

Hadoop Grundlagen

Seminarunterlage

Version: 1.11



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

Inhaltsverzeichnis

1	Agenda	7
1.1	Agenda.....	8
2	Hadoop Überblick.....	9
2.1	Agenda.....	10
2.2	Hadoop.....	11
2.3	1st ORDIX Hadoop Cluster (2015)	12
2.4	Yahoo's Hadoop Cluster (2007).....	13
2.5	Hadoop Kernkomponenten	14
2.6	Hadoop Pseudo Distributed Cluster.....	15
2.7	HDFS - Hadoop Distributed File System	16
2.8	Hadoop FS Beispiele	17
2.9	YARN - Yet Another Resource Negotiator.....	18
2.10	YARN CLI Beispiele	19
2.11	MapReduce Überblick.....	20
2.12	MapReduce - Maximale Temperatur je Jahr.....	21
2.13	MapReduce Job Ausführung.....	22
2.14	Der Hadoop Zoo.....	23
2.15	Fazit.....	25
3	Cloudera Data Platform (CDP)	26
3.1	Agenda.....	27
3.2	Cloudera Distribution.....	28
3.3	Cloudera Manager	29
3.4	Cloudera Manager - Anmeldung.....	30
3.5	Cloudera Manager - Startseite	31
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
4	HDFS.....	39
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	NameNode Web UI	51
4.11	File System Shell	52
4.12	File System Shell - get & put.....	53
4.13	File System Shell - Kopieren, verschieben, löschen.....	54
4.14	File System Shell - Verzeichnismangement	55
4.15	File System Shell - Dateien anzeigen	56
4.16	HDFS Safemode	57
4.17	Snapshots	58
4.18	Fazit.....	59
5	Hadoop Konzepte.....	60
5.1	Agenda.....	61
5.2	Hadoop Design Prinzipien.....	62
5.3	Scale Out Linear	63
5.4	Commodity Hardware	64

5.5	Data Locality	65
5.6	Schema on Read	66
5.7	Hadoop Historie	67
5.8	Unterschiede zwischen Hadoop 1 und Hadoop 2.....	68
5.9	Neues in Hadoop 3	69
5.10	Vergleich RDBMS mit Hadoop.....	70
5.11	Fazit.....	71
6	Hive 101	72
6.1	Agenda.....	73
6.2	Was ist Apache Hive?	74
6.3	Hive Architektur.....	75
6.4	Execution Engines	76
6.5	Hive Verteilung der Rollen	77
6.6	Beeline Shell	78
6.7	Beeline Kommandozeilen Optionen.....	79
6.8	Beeline Kommandos	80
6.9	Hive Datenbanken und Tabellen.....	81
6.10	Datenbank anlegen/verwenden/löschen.....	82
6.11	Tabelle anlegen/löschen	84
6.12	Location.....	86
6.13	Managed Tables und External Tables	87
6.14	Daten Importieren	88
6.15	Daten Exportieren	89
6.16	Wichtige Hive Datentypen.....	90
6.17	Numerische Datentypen.....	91
6.18	String Datentypen	92
6.19	Datum / Zeit Datentypen	93
6.20	Komplexe Datentypen.....	94
6.21	SELECT	95
6.22	Fazit.....	96
7	Hive 102	97
7.1	Agenda.....	98
7.2	UDFs	99
7.3	UDFs	100
7.4	UDAFs.....	103
7.5	UDTFs.....	105
7.6	UDTFs (ff.).....	106
7.7	Views.....	107
7.8	Word Count mit Hive	108
7.9	Partitionen	112
7.10	Hive Partitionierung.....	113
7.11	Partitionierte Tabelle anlegen	114
7.12	Statische Partitionierung	115
7.13	Dynamische Partitionierung	116
7.14	Partitionen hinzufügen und löschen.....	117
7.15	Fazit.....	118
8	Dateiformate.....	119
8.1	Agenda.....	120
8.2	Dateiformate in Hive.....	121
8.3	Text Dateien - Delimited.....	123
8.4	Avro	124
8.5	Parquet.....	125
8.6	ORC	126
8.7	Fazit.....	127
9	Spark.....	128
9.1	Agenda.....	129

9.2	Apache Spark Überblick.....	130
9.3	Spark vs. MapReduce - Kein „Entweder oder“	131
9.4	MapReduce I/O	132
9.5	Spark I/O	133
9.6	Spark vs. MapReduce - Zusammenfassung.....	134
9.7	Apache Spark - Verteilung der Rollen.....	135
9.8	Spark - Architektur	136
9.9	Spark Shell - Master YARN.....	137
9.10	Spark im YARN Resource Manager	138
9.11	Spark History Server	139
9.12	RDD Grundlagen.....	140
9.13	Word Count	141
9.14	RDD erzeugen	142
9.15	Transformationen	143
9.16	Transformationen in Spark.....	145
9.17	Actions.....	146
9.18	Actions – Berechnungen Starten	147
9.19	Spark Modes	148
9.20	Spark - Local Mode (Standalone)	149
9.21	Spark - Client Mode (Master im Cluster).....	150
9.22	Spark - Cluster Mode	151
9.23	Spark Shell.....	152
9.24	Ausführen von Skripten mit der Spark Shell	153
9.25	Spark Submit.....	154
9.26	Fazit.....	155
10	Spark Structured APIs	156
10.1	Agenda.....	157
10.2	Spark - Low Level API.....	158
10.3	Spark DataFrame.....	159
10.4	Spark DataSet.....	160
10.5	Spark SQL.....	161
10.6	DataFrame - Maximale Temperatur je Jahr und Region	162
10.7	DataFrame Temperatur - Daten lesen	163
10.8	DataFrame Temperatur - Spalten transformieren.....	164
10.9	DataFrame Temperatur - Spalten auswählen.....	165
10.10	DataFrame - Max Temperatur je Jahr und Station	166
10.11	DataFrame Station - Daten lesen	167
10.12	DataFrame Station - Spalten transformieren	168
10.13	DataFrame Station - Spalten auswählen	169
10.14	DataFrame - Temperatur JOIN Station.....	170
10.15	DataFrame - Maximale Temperatur je Region + Jahr	171
10.16	DataFrame - Speichern im HDFS.....	172
10.17	Spark SQL & Hive	173
10.18	Spark SQL - TempView	174
10.19	Fazit.....	175
11	YARN (mit Spark)	176
11.1	Agenda.....	177
11.2	YARN - Yet Another Resource Negotiator.....	178
11.3	Spark on YARN Architektur.....	179
11.4	Spark Shell on YARN.....	180
11.5	Cluster Ressourcen.....	181
11.6	YARN Web UI	182
11.7	Nodemanager Konfiguration	183
11.8	Nodemanager yarn-site.xml	184
11.9	YARN Tool	185
11.10	Spark Command Line Optionen.....	186
11.11	Spark Dynamic Allocation	187
11.12	Fazit.....	189

12	ZooKeeper	190
12.1	Agenda.....	191
12.2	ZooKeeper	192
12.3	ZooKeeper Verteilung der Rollen.....	193
12.4	ZooKeeper Architektur Überblick.....	194
12.5	znodes.....	196
12.6	ZooKeeper Client	197
12.7	ZooKeeper Kommandos	198
12.8	ZooKeeper Watches	199
12.9	Fazit.....	200
13	HBase	201
13.1	Agenda.....	202
13.2	HBase.....	203
13.3	HBase Verteilung der Rollen.....	204
13.4	HBase Datenmodellierung	205
13.5	HBase Shell	207
13.6	HBase Shell - Namespaces	208
13.7	HBase Shell - Tabellen anlegen	209
13.8	HBase Shell - CRUD.....	210
13.9	HBase Shell - scan table.....	211
13.10	HBase Shell - Tabellen und Namespace löschen	212
13.11	Hive HBase Integration	213
13.12	External HBase Table anlegen	214
13.13	Daten Laden.....	215
13.14	CRUD Operationen	216
13.15	HBase Key Design für Hive.....	217
13.16	Tabellen mit zusammengesetztem Schlüssel.....	218
13.17	Fazit.....	220
14	Sqoop	221
14.1	Agenda.....	222
14.2	Sqoop	223
14.3	Sqoop Verteilung der Rollen	224
14.4	Sqoop 1 Architektur Überblick	225
14.5	Sqoop List Tables	226
14.6	Sqoop Connection Manager	227
14.7	Allgemeine Sqoop Parameter	228
14.8	Sqoop Eval.....	229
14.9	Sqoop Export	230
14.10	Sqoop Import ins HDFS	231
14.11	Sqoop Import in Hive.....	232
14.12	Sqoop Import in HBase	233
14.13	Fazit.....	234
15	Kafka	235
15.1	Agenda.....	236
15.2	Kafka	237
15.3	Kafka Topics	238
15.4	Kafka Verteilung der Rollen	239
15.5	Partitionierung und Replikation	240
15.6	Kafka Topics verwalten.....	241
15.7	Console Producer & Consumer	242
15.8	Consumer Groups.....	243
15.9	Fazit.....	245