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

Hussnain Ahmed

Using Big Data Analytics for Measuring Energy Consumption Patterns:

An end to end data analytics platform in use for energy efficiency

Master's Thesis
Espoo, June, 2014

DRAFT! — May 31, 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
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: An end to end data analytics platform in use for energy efficiency

Date: June, 2014 **Pages:** 27

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].

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 Respose(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"[1]. 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: [4]

- 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" [3].

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 Finding and referring to sources

Never ever copy anything into your theses from somebody else's text (nor your own previously published text). Never. Not even for starting point to be rewritten later. The risk is that you forgot the copied text to your thesis

and end up to be accused of plagiarism. Plagiarism is a serious crime in studies and science and can ruin your career even its beginning. To repeate: never cut and paste text into your thesis!

2.2.1 Finding sources

All work is based on someone else's work. You should find the relevant sources of your field and choose the best of them. Also, you should refer to the original source where a fact has been mentioned first time. Remember source evaluation (criticism) with all sources you find.

Good starting points for finding references in computer science are:

- Nelli Portal (Aalto Library): `http://www.nelliportaali.fi`
- ACM Digital library: `http://portal.acm.org/`
- IEEExplore: `http://ieeexplore.ieee.org`
- ScienceDirect: `http://www.sciencedirect.com/`
- ... although Google Scholar (`http://scholar.google.com/`) will find links to most of the articles from the abovementioned sources, if you search from within the university network

Some of the publishers do not offer all the text of the articles freely, but the library has agreed on the rights to use the whole text. Thus, you should sometimes use computers in the domain of the university in order to get the full text. Sometimes the Nelli Portal can also help getting the whole article instead of just the abstract. The library has also brief instrucions how to find information [**?**].

Instead of normal Google, use Google Scholar (`http://scholar.google.fi/`). It finds academic publications whereas normal Google find too much commercial advertisements or otherwise biased information. Wikipedia articles should be referred to in the master thesis only very, very seldomly. You can use Wikipedia for understanding some basics and finding more sources, but often you cannot be sure if the article is correct and unbiased.

One important part of the sources that you have found is the reference list. This way you can find the original sources that all the other research of the field refer. Often you can also find more information with the name of the researchers that are often referred in the articles.

2.2.2 Referring to sources

The main point in referring to sources is to separate your own thinking and text from that of others. Facts of the research area can be given without

reference, but otherwise you should refer to sources. This means two things: marking the source in the text where it has been used, and listing the sources usually in the end of the thesis in a way that help the reader to find the original source.

There are several bibliography styles, meaning how to form the bibliography in the end of the thesis. Aalto's library has good instructions for many styles [**?**]. You should ask from your supervisor or instructors which style you should use. This thesis template uses the number style that is often used in software engineering. The other style also used in the CS field, e.g. usability, is the Harvard style where instead of numbers, the reference is marked into the text with author's name and publishing year. Other areas use also many other styles for making the lists and marking the references.

In addition to the list in the end of the thesis, you have to mark the source in the text where the source is used. There are three places for the reference: in a sentence before the period, in the end of a sentence after the period, or in the end of a paragraph. All of them have different meaning. The main point is that first you paraphrase the source using your own words and then mark the source. Next, we give short examples that are marked with *emphasised text*.

*Haapasalo [**?**] researched database algorithms that allows use of previous versions of the content stored in the database.* This kind of marking means that this paragraph (or until the next reference is given) is based on the source mentioned in the beginning. Giving the source you should use only the family name of the first author of the article, and not give any hints about what is the type of the article that is referred.

*B+-trees offers one way to index data that is stored in to a database. Multiversion B+-trees (MVBT) offer also a way to restore the data from previous versions of the database. Concurrent MVBT allows many simultaneous updates to the database that is was not possible with MVBT. [**?**]* When the marking is after the period, the reference is retrospective: all the paragraph (or after previous reference marking) is based on the source given in its end. If the content is very broad, you can start with saying *According to Haapasalo*, then continue referring the source with several separate sentences, and in the end put the marking of your source *that shows that CMVBT are the best. [**?**].*

If your paragraph has several sources, the above mentioned styles are not proper. The reader of your thesis cannot know which of your sources give which of the statements. In this case, it is better to use more finegraded refering where the reference markings that are embedded in the sentences. For example, *the multiversion B+-tree (MVBT) index of Becker et al. [**?**] allows database users to query old versions of the database, but the index*

*is not transactional. It's successor, the transactional MBVT (TMVBT), allows a single transaction running in its own thread or process to update the database concurrently with other transactions that only read the database [***?** *]. Further development, titled the concurrent MBVT (CMVBT), allows several transactions to perform updates to the database at the same time [***?** *].* Here, the references are marked before the period in the sentences where they are used.

Finally, direct quotes are allowed. However, often you should avoid them since they do not usually fit in to your text very well. Using direct quotes has two tricks: quotation marks and the source. *"Even though deletions in a multiversion index must not physically delete the history of the data items, queries and range scans can become more efficient, if the leaf pages of the index structure are merged to retain optimality." [***?** *]* Quotes are hard to make neatly since you should use only as much as needed without changing the text. Moreover, you often do not really understand what the author has mentioned with his wordings if you cannot write the same with your own words. Remember also that never cut and paste anything without marking the quotation marks right away, and in general, never cut and paste anything at all!

Sometimes getting the original source can be almost impossible. In an extremely desperate situation, you can refer with structure *mr X [. . .] according to ms Y [. . .] defined that*, if you find a source that refers to the original source. Note also that the reference marking is never used as sentence element (example of how **not** to do it: *[***?** *] describes an optimal algorithm for indexing multiversiond databases.*).

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* [**?**], 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* [**?**] 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* [**?**] plugin for the Eclipse IDE [**?**]. 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.

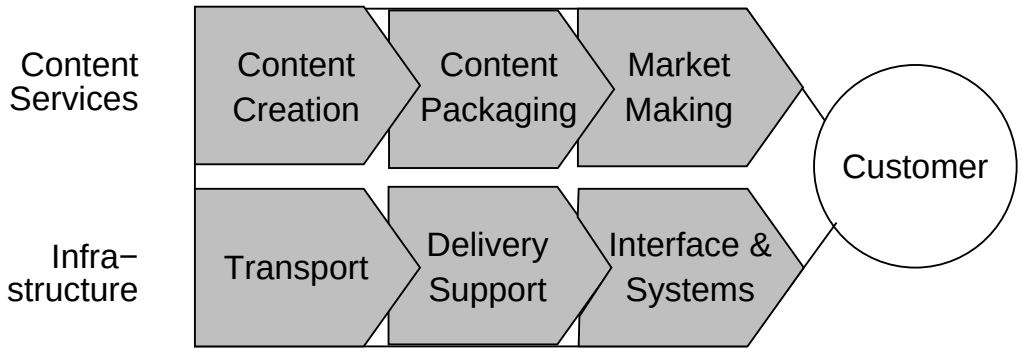


Figure 3.1: The INDICA two-layered value chain model.

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

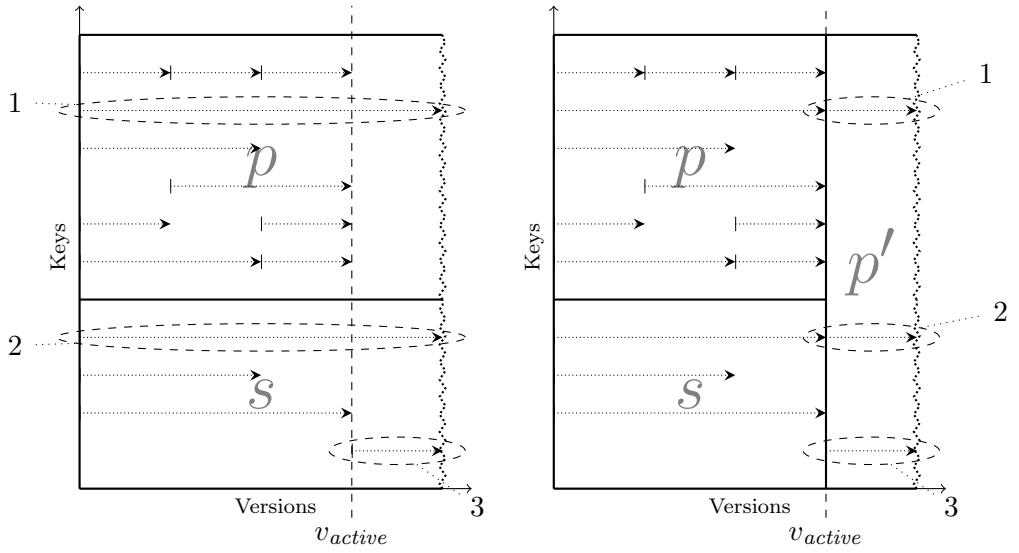
Figure 3.2: Eeyore, or Ihaa, a very sad donkey.

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 ...`.

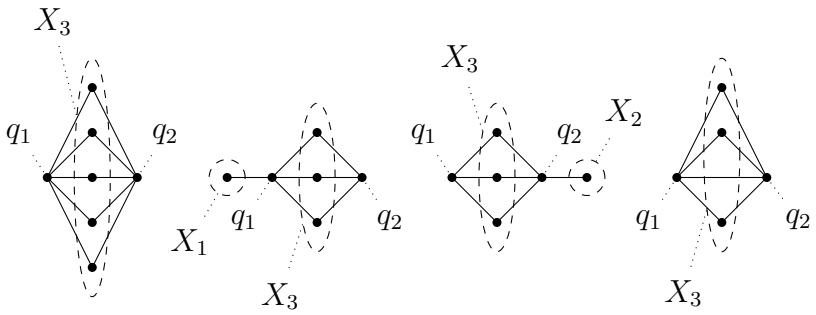
To get the most professional-looking graphics, you can encode them using the TikZ package (TikZ is a frontend for the PGF graphics formatting system). You can create practically any kind of technical images with TikZ, but it has a rather steep learning curve. Locate the manual (`pgfmanual.pdf`) from 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.

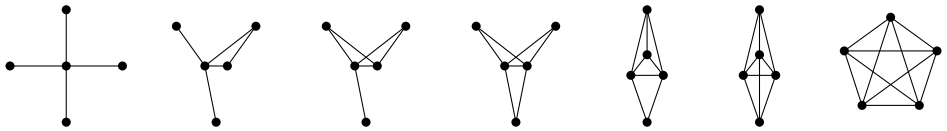


(a) Before (b) After

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



(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 [**?**].

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

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

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

[4] 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.

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