

DATA MANAGEMENT FOR LARGE SCALE POWER QUALITY SURVEYS

ML Peard, ST Elphick, VW Smith, VJ Gosbell, D Robinson

School of Electrical, Computer and Telecommunications Engineering
University of Wollongong, NSW Australia

Abstract

For large scale power quality surveys, the management of the large amount of data generated is a major issue. This paper presents solutions to three main areas of data management, viz. a data interchange format, database design and data processing. Consideration of these issues has come about as a result of the Long Term National Power Quality Survey currently being conducted by the University of Wollongong, and reference is made to that specific application for illustrative purposes.

1. INTRODUCTION

In 2000, the University of Wollongong commenced a pilot power quality survey of Australian electricity distributors in order to get an initial assessment of the state of Australian distribution networks with regard to power quality [1]. This survey was limited in its scope and duration with only eight sites being monitored for one week for each utility. While this initial survey gave a good snapshot of Australian distribution networks, it soon became apparent that a much longer and more extensive survey was required to properly assess networks affected by such a diversity of geographical, load type and seasonal conditions.

Consequently, since 2002, the University of Wollongong has been involved in a routine long term power quality survey known as the Long Term National Power Quality Survey (LTNPQS) [2]. The aim of the survey is to determine overall power quality disturbance levels throughout Australia. This information may be used as benchmarks to determine utility performance and may be of assistance to utilities in negotiating with regulators.

At present there are 9 utilities involved in the survey and this covers the bulk of Queensland, New South Wales, Victoria and Tasmania. Completion of the survey involves utilities submitting data to the University for analysis and reporting. The survey is now into its second year and has grown significantly since its first year both in terms of the number of sites submitted by the utilities as well as the duration of the survey periods from those sites. This year's latest report comprised of some 200 plus sites encompassing voltage levels from low voltage (230V) through to medium and high voltage (11kV, 22kV, 33kV, 66kV) and equating to approximately 2GB of data. Disturbances analysed in the survey are voltage variation, voltage unbalance, voltage harmonics (total harmonic

distortion (THD) and 5th harmonic magnitude), and voltage sags. Plans are in place to continue the survey for at least the next few years.

It is well known that routine power quality surveys from only a few sites generate large amounts of data, the handling of which can be very difficult. When the number of sites is multiplied by 10 or even 100 times, the task of storing and analysing this data is certainly not trivial. Added complexity occurs due to the fact that data is often submitted by utilities in a multitude of different formats from a range of instruments, all of which must be converted to a common format before it can be added to existing data or analysed.

To address the problems of data storage and analysis, a database and data format has been developed by the University of Wollongong. This paper outlines solutions to the three main areas of data management, namely database design, data interchange format and data processing.

2. DATA INTERCHANGE FORMAT

In order for data to be passed from one system to another, as is inherently required by any application which collects data from a variety of sources, the data must be in a format understood by both systems.

7-bit ASCII (plain text) is an encoding method readable by almost all applications and systems, and is therefore a good medium for transferring data. However, reading is not the same as understanding. One system may put voltage readings first, followed by harmonics, and another may start with harmonics and follow them with voltages. It is necessary for systems exchanging data to agree not only on an encoding method, but also on a layout, or format.

CSV (Comma Separated Values) files are plain text files that divide each row of data into fields using a comma. The order of the fields must be predetermined to prevent data being incorrectly interpreted. A specific layout, called the PQ4 format, has been developed for the four types of data collected for the LTNPQS, these being site data, instrument data, continuous measurements and discrete measurements. Continuous measurements are regular, routine samples of voltage and harmonics levels ideally reported at 10 minute intervals as suggested by international standards [3]. Discrete measurements (voltage sags) are triggered by voltage deviation below a predetermined threshold.

The PQ4 format requires specific measurements in a specific order. This standardisation ensures that the systems involved in the transfer process agree on what data goes where and how it is presented.

Unfortunately, the real world isn't an ideal place, and occasionally a data provider is unable to provide the data in the most convenient format. As a result, data has been received in many different formats. A method of converting data from one text-based format to another has been developed to cater for this variety, using text manipulation tools.

The UNIX-based text manipulation utilities *sed*, *grep*, and *awk* are well documented [4], have been around for most of the computing era and are well suited to this sort of task. Using these utilities it is possible to write a simple script to reorder the fields in a CSV file. Sometimes it has been necessary to insert blank fields where the original file did not contain required fields, or to exclude fields completely (such as current which is not considered in the present survey).

Even when the file is in a standardised format, having been either received that way or converted, the receiving database system must still be programmed to recognise what each field signifies, and store it in the right column of the database table. An alternative data format, XML, can overcome both this problem and the problem of missing or extraneous fields.

XML (eXtensible Markup Language) is a technology that has been generating increasing amounts of interest in the computing world in the last few years and also within the power quality community [5]. XML data is self describing, in that each field is labelled, or tagged, as it appears in the plain text file. It allows data to be exchanged between systems with minimal prior agreement on what data should be present and where it should be.

Due to the tags, it is clear from the context which value represents which field.

XML is a subset of SGML (Standard Generalised Markup Language), which is an extremely formal and general markup language, and due to those traits is arguably not particularly useful. XML is more structured and easier to use, and has been adopted by industry at a much higher rate than SGML. HTML (HyperText Markup Language) is also a subset of SGML, however standard HTML limits the tags used to a predefined set, whereas XML allows document authors to define their own tags.

Another advantage of using XML is that the vast majority of modern database management systems understand it, and can work with it easily. It is often a simple matter of pointing a database to an XML file, and clicking "import". A disadvantage is that the XML files are necessarily larger and more verbose than the corresponding CSV files.

An example of continuous data in XML might look like the following:

```
<PhaseA>240.3</PhaseA>
<PhaseB>239.9</PhaseB>
<THD>2.11</THD>
<Fifth>1.85</Fifth>
<PhaseC>242.8</PhaseC>
```

There can be no mistake about which value belongs in which field. Changing the order of the values causes no confusion, and leaving out a value altogether doesn't disrupt the other values. A value with an unrecognised label is simply ignored.

A set of tags has been developed to correspond to field names in the database tables, in order to make it as simple as possible to import XML data into the survey. These tags are short and concise, in an attempt to reduce the overall size of the XML file. They are abbreviations of field names in the PQ4 format, yet remain descriptive enough to be human readable. Examples of these tags can be seen in the field names of the tables shown in Figure 1.

Since data providers at this point are unlikely to provide data in XML, it has been useful to transform data from CSV into XML. This is easier and more sensible than transforming from CSV into another CSV format, because the task is more readily generalised. A utility has been developed, implemented in *awk*, which does exactly that. It takes as arguments a description file and a CSV file, and produces an XML file as its output. This file can be used accurately by any database system that recognises the tags used.

3. DATABASE DESIGN AND PLATFORM CHOICE

Any reasonably sized power quality survey will need to deal with large amounts of data. Only one and a half years into such a survey, the LTNPQS database contains over eight million rows of continuous data, and one million rows of discrete data. As the survey gets into its third and subsequent years, the amount of data will increase exponentially as more and more sites are included in the survey, and data providers become more efficient at providing complete sets of data.

All this data must be filtered, cross referenced and sometimes transformed before meaningful results can be extracted. Results must be calculated only for measurements between specific dates and only measurements within a specific range must be included. Other factors such as the units that the measurements are in must also be taken into account. This places a substantial burden upon the database management system (DBMS). Choosing a DBMS that is capable of performing all the required tasks, now and in the future, was essential to the viability of the survey.

The system chosen needed to be a large scale, relational database system, capable of supporting Structured Query Language (SQL), stored procedures, indexing, and preferably materialised views (described below). Examples of such systems include Oracle, Sybase, Microsoft SQL Server and the latest version of MySQL. Database systems such as Microsoft Access, FileMaker Pro, Microsoft Excel and Visual FoxPro would be unsuitable for such a large scale task [1].

After some deliberation, Microsoft SQL Server was chosen, due to its support for XML, its ability to interface easily with Microsoft Access for presentation purposes, and its ability to handle large amounts of data.

Once that decision had been made, it was necessary to design the database objects to best facilitate the transfer and processing of data. Database objects that required consideration were tables, indexes, views and stored procedures.

To make the task of importing data as simple as possible, the table structure needed to reflect the format of data coming in. To that end, the tables in the database were designed along the lines of the PQ4 format described above, i.e. there are four main data tables: *site*, *instrument*, *continuous* and

discrete. The fields in each table reflect the fields contained in each PQ4 entity.

The relationships between the tables are as follows: each instrument is located at one site, and each measurement (continuous or discrete) is taken by one instrument at one site, as can be seen in Figure 1.

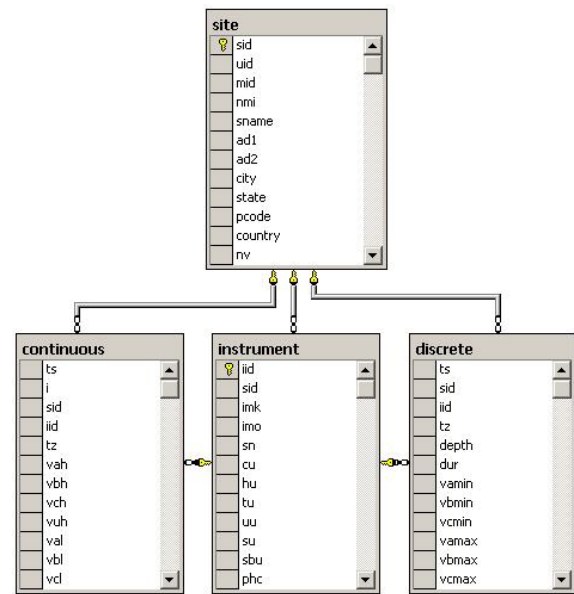


Figure 1: The relationships between the database tables.

Before the data stored therein can be used, it must go through several “normalisation” steps. For voltage data, this includes presenting voltage readings as per unit values and compensating for phase-to-phase vs. phase-to-neutral measurements. Unbalance must also be calculated from the voltage levels, because a measured unbalance isn’t always provided. For discrete measurements, events must be aggregated across phase and time to give a single customer event, and careful distinction must be made between the various units that can be used. These normalisation processes are more fully explained in a recent paper by Gosbell et al. [6]. The current transition period between a 240V standard and a 230V standard in low voltage systems is an additional complication that must also be catered for.

This constitutes a lot of preparation of the dataset before any of the actual power quality calculations can even be started. To make this preparation run more smoothly and in a timely fashion, as well as to facilitate the PQ calculations themselves, database indexes must be used.

A database index is a concatenation of nominated fields for a table (often just one), along with a pointer to the actual data record. Indexes can improve search time by a number of methods. One type of index is an ordered index. Search methods that are applicable to a set of ordered data can be very much more effective than those applied to an unsorted set. Once a set of row pointers has been obtained by this method, only the required rows of data need be retrieved.

Without the use of indexes, the entire table must be scanned. This is called a “table scan” for obvious reasons, and is highly inefficient for large tables such as those in the LTNPQS database. In terms of computer time, disk operations are excruciatingly slow, and so any technique that prevents an exhaustive search for data provides huge gains in efficiency and processing time.

Amongst others, one of the most important indexes in the LTNPQS consists of the SiteID and Timestamp for each row of continuous data. As described above, this allows a set of record pointers to be quickly obtained for readings from a particular site and between two dates. This is significantly more efficient, and hence much faster, than scanning the entire table of eight million rows and checking the SiteID and Timestamp of each row.

Indexes have a disadvantage – they must be updated each time a record is changed in the base table. This slows down the process of importing and changing data. Therefore it is important not to simply put indexes on every field and every field combination. The indexes must be carefully chosen to maximise the efficiency of the processes occurring within the database.

4. DATA PROCESSING

Once the data has been normalised, and is comparable in a meaningful way, the next step is to summarise it to produce an index or indices (not to be confused with database indexes as described above), which characterises the site or utility. This index may take the form of a 95th percentile of a particular set of values, or an RMS average of values, amongst other types of indices [6].

A variety of routines have been developed to calculate these indices. SQL (Structured Query Language) is a language that is used to manipulate database tables, present subsets of data, and match up information between tables. There are two different approaches to extracting summary information: using plain SQL (views) and using

procedural SQL. Procedural SQL adds to SQL the ability to apply a sequence of operations to the data, and flow control mechanisms.

Some routines have been implemented using plain SQL, and some using a stored procedure (procedural SQL). For example, the process of finding the 95th percentile of a site’s voltage readings must be implemented using a stored procedure, as it involves a series of steps, and no built-in function is available in standard SQL to calculate it.

Firstly, the 95th percentile for each phase at a particular site must be calculated. This involves filtering the data to include only readings from that site, only readings between two particular dates, and only readings between particular values (0.8 and 1.2 per unit). Then the data must be sorted, and the number of readings noted. 95% of the number of readings will give you a position number in the ordered set of readings, which will usually not be a whole number. The two readings on either side of that position must be interpolated to get the final value. Once this is done for each phase, the maximum value of the three phases is taken as the value for the site as a whole.

Other indices can be calculated using plain SQL since they only require the data to be presented in a different way.

Database indexes are just as important at this stage of the process as they were in the previous stage. The particular fields used by each of these routines have been taken into consideration, and indexes applied appropriately to get the most efficient processing.

The next phase in optimising the processing of data would involve the use of materialised views. The normalisation procedures described above are implemented using views (i.e. plain SQL). Those views can take time to run, even with the added efficiency provided by the indexes. A materialised view is stored much as other database tables are stored: as actual data. It doesn’t need to be calculated each time it is used. It does, however, need to be updated each time the base table changes, and the overhead required to keep the two sets of data synchronised is non-trivial.

5. CONCLUSION

This paper covers three aspects of data management for long term power quality surveys: data collection, data storage and data processing. When data is collected, it is often in a variety of different

formats, which must be transformed to a common format before being stored in the database. Some methods of performing those transformations are discussed, using a UNIX-based text processing utility called “awk”. The standard text-based format for the University of Wollongong study, entitled the PQ4 format, is described, and the use of XML as an intermediate format is discussed.

The huge amount of data involved in a long term survey is discussed, including the amount collected so far and projections for the future. From only two years of a long term survey, about 5 GB of data has been accumulated. Storing that amount of data requires a large scale relational database management system (RDBMS). Some common database systems are compared. The effect of the database schema design and indexing on query processing is discussed.

Future changes to the database design could include a change in methodology from the current OLTP (OnLine Transaction Processing) design to an OLAP (OnLine Analytical Processing) design. The OLAP methodology involves a multidimensional database schema, known as a star schema, and pre-calculated statistics stored in “cubes”. This approach may well be better suited to this type of application, but is by far a more complicated concept.

6. REFERENCES

- [1] V. Smith, P. Vial, V. Gosbell and S. Perera, “Database Design for Power Quality Survey”, Proceedings of AUPEC 2001, Perth, pp. 79 – 83.
- [2] V. Gosbell, S. Elphick, M. Peard, V. Smith and R Barr. "Long Term National Power Quality Survey of Australian Electricity Distributors - Year 1", Confidential reports, Integral Energy Power Quality Centre, October 2003.
- [3] IEC 61000-4-30 (2003) “Electromagnetic compatibility (EMC) – Part 4-30: Testing and measurement techniques – Power quality measurement methods”.
- [4] M.G. Sobell, “UNIX SYSTEM V: A Practical Guide”, 3rd Ed., Addison-Wesley, 1995.
- [5] J. Braun, V. Gosbell, D. Robinson, “XML Schema for Power Quality Data”, 17th International Conference on Electricity Distribution (CIRED), Barcelona, Spain, May 2003, Session 2, Paper No. 42.
- [6] V. Gosbell, A. Baitch and M. Bollen, “The Reporting of Distribution Power Quality Surveys”, CIGRE/IEEE-PES International Symposium on Quality and Security of Electric Power Delivery Systems, Montreal, Canada, October 2003, Paper 204.