Sondermittelprojekt 2016

Überblick und Ausblick auf die neue Architektur

Workshop für (künftige) DDB-Aggregatoren aus der Sparte Archiv

Frankfurt, 18. April 2016

Stephan Bartholmei
s.bartholmei@dnb.de

DEUTSCHE NATIONAL BIBLIOTHEK

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

Sondermittelprojekt 2016

Ziele

- Modernisierung der IT-Infrastruktur
  - **Performanz**-Steigerung: Datendurchsatz, Dauer (Re)Indexierung, Ingestdauer
  - **Skalierbarkeit**
  - Computing Engine zur Unterstützung neuer Use-Cases, z.B.:
    - Daten**analysen** (→ Datenqualität, Statistik)
    - Datenanreicherung, interne und externe Verknüpfungen
    - nachgelagerte, asynchrone Updates
    - Datendumps
  - Entkopplung Produktiv- und Testsysteme
- Unterstützung dezentraler Szenarien
  - Administrationsoberfläche, Selbstbedienungskomponente, Harvesting
- Verbesserung der Suche

Sondermittelprojekt 2016

Arbeitspakete aus Projektskizze (werden ggf. neu gefasst)

- AP 1 – Überarbeitung der Gesamtarchitektur
  - AP 1.1 – Technologie-Review und –Evaluierung
  - AP 1.2 – Architekturskizze — Mitte/Ende Mai
  - AP 1.3 – Prototyping — 1. Iteration: Ende April
  - AP 1.4 – Feinkonzept — 30.6.
- AP 2 – Softwaretechnische Realisierung
  - AP 2.1 – Umbau der Speicherstruktur (Software)
  - AP 2.2 – Suchmaschinentechnologie und Re-Indexierung
  - AP 2.3 – Ingest- und Update-Prozesse
  - AP 2.4 – Suche und Volltextsuche
  - AP 2.5 – Administrationswerkzeuge
  - AP 2.6 – Selbstbedienungskomponenten
  - AP 2.7 – Harvestingkomponenten
  - AP 2.8 – DDB-Labs
  - AP 2.9 – Aktualisierungen im Portal
- AP 3 – Betriebliche Infrastruktur

Pain points

Status Quo & Lösungsansätze

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

- Datendurchsatz, (Re)Indexierungs- und Ingestdauer
  - Scale out (mehr, nicht schnellere Server), im-memory processing, ...
- Prozess-Schritte werden linear durchlaufen
  - Entkopplung, Parallelisierung
- Anziehen der Binaries = externer Flaschenhals
  - Entkopplung, asynchroner Download, Skalierung & Ingest
- abgebrochene Prozesse können nicht fortgesetzt werden
  - Speicherung von Statusinformationen, Einstieg an jedem Punkt der Prozesskette
- (Historisches Wachstum =>) Daten werden mehrfach „angefasst“
  - Konsolidierung der Prozesse
- Analysen, Datendumps via Suchmaschine unmöglich
  - Computing-Engine

Architektur

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

Status Quo

- 2 Produktionsstränge
  Master/Slave + Sharding
- ETL*-Strecke für XML-Dokumente
  (Transformation & Ingest)
- Solr als Suchindex und
  einzige „Datenbank“
- Filesystem als Repositorium
  für AIPs

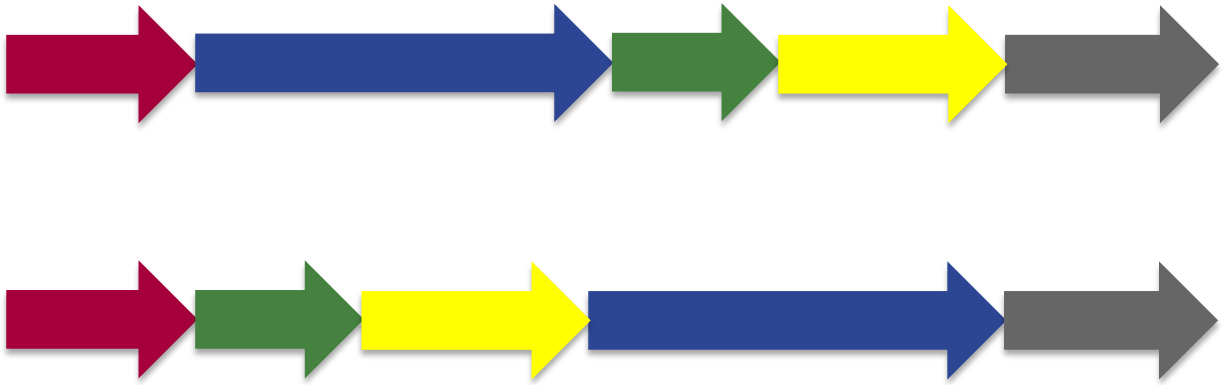
neue Architektur

- Cluster
- ETL-Strecke für XML-Dokumente
  entkoppelt, parallelisiert, asynchron
- Solr als Suchindex und
  NoSQL-Datenbank
- (vorauss.) NoSQL-Datenbank als
  Repositorium für XML-Artefakte und
  ausgewählte EDM-Pfade

*Extract, Transform & Load

Entkoppelung der Prozess-Schritte

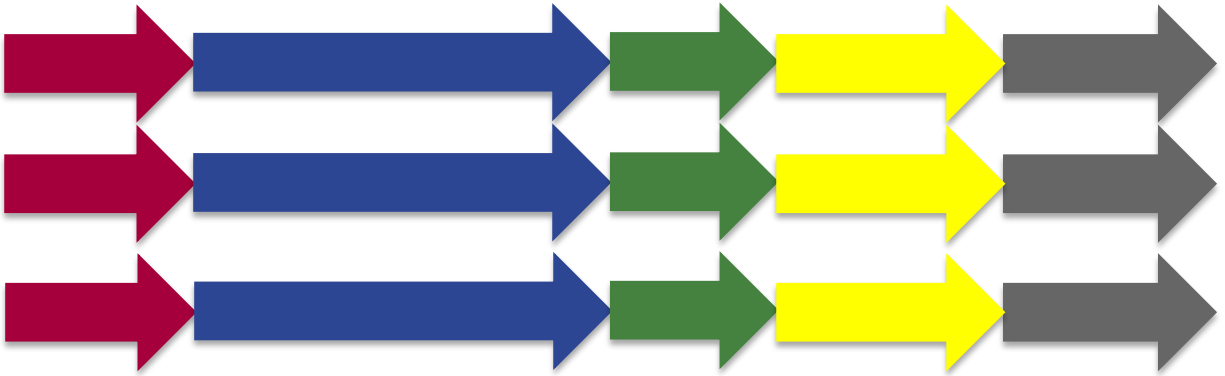
Prä-Ingest/Transformation (the tool formerly known as ASC)



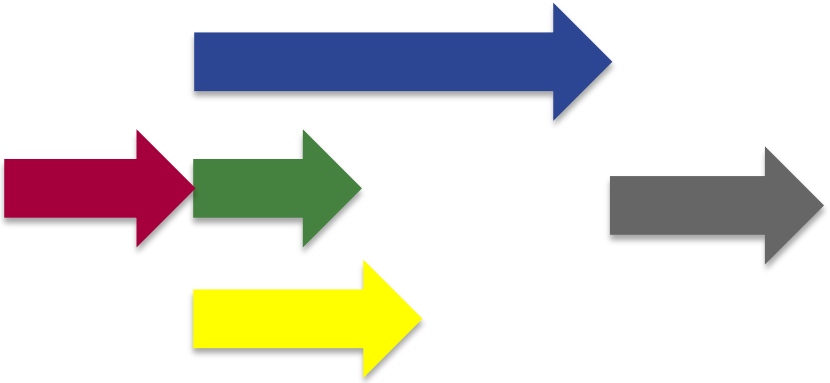
Parallelisierung

DEUTSCHE DIGITALE BIBLIOTHEK
Kultur und Wissen online

Prozesse



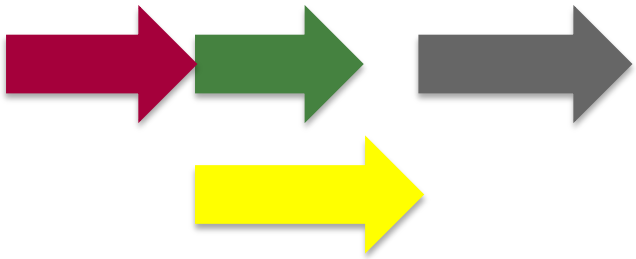
Prozess-
Schritte



Asynchrone Auskopplung

weniger Abhängigkeit von Service Qualität externer Dienste

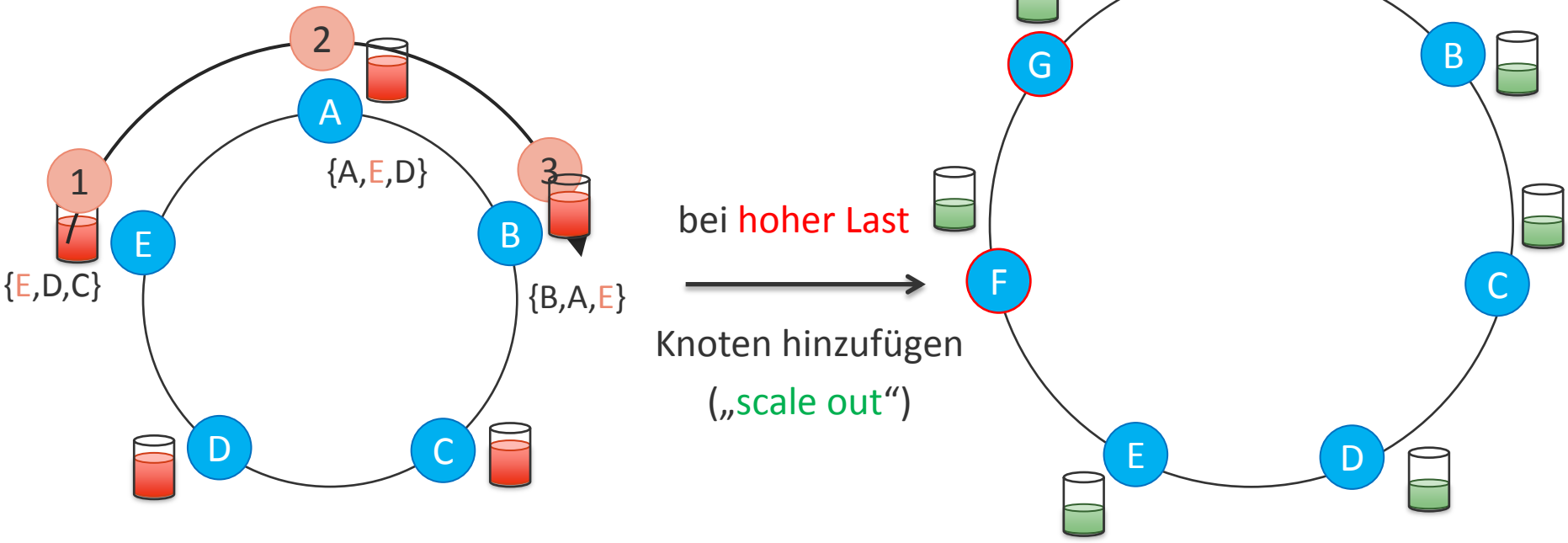
Binaries werden asynchron angezogen und ggf. nach Verfügbarkeit eingespielt



Metadaten sind ingestiert.

Neue Architektur

- Cassandra-Cluster
  - verteilte, nicht-relationale „NoSQL“-DB
  - kein „Master“, Replikation (typ. 3x)
  - hohe Verfügbarkeit, Performanz, Robustheit
  - aber: eventual consistency



Projektsteuerung & Zeitplan

- Projektlaufzeit 1. März 2016 – 31. Mai 2017
- für 1 Jahr 2 neue Kolleginnen in der Projektkoordination
  - Projektleitung: Dr. Tanja Jürs
  - Projektassistenz: Dona-Diana Dworak
- fachliche AP-Verantwortliche stehen im wesentlichen fest
- bis Ende April:
  - z.Zt. Vergleichsmessungen an Referenzsystem (FIZ)
  - Prototypen prä-Ingest (FIZ) und Ingest (IAIS) auf Cluster bei FIZ
- im Mai
  - Auswertung, Architekturskizze, Review (intern/extern), Technologie-Entscheidung
- Bis Ende Juni:
  - Feinkonzept liegt vor, 1. Bericht BKM
  - nicht oder lose von Architektur abhängige APe starten ggf. asynchron