Government Efficiency through Innovative Reform

IBM watsonx.ai

Pricing Document

© Copyright IBM Corporation 2025

Pricing Document

IBM watsonx.ai studio IBM Cloud

Plan	Features and capabilities	Pricing
Lite	**1 authorized user** 10 capacity unit-hours monthly limit Environment = # of capacity units required per hour • 1 vCPU + 4 GB RAM = 0.5 • 2 vCPU + 8 GB RAM = 1 • 4 vCPU + 16 GB RAM = 2 • Decision Optimization + Watson NLP = Environment + 5 • Synthetic Data Generator, 2 vCPU + 8 GB RAM = 7 (requires watsonx.ai Runtime) The Lite plan offers most watsonx.ai Studio data science and AI features with usage restrictions. **Lite plan services are deleted after 30 days of inactivity.**	Free
Professional	**Unlimited collaborators** Unlimited elastic compute environments Environment = # of capacity units required per hour • 1 vCPU + 4 GB RAM = 0.5 • 2 vCPU + 8 GB RAM = 1 • 4 vCPU + 16 GB RAM = 2 • 8 vCPU + 32 GB RAM = 4 • 16 vCPU + 64 GB RAM = 8 • 40 vCPU + 172 GB RAM + 1 NVIDIA V100 (1 GPU) = 68 • 80 vCPU + 344 GB RAM + 2 NVIDIA V100 (2 GPU) = 136 • Decision Optimization + Watson NLP = Environment + 5 • Synthetic Data Generator, 2 vCPU + 8 GB RAM = 7 (requires watsonx.ai Runtime) NVIDIA V100 GPU environments available only in Dallas on IBM Cloud	$1.02 USD/Capacity Unit-Hour

IBM watsonx.ai runtime IBM Cloud

Plan	Features and capabilities	Pricing
Lite	**Service instance** Instance includes: • 20 capacity unit-hours (CUH) per month • 50,000 tokens/data points per month • 100 pages per month --- Foundation models: • Inferencing for text generation consumes tokens (as Resource Units) • Token usage is the sum of input and output tokens • Time series forecasting consumes data points (as Resource Units) • Data point usage is the sum of input and output data points --- Text extraction: • Each document page or image file or .tiff frame is considered 1 Page --- Machine learning training tools: • Compute usage counted as CUH • CUH rate based on training tool, hardware specification, and runtime environment --- Machine learning deployments: • Compute usage counted as CUH • CUH rate based on deployment type, hardware specification, and other factors • Maximum parallel Decision Optimization batch jobs per deployment: 2 • Default deployment jobs retained per deployment space: 100 CUH billing is based on CUH rate multiplied by the hours of consumption. Resource Units billing is based on the rate of the foundation model class multiplied by the number of tokens or data points used. For details on CUH rates, go to https://ibm.biz/wx-plans-ml For details on foundation model classes, go to https://ibm.biz/wx-plans-gen-ai 1 Resource Unit = 1000 tokens/data points **Lite plan services are deleted after 30 days of inactivity.**	Free

Essentials	**Service instance** Foundation models: • Inferencing for text generation consumes tokens (as Resource Units) • Token usage is the sum of input and output tokens • Time series forecasting consumes data points (as Resource Units) • Data point usage is the sum of input and output data points • Compute usage for Tuning Studio is 43 CUH per hour --- Text extraction: • Each document page or image file or .tiff frame is considered 1 Page • Each Page is charged at a flat rate (text extraction category 1 rate) • Text extraction category 2 rates are not applicable at this time --- Machine learning training tools: • Compute usage counted as CUH • CUH rate based on training tool, hardware specification, and runtime environment --- Machine learning deployments: • Compute usage counted as CUH • CUH rate based on deployment type, hardware specification, and other factors • Maximum parallel Decision Optimization batch jobs per deployment: 5 • Default deployment jobs retained per deployment space: 1000	$0.52 USD/Capacity Unit-Hour $0.01 USD/Mistral Large Output Resource Unit $0.038 USD/Page - Text Extraction Category 1 $0.019 USD/Page - Text Extraction Category 2 $0.003 USD/Mistral Large Input Resource Unit $0.0001 USD/Resource Unit (IBM models) $0.0001 USD/Resource Unit (Third party models)
Standard	**Service instance** Instance includes: • 2500 capacity unit-hours (CUH) monthly for compute usage --- Foundation models: • Token usage is the sum of input and output tokens • Inferencing for text generation and unstructured synthetic data generation consumes tokens (as Resource Units) • Time series forecasting consumes data points (as Resource Units) • Data point usage is the sum of input and output data points • Compute usage for Tuning Studio is 43 CUH per hour --- Text extraction: • Each document page or image file or .tiff frame is considered 1 Page • Each Page is charged at a flat rate (text extraction category 1 rate) • Text extraction category 2 rates are not applicable at this time --- Custom Foundation Models (Dallas only): • Hourly rates vary based on the hardware configuration, and include model hosting and inferencing costs --- Machine learning training tools: • Compute usage counted as CUH • CUH rate based on training tool, hardware specification, and runtime environment --- Machine learning deployments: • Compute usage counted as CUH • CUH rate based on deployment type, hardware specification, and other factors • Maximum parallel Decision Optimization batch jobs per deployment: 100 • Default deployment jobs retained per deployment space: 3000	$1,050.00 USD/Instance per month $0.42 USD/Capacity Unit-Hour $0.01 USD/Mistral Large Output Resource Unit $5.22 USD/Hour - Small Model Hosting $10.40 USD/Hour - Medium Model Hosting $20.85 USD/Hour - Large Model Hosting $0.03 USD/Page - Text Extraction Category 1 $0.015 USD/Page - Text Extraction Category 2 $41.70 USD/Hour - Extra Large Model Hosting $0.003 USD/Mistral Large Input Resource Unit $0.0001 USD/Resource Unit (IBM models) $0.0001 USD/Resource Unit (Third party models) $4.43 USD/Hour - Extra Small Model Hosting $83.50 USD/Hour - Very Large Model Hosting $34.30 USD/Hour - Mistral Large Model Hosting Access $0.00625 USD/Resource Unit - InstructLab Data $0.007 USD/Resource Unit - InstructLab Tuning $8.60 USD/Hour - Mistral 1 GPU Model Hosting Access $17.20 USD/Hour - Mistral 2 GPU Model Hosting Access $5.80 USD/Hour - Model Hosting (1A100 80GB) $14.50 USD/Hour - Model Hosting (1H100 80GB) $16.00 USD/Hour - Model Hosting (1H200 141GB) $4.43 USD/Hour - Model Hosting (1L40S 48GB) $11.60 USD/Hour - Model Hosting (2A100 160GB) $29.00 USD/Hour - Model Hosting (2H100 160GB) $32.00 USD/Hour - Model Hosting (2H200 282GB) $8.86 USD/Hour - Model Hosting (2L40S 96GB) $23.20 USD/Hour - Model Hosting (4A100 320GB) $58.00 USD/Hour - Model Hosting (4H100 320GB) $64.00 USD/Hour - Model Hosting (4H200 564GB) $46.40 USD/Hour - Model Hosting (8A100 640GB) $116.00 USD/Hour - Model Hosting (8H100 640GB) $128.00 USD/Hour - Model Hosting (8H200 1128GB)

IBM has a strategic MoU with Crown Commercial Services (CCS) which is applicable to all Public Sector Customers buying via G-Cloud. This provides agreed baseline levels of discounts and details higher levels of discounts for larger volumes or greater durations of contract. The IBM pages describing this are at https://ibm.biz/ccs_mou. Public Sector clients can request a full copy of the IBM MOU from CCS by sending an email to info@crowncommercial.gov.uk

The pricing in this G-Cloud OLering takes account of the MOU baseline discounts. As IBM's portfolio is somewhat complex and there are several factors that aLect the pricing; if your requirements are for more than just one license for a month, we would suggest that you contact us directly to ask us to provide you with pricing with the correct levels of discount for the term and volumes that you require. The email address for G-Cloud related enquiries is
ukcat@uk.ibm.com. Please ensure that you indicate which G-Cloud OLering(s) you are requesting details for.