2022 Tech Sprint on Confidential Data Pooling

Call for Applications

May 16, 2022
Fintech-Innovation Hub

Application deadline: June 3, 2022
Send applications to: techsprint2022@acpr.banque-france.fr

Informal translation. If needed, please refer to the French version, which prevails.

1. Tech Sprint context, objectives and principles

Context and objectives

AML-CFT (Anti-Money Laundering and Countering the Financing of Terrorism) is a privileged area for exploring the use of innovative solutions in the financial sector. The administrative workload and the induced backlog, the volume of false positives triggered by currently implemented procedures are among the main incentives to look for efficiency gains enabled by technical innovations.

In this context, a recurring theme among financial actors pertains to information sharing and data pooling (including customer information and transactional data). However, most projects to date have been limited in scope or scale and their results have failed to materialize.

Nonetheless, when considering the narrower scope of advanced data analytics (AI included) applied to suspicious transaction detection, data sharing or pooling has been largely unexplored thus far even though it might significantly improve the overall performance among financial actors. The FATF (Financial Action Task Force) even dedicated a report to this very theme – and presented its conclusions at an ACPR webinar in March 2021[1]. The report reaffirms the FATF's conviction that data pooling and collaborative analytics would improve the efficacy and efficiency of AML-CFT in general[2].

The ACPR therefore decided to launch **an experimentation** whose main objective is to prove – or infirm – the hypothesis that data pooling enhances the performance of transaction monitoring systems. The objectives, principles and roadmap of the experimentation were presented to the public on March 30. The annex of this document also provides an overview of the experimentation.

Questions raised by the various constraints associated to that primary objective led the ACPR to organize, as part of the experimentation, an event called the **2022 Tech Sprint on Confidential Data Pooling** (CDP). The Tech Sprint event aims at **evaluating the methods and techniques enabling to maintain the confidentiality (and even integrity) of pooled data**. The present Call for Applications is directed at technology providers that would be in capacity to participate in the Tech Sprint.

Principles

Tech Sprint participants are invited to **design and implement a PoC** (Proof of Concept) of their CDP solution, using fictitious data and a demo scenario provided by the ACPR.

The results of their work shall be presented to the public on Sept. 13, 2022. A panel of technical experts in CDP, composed of representatives from the ACPR, Banque de France and FIs participating in the overall experimentation (of which the Tech Sprint is one sequence) shall attend the Tech Sprint Demo Day. Results may be presented in the form of a live demo of the PoC implementation accompanied by a pitch, and shall be followed by a Q&A session.

Providers participating in the Tech Sprint will also be asked:

- To provider a **documentation of their PoC implementation by Sept. 5**.

[1] "AI & Finance Conferences with the ACPR" - Data Sharing in Finance

[2] *STOCKTAKE ON DATA POOLING, COLLABORATIVE ANALYTICS AND DATA PROTECTION*, GAFI 2021. *"By pooling data and using collaborative analytics, financial institutions can better understand, assess, and mitigate money laundering and terrorist financing risks. This will result in a more dynamic, effective and efficient identification of these activities, and help the private sector comply with anti-money laundering and counter terrorist financing requirements in a timelier and less burdensome manner."*

- To **respond to any deep-dive questions** that the panel of technical experts may ask them during the **Demo Day follow-up period (Sept. 13 - Sept. 20)**.

Participation in the CDP Tech Sprint is **free and non-remunerated**.

The Tech Sprint also aims to **elicit partnerships between technology providers and FIs participating in the experimentation**, in view of the implementation and execution of experimental protocols on real data (as per stages 4 to 6 of the experimentation, see annex in Section 6). However, **participation in the Tech Sprint shall not dictate partipating FIs' eventual choice** of their technology provider for those subsequent stages of the experimentation.

Technology providers selected to participate in the Tech Sprint **shall abstain from claiming any endorsement of their technological approach**, their products or services, and more generally of their business.

The ACPR considers **communicating to the marketplace**, in an anonymized form, the main takeaways it will have gleaned from the experimentation in general, and from the Tech Sprint in particular. It also reserves the right to communicate the identity of the organizations who will have participated in any stage of the experimentation, including the Tech Sprint.

2. Detailed description of the CDP Tech Sprint

Timeline

Here is the indicative timeline for the CDP Tech Sprint:

- May 16: **Publication of the Call for Applications (CfA)**.
- May 16 - June 3: **Response to the CfA.** Each interested provider responds with a description of its solution using the application template provided in Section 3 of this CfA.
- June 3 - June 13: **Tech provider selection process.** The ACPR determines which of the candidate providers will be invited to participate in the Tech Sprint.
- June 13: **Communication of the Tech Sprint PoC Requirements Document.** The ACPR sends each pre-selected provider the Tech Sprint PoC Requirements Document (which will include functional and technical requirements related to the PoC implementation, along with a link for downloading fictitious input data).
- June 13 – Sept. 13: **Implementation of the PoC.** Each pre-selected provider implements its solution to the PoC within a 3-month timeframe.
  - June 27: **Q&A session on PoC requirements.** All participating providers are invited to a briefing session hosted by the ACPR and aimed at answering any pending question raised by their reading and analysis of the PoC requirements document.
  - Sept. 5: **Documentation of the PoC implementation.** Each participating provider has sent the documentation of its PoC work to the ACPR (based on a questionnaire provided by the ACPR). This documentation will specifically be used by the panel of technical experts on CDP (composed of representatives from the ACPR and participating FIs) to prepare their evaluation on the Tech Sprint Demo Day.
- Sept. 13: **Tech Sprint Demo Day.** On the Tech Sprint Demo Day, each participating provider presents its solution and PoC implementation (along with any supporting proof), within a 20-minute time slot, followed by 10 minutes of Q&A. The audience will include all FIs participating in the AML-CFT experimentation, representatives from the ACPR and Banque de France, along with selected external guests from the financial sector, academia, public institutions, and technical experts. Besides, a panel of technical experts on CDP (including representatives from participating FIs, the ACPR and Banque de France) will rigorously assess the technical capabilities and limitations of each proposed solution.
- Sept. 13 - Sept. 20: **Demo Day Follow-up.** Each participating provider remains available to answer in-depth questions by participating FIs or the ACPR about their PoC implementation or more generally their technology and solution.

The provisional timeline **post-Tech Sprint** is as follows:

- Sept 13. - Sept. 30: choice of solutions providers. Each participating FI will be invited to choose a CDP technology provider among Tech Sprint participants in order to implement and execute each experimental protocol defined during the experimentation. Each participating FIs shall however retain full discretion in their decision to select or not a given candidate provider, and shall reserve the right not to disclose the reasons of any favorable or negative outcome towards a candidate.
- Sept. 30 - Q1 2023: validation and execution of the experimental protocols. Technology providers who have reached an agreement with participating FIs also participate in the following stages of the ACPR experimentation.

- Report on the CDP Tech Sprint: a synthesis of works by the panel of experts on CDP shall be published by the ACPR, either as part of a general report on the experimentation, or as a stand-alone report.

Relevant technologies

The term “Confidential Data Pooling” (CDP) aims to describe as accurately as possible the type of technical solution sought for the purpose of this ACPR experimentation. In short, the goal is to **enable storing, joining, querying, and feeding to predictive models multiple datasets with confidentiality requirements**. In this case, each dataset contains transactional data from a specific FI, although most CDP technologies are expected to be domain-agnostic.

The scope of CDP is expected to have a significant overlap with PET (Privacy Enhancing Technologies), however a dedicated term is used on purpose in order not to limit the range of techniques that can be proposed by applicants. **The objective pursued is indeed purely functional: to fulfil the confidentially, privacy and integrity requirements** of all stakeholders (first and foremost the participating FIs which will contribute commercially- and potentially privacy-sensitive data).

Nonetheless, CDP solutions considered as likely relevant include most PET approaches: Differential Privacy, Homomorphic Encryption, Secure Multi-Party Computation, Trusted Execution Environments, or any combination thereof.

Selection of candidates for the CDP Tech Sprint

Any interested entity (whether it is legally part of an FI or not) may send its application by June 3 using the application template provided in Section 3 of this document.

The following table provides an overview of the mandatory and optional functionalities required from CDP providers to be pre-selected.

Requirements area	Mandatory requirements	Optional requirements
Confidential data pooling	Automated data pooling with confidentiality guarantees	Integrity guarantees across the data pooling process
Confidential data pooling	API with SQL support for pooling queries	Support for non-SQL query languages
Confidential data pooling	Error checking capability (e.g. to debug queries on encrypted or enclaved data)	Extraction of a sample of pooled data (for spotchecking the results)
Integration with AML-CFT detection models	Model integration API in at least one mainstream language (e.g. Python) or protocol (e.g. REST)	Built-in implementation of certain model classes (ML-based or not, graph-based or not) without requiring integration of an external model
Integration with AML-CFT detection models		ML model training on pooled data
Integration with AML-CFT detection models	Model inference on pooled data, for any class of model (ML or not)	
Integration with AML-CFT detection models	Model performance metrics (custom or ad-hoc) evaluation	
Hosting	Self-hosting by the provider (on-premise, on private cloud, or on public cloud)	Other hosting option (on a participating FI's or a trusted party's data center)

The ACPR will examine all applications and select a list of candidate solution providers. This process will take into account the mandatory and optional functionalities as indicated above, as well as each provider’s estimated ability to conduct the proposed experiment, and other factors including but not limited to: the degree of technical and scientific innovation of the solution proposed, its levels of maturity and robustness, the added value as evidenced by prior production deployments, the cost of the solution, etc.

The ACPR retains full discretion in its decision to select or not a given candidate provider, and reserves the right not to disclose the reasons of any negative or positive outcome.

PoC requirements

The requirements will be described in a CDP Tech Sprint Requirements Document sent to each pre-selected provider on June 13.

Those requirements will mainly consist of implementing a PoC solution according to a pre-defined scenario, using fictitious data. The goal of the PoC will be to convince one or more participating FIs that the solution can, with a limited amount of additional development or customization (i.e. mostly integrating with the FIs' system in stage 4), meet the needs for executing the experimental protocol itself (stage 5, performed on real data). The "P" in "PoC" means quite literally **delivering a *proof* on the Demo Day** (in addition to the demoing the solution itself) that the solution operates as described and meets the confidentiality, performance and other guarantees assessed by the provider.

Since providers participating in the Tech Sprint will only have 3 months to implement the PoC (June 13 - Sept. 13), the definition of the scope and requirements will aim to avoid the tasks that are usually the most time- and resource-consuming for CDP providers:

- Defining the use case: the ACPR will provide a simplified, hypothetical scenario, which will however enable testing both mandatory and optional requirements.
- Data sourcing: the ACPR will provide participants with input data in the form of fictitious datasets.
- Deployment and hosting: to avoid the lengthy, strenuous process (from a legal, technical and organizational perspective) of deploying within an FI's IT environment or even on a trusted third party's data center, participants will be free to use any deployment target. Some are expected to deploy and host their PoC implementation on their own data center, while others might resort to their standard deployment method on a public cloud (likely with no additional security concern given that PoC data and models will be completely non-sensitive).

3. How to apply to the Tech Sprint

Application process

This section describes the modalities for interested solution providers to submit their application to the ACPR, which must be done **by June 3**.

Each interested provider should include in their application the coordinates of a main point of contact within their organization (first and last name, email address) along with the application template below filled as accurately as possible.

Application template

This template is composed of a number of questions and sub-questions. Each of these should be answered in free-form text, as accurately and comprehensively as possible by the provider. Furthermore the provider may be requested to demonstrate any element or argument put forward in its answers.

The scope of each answer should take into account the full perimeter of the experimentation. Beyond the solution currently proposed by the provider, the answers should focus on what can be achieved both:

- for the Tech Sprint PoC (assuming minimal or significant customization of the existing solution but within the 3-month implementation timeframe for the PoC)
- and for a potential experimentation on real data (which will introduce additional scalability, confidentiality, and systems integration requirements).

The template is broken down in three sections: functional characteristics of the solution, technical characteristics, finally the cost and planning aspects.

A. Functional characteristics of the proposed solution

1) **Input data.** Specify any constraints:
   a) on the data types (categorical, numerical, short text) that can be processed by the solution
   b) on the data schemas (e.g. any combination of relational tables)
   c) on the data volumes

2) **Analytical features.** For each of the following functionalities, specify if it is part of the proposed solution. If so, indicate the corresponding *modus operandi* (i.e. what manual or automated steps are necessary compared to a baseline operation on data in clear). Also provide any available metrics on its operational performance (i.e. compare the user experience using your solution vs. just operating on data in clear).
   a) Execution of SQL queries
      i) Loading input data
      ii) Joining input data from multiple FIs
      iii) Drawing a random sample of input data (pre- and post-join)
   b) Built-in implementation of detection models
      i) Creation and configuration of a rule-based model on tabular data
      ii) Creation and configuration of a rule-based model on graph (network) data
      iii) Creation and training of an ML model on tabular data
      iv) Creation and training of an ML model on graph (network) data
   c) Integration of third-party ML models
      i) Loading an untrained model into the Confidentiality Perimeter

   ii) Training a model inside the Confidentiality Perimeter
   iii) Running inference in batch mode inside the Confidentiality Perimeter
   iv) Post-inference: computing and retrieving performance metrics (minimally descriptive statistics)
   v) Post-inference: drawing a random sample of input (joined) data along with their output predictions
   vi) Downloading the trained model
   vii) Which ML model classes are supported? What other constraints on the model?

d) Other types of analytics (i.e. besides SQL queries and ML models)
   i) Which types are available?
   ii) In which programming language or format?

3) Confidentiality and privacy guarantees. What provable guarantees does your solution offer on the following sensitive elements?

a) On datasets uploaded to the Confidentiality Perimeter, then pooled with other input datasets (using SQL joins or another Pooling Mechanism), and possibly fed to ML models as training and test data? For example:
   i) Does it enable synthetic data generation with Differential Privacy guarantees, so as to enable training an ML model outside the Confidentiality Perimeter?

b) On the results of SQL queries? For example:
   i) Can the query rate be throttled? Can the result type be limited to restrict queries to computing aggregated values e.g. descriptive statistics?
   ii) Is this achieved using Differential Privacy on the result dataset or another scheme?

c) On the pre-trained ML model? For example:
   i) Does it provide Differential Privacy guarantees on ML model parameters (such as weights in the case of a linear or logistic regression)?
   ii) Does it at least enable computing prediction metrics, and if so which ones (precision, recall, etc.)?
   iii) Does it enable downloading a sample of test data along with the predictions?

4) **Threat model.** Describe as accurately as possible the threat (and attacker) model under which your solution operates.

5) Integrity guarantees. What integrity guarantees does your solution offer (in order to prove that a given piece of code was executed, that data have not been tampered with, etc.) against the following sensitive elements?

a) On all data hosted in the Confidentiality Perimeter (input datasets and pooled data)?
b) On any SQL queries executed inside the Confidentiality Perimeter?
c) On the source code used for creating, training and testing the ML models?

B. Technical characteristics of the proposed solution

6) **Solution hosting.** Specify any constraints:

a) On the hosting environment (on-premise only or not, which host OS(es) are supported, with mandatory or optional containerization, with mandatory or optional orchestration, etc.)
b) On the storage space (limits on source data, incurred overhead, etc.)
c) On the environment monitoring (live logging capability, post-hoc audit, etc.)

7) **Integration.** Describe the integration capabilities with a target TMS (Transaction Monitoring Systems):
   - What TMS types are currently supported by the solution?
   - What additional TMS types can be supported (within the time and resource limits imposed by the ACPR experimentation)?

8) **Type of technology.** What type(s) of technology are leveraged by the solution?
   a) HE/FHE (Homomorphic Encryption)
   b) SMPC (Secure Multi-Party Computation)
   c) DP (Differential Privacy)
   d) TEE (Trusted Execution Environment) and if so, on which platform(s)
   e) ZKP (Zero-Knowledge Proof)
   f) FL (Federated Learning)
   g) A combination of any of those?
   h) Another method?

9) **Implementation**
   a) Open or closed source?
   b) Based on patented technology or public algorithms?
   c) What algorithms?

C. Cost and planning of the proposed solution

10) **Timeline.** What is the expected timeframe for the following phases?
   a) For implementing the Tech Sprint PoC (i.e. using your solution to execute a fictitious scenario on fictitious data, both provided by the ACPR as described in Section 2)
   b) For potentially running the experimentation on real data (i.e. integrating your solution with a partner FI's systems, then deploying it, finally executing the experimental protocol on their real data).

11) **Financial cost.** What is an initial estimation of the financial cost for adapting your solution to potentially running an experimentation on real data? (Note that the CDP Tech Sprint, including the PoC implementation, is expected to be realized cost-free by all candidate providers.)

12) **Partnership resources**. What resources (technical, functional, logistic or otherwise) are expected:
   a) From partner FIs?
   b) From the ACPR as facilitator of the experimentation?

13) **Internal resources.** For information only: what resources will be provisioned by the technology provider
   a) To implement the Tech Sprint PoC?
   b) To adapt, if needed, your solution to a potential experimentation on real data, then execute a partner FI's experimental protocol?

Processing of personal data

The ACPR shall manage the list of Tech Sprint candidates, with the explicite purpose of organizing the Tech Sprint described in the present Call for Applications. This data processing is based on legitimate interests and complies with the relevant legal and regulatory dispositions, namely European Regulation n° 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data ("GDPR") and French Law n° 78-17 of 6 January 1978, as amended, relating to information technology, files and freedoms.

The purpose of the information requested within the application process is exclusively limited to processing the list of candidates during the Tech Sprint organization, and is solely intended for the ACPR administration. Within this scope, the Authority collects personal data: first and last name, email address of representatives from candidate organizations. Those data shall be retained for at most 2 years.

Only members of the Fintech-Innovation Hub, internal controls and internal audit teams shall be authorized to access the information pertaining to Tech Sprint candidates.

Legal claims (access, rectification, erasure, right to object) may be exercised by emailing the ACPR at fintech-innovation@acpr.banque-france.fr.

A complaint may also be sent to the CNIL (data protection authority). Banque de France's CNIL-registered data protection officer can be reached at 1200-DPD-delegue-ut@banque-france.fr.

4. Next steps for interested providers

This section describes, at a high level, what would be asked in practice from respondents to this call for applications.

- To determine, on the basis of this document, **whether they are interested** and have the capacity to respond to the call for applications.
- If so, to fill in the **application template** (Section 3) as precisely as possible so that the ACPR will be able to determine whether the provider is a relevant and reliable candidate for the experimentation.
- If the respondent is selected, to participate in the CDP Tech Sprint which essentially constitutes a **PoC** (Proof of Concept) of their solution, concluded by a Tech Sprint Demo Day during which they will be able to present their solution.

As for **potential follow-ups** to the Tech Sprint, the next steps will be as follows:

- Following the Tech Sprint presentations, FIs participating in the experimentation may decide to **choose a solution provider.** Technology providers and participating FIs shall then define the details of their collaboration (contractual agreement, cost of the implementation, etc.)
- The provider will **work with the rest of the team** on the remainder of the experimentation: finalizing the experimental protocol and implementing/tailoring its solution to fit that protocol, running the experimentation on real data, and evaluating the experimental results.

5. Glossary

Confidential Data Pooling (CDP): as described in Section 2 of this document, this term (coined by the ACPR for the experimentation) aims to describe any technology which enables storing, joining, querying, and feeding to AML-CFT models a number of sensitive datasets. Sensitive datasets in this context mean any type of data - primarily transactional - with confidentiality, privacy and integrity requirements by all stakeholders in the experimentation. The scope of CDP is expected to have a significant overlap with PET (Privacy Enhancing Technologies), however a dedicated term is used on purpose in order not to limit the range of techniques that can be proposed by applicants.

Confidentiality Perimeter: for the purposes of describing a CDP solution, this perimeter generically refers to a physical or logical data storage and analysis area that provides appropriate confidentiality guarantees. For example in the case of end-to-end encryption this would comprise the entire experimental protocol (both storage area and data flows) following initial encryption of sensitive data.

Differential Privacy (DP): one possible family of techniques for performing CDP. See Wikipedia entry.

Federated Learning (FL): one possible family of techniques for performing CDP. See Wikipedia entry.

Financial Institution (FI): each entity with AML-CFT obligations which participates in the CDP Tech Sprint. Those entities primarily belong to the Credit Institution, Payment Institution, and Insurer categories. Different categories will correspond to different types of data and thus different experimental protocols, but different teams of FIs within the same category might also opt for different protocols (e.g. studying Random Forest models on the retail banking segment in one team, and studying rule-based models on the corporate banking segment in another team).

Fully Homomorphic Encryption (FHE): one possible family of techniques for performing CDP. See Wikipedia entry.

Machine Learning (ML): any AML-CFT detection model relying on any class of ML (linear or logistic regression, random forest or gradient boosted trees, etc. – although probably not deep neural networks) or any combination of such classes.

Pooling Mechanism: the method, and its implementation, used in the ACPR experimentation by a team of co-participating FIs to pool their respective datasets. In order to assess the entire improvement margin of data pooling, the ACPR will encourage participants in the experimentation to explore the more sophisticated pooling mechanisms. Those might for example be based on any number of complex SQL queries including joins, and potentially fuzzy joins, as opposed to more traditional pooling techniques (e.g. simply multiple datasets with identical schemas, or sharing minimal information such as Boolean values representing a “suspicion” flag).

Secure Multi-Party Computation (SMPC): one possible family of techniques for performing CDP. See Wikipedia entry.

Trusted Execution Environment (TEE): one possible family of techniques for performing CDP. See Wikipedia entry.

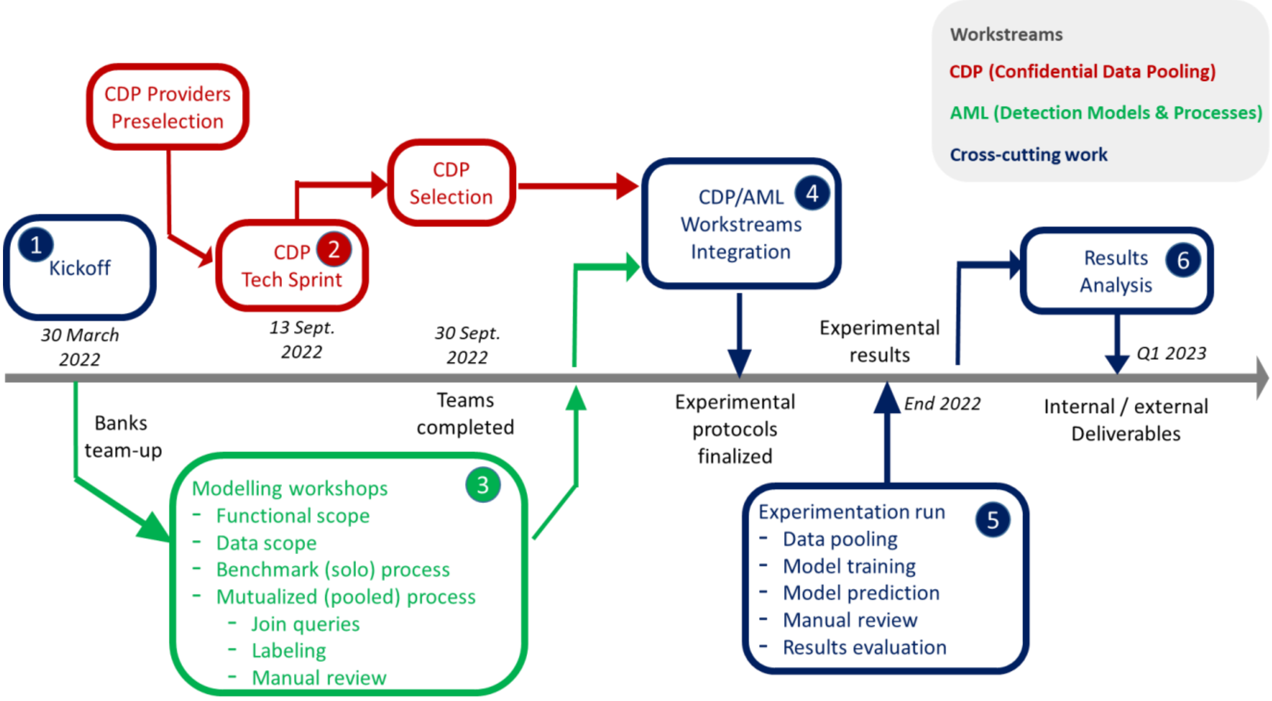
Zero-Knowledge Proof (ZKP): one possible family of techniques for performing CDP. See Wikipedia entry.

6. Annex: experimentation overview

This section provides context relevant to this Call for Applications in the form of an overview (already presented at the kickoff meeting on March 30) of the entire experimentation.

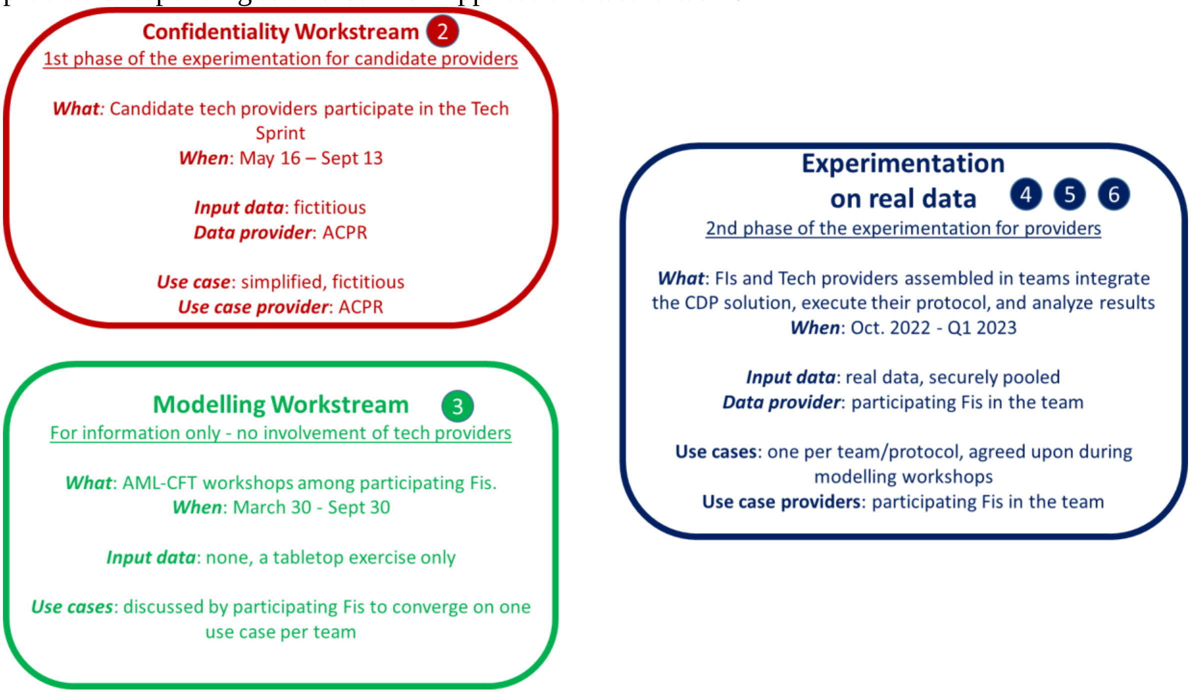
Overall roadmap

The following diagram summarizes the roadmap for the entire experimentation. Each of the six main stages is then described in more detail in the remainder of this section. For technology providers considering responding to this CfA, stages 2 (called the CDP Tech Sprint) and stages 4-6 (called the experimentation on real data) constitute the main areas of involvement.



Roadmap for technology providers

The following diagram summarizes the scenarios (input data and use cases) on which technology providers responding to this Call for Applications would work:



Confidentiality Workstream 2

1st phase of the experimentation for candidate providers

What: Candidate tech providers participate in the Tech Sprint

When: May 16 – Sept 13

Input data: fictitious

Data provider: ACPR

Use case: simplified, fictitious

Use case provider: ACPR

Experimentation on real data 4 5 6

2nd phase of the experimentation for providers

What: FIs and Tech providers assembled in teams integrate the CDP solution, execute their protocol, and analyze results

When: Oct. 2022 - Q1 2023

Input data: real data, securely pooled

Data provider: participating Fis in the team

Use cases: one per team/protocol, agreed upon during modelling workshops

Use case providers: participating Fis in the team

Modelling Workstream 3

For information only - no involvement of tech providers

What: AML-CFT workshops among participating Fis.

When: March 30 - Sept 30

Input data: none, a tabletop exercise only

Use cases: discussed by participating Fis to converge on one use case per team

Kickoff meeting

The kickoff took place on March 30, with an audience comprising all FIs and CDP technology providers who had already expressed an interest and were available to attend the kickoff meeting. The ACPR acted as organizer and presented the objectives pursued by the experimentation, the overall roadmap, and each of its six main stages in a moderate level of detail. A Q&A session allowed to clarify the most pressing issues.

CDP Tech Sprint

Planned timeframe: May-Sept 2022 (13 Sept. for the Tech Sprint Demo Day)

Participants: all CDP providers who responded to the present call for applications. FIs participating in the experimentation will attend the Tech Sprint Demo Day; they will be represented within the panel of technical experts on CDP so as to integrate, as appropriate, technology providers into the teams in charge of executing the experimental protocols.

Description. The objective of this stage is to evaluate methods and techniques for maintaining confidentiality and integrity of pooled data. This stage should also allow to elicit partnerships between FIs (who may have already assembled in teams of two or more) and technology providers. The providers will thus be invited to establish a PoC of their solution, including how it can be tailored to the proposed experiment.

AML-CFT modeling workshops

Planned timeframe: March-Sept. 2022

Participants: all participating FIs, assembled in teams, with the ACPR as facilitator.

Description. A functional workstream on AML-CFT modelling will be set up in parallel with the workstream around the CDP Tech Sprint. This workstream will have a very different format, namely a series of modelling workshops aiming to prepare the experimentation itself. They will gather FIs organized as teams of two or more, not including tech providers, therefore this stage is not described in detail in this document.

In essence, the objective of this stage is for each team to define a “logical protocol”[3] for the experimentation, i.e. one that can be precisely articulated on paper, without any implementation work nor real data processing (the implementation will only come in stages 4 and subsequent, once some technology providers have potentially been chosen by FIs).

AML / CDP workstreams integration

Planned timeframe: Oct.-Nov. 2022

Participants: each team (participating FIs, with some CDP technology providers potentially integrated into each team) working on its protocol definition and implementation, with the ACPR facilitating as needed and as available.

Description. Once the CDP and AML-CFT workstreams (stages 2 and 3) have converged, i.e. once each bank has had the possibility to choose a technology provider, this stage will consist for each team to reach a two-fold agreement. Firstly, each team will define its modus operandi for the experimentation: with the “logical protocol” finalized, material conditions will be specified (including the question of hosting each CDP solution and the data they will be fed), as well as a calendar for the remainder of the experimentation. Secondly, a legal agreement between team members will be carved into a convention specifying each partner’s responsibility (in terms of data and model management, splitting costs of the experimentation, etc.)

The ACPR will here as well act as facilitator of both objectives: in order to ensure consistency across experimentations set up by each team and with the overall objective of the experimentation, as well as adequacy with the ACPR roadmap, it will validate both the experimental protocols and the timelines proposed. From a legal perspective, without impeding on the each entity’s responsibilities and duties, it may be party to the legal agreements between banks and providers as a trusted third-party.

Lastly, this stage will lead to developing the solution (software, hardware or hybrid) on which each experimental protocol will be executed – which may be seen as the “physical protocol” corresponding to the “logical protocol” defined in stage 3. This implementation should include on the one hand the AML-CFT components defined in the modelling workshops, on the other hand the integration between the provider’s CDP solution and those components.

Experimentation run

Planned timeframe: Dec. 2022

Participants: each team (participating FIs + selected technology providers) executing its protocol, with the ACPR facilitating as needed and as available.

Description. This stage will follow the finalization of experimental protocols. Each team will execute its protocol, which typically will include the following sequence of steps:
- Data pooling according to the mechanism agreed upon
- ML Learning in the case of detection models based on ML
- Batch prediction using both processes (benchmark process / mutualized process)

[3] To use an analogy with logical data schemas in the realm of relational databases

- Manual review of a sample of outputs from the mutualized process
- Computation of evaluation metrics.

Results analysis

Planned timeframe: Q1 2023

Participants: each team for the most detailed sharing sessions; all teams assembled for sharing information across experimental protocols; and the broader public for external communication.

Description. The restitution will be delivered at 3 different levels:

- For each team, an in-depth post-mortem analysis of the experimentation
- An event including all participants
- A public event, followed by a summary report, to communication any information that can be shared publicly.