Towards Integrating Pervasive DataRush and KNIME: Parallelizing Data Mining

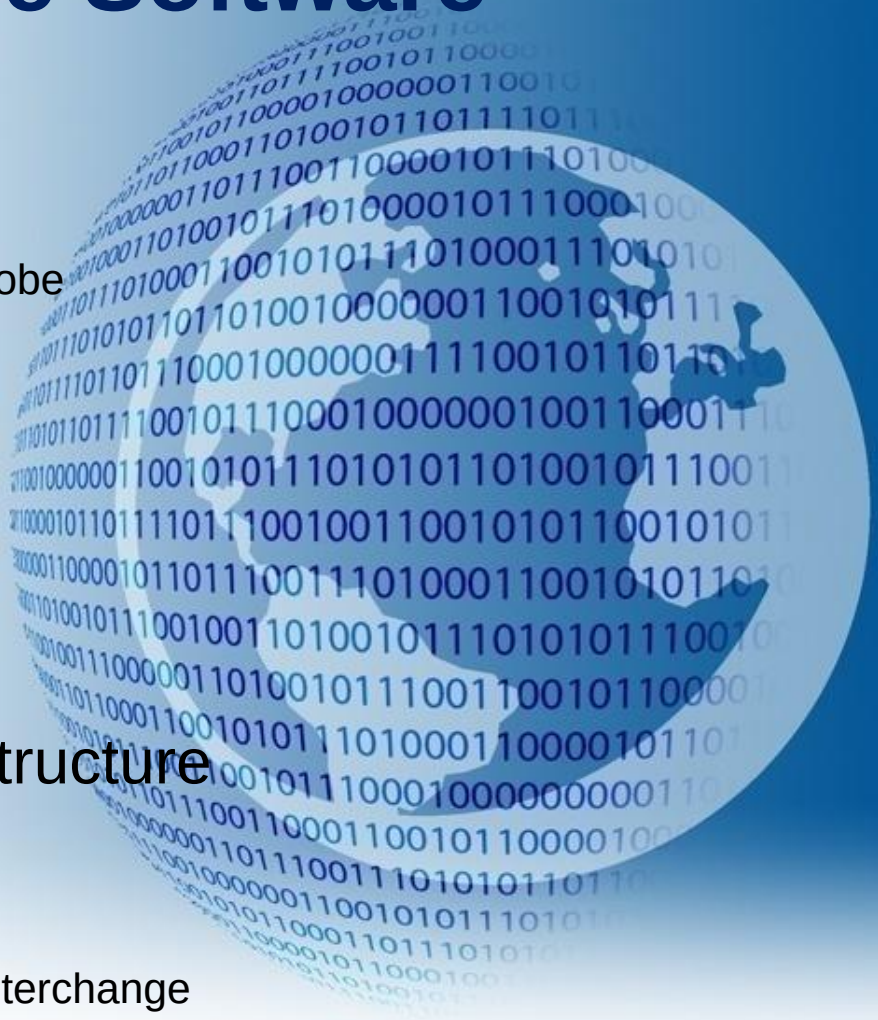
February 2010

PERVASIVE® DATARUSH™: UNLEASH THE POWER OF YOUR DATA

1-866-980-RUSH (7874) or +1-512-231-6818
www.pervasivedatarush.com

PERVASIVE

Introducing Pervasive Software



Global Software Company

- Based in Austin, TX
- Tens of thousands of users across the globe
- Americas, EMEA, Asia

Strong Financials

- $47 million revenue (fiscal 2009)
- 36 consecutive quarters of profitability
- NASDAQ:PVSW since 1997

Leader in Embedded Data Infrastructure

- Data Management
- Data and Application Integration
- Web-based Business-to-Business Data Interchange
- Revolutionary Next-Generation Analytics

Integration Reach

Enterprise Integration Fabric

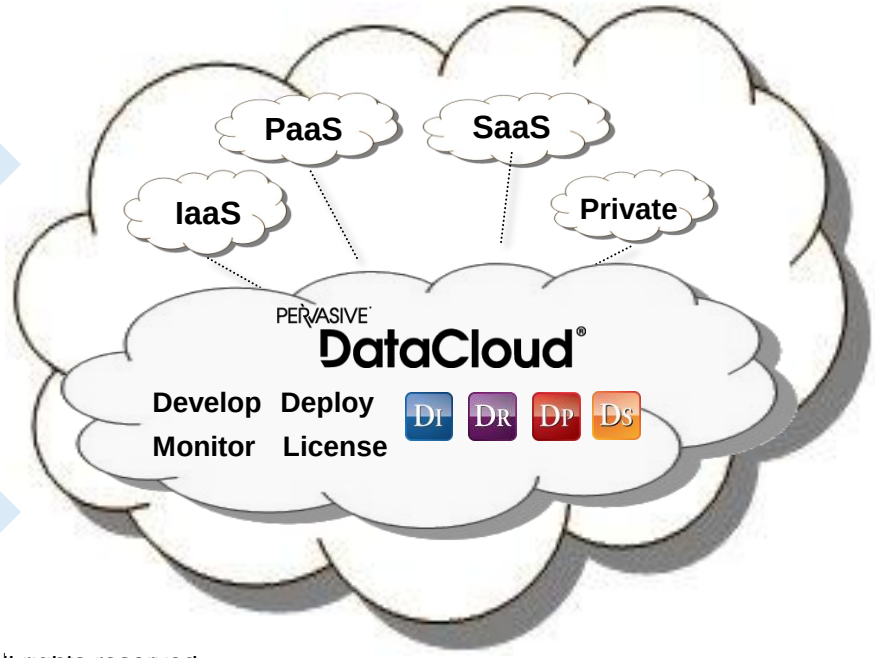
- PERVASIVE Data Integrator™
- PERVASIVE Metadata Management™ for Pervasive Data Integrator
- PERVASIVE Data Profiler™
- PERVASIVE DataRush™
- PERVASIVE Integration Agents

Custom Solutions

- PERVASIVE Integration Hub™ for Pervasive Data Integrator
- PERVASIVE Data Matcher™ for Pervasive Data Integrator
- PERVASIVE Integration Agents

Packaged Solutions

- PERVASIVE Business Xchange™
- PERVASIVE DataSolutions™



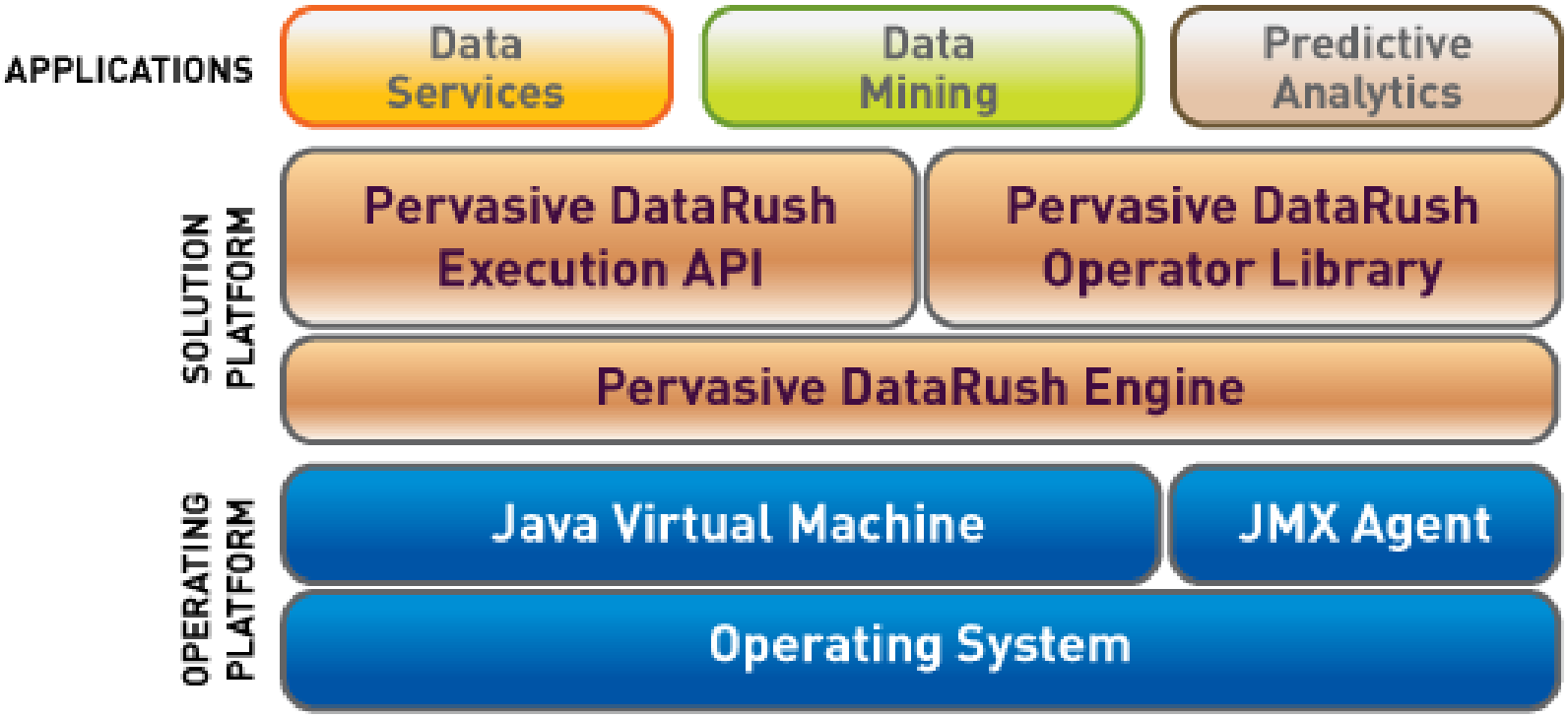
Introducing Pervasive DataRush™

Pervasive DataRush is a platform for building high performing, scalable, data-intensive applications. Implemented with a dataflow architecture, DataRush exposes the power of multicore processors to all developers through its easy to use interface.

DataRush is:

- **Easy to Implement:** DataRush provides several easy to use Java API's that hide the complexity of multi-thread programming.
- **100% Java:** Taking advantage of the Java platform enables multi-platform support and easy integration with other platforms such as KNIME
- **Scalable:** Dynamically scalable and transportable across different hardware platforms and operating systems
- **Next Generation**: Built to take advantage of multicore processors, DataRush shows incredible performance improvements over older, single threaded code.

Pervasive DataRush Platform Architecture



Data Services

- Extract
  - Hundreds of connectors to commercial data stores and multiple data formats
- Aggregate
  - Consumes vast amounts of data to distill the most relevant pieces of information.
  - Group/Summarize records on multiple keys
- Profile
  - Define data metrics such as frequency distribution, average, mean and median.
  - Define quality thresholds such as valid ranges, not null and value outliers to assess data quality.
  - Vertically partitions data for fastest execution
- Cleanse
  - Fuzzy matching (Pervasive DataMatcher) for data de-duplication or record linkage
  - Remediation with connectivity to data dictionaries - Address/Postal code
- Enrich
  - Geocode, Address/Zip/Demographic
  - Relationship analysis
- Consolidate
  - Fuzzy matching for record linkage
  - Federate data

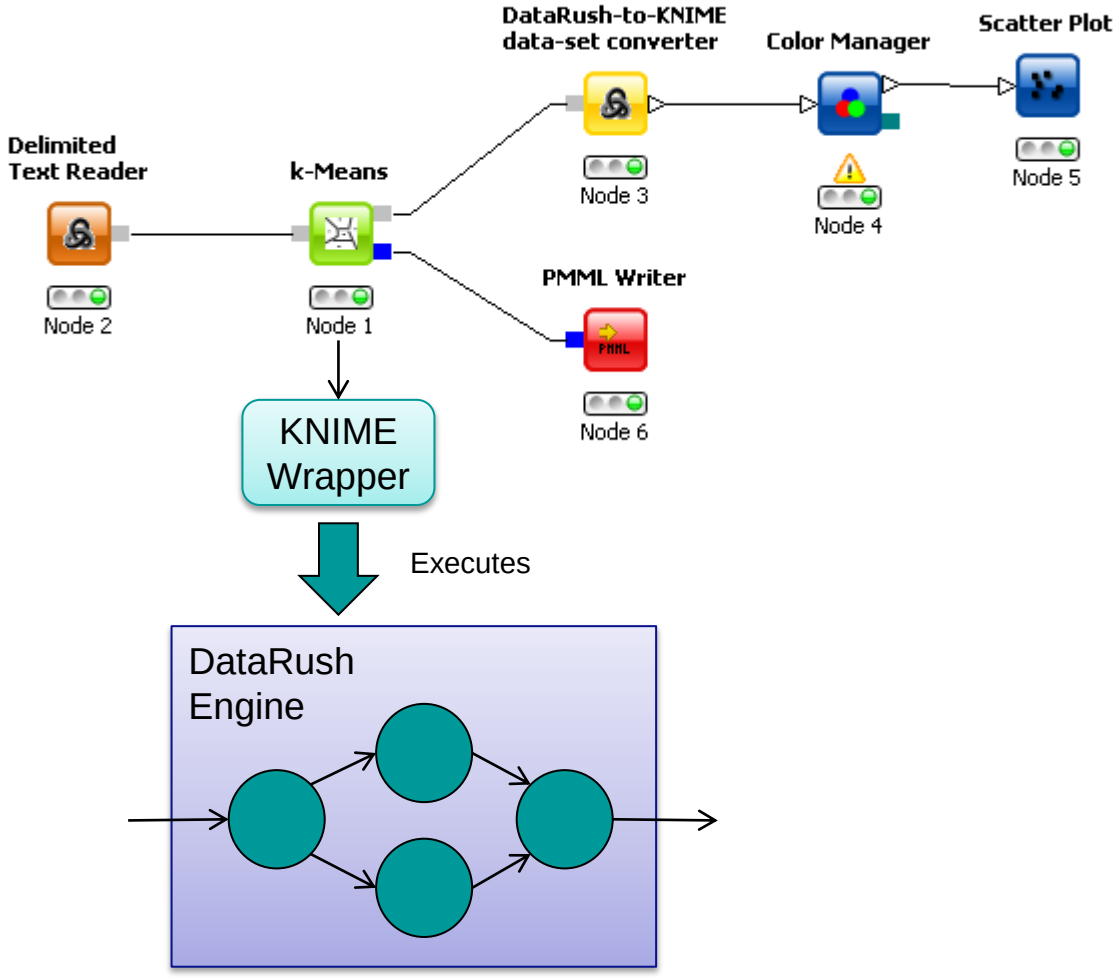
Data Mining and Predictive Analytics

- Clustering
  - K-means
  - Co-clustering
- Classifiers
  - C4.5
  - CART
  - Naïve Bayes
  - KNN
- Feature Selection
  - Principal Component Analysis (PCA)
- Predictive
  - Collaborative Filtering (Rush Recommender)
- Regression
  - Linear
  - Multiple
  - Polynomial
  - Logistic
- Association Rule Mining
  - FP-growth

KNIME & DataRush Integration

- Design Approach
  - Create node extensions within KNIME exposing DataRush functionality
  - DataRush nodes should integrate with standard KNIME nodes
  - Focus on data mining nodes first but also include some data processing/transformation functionality as needed
  - Create common “wrappers” for DataRush framework within KNIME to ease porting effort
  - Each node executes DataRush engine internally allowing thread-level parallelization of data mining algorithms
- Issues
  - A few issues with KNIME framework, turned around quickly by KNIME team
  - BufferedDataTable performance issues
    - BufferedDataTable provides inter-node communication
    - Highly functional (cell level interface)
  - Work around
    - Implement “DataRush” port type; utilizes proprietary DataRush data staging scheme
    - Provide converter nodes from DataRush->KNIME and KNIME->DataRush

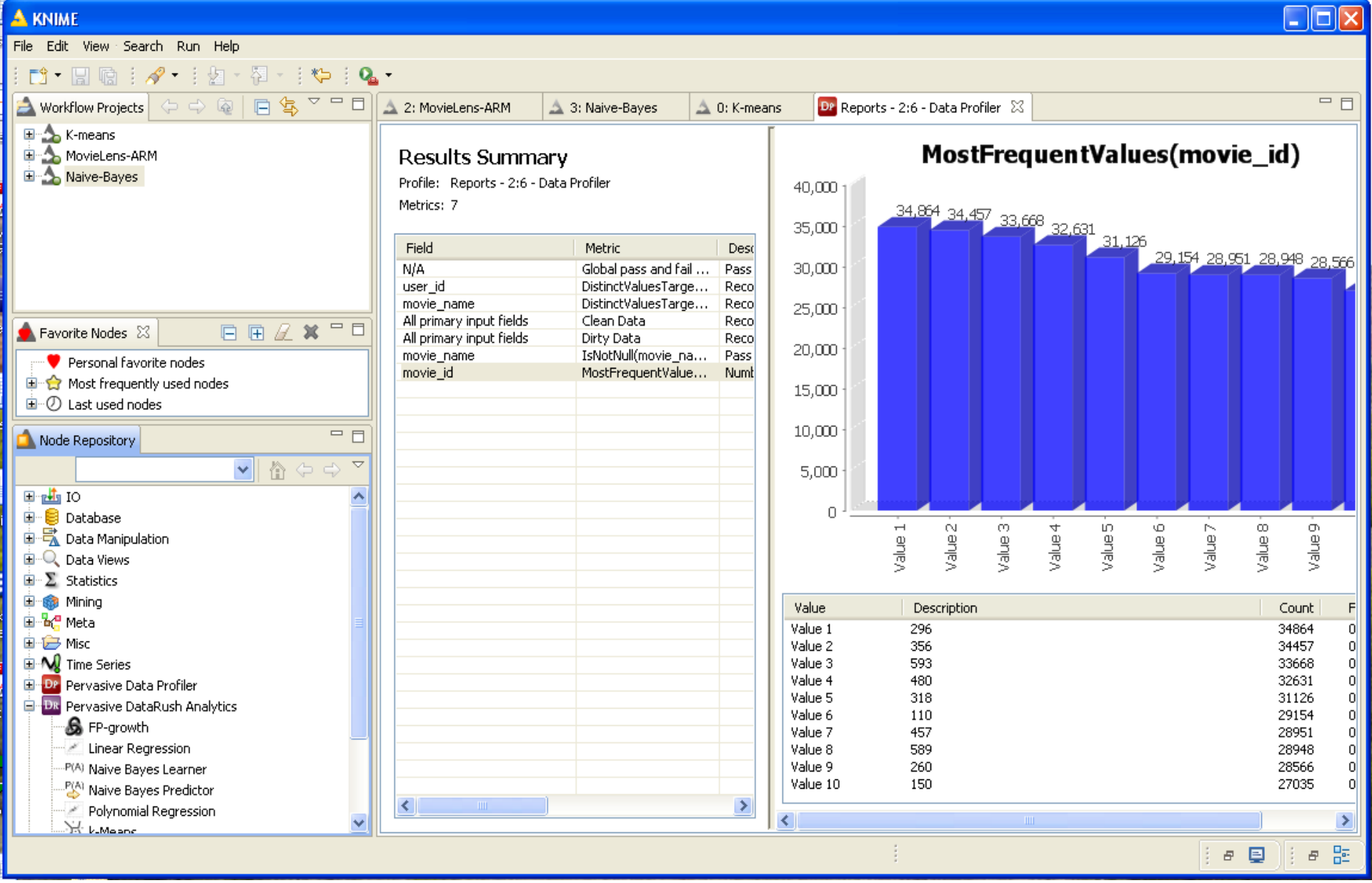
KNIME-DataRush Integration



KNIME & DataRush Integration

- Nodes Implemented
  - *K-means clustering
  - *Naïve Bayes (learner & predictor)
  - *Regression
  - *FP-growth (association rule mining)
  - Data profiler
  - Transformations (lookup, regex)
  - I/O (text, DataRush staging, KNIME converters)
- In work
  - K-nearest neighbors (KNN)
  - Fuzzy matching
  - Collaborative filtering
  - *Decision trees

* **PMML supported**


KNIME
File Edit View Search Run Help
Workflow Projects
K-means
MovieLens-ARM
Naive-Bayes
Favorite Nodes
Personal favorite nodes
Most frequently used nodes
Last used nodes
Node Repository
IO
Database
Data Manipulation
Data Views
Statistics
Mining
Meta
Misc
Time Series
Pervasive Data Profiler
Pervasive DataRush Analytics
FP-growth
Linear Regression
Naive Bayes Learner
Naive Bayes Predictor
Polynomial Regression
k-Means
2: MovieLens-ARM
3: Naive-Bayes
0: K-means
Reports - 2:6 - Data Profiler
Results Summary
Profile: Reports - 2:6 - Data Profiler
Metrics: 7
Field Metric Desc
N/A Global pass and fail ... Pass
user_id DistinctValuesTarge... Reco
movie_name DistinctValuesTarge... Reco
All primary input fields Clean Data Reco
All primary input fields Dirty Data Reco
movie_name IsNotNull(movie_na... Pass
movie_id MostFrequentValue... Numb
MostFrequentValues(movie_id)
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
34,864
34,457
33,668
32,631
31,126
29,154
28,951
28,948
28,566
Value 1
Value 2
Value 3
Value 4
Value 5
Value 6
Value 7
Value 8
Value 9
Value Description Count F
Value 1 296 34864 0
Value 2 356 34457 0
Value 3 593 33668 0
Value 4 480 32631 0
Value 5 318 31126 0
Value 6 110 29154 0
Value 7 457 28951 0
Value 8 589 28948 0
Value 9 260 28566 0
Value 10 150 27035 0

Performance Test

- Naïve-Bayes Use Case:
  - Dataset: 10M rows, 100 columns of strings or doubles (8GB total)
  - Test machine: 24-core box, 4P AMD Opteron 2.6GHZ
  - Windows 2008; 24 GB memory
  - RAID level 0 filesystem

Test	Data Type	DataRush*	Stand-alone*
Naïve-Bayes Learner	Categorical	3.6	436
Naïve-Bayes Predictor	Categorical	7.8	
Naïve-Bayes Learner	Numeric	20	1072
Naïve-Bayes Predictor	Numeric	17	

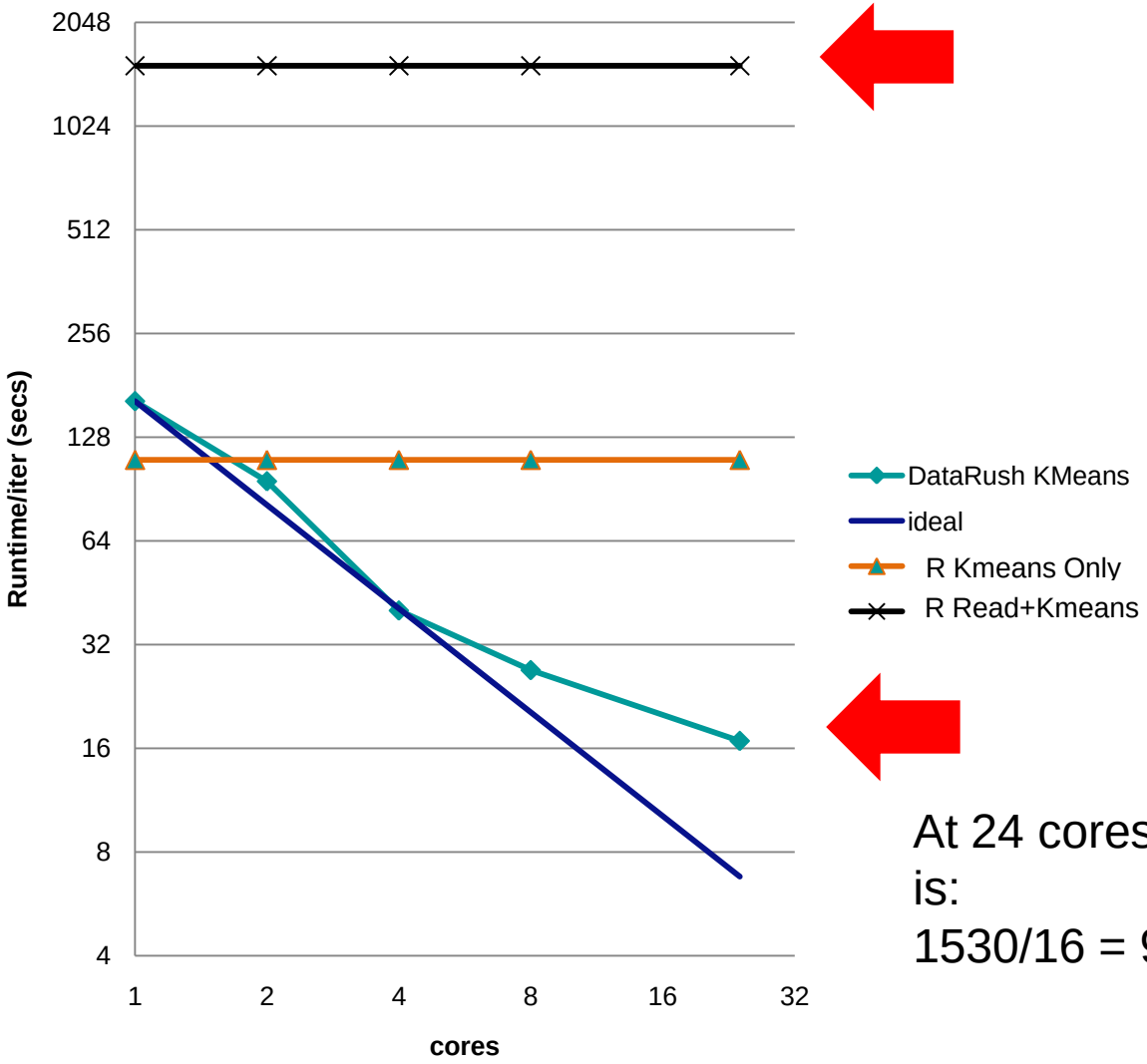
*Runtimes in seconds

Performance Test

- K-means Use Case:
  - Dataset: 400K rows, 1k columns, 3.2GB overall (generated using KNIME cluster generator)
  - Test machine: 24-core box, 4P AMD Opteron 2.6GHZ
  - Windows 2008; 24 GB memory
  - RAID level 0 filesystem

Test	Iterations	DataRush
K-means	20+	212 secs

Kmeans Clustering
100% Netflix Prize dataset



At 24 cores DataRush is:
1530/16 = 91x faster

DataRush Performance

- Reasons for drastic speedups
  - Dataflow architecture
    - High processor core utilization
    - Memory usage patterns optimized to reduce processor cache misses
    - Algorithms re-coded with parallelization in mind
  - Parallelized I/O
    - Native data formats (reduce text conversions)
    - Large, sequential I/O (best fit for today's disks)
    - Encode/decode in parallel

Summary

- Purpose of Integration
  - Parallelize data mining algorithms using DataRush
  - Provide end user experience utilizing KNIME
    - Highly extensible development environment
    - Plug-in capability for deployment
  - Put power of multicore into data miner's hands
- Results
  - Excellent node-level parallelization achieved using DataRush
  - Common data mining algorithms and data transformations fully utilize multicore
  - I/O is also an important factor to performance and scalability

Discussion/Q&A

PERVASIVE® DATARUSH™: UNLEASH THE POWER OF YOUR DATA

1-866-980-RUSH (7874) or +1-512-231-6818
www.pervasivedatarush.com

PERVASIVE