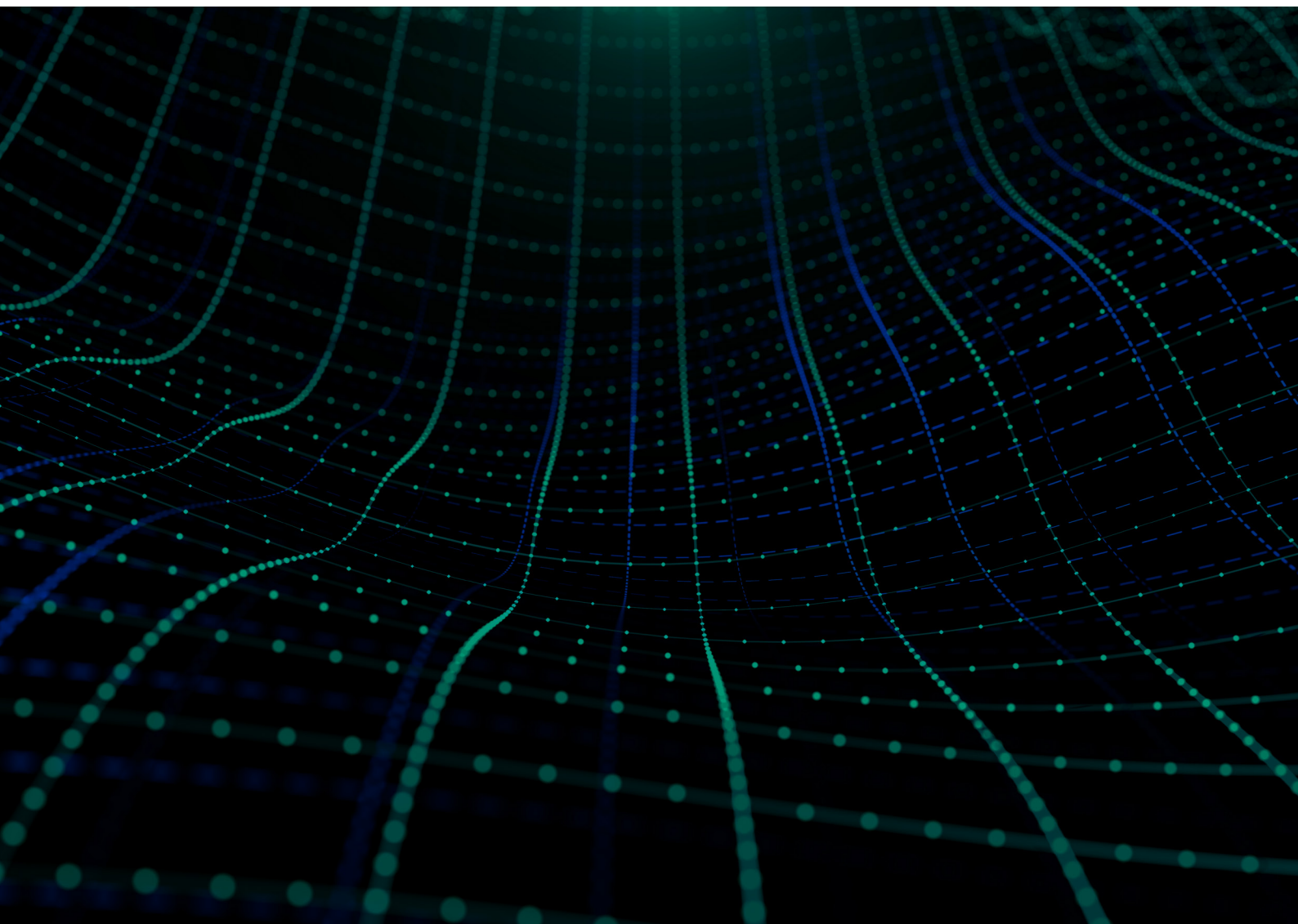


CAPCO

DO YOU STILL FISH IN MURKY WATERS?

DATA LAKE INSIGHTS TO HELP YOU LAND BIG-DATA FISH



A DATA LAKE

A centralized data repository - enables firms to connect data points as well as aggregate external data, to enable the use of powerful analytics tools and ultimately make faster and better evidence-based business decisions.

Today, both FinTechs and RegTechs offer financial institutions effective digital solutions for cumbersome requirements enforced by regulations. But which solution is the most suitable for your company? Most artificial intelligence (AI), cloud-based and digital solutions are only as good as the data they are fed with. Hence, major questions arise:

- What data do I have and where is it located?
- What is the quality of my data?
- Is my Swiss data safe in a cloud-based solution?

Data lake technology can support a bank's daily business by unifying internal and external data, to enable better analytics today and lay the foundation for AI and machine learning powered next generation analytics. This blog aims to help assess data lake use cases within an organization and provide specific insights into a data lake's added value.

WHY DO I NEED A DATA LAKE?

As a scalable and low-cost data repository, data lakes allow your organization to store various data types from different sets of sources and enable you to analyse your data to inform business decisions.

Today, a plethora of information is stored in an isolated manner and siloed in data warehouses or databases that are situated in multiple locations and jurisdictions. This leads to ineffective information flows and disharmonized databases within an organization. Further, this 'bad quality' data leads to many false positive results in compliance, reduced focus on real cases and ultimately fewer genuinely compliant results.

A data lake breaks down these data silos by centralizing and consolidating all of your company's batch and streaming data assets into one complete and authoritative data store for analytics that is continuously updated. This data repository or system is the primary source of information. With a data lake, not only does internal data get organized to enable better analytics, it can also be enriched with external data and external intelligence. By unifying all your data and upgrading it with external data in a data lake, your organization makes the first step towards being able to harness the power of machine learning and AI powered data analytics within the near future. Nearly all client-centric

organizations can benefit from implementation of a centralized data repository, be it to collect and analyse data from social media sites, customer relationship management (CRM) systems, or other sources. The collected and stored data can be used to gauge customer sentiment, adjust go-to-market strategies, mitigate customer-support problems, and even extend targeted offers to existing customers and prospects based on these insights.

As a centralized repository, data lakes enable you to store massive amounts of raw data in its native form, in a single location.

The main difference between a data lake and a data warehouse is that the latter cannot accommodate large data projects or fast-paced acquisition models in the same way that data lakes can. As such, data lakes provide the ideal solution to gather, capture, store, access and analyse enormous amounts of data in one location. In addition, traditional data warehouses rely on their own schema, i.e. a table row-column definition that determines how the data is organized upfront. This means that the attributes of the data must be known beforehand. Data lakes in contrast can store semi-structured data types in their native formats. The most important thing to understand in this context is not

necessarily how data lakes are constructed, but rather what they enable. In essence, data lakes are a comprehensive way to explore, refine and analyse petabytes of information that arrive from multiple data sources at a steady frequency.

Combining the power of analytics with the flexibility and agility of big data models will help unleash almost limitless resources of today's cloud-based solutions.

To exploit the advantages of a data lake to their fullest, the data lake must allow a company to easily store data in its native format, facilitate user-friendly exploration of data, automate routine data management activities, and support a broad range of analytics use cases. Hence, without adequate data quality and data governance, even well-constructed data lakes can

quickly become data swamps – unorganized pools of data that are difficult to use, understand and share with business users. Before diving deeper into the preliminary steps for building and implementing a data lake, let's first look at how data lakes are made.

HOW DO I BUILD A DATA LAKE?

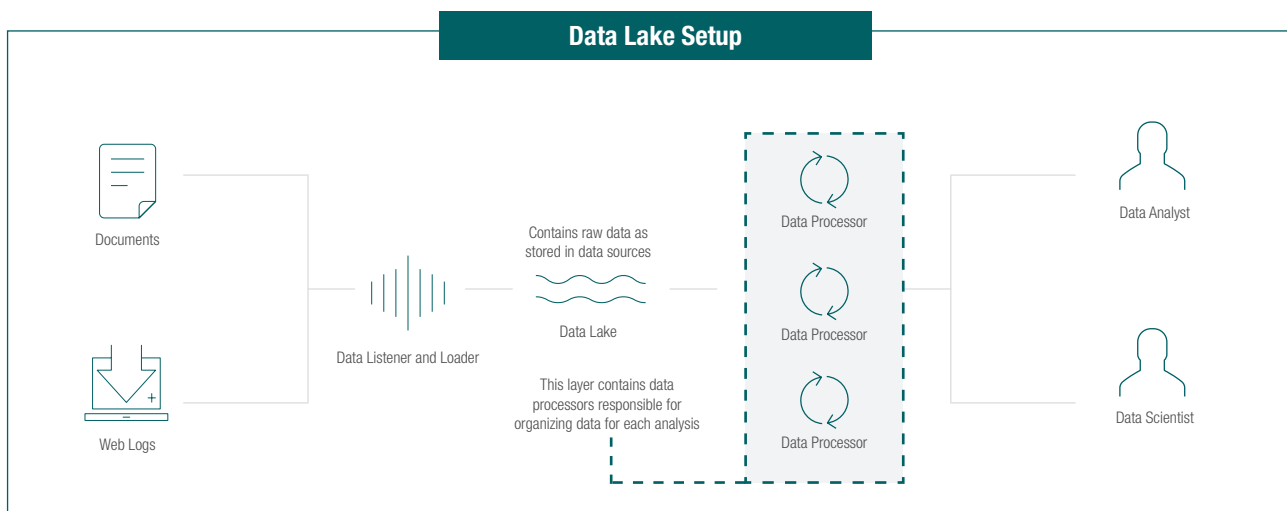
As mentioned earlier, a data lake can be described as a central data repository that can store traditional structured as well as unstructured, non-tabular raw data in its native format. Data lakes work by leveraging inexpensive object storage and open formats to enable various applications to take advantage of the data.

While some organizations benefit from interim data lake solutions from vendors such as Amazon Simple Storage Service (S3), Microsoft Azure Blob or Google Cloud Storage, other companies create their own data lakes from scratch by leveraging their own data storage environments. These highly elastic and sometimes

cloud-based storage solutions enable customers to store unlimited amounts of data in their native formats and conduct analytics. That way, customers no longer need to manage the hardware stack.

However, they still need to create, integrate, and manage their software environment by setting up procedures to transform data and establishing policies and procedures for identity management, data governance and security.

The final goal is then to understand how customers can obtain high-performance analytics.



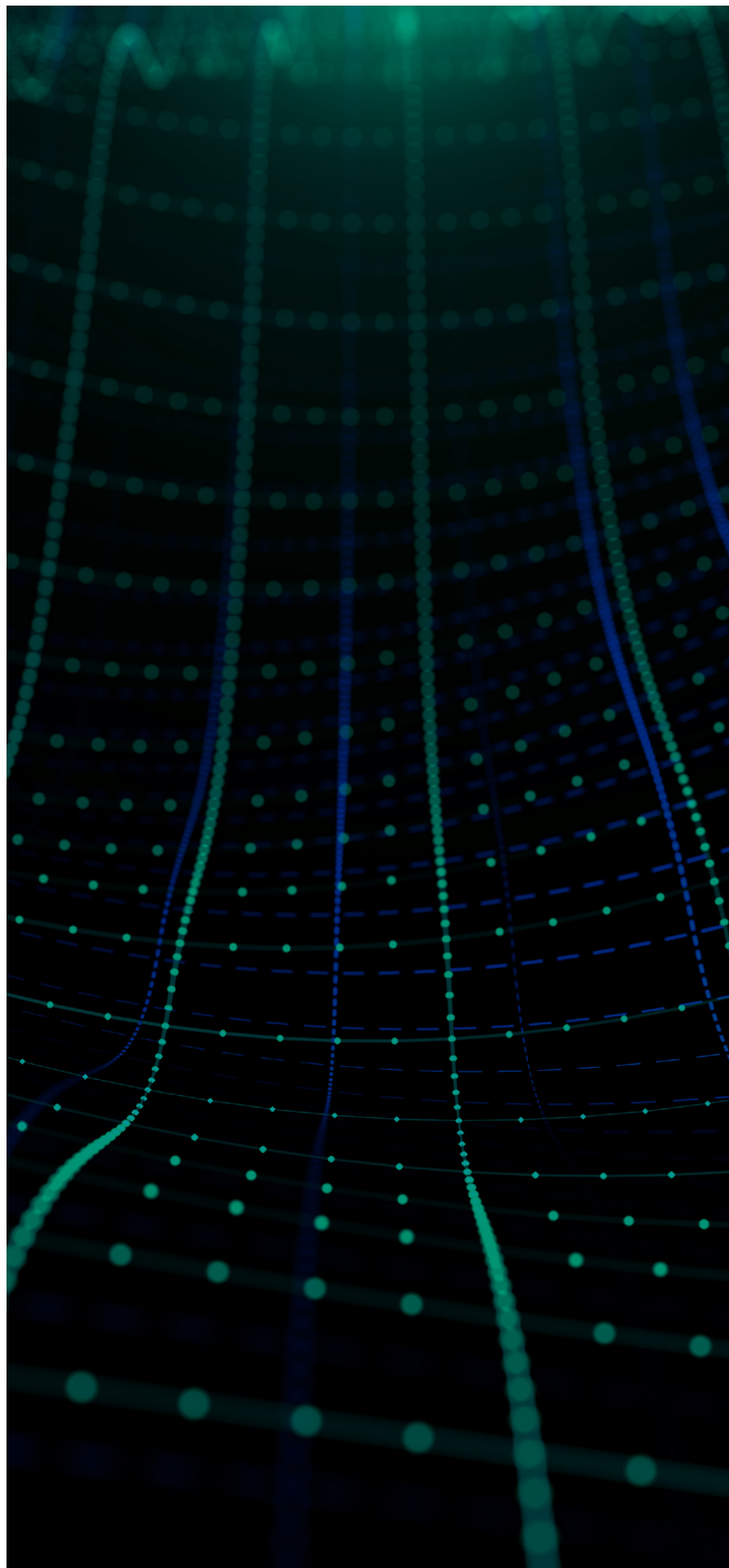
Data quality is a vital prerequisite for getting the most from a data lake.

Prior to planning the implementation of a data lake, it is essential to ensure good data quality. This can be achieved by making sure that the data that is fed into the data lake is collected, described, cleaned, and investigated in an appropriate and precise manner. These four steps form the data pre-processing.

Data lakes offer foundational tenets to overcome challenges such as inherent complexity, poor performance and the lack of governance among existing data warehouses and siloes.

Data lakes key attributes include: (1) instant elasticity, (2) concurrent operations, (3) embedded governance, (4) transactional consistency, (5) no silos, and (6) full management. The first refers to the fact that any amount of computing resources can be supplied to any user or workload and that the size of a compute cluster can be changed dynamically without affecting running queries. Also, the service can be scaled easily to include additional compute clusters and complete intense workloads faster. The second describes the possibility to deploy to an almost unlimited number of users and workloads to access a single copy of your data without affecting performance. Embedded governance means that users are presented with fresh and accurate data and that the focus is placed on collaboration, data quality, access control and metadata management. The fourth term includes the trustfully combination of data to enable both multi-statement transactions and cross-database joins. The fifth point refers to the fact that data lakes are easily capable of ingesting petabytes of structured, semi-structured, and unstructured data into one single repository. Lastly, data lakes enable full management, meaning that with software-as-a-service (SaaS) solutions, data platforms themselves largely manage and handle provisioning, data protection, security, backups, and performance tuning. This in turn allows you to focus on analytic endeavours rather than on managing hardware and software.

Data lakes, however, do not come in one single shape and format. Depending on the organization's preferences and the data to be evaluated, data lakes offer various solutions which accommodate different needs and requirements. The next section focuses on the important decisions that must be taken when implementing a data lake.



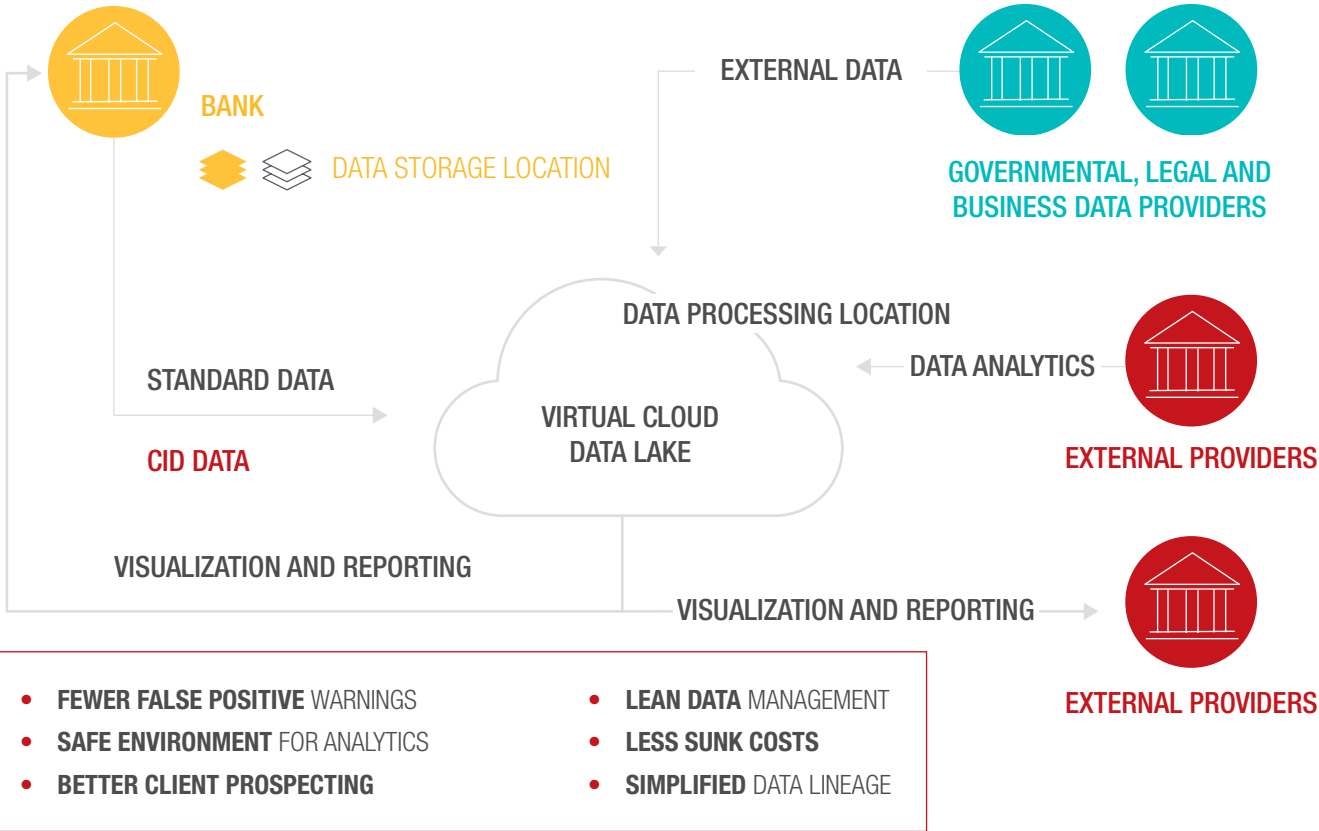
IMPORTANT DECISIONS FOR DATA LAKE IMPLEMENTATION

From a Swiss perspective, data security is key.

Experience shows that NextGen solutions from RegTechs and FinTechs realize their full potential only with the input of the right data. As outlined in earlier sections, this can be addressed with a data lake, however questions remain regarding what kind of data lake an organization should build to fit the purpose. From a Swiss perspective, one key consideration for this

decision is data security, especially when it comes to data for identifying clients which is seen as ‘the gold’ in the 21st century.

A virtual, cloud-based data lake provides a solution for securely enriching internal data with external data and having it evaluated outside of the organization, to benefit from wider data sets and advanced analytics tools.



A possible set-up for a virtual cloud data lake

A virtual data lake retrieves data from its original storage and processes it only for a specific purpose (e.g. analytics). It enables a bank’s data to be enriched with external data (e.g. AML measures) but without being permanently stored in any location other than the original storage. This means that your organization can use sensitive data within a global environment, benefiting from modern analytics but keeping the data safe on its own premises.

Another key consideration is the location of data within your organization. To work within a virtual data lake, your data can remain heterogeneous and in any location, which means high

maintenance for each storage and extreme maintenance to keep all data flows up to date. Here, we recommend creating a well-organized data structure via a cloud-based virtual data lake. A cloud-based solution allows homogenous database within your entire organization, with lower maintenance.

The final dilemma is where to build the cloud, i.e. on or off premises? This decision boils down to a cost versus risk decision. On-premise cloud allows higher control but comes at a higher cost, particularly in Switzerland, hence each organization needs to evaluate the options based on their own parameters.

To conclude, we believe that cloud-based virtual data lakes help organizations enable NextGen analytics while keeping valuable and sensitive data safe. Not only does this boost KYC and AML functions, but it also helps take customer intelligence and client prospecting to a whole new level.

Contact us to learn more about data lakes, their benefits, use cases and implementation best practice. Capco can help your firm assess its data lake readiness and provide a strategic implementation roadmap and implementation support.

EXPERTS & AUTHORS

EUGENIA COLLOVÀ

Data Scientist

M: +41795325893

E: eugenia.collova@capco.com

MARC BERNER

RegTech Expert

M: +41791358521

E: marc.oliver.berner@capco.com

CONTACT



DR. INGO RAUSER

Partner

M: +41444343515

E: ingo.rauser@capco.com

ABOUT CAPCO

Capco is a global technology and management consultancy dedicated to the financial services industry. Our professionals combine innovative thinking with unrivalled industry knowledge to offer our clients consulting expertise, complex technology and package integration, transformation delivery, and managed services, to move their organizations forward.

Through our collaborative and efficient approach, we help our clients successfully innovate, increase revenue, manage risk and regulatory change, reduce costs, and enhance controls. We specialize primarily in banking, capital markets, wealth and asset management and insurance. We also have an energy consulting practice in the US. We serve our clients from offices in leading financial centers across the Americas, Europe, and Asia Pacific.

To learn more, visit our web site at www.capco.com, or follow us on Twitter, Facebook, YouTube, LinkedIn, Instagram and Xing.

WORLDWIDE OFFICES

APAC

Bangalore
Bangkok
Gurgaon
Hong Kong
Kuala Lumpur
Mumbai
Pune
Singapore

EUROPE

Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
London
Munich
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA

Charlotte
Chicago
Dallas
Hartford
Houston
New York
Orlando
Toronto
Tysons Corner
Washington, DC

SOUTH AMERICA

São Paulo

WWW.CAPCO.COM



CAPCO