Leibniz Universität IT Services

Leibniz Universität Hannover

Distributed Research Data Management

eScience-Tage Heidelberg 2017

Reiko Kaps [kaps@luis.uni-hannover.de]

Leibniz Universität IT Services (LUIS)

16.03.2017

über mich

- 2000 - 2006 ... DevOp, Entwickler

über mich

- 2000 - 2006 ... DevOp, Entwickler
- 2006 - 2015 ... Redaktion c’t und Heise online

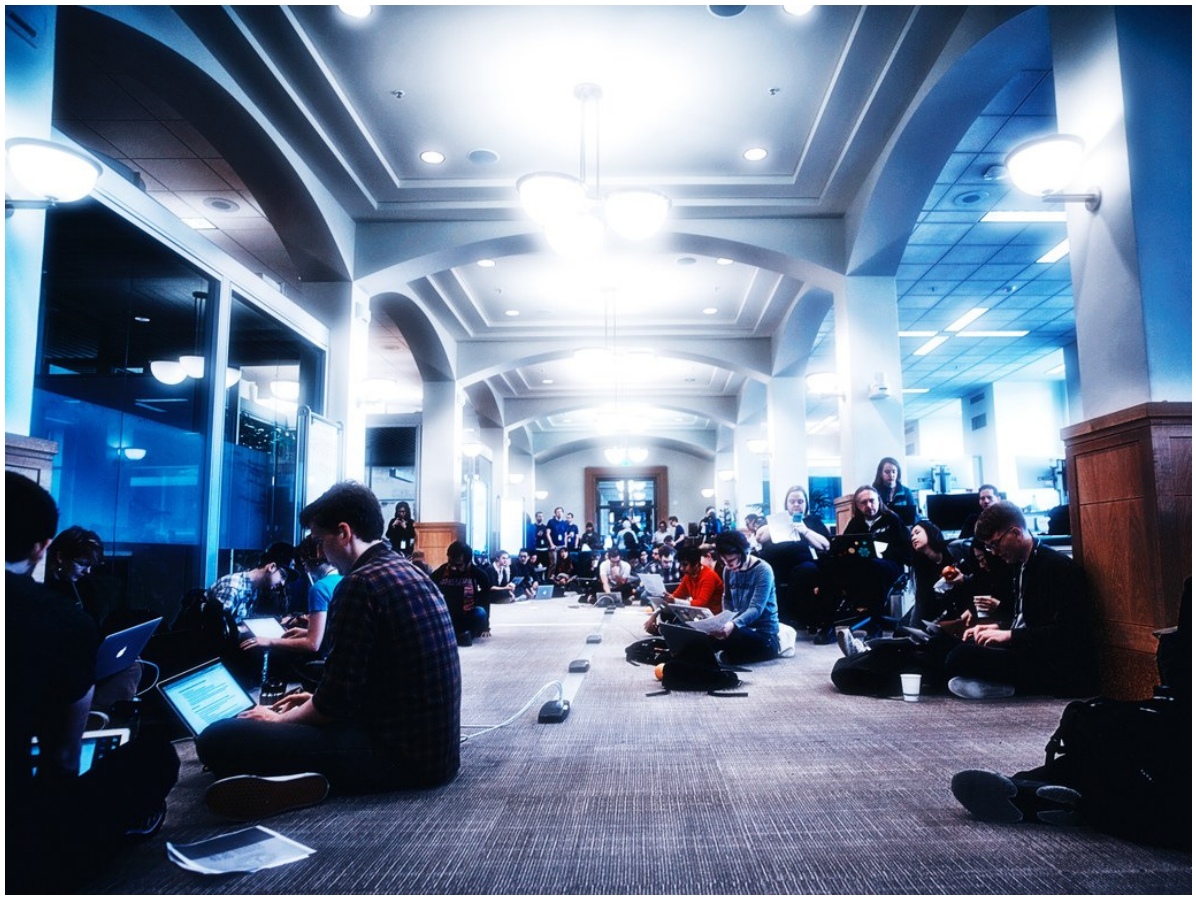
über mich

- 2000 - 2006 ... DevOp, Entwickler
- 2006 - 2015 ... Redaktion c’t und Heise online
- 2015 - 2016 ... Mitarbeit im Forschungsdaten-Projekt der Leibniz Universität Hannover

über mich

- 2000 - 2006 ... DevOp, Entwickler
- 2006 - 2015 ... Redaktion c’t und Heise online
- 2015 - 2016 ... Mitarbeit im Forschungsdaten-Projekt der Leibniz Universität Hannover
- seit 06/2016 .. Umsetzungsphase, Aufbau eines institutionellen Daten-Repositoriums

#Datarefuge



But instead of enjoying the beautiful day, 200 adults had willingly sardined themselves into a fluorescent-lit room in the bowels of Doe Library to rescue federal climate data.
Wired, Februar 2017

Warum also dieser Aufwand?

Warum also dieser Aufwand?

Die erste Antwort

hängt mit diesem Mann
zusammen:

Warum also dieser Aufwand?

Die erste Antwort

hängt mit diesem Mann zusammen:

Warum also dieser Aufwand?

Die erste Antwort

hängt mit diesem Mann zusammen:

Die zweite Antwort

ist leider komplizierter!

Client-Server-Problem I

- IT-Angebote für Forschungsdaten orientieren sich am Client-Server-Modell:

Client-Server-Problem I

- IT-Angebote für Forschungsdaten orientieren sich am Client-Server-Modell:
  - das zentral Daten aufnimmt, veröffentlicht und vorhält (Single point of failure)

Client-Server-Problem I

- IT-Angebote für Forschungsdaten orientieren sich am Client-Server-Modell:
  - das zentral Daten aufnimmt, veröffentlicht und vorhält (Single point of failure)
  - das Redundanz und Verfügbarkeit teuer erkaufen muss (Skalierbarkeit)

Client-Server-Problem I

- IT-Angebote für Forschungsdaten orientieren sich am Client-Server-Modell:
  - das zentral Daten aufnimmt, veröffentlicht und vorhält (Single point of failure)
  - das Redundanz und Verfügbarkeit teuer erkaufen muss (Skalierbarkeit)
  - dessen Integrität vom Betreiber abhängt

Client-Server-Problem II

- in Form von Dienste-Inseln, die sich an kommerziellen IT-Angeboten orientieren

Client-Server-Problem II

- in Form von Dienste-Inseln, die sich an kommerziellen IT-Angeboten orientieren
- das Nutzer als Kunden betrachtet

Client-Server-Problem II

- in Form von Dienste-Inseln, die sich an kommerziellen IT-Angeboten orientieren
- das Nutzer als Kunden betrachtet
- das den Ausstellungscharakter der Daten betont

Client-Server-Problem II

- in Form von Dienste-Inseln, die sich an kommerziellen IT-Angeboten orientieren
- das Nutzer als Kunden betrachtet
- das den Ausstellungscharakter der Daten betont
- das auf HTTP(S) aufsetzt (Bandbreitenverschwendung)

Client-Server-Problem II

- in Form von Dienste-Inseln, die sich an kommerziellen IT-Angeboten orientieren
- das Nutzer als Kunden betrachtet
- das den Ausstellungscharakter der Daten betont
- das auf HTTP(S) aufsetzt (Bandbreitenverschwendung)
- Nachteile lassen sich durch Dezentralisierung mindern

Neue Herausforderungen

nicht nur für Forschungsdaten

- dauerhaftes Bereitstellen und Verteilen riesiger Datensätze

Neue Herausforderungen

nicht nur für Forschungsdaten

- dauerhaftes Bereitstellen und Verteilen riesiger Datensätze
- Verarbeiten großer Datensätze über Organisationsgrenzen hinweg

Neue Herausforderungen

nicht nur für Forschungsdaten

- dauerhaftes Bereitstellen und Verteilen riesiger Datensätze
- Verarbeiten großer Datensätze über Organisationsgrenzen hinweg
- Versionierung und Verknüpfung großer Datensätze

Neue Herausforderungen

nicht nur für Forschungsdaten

- dauerhaftes Bereitstellen und Verteilen riesiger Datensätze
- Verarbeiten großer Datensätze über Organisationsgrenzen hinweg
- Versionierung und Verknüpfung großer Datensätze
- Verlust wichtiger Dateien verhindern

Neue Herausforderungen

nicht nur für Forschungsdaten

- dauerhaftes Bereitstellen und Verteilen riesiger Datensätze
- Verarbeiten großer Datensätze über Organisationsgrenzen hinweg
- Versionierung und Verknüpfung großer Datensätze
- Verlust wichtiger Dateien verhindern
- Integrität, Nachvollziehbarkeit sicherstellen

Forderungen

- permanente Datenspeicherung

Forderungen

- permanente Datenspeicherung
- robuster und verlässlicher Daten-Zugang (Offline-Nutzung)

Forderungen

- permanente Datenspeicherung
- robuster und verlässlicher Daten-Zugang (Offline-Nutzung)
- Nachvollziehbarkeit, Transparenz und Sicherheit

Forderungen

- permanente Datenspeicherung
- robuster und verlässlicher Daten-Zugang (Offline-Nutzung)
- Nachvollziehbarkeit, Transparenz und Sicherheit
- Offener Zugang und Nutzung durch jedermann

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

- Verteilen von Inhalten (Bittorrent)

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

- Verteilen von Inhalten (Bittorrent)
  - verteilt und signiert große Dateien

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

- Verteilen von Inhalten (Bittorrent)
  - verteilt und signiert große Dateien
  - Entkoppelt die Inhalte von ihrer Quelle (Content Addressing)

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

- Verteilen von Inhalten (Bittorrent)
  - verteilt und signiert große Dateien
  - Entkoppelt die Inhalte von ihrer Quelle (Content Addressing)
  - nutzt Netzwerk-Infrastruktur effektiv aus (Schwarm, Peers)

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

- Verteilen von Inhalten (Bittorrent)
  - verteilt und signiert große Dateien
  - Entkoppelt die Inhalte von ihrer Quelle (Content Addressing)
  - nutzt Netzwerk-Infrastruktur effektiv aus (Schwarm, Peers)
  - steuert die Last beim Inhalteanbieter

Auswege

finden sich bei den Peer-to-Peer-Konzepten der vergangenen 20 Jahre

- Verteilen von Inhalten (Bittorrent)
  - verteilt und signiert große Dateien
  - Entkoppelt die Inhalte von ihrer Quelle (Content Addressing)
  - nutzt Netzwerk-Infrastruktur effektiv aus (Schwarm, Peers)
  - steuert die Last beim Inhalteanbieter
  - aktuelle Nutzung: Netflix, Windows Update

Auswege II

- Versionierung von Inhalten (Git)

Auswege II

- Versionierung von Inhalten (Git)
  - protokolliert Änderungen der Dateiinhalte

Auswege II

- Versionierung von Inhalten (Git)
  - protokolliert Änderungen der Dateiinhalte
  - erlaubt verteilte, nicht-lineare Arbeitsabläufe

Auswege II

- Versionierung von Inhalten (Git)
  - protokolliert Änderungen der Dateiinhalte
  - erlaubt verteilte, nicht-lineare Arbeitsabläufe
  - Online- und Offline-Nutzung

Auswege II

- Versionierung von Inhalten (Git)
  - protokolliert Änderungen der Dateiinhalte
  - erlaubt verteilte, nicht-lineare Arbeitsabläufe
  - Online- und Offline-Nutzung
  - gewährleistet Nachvollziehbarkeit

Auswege III

- Nachverfolgung von Inhalten und Transaktionen (Blockchain)

Auswege III

- Nachverfolgung von Inhalten und Transaktionen (Blockchain)
  - Blockchain: Datenbank mit mathematischem Integritätsansatz. Der Hashwertes (Integritätsgarant) eines Datensatzes wird im Hashwert des jeweils nachfolgenden gesichert.

Auswege III

- Nachverfolgung von Inhalten und Transaktionen (Blockchain)
  - Blockchain: Datenbank mit mathematischem Integritätsansatz. Der Hashwertes (Integritätsgarant) eines Datensatzes wird im Hashwert des jeweils nachfolgenden gesichert.
  - Blockchain-Verfahren ist die Grundlage für Kryptowährungen, seine Funktion ähnelt dem Journal in der Buchführung.

Auswege III

- Nachverfolgung von Inhalten und Transaktionen (Blockchain)
  - Blockchain: Datenbank mit mathematischem Integritätsansatz. Der Hashwertes (Integritätsgarant) eines Datensatzes wird im Hashwert des jeweils nachfolgenden gesichert.
  - Blockchain-Verfahren ist die Grundlage für Kryptowährungen, seine Funktion ähnelt dem Journal in der Buchführung.
  - vereinfacht und verbessert Transaktionssicherkeit in verteilten Systemen

IPFS I

- IPFS == Interplanetary File System
  (Referenz an Lickliders Intergalactic Computer Network, 1963)
  `https://ipfs.io`

IPFS I

- IPFS == Interplanetary File System
  (Referenz an Lickliders Intergalactic Computer Network, 1963)
  `https://ipfs.io`
- Vereint Web,- Bittorrent-, Git- und Blockchain-Funktionen in einem verteilten Dateisystem

IPFS I

- IPFS == Interplanetary File System
  (Referenz an Lickliders Intergalactic Computer Network, 1963)
  `https://ipfs.io`
- Vereint Web,- Bittorrent-, Git- und Blockchain-Funktionen in einem verteilten Dateisystem
- 2014, Juan Benet `https://twitter.com/juanbenet`

IPFS I

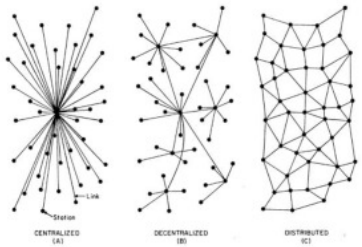
- IPFS == Interplanetary File System
  (Referenz an Lickliders Intergalactic Computer Network, 1963)
  `https://ipfs.io`
- Vereint Web,- Bittorrent-, Git- und Blockchain-Funktionen in einem verteilten Dateisystem
- 2014, Juan Benet `https://twitter.com/juanbenet`
- Whitepaper `https://github.com/ipfs/papers/raw/master/ipfs-cap2pfs/ipfs-p2p-file-system.pdf`

IPFS I

- IPFS == Interplanetary File System
  (Referenz an Lickliders Intergalactic Computer Network, 1963)
  https://ipfs.io
- Vereint Web,- Bittorrent-, Git- und Blockchain-Funktionen in einem verteilten Dateisystem
- 2014, Juan Benet https://twitter.com/juanbenet
- Whitepaper https://github.com/ipfs/papers/raw/master/ipfs-cap2pfs/ipfs-p2p-file-system.pdf
- Quelloffene Referenzimplementierung in Go
  https://github.com/ipfs/ipfs

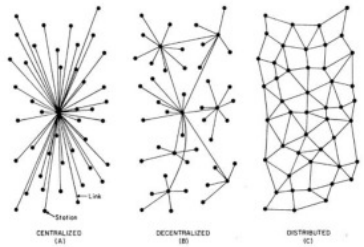
IPFS II

- vollständig verteiltes Netzwerkdateisystem



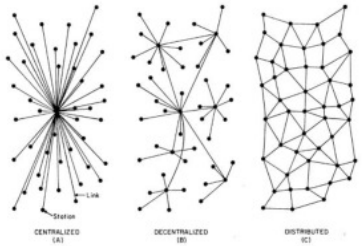
IPFS II

- vollständig verteiltes Netzwerkdateisystem
- Knoten/Nodes: agieren *als Server und Client*



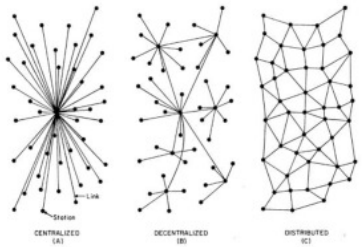
IPFS II

- vollständig verteiltes Netzwerkdateisystem
- Knoten/Nodes: agieren *als Server und Client*
- Einzelne Node ist mit jeder anderen verbunden



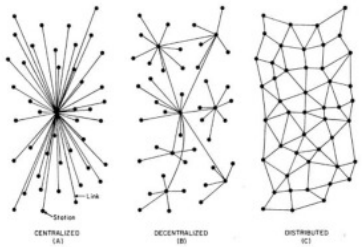
IPFS II

- vollständig verteiltes Netzwerkdateisystem
- Knoten/Nodes: agieren *als Server und Client*
- Einzelne Node ist mit jeder anderen verbunden
- adressiert Inhalte über Hashes (Merkle-Links)



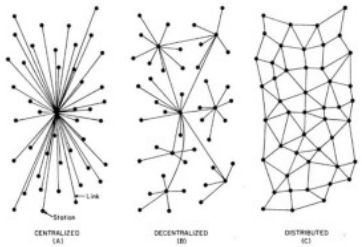
IPFS II

- vollständig verteiltes Netzwerkdateisystem
- Knoten/Nodes: agieren *als Server und Client*
- Einzelne Node ist mit jeder anderen verbunden
- adressiert Inhalte über Hashes (Merkle-Links)
- dedupliziert Dateien innerhalb des IPFS-Netzes



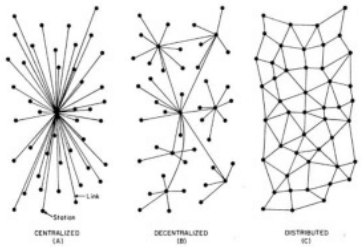
IPFS III

- Lädt Inhalte parallel von mehreren, möglichst nahegelegenden Nodes (wenn verfügbar)



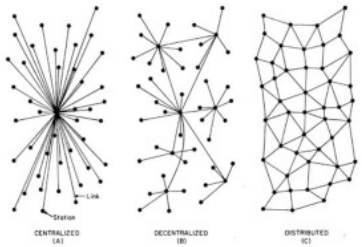
IPFS III

- Lädt Inhalte parallel von mehreren, möglichst nahegelegenden Nodes (wenn verfügbar)
- Knoten können fremde Inhalte dauerhaft vorhalten (pinning)



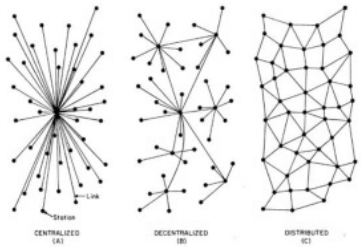
IPFS III

- Lädt Inhalte parallel von mehreren, möglichst nahegelegenden Nodes (wenn verfügbar)
- Knoten können fremde Inhalte dauerhaft vorhalten (pinning)
- protokolliert die Änderungsgeschichte jeder Datei



IPFS III

- Lädt Inhalte parallel von mehreren, möglichst nahegelegenden Nodes (wenn verfügbar)
- Knoten können fremde Inhalte dauerhaft vorhalten (pinning)
- protokolliert die Änderungsgeschichte jeder Datei
- Dateien über lesbare Namen auffindbar (IPNS)



Wer kann davon profitieren?

- Inhalteproduzenten und -anbieter

Wer kann davon profitieren?

- Inhalteproduzenten und -anbieter
- Forschende

Wer kann davon profitieren?

- Inhalteproduzenten und -anbieter
- Forschende
- Infrastrukturbetreiber

Wer kann davon profitieren?

- Inhalteproduzenten und -anbieter
- Forschende
- Infrastrukturbetreiber
- Bibliotheken und Archive

Wer kann davon profitieren?

- Inhalteproduzenten und -anbieter
- Forschende
- Infrastrukturbetreiber
- Bibliotheken und Archive
- ...

IPFS und FDM

- Data Federation: globale Infrastruktur für Forschungsdaten

IPFS und FDM

- Data Federation: globale Infrastruktur für Forschungsdaten
- macht jeden Nutzer zum Teil der Infrastruktur

IPFS und FDM

- Data Federation: globale Infrastruktur für Forschungsdaten
- macht jeden Nutzer zum Teil der Infrastruktur
- hohe Zensur- und Ausfallsicherheit

IPFS und FDM

- Data Federation: globale Infrastruktur für Forschungsdaten
- macht jeden Nutzer zum Teil der Infrastruktur
- hohe Zensur- und Ausfallsicherheit
- optimiert die Netzsicherheit

IPFS und FDM

- Data Federation: globale Infrastruktur für Forschungsdaten
- macht jeden Nutzer zum Teil der Infrastruktur
- hohe Zensur- und Ausfallsicherheit
- optimiert die Netzsicherheit
- eigene Anwendungen on top

IPFS und FDM

- Data Federation: globale Infrastruktur für Forschungsdaten
- macht jeden Nutzer zum Teil der Infrastruktur
- hohe Zensur- und Ausfallsicherheit
- optimiert die Netzsicherheit
- eigene Anwendungen on top
- Inhalte: Open Access zu Commons

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen
- IPFS:

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen
- IPFS:
  - lässt sich sehr einfach an andere Teilnehmer delegieren

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen
- IPFS:
  - lässt sich sehr einfach an andere Teilnehmer delegieren
  - große Freiheit bei der Location-Wahl (anderer Kontinent)

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen
- IPFS:
  - lässt sich sehr einfach an andere Teilnehmer delegieren
  - große Freiheit bei der Location-Wahl (anderer Kontinent)
  - Denkbar: Gegenseitigkeitsprinzip (Peering) aber auch Mietmodelle
  - ...

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen
- IPFS:
  - lässt sich sehr einfach an andere Teilnehmer delegieren
  - große Freiheit bei der Location-Wahl (anderer Kontinent)
  - Denkbar: Gegenseitigkeitsprinzip (Peering) aber auch Mietmodelle
  - ...

Beispiel: Zweitkopie

- verdoppelt Kosten für Speichermedien und Standort
- Problem für kleinere wissenschaftliche Organisationen
- IPFS:
  - lässt sich sehr einfach an andere Teilnehmer delegieren
  - große Freiheit bei der Location-Wahl (anderer Kontinent)
  - Denkbar: Gegenseitigkeitsprinzip (Peering) aber auch Mietmodelle
  - ...

IMHO: ausreichend Gründe für einen Feldtest

Call for Participation

- Gemeinsames Testbed für eine
  `Föderierte Forschungsdaten-Infrastruktur (F2DI)`

Call for Participation

- Gemeinsames Testbed für eine
  `Föderierte Forschungsdaten-Infrastruktur (F2DI)`
- Wissenschaftliche Einrichtungen bringen alles nötige mit:
  - Infrastruktur und Anbindung
  - Kritische Masse an Teilnehmern
  - ausreichend große und wichtige Datenbestände
  - Know-how bei Entwicklung und Betrieb
  - Potenzial der Blockchain bereits erkannt

Fragen?

Fragen?

Kontakt

- Mail: kaps@luis.uni-hannover.de
- Twitter: `https://twitter.com/reik_kaps`

Vielen Dank für die Aufmerksamkeit!

Inhalt

Lesenswertes

- Juan Benet, IPFS Whitepaper
  https://arxiv.org/abs/1407.3561
- Geocities@IPFS https://ipfs.io/ipfs/QmNhFJjGcMPqpuYfxL62VVB9528NXqDNMFXiqN5bgFYiZ1/its-time-for-the-permanent-web.html
- Internet Archive: Call for a distributed web

  http://blog.archive.org/2015/02/11/locking-the-web-open-a-call-for-a-distributed-web/