Aufbau einer Plattform zur forensischen Analyse basierend auf dem Apache Hadoop® Framework

VON JOHANNES BUSAM

VERTEIDIGUNG DER MASTERTHESIS

26.11.2018

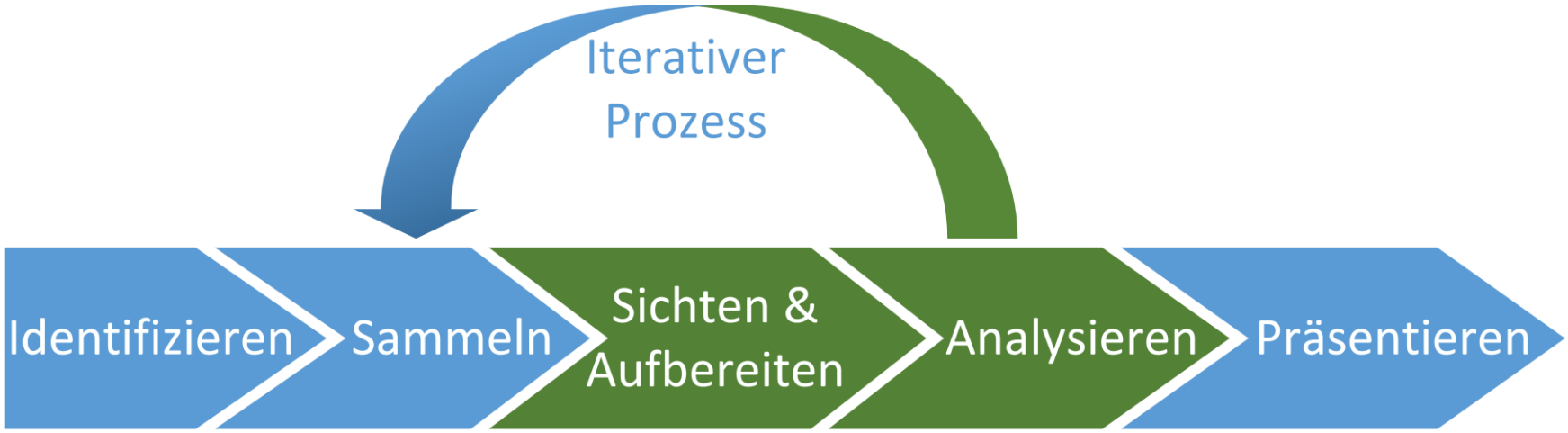
Agenda

- Motivation
- Grundlagen
- Datenspeicherung
- Datenverarbeitung
- Fazit
- Diskussion

Motivation

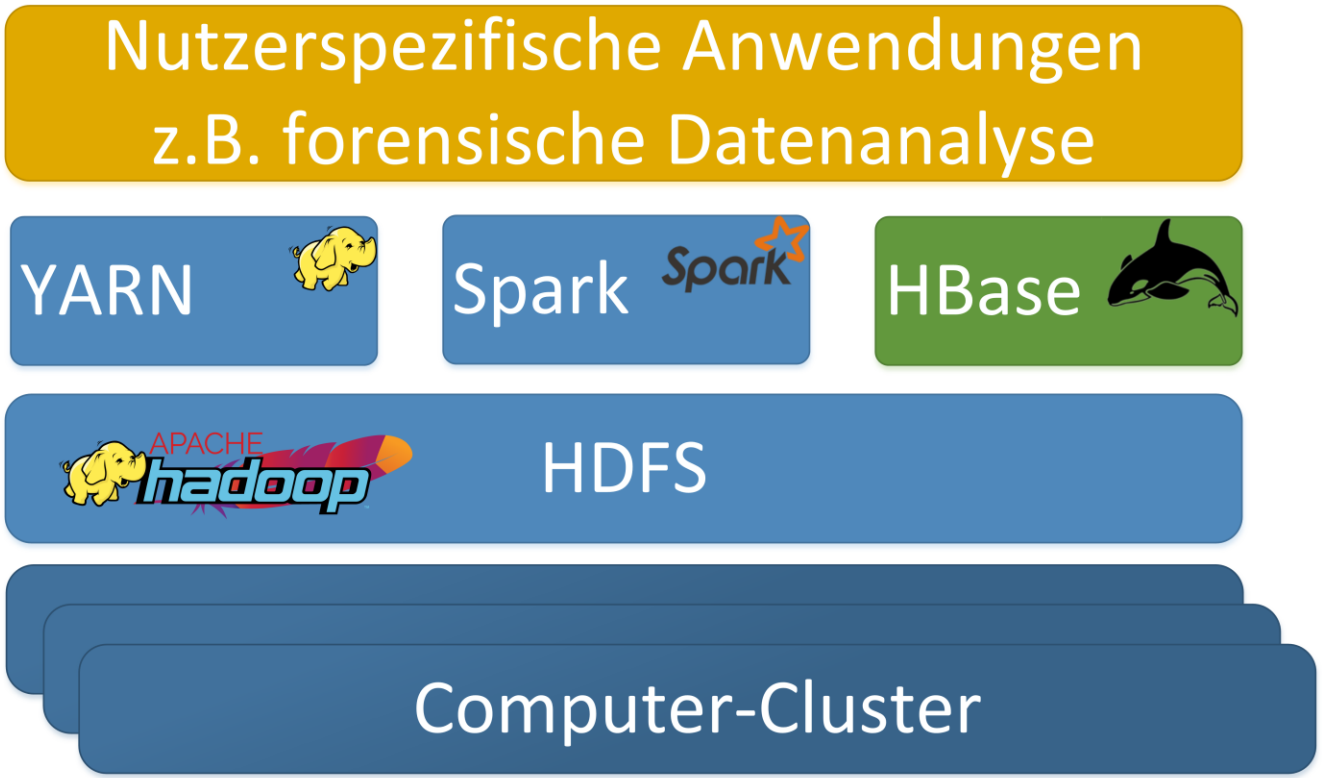
- Datenverarbeitung von großen Datenmengen
- Beschleunigung der forensischen Auswertung

Forensischer Analyseprozess



Grundlagen

Apache Hadoop® Ökosystem

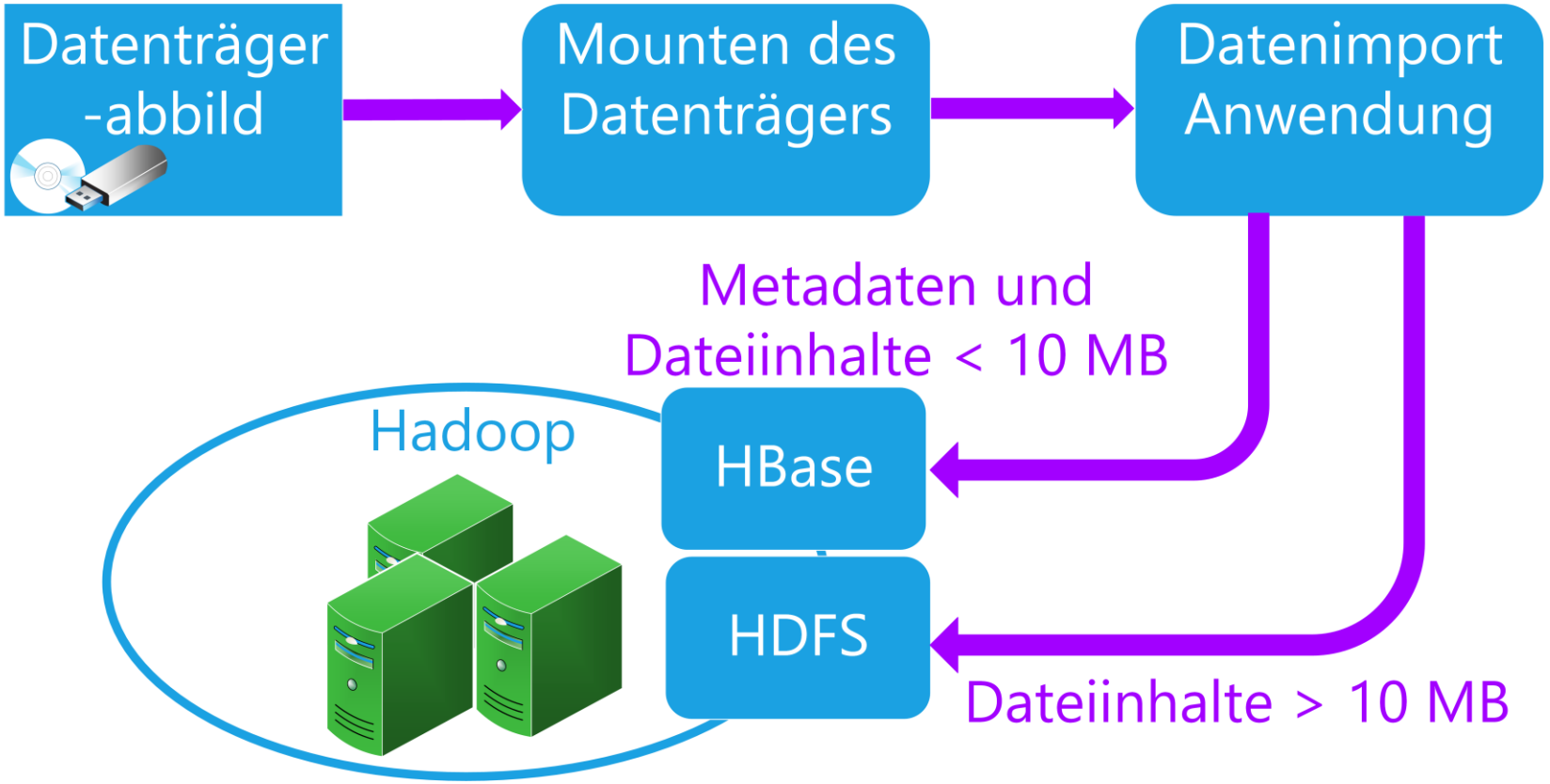


Daten-speicherung

Datenimport - Varianten

- Speicherung beliebiger Dateien und deren Metadaten
- Varianten:
  1. Datenträgerabbilder speichern
  2. Dateien im HDFS speichern
  3. Dateien in Dateicontainer speichern

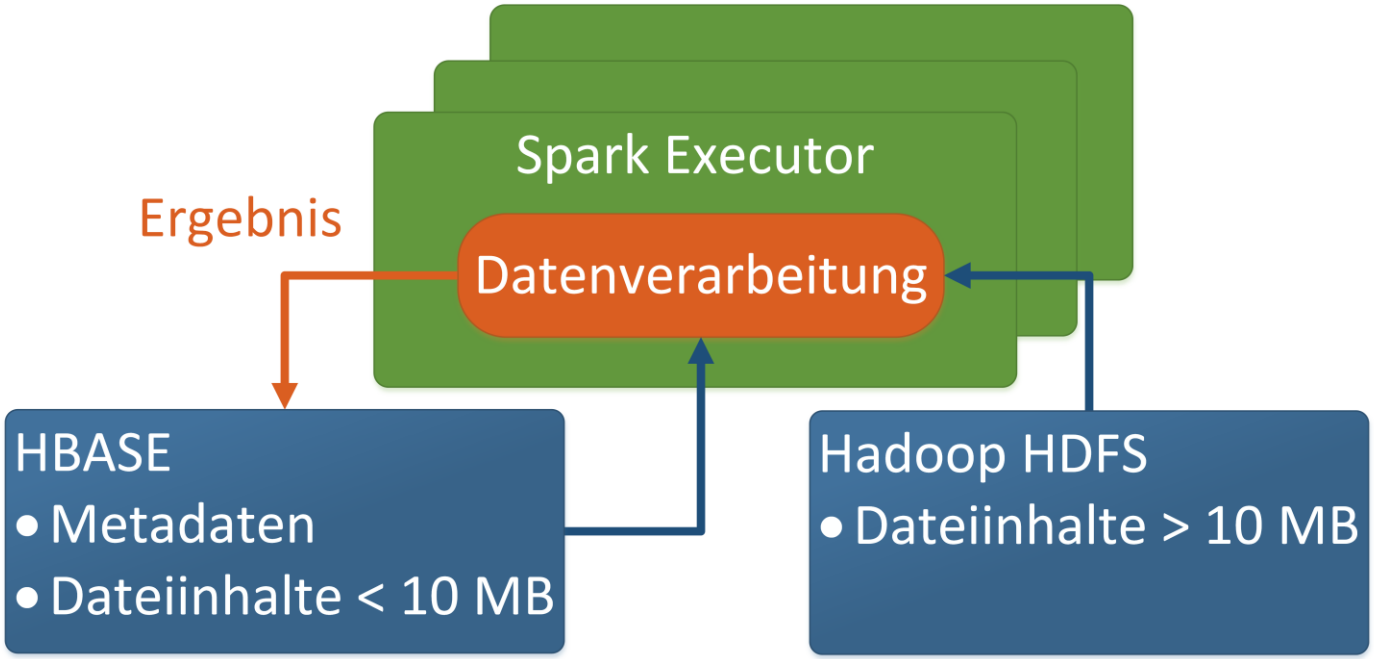
Datenimport – 4. Variante



Daten-verarbeitung

Datenverarbeitung mit Spark

- Ermittlung von Hashsummen und Medientypen



Volltextsuche / Indexierung

- Problemstellung: ***Datenzugriff***
  - Performanz ist wichtig
- Mögliche Lösung: **Apache Solr**
  - Integration in Hadoop-Cluster
  - Zugriff über REST-Schnittstelle

Fazit

Ergebnis

- Speicherung beliebiger Daten zur forensischen Auswertung
  - Vorteile von HDFS und HBase
- Verarbeitung der Daten mit Apache Spark
- Indexierung in Apache Solr
- Visualisierung über Web-Oberfläche

Ausblick

- ***Streaming-Data*** Ansatz
- Optimierungen beim Datenimport
- Erweiterung der Datenverarbeitung
- Sicherheit (Kerberos, Verschlüsselung)
- Datenvisualisierung

Vielen Dank für Ihre Aufmerksamkeit

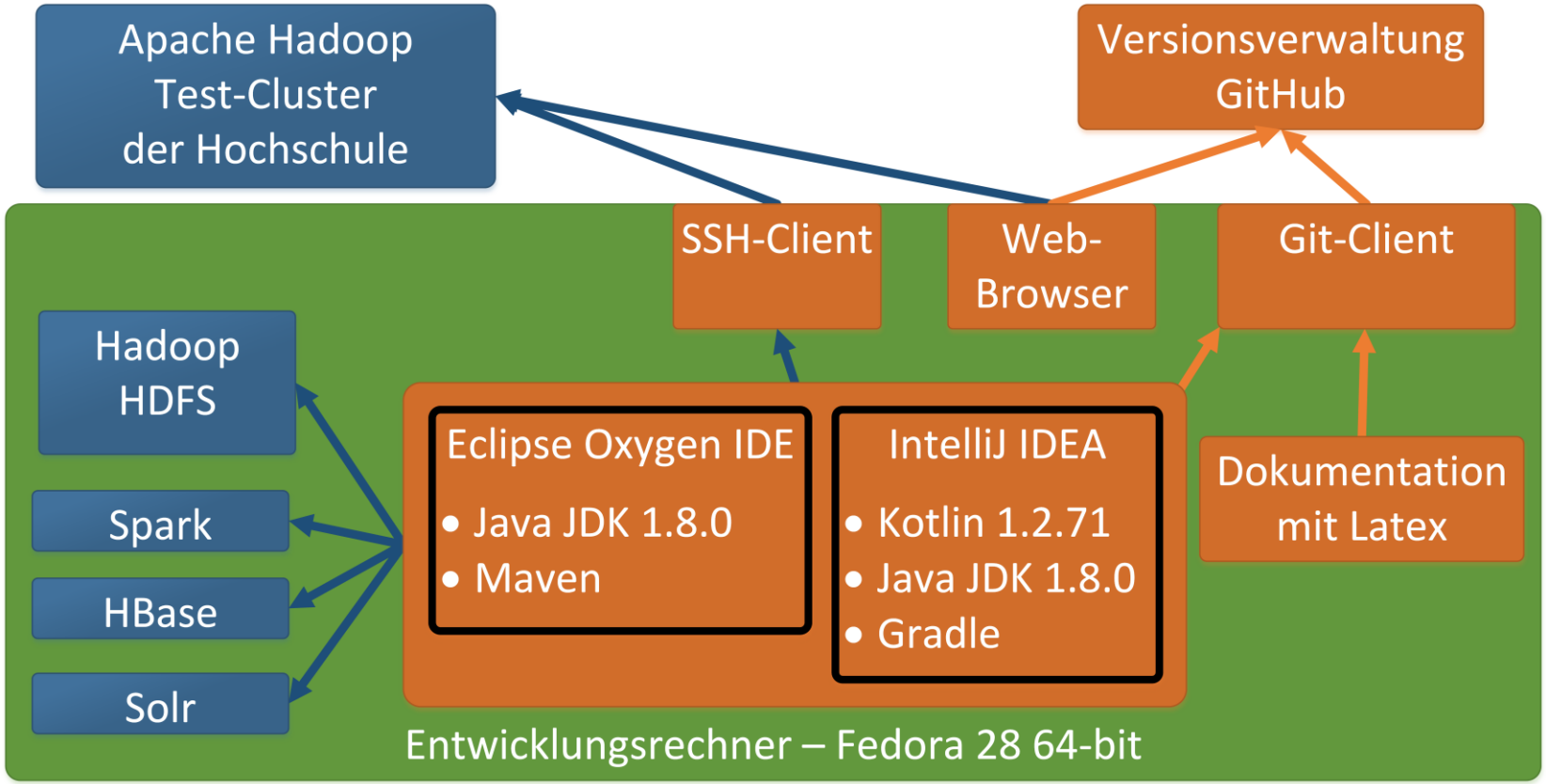
www.github.com/jobusam

Literaturverzeichnis

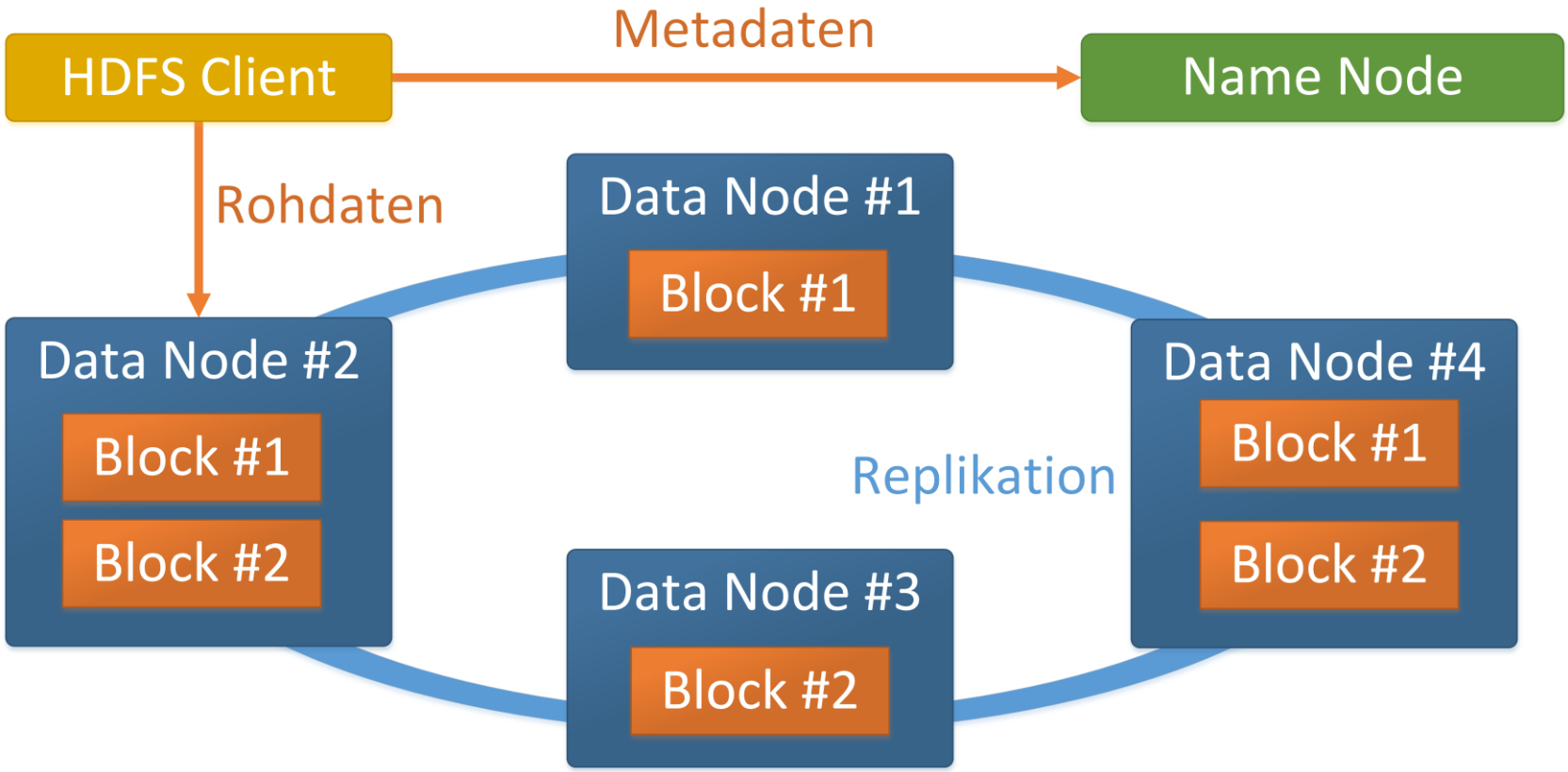
- Folie 4: Forensischer Analyseprozess modifiziert nach André Årnes. Digital Forensics.
  1. Auflage. Wiley, 2017
- Folie 6: Hadoop Ökosystem modifiziert nach
  Jonas Freiknecht. Big Data in der Praxis. 1. Auflage. Hanser Verlag, 2014.
  und Sam R. Alapati. Expert Hadoop Administration. 1. Auflage. Addison Wesley, 2016.

Zusatzinhalte

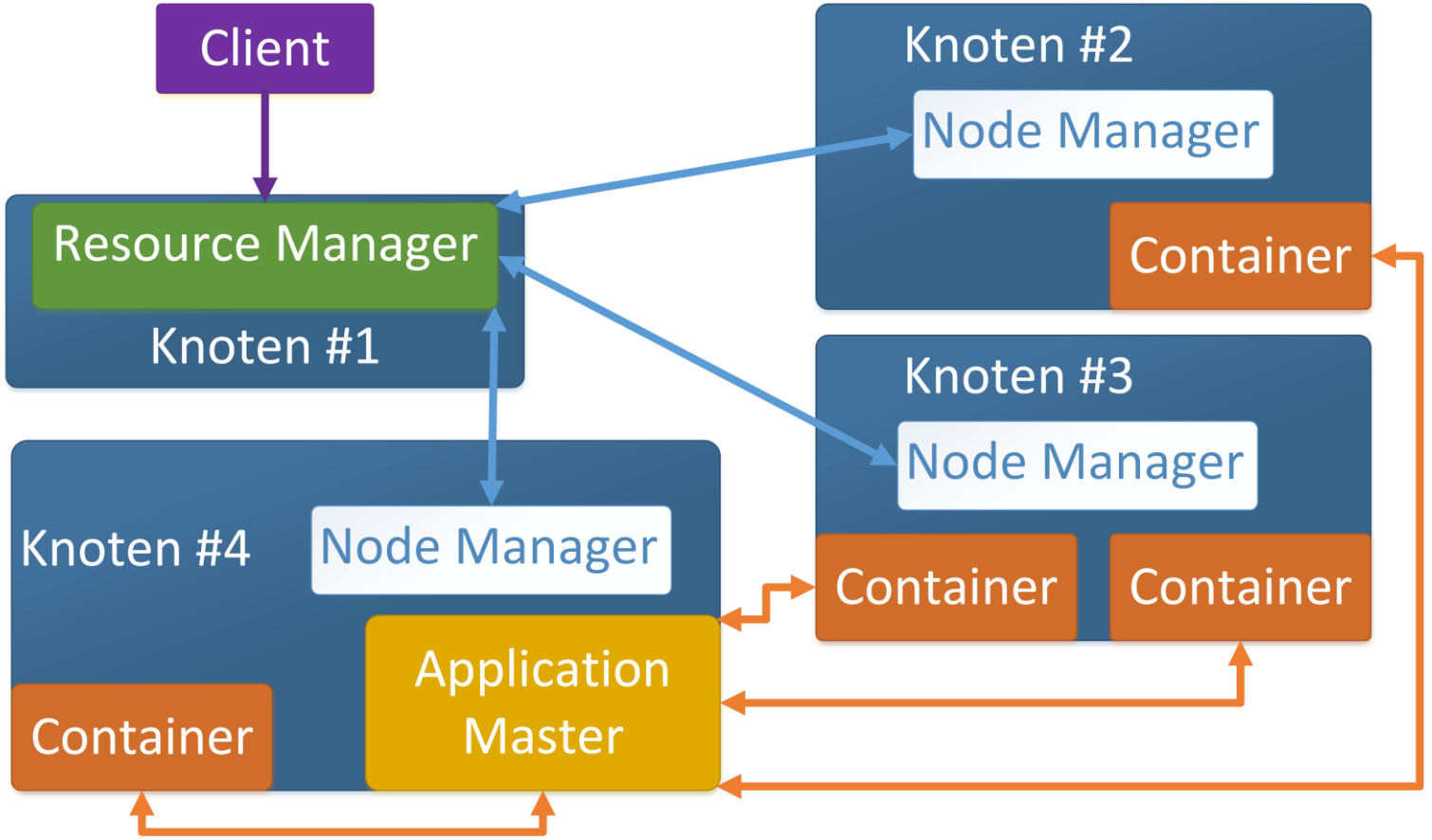
Entwicklungsumgebung



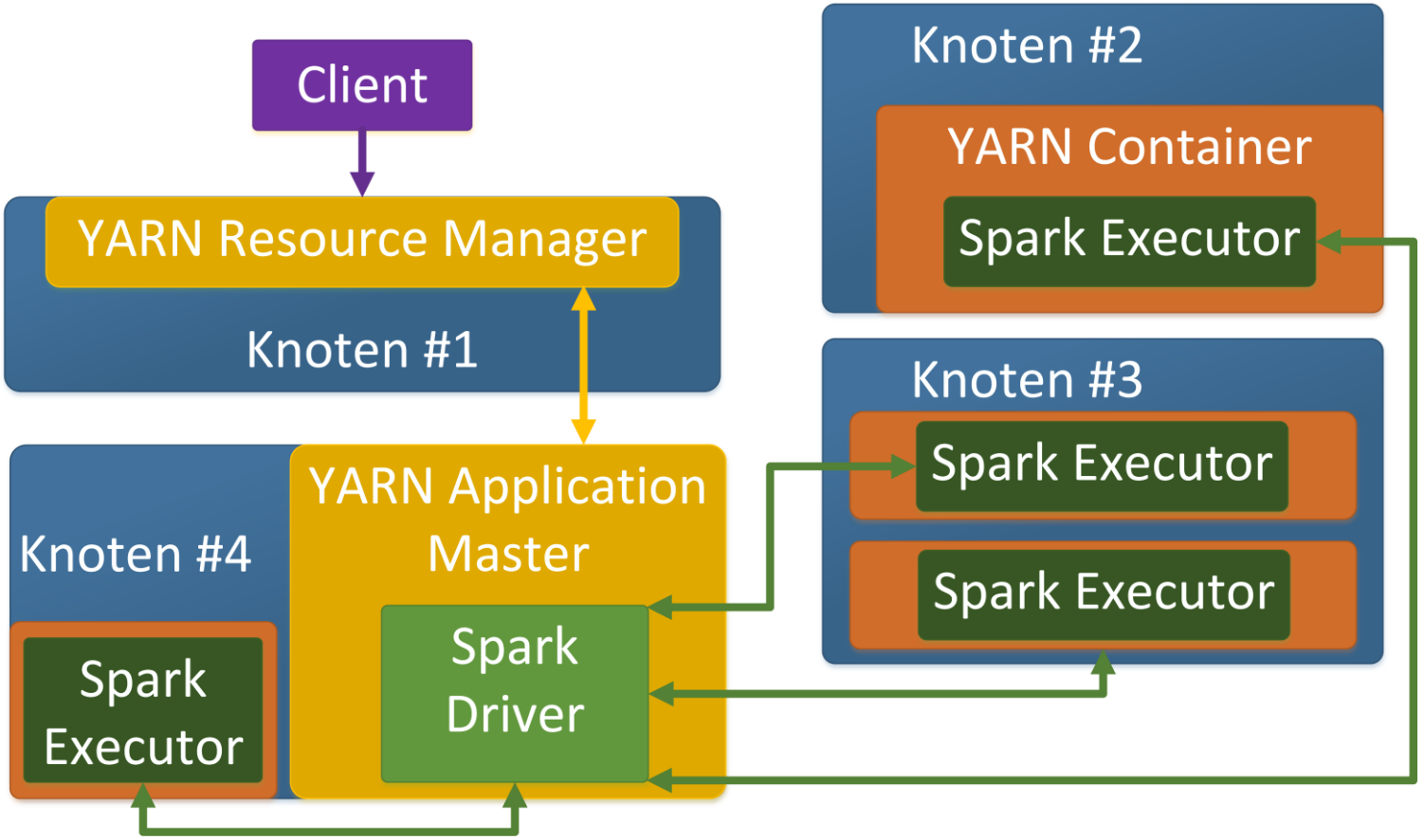
Apache Hadoop HDFS



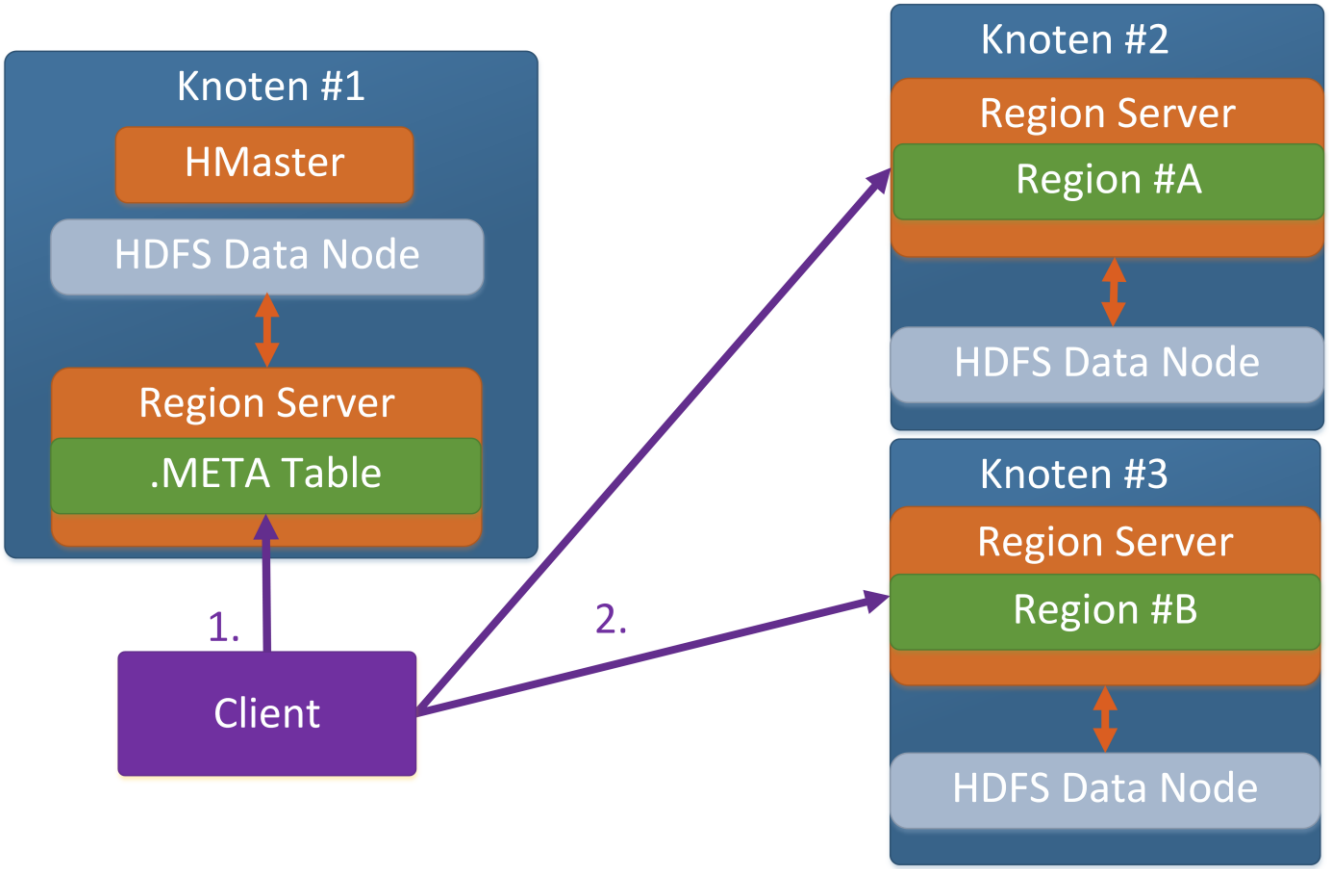
Apache Hadoop YARN



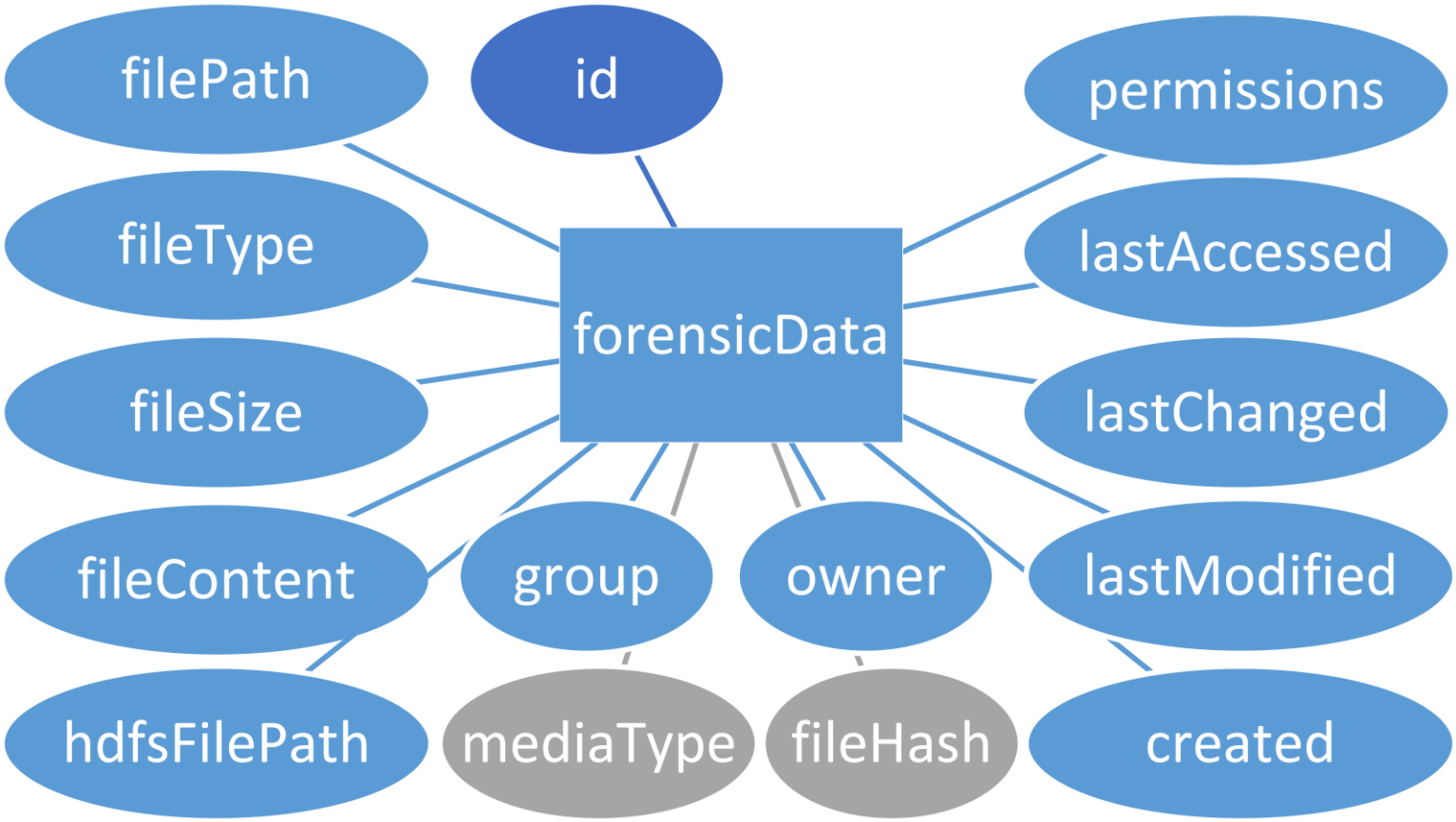
Apache Spark



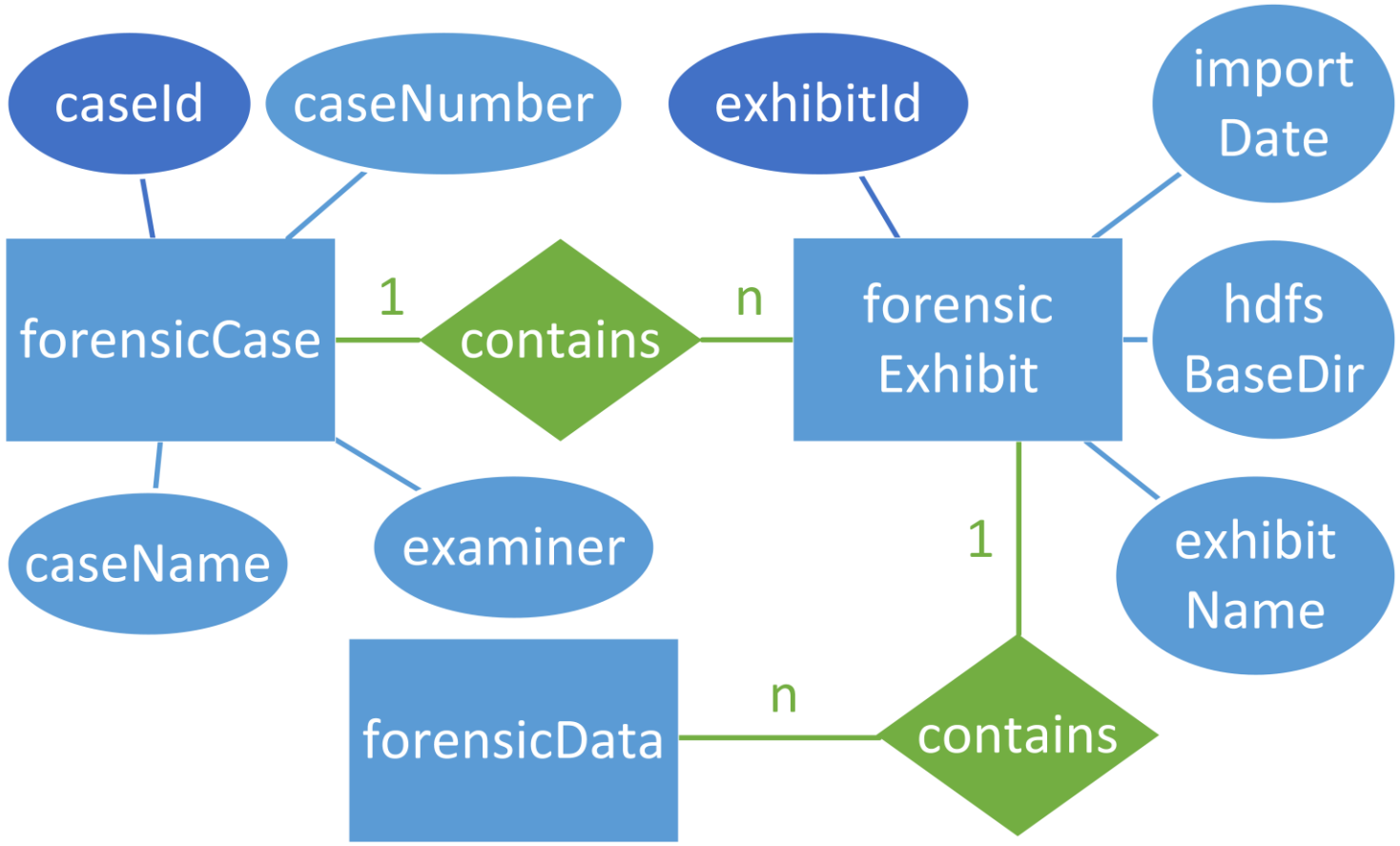
Apache HBase



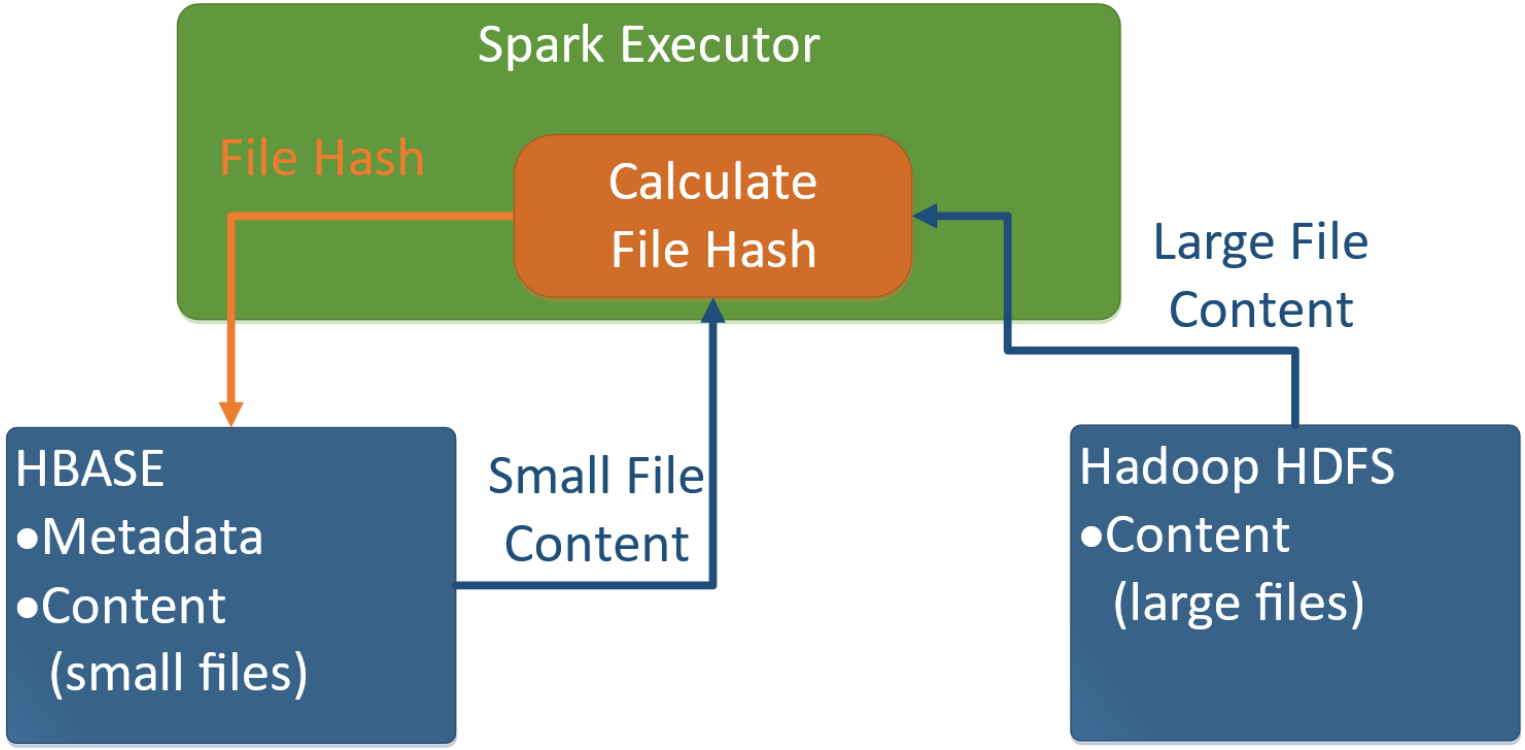
Datenmodell



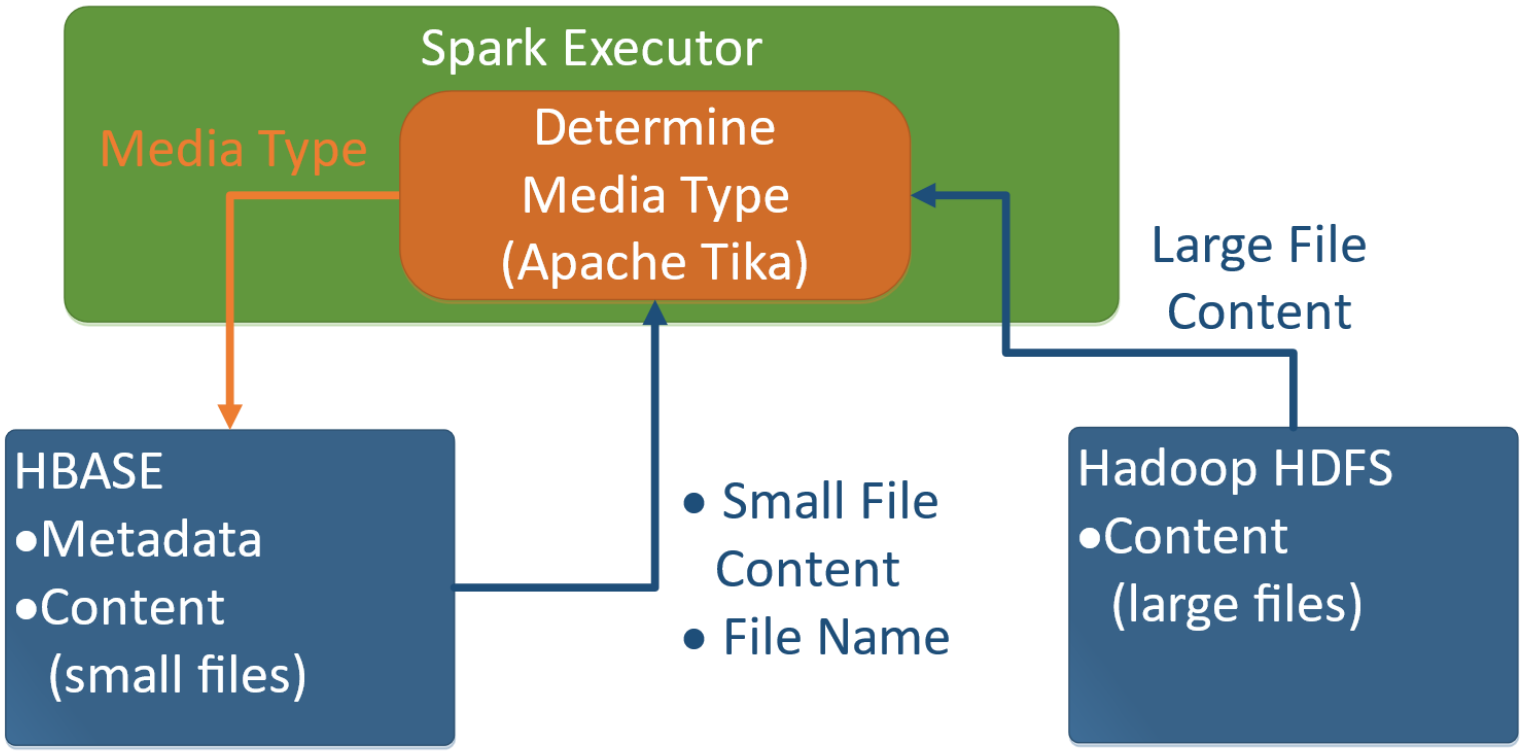
Datenmodell - Fallverwaltung



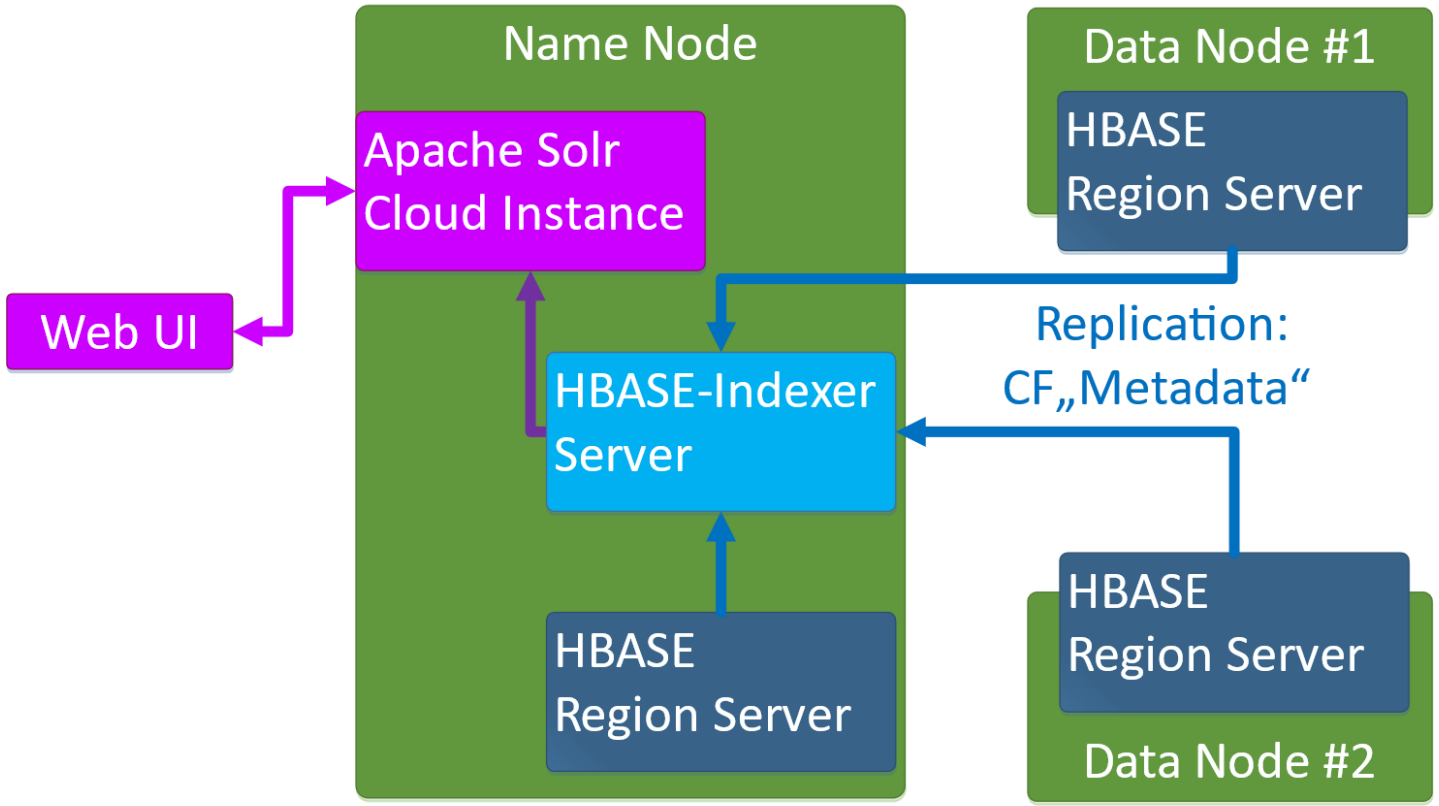
Hashsummen berechnen



Medientypen ermitteln



Indexierung im Testcluster



Indexierung – Technologien #1

- Elasticsearch
  - Sehr leichter Einstieg (Tutorials, Support)
  - Spark-Connector basierend auf RDDs
  - Kerberos-Absicherung kostenpflichtig
  - Skalierung über eigenen Mechanismus
  - Kein Integration in HDP / Ambari

Indexierung – Technologien #2

- Apache Solr
  - Kerberos-Absicherung kostenfrei
  - Solr-Cloud Skalierung via ZooKeeper
  - Integration in HDP / Ambari (Hadoop-Search)
  - HBASE-Indexer (HBASE -> Solr)
  - Solr Content Extraction Library (CELL)