Getting Started with DEPA

DEPA : Data Empowerment And Protection Architecture

Table of Contents

Table of Contents **1**
Data Economy **1**
Background **3**
India transition to a “Data Modelling Economy” 3
Data Sharing 3
Need for DEPA 4
DEPA Genesis 5
DEPA 2.0 Use cases **7**
Health 7
E-Commerce 8
DEPA 2.0 Ecosystem **10**
DEPA 2.0 Primitives **11**
Electronic Contracts (EC) 12
Confidential Compute (CC) 15
DEPA Training Architecture **17**
Workflows **18**
Participant registration 18
Dataset registration 18
Dataset discovery 18
Payments 18
Conclusions **18**
References **19**

Data Economy

What is it ?
Why is important in the current and future times
What is India doing?
DEPA, Sahamti, Use cases, process flow, Inference cycle, Lending use case, Current Statistics & Links

DEPA 2.0 - Data Collaboration Economy (DCE)

What does Data Collaboration mean ?

Not sharing - then you will not have control over the copy of data given

Safe space to bring data and Model together that win-win partnership can be created

Why should India be creating a DCE

Value in the data modeling economy

India is the largest Democracy, and our data feeds the DE and we don't get any value from it.

Threat of Data Colonisation

WHat are current challenge in DCE and how incentives are aligned for big-tech to not collaborate

How DEPA 2.0 looking to solve this DCE problem

- DP
- MC
- CC
- EC

How will the playground look, trust be built, roles and responsibilities

Use Case

- Health Data Collaboration
- Travel Data
-

Technical Artifacts:

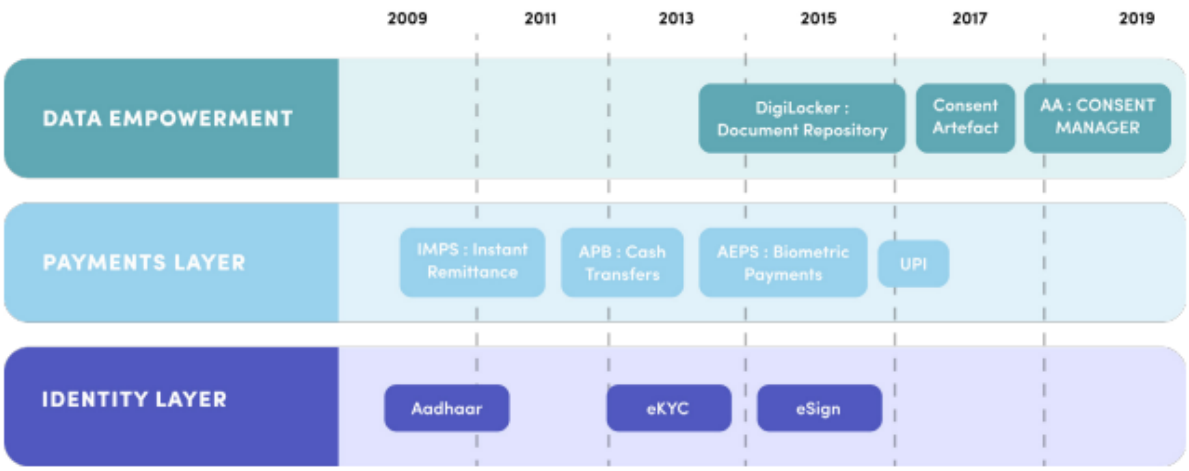
- Architecture
- Code
- Flowcharts
- Etc

Background

India transition to a “Data Modelling Economy”

Data is driving the new age economy. Customer (and their data) is king. India as the largest democracy in the world has tremendous capability to be the ML model capital in the world. This takes it beyond just the data (aka ‘data is new the oil’) to modeling(aka ‘modeling is the new economy’). This transition requires us to provide a comprehensive strategy of an aligned market ecosystem and technical prowess. DEPA is one such attempt to facilitate the transition.

DEPA as a layer of secure digital data sharing through consent forms the final layer of India Stack - a series of digital public goods designed to enable private market innovators to improve digital services for India across a range of sectors [2].



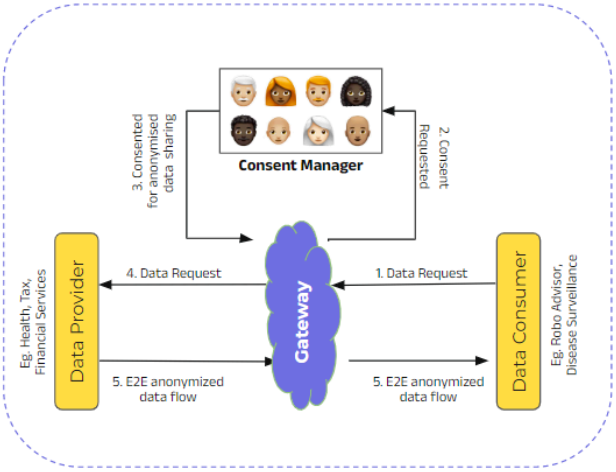
DEPA: Final Layer [2]

Data Sharing

Use of customer data can be anywhere between the extremes such as:

- **Fully Open**: customer shares login-password, data is used as is without anonymization, i.e. in the raw form, close to data stealing? Is it not better to call it transparent, rather than a pejorative term like data stealing?
- **Fully Closed**: Customer does not allow any sharing, no consent what so over, close to black box.

Historical Data sharing : Anonymized / Encrypted Data

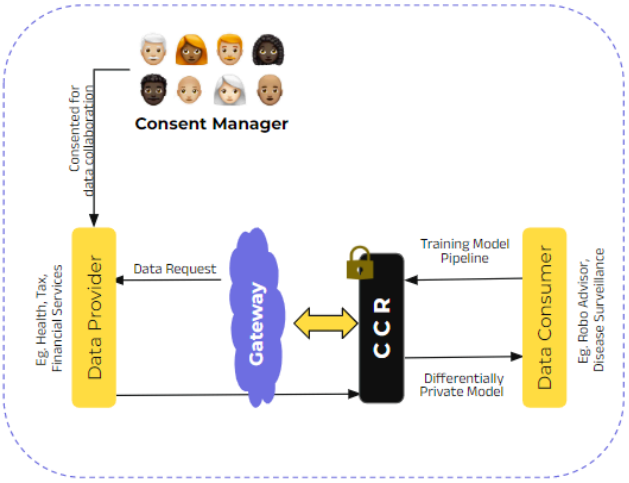


"Transfer" of data from Data Provider to Consumer

- Concerns with "raw" data
  - Privacy
  - Time/purpose limitation
  - Re-sharing
- Concerns with "transformed" or encrypted data
  - Lower data utility
  - New uncertain modeling paradigms

Both extremes are not useful as many new age applications/functionality need data (Why would a fully open model not be useful? This statement may need restatement to be complete or consistent) . Without that, one can not build super-useful applications such as recommendations engines, sales predictions, etc. (A sentence about the importance and use of models in society would be useful here or in the section about model building above.

Future: CCR-enabled Data Collaboration



"No data transfer" to Data Consumer

- "Model Certification" to prevent data leakages due to memorization
- Consumer has "raw" but secure access to data

Since the next section talks about privacy, the need to have privacy in a model based world must be motivated here.)

Need for DEPA

Central to the privacy laws of most countries is a set of principles that define how personal data are collected, shared and processed. However, consumers often do not know the benefits and

costs of the data that pertain to them. And even when they do, consumers find it difficult to assert their rights regarding the collection, processing and sharing of their data [1]. DEPA attempts to provide protocol for the data governance.

In practical terms, there is a need to find a balance between these two extreme usages of data, ie ‘Fully Open’ and ‘Fully Closed’. That's where DEPA comes in, customer empowerment as well as protection of the data. So, in nutshell, to address trust deficit, incentive misalignment and privacy concerns, DEPA specifies/demonstrates prescriptive transfer modalities. DEPA is a protocol essentially democratizing access to data.

DEPA is founded on the premise that individuals themselves are the best judges of the ‘right’ uses of their personal data, rather than competing institutional interests. They should not struggle to access and share their data.[2]

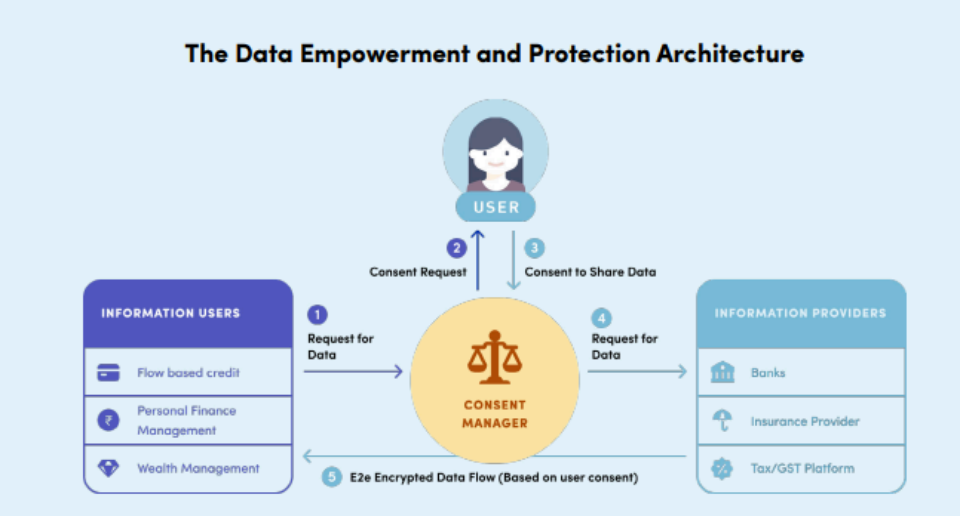
DEPA Genesis

So far healthcare treatments were predominantly based on symptoms. But there is a strong shift towards treatments based diagnostics evidence i.e. biomarker based. Patients will have tests done at various diagnostic labs and hospitals. But for a doctor to get 360^{o} view of the patient there is need to collate all the diagnostic reports. The current form of sharing the diagnostic data via paper reports is an archaic and insecure way. In the digital world, we would need a platform to collect the data in a secure and efficient way so that the doctor gets a holistic view of the patient gathered from data sources from diverse agencies. This is like account aggregation for healthcare data. That's individual usage. That's DEPA 1.0.
The healthcare data once collected in large numbers can be a boon to develop AI/ML based models for wider use, not just for an individual. That's DEPA 2.0. Getting data from multiple sources and of multiple people needs more confidentiality ensuring primitives. These, like DIfferential Privacy, Confidential Clean Rooms, Model Certifications are introduced in DEPA 2.0.

Here is the DEPA progressions with specific constructs used:

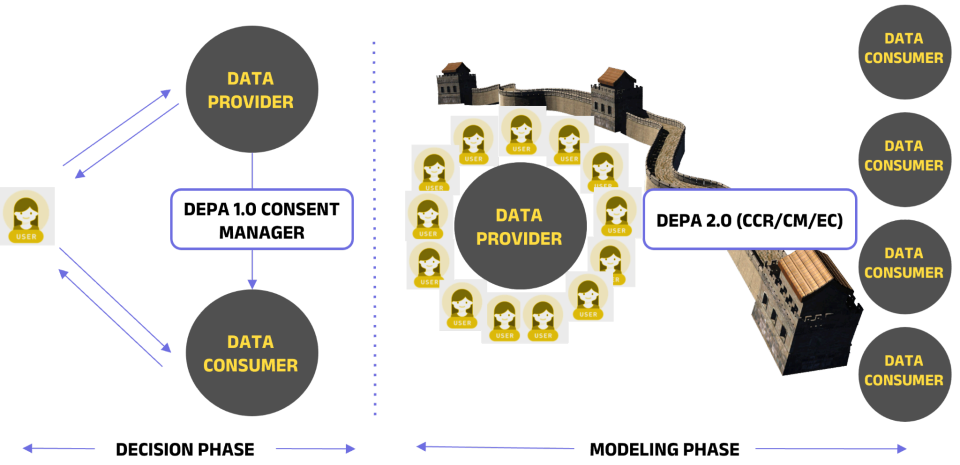
1. User data is typically scattered over at different places, e.g. medical data for a person may be lying in different diagnostics labs, hopsitas, etc. For effective diagnosis, the doctor will need a 360 degree view of all the data. Bringing the data to a central location, with privacy guarantees (you will not see, join,etc) is hard to enforce. So, DEPA 1.0 protocol suggests that:
    a. Let the data reside where it is (golden copy)
    b. Users give Electronic Consent for aggregation of data from various labs to generate a 360 degree view (silver copy)
    c. Doctors use the silver copy data for bio-markers based (as against symptoms based) diagnosis
    d. Purpose limitation is enforced legally.

    So, essentially DEPA 1.0 is an Inference-only cycle with individual usage and is already released on depaworld.com. ‘Sahamati’ is one such reference implementation.



Overall Architecture [2]

2. As users grow, inferencing cannot remain manual, but would need to be Machine Learning (ML) based, e.g. a doctor can have a 360^{o} degree view of say 20 patients a day, but not 1000. For ML training, one needs data from a variety of users. More the merrier. So, using the silver copies described in DEPA 1.0, one can train the ML models. But as this is no longer a personal use, the user needs a guarantee that their data is not seen, joined etc. Deanonymization could be done but does not inspire a lot of confidence. So, DEPA 2.0 brings additional primitives such as Electronic Contracts, Confidential Compute (CC) [which in turn involves, Model Certification, Differential Privacy and Confidential Clean Room (CCR]), to provide mathematical guarantee that the risk of data breaches is reduced significantly.. So, in DEPA 2.0 a certified model gets trained on silver copies in a central Confidential Compute (CC) guaranteeing privacy. DEPA 2.0 is essentially DEPA 1.0 + confidential Compute of (training) ML models. It's a work in progress.

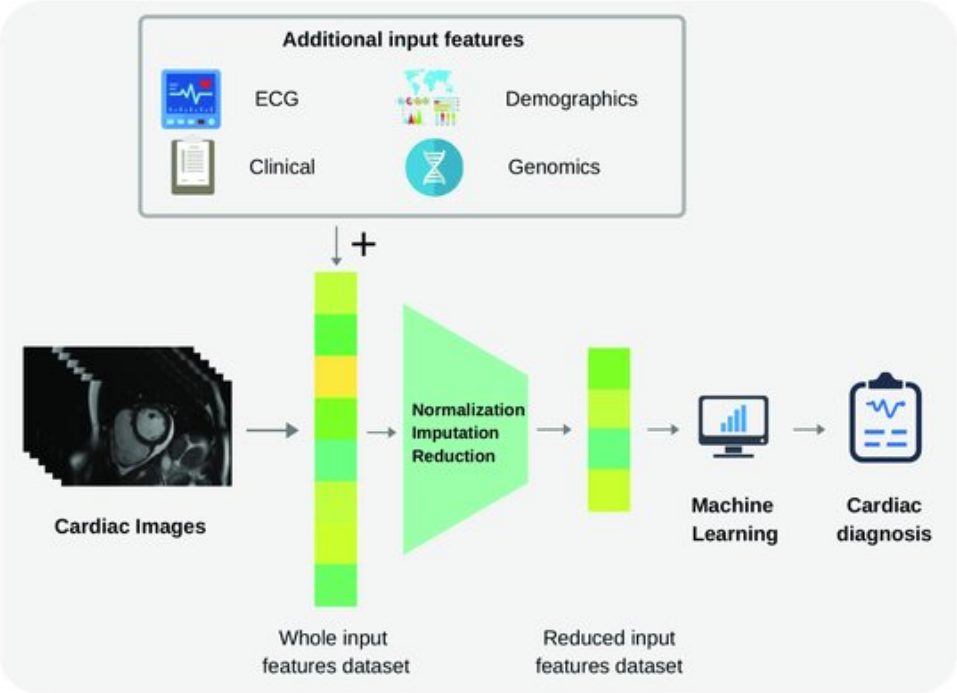


3. Although DEPA 1.0 had purpose limitation, it was enforced only legally which may not be very effective. DEPA 3.0 brings, DEPA 2.0's confidential compute of ML models, but with a technological guarantee on purpose limitations. DEPA 3.0 is still under discussion.

DEPA 2.0 Use cases

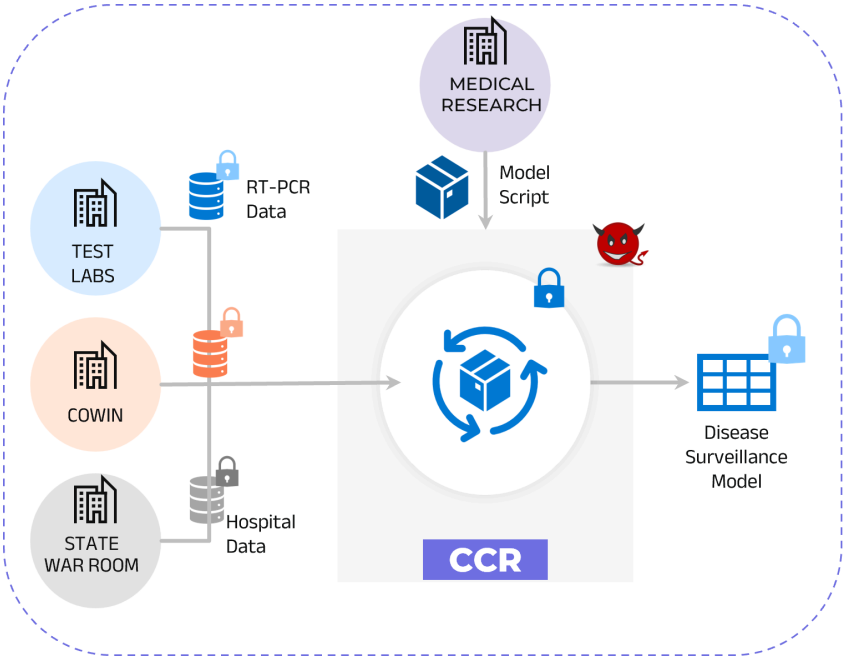
Health

Healthcare treatments are changing from symptoms-based to biomarkers-based approaches. Instead of relying just on the symptoms, doctors are increasingly trusting diagnostic lab reports for identification of the issues and devising the treatment plan. Hospitals and Labs are thus going to be plush with reports such as X-rays for Tuberculosis (TB) or Cardiac images for detecting heart issues. These reports,images, are valuable data for building models to predict various diseases.



"Image-Based Cardiac Diagnosis With Machine Learning: A Review", Carlos Martin-Isla et al

Such machine-learned-based models for healthcare diagnostics are getting approvals from regulatory bodies FDA (Food and Drug Administration) to be used in the field. Instead of using ready models from other countries (which are built on data for their population), it would be appropriate to build models based on local data for more effectiveness and also to avoid the possibility of digital colonization by a foreign-few.



(Example: Covid use case to demonstrate overall workflow)

DEPA helps by specifying protocols to build such data collaboration, model training and inference systems. Following is the way DEPA models the workflow:

- Patients get their X-rays done at the pathology labs or hospitals.
- Such labs and hospitals become Training Data Providers (TDP)s.
- A HealthTech company wants to build a TB prediction model using the X-rays available with these hospitals and labs. It is Training Data Consumer (TDC).
- They enter into an electronic consent-contract allowing such use. The original/golden data stays with the TDPs only. Only specific data (X-rays and some metadata like age, blood reports) can go to the HealthTech as a silver (anonymized?) copy for model training.
- HealthTech hires the service of SafeCompteInc.com which specializes in building machine learning models using Confidential Computing (involves, Differential Privacy, Model Certification and Confidential Clean Room) guaranteeing privacy preserving machine learning (PPML) training.
- Once the PPML model is ready, it is deployed at various hospitals, even the ones that did not provide the data.
- These models are typically used by lab assistants to infer from new X-rays. Predictions provided by the models give the first level of diagnosis. The prediction is further confirmed by certified medical practitioners in case of ambiguities. PPML models provide scalability,time-saving as well as accuracy in many cases.

E-Commerce

A typical travel booking site collects a large amount of data. Apart from building their own models, say for demand prediction, pricing, etc. these sites could be interested in monetizing the data for external applications. For example, BookMyTrip.com (BMT) is a portal for booking

travels by Bus, Train, Flights in both modes, domestic (40%) and International (60%). BMT collects demographic info from each user on the site, when they sign in. At the signing-up the user also gives consent on the ways in which their information can be used. Some users willingly opt for use of their data for recommendations such as travel insurance.

We can segment users based on transactions/ trends such as ‘age’, ‘mode of travel ,’ type of travel - holiday, business’,etc.Some specific segmentation for cross-sell such as

- Young-adults, looking for more adventure international
- Senior citizen seem to prefer pilgrimage and domestic
- Other business relevant /segmentation driven use cases

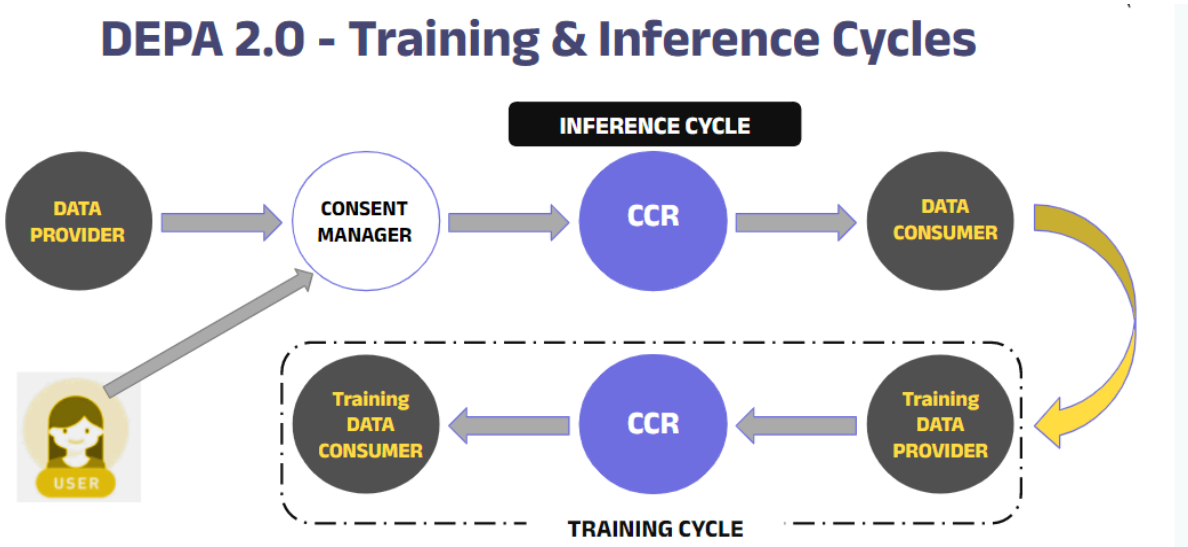
BMT has found out that it's not just selling tickets that will help them expand the offering by cross selling for specific segments with highly relevant offerings. This would require specific partnerships with manufacturers (like insurance, Bfsl products etc.). These opportunities would provide tangible revenue streams as well as value add to customers. Cross Selling products need to be targeted based on the segmentation mentioned above. Typical cross selling products are:

- **Travel Insurance**: The desire for having travel insurance is born due to their past negative experience on flight cancellations, lost baggage, etc. Users do not want to take chances anymore
- **Co-branded credit cards**: The desire for having a co-branded international credit card has become a very much a necessity for international travelers, as it offers seamless purchase and repayment experience.

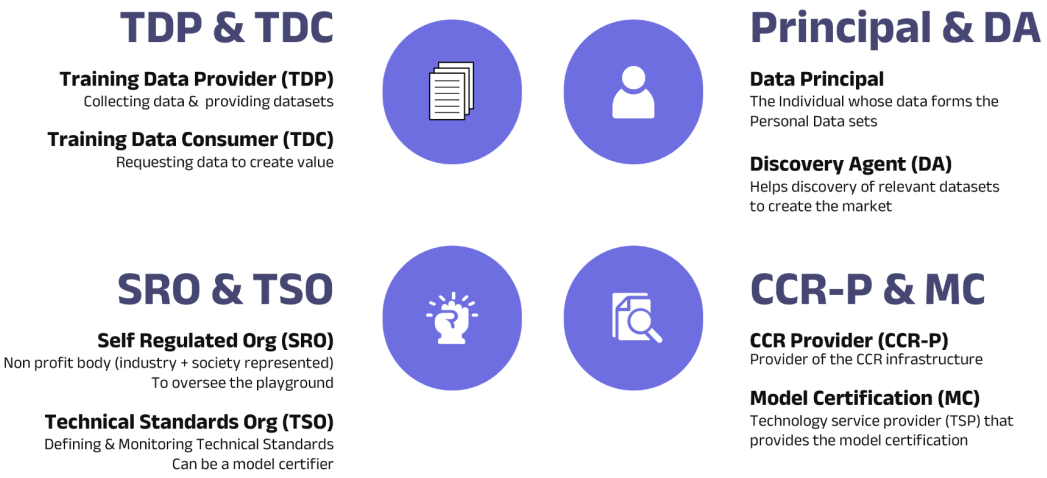
Users are willing to share not just profile data but past transaction data as well, so that they get very precise product recommendations. BMT feels that the data is just lying around and wishes to find appropriate uses of it, to grow his own business, basically to monetize it.

- International young adults for adventure travel offer highly relevant travel insurance with coverages that include adventure tourism, etc.
- For domestic holiday tourism offer comprehensive health insurance products like health cards, pharmacy cards, etc.
- For international travel explore powerful co-branded credit cards with relevant features like travel points, interns lounge, etc

DEPA 2.0 Ecosystem



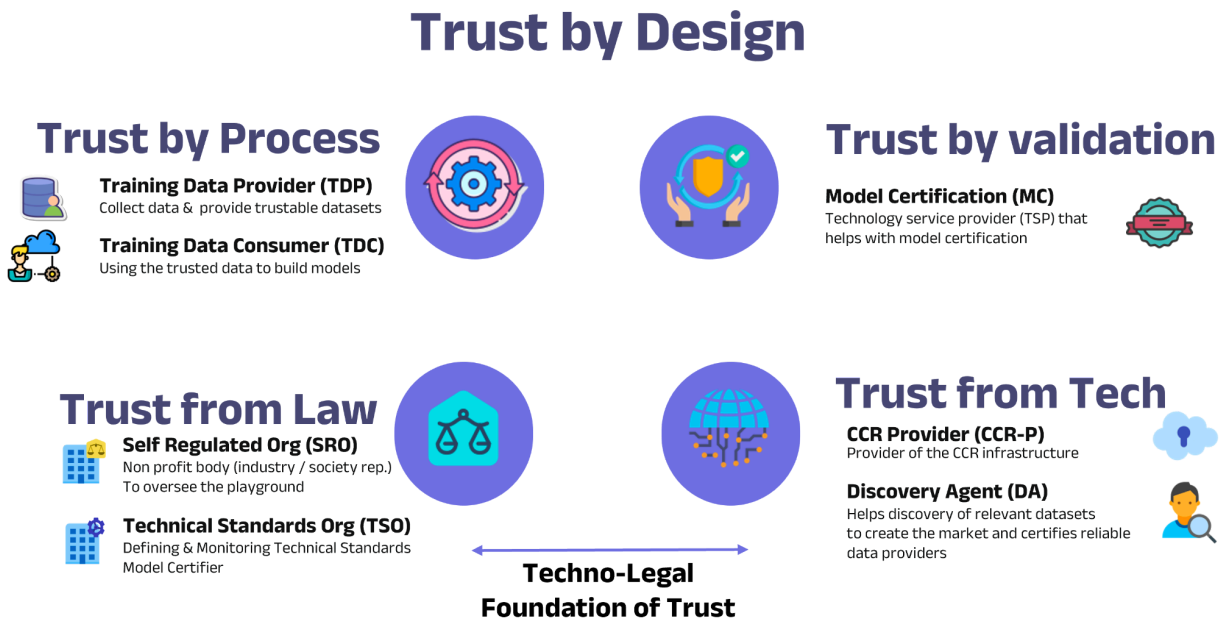
Main stakeholders are:



Role	Medical Use case	E-commerce Use case
TDP (Training Data Provider)	Hospitals having X-Rays	BookMyTrip.com
TDC (Training Data Consumer)	TB detection App company	InsureMyTrip.com
CCP (Confidential Compute Provider)	SafeComputeInc.com	SafeComputeInc.com

Users	Medical Lab Assistants first then doctors	Sales folks at IMT

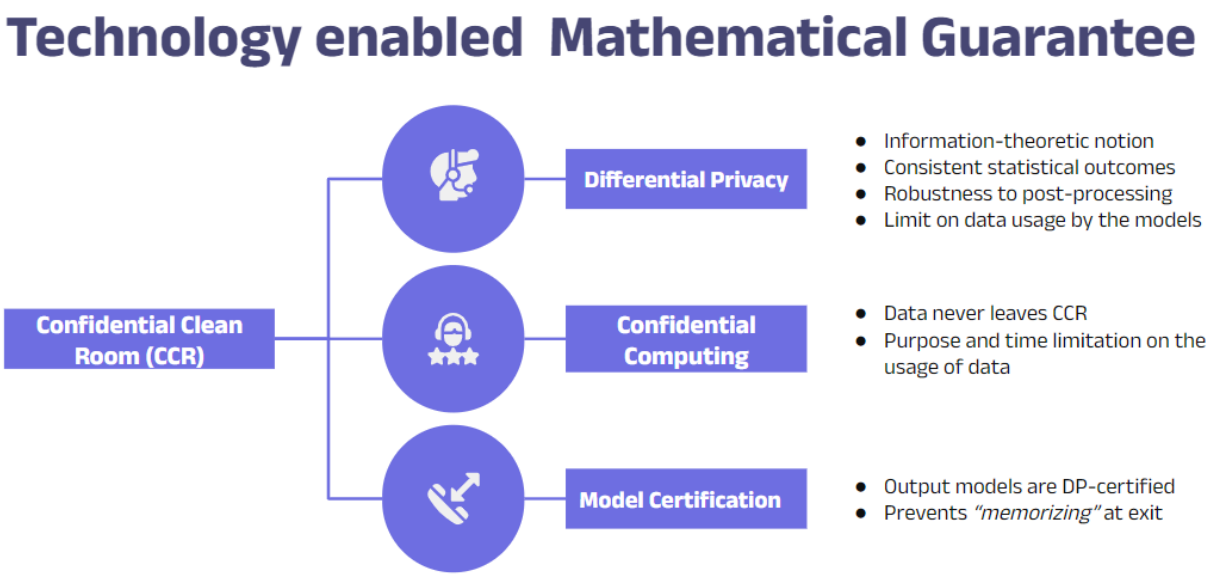
DEPA brings an element of Trust to the data collaboration ecosystem.



DEPA 2.0 Primitives

To ensure Data and Model confidentiality requirements, DEPA has been built on the following primitives (building blocks):

- Electronic Contract Artifact leveraging Smart Contracts technology
- Model Certification leveraging the science behind Differential Privacy
- Confidential Clean Room leveraging advances in Confidential Computing



Technologies that enable safe and secure data sharing are constantly evolving. Furthermore, even specific technologies such as differential privacy require constant re-evaluation. We want our design to be simple and easily evolvable: it should be possible for capabilities to be built incrementally while allowing for rapid adoption in today's world.

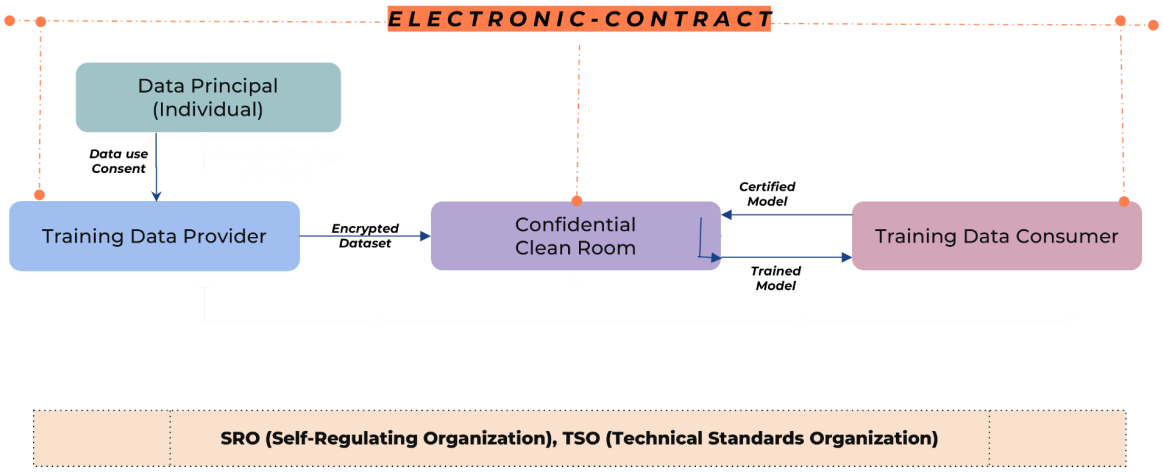
So, though these primitives are mandatory their actual implementation can vary, as long as they conform to the APIs (Application Programming Interface/Protocols) amongst them.

Electronic Contracts (EC)

Need to make sure that the data collaboration happens under contractual obligation, to prevent malpractices and to make legal recourse available in case of breach.

EC specifies monetary transactions amongst the parties, is techno-legal in nature and essentially orchestrates movements of data, money, models, etc between the stakeholders. Contract Service is set up by the SRO that authorizes purpose-limited transfer of training datasets to CC providers on the basis of contracts negotiated between TDPs and TDCs. Metadata and participant signatures are added to a contract via a COSE(Contract Service) envelope by the participants to produce a signed contract. An Envelope contains the identity of the participants and other information to help components responsible for validation that are part of a contract service. In essence, a signed contract is a COSE Envelope wrapped around a contract binding the metadata included in the Envelope to the contract. In COSE, an envelope consists of a protected header (included in the participant signatures) and an unprotected header (not included in the participant's signature).

The Playground



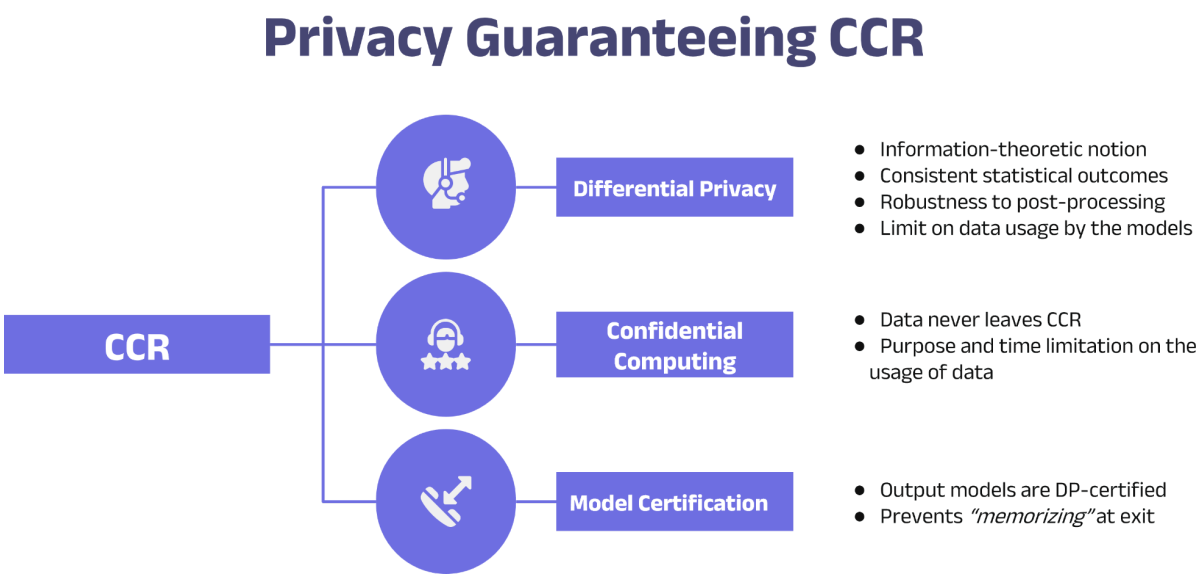
TDP and TDC enter into contract that specifies:

- Which data tables TDP is ready to provide, which JOINs are permitted.
- Purpose of data usage specificity
- IMT agrees and confirms the above, describes the payment model etc.
- The agency, called SafeComputeInc.com (SCI) where the confidential computing/training/app-development happens which ensures privacy while building the recommendation app, called CC (Confidential Compute)

```
{
    "id": "f4f72a88-bab1-11ed-afa1-0242ac120002",
    "schemaVersion": "0.1",
    "startTime": "2023-03-14T00:00:00.000Z",
    "expiryTime": "2024-03-14T00:00:00.000Z",
    "tdc" : "did:web:kapilvgit.github.com/biomedical.iisc.ac.in",
    "tdps" : [
      "did:web:kapilvgit.github.io/icmr.gov.in",
      "did:web:kapilvgit.github.io/cowin.gov.in",
      "did:web:kapilvgit.github.io/swr.kar.nic.in"
    ],
    "ccrp": "did:web:kapilvgit.github.io/ccrp.cloud.com",
    "datasets": [
      {
        "id" : "19517ba8-bab8-11ed-afa1-0242ac120002",
        "name": "icmr-infection",
        "url" : "https://ccrcontainer.blob.core.windows.net/icmr/data.img",
        "provider": "did:web:kapilvgit.github.io/icmr.gov.in"
      },
      {
        "id" : "216d5cc6-bab8-11ed-afa1-0242ac120002",
        "name": "cowin",
        "url" : "https://ccrcontainer.blob.core.windows.net/cowin/data.img",
        "provider": "did:web:kapilvgit.github.io/cowin.gov.in"
      },
      {
        "id" : "2830a144-bab8-11ed-afa1-0242ac120002",
        "name": "state-war-room-hospitalization",
        "url" : "https://ccrcontainer.blob.core.windows.net/swr/data.img",
        "provider": "did:web:kapilvgit.github.io/swr.kar.nic.in"
      }
    ],
    "purpose": "TRAINING",
    "constraints": [
      {
        "privacy": [
          {
            "dataset": "19517ba8-bab8-11ed-afa1-0242ac120002",
            "epsilon_threshold": "1.5",
            "noise_multiplier": "2.0",
            "delta": "0.01",
            "epochs_per_report": "2"
          },
          {
            "dataset": "216d5cc6-bab8-11ed-afa1-0242ac120002",
            "epsilon_threshold": "1.5",
            "noise_multiplier": "2.0",
            "delta": "0.01",
            "epochs_per_report": "2"
          },
          {
            "dataset": "2830a144-bab8-11ed-afa1-0242ac120002",
            "epsilon_threshold": "1.5",
            "noise_multiplier": "2.0",
            "delta": "0.01",
            "epochs_per_report": "2"
          }
        ],
        "compute": [
          "ccr_security_policy": ""
```


Both TDP and TDC need to have a contract with the CC agency also to make sure no data breach happens. There could be an overarching regulator (called Self Regulatory Organization - SRO) to make sure all the above contracting falls in the approved line-of-functioning.

Confidential Compute (CC)



*The above diagram should change, CC (Confidential Compute) should be at the place of 'CCR', which has 3 children as: Differential Privacy which guarantees training data privacy, CCR which provides secure hardware for training and Model Certification provides guarantees regarding sanity of the model.

Differential Privacy

There are multiple approaches like federated learning, anonymisation techniques (k,l anonymity,etc.) and differential privacy. Here are a few:

- Pseudo anonymisation: Some transformation is applied (say PAN number, 'ABCD' becomes 'BCDE'). Masking PII (Personal Identifiable Information) is also done. But many a times, reverse engineering original information is possible directly or indirectly (via JOINs or external knowledge base)
    - Trends: Instead of raw data, store trends, central/statistical numbers or ranges. Say, age is specified only in 3 buckets, 'Young (< 18), Adult (18-60), Seniors (60+). So no actual age is specified but just the label.
- Generation: a new synthetic data is generated from the raw data, having similar distributions, which then can be used for model training.

DEPA has selected Differential Privacy (DP): Random noise is added to raw data to anonymise it; in addition application of differentially private algorithms with privacy budget tracking. Model training happens in CCR. DP provides mathematical guarantees about privacy (subject to allowed budget).

Model Certification

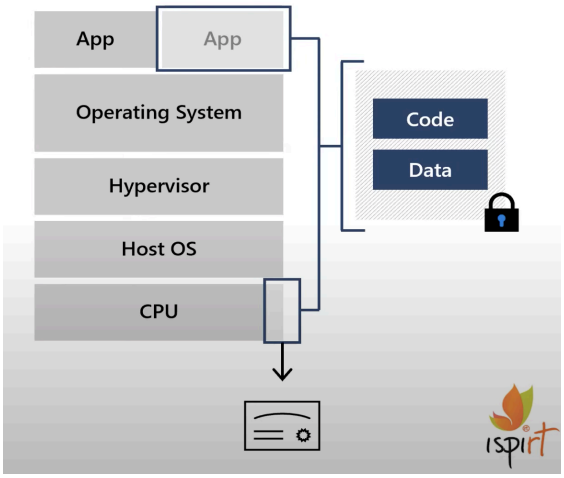
Model Certification certifies model abides with differential privacy and the privacy budget. CC provides evidence for model certification. It provides Mathematical Guarantees such that Model is good for usage and there is no data leakage

TDC needs to submit its current Prediction Model to CC for further retraining using data from TDP. This retraining has to happen in the Privacy Preserving model and thus inside CCR. Before CCP starts using the model, it needs to make sure the validity of it, so as to avoid any malpractices from ill-intentioned parties.

```
{
    "model_input":"/mnt/remote/input_model/ccr_depa_input_model.onnx",
    "model_output":"/mnt/remote/output_model/ccr_depa_trg_model.onnx",
    "model_cert":"/mnt/remote/output_model/ccr_depa_trg_model_cert.cert",
    "batchsize":2,
    "total_epochs":2,
    "l2_norm_clip":2,
    "learning_rate":2,
    "loss_function":"tf.keras.losses.CategoricalCrossentropy(from_logits=True,reduction=tf.losses.Reduction.NONE)",
    "optimizer":"tensorflow_privacy.DPKerasSGDOptimizer",
    "metrics":["accuracy","precision","recall"],
    "eps_function": "compute_dp_sgd_privacy.compute_dp_sgd_privacy"
}
```

Confidential Clean Rooms

Confidential clean rooms (CCRs) are built using hardware-protected secure computing environments where sensitive data can be processed while limiting the purpose for which it can be used. These environments are remotely verifiable by the hardware which can issue certificates. CCRs are based on an emerging technology broadly called confidential computing which is already implemented by major hardware manufacturers such as Intel Corp and AMD and supported by all major cloud providers and infrastructure hyperscalers.

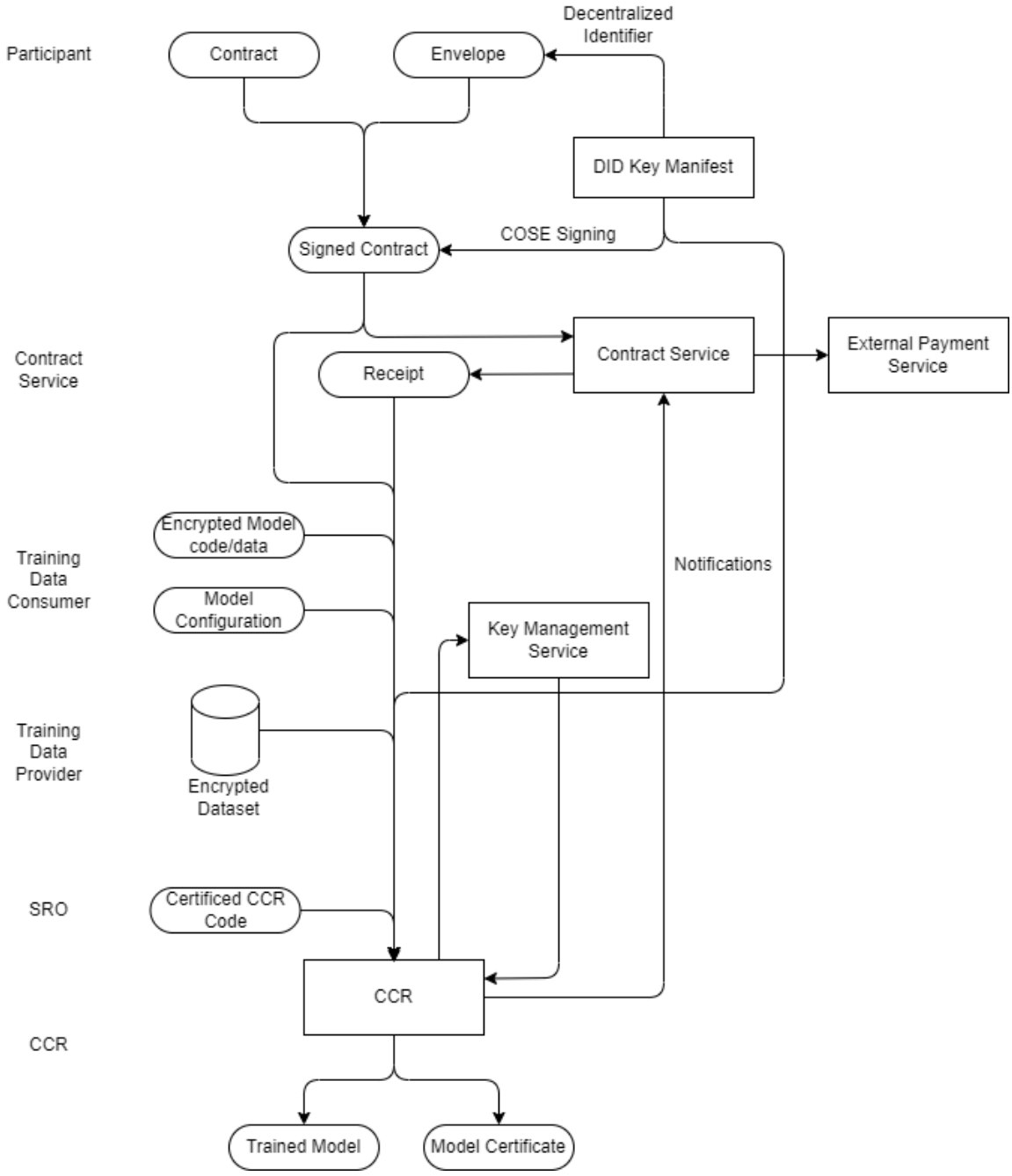


Overall CC Workflow

- TDC to save the current model (m) in h5 format (ONNX coming)
- CCP will take the model, run Model Certification to check its sanity
- Once passed, CCP will (re)populate the model inside CCR
- Currently Tensorflow privacy is used to change the Optimizer to Privacy Preserving Machine Learning (PPML) model.
- Parameters like epsilon, are read from Model Configuration, signed in the Electronic contract
- With data from TDP, ‘m’ gets (re)trained in PPML model and returned as m_p
- ‘m_p’ is then used by TDC for inferencing.

DEPA Training Architecture

The architecture for DEPA training cycle consists of a very loose federation of contract services and CCs, and a set of common formats and protocols for sharing datasets based on signed contracts.



Workflows

Participant registration

All participants, including TDPs, TDCs and CC providers are required to register with a central registry. As part of the registration flow, each participant registers their long term identity (e.gX.509 certificate or DID issued by one of recognized issuers) with the central registry. [figure]

Dataset registration

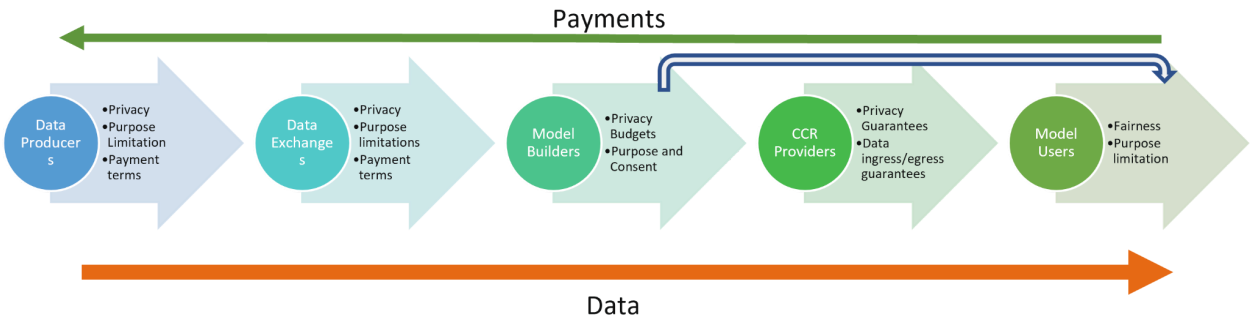
[figure]

Dataset discovery

[figure]

…
…
<TBD>

Payments



Conclusions

In a nutshell, DEPA empowers people to seamlessly and securely access their data and share it with third party institutions [2]

<TBD>

- Upsides & Downside (in absence of collaboration)
- DEPA ensures that customers control their data
- Various parties can leverage this data only by conscious consent
- Consent can be atomic for the limited purposes only
- All the operations are performed under privacy preserving mechanisms.

References

1. "The design of a data governance system" by Siddharth Tiwari, Sharad Sharma, Siddharth Shetty and Frank Packer, BIS Papers No 124, Monetary and Economic Department, May 2022 (revised July 2022) [link]
2. "Data Empowerment And Protection Architecture", Draft for Discussion, AUGUST 2020 [link]
3. “DEPA as a Global Standards” [link]