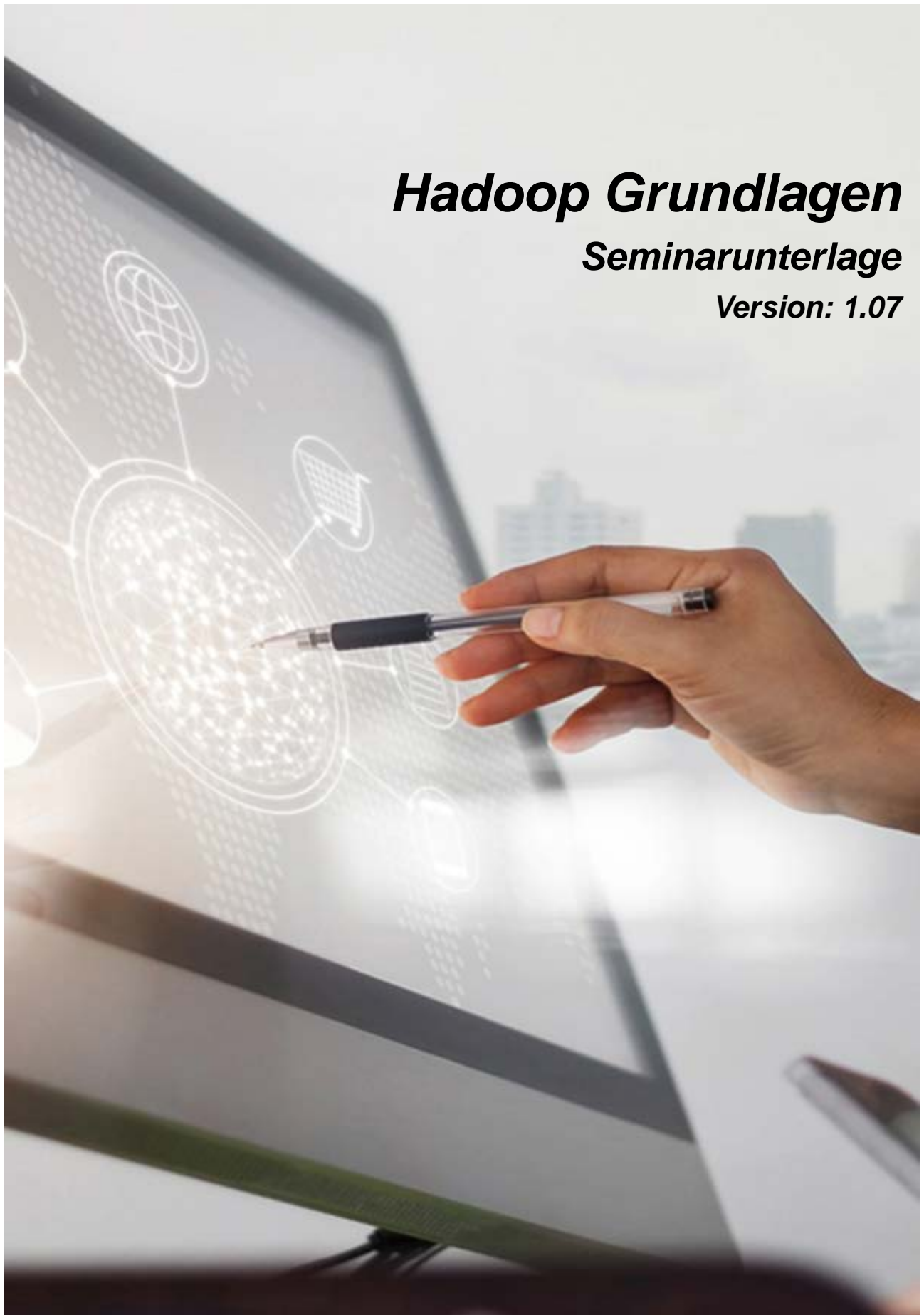


# ***Hadoop Grundlagen***

***Seminarunterlage***

***Version: 1.07***



Dieses Dokument wird durch die ORDIX AG veröffentlicht.

Copyright ORDIX AG. Alle Rechte vorbehalten.

Alle Produkt- und Dienstleistungs-Bezeichnungen sind Warenzeichen oder eingetragene Warenzeichen der jeweiligen Firmen und beziehen sich auf Eintragungen in den USA oder USA-Warenzeichen.

Weitere Logos und Produkt- oder Handelsnamen sind eingetragene Warenzeichen oder Warenzeichen der jeweiligen Unternehmen.

Kein Teil dieser Dokumentation darf ohne vorherige schriftliche Genehmigung der ORDIX AG weitergegeben oder benutzt werden.

### **Adressen der ORDIX AG**

Die ORDIX AG besitzt folgende Geschäftsstellen

ORDIX AG  
Karl-Schurz-Straße 19a  
D-33100 Paderborn  
Tel.: (+49) 0 52 51 / 10 63 - 0  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
An der alten Ziegelei 5  
D-48157 Münster  
Tel.: (+49) 02 51 / 9 24 35 – 00  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Welser Straße 9  
D-86368 Gersthofen  
Tel.: (+49) 08 21 / 507 492 – 0  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Kreuzberger Ring 13  
D-65205 Wiesbaden  
Tel.: (+49) 06 11 / 7 78 40 – 00  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Wikingerstraße 18-20  
D-51107 Köln  
Tel.: (+49) 02 21 / 8 70 61 – 0  
Fax.: (+49) 01 80 / 1 67 34 90

ORDIX AG  
Gewerbegebiet Süd-West Park  
Südwestpark 67/2  
D-890449 Nürnberg  
Tel.: (+49) 0 52 51 / 10 63 - 0  
Fax.: (+49) 01 80 / 1 67 34 90

Internet: <http://www.ordix.de>

Email: [seminare@ordix.de](mailto:seminare@ordix.de)

## Inhaltsverzeichnis

<b>1</b>	<b>Agenda .....</b>	<b>8</b>
1.1	Agenda.....	9
<b>2</b>	<b>Hadoop Überblick.....</b>	<b>10</b>
2.1	Agenda.....	11
2.2	Hadoop.....	12
2.3	1st ORDIX Hadoop Cluster (2015) .....	13
2.4	Yahoo's Hadoop Cluster (2007).....	14
2.5	Hadoop Kernkomponenten .....	15
2.6	Hadoop Pseudo Distributed Cluster.....	16
2.7	HDFS - Hadoop Distributed File System .....	17
2.8	Hadoop FS Beispiele .....	18
2.9	YARN - Yet Another Resource Negotiator.....	19
2.10	YARN CLI Beispiele .....	20
2.11	MapReduce Überblick.....	21
2.12	MapReduce - Maximale Temperatur je Jahr.....	22
2.13	MapReduce Job Ausführung.....	23
2.14	Der Hadoop Zoo.....	24
2.15	Fazit.....	26
<b>3</b>	<b>Cloudera Data Platform (CDP) .....</b>	<b>27</b>
3.1	Agenda.....	28
3.2	Cloudera Distribution.....	29
3.3	Cloudera Manager .....	30
3.4	Cloudera Manager - Anmeldung.....	31
3.5	Cloudera Manager - Startseite .....	32
3.6	Cloudera Manager Konfiguration .....	33
3.7	Cloudera Manager - Restart YARN .....	34
3.8	Cloudera Manager Architektur .....	35
3.9	Cloudera Linux Dienste.....	36
3.10	Cloudera Linux Dienste starten und stoppen.....	37
3.11	Fazit.....	38
<b>4</b>	<b>HDFS.....</b>	<b>39</b>
4.1	Agenda.....	40
4.2	HDFS - Überblick .....	41
4.3	HDFS Architektur .....	42
4.4	NameNode .....	43
4.5	Secondary NameNode.....	44
4.6	DataNode .....	45
4.7	HDFS Schreiben .....	46
4.8	HDFS Lesen.....	48
4.9	HDFS Besonderheiten .....	50
4.10	HDFS Schnittstellen .....	51
4.11	NameNode Web UI .....	52
4.12	File System Shell .....	53
4.13	File System Shell - get & put.....	54
4.14	File System Shell - Kopieren, verschieben, löschen.....	55
4.15	File System Shell - Verzeichnismangement .....	56
4.16	File System Shell - Dateien anzeigen .....	57
4.17	HDFS Safemode .....	58
4.18	Snapshots .....	59
4.19	HDFS Zugriffsrechte .....	60
4.20	Fazit.....	61
<b>5</b>	<b>Hadoop Konzepte.....</b>	<b>62</b>
5.1	Agenda.....	63
5.2	Hadoop Design Prinzipien.....	64

5.3	Scale Out Linear .....	65
5.4	Commodity Hardware .....	66
5.5	Data Locality .....	67
5.6	Schema on Read .....	68
5.7	Write once, read many (WORM).....	69
5.8	Hadoop Historie .....	70
5.9	Unterschiede zwischen Hadoop 1 und Hadoop 2.....	71
5.10	Neues in Hadoop 3 .....	72
5.11	Vergleich RDBMS mit Hadoop.....	73
5.12	Fazit.....	74
<b>6</b>	<b>Hive 101 .....</b>	<b>75</b>
6.1	Agenda.....	76
6.2	Was ist Apache Hive? .....	77
6.3	Hive Architektur.....	78
6.4	Execution Engines .....	79
6.5	Hive Verteilung der Rollen .....	80
6.6	Beeline Shell .....	81
6.7	Beeline Kommandozeilen Optionen.....	82
6.8	Beeline Kommandos .....	83
6.9	Beeline Hive Kommandos.....	84
6.10	JDBC Interface.....	85
6.11	Hive Datenbanken und Tabellen.....	86
6.12	Datenbank anlegen/verwenden/löschen.....	87
6.13	Tabelle anlegen/löschen .....	89
6.14	Location.....	91
6.15	Managed Tables und External Tables .....	92
6.16	Daten Importieren .....	93
6.17	Daten Exportieren .....	94
6.18	Wichtige Hive Datentypen.....	95
6.19	Numerische Datentypen.....	96
6.20	String Datentypen .....	97
6.21	Datum / Zeit Datentypen .....	98
6.22	Komplexe Datentypen.....	99
6.23	SELECT .....	100
6.24	Fazit.....	102
<b>7</b>	<b>Hive 102 .....</b>	<b>103</b>
7.1	Agenda.....	104
7.2	UDFs .....	105
7.3	UDFs .....	106
7.4	UDAFs.....	109
7.5	UDTFs.....	111
7.6	Views.....	112
7.7	Word Count mit Hive .....	113
7.8	Partitionen .....	117
7.9	Hive Partitionierung.....	118
7.10	Partitionierte Tabelle anlegen .....	119
7.11	Statische Partitionierung .....	120
7.12	Dynamische Partitionierung .....	121
7.13	Partitionen hinzufügen und löschen.....	122
7.14	Fazit.....	123
<b>8</b>	<b>Dateiformate.....</b>	<b>124</b>
8.1	Agenda.....	125
8.2	Dateiformate in Hive.....	126
8.3	Text Dateien - Delimited.....	128
8.4	SequenceFile .....	129
8.5	Avro.....	130
8.6	Parquet.....	132

8.7	ORC .....	134
8.8	Fazit.....	136
<b>9</b>	<b>Spark.....</b>	<b>137</b>
9.1	Agenda.....	138
9.2	Apache Spark Überblick.....	139
9.3	Spark vs. MapReduce - Kein „Entweder oder“ .....	140
9.4	MapReduce I/O .....	141
9.5	Spark I/O .....	142
9.6	Spark vs. MapReduce - Zusammenfassung.....	143
9.7	Apache Spark - Verteilung der Rollen.....	144
9.8	Spark - Architektur .....	145
9.9	Spark Shell - Master YARN.....	146
9.10	Spark im YARN Resource Manager .....	147
9.11	Spark History Server .....	148
9.12	RDD Grundlagen.....	149
9.13	Word Count .....	150
9.14	Resilient Distributed Datasets - RDD .....	151
9.15	Transformationen in Spark.....	152
9.16	Transformationen .....	153
9.17	Actions – Berechnungen Starten .....	155
9.18	Actions.....	156
9.19	Spark Modes .....	157
9.20	Spark - Local Mode (Standalone) .....	158
9.21	Spark - Client Mode (Master im Cluster).....	159
9.22	Spark - Cluster Mode .....	160
9.23	Spark Shell.....	161
9.24	Ausführen von Skripten mit der Spark Shell .....	162
9.25	Spark Submit.....	163
9.26	Fazit.....	164
<b>10</b>	<b>Spark Structured APIs .....</b>	<b>165</b>
10.1	Agenda.....	166
10.2	Spark - Low Level API.....	167
10.3	Spark DataFrame.....	168
10.4	Spark DataSet.....	169
10.5	Spark SQL.....	170
10.6	DataFrame Temperatur.....	171
10.7	DataFrame Temperatur - Extract .....	172
10.8	DataFrame Temperatur - Transform .....	173
10.9	DataFrame Temperatur - Load .....	174
10.10	DataFrame Temperatur - Schema .....	175
10.11	DataFrame API - Daten Aggregation .....	176
10.12	DataFrame Station .....	177
10.13	DataFrame Station - Extract.....	178
10.14	DataFrame Station - Transform .....	179
10.15	DataFrame Station - Load.....	180
10.16	DataFrame Station - Schema .....	181
10.17	DataFrame Join.....	182
10.18	DataFrame API - Pivot .....	183
10.19	Ergebnis speichern - CSV.....	184
10.20	Spark SQL in Hive speichern .....	185
10.21	Spark SQL.....	186
10.22	Fazit.....	187
<b>11</b>	<b>YARN (mit Spark) .....</b>	<b>188</b>
11.1	Agenda.....	189
11.2	YARN - Yet Another Resource Negotiator .....	190
11.3	Spark on YARN Architektur.....	191
11.4	Spark Shell on YARN.....	192

11.5	Cluster Ressourcen.....	193
11.6	YARN Web UI .....	194
11.7	Nodemanager Konfiguration .....	195
11.8	Nodemanager yarn-site.xml .....	196
11.9	YARN Tool .....	197
11.10	Spark Command Line Optionen.....	198
11.11	Spark Dynamic Allocation .....	199
11.12	Fazit.....	201
<b>12</b>	<b>ZooKeeper.....</b>	<b>202</b>
12.1	Agenda.....	203
12.2	ZooKeeper .....	204
12.3	ZooKeeper Verteilung der Rollen.....	205
12.4	ZooKeeper Architektur Überblick .....	206
12.5	znodes.....	208
12.6	ZooKeeper Client .....	209
12.7	ZooKeeper Kommandos .....	210
12.8	ZooKeeper Watches .....	211
12.9	Fazit.....	212
<b>13</b>	<b>HBase .....</b>	<b>213</b>
13.1	Agenda.....	214
13.2	HBase.....	215
13.3	HBase Verteilung der Rollen.....	216
13.4	HBase Datenmodellierung .....	217
13.5	HBase Shell .....	219
13.6	HBase Shell - Namespaces .....	220
13.7	HBase Shell - Tabellen anlegen .....	221
13.8	HBase Shell - CRUD.....	222
13.9	HBase Shell - scan table.....	223
13.10	HBase Shell - count .....	224
13.11	HBase Shell - Tabellen und Namespace löschen .....	225
13.12	Hive HBase Integration .....	226
13.13	External HBase Table anlegen .....	227
13.14	Daten Laden.....	228
13.15	CRUD Operationen .....	229
13.16	HBase Key Design für Hive.....	230
13.17	Tabellen mit zusammengesetztem Schlüssel.....	231
13.18	Fazit.....	233
<b>14</b>	<b>Sqoop .....</b>	<b>234</b>
14.1	Agenda.....	235
14.2	Sqoop.....	236
14.3	Sqoop Verteilung der Rollen .....	237
14.4	Sqoop 1 Architektur Überblick .....	238
14.5	Sqoop List Tables .....	239
14.6	Sqoop Connection Manager .....	240
14.7	Allgemeine Sqoop Parameter .....	241
14.8	Sqoop Eval.....	242
14.9	Sqoop Export - INSERT ONLY .....	243
14.10	Sqoop Export - UPDATE (und INSERT).....	244
14.11	Sqoop Import ins HDFS .....	245
14.12	Sqoop Import in Hive.....	246
14.13	Sqoop Import in HBase .....	247
14.14	Weitere Sqoop Kommandos .....	248
14.15	Fazit.....	249
<b>15</b>	<b>Kafka.....</b>	<b>250</b>
15.1	Agenda.....	251
15.2	Kafka .....	252

---

15.3	Kafka Topics .....	253
15.4	Kafka Verteilung der Rollen .....	254
15.5	Partitionierung und Replikation .....	255
15.6	Kafka Topics verwalten .....	256
15.7	Console Producer & Consumer .....	257
15.8	Consumer Groups .....	258
15.9	Kafka Events .....	260
15.10	Key-Value Paare mit Console Producer & Consumer .....	261
15.11	Fazit.....	262