CLOUDERA

Taking Your Data Lifecycle to the Next Level

Why CDP Data Engineering is the Solution for Large-Scale Data Processing and Analytics Success

Table of Contents

The Rising Demand for Effective Data Engineering at Scale

Apache Spark™ is the framework of choice for extract/transform/load (ELT) jobs—but for enterprise users wanting to move those jobs into production at scale...there's a catch.

Data engineers rely on Apache Spark to process large data volumes in near real time. Its ability to accelerate the ingestion, exploration, modeling, curating, and cataloging of data types from multiple sources lets users quickly build batch or streaming pipelines with relative ease.

But for all of its processing prowess, Spark still requires significant manual work under the hood. ETL jobs inherently have resource-intensive and time-consuming requirements that can impact your analytical workflows down the line. When a data pipeline is ready for deployment, you have to provision it with adequate resources, account for the job in your capacity planning, and schedule it. And even after deployment, you have to ensure that the right dependencies are carried over into production, and then continuously monitor the job for problems.

On top of that, having to actually debug or performance-tune the job will only result in lost time, wasted resources, and additional headaches. For example, you have to manually gather and scrutinize all of the necessary logs in hopes of finding a bottleneck or underlying issue. And when it comes time to upgrade to the latest version of Spark, the entire cluster—that could stretch across any number of teams—has to come down, bringing work to a halt.

So, while Spark can process large data volumes at incredible speeds, it struggles with effective data engineering in production at scale—and this shortcoming ultimately has a negative impact on your advanced analytics initiatives.

Effectively and efficiently building, deploying, and managing data pipelines require a secure, integrated, and streamlined approach to data integration, modeling, optimization, quality, governance, security, and reusability.

"Making the right data available for experimentation, and transitioning from experiment to production, are becoming increasingly complex tasks. It is also becoming obvious that the creation and maintenance of these data pipelines won't take care of itself; it must be someone's job."[1]

Gartner "Data Engineering Is Critical to Driving Data and Analytics Success," Robert Thanaraj, et al, 18 December 2019

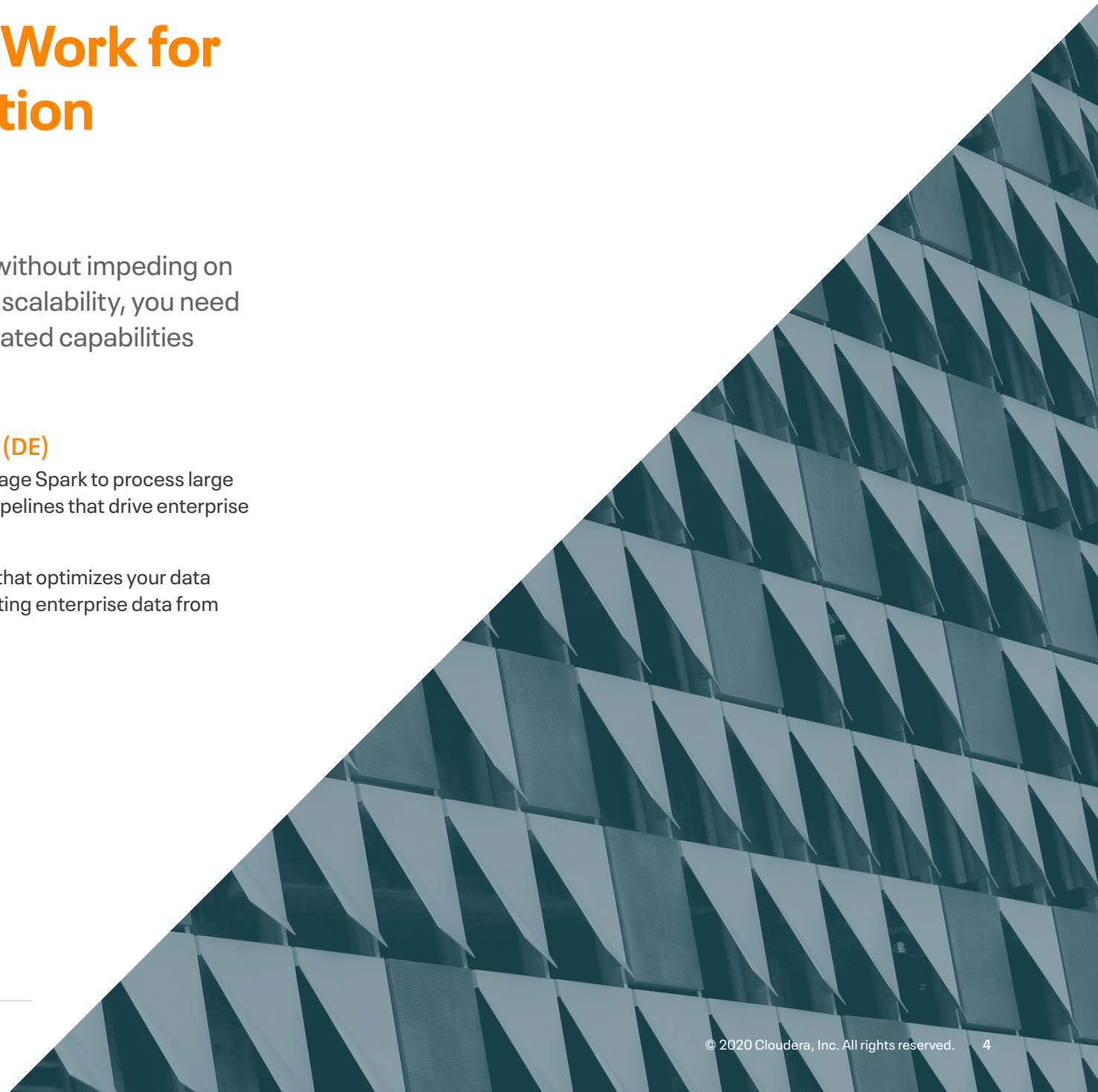
Making Spark Work for Your Organization

To harness the power of Spark without impeding on data engineering flexibility and scalability, you need secure, streamlined, and integrated capabilities and tools.

Introducing CDP Data Engineering (DE)

CDP Data Engineering (DE) lets you leverage Spark to process large data volumes and streamlines the data pipelines that drive enterprise analytics and machine learning success.

DE is a modern data engineering service that optimizes your data pipeline management lifecycle, accelerating enterprise data from ingest to insight at scale.



Deliver Quality Data Sets

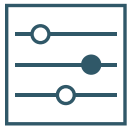
Data sets are the backbone of any advanced analytics initiative. DE lets you deliver curated, quality data sets while also ensuring security and governance compliance—all with open industry standard tools like Apache Airflow.

Manage Resources & Users Effectively

Platform administrators can use DE as a centralized way to control costs, security, and governance. DE empowers users to deliver agile, self-service data engineering for governed data while also regulating provisions and ensuring isolation across business stakeholders.

Deploy With Ease & Control Costs

DE is a containerized, scalable, and portable service, so it's easy to deploy in cloud environments. It also features on-demand automatic resource scaling, so you only pay for what you use and maintain efficient resource utilization.

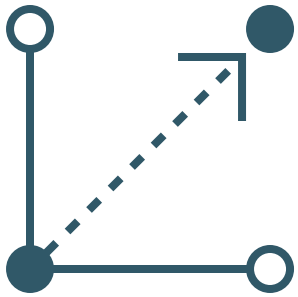


Monitor Jobs & Troubleshoot With Clarity

Data engineers can use DE to efficiently deploy and monitor the lifecycle for every job and then use visual troubleshooting tools to quickly solve issues. Through a single pane of glass, users can make adjustments and review visualized analytics, which helps deliver and continuously maintain production-ready data pipelines.

Taking the Manual Effort out of Data Engineering

To move data pipelines into production at scale, data engineers can't be bogged down by time-consuming tasks that could be automated or otherwise eliminated.



Choose Your Preferred Version of Spark

Because CDP Data Engineering (DE) removes the need for version uniformity within and across teams, data engineers using different versions of Spark can seamlessly collaborate. And, even when a version upgrade is implemented, it takes place behind the scenes without disruption, so your work never skips a beat or comes to a standstill.

Scale Without Disruption

Jobs can scale in DE automatically to match demand without disrupting your other workloads. This frees you from the time-consuming process of requesting additional resource provisioning—or worrying that a workload won't have the needed resources if faced with a sudden spike in demand.

Use Multiple Languages

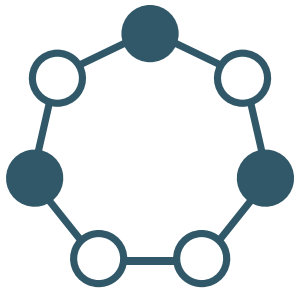
DE gives you the coding flexibility to use whatever language you prefer for a given task. Choose from Scala, Java, Python, and others.

Benefit From Rich APIs for Automation & Services

Use APIs to automate the lifecycle management of clusters, applications, and more. For example, with the Apache Airflow programmatic API, you can automate the orchestration of more complex pipeline scenarios. You can also integrate with the rest of CDP's portfolio and its partner services for additional authoring, scheduling, and monitoring functionality.

Granting Self-Service Access and Autoscaling to Meet Demand

Data engineers and business users need easy access to data, regardless of where it resides. Platform administrators, however, have to make sure that data everywhere is secure, governed, and properly managed.



With Cloudera Shared Data Experience (SDX), CDP Data Engineering (DE) is a secure and portable, multi-function platform with a single SDX across on-premises, hybrid, and multi-cloud environments.

Meanwhile, data engineers can access raw data wherever it resides to create high-quality, production-ready pipelines and power analytics workstreams.

Platform Integration

Through DE's seamless data integration, platform administrators can deliver agile, controlled self-service data engineering for governed business data and analytics pipelines. Not only can administrators make data widely accessible, they can do so while ensuring security, governance, and the minimizing of data copies across all environments.

Data Pipeline Management & Scale

Because DE is containerized and multi-cloud portable, it can ingest and process data regardless of scale. It can also autoscale workload resources, so there's no need for manual provisioning. This not only controls costs, it ensures that jobs have the resources needed during spikes in demand.

Gaining Operational Visibility and Control

Finding and troubleshooting performance issues, tracking lineage, managing jobs, and controlling costs and resources become increasingly complex when attempting to operationalize data pipelines at scale.

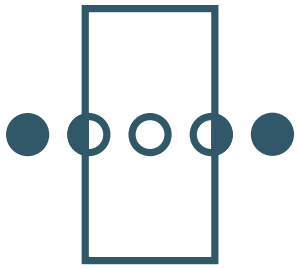
CDP Data Engineering (DE) offers a suite of operational control and visibility features for automated orchestration, self-service troubleshooting, automatic lineage capture, and more.

Troubleshoot, Tune & Debug With In-depth Visual Analysis & Self-Service

Through integration with Safari and WXM, DE provides a visual analysis for each stage of every job's lifecycle. By simply selecting a job from the DE dashboard, a data engineer can check its utilization rate, CPU and memory usage, and review other important metrics, such as how long the job takes to run.

DE analytical capabilities also let data engineers dig deeper to find problems and find correlations. Users can review logs, see a historical analysis of jobs, and even safely experiment with tweaking parameters to see how jobs could be impacted.

Additionally, data engineers can take advantage of DE's intelligent, automated troubleshooting capabilities for Spark. Patterns in job failures and performance issues are recognized over time and encoded into DE. This means that DE can identify circumstances that lead to job issues before they occur and offer corrective recommendations based on an analysis of previous actions and outcomes.



Provide Services Faster While Safeguarding the Platform

From a centralized interface, platform administrators can quickly provision new workloads while easily monitoring capacity and visualizing resource usage over time, giving them total cost transparency.

To further control costs and manage resources, platform administrators can also implement guardrails. So, workloads get the resources they need without taking more than they should. And through integration with Spark Atlas, DE automatically generates and captures lineage information.

Are You Ready to Operationalize Your Data Pipelines?

It's time to simplify the data management lifecycle by taking a modern approach to data engineering. Accelerate your enterprise data from ingest to insight at scale with CDP Data Engineering's ability to create, productionalize, and maintain quality data sets.

Visit our website to learn more about CDP Data Engineering.

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

Sources

[1] Gartner "Data Engineering Is Critical to Driving Data and Analytics Success," Robert Thanaraj, et al, 18 December 2019.

CLOUDERA