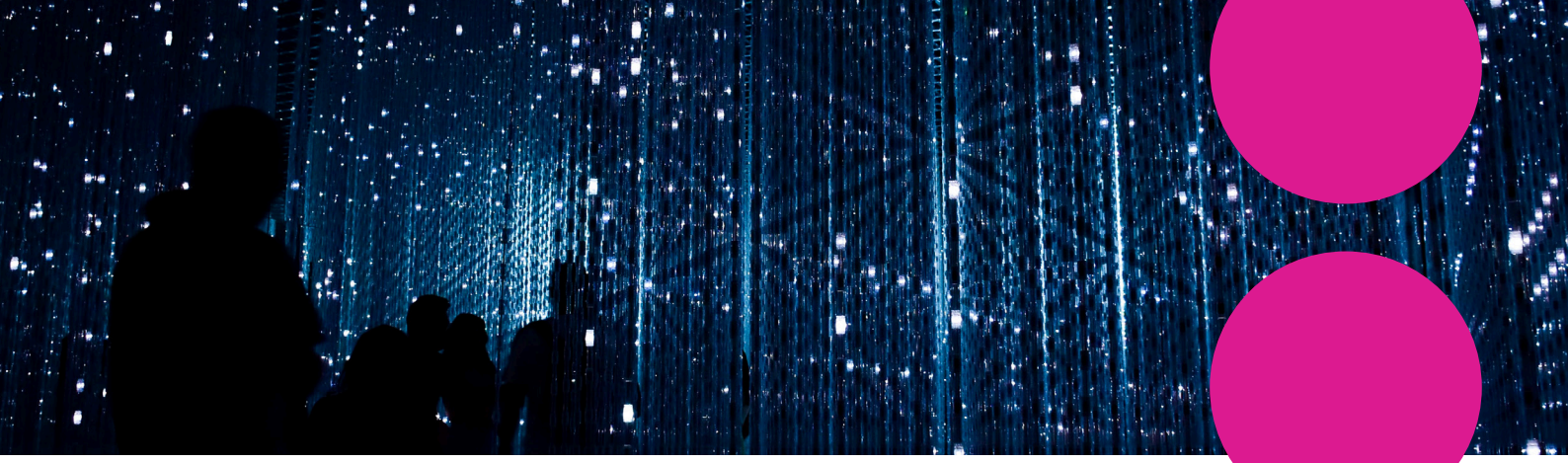


AI の計算性能を向上させ、データサイエンスのイノベーションを加速

Run:ai アトラス　＃ 1 AI 計算プラットフォーム

課題：AI は実行に問題あり

AI の研究イニシアチブのほとんどは、生産に至ることはありません。なぜでしょうか。その原因は、研究者が深層学習（DL）アルゴリズムの構築と学習、そして AI イニシアチブを生産に持ち込むために使用するグラフィクスプロセシングユニット（GPU）の静的な割り当てにあります。これらの GPU リソースは、研究者に静的に割り当てられています。すなわち、ある研究者に割り当てられた高価な計算リソースは、別の研究者が GPU を待っている間でも、アイドル状態になることが多いのです。実際、企業内のほとんどの AI チームは、平均して GPU インフラの 25% しか使用していません。静的な計算割り当てがプロセスを制限してしまうため、モデルを生産に移すのに時間がかかるのです。リソースを効率的に利用できないことが実験を遅くしてしまい、それがほとんどの企業が AI から ROI を得られない主な理由の 1 つとなっています。

ソリューション：AI 基盤の構築

Run:ai をご利用のお客様は、Run:ai を導入していないお客様よりも 10 倍早くモデルを生産に移しています。Run:ai アトラスのソフトウェアプラットフォームを使用することで、企業はあらゆるインフラストラクチャ（オンプレミス、エッジ、クラウド）において、AI アプリケーションの開発、管理、スケーリングの効率を上げられます。研究者は、あらゆる AI ワークロードに対して用意された GPU と CPU リソースに、オンデマンドでアクセスすることができます。革新的でクラウドネイティブなオペレーティングシステムにより、一部の GPU から大規模な分散トレーニングまで、あらゆるものを IT で管理できます。効率が大幅に上がることで、モデル化も早まります。たとえば Run:ai を導入したあるお客様は、最近、6,700 個のハイパーパラメータチューニングジョブを並行して実行し、記録的な速さでモデル化を完了させました。

このソリューションの利点：

より早くイノベーションへ

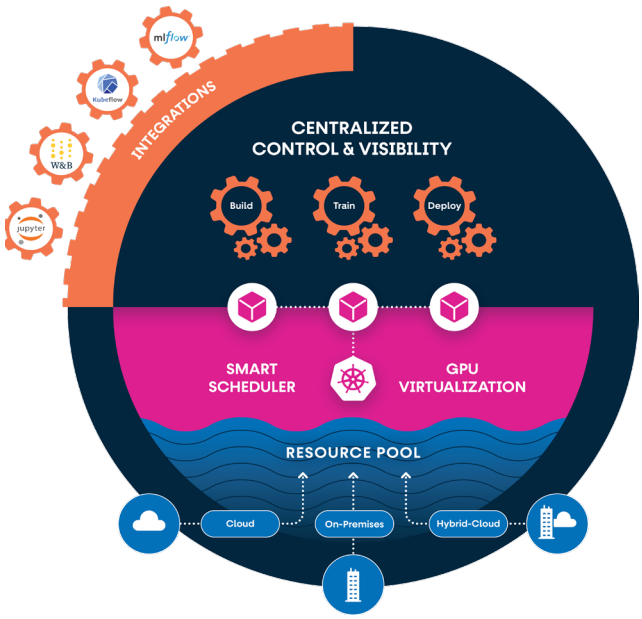
Run:ai のリソースプール、キューイング、優先順位付けのメカニズムを利用することで、研究者はインフラ管理の煩わしさから解放され、計算上のボトルネックなしに必要なだけのワークロードを実行し、データサイエンスのみに集中することができます。

少ないリソースでより多くのことを実行

Run:ai の公平性アルゴリズムにより、すべてのユーザーとチームがリソースの公平な分配を受けられます。優先的なプロジェクトに関するポリシーは事前に設定できます。このプラットフォームでは、ユーザーやチームから別のユーザーやチームへのリソースの動的な割り当てが可能なので、すべてのユーザーが、必要な GPU リソースにタイムリーにアクセスできます。

AI のより速いデプロイへ

Run:ai アトラスでは、Kubernetes 上での分散学習に、GPU の一部、複数個の GPU、GPU の複数ノードを簡単に利用することができます。このように、AI ワークロードは容量ではなくニーズに基づいて実行されます。
データサイエンスチームは、同じインフラ上でより多くの AI の実験を行えるようになります。



Run:ai アトラス

- **リソースプール：**オンプレミスやクラウドなどの場所を問わずに異種ハードウェアをプールし、最適なパフォーマンスと効率性を実現します。

- **オペレーティングシステム：**OS は、GPU 抽象化レイヤーと Kubernetes ベースの AI ワークロードスケジューラーで構成されています。クラウドネイティブなオペレーティングシステムにより、構築、学習、推論など、あらゆる AI ワークロードをスケジュールし管理します。動的なスケジューリング機能により、マルチノード分散学習に GPU の一部を自動で使用することができます。

- **コントロールプレーン：**複数のクラスタがどこにあっても、それらを一元的に可視化、制御できます。ユーザー、ジョブクラスター、プロジェクトを示すリアルタイムおよび過去の分析結果を見ることができます。管理者は、インフラ全体のユーザーやチームに対して、優先順位やポリシーを設定することができます。SSO や高度なエンタープライズ機能を搭載しています。

- **アプリケーションとインテグレーション：**Run:ai の統合ワークフローやお好みの AI/ML ツールを使用して、加速したインフラ上で AI アプリケーションを開発、実行します。
  Run:ai は、Pytorch、Tensorflow、および多くの DS ツール、さらに Kubeflow、MLflow、Seldon、Weights & Biases、その他の MLOps ツールのホストと統合されています。

Kubernetes ベースのアーキテクチャで AI ワークロードのスケジューリングを自動化

今日の一般的な慣習では、コンテナや Kubernetes を中心に深層学習インフラを構築しています。Run:ai は、IT やデータサイエンスのチームの学習曲線を簡素化してインフラの効率を向上させるために、Kubernetes に AI ワークロードのための高性能なスケジューラを直接構築しました。

ビジネス目標に沿った AI インフラの管理

Run:ai アトラスは、Kubernetes の上で複数のキューを使って、タスクをバッチプロセスとして管理します。コントロールプレーンでは、システム管理者がビジネスの優先度に基づき、それぞれのキューに対して異なるルール、ポリシー、要件を定義することができます。クォータベースのシステムおよび設定可能な公平性ポリシーと組み合わせることで、管理者がリソースの割り当てを自動化し、クラスタリソースを最大限に活用できるように最適化することができます。

46 日かかっていた実験が Run:ai でわずか 2 日に短縮

Run:ai アトラスを利用している企業は、より優れたオーケストレーションと計算リソース管理により、AI モデルを生産に移行させるのにかかる時間を大幅に短縮しています。あるお客様は、実験にかかる時間を従来の平均 46 日から現在の平均 1 日半に短縮し、3000%の改善を達成しました。