

DeNA

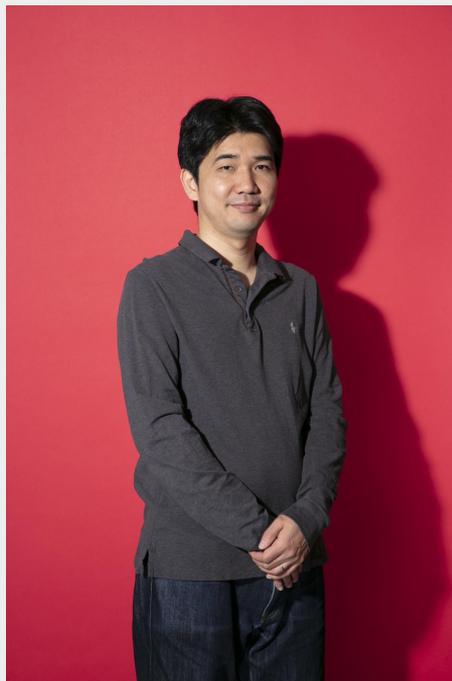
Pocochaにおけるデータマネジメント

長谷川了示

データ本部データ基盤部データエンジニアリング第一グループ

株式会社ディー・エヌ・エー

自己紹介



長谷川 了示

DeNA データ本部データ基盤部

データエンジニアリング第一グループ

グループリーダー

- 新卒 -> コンサルティングファーム
- 製品開発@分析系SaaS ベンダ
- 2016年9月 DeNA 入社
 - 全社データプラットフォームの設計・構築・運用
 - 直近は pococha を含むライブストリーミング事業のデータ基盤を担当

目次

- 1 | pococha のご紹介とデータエンジニアの関わり方
- 2 | データエンジニア参画以前に生じていた課題
- 3 | データマネジメントの取り組み
- 4 | 今後の展望

pococha のご紹介と データエンジニアの関わり方

Pococha について

一般の人も多く参加するライブコミュニケーションアプリ
後発ながら業界トップクラスに成長

07. 事業について Part.1- What We Do

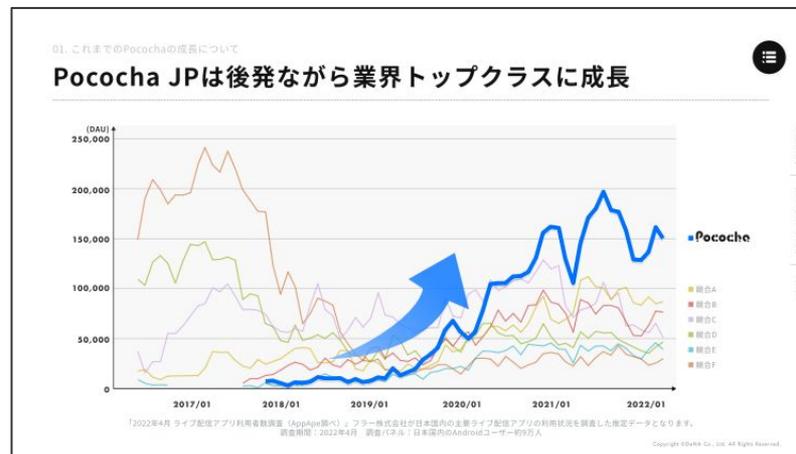
Pococha とは

Pocochaはスマホひとついつでも・だれでも・どこでもライブ配信を通して誰もが誰かの特別な存在になれる。そんな自分に気付けるアプリです。

No.1 美しく輝ける
No.1 一般人が活躍している
No.1 リスナーが楽しい

出典: <https://dena.com/jp/press/4723/>

Copyright © DENA Co., Ltd. All Rights Reserved.



<https://speakerdeck.com/daisukemzt/pococha-culturedeck>

DeNA におけるデータエンジニアと事業の役割分担パターン

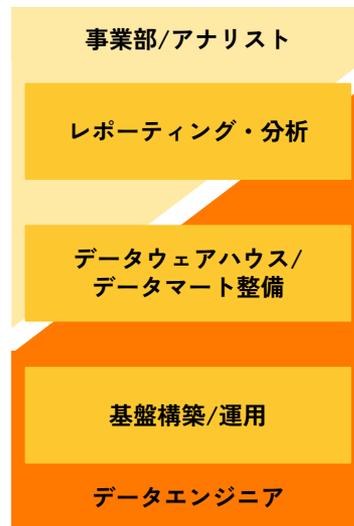
多くの事業を展開しており、事業の状況に応じて役割分担を変えている

新興事業



- データエンジニアは分析基盤のみ提供し、事業部側でデータ/レポート整備
- データの規模/複雑度がまだ小さくなく、事業部側でなんとかできる

主要事業



- データが大規模化、複雑化し、使いやすい形に整える作業が重要
- データエンジニアは分析基盤に加え、データ/レポート整備もアナリストと協働して推進する

Pococha は新興事業として始まった

新興事業

事業部/アナリスト

レポート・分析

データウェアハウス/
データマート整備

基盤構築/運用

データエンジニア

- データエンジニアは分析基盤のみ提供し、事業部側でデータ/レポート整備
- データの規模/複雑度がまだ小さくなく、事業部側でなんとかできる

主要事業

事業部/アナリスト

レポート・分析

データウェアハウス/
データマート整備

基盤構築/運用

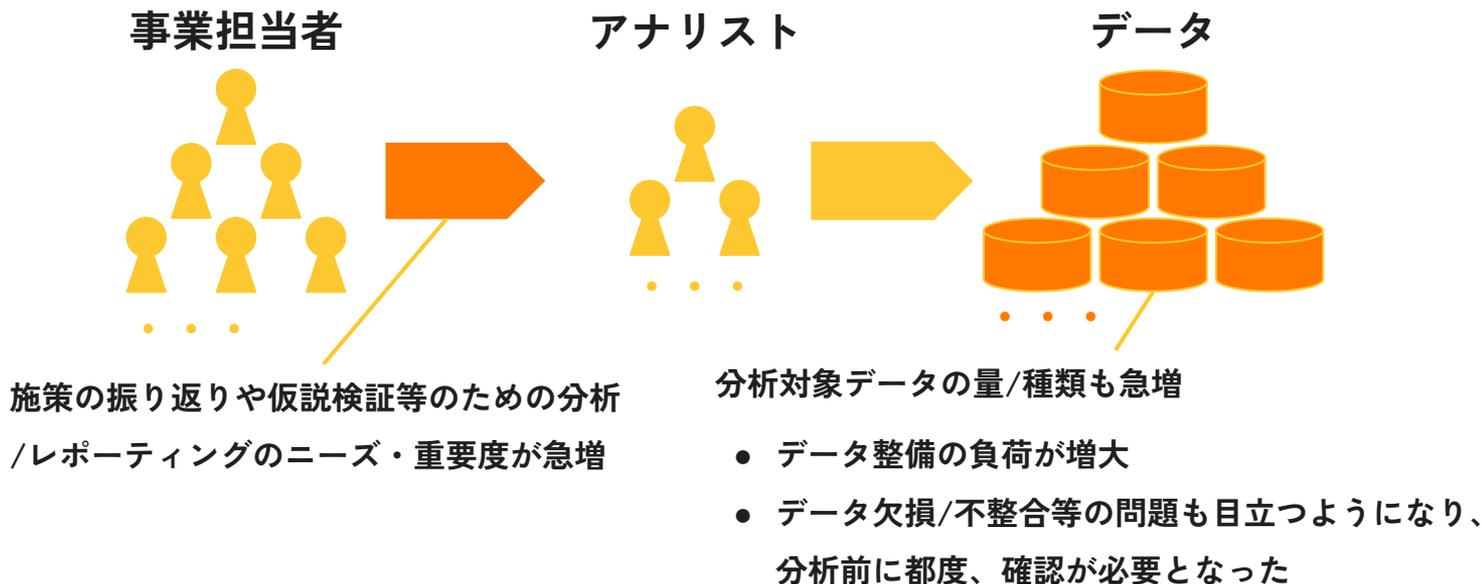
データエンジニア

- データが大規模化、複雑化し、使いやすい形に整える作業が増大
- データエンジニアは分析基盤に加え、データ/レポート整備もアナリストと協働して推進する

データエンジニア参画以前に 生じていた課題

pococha のデータ基盤で生じた課題

事業が急成長するのに伴い、アナリストの業務負荷が増大し、事業担当者の間に対して迅速に回答を出すことが困難になっていった



データエンジニアをアサインし、主要事業パターンへ移行

新興事業



ライブストリーミング事業担当の
データエンジニアがアサインされ
主要事業パターンへ移行を開始



主要事業



データエンジニアが行った 取り組み

大きく3つの領域で改善の取り組みを推進中

1. 体制強化

1. 集計バッチ運用の巻取り
2. データウェアハウス/データマート構築における協働作業
3. アナリストの daily scrum 等へ参加し、常に状況を共有

2. 自動化推進

1. ELT のシステム構成見直し/統合

本日のスコープ

3. データマネジメント推進

1. データ/レポートの設計標準策定
2. データ品質チェック
3. 利用状況のモニタ

データマネジメント推進の背景

データ品質に関する問題が顕在化していた。

- プロダクト側のデータモデル変更に起因する問題がしばしば生じていた
 - バージョンアップと共に新しいテーブルに移行される、など...
 - 分析への影響を事前に把握しきれず、問題が起きてから気づくというケースが発生
 - 関連する集計テーブル、レポートを洗い出し、修正する手間も小さくない
- アナリストがデータウェアハウスやデータマートを構築していたが、品質担保が十分でなかった

取り組み内容

1. データ/レポートの設計標準策定
2. データ品質チェック
3. 利用状況のモニタ

1 | データ/レポートの設計標準策定: データとレポートの分類

データ基盤(BigQuery)のデータを3階層に分類

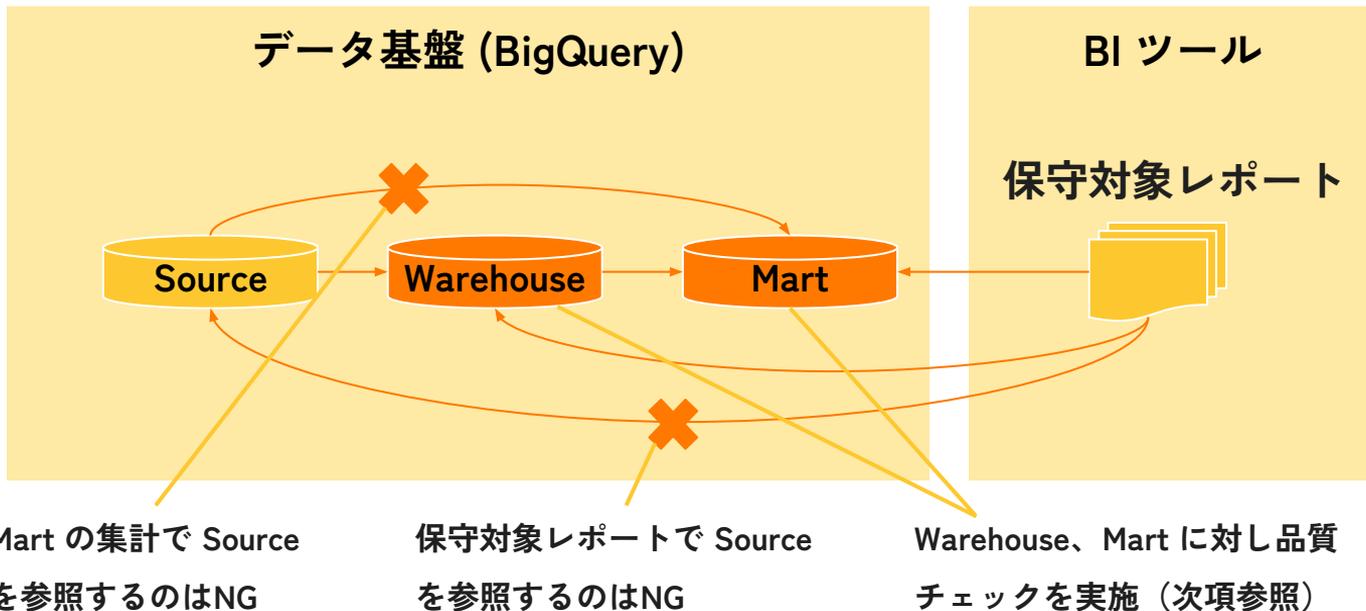
Source	プロダクトのDBテーブルのコピーやログ等の生データ
Warehouse	Source を分析で使いやすい形式に一次加工したデータ
Mart	Warehouse を更に使い勝手のよい形に加工したデータ

BIツール上のレポートを、保守対象とそうでないものに区分

保守対象 レポート	分析チームが品質に責任を持つレポート 定常的に多くのメンバが確認するものを対象とする
保守対象外 レポート	アドホックな用途等で作成されたレポート等 日々、様々なメンバにより多数作成されている

1 | データ/レポートの設計標準策定: 依存関係のルール化

- Source へ依存する箇所をなるべく Warehouse に限定する
- Warehouse、Mart 層でデータ品質を担保する



2 | データ品質チェック: チェック観点

Warehouse、Mart について、「期待される制約条件を満たしているか？」

「テーブル間で整合しているか？」といった観点でチェック

制約条件のチェック

- Primary Key は想定通りになっているか？
- 全行NULLになっているカラムはないか？

等

ID	amount
1	NULL
2	NULL
3	NULL
3	NULL
...	...

テーブル間の整合性のチェック

例: 会計用に pococha 側で作成しているテーブルと warehouse で売上の合計が一致していることを確認

pococha システム

Date	Transaction ID	amount
2022/xx/yy	001	¥ 100
2022/xx/yy	002	¥ 100
2022/xx/yy	003	¥ 100
2022/xx/yy	004	¥ 100
...

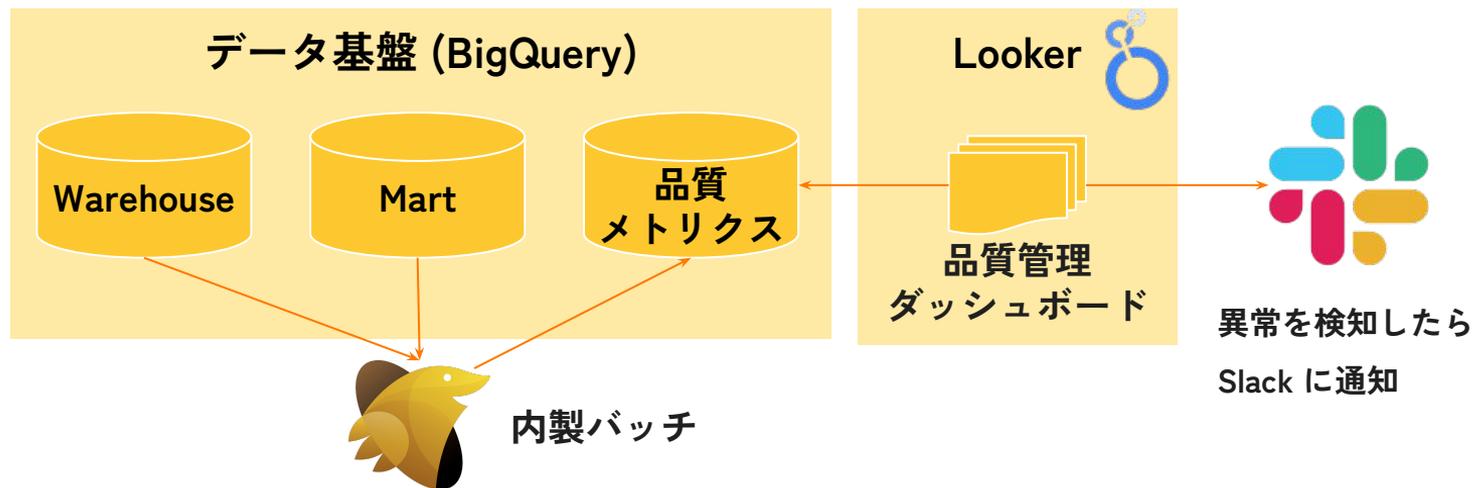
BigQuery

Date	User ID	...	amount
2022/xx/yy	a		¥ 100
2022/xx/yy	b		¥ 100
2022/xx/yy	c		¥ 100
2022/xx/yy	d		¥ 100
...

2 | データ品質チェック: 仕組み

日次でシステム監視

- 内製のバッチでメトリクスを収集し BigQuery に保存した上で Looker で可視化
- 異常を検知したら Looker の通知機能で slack 通知

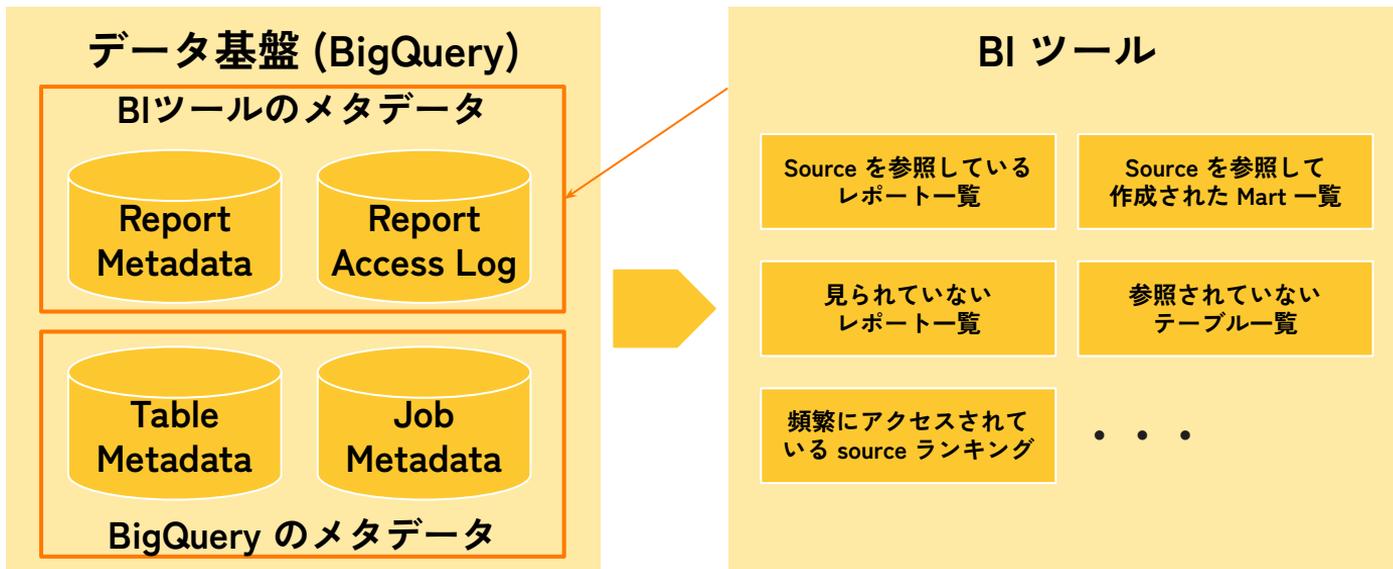


3 | 利用状況のモニタ: モニタ観点

- 設計標準に準拠しているか？
 - ルールを決めるだけでは守られない
- どれくらい頻繁に利用されているか？
 - 使われていないテーブル/レポートは廃止し、コストを削減
 - 頻繁に使われている生データがあれば、warehouse/mart 化を検討する

3 | 利用状況のモニタ: 仕組み

BigQuery、BIツールのメタデータを収集/活用し、可視化している



取り組みの成果

アナリストがその日の業務を開始する時点で、品質の担保された最新のデータが揃っている状態を実現できた

「サービスから出てくる膨大なデータをテーブルに流し込む過程で、バグが発生することがあり、慎重に検証しながら作業していましたが、その時間が、0分になったんです。

朝、分析に取り掛かろうとする段階でちゃんと必要なものが揃っている状態になりました。」

<https://dena.ai/story/pococha-data-management>

今後の展望

より高度なデータマネジメントへ

1. データ品質チェックの仕組みの改善
2. クエリの複雑度のモニタリング

1 | データ品質チェックの仕組み改善

現状

- メトリクスを収集する部分はフル内製
- テーブル/カラム毎に全てのチェック項目を設定する必要あり
- メトリクス収集(内製バッチ)とアラート(Looker)をそれぞれ設定している

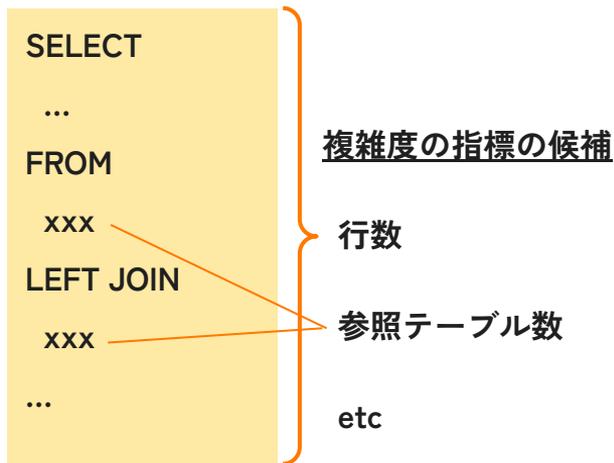


今後の展望

- ツールいろいろ出てきたので内製でなくてもよさそう
 - Dataplex data quality、Dataform、dbtあたりを検証・検討中
- NULLチェック等、全てのテーブル/カラムで実行したい項目は、設定せずとも自動でチェックが回るようにしたい
- メトリクス収集とアラートを一元管理できるようにしたい

2 | クエリの複雑度のモニタリング

データの参照のされ方をより詳細に分析することで、より使いやすい warehouse /mart 設計に活かせるのでは？



頻繁に類似した複雑なクエリを発行している場合、warehouse/mart のリファクタリングや新規追加をすることで改善できるのでは？

まとめ

1. Pococha では事業の急成長に伴い、アナリストの業務負荷が増大し、事業担当者の問に対して迅速に回答を出すことが困難になっていった
2. 課題を解決するために、データエンジニアがより深く支援する体制に移行した
3. その中から今回はデータマネジメントの推進をご紹介した
4. 「データ/レポートの設計標準策定」「データ品質チェック」「利用状況のモニタ」といった取り組みを通じ、アナリストが仕事のしやすい状態を実現できた
5. 今後もより高度なデータマネジメントを目指し、活動していく