PROCEEDINGS OF SCIENCE

High Frequency Financial Data in Egrid

Alessandra Tedeschi*
CNR-INFM - Universita' di Roma La Sapienza
P.zza Aldo Moro 2, 00185 Rome, Italy
E-mail: alessandra.tedeschi@roma1.infn.it

We describe the type of financial data acquired in the framework of the project *Dynamics of ultra-high frequency data in financial markets* and their structure over the EGRID platform. We also illustrate the pre-processing that is being carried out on the data and the new data formats generated. The role of EGRID platform in solving different kind of problems in the treatment of financial data will be highlighted.

Grid Technology for Financial Modeling and Simulations
Palazzo Steri, Villa Zito, Palermo, Italy
February 3 – 4, 2006

* Speaker

PoS(GRID2006)016

1. Introduction

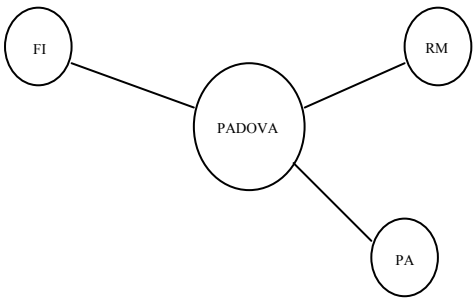
One of the main goals of the Italian research project *Dynamics of ultra-high frequency data in financial markets* is the setup of a national facility for the storage and management of high frequency financial data. This facility is based on grid technology, thanks to the collaboration with EGRID project. The EGRID platform proved to be very useful, helping to assess some common items of the treatment of large amount of data. In the following we will first introduce the research motivations and the organization of the project. We will then give a sketch of the grid implementation for our case and finally we will focus on the data acquired for the project, giving a brief description of the different datasets and describing the pre-processing procedures performed on the "original" data.

1.1 The Project and the Partners

The project *Dynamics of ultra-high frequency data in financial markets* is financed for 3 years (2003-2006) by the Italian Ministry for University and Research (MIUR) and is carried out by different research groups all over Italy: CNR – INFM Palermo, CNR – INFM Roma, CNR – INFM Trieste, Università del Piemonte Orientale, Università Politecnica delle Marche, Dipartimento di Statistica – Università di Firenze, CNR - IAC Roma.
The project main goals are the study of efficiency and structure in financial markets and the quantitative modelling of portfolio choices in the presence of technological innovations.

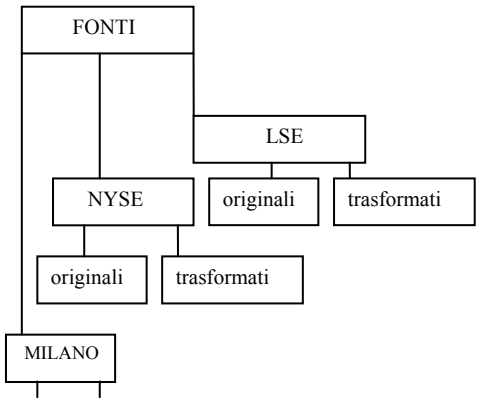
1.1 The EGRID structure

EGRID is organized in a "starred" architecture (see figure 1). The main Storage Element (SE) is located in Padova (2,6 TB of storage) using the INFN infrastructures, while other SE are already active in three locations (Palermo, Firenze, Roma) and more of them will be activated in the future.



The data organization in the filesystem of each SE is depicted in figure 2, guaranteeing efficiency and security.

There are three main directories: *fonti*, where the bought data are stored, *progetti,* that contains data transformed for different research purposes and *utenti*, with the personal home of each researcher.
The directory *fonti* further contains different directories for the different contracts (LSE, NYSE, EMID ….), and in each of them there are two subdirectories: *originali* and *trasformati.* The first subdirectory contains the original data as received from the Exchange or vendors; the second subdirectory contains pre-processed data useful for all the researchers.



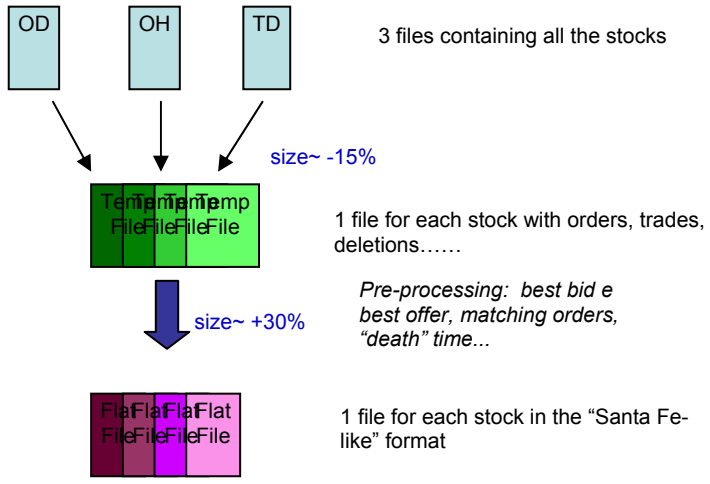
1.2 Dataset Description

Here we present a brief description of the different kind of data purchased by the project.
- London Stock Exchange: Rebuild Order Book 2002
The ROB has 3 type of files: t_OrderDetail.csv, t_OrderHistory.csv, t_TradeReports.csv. Each file is produced in a comma separated format.
The first file contains details of orders that enter the Order Book and of orders that completely execute on entry. The second file contains a history of changes to each order and the method by which it is removed. For each order there are one or more rows shown with an *order action type* for each record. The order action type indicates if the order was partially or fully executed against, deleted, expired etc ...The trade report file contains details of every automatic and manual trade that has taken place, except for trades that have not been published.
- Borsa Milano: Trades, Quotes, 5 Best Offers 2002
- S&P: Index 1983-2004, Futures 1982-2004
- New York Stock Exchange: Trades And Quotes 1995-2003
The Trade and Quote (TAQ) database contains intraday transactions data (trades and quotes) for all securities listed on the NYSE and AMEX, as well as Nasdaq National Market System and SmallCap issues.
- New York Stock Exchange: Openbook 2002
NYSE OpenBook provides a detailed view of the Exchange's limit-order book for all NYSE-traded securities. It tracks the aggregate limit-order volume at every bid and offer price throughout the trading day.
- Tokyo Stock Exchange: Trades 2002

- E-Mid: Interbank Deposits 2002
e-MID is the electronic organised Market for Interbank Deposits (in euros and American dollars). e-MID was founded in Italy in 1990 thanks to the expertise of the main Italian treasurers and money market dealers.
- E-Mid: Overnight Swaps On The Interbank Deposits 2002 (e-MIDER)
e-MIDER is the electronic organised Market for Overnight Indexed Swaps (in euros and American dollars). In a money market industry turning more and more to derivative instruments for other than very short term positions, e-MIDER is the natural complement to the e-MID deposit market, serving both hedging and trading purposes.
- Euronext: Trades Paris, Amsterdam, Brussels 2002
Euronext Data has 4 types of files: the Adjustment Coefficient File, the Cash Intraday File, the Shares Intraday File and the Dividend File. The market quotes shown are retrieved from the market data feed not more than once, every second. In the case that multiple prices were traded during this time frame, only one price is shown. The market overview is designed to give a good impression of the trading day but should not be considered a full recording of the trading day.
- Mts –Time Series 2003-2004
High-frequency data on European bonds.

1.3 Pre-processed Data

We have produced different kinds of "transformed" data.
The first and simpler transformation is the generation of ASCII files from binary files containing some type of data (as the NYSE datasets). The second kind of pre-processing is the generation of "new formats" that are more clear and useful for our research purposes and that contain more explicit information.
Two examples of pre-processed files from the LSE order book are the "Santa Fe-like" flatfiles and the "sec-by-sec" files.
In figure 3 we present a sketch of the method used in order to create the "Santa Fe-like" flatfiles.



1.4 Legal aspects

An important characteristic of our project is that the contracts drawn up with Stock Exchanges and data vendors have different conditions about ownership and sharing of financial data, i.e.

	Property	Duration	Extension to other research groups joining the project	Property at the end of the project
LSE, TSE, Borsa Milano, NYSE OPENBOOK, TAQ (2003)	All groups	3 years (renewable)	yes	INFM
TAQ (1995-2002)	Extended to all the research groups for the products previously acquired by Palermo (1995-2000) and Firenze (2000-2002)	3 years (renewable)	yes	Original owners
E-Mid, Euronext	All groups	Unlimited	no	All the groups
MTS	All groups	3 years	no	INFM
S&P	INFM (PA,TS,RM) Universita'di Firenze	Unlimited	no	All the groups

The EGRID infrastructure permits an efficient organization of datasets with different contractual clauses for a large group of researchers, belonging to different institution and university all over Italy.
The EGRID users, thanks to the EGRID system of certifications and permissions, gain access just to the dataset they can use accordingly to the contracts they have signed, but it is still possible to create temporary research groups of users from different location, working on the same data.
The whole structure can be easily modified and enriched with new users, research groups or data, guaranteeing a high degree of transparency and security.

1.5 Conclusion

The EGRID platform helps us in the storage and in the management of this very large (order of 2 TB) amount of high frequency financial data and has significantly improved the performances of the applications that process and analyze the data themselves. Some specific EGRID services have been developed, in collaboration with EGRID group, with the purpose of optimizing our data analysis and sharing.

Surprisingly enough, the grid protocol, thanks to its security services and to the fact that data are centralized in a single, controlled structure, has allowed us to purchase financial data more easily and at good conditions. Data vendors, in fact, show great confidence and trust in the reliability of grid data organization and this makes the collaboration much easier.
Moreover, due to the flexibility of the Grid, we can organize the data respecting all legal terms and conditions signed in the different contracts.