big data platform to perform analysis on real life use case and generate useful insights. The model that we present in this thesis is based on open source software components available free of charge. There are other closed source software alternatives that can fit into the presented model but they are not discussed in this scope of this thesis.

This thesis is also part of European Union CIVIS- Cities as drivers of social change project under 7th framework. CIVIS project focuses on adaption of ICT tools and techniques for low carbon smart energy grid, distributed energy and information flow. The use of pervasive ubiquitous computing is driving the smart energy solutions. Combined with internet of things (IoT) for home/building automation, smart commuting, and remote monitoring is becoming the basis for energy conservation and energy efficiency. All the smart energy devices as part of this ecosystems generates high volumes of data, that needs to be instantaneously transferred, stored, analysed and visualized for knowledge discovery and improvements of services for the goal of achieving high energy efficiency. The platform that was developed as part of this thesis has the capability to automate the whole process.

Energy usage pattern detection, classification of buildings on basis of energy efficiency and a prediction model for energy consumption per household will be the use cases for validating the developed big data analytics platform. These use cases also provide the basis for designing, planning and implementing schemes for improving energy related services for sake of achieving higher efficiency in both production and usage that contributes to cause of green environment in terms of less C02 emissions. The insight generated from these use cases can also help in educating the consumer about benefits of energy conservation and spread the awareness about behavioural changes that can benefit society as well as individuals themselves.

This master thesis is also supported by VTT, Technical Research Centre of Finland as part of their Green Campus initiative that focuses on use of ICT based solutions for innovative energy management and control systems capable to optimize the consumption without compromising the indoor environment. VTT is also a supporter and partner of CIVIS project. VTT has installed specialized smart devices in selected test sites that are the buildings owned by Aalto University. VTT has contributed to this thesis by providing the data generated by these smart devices. VTT has also helped in scoping for the use cases for energy efficiency by the experience and the knowledge they have from the related projects and research.

In a nutshell, this thesis focuses on providing a solution for collecting, storing, analysing and visualizing data generated by smart energy device for generating insights about energy consumption patterns and discovering the performance of different building units in terms of energy efficiency. This thesis also provides the models for knowledge discovery that can be used to improve energy efficiency at both producers and consumers ends. The big data analytics platform developed as part of this thesis is not limited to be used only for energy efficiency. It has the capability of handling other big data uses cases as well but we shall discuss its use for energy pattern detection and usage efficiency only in scope of this thesis report.

1.1 Problem statement

Energy conservation is required to reduce C02 emissions from energy production and usage. To achieve this goal we need to understand and improve the energy efficiency on both producer and consumer end. ICT enabled smart energy grids and devices are being rolled out globally to measure energy consumption and improve on energy efficiency. These smart devices produce high volumes of data that may or may not be predicted and planned at time of setting up the infrastructure. The data generated by different devices comes in different formats. For knowledge discovery from this data it is required to collect, store analyse the data and then visualize the generated insights so the information can be understood efficiently. The challenge gets even tougher when data needs to be collected and analysed in real time. Then with the time, volume of data and scope of analysis is expected to increase. So to cater for all this a highly scalable and flexible data analysis platform is required that can automate the whole process. This platform needs to be very cost effective for global adaptation.

In scope of this research we provide a model for big data analytics platform that can provide the solution for these requirements. We also implement the proposed model and test it with real life energy smart devices data and use cases. The proposed solution is based on open source components that can be deployed on general purpose commercially available, hence it is very cost effective. The proposed platform can be scaled according to data volumes and additional functional components can be integrated as per the scope of analysis.

1.2 Helpful hints

Read the information from the university master's thesis pages before starting the thesis. You should also go through the thesis grading instructions together with your instructor and/or supervisor in the beginning of your work.

1.3 Structure of the Thesis

You should use transition in your text, meaning that you should help the reader follow the thesis outline. Here, you tell what will be in each chapter of your thesis.

Chapter 2

Background

The problem must have some background, otherwise it is not interesting. You can explain the background here. Probably you should change the title to something that describes more the content of this chapter. Background consists of information that help other masters of the same degree program to understand the rest of the thesis.

Transitions mentioned in Section 1.3 are used also in the chapters and sections. For example, next in this chapter we tell how to use English language, how to find and refer to sources, and enlight different ways to include graphics in the thesis.

2.1 Smart grids

Energy industry across the globe is facing numerous challenges. There is a huge pressure from regulatory authorities and environmental organizations to reduce carbon foot print, expand their renewable energy portfolios, and take energy conservation measures. The demand response (DR)[1] and its impacts on consumer behaviour requires rapid adaptations in energy service providers business models. According to United States Federal Energy Regulatory Commission (FERC) , “Demand response can provide competitive pressure to reduce wholesale power prices; increases awareness of energy usage; provides for more efficient operation of markets; mitigates market power; enhances reliability; and in combination with certain new technologies, can support the use of renewable energy resources, distributed generation, and

[1] Demand Response(DR); Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized.

advanced metering. Thus, enabling demand-side resources, as well as supply-side resources, improves the economic operation of electric power markets by aligning prices more closely with the value customers place on electric power"[6]. Traditionally, power system participants have been strictly producers or consumers of electricity. The demand response and reliability issue with conventional electric power distribution models on consumer side are causing a major trend in motivating consumers to produce electricity at domestic level mostly using the renewable energy production methods. " Prosumer" is an emerging term used for an economically motivated entity that: [12]

- Consumes, produces, and stores power,
- Operates or owns a power grid small or large, and hence transports electricity, and
- Optimizes the economic decisions regarding its

The current energy grids support unidirectional distribution models and are centralized in nature. They are very limited to handle the prosumer needs. Line loses and hierarchical topology makes them less reliable. They usually become bottle neck when rapid adaptations are required for demand response. Farhangi, 2010 define smarts grids as "The next-generation electricity grid, expected to address the major shortcomings of the existing grid. In essence, the smart grid needs to provide the utility companies with full visibility and pervasive control over their assets and services. The smart grid is required to be self-healing and resilient to system anomalies. And last but not least, the smart grid needs to empower its stakeholders to define and realize new ways of engaging with each other and performing energy transactions across the system" [9].

In our research, we used data collected from smart metering devices as part of a pilot smart grid project. The data was used to generate analysis that recommends improvement for both demand and supply side to achieve energy efficiency as well as provide understanding to enable correct decision to adapt for demand response.

2.2 CIVIS project

CIVIS is the abbreviated name for "Cities as drivers of social change" project under European Union 7th framework. It is a part of the programme for optimising energy systems in smart cities. CIVIS project is a collaborative effort of 10 European universities) [2]. It aims to embed the social aspect into the advancements of energy technology. To unleash the full potential of this vision, smart grids need to be coupled with broader social and cultural considerations and understood as complex socio-techno-economic systems with multiple decision making layers that are in effect at the physical, cyber, social, and policy [8].

ICT acts as one of the main enabler of smart grids, distributed and bidirectional information flow models. On the other hand ICT also provides a lot of new mediums for social aggregation e.g. internet based social media. CIVIS projects tends to connect these two different dimensions with innovative ICT solutions. An integrated approach to energy efficiency is the basic manifesto of CIVIS project. [8]

Understanding energy usage patterns and benchmarking energy efficiency performance of small units within cities are some preliminary items in list of CIVIS objectives. Within scope of our research we analyze energy data to understand the consumption patterns and try to evaluate various factors that can effect directly or indirectly on the usage patterns. We also try to classify the building on basis of energy efficiency and try to test the sensitivity of energy efficiency with respect to factors that can cause shift in usage patterns. For the CIVIS project aim of social aspect integration, we also present an ICT application framework that can be used to collect and analyse social media data. However the analysis of that data is not within the scope of this research.

2.3 Green campus initiative

Green campus initiative is a project by VTT “Technical Research Centre of Finland” . It is part of EcoCampus 2030 program. EcoCampus is an attempt to contribute to increased energy efficiency in districts and buildings by innovative management and control systems capable to optimize the local consumption without compromising the indoor environment, occupant comfort and building performance, and by introducing new ICT enabled business models [18]. The vision of the program is to realize a net zero energy model for a world class research, development and educational facility. Program focuses on co-designing this model with user by educating them and then

[2] 1. Associazione Trento RISE, Italy 2. Aalto university, Finland 3. Imperial College London, UK 4. ENEL Foundation, Italy 5. Instituto Superior TÃ©cnico, Portugal 6.Karlsruhe Institute of Technology, Germany 7.Kungliga Tekniska Hogskolan, Sweden 8.SANTER REPLY SpA Italy 9.Nederlandse Organisatie voor toegepast Natuurwetenschappelijkonderzoek, Netherlands 10. Delft University of Technology,Netherlands

collecting feedbacks for improvement. The main aim is to gain energy efficiency by building infrastructure in the building units that can make them self sustain for future requirements. The aim is build to build a performance based ecosystem that can help both consumers and producers to adapt with demand response.

Green campus initiative is a pilot project for EcoCampus program in which VTT has installed smart devices inside Aalto University, Finland campus building in cities of Espoo and Helsinki. These specialized devices contained smart metering for energy consumption and indoor environment monitoring sensors. The data used for analysis in our research was collected from 100 buildings as test sites. The data includes hourly consumption of electricity and electricity used for heating. For one of the test sites VTT provided us the data with the details up to use of respective electric devices used in that site. This was achieved using smart NIALM [3][15] meters that can distinguish between different electric devices used on basis of their signal thumb print.

Apart from providing the data, VTT green campus researchers have also helped us in formulating the use cases for this thesis research.

2.4 Big data analytics

Big data analytics is application of advance data analytics techniques on large volumes of data. Advance analytics is a generalized term used for data analysis techniques like statistical analysis, data mining, machine learning, natural language processing, text mining and data visualization etc [26]. Although volume of the data is a widely used factor for qualification of a data set as big data but when it come to big data analytics there few other important attributes i.e. variety, velocity, valuation and veracity. The concept of 3Vs (volume, variety and velocity) of data was first given by an analyst, Doug Laney from Gartner in a 2001 MetaGroup research publication, "3D data management: Controlling data volume, variety and velocity"[20]. Gartner used this concept to formulate a data magnitude index that can support decision making for selection of the solutions for tackling big data challenge on use case base. This concept is shown in figure 2.1 below

Number 0 to 3 represents the scale of data that you perceive on each dimension. Adding them together for a big data case can provide the data magnitude index. This method provides some basis for quantifying the data

[3] NIALM stands for non-intrusive appliance load monitoring, is a process for analysing changes in the voltage and current going into a house and deducing what appliances are used in the house as well as their individual energy consumption

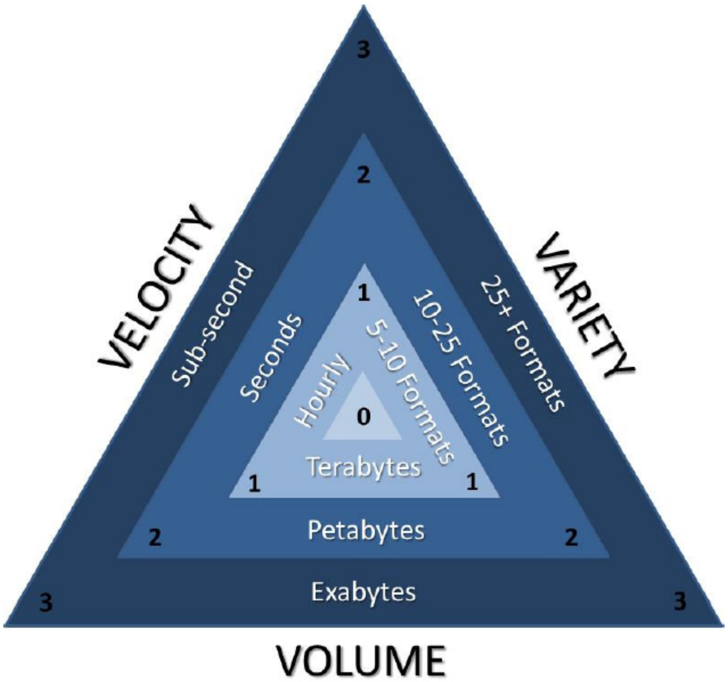


Figure 2.1: Gartner 3Vs of data and data magnitude index [20].

for big data qualification, However it is not providing a definitive model as it allows presumptions to scale the data. Valuation and veracity are two other factors that are being used widely along with Gartner's 3V. Valuation supports the decision making by considering the value of outcomes against the efforts required to collect, manage, process and analyse huge amounts of data. While veracity refers to ambiguity in the data that can cause complexity. There is no standard definition of big data but most of the attempts to define big data can be associated with these five factors that we have discussed.

As a matter of fact, we are not attempting to provide a definition of big data as part of this thesis or stating any criteria for qualification of a data set as big data. Instead we shall be proposing an advance analytics model that should be capable enough to handle big data as well other smaller data sets on need basis. The modular architecture of the model platform can be tweaked to handle volume, variety, velocity, and veracity on need basis while trying to maximize the valuation for the use case. In following subsections we shall discuss the some of the relevant technological advancements that enables to handle the mentioned challenges of big data analytics. These concepts, tools and techniques are also used in developing the data analytics platform and performing the analysis for our thesis research.

2.4.1 Parallel batch processing with MapReduce and Hadoop

It is hard to predict the size of data and computing power required to pocess it when dealing with big data. Scaling up [4]is an option that is always bound by some maximum capacity limits. Also specialized hardware to scale up for higher capacity usually gets very expensive. So the viable option is to scale out [5] using required number of smaller machines with relatively low computing resources in parallel. We need a system that can handle large scale parallelization. From programming point of view managing parallel running processes on different machines while ensuring low failure rate is a tough job. So the system should provide programmers an abstraction from lower level system details to enable rapid and fault tolerant development for big data processing. MapReduce is a parallel batch processing framework developed at Google for the purpose of web indexing. The concept of MapReduce was published by Jeffrey Dean and Sanjay Ghemawat in 2008 within their research paper "MapReduce: simplified data processing on large clusters" [7]. This paper describes MapReduce as a "programming model provides a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required intermachine communication".

Hadoop is the open source implementation of MapReduce developed by Doug Cutting and Mike Cafarella. It was initially created in 2005 to support an open source search engine but then adapted to the published MapReduced framework [7]. It was released by Apache foundation. Apache foundation has also built many supporting tool around Hadoop framework to support end to end big data analytics ecosystem e.g. Apache flume for data collection, Hadoop File system (HDFS) for storing, Apache Pig and Hive for processing,

[4] When the need for computing power increases, a single powerful computer is added with more CPU cores, more memory, and more hard disks and used in parallel.

[5] When the need for computing power increases, the tasks are divided between a large number of less powerful machines with (relatively) slow CPUs, moderate memory amounts, moderate hard disk counts.

Apache Mahout for machine learning etc. We have used some of these tools within scope of thesis research.

MapReduce and Hadoop are batch processing frameworks that empower processing of large volumes of data using commercial grade low cost computing infrastructure. So it supports volume and valuation directly. Variety can also be supported with support of all format files into associated files system e.g. HDFS. Veracity is subjected to supported tools like data collection or data mining tools. Support for such tools is available in Apache hadoop e.g. Flume, Mahout etc. Velocity however is the only feature that a batch processing framework like MapReduce and Hadoop cannot handle. The next subsection answers the question of velocity.

2.4.2 Real time big data processing

Real time data processing is generally associated with live streams of data. Real time data can be processed and analyzed on arrival or it can be buffered for small intervals to provide near to real time analysis. However in many modern data applications instantaneous data need to be analysed in context to large volumes of historic data. To apply advance analytics models like machine learning active feedback loops are also necessary. Even for stored (non live data) big data, applications require data processing system to answer queries very fast. To fulfil these industry driven requirements technology is in rapid advance mode. In last twelve to eighteen months we have seen softwares like YARN (hadoop 2.0), storm, spark, shark , cloudera impala etc with near to real time processing capabilities. On top of it tools like Mlbase and cloudera oryx have started to enable real time advance analytics. Most of these system, frameworks and tools are being developed as the evolution path for MapReduce and Hadoop. All of them have their own purpose, strengths , and limitations. They are mostly used in combinations based on use cases. We shall not be discussing or comparing these systems and tool. Instead, in this article we shall be briefly discussing the two prevailing architectural constructs that can enable real or near to real time big data processing.

2.4.2.1 Lambda architecture

Lambda architecture presents a hybrid model by using fast stream processing together with relatively slow parallel batch processing. It was developed by Nathan Marz on the basis of knowledge and experience he gained from his work with large data sets at Twitter Inc. His approach decompose data processing system into three layers i.e. a batch layer, a serving layer and a speed layer. The stream of data is dispatched to both the batch and speed layers. Batch layer manages the data set and pre-compute the batch views. Serving layer indexes the batch views so the queries can be served with low latency as compared to traversing through complete data set. Speed layer deals with the recent data thus compensates for the change of data sets during updates of serving layer. An answer to the query is the merged view batch view and the real time view.[24][3]

Figure 2.2 below show the Lambda architecture. Lambad architecture can be implemented using combination of systems and tools e.g.Apache Hadoop along with Apache Storm.

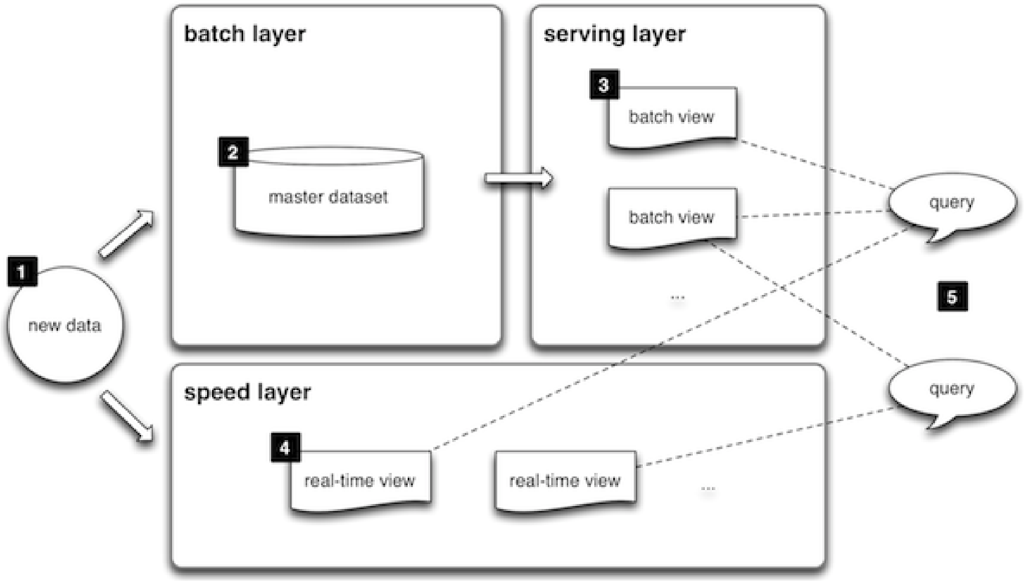


Figure 2.2: Lambda Architecture [3].

2.4.2.2 Massively parallel processing - MPP databases and query engines

MPP based architectures use multiple independent computing resources like severs, processors and storages to execute processing jobs in parallel. Most of the MPP based database approaches implements shared nothing (SN) architecture i.e. a distributed computing architecture in which each node is independent and self sufficient and there is no point of contention across the system [31]. The SN concept for databases was first presented by Michael Stonebraker at University of California Berkeley in 1986 [27].. The SN databases have been very popular in commercial application primarily because of the high scalability offered by this architecture. Teradata warehousing solutions has been using SN database architectures extensively. Greenplum is an example for open source SN database.

Despite high scalability and other positive aspects, SN databases needs a lot of manual work in terms of partitioning the data, tuning the data and load balancing etc. So usually maintenance such database systems is expensive. MapReduce and Apache Hadoop ecosystem provides high level of automation along with scalability, flexibility and fault tolerance. However parallel batch processing is not as fast SN based MPP databases. Merging both the models solves can solve all these issues. Cloudera Imapala is one of the example of a MPP based online query engine that runs natively on top of Hadoop [1]. It can provide MPP like query response time performance with processing power and flexibility of Hadoop. For our research we have used Cloudera Impala for handling near to real time velocity for big data processing.

2.5 Energy efficiency and eco-effeciency

In previous sections of this chapter, we have highlighted the importance of energy conservation. We discussed the advancements in pervasive smart energy device and grids and their role in improving energy efficiency. We have also discussed the need for collecting and processing large volumes of data from smart energy devices and the available solutions. In this section we shall explain the main motivation and the theoretical concept behind data analysis part of our research.

Unprecedented challenges arising from increasing dependency on conventional energy are part of a global phenomenon. Improving energy efficiency is an important mean to tackle these challenges. Like other economies, European Union is also putting a lot of focus on energy efficiency to ensure energy supply security by reducing primary energy consumption and decreasing energy imports. It helps to reduce greenhouse gas emissions in a cost- effective way and thereby to mitigate climate change [2]. Member states agreed to reduce 20% of the EU's primary energy consumption by 2020 in European Union of council March, 2007. EU's Energy Efficiency Directive 2012 [2] defines energy efficiency as the ratio of output of performance, service, goods or energy, to input of energy. This definition was first discussed in 2006 in European commission action plan for energy efficiency. This generic definition covers all major aspects of the energy efficiency i.e. production, distribution, consumption and the value created in comparison to the resources consumed during the whole process. However, To develop a methodology for measuring energy efficiency and to evaluate the saving, project "Measuring and potentials of energy efficiency (EPO)" was started in 2008[4]. As part of this project VTT published a report "Measuring energy efficiency Indicators and potentials in buildings, communities and energy systems"[11]. This report presents the model for calculating energy efficiency and its correlation with environmental factors. VTT's research presented in this report considers energy efficiency as a subset of larger eco-efficiency. The ecological factors that can affect energy efficiency are e.g. Temperature, CO2, NOx,SO2 etc. The ecological efficiency itself is a way of measuring sustainable development. VTT summarizes the whole ecosystem in Figure 2.3 below

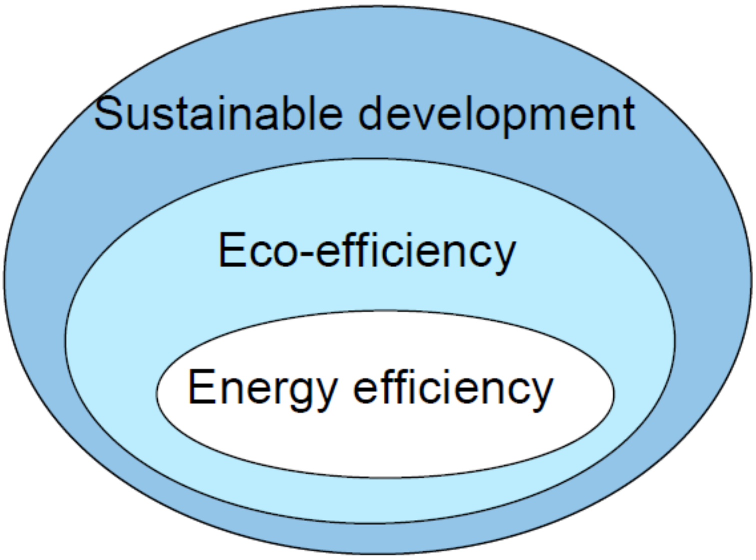


Figure 2.3: Energy efficiency, eco-efficiency and sustainability[11].

The concept of eco-efficiency provides the basis for data analysis in our research. We have applied basic and advanced analytics techniques on data sets collected from building units that are part of VTT's green campus initiative pilot project with consideration of eco-efficiency model presented in VTT's report. We calculated energy efficiency of the buildings on basis of formula deduced in Chapter 5 (equation 5.1 and 5.2) of the VTT's report [11].

$$Energy\ efficiency\ of\ a\ building = \frac{Energy\ consumed}{Built\ area} \tag{2.1}$$

In case of a specific energy consumption (SEC) [11] equation 2.1 can be written as

$$SEC = \frac{Q}{A} \tag{2.2}$$

Where Q denotes the consumption for a single energy type for example electricity and A is the built area in meter square.. In subsequent sections we shall be referring to these equations when we try to identify the usage patterns on building level, discuss the relevance of energy efficiency with these patterns and then discuss a model for classifying buildings on energy efficiency .

2.6 Daily consumption patterns, base load and user load

Daily consumption pattern of a building unit corresponds to the respective usage of the building. Understanding daily usage patterns can help in identifying the optimization point for improving the energy efficiency of that building unit. Base load of a building is one important metric that can be detected through observing the daily consumption. Base load is the consumption that takes place regardless of the actual use of the building and of the user's energy consumption[11]. It is the permanent minimum load that a power supply system is required to deliver. The base load is usually caused by the continuous consumption for building maintenance like air conditioning, ventilation, or night time lighting. Sometimes base load also include some energy consumption by functional components inside building like computer servers, lab equipments, and refrigerators etc. However VTT differentiate this load from user energy load that is characterized by the direct involvement of the users of a building. For example an office building that has peak load during day time because user are using various additional appliances like personal computers, coffee makers, lights etc compared to base load that is generated during night time when office building is not in use. Figure 2.4 illustrates the concept of base load and user load.

Energy efficiency of base consumption and energy efficiency of use shown in figure 2.4 can be calculated using equation 2.1 or 2.2. This provides a weighted metric that can be benchmarked and compared. It can help to narrow down scope of research by referring to problematic buildings and their issues.

2.7 Energy consumption seasonal patterns

Energy consumption has high dependency on seasonal factors like weather. The energy consumption trends vary with outside temperature. Among other things electricity or other energy types required for the air conditioning in

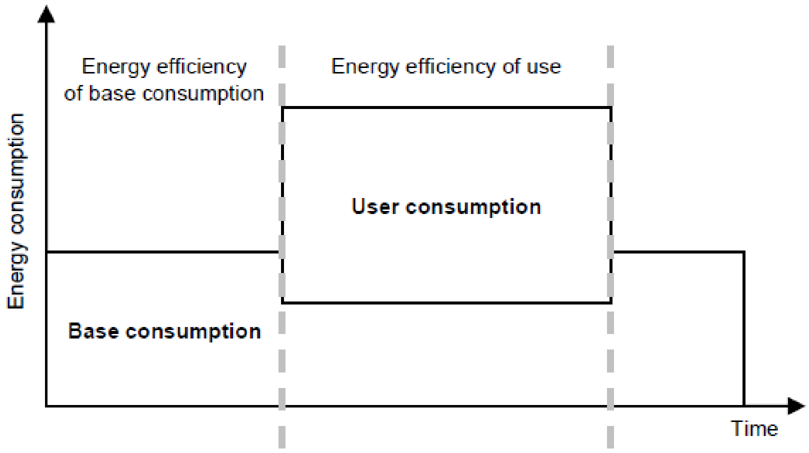


Figure 2.4: Base load , user load and energy efficiency [11].

the buildings is major variable factor dictating the trends. Due to regional weather differences the seasonal energy consumption patterns are also different for different regions e.g. in cold regions of the world energy consumption surges in winters while in warmer regions energy consumption increase is expected in summers because of the air conditioning requirements. Energy service providers usually conduct demand planning with consideration of seasonal trends.Considering seasonal trends is also very important while optimising for gaining energy efficiency.

In scope of our research we have also analysed the seasonal trends. It was not hard for us to perceive the trends while knowing the weather trend for localities of our test building. However, the interesting use case in our research was to check the sensitivity of other consumption patterns and analysis results against the seasonal trend. This will be more explained in the later part of document where we shall discuss the results of our analysis.

Previously, there have been many studies for both daily and seasonal trends in energy consumption. Due to regional differences in trends, many of these studies focused on consumption patterns within a country. Geoffrey K.F. Tso et. all, 2003[30] and Yigzaw G. Yohanis et. all 2007[32] study the energy consumption pattern in Hong Kong and United Kingdom respectively. Buildings units e.g. residential houses apartments and commercial offices etc were considered as basic unit of analysis. Yigzaw G. Yohanis et. all methodology resembles most to our approach as they considered ecological factor along with energy efficiency calculated in similar way as equation 2.1 and 2.2. As discussed before, the main purpose of VTT's green campus initiative under EcoCampus 2030 plan is to develop a highly efficient model ecosystem for energy production, distribution and consumption that can be expanded further to any scale. Alligned to this goal, we have attempted to provide a data analysis model that is not specific to certain geographic locations. However detailed study is required for adapting such generic models to region specific requirements. In our research we have also attempted to classify the buildings on basis energy efficiency that is explained in next section.

2.8 Classification of buildings based on energy efficiency

Earlier we mentioned that quantifiable energy efficiency through equation 2.1 and 2.2 can be used as a metric for benchmarking and comparison. For energy service providers, governmental energy regulatory agencies or research institute like VTT, it is very important to identify the problematic consumption units in larger number of highly optimized or average performing consumption units.classification of these units into similarly performing groups can help them to narrow down the focus to only problematic units. Sometimes it can also help in understanding the good practices applied by certain consumptions unit that has improved their energy efficiency performance.

Classification for fault detection analysis of a building energy consumption has been used previously as well. Xiaoli Li et. all, 2010 used classification along with outlier detection mechnism to identify the energy inefficient building [21].They provide a step wise approach to extract the features (types of energy, trends etc) from the data collected as a time series. Then detect identify the daily usage patterns using auto regression technique and pass the results to benchmark against any outlying data point that can refer to faulty behaviour. Imran Khan et all. 2013, proposes different clustering techniques to group building with similar level of energy efficiency together[19]. In our research we used a hybrid method using feature extraction and trend detection techniques like [21] and then applied a clustering technique proposed in [19]. The clustering technique that used is called K-means clustering. It is explained in the next subsection of this article.

2.8.1 K-means clustering

K-means is an algorithm for cluster analysis. In context to machine learning cluster analysis or clustering is an unsupervised task of grouping a set of objects in a way that objects in same group are similar to each other more than the objects in other group. K-means algorithm clusters the set of objects i.e. energy efficiency values in our case into predefined number of classes. We shall term these values as data points. K represents the number of cluster and groups that we can set in start of the process. K-means means algorithm was first proposed by Stuart Llyod in 1957[22] but the k-means term was first used by James Mcqueen in 1967[23]. There have many adaptations and optimizations in Lloyd's basic algorithm. K-means algorithm today has many variants like Fuzzy C-means clustering, k-medoids and spherical means etc. Even for original Lloyd's algorithm there has been some modification in methodology. Two very commonly used methods are Forgy method [10] and Hartigan-Wong method[16]. In our approach we are using Hartigon-Wong method. We shall also use some references from Forgy method when explaining the K-means algorithm.

K-means groups the data points in cluster with a logical centre point. The aim of the K-means algorithm is to divide data points in certain dimensions into K clusters so that the within-cluster sum of squares is minimized [16]. Lets assume if we want have K cluster for data points D = $\{x_1, x_2, \ldots, x_n\}$ in d dimensions then

$$x_i \in R^d$$

K-means algorithm uses following steps to cluster data into groups[25].

1. Initialize the centroids randomly for each K i.e. for each group.
2. Data points are assigned to closest centroid.
3. Move the centroids to the mean of the data points assigned to that cetroid in step 2.
4. Repeat 2 and 3 till convergence Convergence means that values stop changing in further iterations.

Mathematically randomly initialized centroid are

$$\mu_1, \mu_2, \ldots, \mu_k \in R^n$$

If c^i is the distance of centroid to assigned data point then Step 2 and 3 with recursive distance minimization and mean adjustment can be explained as

For every i, set

$$c^i := \arg\min_j ||x^i - \mu_j||^2 \tag{2.3}$$

For every j, set

$$\mu_j := \frac{\sum\limits_{i=1}^{n} 1\{c^i = j\}x^i}{\sum\limits_{i=1}^{n} 1\{c^i = j\}} \tag{2.4}$$

The input to k-means is a set of feature vectors along with the number of clusters required. In our case we shall be have two feature hence two dimensional matrix of energy efficiency values for electricity and electricity used for heating. Before inserting data to k-means it is required to set the similar scale for features as well set the standard variance to avoid errors in the results. We required to classify pilot site buildings into four groups with High efficiency, moderate efficiency, low efficiency and poor efficiency classes. So we have set K value as 4.

2.8.2 Forecasting the energy consumption

Estimating equipment specific energy consumption has been a key focus area for energy service providers. It can help in demand planning, load forecasting, and understanding end user behaviour. Based on this such estimating energy service providers can design better service offerings for their consumers. Unit energy consumption (UEC) is a term generally used for estimating equipment specific energy consumption. It is the average annual amount of energy consumed by a user device. Conditional demand analysis (CDA) model has been one the most commonly used method for UEC estimations. K. H. Tiedemann, 2007 explain CDA as a multivariate regression technique which combines utility billing data with weather information and customer survey data to produce robust end-use energy consumption estimates.

Chapter 3

Environment

A problem instance is rarely totally independent of its environment. Most often you need to describe the environment you work in, what limits there are and so on. This is a good place to do that. First we tell you about the LaTeX working environments and then is an example from an thesis written some years ago.

3.1 LaTeX working environments

To create LaTeX documents you need two things: a LaTeX environment for compiling your documents and a text editor for writing them.

3.1.1 Environment

Fortunately LaTeX can nowadays be found for any (modern) computer environment, be it Linux, Windows, or Macintosh. For Linuxes (and other Unix clones) and Macs, I'd recommend *TeX Live* [28], which is the current default LaTeX distribution for many Linux flavors such as Fedora, Debian, Ubuntu, and Gentoo. TeX Live is the replacement for the older *teTeX*, which is no longer developed.

TeX Live works also for Windows machines (at least according to their web site); however, I have used *MiKTeX* [5] and can recommend it for Windows. MiKTeX has a nice package manager and automatically fetches missing packages for you.

3.1.2 Editor

You can write LaTeX documents with any text editor you like, but having syntax coloring options and such really helps a lot. My personal favourite for editing LaTeX is the *TeXlipse* [17] plugin for the Eclipse IDE [29]. Eclipse is an open-source integrated development environment (IDE) initially created for writing Java code, but it currently has support for editing languages such as C, C++, JavaScript, XML, HTML, and many more. The TeXlipse plugin allows you to edit and compile LaTeX documents directly in Eclipse, and compilation errors and warnings are shown in the Eclipse *Problems* dialog so that you can locate and fix the issues easily. The plugin also supports reference traversal so that you can locate the source line where a label or a citation is defined.

Eclipse is an entire development environment, so it may feel a bit heavy-weight for editing a document. If you are looking for a more light-weight option, check out TeXworks. TeXworks is a LaTeX editor that is packaged with the newer MiKTeX distributions, and it can be acquired from `http://www.tug.org/texworks/`.

And if you are attached to your *emacs* or *vim* editor, you can of course edit your LaTeX documents with them. Emacs at least has syntax coloring and you can compile your document with a key binding, so this may be a good option if you prefer working with the standard Linux text editors.

3.2 Graphics

When you use `pdflatex` to render your thesis, you can include PDF images directly, as shown by Figure 3.1 below.

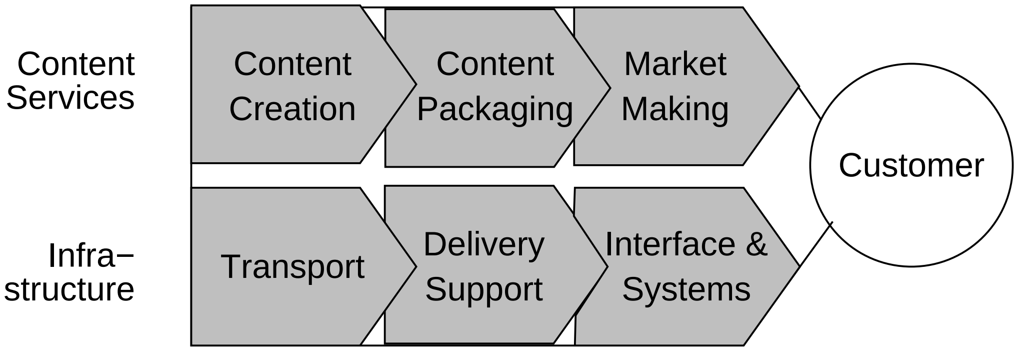


Figure 3.1: The INDICA two-layered value chain model.

You can also include JPEG or PNG files, as shown by Figure 3.2.

You can create PDF files out of practically anything. In Windows, you can download PrimoPDF or CutePDF (or some such) and install a printing

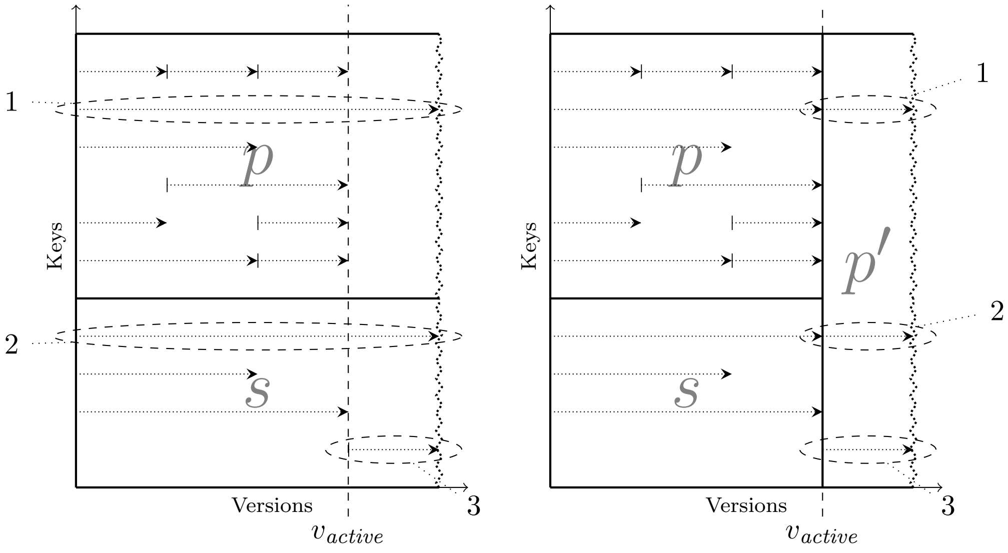
Figure 3.2: Eeyore, or Ihaa, a very sad donkey.

driver so that you can print directly to PDF files from any application. There are also tools that allow you to upload documents in common file formats and convert them to the PDF format. If you have PS or EPS files, you can use the tools `ps2pdf` or `epspdf` to convert your PS and EPS files to PDF.

Furthermore, most newer editor programs allow you to save directly to the PDF format. For vector editing, you could try Inkscape, which is a new open source WYSIWYG vector editor that allows you to save directly to PDF. For graphs, either export/print your graphs from OpenOffice Calc/Microsoft Excel to PDF format, and then add them; or use `gnuplot`, which can create PDF files directly (at least the new versions can). The terminal type is *pdf*, so the first line of your plot file should be something like `set term pdf ...`.

To get the most professional-looking graphics, you can encode them using the TikZ package (TikZ is a frontend for the PGF graphics formatting system). You can create practically any kind of technical images with TikZ, but it has a rather steep learning curve. Locate the manual (`pgfmanual.pdf`) from your LaTeX distribution and check it out. An example of TikZ-generated graphics is shown in Figure 3.3.

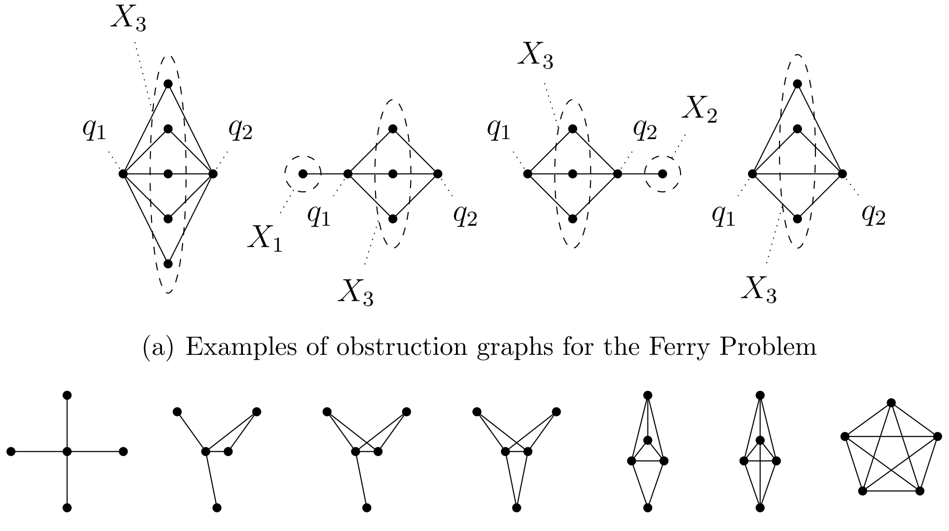
Another example of graphics created with TikZ is shown in Figure 3.4. These show how graphs can be drawn and labeled. You can consult the example images and the PGF manual for more examples of what kinds figures you can draw with TikZ.



(a) Before

(b) After

Figure 3.3: Example of a multiversion database page merge. This figure has been taken from the PhD thesis of Haapasalo [14].



(a) Examples of obstruction graphs for the Ferry Problem

(b) Examples of star graphs

Figure 3.4: Examples of graphs draw with TikZ. These figures have been taken from a course report for the graph theory course [13].

Chapter 4

Methods

You have now stated your problem, and you are ready to do something about it! *How* are you going to do that? What methods do you use? You also need to review existing literature to justify your choices, meaning that why you have chosen the method to be applied in your work.

If you have not yet done any (real) metholodogical courses (but chosen introduction courses of different areas that are listed in the methodological courses list), now is the time to do so or at least check through material of suitable methodological courses. Good methodologial courses that consentrates especially to methods are presented in Table 4.1. Remember to explain the content of the tables (as with figures). In the table, the last column gives the research area where the methods are often used. Here we used table to give an example of tables. Abbreviations and Acronyms is also a long table. The difference is that longtables can continue to next page.

Code	**Name**	**Methods**	**Area**
T-110.6130	Systems Engineering for Data Communications Software	Computer simulations, mathematical modeling, experimental research, data analysis, and network service business research methods, (agile method)	T-110
Mat-2.3170 Simulation (here is an example of multicolumn for tables)		Details of how to build simulations	T-110
S-38.3184	Network Traffic Measurements and Analysis	How to measure and analyse network traffic	T-110

Table 4.1: Research methodology courses

Chapter 5

Implementation

You have now explained how you are going to tackle your problem. Go do that now! Come back when the problem is solved!

Now, how did you solve the problem? Explain how you implemented your solution, be it a software component, a custom-made FPGA, a fried jelly bean, or whatever. Describe the problems you encountered with your implementation work.

Chapter 6

Evaluation

You have done your work, but that's[1] not enough.

You also need to evaluate how well your implementation works. The nature of the evaluation depends on your problem, your method, and your implementation that are all described in the thesis before this chapter. If you have created a program for exact-text matching, then you measure how long it takes for your implementation to search for different patterns, and compare it against the implementation that was used before. If you have designed a process for managing software projects, you perhaps interview people working with a waterfall-style management process, have them adapt your management process, and interview them again after they have worked with your process for some time. See what's changed.

The important thing is that you can evaluate your success somehow. Remember that you do not have to succeed in making something spectacular; a total implementation failure may still give grounds for a very good master's thesis—if you can analyze what went wrong and what should have been done.

[1] By the way, do *not* use shorthands like this in your text! It is not professional! Always write out all the words: "that is".

Chapter 7

Discussion

At this point, you will have some insightful thoughts on your implementation and you may have ideas on what could be done in the future. This chapter is a good place to discuss your thesis as a whole and to show your professor that you have really understood some non-trivial aspects of the methods you used…

Chapter 8

Conclusions

Time to wrap it up! Write down the most important findings from your work. Like the introduction, this chapter is not very long. Two to four pages might be a good limit.

Bibliography

[1] Cloudera impala.

[2] Directive 2012/27/eu of the european parliament and of the council of 25 october 2012 on energy efficiency. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:315:0001:0056:EN:PDF`. Accessed: 2014-06-10.

[3] Lambda Architecture what is the lambda architecture. Accessed: 2014-06-09.

[4] Arundel, A., and Kemp, R. Measuring eco-innovation. *United Nations University Working Paper Series*, 2009/017 (2009), 1–40.

[5] Christian Schenk. MiKTeX, 2010. `http://miktex.org/`. Accessed 25.2.2011.

[6] Commission, F. E. R., et al. Assessment of demand response and advanced metering.

[7] Dean, J., and Ghemawat, S. Mapreduce: simplified data processing on large clusters. *Communications of the ACM 51*, 1 (2008), 107–113.

[8] European Union 7th framework programme. Proposal part b - cities as drivers of social change civis project - ict-2013.6.4 optimising energy systems in smart citiese, 2013.

[9] Farhangi, H. The path of the smart grid. *Power and Energy Magazine, IEEE 8*, 1 (2010), 18–28.

[10] Forgy, E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics 21* (1965), 768–769.

[11] Forsström, J., Lahti, P., Pursiheimo, E., Rämä, M., Shemeikka, J., Sipilä, K., Tuominen, P., and Wahlgren, I. Measuring energy efficiency.

[12] Grijalva, S., and Tariq, M. U. Prosumer-based smart grid architecture enables a flat, sustainable electricity industry. In *Innovative Smart Grid Technologies (ISGT), 2011 IEEE PES* (2011), IEEE, pp. 1–6.

[13] Gï¿½ï¿½s, M., Haapasalo, T., and Moisio, L. Finding hard instances of the ferry problem. Course report for the course T-79.5203/S-72.2420 Graph theory, Aalto SCI, 2010.

[14] Haapasalo, T. *Accessing Multiversion Data in Database Transactions*. PhD thesis, Department of Computer Science and Engineering, Aalto University School of Science and Technology, Espoo, Finland, 2010. `http://lib.tkk.fi/Diss/2010/isbn9789526033600/`.

[15] Hart, G. W. Nonintrusive appliance load monitoring. *Proceedings of the IEEE 80*, 12 (1992), 1870–1891.

[16] Hartigan, J. A., and Wong, M. A. Algorithm as 136: A k-means clustering algorithm. *Applied statistics* (1979), 100–108.

[17] Hupponen, T., Karlsson, K., Laitinen, J., Ojala, O., Pirinen, A., Seuranen, E., Takkinen, L., von Loesch, B., and Vestbï¿½, T. A. TeXlipse plugin for Eclipse, 2010. `http://texlipse.sourceforge.net/`. Accessed 25.2.2011.

[18] Janne Peltonen. Presentation on vtt otaniemi greencampus summary, 2013.

[19] Khan, I., Capozzoli, A., Corgnati, S. P., and Cerquitelli, T. Fault detection analysis of building energy consumption using data mining techniques. *Energy Procedia 42* (2013), 557–566.

[20] Laney, D. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note 6* (2001).

[21] Li, X., Bowers, C. P., and Schnier, T. Classification of energy consumption in buildings with outlier detection. *Industrial Electronics, IEEE Transactions on 57*, 11 (2010), 3639–3644.

[22] Lloyd, S. Least squares quantization in pcm. *Information Theory, IEEE Transactions on 28*, 2 (1982), 129–137.

[23] MacQueen, J., et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (1967), vol. 1, California, USA, p. 14.

[24] Marz, N., and Warren, J. *Big Data: Principles and best practices of scalable realtime data systems.* O'Reilly Media, 2013.

[25] Ng, A. Cs229 lecture notes. *CS229 Lecture notes 1*, 1 (2000), 1–3.

[26] Russom, P., et al. Big data analytics. *TDWI Best Practices Report, Fourth Quarter* (2011).

[27] Stonebraker, M. The case for shared nothing. *IEEE Database Eng. Bull. 9*, 1 (1986), 4–9.

[28] TeX Users Group. TeX Live, 2010. `http://www.tug.org/texlive/`. Accessed 25.2.2011.

[29] The Eclipse Foundation. Eclipse, 2011. `http://eclipse.org/`. Accessed 25.2.2011.

[30] Tso, G. K., and Yau, K. K. A study of domestic energy usage patterns in hong kong. *Energy 28*, 15 (2003), 1671–1682.

[31] Wikipedia. Shared nothing architecture — wikipedia, the free encyclopedia, 2014. [Online; accessed 9-June-2014].

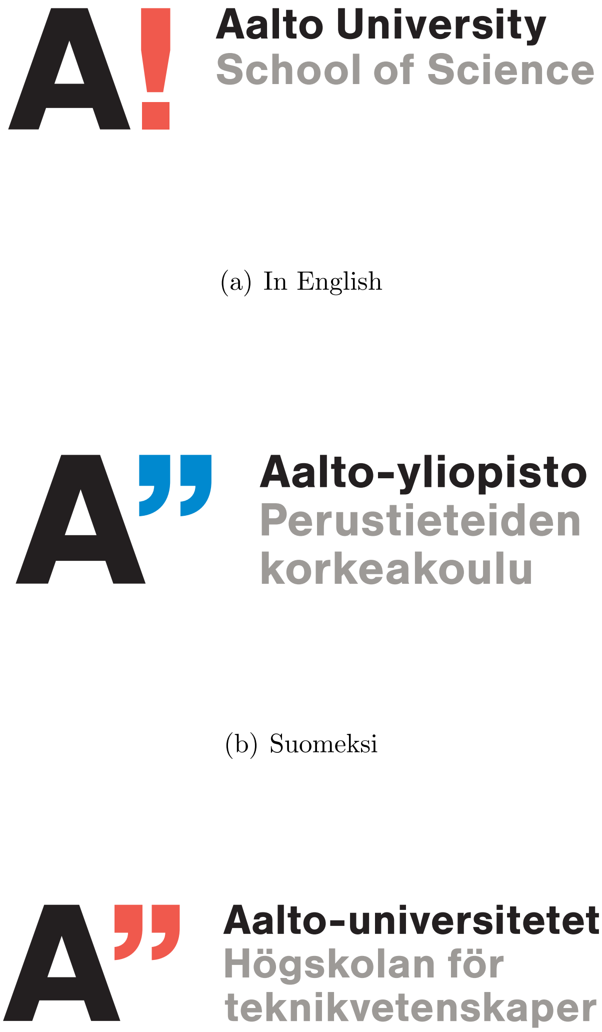
[32] Yohanis, Y. G., Mondol, J. D., Wright, A., and Norton, B. Real-life energy use in the uk: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings 40*, 6 (2008), 1053–1059.

Appendix A

First appendix

This is the first appendix. You could put some test images or verbose data in an appendix, if there is too much data to fit in the actual text nicely.

For now, the Aalto logo variants are shown in Figure A.1.



(a) In English

(b) Suomeksi

(c) På svenska

Figure A.1: Aalto logo variants