IPFS Large Volumes UX

Research for “Large Volumes” Use Cases (50+ TB of Data)

Kelani Nichole
Protocol Labs
June 26th, 2018

Contents

Overview

Qualitative research project to understand the user needs and opportunities for IPFS ’Large Volumes’.

Project Goals:

IPFS Large Volumes

- *Identify critical paths for users to successfully navigate our documentation, APIs, and support channels when managing large volumes of data on IPFS*
- *Identify a critical list of features and a critical path for implementation that will satisfy complete journeys*
- *Collect information to help IPFS make more user-centered decisions, produce more user-centered designs, and communicating with more cohesive narratives*
- *Identify where further user research is necessary*

Research Goals:

Large Volumes Use Cases

- *Gain deeper knowledge of user needs around handling large volumes of data (50+ TB)*
- *Uncover blockers and knowledge gaps*
- *Identify end user pain points*

The Process

Identify Open Questions through Stakeholder Interviews

Develop a Research Plan and Script. Recruit Participants

Moderate Sessions, Debrief in real-time and transcribe notes. Highlight Findings

Sketch through journeys and synthesize findings, quotes and themes.

Research Plan

PRODUCT

What's being researched?

Data Managers handling IPFS ‘Large Volumes’ integrations

LOGISTICS

When and where will the test be conducted?

- Remote One-on-one (1Hr) Sessions, Recruitment by IPFS Team
- Recorded in Zoom, Observers Welcome
- 30 Minute Debrief, Shared Summary Notes

GOALS

What are the main goals of the test? What specific questions do we hope to answer?

- Support Pilot launch for Large Volumes
- Validate Differentiators
- Document Workflows and Needs
- Identify Areas of Opportunity
- Build a Base-level Experience Architecture for the IPFS User Journey

PARTICIPANTS

How many participants will be recruited? What are the key behaviors we're targeting?

‘Large Volumes’ Use Cases (Handling 50+ TB)

Professional Data Manager

(Gov’t Org)

Software Developers

(University Library / Archive), (Genetics Researcher), (Startup)

Community Data Stewards

(Research Network / Initiative)

TASKS & TOPICS

What are the key tasks to discover? What hypothesis do we have? Which topics should be covered?

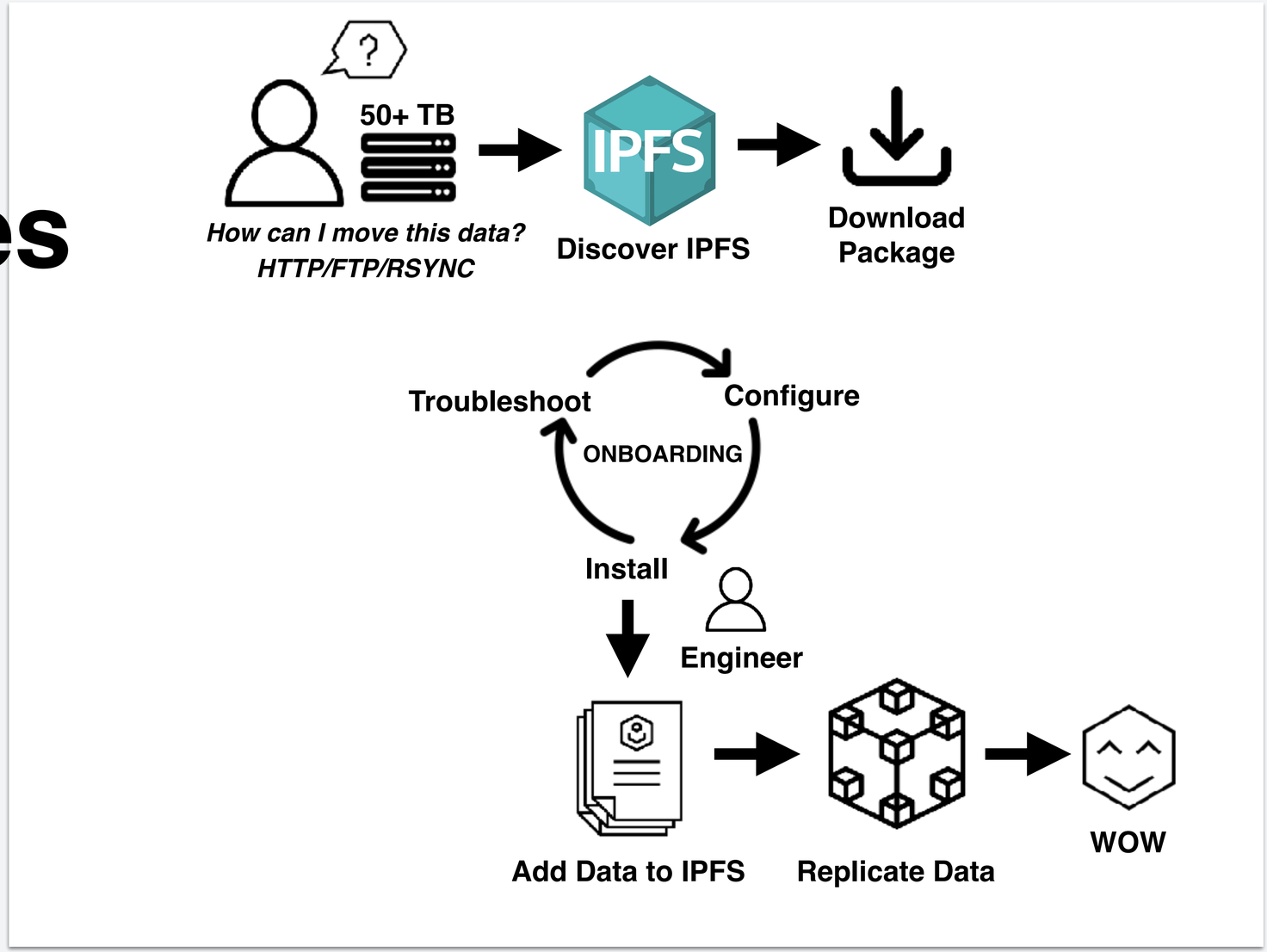
- **Participant’s Goals**
- Stakeholders and Users
- Infrastructure/Approach
- **Storage, Versioning and Access**
- Challenges and Needs
- **Primary Data Management Tasks**
- Perception of IPFS
- **Differentiators**

PROCEDURE → WELCOME/INTRO → PROFILE/NEEDS → PRIMARY TASKS → IPFS JOURNEY → RATINGS/THANKS!

What we Want:

Large Volumes ‘Happy Path’

Where/how can IPFS better achieve this simple win in different ‘Large Volumes’ contexts?



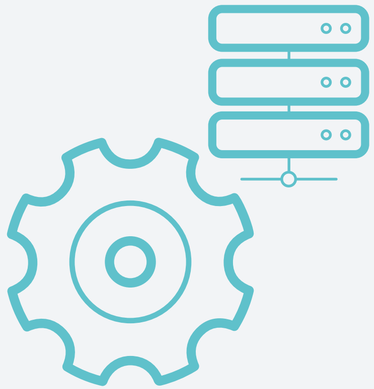
Who We Talked To

Profiles for ‘Large Volume’ use cases, including general attributes and specific details of our participant’s technology landscape.

Who We Talked To:

Large Volumes Users

Producer



Provider

Consumer

Three primary user journeys were identified in our first round of research: Producers and Providers were our primary focus.

Profile:

Producers

We spoke with researchers and scientists who are producing, processing and publishing large volumes of data.

Primary Data Generators

Moving, Analyzing, and Versioning Data

Producing Insights

Landscape:

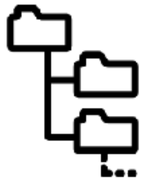
Genetics Researcher

"The future is to respect the distributed nature of the data."

Centralized Data
(Reluctance for Sharing)

FTP Service in Virtual Machines

Metrics in Log Files



Naming Conventions
(Mapped with MD5 Hashes)

- Describe Data
- Read/Sequence Data
- Link Data

"They will need a signing off to make this a proper public service, because it will require a VM and getting a VM and allocating CPU will need some kind of sign-off."

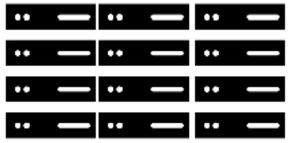
Versioning:
'Variance Call File'
(Lots of CPU is Required)

Landscape:

Startup Enterprise

"Self-sovereign"

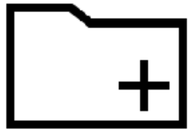
Onsite Servers



Storing Hashes of Genomes

Firewall Rules *(I/O Chatter)*

Bioinformatics Pipeline *(Mongo DB Backend)*



Patient Health Records *(Key Management / Rotation)*

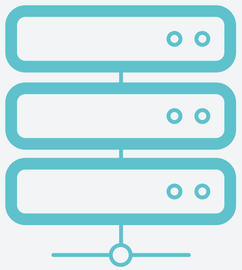
"Those species are publicly know and maybe we're doing a public good by providing them, I'm not terribly worried about the storage cost, that's a data together type of sharing."

Profile:

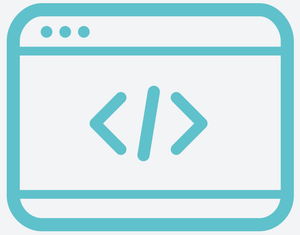
Providers

We spoke with data engineers providing tools for access and archiving large volumes of data.

Building Tools for Access



Archiving and Moving Data

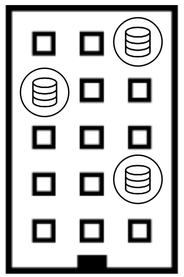


Metadata Bottleneck

Landscape:

National Scientific Agency

"Improved Availability is a Win"



Grassroots Storage
(Duplication)

Cloud Storage Collaborators:
(AWS, Google, IBM, Microsoft, Open Commons Consortium)

Robotic Tape Archives

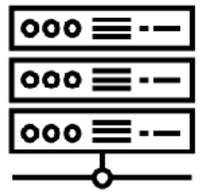
Email/FTP
(Moving Data)

"Across the org storage happened in a grassroots way at the level of a program or project, but not strategic or enterprise level"

Landscape:

University Library / Archive

"How can we preserve data across organizations?"

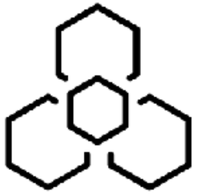


Virtualized Hardware
(Spinning disk)

Index Maps & Monitoring Systems

Repository
(Fedora app links metadata, access rules)

Versioning: Run Fixity Checks
(Backup of subsequent versions)



APIs and Web Services for Content

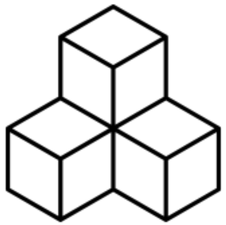
Wikis/ Documentation
(for end users)

"We have virtualized hardware for storage to be portioned out by our systems team when we need it. We have some level of flexibility there. The majority of this is on spinning disk, unfortunately a lot of it is quite slow."

Landscape:

Open Research Initiative

"Shotgun Shells and Duct Tape"



Container Environment *(RDS Instances holding pointers to hashes)*

Shared Data Stewardship

Construct Hash, Store on IPFS

IPFS Daemons point to Kubernetes Cluster

Custom Crawlers *(Identify assets PDF, XLS, SQL)*

"IPFS is running in a container environment. When it fills up we swap it with another. It works. It took us a while to arrive at this configuration. "

What We Heard

From 5 participant session, themes began to emerge in our discussions around ‘Large Volumes’ data management.

Theme:

Metrics Matter

- Metrics around data consumption are key to success in ‘Large Volumes’ use cases. People must prove value to sustain their work.

- *Opportunity: Provide great tooling for showing usage of data across the network (beat HTTP access logs)*

Quotes

“Projects are significantly more successful when we have buy-in from the line-office/ scientist expertise.”

“High impact projects are recognized by the community for being impactful, citation counts. They want a metric how USED the data is, So how does that impact grants? Someone will have to be the first to take a hit on metrics.”

”Our influencers I guess are other teams across the library, like the earth sciences and geospatial team. They drive a lot of our product and goals for the year.”

Theme:

Authenticity Adds Value

- Verifying the source and authenticity of data is a common need across participants for accurate citations, reliable versions, and enhanced performance in addition to security.

- *Opportunity: The way IPFS uses content addressing gracefully addresses these issues. Provide people with the tools to wield that power in these cases.*

Quotes

“What we need to ensure is that someone could audit it and verify we didn’t do anything with the bytes. We try to make sure the number of hands involved are very few, so we automate as much as possible.”

“Content addressing is really interesting to me, one thing we are concerned with is data authenticity.”

“If you've got the hash going in and the hash going out, and you refer all that with the root node in your merkle dag in your publication then all of a sudden you’ve cryptographically verifiable, mathematically confident link to everything being correct.”

Theme:

Deduplication

- There is a clear need to reduce redundancy, manage sub-sets of data, and create efficiency in data management processes.

- *Opportunity: Align messaging about ‘deduplication’ to the key painpoints and provide better tooling.*

Quotes

“We know we have a deduplication problem with other orgs, but we don’t know how to address it. Content addressing could help with that, but it hasn’t risen to the top of the priorities.”

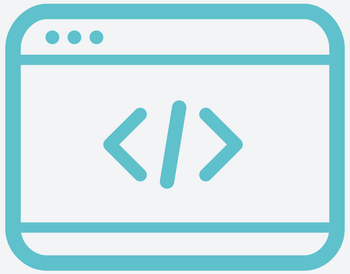
“How many copies of this same data are we storing locally? More importantly how much is the taxpayer spending to store all copies of this same data locally”

“In the cluster we don’t have to worry about which server has that file, and now more and more places are putting genomes from NCBI onto IPFS

We are leveraging the natural deduplication properties of the content addressing where, when the PDF doesn’t change, the hash doesn’t change, and we add nothing to our overhead.

Theme:

Metadata



- Producing metadata is a substantial overhead. In particular, due to the proliferation of schemas and identifiers, people struggle to connect and reuse data.

- *Opportunity: IPLD is positioned to solve this, but successfully applying it in this domain will require grappling with the complexities and nuance of metadata.*

Quotes

“The primary bottleneck is management and prioritization of getting metadata created.”

“The question is how do I refer to those files? The archive is centralized so I have to use an ID, it’s an arbitrary number. The search engine has discovered it but then there’s the problem of how do you talk to other people about the data. IPFS solves this because you have a global namespace, a method of talking to data, your multihash will be consistent so I think it can be helpful there.”

“A lot is modeling of IPLD that surrounds the storage of the metadata. I don’t think IPFS will know that a string of bytes is a radiology image or a genome, so modeling the metatdata surrounded around that, and tying that to the code, So that’s an issue just in the piping, and how do you use a standard, there’s not one, so we need to create a standard and publish it.”

Insights:

Perception of IPFS

Participant sentiment was gathered on a few key attributes of IPFS ‘Large Volumes’.

This section contains a subjective indicator of current sentiment to evaluate areas of opportunity.

How Reliable is IPFS?

PROVIDERS: Very Good

“It’s just the frequency of change, we aren’t at a 1.0 so no 10 until it’s a stable API”

“There’s positive buzz, it’s a fully uneducated rating, based on what i've seen on twitter”

PRODUCERS: Good

“It’s been running for the last year other than when I’m doing updates.”

“Because there are certain areas that come up in the running instance and publishing the namespace stuff takes awhile and I don’t have certainty around that. If it queries for a remote file that doesn't exist it takes a long time.”

How Secure is IPFS?

PROVIDERS: Average

“Private keys in the config file, not so good.”

“I haven’t looked into it.”

“I’d rank FTP and SFTP ‘very poor’ on this scale. Ability to verify the data helps us in the security area. Cleaner and more well-defined connection between two points vs the feeding trough of FTP”

PRODUCERS: Very Good

“So far as I know it’s secure. I’ve been digging deep into Go for key storage, and there are limitations that are beyond me, I trust the code works.”

“I have a lot of confidence the code does what it says it does …I just looked at the people who built it, the way your guys think and what you are trying to achieve and realizing that the things I deal with don’t have much security concern so I just went with it.”

How is the Performance of IPFS?

PROVIDERS: Average

“Good. Go. Not a 10 because, connection closing.”

“When working with large data sets it seemed like I was running into performance issues, I was able to RSYNC files over to another server faster than I was able to run IPFS Pack on them, so I wasn't having a real gain there.”

PRODUCERS: Very Good

“To be fair – TB of data takes a long time, I think the genomes hashing took awhile, but overall performance is good.”

“Multi hashing is really really quick it always surprises me. All the API calls seem quick, I know publishing to the namespace stuff you guys are working on that”

How Scalable is IPFS?

PROVIDERS: (Split Perception)

“IPFS changes the game for scalability, this is a totally different paradigm.”

“I’ve had some experience on small IPFS things where it worked well, but when I scaled up and was trying to use IPFS PAC I couldn’t get it to work and I had to use a different solution.”

“In my experience with data.gov I ran into scale issues, with software and hardware, and I couldn’t even diagnose it…”

PRODUCERS: Excellent

“I haven’t played around with IPFS cluster yet. But it’s scalable and reliable and backed up, that’s the entire point of IPFS it’s the new decentralized web.”

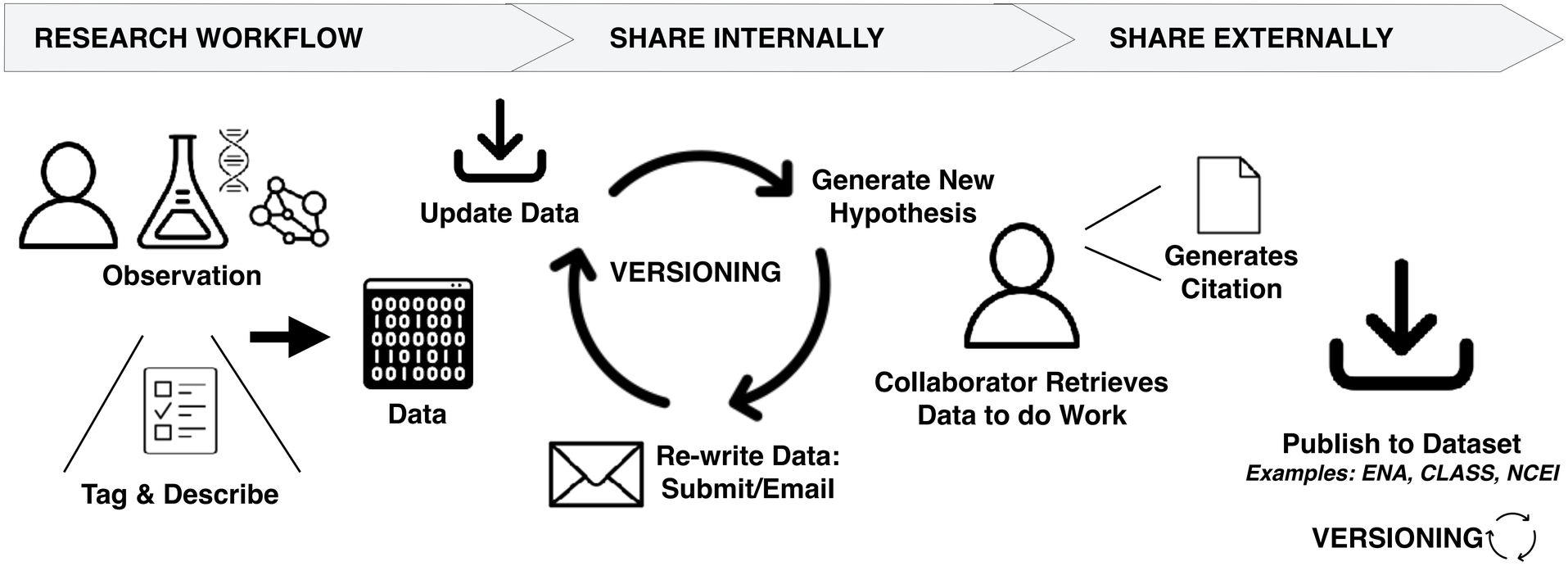
“I haven’t tested it with very large scale data I know it holds 100s MB in the private directory in the home folder so I haven’t noticed scaling problems.”

The User Journey

Illustrations to visualize the journey and areas of opportunity for IPFS ‘Large Volumes’

Mapping the Journey:

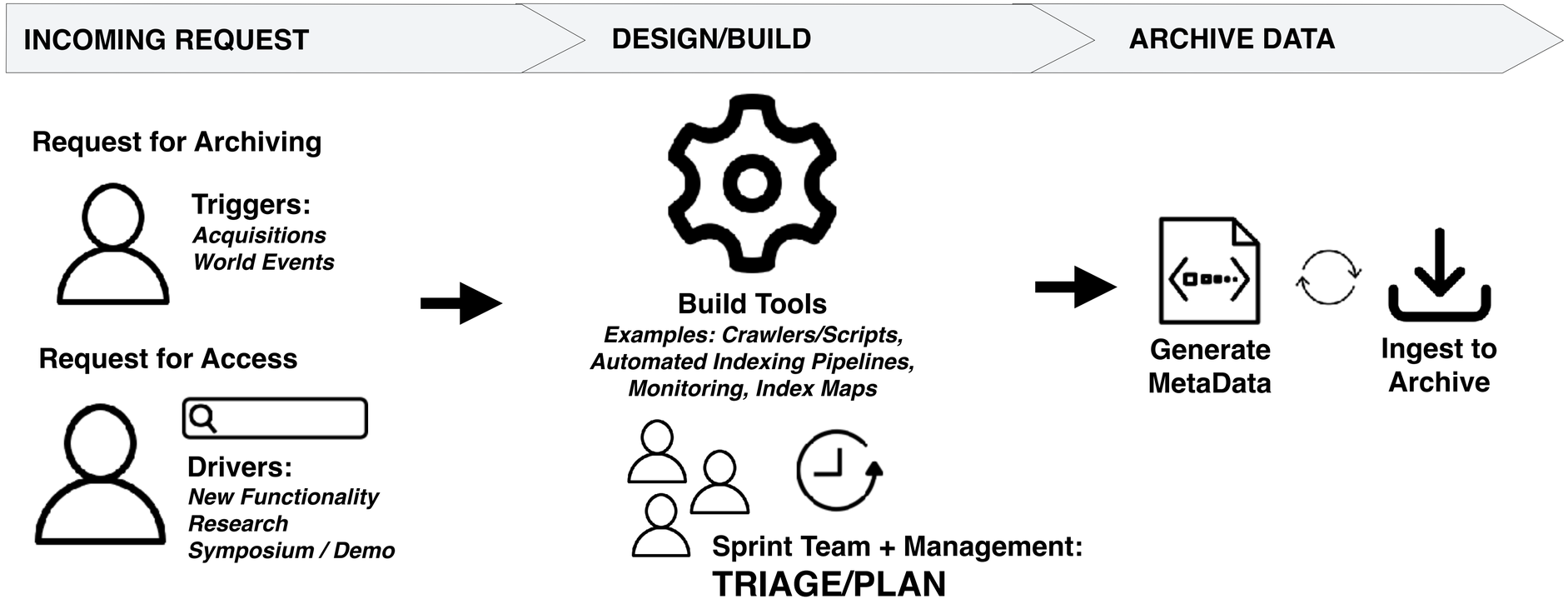
Producers



Touchpoints *(How people do this now)*	**Lab Software** **LIMS** **Jupyter Notebooks** **XML Markup/Metadata**	**FTP** **RSYNC** **Email** **Google Drive** **Databases**	**Citation Network (Mendeley)** **Patents**
Opportunity for IPFS	*How might we better integrate popular tools with IPFS?*	*How might we make versioning and internal sharing in IPFS good enough to replace existing tools?*	*How can we make it clear when a data set is updated?*

Mapping the Journey:

Providers



Touchpoints *(How people do this now)*	**Email** **Stakeholders** **Partners**	**FTP** **Email** **Google Docs** **JIRA/Product Planning** **Scripting/Coding Platforms**	**XML Documents** **Datastore** **Data Sharing Platforms (Figshare)** **AWS Public Data Sets**
Opportunity for IPFS	*How might we make indexing and searching through distributed content easier?*	*How might we make it easier for the people building these tools to use IPFS?*	*How might we support the activities of archiving and stewardship of data sets?*

What's Next?

Workshop it.

How can we work together to apply these findings to our work? What are the areas for further inquiry?

Appendix

Additional Resources

	Participant 1 EDGI	Participant 2 Stanford	Participant 3 Jonny / Transcendex
Reliability	**7.5** * “Just the frequency of change, we aren’t at a 1.0 so no 10 until it’s a stable API”	**4** * “I’ve had some experience on small IPFS things where it worked well, but when I scaled up and was trying to use IPFS PAC I couldn’t get it to work and I had to use a different solution.”	**9.5** “It’s been running for the last year other than when I’m doing updates.”
Security	**6** “Private keys in the config file, not so good.”	**?** “I haven’t looked into it.”	**9** * “So far as I know it’s secure. I’ve been digging deep into Go for key storage, and there are limitations that are beyond me, I trust the code works.”
Performance	**7 or 8** “Good. Go. Not a 10 because, connection closing.”	**5** “When working with large data sets it seemed like I was running into performance issues, I was able to RSYNC files over to another server faster than I was able to run IPFS PAC on them, so I wasn't having a real gain there.”	**7** “Fair – TB of data takes a long time, I think the genomes hashing took awhile, but overall performance is good.”
Scalability	**10** “IPFS changes the game for scalability, this is a totally different paradigm.”	**2** “In my experience with data.gov I ran into scale issues, with software and hardware, and I couldn’t even diagnose it..”	**10** “I haven’t played around with IPFS cluster yet. But it’s scalable and reliable and backed up, that’s the entire point of IPFS it’s the new decentralized web.”
	** Indicates the participant’s #1 priority*		

	Participant 4 NOAA	Participant 5 ENA
Reliability	**7** “There’s positive buzz, it’s a fully uneducated rating, based on what i've seen on twitter”	**7** “Because there are certain areas that come up in the running instance and publishing the namespace stuff takes awhile and I don’t have certainty around that. If it queries for a remote file that doesn't exist it takes a long time.”
Security	**7 or 8** * I’d rank FTP 1 and SFTP 1.5 on this scale. Ability to verify the data helps us in the security area. Cleaner and more well-defined connection between two points vs the feeding trough of FTP	**?** “I have a lot of confidence the code does what it says… I just looked at the people who built it, the way your guys thins and what you are trying to achieve and realizing that the things I deal with don’t have much security concern so I just went with it.”
Performance	**6** “Uneducated guess.”	**10** “Multihashing is really really quick it always surprises me. All the API calls seem quick, I know publishing to the namespace stuff you guys are working on that”
Scalability	**6** “Equally uneducated guess.”	**?** * “I haven’t tested it with very large scale data I know it holds 100s MB in the private directory in the home folder so I haven’t noticed scaling problems.”
	** Indicates the participant’s #1 priority*	