Informatica™

ビッグデータ ワークブック

初めてのビッグデータプロジェクトを立ち上げるための実践ガイド

目次

ビッグデータへの対応 **03**

パート1はじめに **04**
知っておくべき事 05
– 多くの組織がビッグデータプロジェクトを導入する理由 06
– ビッグデータ プロジェクトが失敗する原因 07
– ビッグデータ プロジェクトを機能させる方法 08
適正なプロジェクトの選択 10
– 適正なプロジェクトのあるべき姿 11
– 影響を考慮する 13
– 実践的なビッグデータプロジェクト:事例 15

パート2戦略 **16**
目標を策定する 17
– ビジネス目標 18
– IT 目標 20
データ要件を定義する 22
– 必要なデータは何か? 23
– データに関する 5 つの重要な検討事項 26

パート3リーンなビッグデータサプライチェーン **28**
チーム 29
– チーム編成における重要な5 つの教訓 30
– データ ガバナンスを設定する 33
– 必要なスキルと既存のスキル 35
– ビッグデータのツールについて 37
プロセス 38
– ビッグデータの8 つのステップ 39
アーキテクチャ 41
– 最初のステップ: サンドボックス (テスト環境) 42
– ビッグデータの理想的なアーキテクチャ 43
プロジェクト計画 44
– 始める 47

次のステップ **48**

インフォマティカについて **49**

クリックで各セクションに移動できます

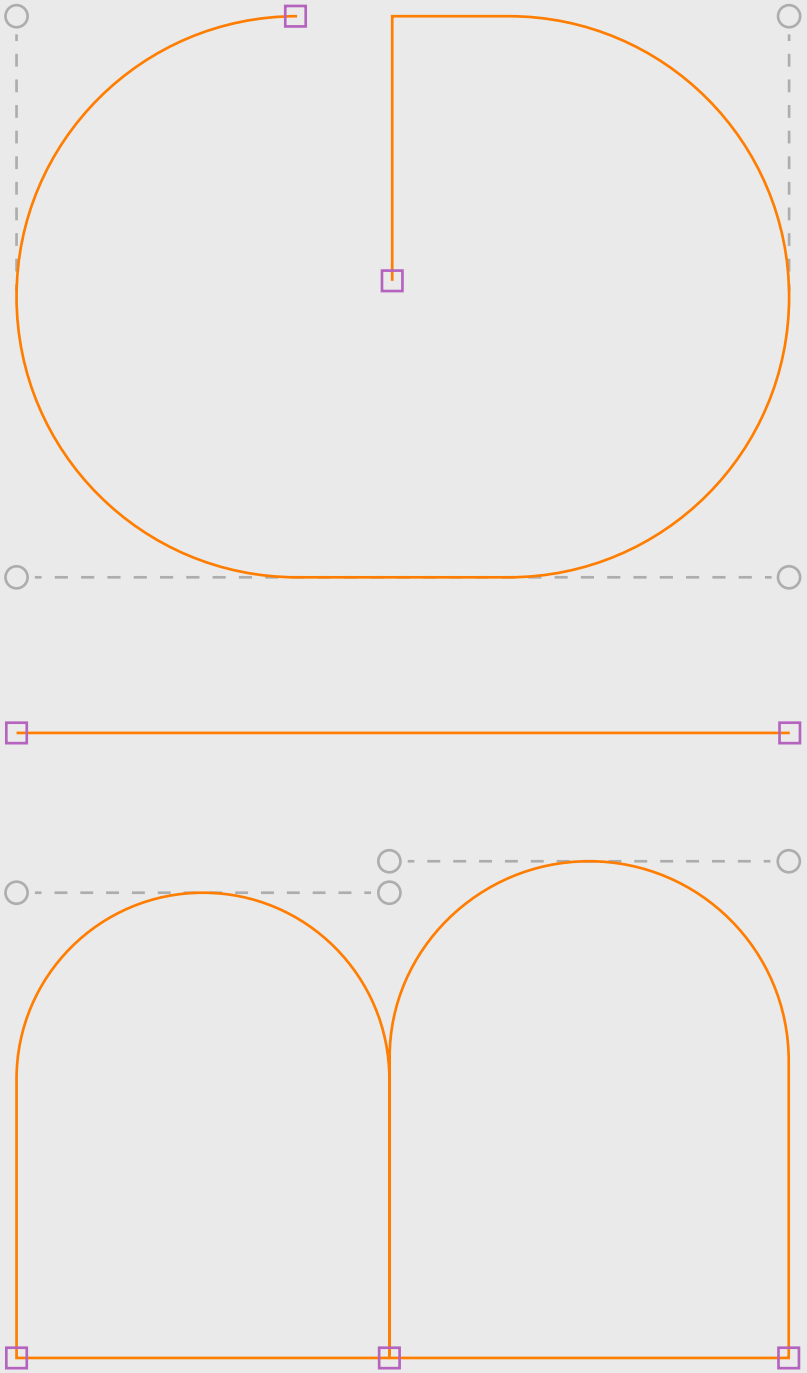
ビッグ
データへの対応

これまでに、ビッグデータほど誇大に言いはやされてきたテクノロジー トレンドは、ほとんどありませんでした。

同時に、これほど多くの変革の可能性を企業にもたらしたテクノロジートレンドもありません。世紀の変わり目にソフトウェアがビジネス プロセス全体を包含するようになって以来、明白な事があります。それは、データは、私たちのワークスタイルを変えるという事です。

そして貴社の戦略を守りつつ、実践的に実行する方法を説明します。地域限定的で戦術的なイニシアチブに着手するのか、あるいは企業全体の根幹に関わるようなプロジェクトなのかに関係なく、本ワークブックはプロセスを進める上での実践的なガイドとして役立ちます。

では、始めましょう。



パート1

はじめに

このワークブックは３つのパートに分かれています。最初のパートでは、ビジョンを明確にして、適正なプロジェクトを選択できるようになりましょう。

知っておくべき事

プロジェクトの詳細を掘り下げる前に、多くのビッグデータ実践者がプロジェクト開始前に知っておきたいと思う教訓をいくつかここでご紹介します。

多くの組織がビッグデータプロジェクトを導入する理由

一般的に、企業がビッグデータに取り組む理由は次のいずれかです。

より詳細な分析を実施するためには、分析対象としてかなりのデータ量が必要であるという認識。通常は、業務部門 (マーケティングなど) が、こういったイニシアチブを開始します。

自社製品を顧客サービスに役立てるためには、顧客がより効率的かつ効果的に製品を使えるような分析情報を (多くの場合、リアルタイム) に提供する必要があるという認識。

全ての意思決定者がビッグデータを活用することで、特定の業務部門やプロセスについての課題を、より迅速かつスマートに、コストをかけずに決定したいと思っています。

全ての部門にとってビッグデータは重要であり、自社のビジネス全体をデータ中心にするための基盤を築こうとしています。

しかし手遅れになる前にビッグデータに精通する必要があることは理解していても、具体的に何をどうすべきかが分かっていません。学習して実際に試してみることが狙いです。

上記の理由はどれもビッグデータに関心を寄せるには十分です。しかし、プロジェクトがテスト段階にある場合 (さらに複数部門によって精査される場合)には、貴社に最も当てはまる理由はどれかを明確にしておく必要があります。

ビッグデータ プロジェクトが失敗する原因

調査[1]では、全てのビッグデータプロジェクトのうち 55% は完結できず、残りの多くは目的を達成できなかったことが判明しています。

この驚くべき比率は、テクノロジートレンドの初期段階ではよくあることですが、失敗したプロジェクトが教えてくれる教訓を学ばない手はありません。

ビッグデータ プロジェクトが失敗する、主な 4 つの原因を見ていきましょう。

1. 目標が曖昧

調査で最も多く挙げられている失敗の原因は、プロジェクトの「範囲が不正確な事」でした。あまりに多くの企業が、プロジェクトに対する期待を膨らませ、それと同時に、明確な目標のないままプロジェクトを立ち上げてしまうため、何が重要で何がそうでないかを選択するような厳しい場面で、プロジェクトがとん挫してしまうのです。ビッグデータ プロジェクトのために、ビッグデータを追い求めることは、災いの元です。こういった複雑なプロジェクトでは、一定の成果を上げるための強いコミットメントが必要です。確固たる目標がなければ、不可能なのです。

2. 過度の期待

ビッグデータにまつわる誇大な謳い文句のため、プロジェクトの成果に過度な期待をしてしまうのも事実です。短い間に大胆な約束を打ち出すことは魅力的な事でしょうが、プロジェクトから何を期待できるのか、プロジェクトにはどのくらいの時間を要するのか、目標達成のための労力はどの程度なのか、といった現実的な視点を持つことが重要です。効果と洞察に対する期待が膨らみ過ぎると、まだ見ぬ金塊を求めて膨大な量のデータの中をさまようことになるでしょう。成果への期待値が現実的でない場合、全く不可能な期日と予算に追われることになります。

3. プロジェクトの超過と遅延

ビッグデータの導入がどれだけ目新しいものであるかを考えれば、ほとんどのビッグデータプロジェクトが完了する頃には、コストや時間が超過してしまうのも頷ける事です。これは一般的に、過度な期待値をコントロールできなかった事と、スケーラブルなアーキテクチャの構築方法を誤った事の両方が原因です。希少で高価な Hadoop-Java 開発者を雇用し、膨大な量を手作業でコーディングするとなれば、エラーを完全になくして、テスト環境から抜け出すことは不可能なことはすぐにわかります。その結果、ビッグデータプロジェクトは科学的実験としての価値を失い、研究室から出て、日の目を見ることは二度となくなるのです。

4. 拡張性のなさ

長期的な持続可能性ではなく、短期的な便宜を求めてしまうことがあまりに多いのです。このようなトレードオフがあること、それを回避できる方法について言及しないまでも、長期的な観点を持つことの重要性は、いくら強調してもし過ぎることはありません。貴社のデータを適正にセキュリティで保護して管理するためには、プロジェクトの長期的な視点で据える必要があります。

ビッグデータが失敗する 4 つの原因を考えると、不安に思われるかもしれませんが、実は非常に一般的なものです。では、失敗を回避し、持続できる実装方法を見ていきましょう。

[1] InformationWeek, “Vague Goals Seed Big Data Failures.”

ビッグデータ プロジェクトを機能させる方法

多くのビッグデータ プロジェクトが、不明確で、イニシアチブの実践的な側面を明示できずに失敗に終わるなら、それを肝に銘じてプロジェクトにしっかり焦点を当てて実証すべきです。ここで、プロジェクトに取り組み、推進していくために役立つヒントを３つご紹介します。

1. 明確な目標を定め、期待値を管理する

プロジェクトの目標が曖昧なら、既存のデータ インフラストラクチャに設定した目標を検討してみましょう。

もしも、ある業務 (不正の検出または市場分析など) でデータを必要としているなら、それをより良く、価値あるものにするために、ビッグデータをどのように活用できるかを考えます。まずは全く新しい問題に取り組むのではなく、既存のプロセスまたはプロジェクトを改善することを目指します。

明確に照準を定めても、業務ユーザーに明示できるような価値がなければ、プロジェクトは破綻します。

2. プロジェクトの価値を実証するための指標を定義する

目標に合わせた指標を明確に定義すれば、多くの問題を回避することができます。測定可能で現実的な目標を定めれば、誰でも進捗状況を確認できます。

さらに重要な事は、長期的に何を目指しているのかを周囲に知らせることができる点です。そのためには、目標に照らし合わせながら、プロジェクトの効果を目に見える形で測定する方法を考える必要があります。

これは非常に重要な事です。なぜならプロジェクトの実装に伴い、業務ユーザーは短期的に妥協しながらサポートすることになるので、目に見える形で成果が分かる合理的な目標を設定することで、彼らが感じている以上の価値を提供できることを実証できるからです。

3. ツールや手作業コーディングに対する戦略を練る

Hadoop に直接、手作業で全てをコーディングしようという誘惑に負けないようにしましょう。忘れてはならない事。それは、ゼロから手作業で実効性のある導入を目指すのではなく、ビッグデータの価値を組織に提供する事が目標だという事です。

統合、全データセットのクレンジング、全てのアナリティクスを手作業でコーディングする代わりに、プロセスに関わる作業を支援するツールとプロセスを自動化することに目を向けましょう。

さらに重要なのは、希少高価なJava 開発者の能力を無駄にしない事です。ここで成すべき事は、目標を達成するために、少ないリソースを戦略的に配分することです。

今ある ETL のスキルと知識、データ品質、ビジネスインテリジェンスのプロフェッショナルを活用すると同時に、導入チームの生産性を高められるようなツールを採用しましょう。ツール

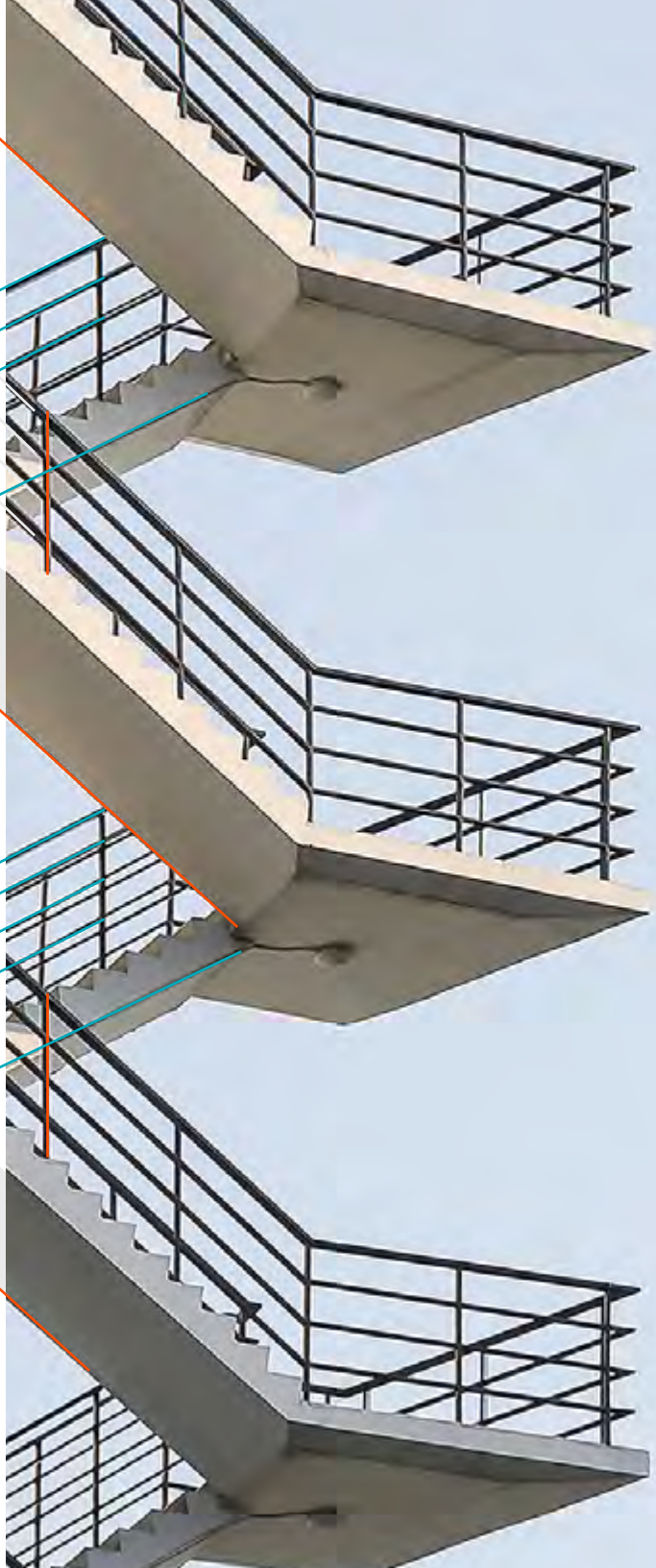
を利用できない特別なロジックは、能力の高い Java 開発者に担当させます。

さらに、Hadoop のようなテクノロジーは日々進化しているため、常に変化する基本テクノロジーの仕様変更を吸収できるような抽象化レイヤも検討の価値があります。

とにかく、必要なスキルは希少ですが、ツールはいつでも利用できることを忘れないで下さい。

適正なプロジェクトの選択

想定される課題を踏まえ、組織にとって適正なプロジェクトを選択するにはどうすべきかを考えてみましょう。



適正なプロジェクトのあるべき姿

組織が変化を求めており、質を高めるためには総合的なデータ ガバナンスのフレームワークが必要であることをすでに理解しているのであれば、本セクションは読み飛ばしても構いません。

逆に、地域限定的な戦術的プロジェクトを、将来的には社内の広範囲に適用する事を検討しているなら、このまま読み進めて下さい。

適正なプロジェクトは、次の 4 つで構成されます。

1. 実証可能な価値

適正なプロジェクトとは、IT 部門と業務部門の両方に均等に価値をもたらします。つまり、明確な価値を目に見える方法で部署、部門、グループに提供しなければなりません。

2. スポンサー

ビジョンの実現に賛同してくれる経営幹部は、プロジェクトの成功に不可欠な存在です。ビッグデータプロジェクトには、プロジェクト作業を弁護してくれるような経営層の支持と出資が必要です。

従って、ロジスティクスに関する優れたアナリティクスを構築できることが分かっていながら、経営層からの賛同者が CMO(マーケティング最高責任者)だけならば、見直す必要があります。マーケティング部門が支持者になるなら、マーケティングの分析要件を満たすプロジェクトにしましょう。誰にも変化を強制することはできません。影響力を借りて、そこから可能な限り最大限の価値を引き出すようにする必要があります。

3. ボウリング ピン効果

最初の戦術的プロジェクトには、戦略的な重要度が求められます。

自身が属する部門にビッグデータが役立つのかという疑念を晴らすだけでなく、その価値を社内全体が簡単に理解できるようにしたいものです。

こうした理由から、最初のプロジェクトは、戦略的に選択する必要があります。

例えば、ビッグデータの価値をマーケティング部門に示せば、あまり乗り気ではなかったロジスティクス部門からの賛同も得やすくなるでしょう。

4. 共有可能なスキル

最後のポイントで述べたように、最初のプロジェクトで得た価値を、他部門を説得するためにも役立てたいものです。そのためには、最初のプロジェクトで、適正なスキル、能力、知見を確実に学びとる必要があります。さらに、その内容を次のプロジェクトに活かせるように全て文書化する必要もあります。成功を目指すのであれば、今後のプロジェクトにも目を向けなければなりません。

さらに多くのプロジェクトを手掛けられるように、拡張のための準備は怠らないようにしましょう。これは、単にクラスタの拡張の問題だけではありません。スキルや業務の拡張にも関わってくることです。さらに多くの優秀な Java/Hadoop 開発者を見つける必要や、既存のリソースをさらに活用できる方法を見つける必要があるかもしれません。

影響を考慮する

次のプロジェクトを検討する際、それが組織にどのような影響を与えるのかも考慮する必要があります。ビッグデータプロジェクトの適正を判断する上で、考慮すべき側面は大きく分けて 3 つあります。

1. コストと破壊的イノベーション

最も基本的な事として、プロジェクトを進めるにあたって必要となる時間と金額がコストの基盤となります。実際には、プロジェクトが引き起こす破壊的イノベーションの影響も考慮する必要があります。

破壊的イノベーションは、例えばデータを所有している部門が、一元化されたデータ ガバナンス フレームワークに移行することで、データに対するコントロール権限を手放すことに難色を示すといった場合もあれば。

また、新しいテクノロジーを既存のインフラストラクチャに統合することで生じるスキルの見直しやアップグレード等の技術的なスキルに関連する場合もあります。

いずれの場合も、このような影響を想定、認識し、その影響を最小限に抑えるか、破壊的イノベーションが高い価値をもたらすことを周知させる必要があります。

2. メリットと影響のタイミング

開始時期が異なる複数のプロジェクトを検討する場合、ビジネスへの影響と改善効果が最も大きいプロジェクトを重視するのは自然な事です。しかし、ビジネスへの影響の本質を考慮することも重要です。価値の大部分を短期間で実現できるのか、または長期間を要するのか?

さらに重要なのは、業務ユーザーがビジネスへの効果を実感できるのはいつなのか?例えば、マスターデータ管理をデータウェアハウスに導入し、ビジネス インテリジェンスの効率性が飛躍的に改善されたとしても、ビジネス アナリストは、財務データをクレンジングする必要がなくなって初めてその価値を実感します。

3. リソースと制約

これまでに分析した 2 つの要素から、利用できるリソースを検討します。詳細は後述しますが、ここではプロジェクトに投資した金額を上回る効果を生み出す事だけにフォーカスしましょう。

この目標を達成すれば、2 つの要素の両方に効果をもたらします。ビジネスへの効果を最大化すると同時に、予算の使い方を戦略的に考える必要もあります。Google 社に比肩し得るデータ サイエンティストのチームを作ろうと思っても、予算的に可能でしょうか?プロジェクトの成功には、使えるツールと人員を考慮して、賢い選択をすることが不可欠です。

実践的なビッグデータ プロジェクト:事例

ビッグデータの応用例は多岐にわたります。それは魅力的な一方で、どのプロジェクトから手を付けるべきか確信が持てない人にとっては簡単な事ではありません。ここでは、弊社のお客様が行っている実践的なビッグデータ プロジェクトの例を紹介します。

まだ組織として行うべきプロジェクトが明確ではない場合、次の例を参考にして、ビッグデータが貴社にもたらす価値を考えてみましょう。

金融業

- リスクおよびポートフォリオ分析
- 投資提案

小売業

- プロアクティブな顧客エンゲージメント
- 地域に根付いたサービス

ヘルスケア業界

- 患者の予後の予測
- 総治療費
- 新薬開発

製造業

- コネクテッド車両プログラム
- 予防保全

公共部門

- 健康保険
- 為替
- 税金の最適化
- 不正の検出

パート2

戦略

それでは、ビッグデータに関する次の(または最初の)プロジェクトの要件を、実践的な観点から考えてみましょう。

目標を策定する

ペンを用意して下さい。すでに説明したように、ビッグデータのプロジェクトが失敗する原因で最も多いのは、明確な目標がないことです。想定しているプロジェクトが曖昧なために失敗しないようにしましょう。

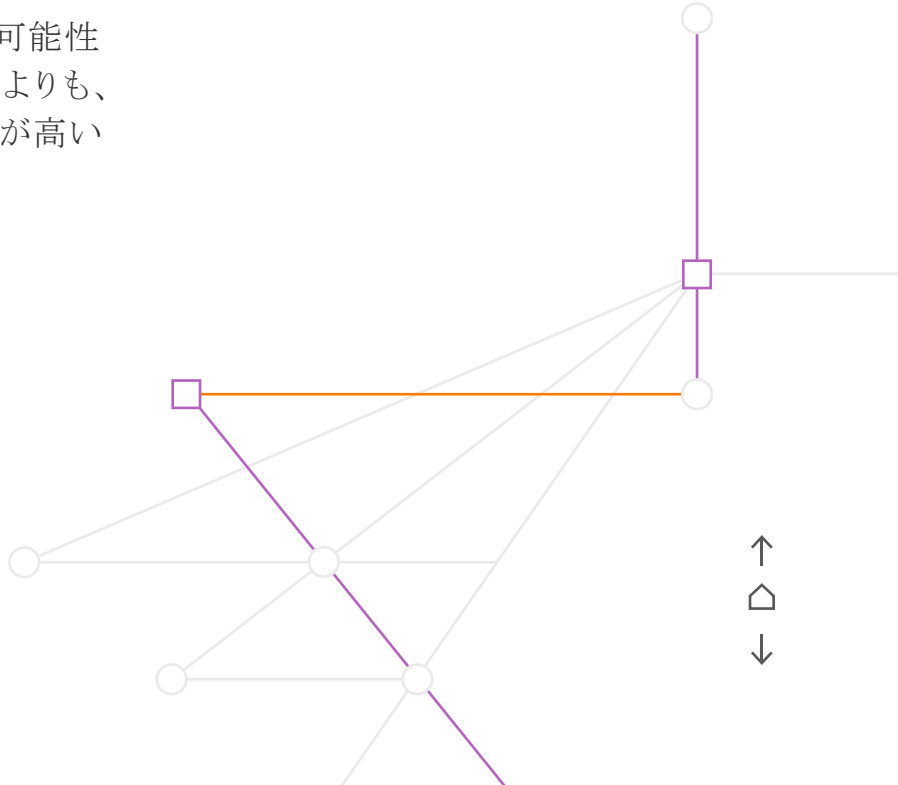
ビジネス目標

最初にビジネス目標を取り上げる理由は、プロジェクトが全面的に承認されれば、IT 目標よりもビジネス目標を優先しなければならないからです。

プロジェクトが達成すべきビジネス目標は、できる限り具体的に定義します。効果を測定できるような目標を設定しましょう。

例えば、顧客離れを予測するカスタマーサービス インターフェイスの場合、プロジェクトの目標は「カスタマー エクスペリエンスの向上」のような曖昧なものにするべきではありません。

目標が明確であるほど、達成の可能性も高くなります。曖昧な目標 1 つよりも、詳細に絞り込んだ目標 5 つの方が高い価値があります。



ビジネス目標

ビジネスや業務ユーザーに関連するビッグデータプロジェクトの目標を、重要度が高い順に記載します(目標の数は、自由に追加・削除可)。

例: 顧客離れ (チャーン) の削減

目標を達成するまでの最短期間と最長期間を記載します。

例: 3 ～ 6 カ月

目標毎に、達成度合いを確認できる基準を記載します。既存の評価基準や計算式が理想的です。

例: 月間の顧客離れ率を X% 削減

ビッグデータ プロジェクトの所要期間はどのくらいか？

ビッグデータ プロジェクトの期間は、その価値を最大限に提供できるまで十分な期間に設定します。経験上、これはプロジェクトの対象範囲によって決まります。

戦術的プロジェクトを 3 カ月未満で完了したお客様がいた一方で、 基盤となるプログラムの構築に 3 年を費やしたお客様もいました。

長期的なプロジェクトでは、6 カ月ごとにプロジェクトの価値を実証できるようにします。またアジャイルなアプローチを採用すれば、個々のフェーズやマイルストーンを小規模なプロジェクトとして提供することができます。

明確なことは、プロジェクトの所要期間を推測に頼って決めてはいけないということです。経験や以前に同様のプロジェクトを実施した他の人の経験を基に期間を見積もりましょう。適任者がいない場合は、弊社にご連絡下さい。

IT 目標

今度は、プロジェクトに関連するIT 目標について考えましょう。

(プロジェクトが IT 部門の作業の向上や迅速化といったことを目標にしている場合、業務ユーザーの説得に苦労することが予想されるため、IT 目標は、業務ユーザーがすでに了承している目標と合わせて伝えるべきです)

IT に関連するビッグデータ プロジェクトの目標を、重要度が高い順に記載します(目標の数は、自由に追加・削除可)。

例:顧客データやクレジットカード使用データ、ソーシャル グラフ データ、顧客離れインジケータ等のデータ集計し、リアルタイムに収集、クレンジング、マスター管理し、保管するプロセスを確立する

一度立ち止まり、他部門とコラボレーションし、耳を傾ける

本ワークブックは、読者が業務部門と IT 部門のどちらに属しているのかに関係なく、ビッグデータ プロジェクトを開始できる事を目的としています。どちらの場合も推測に基づいて目標を設定することはできません。何を目標にすべきかについて具体的なガイドが必要な場合は、必要な経験やノウハウを持つパートナーを探し、彼らとコラボレーションしましょう。

プロジェクトの成功には、戦略的なコラボレーションが不可欠です。

IT 目標

各目標を達成するまでの最短期間と最長期間を記載します。

例: 2 ～ 4 カ月

目標毎に、達成度を確認できる基準を記載します。既存の評価基準や計算式が理想的です。

例: X% の正確な顧客離れ予測率

データ要件を定義する

ビッグデータ プロジェクトの目標の概要は説明しましたので、次にプロジェクトの中心となるデータについて説明します。どのようなプロジェクトでも、必要な情報は何か、ニーズに合うデータセットは何か、これらのデータをどのように取得し、どのように使用するかを戦略的に考える必要があります。

必要なデータは何か?

まず、ビッグデータ プロジェクトの最も基本的な目的、すなわち社内に配信すべき情報について考えます次の質問にできる限り具体的に答えて下さい。

これまで概要説明したようなビジネス目標を達成する場合、業務ユーザーが質の高い情報を基盤に意思決定する上で、何を知る必要がありますか?

例: 最も価値の高い顧客が離れる原因および顧客離れに関連する行動

その知識を得るために使用できるデータは?

例: 顧客の購入履歴、評価データ、購入率、離脱率、直帰率、顧客 QOS

こうしたデータセットが格納されているソースシステムは？

例: 顧客サービスのレコード、製品のパフォーマンス評価基準、顧客行動のデータベース、顧客マスターデータ管理

上記のデータの他に、分析に役立つコンテキストや分析の価値を高める情報は？

例: 顧客サービスのアンケートデータ、競合他社の分析、天候データ、ソーシャル データ

現在は利用できなくても、新たなコンテキストデータが含まれている可能性があるデータセットは？

例: サードパーティのソーシャルデータ、サードパーティの市場データ、天候データ

隠れたデータを探す

現在利用できないデータセットを利用する場合、社外データに限定しない事です。Gartner 社によると、ほとんどの企業は社内にあるデータの 15% しか利用していません2。また、データウェアハウスの利用状況を統計分析している Appfluent 社は、データ ウェアハウスのデータの約30 ～ 70% が休眠中であることを明らかにしました。他のデータは、利用コストが高く、アクセスも検索も困難なデータサイロやレガシーシステムのアーカイブ、データストアの中に隠れています。このようなデータの保存に、コストが発生していること自体が大きな問題です。

必要なデータを探す場合は、すでに社内にあるデータを最初に確認する価値があります。

データに関する 5 つの重要な検討事項

探すべきデータの概要が決まれば、ビッグデータの具体的な課題がより明確になります。

特に、先へ進む前に検討すべき5 つの重要な要素があります。ビッグデータのデータセットと同様に、各データセットに何をすべきかということも、これらの要素によって決まるからです。

1. 大量データへの対応

必要なデータの「量」に対応できるように準備しましょう。複数の次元にわたり、データをその価値(顧客取引など)、使用量(アクセスの頻度)、サイズ(ギガバイト、テラバイト)、複雑さ (マシンデータ、リレーショナル データ、ビデオ…)、アクセス権の保有者 (データ サイエンティストのみ、または一般の業務ユーザー全て) に基づいて分類します。

データのインベントリを徹底的に体系化すれば、あらゆるデータの管理方法の決定に役立ちます。現在のストレージと処理能力を評価して、最も費用対効果が高く、スケーラブルで効率的な方法を探します。

2. 多様性への対応

ビッグデータの最も難しい点は、分析できるようにするために、様々なデータ形式や構造を調整しなければならないことです。既存のソース (リレーショナル、レガシー メインフレーム) に新しいデータタイプと構造 (ソーシャル、センサー、ビデオ) を組み込むためには、多数のソースを統合する必要があります。

統合の度に手作業でコーディングしようとするのは非常に手間がかかり、時間とリソースの全てを浪費します。利用できるデータ統合ツールとデータ品質ツールを最大限活用することで処理を高速化し、作業の価値を高めましょう。

3. スピードへの対応

リアルタイム ストリーミングデータと履歴データを組み合わせれば、予測分能力が高まります。つまり、システムに絶えず流入している場合のみ価値があるデータもあるという事です。

事実、ほとんどのリアルタイム分析はストリーミングデータを基にする必要があり、データのソースと形式は大きく異なっています。あらゆるデータを管理できるように準備する際には、ストリーミング分析テクノロジーと論理インフラストラクチャを利用しましょう。

4. 正確性を確保する

分析がどれほど重要であっても、分析の基になるデータを信頼できなければ、価値がありません。分析データが増えれば、高いデータ品質を維持する重要性も高まります。

データと目的を適合させるには、データの使用目的を知る必要があります。例えばデータ サイエンティストが単に集計した顧客データからパターンを見つけようとしているなら最小限の準備で済みますが、 財務諸表やサプライチェーン データなどの場合は、精度とコンプライアンスのために、高度な管理、クレンジング、認証が必要になります。

生データから、クレンジングされた信頼できる正当なデータを格納する高度に管理されたマスターデータストアに至るまで、必要な準備の大きさに応じてカテゴリを定義しておく必要があります。

5. コンプライアンスに配慮する

データセットの取り扱いには、異なるセキュリティの規定と要件が伴います。データセット別に、セキュリティ ポリシーに基づいてデータを匿名化する必要があります。

全社規模の数百のデータ ストアでは、大量のデータが発生します。

機密データの場所を把握し、暗号化することで、データソースで確実に機密データを保護し、アクセスできるユーザーを制御する必要があります。

機密データをセキュアなスマート アーカイブで管理していたとしても、開発環境やテスト環境に移行する場合は、必ずルールを事前定義してデータをマスキングしましょう。

取り扱う全てのデータセットに5 つの検討事項を適用すれば、より現実的な方法でビッグデータの課題に対処する準備ができます。

パート3

リーンなビッグデータサプライチェーン

従来のビジネス インテリジェンスとデータ ウェアハウスの手法では、ビッグデータ イニシアチブのニーズに合わせて拡張することができません。そこで、ここではチーム、プロセス、インフラストラクチャを拡張する方法を説明します。

チーム

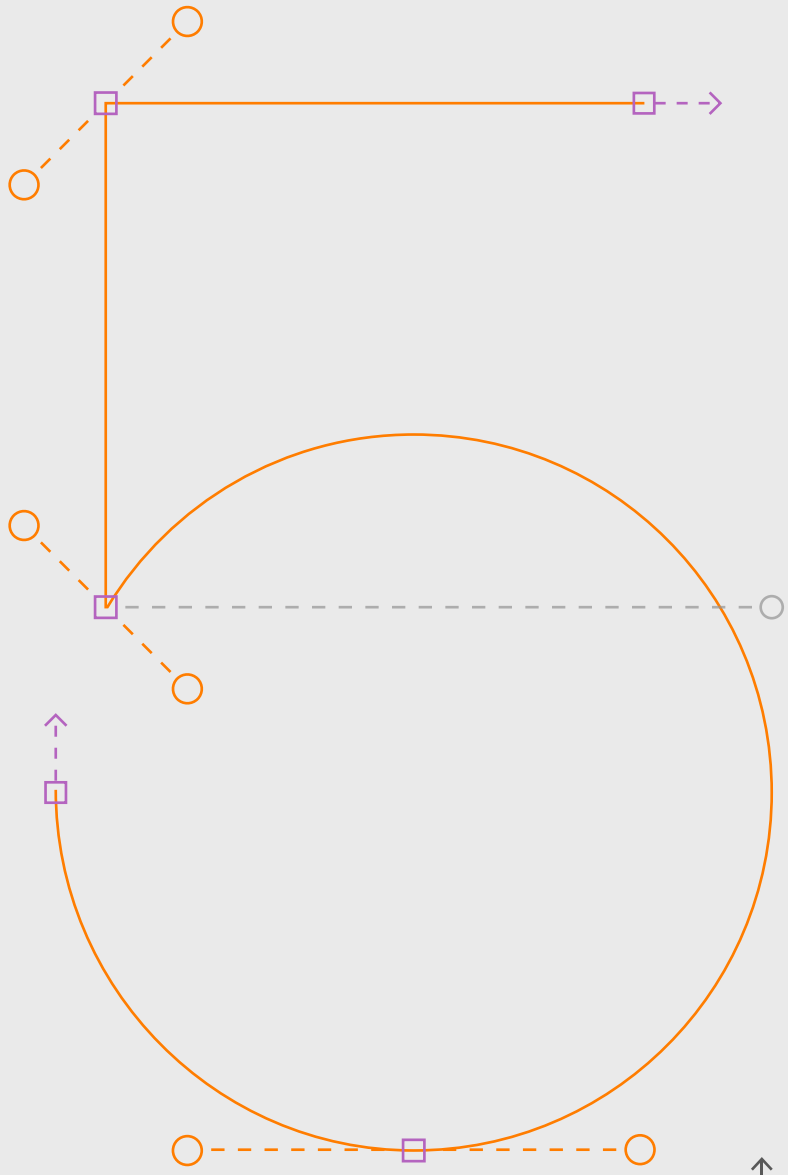
ビッグデータにの前には、最大の課題と最大のチャンスの両方が立ちはだかっています。ビジネス目標を理解している人と、技術要件を実現できる人のバランスを細かく調整する必要があります。

チーム編成における重要な 5 つの教訓

大半の組織は、Hadoopのような新しいテクノロジーの適用を成功させるために必要なスキルレベルを過小評価しています。

分散データ フレームワークは管理が困難です。Hadoop の開発に必要な Java のスキルから新しいデータ サイエンスのスキルに至るまで、プロジェクトを実施するためには多くの新しいスキルを導入する必要があります。[2]

チームの編成を始める場合は、次の教訓を雇用戦略に活かして下さい。



[2] 『Hadoop, Python, and NoSQL lead the pack for big data jobs』、InfoWorld、2014 年 5 月 5 日

1. 雇用目的に合ったスキルを活用する

データ サイエンティストと定量分析アナリストを雇用する場合の最大の過ちの 1 つは、彼らのスキルに見合っていない仕事を割り当ててしまうことです。スキルの高い人材がデータ統合の手作業コーディングやデータのクレンジングに全ての時間を費やせば、彼らが不満を感じるだけではなく、その希少なスキルを活用できません。

得難く、優れたスキルは、そのスキルを必要とするタスクに集中させなければなりません。代わりのツールでこなせる仕事に、最高の人材の時間を浪費させることは避けましょう。

2. チーム構成を戦略的に考える

順調に進めば、プロジェクト範囲も拡大し、リソースも成長します。今こそ、厳しい現実を認識し、戦略的に備えておく必要があります。必要なスキルを有する人材の育成は、シリコンバレーでさえ限界があり、すぐには補充できないプロセスです。

プロジェクト範囲を拡大した場合、どのようなスキルなら合理的にニーズに合った人材をタイムリーに見つけられるでしょうか？例えば、データ サイエンティストを見つけて雇用することは、開発者よりもはるかに困難です。[3]

チームのバランスは非常に重要です。苦労して習得したデータ管理の経験を持つ人と、新しいツールのことを学ぶ意欲を持つ人を適正に混在させるよう努めましょう。さらに、技術的なスキルを持つ人と、適正なモデルの構築に必要な分野の専門知識を持つ人のバランスも取る必要があります。

3. プロジェクトの目標を早期に調整し、チームに伝える

人材の採用時にありがちなミスの 1 つは、プロジェクトの本来の目標を彼らに伝え忘れることです。最初の面接から実際の仕事に着手するまで、業務ユーザーに何を提供しようとしているかを明瞭にする必要があります。ミッションを伝える際は、経営幹部の協力を活用すると共に、問題点と同様に成功事例も共有する必要があります。

プロジェクトのビジネス価値をしっかり把握していない新規採用者の場合、IT 目標だけに囚われるリスクがあります。

[3] 『Big Data’s High-Priests of Algorithms』、Wall Street Journal、2014 年 8 月 8

4. チームが大きくなれば、それだけチームを管理する必要性も高まる

導入、実装、統合を客観的に実現できる新しいテクノロジーとは異なり、新規採用者は、職場、仕事の内容、その理由に慣れる必要があります。新しいチームの管理は、必ず誰かが引き受ける必要があります。

文化や結束などの要素も軽視できません。新規採用者をプロセスに統合する方法を、じっくりと考えましょう。新規採用者にスキルアップのトレーニングは無理かもしれませんが、より良いチームメンバーにする手助けは可能です。

5. チームを立ち止まらせる余裕はない

ビッグデータ テクノロジーは、毎日生まれています。一方で既存のテクノロジーも急速に進化しています。ベスト プラクティスを早期に採用する果断な企業にとっては、素晴らしい時代です。ただし、競合他社よりも先にスタートを切らなければならないというプレッシャーもあります。

チームメンバーは、環境の変化に応じて、スキルを迅速に向上させる必要があります。 時代の先を行くという課題は、優秀な人材にとっては、打ってつけの動機となります。ここでは、チームメンバーとプロジェクト管理者の能力を継続的に高めるために必要なトレーニングとディスカッションを提供することが課題です。

戦略的である事の重要性

自動化ツールを使う能力か、あるいは手作業による統合能力のどちらを確立すべきかということは、何度も繰り返される重要な選択です。

手作業コーディングなら、構築すべき対象を完全かつ正確に制御できます。複雑なスクリプトの記述や、これまで不可能だった方法によるメタデータの抽出などの場合、ハードコーディングの価値が必要とされるケースがよくあります。

一方で、ツールを使用すれば、俊敏性が向上し、同じ処理を継続して繰り返すことができます。データ統合やデータ品質のようなタスクには、ツールが不可欠です。非常にインテリジェントなアナリストやサイエンティストに割に合わない仕事を強制しないためです。

リソースについて現実的に考えましょう。Google 社に比肩し得る大規模で優秀なチームを作ることができない限り、限られたリソースを試行錯誤で浪費する余裕はありません。

データ ガバナンスを設定する

基盤となるようなビッグデータの取り組みを実装する場合 (それが望ましいですが)、データガバナンスの手順に関するフレームワークを導入する必要性が出てきます。実際には、ビッグデータ プロジェクトの目的がたとえある部門にしか価値を提供できなくても、極小規模のデータガバナンス委員会を立ち上げるように検討すべきです。委員会の出席者など、固有の課題への対処方法を確認するためです。

データ ガバナンス委員会は、本質的には、データに対する企業のアプローチを監視するための正式な役員機関です。しかし、これにはデータスチュワードも必要です。データ スチュワードとは、職務または部門固有の人で、特定の部門から受信するデータを管理する責任を負います。

(実際、クライアントの中にはデータ ドメインに基づいてデータ スチュワードの役割を割り当てている企業も存在します。例えば製品データや顧客データなど、データ別に担当者を分けることを意味します)。

データ ガバナンスのフレームワークがデメリットを上回るメリットを確実にもたらすことを意図して、プロセスを作成しましょう。全員が同じ期間に同じ目標の達成に取り組むことを念頭に置いて、形式的な作業に終わることのないように注意して下さい。

データ ガバナンス フレームワークは次の 5 つの特性を目的として構築します。

1. 業務横断型

よく似たロール (役割) を持つメンバーでデータ ガバナンス委員会を構成するのは効果的ではありません。ビッグデータプロジェクトに関わる各部門独自の観点やニーズを代弁できるようなメンバーで構成する必要があります。

2. 円滑なコミュニケーション

異なる業務、部門、分野間のコミュニケーションが不足すると、プロジェクトが形式だけの作業に陥ったり、誤解を生んだりすることになりかねません。このような事態は容易に起こり得ます。あらゆる懸念事項を緩和するか、適正に解決しましょう。

3. 効率的である事

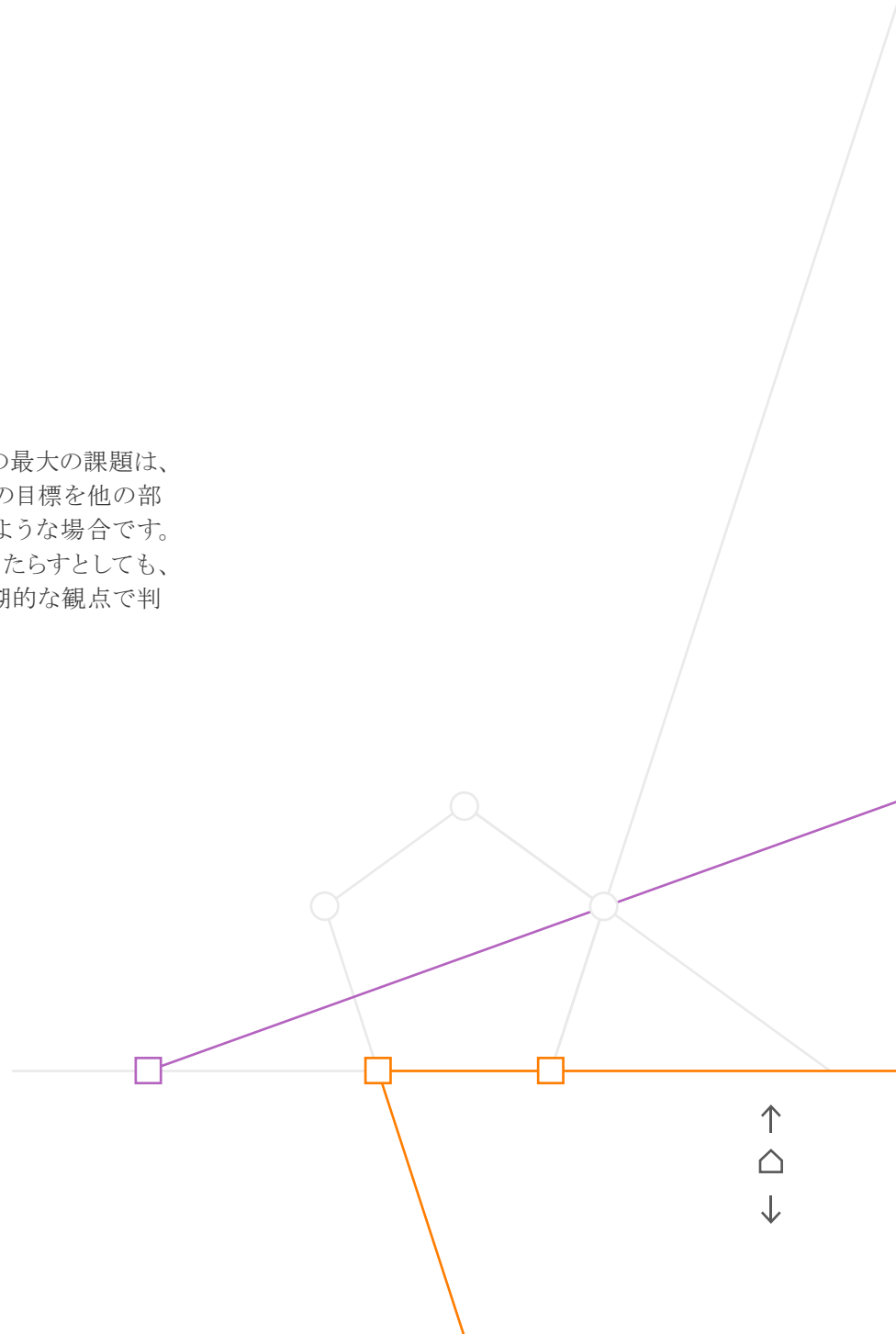
業務横断型のプロセスには壁があるように思うようなら問題です。ビッグデータ プロジェクトの成功には高い俊敏性が必要とされます。そのためには、可能な限り全ての領域に自動化ルールや例外レポート作成ルールを組み込み、コラボレーションツールを導入することで、常に迅速にコミュニケーションが取れるようにしましょう。

4. コミットする事

プロジェクトの主要目標を効果的に伝達し、データガバナンスのフレームワークに関わる全員がその目標の達成に全力を尽くせるようにします。共通の目標を持つことは、適正なガバナンス思考と意思決定を助けます。

5. 一元化する事

データ ガバナンス フレームワークの最大の課題は、委員会の決議において、ある部門の目標を他の部門よりも優先しなければならないような場合です。短期的には 1 つの部門に利点をもたらすとしても、委員会全体の利点となるような長期的な観点で判断しましょう。



必要なスキルと既存のスキル

ここでまたペンを用意して下さい。新しいチームに起こり得る主観的な落とし穴と機会を説明しましたが、次は実際にどのようなチームが必要なのかを説明します。

弊社のお客様が新たに雇用した職種を参考にして、ビッグデータ プロジェクトの役割一覧を次のページに示します。すでに確保している人数とプロジェクトに要する想定期間に基づいて (23 ページから始まるセクションで記入済み)、雇用する必要がある人数をリストに記入します。

役割	この役割を果たせる経験者はいるか?	この役割のために雇用が必要	期間を想定すると X 人が必要
データ サイエンティスト			
ドメイン エキスパート			
ビジネス アナリスト			
データ アナリスト			
データ エンジニア			
データベース管理者			
エンタープライズ アーキテクト			
ビジネス ソリューションアーキテクト			
データ アーキテクト			
データ スチュワード			
ETL (データ統合) 開発者			
アプリケーション開発者			
ダッシュボード開発者			
統計モデラー			
その他			
その他			
その他			
その他			
その他			

統合的な思考の必要性

新たなチーム メンバーを探す場合、目的の資質を備えた人材のみに限定しない事です。適正な資質を備えた人材を見つけることは、かなりの難題です。また、ビジネス目標に合った技術的能力を発揮すること(目標と技能の統合) に意欲的な人材を見つける必要もあります。

弊社のお客様から何度も聞いたのは、ビッグデータ プロジェクトには、ビジネスの現状を理解し、複雑なデータ サイエンスを実践できる人材の参加が重要だという事です。このように統合的な思考を出来る人材は、非常に貴重な存在であり、容易には見つかりません。このような人材には、トレーニングや報奨の機会を与える価値があります。

ビッグデータのツールについて

弊社の経験では、ビッグデータプロジェクトで必要とされるアーキテクチャには次のツールが不可欠です (アーキテクチャについては詳細を後述します)。当然ながら、特定のプロジェクトに必要なテクノロジーの組み合わせは貴社の目標やリソースによって決まります。

分析

生データを変換して、分析対象となるドメインに関する洞察やパターン、予測、計算結果を導き出すツールやプロセスです。

可視化

データや調査結果を簡単に呑み込んで理解できるように提示できますか?

高度な分析

最新の分析アルゴリズムをデータセットに適用して、複雑な計算を実行できますか?

マシン ラーニング

高度なマシン ラーニングアルゴリズムを使って、手作業では対応できないレベルのパターン認識と予測を実行できますか?

必要なツールやテクノロジーのうち、データ統合、データ品質、マスターデータ管理のような一部のツールは、ビッグデータ プロジェクトにおいて基本中の基本なので再構築の必要はありません。これらの機能の構築に要する手間と時間に、ビッグデータ プロジェクトの貴重なスキルと労力を投入する価値はありません。

プロジェクトの目標は、全てをカスタム構築することではないことを思い出して下さい。

プロセス

次に、ビッグデータ プロジェクトで実際に取り組む必要のあるプロセスについて詳しく見ていきましょう。具体的なプロセスは目標や要件によって異なりますが、ここではどのような状況が想定され、そこから何を学ぶことができるのかという点を簡単に説明します。

ビッグデータの 8 つのステップ

弊社の経験から、ビッグデータプロジェクトにとって、アジャイル アプローチは優れた手法であるといえます。この手法により、期待値をコントロールし、失敗から学び、作業の反復によって最善のプロセスを探り出すことが可能になります。しかし、プロジェクトへのアプローチは状況によってまったく異なります。

いずれにしても、ビッグデータサプライチェーンには次の 8 つのステップが重要です。どのような方法で取り組む場合でも、これらのステップを実現する効果的なプロセスをチームで確立する必要があります。

1. データへのアクセス

最初の課題は、必要な全てのデータを取得することです。状況によって、ストリーミング データのキャプチャが必要なこともあれば、データベースからの抽出が必要な場合もあります。取得したデータが使用方法に応じて保存されるように、繰り返し実行可能で適正に管理できるプロセスを確立して下さい。

2. データの統合

ビッグデータは、多様なデータ構造とデータ形式が絡む複雑な課題です。持続的な分析を可能にするには、全てのデータの統合と正規化を行うプロセスを確立する必要があります。その際、手作業によるプロセスをできる限り少なくすることが理想的です。

3. データ クレンジング

信頼できる分析結果を得るには、重複データ、エラーのあるデータ、不正確なデータ、不完全なデータを除外しなくてはなりません。有能なアナリストやサイエンティストが全ての時間をクレンジング作業に費やすことのないようにプロセスを用意する必要があります。

4. データのマスター管理

統合されたクリーン データの信頼できるソースを維持する方法の 1 つとして、データのマスター管理プロセスの確立があります。狙いは、統合した豊富なデータのコレクションを作成し、ドメイン (製品、顧客など) 別に整理して、ビッグデータから得た洞察を付加した上で他のシステムに供給することです。

5. データの保護

ここでは 2 つの基本プロセスを確立します。1 つは、各データセットで必要とされるセキュリティのルールと実践方法の定義です。2 つ目は、ルールとベスト プラクティスを一貫して適用するために、永続的な方法または動的な方法で機密データを検出してマスキングすることです。

6. データの分析

分析プロセスはアナリスト、分析ツール、また目標に応じた要件によって異なります。このプロセスを改善し、迅速性とコスト効率を高め、時間と経験に応じて拡張していくことを考えた場合、発見を繰り返し、改善を継続するという心構えが大切な役割を果たします。

7. ビジネスニーズの分析

このステップは非常に重要である一方で見落とされがちなステップです データを分析する際にも、ビジネス ニーズを分析する明確なプロセスを確立する必要があります。ビジネスの動向を把握していないと、プロジェクトの努力がビジネスに結び付かず、ビジネスへの効果が最小化してしまうというリスクがあるため、このプロセスは重要です。

8. 洞察を業務に活かす

前半で触れたように、ビッグデータ プロジェクトがビジネスにもたらす効果は実感されなくてはなりません。答えを見つけたら、それを最も必要としている業務ユーザーに提供できるような自動パイプラインを作成します。例えば顧客離れ (チャーン) の恐れが高い顧客のデータであれば、ダッシュボードを通じてカスタマー サービスの担当者に知らせる必要があります。フィードバックのループを組み込むことも忘れないで下さい。洞察がどのように受け取られたかを知ることができます。

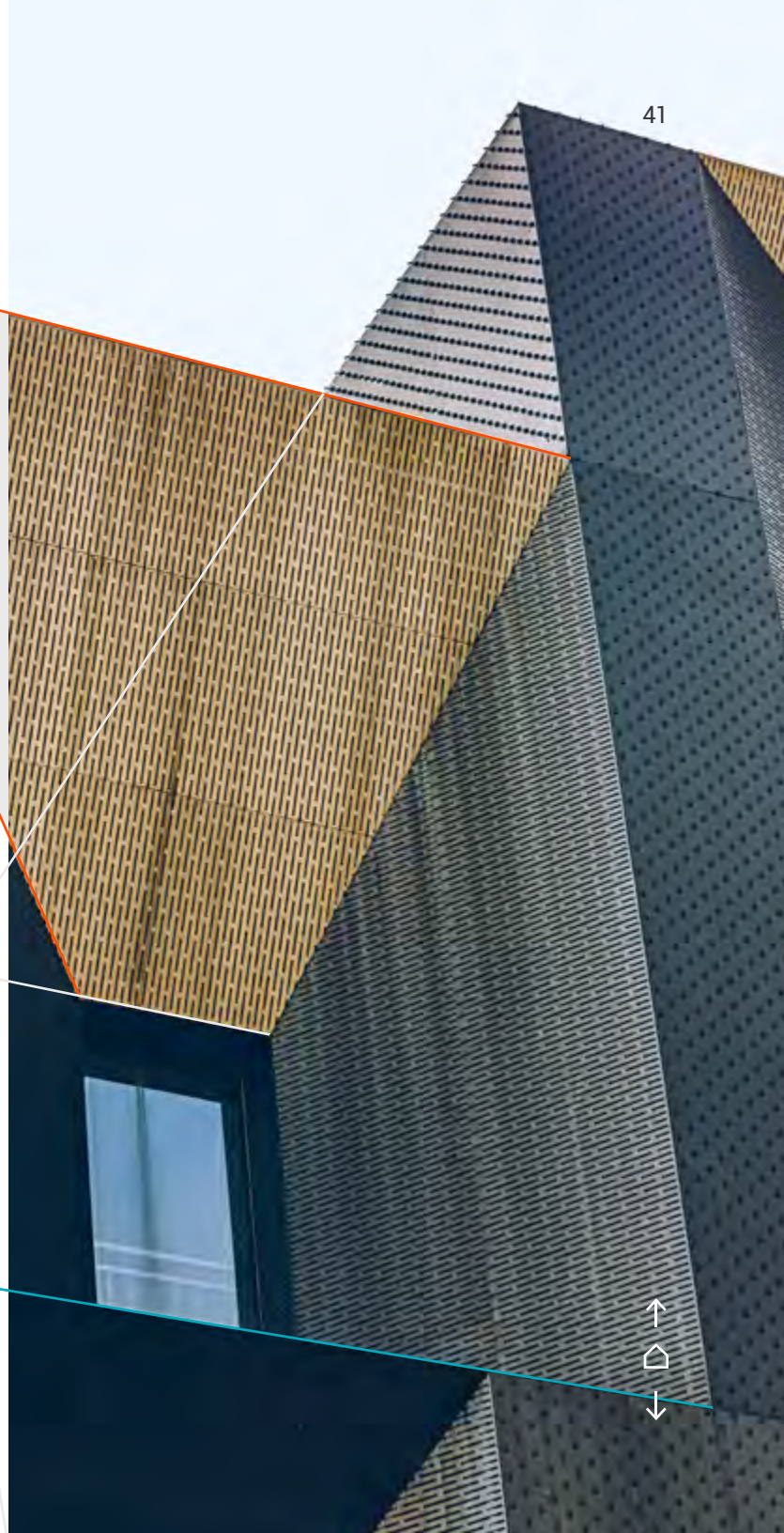
ドキュメントの重要性

これら 8 つのステップをマスターし、ビッグデータ プロジェクトを正しい方向に進めることを目指して下さい。この狙いは、反復可能で拡張性に優れた、継続的に改善していくプロセスを確立することです。そのためには、これらのプロセスを文書化し、改善を続けることが不可欠です。

ビッグデータ プロジェクトのスキル、能力、蓄積した知見は、組織内で共有化し、頻繁にやり取りする必要があります。

アーキテクチャ

ビッグデータのサプライチェーンの無駄をなくして効率化するためには、信頼性が高いアーキテクチャを戦略的に構築する必要があります。このセクションでは、理想的なビッグデータアーキテクチャの概要と、導入の段階的アプローチを説明します。



最初のステップ: サンドボックス(テスト環境)

ビッグデータのアーキテクチャの構築を始めるとき、最も論理的な開始点となるのは、テストデータを使ってアーキテクチャの実用性を確認するためのサンドボックス開発環境をセットアップすることです。 そのためには、次の教訓を検討して下さい。

小さく始める

完全に制御可能で、明確に定義されたサンドボックス環境から着手すれば、作業を反復することで最適な実装方法を探り出すことができます。 可能な限り速やかに稼働させて、実装の反復によって得た知見を文書化する必要があります。

規模の問題

サンドボックス環境と実稼働環境との主な違いは、運用環境のほうがはるかに大規模であるという点です。出力データの取り込み、統合、クレンジング、配信プロセスを自動化する必要があります。このような理由から、実際の運用環境には、はるかに堅牢な構造で、実績あるコンポーネントとプロセスを組み込んだ、信頼性と柔軟性の高いアーキテクチャを構築することになります。

テスト前のマスキング

通常は、運用環境のデータ形式や構造を確認できるように、実際の運用データを加工したものがテストデータとして使用されます。このデータのマスキングを適正に行わないと、安全が確保されていないテスト環境で機密データがリスクに曝されます。

運用環境で迷わないために

ビッグデータ プロジェクトが長期化したり遅延コストが発生したりする最も一般的な原因の 1 つは、サンドボックス環境で見逃した手作業でのコーディングエラーが、アーキテクチャの稼働後に表面化した場合です。従ってアーキテクチャの大半を手作業でコーディングする場合は、運用環境レベルの要件を満たすように、多くのコードをリファクタリングし、それに見合う期待値に設定することです。または、コードエラーのリファクタリングが必要ない生産性ツールや自動化ツールを使用しましょう。

ビッグデータの理想的なアーキテクチャ

次の図は、ビッグデータのテクノロジーとプロセスの理想的なアーキテクチャを構築するための推奨方法を表しています。

データソース

- リレーショナルデータベース
- メインフレーム
- 文書、電子メール
- ソーシャル メディア、第三者データ、ログファイル
- マシン センサー
- パブリック クラウド
- プライベート クラウド

データの取り込み

バッチロード

変更データキャプチャ

データストリーミング

アーカイブ

データ管理

- データ統合
- データ品質
- 仮想データ マシン
- データ セキュリティ
- マスターデータ管理
- 拡張性の高いストレージ (Hadoop など)
- データ ウェアハウス

データ配布

バッチロード

データ統合ハブ

データ仮想化

リアルタイムのイベントベース処理

アプリケーション

- 可視化
- モバイル アプリ
- 分析
- ビジネスインテリジェンス
- リアルタイムダッシュボード

プロジェクト計画

ビッグデータ プロジェクトを様々な側面から分析しました。次のステップでは、本書のプロジェクト計画をひな形として使用して、ビッグデータ プロジェクトの開始から実装までを管理します。

プロジェクト計画

このプロジェクト計画テンプレートを枠組みとして使用して、ビッグデータ プロジェクトに関わる詳細な情報や様々な要素をドキュメントに記録します。作成したドキュメントは、組織内の他のメンバーの支持を得るために使用します。また、外部パートナーと交渉する際にも役立ちます。

段階 1: 戦略	段階 2: データ
業務目標と IT 目標を策定する	必要な情報を特定する
成功の評価基準を定義する	データとそのソースを特定する

段階 3: サプライチェーン

人材

- 必要なスキルの評価
- 既存スキルの評価

プロセス

- データへのアクセス
- データの統合
- データのクレンジング
- マスターデータ
- データの保護
- データの分析
- ビジネスニーズの分析

ツール

- 分散コンピューティング (Hadoop など)
- データ品質
- データ統合
- マスターデータ管理
- データ マスキング
- 可視化
- ストリーミング分析
- 分析
- マシン ラーニング

段階 4: 洞察を業務に活かす

ダッシュボードの開発

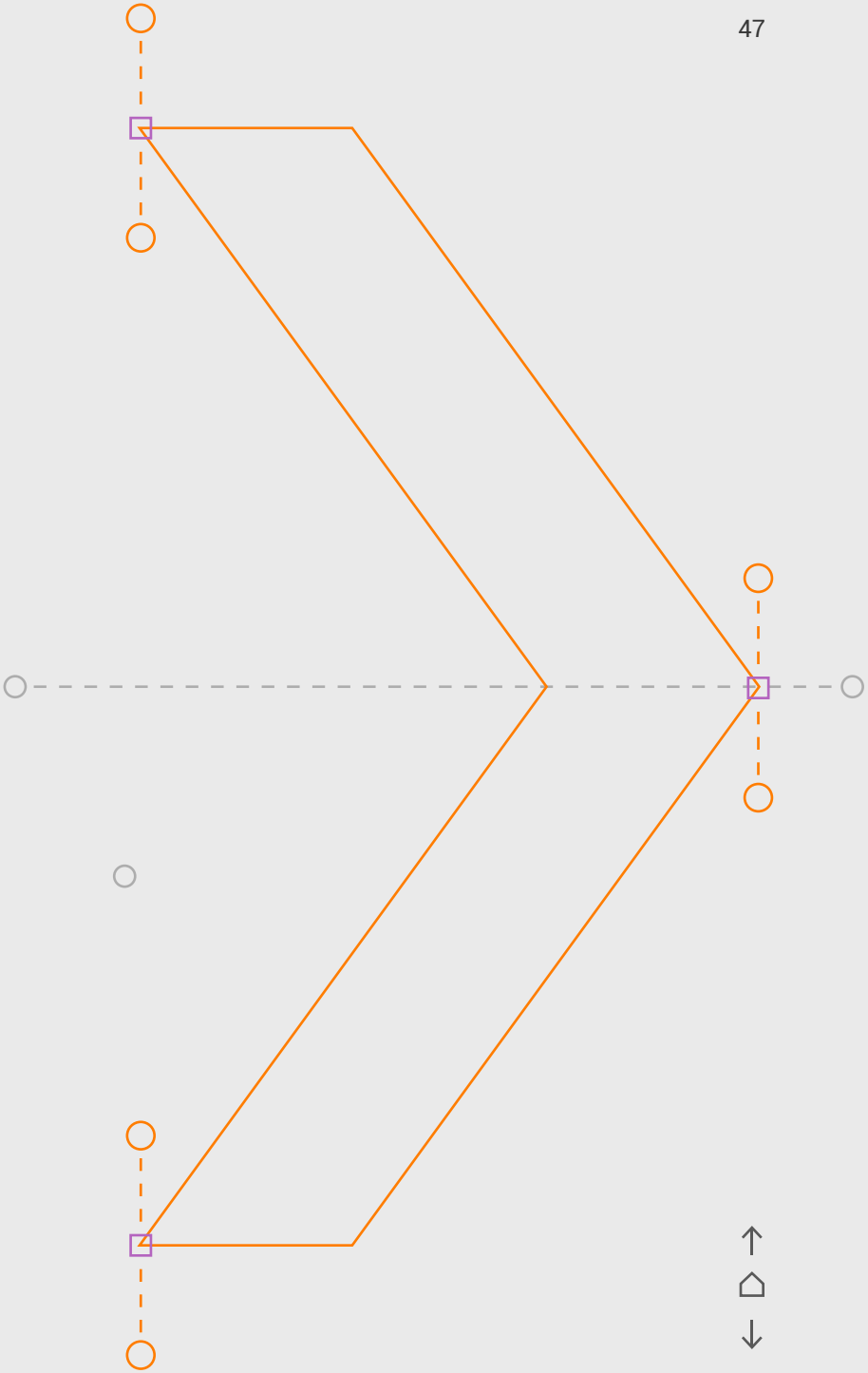
データ配信プロセスの自動化

フィードバック プロセスの確立

始める

本ワークブックではチェックリスト、原則、ガイドラインを紹介して、ビッグデータが貴社にもたらす可能性を簡単に説明しました。プロジェクトの規模に関わらず、ビッグデータプロジェクトで避けられない多くの難題に備えるために本書が役立てば幸いです。

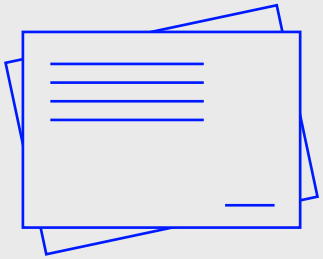
これは、大きな変革となります。



次のステップ

学んだことを実践する準備はできましたか？

インフォマティカの開発者は、Hadoop 開発者としても活躍できます。また、弊社のビッグデータ ソフトウェアの体験版や、コネクターサービスは、貴社の適正な導入パスをサポートします。



ダウンロードはこちら

インフォマティカ
について

デジタルトランスフォーメーションによって我々の期待値が変化しています。より良いサービスを、素早く、便利に、低コストで利用したいという期待が高まっているのです。企業も状況に応じて変化する必要があります。そしてそのヒントは「データ」にあります。

エンタープライズ向けクラウドデータ管理で世界をリードするインフォマティカは、俊敏性の向上、新たな成長機会の獲得、新しいソリューションの開発を実現するための洞察を通じて、あらゆる産業や分野の企業がインテリジェントにビジネスをリードできるよう支援します。インフォマティカは、あらゆるデータを徹底的に重視し、企業の成功に必要とされる汎用性を提供します。

インフォマティカは、企業がこれからのインテリジェントな破壊的イノベーションを推進できるよう、当社が提供するあらゆるサービスを通じてデータの力を継続的に引き出すことを支援します。

〒105-6226
東京都港区愛宕2-5-1 愛宕グリーンヒルズMORIタワー26階
電話：03-6403-7600(代表)
FAX：03-3433-1021

informatica.com/jp
linkedin.com/company/informatica
twitter.com/Informatica

お問い合わせはこちら

IN18-0917-2730

©2017 Informatica LLC. All rights reserved. Informatica®およびPut potential to work™は、米国およびその他の国におけるインフォマティカの商標または登録商標です。その他全ての企業名および製品名は、各社が所有する商号または商標です。

Informatica™