Theia: Visual Signatures for Problem Diagnosis in Large Hadoop Clusters

ELMER GARDUNO, SOILA P. KAVULYA, JIAQI TAN, RAJEEV GANDHI, AND PRIYA NARASIMHAN

Elmer Garduno is the Principal Big Data Architect at UPMC's Technology Development Center. He holds a Master's degree on very large information systems from Carnegie Mellon University. Elmer's main focus is on large scale language technology systems. elmerg@cs.cmu.edu

Soila Kavulya is a PhD student at Carnegie Mellon University. Her research focus is primarily on problem diagnosis in large distributed systems such as Hadoop and Voice over IP systems. She is also interested in fault-tolerance in safety-critical embedded systems, and holds a patent for a fault-tolerant propulsion-by-wire system. spertet@ece.cmu.edu

Jiaqi Tan is a PhD student at Carnegie Mellon University. His research focus is on problem diagnosing in MapReduce and other distributed systems through log-analysis and visualization. He is also interested in program-analysis techniques for understanding software and ensuring security. tanjiaqi@cmu.edu

Rajeev Gandhi is a Senior Systems Scientist at Carnegie Mellon University. His research focuses on problem diagnosis in large-scale distributed systems and video streaming. Rajeev was previously involved in the H.264 video-compression standardization and received a Motorola Outstanding Performance award in 2002 in recognition of his contributions to global standardization. In 2000, Rajeev received his PhD from the University of California, Santa Barbara. rgandhi@ece.cmu.edu

Priya Narasimhan is an Associate Professor at Carnegie Mellon University. Her research interests lie in the fields of dependable distributed systems, embedded systems, mobile systems, and sports technology. Priya is also the CEO and Founder of YinzCam, Inc., a Carnegie Mellon spin-off company focused on mobile live streaming and scalable video technologies to provide sports fans with the ultimate in-stadium mobile experience at NHL/NFL/NBA games. priya@cs.cmu.edu

Visualization tools play an important role in summarizing large volumes of data by revealing interesting patterns such as trends, gaps, and anomalies in the data. Users can leverage visualization tools to identify problems in their programs quickly. In this article, we present novel visualizations that help users diagnose problems in Hadoop applications. These visualizations allow users to identify problematic nodes in the cluster quickly, and distinguish between different classes of problems.

Hadoop is a popular open source system that simplifies large-scale data analysis. Large companies like Twitter use Hadoop to store and process tweets and log files [6]. A typical Hadoop deployment can consist of tens to thousands of nodes. Manual diagnosis of performance problems in a Hadoop cluster requires users to comb through the logs on each node—a daunting task, even on clusters consisting of tens of nodes. We have developed a visualization tool, Theia (named after the Greek goddess of light), that helps users distinguish between application-level problems (e.g., software bugs, workload imbalances),, which they can fix on their own, and infrastructural problems (e.g., contention problems, hardware problems),, which they should escalate to the system administrators.

Each Hadoop job consists of a group of Map and Reduce tasks. Map tasks process smaller chunks of the large data set in parallel and use key/value pairs to generate a set of intermediate results, while Reduce tasks merge all intermediate values associated with the same intermediate key. Theia leverages application-specific knowledge about how MapReduce jobs are structured to generate compact, interactive visualizations of job performance. Theia generates three different types of visualizations: one at the cluster-level that represents the performance of jobs across nodes over time, and two others at the job-level that summarize task performance across nodes in terms of task duration, task status, and volume of data processed.

We describe how Theia works, and use actual problems from a production Hadoop cluster to illustrate how our visualizations can provide users with a better understanding of the performance of their jobs and easily spot anomalous nodes.

Generating Visual Signatures of Hadoop Job Performance

We implemented Theia using a Perl script that gathered data about job execution from the job-history logs generated by Hadoop. These logs store information about the Map and Reduce tasks executed by each job, such as task duration, status, and the volume of data read and written. We store this information in a relational database, and generate visualizations in the Web browser using the D3 framework [1].

	Visual Signatures of Problem Classes		
	Application problem	**Workload imbalance**	**Infrastructural problem**
Time	*Single* user or job over time	*Single* user or job over time	*Multiple* users and jobs over time
Space	Span *multiple* nodes	Span *multiple* nodes	Typically affect *single* node, but correlated failures also occur
Value	Performance degradations and *task exceptions*	Performance degradation and *data skews*	Performance degradations and *task exceptions*

Table 1: Heuristics for developing visual signatures of problems experienced in a Hadoop cluster

We developed visual signatures that allow users to spot anomalies in job performance by identifying visual patterns (or signatures) of problems across the time, space, and value domain. Table 1 summarizes the heuristics that we used to develop visual signatures that distinguish between application-level problems, workload imbalances between tasks from the same job, and infrastructural problems. These heuristics are explained below:

1. *Time dimension.* Different problems manifest in different ways over time. For example, application-level problems and workload imbalances are specific to an application; therefore, the manifestation of a problem is restricted to a single user or job over time. On the other hand, infrastructural problems, such as hardware failures, affect multiple users and jobs running on the affected nodes over time.
2. *Space dimension.* The space dimension captures the manifestation of the problem across multiple nodes. Application-level problems and workload imbalances associated with a single job manifest across multiple nodes running the buggy or misconfigured code. Infrastructural problems are typically limited to a single node in the cluster. However, a study of a globally distributed storage system [2] shows that correlated failures are not rare, and were responsible for approximately 37% of failures. Therefore, infrastructural problems can also span multiple nodes.
3. *Value dimension.* We quantify anomalies in the value domain by capturing the extent of performance degradation, data skew, and task exceptions experienced by a single job. Application-level and infrastructural problems manifest as either performance degradations or task exceptions. Workload imbalances in Hadoop clusters can stem from skewed data distributions that lead to performance degradations.

Detecting Anomalous Nodes

We detect anomalies by first assuming that under fault-free conditions, the workload in a Hadoop cluster is relatively well-balanced across nodes executing the same job—therefore, these nodes are peers and should exhibit similar behavior [4]. Next, we identify nodes whose task executions differ markedly from their peers and flag them as anomalous. Aggregating task behavior on a per-node basis allows us to build compact signatures of job behavior because the number of nodes in the cluster can be several orders of magnitudes smaller than the maximum number of tasks in a job.

We compute an anomaly score using a simple statistical measure known as the z-score. The z-score is a dimensionless quantity that indicates how much each value deviates from the mean in terms of standard deviations, and is computed using the following formula: $z = [(x - \mu)/(\sigma)]$, where μ is the mean of the values, and σ is the corresponding standard deviation. We compute z-scores for each node based on the duration of tasks running on the node, the volume of data processed by the node, and the ratio of failed tasks to successful tasks. For the cluster-level visualization, we estimate the severity of problems by using a single anomaly score that flags nodes as anomalous if the geometric mean of the absolute value of the z-scores is high, i.e., Anomaly-Score=(|z_task_duration|*|z_data_volume|*|z_failure_ratio|) ^ (1/3).

Visualizations and Case Studies

Theia generates three different visualizations that allow users to understand the performance of their jobs across nodes in the cluster. The first visualization is the anomaly heatmap, which summarizes job behavior at the cluster-level; the other two visualizations are at the job-level. The first job-level visualization, referred to as the job execution stream, allows users to scroll through jobs sequentially, thus preserving the time context. The second job-level visualization, referred to as the job-execution detail, provides a more detailed view of task execution over time on each node in terms of task duration and amount of data processed. We analyzed the jobs and problems experienced by Hadoop users of the 64-node OpenCloud cluster for data-intensive research [5] over the course of one month. We use actual problems experienced by users of the cluster to illustrate our visualizations.

Anomaly Heatmap

A heatmap is a high-density representation of a matrix that we use to provide users with a high-level overview of jobs execution at the cluster-level. This visualization is formulated over a grid that shows nodes on the rows and jobs on the columns, as shown

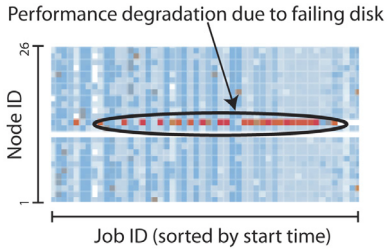

Figure 1: Visual signature of an infrastructural problem using an anomaly heatmap shows succession of anomalous jobs (darker color/red) due to a failing disk controller on a node.

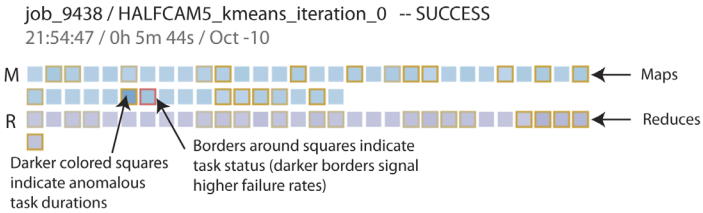

Figure 2: The job execution stream visualization compactly displays information about a job's execution. The header lists the job ID, name, status, time, duration and date. The visualization also highlights anomalies in task duration by using darker colors, and task status by using yellow borders for killed tasks and red borders for failed tasks. The nodes are sorted by decreasing amount of I/O processed.

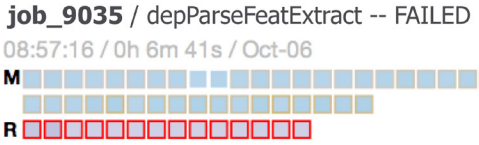

Figure 3: Visual signature of bugs in the Reduce phase. Failures spread across all Reduce nodes (solid dark/red border) typically signal a bug in the Reduce phase.

in Figure 1. The darkness of an intersection on the grid indicates a higher degree of anomaly on that node for that job. By using this visualization, anomalies due to application-level and infrastructural problems can be spotted easily as bursts of color that contrast with non-faulty nodes and jobs in the background.

Figure 1 displays the visual signature of an infrastructural problem identified by a succession of anomalous jobs (darker color) due to a failing disk controller on a node. The data density of the anomaly heatmap is around 2,900 features per square inch on a 109 ppi (pixels per inch) display, using 2x2 pixels per job/node, which is equivalent to fitting 1200 jobs x 700 nodes on a 27-inch display.

Job Execution Stream

The job execution stream, shown in Figure 2, provides a more detailed view of jobs while preserving information about the context by showing a scrollable stream of jobs sorted by start time. In addition to displaying general information about the job (job ID, job name, start date and time, job duration) in the header, this visualization uses variations in color to highlight anomalous nodes.

Figure 4: Visual signature of data skew. A node with anomalous task durations (darker color) and high volume of I/O (nodes are sorted by decreasing order of I/O) can signal data skew.

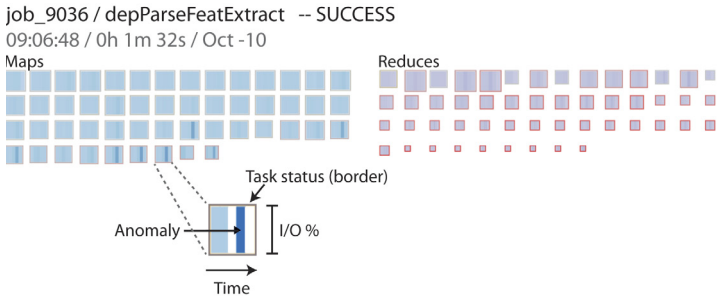

Figure 5: The job execution detail visualization highlights both the progress of tasks over time and the volume of data processed. Each node is divided into five stripes that represent the degree of anomalies in tasks executing during the corresponding time slot; the size of the square represents the proportion of I/O processed by that node.

Because the application-semantics of Map and Reduce tasks are very different, we divided nodes into two sets: the Map set and the Reduce set. We enhanced the representation of each node with a colored border that varies in intensity depending on the ratio of failed tasks to successful tasks, or the ratio of killed tasks to successful tasks; killed tasks arise when the task scheduler terminates speculative tasks that are still running after the fastest copy of the task completed. Killed tasks are represented using a yellow border, which is overloaded by a red border if there are any failed tasks.

The job execution stream visualization allows us to generate signatures for application-level problems, which manifest as a large number of failed tasks across all nodes in either the Map or Reduce phase (see Figure 3). Workload imbalances and infrastructural problems tend to manifest on a single or small set of nodes in the system. For example, the dark left-most node in Figure 4 shows a node whose performance is slower than its peers due to data skew.

Job Execution Detail

The job execution detail visualization provides a more detailed view of task execution and is less compact than the job execution stream. The job execution detail visually highlights both the progress of tasks over time and the volume of data processed as shown in Figure 5. Nodes are still represented as two sets of squares for Map and Reduce tasks; however, given that there is

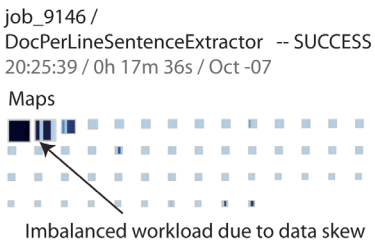

Figure 6: Visual signature of data skew. A single node with high duration anomaly (darker color) and high amount of I/O (larger size) can signal data skew.

additional space available because we are only visualizing one job at a time, we use the available area on each of the squares to represent two additional variables: (1) anomalies in task durations over time by dividing the area of each node into five vertical stripes, each corresponding to a fifth of the total time spent executing tasks on that node; darker colors indicate the severity of the anomaly while white stripes represent slots of time where no information was processed; and (2) percentage of total I/O processed by that node, i.e., reads and writes to both the local file system and the Hadoop distributed file system (HDFS); larger squares indicate higher volumes of data.

Figure 6 shows the visual signature of a data skew where a subset of nodes with anomalous task durations (darker color) and high amounts of I/O (first nodes in the list, large square size) indicate data skew. In this visualization, the data skew is more obvious to the user when compared to the same problem visualized using the job execution stream in Figure 4. Infrastructural problems such as the failed NIC (network interface controller) in Figure 7 can be identified as a single node with high task durations (darker color) or failed tasks (red border), coupled with a low volume of I/O (small square size), which might indicate a performance degradation.

All of our visualizations are interactive, and they provide access to additional information by using the mouse-over gesture. By hovering over the failed node in Figure 7, a user can obtain additional information about the behavior of tasks executed on that node. For example, a user can observe that 50% of the tasks executed on this node failed.

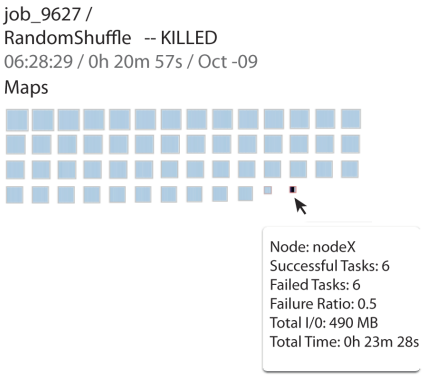

Figure 7: Interactive user interface. This job execution detail visualization shows degraded job performance due to a NIC failure at a node. Hovering over the node provides the user with additional information about the behavior of tasks executed on that node.

Evaluating the Effectiveness of Visualizations

We generated our visualizations using one month's worth of logs generated by Hadoop's JobTracker on the OpenCloud cluster. During this period, 1,373 jobs where submitted, comprising a total of approximately 1.85 million tasks. From these 1,373 jobs, we manually identified 157 failures due to application-level problems, and two incidents of data skew. We also identified infrastructural problems by analyzing a report of events generated by the Nagios tool installed on the cluster. During the evaluation period, Nagios reported 68 messages that were associated with 45 different incidents, namely: 42 disk controller failures, two hard drive failures, and one network interface controller (NIC) failure.

We evaluated the performance of Theia by manually verifying that the visualizations generated matched up with the heuristics for distinguishing between different problems described in Table 1. Table 2 shows that we successfully identified all the application-level problems and data skews using the job execution stream (similar results are obtained using the job execution detail). Additionally, the anomaly heatmap was able to identify 33 of the 45 infrastructural problems (the problems identified by the job execution stream are a subset of those identified by the heatmap). We were unable to detect four of the infrastructural problems because the nodes had been blacklisted. We hypothesize that the heatmap was unable to detect the remaining

Type	Total problems	Diagnosed by heatmap	Diagnosed by job execution stream
Application-level problem	157	0	157
Data-skew	2	2	2
Infrastructural problem	45	33	10

Table 2: Problems diagnosed by cluster-level and job-level visualizations in Theia. The infrastructural problems consisted of 42 disk controller failures, two hard drive failures, and one network interface controller (NIC) failure. The infrastructural problems diagnosed by the job execution stream were a subset of those identified by the heatmap.

eight infrastructural problems because they occurred when the cluster was idle.

Conclusion

Theia is a visualization tool that exploits application-specific semantics about the structure of MapReduce jobs to generate compact, interactive visualizations of job performance. Theia uses heuristics to identify visual signatures of problems that allow users to distinguish application-level problems (e.g., software bugs, workload imbalances) from infrastructural problems (e.g., contention problems, hardware problems). We have evaluated our visualizations using real problems experienced by Hadoop users at a production cluster over a one-month period. Our visualizations correctly identified 192 out of 204 problems that we observed. More details about Theia can be found in our USENIX LISA 2012 paper [3].

References

[1] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics,* vol. 17, no. 12, 2011, pp. 2301-2309.

[2] D. Ford, F. Labelle, F.I. Popovici, M. Stokely, V.-A. Truong, .L. Barroso, C. Grimes, and S. Quinlan, "Availability in Globally Distributed Storage Systems," *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation* (OSDI) (Vancouver, CA, October 2010), pp. 61-74.

[3] E. Garduno, S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, "Theia: Visual Signatures for Problem Diagnosis in Large Hadoop Clusters," USENIX Large Installation System Administration Conference (LISA) (San Diego, CA, December 2012).

[4] X. Pan, J. Tan, S. Kavulya, R. Gandhi, and P. Narasimhan, "Ganesha: Black-Box Diagnosis of MapReduce Systems," *SIGMETRICS Performance Evaluation Review* 37 (January 2010).

[5] Parallel Data Lab. Apache Hadoop: http://wiki.pdl.cmu.edu/opencloudwiki/, September 2012.

[6] The Apache Software Foundation, PoweredBy Hadoop: http://wiki.apache.org/hadoop/PoweredBy, September 2012.

Join other members of the sysadmin community looking to stay ahead in this dynamic industry. Become a part of the SIG today.

Are you getting the most out of the sysadmin community?

We can help . . .

Created by and for system administrators, this USENIX SIG serves the system administration community by:

- Offering conferences and training to enhance the technical and managerial capabilities of members of the profession
- Promoting activities that advance the state of the art or the community
- Providing tools, information, and services to assist system administrators and their organizations
- Establishing standards of professional excellence and recognizing those who attain them

LISA members enjoy a number of benefits, including:

- Discount on registration for LISA, the annual Large Installation System Administration Conference
- Access to the large and growing online library of Short Topics in System Administration books—see reverse for the complete catalog
- A free Short Topics in System Administration book every year: Your choice of any book in print
- Access to the LISA SIG Jobs Board, including real-time email notification of new jobs posted and the ability to post resumes
- The option to join LISA-members, an electronic mailing list for peer discussion and advice
- Student programs, including grants to help fund conference attendance, low membership fees, and a university outreach program
- And more!

Find out more at www.usenix.org/lisa