UNIVERSITÀ DEGLI STUDI DI BERGAMO
Facoltà di Ingegneria

Corso di Laurea Specialistica in Ingegneria Informatica
Classe n. 35/S - Ingegneria dell'Informazione

Database management for an electrical distribution network of intermediate complexity
CERN - Genève

Relatore:
Chiar.mo Prof. Marco Villa

Correlatori:
Gent.ma Sonia Infante
Gent.mo Geroges Burdet

Tesi di Laurea Specialistica
Daniele De Ruschi
Matricola n. 57688

ANNO ACCADEMICO 2009/2010

Printed on July 5, 2010.

Riassunto

L'integrazione dei dati è da decenni una delle sfide più importanti che un ingegnere informatico deve affrontare; con l'avvento di nuovi metodi di interoperabilità dei sistemi informativi e il continuo moltiplicarsi di servizi che necessitano di flussi sempre maggiori di informazioni provenienti da fonti diverse, l'integrazione dei dati ha acquisito un ruolo fondamentale nel mondo informatico. Le differenti metodologie sviluppate in anni recenti indicano che non esiste un'unica soluzione, ma che il contesto del progetto realizzazione e gli obiettivi da raggiungere determinano quale tra gli approcci di integrazione è il più adatto. Al CERN presso il dipartimento di Ingegneria Elettrica Controllo (EN/EL/CO) un team di sei persone gestisce il sistema SCADA che supervisiona, controlla e acquisisce dati per l'intera rete elettrica del centro di ricerca nucleare più grande del mondo; dove più di 5000 fisici e scienziati da tutto il mondo compiono esperimenti d'avanguardia per comprendere l'interazione tra materia e "energia oscura". In questo scenario si è dovuto affrontare per la prima volta una complicata sfida di integrazione dei dati del sistema elettrico, che ha portato a una rappresentazione unificata e affidabile dell'intero network di controllo elettrico e del suo flusso informativo. Inoltre, l'accesso ai dati è stato reso più facile e le informazioni sono state messe a disposizione anche ai centri di gestione degli esperimenti principali (ATLAS, CMS, Alice e LHCb) che nel passato hanno sofferto importanti ritardi per mancanza di informazioni sul sistema elettrico.

Usufruendo della complessa e funzionale infrastruttura informatica del CERN ho ideato una soluzione in grado di raggiungere gli obiettivi di progetto rispettando i vincoli di risorse imposti e in particolare la conclusione del progetto prima dell'avvio dell'acceleratore principale (LHC).

La soluzione realizzata utilizza uno dei metodi applicati all'integrazione dei dati nei sistemi informativi: extract, transform, and load (ETL); questa scelta è stata motivata principalmente dalla sua semplicità e dall'omogeneità dei sistemi sorgente. Nonostante molti prodotti, sia pre-esistenti al CERN sia di carattere commerciale, offrano strumenti per la realizzazione e l'automazione di processi ETL, si è deciso di scrivere del codice ad-hoc principalmente in SQL per non dover dipendere da un framework software che avrebbe interagito con i database del sistema di controllo della rete elettrica.

Il lavoro fatto consiste nell'implementazione di un approccio ETL per l'integrazione dei dati relativi la struttura di controllo della rete elettrica:

1. Estrazione: i dati interessanti vengono selezionati dai sistemi sorgenti e combinati tra loro per mezzo di join relazionali;
2. Trasformazione: i dati vengono controllati verificati ed eventualmente modificati al fine di rispecchiare le caratteristiche imposte dal sistema di destinazione;

3. Inserimento: i dati vengono inseriti nel sistema unificato dove potranno essere facilmente raggiungibili anche da sistemi esterni.

Oltre al codice relativo alle tre fasi del processo ETL, ho realizzato delle applicazioni dimostrative attualmente usate dal dipartimento di controllo per la visualizzazione e navigazione dei dati collezionati e memorizzati nel nuovo sistema informativo.

Keywords: CERN, Data Integration, ETL,
Electrical Network, Database, Data Warehouse, EN/EL/CO.

Preface

This thesis is submitted as the final work for the degree of Master of Science in Engineering of Information that has been taken by the writer at at University of Bergamo, Italy. The report is based on the work conducted by the writer from September 2009 throughout June 2010 on a project assignment given by the department of Engineering Electrical Control at CERN Genève The work performed is a contribution to the GESMAR system of CERN. GESMAR is a CERN made complex platform for support and management of electric network.

In this work is developed an information system for an ETL process. The report presents the design, implementation and evaluation made, prototypes of applications which take advantages of new information inserted in GESMAR are also presented.

I wish to thank prof. Marco Villa for the support to my work from Italy and motivations gave to me along these years. I'm also grateful to my supervisor at CERN Georges Burdet and the section leader Sonia Infante for the excellent support and valuable feedbacks.

Genève, July 7, 2010

Daniele De Ruschi

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 The CERN

Founded in 1954, The European Organization for Nuclear Research (CERN) is an intergovernmental organization with 20 Member States[1]. It has its seat in Geneva but straddles the Swiss-French border. CERN, is the worlds largest and most famous center for scientific research. Its objective is to foster collaboration among European States in the field of high energy particle physics; to this end, CERN designs, constructs and runs the several particle accelerators with the associated detectors and experimental areas.

At CERN, complex scientific instruments are used to study the basic constituents of matter the fundamental particles. By studying what happens when these particles collide, physicists learn about the laws of Nature. The

[1]CERN, the European Organization for Nuclear Research, is the worlds leading laboratory for particle physics. It has its headquarters in Geneva. At present, its Member States are Austria, Belgium, Bulgaria, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Italy, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland and the United Kingdom. India, Israel, Japan, the Russian Federation, the United States of America, Turkey, the European Commission and UNESCO have Observer status

instruments used at CERN are particle accelerators, detectors and computers. Accelerators boost beams of particles to high kinetic energies before they are made to collide with each other or with stationary targets. Detectors observe and record the results of these collisions, while computers manage, control and analyze the information flow. The Large Hadron Collider (LHC) is the biggest particle accelerator, currently under commissioning at the CERN site. The LHC will mainly accelerate and collide 7 TeV proton beams. It is installed in a 27 km circumference tunnel, about 100 m underground, which previously housed the Large Electron Positron Collider (LEP). There are 8 access points, from the surface to the underground infrastructure, spread along the circumference tunnel.

CERN has been playing an important role for the scientific community with a huge impact on the daily life of every one of us. Everybody knows and uses the word wide web, but few knew that it was invented at CERN. Other important results are being used in medicine therapy for cancer, in technological applications for the industry, among others.

Many Nobel Prizes have been awarded for work made at CERN and numerous Nobel Laureates currently use its facilities. At present, more than 5000 physicists from all over the world cooperate in research programs carried out at CERN.

The LHC design is based on superconducting twin-aperture magnets which operate in a super-fluid helium bath at 1.9 K.

1.2 Description of LHC and its experiments

The LHC will produce head-on collisions between two beams of particles of the same kind, either protons or lead ions. The beams will be created in CERNs existing chain of accelerators and then injected into the LHC. Superconducting magnets operating near 4K will guide the beams around the

Figure 1.1: LHC accelerator at CERN

ring. The main accelerators that will operate in cascade are the Proton-Synchrotron (PS), the Super Proton-Synchrotron (SPS) and in the near future the Large Hadron Collider (LHC).

Each beam will consist of nearly 3000 bunches of particles and each bunch will contain as many as 100 billion particles. When bunches cross, there will be only about 20 collisions among 200 billion particles. However, bunches will cross about 30 million times per second, so the LHC will generate up to 600 million collisions per second. LHC will provide collisions at the highest energies ever observed in laboratory conditions. Four detectors, namely ALICE, ATLAS, CMS and LHCb, will observe the collisions so that the physicists can explore new territory in matter, energy, space, and time.

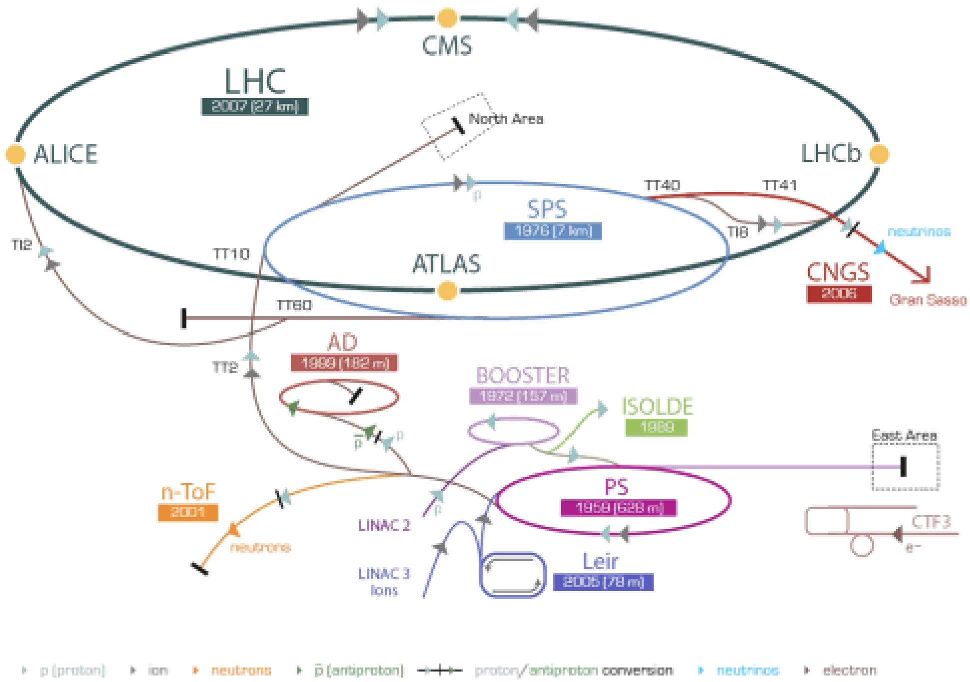

Figure 1.2: CERN Accelerators complex

1.2.1 ALICE

The ALICE Collaboration is building a dedicated heavy-ion detector to exploit the unique physics potential of nucleus-nucleus interactions at LHC energies. The aim is to study the physics of strongly interacting matter at high energy densities, where the formation of a new phase of matter, the quark-gluon plasma, is expected. For this purpose, ALICE intends to carry out a comprehensive study of the hadrons, electrons, muons and photons produced in the collision of heavy nuclei.ALICE will also study proton-proton collisions both as a comparison with lead-lead collisions in physic areas.

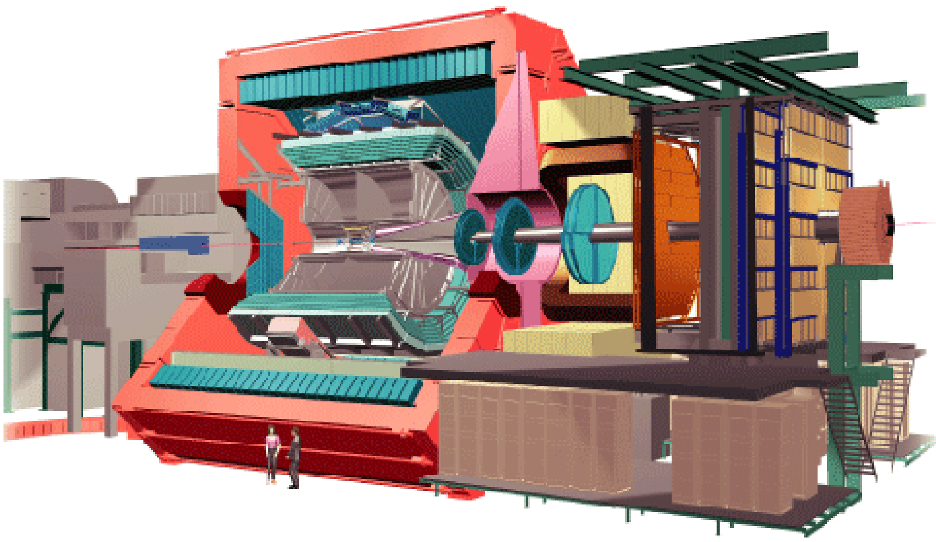

Figure 1.3: Drawing of ALICE detector

1.2.2 ATLAS

ATLAS is a particle physics experiment that will explore the fundamental nature of matter and the basic forces that shape our universe. ATLAS detector will search for new discoveries in the head-on collisions of protons of extraordinarily high energy. ATLAS is one of the largest collaborative efforts ever attempted in the physical sciences. There are 1900 participating physicists (Including 400 students) from more than 164 universities and laboratories.

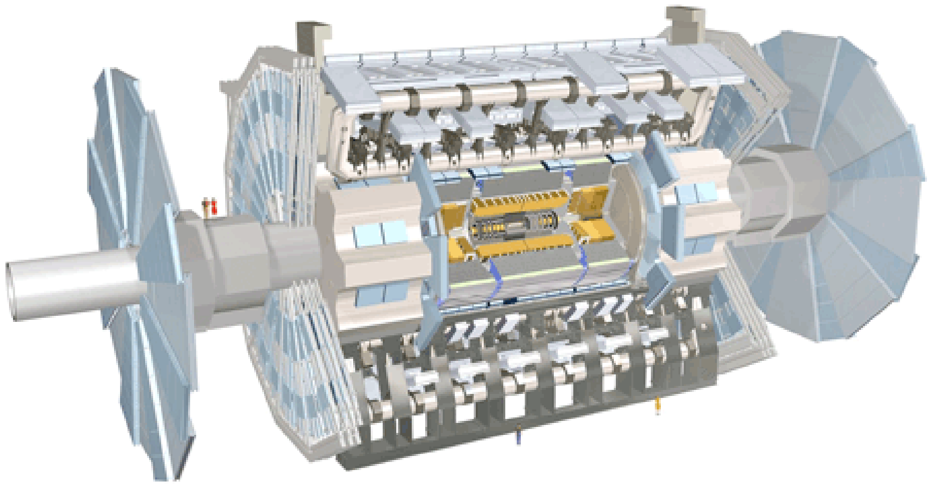

Figure 1.4: Drawing of ATLAS detector

1.2.3 CMS

CMS is a general-purpose detector, capable of studying many aspects of proton collisions at 14 TeV. It contains subsystems to measure the energy and momentum of photons, electrons, muons, and other products of the collisions. It is based on a silicon-based tracker surrounded by a scintillating crystal electromagnetic calorimeter, itself surrounded with a sampling calorimeter for hadrons. The tracker and the calorimeter are compact enough to fit inside the CMS solenoid which generates a powerful magnetic field of 4 T (Tesla).

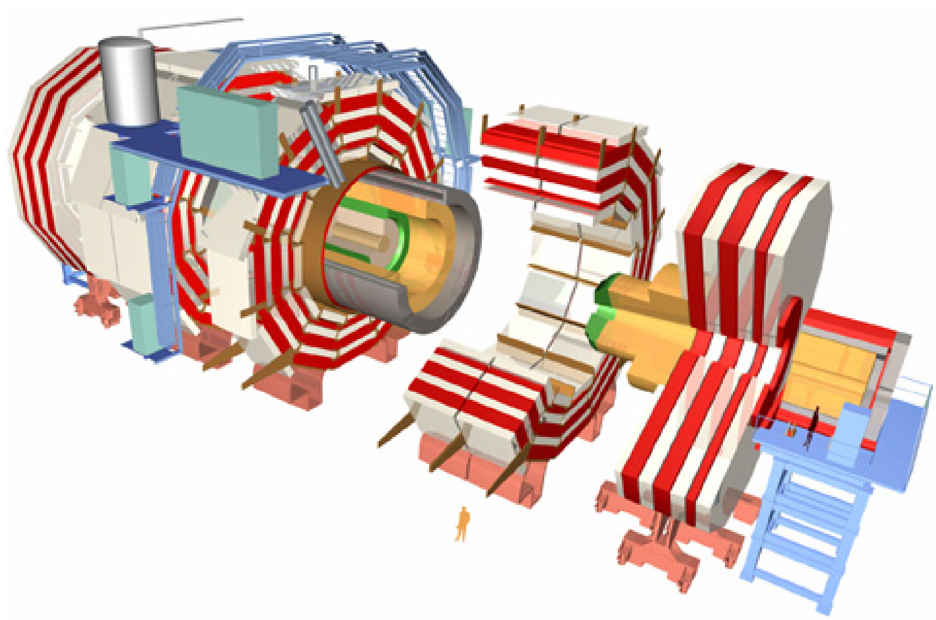

Figure 1.5: Drawing of CMS detector

1.2.4 LHCb

The LHCb detector is a specialized tool designed for B meson physics. It turns out that at the LHC, B mesons are most likely to emerge from collisions close to the beam direction, so the LHCb detector is designed to catch low-angle particles. Its key elements will be its vertex detector, that will measure charged particle tracks, and its RICH detectors (Ring-Imaging Cherenkov), that will identify different kinds of particles.

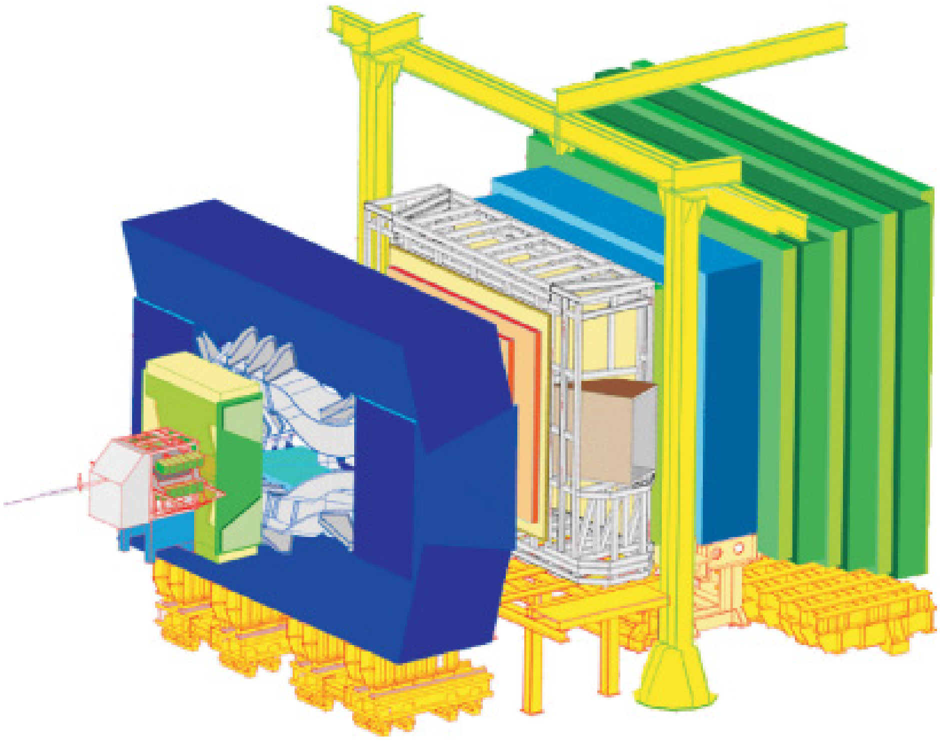

Figure 1.6: Drawing of LHCb detector

1.3 Power Demand & Energy Consumption

Particle accelerators and detectors as well as infrastructure equipment need to be supplied with electricity. A large fraction of the LHC electrical consumption goes into maintaining the superconducting magnets at their operating temperatures. Thanks to the superconducting technology employed for its magnets, the nominal consumption of the LHC is not much higher than that of the Super Proton Synchrotron (SPS), even though the LHC is much larger and with higher energy density. Accelerators are stopped in winter time to perform maintenance tasks when the energy costs are higher. The CERN base load is around 40 MW while the mean power reaches 130 MW in summer time. With the PS and SPS accelerators in operation, the peak power is around 250 MW as these are pulsed machines. With the LHC in operation, the mean power will increase to 230 MW. The nominal annual electricity consumption for full use of all accelerators is about 1000 GWh.

1.4 Electrical distribution network

Due to the complexity of CERN, both in terms of its power demand and internal topology, and due to the severe power requirements of particle accelerators and related experiments, a high performance electric network is needed. CERN is supplied with power from high-voltage networks of France and Switzerland. CERN has installed and is operating its own internal network for transport and distribution to the accelerators and the general facilities and infrastructure, operating at all voltage levels from 400kV to 48V and consisting of about 160000 control points.

The CERN electrical network is supplied from EOS (Energie Ouest Suisse), the Swiss regional grid company, through a 130 kV line and from RTE (R'éseau Transport Électrique), the French national grid company through a 400 kV line. With the 130kV Meyrin and the 400kV Prévessin network connections there are two input substations corresponding to these two sources. The entire CERN network is fed through the 400 kV RTE supply during accelerator operation and fed through the 130 kV SIG/EOS supply during the shut-downs, typically occurring in winter time. The switching from one source to the other is performed without power cuts twice a year. Accelerators operation on the 130 kV supply is not possible. In case of a breakdown of one of the two sources, 400kV or 130kV, CERNs decentralized Automatic Source Transfer (Autotransfer) system automatically detects this situation individually for each major substation, and changes over to the remaining source. The Autotransfer sequence takes about 20s from the loss of power until the completion of the source transfer. The Autotransfer system is limited to 60MVA in total, and it is only used for general services and part of the LHC cryogenics.

The end users are supplied at 18 kV, 3.3 kV and 400 V voltage levels. Within CERN, the 18kV level is the backbone for energy distribution, it covers all area of CERN. It is divided into several parts related to the specific type of loads: pulsed loads, stable loads, Meyrin Site general services, Prévessin Site general service, Super Proton Synchrotron (SPS, another CERNs accelerator that will be used as proton beams injector for LHC) Site, other accelerators and experimental areas, and finally, the LHC areas.

Accelerators generate large fluctuating of active and reactive power and produce large harmonic currents, injected back in the network. As an example, the SPS accelerator has an 18 s cycle pulse where the power consumption varies by 150 MW, when operated at 450 GeV (450 x 109 eV) beam energy fig. 1.7 .

To limit the network pollution (voltage swing and harmonic pollution), the network is divided in many sectors. Sectors containing manly magnet rectifiers are equipped with compensating and filtering systems. Fluctuations of the reactive power consumption are compensated with saturated reactors or thyristor-controlled rectifiers (TCR). Harmonic currents are eliminated with

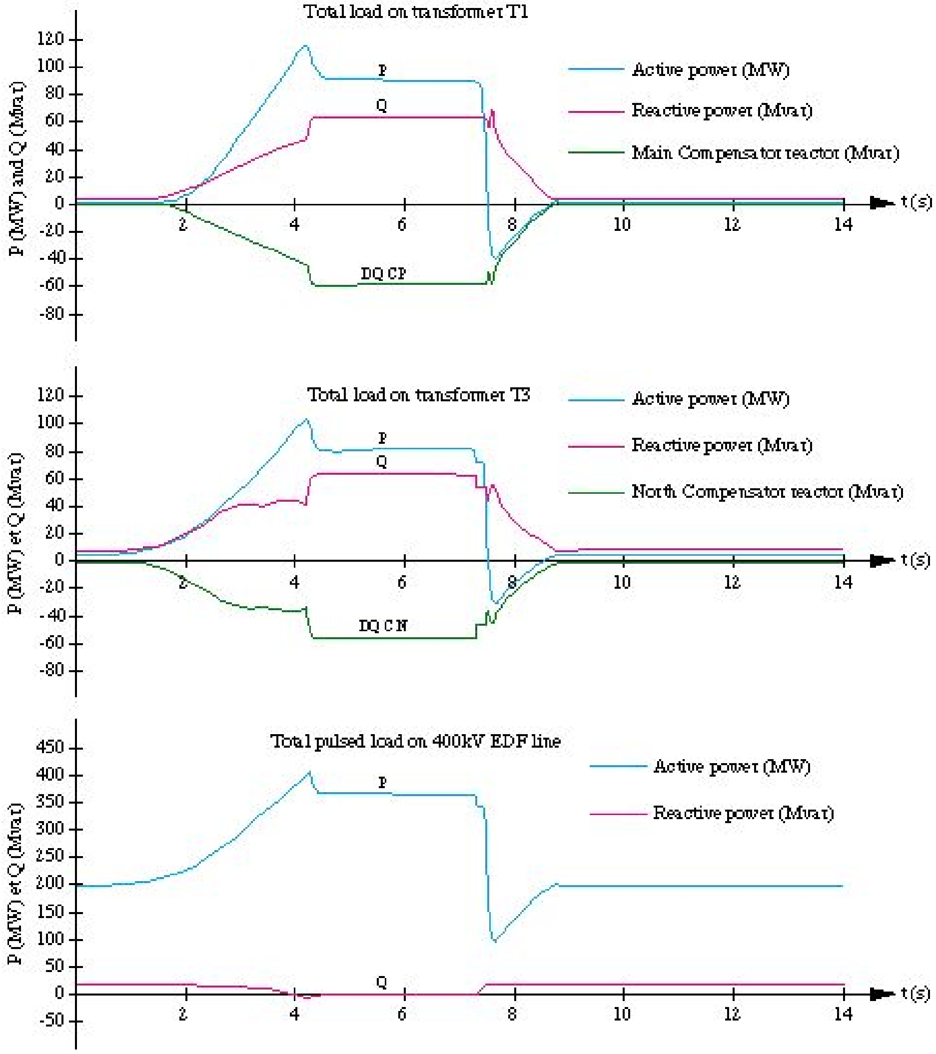

Figure 1.7: SPS Pulse load and its reactive power

harmonic filters, which simultaneously generate the necessary reactive power with capacitor banks.

1.5 The 400 kV and 130 kV Levels

The main 400 kV substation on the Prévessin site (France) fitted with three 400/18 kV 90 MVA and two 400/66 kV 110 MVA transformers, is owned and operated by CERN Electrical Service and dedicated to operation of LHC and SPS machine. The substation is connected with its overhead line to the Génissiat HV station and with the Valais (CH) back-up station. The 130 kV substation, located on the Meyrin site, is fed from the Verbois SIG (Services Industriels de Geneve), the Geneva local power distributor. The substation, fitted with three 130/18 kV transformers, is owned and operated by CERN, is a back-up to maintain critical loads in case of 400kV power failures. It also offers a high degree of flexibility in winter and gives some advantages and potential cost savings when negotiating CERNs energy contracts.

1.6 The 66 kV Level

The 66 kV substation supplies the CERN stable loads: all the general services, the Proton Synchrotron (PS, a CERN accelerator) complex and some LHC areas. Its a double bus-bar substation powered by the two 400/66 kV, 110 MVA transformers, operating in parallel. The 66 kV network supplies the LHC high-power points via cables laid in trenches and 66/18 kV, 38 MVA transformers installed in these points, as well as a 66/18 kV, 70 MVA transformer which is the source for the 18 kV stable network. A second 66/18 kV, 70 MVA transformer supplies the ATLAS experiment and the Meyrin site.

1.7 18 kV Pulsed Loads

An 18 kV substation, devoted to the SPS pulsed loads, consists of three power sections each equipped with a 400/18 kV, 90 MVA transformer, compensator and filter. One 93 MVAr compensator is based on saturated reactor technology. A 150 MVAr compensator is based on the newest thyristor control technology. A second 150 MVAr compensator based on this technology has been in operation since 2008. The 18 kV SPS pulsed network comprises sixteen links to the Auxiliary Buildings of the SPS machine (main magnets power rectifiers), the SPS pulsed loop (auxiliary magnets power rectifiers) passing through all SPS auxiliary building, and the North and West Experimental Areas.

1.8 18 kV General Services and SPS stable network

An 18 kV substation is devoted entirely to the CERN stable loads. It includes the SPS auxiliary buildings, the Prévessin and Meyrin sites, the Assembly Hall, workshops, storage and office buildings, as well as the North Experimental Area and LHC general services. The Prévessin and Meyrin sites may either work together or in tandem configuration.

1.9 18 kV LHC Machine Network

The 18 kV LHC machine network is fed from six 66/18 kV transformers. This network supplies the LHC machine systems (magnet rectifiers, radio frequency generators, cooling and ventilation of the LHC tunnel). In four of the major LHC points, a thyristor-controlled reactor compensator is installed at the 18 kV level.

1.10 18 kV LHC General Services Network

The 18 kV LHC general services network is fed from an 18 kV cable loop following the tunnel of the LHC machine. The loop passes through all LHC points where it supplies 18 kV switchboards, in surface and underground substations.

1.11 18 kV Network on the Meyrin Site

The main Meyrin substation is fed by five 15 MVA links from the EOS/SIG input substation, a 60 MVA link from the LHC point 1 substation and a 30 MVA link from the Prévessin site. Several 18 kV loops are used to feed the Meyrin site: for the physicists and administration buildings zones, the PS accelerator complex and the Isolde/AA/Booster complex, as well as the Computer Centre. A TCR load leveler compensates Booster pulses and eliminates harmonics.

1.12 3.3kV Network

The 3.3kV distribution network supplies mainly compressors for cryogenics, tunnels ventilation, cooling and safety services.

1.13 Diesel-Generator Rescued Network

The diesel-generator rescued network is a part of the CERN electrical network, which, in normal operation, is fed from the normal distribution network. In case of any voltage loss the secured network is automatically isolated from the rest of the network and re-supplied from the diesel generators. This network is divided in four subsystems: Meyrin, SPS/North Area, LHC and

ATLAS. There are a number of diesel generators connected to the 18kV level, acting as a back-up supply for critical loads in case of failure of the main power source. These generators are stand-by units, permanently ready for service and have a start-up time of 45s.

1.14 Low Voltage 400 V Network

As for the 18 kV network, the LV distribution is designed to facilitate network reconfiguration procedures within the hierarchy of networks, such as: pulsed network, machine network, general services network (backed up by auto transfer), rescued network (backed up by auto transfer and by the diesel-generator set).Technical systems like safety transmissions, communications, cryogenics, cryogenic instrumentation and power converter control systems, require uninterrupted power supplies, therefore they are fed through UPS systems.

1.15 48 Vdc Network

Auxiliary systems like HV/LV protection, Remote control, Communication system, Emergency lighting, Emergency stop, etc., are fed by a 48 Vdc source, with battery backup.

Chapter 2

CERN Power Distribution Network Monitoring and Control SYSTEM

2.1 Overview

The CERN electrical distribution network is supervised by the Electrical Supervisor system, it allows monitoring and control of all aspects of the network from the Cern Control Center (CCC) room or any other location on the CERN site.

The requirements of the Electrical service at CERN can be summarized as follows:

- Maintain, extend and improve Electrical Network supervisor system.
- Application and communication software maintenance.
- Specify procedure and install devices.
- Manage software, back-up and spare parts.

- Provide and maintain software for power and energy consumption tracking.
- Provide support for post mortem fault analysis.
- Standby service 24/7/365.
- Remote monitoring of equipment states:
  - Status lists of electrical equipment.
  - Synoptic diagrams - single line diagrams.
  - Alarm lists.
- Data recording for analysis:
  - Events, status changes.
  - Measurement states.
- Process control (local or centralized).
- Automatic procedures via standard languages.
- High-level procedures without programming.

In order to cope with these requirements and improve the user interface of the supervision system, in 1998 CERN launched a tender call regarding the participation of the major Member States SCADA (Supervisory Control and Data Acquisition) players. The supplied solutions are based on well proven solutions, respectively the SCATE X DMS for the Control Centre and 57 concentrators (CLP 500RTU) for acquisition, supervision and control of substations.

2.2 Electrical supervisor SCADA System

The electrical equipment of the CERN Power Distribution Network is monitored with a SCADA X DMS system, which provides the Cern Control Center (CCC) room and the electrical operation personnel with remote monitoring and control facilities. SCADA X DMS is a particular SCADA application, for electric power distribution management. It offers tools like the following, the list not being exhaustive:

- Network topology processing, aiming to identify energized and non-energized branches of the network, network coloring by power sources or by voltage levels, etc.
- State estimation and power flow calculations, aiming to provide calculated data about current flows, voltage levels, power flows, losses, etc., based on real time network state and effective measurements.
- Short Circuit Analysis, aiming to provide studies about the settings of protective high voltage equipment.
- Voltage/VAr Control, for reactive power minimization with impact on loss reduction.
- Fault Detection, aiming to provide faulty network branches isolation and load recovery by network reconfiguration.
- Alarm treatment and event, data logging for human intervention and post mortem analysis.

Typically, each company using SCADA systems, like CERN, has a Control Center providing tools for the network operators. These Control Centers communicate with Remote Terminal Units, placed in strategic points of the network, such as power substations. The network supervision and operation is performed in real time. This allows the network operators to be informed about any abnormal situation, assisting in executing the recovery procedures while minimizing the impact of any faulty situation.

This system manages almost all the equipments located in the CERN electrical network, including substations in surface and underground areas, as well as other network branches existing in technical buildings. The equipments supervised in these installations span all voltage levels from 48 V-DC of battery chargers to 400 kV breakers and transformers. Since the distribution network involves different generations of technology, the SCADA system comprises many heterogeneous hardware and software interfaces that requires different methods of integration.
Actually, the system manages about 180000 signals, which mostly carry status information but comes also from counters and analog instruments. Control output channels are available for remote operation of specialized substation equipment. By the time when all CERNs substation equipment will have been migrated into the new SCADA systems, a total of 200000 signals is expected to be managed by SCATE X DMS.

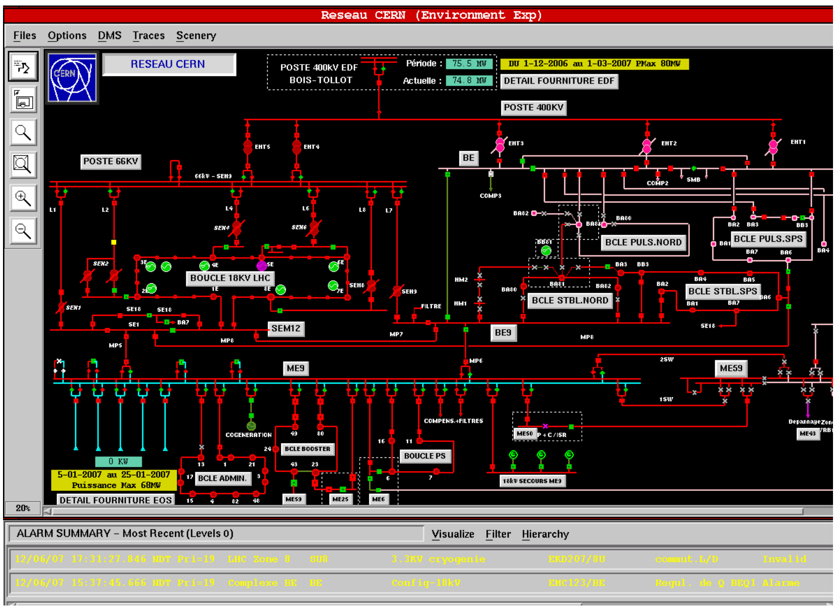

Figure 2.1: Example of a SCATE X DMS synoptic diagram representing CERNs main substations and link interconnections with alarm summary.

2.3 SCADA System Components and Features

2.3.1 Equipments and Electrical Devices

The operation of the electrical power network is maintained by different equipments such as breakers, transformers, switchboards; each equipment is controlled and monitored by the SCADA system. Due to the purchasing policy of CERN the electrical equipments come from different vendor; some of them are in operation since the LEP era while other are new prototype models with latest technology. As a consequence, the electrical network of CERN is composed by a large and heterogeneous variety of equipments, protocols and technologies.

2.3.2 Type of data

There are 4 different types of signals/data:

- Digital: a n bit signals that carry information about the status of equipments.
- Analog: are electric signal representing physical quantities such as voltage, current, temperature, etc etc.
- Counters: they usually increment each time a transition occurs.
- Controls, are used to send commands to the equipments.

Each equipment come with a set of these building blocks with a mean value of 10 data for equipment, although we have devices with just one data point and other (like SEPAM) with hundreds data points.

2.3.3 IED and PLC

Intelligent electronic devices (IED) and PLC are part of the communication and control layer they are programmed in order to react to input values and notify anomalies of controlled equipments.

Figure 2.2: Example of a PLC.

Figure 2.3: Example of a IED.

2.3.4 Acquisition Unit

They are dedicated acquisition boards that collect digital and analog data from electrical devices, they can also send control command to the connected device.

2.3.5 RTU Concentrator

CLP 500RTU is a specific application of CLP 500, EFACECs distributed platform for supervision and control. Basically, the architecture of the CLP 500 platform is based on a Central Unit, an industrial PC based unit, which communicates with a set of distributed Control and Acquisition Units, via a high speed LAN, such as Ethernet and several protocols over serial ports to communicate with IEDs. CLP 500RTU implementation follows CERNs specifications, namely those related to availability and data logging requirements.

RTU are placed in surface buildings span many substations, thats why some of them are so large in terms of data points that represent the full state of the power equipment. As these substations are permanently accessible, CERN decided to install RTUs providing a built-in human machine interface, used locally by any operator with permission to supervise or locally control the substations power process.

Figure 2.4: One RTU located in a surface electrical substation.

Due to limited access conditions of the underground areas, RTUs placed in

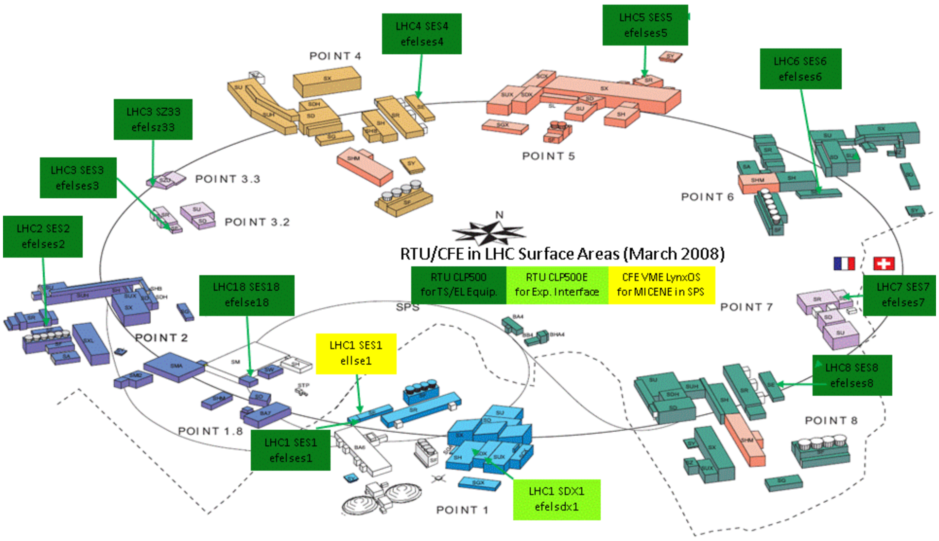

Figure 2.5: Location of RTU in surface buildings.

underground substations have no mechanical parts (they dont have any hard disks nor ventilation fans) to improve its reliability and reduce the necessary maintenance.

Underground substations are distributed along the LHC tunnels, some of them are very close to the LHC experiments distribution substations.

Figure 2.6: One RTU located in an underground electrical substation.

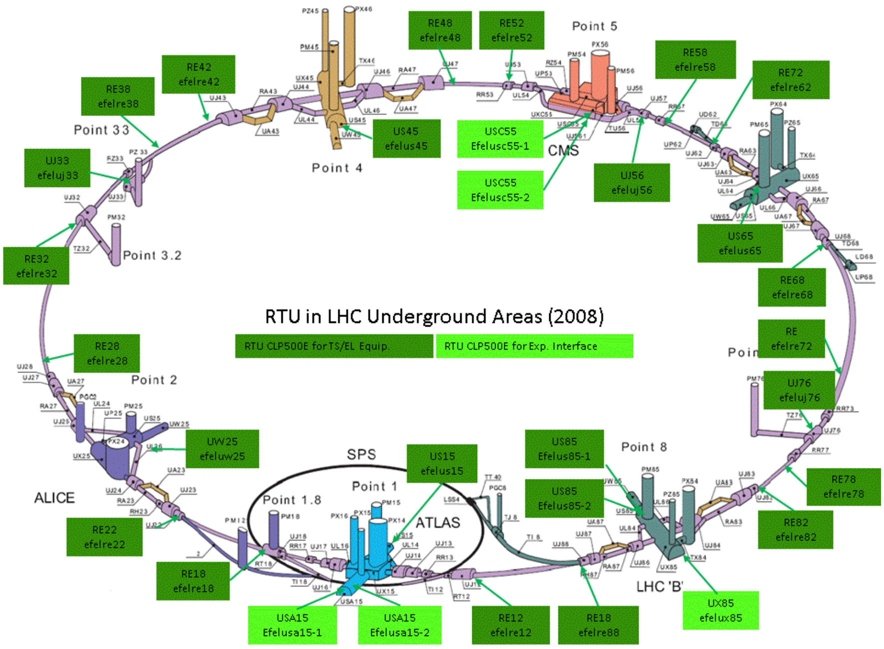

Figure 2.7: Location of RTU in underground.

2.3.6 Front End

Front End is composed of 8 RTU CLP500; its aim is to collect information from the concentrators and foreword them to SCADA servers. They are located in the SCADA servers room in the CERN Control Center building.

Figure 2.8: Front End of control system.

2.3.7 Servers watch dog SAH long term archive

The SCADA system has 2 twins Hewlett-Packard Proliant Type DL380 servers, running LINUX Red Hat Distribution, ORACLE Enterprise as RDBMS system and a SCADA software working in parallel for reliability and redundancy; another similar server is located in a different place in order to deal with disaster recovery.
All the data coming from the field are collected, monitored and verified based on reference values stored in database; in case of anomaly, the status of a particular device update and an alarm is inserted in a priority list.
The three servers are managed by a watchdog that monitor continuously their status and balance the load of the requests.
To keep track of the historical status of the electrical power network of CERN, every 10 minutes the most relevant information are saved to a database.

Figure 2.9: Severs of the SCADA system.

2.3.8 The CERN Control Center

The purpose of the CERN Control Centre (CCC) is to combine the control rooms of the Laboratorys eight accelerators, as well as piloting of cryogenics and technical infrastructures. Several people man the CCC; in their monitors they can visualize and analyze the current status of the Electrical Network and notify the right people in case of anomalies or fault.

Figure 2.10: Monitoring location in the CERN Control Center.

2.4 Main SCADA / Electrical Network Supervisor features

- Data acquisition from the state network, in particular from substations and switches strategically placed in the network. It also provides automation functions as well as control execution with impact on the network state, meaning that an operator can control remotely the networks shape and power profile. Relevant data is presented to the operators, by means of schematic diagrams dynamically representing the network in real time. Lists of alarms and events, as well as graphical curve displays are part of the high level features offered by the SCADA software.
- It tracks, in real time, the state of the network topology, namely knowing which network branches, substations bus-bars and substation feeders are electrically connected. It also keeps track of the circuit breakers state (opened or closed).
- Source of power, energized and non-energized network branches, voltage level
- State Estimation & Power Flow for study mode, it calculates, cyclically, on event or as a result of an operator request, the power flows and currents in all network lines and branches, the power losses, the currents and the voltage at every network node. This calculation is performed using SCADA input data for load power consumptions and power injection feed points.

2.5 Interfaces to external systems

CERN has to buy equipment from member states companies through supply contracts. Due to this, CERN has a very heterogeneous set of different

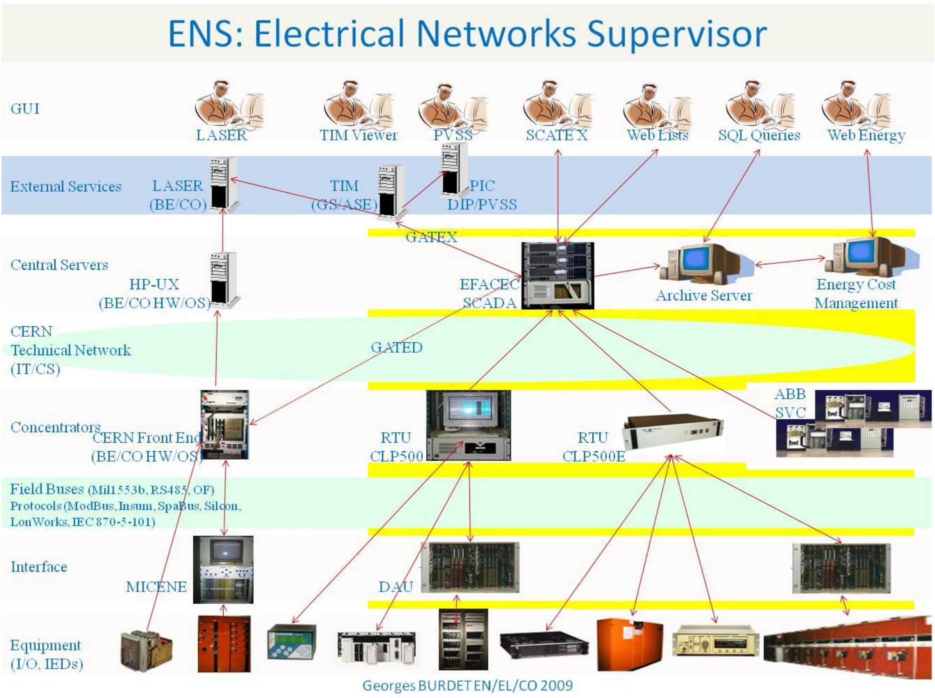

Figure 2.11: Overview of the Electrical Network Supervisor, on the left side there is the legacy control system that is scheduled for removal, while on the yellow stack there is the new control system architecture.

electrical equipments that have to be interfaced to the RTUs, through several proprietary and standard protocols, as well as field buses.

2.5.1 Protocols and Supports (connection)

The following list is not exhaustive, and presents the set of CERNs external systems which the RTUs interface to:

- JBUS digital protection relays for HV and MV substation, Battery Chargers, UPS and other IEDs
- SILCON UPS
- SPABUS Tap Change Controllers
- INSUM LV Breakers
- MLINK
- MODBUS
- LONWORKS
- IEC8705010X
- SIEUPS

2.5.2 Equipments

Different equipments communicating with CLP 500RTU, not restrictive:

- Digital Protection Relays
  - HV & MV SEPAM (SCHNEIDER)
    - JBUS
- Tap Changer Regulator

  - MR TCS 1.10 (MASCHINENFABRIK REINHAUSEN)
    * SPABUS
- PLC
  - PREMIUM, TWIDO (TELEMECANIQUE)
    * JBUS
  - S5 (SIEMENS)
    * JBUS
- LV IEDs
  - MEGAMAX (ABB)
    * INSUM
  - CIT1.32P (HAZEMEYER)
    * JBUS
  - DIRIS (SOCOMEC)
    * JBUS
- UPS
  - EPS2000, EPS5000, GALAXY, COMET (MGE SCHNEIDER)
    * JBUS
  - EDP90, SYNTHESIS (CHLORIDE)
    * JBUS
  - From APC
    * SILCON
- Battery Chargers
  - From PROMEC and ELTEK

* JBUS

2.5.3 NTP (network time protocol)

In order to synchronize the electrical control infrastructure a NTP service is used. A default installation of the NTP daemon on a modern UNIX based operating system generates continuous time synchronization to the order of 1ms. The group runs two Stratum 1 time servers on the Technical Network and two Stratum 2 time servers on the General Purpose Network.

Chapter 3

Project Overview

3.1 Current Limitations

The electrical control system is managing 180000 data points, but the same system is used in order to get detail information for administrative and managing purpose causing overload, security and error issues. Most of the time users interact with the control system just to copy data and paste them in a personal spreadsheet or making screenshots in order to make mathematical operations, statistical analysis and reports.
Expert users deal directly with SQL on the SAH (long term archive database) in order to get the historical information for analysis.

The main issues/problems can be summarized as follows:

- There is no a unique entry point with authentication and authorization managed centrally to retrieve information about the control network except the technical SCADA system.
- Information are on the Technical Network and they are only accessible using a Windows terminal server.

- There is not a clear scheme of the physical hierarchy of the control network; in particular its very difficult to answer to questions such as:
  - How many serial ports are used by a RTU ?
  - Which kind of acquisition units are connected to a RTU ?
  - How many equipments are controlled by a RTU ?
  - Which name should be given to a new equipment ?

  For example, when a new value for an alarm priority has to be set, the operator would like to know what are the RTU affected by this operation, in order to plan and schedule this intervention (access to experimental area is restricted since changes that could impact the machine operation can be done only during a technical stop).

- Make offline analyses. For example, it could be useful combine values and information based on location or same branch of the hierarchy tree or analyze the number alarms for a group of same electrical devices, but its very difficult and not automatic due to the lack of a digital representation of the control infrastructure.

Most of the information needed to solve these kinds of issues is spread around the electrical control system and its not fast and easy to access, especially for non technical users such as managers and analysts and operative staff.

3.2 Requirements

The aim of this project is to extract as much correct information as possible regarding the hierarchy of the control network in order to populate dimension tables of a Data warehouse that will contain information about the physical hierarchy of the control system creating a static representation of it. More these requirements will be respected; more end users will gain confidence in the database while making informed decision [1] and quickly retrieving reliable information.

3.2.1 Data type

Extracted data will be stored in already existing tables of a data warehouse; they have to be of the right type and of proper format.The naming convention will follow the CERN naming conventions; the names of devices seen in the field should be consistent with names given in the SCADA system.

3.2.2 Data accuracy

The data should be validated in order to have a working and trustable system from the beginning. Typographical errors, incomplete records or other dirty data should be correct during the process Semantic conflict should be notified and in case resolved or added to an issues list to increase the general quality of the data in the control system.

3.2.3 Timely data

The data should be fresh and corresponding to the actual state of the network and should be maintained as long as the situation requires. Old data should be updated to reflect real world changes.

3.2.4 Machine operation

The extraction of data from the 60 databases doesnt have to disturb the electrical control system that continuously interacts with these databases

3.2.5 Time constrain

Most of the integration process especially the extraction of information from database should be done before the LHC start in middle of November 2009.

3.3 Understanding the Data Sources

In order to know which kinds of information are available inside the sources database, a study on the RTU database scheme has been done. There are two types of RTU that are currently used at CERN; since both are using the same version of the database, this is a key point to consider for the extraction procedure; dealing with homogenous data sources at scheme level allows to use a simpler approach.

3.3.1 Inside an RTU database

Each RTUs has a database that stores information about the configuration of different devices connected to it. Since the electrical control system is provided by an external company, the database has been developed by them and then modified along 12 years in order to meet the always evolving functional requirements of CERNs electrical control network.
At the beginning of 1998, a new version of database has been released. Since then, the nor the core system nor the the interface to interact with it has ever been modified. The operators are forced to update and delete information directly on the tables, mostly with an error-prone cut and paste procedure.
Custom tables to manage new protocols and new kind of devices have been added and the total number of tables from the basic database have increased to 279 tables and more than 3000 relations fig. 3.1.

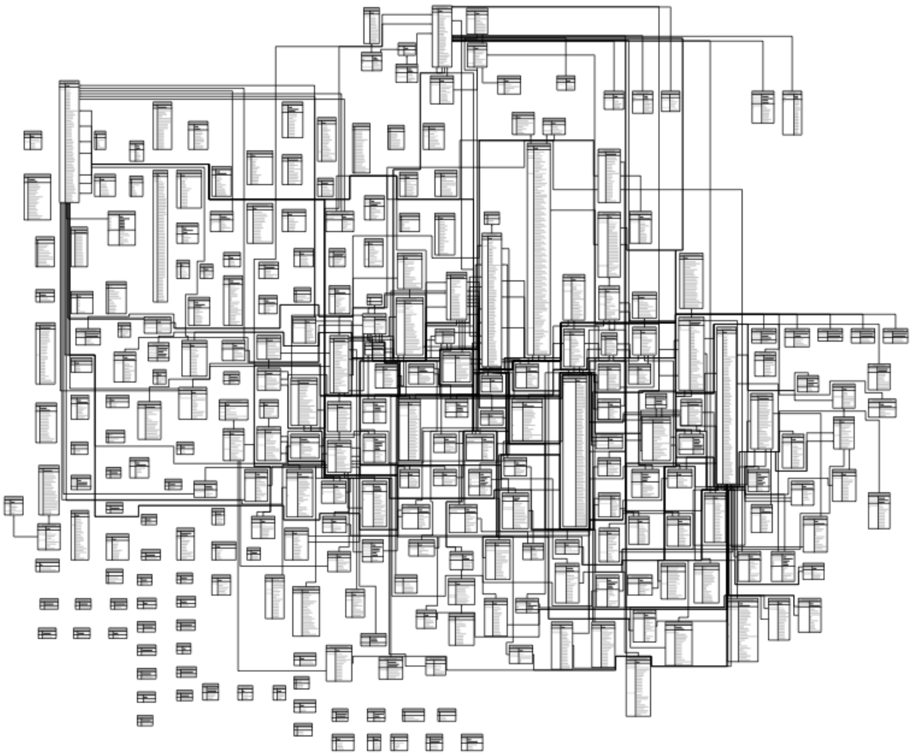

Figure 3.1: Relationship diagram of RTU database

3.3.2 Meaning of the tables

The database is an ACCESS file with the same structure for each RTU, it consists of 279 tables that permit to the RTU control software to retrieve and manage all the possible settings for the different protocols and equipments controlled.
There isnt any documentation for this complex database, but after a long analysis and consultation with operators that use this database regularly, it was possible to extract 45 tables storing data about hierarchy architecture, describing the information flow from the RTU serial port to the datapoint on the electrical device.

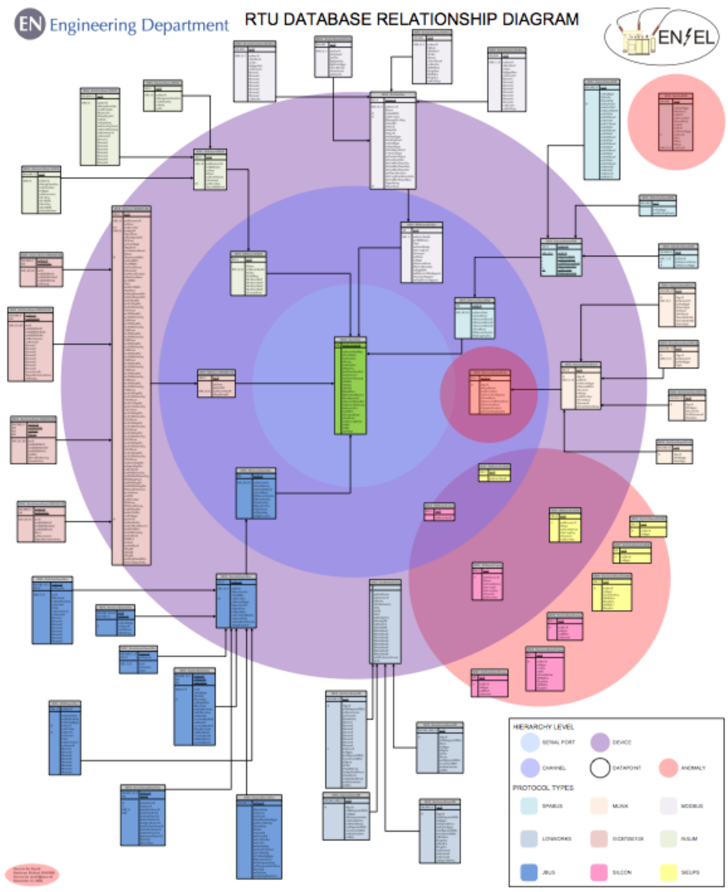

Figure 3.2: Tables to be extracted from the RTU database

From the center of the picture you can see four different hierarchy levels: serial port, acquisition, device, datapoint. Tables referring to the same protocol are filled with the same color; while the red areas denote anomalies in the definition of the database, corresponding to some of the custom addition made for CERN.

Tier 1 (Serial port) : The starting point can be considered the tCNF_SerialPort table; it is used to configure the communication parameters of the serial ports for example baud rate, data bits, parity, stop bits

Tier 2 (Protocol Channel): Eight tables:

- tCNF_CCDOWNModbus
- tCNF_ChannelDownSPA
- tCNF_ChannelDownMLink
- tCNF_CCDownSieUPS
- tCNF_CCDownSilcom
- tCNF_ChannelDownJbus
- tCNF_CCDownIEC870510X
- tCNF_CCDownInsum

are used to describe and configure the communication channels, specifying which serial port is used, the dialect of the protocol and its settings for example: message length, event scan, active channel...

Tier 3 (Device): Nine tables:

- tCNF_ModbusDev
- tCNF_DevicesDownSpa
- tCNF_DevicesDownMlink
- tCNF_RuDownSieUPS

- tCNF_RudownSilcom
- tCNF_LonWorksNodes
- tCNF_DevicesDownJbus
- tCNF_RuDownIEC80510X
- tCNF_INSUM

are used to describe electrical devices settings declaring which channel is used and which slave address for example: index, address, SyncTime, DeviceType, SerialNumber...

Tier 4 (DataPoint): For each kind of device protocol in the database are declared four tables name in the following way:

- tCNF_MedidasDownXXX
- tCNF_DigitaisDownXXX
- tCNF_ContadoresDownXXX
- tCNF_ControlosDownXXX

Where XXX stands for the specific protocol name; those tables declare which kind of datapoints are scanned by the RTU and particular settings in order to read and parse them on the bus.

3.4 GESMAR the meeting point of Electrical information

The information system for electrical information at CERN is called GESMAR: its based on a relational database Oracle and it covers two different domains:

1. Technical management of installations composed by three different modules

    - Object module manages position functional: equipments, tranches and departs electric . Using this module it is possible define all kind of characteristic (physic, functional, structural and classification) for any positions functional and composition of electric switchboards. All the other modules and applications use this as reference to any position functional to assurance cabling, control and maintenance.

    - Module for management of cabling. Using this module it is possible manage installation of all the electrical cables for general and technical services related to experiment machine.

    - Spare parts module. Using this module it is possible manage electrical spare parts to be used by ENEL at CERN. The spare parts are divided in 3 categories, cables (code: J), connectors (code: K) and others (code L).

2. Management of contracts of industrial services, it is composed by tens of modules and the most important are:

    - Case tracking module. Its used for description, monitoring and administrative follow-up for each work requires the contractor.

    - Financial module. Its used for payments and billing for work carried.

3.4.1 Interdependency modules

All the modules used by GESMAR are connected together and largely use reference to dictionaries that define common data. Most of the dictionaries are shared between modules. This permits to anyone who has to work with electrical data to use common definitions for objects. This project uses 5 different dictionaries:

- DI-SCR defines different kinds of physical support (RS485A, RS485C, Ethernet, Optical Fiber).
- DI-PCR defines different kinds of protocols (JBUS, MODBUS, SPABUS, LONWORKS).
- DI-DBI defines different databit values (8, 7).
- DI-PAR defines different parity values (Even, Pair, None).
- DI-VCR define different speed values (12, 1250, 192, 1MB, 24, 384, 48, 96).

3.4.2 Definition of new tables

In order to satisfy requirements and improve functionalities of the information system removing most of the current limitation, new tables have been added to the GESMAR database.
Three tables CT_POR, CT_RPA, CT_RAO correspond to the 3 hierarchy levels of the electrical network control system RTU, acquisition, equipments; other two tables CT_RSP, CT_SMA are used as support to define more details about the communication between levels.

3.4.3 SQL code

CT_SMA is a table used to define how acquisition and equipments are connected based on a particular scheme.

```
(CT_SMA_NO_SMA          NUMBER         NOT NULL     -- PK: Sequence
  SEQ_CT_CONTROLE
,CT_SMA_SCHEME          VARCHAR2(4)    NOT NULL      -- PK: Scheme
,CT_SMA_SUBTYPE         VARCHAR2(1)    NOT NULL      -- {'U' pour
  Acquisition, 'P' pour Position fonctionnelle}
```

```
  ,CT_SMA_VAR_TRAITMT    NUMBER(2)      DEFAULT 0 NOT NULL  -- Variante de
    traitement (measure)
  ,CT_SMA_VAR_CABLAGE    NUMBER(2)      DEFAULT 0 NOT NULL  -- Variante de
    cablage
  ,CT_SMA_LIBELLE        VARCHAR2(20)                       -- Description
    courte
  ,CT_SMA_DESC           VARCHAR2(45)   NOT NULL            -- Description
  ,CT_SMA_PLAN           VARCHAR2(10)                       -- Numero du
    plan
  ,CT_SMA_DAT_CRE        DATE           NOT NULL -- Date de creation
  ,CT_SMA_NO_ACT_CRE     NUMBER         NOT NULL -- FK1: NumerO acteur de
    creation
  ,CT_SMA_DAT_MOD        DATE                    -- Date de modification
  ,CT_SMA_NO_ACT_MOD     NUMBER               -- FK2: NumerO acteur de modif
    .


ALTER TABLE CT_SMA ADD
   CONSTRAINT PK_CT_SMA PRIMARY KEY (CT_SMA_NO_SMA)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_SMA ADD
   CONSTRAINT UQ_CT_SMA_1 UNIQUE (CT_SMA_SCHEME, CT_SMA_SUBTYPE,
     CT_SMA_VAR_TRAITMT, CT_SMA_VAR_CABLAGE)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_SMA ADD
   CONSTRAINT CK_CT_SMA_1 CHECK (CT_SMA_SUBTYPE IN ('U','P'));

ALTER TABLE CT_SMA ADD
   CONSTRAINT FK_CT_SMA_1 FOREIGN KEY (CT_SMA_NO_ACT_CRE)
                 REFERENCES SYS_USERS (USER_NO);

ALTER TABLE CT_SMA ADD
   CONSTRAINT FK_CT_SMA_2 FOREIGN KEY (CT_SMA_NO_ACT_MOD)
                 REFERENCES SYS_USERS (USER_NO);
```

CT_RSP is a table used to define which kind of physical connection its used and which protocol.

```
CREATE TABLE CT_RSP
  (CT_RSP_NO_RSP         NUMBER         NOT NULL -- PK: Sequence
    SEQ_CT_CONTROLE
  ,CT_RSP_CODE_SUPPORT   VARCHAR2(6)    NOT NULL -- Support de communication
     (DI-SCR)
  ,CT_RSP_CODE_PROTOCOLE VARCHAR2(6)    NOT NULL -- Protocole de
    communication (DI-PCR)
```

```
  ,CT_RSP_ADR_MIN        NUMBER                    -- Num ro d'adresse min
  ,CT_RSP_ADR_MAX        NUMBER                    -- Num ro d'adresse max
  ,CT_RSP_NB_MAX_ACQ     NUMBER                    -- Nb max de position
    fonctionnelle de classe acquisition
  ,CT_RSP_DAT_CRE        DATE           NOT NULL -- Date de creation
  ,CT_RSP_NO_ACT_CRE     NUMBER         NOT NULL -- FK1: NumerO acteur de
    creation
  ,CT_RSP_DAT_MOD        DATE                      -- Date de modification
  ,CT_RSP_NO_ACT_MOD     NUMBER                    -- FK2: NumerO acteur de
    modif.


ALTER TABLE CT_RSP ADD
   CONSTRAINT PK_CT_RSP PRIMARY KEY (CT_RSP_NO_RSP)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_RSP ADD
   CONSTRAINT UQ_CT_RSP_1 UNIQUE (CT_RSP_CODE_SUPPORT, CT_RSP_CODE_PROTOCOLE
     )
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_RSP ADD
   CONSTRAINT FK_CT_RSP_1 FOREIGN KEY (CT_RSP_NO_ACT_CRE)
                 REFERENCES SYS_USERS (USER_NO);

ALTER TABLE CT_RSP ADD
   CONSTRAINT FK_CT_RSP_2 FOREIGN KEY (CT_RSP_NO_ACT_MOD)
                 REFERENCES SYS_USERS (USER_NO);
```

CT_POR is a table used to declare each communication port and their settings used by an RTU.

```
CREATE TABLE CT_POR
  (CT_POR_NO_POR         NUMBER         NOT NULL -- PK: Sequence
    SEQ_CT_CONTROLE
  ,CT_POR_CT_PF          VARCHAR2(19)   NOT NULL -- Concentrateur
  ,CT_POR_PORT           NUMBER(4)      NOT NULL -- Num ro du port
  ,CT_POR_NO_RSP         NUMBER                    -- FK: Relation support -
    protocole
  ,CT_POR_VITESSE        VARCHAR2(4)              -- Vitesse de communication
    (DI-VCR) {...,96=9600,192=19200,...}
  ,CT_POR_DATABIT        VARCHAR2(2)               -- Databit (DI-DBI) {7,8}
  ,CT_POR_PARITE         VARCHAR2(1)               -- PARITE (DI-PAR) {P=PAIR,
     I=IMPAIR, S=SANS}
  ,CT_POR_DAT_CRE        DATE           NOT NULL -- Date de creation
  ,CT_POR_NO_ACT_CRE     NUMBER         NOT NULL -- FK1: NumerO acteur de
    creation
```

```
  ,CT_POR_DAT_MOD        DATE                    -- Date de modification
  ,CT_POR_NO_ACT_MOD     NUMBER                  -- FK2: Numer0 acteur de
    modif.
  );

ALTER TABLE CT_POR ADD
   CONSTRAINT PK_CT_POR PRIMARY KEY (CT_POR_NO_POR)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_POR ADD
   CONSTRAINT UQ_CT_POR_1 UNIQUE (CT_POR_CT_PF, CT_POR_PORT)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_POR ADD
CONSTRAINT FK_CT_POR_CT_RSP FOREIGN KEY (CT_POR_NO_RSP)
                REFERENCES CT_RSP (CT_RSP_NO_RSP);

ALTER TABLE CT_POR ADD
CONSTRAINT FK_CT_POR_1 FOREIGN KEY (CT_POR_NO_ACT_CRE)
                 REFERENCES SYS_USERS (USER_NO);

ALTER TABLE CT_POR ADD
CONSTRAINT FK_CT_POR_2 FOREIGN KEY (CT_POR_NO_ACT_MOD)
                 REFERENCES SYS_USERS (USER_NO);

CREATE INDEX ID_CT_POR_1 ON CT_POR (CT_POR_NO_RSP)
TABLESPACE INDX;
```

CT_RPA is a table used to define acquisition units their type and address.

```
CREATE TABLE CT_RPA
  (CT_RPA_NO_RPA         NUMBER         NOT NULL -- PK: Sequence
    SEQ_CT_CONTROLE
  ,CT_RPA_NO_POR         NUMBER         NOT NULL -- FK: Port de
    concentrateur
  ,CT_RPA_AQ_PF          VARCHAR2(19)   NOT NULL -- Acquisition
  ,CT_RPA_NO_SMA         NUMBER         NOT NULL -- FK: Scheme de l'
    acquisition  - Subtype = 'U'
  ,CT_RPA_ADRESSE        NUMBER(3)               -- Adresse de la PF
    acquisition sur le bus
  ,CT_RPA_FAMILLE        VARCHAR2(6)             -- Familly (rseau)
  ,CT_RPA_MEMBRE         VARCHAR2(6)             -- Member  (Accs)
  ,CT_RPA_DON_TECH1      VARCHAR2(6)             -- Donne technique 1
  ,CT_RPA_DON_TECH2      VARCHAR2(6)             -- Donne technique 2
  ,CT_RPA_DAT_CRE        DATE           NOT NULL -- Date de creation
```

```
  ,CT_RPA_NO_ACT_CRE     NUMBER         NOT NULL -- FK1: NumerO acteur de
    creation
  ,CT_RPA_DAT_MOD        DATE                    -- Date de modification
  ,CT_RPA_NO_ACT_MOD     NUMBER                  -- FK2: NumerO acteur de
    modif.


ALTER TABLE CT_RPA ADD
   CONSTRAINT PK_CT_RPA PRIMARY KEY (CT_RPA_NO_RPA)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_RPA ADD
   CONSTRAINT UQ_CT_RPA_1 UNIQUE (CT_RPA_NO_POR, CT_RPA_AQ_PF)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_RPA ADD
   CONSTRAINT FK_CT_RPA_CT_POR FOREIGN KEY (CT_RPA_NO_POR)
                REFERENCES CT_POR (CT_POR_NO_POR);

ALTER TABLE CT_RPA ADD
   CONSTRAINT FK_CT_RPA_CT_SMA FOREIGN KEY (CT_RPA_NO_SMA)
                REFERENCES CT_SMA (CT_SMA_NO_SMA);

ALTER TABLE CT_RPA ADD
   CONSTRAINT FK_CT_RPA_1 FOREIGN KEY (CT_RPA_NO_ACT_CRE)
                 REFERENCES SYS_USERS (USER_NO);

ALTER TABLE CT_RPA ADD
   CONSTRAINT FK_CT_RPA_2 FOREIGN KEY (CT_RPA_NO_ACT_MOD)
                 REFERENCES SYS_USERS (USER_NO);

CREATE INDEX ID_CT_RPA_1 ON CT_RPA (CT_RPA_AQ_PF)
TABLESPACE INDX;

CREATE INDEX ID_CT_RPA_2 ON CT_RPA (CT_RPA_NO_SMA)
TABLESPACE INDX;
```

CT_RAO is a table used to define electrical equipments and their control network parameters.

```
CREATE TABLE CT_RAO
  (CT_RAO_NO_RAO         NUMBER         NOT NULL -- PK: Sequence
    SEQ_CT_CONTROLE
  ,CT_RAO_NO_RPA         NUMBER         DEFAULT 1000 NOT NULL -- FK:
    Acquisition
  ,CT_RAO_PF_PF          VARCHAR2(19)   NOT NULL -- Position fonctionnelle a
     surveiller
```

```
  ,CT_RAO_NO_SMA          NUMBER         NOT NULL -- FK: Scheme de la
    position fonctionnelle - Subtype = 'P'
  ,CT_RAO_ADRESSE         NUMBER(3)               -- Adresse de la PF
    acquisition sur le bus
  ,CT_RAO_FAMILLE         VARCHAR2(6)             -- Familly (r seau)
  ,CT_RAO_MEMBRE          VARCHAR2(6)             -- Member  (Acc s)
  ,CT_RAO_DON_TECH1       VARCHAR2(6)             -- Donn e technique 1
  ,CT_RAO_DON_TECH2       VARCHAR2(6)             -- Donn e technique 2
  ,CT_RAO_DON_TECH3       VARCHAR2(6)             -- Donn e technique 3
  ,CT_RAO_DAT_CRE         DATE           NOT NULL -- Date de creation
  ,CT_RAO_NO_ACT_CRE      NUMBER         NOT NULL -- FK1: NumerO acteur de
    creation
  ,CT_RAO_DAT_MOD         DATE                    -- Date de modification
  ,CT_RAO_NO_ACT_MOD      NUMBER                  -- FK2: NumerO acteur de
    modif.
;

ALTER TABLE CT_RAO ADD
   CONSTRAINT PK_CT_RAO PRIMARY KEY (CT_RAO_NO_RAO)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_RAO ADD
   CONSTRAINT UQ_CT_RAO_1 UNIQUE (CT_RAO_NO_RPA, CT_RAO_PF_PF)
   USING INDEX TABLESPACE INDX;

ALTER TABLE CT_RAO ADD
   CONSTRAINT FK_CT_RAO_CT_RPA FOREIGN KEY (CT_RAO_NO_RPA)
              REFERENCES CT_RPA (CT_RPA_NO_RPA);

ALTER TABLE CT_RAO ADD
   CONSTRAINT FK_CT_RAO_CT_SMA FOREIGN KEY (CT_RAO_NO_SMA)
              REFERENCES CT_SMA (CT_SMA_NO_SMA);

ALTER TABLE CT_RAO ADD
   CONSTRAINT FK_CT_RAO_1 FOREIGN KEY (CT_RAO_NO_ACT_CRE)
               REFERENCES SYS_USERS (USER_NO);

ALTER TABLE CT_RAO ADD
   CONSTRAINT FK_CT_RAO_2 FOREIGN KEY (CT_RAO_NO_ACT_MOD)
               REFERENCES SYS_USERS (USER_NO);


CREATE INDEX ID_CT_RAO_1 ON CT_RAO (CT_RAO_PF_PF)
TABLESPACE INDX;

CREATE INDEX ID_CT_RAO_2 ON CT_RAO (CT_RAO_NO_SMA)
TABLESPACE INDX;
```

3.4.4 Relationship diagram

This is the relationship diagram of the new added tables:

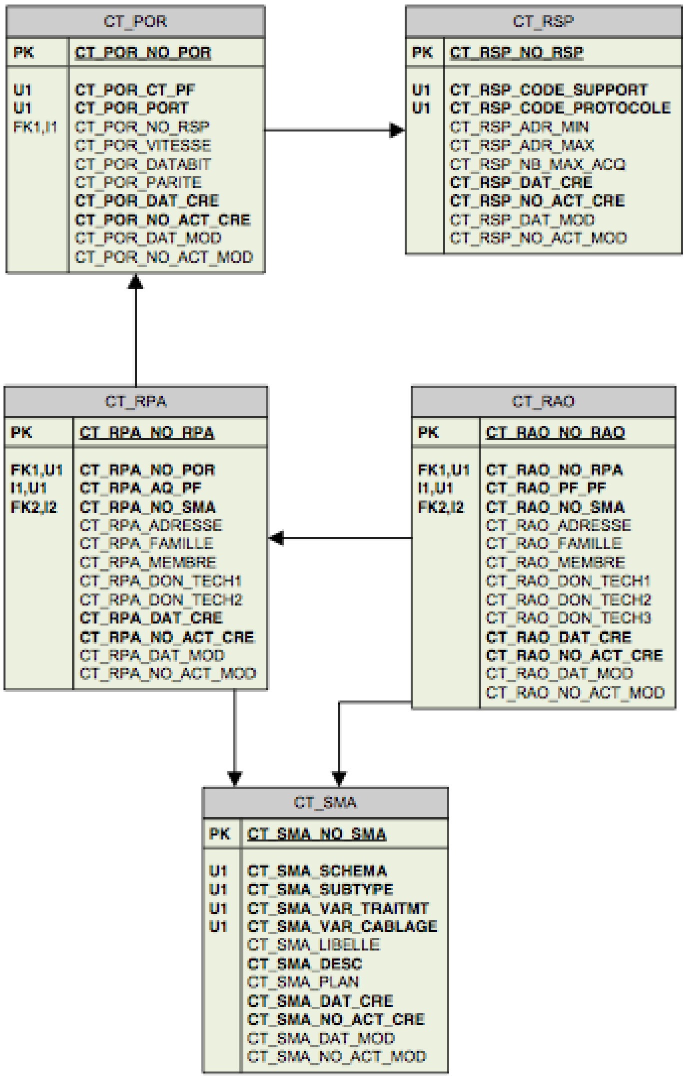

Figure 3.3: Relationship diagram of the new tables in GESMAR

Chapter 4

Data Integration

4.1 General considerations

The aim of data integration is to combine selected systems in order to for a unified one and give to the users the feeling of interact with only one information system. Integration of multiple data sources is an important issue, different information systems interconnect their operations and demand more from database systems; the integration process is hard because structural and semantic conflicts do exist.

Taking in consideration an integration process of multiple heterogeneous sources, most of the times its impossible to create a single homogeneous information system with a single scheme as suggested by a scheme integration approach [2], its often used a data integration approach [3] where integration is carried out by transforming queries or data. To perform this task, some sort of mappings between information systems is required. If this process is entirely carried out on the level of queries and scheme, its usually called query rewriting, while data transformation is used for data transformed and fused in an integration scheme.

Its possible to distingue two different categories of approaches for data integration, those that integrate against a single global scheme (global as view) and those dont (local as view). These approaches widely differ in their strengths, weaknesses and complexity.

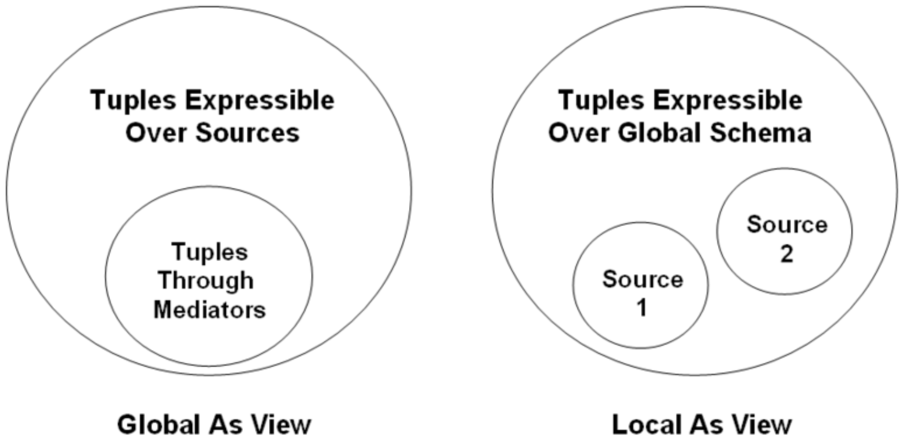

Figure 4.1: Global as view and local as view diagrams

Its clear that to solve integration problems there is not a single way to proceed; even if the goal is the same, the particular task may depend upon the architectural structure, the content and the functionalities of sources system, the level of autonomy, performances, the available resources. Other aspects of heterogeneity have to be considered, for example: different hardware, operating system or database management software, different business rules and integrity constraints.

Some examples of integration approach are:

- **Mediated query systems** form a unified data access providing a single point to query data sources. Mediators use a global query processor that allows sending subqueries to local data sources, the returned dataset are then combined together.

- **Data warehouses** use a unified point for accessing data. Data coming from sources [typically from on line transaction processing system (OLTP)] are extracted, transformed and loaded (ETL) into a data warehouse. When the information are consolidated in a single point its possible to analyze them using on line analytical processing (OLAP).

- **Federated database systems** (FDBMS) are based on logical integration of data sources to constitute a real DBMS that implements its own data model, supporting global queries, transactions and access control.

- **Workflow management systems** (WFMS) create business processes where each step is managed by a different application. WMFS are considered as integration by application approach and can have support for modeling, executing and maintaining of the process used for the integration.

- **Integration by web services** use software components that support interaction between machines, most of the times using XML based messages over network. It can be used to represent uniform data access or just as interface for subsequent integration.

- **Semantic integrations** are new techniques used to overcome the limitation of structural integration applied to semi- and unstructured data where not only data models are different but the semantic is heterogeneous. An ideal approach could be to exhaustively define the real world semantic of all scheme, but the problem is that semantics is embodied in data models, schemes, applications, contents and in the mind of users. For all these reasons, data integration is a much open challenge.

4.2 Data Integration for electrical network control hierarchy

The CERNs Engineering Electrical Control section (EN/EL/CO) needs to have an overview of the hierarchy control system which should be accessible by querying a single system (the data warehouse GESMAR) while focusing on the technical constrains. Since the RTUs databases are homogeneous sources of information and there is full access to the source data, this process has to be execute only once. Therefore, the ETL process seems to be a reasonable approach to solve this data integration case.
An ETL process consists in three steps processes that extract information from sources; transform them as needed corresponding to the correct type on the destination table, and load the verified data in a new table. In this way its possible obtain a single scheme of query in the data warehouse. There are several tools commercial and open source that can help in defining a ETL process, but since none of them is supported at CERN, I have developed a software ad-hoc using SQL and VBA.

4.3 Extraction

4.3.1 Extraction method

Knowing which are the information needed in GESMAR and knowing where they are stored in the RTU databases, the extraction procedure consists in a cascade of queries that joins different tables creating a unified tables with all pertinent information. This is initially done querying from an Access database a RTU Access database saved on the local hard disk; then an automatic batch program has been written which retrieve all the databases and save them on local disk.
For each database starting from the main table in the first tier tCNF_SerialPort

and following the relationships with others tables it possible to build a consistent view in one single table.

4.3.2 Extraction testing

When all the queries for a single database have been written, it was possible to test them against other databases. Due to the homogeneity in the databases, the queries has been tested on particular databases; for example the oldest one, the newest one, the largest and the smallest ones. To accomplish this, an ad hoc software in VBA has been developed in order to automate this procedure and collect information about number of data extracted, error count, time elapsed.
The software consists of four modules, each performing a specific task:

SP_CH_DE collects information regarding serial ports, communication channels, protocols and devices and store them in a temporary table.

CD_CO_ME_DI collects information regarding all the kind of data-points: counters, controls, analogs and digitals.

HI_DE_DP collects information about the acquisition used to acquire information from the device and join the two previous temporary tables.

EXTRA_FIELDS collects additional information for analogs and digitals not used in the destination data warehouse but considered useful for future analysis.

4.3.3 Extraction errors

At the beginning of this extraction process several anomalies have been encountered, which are related with inconsistent design of the original database and the bad implementation of the customization made for CERN. First of all, we have the lack of constrains on important fields such as hierarchy levels

SITE	SECTOR_POIN	LOCATION	FILE_NAME	DATE	VERSION	TEST	RESULT	ADAM	PATH
LHC	LHC1	RE12	CLP500E_RE12.mdb	15/09/09	??		NO	NO	https://dfs.cern.c
LHC	LHC1	RE18	CLP500E_RE18 v4.3.zip	14/09/09	4.3		NO	NO	https://dfs.cern.c
LHC	LHC1	SDX1	CLP500E_SDX1 v4.5.4.zip	16/09/09	4.5.4		NO	NO	https://dfs.cern.c
LHC	LHC1	SES1	CLP500E_SES1.mdb	18/09/09	?		NO	YES	https://dfs.cern.c
LHC	LHC1	US15	CLP500E_US15 v4.6.zip	17/09/09	4.6		NO	NO	https://dfs.cern.c
LHC	LHC1	USA15	CLP500E_1 v4.3.zip	17/09/09	4.3		NO	NO	https://dfs.cern.c
LHC	LHC1	USA15	CLP500E_2 v4.7.zip	24/11/08	4.7		NO	NO	https://dfs.cern.c
LHC	LHC18	SE18	CLP500_SM18_v4.7.zip	16/07/09	4.7	23/09/09	NO	YES	https://dfs.cern.c
LHC	LHC2	RE22	CLP500E_RE22 v4.3.zip	14/09/09	4.3		NO	NO	https://dfs.cern.c
LHC	LHC2	RE28	CLP500E_RE28 v4.3.zip	14/09/09	4.3		NO	NO	https://dfs.cern.c
LHC	LHC2	SES2	CLP500_LHC2.mdb	18/09/09	?	23/09/09	NO	YES	https://dfs.cern.c
LHC	LHC2	UW25	CLP500E_UW25.mdb	17/09/09	4.63		NO	NO	https://dfs.cern.c
LHC	LHC3	RE32	CLP500E_RE32 v4.3.zip	14/09/09	4.3		NO	NO	https://dfs.cern.c
LHC	LHC3	RE38	CLP500E_RE38_v4.2.zip	25/05/09	4.2		NO	NO	https://dfs.cern.c
LHC	LHC3	SES3	CLP500_LHC3 v5.1.2.zip	24/07/09	5.1.2		NO	YES	https://dfs.cern.c
LHC	LHC3	SZ33	CLP500_LHCZ33 v5.0.zip	27/07/09	5.0		NO	YES	https://dfs.cern.c
LHC	LHC3	UJ33	CLP500E_UJ33.mdb	5/8/2009	?		NO	NO	https://dfs.cern.c
LHC	LHC4	RE42	CLP500E_RE42_v4.2.zip	25/05/09	4.2		NO	NO	https://dfs.cern.c
LHC	LHC4	RE48	CLP500E_RE48 v4.2.zip	2/7/2009	4.2		NO	NO	https://dfs.cern.c
LHC	LHC4	SES4	CLP500_LHC v5.1.6.zip	22/07/09	5.1.6		NO	YES	https://dfs.cern.c
LHC	LHC4	US45	CLP500E_US45 v4.2.zip	09/06/09	4.2		NO	NO	https://dfs.cern.c
LHC	LHC5	RE52	CLP500E_RE52 v4.1.zip	2/6/2009	4.1		NO	NO	https://dfs.cern.c
LHC	LHC5	RE58	CLP500E_RE58 v4.1.zip	16/09/09	4.1		NO	NO	https://dfs.cern.c
LHC	LHC5	SES5	CLP500_LHC5.mdb	16/09/09	?		NO	YES	https://dfs.cern.c
LHC	LHC5	UJ56	CLP500E_UJ56 v4.4.zip	5/6/2009	4.4		NO	NO	https://dfs.cern.c
LHC	LHC5	USC55-1	CLP500E_USC55-1.mdb	31/08/09	?		NO	NO	https://dfs.cern.c
LHC	LHC5	USC55-2	CLP500_USC55-2_V4.41.zip	10/8/2009	4.41		NO	NO	https://dfs.cern.c
LHC	LHC6	RE62	CLP500E_RE62 v4.1.zip	16/06/09	4.1		NO	NO	https://dfs.cern.c
LHC	LHC6	RE68	CLP500E_RE68 v4.1.zip	22/06/09	4.1		NO	NO	https://dfs.cern.c
LHC	LHC6	SES6	CLP500_LHC6.mdb	17/09/09	?		NO	YES	https://dfs.cern.c
LHC	LHC6	US65	CLP500_US65_v4.0.zip	6/7/2009	4.0		NO	NO	https://dfs.cern.c
LHC	LHC7	RE72	CLP500E_RE72 v4.3.zip	14/09/09	4.3		NO	NO	https://dfs.cern.c
LHC	LHC7	RE78	CLP500E_RE78 v4.2.zip	6/7/2009	4.2		NO	NO	https://dfs.cern.c
LHC	LHC7	SES7	CLP500_LHC7 v5.1.zip	12/8/2009	?		NO	YES	https://dfs.cern.c
LHC	LHC7	UJ76	CLP50E_UJ76.mdb	16/06/09	?		NO	NO	https://dfs.cern.c
LHC	LHC8	RE82	CLP500E_RE82 v4.2.zip	6/7/2009	4.2		NO	NO	https://dfs.cern.c
LHC	LHC8	RE88	CLP500E_RE88 v4.2.zip	7/7/2009	4.2		NO	NO	https://dfs.cern.c
LHC	LHC8	SES8	CLP500_LHC8.mdb	8/9/2009	?	24/09/09	NO	YES	https://dfs.cern.c
LHC	LHC8	US85-1	CLP500E_US85_1 v4.2.zip	15/12/08	4.2	23/09/09	NO	NO	https://dfs.cern.c
LHC	LHC8	US85-2	CLP500E_US85_2 v4.2.zip	7/7/2009	4.2		NO	NO	https://dfs.cern.c
LHC	LHC8	UX85	CLP500E_UX85_V4.27.zip	11/8/2009	4.27		NO	NO	https://dfs.cern.c
MEYRIN	MEYRIN	B54	CLP500_Labo_Sepam80.mdb	24/04/09	?		NO	NO	https://dfs.cern.c
MEYRIN	MEYRIN	Labo54	CPL500_Labo_Sepam80.mdb	7/4/2009	?	22/09/09	NO	NO	https://dfs.cern.c
MEYRIN	MEYRIN	ME04					NO		
MEYRIN	MEYRIN	ME06	CLP500_ME6 v4.6.zip	14/08/09	4.6		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME09	CLP500_ME9_v4.6.zip	30/06/09	4.6		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME09B	CLP500_ME9B_V4.2.mdb	28/08/09	4.2		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME11	CLP500_ME11.mdb	17/09/09	?		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME23	CLP500_ME23.mdb	2/7/2009	?		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME25	CLP500_ME25_v4.2.zip	29/06/09	4.2		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME43	CLP500_ME43.mdb	17/09/09	?		NO	YES	https://dfs.cern.c
MEYRIN	MEYRIN	ME59					NO		
PREVESSI	PREVESSIN	BA80					NO		
PREVESSI	PREVESSIN	BA81					NO		
PREVESSI	PREVESSIN	BA82					NO		
PREVESSI	PREVESSIN	BE	CLP500_BE V5.36.zip	21/08/09	5.36		NO	YES	https://dfs.cern.c
PREVESSI	PREVESSIN	BE9	CLP500_BE9 v4.6.5.zip	10/8/2009	4.6.5		NO	YES	https://dfs.cern.c
PREVESSI	PREVESSIN	BEM9	CLP500_BEM9.mdb	31/08/09	?		NO	NO	https://dfs.cern.c
SPS	SPS	BA1	CLP500_BA1 v5.2.zip	12/6/2009	5.2		NO	YES	https://dfs.cern.c
SPS	SPS	BA2	CLP500_BA2_v5.6.zip	16/07/09	5.6		NO	YES	https://dfs.cern.c
SPS	SPS	BA3	CLP500_BA3.mdb	3/7/2009	?		NO	YES	https://dfs.cern.c
SPS	SPS	BA4	CLP500_BA4.mdb	18/09/09	?		NO	YES	https://dfs.cern.c
SPS	SPS	BA5	CLP500_SPS_BA5 v5.3.1.zip	24/06/09	5.3.1		NO	YES	https://dfs.cern.c
SPS	SPS	BA6	CLP500_BA6.mdb	1/9/2009	?		NO	YES	https://dfs.cern.c
SPS	SPS	BA7	CLP500_BA7.mdb	1/9/2009	?		NO	YES	https://dfs.cern.c

Figure 4.2: A spreadsheet file generated to track the testing results.

or device name, second the lack of a user interface which validates the user input or at least help to uniform them.

In particular, the anomalies can be caused by:

- **Partial updates**: when the same information is stored in more than a table, one had to manually update all the occurrences, which are seldom known
- **Insertions**: it is sometimes impossible to have all the information needed for a particular row (for example when an electrical equipment is inserted before a physical installation).
- **Deletions**: when deleting a particular row containing information about an instance, we should avoid deleting information about related instances.
- **Freshness**: an old information, for example that of a dismantled equipment, should be properly and timely replaced

Some errors found in RTU database were conflicts caused by duplicated datapoints with different settings (like protocol or channel used), old information about replaced systems (like SAINCO emergency system), use of non conform name for hierarchy level three (acquisition); the most relevant error discovered was the mistyping of O instead of 0 in a collection of datapoints causing strange behavior in the SCADA system.

4.3.4 Extraction optimization

Before execute the same queries on all databases to collect information, some optimization principles have been applied to the sql codes and VBA code to obtain a faster extraction process.

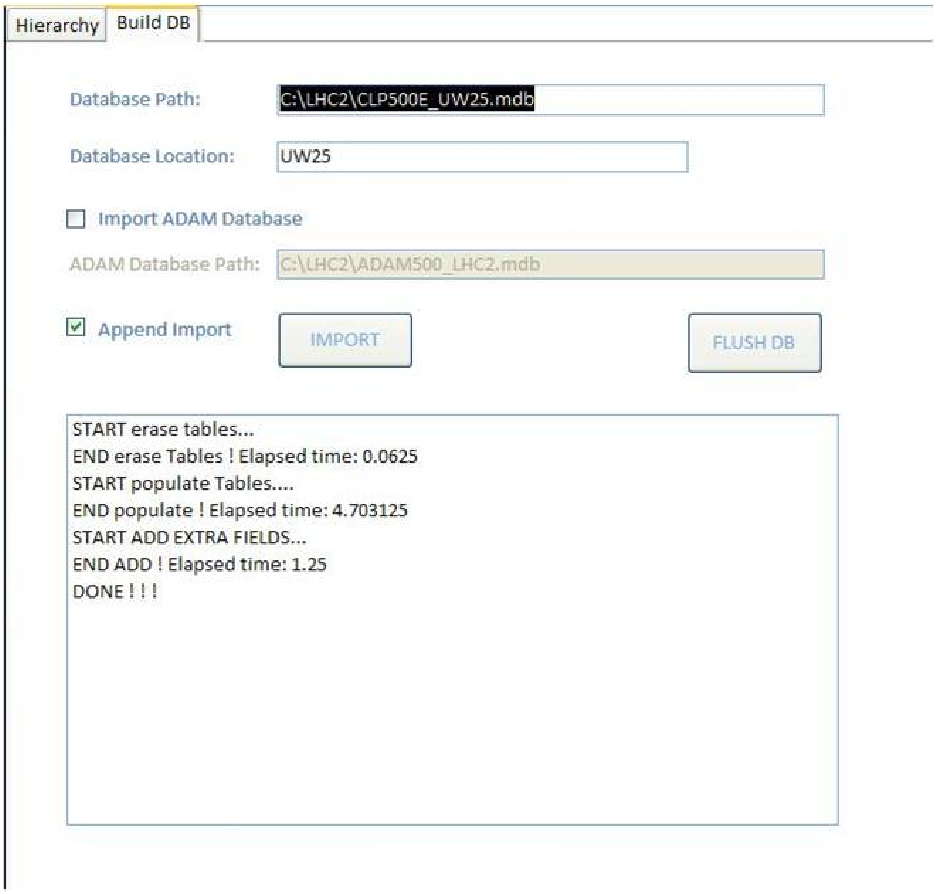

Figure 4.3: VBA interface for extraction of data from RTU database

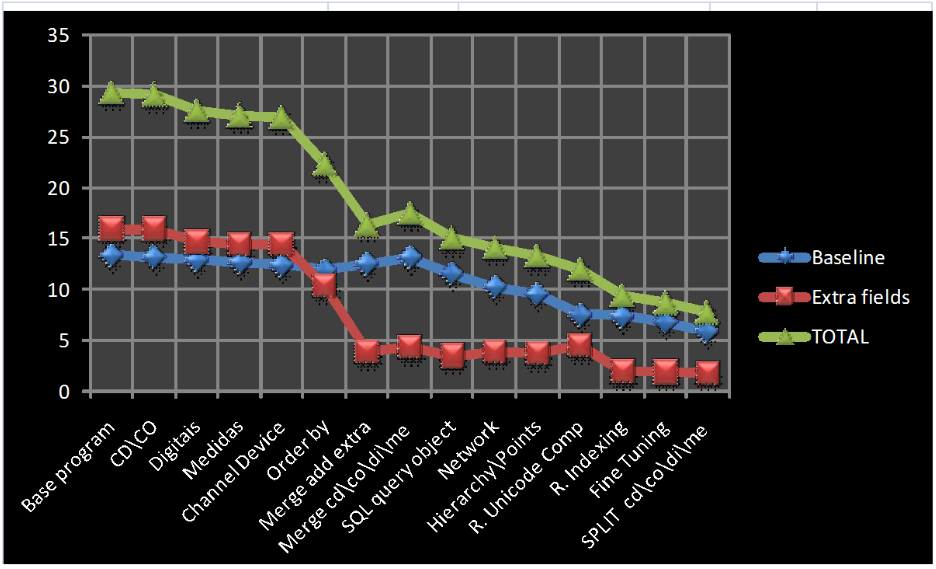

Figure 4.4: Optimization steps

From the image the is possible to see that the total time of 30s for database extraction has been reduced to 8s; mainly due to the removing of Order by command and Unicode Compression, while from theVBA point of view the improvements come from rearranging the source code in string composition and using a faster query object.

4.3.5 Extraction results

After extraction from all the RTUs database a single table of 180000 rows with all the hierarchy information from the datapoint to the RTU has been created. This is the starting point for the next step. To quickly verify the extraction result the software developed allows navigate through the hierarchy displaying related information.

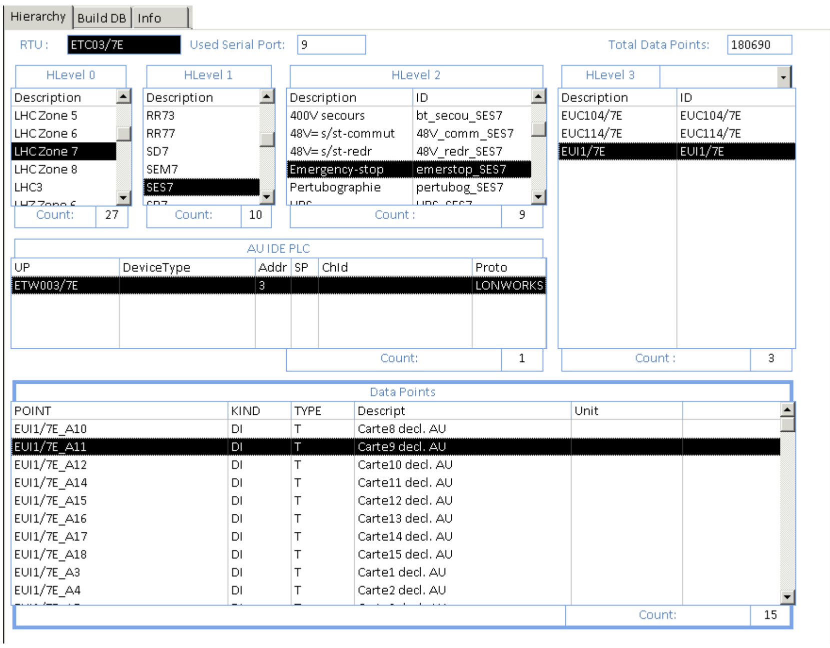

Figure 4.5: VBA interface for navigate in the extracted data from RTU database

4.4 Transformation

Due to the anomalies founded in the extraction process the table containing 180000 rows presents a number of anomalies that have to be fixed or dropped before load them in the data warehouse. Data have to be carefully cleaned and verified before; the idea is to extract three tables from the big table respectively for CT_POR, CT_RPA, CT_RAO.
The destination tables have the same field properties in order to verify that data type are correct.
Using auxiliary mapping tables its possible to convert values from the RTU conventions to the GESMAR conventions; for example a speed marked as 1200 in RTU, becomes 12 in the GESMAR database since a different dictionary is used.

DI-VCR	
speed_rtu	**speed_gesmar**
1200	12
125000	1250
19200	192
1MB	1MB
2400	24
-27136	384
4800	48
9600	96

Figure 4.6: Mapping table for DI-VCR dictionary

Other more complex mappings tables have been created manually to combine information with data that are not present in RTUs database, for example the host name and the building number:

A total number of nine mappings tables have been created and used during the transformation process.

AUX_CONCENTRATOR			
concentrator	host	building	description
ETC04*9	***09B	212	RTU JURA DIESELS
ETC03*9	***09	212	RTU MEYRIN JURA
ETC03*23	***23	214	RTU MEYRIN ME23
ETC03*11	***11	251	RTU MEYRIN PS
ETC03*25	***25	271	RTU MEYRIN BOOSTER

Figure 4.7: Mapping table for concentrators

Due to some errors using the codification of equipments from the GESMAR domain to the RTU domain, manual fixes have been done, a report file of all the changes is present on the documents management system of CERN [4] .

4.5 Loading

Since the destination data warehouse is based on Oracle and the extraction process has been done in Access, the transformation procedure store its results in Oracle using a technique available in Access called linked tables, this allows a database Access to have read and write access to data in a remote Oracle database [5] Thanks to this, during the transformation data are preloaded in temporary tables TEMP_CT_POR, TEMP_CT_RPA and TEMP_CT_RAO similar to the real one created in the data warehouse, but without indexing and constrains. Its possible to use the more powerful Oracle environment for the final loading.

TEMP_CT_POR					
CT_PF	PORT	NO_RSP	VITESSE	DATABIT	PARITE
ETC03*11	10	1021	48	8	S
ETC03*11	1000	1023	1MB		
ETC03*11	11	1021	96	8	S
ETC03*11	12	1021	192	8	S

Figure 4.8: Extraction of the TEMP_CT_POR table containing 476 rows

TEMP_CT_RPA			
NO_POR	AQ_PF	NO_SMA	ADRESSE
1561	ETZ507*43	1690	7
3706	ETZ6/8E	1690	6
3706	ETZ13/8X	1690	13
3706	ETZ9/8E	1690	9
1561	ETZ509*43	1690	9

Figure 4.9: Extraction of the TEMP_CT_RPA table containing 2034 rows

TEMP_CT_RAO			
NO_RAO	NO_RPA	PF_PF	NO_SMA
17079	4840	EBD102/73	5473
17080	4843	EBD102/76	5473
17081	4840	EBD102/77	5473
17082	4830	EBD102/78	5473

Figure 4.10: Extraction of the TEMP_CT_RAO table containing 10513 rows

4.5.1 Loading CT_POR sql code

Using the SQL Plus console the following script has been executed, it checks constrains and for already existing data and creates a report of the insert/update result.

```
SET TERMOUT ON ECHO OFF
SPOOL CT_POR.TXT
PROMPT "Nb ENREGISTREMENTS"
SELECT count(*)
FROM   GESMAR_PUB.TEMP_CT_POR A
;

PROMPT "Duplication UQ"
SELECT count(count(*))
FROM   GESMAR_PUB.TEMP_CT_POR A
GROUP BY
 A.CT_POR_CT_PF
,A.CT_POR_PORT
HAVING COUNT(*) > 1
;

PROMPT "UQ existe  d j "
SELECT
 COUNT(*)
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE EXISTS (SELECT NULL
              FROM CT_POR B
              WHERE B.CT_POR_CT_PF = A.CT_POR_CT_PF
              AND   B.CT_POR_PORT  = A.CT_POR_PORT);

PROMPT "UQ gesmar n'existe pas (temp)"
SELECT
```

```
 B.CT_POR_CT_PF
,B.CT_POR_PORT
FROM CT_POR B
WHERE NOT EXISTS (SELECT NULL
              FROM GESMAR_PUB.TEMP_CT_POR A
              WHERE A.CT_POR_CT_PF = B.CT_POR_CT_PF
              AND   A.CT_POR_PORT  = B.CT_POR_PORT);

PROMPT "Verification PF"
SELECT A.CT_POR_CT_PF
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE NOT EXISTS (SELECT NULL FROM OBVOBJ B
                  WHERE  B.OBVOBJ_CODE = A.CT_POR_CT_PF
                  AND    B.OBVOBJ_OBTYPE = 'P')
;

PROMPT "VITESSE n'existe pas"
SELECT
 DISTINCT A.CT_POR_VITESSE
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE NOT EXISTS (SELECT NULL FROM SYS_LOOKUPS B
                  WHERE B.LOOK_TABLE = 'DI-VCR'
                  AND   B.LOOK_CODE  = A.CT_POR_VITESSE)
;

PROMPT "DATABIT n'existe pas"
SELECT
 DISTINCT A.CT_POR_DATABIT
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE NOT EXISTS (SELECT NULL FROM SYS_LOOKUPS B
                  WHERE B.LOOK_TABLE = 'DI-DBI'
                  AND   B.LOOK_CODE  = A.CT_POR_DATABIT)
;


PROMPT "PARITE n'existe pas"
SELECT
 DISTINCT A.CT_POR_PARITE
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE NOT EXISTS (SELECT NULL FROM SYS_LOOKUPS B
                  WHERE B.LOOK_TABLE = 'DI-PAR'
                  AND   B.LOOK_CODE  = A.CT_POR_PARITE)
;

PROMPT "Relation support - protocole N'EXISTE PAS"
SELECT
 A.CT_POR_CT_PF
```

```
,A.CT_POR_PORT
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE NOT EXISTS (SELECT NULL FROM CT_RSP B
                  WHERE B.CT_RSP_NO_RSP = A.CT_POR_NO_RSP)
;

PROMPT "UPD CT_POR existants"
UPDATE CT_POR B SET
 (CT_POR_NO_RSP
 ,CT_POR_VITESSE
 ,CT_POR_DATABIT
 ,CT_POR_PARITE) =
(SELECT
 A.CT_POR_NO_RSP
,A.CT_POR_VITESSE
,nvl(A.CT_POR_DATABIT,'*')
,nvl(A.CT_POR_PARITE,'*')
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE A.CT_POR_CT_PF = B.CT_POR_CT_PF
AND   A.CT_POR_PORT  = B.CT_POR_PORT)
WHERE EXISTS (SELECT NULL
              FROM GESMAR_PUB.TEMP_CT_POR A
              WHERE A.CT_POR_CT_PF = B.CT_POR_CT_PF
              AND   A.CT_POR_PORT  = B.CT_POR_PORT)
;

INSERT into CT_POR
(CT_POR_NO_POR
,CT_POR_CT_PF
,CT_POR_PORT
,CT_POR_NO_RSP
,CT_POR_VITESSE
,CT_POR_DATABIT
,CT_POR_PARITE
,CT_POR_DAT_CRE
,CT_POR_NO_ACT_CRE
,CT_POR_DAT_MOD
,CT_POR_NO_ACT_MOD
)
select
 SEQ_CT_CONTROLE.NEXTVAL
,A.CT_POR_CT_PF
,A.CT_POR_PORT
,A.CT_POR_NO_RSP
,A.CT_POR_VITESSE
,NVL(A.CT_POR_DATABIT,'*')
,NVL(A.CT_POR_PARITE,'*')
```

```
,SYSDATE
,3048018
,NULL
,NULL
FROM GESMAR_PUB.TEMP_CT_POR A
WHERE NOT EXISTS (SELECT NULL
                  FROM CT_POR B
                  WHERE B.CT_POR_CT_PF = A.CT_POR_CT_PF
                  AND   B.CT_POR_PORT  = A.CT_POR_PORT);
SPOOL OFF
SET TERMOUT OFF ECHO OFF
```

4.5.2 Loading CT_POR sql result

"Nb ENREGISTREMENTS" 476

"Duplication UQ" 0

"UQ existe deja" 238

"UQ gesmar n'existe pas (temp)" 7 rows selected.

CT_POR_CT_PF CT_POR_PORT

————————- ————

ETC03/1E 23

ETC03/2E 24

ETC03/E9 13

ETC04*9 7

ETCXX/XX 1

ETC03/E18 1

ETC03/E9 19

"Verification PF" no rows selected

"VITESSE n'existe pas" no rows selected

"DATABIT n'existe pas" no

"PARITE n'existe pas" no

"Relation support - protocole N'EXISTE PAS" no rows selected

"UPD CT_POR existants" 238 rows updated.

238 rows created.

4.5.3 Loading CT_RPA results

A similar script has been executed for the acquisitions level. The report is the following:

"Nb ENREGISTREMENTS" 2034
"Duplication UQ" 0
"UQ existe deja" 1174
"UQ gesmar n'existe pas (temp)" 135
"Verification PF" no rows selected
"Schma inexistant" no rows selected
"UPD scheme et adresse CT_PRA existants" 1174 rows updated.
"TAG CT_PRA inexistants dans TEMP_CT_RPA (famille = 'OLD?') 135 rows updated.
"Insertions" 860 rows created.

4.5.4 Loading CT_RAO results

A similar script has been executed for the equipments level, this is the report:
"Nb ENREGISTREMENTS" 12437
"Duplication UQ" 0
"UQ existe deja" 0
"UQ gesmar n'existe pas (temp)" 443
"PF gesmar n'existe pas (temp)"441
"PF gesmar existe (temp)"
CT_RAO_PF_PF
————————- ESU201/E18
ESU101/E18

"Verification PF" 2365
"Schma inexistant" no rows selected
"Insertion DES 2 REFERENCES DEJA EXISTANTES" 2 rows created.
"Insertion" 10070 rows created.

4.5.5 Loading CT_PORT, CT_SMA

These tables have been manually created before the loading process, because they contain information that is not present in the RTU databases.

4.5.6 ETL result

After the loading step the hierarchy of the control system of the electrical network of CERN is stored in the data warehouse. In summary: 57 RTUs are using 476 channels with 19 different combination of support and protocol to communicate with 2169 acquisition units, intelligent electrical devices and PLC that control and monitor 10513 electrical devices.
Note: currently 2365 electrical devices are registered with a wrong name at the RTU level; at the moment they have been ignored pending a decision from technicians and management.

Chapter 5

Sample applications

5.1 Applications using new tables in the data warehouse

Data are useful when they can be used for analysis or in application, helping final user to accomplish decisions with a better knowledge, a set of sample programs have been developed in order to demonstrate some useful application using data from the datawarehouse.

5.1.1 Hierarchy navigation using Oracle form

This is the simplest way to access the data using an Oracle form and navigate through the hierarchy fig. 5.1.

5.1.2 Hierarchy visualization using dot language

Querying the data warehouse and exporting the result in a dot language file, its possible to create representations of the hierarchy; changing settings in the exported file can focus attention on different details fig. 5.2.

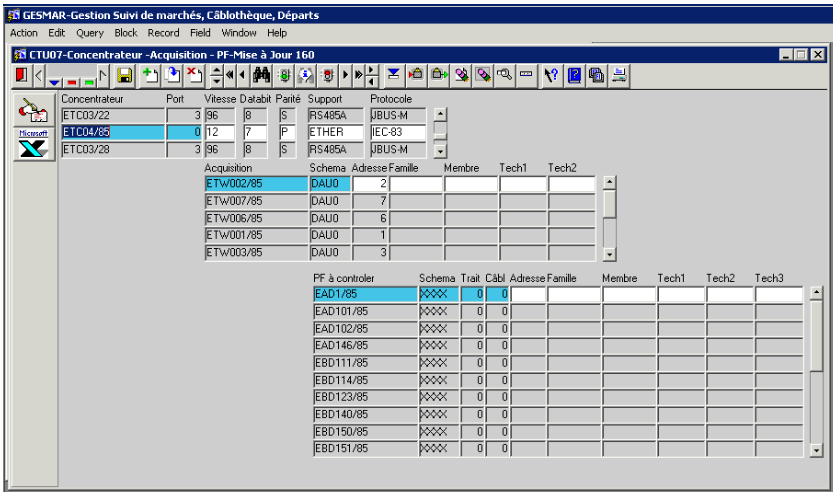

Figure 5.1: Sample application using Oracle forms

5.1.3 Hierarchy and current status using Geographic Information System

Combing data coming from the GESMAR data warehouse about the hierarchy and buildings position with the live data coming from the SCADA system representing status of devices, is possible to create a global overview of the control system in a GIS environment fig. 5.3.

5.1.4 Sample web application

The same information can be displayed on web pages using a web server and some tools like Google Visualization API (reference) fig. 5.4.

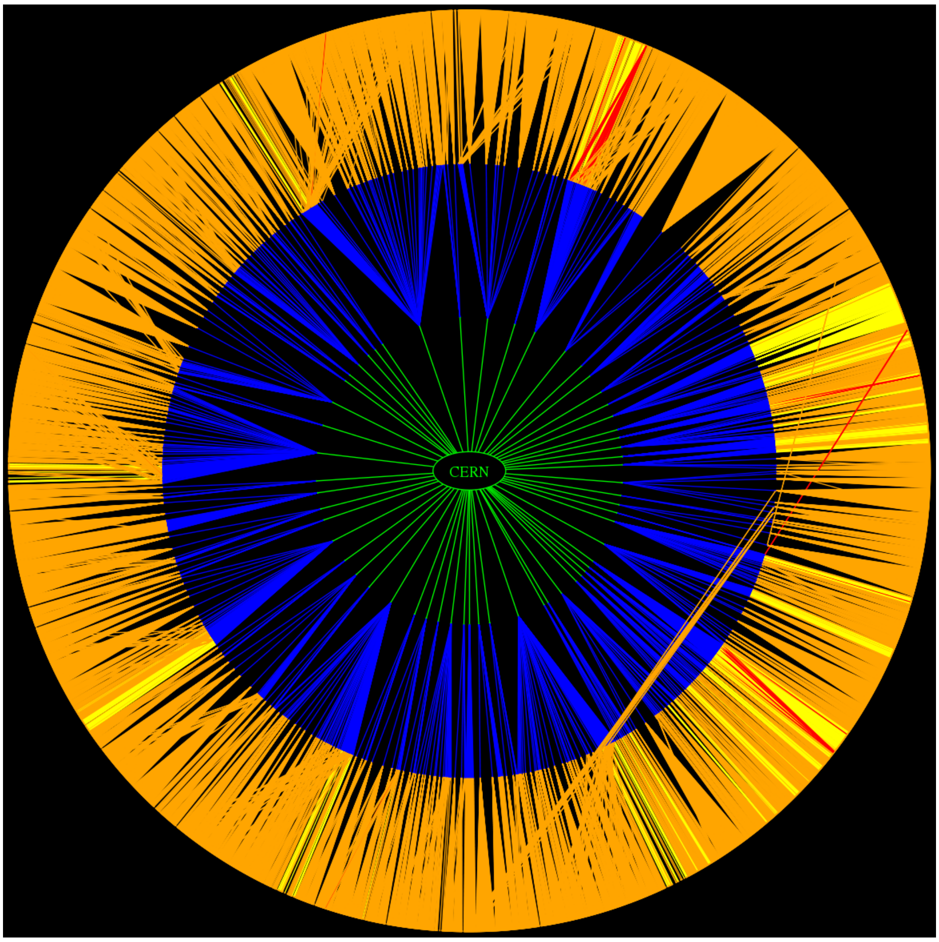

Figure 5.2: Sample application using Oracle forms

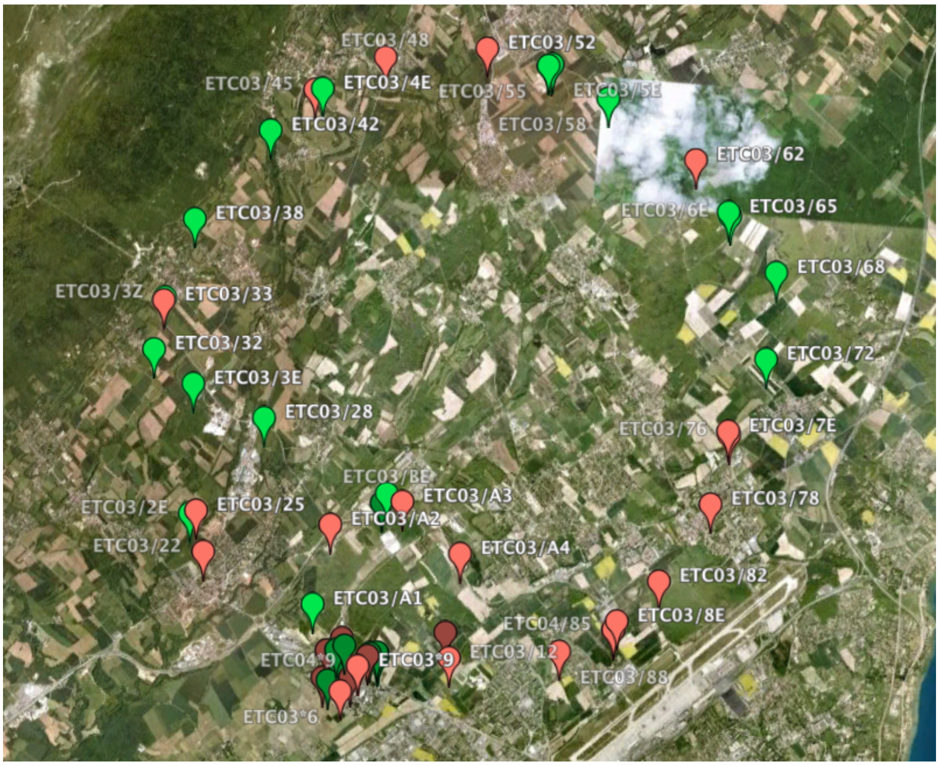

Figure 5.3: Sample application using a Geographic Information System

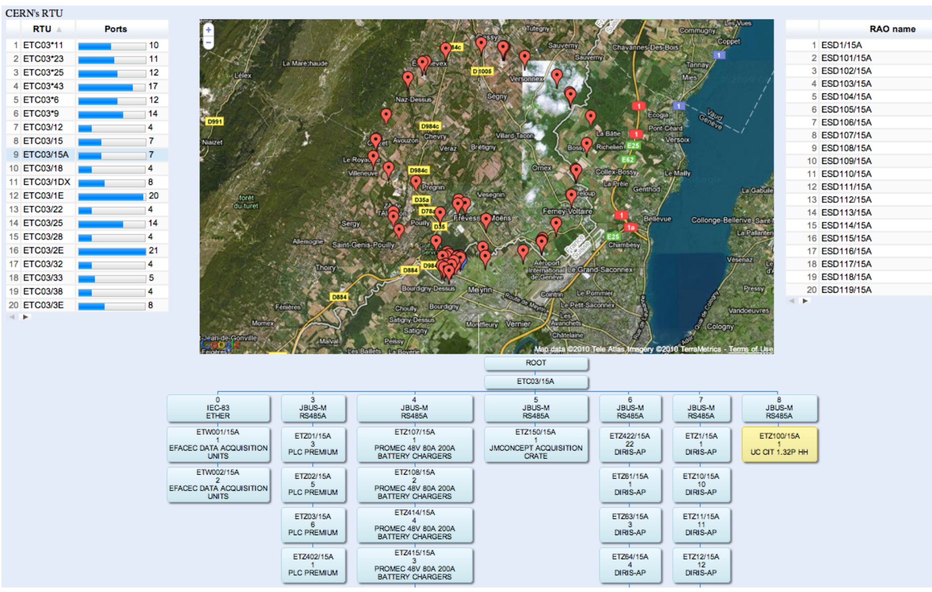

Figure 5.4: Sample web application using Google visualization API

Chapter 6

Final Considerations

6.1 Conclusion

Data integration is an important challenge not only to consolidate a legacy system, but also to expand capabilities and functionalities. In relation with integration of the electrical control network at CERN, the solution adopted by us is one of the most simple but the overall project was very complex. A lot of difficulties have been encountered due to the heterogenity of the system and the presence of legacy exceptions coming from the old control system, with same occurrences presented in different ways by different users. Thanks to a great team work of experts in the electrical domain at CERN most of the anomalies have been fixed and the project completed before the official first LHC collision day 30 March 2010 [6].

References

[1] Redman, T.C., Improve data quality for competitive advantage, Sloan Management Review, Winter, 1995, Vol. 36, No. 2, pp. 99-107.

[2] C. Batini and M. Lenzerini, A comparative Analysis of Methodologies for database Scheme integration 1986.

[3] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. Fundamentals of Data Warehouses. Springer-Verlag, 2000.

[4] Daniele De Ruschi, ELG-INFOSYS-QN-0001, 2010.

[5] Tom Erket, Oracle tables to Microsoft Access, 2000.

[6] "LHC First Physics" Cern.ch http://press.web.cern.ch/press/lhc-first-physics/ *Web source*, 2010.

General bibliography

- "CERN in a nutshell" Cern.ch. http://public.web.cern.ch/public/en/About/About-en.html *Web source*, 2008.
- Reva Bash's, Database Searcher, Measuring the Quality of the Data, 1990.
- Jeff Reagan, Decomposing Data Integration: Best Practices in Accelerating Pre-ETL Processes, International Journal of Cooperative Information Systems, 2001, Vol.10,No. 3 237-271

- D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi and R. Rosati, Data Integration in Data Warehousing, 2001
- Chrisoph Koch, Data Integration against Multiple Evolving Autonomous Schemata, 2001.