Standardizing, Structuring and Linking: Solving Alternative Data Challenges with S&P Global Marketplace

Jay Bhankharia
Senior Director

Laura Ieni
Director

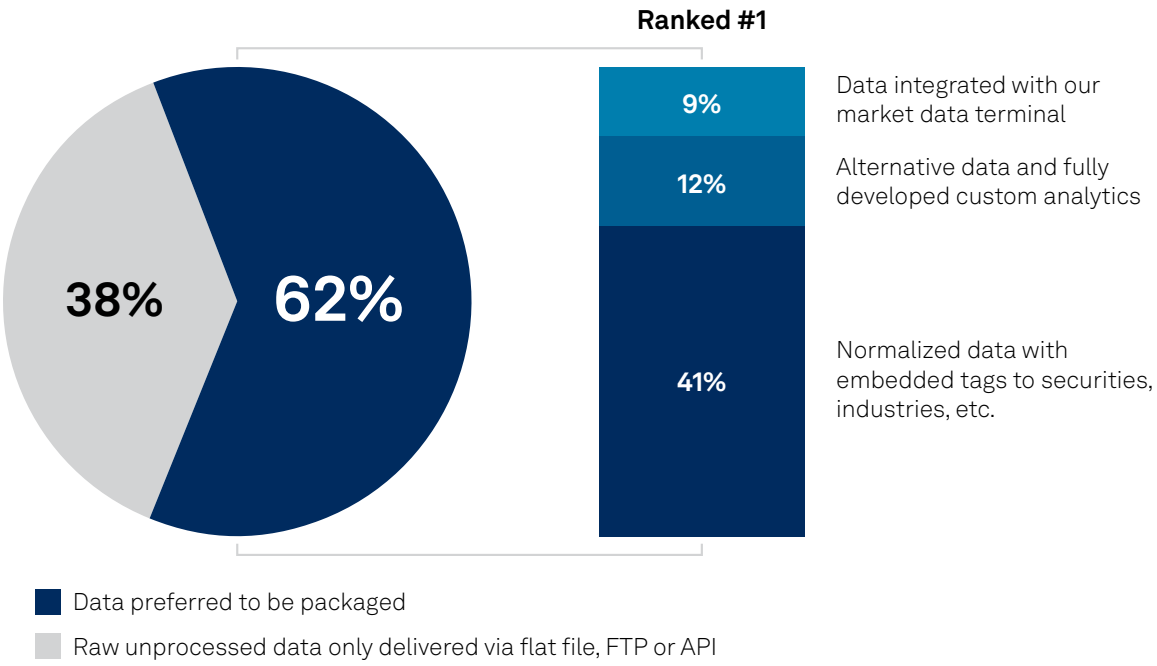
April 2021

S&P Global
Market Intelligence

Standardizing, Structuring, and Linking: Solving Alternative Data Challenges with S&P Global Marketplace

The amount of data available in the market is expected to grow tenfold by 2025, with alternative data leading the way[1]. As alternative data becomes more mainstream, businesses are discovering the value and actionable insight it can provide — allowing for more informed decisions in today's quickly evolving markets. However, one of the largest barriers in the adoption of alternative data is the time commitment to making the data usable. Often times, alternative data is messy and unstructured, so it needs to be validated, structured, and linked before meaningful analysis can begin. In a recent study conducted by Greenwich Associates, 62% of respondents wanted their data pre-packaged and structured to make it more easily consumable. 48% of respondents also cited that difficulty cleaning and integrating data was a key obstacle to using alternative data[2]. This is where S&P Global Marketplace, a platform for exploration and evaluation of data from S&P Global and select third party providers, comes in. Understanding these challenges, S&P Global does the heavy lifting of structuring and linking datasets upfront to enable our customers to spend more time on analytics. Below, we'll take a deeper dive into how S&P Global standardizes, structures, and links data using a few examples of data offerings available on the S&P Global Marketplace.

Data Delivery Preference



Note: Based on 34 respondents.
Source: Greenwich Associates 2019 Alternative Data Study.

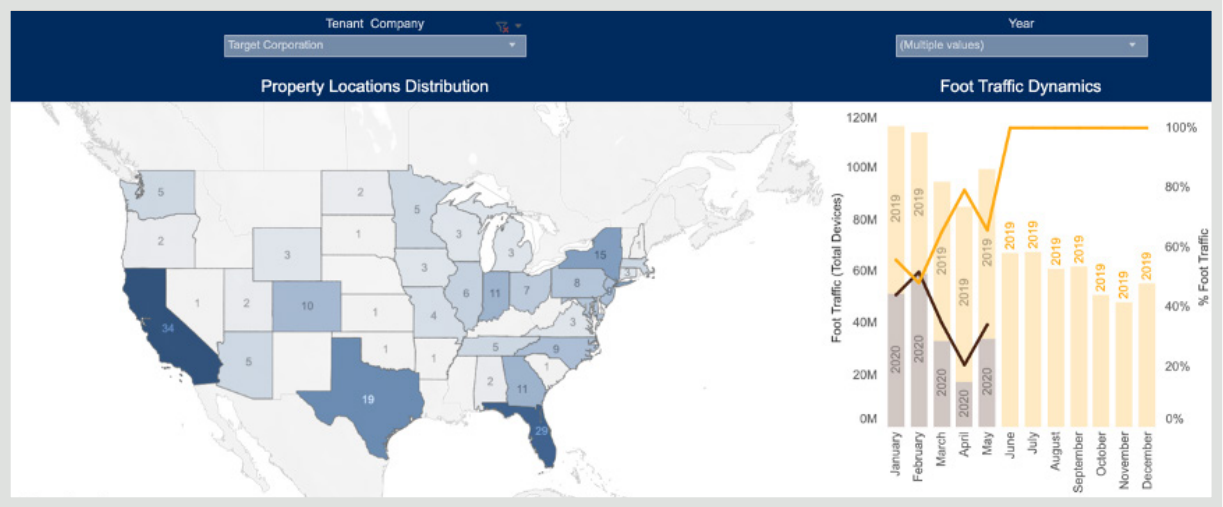
[1] Deloitte, "InFocus Alternative data adoption", 2018.
[2] Greenwich Associates, "Demystifying Alternative Data", May 22, 2019.

Data Standardization of AirSage Real Estate Foot Traffic Data

One of the main challenges with alternative data is that it's all created and organized differently depending on who is collecting the data. For instance, something as small as characters, spacing or naming of a specific field can add hours of data cleansing work before being able to begin analysis of the data. With awareness of this challenge, S&P Global has ensured that all 3rd party data available on S&P Global Marketplace is standardized to be seamlessly integrated with other core S&P Global datasets.

The data standardization process is demonstrated with one of the recent products added to S&P Global Marketplace, S&P Global Real Estate Foot Traffic. This product was developed in partnership with AirSage, a leader in location data intelligence, and offers anonymized foot traffic data collected from mobile devices, mapped to S&P Global Real Estate properties, alongside aggregated demographic information of location visitors. Investors can use foot traffic data to understand the health of a business based on an increase or decrease in visitors, as well as evaluate the risk of REIT vacancy. Foot traffic data has become a major focus as investors and corporations try to gain an understanding of movement behavior and patterns in light of COVID-19 business closures and the subsequent re-opening of economies across the United States.

Nordstrom Property Locations and Foot Traffic 2019 vs. 2020



The S&P Global Marketplace product team works alongside AirSage to sync their foot traffic data with the S&P Global real estate properties database on a monthly basis. This ensures that the product always reflects the most up to date information as new properties emerge or change hands. After S&P Global receives the data from the vendor, we run standardization processing that transforms the data received into formatting consistent with other S&P Global data products. In the example below, we've reformatted the date fields and property ID column headers to align to our standards. This is helpful for our clients because it ensures consistency and uniformity across all data products purchased from S&P Global and helps to minimize the learning curve when using new datasets.

Examples of AirSage Standardization

- Date field (date, asofdate) formatting is standardized to YYYY-MM-DD format — "20200617" → "2020-06-17"
- Schema standardization for consistent column naming, special characters are removed
  - "#poi" → "propertyId"
  - "as:ofdate" → "asOfDate"
  - "extrapoloated_0000_0600" → "devices12AMto6AM

Source File (vendor-Airsage)

#poi	da,te	as:ofdate`	extrapolated_device_total	extrapolated_0000_0600
4195	20170102	20170102	1988.851084	0.000000
4195	20170103	20170103	1222.462004	0.000000

Raw Table (vendor-Airsage)

raw_airsageweeklysummary.poi	raw_airsageweeklysummary.date	raw_airsageweeklysummary.asofdate	raw_airsageweeklysummary.extrapolated_device_total	raw_airsageweeklysummary.extrapolated_0600_1000
4195	20170102	20170102	1988.851084	0.000000
4195	20170103	20170103	1222.462004	0.000000

Data in XF – Client Facing Table

propertyId	asOfDate	devicesTotalPerDay	devices12AMTo6AM	devices6AMTo10AM
4195	2017-01-02 00:00:00.000	1988.851084	0.000000	155.001843
4195	2017-01-03 00:00:00.000	1222.462004	0.000000	266.507939

The standardization of the AirSage data makes it easy to further analyze a property's information with other S&P Global datasets. One use case is using the Real Estate Foot traffic package to zero in on a specific property or list of properties and return all tenants. One can then view each tenant's recently reported financials and credit health by marrying the data with core datasets like S&P Capital IQ Fundamentals and RatingsXpress®.

Data Structuring of FDA Data

Although many datasets are available online for free via public access, they lack the structure and time variant consistency to be usable without significant maintenance. Even if the data is valuable, many companies do not have the resources or time to manage and structure the data. A good example of this is data provided by the U.S. Food and Drug Administrations (FDA). FDA data is publicly available via the U.S. government website, however, it lacks consistency due to file formats changing over time, and requires significant structuring to become readily consumable. However, S&P Global has designed a suite of FDA data products from the FDA that addresses these challenges. Similar to our other data offerings, these datasets have been thoughtfully structured and linked back to S&P Global company IDs, so clients can seamlessly use this data in conjunction with our other core and alternative datasets.

The three FDA datasets that make up the FDA product suite provide valuable insight into the healthcare space — the offering includes recently filed medical device applications (510k), pharmaceutical drug patents (Orange Book), and medical adverse events (FAERS).

Dataset	Description	Usage
510k Filings	Issued when there is intent to bring a medical device to market within the next 90 days.	Used to notify the FDA of their intent to market medical devices up to 90 days before coming to market. The 510k notice may be available via the FDA before the product is approved.
Orange Book	Identifies drug products approved on the basis of safety and effectiveness by the FDA.	Used to analyze a company's performance correlated to their products, drug patents, and exclusivity to market the pharmaceutical.
FDA Adverse Event Reporting System (FAERS)	Contains records of adverse event reports, medication error reports and product quality complaints resulting in adverse events that were submitted to the FDA.	Used to understand risk and impact of recalls of drugs (Tylenol), ingredients (acetaminophen), or devices (pacemaker).

For these FDA datasets, we have gathered historical data to ensure its suitability for backtesting and historical analysis. The S&P Global Alternative Data Vetting team, which scouts the market for interesting and unique content providers to partner with, filed a Freedom of Information Act (FOIA) request to acquire the historical data, ensuring a robust history. After reviewing the files, we observed that historical files lacked column headers and schemas and had metadata changes. The product file, for example, had five versions over time with varying column headers:

File Type	Schema
productv4	Ingredient~DF;Route~Trade Name~Applicant~Strength~NDA~Product~TE Code~Approval Date~RLD~Type
productv3	DF;Route~Trade Name~Applicant~Strength~NDA~Product~TE Code~Approval Date~RLD~Type~Applicant Full Name
productv2	DF;Route~Trade_Name~Applicant~Strength~Appl_Type~Appl_No~Product_No~TE_Code~Approval_Date~RLD~Type~Applicant_Full_Name
productv1	Ingredient~DF;Route~Trade_Name~Applicant~Strength~Appl_Type~Appl_No~Product_No~TE_Code~Approval_Date~RLD~Type~Applicant_Full_Name
product	Ingredient~DF;Route~Trade_Name~Applicant~Strength~Appl_Type~Appl_No~Product_No~TE_Code~Approval_Date~RLD~RS~Type~Applicant_Full_Name

With this understanding, processes were built to account for file variations over time and checks were added for future file deviations to be accounted for in the data ingestion pipeline.

The data has also been structured into a schema that is easier to query. One example of re-structuring is creating reference tables that convert text, which can be inconsistent and is case sensitive, to look-up codes. Specifically, for the FAERS dataset a reference table was created that matches patient outcomes to an outcome code. The user can then construct a query to look at adverse event cases that resulted in a specific outcome, such as hospitalization.

Reference Lookup Table for PatientOutcomeCode

keyfaerspatientoutcomecode [PK] integer	mpfdapatientoutcomecode character varying (10)	mpfdapatientoutcomecodedesc character varying (4000)
1	DE	Death
2	LT	Life-Threatening
3	HO	Hospitalization (Initial or Prolonged)
4	DS	Disability
5	CA	Congenital Anomaly
6	RI	Required Intervention to Prevent Permanent Impairment/Damage
7	OT	Other Serious (Important Medical Event)

Below is an example query that looks at a specific adverse event case reported when taking a 1.5 mg dose of Exelon, a product manufactured by Novartis, that required hospitalization. The query also includes information about the specific reaction that occurred, such as abnormal white blood cell counts, pneumonia, etc.

Example query results of FAERS data

mpfdacaseid bigint	mpfdamfrsndr character varying (300)	mpfdadrugname character varying (500)	mpfdadosevbm character varying (300)	mpfdadoseamt character varying (15)	mpfdapt character varying (500)	keyfaerspatientoutcomecode integer
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	White blood cell count abnormal	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	White blood cell count abnormal	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	Dyspepsia	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	Dyspepsia	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	Pneumonia	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	Pneumonia	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	Abdominal pain	3
3567656	NOVARTIS	EXELON	1.5 MG, BID	1.5	Abdominal pain	3

Data Linking of Patent and Trademark Data

While information is beneficial in decision making, it only becomes actionable when this information can be combined with other data and information to see the complete picture. With that, linking of data is one of the greatest challenges for users.

Linking is a challenge for any analysis requiring more than one dataset. To address this within S&P Global products we brought in Kensho, the AI arm of S&P Global, to apply machine-learning solutions to solve big data linking challenges. Kensho Link, which provides a S&P Global identifier to user submitted company lists, has been developed over the past three years based on the corpus of over 10 million+ private and public entities within S&P Global Market Intelligence's databases. The product is able to recognize parent child relationships as well as common abbreviations and aliases. Kensho Link can link, on average, 100,000 entities per hour, which is superior to similar linking offerings in processing volume and timeliness. Linking to S&P Global identifiers also opens up the Cross Reference Services universe, which includes IDs such as LEI and DUNS.

IPqwery Patent and Trademark is a compelling case study for the power of linking. This data is gathered by the vendor from several government IP offices and can provide insight into new products before press releases hit newswires. Users can utilize this data to understand IP industry leaders and to perform comparative analysis across companies, competitors, and industries. Linking IPqwery data to the millions of public and private companies covered in the S&P Global universe unlocks powerful analysis that can be performed when this data is combined with S&P Global core datasets, like S&P Global Capital IQ Fundamentals.

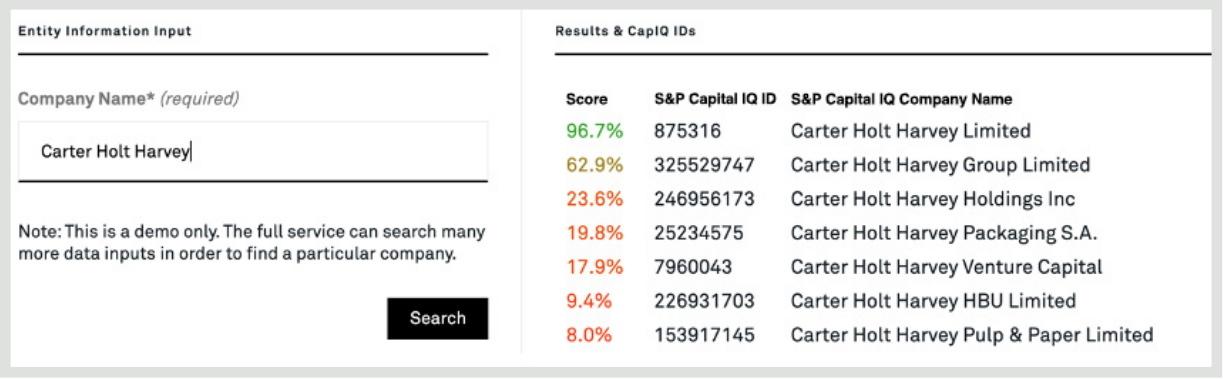
Below you can see an example of a proprietary owner ID sent to us from IPqwery along with the owner's patents and trademarks. Using the Kensho algorithm we can match Carter Hold Harvey to Capital Q ID 875316, which can be used to link to other S&P Global core datasets.

Kensho Linking Example

Raw file from vendor


```
{
    "ownerId": 230291,
    "ownerName": "Carter Holt Harvey Limited",
    "address": "640 Great South Road, Manukau City, Auckland",
    "state": "",
    "country": "New Zealand",
    "countryCode": "NZ",
    "crunchbaseUuid": "c21ab203-da36-af02-5aa8-92f58ff311db"
}
```


Linked to S&P Capital IQ ID



One can use intellectual property data to uncover when a company began filing a specific type of patent and link that activity back to spending on R&D as a potential predictor of future revenue and stock returns. In the example below, we can take a look recently filed patents by Apple Inc. to get a better idea of where their product focus is.

Recently Filed Apple Inc Patents

COMPANYNAME	COMPANYID	OWNERID	STOCKTICKER	IPID	PATENTFILINGDATE	ISSUEREGISTRATIONDATE	PRISTATUS	PATENTTITLE
Apple Inc.	24937	24	AAPL	PTUS_G10769415	8/31/18	9/8/20	Registered	Detection of identity changes during facial recognition enrollment process
Apple Inc.	24937	24	AAPL	PTUS_G10769415	8/31/18	9/8/20	Registered	Detection of identity changes during facial recognition enrollment process
Apple Inc.	24937	24	AAPL	PTUS_G10769415	8/31/18	9/8/20	Registered	Detection of identity changes during facial recognition enrollment process
Apple Inc.	24937	24	AAPL	PTUS_G10769415	8/31/18	9/8/20	Registered	Detection of identity changes during facial recognition enrollment process
Apple Inc.	24937	24	AAPL	PTUS_16883031	5/26/20		Pending	IN-EAR HEADPHONE
Apple Inc.	24937	24	AAPL	PTUS_16883031	5/26/20		Pending	IN-EAR HEADPHONE
Apple Inc.	24937	24	AAPL	PTUS_16530708	8/2/19	9/8/20	Registered	Reducing the need for manual start/end-pointing and trigger phrases
Apple Inc.	24937	24	AAPL	PTUS_16530708	8/2/19	9/8/20	Registered	Reducing the need for manual start/end-pointing and trigger phrases
Apple Inc.	24937	24	AAPL	PTUS_16530708	8/2/19	9/8/20	Registered	Reducing the need for manual start/end-pointing and trigger phrases
Apple Inc.	24937	24	AAPL	PTUS_16530708	8/2/19	9/8/20	Registered	Reducing the need for manual start/end-pointing and trigger phrases
Apple Inc.	24937	24	AAPL	PTUS_16530708	8/2/19	9/8/20	Registered	Reducing the need for manual start/end-pointing and trigger phrases

Bring it all together with Trucost Climate Change Physical Risk

Trucost Physical Risk, part of the broader Climate and ESG data available from S&P Global, assesses a company's exposure to climate-related physical risks at the asset-level. A company's exposure is based on seven climate change physical risk indicators, including heatwaves, water stress, wildfires, and sea level rise, using high/low/moderate scenario analysis. The dataset also provides estimates of future climate change physical impact.

The Physical Risk offering brings together a large number of different datasets (companies, assets, indicators, scenarios, timeframes) under the lens of climate risk and is very relevant when making investment and business decisions. It can be used by companies and investors to better understand climate related risk exposure, identify opportunities to mitigate risk, and to provide more transparency in line with mandatory and voluntary reporting requirements such as the Task Force on Climate-related Financial Disclosures (TCFD).

As this dataset is comprised of data sourced from several places, the important process of data standardization, structuring and linking was applied. The team needed to standardize the physical risk data at different geo-spatial scales, for the seven climate hazards and three different climate change scenarios. Following the standardization process, the data needed to be structured in order to calculate the physical risk scores at the asset level, which involves combining asset location information and physical risk data for 500k+ corporate assets from S&P Global databases. Finally, the corporate assets were linked to 15,000+ ultimate parents in the S&P Global Market Intelligence database. This is done so the user can understand the aggregated risk and impact at a company level.

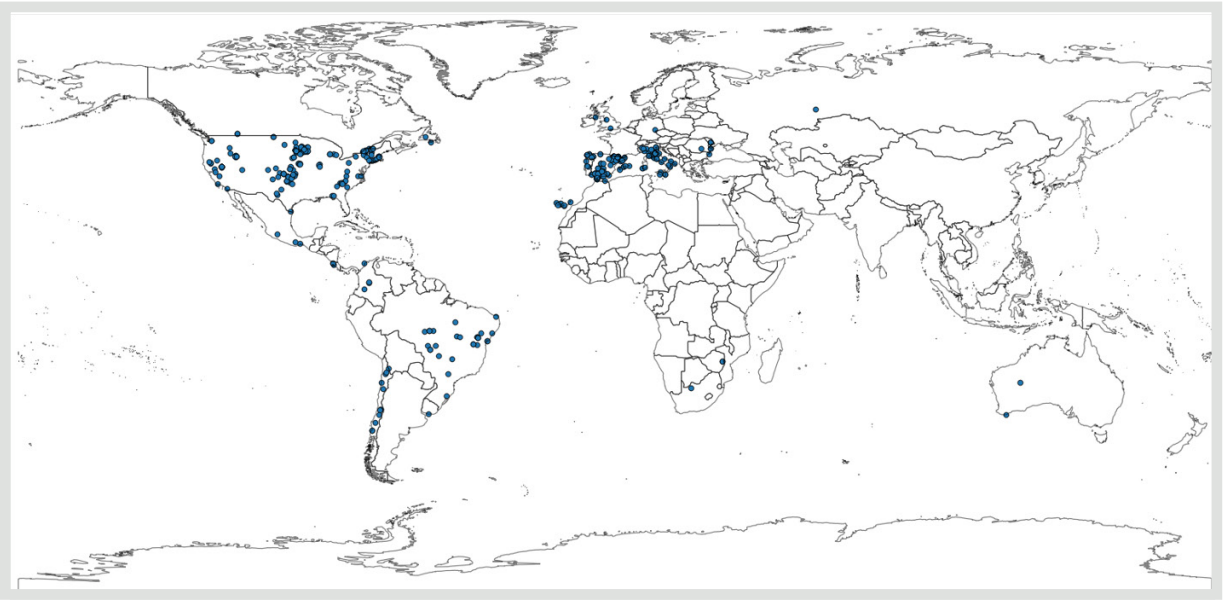
Nextera Energy Assets Alongside Their Physical Risk Scores

Asset Name	MI Asset ID	Country	Facility Type	Owner Name	Owner CIQ ID	Ultimate Parent Name	Ultimate Parent CIQ ID	GICS Sector	High Scenario 2050							
									Wildfire	Coldwave	Heatwave	Water Stress	Flood	Hurricane	Sea Level Rise	Composite Score
Adelaide Wind Energy Centre	-167381	Canada	Power Plant - Wind Turbine	NextEra Energy Canada, ULC	270586	NextEra Energy, Inc.	270586	Utilities	30	18	29	10	29	1	1	56
Adelanto I Solar Farm	-171391	USA	Power Plant - Solar	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	100	22	16	88	5	1	1	80
Adelanto II Solar Farm	-172207	USA	Power Plant - Solar	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	100	22	16	88	5	1	1	80
Aiken County Solar Facility (Shaw Creek Solar)	-151391	USA	Power Plant - Solar	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	59	29	34	50	100	1	1	85
AMP Solar Phase II	-172366	USA	Power Plant - Solar	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	39	30	30	100	7	1	1	77
Antelope Hills Wind Farm (Western Branch)	-121762	USA	Power Plant - Wind Turbine	NextEra Energy, Inc.	270586	NextEra Energy, Inc.	270586	Utilities	12	43	17	70	20	1	1	68
Arlington (Riverside County Solar)	-143669	USA	Power Plant - Solar	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	100	20	22	11	1	1	1	66
Armadillo Flats Wind Project	-171984	USA	Power Plant - Wind Turbine	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	3	43	25	10	8	1	1	47
Ashtabula I - NextEra Wind Farm	-167240	USA	Power Plant - Wind Turbine	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	27	31	18	100	82	1	1	84
Ashtabula II Â– NextEra	-163441	USA	Power Plant - Wind Turbine	NextEra Energy Resources LLC	270586	NextEra Energy, Inc.	270586	Utilities	27	31	18	100	82	1	1	84
Ashtabula III-NextEra Wind Farm	-118982	USA	Power Plant - Wind Turbine	NextEra Energy Equity Partners, LP	270586	NextEra Energy, Inc.	270586	Utilities	27	31	18	100	82	1	1	84
Ashtabula IV	-105084	USA	Power Plant - Wind Turbine	NextEra Energy, Inc.	270586	NextEra Energy, Inc.	270586	Utilities	27	34	18	100	82	1	1	84
Athol Solar Project	-169493	USA	Power Plant - Solar	Smart Energy Capital, LLC	270586	NextEra Energy, Inc.	270586	Utilities	17	19	26	56	7	1	1	59
Babcock Ranch Solar Energy Center	-167973	USA	Power Plant - Solar	Florida Power & Light Company	270586	NextEra Energy, Inc.	270586	Utilities	44	27	66	100	15	61	1	89
Baldwin Wind Energy Center	-118960	USA	Power Plant - Wind Turbine	NextEra Energy Equity Partners, LP	270586	NextEra Energy, Inc.	270586	Utilities	12	38	17	70	20	1	1	67

On the next page you can see a more granular view of how one climate hazard score is calculated for one of the corporate assets owned by a global energy company.

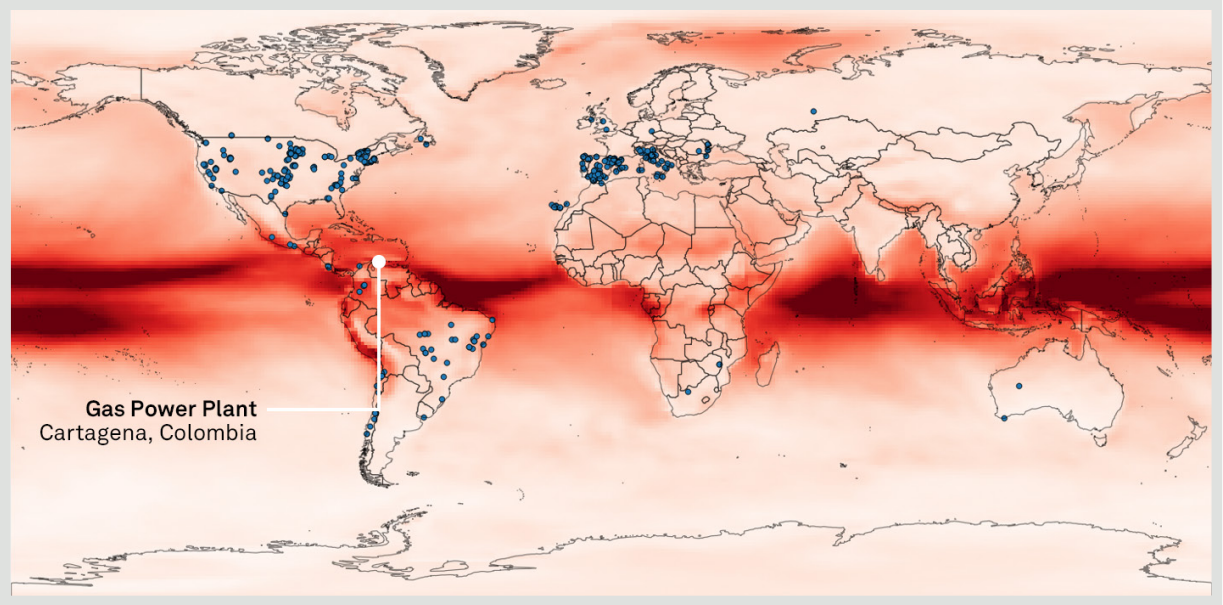
Asset Distribution for a Global Energy Company

Global operations of a global energy company with activities spanning the USA, Europe, Latin America, Africa and Australia



Source: Trucost analysis (2019). For illustrative purposes only.

Asset Analysis



Source: Trucost analysis (2019) based on CMIP5 and S&P Global asset data. For illustrative purposes only.

Corporate Asset: Gas Power Plant
Location: Cartagena, Colombia
Climate Hazard: Heatwaves
Scenario: High — Year 2050

Step 1: Overlay Asset Locations on Climate Hazard Map

Geospatial analysis is used to overlay asset locations on a climate hazard map. In this case the hazard map represents heatwave risk, defined as the expected number of heatwave days per annum under a high climate change scenario in year 2050.

Step 2: Sample Hazard Level at Asset Location

Climate hazard data is mapped in the form of spatial grids with estimates of the number of heatwave days in each grid cell. Assets are assigned a heatwave risk estimate based on the risk level of the cell in which it is located.

Step 3: Score Normalization

Asset level scores are normalized relative to the range of hazard levels for each indicator globally to enable simpler comparison of the risk exposure of multiple assets, and the risk level of any asset relative to global conditions.

Conclusion

Cleaned, structured and linked data is essential for efficient and beneficial data analysis. S&P Global Marketplace stresses these qualities for all alternative datasets available via FTP, Cloud or API. We've built out robust data checks and standardization processes to ensure high quality data that's clean and conforms to S&P Global formatting standards. Our product managers, alongside our technology team, design the best structure possible for each dataset so that it's easy to use and understand, especially since many of these datasets are new and may be unfamiliar. Lastly, Marketplace datasets are linked to S&P Global identifiers. This ensures that alternative datasets being brought to market can be easily analyzed alongside other S&P Global content. Above all, S&P Global Marketplace has selected to offer alternative data, that when enriched and combined with our core data, leads to informative and actionable insights so our customers can make decisions with conviction.

About S&P Global Market Intelligence

At S&P Global Market Intelligence, we understand the importance of accurate, deep and insightful information. We integrate financial and industry data, research and news into tools that help track performance, generate alpha, identify investment ideas, perform valuations and assess credit risk. Investment professionals, government agencies, corporations and universities around the world use this essential intelligence to make business and financial decisions with conviction.

S&P Global Market Intelligence is a division of S&P Global (NYSE: SPGI), the world's foremost provider of credit ratings, benchmarks and analytics in the global capital and commodity markets, offering ESG solutions, deep data and insights on critical business factors. S&P Global has been providing essential intelligence that unlocks opportunity, fosters growth and accelerates progress for more than 160 years. For more information, visit www.spglobal.com/marketintelligence.

CONTACT US

The Americas
+1-877-863-1306

Europe, Middle East & Africa
+44-20-7176-1234

Asia-Pacific
+852-2533-3565

References

2018, "InFocus Alternative data adoption", www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-infocus-alternative-data-in-im-2018.pdf

May 22, 2019, "Demystifying Alternative Data", www.greenwich.com/asset-management/demystifying-alternative-data