Microsoft Azure Databricks for data engineering

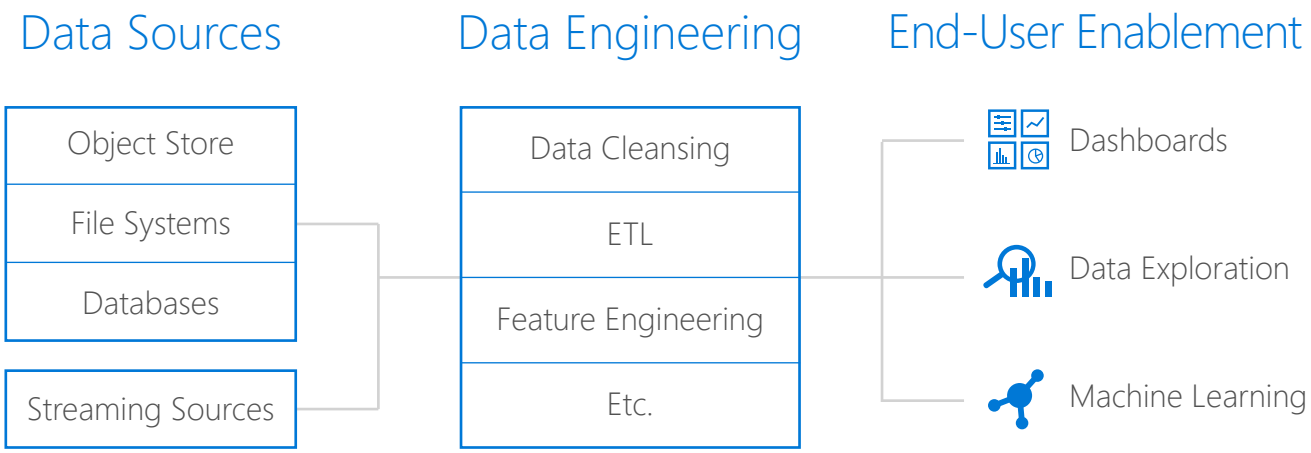
Building production data pipelines with Apache® Spark™ in the cloud

Microsoft

Azure Databricks

As companies continue to set their sights on making data-driven decisions or automating business processes with intelligent algorithms, mastering data engineering is a business necessity.

Data engineers perform mission-critical data cleansing, transformations, and manipulations to make business-use cases such as real-time dashboards or fraud detection possible. Common data engineering tasks include Extract, Transform, and Load (ETL) of unstructured data into a data warehouse or feature engineering to train machine learning algorithms.



Key data engineering requirements

Production readiness

Business-critical data pipelines, whether ETL or feature engineering, must not go down. Every outage has the potential to cost millions of dollars in lost revenue. If an outage were to occur, the data engineering team must be able to easily determine the root cause and remediate the problem to limit the damage. The data infrastructure must also meet the necessary data protection and compliance standards.

Any scale performance

As the volume and variety of data grows to support more sophisticated needs, the data infrastructure must be easy to scale up from small volumes of structured data to large volumes of unstructured data. At the same time, transient heavy loads (i.e., due to seasonal business needs) should not translate to permanently high infrastructure costs. The infrastructure should incur minimal costs while idle.

Compatibility with a wide range of data stores and tools

Data often exists in silos. The data engineering team must be able to easily connect to wherever valuable data resides, whether it's an object store in the cloud, a traditional data warehouse, or a streaming data source. In addition to data stores, data engineers also use a wide variety of software tools for continuous integration (e.g., Jenkins) and workflow management (e.g., Airflow). Consequently, they need a solution that can easily integrate with other technologies using well-defined interfaces and APIs.

Data engineering and Apache Spark

Apache Spark is the largest open source project in data processing, with 1,000+ contributors from 250+ organizations, and it's the data engineering technology of choice for market leaders such as Facebook, Uber, Netflix, and Tencent. These organizations, and many more like them, turn to Apache Spark on Azure Databricks for a variety of reasons, including:

Flexibility

Spark is capable of applying SQL, machine learning, and graph processing to big and small data. It can process data in batches as well as stream in real time, and developers can use the language they are most comfortable or familiar with such as SQL, R, Java, Python, or Scala.

100x Performance

Spark is 100x faster than the legacy-distributed computing framework MapReduce. This means not only faster data analysis but faster, more relevant insights for your business.

Developer-centric

It is much easier to analyze data and write Spark applications because Spark's API is far simpler to use than legacy frameworks such as MapReduce. The Microsoft Azure Databricks platform provides essential security, reliability, usability, and management services around highly performant Spark versions in the cloud to simplify data engineering.

Azure Databricks: the power you need for Spark-based analytics

Turnkey solution

Deploying production Spark pipelines with Microsoft Azure Databricks does not require specialized tools or resources. Using Spark, you can automatically recover from failures without human oversight and diagnose and solve outages and performance degradations. The platform's high-performance processing helps accelerate productivity by simplifying and streamlining processes and workflows. As additional backup, Azure Databricks' Spark committers are on standby to resolve the most challenging reliability and performance issues with your Spark data pipeline.

Apache Spark clusters optimized for the cloud, tuned and supported by committers

The Azure Databricks Spark clusters are tuned, maintained, and supported by Spark committers specifically for the cloud. We have also built extensive performance, scalability, and reliability features around Spark to take advantage of the cloud environment. Our innovations enable you to run multiple Spark versions on a wide variety of instance types and automatically scale clusters based on load.

Seamless integration with tools and data stores

Databricks is ready to be plugged into a wide variety of Azure tools and data stores such as Azure Storage Blob, Azure Data Lake Store, Azure Event Hubs, and Azure Cosmos DB. Integration with various client tools like Power BI, Tableau, and plugins for IDEs is also available.

Lower total cost of ownership

Azure Databricks' performance-tuned Apache Spark clusters allow you to complete jobs faster, reducing cloud compute costs. The fully managed Azure Databricks Spark clusters enable you to further reduce costs by avoiding time-consuming tasks to build, configure, and maintain complex Spark infrastructure.

Key data engineering features

	Feature	Benefit
Production readiness	**Turnkey security and compliance standards**	Launch Spark clusters with end-to-end encryption in an environment instantly.
	Fault-tolerant Spark jobs	Run JARs and notebook jobs that automatically alert you and recover from failure without human intervention.
	Detailed and easily accessible logs	Access end-to-end logs easily for debugging.
	Spark support from committers	Escalate bugs to Spark committers for deep technical support.
Performance optimization	**Tuned and optimized Spark versions**	Tune spark committers for a wide variety of instance types, including compute, memory, and storage-optimized types as well as GPUs.
	Autoscaling clusters	Scale up Spark clusters automatically to match a surge in demand, and scale down to avoid idle resources.
Integration	**Comprehensive API**	Launch clusters, jobs, and everything necessary for data engineering with REST API.
	Secure SQL server	Connect to tools such as Power BI and Tableau via an encrypted ODBC interface.
	Data sources catalog	Create a central repository of Spark data sources, making every data source immediately available to all Azure Databricks users without duplicating data ingest work.
Manage TCO	**Cluster tagging**	Associate cluster usage with user groups to track resources used and attribute costs.
	Automatic patching / upgrades	Get new Spark versions and bug fixes automatically without incurring DevOps effort.

Azure Databricks

Learn more