THE CHALLENGE

Infrastructure and capacity managers are constantly using a variety of different tools to understand the performance of clusters to manage and monitor resource capacity, improve efficiency and productivity, maximize existing infrastructure investments, and forecast resource needs. Similarly, developers need to understand the performance of their applications to ensure SLAs, avoid failure, improve efficiency, and monitor resource capacity. But nobody has an accurate picture because without Pepperdata, they don’t have access to the right data.

Pepperdata continuously collects extensive data on over 350 real-time metrics from all of your applications and infrastructure resources — metrics about CPU, RAM, disk I/O, and network usage for every job, task, user, host, workflow, and queue. This data — not available with any other tool — empowers developers and capacity and infrastructure managers to understand and improve application performance, reduce mean time to problem resolution and increase capacity utilization by 30-50% without adding new hardware. In addition to surfacing performance bottlenecks, Pepperdata also provides automatic tuning for recurring applications, delivers app-specific recommendations and allows users to set up alerts on specific behaviors and outcomes to avoid the risk of failure.

BENEFITS

Reduce time to problem resolution using comprehensive and detailed performance data.

With its vast collection of real-time hardware metrics, Pepperdata allows developers and operations staff to create custom alerts and queries and filter and compare by different dimensions (application name, job group, queue, etc.) to quickly identify which app is causing a problem and which user submitted it.

Increase capacity utilization by 30-50% without adding hardware.

Pepperdata tracks actual hardware resource usage of every node on the cluster in real time and allows more tasks to be run on nodes that have available free resources at the moment. This provides an immediate boost in the number of concurrently-running containers and overall throughput, thereby increasing capacity without adding hardware.

Advises with application and platform understanding and improvement.

Pepperdata generates targeted recommendations to improve application performance, highlights applications that require attention, automatically identifies bottlenecks, and alerts on duration, failure conditions, and resource usage.

Advises you make informed capacity and resource decisions.

Pepperdata generates comprehensive planning reports containing real-time and historical data about the cluster so you can plan capacity for predictable performance. Use these reports to identify wasteful system resource usage to attribute accurate costs to users and business units.
FOR INFRASTRUCTURE AND CAPACITY MANAGERS

PLATFORM SPOTLIGHT

You’re constantly using a variety of different tools to understand the performance of your clusters in order to manage and monitor resource capacity, improve efficiency and productivity, maximize your existing infrastructure investment, and forecast resource needs, but it’s impossible to have an accurate view without access to the right data. Your operators are challenged with configuring and sizing critical resources running on multi-tenant clusters with mixed workloads, and you receive alerts without enough specific detail to isolate and resolve problems. Maintaining performant clusters managing capacity successfully requires an understanding of dozens of performance metrics and tuning parameters.

Pepperdata Platform Spotlight continuously collects extensive unique data—that nobody else collects—about your hosts, queues, users, applications and all relevant resources, providing you with a 360° cluster view that enables you to quickly diagnose performance issues and make resource decisions based on user priorities and needs. Platform Spotlight leverages AI-driven resource management to automatically tune clusters and recapture wasted capacity, allows you to create alerts to identify root causes, and provides recommendations to rightsize containers, queues and other resources.

HOW PLATFORM SPOTLIGHT CAPACITY OPTIMIZER AUTOMATICALLY INCREASES CLUSTER THROUGHPUT BY 30-50%

Operators who spend significant time tuning their Hadoop deployments may be skeptical of Capacity Optimizer’s ability to automatically increase cluster throughput by 30-50%. How does Pepperdata’s unique capacity optimization improve performance on a cluster that has been tuned using industry-standard best practices? Platform Spotlight uses Capacity Optimizer to identify “holes” where a node can temporarily do more work, and fills those holes with additional tasks, all the while ensuring cluster reliability and safety. Capacity Optimizer automatically monitors and adjusts hardware resource usage at the process level in real time. On a typical cluster, Capacity Optimizer makes hundreds or thousands of decisions per second. Even if Hadoop provided a mechanism to do so, the most talented dedicated operator or outside consultant cannot make manual configuration changes with the precision and speed of Capacity Optimizer.

By contrast, standard Hadoop configurations affect only up-front static resource reservations, so Hadoop must assume peak resource usage by every task, which typically wastes a significant amount of the cluster’s hardware resources. Additionally, YARN cannot engage in active resource management after container launch, except to kill jobs under certain conditions.

Platform Spotlight.
**360° PLATFORM VIEW**

Create a holistic source of operational and performance truth across your clusters.

**A single source of operational and performance truth.**
Access real-time and historical data about the cluster, including system demand, abusive users, and wasteful applications, as well as queue and container sizes. Drill down or zoom out to analyze any application to understand its performance in the context of the entire multi-tenant cluster.

**Complete instrumentation of your Big Data ecosystem.**
Monitor any process including Kafka, Impala, HBase, Hive, MapReduce, Tez, Spark, IBM BigSQL, and LLAP.

**Quantify the impact applications have on the cluster.**
Identify the impact that applications have on cluster-wide resources such as CPU, disk, memory, network, NameNode and HDFS.

**Quantify the impact the cluster has on applications**
Identify which application slowdowns were caused by disk, network or node failures/congestion.

**Attribute accurate costs to users and business units.**
Pepperdata provides you with detailed usage data of every resource, which when combined with your cost model, generates the most accurate reports to identify the most costly users and business units.

---

**PLATFORM TUNING**

*Improve the performance and efficiency of your cluster.*

**Maximize your existing Infrastructure Investment.**
Increase hardware utilization and eliminate or reduce millions of dollars in hardware costs.

**Auto-scale cluster resources for peak efficiency.**
Rightsize cluster resource allocation based on real-time capacity to run more jobs faster. Using active resource management, Pepperdata continuously and dynamically eliminates inefficiencies and bottlenecks without manual intervention for job or cluster tuning.

**Works with the Existing YARN Scheduler.**
Be confident that all your YARN capabilities will remain intact.

---

**PLATFORM RECOMMENDATIONS**

*Achieve optimal application and cluster performance on multi-tenant systems.*

**Size queues appropriately based on criticality.**
Get recommendations for queue sizing.

**Size containers for optimal resource consumption.**
Get self-service recommendations for optimal container sizes for workloads.

**Accurately forecast resource needs**
Identify growth trends across resource groups.
**PLATFORM ALERTING**

Create and receive alerts that identify root causes of performance issues and operational inefficiencies.

**Identify who's blowing up the cluster.**
Use Pepperdata's rich data to track any exhaustible resource, including CPU, memory, and IO, to alert on rogue users.

**Identify which applications will miss their SLAs.**
Use Pepperdata's data on activity such as duration, amount of data processed or other metric to alert on rogue applications.

**Identify resource contention.**
Use Pepperdata's data to alert when problems arise from multiple users or applications contending for resources.

**Identify underperforming or failing hardware.**
Use Pepperdata's rich and robust platform and infrastructure data to identify network errors, disk latency and potential capacity shortages.

---

**360° REPORTS**

Make informed capacity and resource decisions.

**Accurately plan for growth.**
Access reports with all relevant real-time and historical data about the cluster, including system demand, users, and application.

**Identify wasteful use of system resources.**
Identify users and applications wasting system CPU and memory, enabling you to prioritize and tune the platform appropriately.

**Attribute accurate costs to users and business units.**
Get accurate reports on which users and business units are the most costly based on resource utilization.

---

**CHARGEBACK REPORTING**

Operators of multi-tenant environments often want to measure cluster usage by different tenants and to charge those tenants accordingly for their share of the common resource pool. Platform Spotlight provides historical and trending data for actual hardware usage (CPU, RAM, disk I/O, and network) to present a comprehensive picture of the overall cost of each tenant. This information can be exported in common file formats that can be used in financial modeling applications to provide a strong foundation for internal chargeback models. These capabilities can also facilitate capacity planning and accurate budgeting for growth of the shared platform.
Using HBase and Impala

**HBase**

Apache HBase is a column-oriented key value store built to run on top of the Hadoop Distributed File System (HDFS). HBase is a distributed, NoSQL, non-relational database designed for Big Data use and often employed in Hadoop production clusters. Many enterprises that collect streaming real-time event data use a variety of technologies, including Kafka, Apache Storm, HBase, MapReduce, Hive, and Parquet; these often run on a single shared cluster. The relative performance impacts of these components fail when volume, velocity, and variety of data impact a shared Hadoop cluster. Even for the most talented Hadoop experts, manual allocation to specific jobs fails when volume, velocity, and variety of data impact a shared Hadoop cluster.

Pepperdata automatically collects CPU, RAM, JVM garbage collection stats, HDFS I/O, and network usage from each RegionServer. These HBase statistics are available alongside statistics for other jobs in Platform Spotlight so that operators can measure the HBase resource utilization in comparison with other applications on the cluster.

**Impala**

Impala is an open source, massively parallel and interactive SQL-like query engine that runs on top of HDFS. Platform Spotlight supports process-level monitoring and query-level metrics for Impala. Memory, CPU, disk I/O, and network usage are presented in fine granularity for comparison against other user-defined applications and their performance statistics. This context helps operators understand contention with other user-level jobs run via YARN.

For query-level metrics, Platform Spotlight displays memory and CPU details on a per-query basis and provides table-level dimension aggregates for analysis of performance impacts and resource utilization. Pepperdata displays which tables and which users are using specific resources and can also provide chargeback reports. Furthermore, Impala usage is taken into account by Capacity Optimizer, allowing dynamic scheduling of additional containers to improve overall capacity utilization at the cluster level.

**Doesn’t YARN Already Do This?**

YARN (“Yet Another Resource Negotiator”) was introduced as part of Hadoop 2.0 in 2012. YARN takes the resource management capabilities of MapReduce and packages them for use by new engines. YARN enables batch, interactive, and streaming jobs to run simultaneously on the same Hadoop cluster. This allows enterprises to deploy Hadoop for new and different applications and use cases. YARN coordinates consumption and usage reservations in an attempt to ensure resources are allocated fairly. However, YARN does not track containers once they start running. This means that YARN must be conservative in its assumptions about memory usage, and assume the worst case, instead of monitoring and adjusting based on actual usage. The Pepperdata solution solves these problems by monitoring per-task hardware usage as jobs run and maximizing resource utilization.
You’re constantly looking at different tools to understand the performance of your applications to ensure SLAs, avoid failure, improve efficiency, and monitor resource capacity, and you still don’t have an accurate picture because you don’t have access to the right data. You submit queries and wait. You have no insight into when a query or app starts running, if there are delays, or whether it will finish. Developing fast and efficient applications requires an understanding of dozens of performance metrics and tuning parameters.

Application Spotlight provides you a 360° view of all this data in one place, so you can gauge performance in the context of the entire cluster, quickly diagnose application performance issues up to 90% faster, and improve overall efficiency. Pepperdata also provides automatic tuning for recurring applications, delivers job-specific recommendations and allows you to set up alerts on specific behaviors and outcomes to avoid the risk of failure.

- Identify the lines of code and the stages that cause performance issues related to CPU, memory, garbage collection, network, and disk I/O
- Easily disambiguate resources used during parallel stages
- Understand why run-time variations occur for the same application
- Determine whether performance issues are due to the application or other workloads on the cluster
- Receive actionable recommendations for tuning jobs
- Validate tuning changes made to applications with before and after comparisons
- View the highlights worst performing phases of jobs
- Improve MapReduce and Spark developer productivity
- Improve cluster efficiency based on clear recommendations on how to modify workloads and configurations
**360° APPLICATION VIEW**

Create a holistic source of application performance truth across your clusters.

**Self-service access to your application performance Data.**
Streamline the effort to profile and optimize application performance via recommendations and key performance indicators.

**Complete instrumentation of your application.**
Fine tune your application by understanding exactly what CPU and memory resources it requested, what it needs, what it used, and what it wasted. Pepperdata Application Spotlight analyzes all Hadoop and Spark jobs running on the cluster and distills the information into the most important data points that developers should pay attention to. Application Spotlight provides technical insights on how each job performed, as well as recommendations for addressing the issues that are discovered.

**Quantify the impact the cluster has on your applications.**
Identify the impact that queue congestion and bottlenecks and hardware failures have on application performance.

**APPLICATION TUNING**

Improve performance and efficiency of your applications.

**Single source for application performance data presented in context of the cluster.**
Self-service access to all of the data on your applications in one place. The ability to distinguish whether performance issues were caused by your application or other applications on the cluster.

**AI-driven optimization of recurring applications.**
Recurring applications account for a substantial portion of workload. These applications typically need to meet constantly changing SLAs. Pepperdata auto tunes configurations to optimize resource utilization or runtime.

**APPLICATION RECOMMENDATIONS**

Achieve optimal application performance.

**Change configuration parameters to optimize performance.**
Get self-service recommendations on data partitioning and serialization.

**Tune CPU and member reservations based on actual consumption.**
Get self-service recommendations on container sizes and heap reservations.

**Change queue selection or launch time based on cluster activity.**
Identify the best queue and launch time for applications based on current workloads to improve runtime and maximize performance.
APPLICATION ALERTING

Create and receive alerts about events that interfere with application performance.

**Identify cluster bottlenecks.**
Use Pepperdata’s rich data to identify resource bottlenecks, including CPU, memory, and IO.

**Identify application bottlenecks.**
Use Pepperdata’s data to pinpoint straggling tasks or poor parallelization that can significantly impact runtime.

**Identify applications at risk of missing their SLAs.**
Use Pepperdata’s data to alert on duration, amount of data processed or other metrics.

There is no tolerance for downtime in healthcare, which is why we bought Pepperdata. We started using Pepperdata on day one because Pepperdata instruments and monitors the resources as well as the applications running on the Clearsense Platform. No else does that. **We couldn’t do what we do without Pepperdata.**

– Charles Boicey, Chief Innovation Officer at Clearsense
OPERATIONAL EXPERIENCE

Leverage Pepperdata’s field-tested methodology and network of Big Data experts to develop solutions that address your challenges and strategic initiatives, providing you with a complete, transformational Big Data journey that modernizes your business. Tap into our experience and expertise to achieve more uptime, better performance, improved capacity planning, faster case resolution, and proactive issue prevention.

DEEP EXPERTISE

Achieving your Big Data goals can prove to be a moving target, with ever-changing technologies, workloads and success measures. With proven experience on hundreds of clusters at Fortune 1000 companies, Pepperdata is a trusted advisor helping enterprises establish and follow best practices, and our expert consultants can provide you with guidance on the best architecture using real-world experience derived from some of the world’s biggest clusters.

BEST PRACTICES

With proven experience on hundreds of clusters at Fortune 1000 companies, Pepperdata is a trusted advisor helping enterprises establish and follow best practices.

Get Big Data Right
Leverage Pepperdata’s field-tested methodology and network of Big Data experts to develop solutions that address your challenges and strategic initiatives, providing you with a complete, transformational Big Data journey that modernizes your business.

Keep Big Data Working
Tap into the knowledge of our experts to achieve more uptime, better performance, faster case resolution, and proactive issue avoidance and prevention.

PERFORMANCE PLANNING

Pepperdata continuously collects and integrates extensive data with existing data sources and provides expertise to help you make changes and tune both your platform and applications for speed and efficiency.

Maximize your existing infrastructure investment
Pepperdata helps you increase hardware utilization and eliminate or reduce millions of dollars in hardware costs.

Understand application performance
Configure and tune applications to maximize resource allocation with insights from Pepperdata’s proven experience on hundreds of clusters at Fortune 1000 companies.

CAPACITY PLANNING

Pepperdata experts help you squeeze the most performance out of your existing capacity and help in forecasting capacity needs.

Stay ahead of capacity needs
Leverage Pepperdata support to complement in-house knowledge when making decisions to ensure the right resources are available at the right time.

Accurately forecast resource needs
Rely on Pepperdata consultants to uncover the relationship between workloads and infrastructure to make intelligent hardware purchases.

ARCHITECTURE DESIGN

Pepperdata’s expert consultants provide you with guidance on the best architecture using real-world experience derived from some of the world’s biggest clusters.

Understand trade-offs
Understand trade-offs resulting from node hardware, network choices, and storage sub-systems as well as trade-offs associated with Big Data distribution vendors.

Run what-if scenarios
Leverage Pepperdata to conduct analyses and run what-if scenarios to determine precise requirements, avoiding both cost and risk.
**OPERATOR TESTIMONIALS**

“Before Pepperdata, we experimented with various approaches to solve our performance issues, but we couldn’t see deep enough into the cluster. Pepperdata shined a bright light into our Hadoop environment and provided the detailed data that helped us isolate and resolve the problem.”

– Software Architect at Philips Wellcentive

“At Rubicon Project, having the appropriate visibility and insight into our Big Data applications is extremely important when delivering detailed reports to our clients and meeting our SLA. We challenged Pepperdata to find a solution to profile our applications before going to production that would help us maintain our customer SLA as we introduce new applications. Pepperdata listened to us and quickly understood the problem we were trying to address.”

– Senior Systems Engineer at Rubicon Project

**DEVELOPER TESTIMONIALS**

“I develop a lot of complex Spark code to perform ETL on Hadoop clusters. In these complex, large-scale systems, you must be able to understand where the performance bottlenecks are. Pepperdata Application Spotlight gives developers detailed time-series performance data for things like CPU, JVM memory, and I/O usage overlaid against Spark job stages. I’m excited about the direction Pepperdata is moving — letting developers quickly see problems in time-series views and tie them back to their actual Spark application code will be a very useful tool for developers working on production Spark applications.”

– Software Engineer at Stripe and Pepperdata Technology Advisory Board Member

“Chartboost is the world’s largest mobile games-only advertising platform, reaching one billion active players around the world every month. Chartboost utilizes Apache Spark on large Amazon EC2 Hadoop clusters for machine learning and ETL workflows. Understanding Spark application performance in these complex environments is always a challenge. As a current user of Pepperdata Platform Spotlight, it has been great to work with Pepperdata on the development of the Application Spotlight self-service portal software. It will give us a comprehensive insight into Spark jobs.”

– Manager of Data Engineering at Chartboost

**PRODUCT INSTALLATION**

The Pepperdata software suite installs in about 20 minutes on an existing Hadoop cluster without any modifications to scheduler, workflow, or job submission processes. Pepperdata supports all Hadoop distributions, including Cloudera, Hortonworks, MapR, IBM BigInsights, Pivotal PHD, and Apache Hadoop, as well as Amazon EMR.

**ABOUT PEPPERDATA**

Pepperdata is the leader in Application Performance Management solutions and services for Big Data success. The company partners with its customers to provide proven products, operational experience, and deep expertise to deliver predictable performance, empowered users, managed costs and managed growth for their big data investments, both on-premise and in the cloud. Leading companies like Comcast, Philips Wellcentive, and NBC Universal depend on Pepperdata to manage and improve the performance of their Big Data infrastructures by troubleshooting performance problems, increasing cluster utilization, and enforcing policies to support multitenancy.