

Willkommen beim #GAB 2015!



Spark  on Azure
mit HDInsight & Script Actions

Hans-Peter Grahl
Netconomy | Entwickler & Berater | FH CAMPUS 02

Twitter: @hpgrahl 

Lokale Sponsoren:



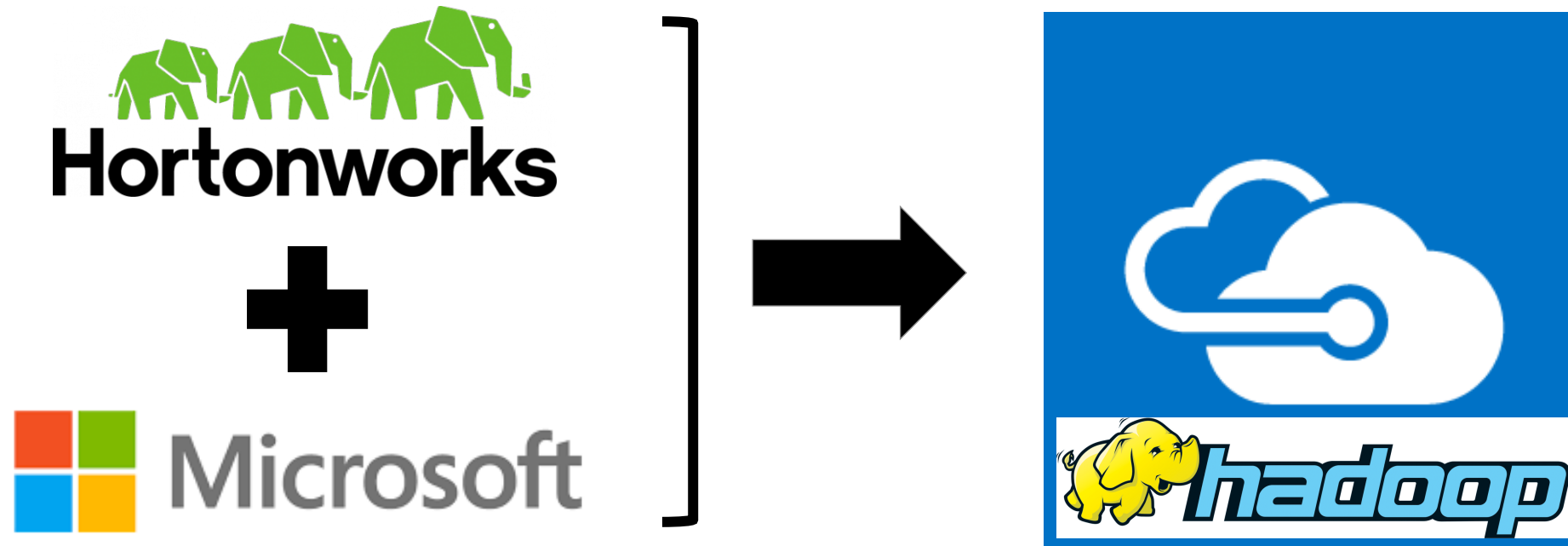
Überblick Inhalte

- Was ist HDInsight?
- Wozu überhaupt Spark?
- Customizing mit Script Actions
- Demo: Spark on Azure

HDInsight

▪ Was ist HDInsight?

- 100% auf Apache Hadoop basierendes **Azure Cloud Service**
- entstanden aus enger Zusammenarbeit von **Microsoft & Hortonworks**
- **Hortonworks HDP** ist die **on-premise** Version für **Windows Server** Umgebungen



HDInsight

■ Was ist HDInsight?

- beinhaltet viele **Komponenten des Hadoop Ökosystems**

Pig, Hive, Sqoop, Oozie, Mahout, ...

- ergänzende HDInsight Services: **HBase od. Storm**

- „Versionsdschungel“: **HDInsight Version => HDP Version => Hadoop Version**

COMPONENT	HDINSIGHT VERSION 3.2	HDINSIGHT VERSION 3.1 (DEFAULT)	HDINSIGHT VERSION 3.0	HDINSIGHT VERSION 2.1
Hortonworks Data Platform	2.2	2.1.7	2.0	1.3
Apache Hadoop & YARN	2.6.0	2.4.0	2.2.0	1.2.0

<http://azure.microsoft.com/en-us/documentation/articles/hdinsight-component-versioning/>

Hadoop auf einer Folie...

▪ Was ist Hadoop?

- verteiltes System zur Speicherung & Analyse von Daten
- typischerweise große unstrukturierte Datenmengen

⇒ 2 Hauptkomponenten

- **HDFS:** redundante verteilte Datenspeicherung
Hadoop Distributed File System
 - **MapReduce:** fehlertolerantes skalierbares Programmier-Paradigma
inkl. Ressourcen Verwaltung und Job Scheduling
-
- **Datenlokalität:** Berechnungen laufen auf jenen Knoten im Cluster
wo Daten gespeichert sind (bzw. in maximaler Nähe dazu)



Wozu überhaupt Spark?



- **Hadoop MapReduce** gilt seit Jahren als **de-facto Standard... ABER**

- 1. keine high-level Abstraktion** hinsichtlich fehlertoleranter & verteilter **in-memory Datenstrukturen**

- *sämtliche Datenverarbeitung mittels MapReduce ist mühsam*
- *Wiederverwendung von Daten nur mittels temp. Persistenz*

- 2. im Kern primär Batch-Verarbeitung ruhender Daten**

- *iterative Analyseverfahren?*
- *Data Mining & Machine Learning?*
- *interaktive Auswertungen und Stream Verarbeitung?*

Verallgemeinerung mittels Spark

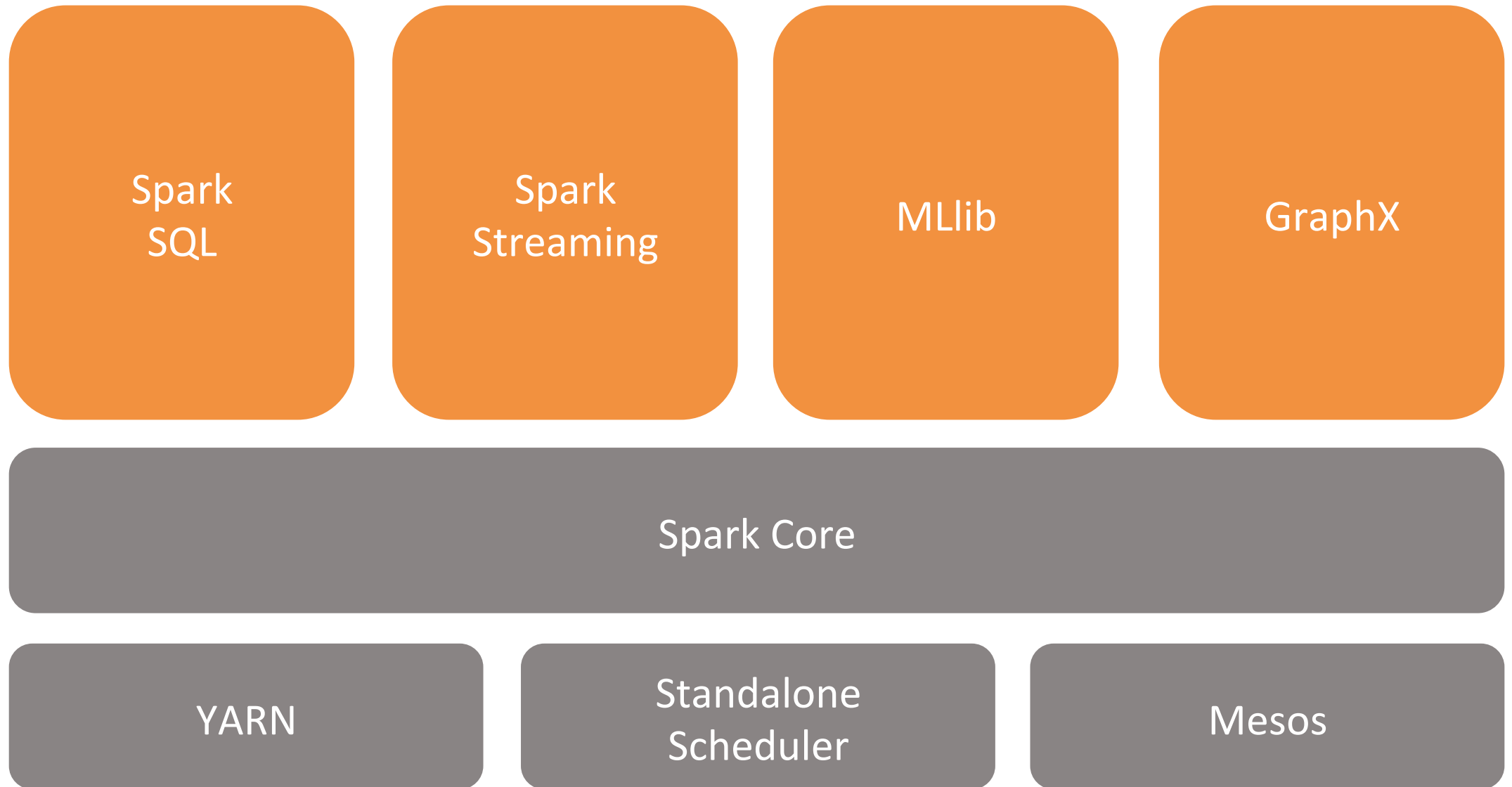


- beide Aspekte werden von Spark adressiert

Apache Spark™ is a fast and general engine for large-scale data processing.

- **verteilte & fehlertolerante in-memory Datenstrukturen**
- **generische Abstraktionen für diverse Anwendungsszenarien**
- Implementierungssprache: Scala (Language Bindings Java & Python)

Spark Stack



Spark on Azure ?

HDInsight
+
Script Actions

=

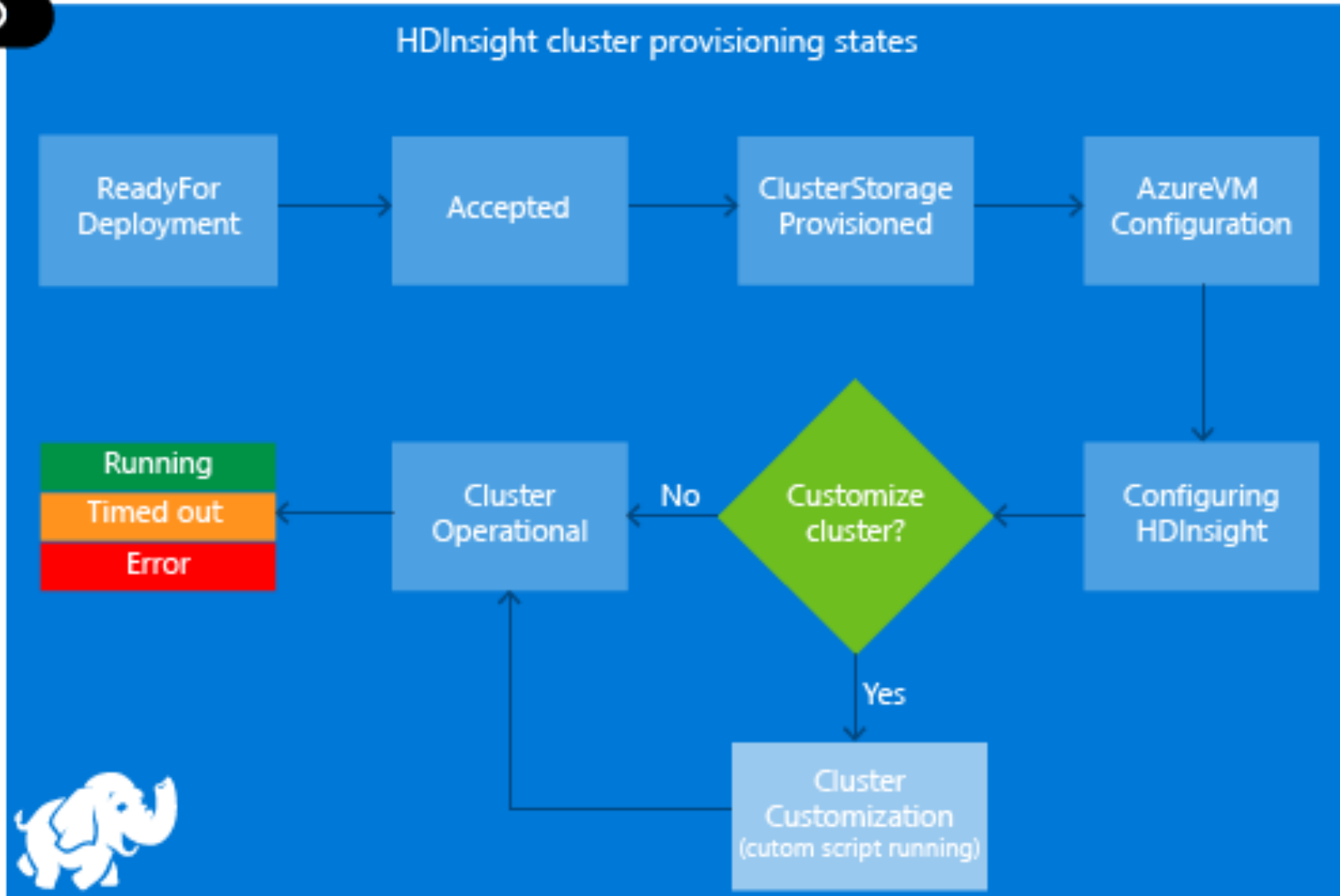


Script Actions ?

- **individuelle Anpassung** von HDInsight Clustern
 - anwendbar auf Head / Worker / alle Nodes

- **2 Hauptanwendungsfälle:**
 - weitere Software Pakete & Frameworks installieren
 - Konfiguration bestehender Komponenten ändern

Script Actions



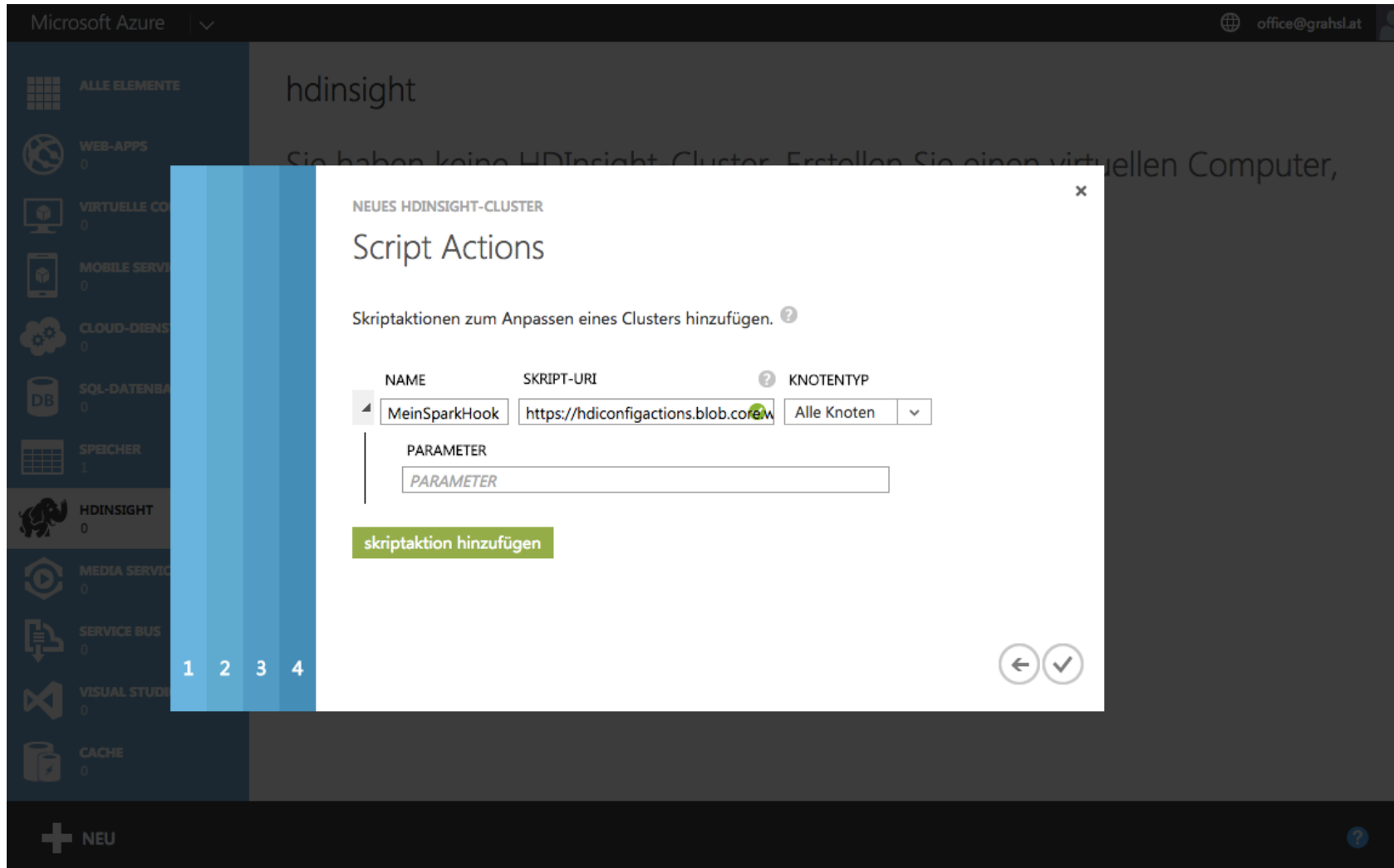
Script Actions

- Script Actions als „*Post-Installation-Hooks*“
 - nach Standard HDInsight Konfiguration der Nodes
 - laufen mit Admin(!) Privilegien
 - Reihenfolge für mehrere Script Actions definierbar

- **3 Roll-Out Möglichkeiten** für Script Actions
 - Konfigurations-Wizard im Azure Management Portal
 - Azure PowerShell cmdlets => Add-AzureHDInsight**ScriptAction**
 - HDInsight .NET SDK

Script Actions

- via Azure Portal & Custom Config Wizard



The screenshot shows the 'NEUES HDINSIGHT-CLUSTER' wizard in the Azure Portal. The 'Script Actions' step is active, allowing users to add custom scripts to the cluster. The interface includes a table for defining script actions and a 'skriptaktion hinzufügen' button.

NAME	SKRIPT-URI	KNOTENTYP
MeinSparkHook	https://hdiconfigactions.blob.core.w	Alle Knoten

PARAMETER

PARAMETER

skriptaktion hinzufügen

Navigation: 1 2 3 4

Script Actions

- via Azure Powershell

```
$config = Add-AzureHDInsightScriptAction -Config $config -Name "MeinSparkHook"  
-ClusterRoleCollection HeadNode -Uri <URL_TO_PS1_SCRIPT>
```

- via HDInsight .NET SDK

```
clusterInfo.ConfigActions.Add(new ScriptAction(  
    "MeinSparkHook",  
    new ClusterNodeType[] { ClusterNodeType.HeadNode},  
    new Uri(<URL_TO_PS1_SCRIPT>),  
    null //keine Parameter erforderlich  
));
```

Script Actions Beispiele

- Beispiele für PowerShell Script Actions



Script Actions Helper

- viele Hilfsmethoden zur Erstellung eigener Skripts vorhanden z.B.
 - Download von Dateien
 - Archive entpacken
 - Hadoop Version feststellen
 - laufende Dienste inspizieren
 - wichtige XML Konfigurationsdateien anpassen
 - etc.

<http://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-script-actions/#helpersmethods>

Script Actions Beispiele

- Vorgefertigte Script Actions von Microsoft als Basis für eigene

Demo: Wie sieht Script Action z.B. für Spark Customization aus?

<https://hdiconfigactions.blob.core.windows.net/sparkconfigurationv03/spark-installer-v03.ps1>

```
12 param (  
13     # The binary location for Spark in zip format.  
14     [Parameter()]  
15     [String]$SparkBinaryZipLocation,  
16  
17     # The name of the folder for Spark root.  
18     [Parameter()]  
19     [String]$SparkRootName)  
20  
21  
22 # Download config action module from a well-known directory.  
23 $CONFIGACTIONURI = "https://hdiconfigactions.blob.core.windows.net/configactionmodulev03/HDInsightUtilities-v03.psm1";  
24 $CONFIGACTIONMODULE = "C:\HDInsightUtilities.psm1";  
25 $webclient = New-Object System.Net.WebClient;  
26 $webclient.DownloadFile($CONFIGACTIONURI, $CONFIGACTIONMODULE);
```

Script Actions Best Practices

■ HDInsight bzw. Hadoop **Version prüfen**

- Unterstützung für Anpassungen erst ab *HDI 3.1 == Hadoop 2.4*
- man benötigt z.T. versch. Versionen der zu installierenden Komponenten

■ Script & Ressourcen **Bereitstellung über permanente Links**

- wichtig z.B. für re-imaging von Node
- am besten über Azure Storage Account verlinken

■ **geeigneter Installationsort** für Komponenten

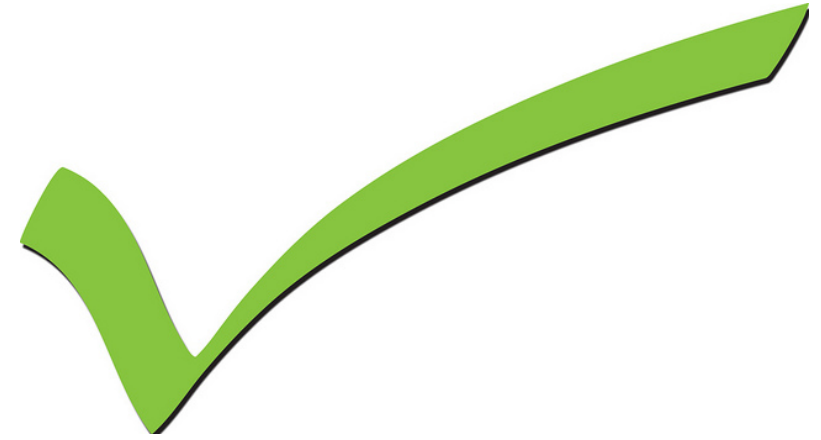
- typischerweise unter C:\apps (\dist) oder D:\

Script Actions Best Practices

- Einstellung des **Hochverfügbarkeitsmodus berücksichtigen**
 - per default kein auto-failover für nachinstallierte Komponenten
- Scripts sollten **idempotent** sein
 - relevant bei mehrmaliger Ausführung z.B. bei re-imaging von Node
- **Azure Blob Storage** Konfiguration
 - HDInsight Cluster kann von Haus aus HDFS + WASB
 - Ökosystem Komponenten per default auf HDFS ausgerichtet
 - ⇒ z.B. muss Spark explizit für WASB konfiguriert werden

Script Actions Testläufe

- Testläufe mittels **HDInsight Emulator**



- **Variante 1: auf lokaler Instanz**

=> Installation je nach Windows Version leider nicht immer reibungslos

- **Variante 2: auf Azure VM**

=> am besten mit Windows Server 2012 R2 Image

Script Actions Troubleshooting

- Fehlersuche bei Problemen



- **Logs in Azure Table Storage**

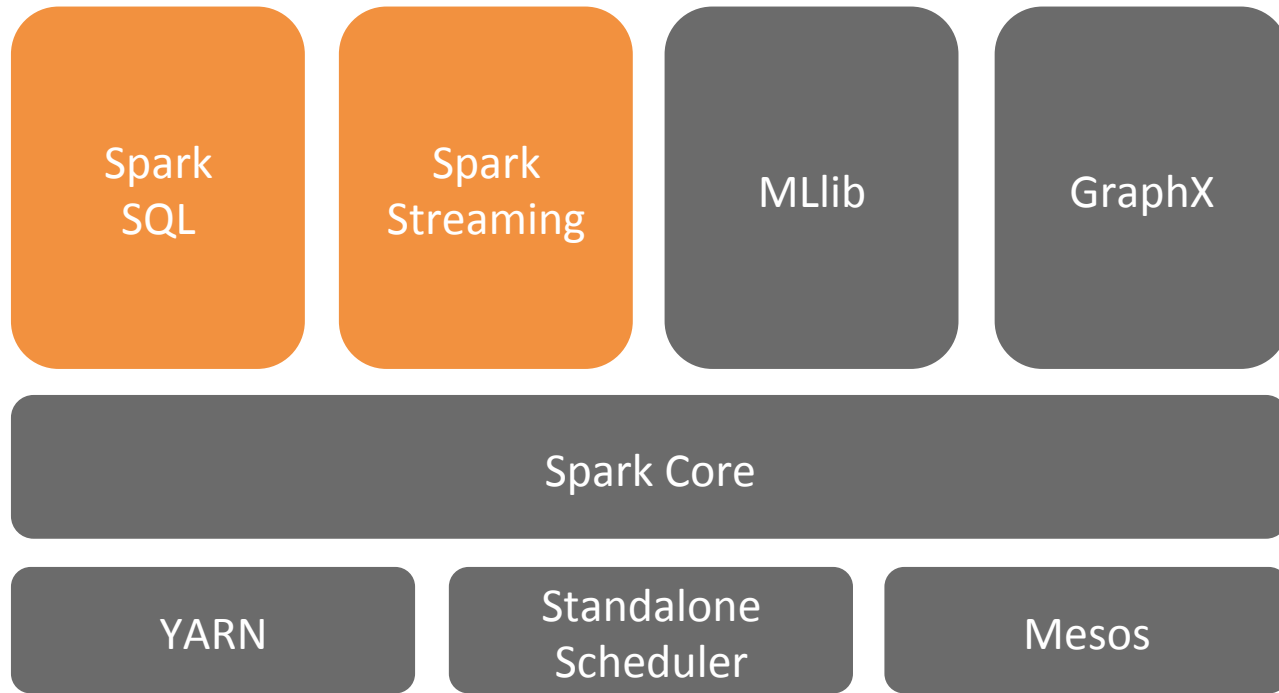
=> output und error logs im Storage Account des Clusters aller Nodes

- **lokale Log-Files auf einzelnen Cluster Nodes**

=> C:\HDInsightLogs\DeploymentAgent.log

Running Spark Applications on Azure

- Demo Time – Spark SQL und Spark Streaming Showcase



- **Spark WebUI** und/oder **YARN WebUI** am Head Node

z.B. <http://headnode0.hpghdi15.f8.internal.cloudapp.net:4040>

Aktuelle Herausforderungen

▪ TOP 3 Herausforderungen: meine persönliche Liste

1. derzeit keine Möglichkeit für Remote Job Submission

- nur am Head Node (RDP) in CmdPrompt mittels *spark-submit* und einem lokalen JAR File

2. aktuell nur Zulu OpenJDK 1.7 d.h. kein Java 8 Support ☹

- Spark ohne Lambdas & Co macht einfach wenig(er) Spaß

3. für neuesten Spark Versionen (1.3+) etwas Handarbeit nötig

- derzeit bis Spark 1.2.1 inkl. Script Action & WASB Config alles vorbereitet



Zusammenfassung



▪ Spark on Azure:

- sinnvolle **Ergänzung zu HDInsight** Standard Komponenten
- bietet **high-level APIs** und verteilte sowie fehlertolerante **in-memory Datenstrukturen**
- unterstützt beliebige Kombinationen aus **SQL, Graph & Stream Verarbeitung** sowie **Machine Learning** innerhalb einer **Anwendung**
- ist **durch vorgefertigte Script Actions** auf Knopfdruck verfügbar

Kontakt

Hans-Peter Grahsl

hanspeter@grahsl.at

+43 650 217 17 04



@hpgrahsl



https://www.xing.com/profile/HansPeter_Grahsl



hans_peter_g