

Apache NiFi A Dataflow Solution



Agenda

- 1. Einführung
- 2. Funktionsweise
- 3. Installation
- 4. Verwendung
- 5. Fazit
- 6. Übung





Einführung

Herausforderungen, Einleitung in NiFi und Historie

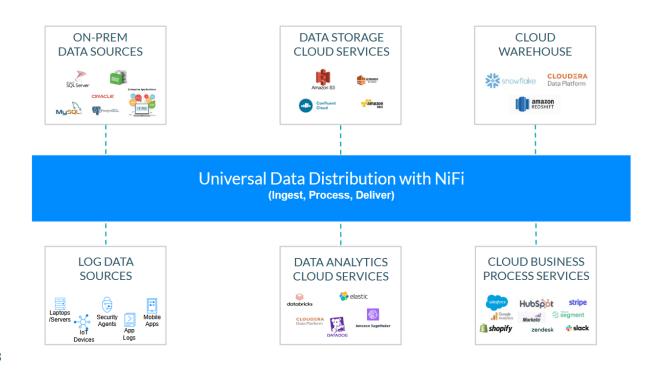
Herausforderungen

- Wachsende Datenmengen aus verschiedenen Quellen
- Unterschiedliche Datenformate (z. B. JSON, XML, CSV)
- Bedarf an Echtzeit- und Batch-Verarbeitung

Bisherige Lösungen

- Bspw. manuelle ETL-Prozesse oder Custom Middleware-Lösungen
- Komplex
- Schlecht skalierbar
- Hohe Kosten, initial und bei Wartung

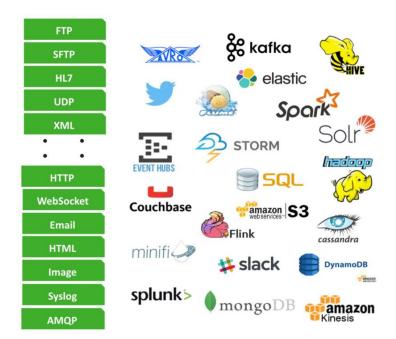
Apache NiFi

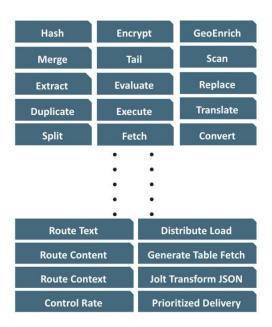


Apache NiFi

- Visuelle Dataflows
- Keine Programmierkenntnisse notwendig
- Schnelle Integration und Automatisierung
- Einfache Anpassung an neue Gegebenheiten

Apache NiFi





Unternehmen, die NiFi einsetzen





[5]



[6]



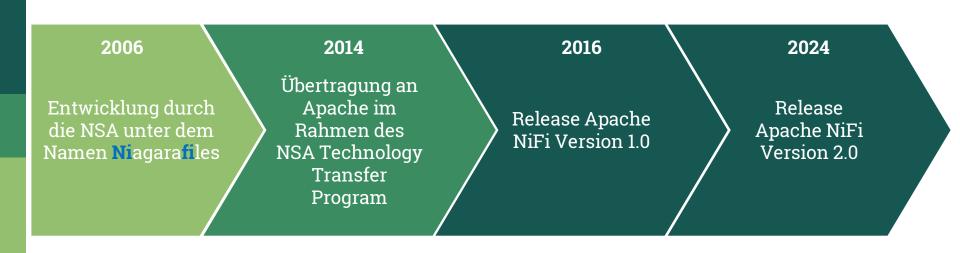
[3] https://www.telekom.sk/about [4] https://home.kuehne-nagel.com [5] https://www.audi-mediacenter.com/de/fotos/detail/porsche-logo-28410 [6] https://www.micron.com/about/press/image-gallery/micron-logos

Unternehmen: Porsche



- Aufbau einer zentralen Data Streaming-Plattform
- Findet Anwendung in verschiedenen Applikationen bei Porsche
- Nutzung von Apache Kafka und Apache NiFi

Historie



NiFi 2.0

- Java 21
- Kein Support von XML Dataflows
- Unterstützung von Python-Processors



Funktionsweise

Konzept, Architektur, Komponenten und Clustering

Designkonzept

- Starke Ähnlichkeiten zu den Kernideen vom "Flow Based Programming" (FBP)
- Fokus:
 - Verbindung von Prozessen (Blackboxen)
 - Datenfluss
 - Scheduling

Terminologie

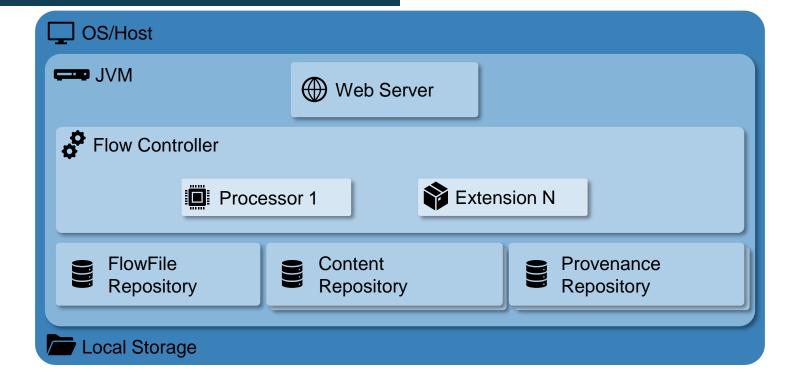
NiFi Term	FBP Term	Bedeutung
FlowFile	Information Packet	Repräsentieren Objekte im System. Beinhalten Meta- Informationen und assoziierte Nutzdaten.
FlowFile Processor	Black Box	Ausführung der eigentlichen Arbeit. Enthält tatsächliche Logik.
Connection	Bounded Buffer	Verbindung zwischen Processors. Agieren als eine Art Queue/Buffer.
Flow Controller	Scheduler	Beinhaltet den Aufbau der Processors und deren Verbindungen. Verantwortlich für die Verteilung.
Process Group	Subnet	Gruppierung von Processors und Connections, die Inputs und Outputs haben. Ermöglichen Erweiterung durch Komposition.

FlowFile

- Datenfluss-Objekt
- Speicherung in einer Hash-Map (In Memory)
 - Keine Persistenz
- Beinhaltet Pointer zum Content (Nutzdaten)
- Beinhaltet Meta-Informationen:
 - □ **UUID**: Unique Identifier
 - □ **FILENAME**: Dateiname
 - □ **PATH**: Relativer Pfad zum Content
 - □ **MIME_TYPE**: Dateityp
 - □ **PRIORITY**: Prioritätszahl
 - □ Und weitere...

```
{
    "uuid": "a1b2c3d4-5678-90ab-cdef-1234567890ab",
    "filename": "example.json",
    "path": "/relative/path/to/content/",
    "mime_type": "application/json",
    "priority": 5
}
```

Architektur







- Neuste Version läuft auf Angular 18
 - NiFi 1.X auf AngularJS
- HTTP-basierte Verwendung
- Bietet grafische Oberfläche
- Alternative zur NiFi CLI

Flow Controller



- "Gehirn" von NiFi
- Bietet Umgebung für Extensions an
- Verwaltet Ressourcenverteiltung (Scheduling)
- Beinhaltet die Flows

FlowFile Repository



- Speicherort der FlowFiles: Beinhaltet aktuell verwendete FlowFiles.
- Metainformationen:
 - Aktueller Status
 - Informationen zum Prozess
- Standardimplementierung: Persistenter Write-Ahead Log (WAL). Sicherstellung, dass Änderungen wiederherstellbar sind.

Content Repository



- Speicherort der Nutzdaten: Beinhaltet eigentliche Dateninhalte.
- FlowFiles Content:
 - Beinhaltet assoziierte Dateninhalte.
 - "Bits und Bytes"
- Standardimplementierung: Speicherung im Dateisystem unter einem oder mehreren physischen Pfaden, um Last zu reduzieren.

Provenance Repository

- Speicherort aller Ereignisdaten: Sämtliche Ereignisdaten, die den Lebenszyklus von FlowFiles betreffen.
- Implementiert eine Historie:
 - Detaillierte Nachverfolgbarkeit der Daten
 - Analysezwecke
- **Standardimplementierung**: Speicherung im Dateisystem, unter einem oder mehreren physischen Pfaden. Daten werden indexiert und sind somit effizienter durchsuchbar.

Processor



- Steuerung über den Flow Controller
- Ausführung der eigentlichen Aufgaben
- Viele unterschiedliche Processors vorhanden, beispielsweise:
 - **GetFile**: Erstellt FlowFiles aus Dateien im Dateisystem
 - PutFile: Schreibt Content aus FlowFiles in das Dateisystem

Exkurs: SPI Muster



- Service Provider Interface
- Basiert auf dem Konzept der "Replaceable component models"
- Beschreiben die Erweiterbarkeit einer Anwendung ohne Bearbeitung des Originalquellcodes
- Funktionale Schnittstellen-Spezifikation

E SPI



Komponenten:

- Service: Definiert die Funktion (Anforderungen).
- **SPI**: Ansammlung öffentlicher Interfaces und abstrakter Klassen, die die Funktion des Service definieren.
- Service-Provider: Implementierung eines SPIs

SPI Beispiel: CustomProcessor

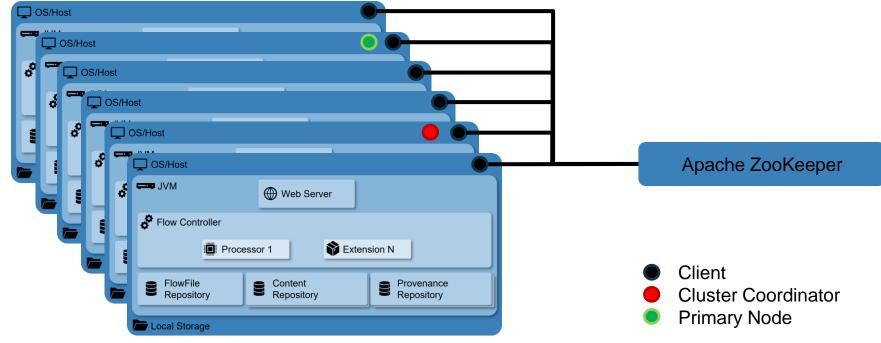


- 1. Klasse erweitern: AbstractProcessor
- 2. Benötigte Methoden Implementieren
 - onTrigger(ProcessContext, ProcessSession)
- 3. Als NAR (NiFi Archive) verpacken
- 4. Im NIFI_HOME/lib Verzeichnis ablegen
- **5.** NiFi neustarten

Alle NiFi Extensions müssen thread-safe sein.

NiFi - Clustering





[9] In Anlehnung an https://nifi.apache.org/docs/nifi-docs/html/images/zero-leader-cluster.png

Apache ZooKeeper



- Koordinationssystem f\u00fcr verteilte Anwendungen
- Zentraler Austauschpunkt
- Bestimmt Cluster Coordinator und Primary Node

Cluster Coordinator

- Übernimmt alle Managementaufgaben
- Bekommt "Heartbeats" von Nodes
- Zuständig für:
 - Trennung von Nodes
 - Verbindung neuer Nodes
 - Synchronisation von Flows an neue Nodes

Primary Node



- Genau einer pro Cluster
- Erlaubt die Ausführung von "Isolated Processors"
- Isolated Processors:
 - Werden nicht auf allen Nodes ausgeführt
 - Einmalige Ausführung
 - Beispiel: Initialisierung oder Kommunikation mit externen Systemen

Cluster



- Zero-Leader Clustering Paradigma:
 - Jede Node besitzt den gleichen Flow
 - Jede Node führt die gleiche Arbeit aus, aber auf unterschiedlichen Daten
 - Daten werden vom Cluster automatisch verteilt.
- Bei Ausfall wird Arbeit nicht automatisch neu verteilt
 - Aufgabe des Entwicklers!
- Lokale Veränderungen werden synchronisiert:
 - Flow wird über alle Nodes repliziert

Apache NiFi Ecosystem

- NiFi Toolkit
 - Ansammlung an Werkzeugen
- MiNiFi
 - Ressourcensparende Version
- NiFi Registry
 - Verwaltung und Versionierung von Flows
- NiFi Flow Design System
 - Ansammlung an UI-Elementen



Installation

Bereitstellung und Konfiguration

Systemvoraussetzungen

- Erfordert mindestens Java 8
- Unterstützung aller gängigen Betriebssysteme und Browser
- Benötigte Hardwareressourcen sind stark von verwendeten Dataflows abhängig

Installationsmöglichkeiten

Bare Metal



Docker



[11]

Installation (Bare Metal)

Download auf der Website

NiFi 1.28.1 Release Notes

Apache NiFi 1.28 is the last minor release of the version 1 series. The project management committee may consider critical bug fixes for essential framework features on an exceptional basis. Critical bug fixes do not include upgrading project dependencies. Multiple fundamental dependencies in NiFi 1 cannot be upgraded. These dependencies include Jetty 9.4, Spring Framework 5.3, and AngularJS 1.8. The Apache NiFi Team strongly encourages users to upgrade to NiFi 2.

End of Support: 2024-12-08

Source 1.28.1 OpenPGP SHA-512

NiFi Standard 1.28.1 OpenPGP SHA-512

Installation (Bare Metal)

Nutzung von /bin/nifi.cmd (Windows) bzw. /bin/nifi.sh (Linux) mit Parametern

- start
- stop
- status

Installation (Bare Metal)

 Standardnutzer wird beim ersten Start angelegt

```
Generated Username [9773ad05-c65a-4822-baa6-a1be34fb9406] Generated Password [2UH0gTbhrgrWNmmolmQInL2KMR2FvPxh]
```

Änderung mit Befehl möglich:
 /bin/nifi.sh set-single-user-credentials
 <username> <password>

Installation (Docker)

Offizielles Apache Docker-Image verfügbar



```
apache/nifi 

Sponsored OSS
```

By The Apache Software Foundation • Updated 11 days ago

Apache NiFi unofficial binary build

IMAGE

☆297 <u>↓</u>10M+

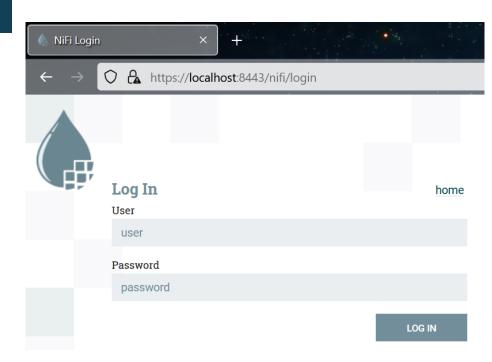
Installation (Docker)

- Docker Compose:
 - Vereinfachte Verwendung in Multi-Container Umgebungen
- Auszug aus der von uns erstellten docker-compose.yml

```
# Cluster
nifi:
  image: apache/nifi:1.19.1
  restart: always
  expose:
    - ${NIFI_HTTP_PORT}
    - ${PROCESSOR HTTP PORT}
  deploy:
    mode: replicated
    replicas: ${NIFI NODE COUNT}
    endpoint_mode: vip
    restart policy:
      condition: on-failure
      delay: 5s
      max_attempts: 3
      window: 120s
```

3 Installation

Öffnen der Benutzeroberfläche unter https://localhost: 8443/nifi



Konfiguration

Die Konfiguration erfolgt hauptsächlich in der Datei nifi.properties.

```
nifi.web.https.host=localhost
nifi.web.https.port=8443
nifi.web.https.network.interface.default=
nifi.web.https.application.protocols=h2 http/1.1
nifi.web.jetty.working.directory=./work/jetty
nifi.web.jetty.threads=200
nifi.web.max.header.size=16 KB
nifi.web.proxy.context.path=
nifi.web.proxy.host=
nifi.web.max.content.size=
nifi.web.max.requests.per.second=30000
nifi.web.max.access.token.requests.per.second=25
nifi.web.request.timeout=60 secs
nifi.web.request.ip.whitelist=
nifi.web.should.send.server.version=true
```



Verwendung

Nutzung von UI, CLI und API

Verwendungsarten



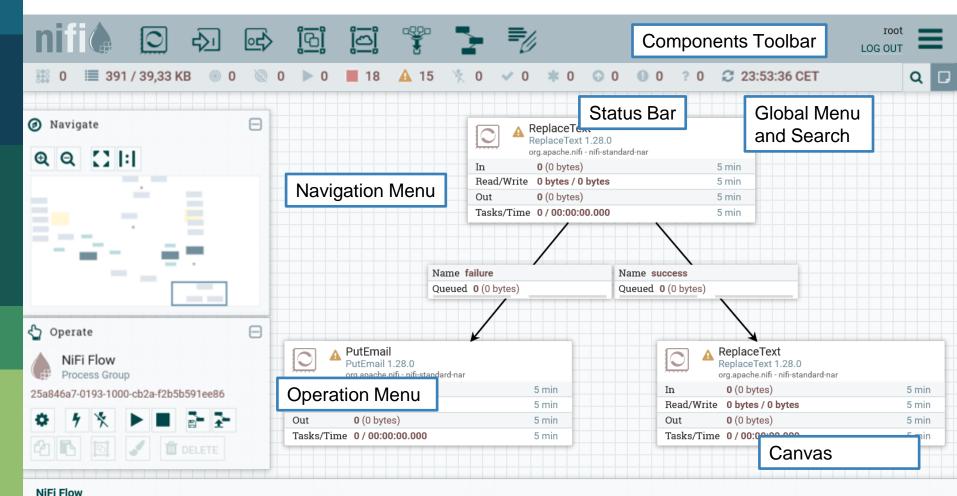


CLI



API





Processor



Displaying 364 of 364		Filter
Type ▲	Version	Tags
AttributeRollingWindow	1.28.0	rolling, data science, Attribute
AttributesToCSV	1.28.0	flowfile, csv, attributes
AttributesToJSON	1.28.0	flowfile, json, attributes
Base64EncodeContent	1.28.0	encode, base64
CalculateParquetOffsets	1.28.0	break apart, cluster, split, partit
CalculateParquetRowGrou	1.28.0	break apart, cluster, split, partit
CalculateRecordStats	1.28.0	stats, record, metrics
CaptureChangeMySQL	1.28.0	cdc, restricted, jdbc, mysql, ev
CompareFuzzyHash	1.28.0	fuzzy-hashing, hashing, cyber
CompressContent	1.28.0	compress, snappy framed, bro
ConnectWebSocket	1.28.0	subscribe, consume, listen, We 🗸

AttributeRollingWindow 1.28.0 org.apache.nifi - nifi-stateful-analysis-nar

Track a Rolling Window based on evaluating an Expression Language expression on each FlowFile and add that value to the processor's state. Each FlowFile will be emitted with the count of FlowFiles and total aggregate value of values processed in the current time window.



Processor



R	ReplaceText ReplaceText 1.28.0 org.apache.nifi - nifi-standard-nar		
In	0 (0 bytes)	5 min	
Read/Write	0 bytes / 0 bytes	5 min	
Out	0 (0 bytes)	5 min	
Tasks/Time	0 / 00:00:00.000	5 min	

Processor



SETTINGS	SCHEDULING	PROPERTIES	RELATIONSHIPS	COMMENTS	
----------	------------	------------	---------------	----------	--

Required field

Property		Value
Replacement Strategy	0	Regex Replace
Search Value	0	(?s)(^.*\$)
Replacement Value	0	\$1
Character Set	0	UTF-8
Maximum Buffer Size	0	1 MB
Evaluation Mode	0	Line-by-Line
Line-by-Line Evaluation Mode	0	All

NiFi Expression Language

- Zugriff auf FlowFile-Attribute
- Ähnlich zu template literals
- Vielzahl an Methoden, wie replace(), contains()
- Bsp.: \$\filename\}-\$\{now():\format('yyyy-MM-dd')\}
 => report-2024-12-11

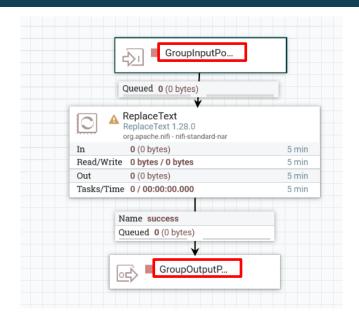
Process Group



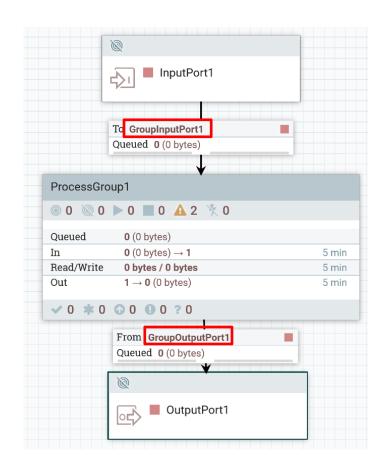
ProcessGrou	ıp1	
◎ 0 ◎ 0	0 0 0 0 0 0	
Queued	0 (0 bytes)	
In	0 (0 bytes) → 0	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 → 0 (0 bytes)	5 min
✓ 0 * 0	000070	

Process Group: Input / Output Port





Ansicht in Process Group



Globale Ansicht Flow

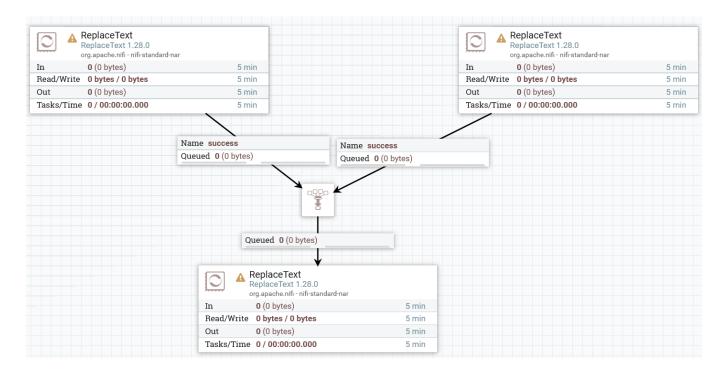


Add Remote Process Group

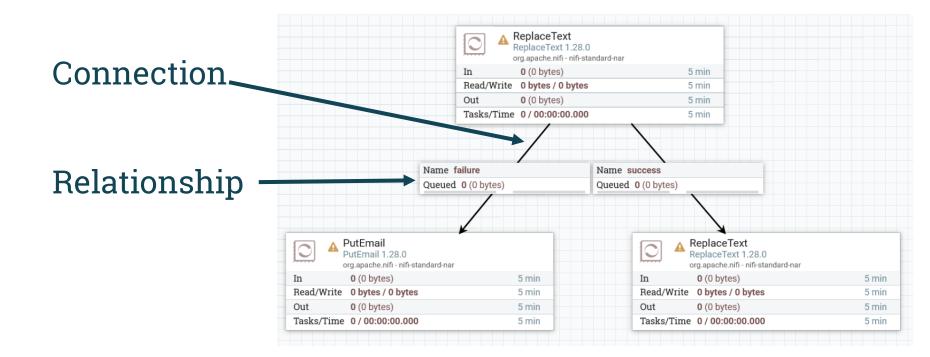
URLs ?	
https://remotehost:8443/nifi	
Transport Protocol 😯	Local Network Interface 😯
RAW 🗸	
HTTP Proxy Server Hostname 🔞	HTTP Proxy Server Port ②
HTTP Proxy User 🔞	HTTP Proxy Password ②
Communications Timeout ②	Yield Duration ②
30 sec	10 sec

Funnel





Connections und Relationships



4 CLI

- Verwaltung und Automatisierung
- Überwachung der Flows
- Eingeschränkte Funktionalität, ermöglicht z. B. keine Bearbeitung von Flows

4 API

- REST-API
- Vielfältige Interaktionsmöglichkeiten, alle Funktionalitäten der UI vorhanden
- Integration in beliebige externe Systeme



Fazit

5 Fazit

- Skalierbarkeit
- Individuell anpassbar
- Unterstützung großer Datenmengen

Aber:

- Einarbeitung in NiFi notwendig
- NiFi bringt eigene Komplexität mit

Quiz

Welche Organisation hat NiFi ursprünglich entwickelt??

1. CIA

2. NSA

3. NASA

4. FTC

Welche Organisation hat NiFi ursprünglich entwickelt??

1. CIA 2. NSA

3. NASA 4. FTC

Die NiFi Expression Language berechnet statistische Analysen von Daten.

Korrekt Inkorrekt

Die NiFi Expression Language berechnet statistische Analysen von Daten.

Korrekt

Inkorrekt

Welche Programmiersprache muss ein NiFi Anwender beherrschen?

1. Java 2. C

3. Keine 4. JavaScript

Welche Programmiersprache muss ein NiFi Anwender beherrschen?

1. Java 2. C

3. Keine 4. JavaScript

Welche Betriebssysteme werden von NiFi unterstützt?

1. Windows

2. Linux

3. MacOS

4. Alle genannten Betriebssysteme

Welche Betriebssysteme werden von NiFi unterstützt?

1. Windows

2. Linux

3. MacOS

4. Alle genannten Betriebssysteme

In welcher Programmiersprache ist NiFi geschrieben?

1. Java

2. C

3. C++

4. JavaScript

In welcher Programmiersprache ist NiFi geschrieben?

1. Java

2. C

3. C++

4. JavaScript

NiFi 2.0 unterstützt das Programmieren eines eigenen Processors in Python.

Korrekt

Inkorrekt

NiFi 2.0 unterstützt das Programmieren eines eigenen Processors in Python.

Korrekt

Auf welchem Kernkonzept baut NiFi auf?

1. Flow Based Programming

2. Dynamic Programming

3. Procedural Programming

4. Functional Programming

Auf welchem Kern-Konzept baut NiFi auf?

1. Flow Based Programming

2. Dynamic Programming

3. Procedural Programming

4. Functional Programming

Was ist der Fokus von Flow Based Programming?

 Erstellung von Benutzeroberflächen 2. Verwaltung von Datenbanken

3. Verbindung von Processors

4. Verwaltung von Netzwerkprotokollen

Was ist der Fokus von Flow Based Programming?

1. Erstellung von Benutzeroberflächen 2. Verwaltung von Datenbanken

3. Verbindung von Processors

4. Verwaltung von Netzwerkprotokollen

Was sind **keine** Komponenten der NiFi Architektur?

1. Processors

2. Netzwerkkarte

3. Flow Controller

4. Datenbankserver

Was sind **keine** Komponenten der NiFi Architektur?

1. Processors

2. Netzwerkkarte

3. Flow Controller

4. Datenbankserver

FlowFiles speichern die eigentlichen Nutzdaten direkt innerhalb ab!

Korrekt Inkorrekt

FlowFiles speichern die eigentlichen Nutzdaten direkt innerhalb ab!

Korrekt

Was ermöglicht den Zugriff auf die Admin Funktionen von NiFi?

1. UI

2. CLI

3. FTP

4. API

Was ermöglicht den Zugriff auf die Admin Funktionen von NiFi?

1. UI

2. CLI

3. FTP

4. API

Wie werden FlowFiles intern in NiFi abgelegt?

1. Linked-List

2. Array

3. Binary-Tree

4. Hash-Map

Wie werden FlowFiles intern in NiFi abgelegt?

1. Linked-List

2. Array

3. Binary-Tree

4. Hash-Map

Wie kann der Flow Controller auch genannt werden (FBP)?

1. Bounded Buffer

2. Scheduler

3. Subnet

4. Black Box

Wie kann der Flow Controller auch genannt werden (FBP)?

1. Bounded Buffer

2. Scheduler

3. Subnet

4. Black Box

Die Processors sind für die Verteilung der FlowFiles verantwortlich!

Korrekt

Die Processors sind für die Verteilung der FlowFiles verantwortlich!

Korrekt

FlowFiles können nicht ohne FlowFile Repository persistent gespeichert werden!

Korrekt

FlowFiles können nicht ohne FlowFile Repository persistent gespeichert werden!

Korrekt

Was bedeutet SPI im Zusammenhang mit NiFi?

1. Service Provider Interface

2. Simple Programming

Interface

3. Secure Process Integration

4. Single Page Implementation

Was bedeutet SPI im Zusammenhang von NiFi?

1. Service Provider Interface

2. Simple Programming Interface

3. Secure Process Integration

4. Single Page Implementation

Worüber wird ein Processor implementiert?

1. Einen Service

2. Einen Service-Provider

3. Ein Service Provider Interfaceation

4. Es können keine Processors implementiert werden

Worüber wird ein Processor implementiert?

1. Einen Service

2. Einen Service-Provider

3. Ein Service Provider Interfaceation

4. Es können keine Processors implementiert werden

In welchem Format werden Extensions gespeichert?

1. WAR 2. ZIP

3. NAR 4. TAR

In welchem Format werden Extensions gespeichert?

1. WAR 2. ZIP

3. NAR 4. TAR

Was passiert bei Ausfall des Cluster Coordinators?

1. Nichts

2. Neuer wird bestimmt

3. Cluster läuft in Deadlock

4. Cluster stürzt ab

Was passiert bei Ausfall des Cluster Coordinators?

1. Nichts

2. Neuer wird bestimmt

3. Cluster läuft in Deadlock

4. Cluster stürzt ab

Wie wird die Node genannt, die "Isolated Processors" ermöglicht?

1. Main Node

2. Secondary Node

3. Primary Node

4. Cluster Coordinator

Wie wird die Node genannt die "Isolated Processors" ermöglicht?

1. Main Node

2. Secondary Node

3. Primary Node

4. Cluster Coordinator



Bildquellen

- [1] https://nifi.apache.org
- [2] https://f.io/I9DIDsP3
- [3] https://www.telekom.sk/about
- [4] https://home.kuehne-nagel.com
- [5] https://www.audi-mediacenter.com/de/fotos/detail/porsche-logo-28410
- [6] https://www.micron.com/about/press/image-gallery/micron-logos
- [7] https://nifi.apache.org/docs/nifi-docs/html/images/zero-leader-node.png
- [8] https://drive.google.com/drive/folders/ 1IgcAwLDVZUz8ycnFa7T4_H6B4V4LhYUQ://f.io/I9DIDsP3
- [9] https://nifi.apache.org/docs/nifi-docs/html/images/zero-leader-cluster.png
- [10] https://www.flaticon.com/de/kostenloses-icon/pc_4703650
- [11] https://www.docker.com/company/newsroom/media-resources

Alle Icons stammen von Microsoft PowerPoint.

Quellen

- **NiFi Dokumentation:** https://nifi.apache.org/documentation/v1
- **ZooKeeper:** https://zookeeper.apache.org
- **Porsche:** https://www.automotiveit.eu/technology/wie-porsche-dem-datenberg-herr-wird-280.html https://www.porsche-digital.hr/projects/streamzilla
- **NiFi Unternehmen:** https://nifi.apache.org/community/powered-by/
- Replaceable Components: Robert C Seacord. "Replaceable components and the service provider interface". In: COTS-Based Software Systems: First International Conference, ICCBSS 2002 Orlando, FL, USA, February 4–6, 2002 Proceedings. Springer. 2002, S. 222–233. ISBN: 978-3-540-45588-2. DOI: 10.1007/3-540-45588-4 21
- SPI: https://docs.oracle.com/javase/tutorial/ext/basics/spi.html
- **Template:** https://www.slidescarnival.com