StarQube

INVESTMENT DECISION IN A BOX

StarQube ESG #2

Collecting, cleansing, hosting, transforming raw ESG data

Collecting, cleansing, hosting, transforming raw ESG data

In this series of articles, we highlight the practical challenges that asset managers face with the proliferation of ESG data.

Our previous post was devoted to the management of correspondence tables – in particular between issuers and financial instruments. We will focus here on the challenges in terms of collecting, cleansing, hosting and transforming raw ESG data.

At the scale of a portfolio of 200 stocks and with a single ESG data provider, the integration of ESG data may seem simple and manually manageable on an Excel sheet. But this is rarely the reality of asset managers – who are quickly confronted with issues in terms of volume requiring the implementation of an industrial ESG process.

Take the example of an asset manager who wishes to have as exhaustive an ESG coverage as possible on his global and cross-asset investment universe. His "ESG volumetry" will depend on the following parameters (the following figures are orders of magnitude):

Issuer / instrument coverage: ~10,000 issuers / 500,000 instruments

- 8,500 issuers in the MSCI ESG universe, 11,000 in the S&P ESG universe.
- MSCI indicates that it covers 680,000 financial instruments.

Number of data fields per issuer: 50+ indicators (even 1,000+ fields)

- *Based on the ESG scores of external providers*: 1 composite ESG score + 3 scores per pillar (E, S and G) + 10 scores per theme + 35 scores per key subject for MSCI, i.e. at least 50 indicators per issuer x number of ESG data providers to which the asset manager subscribes.
- *Based on a direct collection of data from issuers*: MSCI or S&P indicate that they collect more than 1,000 data fields from issuers.
- The transparency requirements of asset managers and their need to have a detailed perception of the ESG risks that weigh on issuers are increasingly pushing them to collect raw ESG data rather than contenting themselves with external scores.

Historical depth: ~15-20 years + need to increase collection frequency

- 17 years of historical depth for MSCI.
- If the scores are currently constructed on an annual frequency, it is likely that in the future investors will wish to have much more dynamic scores, which adjust during the year with the news flow; in our next paper, we will explain why it is already essential to organize ESG data in order to build dynamic and time-stamped ("point-in-time") indicators.

And of course, all these scores (at least composite scores) must be distributed across all of the asset manager's portfolios for portfolio management and (internal, commercial, regulatory) reporting purposes.

It is therefore easy to understand that it is impossible to build an industrial ESG process without extremely solid data foundations, which facilitate:

1. Data collection
    - ESG data updates, usually from multiple providers and delivered in various formats, should be collected in the most automatic way possible.
    - Even though ESG indicators are most often disclosed by issuers on an annual basis, not all issuers communicate in a synchronized manner, so the raw ESG data (and the composite scores calculated by the agencies) are subject to updates throughout the year.
2. Cleansing incoming data
    - The files transmitted by data providers are never 100% clean: missing data, inconsistent data (leaps in data compared to the previous information), etc. Asset managers must put in place automatic processes for identifying and correcting most obvious errors in the data collected.
3. Data hosting
    - The choices relating to the infrastructure and hosting of ESG data are very structuring in terms of the ability of operational teams (ESG analysts, quantitative researchers, portfolio managers, reporting officers, etc.) to access them and use them effectively.
4. Data transformation and enrichment
    - In the absence of a standard ESG framework, ESG score providers have developed proprietary classifications: MSCI's ESG scores range from AAA (for the best issuers) to CCC (for the worst), while the S&P scores are constructed on a scale of 0 – 100 (100 for the best). Asset managers must therefore at least standardize the ESG scores of their providers to build consistent reports or create composite scores.
    - But most asset management companies go beyond collecting ESG scores from external providers and prefer to develop a proprietary ESG methodology that reflects their environmental, social and governance priorities. Without a flexible data infrastructure, the implementation of a proprietary ESG methodology is quickly complex for large investment universes.
    - The data must also be enriched – for example by replacing non-available information on an issuer with a sector average or median.

WHAT STARQUBE OFFERS

StarQube is optimized for the management (and use) of large volumes of data and offers the following capabilities:

1. *Collection/Cleansing:* StarQube has "off-the-shelf" connectors with most (financial or extra-financial) data providers. These connectors make it possible to retrieve new files made available over time (regardless of their format and delivery method), to format them and to insert all completeness / consistency / correction tests desired. New connectors can be built very quickly by StarQube according to the needs of users, who are also free to build their own connectors based on the models and documentation provided. The connectors are scheduled to automatically retrieve updated data, without manual triggering from users.
2. *Hosting*: the data is hosted in a NoSQL database optimized for heavy calculations (backtests, portfolio optimizations) on large volumes of data, and natively point-in-time (all data is timestamped – read our next article on the importance of timestamping your data). Most calculations performed "on the fly" are instantaneous; the most complex backtests can take a few tens of seconds, a few minutes at most.
3. *Organization:* the data is organized around (1) a single repository (unique identifier per issuer and per instrument) which facilitates navigation between datasets, and (2) a library of group org charts which has all the correspondences between parent companies and subsidiaries, and ensures the correspondences between the tables of issuer identifiers and the tables of instrument identifiers – read our previous article on "mapping".
4. *Transformation / Enrichment:* StarQube has a "low-code" language that offers total flexibility to transform raw ESG data into proprietary scores or to enrich the data by completing missing information (e.g. with sector/country averages or medians ...).

LINK TO FULL ARTICLE

About StarQube

Founded in 2013, StarQube develops an innovative and modular solution for asset management companies based on two pillars. The **data management** pillar industrializes the collection, cleansing and organization of all types of data useful for the investment process within a centralized NoSQL database. The **portfolio construction and management** pillar offers modules to analyze the research universe, build proprietary risk models, create model indices or portfolios, model and backtest investment strategies, optimize and rebalance portfolios. Graphical interfaces make it possible to view, analyze and manage portfolios using customizable screens to display the information which is relevant according to the investment style.

Contact

StarQube
Rue des Corps-Saints 4
1201 Genève

Sales contact: Guillermo Albiñana Arias
Phone: +33 6 52 33 80 33
Mail: guillermo.albinana@starqube.com
Web: www.starqube.com
LinkedIn: https://www.linkedin.com/company/starqube