



KNIME HPC INTEGRATION VIA UNICORE

Richard Grunzke*, Florian Jug*, Bernd
Schuller, René Jäkel, Gene Myers,
Wolfgang E. Nagel

www.scads.de

- Introduction
- KNIME
- UNICORE Data Oriented Processing
- KNIME HPC Integration via UNICORE
- Conclusion and Outlook

- **Introduction**
- KNIME
- UNICORE Data Oriented Processing
- KNIME HPC Integration via UNICORE
- Conclusion and Outlook

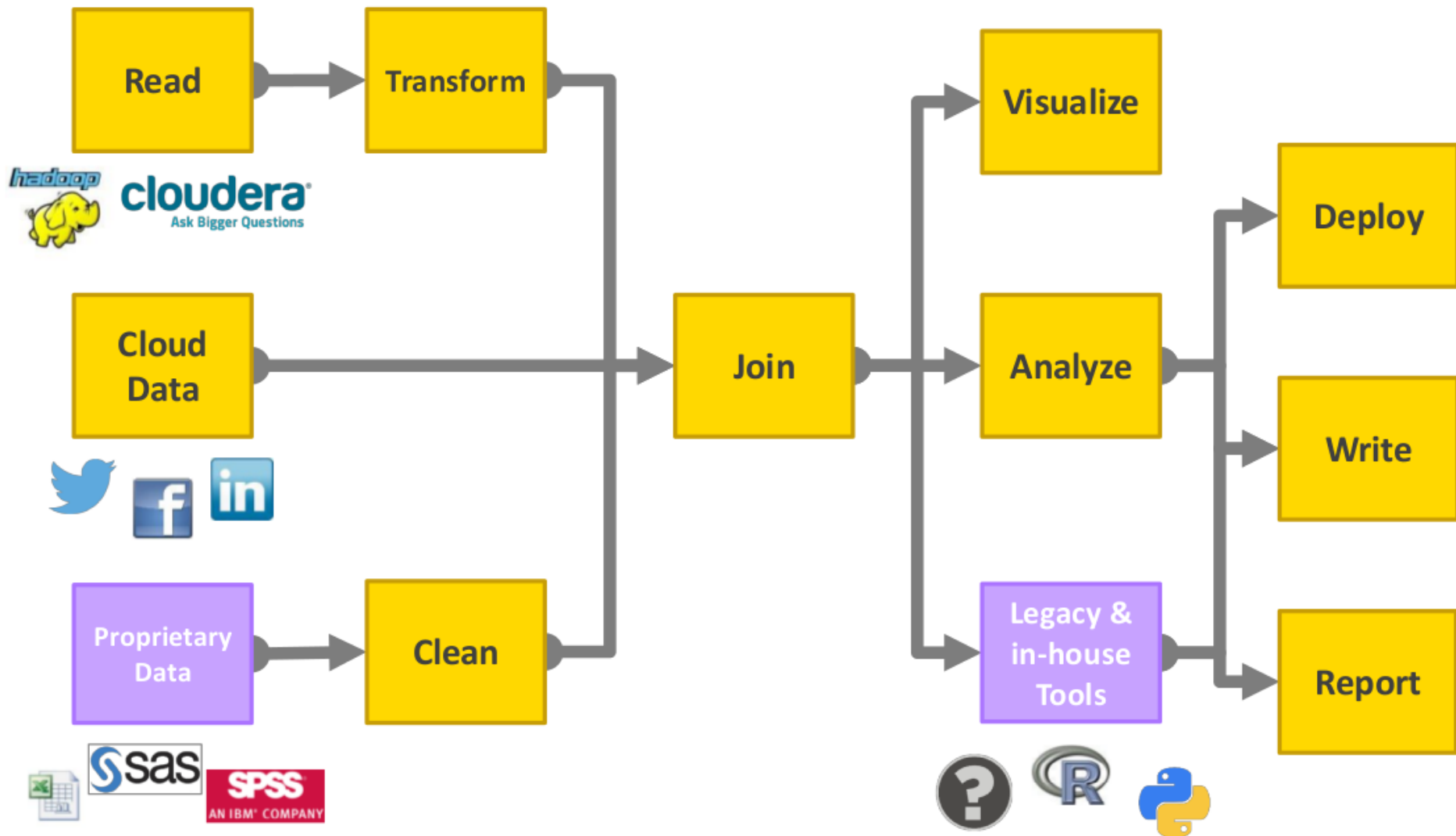
- Varying user data sets and workflows in KNIME
- Challenge of parallelization and HPC integration
- KNIME for workflow creation and execution
 - But lacks generic HPC integration
- UNICORE as middleware for generic HPC access
- HPC Integration of KNIME via UNICORE

- Evaluation within image analysis use case
- In cooperation with
 - MPI-CBG (Jug, Myers)
 - JSC Jülich (Schuller)

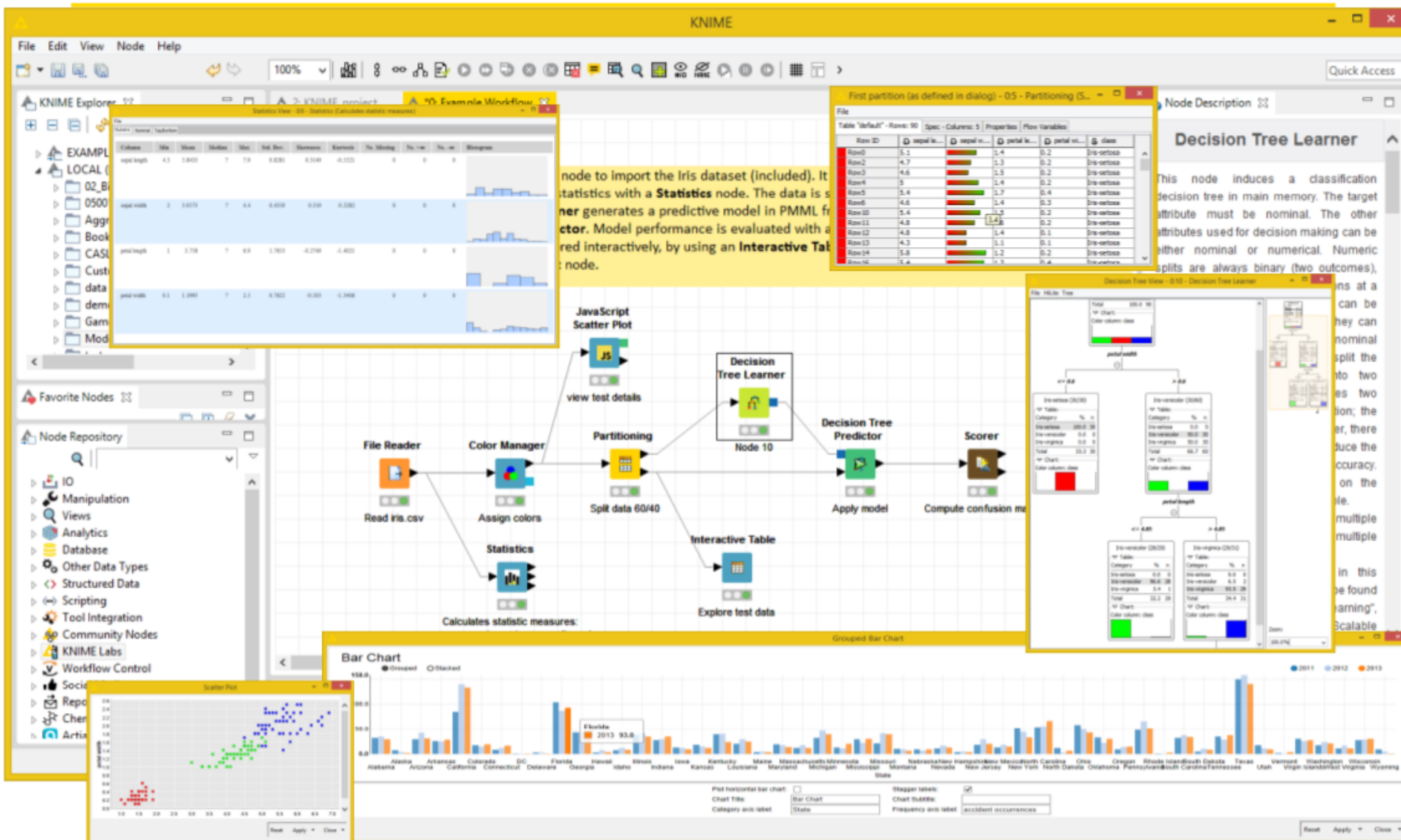
- Introduction
- **KNIME**
- UNICORE Data Oriented Processing
- KNIME HPC Integration via UNICORE
- Conclusion and Outlook

- KNIME – The Open Analytics Platform
- Flexibly analyze data with workflows
- Re-use of sub-workflows
- Graphical ease-of-use
- Widely used in academia and industry

KNIME: Integrating Data and Tools

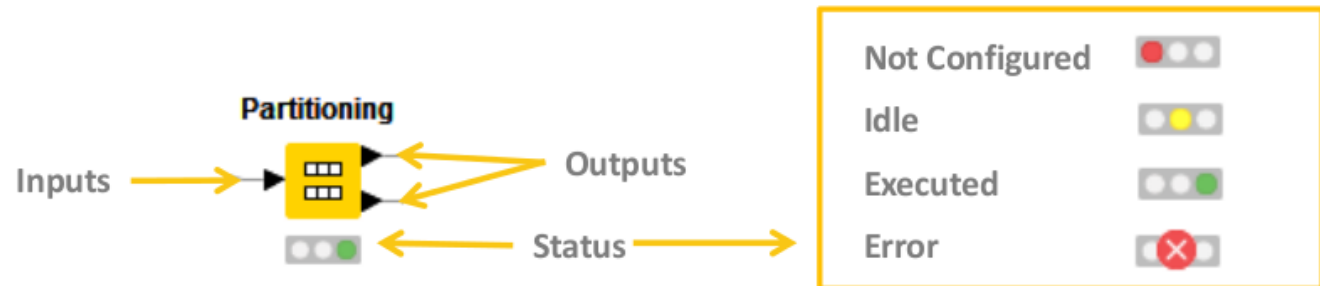


The KNIME Analytics Platform

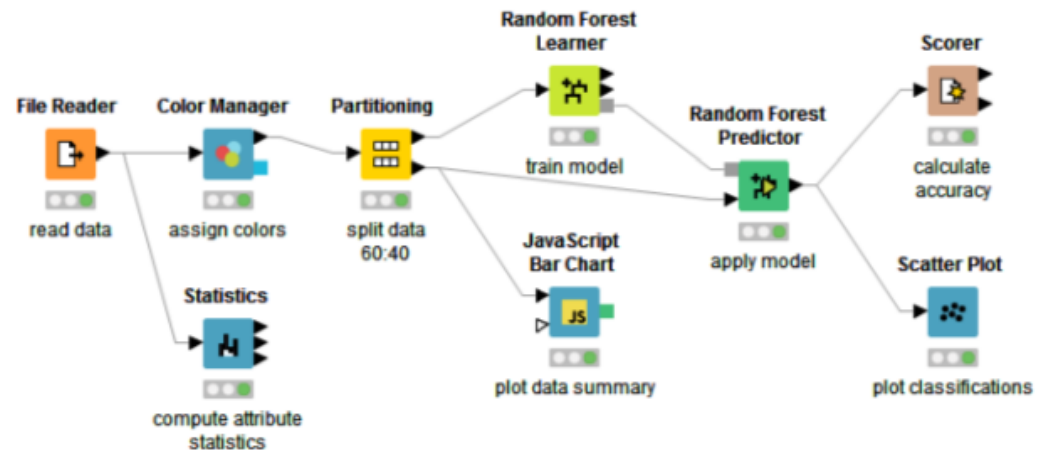


Visual KNIME Workflows

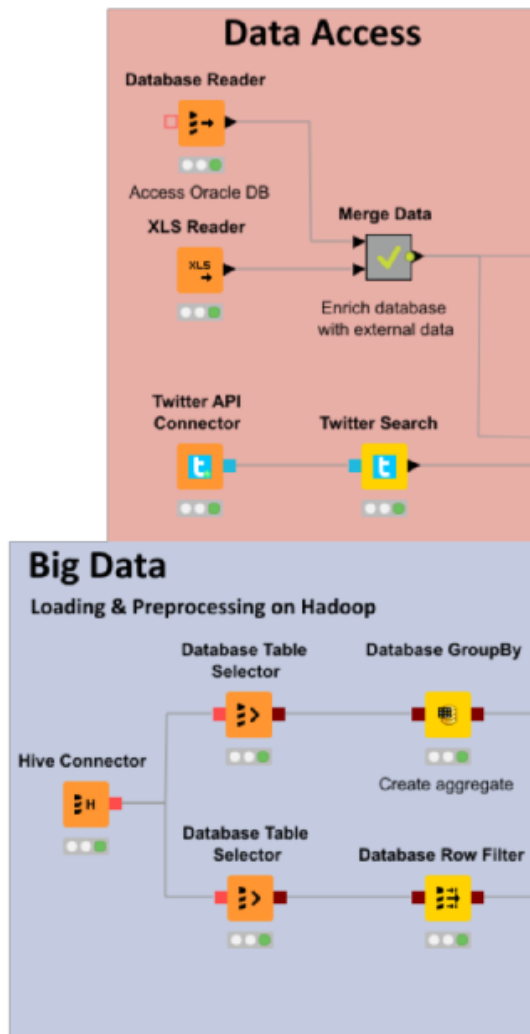
NODES perform tasks on data



Nodes are combined to create
WORKFLOWS

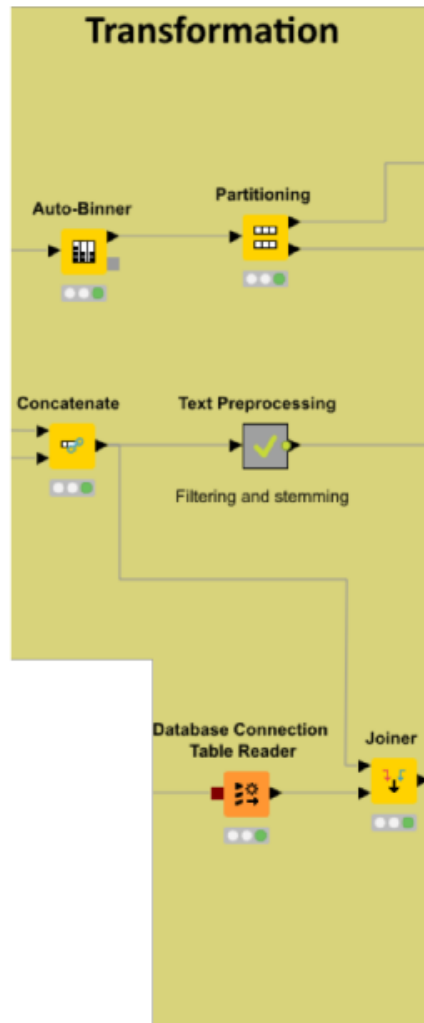


Data Access



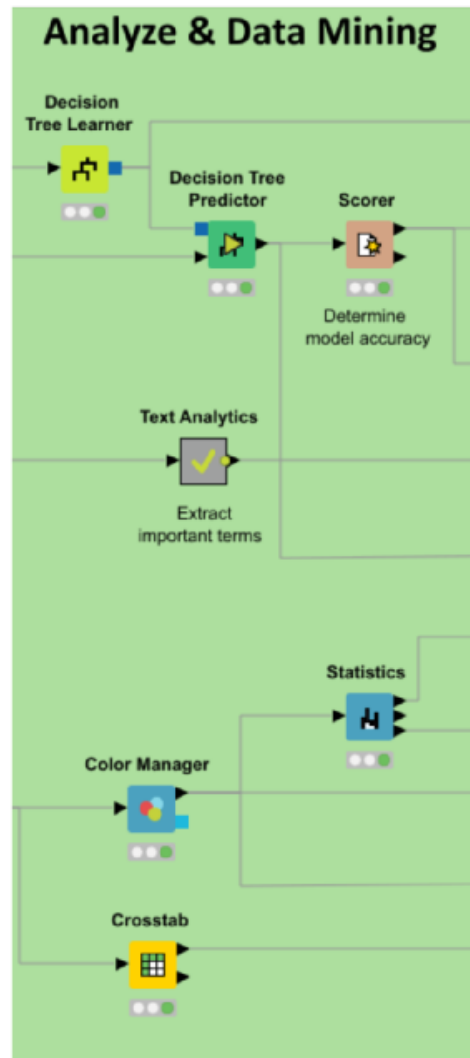
- **Databases**
 - MySQL, PostgreSQL
 - any JDBC (Oracle, DB2, MS SQL Server)
- **Files**
 - Csv, txt
 - Excel, Word, PDF
 - SAS, SPSS
 - XML
 - PMML
 - Images, texts, networks, chem
- **Web, Cloud**
 - REST, Web services
 - Twitter, Google

Transformation



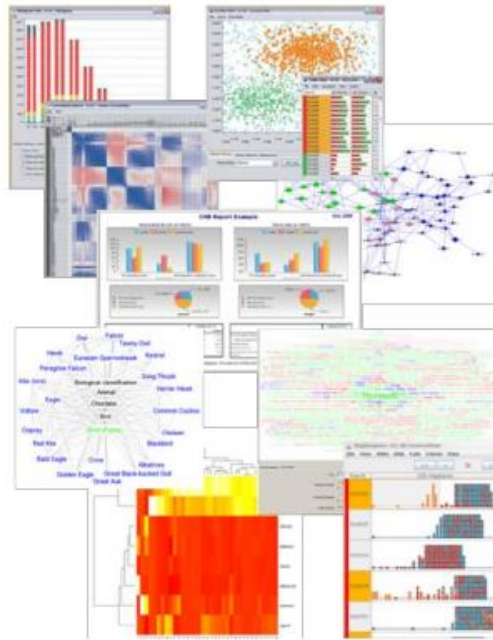
- Preprocessing
 - Row, column, matrix based
- Data blending
 - Join, concatenate, append
- Aggregation
 - Grouping, pivoting, binning
- Feature Creation and Selection

Analyze & Data Mining



- Regression
 - Linear, Logistic
- Classification
 - Decision tree, ensembles, SVM, MLP, Naïve Bayes
- Clustering
 - k-means, DBSCAN, hierarchical
- Validation
 - Cross-validation, scoring, ROC
- Misc
 - PCA, MDS, item set mining
- External
 - R, Weka

Visualization



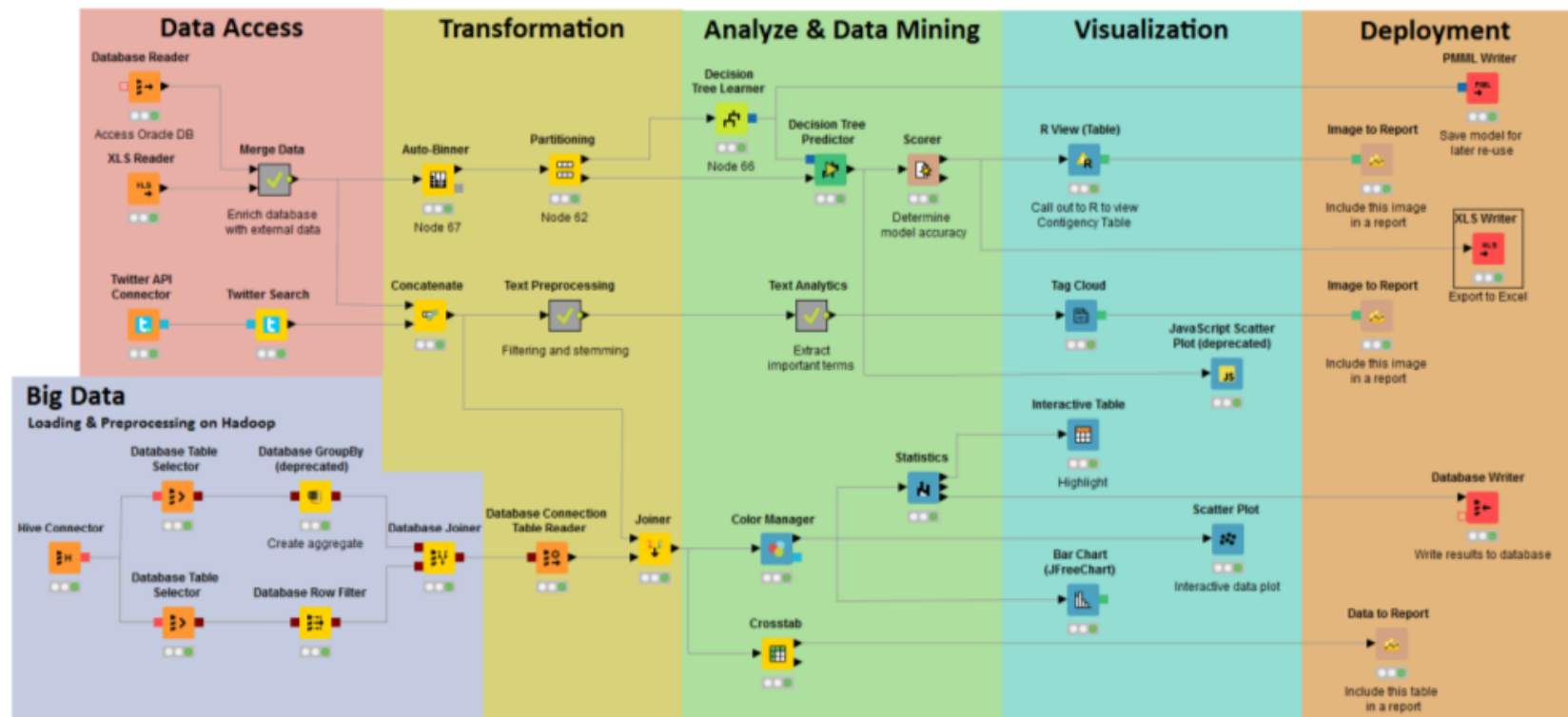
- Interactive
 - Scatter plot, histogram, pie charts, box plot
 - Highlighting (brushing)
- JFreeChart
- JavaScript
- Misc
 - Tag cloud, open street map, networks, molecules
- External
 - R

Deployment



- Database
- Files
 - Excel, csv, txt
 - XML
 - PMML
 - to: local, KNIME Server, SSH-, FTP-Server
- BIRT Reporting

Over 1000 native and embedded nodes included:



Data Access

MySQL, Oracle, ...
SAS, SPSS, ...
Excel, Flat, ...
Hive, Impala, ...
XML, JSON, PMML
Text, Doc, Image, ...
Web Crawlers
Industry Specific
Community / 3rd

Transformation

Row,
Column
Matrix
Text, Image
Time Series
Java
Python
Community / 3rd

Analysis & Mining

Statistics
Data Mining
Machine Learning
Web Analytics
Text Mining
Network Analysis
Social Media Analysis
R, Weka, Python
Community / 3rd

Visualization

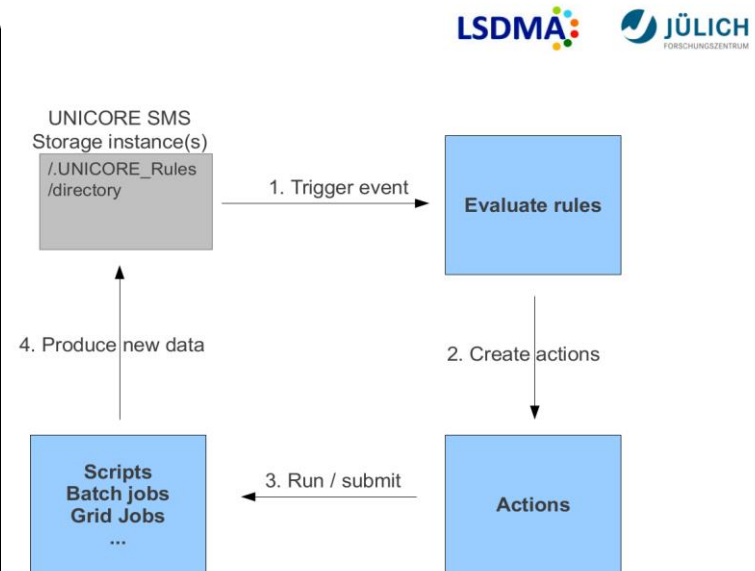
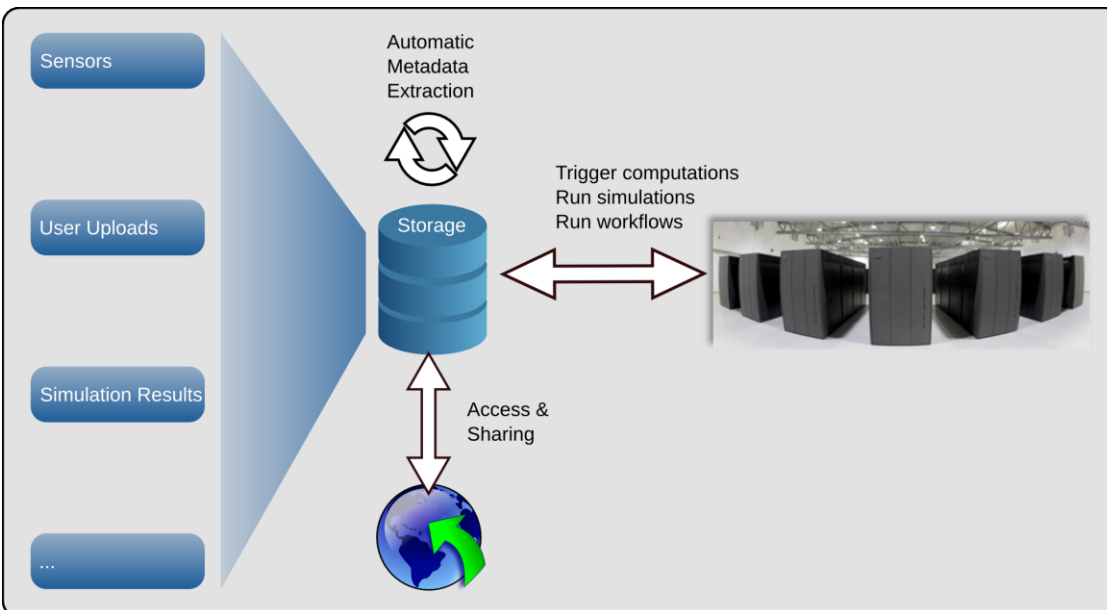
R
JFreeChart
JavaScript
Community / 3rd

Deployment

via BIRT
PMML
XML, JSON
Databases
Excel, Flat, etc.
Text, Doc, Image
Industry Specific
Community / 3rd

- Introduction
- KNIME
- **UNICORE Data Oriented Processing**
- KNIME HPC Integration via UNICORE
- Conclusion and Outlook

- UNICORE monitors directory
- Pre-defined rule trigger actions for new files
- Incoming data directly triggers analysis tasks
- E.g. for preprocessing, metadata extraction, compression, ...



An example



- Goal: calculate checksums (md5) of PDF files in a certain directory using batch jobs
- Rule (job is in UCC syntax!)

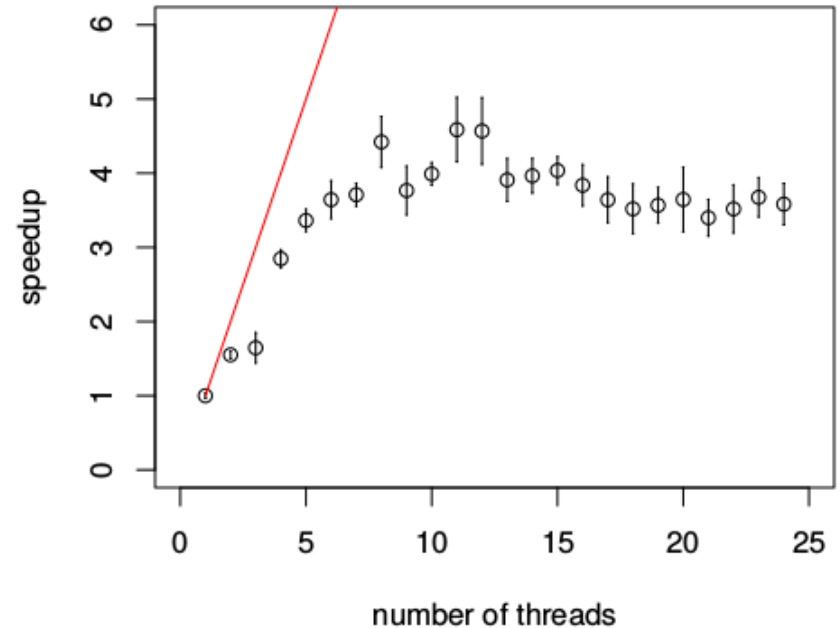
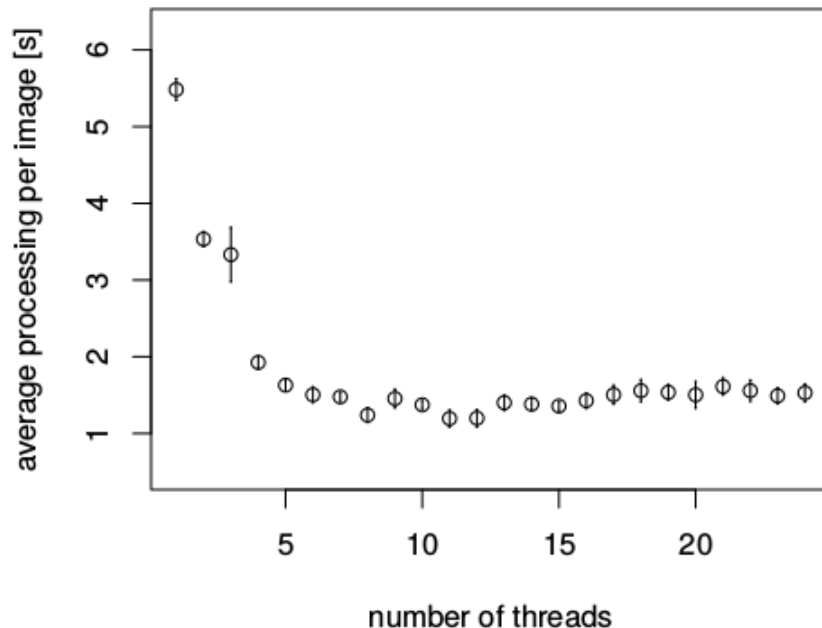
```
Name: computeMD5Sum, Match: ".*\\.pdf",
Action: {
  Type: BATCH,
  Job: {
    Executable: "/usr/bin/md5sum",
    Arguments: ["${UC_FILE_PATH}"],
    Exports: [
      {From: "stdout",
       To: "file://${UC_BASE_DIR}/checksums/${UC_FILE_NAME}.md5"},
    ],
  }
}
```

- Introduction
- KNIME
- UNICORE Data Oriented Processing
- **KNIME HPC Integration via UNICORE**
- Conclusion and Outlook

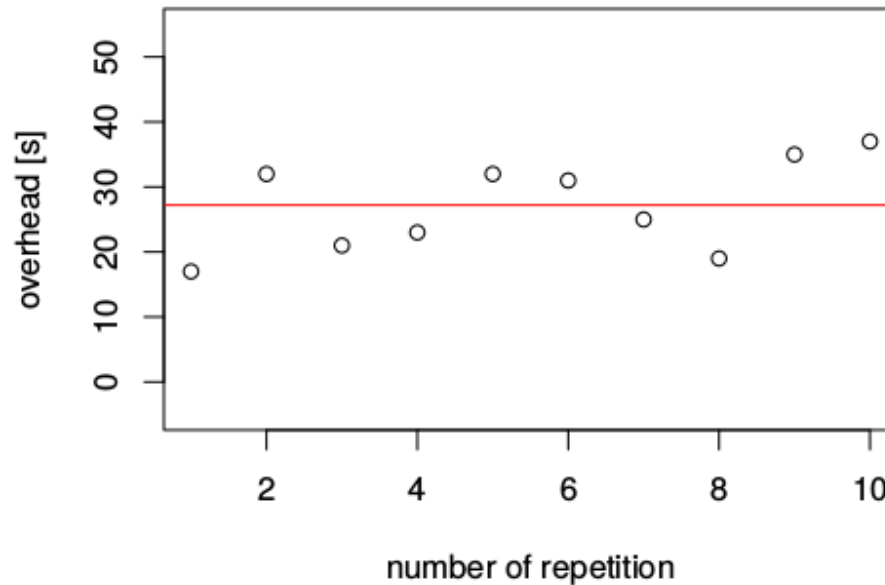
- Configuration of UNICORE DOP feature for remote directory
- User graphically exports workflow in KNIME to directory
- Automatic execution of workflow on HPC system
- Existing data sets on cluster extended with results
- User result access via directory
- Easy to use

```
{
  DirectoryScan: { IncludeDirs: ["."], "Interval": "30", },
  Rules: [ {
    Name: BioHPCMeasurements_2880_1,
    Match: ".*.zip",
    Action: { Type: BATCH, Job: {
      Name: knime_headless,
      Imports: [{ From: "file://${UC_FILE_PATH}",
        To: workflow.zip },],
      ApplicationName: knime,
      ApplicationVersion: 2.11.3_headless,
      Parameters: {
        l: "/lustre/ssd/grunzke/2880_1",
        k: { From: 0, To: 9, Step: 1 },
        n: "10",
        w: "../workflow.zip",
        t: "8",
        tr: "${UC_FILE_PATH}", },
      Resources: {
        Memory: 20664M, CPUs: 8,
        Runtime: 1h, Queue: haswell,
        CPUsPerNode: 8, }, }, },
    }, ... ], }
```

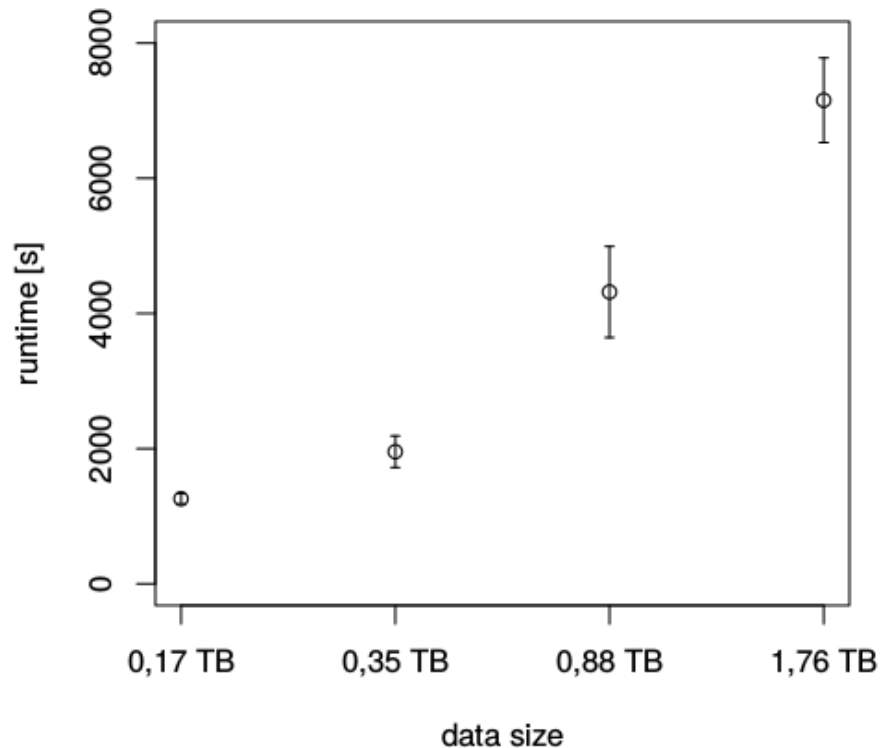
- Mean processing time and speedup with varying # of threads
- Balance at about 8 threads



- UNICORE induced overhead per workflow execution
- Average of 27s out of 10 repetitions



- Measurements with up to 1,76 TB in 7,488 M files
- Overall runtime previously 17d - now 2h → 400x faster



- Introduction
- KNIME
- UNICORE Data Oriented Processing
- KNIME HPC Integration via UNICORE
- **Conclusion and Outlook**

- Challenge of parallelization and HPC integration
- KNIME as widely used workflow system
- UNICORE for HPC and Big Data capabilities
- Integration of both
- KNIME now HPC-enabled in generic and easy way

CONTACT

Dr. Richard Grunzke

Technische Universität Dresden
Zentrum für Informationsdienste und
Hochleistungsrechnen
01062 Dresden

Telefon: +49 351 463-35448
E-Mail: richard.grunzke@tu-dresden.de



GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung



DLR Projektträger

**SCIENTIFIC
COORDINATOR**

Prof. Dr. Wolfgang E. Nagel

Technische Universität Dresden
Zentrum für Informationsdienste und
Hochleistungsrechnen
01062 Dresden

Telefon: +49 351 463-35450
E-Mail: wolfgang.nagel@tu-dresden.de

MANAGEMENT DIRECTOR

Dr. René Jäkel

Technische Universität Dresden
Zentrum für Informationsdienste und
Hochleistungsrechnen
01062 Dresden

Telefon: +49 351 463-42331
E-Mail: rene.jaekel@tu-dresden.de

**DEPUTY SCIENTIFIC
COORDINATOR**

Prof. Dr. Erhard Rahm

Universität Leipzig
Fakultät für Mathematik und Informatik
Augustusplatz 10
04109 Leipzig

Telefon: +49 341 97-32221
E-Mail: rahm@informatik.uni-leipzig.de