AI でビジネス課題を解決 -AWS におけるインテルのパワーで実現

アカンクシャ・バラニ インテルグローバル AI GTM リード AWS APJ Alliance ヘッド



アジェンダレイアウト

- AI の現在
- インテル & AWS が選ばれる理由
- AWS でのインテル AI
- Habana Gaudi インスタンスのご紹介
- ・ 学習 | エンゲージメント | インテルでの AI のイノベーション



AIによる変革

















産業機器



民生機器

スマート・アシスタント チャットボット パーソナライズを検索 拡張現実 ロボッツ

ヘルス

診断の強化 創薬 患者ケアの研究 センサー 支援

財務

アルゴリズム取引

不正検出研究 体駅 パーソナル・ ファイナンス マーチ リスクの軽減 ダイジ ロイヤル サプライヲ

小売

サポート

体験 マーケティング マーチャン ダイジング ロイヤルティー サプライチェーン セキュリティー

行政機関

データインサイト 安全性 & セキュリティー 居住者の エンゲージメント よりスマートな 都市

防御

エネルギー

石油 & ガスの採掘 スマート グリッド 運用の改善 保全

輸送

自動運転車

自動追跡

航空宇宙産業

輸送

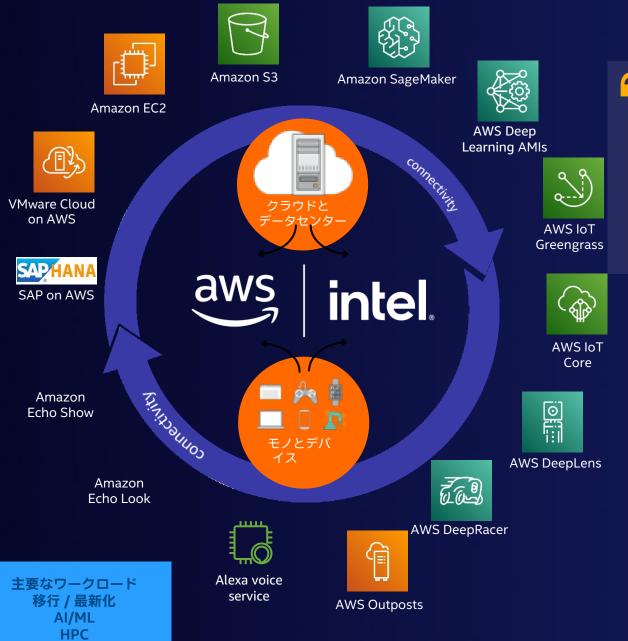
捜索 & 救出

ファクトリー・ オートメーション 予測メンテナンス 精密農業 フィールド・オー トメーション

その他

広告 教育 ゲーム プロフェッショナル & IT サービス 電気通信 / メディア スポーツ

インテルが Amazon Web Services と協働で行っている内容



インテルは、AWS と深い絆でつながっているパートナーであり、今後もずっとそうあり続けるでしょう。 このパートナーシップが変わることはありません。

元AWS CEO **Andy Jassy氏** 共通の歴史と価値

14 年以上にわたるエンジニアリングの パートナーシップ

デジタル・トランスフォーメーション

顧客のロイヤルティーを共有 ハイパフォーマンス + ローコスト

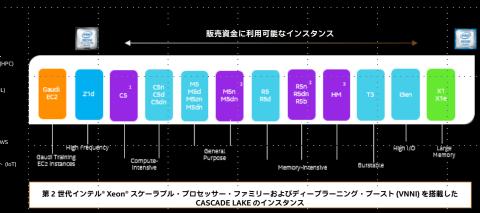
世界屈指のサプライチェーン

極めて優れた多様さと可用性で、 お客様のグローバルなワークロードのニーズに対応

aws intel

250 以上のインテル®のインスタンス

最**適化されたストレージ。** |3||3en||D3/D3en



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ハイブリッド / エッジ

インテル® アーキテクチャーのインスタンス・タイプ 220 以上のインテル® インスタンス

汎用

汎用インスタンスにより、 コンピューティング、メモ リー、ネットワーキング・ リソースのバランスを実現し、 さまざまなワークロードに 使用可能です。

M6i

HM5dn F

T2 **E**

M5nE

Gaudi

Habana

Ice Lake

Cascade Lake

Skylake および Cascade Lake

Skylake

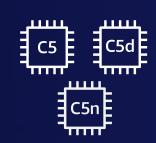
Broadwell

Haswell

コンピューティング最適化

コンピューティング最適化の インスタンスは、ハイパ フォーマンス・プロセッサー のメリットを活かすコン ピューティング・バウンド・ アプリケーションに最適です。







メモリー最適化

メモリー最適化のインスタン スは、メモリーで大規模な データセットを処理する ワークロードに対して高速パ フォーマンスを提供するよう に設計されています。

高速化コンピューティング

高速化コンピューティングの インスタンスは、ハード ウェア・アクセラレーター またはコプロセッサーを 使用して、機能をより効率的 に実行します。



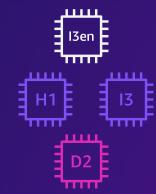


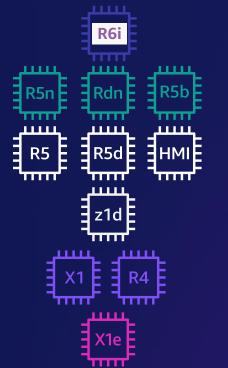


ストレージを最適化

ストレージ最適化のインスタンスは、ローカルストレージ上の非常に大規模なデータセットへの高速な連続読み出し/書き込みアクセスを必要とするワークロード向けに設計されています。









過去1年間のハイライト

パンデミックの間、Career Launcher の ラーニングポータルが迅速に拡大 インテルと AWS のコラボレーションにより、 AWS で 2 カ月以内にインドの 16 万人以上 の学生にサービスを提供します。¹

山火事の並列シミュレーションが 65 倍増加

インテルと AWS は、オーストラリアで の消火活動の効率化支援を拡大するため、 RONIN との連携を行っています。⁴





AWS ParallelCluster

AWS は、検証済みのインテル® Select ソリューション初の CSP ²

Amazon EC2 M5zn インスタンス – クラウドで最速のインテル® Xeon® スケーラブル・プロセッサー ⁵ 最大 4.50GHz の動作周波数で、最高のオールコアターボ CPU のパフォーマンス。

AWS は DL1、M6i、C6i、DL1 を発表 Habana Gaudi 搭載 40% 優れた価格 / パフォーマンスの AI インスタンス ³ 最大 8 個の Gaudi アクセラレーターを 使用した、インテルの Habana & AWS 共同開発ソリューション



- [1] https://www.intel.co.jp/content/www/jp/ja/customer-spotlight/stories/career-launcher-customer-story.html
- [2] https://docs.aws.amazon.com/parallelcluster/latest/ug/intel-select-solutions.html
- [3] https://aws.amazon.com/ec2/instance-types/habana-gaudi/
- [4] https://dpgresources.intel.com/asset-library/intel-aws-the-csiro-spark-intel-poc-summary/
- [5] https://aws.amazon.com/blogs/aws/new-ec2-m5zn-instances-fastest-intel-xeon-scalable-cpu-in-the-cloud/





新しい第3世代インテル®Xeon®スケーラブル・プロセッサー

- 高性能なワークロード・パフォーマンス
- 大規模展開で信頼性を実現する設計
- 新しい暗号化の高速化
- 高度なセキュリティー機能
- トータル・メモリー・エンクリプション (TME)

1.58x ウェブ・マイクロサービスの **1.58x** パフォーマンスが向上

最大 パフォーマンスの向上
40% (SPECrate2017_int_base)
新しい Ice Lake* SKU 製品と Cascade Lake** の比較

最大 プロセッサー当たりのコア数が多い
1 42 x 40 コア Ice Lake* と 28 コア Cascade Lake** の比較

*lce Lake: 第 3 世代インテル® Xeon® スケーラブル・プロセッサー **Cascade Lake: 第 2 世代インテル® Xeon® スケーラブル・プロセッサー





Habana Gaudi 搭載インスタンスのご紹介

インテルの新しい HABANA GAUDI プロセッサーを活用した ML トレーニング



MLトレーニング専用の新しい Amazon EC2 インスタンス。 インテルの新しい Habana Gaudi プロセッサーを最大 8 個搭載

GPU 搭載インスタンスと比較して、 ディープラーニング・モデルのト レーニング・コストを最大 40% 削減 お客様がモデルのイタレーション とトレーニングをより頻繁に行う ことができる

フルスタックの Amazon EC2 サービス - AWS ディープラーニング AMI、コンテナ化されたアプリケーション用のディープラーニング・コンテナー、最終的には Amazon SageMaker を利用可能

デベロッパーは、コンテナ化されたアプリケーション用の Amazon ECS および EKS を介して Gaudi 搭載インスタンスを実装できる

TensorFlow、PyTorch などの一般的なフレームワークをサポート

NLP、画像分類、オブジェクト検出、 レコメンデーション・システムなど、 幅広い ML ワークロードに対応

複数の Gaudi を搭載した Amazon EC2 インスタンスを効率的にスケー リングするために、AWS Elastic Fabric Adapter をサポート



データ分析のポートフォリオ



ソリューション ソリューション・アーキテクト



フォーム



















OpenVINO™ ツールキット

TF、Caffe*、MXNet* を使用してCPU、プロセッサー・グラフィック ス、FPGA & VPU 上の推論環境向けの OpenVINO™ ツールキット

ディープラーニング・デベロッパー・ツールキット

TensorFlow* と Caffe* を使用したすべての インテル® Movidius™VPU 向けに最適化された推論環境



ライブラリー

マシンラーニング・ライブラリー

Python

NumPv

Scikit-learn

・カート Pandas RandomForest

分散型 • MlLib (on Spark)

Mahout

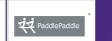
mxnet





ディープラーニング・フレームワーク





TensorFlow* MXNet* Caffe* BigDL / Spark*

Caffe2* PvTorch* PaddlePaddle³



ライブラリー・デベロッパー

分析、マシン & ディープラーニング・プリミティブ

マシンラーニング向けに最適化された インテル® ディストリビューション

Python

•E1071

インテル® DAAL

インテル® データ・アナリティクス・ アクセラレーション・ライブラリー (マシンラーニング向け)

MKL-DNN

clDNN

CPU、プロセッサー・グラフィックス向けのオープンソース・ ディープ・ニューラル・ネットワーク機能



ハードウェア IT システム・アーキテクト

基盤











アクセレレーター







*その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。 すべての製品、コンピューター・システム、日付、および数値は、現在の予想に基づくものであり、予告なく変更されることがあります。



インテル製品搭載 ML 向け Amazon EC2 インスタンス

M5 M5n M5d M5dn 第2世代インテル®Xeon®スケーラブル・プロセッサー

推論向けインテル® DL ブースト、シングル・ノード・トレーニング





M5、C5 インスタンスは、すべての コンピューター・ビジョン、ML および DL 推論のワークロードに適している R5 インスタンスは、メモリー要件が 192GB を 超える 3D-CNN/BERT-large/T5 トポロジーを使 用するメモリー集約型ワークロードに適している

T3 インスタンスは、ML アプリケーション および大量のコンピューティングを必要とし ない DL 推論アプリケーションに適している

C5n インスタンスは、ノード間通信に必要な NW パフォーマンスが高いため、分散型ディー プラーニング・トレーニングに適している

HVM ベースのインスタンスはパフォーマンス上の オーバーヘッドが最大 10% 増加するため、ベアメタル・ インスタンスは大規模なトポロジーに適している

*lce Lake: 第 3 世代インテル® Xeon® スケーラブル・プロセッサー **Cascade Lake: 第 2 世代インテル® Xeon® スケーラブル・プロセッサー



インテルによる AWS データ高速化

AWS C5 ディープラーニング AMI インテル[®] CPU に合わせて最適化 (トレーニング)

入手可能な Tensorflow 1.6 バイナリでの トレーニングと比較して

7.4 倍高速

AWS C5 ディープラーニング AMI インテル® CPU に合わせて最適化 (推論)

> MxNet による NMT 推論の デフォルト構成と比較して

12 倍高速

Amazon SageMaker 機械学習 インテル® CPU に合わせて最適化 (トレーニング & 推論)

IA CPU 向けに最適化された マシンラーニング・アルゴリズムの

10 倍高速









Amazon Forecast

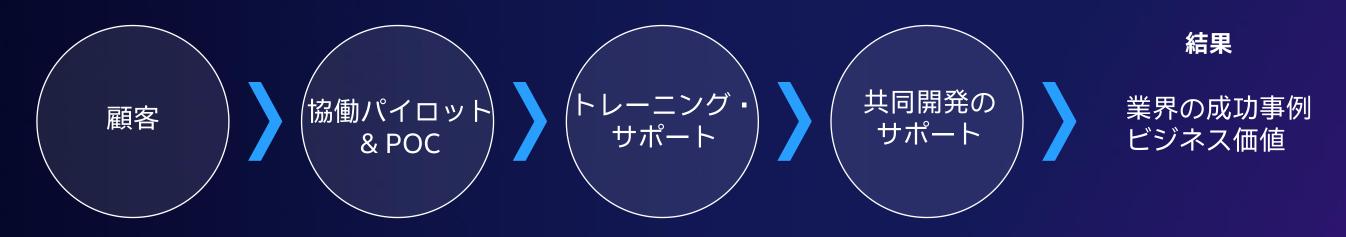


Amazon Personalize

AWS Marketplace: ディープラーニング AMI | OneDNN に最適化されたフレームワークを備える複数コンテナー



インテルの AI / ML 顧客プログラム



エンゲージメント方法

AWS Summit でインテルセッションに参加する

AWSMarketplace でインテルが最適化した 16 のソフトウェア・ライブラリーとフレームワークの詳細を確認する

FootAsylum、Thorn、Zignal Labs、Krispy Kreme、GE、Thomson Reuters、ASIC などの他のお客様から 学習する: <u>https://aws.amazon.com/jp/solutions/case-studies-jp/</u>

DataRobot、Data Bricks、IntellHQ、Intellify、Peak.AI、C3.AI、H2O.AI、Slalom などのパートナーとつながる



AIの成功事例



小麦の多様性研究のため、3TB の 植物育種と順化のゲノミクスデータを 数時間で分析しました。

EnglishHelper

デジタル装備の公立学校における ReadToMe® ソフトウェア。 学生のインタラクティブな学習を可能にし、 1 ~ 12 学年の教師の有効性を向上させる AI 対応の多感覚テクノロジーを搭載



オーストラリアの CSIRO 研究機関が、Amazon EC2 C5 ベースの large インスタンスを 98% 使用して、山火事の並列シミュレーション件数 を 60 倍増加させることができました。



研究者が絶滅危惧種のための ゲノム解読データを民主化し、 ML モデルを向上させました。



匿名化された政府および公開の新型コロナウィルスのデータポイントを使用して、 100以上のダッシュボードを提供し、新型コロナウィルスの感染を防止しました。

GRAYMATICS

インテル® ディストリビューションの OpenVINO™ ツールキットを使用した AI ソリューション。 シンガポールの海岸沿いで危険物を リアルタイムで識別、検出し、 対応できるようになります。



AWS顧客事例

THORN &



重要な時間の節約

65%

インテル® Xeon® スケーラブル・ プロセッサー・ファミリーに Amazon EC2 C5 インスタンスと Amazon S3 および Amazon Rekognition を搭載して、法執行 機関が児童の人身売買と戦うのに 役立ちます。 18,119

特定された被害者

5,791

特定された児童

6,553

特定された人身売買業者

Thorn が人身売買の 被害者をより高速に 見つける

法執行機関との管轄を超えた協力 を強化 ニーズ: 虐待者は子供たちを容易に搾取するため、高度なテクノロジーを使用しており、 法執行機関は状況を好転させ、子供たちをより迅速に見つける必要がありました。 ソリューション: Spotlight ソフトウェアが、1日に10万件以上のオンラインエスコート 広告を取り込み、Amazon S3に保存します。Amazon Rekognition は、運用分析のため の拡張可能なデータベースである MemSQL と組み合わせ、画像検索エンジンに勝つため に編集された写真の検索に役立ちます。Spotlight の ML モデルは、インテル® Xeon® ス ケーラブル・プロセッサー・ファミリー搭載の Amazon EC2 C5 インスタンスで実行され ます。

評価: Spotlight は、米国とカナダで 18,119 人の被害者と、5,791 人の子供および 6,553 の人身売買業者の特定を支援するために 65% の時間を短縮しました。

「Spotlight は、警察官が従来の筆記方法よりもはるかに高速に、子供の性的人身売買の広告を特定できます。」

ブルック・イストック Thorn 戦略および運用部門 ディレクター

製品とソリューション

Amazon Rekognition
Amazon S3

<u>インテル® Xeon® スケーラブル・プロセッサー</u>

業種 AI-ML / 画像分析 組織の規模 1 ~ 50

国 アメリカ合衆国 パートナー AWS

Thorn
MemSQL

詳細

<u>パートナーストーリー</u>





AI 搭載の顧客パーソナライズにより、 販売を促進

Footasylum と Peak は AWS を 活用し、広告費用対効果を劇的に 増加させることができます。

製品とソリューション

Peak.ai

Amazon EMR - Apache Spark を使用したビッグ・データ・プラットフォーム

インテル® Xeon® スケーラブル・プロセッサー

AWS アドバンスト・コンサルティング・パートナーの Peak は、AWS を利用したソリューションを実装し、Footasylum が AI 搭載の顧客セグメンテーションを通じてマーケティング・コミュニケーションの高度なパーソナライズを実現しています。

ニーズ: パーソナライズされたオンラインでの小売体験を提供し、メール・マーケティングを通じて広告収益を向上させることができます。

ソリューション: 顧客の購買データは、Amazon EMR で実行されている Apache Spark を使用して Amazon S3 に保存され、処理されます。インテル® Xeon® スケーラブル・プロセッサー・ファミリー搭載の AI システムは、Amazon EC2 C5 インスタンスで実行され、ML モデルをトレーニングし、モデルに基づいて予測を実行します。

評価: 予測と規定を利用し、8,400% の広告対費用 (ROAS) とメール・マーケティングによる 28% の収益向上を達成する方法で顧客グループをターゲットにしています。

リューションに必要な 計算能力を提供できま す。これは、私たちの ビジネスにとって重要 なものです。」

「インテルは拡張能力

を提供し、我々は AI ソ

Peak AI 小売責任者 Mylo Portas

業種 AI-ML / オンライ ン・マーケティング

1 ~ 50

組織の規模

ロ イギリス パートナー AWS Footasylum

<u>パートナーストーリー</u>

詳細



AWS 顧客事例







2500 万件

ワクチン投与を 1 日でサポート

クラウドネイティブな 国民スケールのプラッ トフォームで

インドの COVID-19 ワクチン投与の 計画、実施、監視と評価を実現

製品とソリューション

第 2 世代インテル[®] Xeon[®] スケーラブル・プロセッサー・ ファミリー インテル[®] ディープラーニング・ブースト インテル[®] アドバンスト・ベクトル・エクステンション 512 CoWIN (Covid Vaccine Intelligence Network) は、インドの保健家族福祉省が所有、運用する COVID-19 ワクチン接種登録のためのインド政府のウェブポータルです。このアプリケーションの開発は、Cloudfront、API Gateway、DynamoDB、さらにインテル製品による EKS、Aurora MySQL、Redshift と ElastiCache のインスタンスを活用した AWS クラウドネイティブなアーキテクチャーを利用して行われました。利用する全サービスに対して Active-Active のマルチ AZ 設定を行うことで、データセンター・レベルでの障害が発生してもアプリケーションの可用性に影響を出さず、3 つの AZ 全体におけるリソースのスケーリングも可能となり、同時にほぼゼロの RTO / RPO を提供してあらゆるデータ損失を回避します。このプラットフォームは、1 日で 20 億件の API ヒットに対応しました。

業種 ヘルスケア **組織の規模** インドの 10 億人 以上の市民 **国** インド パートナー MoHFW、MeiTY、 Trigyn 詳細

このプラットフォームの堅 牢性と AWS からのサポート により、たったの数週間で アプリケーションを提供し、 プラットフォームでは 1 日 当たり何十億件という前代 未聞のスケールに対応する ことができました。」

Digital India Corporation CTO デバブラタ・ナヤク





EnglishHelperは、 英語リテラシーの 向上と、より良い 学習成果の実現に 取り組んでいます。 言語学習テクノロジー企業、社会的企業である EnglishHelper (EH) は、イン ドの公立学校の学生が英語の学習や習熟で直面している課題に気づきました。 ウェアである ReadToMe が開発されました。EH は、第2世代インテル® ド全州の公立学校にソフトウェアを提供しました。

製品とソリューション

第2世代インテル®Xeon®スケーラブル・プロセッサー・ ファミリー

インテル[®] ディープラーニング・ブースト

インテル®アドバンスト・ベクトル・エクステンション 512

同社は長期的な解決策のために Schoolnet India Ltd と提携しました。言語は 多感覚環境での学習が最良であるという前提のもとに構築された、多感覚の 読解と理解のためのテキスト読み上げ音声テクノロジーを搭載した AI ソフト Xeon® スケーラブル・プロセッサー・ファミリーを搭載した AWS Elastic Compute Cloud (EC2) M5 のインスタンスを選び、拡張性を実現して、イン

パートナー 組織の規模 詳細 インド eラーニング 51 ~ 200 **AWS Schoolnet India** 導入事例 RightToRead プログラムは、十分 なサービスを受けていないインド (および世界中)の人々の間で、 英語リテラシーと習熟度を高める ことに集中する機会を与えてくれ ます。インテルと AWS は、クラ ウドを通して拡張性と適応性を提 供し、インドの最も遠い地域まで プログラムを拡大して、学生と教 師の両方に変化をもたらしました。

EnglishHelper CEO Sanjay Gupta



AWS顧客事例

intel. 3DAT



Habana Gaudi DL1 インスタン スで価格パフォーマ ンスが 32% 向上

インテル[®] 3DAT: 人間の動きに関 する理解の向上

オリンピックの試合に備えて、アメリカの多国籍企業であり世界テクノロジー企業の 1 つであるインテルが、3D アスリート・トラッキング (3DAT) についてのコンセプトを開発しました。3DAT は、競技中のアスリートのデジタルモデルをリアルタイムに作成し、放送中のファンのエンゲージメントを高めるための ML ソリューションです。コーチングや一流のアスリートのトレーニング目的で、インテルはテクノロジーの活用を目指していました。

3D ポーズ再編のための古典的なコンピューター・ビジョン方法は、ほとんどの科学者にとって面倒なものであることが証明されています。これらのモデルが、3D ラベルやモデルがないことから、アスリートに追加のセンサーを組み込むことに依存しています。通常の携帯電話では、シームレスなデータ収集機を導入できますが、2D ビデオには情報の奥行きがないため、2D ビデオデータを使用した 3D モデルの開発は困難です。インテルの 3DAT チームは、Amazon Machine Learning Solutions Lab (Amazon MLSL) を使用して、2D ビデオ上で 3D 人間のポーズ推定技術を開発し、コーチがアスリートのパフォーマンスのバイオメカニクスや、他の指標を抽出するための軽量ソリューションを作成しました。

製品とソリューション インテル 3DAT 動画について詳しくはこちら 3DAT re:Invent トラックセッション 3DAT + AWS Sagemaker ブログ

3DAT は、第3世代インテル® Xeon® スケーラブル・プロセッサー・ファミリーを搭載した Amazon の EC2 M6i インスタンスを活用し、コンピューティング価格パフォーマンスを 18% 向上することができます。Habana Gaudi Al プロセッサー搭載の Amazon EC2 DL1 インスタンスを使用して、ディープラーニング・モデルのトレーニングの速度と効率を向上させることができます。DL1 インスタンスは、価格 / パフォーマンスを最大 32% 向上しました。



その他の Habana DL1 参考資料:











「Habana Gaudi アクセラレーター搭載の Amazon EC2 DL1 インスタンスは、コスト面とパフォーマンス面での大きなアドバンテージに加え、将来的にはAWS コンピューティング・クラスターに追加されることを期待しています。Habana Labs は進化し、オペレーターの幅広いカバレッジを実現し続けるにつれ、更なる企業のユースケースの拡大が可能になり、それにより、更なるコスト削減を実現する可能性があります。」

Darrell Louder
Seagate オペレーションおよびテクノロジーアドバンスト分析担当 シニア・エンジニアリング・ディレクター AI とディープラーニングは、マシンビジョン機能の中核で、 私たちが提供する業界全体で、お客様により良い意思決定を 進めることができます。精度を向上するためにデータセットは、 より複雑になりつつあり、より大型で複雑なモデルが必要にな ります。これにより、コンピューティング価格パフォーマンス を向上させる必要性が高まっています。新しい Amazon EC2 DL1 インスタンスは、GPU ベースの Amazon EC2 インスタ ンスと比較して、大幅に低コストのトレーニングを約束します。 これにより、クラウドの AI モデルのトレーニングが、以前よ りはるかにコスト競争力があり、幅広いクライアントにとって アクセスしやすいものになることを期待しています。」

Fractal グループ CEO Srikanth Velamakanni Leidos とそのお客様がディープラーニング・モデル向けに、迅速で簡単で、コスト効果の高いトレーニングを必要としていることを考えると、Habana Gaudi AI プロセッサー搭載の Amazon EC2 DL1 インスタンスを使用するために、インテルとの旅を始めたことを嬉しく思います。 DL1 インスタンスを使用して、モデル・トレーニングの速度と効率が向上し、その後の研究開発のリスクとコストが削減されることを期待しています。」

Leidos Health and Human Sciences CTO Chetan Paul



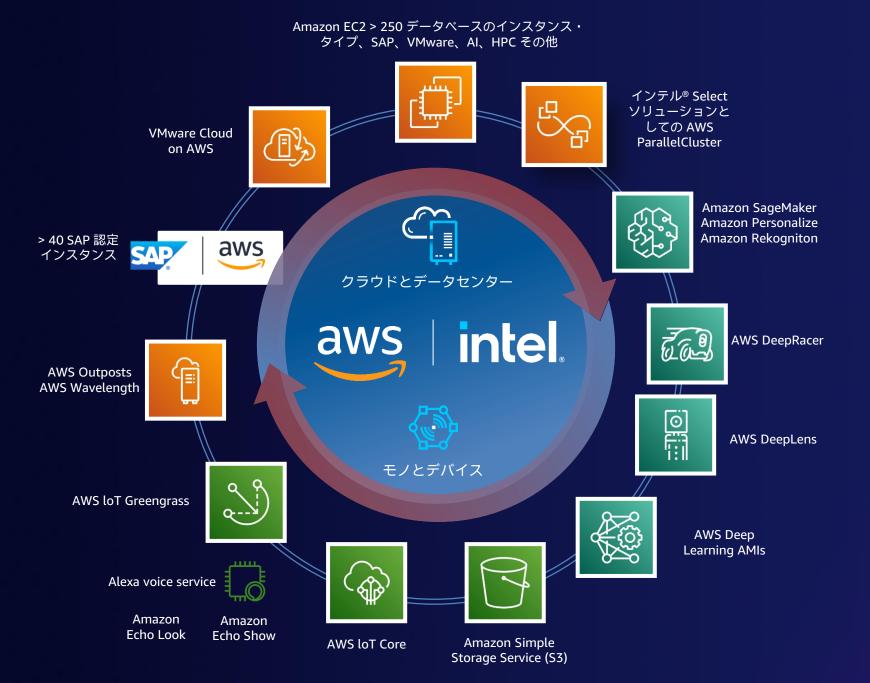
<u>Habana DL1 インスタンス参考資料</u>

インテル / AWS 素晴らしい組み合わせ

https://youtu.be/LfhSaaULoe4



サマリー



- インテルと AWS 間の密接なコラボレーションは、 クラス最高のエンドユーザー体験と顧客の成功を もたらしました。
- インテルで最高の TCO により、さまざまな種類や ワークロード全体で顧客のアプリケーションを 高速化します。
- 多くの成功した成果で、導入のための既存のソ リューションが、ハイ・パフォーマンスとコスト 削減の両方を実現します。

学習 | もっと知る | インテル AI のエンゲージメント



学習

AWS およびインテルに関する詳細を https://aws.amazon.com/jp/intel/で学習



詳しくは こちら インテル® Xeon® スケーラブル・プロセッサー 搭載の AWS の新しいインスタンス (M5、C5、C5n、R5、T3)



ニンゲー

インテル® AI や POC の機会 / ケーススタディー については、インテル / AWS の担当者までお問い 合わせください



世界を変える技術を創造し、地球上のすべての人の生活を豊かにします



Thank you!

アカンクシャ・バラニ インテルグローバル AI GTM リード AWS APJ Alliance ヘッド

