Cloud TPU で実現する AI SaaS プロダクトづくり

ストックマーク 取締役CTO 有馬

Google Cloud

Google

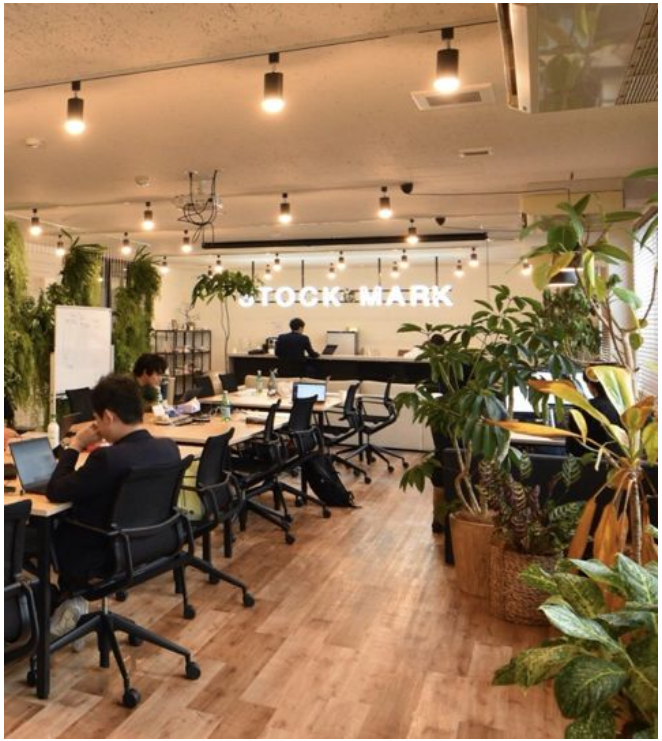
アジェンダ

1. 会社及びプロダクトの紹介
2. Cloud TPUを活用したAI SaaS アーキテクチャの紹介
3. Cloud TPUの活用インパクトの紹介

会社及び
プロダクトの紹介

01

会社紹介



会社名	ストックマーク株式会社
オフィス	東京都港区南青山1丁目12-3 LIFORK MINAMI AOYAMA S209
設立	2016年11月15日
創業者	代表取締役CEO 林 達 取締役CTO 有馬 幸介
事業内容	自然言語処理技術技術を活用した ビジネス意思決定サポートサービスの提供
従業員数	60名

今すぐ使える3つのAI SaaS(Aseries) を提供中

A news

全部署
新規事業

ビジネスのミッションに直結するニュースを
AIが収集し流通させ、全社員の現場力を向上

A strategy

経営企画
新規事業

ニュースや社内外の情報を解析しマーケット動向や
競合をAIが解析し、ビジネスの先手を打つ

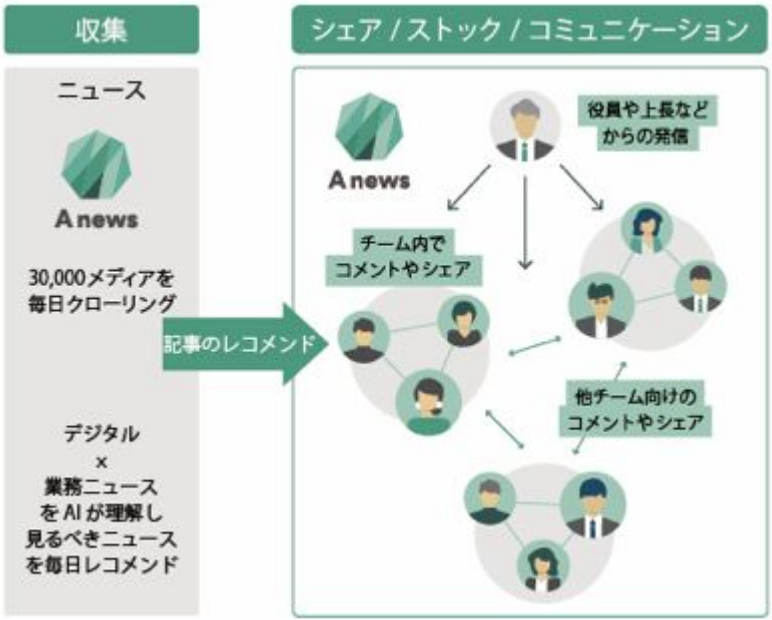
A sales

営業企画
営業現場

商談やメールのテキスト情報から営業の勝ち
パターンをAIが解析し、より売れる営業組織に

Anews

必要なビジネスニュースが毎日届く
ナレッジシェア活性化による組織の情報感度向上



1. 国内外3万ソースから、組織のミッションに即したニュースをAIがレコメンド
2. コメント機能で簡単にアイデア共有ができ、チームの情報感度/ナレッジシェアを促進

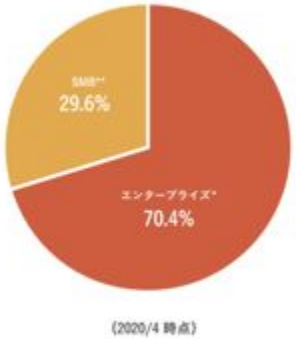
Google Cloud

Google

Anews 導入実績

Anewsは 2017 年 4 月リリースから累計 **1500** 社でご利用

顧客分布

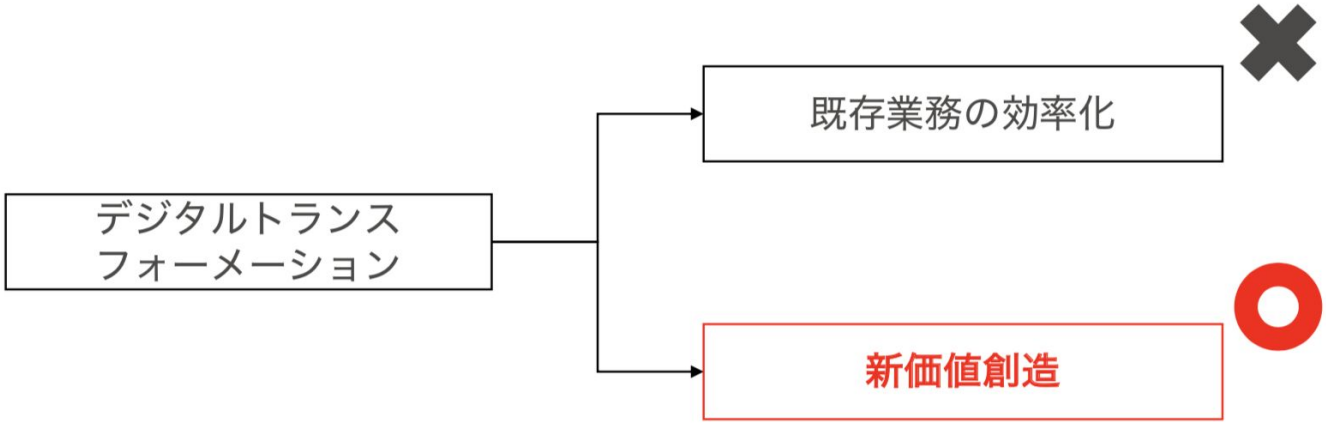


*エンタープライズ：従業員数1,000名以上の企業
**SMB（スモールトゥミディアムビジネス）：1,000名未満の中小企業

MIZUHO MUFG セブン銀行 三菱商事 住友商事
あなたの未来を強くする 住友生命 かんぽ生命 SOMPO ひまわり生命 TOPPAN TOTO
LIXIL TEIJIN SAPPORO SUNTORY Asahi アサヒ飲料
大日本住友製薬 HITACHI Inspire the Next NS Solutions SoftBank ROYAL
JTB Benefit KANDAI Me RISE WOWOW ZACROS

DXの中でAseriesが目指すもの

- Aseriesは業務の効率化ではなく新価値創造にフォーカスしている



非構造化データを活用しヒトの攻撃力を高める

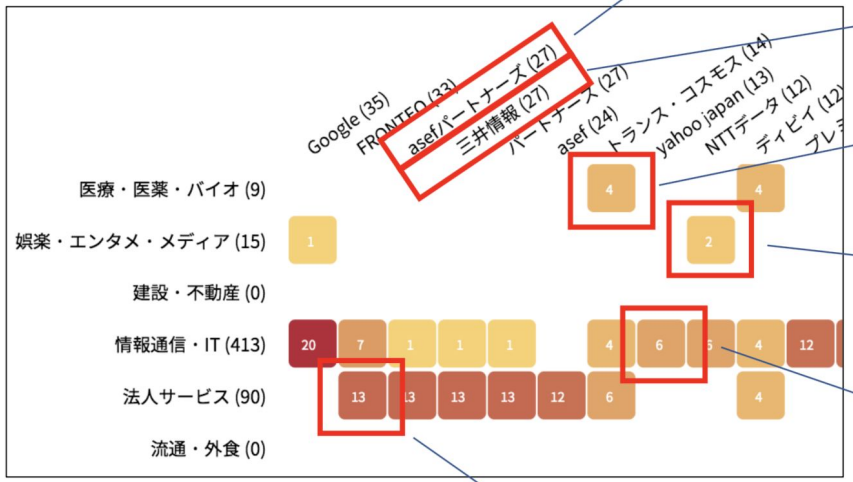
- 世界の90%を占める非構造化データを構造化し、ビジネスマンの攻撃力を高めるAI SaaS

コンサル/新規事業企画者
社外に発散しているオープンデータを構造化し、企画力を高める

営業担当者
社内に眠るクローズドデータを構造化し、提案力を高める

オープンデータを構造化し、企画力を高める

Astrategy



未知の有力ベンチャーの発見

未知の有力子会社の発見

大手企業の新規領域への参入事例の発見

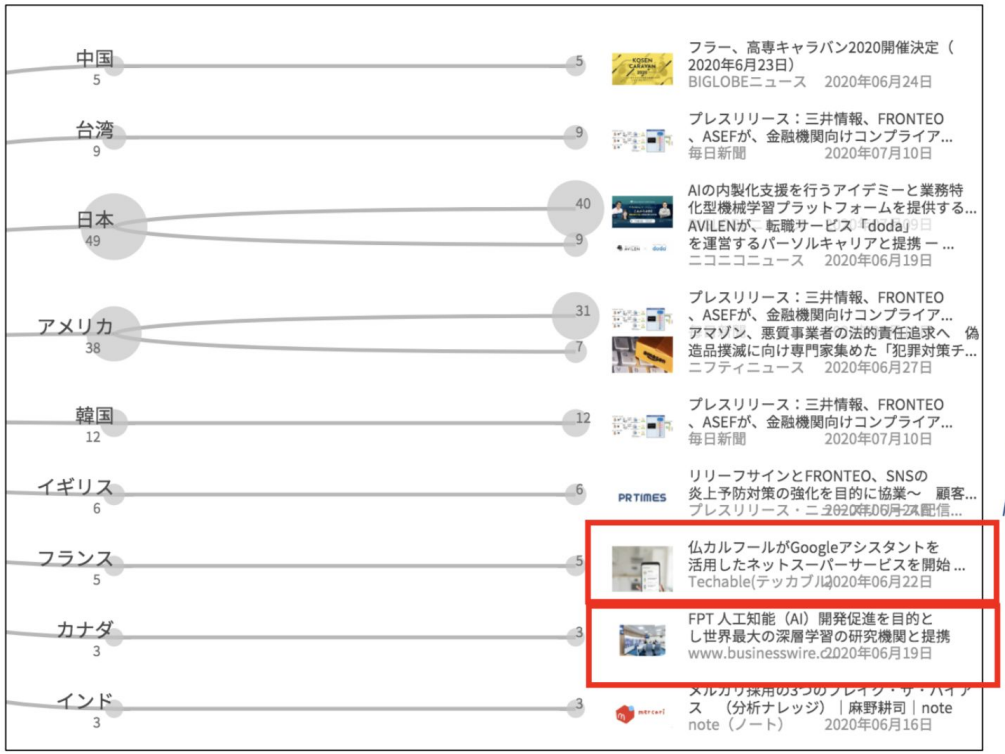
大手企業の近隣領域への参入事例の発見

提案先業界の先行事例の発見

提案先近隣業界の先行事例の発見

オープンデータを構造化し、企画力を高める

Astrategy



[未知のビジネスモデルの発見]

仏カルフールがGoogleアシスタントを活用したネットスーパーサービスを開始

[未知の有力機関の発見]

FPTがケベックのディープラーニングに特化した世界最大の学術研究所Milaと提供

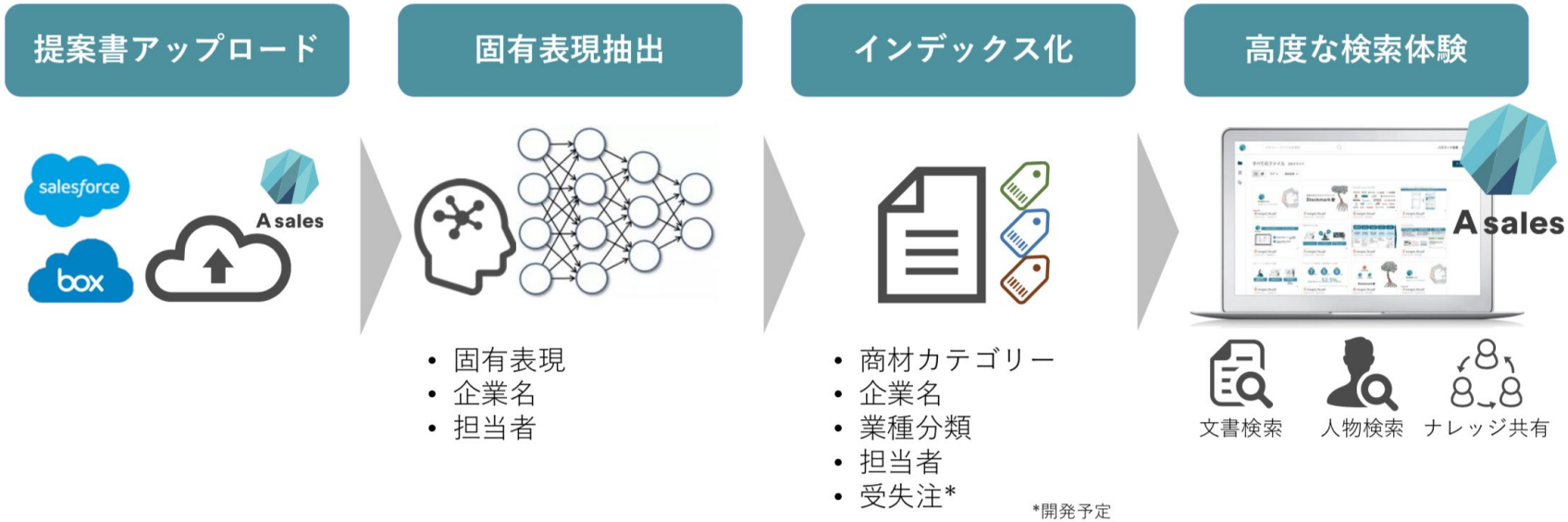
Google Cloud

Google

クローズドデータを構造化し、営業力を高める

A sales

- アップロードされた提案書群をAIがインデックス化
- 社内の全提案を構造的に俯瞰し、顧客訴求力の高い提案が作成可能



Cloud TPUを活用したAI SaaSアーキテクチャ

02

高い構造化精度を得るためには
先端的な重量級AIモデルの活用が必須

- BERT等の先端AIモデルは数億パラメータで構成されており
  通常の計算ユニットでは現実的な時間で処理が終了しない

重量級AIモデルをSaaSビジネスで活用するには Cloud TPUの活用が(実質的に)必須

- 巨大なデータ量の学習処理が必要
  - 弊社AIモデル(BERT)は 300 万ビジネス記事を学習
- 巨大なデータ量の推論処理が必要
  - Astrategyでは日次で数 10 万ビジネス記事を推論
- AIの処理に特化した演算ユニットであるCloud TPUの活用が必要に

TensorFlow Research Cloudプログラムにも参加

- TFRCプログラムに参加し、TPUの有用性を確認。SaaS内運用に着手

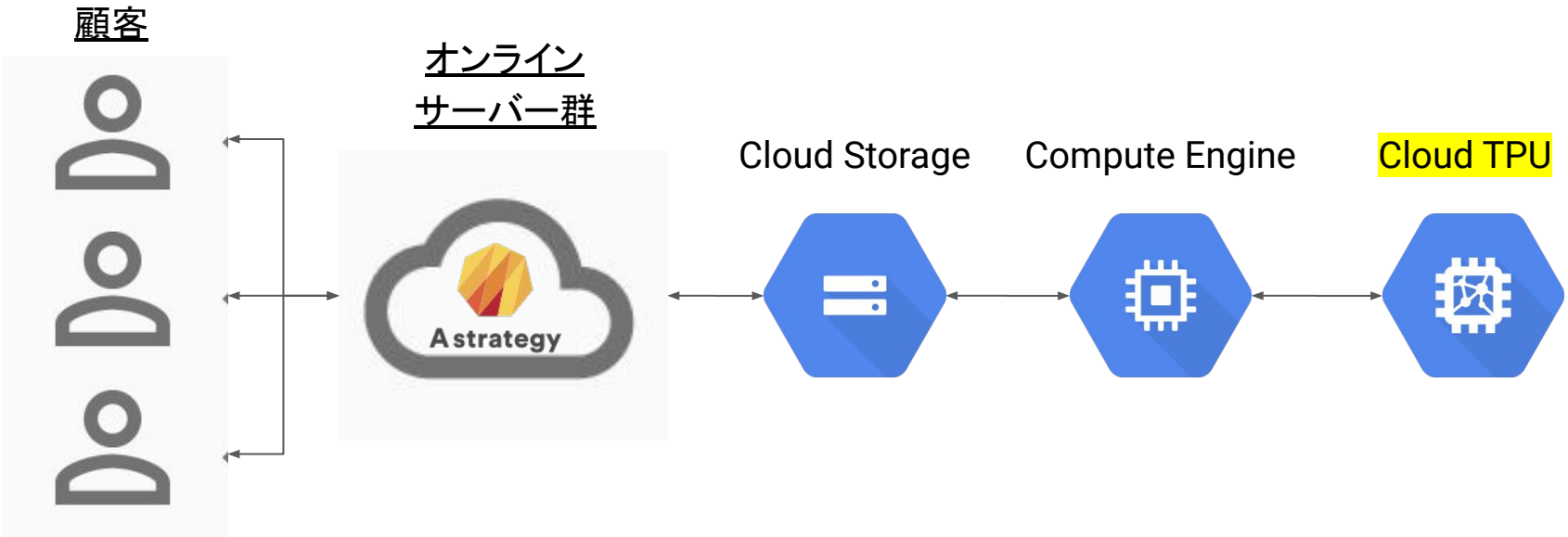
TensorFlow Research Cloud（TFRC）プログラムとは

TensorFlow Research Cloud（TFRC）プログラムに申し込むと、1,000 個以上の Cloud TPU から成るクラスタにアクセスできるようになります。このクラスタには、計 180 PFLOPS 以上の処理能力があります。TFRC プログラムに参加するとこれらの Cloud TPU を無料で使用できるため、研究を大きく前進させることができます。

TFRC プログラムの参加者は、TFRC を利用した研究結果を、査読を受けた論文、オープンソース コード、ブログ投稿などの形で全世界に公表することが求められます。また、具体的なフィードバックを Google に提供することで、TFRC プログラムおよびその基盤となる Cloud TPU プラットフォームの将来的な発展に貢献することが期待されています。プログラム参加者は、Google の利用規約に同意し、さらに、参加者自身の情報が Google のプライバシー ポリシーに従って使用されること、および、Google の AI に関わる基本理念に沿って研究をすすめることに同意する必要があります。

Cloud TPU を活用した Astrategy のアーキテクチャ

- 各AI処理ごとにCloud TPUをコールしCloud Storageを経由して計算結果をSaaSで顧客へご提供



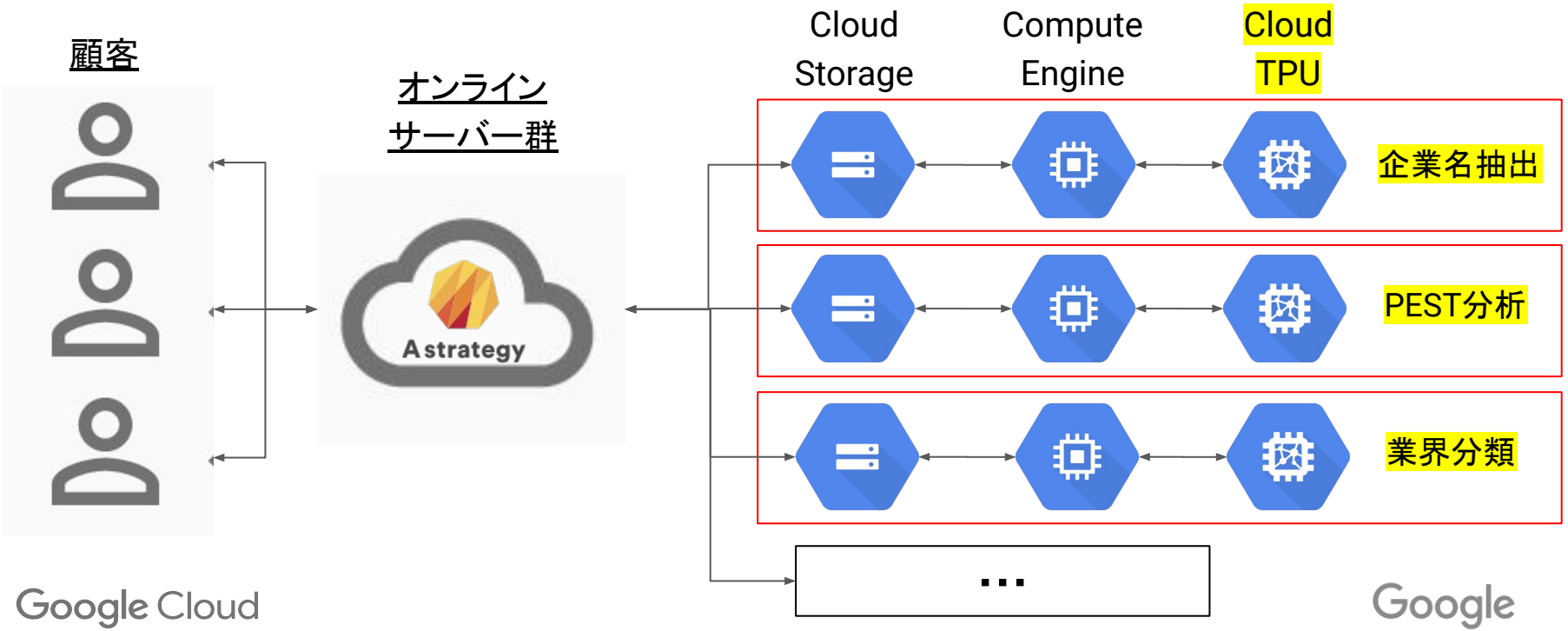
Astrategy が行なっているAI処理の一覧

以下の各種AI処理を日次で数 10 万ビジネス記事に対して実施

- 企業名抽出
- PEST分析
- 記事質の分類
- 業界分類
- 地域別分類
- トレンドキーワード抽出
- 特徴事象のクラスタリングと抽出
- 類似事象の集約
- etc, etc...

Cloud TPU を活用した Astrategy のアーキテクチャ

- 各AI処理ごとにCloud TPUをマイクロサービス的にコールしCloud Storage を経由して計算結果をSaaSで顧客へご提供



Cloud TPUの
活用効果の紹介

03

Cloud TPU の実運用性能のご紹介(学習編)

学習 (pre-training) 処理の性能評価

- 学習データ: 約 300 万ビジネス記事
- (* GPUでもNVLinkによるスケールアウトで時間短縮は可能と思料)

	処理時間	費用
GPU (NVIDIA V100)	概算60日以上程度	$4400以上
Cloud TPU	5日程度	$530程度

Cloud TPU の実運用性能のご紹介(推論編)

推論(企業名抽出)処理の性能評価

- 推論データ:約 50 万ビジネス記事
- Cloud TPUにて「安くて速くて美味い」SaaSが実現可能

	処理時間	費用
GPU (NVIDIA V100)	103分	$4.3
Cloud TPU (v2-8)	29分	$2.2

Cloud TPU があることで可能となったこと

1. 学術界レベルの高い精度のAI処理をSaaSに組み込み顧客へ提供可能になった
2. リアルタイムに近い形でのAI処理が提供可能になった

タイムリーかつプロレベルの業界分析

新鮮な顧客ニーズの発見

今後ますます高まる Cloud TPU の重要性

- 先端的なAIモデルはますます巨大化して行っている
  - GPT-3などは1750億パラメータ！

- Cloud TPUのような高効率AIプロセッサークラスターの活用の重要性は、SaaS運営において今後ますます高まっていく

Thank you