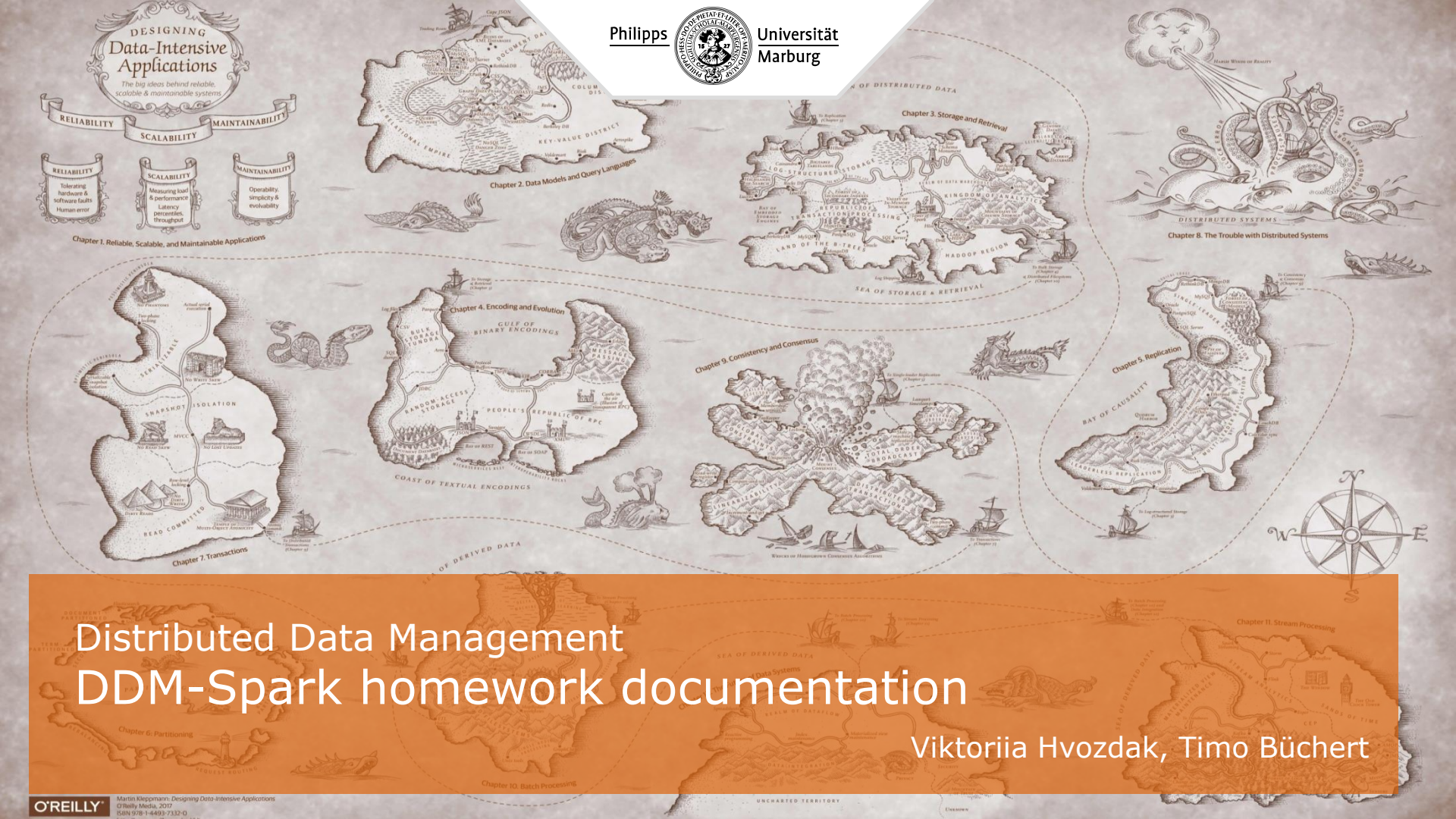

Philipps Universität Marburg

Distributed Data Management

DDM-Spark homework documentation

Viktoriia Hvozdak, Timo Büchert

O'REILLY
Martin Kleppmann: *Designing Data-Intensive Applications*
O'Reilly Media, 2017
ISBN 978-1-4493-7332-0

DDM-Spark homework documentation

Transformation pipeline

Philipps Universität Marburg

The code begins by reading input data and flattening it into key-value pairs. These pairs, representing column names and cell values, are then merged from multiple sources into a single dataset. The merged data is grouped by cell values to identify distinct groups.

Inclusion dependencies (INDs) are then found within each group by intersecting attribute values. The resulting dependencies are sorted by attribute names and printed out. Pairs with empty sets of values are excluded from the output for clarity and relevance.