Databricks Features

A Primer

databricks

Jobs

FEATURE	FUNCTION	BENEFIT
Flexible Scheduler	Execute jobs for production pipelines on a specified schedule	Ability to schedule jobs at minute to monthly intervals in different time zones; includes cron syntax.
Notifications	Send an email to a set of users whenever a production job starts, fails, and/or completes.	Notification on events (e.g. failures) including third party production pager integration; zero human intervention
Run Notebooks as Jobs	Existing notebooks can be executed as jobs.	Ability to run notebooks as well as standalone Apache® Spark™ applications and seemlessly migrate between exploration and production
Add Custom Spark Libraries and Applications to your Jobs	Existing Spark libraries and applications can be included in jobs.	Enhance your notebook jobs by including custom and legacy Spark libraries and applications.
Run Spark JARs as Jobs	Existing Spark JARs can be executed as jobs.	Reduce engineering efforts by the ability to run jobs against an existing Spark JAR.
Flexible Cluster Support	Ability to re-use existing clusters or launch on-demand instances (including spot instances).	Allocate desired resources to execute your job whether running on-demand clusters or making use of an existing one.
Job Retries	Jobs that fail can be configured to be automatically relaunched.	Enables production deployments, especially long-running applications such as streaming, to be automatically relaunched whenever a failure happens.
Run Log History	Historical run logs are saved and retrievable so even after the job has completed and the Spark Clusters have been shut down.	Allows you to debug, assess, and tune your code.

Account Management

Role-Based Access Control	Through the Accounts UI, easily create add or remove users to your Databricks account. Users can be marked as Administrators to manage deployment.	Account management is a foundational component for security and governance, risk, and compliance (GRC)

Notebooks

FEATURE	FUNCTION	BENEFIT
Interactive Workspace	User-friendly UI to apply notebooks to cluster and/or jobs.	An interactive workspace without investing time to integrate and maintain a 3rd party tool. A more seamless workflow providing efficacy and efficiencies.
Pipeline Workflows	Notebooks can be used for control workflow of pipelines in production jobs.	Data Scientists and Data Engineers can use the same notebook for creating models to production jobs.
One-click Visualizations	Provide a wide range of visualizations including full big-data pivoting, histograms, scatterplots, maps, etc	Easily create and embed within your notebooks your models and associated visualizations.
Collaboration	Designed for collaboration, notebooks contain features such as comments, viewer log, and history.	Allows team members to provide feedback and includes usage tracking.
Live Sharing and Editing	Real time collaboration among team members performining data modeling or analysis	Multiple team members can collaborate together across different geographic locations and time zones on the same notebook
Notebooks as Scripts	Ability to invoke notebooks that can invoke each other.	Allows you to build up notebook workflows where the results of one notebook is a requirement for subsequent notebooks
Autocomplete	Automatically completes function names and invocations within the notebook	Allows the developer, engineer, or data scientist to find and use the correct function faster.
Parameterized Queries	Parameterized Queries for notebooks	More easily re-use notebooks by utilizing parameterized queries.
Extensibility	Import any JAR or egg to be used by your notebook or job	Make use of popular libraries within your notebook or job such as scikit-learn, nltk ML, pandas, etc.
Multi-Language Support	Notebooks can be created using Python, Scala, SQL, or R (including markdown)	Make use of your favorite Data Sciences, Data Engineering, or Analyst language to create your notebook
One Click Publishing from Notebooks	Create shareable dashboards from notebooks with a single click. One notebook can be tailored into multiple dashboard views.	Eliminate additional steps and complex tools to share results with different audiences.

Notebooks (continued)

FEATURE	FUNCTION	BENEFIT
Collaboration	Designed for collaboration, notebooks contain features such as comments, viewer log, history, and github integration.	Allows team members to provide feedback and includes usage tracking.
Multi-Language Support	Notebooks can be created using Python, Scala, SQL, or R (including markdown). Databricks enables commands across language in cells.	Make use of your favorite Data Sciences, Data Engineering, or Analyst language to create your notebook
Schedule Notebooks	Execute jobs for production pipelines on a specified schedule directly from a notebook.	Ability to schedule jobs at minute to monthly intervals in different time zones; includes cron syntax.
Notebook Widgets	The widget API consists of calls to create different types of input widgets, remove them, and get bound values. Widgets can be created in SQL/Scala/Python/R.	Input widgets allow you to parameterize your notebooks.
Spark progress reporting and Spark UI integration	View real time progress of all the jobs and stages of a Spark command directly from the progress bar of a command run within a notebook.	Allows you to monitor Spark progress directly from your notebook.
One Click Publishing from Notebooks	Create shareable dashboards from notebooks with a single click. One notebook can be tailored into multiple dashboard views.	Eliminate additional steps and complex tools to share results with different audiences.

Dashboards

Continuous Dashboard Updates	Publish dashboards and schedule the content to be updated continuously.	Easily build dashboards to monitor critical operations.
Parameterized Dashboards	Provide drop-downs in the dashboards to enable changing input parameters to dashboard values.	Enable non-technical users to perform scenario analysis directly from published dashboards instead of manipulating code.
Schedule Dashboards	Execute jobs for production pipelines on a specified schedule directly from a dashboard.	Ability to schedule jobs at minute to monthly intervals in different time zones; includes cron syntax.
Dashboard Widgets	The widget API consists of calls to create different types of input widgets, remove them, and get bound values. Widgets can be created in SQL/Scala/Python/R.	Input widgets allow you to parameterize your dashboards.

Clusters

FEATURE	FUNCTION	BENEFIT
Easy-to-Use Cluster Management	User-friendly user interface simplifying the creation, restarting, and termination of clusters	Increase visibility to your clusters for easier manageability and help control costs.
High Availabiilty	If a worker instance is revoked or crashes, the Databricks cluster manager will relaunch it — transparent to the user.	Ensure your service is always up and running without the need to manage it yourself.
On-Demand Clusters	Build On-Demand clusters in minutes with a few clicks.	Build On-Demand clusters easily to meet the needs of your team or of your service.
Tuned and Optimized Out of the Box	Databricks Spark clusters are tuned for best performance by the experts who created Spark.	Spend time working on data models and analysis and little-to-no time optimizing the performance of your cluster(s).
Elasticity	Scale up or down your clusters based on your current needs.	Easy reconfiguration of your clusters ensure that your cluster will be the right size for your needs.
Spot Instances	Choose to create Spot instances for your clusters (and jobs)	Use spot instances for your cluster and/or jobs to help reduce costs
100% Spark Version Compatibility	100% Compatibility with Spark and support for multiple Spark versions.	Choose the version of Spark you want to use and feel safe knowing that there are no compatibility issues. As well, legacy jobs can continue to run without disruption since those can remain on previous versions of Spark.
Automatic Upgrades	Automatically upgrades your Spark clusters so you don't have to.	Get the latest version of Apache Spark hassle free.
Multiple Instance Types	Databricks provides multiple instance types including memory-optimized, compute-optimized, and GPU-accelerated instances.	Optimize your clusters for the profile of your workload (e.g. use compute-optimized instances for your machine learning workloads)
AWS Tag Support	Users can use AWS tags to assign metadata to each Spark cluster.	Easily track and attribute the usage of your AWS EC2 clusters to different groups.
Encrypted AWS Elastic Block Storage (EBS)	Users can attach encrypted EBS volumes to Spark clusters.	Long-running clusters will have additional storage, which provides more stability.

FEATURE	FUNCTION	BENEFIT
Data Management		
Data Sources Catalog	Central respository for your Spark data sources	As Data Engineers add or modify available data sources, these sources are immediately available to all users of the clusters.
Utilize Databricks File System	DBFS mounts are pointers to remote S3 paths.	Access S3 objects as if they were on the local file system.
Cache Tables Configuration	Cache your tables right from the Tables UI.	Improve query and processing performance by caching your tables to memory.
Import Data Seamlessly	Native integration with many data sources and data formats including CSV, many popular databases, JSON, Parquet, etc.	Easily import data from multiple data sources so your data engineers and data scientists can focus on the models instead of the data extraction.
Encryption at rest	Data is encrypted while stored in non-volatile memory.	Encryption at rest is a foundational component for security and governance, risk, and compliance (GRC)
Integration		
Integrate with Databricks Using REST APIs	Databricks provides a rich set of REST APIs cluster management, DBFS, jobs, and libraries.	Programmatically interact with the Databricks platform using REST APIs to integrate your tools or services with the Databricks platform.
Integrate with Your Favorite 3rd Party Tools	Integrate with your favorite 3rd party tools including Tableau, Pentaho, Qlik, PanTera, TIBCO Jaspersoft, and ZoomData	Use the Databricks platform with your favorite 3rd party tool.
Integrate with Your Favorite Data Sources	Integration with your favorite Data Sources including: - SQL stores (JDBC/ODBC) - NoSQL stores (Cassandra, HBase) - Columnar stores (Redshift, Vertica) - Document-oriented stores (MongoDB) - Hadoop and Hive including custom UDFs, UDAFs, and UDTs - File stores (S3, AWS/EFS coming soon) - File formats (CSV, JSON, Parquet, SequenceFile, Avro, RCFile, ORCFile) - Search engines (Lucene, SOLR, ElasticSearch)	Easily work with your favorite Data Sources within Databricks and Apache Spark.

FEATURE	FUNCTION	BENEFIT
Support		
Support and Community Forums	Dedicated support and community forums for the Databricks Platform	Engage through forums with the Databricks field engineers supported by the Databricks engineers leading the development of Apache Spark.
Direct Premium Support	Dedicated Databricks field engineers to support you with your solutions.	Engage directly with the Databricks field engineers supported by the Databricks engineers leading the development of Apache Spark.
Continuously updated Databricks Guide	The Databricks Guide is constantly updated with the very latest examples, tutorials, and datasets.	Central resource to reference all of the very latest Apache Spark and Databricks has to offer.
Sample Applications & Tutorial in Databricks	Databricks is continually updated with sample applications and training.	Central resource for sample applications and the latest Apache Spark training by Databricks.
Security	*Databricks has completed SOC 2 Type 1 certification, and can offer a HIPAA-compliant service*	
Role-Based Access Control	Through the Accounts UI, easily create, add, or remove users to your Databricks account. Users can be marked as Administrators to manage deployment.	Account management is a foundational component for security and governance, risk, and compliance (GRC)
AWS IAM Roles Integration	Users can use their IAM credentials to access AWS resources from Databricks.	Provides secure access to AWS resources to diverse user groups in the same organization.
Single Sign-On	A central authority, such as the CISO, will also be empowered to revoke access via a SAML 2.0 compatible Identity Provider service to safeguard enterprise resources as needed.	Simplify security context by reducing the amount of passwords users have to remember and provides convenience to the users.
Notebook Access Control List	Whether individuals have permissions to read, run, create, or delete notebooks.	Provide mechanisms for users to protect confidential information while allowing code reuse.
Cluster Access Control List	Whether individuals have permissions to create, attach to, terminate, or invite other users to Spark clusters that are launched.	Gives a central authority the ability to protect the access to production compute resources, or to limit the expenditure associated with launching new resources to a few trusted entities.
Audit logs	A complete record of which individuals performed what action on the Databricks platform.	Allows Databricks customers to meet compliance standards such as SoC2, and to monitor or investigate detailed usage patterns of Databricks as the business requires.