Managing Heterogeneity in Big Data Integration in Data Lake Systems

PD Dr. Christoph Quix

Fraunhofer Institute for Applied Information Technology FIT
St. Augustin, Germany

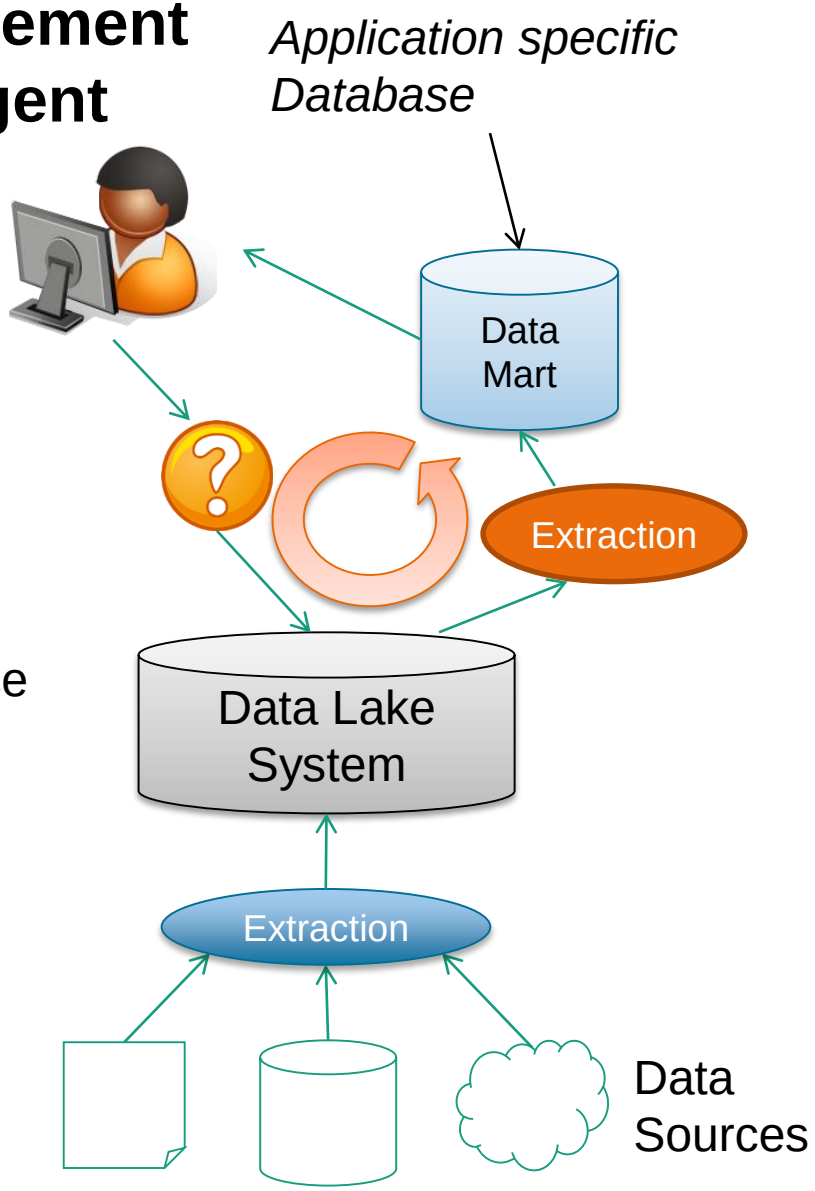
http://fit.fraunhofer.de
christoph.quix@fit.fraunhofer.de

Also affiliated with:
Informatik 5 (Information Systems)
RWTH Aachen University
Aachen, Germany
http://dbis.rwth-aachen.de
quix@dbis.rwth-aachen.de

Fraunhofer FIT

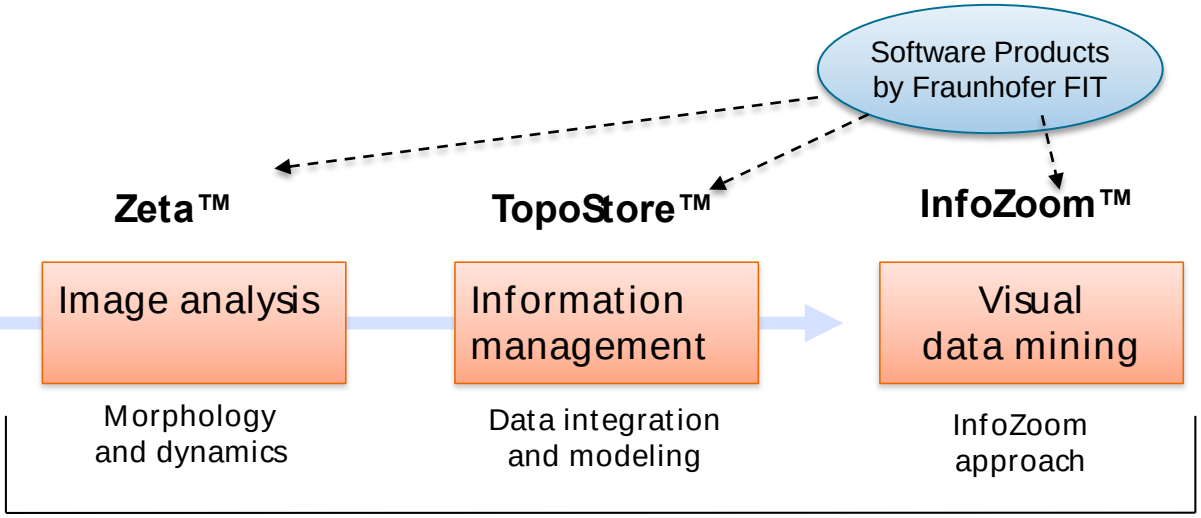
Our Approach to Big Data Management
Integrated – Incremental – Intelligent

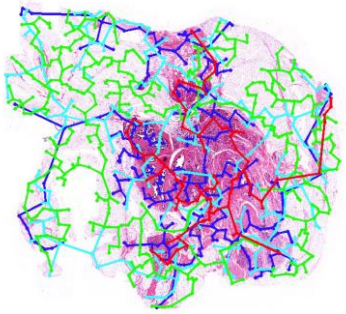
- Adhoc-Integration of heterogeneous data is frequently required, especially in scientific applications
- Heterogenity is immanent in Big Data;
  ➔ Manage it, instead of trying to remove it
  ➔ Data Lake System keeps data in its original format and provides uniform interface to access heterogeneous data
- Metadata are the key for data exploration
- Extract detail data only on demand
- Create application specific data collections instead universal, integrated database

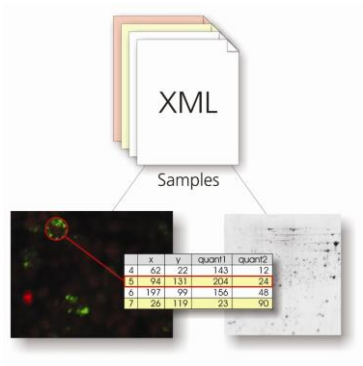


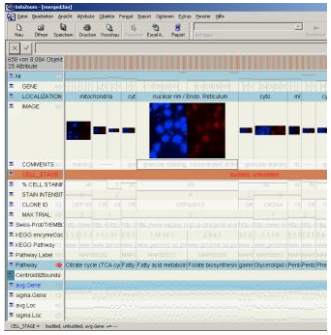
Fraunhofer FIT

Big Data in Life Sciences









- High-Content-Analysis
- Systematic Analysis of huge image sets
- Automated image analysis
- Meta data extraction from multimedia data
- Data management not only in life sciences → Scientific Data Management
- Workflow integration
- Management of heterogeneous data

Fraunhofer FIT