AI for Data Scientist

1. **Vertex AI:**

unified machine learning platform will help you build, deploy and scale more effective AI models.

Benefits:

- Accelerate ML with a unified data and AI platform and tooling for pre-trained and custom models
- Build generative AI apps quickly and responsibly with Model Garden and Generative AI Studio
- Implement MLOps practices to efficiently scale, manage, monitor, and govern your ML workloads
- Reduce training time and costs with optimized infrastructure.

Key Features:

- Choose a model that fits your needs (using Model Garden and Generative AI studio)
- End-to-end MLOps (train, test, monitor, deploy, and govern ML models at scale)
- Low-code and no-code tooling
- Open and flexible AI infrastructure

All features:

- **Vertex AI Model Garden:** A single place to search, discover, and interact with a wide variety of foundation models from Google and Google partners, available on Vertex AI
- **Vertex AI Generative AI Studio:** A managed environment in Vertex AI that makes it easy to interact with, tune, and deploy foundation models to production
- **AutoML:** Easily develop high-quality custom machine learning models without writing training routines. Powered by Google's state-of-the-art transfer learning and hyperparameter search technology.

- **Deep Learning VM Images:** Instantiate a VM image containing the most popular AI frameworks on a Compute Engine instance without worrying about software compatibility.
- **Vertex AI Workbench:** Vertex AI Workbench is the single environment for data scientists to complete all of their ML work, from experimentation, to deployment, to managing and monitoring models. It is a Jupyter-based fully managed, scalable, enterprise-ready compute infrastructure with security controls and user management capabilities.
- **Vertex AI Matching Engine:** Massively scalable, low latency, and cost-efficient vector similarity matching service.
- **Vertex AI Data Labeling:** Get highly accurate labels from human labelers for better machine learning models.
- **Vertex AI Deep Learning Containers:** Quickly build and deploy models in a portable and consistent environment for all your AI applications.
- **Vertex AI Feature Store:** A fully managed rich feature repository for serving, sharing, and reusing ML features.
- **Vertex ML Metadata:** Artifact, lineage, and execution tracking for ML workflows, with an easy-to-use Python SDK.
- **Vertex AI Model Monitoring:** Automated alerts for data drift, concept drift, or other model performance incidents which may require supervision.
- **Vertex AI Neural Architecture Search:** Build new model architectures targeting application-specific needs and optimize your existing model architectures for latency, memory, and power with this automated service powered by Google's leading AI research.
- **Vertex AI Pipelines:** Build pipelines using TensorFlow Extended and Kubeflow Pipelines, and leverage Google Cloud's managed services to execute scalably and pay per use. Streamline your MLOps with detailed metadata tracking, continuous modeling, and triggered model retraining.
- **Vertex AI Prediction:** Deploy models into production more easily with online serving via HTTP or batch prediction for bulk scoring. Vertex AI

Prediction offers a unified framework to deploy custom models trained in TensorFlow, scikit or XGB, as well as BigQuery ML and AutoML models, and on a broad range of machine types and GPUs.

- **Vertex AI Tensorboard:** This visualization and tracking tool for ML experimentation includes model graphs which display images, text, and audio data.
- **Vertex AI Training:** Vertex AI Training provides a set of pre-built algorithms and allows users to bring their custom code to train models. A fully managed training service for users needing greater flexibility and customization or for users running training on-premises or another cloud environment.
- **Vertex AI Vizier:** Optimized hyperparameters for maximum predictive accuracy.

2. **Vertex AI Workbench:**

The single development environment for the entire data science workflow(data analytics and machine learning workflows).

AI for Developer

1. **AutoML:**

AutoML enables developers with limited machine learning expertise to train high-quality models specific to their business needs. Build your own custom machine learning model in minutes.

- **Types of models you can build using AutoML:**

Data type	Supported objectives
Image	Classification, Object Detection
Tabular	Classification(Binary, multi-class), Regression, forecasting
Text	Classification, entity extraction, sentiment analysis
Video	Action recognition, Classification, Object tracking

2. **Natural Language AI:**

Derive insights from unstructured text using Google machine learning.

Key features (Three natural language solutions that work with your text):

2.1AutoML
2.2Natural Language API:

The powerful pre-trained models of the Natural Language API empowers developers to easily apply natural language understanding (NLU) to their applications with features including sentiment analysis, entity analysis, entity sentiment analysis, content classification, and syntax analysis.

2.3Healthcare Natural Language API

The Healthcare Natural Language API parses unstructured medical text such as medical records or insurance claims. It then generates a structured data representation of the medical knowledge entities stored in these data sources for downstream analysis and automation. For example, you can Extract information about medical concepts like diseases, medications, medical devices, procedures, and their clinically relevant attributes

	AutoML	NL API
accessible via REST API	✅	✅
Syntax analysis		✅
Multi-language	✅	✅
Custom model train	✅	
Large dataset support	✅	

3. Speech-to-Text:

Accurately convert speech into text using an API powered by Google's AI technologies.

- Speech adaptation: Provide hints to boost the transcription accuracy of rare and domain-specific words or phrases. Use classes to automatically convert spoken numbers into addresses, years, currencies, and more.
- Domain-specific models
- Easily compare quality
- Speech On-Device
- Foundation model for Speech-to-Text (powered by Chirp)

4. Text-to-Speech:

Convert text into natural-sounding speech using an API powered by Google's AI technologies.

- Neural2 voices
- Studio voices (Preview)
- Custom Voice
- Voice tuning
- Text and SSML support

5. **Timeseries Insights API**

   Large-scale time series forecasting and anomaly detection in real time.

6. **Translation AI**

- Translation Hub (135 languages)
- AutoML Translation
- Translation API
- Media Translation API

7. **Video AI**

- AutoML Video Intelligence
- Video Intelligence API

8. **Vision AI**

- AutoML Vision
- Vision API

AI infrastructure

1. **Deep Learning Containers**

   Preconfigured and optimized containers for deep learning environments. These Docker images use popular frameworks and are performance optimized, compatibility tested, and ready to deploy

   - Consistent environment
   - Fast prototyping
   - Performance optimized: Accelerate your model training and deployment with the latest framework versions and NVIDIA® CUDA-X AI libraries.

2. **Deep Learning VM Image**

   Preconfigured VMs for deep learning applications.

   - Broad support
   - Optimized for performance: optimized with the latest NVIDIA® CUDA-X AI libraries and drivers and the Intel® Math Kernel Library.
   - Fast prototyping
   - Integrated notebook experience

NVIDIA® CUDA-X is a suite of software development tools and libraries provided by NVIDIA, designed to enable developers to harness the power of NVIDIA GPUs (Graphics Processing Units) for accelerated computing. It includes:

- CUDA Toolkit for GPU programming
- cuDNN for deep neural networks
- TensorRT for optimized model deployment
- NCCL for multi-GPU communication
- Nsight for debugging and profiling
- CUDA-X AI for accelerating AI workloads: it includes libraries like cuML for machine learning algorithms, cuGraph for graph analytics, and cuDNN for deep learning.

----

3. GPUs

High-performance GPUs on Google Cloud for machine learning, scientific computing, and 3D visualization

- GPU types: NVIDIA L4, P100, P4, T4, V100, and A100
- Flexible performance: Optimally balance the processor, memory, high performance disk, and up to 8 GPUs per instance for your individual workload. All with the per-second billing, so you only pay only for what you need while you are using it.

4. TensorFlow Enterprise

Reliability and performance for AI applications with enterprise-grade support and managed services.

5. TPUs

Cloud TPUs optimize performance and cost for all AI workloads, from training to inference