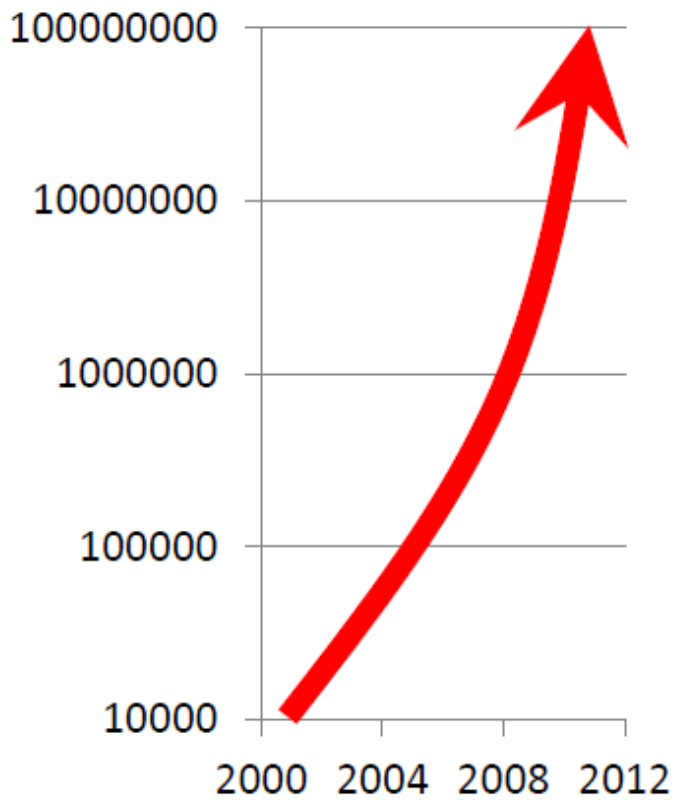
Rule Based Classification on a Multi Node Scalable Hadoop Cluster

Shashank Gugnani
Devavrat Khanolkar
Tushar Bihany
Nikhil Khadilkar

BITS Pilani
K K Birla Goa Campus

Data Hypergrowth

- Reuters-21578: about 10K docs (ModApte)
  - Bekkerman et al, SIGIR 2001
- RCV1: about 807K docs
  - Bekkerman & Scholz, CIKM 2008
- LinkedIn job title data: about 100M docs
  - Bekkerman & Gavish, KDD 2011
- Common Crawl Corpus: 5 Billion docs
  - Common Crawl Foundation, 2014



New Age of Data

- The world has gone mobile
  - 5 billion cellphones produce daily data
- Social networks have gone online
  - Twitter produces 200M tweets a day
- The web is growing
  - 1M new websites created everyday

Source: mediapost.com, bigdatainsightsgroup.com, bbcnews.com

What is MapReduce?

- Data-parallel programming model for clusters of commodity machines
- Pioneered by Google
  - Processes 20 PB of data per day
- Popularized by Apache Hadoop project
  - Used by Yahoo!, Facebook, Amazon, …
- Scalable to large data volumes
  - Scan 100 TB on 1 node @ 50 MB/s = 24 days
  - Scan on 1000-node cluster = 35 minutes

What is MapReduce?

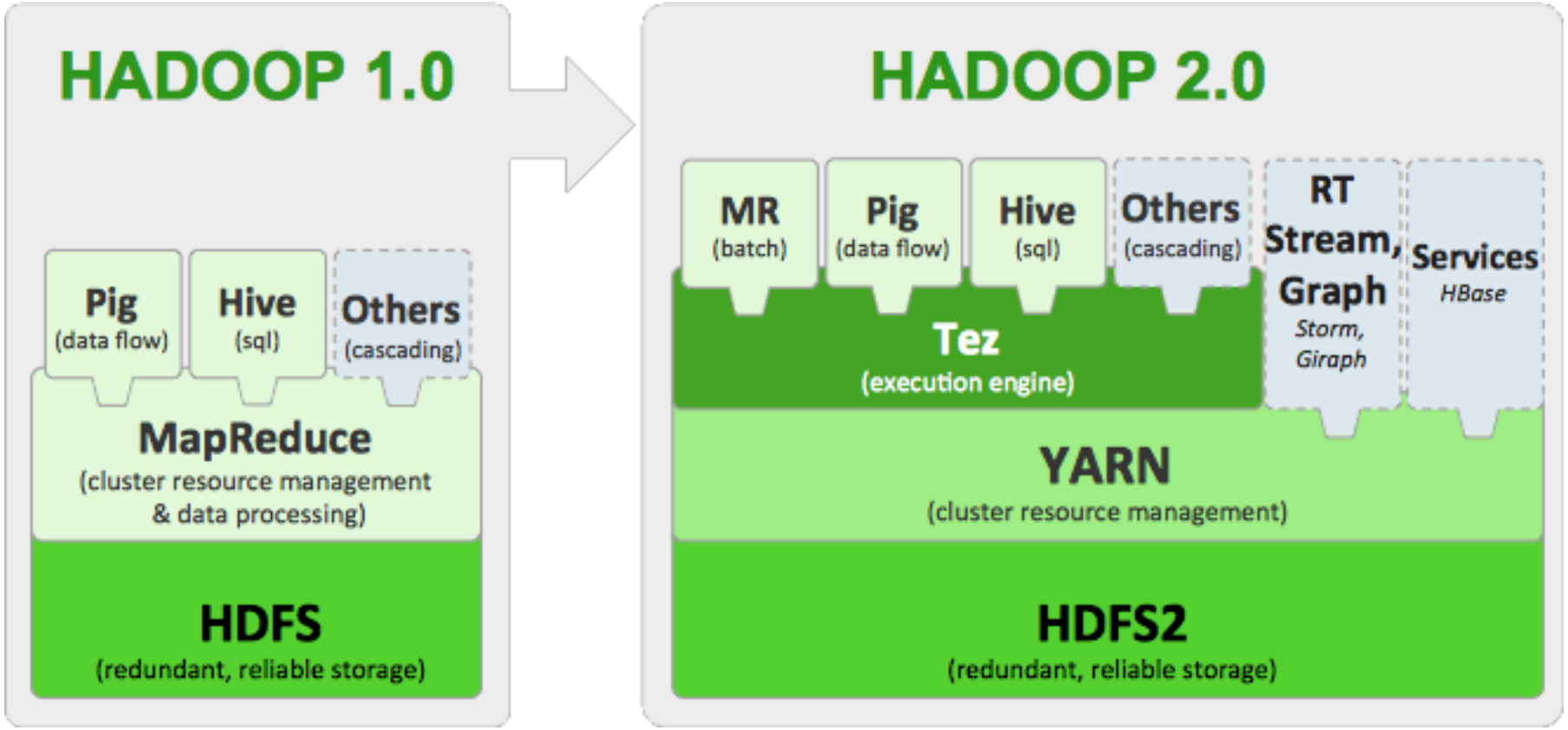
Map function:

$$(K_{in}, V_{in}) \rightarrow list<(K_{inter}, V_{inter})>$$

Reduce function:

$$(K_{inter}, list<V_{inter}>) \rightarrow list<(K_{out}, V_{out})>$$

Hadoop


HADOOP 1.0
Pig
(data flow)
Hive
(sql)
Others
(cascading)
MapReduce
(cluster resource management
& data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MR
(batch)
Pig
(data flow)
Hive
(sql)
Others
(cascading)
Tez
(execution engine)
RT
Stream,
Graph
Storm,
Giraph
Services
HBase
YARN
(cluster resource management)
HDFS2
(redundant, reliable storage)

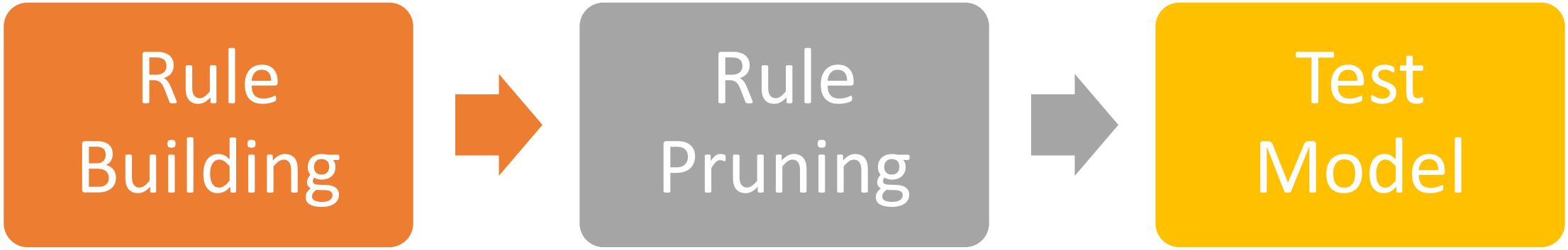
Rule Based Classification

- Classification method in which classifier consists of rules
- Rule: (Condition) → y Where,
  - Condition is a conjunction of attribute tests
  - (A1 = v1) and (A2 = v2) and … and (An = vn)
  - y is the class label
- LHS: rule antecedent or condition
- RHS: rule consequent
- Eg. (Blood Type = warm) ∧ (Lays Eggs = yes) → Birds
- Eg. (Give Birth = no) ∧ (Live in water = yes) → Fishes

RIPPER

- Repeated Incremental Pruning for Error Reduction
- Builds rules by adding attribute tests one by one to condition
- Uses FOIL's information gain to select best attribute test to add
- FOIL's information gain = $p_1 \times (\log p_1/(p_1 + n_1) - \log p_0/(p_0 + n_0))$
- Rules are pruned using pruning metric
- Pruning metric = $(p - n)/(p + n)$

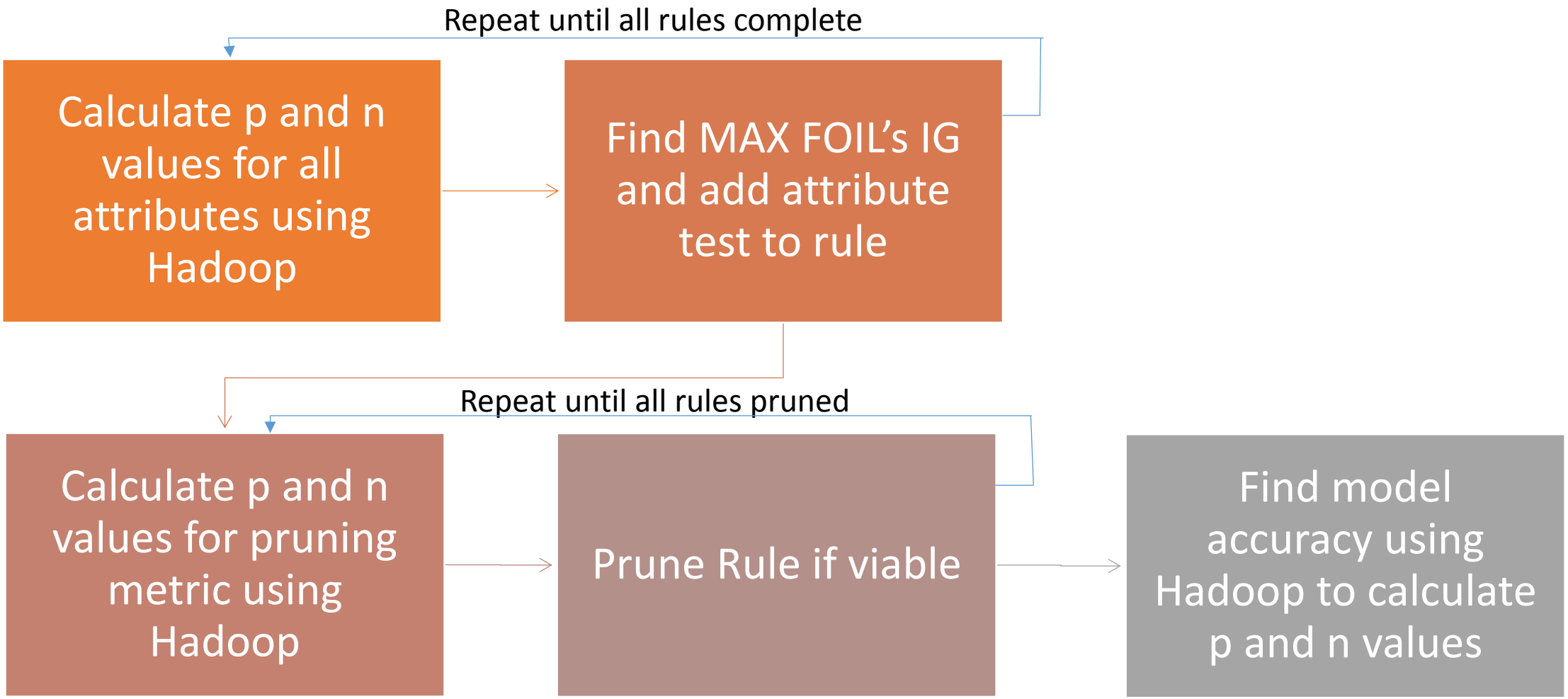
RIPPER



RIPPER with Hadoop

- Each step requires calculation of p and n values which means going over the whole dataset
- Could take a lot of time if dataset large
- Use Hadoop to parallely calculate p and n values
- Use p and n as key values in Map and Reduce functions
- Significant time reduction

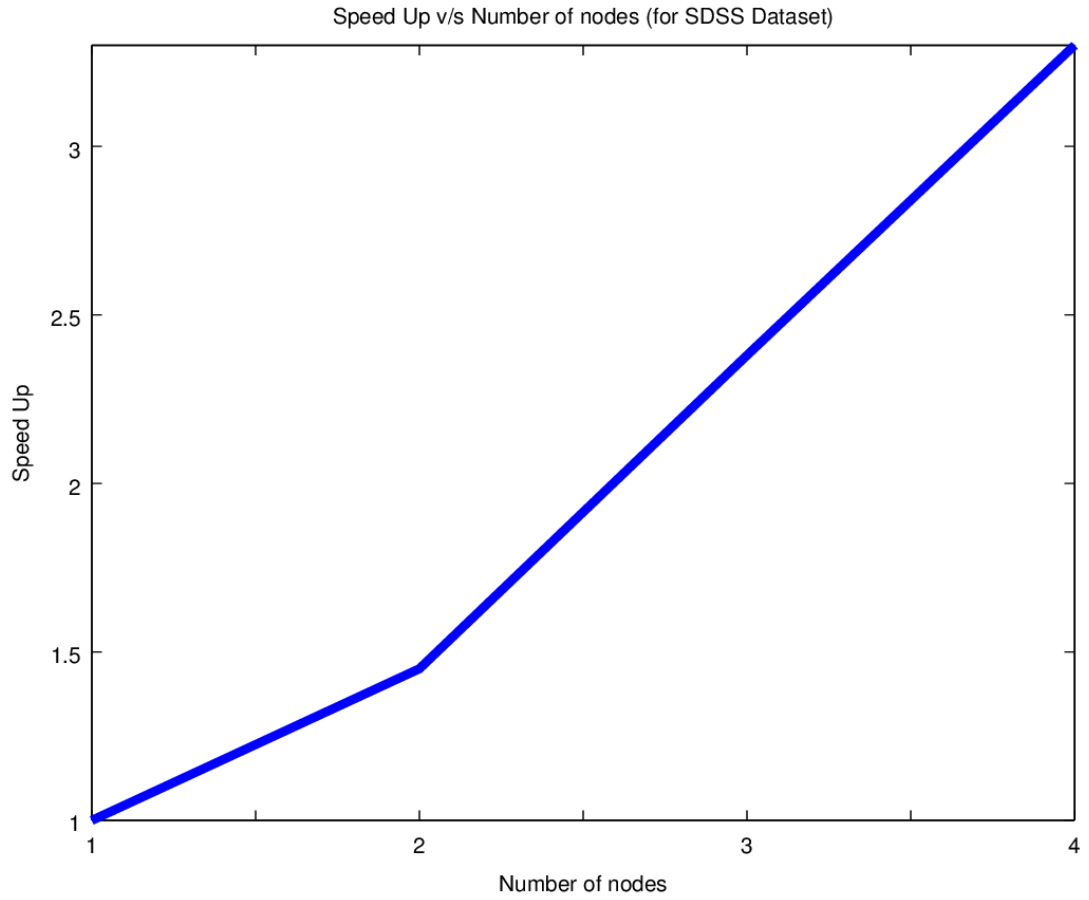
RIPPER with Hadoop



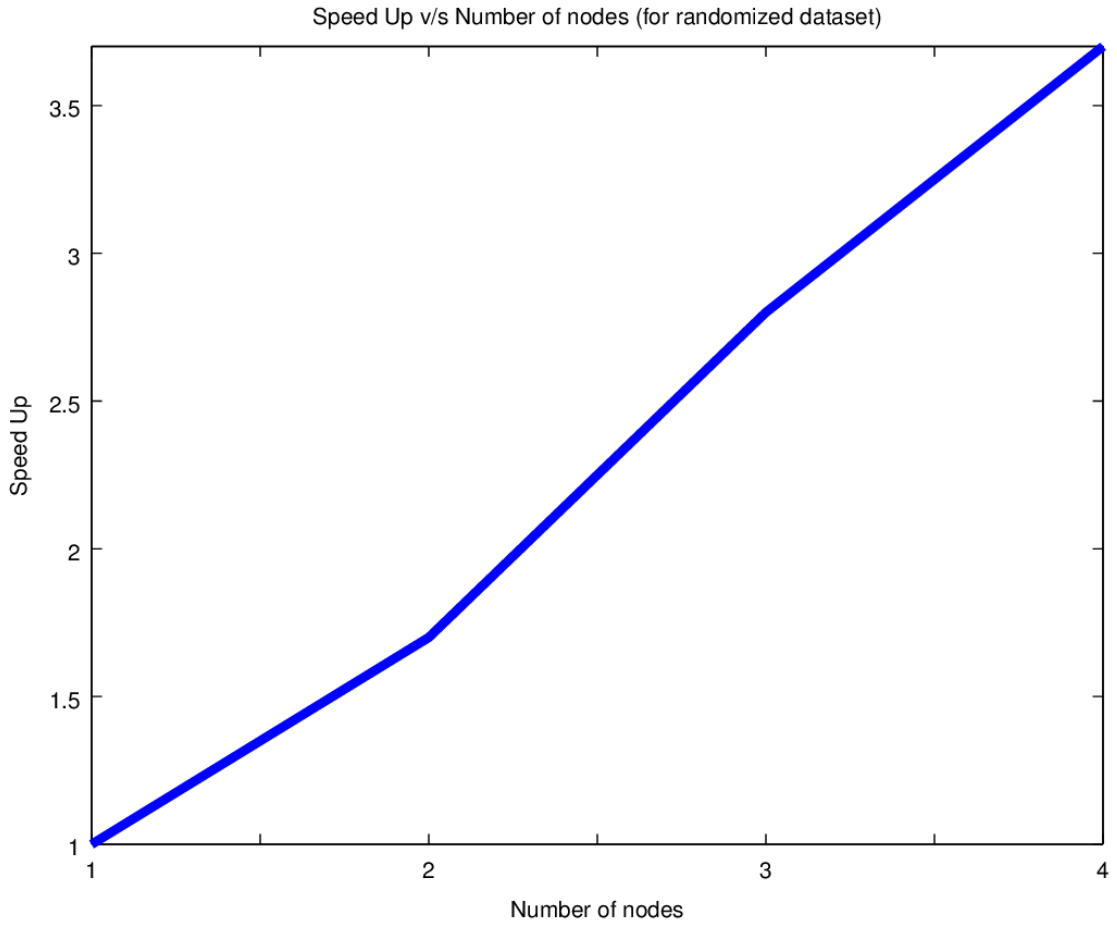
Experiments

- Two Datasets used
  - Randomly generated dataset – 100M Records, 22 Attributes, 2 classes
  - Sloan Digital Sky Survey (SDSS) Dataset – 2.5M Records, 6 Attributes, 2 classes
- Cluster Configuration
  - 4 nodes
  - Hadoop 1.0
  - Gigabit Ethernet
- Experiments run on both datasets
  - Vary number of nodes in cluster
  - Speed up almost linear with number of nodes
  - Algorithm scalable

Results


Speed Up v/s Number of nodes (for SDSS Dataset)
Speed Up
Number of nodes
1
1.5
2
2.5
3
1
2
3
4

Results


Speed Up v/s Number of nodes (for randomized dataset)
Speed Up
Number of nodes
3.5
3
2.5
2
1.5
1
1
2
3
4

References

1. Bekkerman, Ron, et al. "On feature distributional clustering for text categorization." *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001.
2. Bekkerman, Ron, and Martin Scholz. "Data weaving: Scaling up the state-of-the-art in data clustering." *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008.
3. Bekkerman, Ron, and Matan Gavish. "High-precision phrase-based document classification on a modern scale." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011.
4. Apache Hadoop. http://hadoop.apache.org/. Accessed 18/09/2014.
5. Cohen, William W. "Fast Effective Rule Induction." *Proceedings of the Twelfth International Conference on Machine Learning, Lake Tahoe, California*. 1995.
6. Sloan Digital Sky Survey DR 10. http://skyserver.sdss3.org/dr10/en/home.aspx. Accessed 18/09/2014.