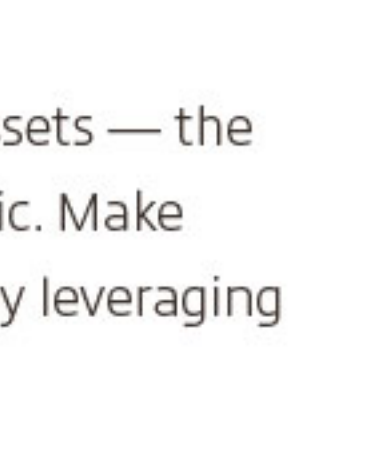


# How to Make the Most of DATABRICKS

Whether you're preparing to deploy Databricks® for the first time or looking to maximize the impact of your investment, use this checklist to ensure you're taking full advantage of the capabilities, features and tools this technology has to offer.



## 1 Are you properly securing your data environment?

Your organization's data is one of its most valuable assets — the consequences unauthorized access can be catastrophic. Make sure to create a highly secure analytics environment by leveraging Defense-in-Depth (DiD) security principles.

- Leverage firewalls and network security.
- Implement ACLs and identity management.
- Deploy securely at scale with integrated DevSecOps.
- Elevate security as one of your core design principles.
- Reduce your attack surface with a single, unified analytics tool like Databricks.
- Enable secure and transparent collaboration between data engineering and data science teams.



## 2 Are you properly governing your data environment?

Implementing an effective data governance solution can help your company ensure that effective rules are in place for data privacy, confidentiality, cost controls and data auditability.

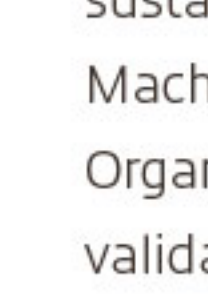
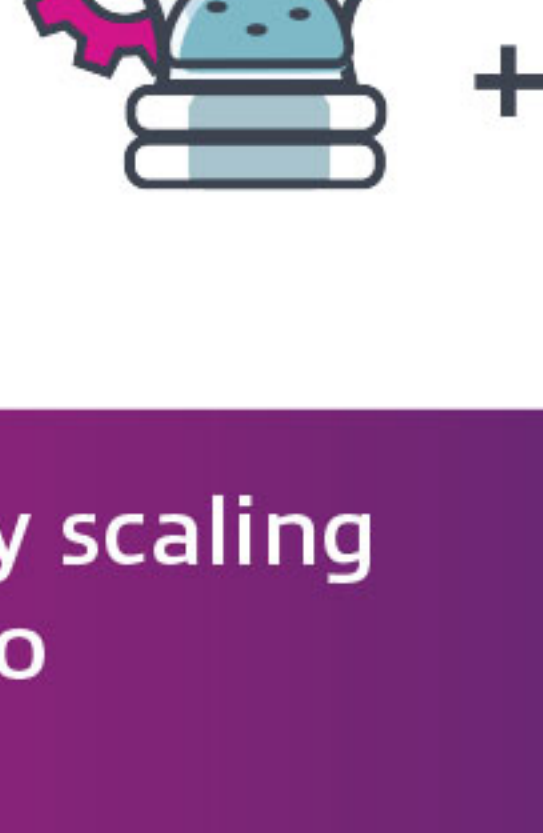
- Utilize cluster policies to manage cost and enforce security and compliance.
- Implement the Unity Catalog.
- Apply Data Lakehouse access control through a combination of security principals and carefully applied ACLs.
- Take advantage of cluster and workspace access control policies to grant appropriate access to team members.
- Leverage table access control.
- Use credential passthrough for high-concurrency clusters.
- Integrate auditing logging into your platform and overall development process.



## 3 Do you have observability into the health of your data environment and solution?

It's critical to understand the health of your environment. Being notified when things go wrong, with an understanding what happened, is key to supporting a sustainable solution. Ensure that you are following best practices receive high-quality alerts and log information.

- Configure Databricks workspaces to send statistics to logging solutions.
- Enable alerts for critical events (job failure, security violations, etc.).
- Consistently instrument your notebooks with logging statements.
- Bring logging and monitoring information together.



## 4 Are you successfully scaling machine learning to production?

Establishing and maintaining a sustainable, scalable solution for Machine Learning (ML) is hard work. Organizations need way to quickly validate data to be used for models, identify and track the best performing models, deploy to production and reuse common feature transformations.



Streamline and simplify your ML pipeline with:

- MLFlow
- AutoML™
- The Databricks Feature Store
- Support for AI/ML Ops



## 5 Do you have a scalable data framework?

As new platforms become successful, data analysts and scientists will require new datasets faster and will expect a high level of consistency in how the data is presented. Building frameworks for ingesting and processing becomes critical to be able to quickly bring on new datasets while building in that desired consistency. Many of these are configuration driven, which makes bringing on new data simple, without spending development cycles.

Use Databricks to:

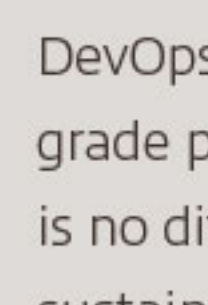
- Stay configuration driven.
- Respond quickly to new analytic needs.
- Expand who can onboard new data.
- Provide constancy.



## 6 Is your current Databricks solution performing as well as you'd like?

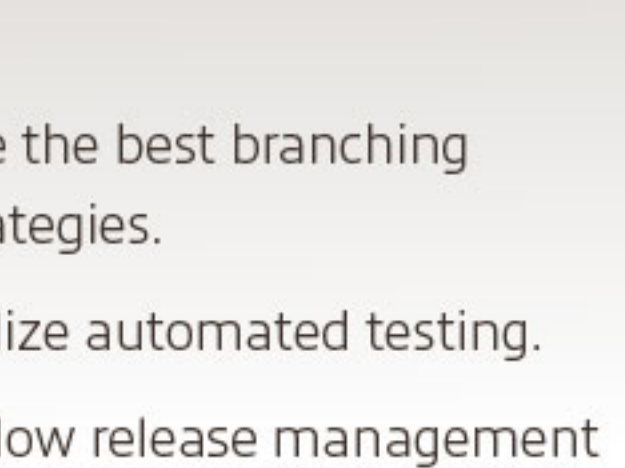
Simply implementing Databricks should vastly increase the performance of a data solution, but you can do more to increase the performance by utilizing all the features that Databricks makes available. To get the most performance and reduce costs, it's essential to follow best performance practices.

- Utilize the correct Spark settings and values.
- Correctly partition your data.
- Understand how to balance usage of user defined functions vs. native Spark operations.
- Understand when to use Python® vs. Scala.
- Choose the right storage formats,
- Identify the right times to use caching.

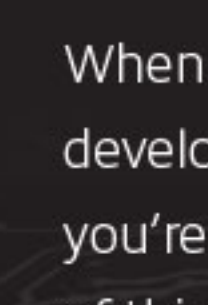


## 7 Are you set up for success by following proper DevOps practices?

From managing infrastructure and code to continuous deployment, following proper DevOps procedures is critical to any enterprise-grade platform. Your Databricks environment is no different. To ensure a healthy and sustainable solution, while enabling continuous integration, make sure that you are setup for success with your DevOps practices.



- Utilize Infrastructure as Code (IaC) to deploy your environments.
- Implement CI/CD best practices.
- Correctly manage your code.
- Use the best branching strategies.
- Utilize automated testing.
- Follow release management best practices.



## 8 Are you using all the development tools and practices at your disposal?

When most businesses get started with Databricks, the focus is on development using the web-based notebook environment. Are you sure that you're using all the features and processes available to maximize the value of this environment? Are there new design patterns in the tool that you want to utilize? There may be other processes and tools that can help you get the most out of your investment.

Be sure to explore and evaluate the following development options:

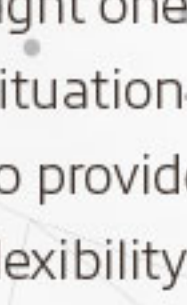
- VS Code™, PyCharm®
- Python vs. Scala
- Wheel files
- Code management, Git®
- Referenceable notebooks
- Functions
- Libraries



## 9 Are you using all the capabilities of streaming?

With Databricks, stream processing is highly flexible and scalable, but there is more to stream processing than reading from a queueing service.

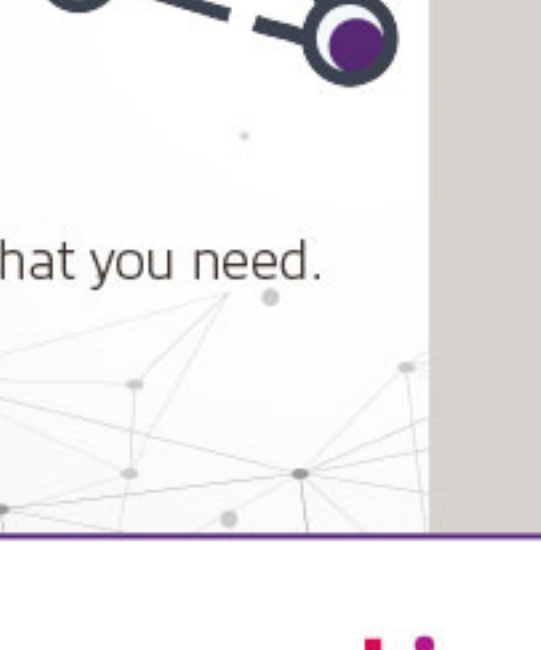
- Utilize the state tracking of streaming to prevent batch jobs from having to track input state.
- Save on compute costs by using streaming as needed — rather than using it continuously.
- Stream from cloud files using Auto Loader.
- Provide live data to BI tools to provide immediate intelligence.



## 10 Are you successfully scaling machine learning to production?

Are you storing your data in the right formats?

Spark supports a large number of data formats. Using the right one for the right situation is essential to provide the right flexibility performance that you need.



- Understand the advantages and disadvantages of using Delta Lake and other data formats.
- Choose the right compression for your data.
- Learn how to connect Spark to data storage that may not be supported out of the box (to avoid the cost of conversion).