ABP Privacy Infra, Long Range Investments [A/C Priv]

Goals of this Review

- This document was written to advise leadership on preparedness, investments and technology plans with respect to inbound regulations.
- ABP is seeking continuous, multi-year investment and support from Data Infrastructure, Core Data and Central Privacy for the plans laid out in this document.

  NOTE (1) This document is an abridged version of ABP Privacy Infra 2021 (2) the follow up reading of this document in ABP Privacy Infra: WWW [A/C priv] which expands on the base understanding provided here.

Executive Summary

- We were surprised in 2021 with regulatory changes in the EU and India that will restrict 1P data use; setting the stage for a global regulatory push toward consent for 1P data use in Ads.
- Our past policy enforcement plans were already insufficient (on any timeframe) for handling 2PD concerns. The gap with success (scalably enforcing policy) is now an order of magnitude larger with the increased 1P governance.
- We do not have an adequate level of control and explainability over how our systems use data, and thus we can't confidently make controlled policy changes or external commitments such as "we will not use X data for Y purpose." And yet, this is exactly what regulators expect us to do, increasing our risk of mistakes and misrepresentation.
- Addressing these challenges will require additional multi-year investment in Ads and our infrastructure teams to gain control over how our systems ingest, process and egest data. This new investment is needed in addition to the ongoing Purpose Policy Framework investments.
- The remainder of this document is structured as follows

Motivations

- Regulatory Landscape
- Fundamental Problems
- Gaps in Purpose Policy Framework

Investments Needed

- Curated Data Sources

- 1PD Controls
- Timelines & Goals
- Resource Napkin and Funding Gaps

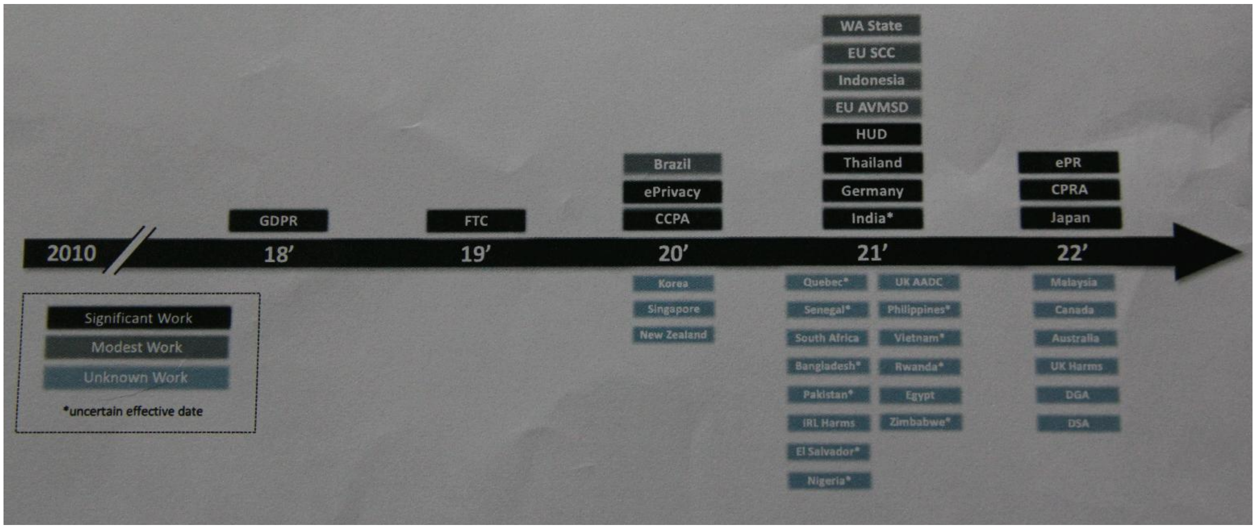
Nascent Investments

- Competition
- WWW Isolation and Control

Motivations

Regulatory Landscape

In 2021 regulatory restrictions will continue to expand around the globe as we shift toward consent. Under a consent regime, by default we are not allowed to use personal data for ads. We are anticipating impactful regulations from India, Thailand, South Korea, South Africa, Egypt, and many other jurisdictions (see image below). We also expect the US to make progress on federal privacy legislation, though the effective data will likely not be in 2021. Key point: historically regulations have been major thrashing changes for the company, but we've had the "luxury" of addressing one at a time (GDPR in 2018, FTC suit in 2019, CCPA in 2020). This is no longer the case. We face a tsunami of inbound regulations that all carry massive uncertainty.

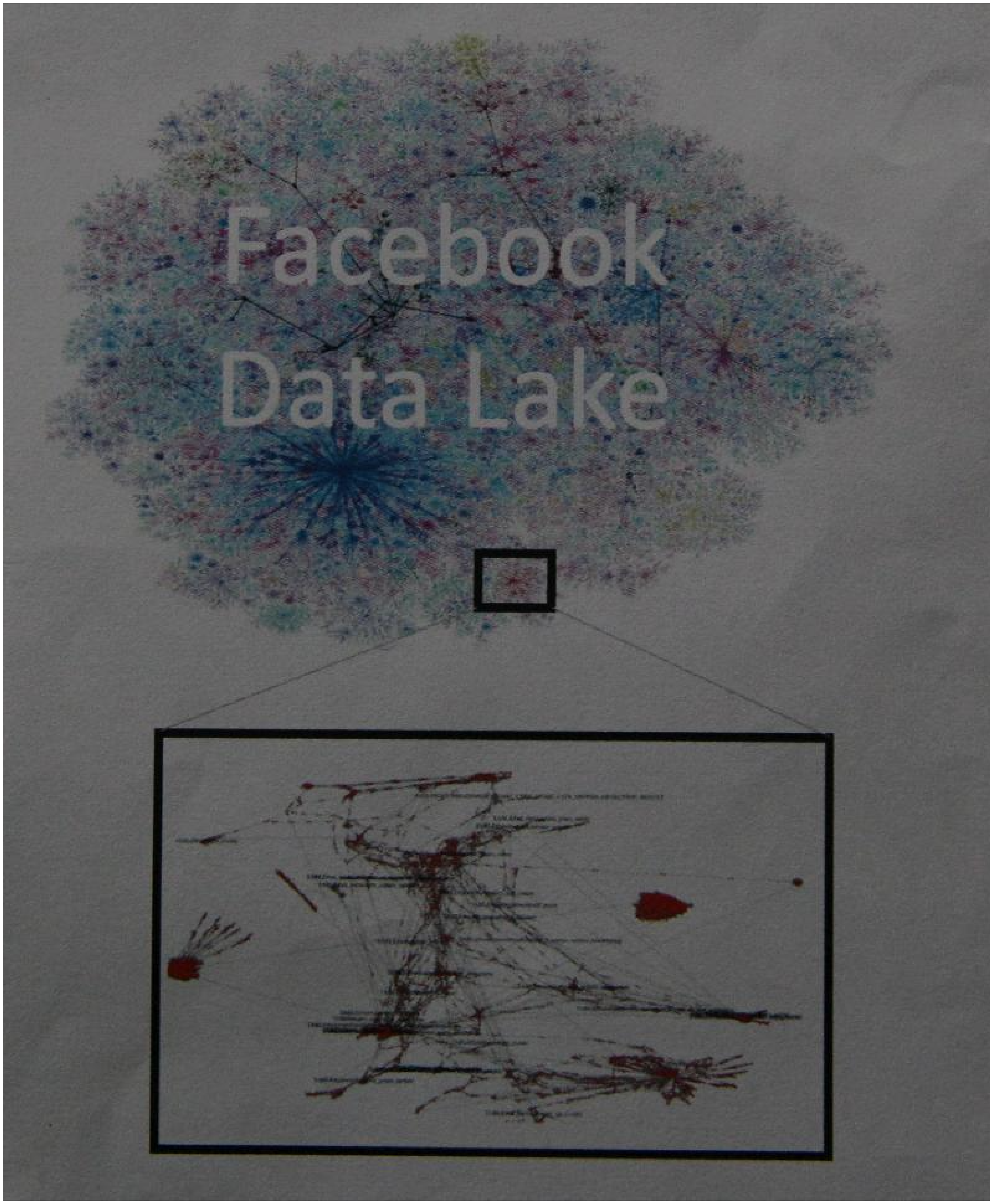


Fundamental Problems

From an infrastructure point of view, the heart of our challenge is a lack of "closed form systems." In this context, "closed form" refers to the union of these desirable properties of a system: (1) All data types ingested, derived and egested by the system can be enumerated and controlled; (2) All data uses within the system can be enumerated and controlled; (3) All data uses of egested data by consumers downstream, can be enumerated and controlled. These properties are what make a system tractable, and controllable. If we can't enumerate all the data we have - where it is; where it goes; how it's used - then how can we make commitments about it to the outside world?

We fundamentally lack closed-form properties in Facebook systems. For more than a decade, openness and empowering individual contributors has been part of our culture. We've built systems with open borders. The result of these open systems and open culture is well described with an analogy: Imagine you hold a bottle of ink in your hand. This bottle of ink is a mixture of all kinds of user data (3PD, 1PD, SCD, Europe, etc.) You pour that ink into a lake of water (our open data systems; our open culture) … and it flows … everywhere. How do you put that ink back in the bottle? How do you organize it again, such that it only flows to the allowed places in the lake?

To make this understanding a bit more concrete, consider this: There are 15K features used in ads models. The graph to the right shows the dependency chain of actual tables used to produce **just one single feature**. In total, ~6K tables (the red dots) were used to produce "user_home_city_moved"



Gaps in Purpose Policy Framework

In 2019 ABP accepted the vision from Central Privacy that in the longer term, most, if not all of our challenges would be solved by the Purpose Policy Framework (PPF). The basic idea of PPF was that we could attach a policy, at a very granular level, to every piece of data as it gets ingested or created in Facebook. Then, as the data flowed from system to system, the policy would flow with it. Then, at every place the data was being processed or used, it would get properly handled, because every processor would have a declared purpose. If the declared purpose wasn't allowed, the data would get filtered out by the underlying infrastructure…And all this would happen magically and silently so that engineers wouldn't need to restructure all of its data flows to gain control over its data use.

Key point: The beauty of (Purpose Policy Framework) was that -in theory- ABP wouldn't need to restructure all of its data flows to gain control over its data use.

In late 2020 PPF was folded under the branding of Privacy Aware Infra (PAI), and while it carries forward many of the challenges characteristic of large infrastructure projects (delays technological gaps, misaligned cross-org investment levels), it is still one of our most necessary and promising investments. **ABP remains committed to PPF adoption. However we now see that the original version wasn't a full solution.**

Key Gaps:

1. **Investment gap** → Policy Annotation. There are not aggressive and proportionate investments among the Family of Apps for attaching policies to their tens-of-thousands of root data sources (e.g. APIs in WWW, Tables in Hive, etc.)
2. **Technological Gap** → Policy Propagation. Even after policies are attached, they need to be handled through complex operators (joins, unions, concatenations, transformations, etc.) This is a very technologically difficult to do for several classes of operators that are frequently used throughout ads pipelines. Contrary to the original vision, PPF won't be able to support major classes of Ads operators, on any timeframe.
3. **Technological Gap** → System Coverage. Policies need to be handed off from one data platform to another, passing all the way from root systems to target systems without getting dropped (e.g. www → thrift → scribe → hive → etc.) There are more than 140 data processing systems (Asset Classes), and even for some of the major ones like WWW, PPF won't be supported until 2023 or later.

   … summing these gaps: ABP has been getting ready to receive a pitch, but no one will be ready to throw the ball.

Key Point: In light of these gapos, and pressured by accelerated regulatory timelines, we can't wait for PPF to achieve the "closed form" system properties we need while holding our core ads infrastructure as invariant. We need to restructure our data flows so that they are closer to the goal by design.

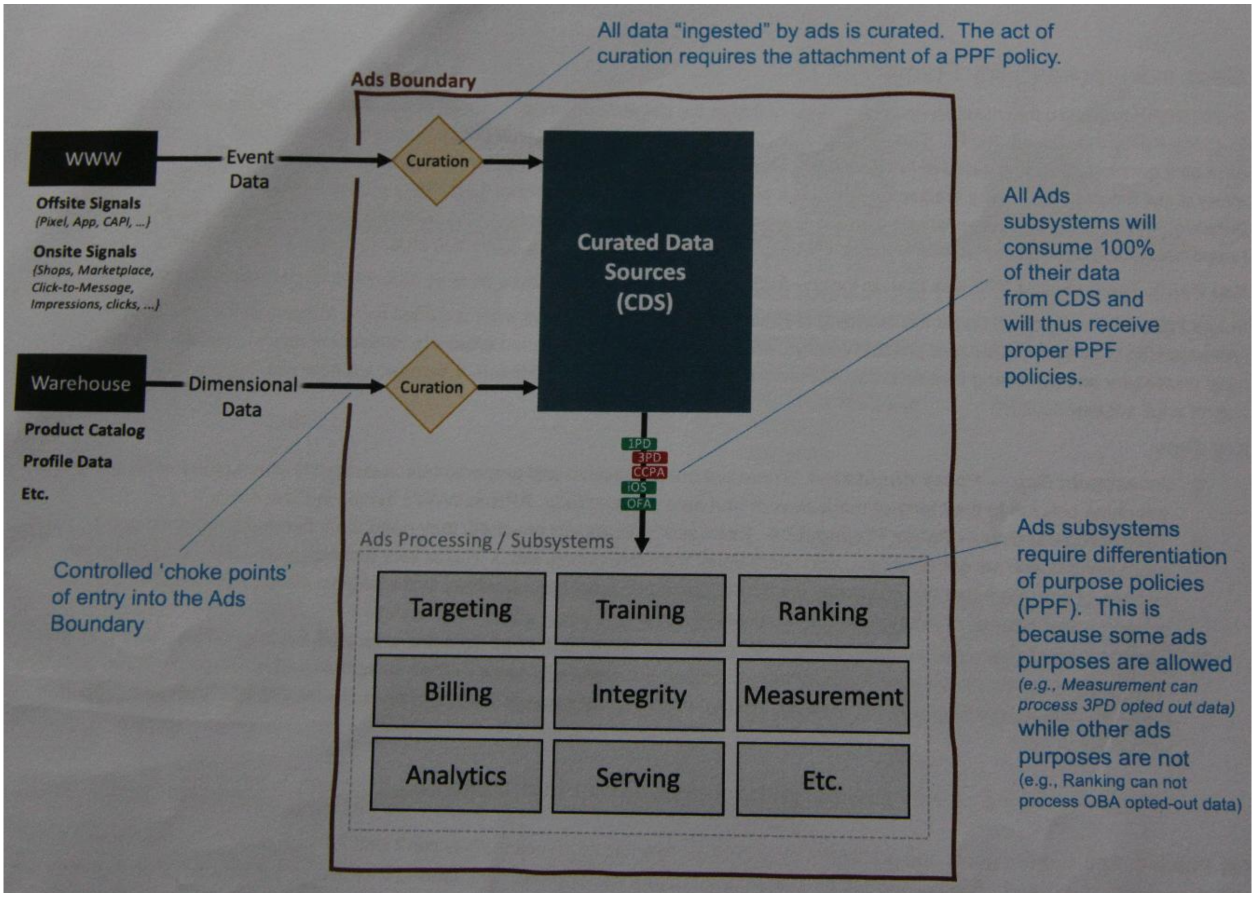
Investments Needed

In this section we'll discuss (1) Curated Data Sources and (2) 1P Data Controls. These are the most major infrastructure investments needed, and they have the most developed plans.

Curated Data Sources (CDS)

There are tens-of-thousands of uncontrolled data ingestion points into Ads systems today. We will rewrite and migrate all of these onto a controlled 'choke point' where we can systematically annotate the data with PPF policy. All of the Ads subsystems will only consume data from [Curated Data Sources]. Thus we will know exactly what data is being consumed by Ads subsystems, where it comes from, how it's generated, and what policies apply. Furthermore, CDS will only use computational operators (joins, unions, etc.) that PPF can support. With CDS we will be able to make product commitments and comply with regulatory requirements.

The below diagram provides a 30K foot architecture. A more detailed architecture can be found here. [URL redacted]



All data "ingested" by ads is curated. The act of curation requires the attachment of a PPF policy.

All Ads subsystems will consume 100% of their data from CDS and will thus receive proper PPF policies.

Controlled 'choke points' of entry into the Ads Boundary

Ads subsystems require differentiation of purpose policies (PPF). This is because some ads purposes are allowed *(e.g., Measurement can process 3PD opted out data)* while other ads purposes are not (e.g., Ranking can not process OBA opted-out data)

Why do we need Curated Data Sources if we have Purpose Policy Framework?

Two reasons:

1. All data consumed by Ads must have an attached policy. The act of curation into CDS is the attachment of a PPF policy (by ads engineers). The alternative of waiting for FOA [Family of Apps] to decorate all their data with policy is **many** years away, at best.
2. PPF technology will never be capable of handling certain classes of operators and processors that are common among Ads pipelines. This, even if FOA were to properly decorate all their data with policies, Ads would still need to refactor most of its data flows to make them PPF compatible.

Why do we need Purpose Policy Framework if we have Curated Data Sources?

Two reasons:

1. Subsystems within Ads may or may not be allowed to consume certain data from CDS, subject to privacy policy. For example, we may allow certain opt-out data to be used in training, but not in ranking or targeting. To support these different levels of access among CDS consumers, we need a policy enforcement mechanism within the Ads boundary. Ads could build a new custom enforcement - why would we? - This is exactly what PPF is for, and Ads intends to take a low-risk dependency on PPF's basic capabilities.
2. PPF policies are the standard way to express a purpose policy today, and will generalize for other policy types as part of PAI going forward. Data arriving at and leaving Ads boundary will need to carry such a policy to remain compatible with the rest of the infra ecosystem.

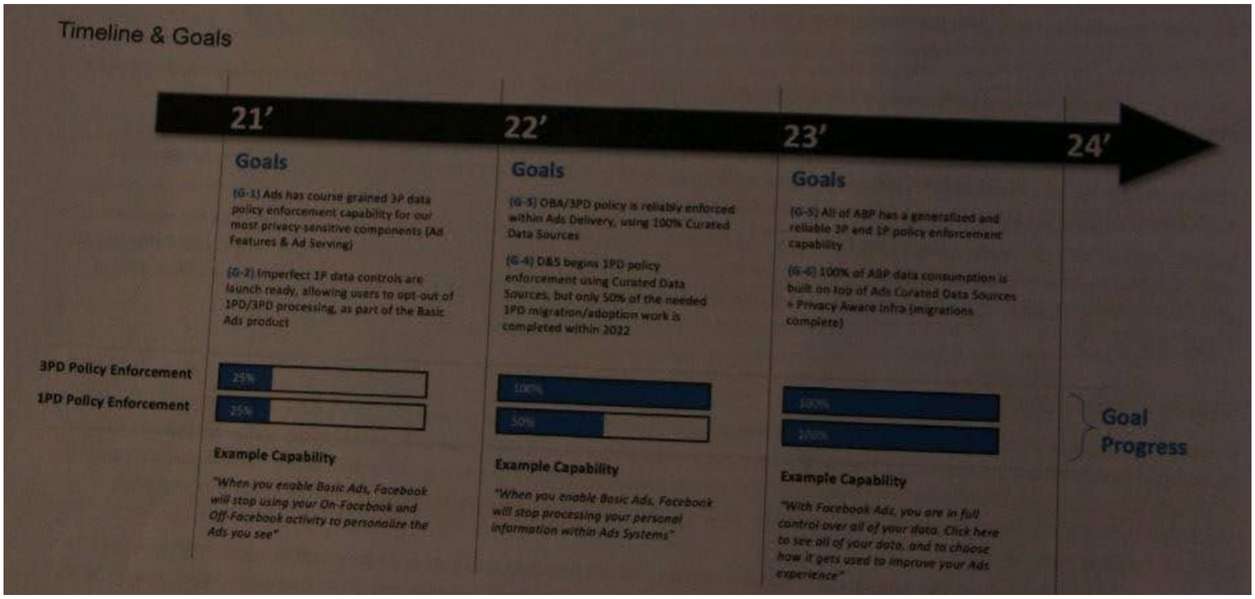
1P Data Controls

There is a global regulatory trend pushing Facebook Ads toward legitimate interest or consent for 1P data use. In either legal construct (legitimate interest or consent) Ads must "immediately" halt processing of an opted-out user's data; including user data that was accumulated before the opt-out occurred.

Long term: A major part of our long term solution for halting processing in ads is CDS+PPF, as described above in this doc. However, even CDS+PPF will not be sufficient, because we will not be able to retroactively apply a user opt-out toi data that has already been consumed by Ads systems. There is strong consensus among Ads and Data Infrastructure TL circles that to fully stop processing (including data that has already entered ads systems) we'll need to efficiently filter opted-out users in the underlying Data Infrastructure.

Specifically, Data Infra will need to have some semantic awareness of its data assets (hive tables, scribe streams, etc.) such that it can know if the data it contains includes an opted-out user information (a UserID) - in which case, the infra would be silently and efficiently filter that opted-out user from further processing. Building this will take multiple years, and will require DI investment. We are still in the early stages of joint scoping and estimations.

Short Term: Our short term response to these requirements is to build “Basic Ads.” A new product initiative that needs to be launch-ready in Europe by January, 2022. When launched, Facebook users will be able to “opt-out” from having almost all of their 3P and 1P data used by Ads systems - page likes, posts, friends list, etc. To build this, Ads will still require efficient Data Infra filtration, however the work will be more narrowly scoped - we will only filter data at the boundary of Ads warehouse systems (Uhaul jobs, etc.) Data infra scoping and estimation work for this is underway.



Timeline & Goals

21’ Goals

(G-1) Ads has course grained 3P data policy enforcement capability for our most privacy-sensitive components (Ad Features & Ad Serving)

(G-2) Imperfect 1P data controls are launch ready, allowing users to opt-out of 1PD/3PD processing, as part of the Basic Ads product

Example Capability

"When you enable Basic Ads, Facebook will stop using your On-Facebook and Off-Facebook activity to personalize the Ads you see"

22’ Goals

(G-3) OBA/3PD policy is reliably enforced within Ads Delivery, using 100% Curated Data Sources

(G-4) D&S begins 1PD policy enforcement using Curated Data Sources, but only 50% of the needed 1PD migration/adoption work is completed within 2022

Example Capability

"When you enable Basic Ads, Facebook will stop processing your personal information within Ads Systems"

23’ Goals

(G-5) All of ABP has a generalized and reliable 3P and 1P policy enforcement capability

(G-6) 100% of ABP data consumption is built on top of Ads Curated Data Sources + Privacy Aware Infra (migrations complete)

Example Capability

"With Facebook Ads, you are in full control over all of your data. Click here to see all of your data, and to choose how it gets used to improve your Ads experience"

Resourcing Napkin

Ads will require substantial investment for the next several years in order to deliver on our aggressive timeline. A detailed resourcing plan is here. In brief:

- Estimate: 450-750 eng years. For simplicity, yearly breakdowns assume the mid-point of 600 eng years.
- Assume: Execution timeline of 3 years.
- High-level execution plan and resourcing needs by year:

2021

- +150 HC (150 in total)
- Fully staff ingestion for CDS and 1P data controls
- Partially staff consumption of CDS and frontload key tech investments

- 150 eng years of work is completed.

  Fully Funded
  ✓50 HC from PAI pool
  ✓40 HC from ABP pool
  ✓60 people from ABP repri

2022

- +200 HC (350 total)
- Fully staff consumption of CDS - migrate Delivery systems to use only curated data
- Egin investments outside Ads Delivery (Business Integrity, Measurement, etc.)
- 500 eng years of work is completed.

  Funding Source Needed

2023

- ~250 HC (100 total)
- Drop the investment for curated data sources since most work should be completed

  Funding Source Needed

Nascent Investments

In this section we'll discuss (1) Competition and (2) WWW isolation. These are two relatively more nascent challenges that will drive additional investment needs over the next several years.

Competition

Our legal teams anticipate negative competition judgments from the European Commission within 2021. This will likely have contagion effects for other jurisdictions (United States) and other FB product areas (Shops).

To oversimplify, ‘Privacy’ is concerned with controlling data flowing *from our Family of Apps into Ads*; whereas ‘Competition’ is concerned with the reverse flow - i.e. competitively sensitive data flowing *from Ads into our Family of Apps*. But the fundamental challenge is exactly the same. That is, data isolation and control among FACEBOOK properties.

In 2020 Ads wrote a joint proposal with the Choice and Competition team that would stem risks associated with competitively sensitive data use. It called for accelerations of PPF adoption around the company and a massive surge in headcount split among Central

Privacy, Ads and FOA → see [Occam] Avoiding Competitive Data Use Mistakes [A/C priv]. The Occam proposal remains unfriended in this writing (April, 2021)

WWW Isolation and Control

From an infrastructure point of view, the aforementioned investments are about 'isolation and control' of data and processing. That is, we need to isolate Ads from the rest of the Family of Apps, and build well controlled interfaces at the boundaries between Ads and other Facebook properties. Creating this isolation in WWW wil be especially difficult due to its monolithic nature. There are hundreds of thousands of controllers and call sites, and there is no clear solution for defining an "ads boundary", short of doing the hard work of manually visiting each of them.

Our overall strategy for solving this problem is to migrate as many of the Ads data dependencies as possible out of WWW, in favor of Warehouse data sources, where boundary definitions, isolation and control systems are relatively more available. We will execute on this strategy (migrating to Warehouse) as part of our Curated Data Sources investments. But this will not be enough. There will remain substantial Ads code in WWW that is entangled with FACEBOOK organic processing. The legal constructs of legitimate interest and consent require us to also stop this WWW processing.

We will continue working with Privacy Infra to evaluate new WWW lineage tools, boundary annotations and enforcement technologies such as CIPP. However, there is no obvious solution yet, which doesn't involve dramatic investment by Ads engineers. At present, it appears we will need to manually visit tens-of-thousands of call sites and code paths to define and enforce an Ads boundary. And in many cases we will need to manually untangle and rebuild Ads scenarios.

Key point: Just like PPF isn't a sufficient solution without heavy Ads refactoring and adoption work in warehouse, WWW enforcement technologies such as CIPP will not magically solve these problems without enormous Ads and FOA side investments for adoption.

Appendix-1: Point of view on readiness and uncertainty of solutions

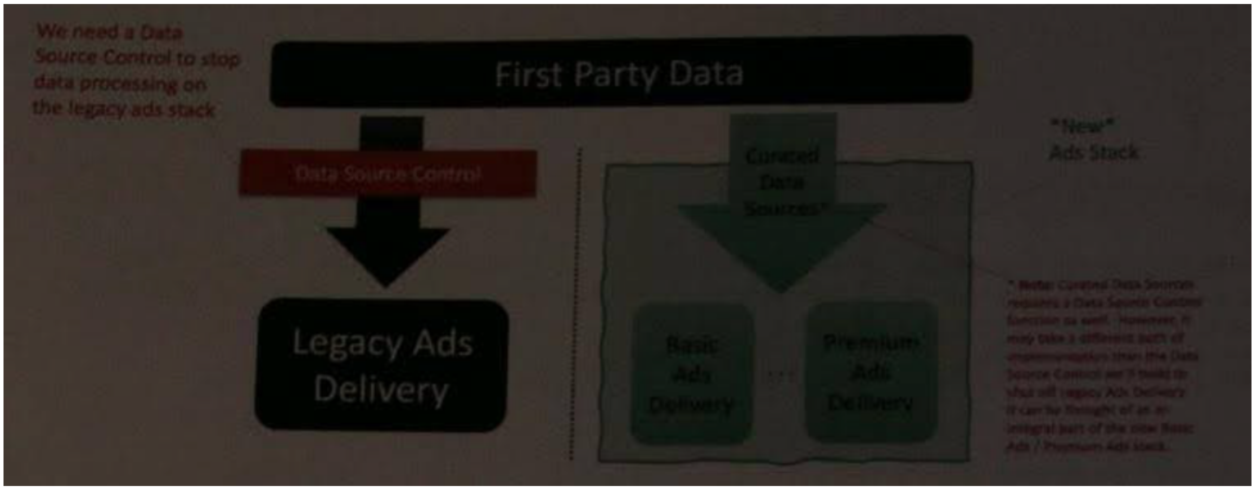
The below grid shows the dimensions of requirements we described above. For each of these, we are assigning one of 3 colors that represents our level of confidence in our proposed infrastructure solutions. This is obviously subjective, and it oversimplifies things. Nevertheless, we hope it can level set perspectives on uncertainty.

High Confidence	Low Confidence	Very Low Confidence

	Necessary?	**Sufficient?**	
	How confident are we that our infra investments address long term fundamental problems and are necessary?	How confident are we that our infra solutions are sufficient for 2022?	**Explanation**
Jurisdiction	🟩	🟩	We have a reasonable handle on which jurisdictions are going to drive significant requirements and there is a Central Privacy team monitoring global changes.
Sensitive Categories	🟩	🟥	With inferences counting as SCDs, this problem space expands to ML. Our solutions in this area are immature, and have not been subjected to regulatory scrutiny. We will likely discover new work (e.g. Youth requirements).
Purpose Limitation & Legal Basis	🟩	🟨	The legal constructs in this area are well understood and similar around the world. We've had a lot of time to consider the problem space. However, we are making risk based decisions around legal basis, and we may be challenged by regulators.
ML/AI	🟨	🟥	This is a new area for regulation and we are very likely to see novel requirements for several years to come. We have very low confidence that our solutions are sufficient.
Data Provenance and Responsibilities	🟩	🟨	Similar to purpose limitations, we understand the legal constructs in this area, and they are patterned around the globe. And yet, we have risk based

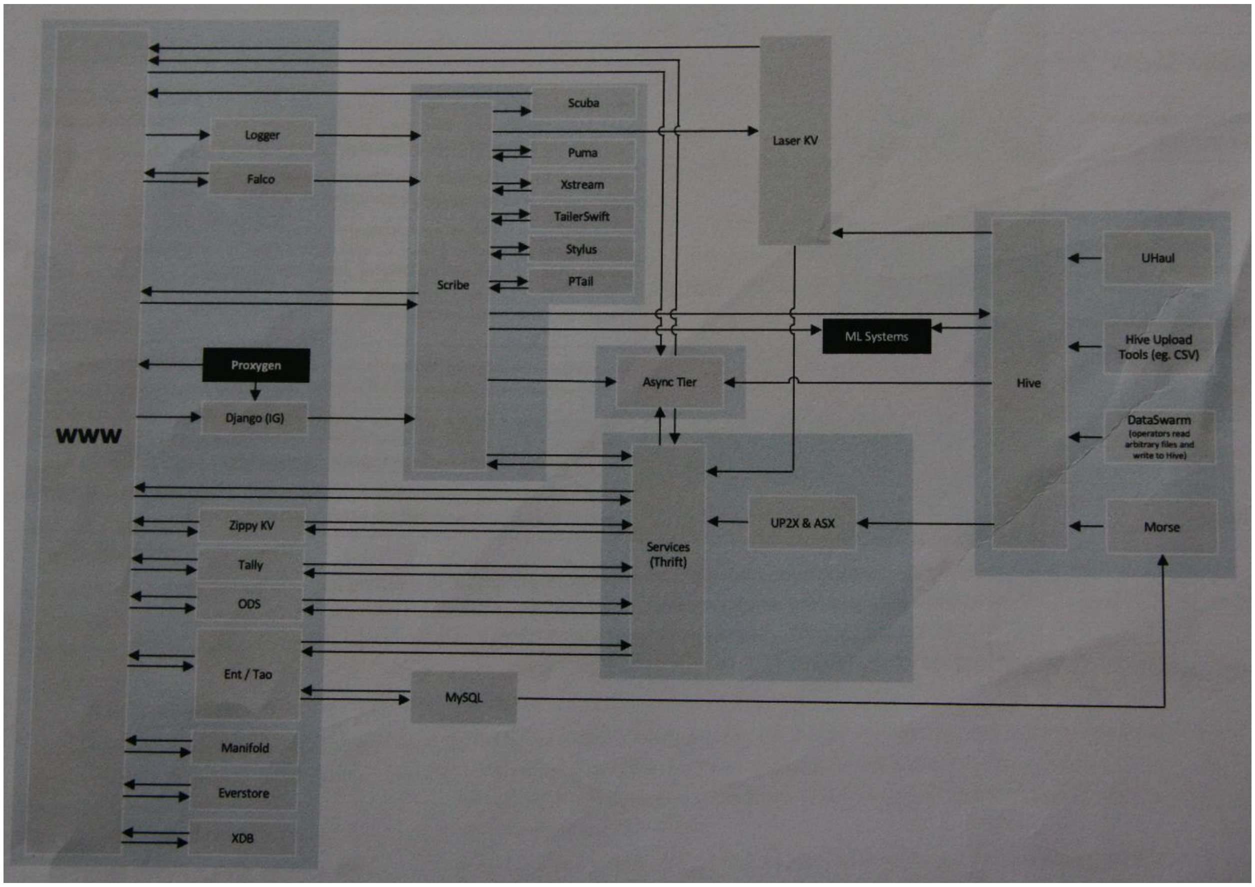
	🟩	🟨	tradeoffs and there is room for regulatory surprise and novel requirements, particularly from the USA.
Transparency & Control	🟩	🟨	Transparency and Control over machine learning is likely to reveal new requirements.
Localization	🟥	🟥	Ads is largely in the dark here. We are not aware of any ads specific requirements, and are awaiting company wide solutions that will most likely be owned by core infra teams.
Minimization	🟩	🟩	This area is largely understood and similar around the world. The infra solutions for controlling retention policies are in place and continue to improve coverage through Central Privacy Waves. There is room for surprise here as ML regulations progress.
Timing	🟩	🟨	Our infra solutions have not, and most likely will not be ready before significant regulations take effect. Of most concern: India will likely mandate Consent within 2023.

Appendix-2: Collection of diagrams for potential Q/A



We need a Data Source Control to stop data processing on the legacy ads stack

* Note: Curated Data Sources requires a Data Source Control function as well. However, it may take a different path of implementation than the Data Source Control we'll build to shut off Legacy Ads Delivery. It can be thought of as an integral part of the new Basic Ads / Premium Ads stack.



Appendix-3: Examples and Understanding of Ads WWW Footprint and Entanglement

Question from [NAME REDACTED]: “One thing I’d like to understand better is how Ads uses www. If possible, it would be great if you could pull together some specific examples of each of the three categories you listed in the architecture diagram in the appendix (organic processing, Entangled Ads processing and signal processing) along with where in www this code lives and what it depends on."

Answer: We don't have an organized understanding of all the Ads workloads and data flows in WWW. Even a specific and commonly used example -the AdRequest- has unboundedness and intractability in its dependency graph. As you read this, keep in mind this is probably our most well understood and controlled example, and that other ads workloads are similarly gnarly (measurement, targeting, business integrity etc… etc…)

Macro view: Ads Footprint and Entanglement in WWW

APIs

- There are about 350K endpoints in the entire company. Of these, about a quarter (~80K) log to scribe channels that are owned by Ads OnCalls.
- According to OnCall associations, Ads only owns about 30K of the 80K endpoints that send data into Ads → Restated: 60% of the WWW APIs that Ads uses are not owned by Ads. Let's call these "entangled" APIs.
- Presumably most of the entangled APIs have alternate or original purposes that are 'organic' in nature.

Files and Directories

- There are 800K directories in WWW, containing 4.8 million files.
- Of these, at least 6% of the file structure (50K directories, containing 300K files) are in a directory that starts with "ads" in the name.
- Important note: We don't have a useful structure in WWW for determining what is ads vs what is not ads. Let alone a useful way to substructure the WWW workloads within Ads. This is, in part, the nature of our WWW entanglement problem. The heuristics described above, such as *'includes' "Ads" in the filename'*, or *'has an Ads oncall attached'* only give an approximation.

Zoom in on a specific Ads workload: The AdRequest

- Let's consider the "ad request"; one of our more "standard" interfaces between Ads & FOA.
- There are at least 60 different end points that generate 680 distinct code paths which eventually lead to an AdFinder call.
- In each ad request there are about 200 fields passed into AdFinder. These fields can be objects and complex structures. Each AdRequest is stuffed with *whatever* data placement teams and organic surfaces around FACEBOOK decided to compose in WWW. This would include unbounded and uncontrolled fetching of data from Ents and organic data stores from around the company (zippy, laser, … or whatever).
- When calling AdFinder, the min—average—max call stack depth is 19—90—127. That is, on average there are 90 WWW files used before AdFinder is even called.

Again, this is organic WWW processing running arbitrary code, and eventually building an arbitrary AdRequest using %whatever% data from around the company that they want.