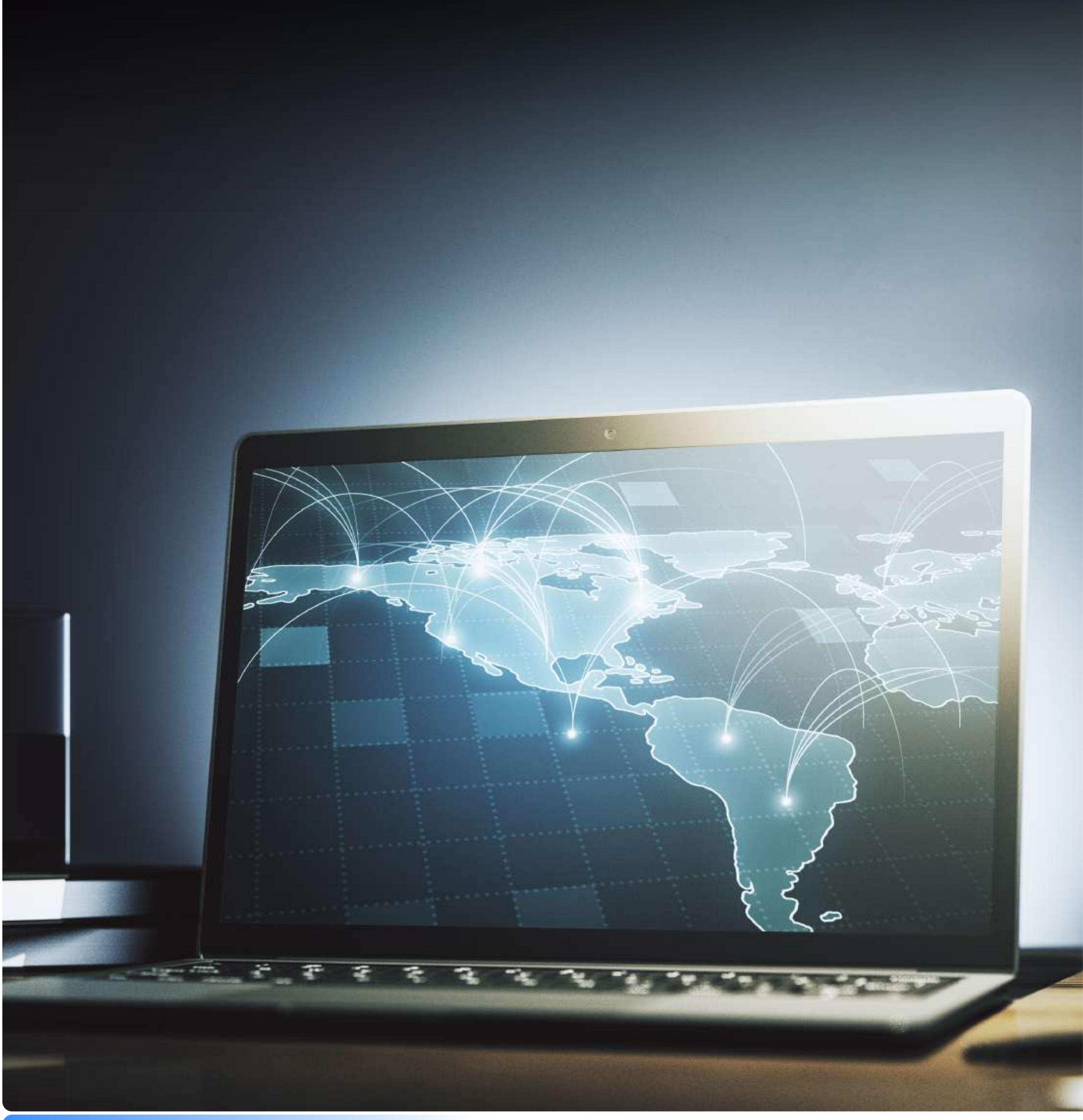
HCLTech | Supercharging Progress™

Effective location data monetization: Strategic and technical enablers



Abstract

With a smartphone in virtually everyone's pocket, businesses have more opportunities than ever to easily understand and effectively leverage traffic and movement patterns. Providers of mobile applications and businesses involved in industries like telecommunications and transportation often have rich sets of location data that can be a treasure trove for a multitude of different firms and for a variety of use cases. From measuring foot traffic at retail stores to identifying commuting patterns, geolocation data has wide applicability for improving products, services and operations.

This whitepaper discusses specific applications and techniques for enabling an effective and sustainable enterprise strategy to provide geolocation data as an additional source of revenue.

Introduction

Geolocation data is data about the physical location of electronic devices, such as smartphones, tablets, laptops, etc. and consequently, people. GPS, mobile cell towers, Wi-Fi access points, transactions and Point of Sale (PoS) terminals or any combination of similar sources are constantly generating geolocation data that can be aggregated or used on the individual level.

This is very sensitive information, and therefore security and strict regulatory compliance must be a priority when collecting and using this data. Legal requirements vary across different countries and are markedly different between the EU and the US, but collecting and using location data almost always requires consent on the customer side.

If a company generates geolocation data as part of its core operations, it is often obvious how this can be used to improve its business. Think of food delivery or car sharing, for example. But location data can also be valuable to companies that do not have the data already available as a "byproduct" of their own operations. This also means that those firms who have the data are presented with an opportunity to monetize it by selling either the raw data or insights based on it to external partners. Location intelligence companies, who buy location data from various sources and sell analyses derived from it. Also sometimes act as middlemen between firms possessing and firms looking for location insights.

The range of industries in which location data is valuable extends far beyond the obvious. Virtually every retailer with brick-and-mortar stores can use foot traffic data to assess potential sites, measure volume at competitors or understand how customers fit store visits into their daily routines. Location data also enables customer profiling and, in turn, targeted advertising for both the virtual and the physical world. It enables measuring traffic and load on various pieces of infrastructure, making it a useful tool for not only private companies but governmental agencies as well.

Leveraging location data

So, how exactly can location information be turned into valuable business insights? What do the use cases outlined above look like in detail? To answer these questions, it is important to understand the distinction between aggregated and individual-level geolocation data. Generally, aggregated data provides insight into places or spatial relations, while the individual-level data gives insight about specific people. Aggregated data is usually computed from individual-level data and is much easier to acquire due to being less sensitive.

One key use of aggregated data is assessing the potential performance of a physical location, be it for a store, an advertising site or any other venue. This is done by measuring "foot traffic" at specific locations. On a basic level, places with more traffic are likely to have a higher sales potential, but aggregated geolocation data can be much more detailed. Traffic around a point can be broken down into demographic categories with additional data or along other spatial features. What if a business wants to target people living in specific areas, commuting around specific lines or visiting specific locations? Aggregated data can answer these questions if the underlying source is granular enough. Even for existing locations, measuring traffic can help identify peak periods or optimal opening hours.

Similar demand-measurement techniques can be applied to infrastructure as well. Traffic can be measured along lines, as well as around spaces. This is traditionally called origin-destination (O-D) analysis and is often done by hand; however, location data is more precise, scalable and flexible. If one is interested in whether a new tram line would attract commuters or whether extending the spatial scope of a shared mobility service would attract additional users, location data can provide solid ground for inference. In addition to measuring traffic between two points or areas, it is possible to break this traffic down into categories based on speed or even mode of transportation.

One can also measure traffic around existing stores. It might seem redundant, as the number of purchases, or any similar measure, can be tracked at stores more directly. Yet there are situations in which this is not possible, such as large public events, tourist sites, etc. This approach is also useful for proxying competitor performance by measuring traffic around locations of competitors' stores.

Aggregated data can also contribute to understanding customer behavior. After identifying customers around a store, it can describe their movement patterns. Where do customers come from? Which other places do they also visit? Similar questions can be combined innovatively to even measure the effectiveness of individual actions, like seeing an ad. This would involve identifying those who were in the proximity of the advertisement and measuring whether they visit specific locations or perform specific actions afterwards; the result of this analysis is still aggregated.

When it comes to individual-level data, its main uses are customer segmentation and targeting. Location data enables us to build rich customer profiles; it can be used to both predict general demographic information and to define very specific customer features. A typical such feature would be "frequent visitor of a specific venue," but any combination of spatiotemporal patterns can be used for segmentation e.g., "people visiting computer stores" or "people visiting an airport."

This kind of data is clearly very sensitive and is only to be used with explicit consent from the user. Such profiling is mostly used to target customers with offers or to assign them measures of risk.

A potential use case for this is targeting those who visited competitors (proximity marketing); this is again something that location data can identify. Measuring advertising effectiveness was already mentioned above and it is naturally even more straightforward with individual data.

Sourcing location data

Geolocation data can be obtained in various ways, each with its own advantages and limitations. Although it is relatively easy to gather some location data, acquiring an accurate, representative and scalable data source can be a challenge.

The most typical source of location data is mobile devices, either through applications or telecommunications providers. Several applications ask for the user's permission to track the phone's location, and many of those applications aim to monetize this information. This is generally done by integrating standard software development kits (SDKs) into applications. Such information is very accurate because it relies on GPS signals, the "gold standard" of location data.

Applications can often also provide additional behavioral data to go with GPS tracking. This can include information like spending patterns or content preferences. However, these sources are lacking in representativeness and scalability. The user bases of these applications are quite narrow and often very specific in terms of demographics or consumption habits; additionally, some applications only provide data for the periods while the application is actively used. Therefore, for example, a fitness app might be a perfect tool for understanding the exercise habits of devoted runners, but ill-suited to measure foot traffic around supermarkets.

Telecommunications providers, on the other hand, manage enormous fleets of devices with extended user bases, which makes them ideal data sources in terms of scalability and representativeness. Their data, however, does not use GPS for location tracking but, often, another form of trilateration, the measurement of distances between a device and multiple mobile cells. This is a remarkably accurate method, especially for 4G and 5G connections, reaching a theoretical precision of a few tens of meters. Telecommunications companies are also able to enrich their data with contractual and demographic information.

A further, more limited source of location data is fixed-location physical hardware: Wi-Fi access points, point of sales terminals and beacon transmitters. Wi-Fi routers can identify devices nearby, and even estimate their distance: this solution is limited in scalability but is a very good option indoors where both GPS and cell signal can be obstructed. PoS terminals only track purchases, which of course always correspond to actual visits, but an actual visit is never guaranteed to result in a purchase. Beacons are devices installed with the explicit intent to sense other devices in their proximity. This makes them suitable for measuring traffic at fixed locations, but they are unable to track patterns across larger areas.

Making sense of location data

Location data, at a fundamental level, is made up of points. These points also generally have timestamps, device or customer IDs and, possibly, additional features. Even if a company is only interested in purchasing aggregated data products, it is essential to develop a clear understanding of them to get a clear picture of what kind of aggregated data is both feasible and suitable to the problem at hand.

Points are represented by coordinates and sometimes additional altitudes, though those are not key to most applications. The most common way to express coordinates are latitudes and longitudes, which identify a point anywhere on the surface of the earth. This also means that latitude-longitude pairs are not located on a plane, so naive ways to measure distances and areas cannot be used when analyzing them.

But there are other projections as well, alternatives to the World Geodetic System that use latitudes and longitudes. These are often only applicable locally (e.g., for a country or state) but can have a huge advantage if they project the 3D sional surface to two dimensions: calculations become easier, faster and more intuitive. Another option is "geohashing": this method converts any location into an alphanumeric string with a chosen precision (the more precise the location, the longer the string). This can greatly improve the speed of looking up or sorting locations/points.

Points, timestamps and IDs should, of course, be aggregated into meaningful measurements. For example, traffic measures at a location. For this, locations must be converted into a polygon or multiple polygons, a practice called geo-fencing. This can be done manually or using openly available data, like that of OpenStreetMap. It can also be automated, approximating locations with circles or rectangles based on point density around them.

The number of points, or unique devices/customers, can also be counted under rasters or hexagons that cover the plane without gaps. This, along with interactive heatmaps, is a fundamental way to capture overall patterns in location data. There are various tools for creating such visualizations including GIS, Kepler, Folium, DataShader, etc.

A more complex challenge is to understand a sequence of points originating from a single device. Location data has an inherent variance or noise, the magnitude of which depends on the source, that complicates this task. One needs to get rid of "unreal" data points and be able to differentiate between noise in the data and the underlying signal. For this, various smoothing and outlier detection techniques exist, like using a simple moving average, the Kaman-filter or the Savitzky–Golay filter.

Smoothing and outlier detection are essential for classifying location signals into "stays" and "trips." Stays constitute time spans during which the location of a device is relatively constant. These are necessary for measuring time spent at venues and differentiating actual visits from simply passing by a location. Clustering algorithms, mainly DBSCAN and its modifications, can also be used to identify stays.

Trips are movements between stays and are also valuable themselves as they can be used to identify the mode of transportation based on route and speed, which is invaluable to many origin-destination analyses. Working in an environment that enables using these methods at scale is essential for dealing with massive location data.

Managing location data

There are two challenges that need to be handled if one wants to store and analyze location data. The first is the sheer amount of data. In the case of mobile network data, transactions — points — per day can easily be in the billions. The second is managing a special data type called "geometry" data type, which encompasses polygons, lines and points, but in some use cases it might even be necessary to describe 3D objects as well. The geometry data type is available by a GIS extension in most popular database systems. Python also has its own packages specifically designed to manage spatial data. This geometry data type is mainly required to be able to use related functions like "within" or "overlaps" without having to develop them from scratch.

Organizations must consider how they are affected by these challenges in case of their own implementation and how they can align their current environment with these new requirements. Processing small amounts of location data is relatively straightforward from a data architecture perspective, but those organizations that access a large amount of such data through their operations need to tailor their systems accordingly. In the case of enormous amounts of incoming location data, the organization likely to implement a distributed database system that enables massively parallel processing. Additionally, to decrease latency between generating data and performing analysis built on them, the company might want to add streaming functionality to their platform instead of batch processing.

In large sources of location data, like mobile applications or telecommunications networks, data is generally captured by an application and gets forwarded to the streaming platform. It is then processed by a Spark Streaming API or a similar alternative and stored in a data lake. This data consumes a lot of storage, which is why it is necessary to compress it. One of two options for this is choosing another file type or changing the compression method of the file type. The other option is to identify those events that are generated at the same point, which makes it possible to shrink the events and save storage.

In general, compression saves disk space, but to use this data later, it must be decompressed every time a user wants to read it. This can make data queries more Central Processing Unit(CPU)-intensive, but also less IO-intensive. Different compression types have different compression ratio and CPU demand, so when the compression type is picked from a wide range of possible options, the user always must find the balance between these two parameters to be able to squeeze the best performance out of the available hardware under their system. The other important factor that can save disk space is the chosen data file format; for example, using Apache Parquet instead of raw CSV files can compress the file size to a tiny fragment of the original one. Parquet achieves this by creating its own metadata-like dictionaries of the values occurring multiple times, so these must be stored only once.

Distributed databases come with a few drawbacks that organizations must consider. Redundancy of the data is a must, as the system can only work efficiently when the same file can be reached from multiple disks. This means an increase in the cost of hardware, but also results in a robust system that is safe from data losses in case of malfunctioning disks. Due to the expected large amount of data, these systems are designed for mass writing or reading of files, but not for modifying them. This necessitates the use of insert-only tables, and modifications require the complete rewriting of the files concerned. And while the option to create updatable tables seems enticing, these are currently inefficient for working with terabytes of data.

Legal challenges

Location data is potentially very sensitive, so it always needs to be approached with the utmost caution. This requires different actions from companies that generate such data, as well as from those who wish to acquire location data or insight into the market. This potential sensitivity is most evident with individual-level data. Even without any additional personal information, a sequence of locations can be used to identify people and draw further conclusions about them. Individual-level data must be stored using pseudo-anonymization: assigning meaningless IDs to each individual in the data and regularly updating such IDs.

Even this, however, does not fully solve the problem outlined above. Hence, regulators often limit the commercial availability of individual-level data, especially in Europe. The regulations further increase the importance of extracting information from aggregated data. Similar limitations are common regarding the purpose of data collection, like limiting firms to only collect data to serve their core business purposes, i.e., their legitimate interests.

Aggregating the individual observations still does not completely solve the identification problem. Appropriate lower thresholds need to be applied to aggregate values to make sure no individual can be identified. The specifics of this are often not codified in the relevant regulations, so in addition to adhering to the relevant local policies like GDPR, CCPA, etc., companies must establish transparent internal guidelines for these practices.

Conclusion

Location data can provide invaluable insight into a wide range of industries and use cases. The scope of location analytics is immense, and new ways to incorporate it are already on the horizon. These emerging solutions are worth watching whether a company can directly extract location data via its operations.

Those who generate and possess location data can sell it on the market, while those who do not can obtain it when necessary. To decide if location analytics is a worthwhile investment, companies need to answer two questions: can they monetize location analytics and can location analytics contribute to higher-level business goals.

To answer any of these questions, one must understand location data: where it comes from, what it is capable of and how to handle it safely and efficiently. This includes the ecosystem of data sources, aggregators and companies offering data engineering or analytics services. This is still a somewhat turbulent and volatile market, where finding the right partners, to purchase data from, to build a data processing infrastructure, to turn raw data into business insight or to monetize one's own data, is also key to realizing the potential of this opportunity and warrants careful inquiry.

About the author

Benedek Tóth

Working as a Data Scientist at HCLTech, Benedek extracts actionable insights from structured and unstructured data. His background is in statistics and economics, and he has previously worked at a business consulting firm researching financial technologies. At Starschema, he has developed time series and content recommendation models for enterprises in various industries and helped a major telecommunications company transform their location data to marketable analytics services.

HCLTech | Supercharging Progress™

hcltech.com

HCLTech is a global technology company, home to 222,000+ people across 60 countries, delivering industry-leading capabilities centered around Digital, Engineering and Cloud powered by a broad portfolio of technology services and software. The company generated consolidated revenues of $12.3 billion over the 12 months ended December 2022. To learn how we can supercharge progress for you, visit hcltech.com.