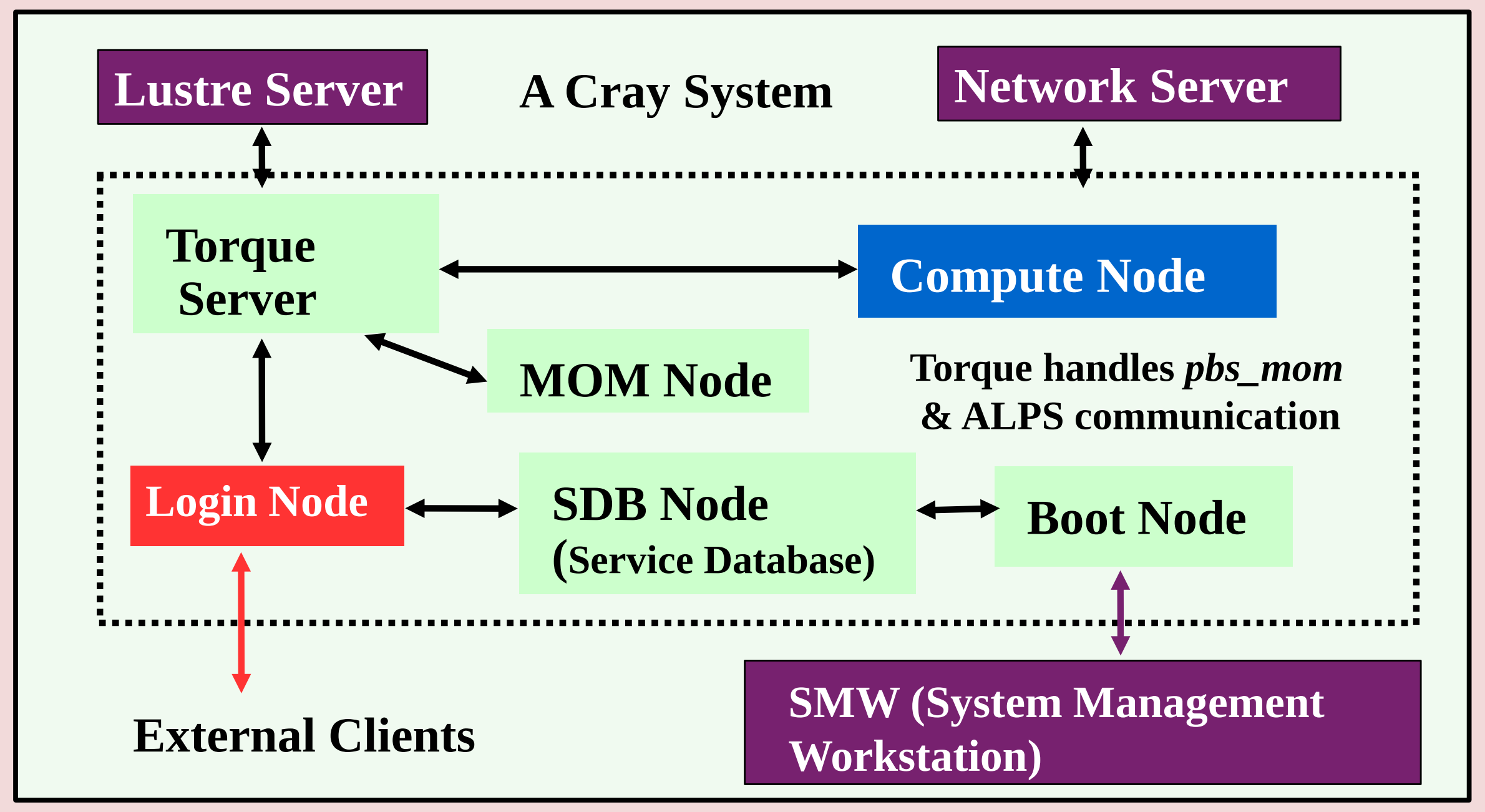
Pin-pointing Node Failures in HPC Systems

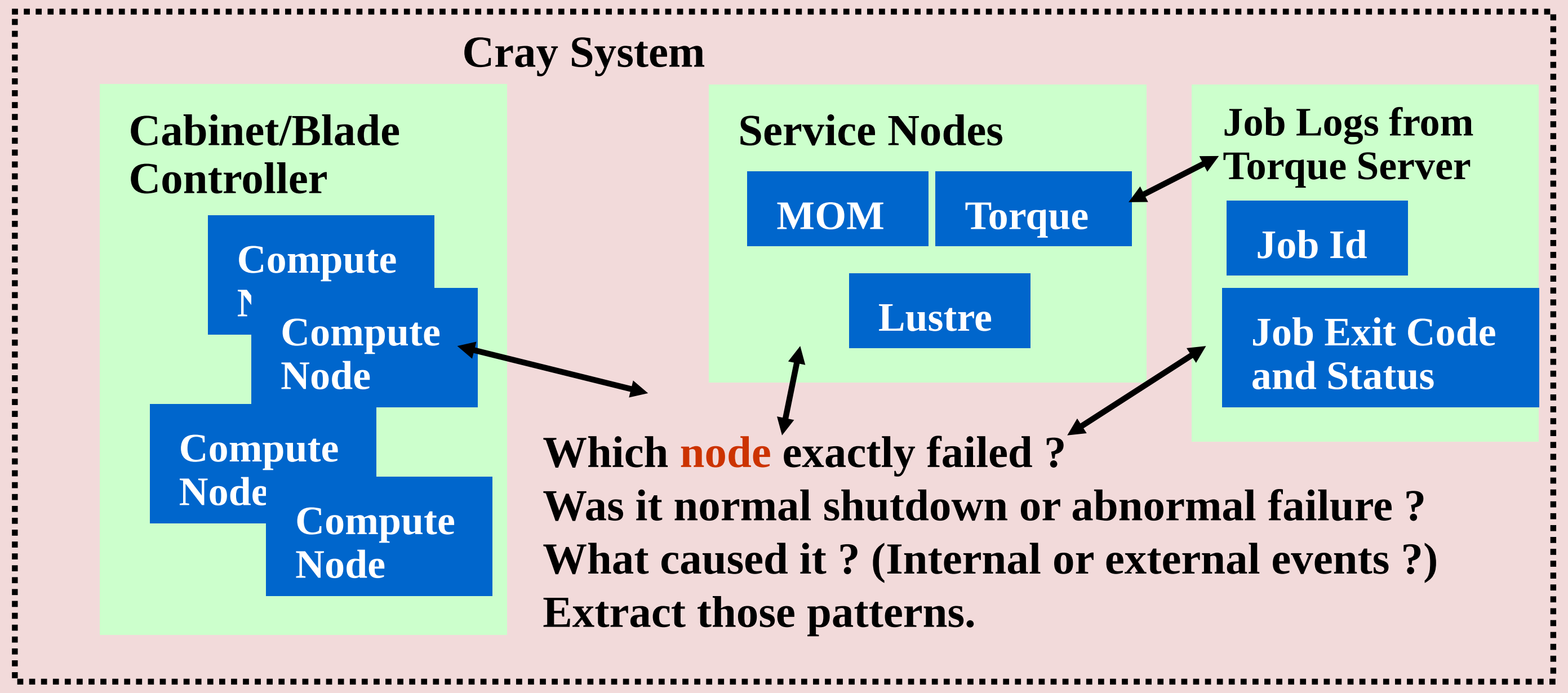
Anwesha Das, Frank Mueller (North Carolina State University)
Paul Hargrove, Eric Roman (Lawrence Berkeley National Lab)

Motivation



*Problems ? Overwhelming raw logs from **several sources**, diverse & complex, Finding **infrequent** node failures is painful, How to detect node failures ?*
Aim - Quantify node failures, Devise a way to automate data processing and node failure identification.

Problem



Which node exactly failed ?
Was it normal shutdown or abnormal failure ?
What caused it ? (Internal or external events ?)
Extract those patterns.

*Challenges ? Detecting faults **independently** without **pin-pointing** node failures is less effective for node resilience, Correlation extraction is hard.*
Goal - Can automated Machine Learning Techniques help us? What features are required to extract node failures? Study logs to extract required patterns.

Contributions

Identification of patterns for indicating node failures distinguishing from mass service shutdown for maintenance based on size and time.

Leveraged TOT - Topics Over Time (continuous time based LDA - Latent Dirichlet Allocation algorithm) to estimate dynamic change in log messages.

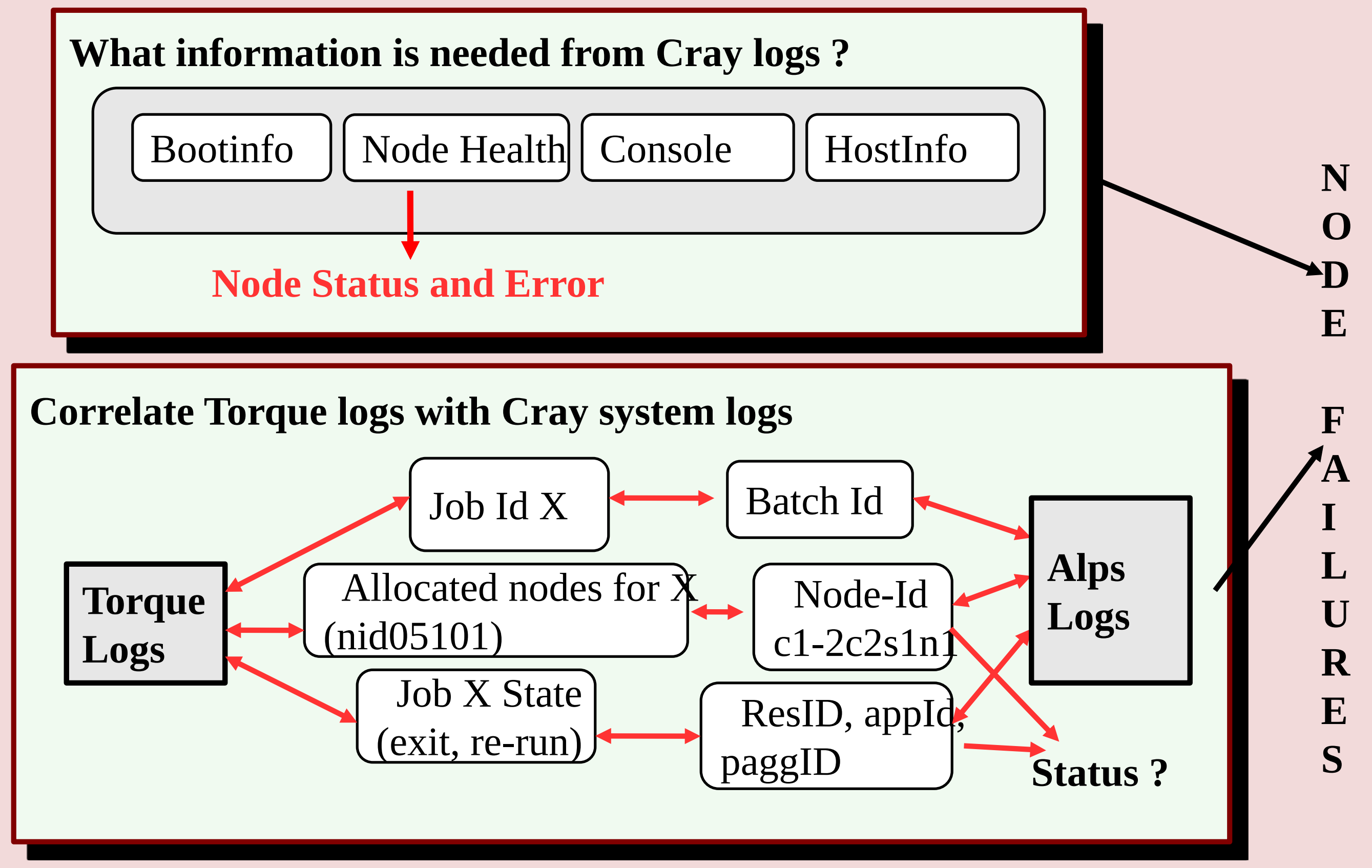
Derivation of ways to correlate Torque (Job) logs and Cray system logs to pin-point node failures.

Table 1: Some Typical Node Failures

NodeId	Node-Type	Error
c0-0c0s0n2	Service	Node BIOS communication error
c4-0c2s0n1	Service	NMI Fault
c1-0c0s1n2	Service	Disk Queue Fatal Error
c2-0c0s7n0	Compute	Lustre Error
c3-0c2s13n3	Compute	LNET Router Error

Solution Approach

- ✓ **Node Status and Events**
- ✓ **Torque (Job) & System Data Correlation**



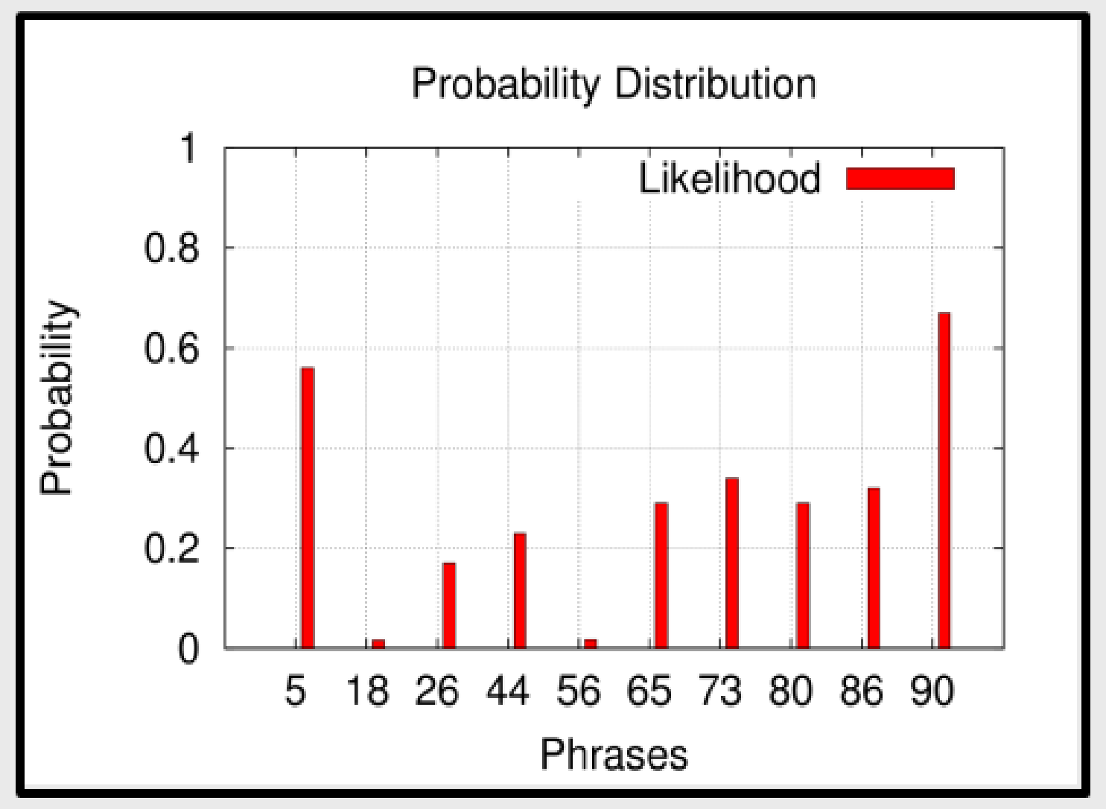
Insights and Findings

- ***Normal node shutdowns very frequent for maintenance (multiple chassis going down in groups of 4) & periodic reboots, compute node failures relatively infrequent compared to service nodes (Job failures & error logs)***
  - ***Service nodes – 12 times a month***
  - ***Compute Nodes – 3 per week***
- ***Some Key phrases of interest for node failures:***
  Failing node c1-0c0s1n2, node_unavailable, node status down, Errors, Fatal, exit codes, allocated nodes for Jobs, etc.
- ***Leverage Job Id & state coupled with node Id & state to correlate Job logs and Cray system logs for pin-pointing node failures.***

(2013-04-26T00:00:41.948135-05:00, AER_BAD_TLP, 0.064),
(2013-04-26T00:00:41.948135-05:00, ec_hw_error, 0.56)

TOT provides the dynamic phrase distribution over continuous time-series data.

Results

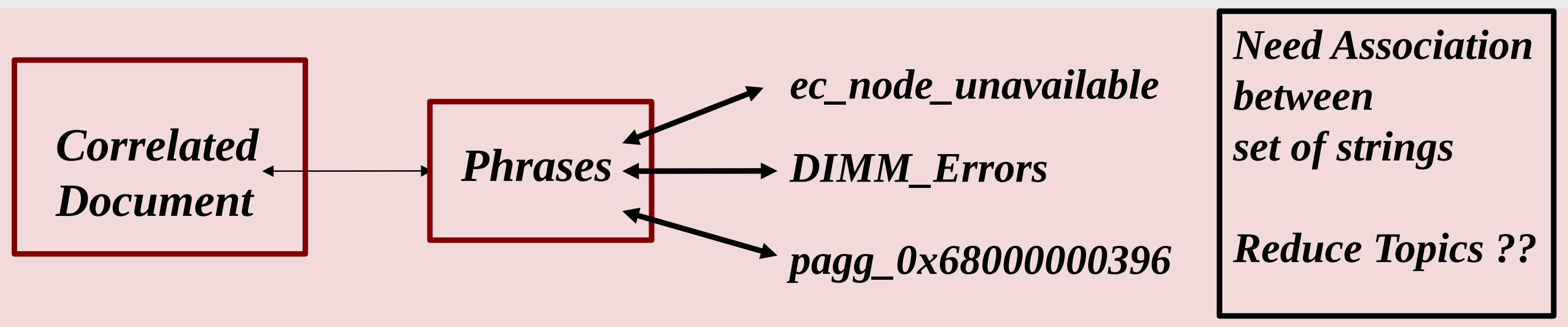


Log Data Details

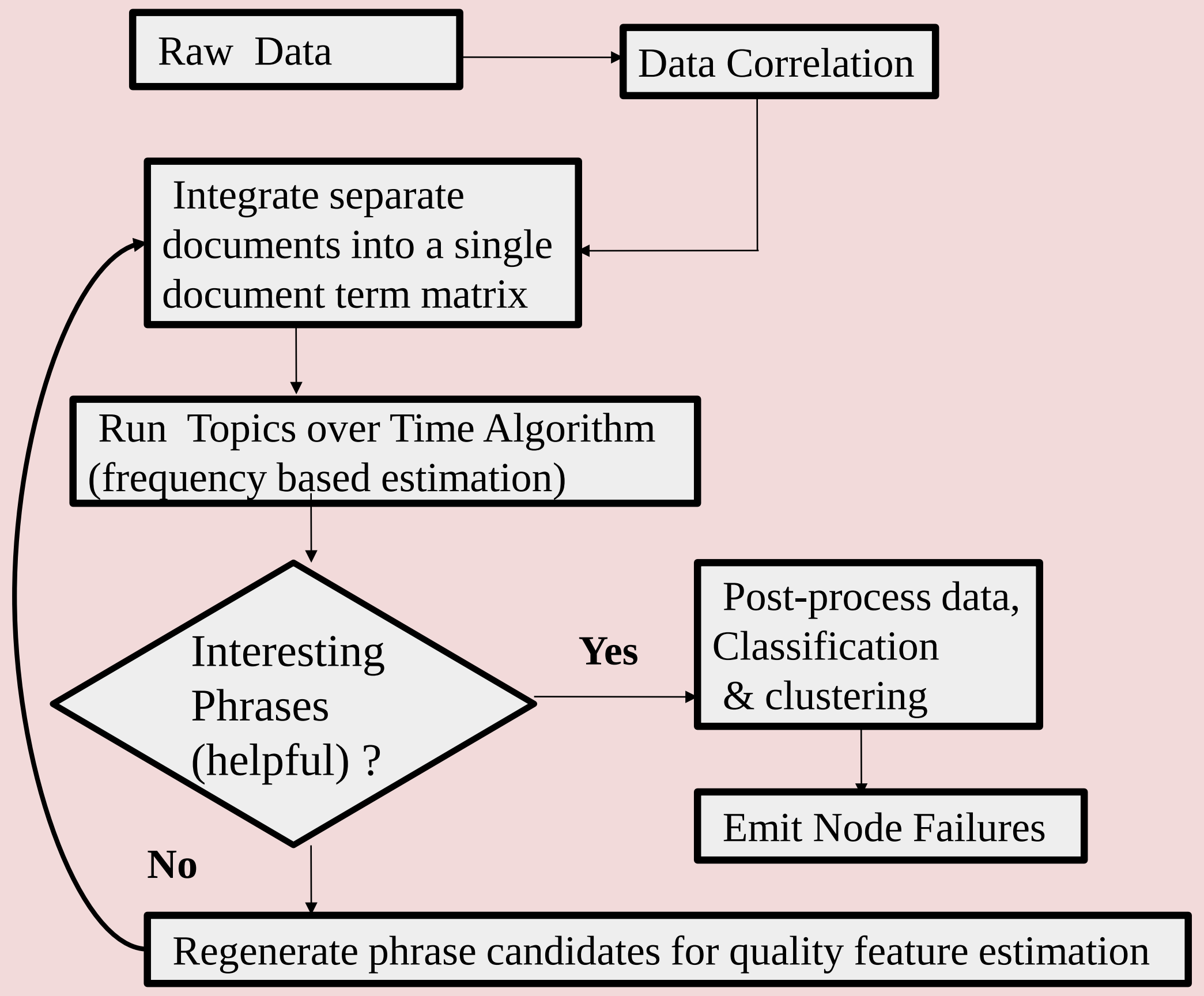
Edison, Hopper, Cori based Cray logs – Approx 3698961 files, more than 600 GB data.

Factorie toolkit, scikit-learn python packages for various libraries.

LogDiver Tool for high-level data analysis.



Overall Methodology


Raw Data
Data Correlation
Integrate separate documents into a single document term matrix
Run Topics over Time Algorithm (frequency based estimation)
Interesting Phrases (helpful) ?
Yes
No
Post-process data, Classification & clustering
Emit Node Failures
Regenerate phrase candidates for quality feature estimation

Node failures

- ✓ **Internal Errors**
  - **Job related**
  - **System/Kernel**
  - **Disk based**
- ✓ **External Errors**
  - **Compute Node OK**
  but something wrong in the Blade or Cabinet
  - **Aries/Gemini**
  Interconnect issues
  - **Routing Errors**
  - **Lustre Errors**
  - **Entire Blade going down**

Conclusion

:

- ***Extracted patterns of distinction between normal mass shutdown versus an infrequent single compute node failure.***
- ***Devised correlation between job logs with system logs.***
- ***Performed continuous time likelihood estimation of topics from the preprocessed document for subsequent outlier detection and prediction.***
- ***Employed the idea of long-term correlation using probability distribution of more likely events.***

Future Work

- ***Investigate unsupervised temporal and spatial log analysis alternatives suitable for failure pattern detection.***
- ***Study of efficient techniques to pre or post process raw data aiding Machine Learning tools for fault extraction.***
- ***Perform data training and testing to validate efficiency.***
- ***Devise ways to predict failures before node goes down.***

Acknowledgments: Dr. Frank Mueller, Dr. Paul Hargrove, Dr. Eric Roman for insightful guidance. Nick Wright, Tina Butler, James Botts and Annette Greiner from NERSC for helping with data and valuable information. This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.