

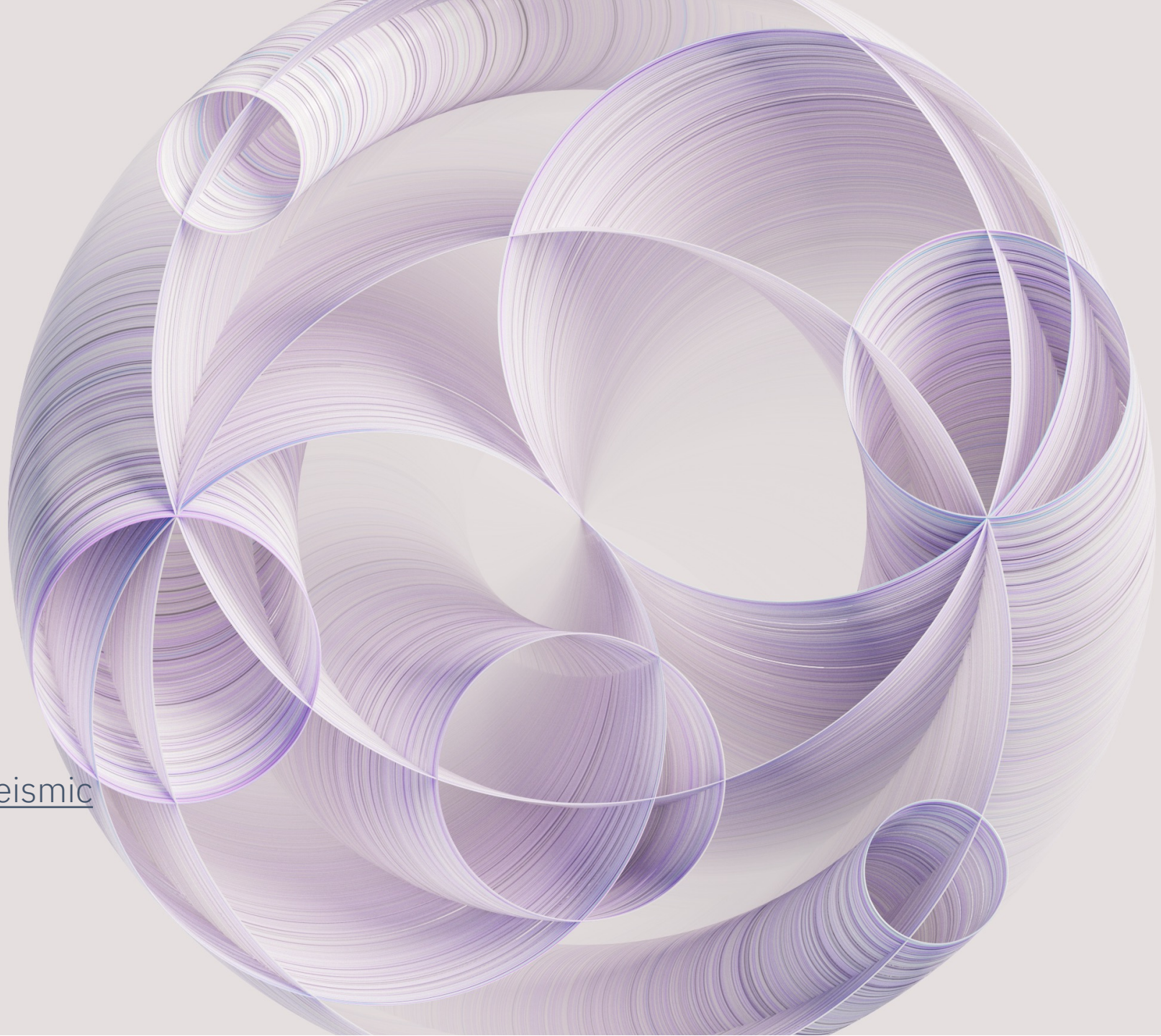


Parts & Pricing Enablement

July 17, 2023

[Find the latest version of this deck on Seismic](#)

IBM and Business Partner
Internal Use Only



Seller guidance and legal disclaimer

IBM and Business Partner
Internal Use Only

Slides in this presentation marked as **"IBM and Business Partner Internal Use Only"** are for IBM and Business Partner use and should not be shared with clients or anyone else outside of IBM or the Business Partners' company.

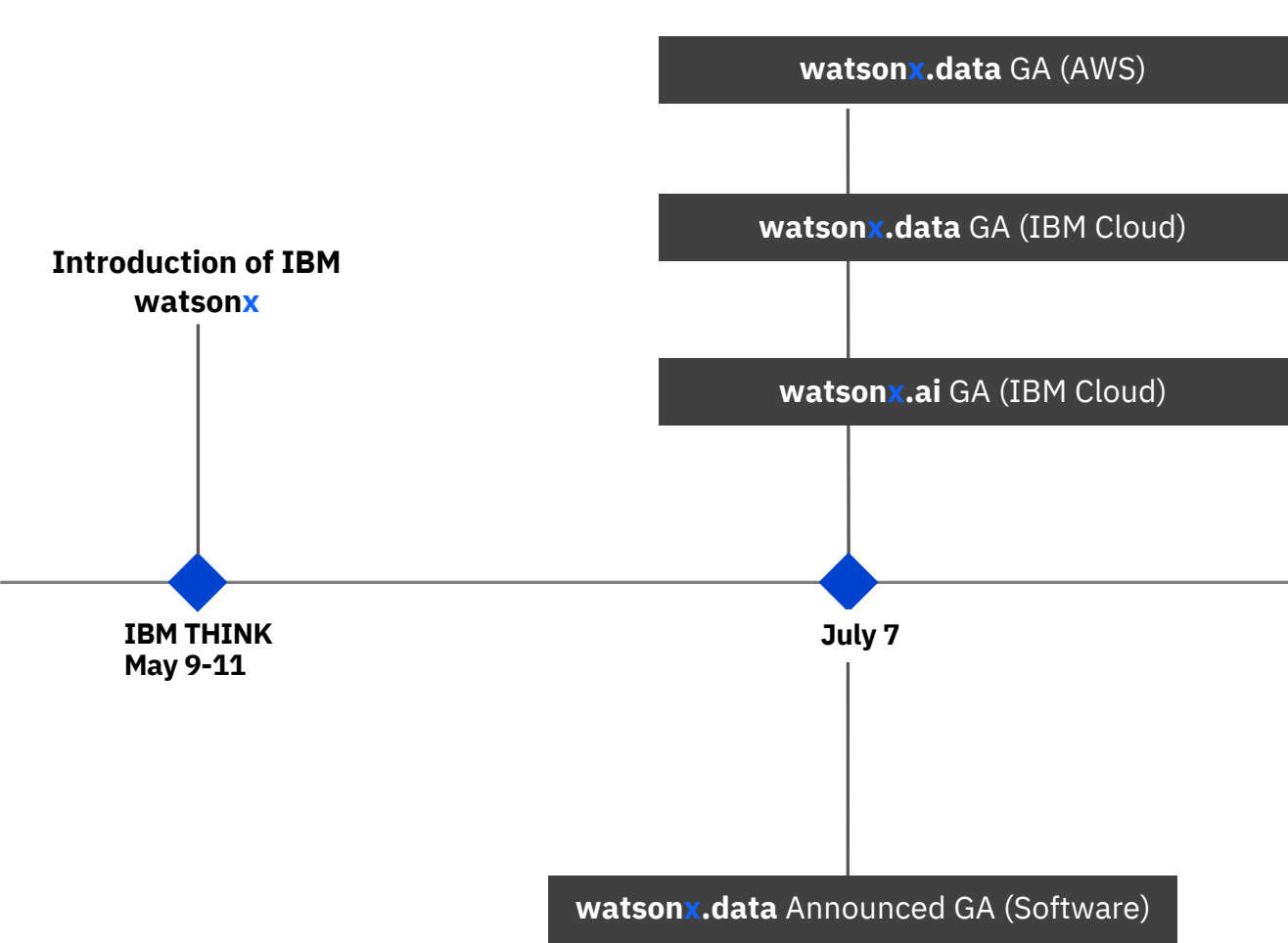
© IBM Corporation 2023.
All Rights Reserved.

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth, or other results.

All client examples described are presented as illustrations of how those clients have used IBM products and the results, they may have achieved. Actual environmental costs and performance characteristics may vary by client.

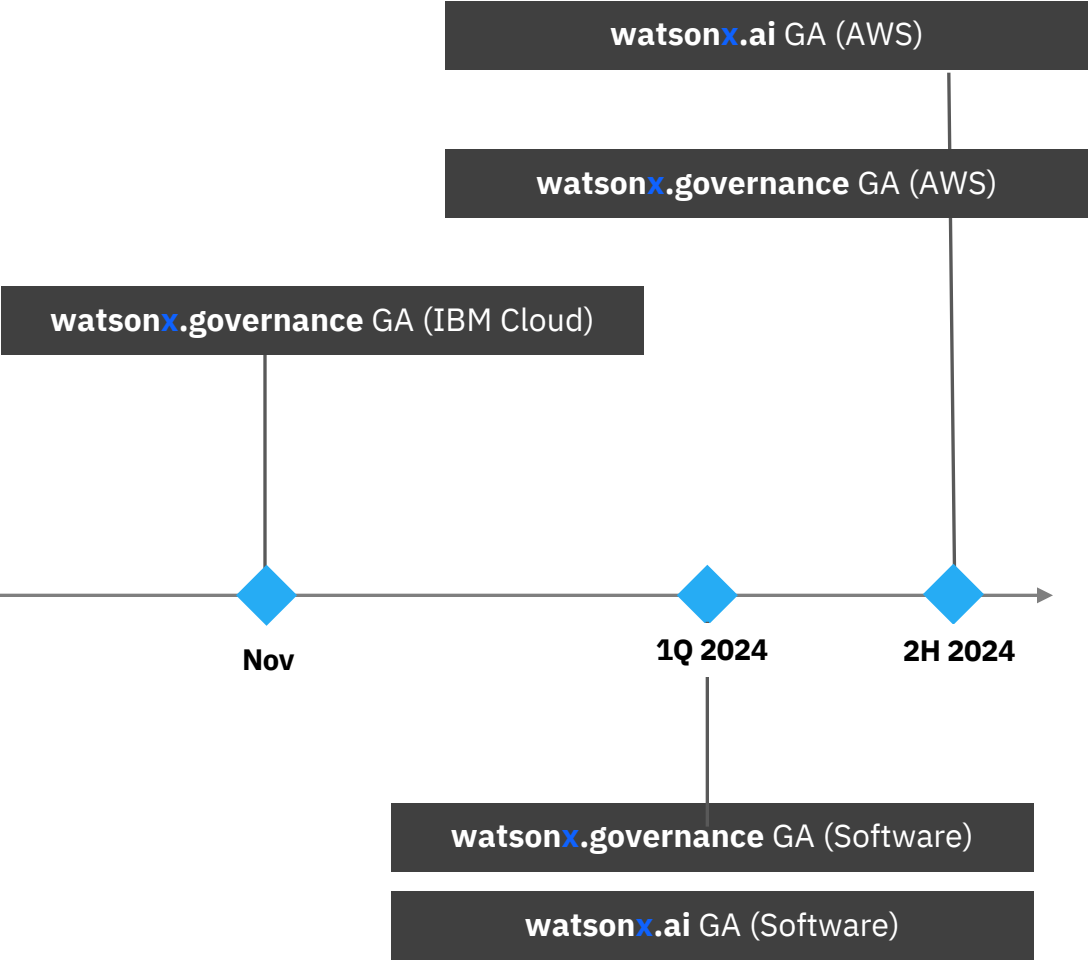
Available!



Complete

Target


Coming soon!



watsonx SaaS Launch (Tier Fee + Consumption)

Trials available via IBM Cloud

watsonx Tier Fee			
Tier fee is a fixed per instance charge for access to watsonx. Higher tiers grant access to additional capabilities. Tier fee does not cover usage fees.			
	Essentials: \$0 / Month	Standard: \$1500 / Month	Premium: TBD / Month
watsonx .ai	<ul style="list-style-type: none">ML functionalityInferencingPrompt LabOpensource models <p>Mapped to WML Standard at launch</p>	<ul style="list-style-type: none">ML functionalityInferencingPrompt LabOpensource models <p>Note: Mapped to WML Professional, with \$1050 tier fee at launch</p> <ul style="list-style-type: none">Prompt Tuning *Model hosting (limit 1 model) *BYOM *Synthetic Data: Paid Add-on *	Development Underway
watsonx .data		<ul style="list-style-type: none">Hive metastore and Iceberg catalogInfrastructure manager and query editorAhana Presto query engine and connectorsSpark, Db2Wh, and Netezza integration <p>Note: No Tier fee at Launch</p>	Development Underway
watsonx .gov		Development Underway	Development Underway

 = available at launch  = Not at launch

* Rolling delivery through 2H, not on 7/7

watsonx SaaS Subscription

The SaaS parts you will sell to watsonx.ai and watsonx.data SaaS Customers

How the watsonx SaaS Subscription works

There are two types of parts

- Per Month parts for regular usage
- Overage parts for overage usage and fees
- AI and IA sellers are compensated on the subscription

Per Month Parts

- Think of these as a gift card. Customers commit to a minimum spend towards watsonx* for a time period
- Customers usage of watsonx services will count against (deduct from) their gift card spend at rates designated by the product
- Committed spend is billed at a monthly
- Discounting available for spend
- Parts are priced in local currency values
- Spend is retained for the length of the subscription term

Overage Parts

- When customers run out of their Subscription spend, they are charged for overage usage. These parts are required to cover those situations
- Parts are priced in local currency values
- Discounting not available for overage parts

How to use the watsonx SaaS Subscription

- Quote the watsonx as a Service per Month part in your customer's local currency for their desired amount of spend
- Quote must also include a single unit of the Overage part in customer's local currency to cover potential overage fees
- All parts listed [here](#)

How can existing SaaS customers use watsonx

- watsonx saas services can be consumed with available IBM cloud credits previously purchased by other subscriptions

What's available on Day 1 (more details)

watsonx.ai (SaaS)

- **Opensource generative models:** flan-ul2-20b, gpt-neox-20b, mt0-xxl-13b, flan-t5-xxl-11b, mpt-instruct2-7b.
- **IBM Slate Family**, fine-tuned for entity extraction, relationship detection, and sentiment analysis.
- **Prompt engineering with Prompt Lab** (both GUI and REST API) that enables inferencing on generative models. Support for 5 key IBM generative AI tasks (Q&A, summarization, content generation, fact extraction, classification)
- **Fine tuning** for Developer persona via notebook and APIs for IBM Slate models for classification and entity extraction use cases
- **Data Refinery** - Visual data wrangling tool for the Data Scientist persona
- **Feature Store** - Create a store of features relevant to Data Science
- **AutoAI** - Automate data preparation, model selection, training, hyper parameter tuning and evaluation.
- **Modeler Flow** - Visual modeling environment [SPSS Modeler]
- **Jupyter notebook** - Coding environment in Notebooks
- **RStudio** - R Coding IDE
- **Decision Optimization** - Optimize solving business problem scenarios [based on CPLEX]
- **Federated Learning** - Train models on remote parties without sharing data.
- **Pipelines** - Automate end-to-end data science process.
- **Core capabilities** (jobs, parameter sets, projects, deployment spaces)

watsonx.data (SaaS and Software)

- **Multi-cloud, hybrid cloud availability:** Supporting both SaaS and self-managed software deployment models, or a combination of both,
- **Ahana Presto engine:** Ahana Presto is an open-source, fast, reliable, and highly scalable SQL query engine
- **Multi-engine integration:** Ability to associate Db2 Warehouse, Netezza and Spark with watsonx.data so these engines can access the same data and metadata as well as work with the Iceberg table format
- **Open data and table format support:** Support for Apache Iceberg. Iceberg time travel and rollback features in watsonx.data will enable the ability to examine table changes over time and rollback quickly to a previous state.
- **Enterprise compliance and security:** Basic access control and governance through Casbin
- **Easy to use, integrated data console:** Brand new and modern UI that enables users to easily navigate and manage their data on the new lakehouse architecture
- **Ecosystem integrations:**
 - Presto connectors to CPD
 - Integration with Db2 warehouse will also enable Z data to flow into the watsonx.data ecosystem via data gate

Major FAQs

Will customers be able to draw down their IBM Cloud spend on this service

YES!

Can customers draw down their AWS Credits via the AWS marketplace for watsonx.data?

YES for watsonx.data!

Is the Tier fee per product or across watsonx?

The tier fee spans watsonx – i.e. a customer paying for Standard will cover their tier fee for both .ai and .data

Is there watsonx.ai pricing available On-Prem?

Work in progress! Pre-announce pricing via L112 not yet available.

Since CP4DaaS users get access to watsonx, does that mean CP4D subscription credits can be used towards watsonx or something else?

YES!

Are there trade ups for software customers who are already using CP4D or products on CP4D?

These are WIP. More news to come.

Is RHOS included?

For watsonx.data SW, YES. For SaaS, N/A.

watsonx.ai

watsonx Pricing

watsonx.ai (SaaS ONLY at GA)

- Foundation models are typically charged by 3 components:
- **Inference**= using a GAI model to process natural language text
 - **Tuning** = using GPUs to tune your GAI model
 - **Hosting** = hosting a stable, trained version of a GAI model

Pricing of FMs (Opensource + BYO model)

Model Size (parameters)	Inference per 1K tokens per model (USD)	Tuning per compute hour (USD)	Hosting per hour (USD)
Class 1 (<10B)	OS: \$0.0006	\$22	\$1.00
Class 2 (10B-19B)	OS: \$0.0018	\$24	\$3.50
Class 3 (20B+)	OS: \$0.0050	\$26	\$6.00

IBM models will be offered in future watsonx.ai releases and charged at a premium versus OS models. Use-case specific models may be developed and monetized at specific rates.

Pricing for ML Functionality: \$ per Capacity Unit Hour (CUH)

- **Machine Learning** (formerly WML)
 - If on **essentials plan**- pay as you go at **\$0.52/CUH**
 - If on **standard plan**- first 2500 CUH no charge, after 2500 CUH **\$0.42/CUH**
- **ML Tools** (formerly Watson studio)
 - Pay as you go **\$1.02 per CUH**

Back of the napkin pricing for foundation models

- **1 token** ~= 4 characters or 3/4 of a word
- **100 tokens** ~= 75 words
- **Reference point:** the collected works of Shakespeare ~ 900,000 words or 1.2M tokens.
- **Total token count** = (total use case wordcount * 1.33)

Calculating token count example:

- A marketer using a class 3 model to create content and wants to generate 25 social media posts per month
- Uses ~15 word prompts to generate ~75 words of content
- Find number of tokens per post: (15+75=90 words, 90*1.33 = ~120 tokens)
- Calculate against monthly posts (25) to get to total tokens per month= 3000



What drives compute for Machine Learning:

- Auto AI
- DO training and deployments
- ML deployments
- ML models via API (training, evaluating, or scoring)
- Pipelines if invoked by one of the above services

What drives compute for ML Tools:

- Notebook editor
- Data Refinery
- SPSS Modeler
- RStudio IDE
- NLP library/models
- Pipelines if invoked by one of the above services

Note: Prompting does not consume CUH for ML or ML Tools

watsonx.ai SaaS sample T-Shirt Sizes

	Small	Medium	Large	X-Large
Hosting	Not Available in Essentials	\$732 / month 24/7 hosting of a Class 1 model	\$4,392/ month 24/7 hosting of a Class 3 model	\$4,392/ month 24/7 hosting of a Class 3 model
Tuning	Not Available in Essentials	\$66 / month 3 hours of tuning per month, on a Class 1 model	\$312 / month 12 hours of tuning per month, on a Class 3 model	\$312 / month 12 hours of tuning per month, on a Class 3 model
Inferencing	\$420 /month >80M tokens used	\$1,868 / month >3.1B tokens used	\$2,129 / month >425M tokens used	\$14,596 / month >2.9B tokens used
Tier Fee	\$0 Essentials	\$1,500 / month Standard	\$1,500 / month Standard	\$1,500 / month Standard
Total	\$5,000 / year ~\$420 / month	\$50,000 / year ~\$4,200 / month	\$100,000 / year ~\$8,300 / month	\$250,000 /year ~\$20,800 / month
Notes / Sample Use-case	<ul style="list-style-type: none">Marketing agency using GAI to generate social media contentZero-shot prompting (i.e. not training) against OOTB a large open-source modeUsing Essentials	<ul style="list-style-type: none">Customer support center using >3B tokens per month to summarize internal documents for agentsUsing a trained model for support staff 24/7, using Standard for Hosting + Tuning capability	<ul style="list-style-type: none">Bank using >425M tokens per month to summarize industry reports, extract key terms and entities, to arm wealth managers for client conversationsUsing Standard for Hosting + Tuning capability to iteratively improve model outputs	<ul style="list-style-type: none">Same Bank as Large – scaling up their use case to automatically summarize key financial news to end customers via their banking mobile appUsing Standard for Hosting + Tuning capability to iteratively improve model outputs

Pricing of FMs (Opensource + BYO model)

Model Size (parameters)	Inference per 1K tokens per model (USD)	Tuning per compute hour (USD)	Hosting per hour (USD)
Class 1 (<10B)	OS: \$0.0006	\$22	\$1.00
Class 2 (10B-19B)	OS: \$0.0018	\$24	\$3.50
Class 3 (20B+)	OS: \$0.0050	\$26	\$6.00

GA Models

IBM Foundation Models

Slate (encoder-only) Natural Language Processing Models

CLASS 1

Slate
153 million params
multilingual distilled

Fine Tuning Required to support:

Extract

Classify

Additional IBM Models WIP

Open-Source Large Language Models



Encoder/decoder & decoder-only Large Language Models available in *Prompt lab*
(Tuning *NOT* required for most tasks)

CLASS 3

flan-ul2-20b
20 billion params
encoder/decoder

gpt-neox-20b
20 billion params
decoder only

Q&A

Generate

Extract

Summarize

Classify

Q&A

Generate

CLASS 2

mt0-xxl-13b
13 billion params
encoder/decoder

flan-t5-xxl-11b
11 billion params
encoder/decoder

Q&A

Generate

Extract

Summarize

Classify

Q&A

Generate

Summarize

Classify

CLASS 1

mpt-instruct2-7b
7 billion params
decoder only

Q&A

Generate

Language Tasks

Q&A

Model responds to a question in natural language

Generate

Model generates content in natural language

Extract

Model extract entities, facts, and info. from text

Summarize

Model creates summaries of natural language

Classify

Model classifies text (e.g., sentiment, group)

Model variety to cover enterprise use cases and compliance requirements

[Full model roadmap available on Seismic](#)

Note: Slate models can be fine-tuned via notebooks and APIs

watsonx.ai SaaS sales scenario (inference)

- Know that 750 words roughly equals 1k tokens
- Ask your customer the average amount of words they might ask or want for their use case
- If the customer is unsure of the number of words they might use, introduce the trial.
- Customers will be able to look at the cumulative total of their trial on the resource usage page. Specifically, the customer will be able to see what model they used and identify its class.
- We categorize models into three buckets

Model Size (parameters)	Inference per 1K tokens per model (USD)
Class 1 (<10B)	OS: \$0.0006
Class 2 (10B-19B)	OS: \$0.0018
Class 3 (20B+)	OS: \$0.0050

- Calculate number of tokens per 1000 tokens using respective cost.
Ex: If the resource page reflects 5 Million tokens for a class 3 model
 - $5,000,000 \text{ tokens} / 1,000 = 5,000$
 - $5,000 * .005 = 25 \text{ dollars per month}$

Prompt Lab

New (unsaved)

New prompt + Save work v

Sample prompts < Structured Freeform Model: flan-ul2-20b </> ⚙

Summarization

Meeting transcript summary
Summarize the discussion from a meeting transcript.

Earnings call summary
Summarize financial highlights from a quarterly earnings call.

Classification

Scenario classification
Classify scenario based on

20

John Doe 00:04:47.764 --> 00:04:48.664

On the same dataset.

Summary:
John and Jane are trying to replicate the results from the last analysis. They found out that the testing of the downstream classifier was done on the training data. They want to set up a consistent evaluation protocol that they can replicate on their side.

Stop reason: End of sequence token encountered
Tokens: 1623 input + 51 generated = 1674 out of 4096
4 seconds

Generate →

Stop reason: End of sequence token encountered

Tokens: 57 input + 54 generated = 111 out of 4096 | Seed: 111

3.37 seconds

Customer Usage Examples

Customers will pay tier charges for access to features and then pay as you go based on consumption

***Tier charge only enables access to the tier features and isn't applied against the consumption based charges

Sample Essential Tier Spend: \$7.50

Tier Charge: \$0 / month

+

.ai: \$7.50 / month

- Who-Marketing agency
- Use Case- Class 3 OS model in watsonx.ai for content generation (inferencing)
- Scope-supporting 30 clients, creating 30,000 words per client per month, with some uplift for prompting
- Inference: 1.5M tokens / month = \$7.5/ month
- Tuning: N/A
- Hosting: N/A
- ML Functionality= \$0

Sample Standard Tier Spend: \$3084.2

Tier Charge: \$1.5K / month

+

.ai: ~\$2,871.2 / month

- Who-Legal client
- Use Case- using a Class 3 OS model for a summarization (inferencing)
- Scope-5,000 documents per month, 3,000 words per document; hosting at 50% utilization.
- Inference: 22M tokens/month = \$110 / month
- Tuning: \$24 x 6.5 hrs / month = \$156 / month
- Hosting: \$3.5 x 732 hrs / month = \$2,562 / month
- ML Functionality: 30 CUH / month = \$43.2

watsonx.data

watsonx Pricing (for 7/7 GA)

watsonx.data (SaaS)

Consumption-based charges, driven off 3 Price metrics*

- Quantity of nodes deployed
- Duration of nodes deployed
- Support services deployed (e.g., Hive, Ranger)

*watsonx.data only available for Standard and Premium Tier Customers

Pricing

	Price (USD) / Hour
Coordinator node per hour	\$2.80
Cache node cost per hour	\$2.80
Computer node cost per hour	\$6.50
Supporting services per hour	\$3.00

watsonx.data (On-Prem)

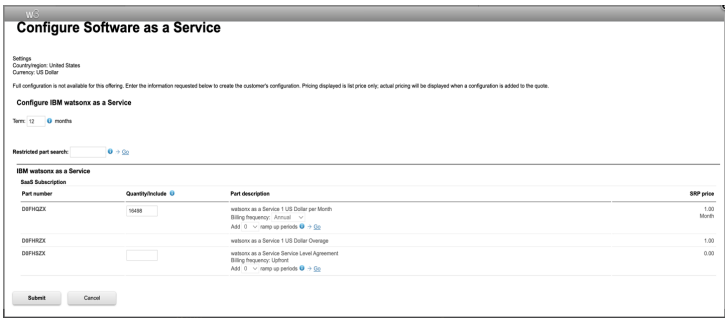
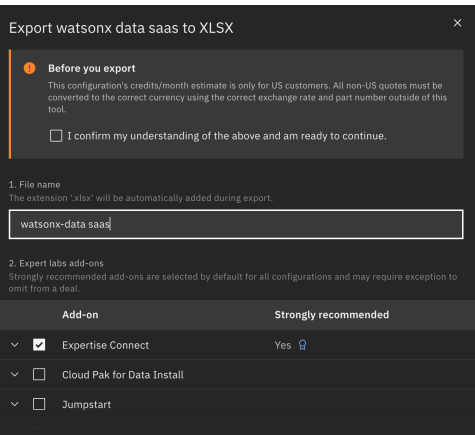
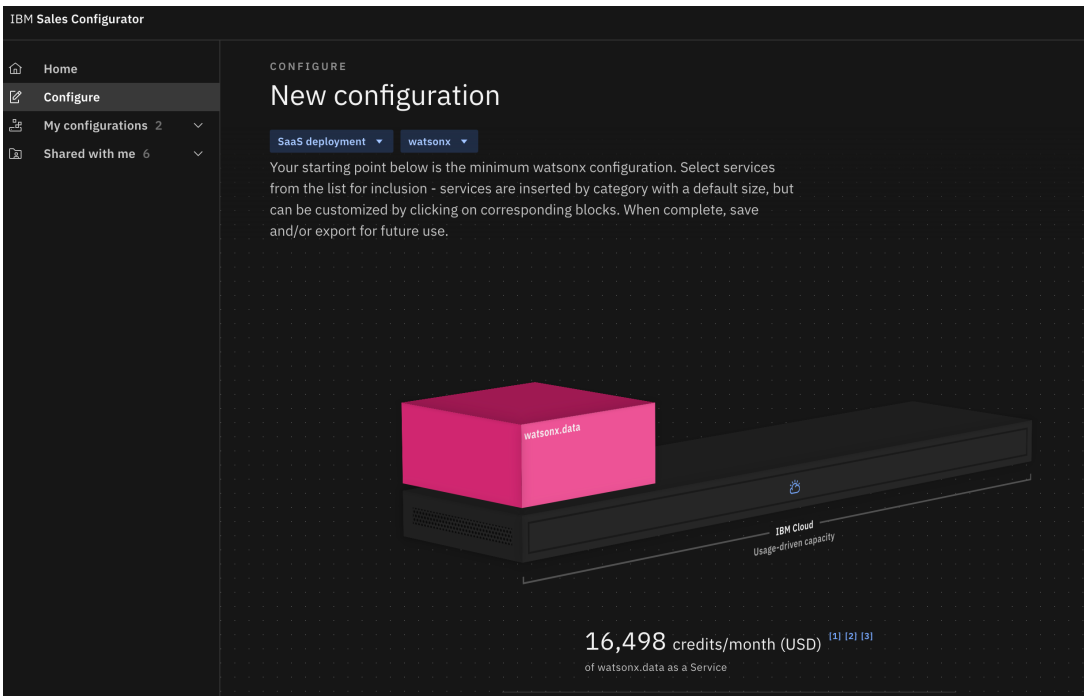
- Price Metric: VPCs (16 VPCs/node) @ \$7500/VPC
- Minimum of 10 nodes (160 VPCs) recommended
- Production versions available in Perpetual, Subscription and Monthly Licenses. Options for Non-production, reserved, and various support tiers also available.

Pricing

License Type	Price (USD) / VPC
Perpetual	\$7,500
Subscription	\$250 / month
Monthly	\$312.50 /month

On-prem Parts are listed [here](#)

Leverage the [sales configurator](#) for sizing- SW and SaaS



Note: Recommend engaging Expert Labs Services and Client Engineering to accelerate productive use

watsonx.data SaaS Pricing – What you need to know

Consumption-based charges, driven off 3 Price metrics*

- Quantity of nodes deployed
- Duration of nodes deployed
- Support services deployed (e.g., Hive, Ranger)

*watsonx.data only available for Standard and Premium Tier Customers

Consumption measured in Resource Units (RU). 1 RU = \$1 USD

Pricing	RU / Hour
Cache node cost per hour	\$2.80
Computer node cost per hour	\$6.50
Supporting services per hour	\$3.00

Key Takeaways

- Usage can occur on IBM Cloud or AWS infrastructure
- Pricing depends on type of nodes, number of nodes used, and usage of nodes
 - AWS infrastructure provides two node types (Cache and Compute Optimized), IBM Cloud infrastructure has one node type (Cached Optimized). Prices are the same for AWS and IBM Cloud
 - Number of nodes relates to T-shirt sizes.
 - Usage relates to uptime of services.
- Pricing is measured in Resource Units. 1 RU = \$1 USD

Estimated T-Shirt Sizes for IBM Cloud

Sizing	Nodes	Total vCPU	Total RAM (GiB)	Resource Unit per Month*
Starter	2 nodes	32 vCPU	256 GB	•Using service 100% of the time: 6278 •Using service 70% of the time: 5052 •Using service 35% of the time: 3621
Small	4 nodes	64 vCPU	512 GB	•Using service 100% of the time: 10366 •Using service 70% of the time: 7913 •Using service 35% of the time: 5052
Medium	7 nodes	112 vCPU	896 GB	•Using service 100% of the time: 16498 •Using service 70% of the time: 12205 •Using service 35% of the time: 7198
Large	12 nodes	192 vCPU	1536 GB	•Using service 100% of the time: 28762 •Using service 70% of the time: 20790 •Using service 35% of the time: 11490

*Resource Unit Values in the table can be converted to US Dollar value at 1:1 Ratio

Standard (on-Prem) T-shirt Sizing

T-shirt size for a single cluster with performance characteristic: If customer wants multiple identical environments for different departments or Dev+Prod, multiple across
This assumes a worst-case scenario - 100% of your data is in active memory. **Note:** These sizes below are each for 1 cluster

	Description	License (VPC) for 1 cluster of this profile	Base Entitlement (UI,HMS, Etc)	Total Entitlement	Total HW requirements
Small	1 Coordinator node, 3 worker nodes Each node assumes 16vCPU, 128Gb of memory, and 2 x 1900GB NVMe, up to 10 GbE. Example of estimated workload supported by the configuration: <ul style="list-style-type: none">1 to 5 queries processing 100GB → 400 GB expanded/uncompressed active data in memory or up to 50 queries processing up to 8GB each100-200GB Active data compressed on disk required to be processed.	64	12	76	76 vCPU Total Memory:608Gb
Medium	1 Coordinator node, 9 worker Each node assumes 16vCPU, 128Gb of memory, and 2 x 1900GB NVMe, up to 10 GbE. Example of estimated workload supported by the configuration: <ul style="list-style-type: none">1 to 5 queries processing 250GB → 1125 GB expanded/uncompressed active data in memory or up to 50 queries processing up to 22GB each250-500GB Active data compressed on disk required to be processed.	160	24	184	184vCPU Memory: 1472Gb
Large	1 Coordinator node 19 worker Each node assumes 16vCPU, 128Gb of memory, and 2 x 1900GB NVMe, up to 10 GbE. Example of estimated workload supported by the configuration: <ul style="list-style-type: none">1 to 5 queries processing 500GB → 2250 GB expanded/uncompressed active data in memory or up to 50 queries processing up to 45GB each500-1000GB Active data compressed on disk required to be processed.	320	48	368	368vCPU Memory: 2944Gb
X-Large	70 Nodes to configure as desired (i.e 10 clusters of 1 head node and 6 worker nodes or 1 large cluster with 70 nodes). Each node assumes 16vCPU, 128Gb of memory, and 2 x 1900GB NVMe, up to 10 GbE. Example of estimated workload supported by the configuration: <ul style="list-style-type: none">1 to 5 queries processing 1750GB → 7875 GB expanded/uncompressed active data in memory or up to 175 queries processing up to 45GB each1750-3500GB Active data compressed on disk required to be processed.	1120	48	1168	1168vCPU Memory: 9344Gb
XX-Large	200 Nodes to configure as desired (i.e 10 clusters of 1 head node and 19 worker nodes or 1 large cluster with 200 nodes). Each node assumes 16vCPU, 128Gb of memory, and 2 x 1900GB NVMe, up to 10 GbE. Example of estimated workload supported by the configuration: <ul style="list-style-type: none">1 to 5 queries processing 5000GB → 22,500 GB expanded/uncompressed active data in memory or up to 500 queries processing up to 45GB each5000-10,000GB Active data compressed on disk required to be processed.	3200	48	3248	3248vCPU Memory: 25600Gb

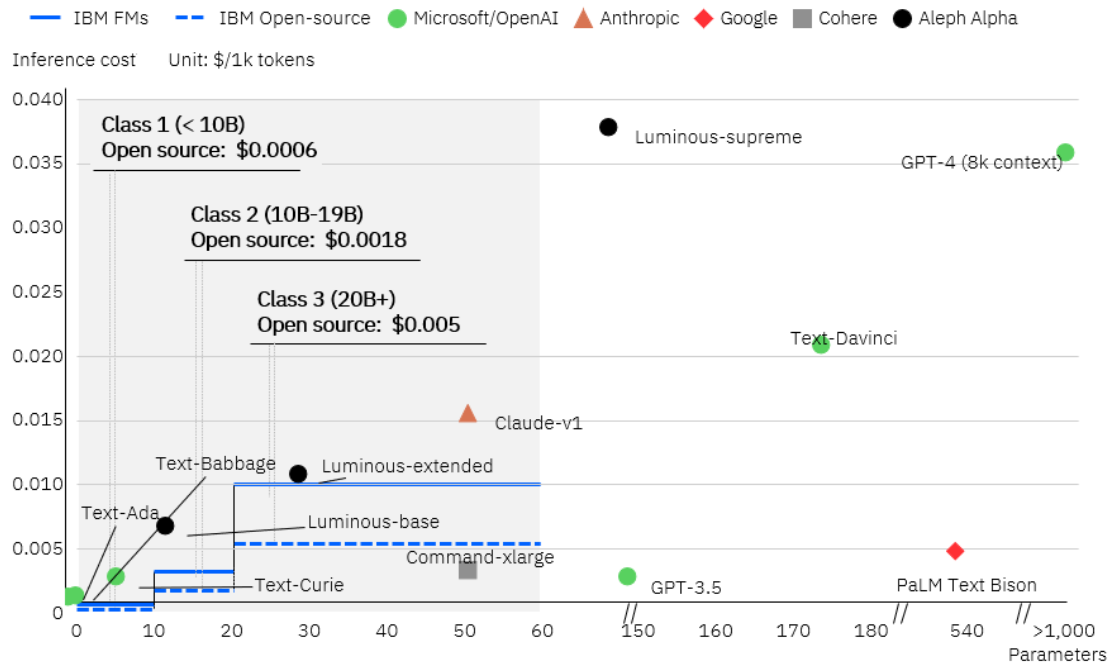
Get started with watsonx

	watsonx.data	watsonx.ai	watsonx.governance
Trials	Available – <u>get started today via IBM Cloud or AWS</u>	Available – <u>get started today!</u> If you need additional tokens, please see guidance on account upgrades	WIP – Get ready for GA later this year!
IBM Tech Zone	Available – <u>get started today!</u>	WIP – More news coming!	WIP – Get ready for GA later this year!
Other	<u>Request a POC</u>	Tech Preview available for select clients – <u>work with Client Engineering to get started</u>	WIP – Get ready for GA later this year!

Competitive

watsonx.ai pricing competitive with market

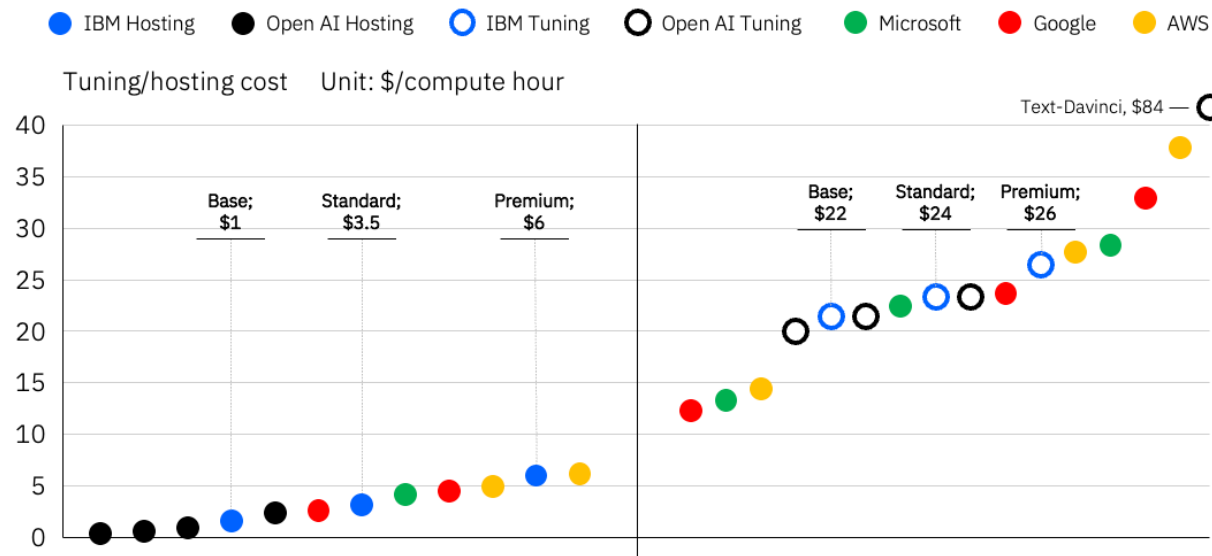
Inference Pricing



- IBM models expected to command a higher premium (up to 2x) to market due to 'responsible' differentiators
- CSP competitors typically offer a further discount for existing CSP customers
- Domain-specific models such as Bloomberg's and Morgan Stanley's are expected to enter and provide more cost-efficient offerings for their domains
- Competitors are changing pricing on a near-weekly basis, updates will be required

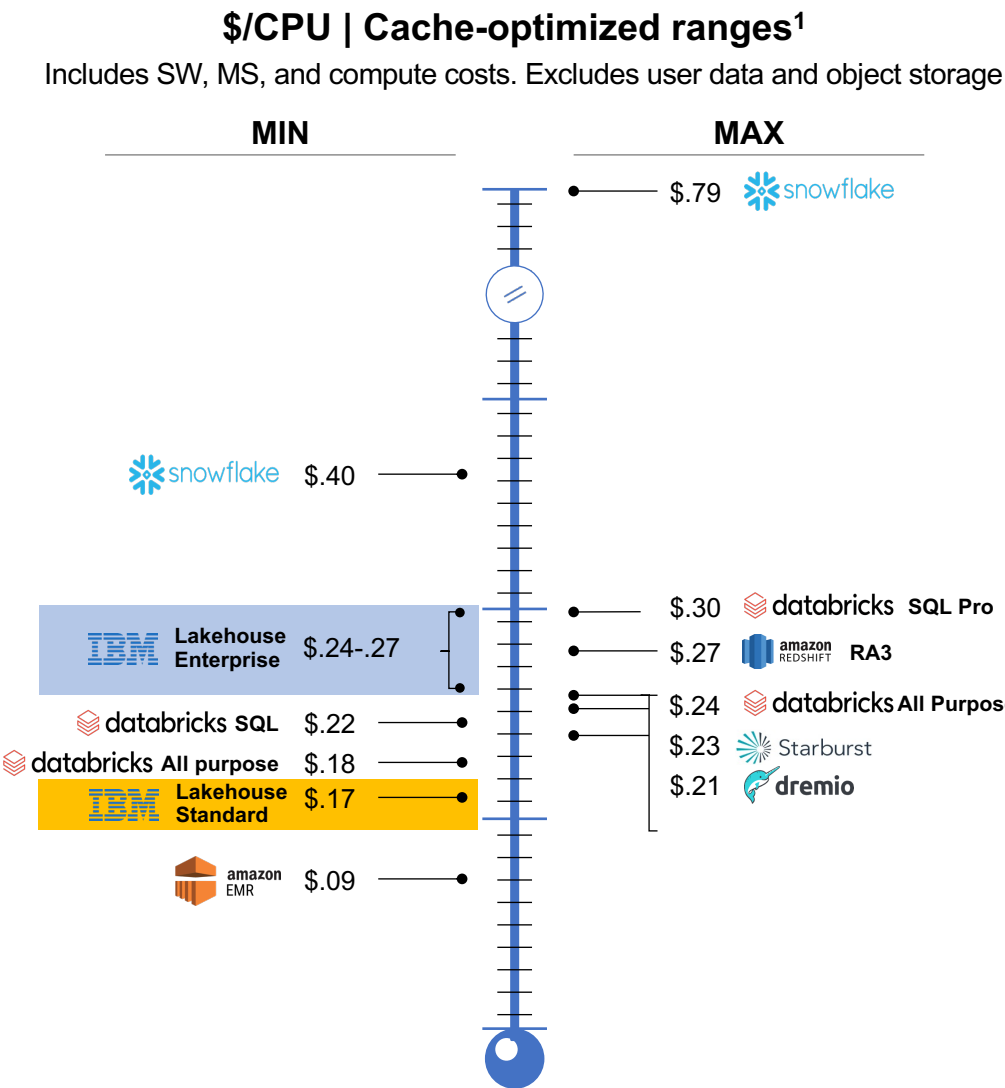
Hosting Pricing

Tuning Pricing



- Microsoft, Google, and AWS let customers select their instance; IBM will designate specific instances depending on model size
- Startup competitors (e.g., Lambda Labs) are emerging that provide significantly lower prices through data centers purpose-built for GenAI
- Leading startup competitors tend to charge an additional increased inference price when their models have been fine-tuned
- Cohere is the only top competitor offering on-prem hosting for its foundation models, and their pricing is based on a monthly fee

.data SaaS pricing is strategically set to maintain competitiveness while preserving margin



Min and max prices represent pricing for each competitors lowest (e.g., Standard) and highest (e.g., Enterprise) paid tier

IBM Lakehouse is **cheaper than Snowflake and Databricks**

- IBM will be over 50% cheaper than **Snowflake's lowest tier** product offering
- IBM is ~5% cheaper than Databrick's lowest tier of All Purpose Compute
- IBM is ~20% cheaper than Databrick's lowest tier of SQL

We expect Snowflake and Databricks to be relatively inflexible due to:

- **A low % open-source code-base vs. 90% open source for IBM**
- **High SG&A costs** from hyper-growth phase

IBM can potentially increase margins or be able to respond to competitive actions by reducing costs by an additional 12-20% if:

- Lakehouse on AWS can **run without IBM Cloud** layered in
- **ROSA/HyperShift is relaxed as a mandated platform on AWS**

1. Min rates represent a company's base non-free offering, while max rates represent a company's top tier | 2. IBM, Dremio, and Starburst have no available tiering information, therefore only one rate is shown per node profile | 3. Databricks SQL rates are shown for SQL Compute and SQL Pro Compute respectively. Only All-purpose has differentiated price levels by tiering