unravel™

SOLUTION GUIDE

ARCHITECT'S SOLUTION GUIDE

Optimising Data Pipelines with the power of AI

TABLE OF CONTENTS

WHAT IS DATA OPERATIONS AND WHY DO ENTERPRISES NEED IT?

Data Operations and Application performance management is not a new discipline, but it is a new best practice for big data – adopting an application-first approach to guarantee full-stack performance, maximize utilization of cluster resources, while minimizing the TCO of the infrastructure.

For architects, it means that the big data architecture has to be designed to meet new business needs for speed, reliability, and cost-effectiveness, as well as align with architecture standards for performance, scalability, and availability.

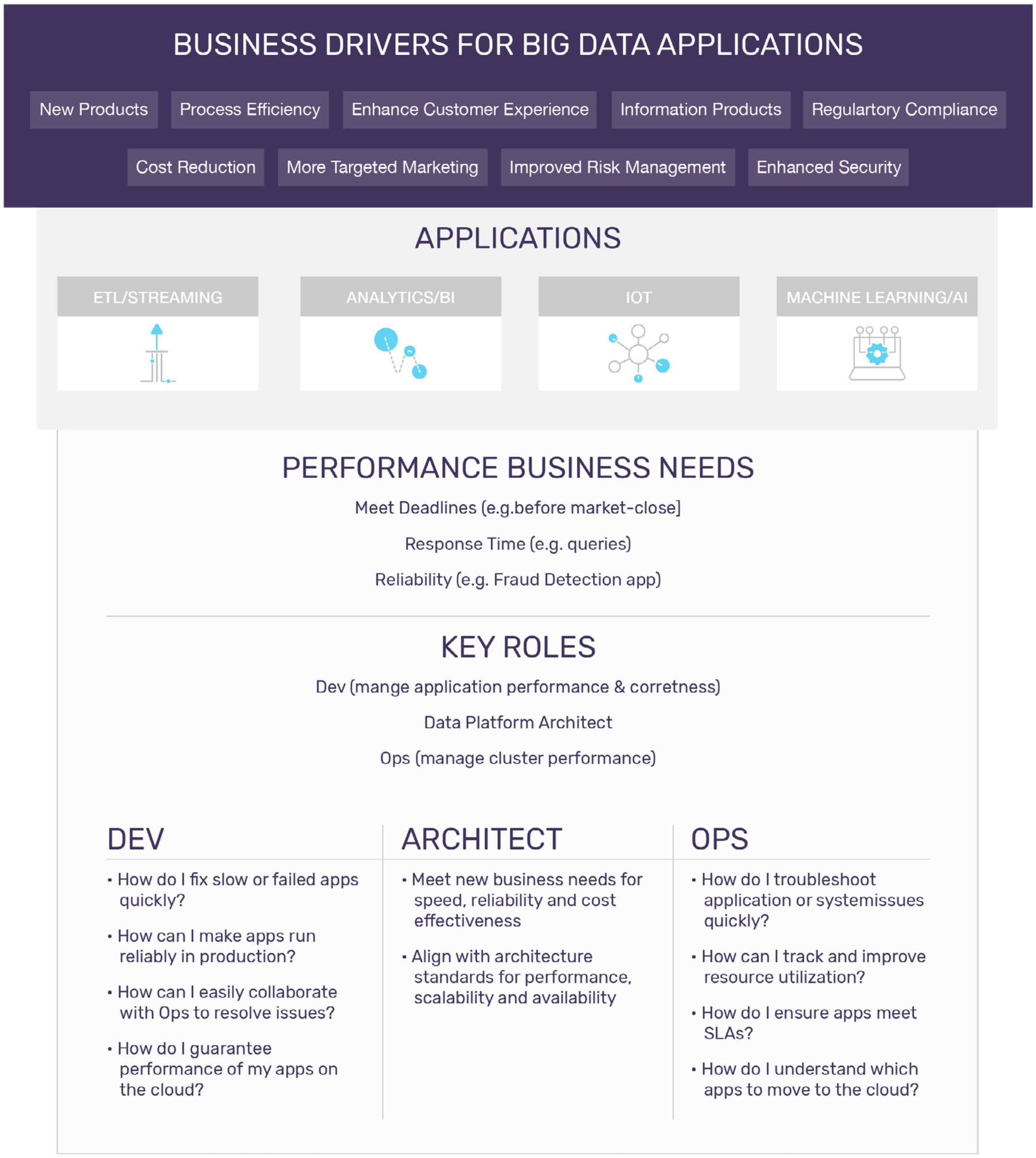
Troubleshooting and tuning distributed applications is difficult and expensive.

CONTAINER
SIZES
BAD
JOINS
FILE
FORMATS
CONFIG
SETTINGS
MACHINE
DEGREDATION
BUGS
SCHEDULER
SETTINGS
DATA
LAYOUT
NETWORK
SETTINGS

Big data is trending from experimental projects to becoming a mission-critical data platform offering a range of big data applications. Enterprises look to these big data applications (e.g., ETL offload, Business Intelligence, Analytics, Machine Learning, IoT, etc.) to drive strategic business value.

As big data applications have moved to production, performance expectations have changed and now need to be production grade.

The business needs answers in seconds and not hours, hardware and resources need to be continuously optimized for cost, and deadlines/SLAs need to be guaranteed. This means that APM needs to become a strategic component of a big data architecture in order to eliminate risks and costs associated with poor performance, availability, and scalability.



BUSINESS DRIVERS FOR BIG DATA APPLICATIONS

New Products | Process Efficiency | Enhance Customer Experience | Information Products | Regulartory Compliance

Cost Reduction | More Targeted Marketing | Improved Risk Management | Enhanced Security

APPLICATIONS

ETL/STREAMING | ANALYTICS/BI | IOT | MACHINE LEARNING/AI

PERFORMANCE BUSINESS NEEDS

Meet Deadlines (e.g.before market-close]

Response Time (e.g. queries)

Reliability (e.g. Fraud Detection app)

KEY ROLES

Dev (mange application performance & corretness)

Data Platform Architect

Ops (manage cluster performance)

DEV

- How do I fix slow or failed apps quickly?
- How can I make apps run reliably in production?
- How can I easily collaborate with Ops to resolve issues?
- How do I guarantee performance of my apps on the cloud?

ARCHITECT

- Meet new business needs for speed, reliability and cost effectiveness
- Align with architecture standards for performance, scalability and availability

OPS

- How do I troubleshoot application or systemissues quickly?
- How can I track and improve resource utilization?
- How do I ensure apps meet SLAs?
- How do I understand which apps to move to the cloud?

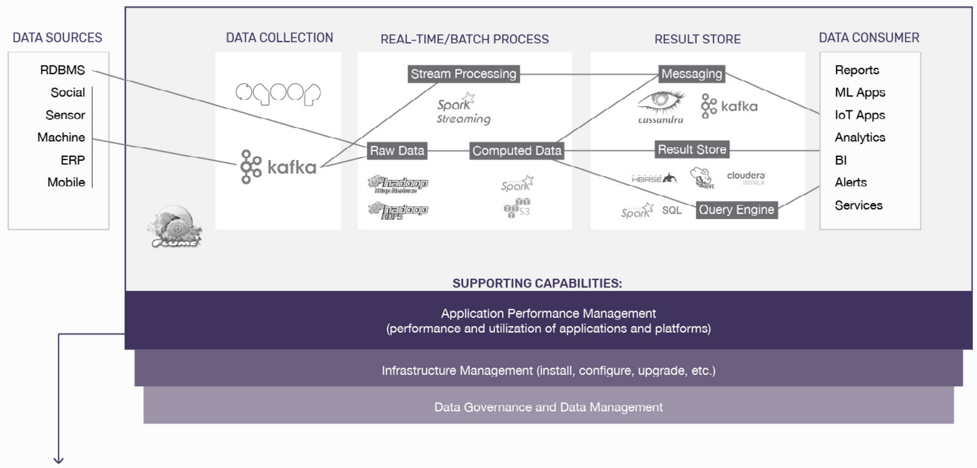
CURRENT CHALLENGES

The fundamental problem is that the big data stack is complex due to its' distributed nature where infrastructure, storage and compute are spread across many layers, components, and heterogeneous technologies.

This problem exists regardless of the specific architecture (i.e., traditional, streaming analytics, Lambda, Kappa, or Unified). From the perspective of the production big data platform and the applications that run on it, everything must run like clockwork: ETL jobs must happen at fixed intervals; users expect dashboards to be up-to-date in real time; user-facing data products must work constantly. But from the perspective of the underlying platform, the application is not an isolated job, but rather a set of processing steps that are threaded through the big data stack.

For example, a fraud detection application (i.e., Data Consumer) would be comprised of a chain of many systems from Spark SQL, Spark Streaming, HDFS, MapReduce, and Kafka, as well as many processing steps within each system. As a result, the entire process of managing performance and utilization across all these layers is exponentially complex.

From the perspective of the production big data platform and the applications that run on it, everything must run like clockwork: ETL jobs must happen at fixed intervals; users expect dashboards to be up-to-date in real time;user-facing data products must work constantly.



Example of Big Data Architecture

Performance Info Levels	Current Approaches Aren't Enough
Recommendations	None
Insight	None
Correlated Information	Manual process or coding
Logs	Look at component logs
Jobs KPIs	Not always available

This complexity makes it very hard to implement Application Performance Management services that provide a single view to manage performance and utilization across the full-stack. In particular, there is no rationalized instrumentation across the stack to enable a holistic approach to guarantee performance and maximize utilization. Instead, performance and utilization information is scattered across disjointed metrics, buried in logs, or spread across performance monitoring / management tools that only provide an incomplete infrastructure view as opposed to a full stack view.

Challenges:

Lack of reliability
Missed SLA and revenue

Sub-optimized resources
High infrastructure and project costs

Inefficient resolution
Long MTTR (mean time to resolution)

CHALLENGES FOR ARCHITECTS

As a result, the process of planning, operationalizing, and scaling the performance and utilization across applications, systems, and infrastructure is not production-ready.

This challenge is called out in Gartner's March 2017 Market Guide for Hadoop Operations Providers. The report states "scaling Hadoop from small, pilot projects to large-scale production clusters involves a steep learning curve in terms of operational know-how that many enterprises are unprepared for."

The lack of production readiness spans multiples areas across business units, developers, and operations. At its core, it makes it impossible to implement a multi-tenant cluster model, where a small Ops team needs to support a large number of applications, business units, and blended workloads with a combination of SLA-bound jobs vs. data discovery. The ultimate impact affects adoption of big data and business value realization.

Plan

- No understanding of dependencies across the big data stack
- Lack of understanding of application usage
- Guesswork to size and tune clusters

Operationalize

- Manual troubleshooting
- Trial-and-error resolution
- Escalating support tickets

Scale

- Lack of control to maximize utilization
- Lack of self-service to troubleshoot issues
- Lack of governance to optimize storage and compute utilization

ARCHITECT'S CHECKLIST

The architect should play a pivotal role to ensure that the big data platform is designed for production. The architect can ensure that the big data platform will meet the needs of the business within time and budget constraints, as well as ensure the architecture will adapt to new business needs as they evolve over time. The architect's checklist can be used in planning, operationalizing and scaling the big data platform in order to manage performance, utilization, and cost.

Planning

1. What types of applications is the business trying to build and deploy?
2. Will applications be SLA-bound or ad-hoc? How will workloads be prioritized cost effectively?
3. Which systems are best suited for the applications (e.g., Spark, Hadoop, Kafka, etc.)?
4. Which architecture approach is best suited (e.g., Lambda, etc.)? Will the cluster be on-premise, in the cloud, or hybrid?
5. How many concurrent users need to run on the same cluster without running out of resources? How many applications need to run on the same cluster within 24 hours? How will throughput be optimized?
6. How should storage be tiered?
7. How many nodes will the cluster need? What infrastructure capabilities need to be in place to ensure scalability, low latency, and performance, including computing storage and network capabilities?
8. What data governance policies need to be in place?
9. How will dev, QA, and production be staged?
10. How much will the cluster cost to run? How will the business be charged back?

Operationalizing

The operationalizing phase can be broken into 4 stages. The staged approach helps to gradually shape and scale the successful implementation and ROI of big data applications.

Experimentation
Understanding the capabilities of a big data platform

Implementation
Developing first production use cases

Expansion
Expanding to multiple use cases across the company

Optimization
Optimize and integrate apps on the converged data platforms

For each stage the following set of key questions apply:

1. What are the SLAs for applications? How can they be guaranteed?
2. What are the latency targets for applications? How will they be met?
3. How will Ops be able to support business units and users in a multi-tenant cluster? How will dev be able to monitor applications in a self-service fashion? How will Ops troubleshoot issues?
4. When do users typically log in and out? How frequently?
5. Do different groups of users behave differently? How do activity profiles of users change over time?
6. How do I track costs in a multi tenant cluster? How do I assign them to projects, business units, users, applications, etc.?
7. How will data governance policies be enforced?

Scaling

In order to automatically correlate, root-cause, and provide fast and accurate answers to the following questions, you need full stack data (apps, datasets, resources, users) and intelligence.

1. How will a test job with sample data run with a full load in production?
2. How many nodes to add?
3. How to keep costs low?
4. How to scale fast but still maintain enterprise quality and reliability?
5. How to improve by:

- **Scaling down:** the amount of data processed or the resources needed to perform the processing
- **Scaling up:** the computing resources on a node, via parallel processing and faster memory/storage
- **Scaling out:** the computing to distributed nodes in a cluster/cloud or at the edge

Best Practices

Managing a big data platform in production requires a set of best practices. The following is a list of best practices that pertain to application performance management for production big data applications and clusters.

1. Tear Down Silos and Be Proactive

- Adopt a lifecycle and 24/7 production approach. It will enable you to notice changes in data and compute distribution over time. In addition, a lifecycle approach will allow you to immediately pinpoint any negative changes introduced by new applications, users, or changes in the big data platform.
- Don't just wait to fix problems in production. The ultimate goal is to become proactive by fixing issues in development or testing—before they spill over into production.
- Instead of relying on fragmented and incomplete sets of tools, use one single solution that is designed to be used across the application lifecycle, from development to production, as well as all the systems the application will run on. It will make it easier to share application performance data between lifecycle stages and understand the dependencies across all the underlying systems and cluster resources.

2. Down to "Source Code" Troubleshooting

- Knowing if it was the code or infrastructure or something in between that is causing poor performance of applications is key to expedient troubleshooting. An APM solution needs to lower the Mean-Time-To-Resolution (MTTR) from as much as several days down to minutes. In order to do so, fullstack information needs to be available and correlated in order to pinpoint the root-cause quickly and confidently.

3. Centralize and Correlate Data Collection of Performance Information

- It is not feasible to manually collect different types of performance information from different big data systems and clusters. It would require multiple monitoring tools, logs and custom data transformation to create a single view that can be analyzed – this approach is both expensive and slow to implement and requires a large number of experts. But the biggest drawback is that the data from different tools is either incomplete or doesn't "line up."

- Instead of this approach, use a single view of performance and utilization across the stack to all key stakeholders from Ops, to Dev, and support teams.
- Correlating, storing, and accessing performance data from a single, centralized repository enables fast and powerful analysis and visualization with a correlated view of performance metrics.

4. Ability to Extend and Integrate

- The big data stack is an ever-changing landscape of new systems and environments.
- An APM solution must be able to adapt to integrate with the big data stack as it evolves (e.g., support for new technologies like Impala, Tensorflow, Kudu, etc.).

5. Don't Just Throw More and More Hardware at the Problem

- Although hardware is cheaper, it is not cheap. But you also have to consider the operational drag. Every node you add makes the cluster more complicated. Instead, ensure that you can understand and optimize all jobs and workflows across SLA-bound and "rogue" applications to reduce both the time and resources it takes to run them.

6. Institutionalize a Performance Best-practice

- Implement APM as a discipline to drive an application-driven mindset in the organization. Start by implementing a Performance Center of Excellence. This will serve as the foundation to help the organization move from reactive performance troubleshooting to proactive performance prevention.

CRITICAL CAPABILITIES FOR DATA OPERATIONS AND APPLICATION PERFORMANCE MANAGEMENT

There are critical capabilities required to manage performance and utilization of big data applications and the platform they run on. These capabilities span both business and technical requirements to ensure all needs are met between business and IT.

Full Stack Support

Big data applications don't execute in isolation- everything is part of some larger workflow. The ability to support full-stack means that performance and utilization need to be managed not just for one job, but correlated across all the jobs the application is dependent on. This capability is essential in order to thread an end-to-view from application down to infrastructure, because jobs depend on other jobs, and code contributions come from several development teams which then need to be orchestrated by a different infrastructure team.

SLA Management

SLA management ensures that business needs in terms of responsiveness and uptime are defined, agreed-upon, and will be met. This includes identifying adherence to contracted service-level application availability and performance thresholds. SLA metrics provide visibility to LOB and application owners, as well as IT operations and DevOps teams, in order to have a shared view of key business needs.

Anomaly Detection and Automated Actions

The ability to detect and rapidly respond to situations that begin to fall outside normal operating parameters. The earlier this detection takes place, the better, in order to prevent cascading problems. Anomaly detection is critical for IT operations, DevOps and even application support teams. Moreover, the ability to take automated actions further improves the speed with which downstream problems can be prevented, as well as frees up a small Ops or support team to focus on more critical issues.

Automated Troubleshooting and Insights

There are many potential sources of application performance degradation, and root cause analysis can span an exponential number of parameters. As such, manual troubleshooting cannot scale in a multi tenant production environment. The ability to automatically identify the source of the degradation, as well as provide insight as to the causes, can dramatically improve the productivity of DevOps and support teams while drastically reducing the Mean-Time- To-Resolution.

Automated Recommendations

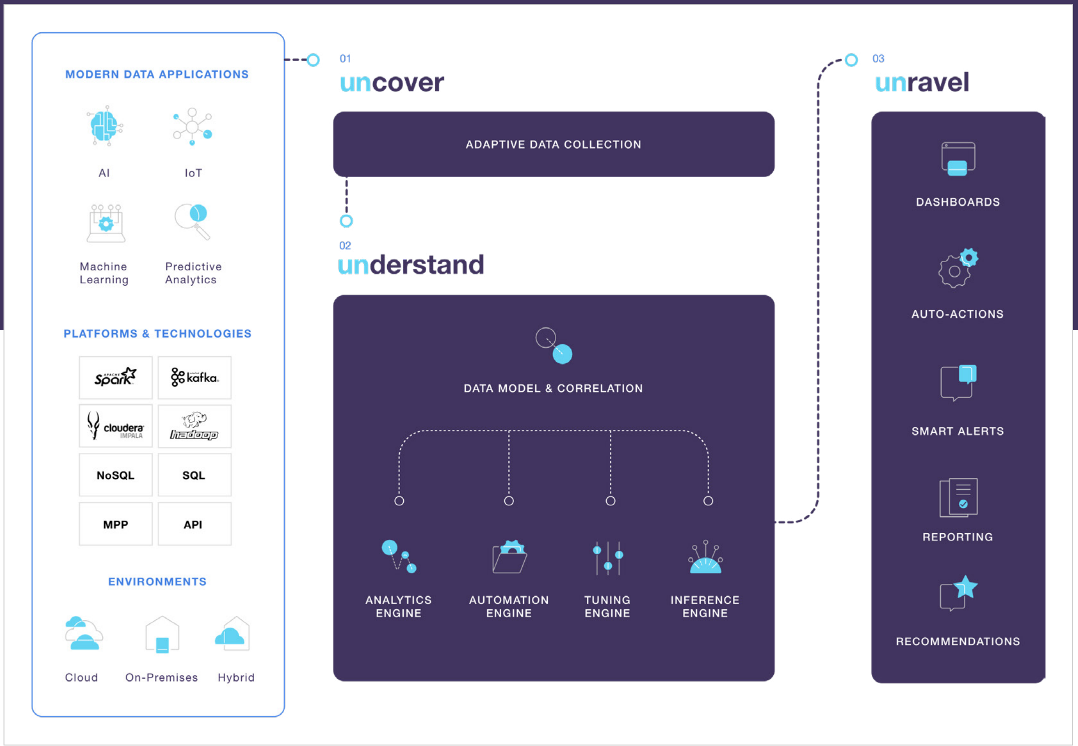
Automated recommendations go one step further than automated troubleshooting and insights. Automated recommendations eliminate the need for trial-and-error in fixing problems during production, by providing a specific fix which DevOps or support teams can more readily deploy. Automated recommendations can also accelerate debugging during the application development process.

Workload Planning and Chargebacks

Storage and compute utilization need to be optimized to maximize performance as well as reduce cost. It is key to understand utilization with an application first perspective in order to track all the processing steps the application depends on across systems. This capability provides DevOps with the ability to plan and optimize cluster resources, tell business stakeholders of the potential impact for any planned changes as well as accurate chargebacks.

THE UNRAVEL SOLUTION

Unravel offers solutions for [AI-powered] data operations, helping Architects, Operations and Development teams get the most out of the modern data applications that power your business by radically simplifying how you monitor, manage, and optimize the performance of your data workloads.



Modern data pipelines are inherently complex, with many different, disparate components that all need to work together seamlessly and reliably. Until now, seeing across all of those layers – how they work together, where the issues are, what can be improved – has been challenging.

Unravel dramatically simplifies and streamlines this process, offering a single, unified view across your stack, so you can understand the complete picture – with real-time intelligence into the issues that can hinder performance.

But Unravel doesn't stop there. Using AI, machine learning, and predictive analytics, Unravel provides tangible, actionable advice and recommendations, so you don't just know what's broken or what's possible – you know exactly what steps to take to get the most out of your apps, systems, and infrastructure. You can even automate these recommendations, for the continuous improvement of your application environment.

By fully operationalizing how you monitor and optimize performance, Unravel empowers your team to quickly and easily tackle the thorniest and most elusive problems plaguing your applications.

Unravel was designed specifically for big data platforms to be full-stack (centralize and correlate all performance data from infrastructure, services, to apps), intelligent (generate reports, insights and recommendations to optimize, troubleshoot, and analyze performance and utilization), and autonomous (e.g., automatically detect and resolve issues).

What makes Unravel different?

Full Stack Coverage

Unravel works across your entire ecosystem to demystify and simplify operations.

- Get 360° visibility into apps, systems, services, users, infrastructure, and more.
- Gain clarity into code, configurations, containers, resource constraints, and dependencies.
- Unravel's agentless design and lightweight micro-sensors put minimal load on clusters.

AI-driven Recommendations

Unravel does more than monitor – it shows you how to make things better.

- Uncover the root cause of issues and specific recommendations for enhancing performance.
- Map dependencies among applications, services, resources, and users.
- Optimize cloud platform instances by forecasting application needs.

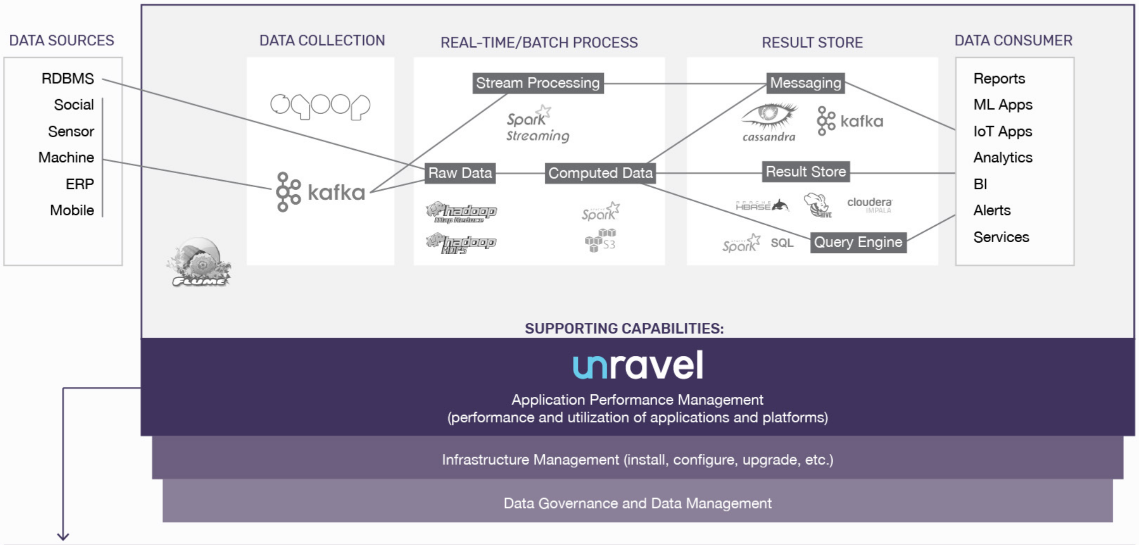
Automated Tuning Remediation

Unravel operationalizes big data by automating it.

- Leverage Auto-Actions to automatically improve performance, resource usage, and reliability.
- Automatically detect and correct performance bottlenecks and failures.
- Autonomously find and eliminate rogue or resource-wasting applications and users.

Intelligent and Autonomous APM with Unravel

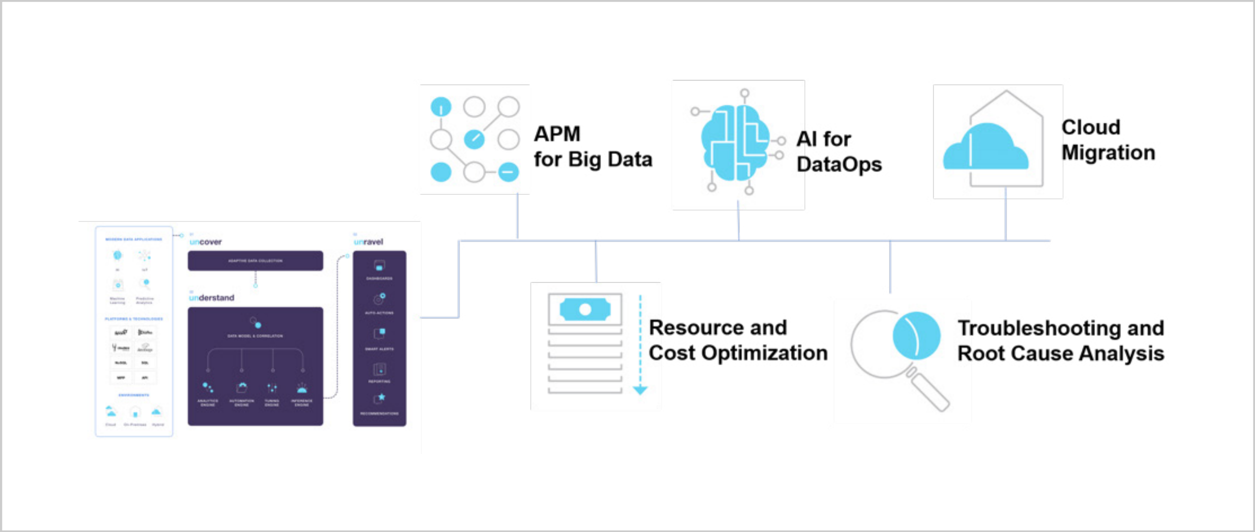
Unravel leverages Artificial Intelligence and Machine Learning to achieve the highest level of maturity for an APM solution to automatically deliver insights and recommendations just like an expert would.



Performance Info Levels	Without Unravel	unravel
Recommendations	None	AI
Insight	None	Machine Learning
Correlated Information	None	Automated Correlation
Logs	Manually Sifting Through Logs	Automated Parsing
Jobs KPIs	Fragmented Metrics	Unified KPIs
Basic Performance KPIs	Platform-specific Monitoring Tool	Unified KPIs

How can Unravel help?

Unravel provides the insight, intelligence, and guidance to address a number of use cases across your business.



APM for Big Data

The size and complexity of the big data ecosystem makes it challenging to manage application performance. Unravel sets a new standard for big data APM, helping you not only monitor performance, but optimize it – for applications that are faster, more reliable, and cheaper to run. .

AI for DataOps

Today's data is too complex to manage manually. Unravel uses AI, machine learning, and predictive analytics to automatically monitor, diagnose, and troubleshoot problems. Get straightforward answers. Quickly resolve issues. Improve the productivity of your team.

Cloud Migration

Get the most out of your cloud investment – Unravel can help you perfectly plan and execute the migration of your apps to the cloud, helping you reduce friction, maximize performance, and minimize resource usage costs.

Resource and Cost Optimization

Inefficient clusters can cost your business big. Unravel reveals exactly how apps are consuming cluster resources, providing the insight you need to allocate and use resources the right way. If, for example, you typically run 100 apps in a day, Unravel can tune your environment to run 200 – with the same performance and stability – and a 50% reduction in costs.

Troubleshooting and Root-cause Analysis

Complicated apps and pipelines make it hard for teams to find the right solutions. Unravel helps you quickly uncover what's causing bottlenecks and issues – from code to infrastructure – so you can quickly resolve them. We can even find and fix these problems automatically.

UNRAVEL MAKES DATA WORK.

With all the apps, systems, and teams working in the modern data ecosystem, it can be difficult to understand what's affecting performance – and how to improve things.

Unravel sheds new light on your apps and environment, providing every team with the visibility, the intelligence, and the guidance to drive better performance – and better results – in the applications that power your business.

Greater Productivity

98% reduction in troubleshooting time.

Guaranteed Reliability

100% of apps delivered on time.

Lower Costs

60% reduction in costs.

UNRAVEL SUPPORTS THE TECHNOLOGIES YOU RELY ON MOST.

Supported Cloud Environments

A full-featured trial version of Unravel is available as on both Amazon and Azure clouds.

aws	MAPR	databricks	HORTONWORKS
cloudera	Qubole	Google Cloud	Azure

Unravel for Amazon AWS and EMR

Unravel is available on Amazon AWS and Amazon EMR, supporting a variety of cloud services.

Amazon EC2	amazon REDSHIFT	amazon web services	Amazon Athena
amazon web services S3	Amazon S3	amazon EMR	

Unravel for Microsoft Azure

A Microsoft Co-Sell Partner, Unravel is available in the Azure Marketplace and supports multiple cloud services.

Microsoft Azure	Azure (iaas)	APACHE Spark	databricks
Azure ADLS	HDInsight		

SUPPORTED TECHNOLOGIES

Big Data Ecosystem

SYSTEMS AND ENGINES

APACHE Spark	hadoop	cloudera IMPALA	APACHE kafka	cassandra	cascading	TEZ	APACHE HBASE

WORKFLOW SCHEDULERS

OOZIE	Airflow	Control-M	cron	HUE	H_2O	Apache Zeppelin

PLATFORMS

databricks	MAPR	cloudera	aws	HORTONWORKS	Azure	Google Cloud	Qubole
ORACLE DBA							

Environment

MICROSERVICES

MESOS	kubernetes	docker

INFRASTRUCTURE ENVIRONMENTS

Azure ADLS	Amazon EC2	amazon EMR	Amazon S3	HDInsight

SECURITY AND ACCESS CONTROL

LDAP	okta	Ping Identity	Microsoft AD	SAML	KERBEROS DIGITAL DIVISION	Azure Active Directory

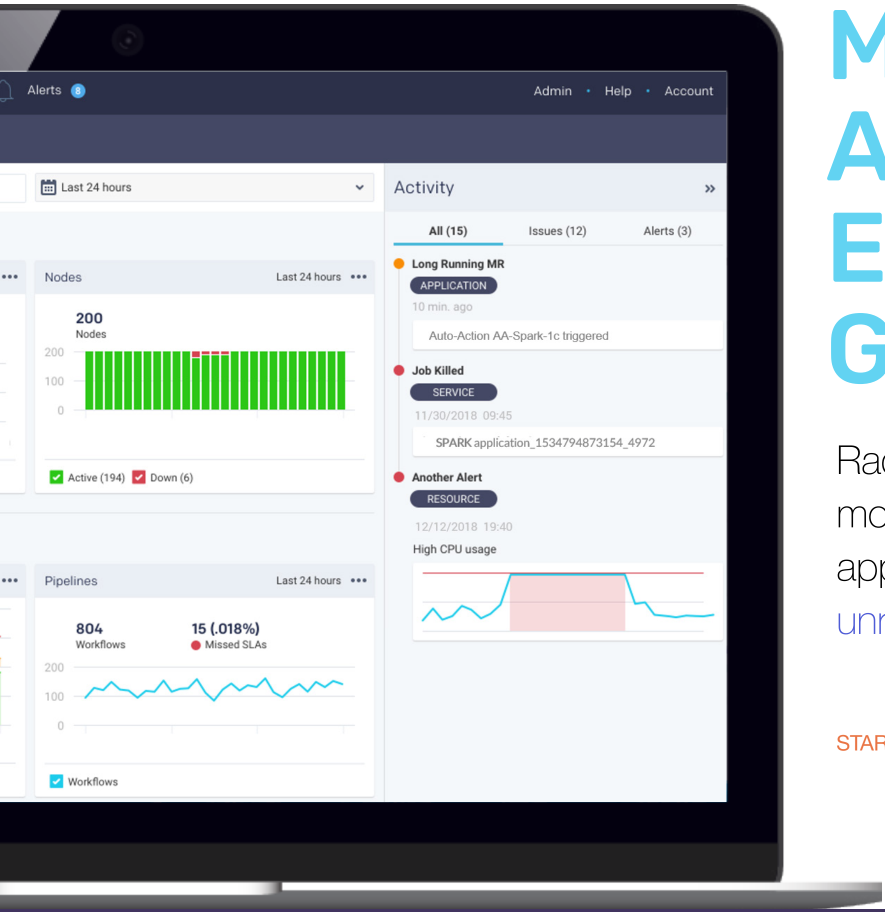
Other Tools

MONITORING

cloudera manager	HORTONWORKS Ambari	MAPR Control System	Spark UI	DATADOG	Grafana	splunk>

COLLABORATION

servicenow	slack	Jira	pagerduty



Long Running MR
10 min. ago
Auto-Action AA-Spark-1c triggered

Job Killed
11/30/2018 09:45
SPARK application_1534794873154_4972

Another Alert
12/12/2018 19:40
High CPU usage

READY TO MAKE YOUR MODERN DATA APPLICATIONS ENTERPRISE GRADE?

Radically simplify the way you troubleshoot, monitor, and optimize your data-driven applications – try Unravel for free at unraveldata.com.

START YOUR FREE TRIAL →

About Unravel

Unravel radically simplifies the way businesses understand and optimize the performance of their modern data applications – and the complex pipelines that power those applications. Providing a unified view across the entire stack, Unravel's AI-powered data operations platform leverages AI, machine learning, and advanced analytics to offer actionable recommendations and automation for tuning, troubleshooting, and improving performance – both today and tomorrow. By operationalizing how you do data, Unravel's solutions support modern big data leaders, including Kaiser Permanente, TIAA, Adobe, Deutsche Bank, Wayfair, and Neustar. The company is headquartered in Palo Alto, California, and is backed by Menlo Ventures, GGV Capital, M12, Data Elite Ventures, and Jyoti Bansal. To learn more, visit unraveldata.com.

unravel

unraveldata.com | hello@unraveldata.com