

# Obsah

|  |    |
|--|----|
| O autorech .....   | 13 |
| Předmluva .....  | 15 |
| <b>I. Pojem Big Data a principy distribuovaného zpracování dat</b> |    |
| 1. Úvod .....  | 19 |
| 1.1 Jak velká jsou Big Data? .....                                 | 20 |
| 1.2 Historie a vznik NoSQL databází .....                          | 22 |
| 1.2.1 Konec relačních databází? .....                              | 25 |
| 1.3 O čem bude kniha .....   | 27 |
| 2. Datové formáty .....  | 29 |
| 2.1 JSON .....   | 30 |
| 2.1.1 JSON schéma .....  | 33 |
| 2.2 XML .....  | 35 |
| 2.2.1 XML schémata .....   | 37 |
| 2.3 YAML .....   | 39 |
| 2.4 Formáty Linked Data .....                                      | 41 |
| 2.4.1 RDF/XML .....  | 41 |
| 2.4.2 JSON-LD .....  | 42 |
| 2.5 CSV .....  | 43 |
| 2.6 Optimalizace ukládání a přenosu dat .....                      | 44 |
| 2.6.1 Protocol Buffers .....                                       | 44 |
| 2.6.2 Apache Thrift .....  | 45 |
| 2.6.3 BSON .....   | 45 |
| 2.6.4 EXI a FastInfoset .....                                      | 45 |
| 2.6.5 ASN.1 .....  | 46 |
| 2.7 Jaký formát vybrat .....                                       | 46 |
| 3. Základní principy .....   | 47 |
| 3.1 Škálovatelnost .....   | 48 |
| 3.2 Konzistence .....  | 49 |
| 3.2.1 Souběh transakcí .....                                       | 50 |
| 3.2.2 CAP teorém .....   | 52 |
| 3.2.3 Občasná konzistence .....                                    | 54 |

|           |  |           |
|-----------|--|-----------|
| 3.3       | Distribuce .....                             | 56        |
| 3.3.1     | Rozdělení dat (sharding) .....               | 57        |
| 3.3.2     | Master-slave replikace .....                 | 58        |
| 3.3.3     | Peer-to-peer replikace .....                 | 59        |
| 3.3.4     | Replikace + sharding .....                   | 61        |
| <b>4.</b> | <b>Zpracování dat pomocí MapReduce .....</b> | <b>63</b> |
| 4.1       | Funkce Map a Reduce .....                    | 66        |
| 4.1.1     | Další příklady .....                         | 68        |
| 4.2       | MapReduce framework .....                    | 68        |
| 4.2.1     | Další vlastnosti .....                       | 70        |
| 4.3       | Hadoop .....                                 | 72        |
| 4.3.1     | HDFS .....                                   | 72        |
| 4.3.2     | Hadoop MapReduce .....                       | 74        |
| 4.3.3     | Další nastavení systému Hadoop .....         | 76        |
| 4.4       | Kritika a ústup od MapReduce .....           | 82        |

## II. NoSQL databáze

|           |   |            |
|-----------|---|------------|
| <b>5.</b> | <b>Základní principy NoSQL databází .....</b> | <b>87</b>  |
| 5.1       | Společné principy NoSQL databází .....        | 88         |
| 5.2       | Datové modely v NoSQL databázích .....        | 89         |
| 5.3       | Typologie NoSQL databází .....                | 93         |
| <b>6.</b> | <b>Databáze typu klíč-hodnota .....</b>       | <b>95</b>  |
| 6.1       | Principy .....                                | 96         |
| 6.1.1     | Základní operace a práce s klíči .....        | 96         |
| 6.1.2     | Jmenné prostory klíčů .....                   | 97         |
| 6.1.3     | Druhy úložišť typu klíč-hodnota .....         | 98         |
| 6.2       | Realizace a vlastnosti .....                  | 99         |
| 6.2.1     | Distribuce dat .....                          | 99         |
| 6.2.2     | Konzistence a dostupnost dat .....            | 102        |
| 6.2.3     | Lokální organizace dat .....                  | 105        |
| 6.3       | Práce s daty .....                            | 105        |
| 6.3.1     | Sekundární indexy .....                       | 106        |
| 6.3.2     | Redis .....                                   | 106        |
| <b>7.</b> | <b>Dokumentové databáze .....</b>             | <b>109</b> |
| 7.1       | Datový model „dokument“ .....                 | 109        |
| 7.2       | Dotazování a manipulace s daty .....          | 115        |
| 7.2.1     | Dotazy .....                                  | 115        |
| 7.2.2     | Modifikace databáze .....                     | 116        |
| 7.2.3     | Agregované dotazy a MapReduce .....           | 117        |
| 7.2.4     | Shrnutí .....                                 | 117        |

|           |   |            |
|-----------|---|------------|
| 7.3       | Vlastnosti dokumentových databází .....       | 118        |
| 7.3.1     | Indexy .....                                  | 118        |
| 7.3.2     | Replikace dat a dostupnost systému .....      | 119        |
| 7.3.3     | Rozdělení dat .....                           | 122        |
| 7.3.4     | ACID pro jednotlivé operace a transakce ..... | 123        |
| 7.4       | Závěr .....                                   | 124        |
| <b>8.</b> | <b>Sloupcové databáze .....</b>               | <b>127</b> |
| 8.1       | Datový model .....                            | 128        |
| 8.2       | Cassandra: datový model sloupců v praxi ..... | 132        |
| 8.2.1     | Data jako multidimenzionální pole .....       | 133        |
| 8.2.2     | Data jako řádké tabulky .....                 | 134        |
| 8.3       | Struktura a vlastnosti systému .....          | 136        |
| 8.3.1     | Distribuce a replikace dat .....              | 136        |
| 8.3.2     | Lokální organizace dat .....                  | 137        |
| 8.4       | Dotazy, indexy a transakce .....              | 138        |
| 8.4.1     | Dotazy .....                                  | 139        |
| 8.4.2     | Indexy .....                                  | 140        |
| 8.4.3     | Transakce .....                               | 141        |
| <b>9.</b> | <b>Grafové databáze .....</b>                 | <b>143</b> |
| 9.1       | Typy grafů a související pojmy .....          | 145        |
| 9.2       | Databáze Neo4j .....                          | 146        |
| 9.2.1     | Datový model Neo4j .....                      | 146        |
| 9.3       | Přístup k databázi Neo4j .....                | 147        |
| 9.3.1     | Java API .....                                | 147        |
| 9.3.2     | Gremlin .....                                 | 149        |
| 9.3.3     | Cypher .....                                  | 152        |
| 9.4       | Pokročilé rysy Neo4j .....                    | 156        |
| 9.4.1     | Neo4j HA .....                                | 156        |
| 9.4.2     | Transakce .....                               | 157        |
| 9.4.3     | Indexy .....                                  | 158        |
| 9.5       | Další grafové databáze .....                  | 162        |
| 9.5.1     | Sparksee .....                                | 163        |
| 9.5.2     | InfiniteGraph .....                           | 163        |
| 9.5.3     | OrientDB .....                                | 163        |
| 9.5.4     | Titan .....                                   | 164        |
| 9.6       | RDF databáze .....                            | 164        |
| 9.7       | Srovnání úložišť pro grafy .....              | 165        |
| 9.8       | Závěr .....                                   | 167        |

### III. Pokročilé aspekty zpracování Big Data

|   |     |
|---|-----|
| 10. Další aspekty zpracování Big Data .....           | 171 |
| 10.1 Analytické zpracování Big Data .....             | 172 |
| 10.1.1 Schéma dat .....                               | 172 |
| 10.1.2 Tvorba datových skladů .....                   | 175 |
| 10.1.3 Analytické zpracování .....                    | 176 |
| 10.2 Vizualizace Big Data .....                       | 178 |
| 10.2.1 Vizualizace propojených dat .....              | 179 |
| 10.2.2 Nástroje pro vizualizaci .....                 | 181 |
| 10.3 Invertovaný index jako databáze .....            | 182 |
| 10.3.1 Apache Lucene a jeho nastavy .....             | 183 |
| 10.3.2 Zpracování logů .....                          | 186 |
| 10.4 Cloud computing .....                            | 187 |
| 10.4.1 Cloud computing a Big Data .....               | 189 |
| 11. Dotazování nad NoSQL databázemi .....             | 193 |
| 11.1 Přímý přístup pomocí programového rozhraní ..... | 194 |
| 11.2 MapReduce .....                                  | 196 |
| 11.3 Specifické dotazovací jazyky .....               | 196 |
| 11.3.1 Elasticsearch Query DSL .....                  | 197 |
| 11.4 Univerzální dotazovací jazyky .....              | 199 |
| 11.4.1 Deriváty SQL .....                             | 199 |
| 11.4.2 Rozšíření SQL .....                            | 199 |
| 11.4.3 XQuery .....                                   | 202 |
| 11.4.4 JSONiq .....                                   | 203 |
| 11.4.5 SPARQL .....                                   | 203 |
| 11.5 Závěr .....                                      | 204 |
| 12. Transakce v distribuovaném prostředí .....        | 205 |
| 12.1 Vlastnosti CAP podrobněji .....                  | 205 |
| 12.2 Základní transakční modely .....                 | 206 |
| 12.2.1 Ploché transakce .....                         | 207 |
| 12.2.2 Zřetěžené transakce .....                      | 207 |
| 12.2.3 Hnízděné transakce .....                       | 207 |
| 12.3 Transakce v distribuovaném prostředí .....       | 208 |
| 12.3.1 2PC protokol .....                             | 209 |
| 12.3.2 3PC protokol .....                             | 210 |
| 12.4 Optimistické a pesimistické off-line zámky ..... | 210 |
| 12.4.1 Optimistický přístup .....                     | 211 |
| 12.4.2 Pesimistický přístup .....                     | 211 |
| 12.5 Uspořádání časových razítek .....                | 213 |
| 12.5.1 Pesimistické uspořádání .....                  | 213 |
| 12.5.2 Optimistické uspořádání .....                  | 214 |

|                                 |   |            |
|---------------------------------|---|------------|
| 12.6                            | MVCC .....  | 215        |
| 12.7                            | Závěr .....   | 216        |
| <b>13.</b>                      | <b>Pokročilé aspekty grafových databází .....</b>   | <b>217</b> |
| 13.1                            | Reprezentace grafů .....                            | 217        |
| 13.1.1                          | Matice sousednosti .....                            | 218        |
| 13.1.2                          | Seznam sousedů .....                                | 218        |
| 13.1.3                          | Matice incidence .....                              | 219        |
| 13.1.4                          | Laplaceova matice .....                             | 219        |
| 13.2                            | Lokalita dat .....                                  | 220        |
| 13.3                            | Distribuce grafu .....                              | 221        |
| 13.4                            | Dotazování nad grafy .....                          | 223        |
| 13.4.1                          | Typy dotazů .....                                   | 224        |
| 13.4.2                          | Vyhodnocování dotazů a indexace grafových dat ..... | 225        |
| 13.4.3                          | Dotazovací jazyky pro grafy .....                   | 234        |
| 13.5                            | Závěr .....   | 242        |
| <b>14.</b>                      | <b>Další databáze pro Big Data .....</b>            | <b>243</b> |
| 14.1                            | Hybridní databáze .....                             | 243        |
| 14.1.1                          | PostgreSQL .....                                    | 244        |
| 14.1.2                          | MarkLogic .....                                     | 249        |
| 14.2                            | Databáze ve webovém prohlížeči .....                | 250        |
| 14.2.1                          | Web Storage .....                                   | 250        |
| 14.2.2                          | Indexed Database .....                              | 252        |
| 14.3                            | NewSQL databáze .....                               | 254        |
| 14.3.1                          | VoltdB .....  | 255        |
| 14.4                            | Array databases .....                               | 256        |
| 14.4.1                          | SciDB .....   | 257        |
| <b>Závěr</b> .....              | <b>261</b>  |            |
| <b>Použitá literatura</b> ..... | <b>263</b>  |            |
| <b>Rejstřík</b> .....           | <b>273</b>  |            |