オープンソースカンファレンス 2015 Tokyo/Spring

2日目：ライトニングトーク

「ビッグデータ分析基盤を支えるOSSたち」

髙橋達

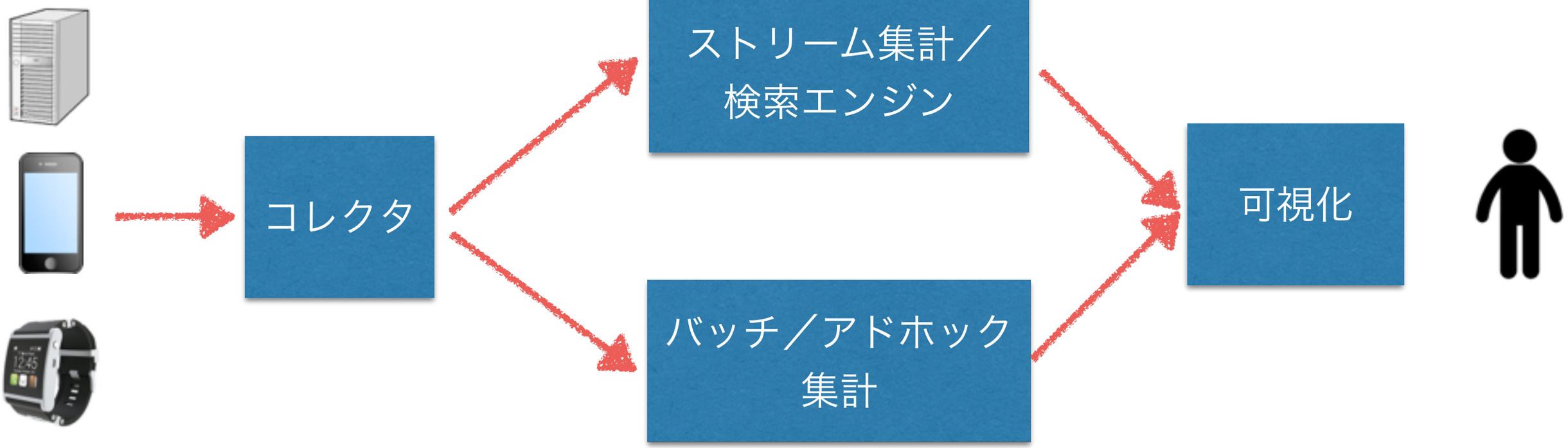
サイオステクノロジー株式会社

2015/02/28

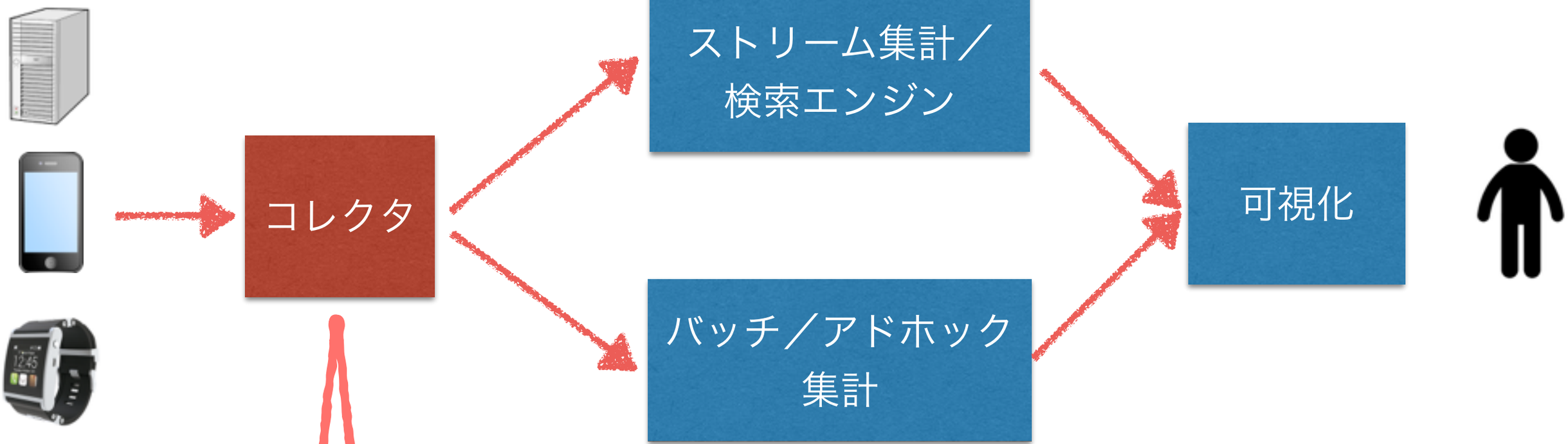
ビッグデータ分析基盤とは？

- データを収集・保存と分析・可視化するためのシステム
- なぜ分析基盤が必要？
  - “ビッグデータ”という流行と共に、データに対して多くの人が興味を持ちはじめ、データを元に意思決定をするようになりつつあるから
- なぜOSSが必要？
  - データを分析したからといって、儲かるわけではない
    - 高価な分析システムを利益もわからず購入するのが難しい
  - 目的に応じて色んな組み合わせをする必要がある
    - カスタマイズもOSSだからこそ可能

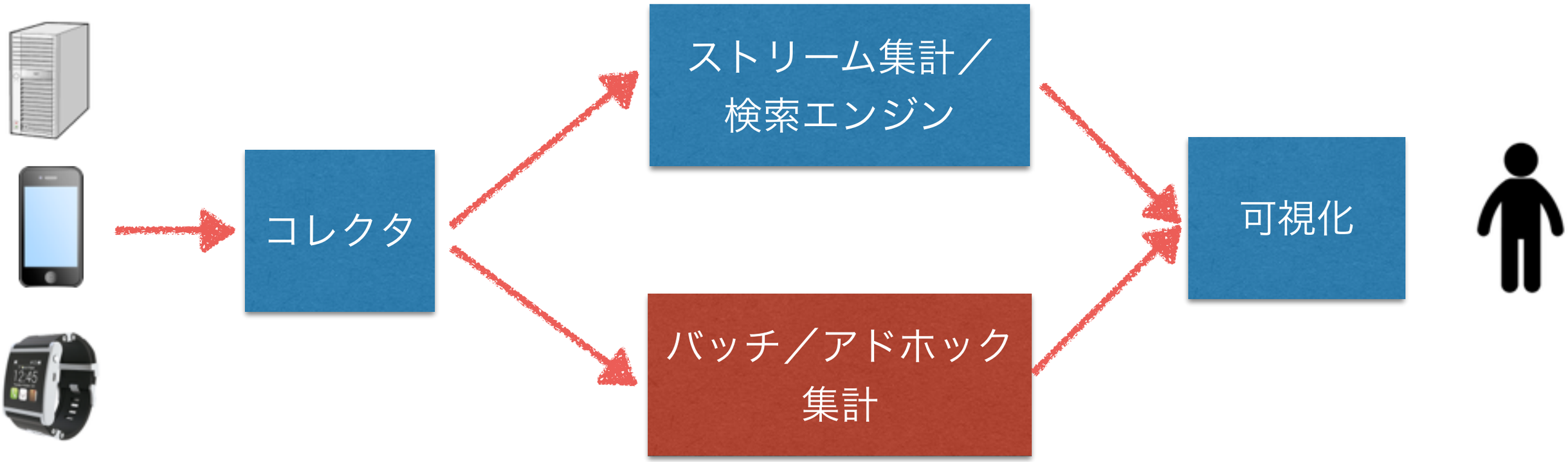
ビッグデータ分析基盤の基本構成（1/6）



ビッグデータ分析基盤の基本構成（2/6）



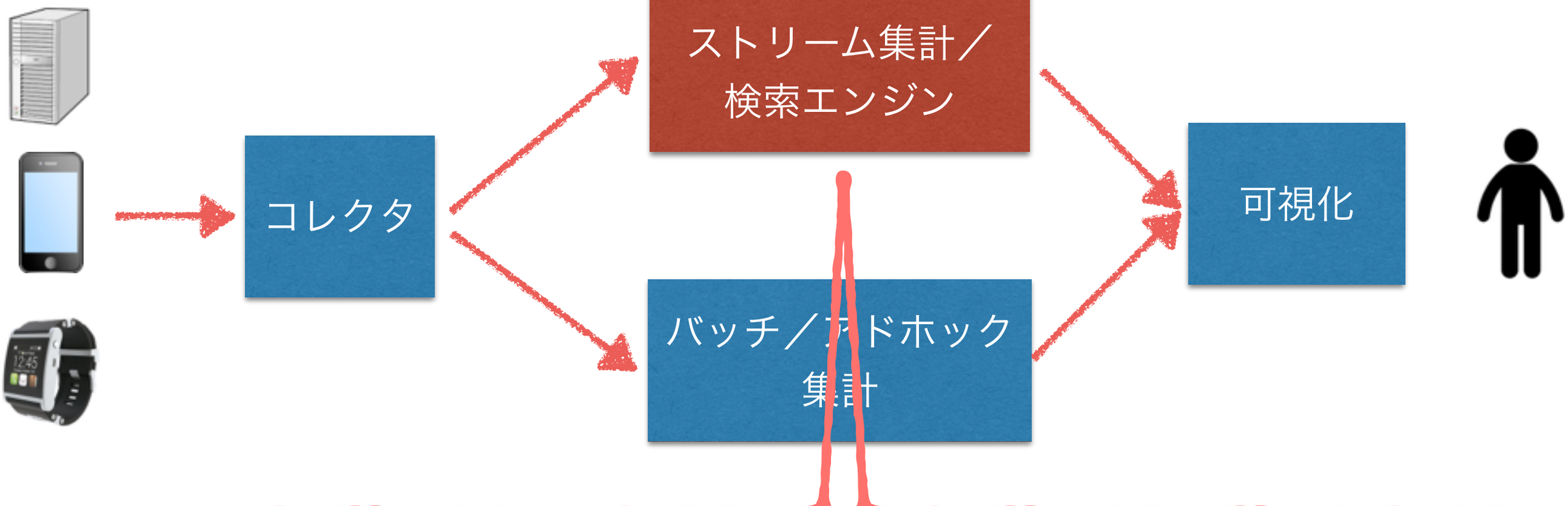
ビッグデータ分析基盤の基本構成（3/6）



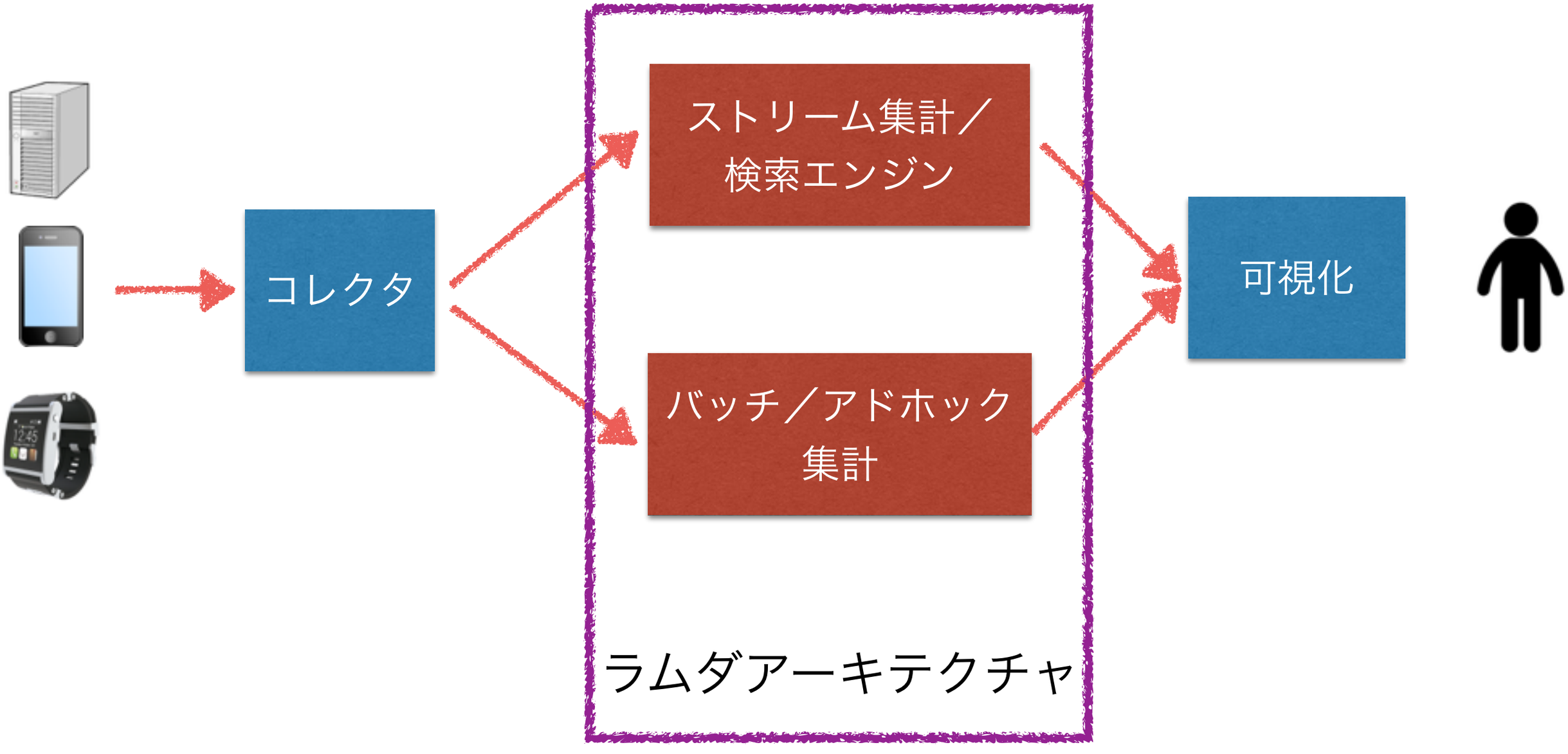
何十億何百億件というデータを一気に計算して
集計をするためのソフトウェア

Ex. Hadoop, Presto, …

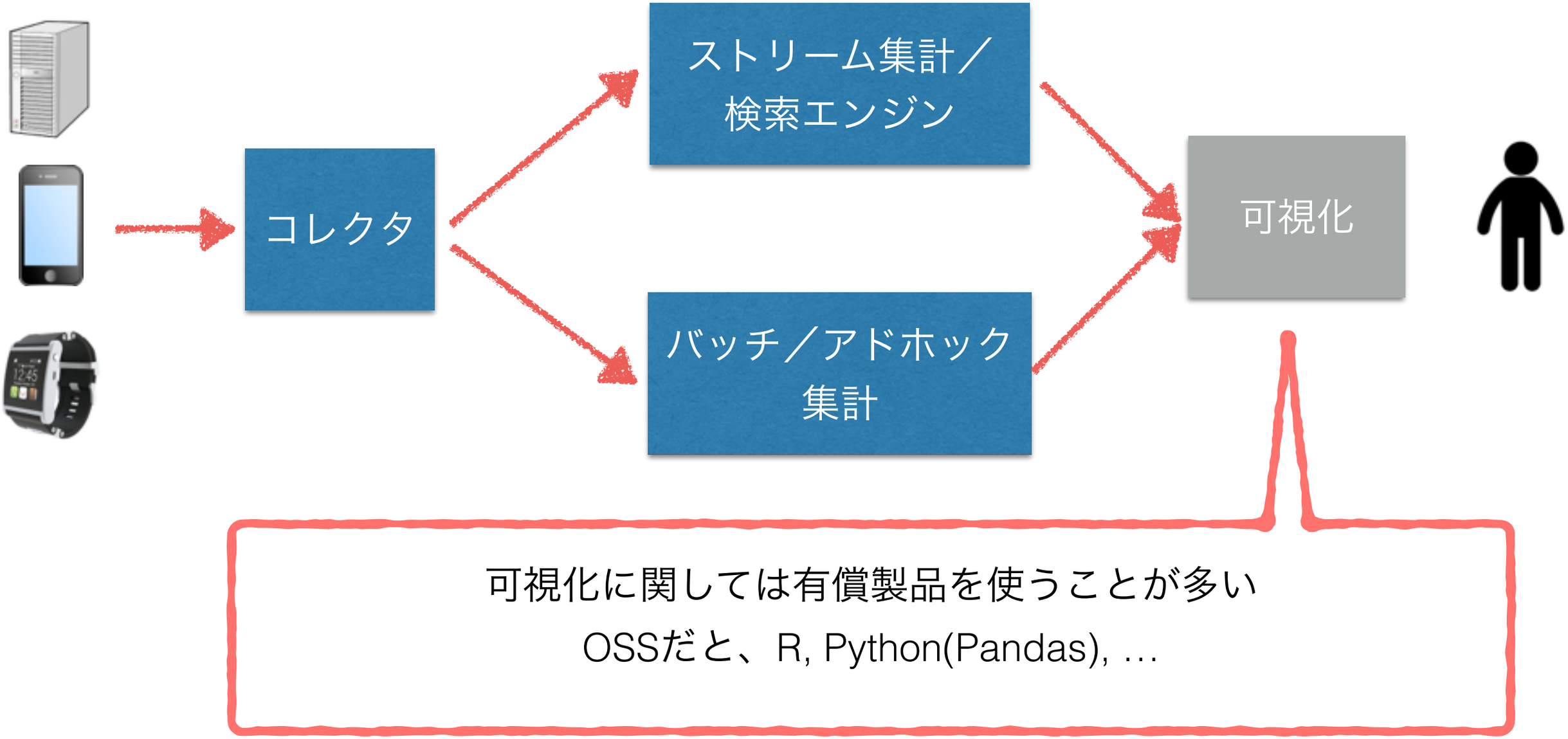
ビッグデータ分析基盤の基本構成（4/6）



ビッグデータ分析基盤の基本構成（5/6）

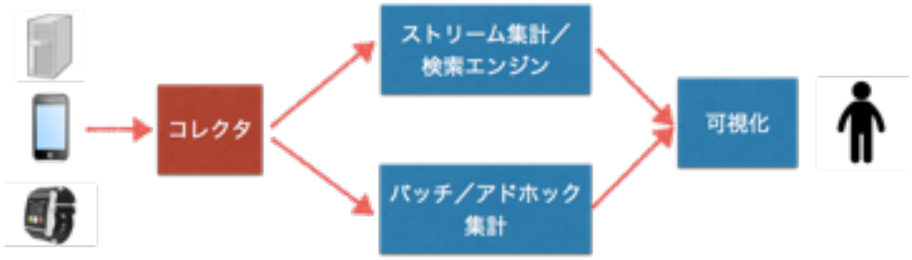


ビッグデータ分析基盤の基本構成（6/6）



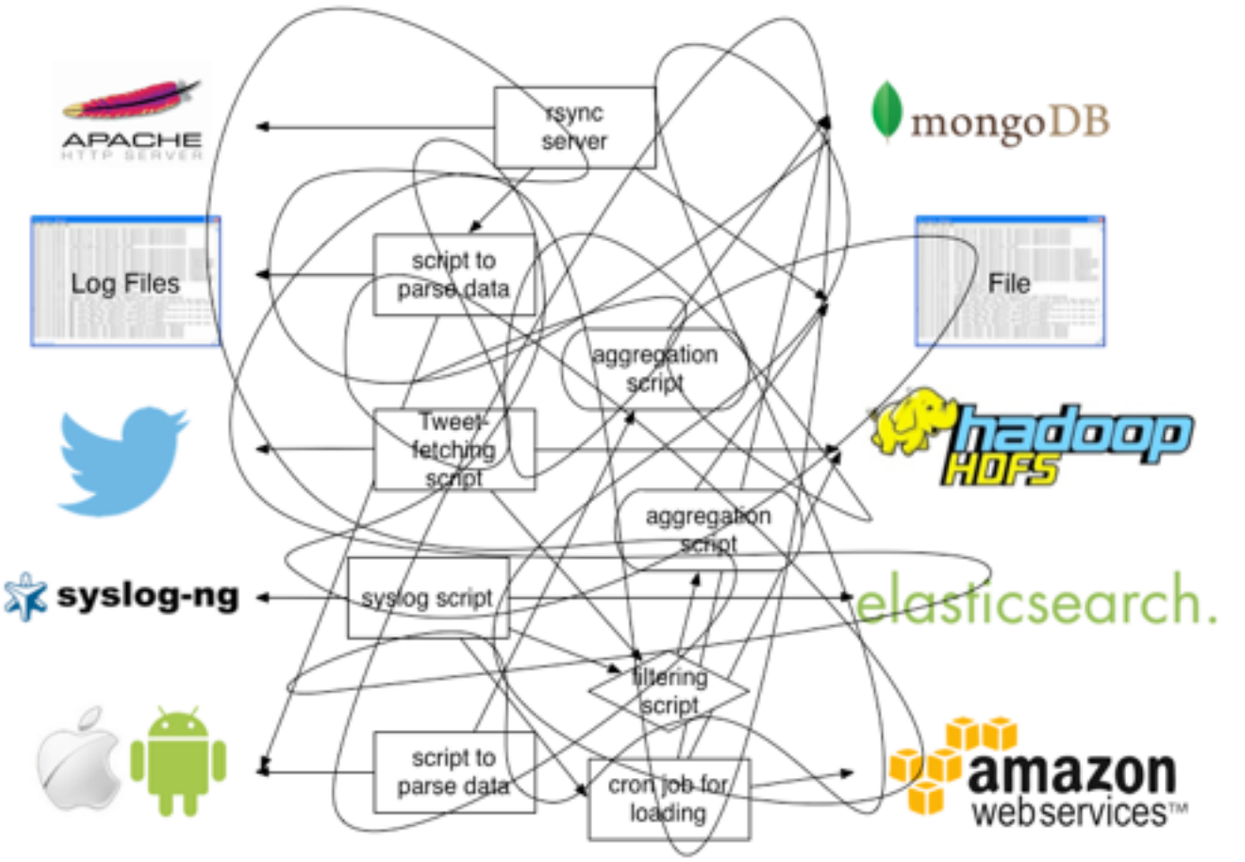
可視化に関しては有償製品を使うことが多い

OSSだと、R, Python(Pandas), …

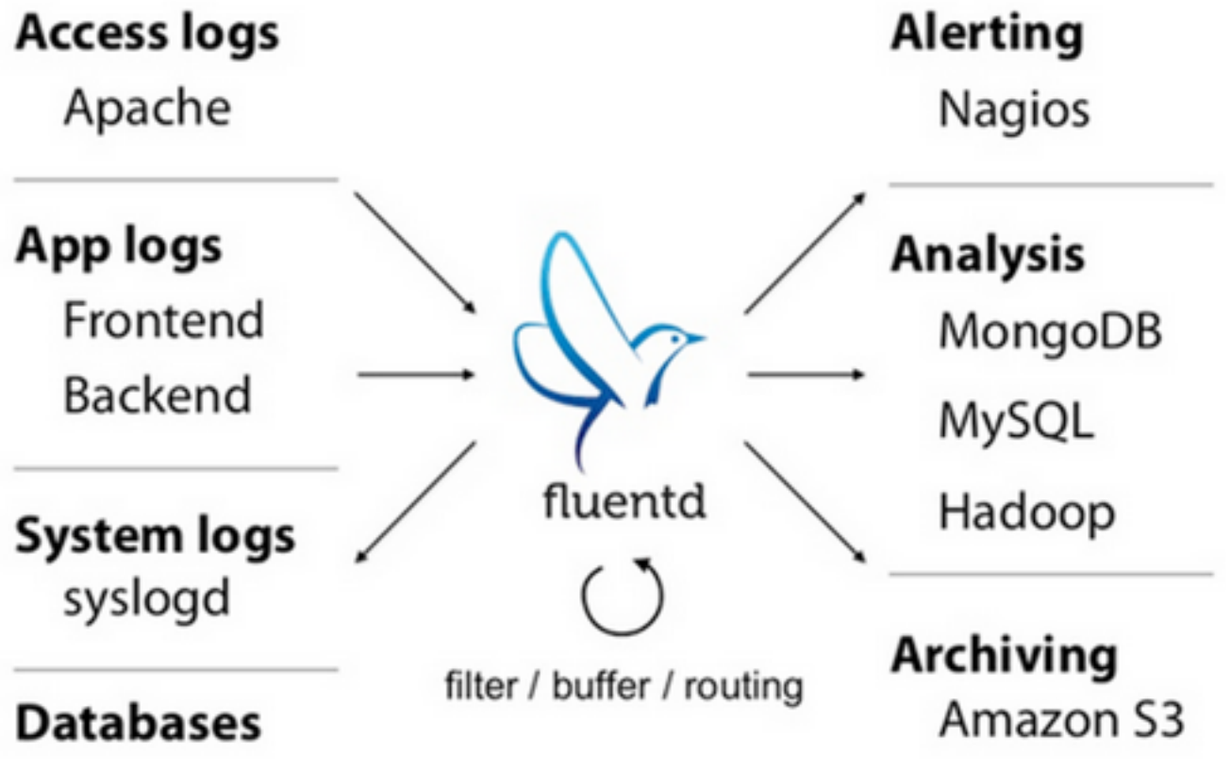


Fluentd
ストリーミング処理のデータの流れをシンプルに

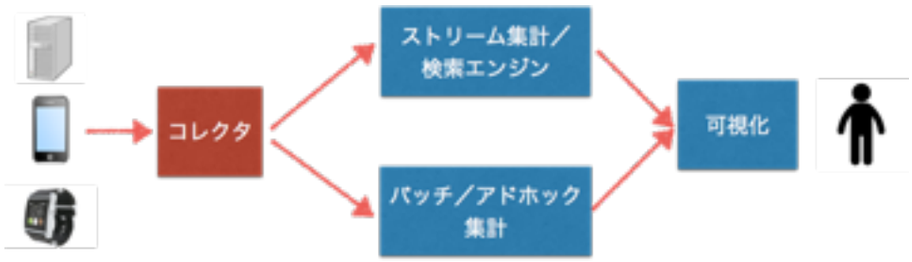
Before



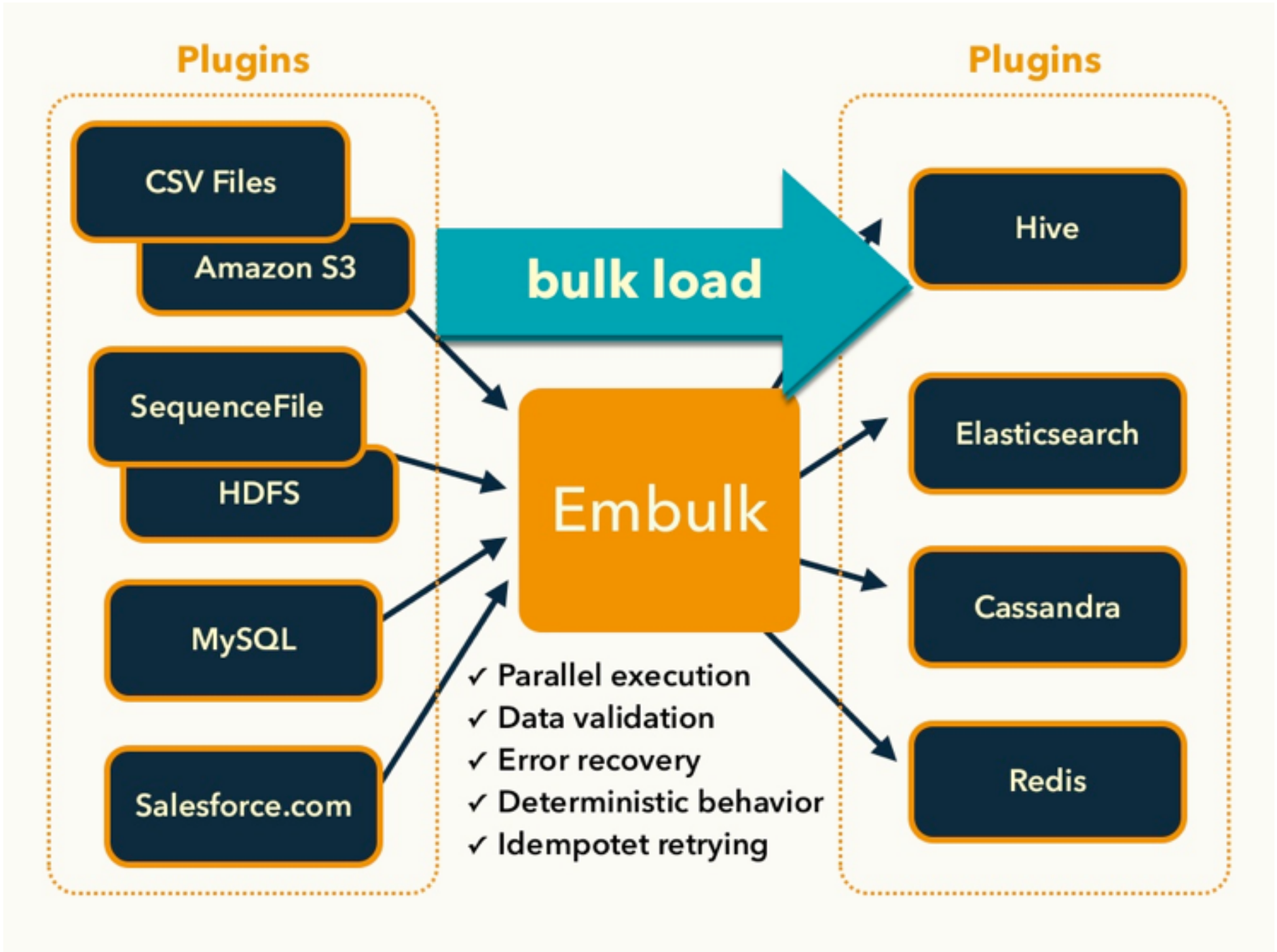
After



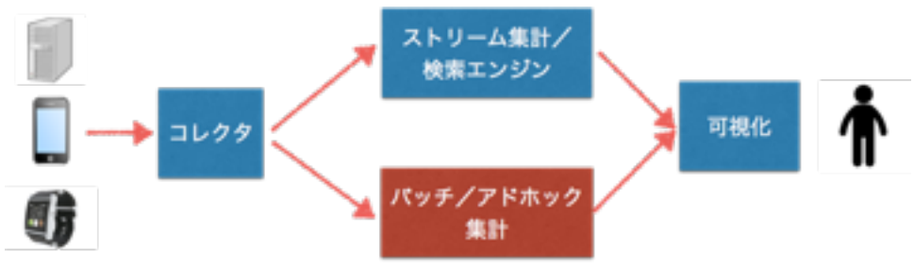
引用：http://www.fluentd.org/architecture



Embulk
バッチ処理のデータの流れをシンプルに



引用：http://www.slideshare.net/frsyuki/embuk-making-data-integration-works-relaxed

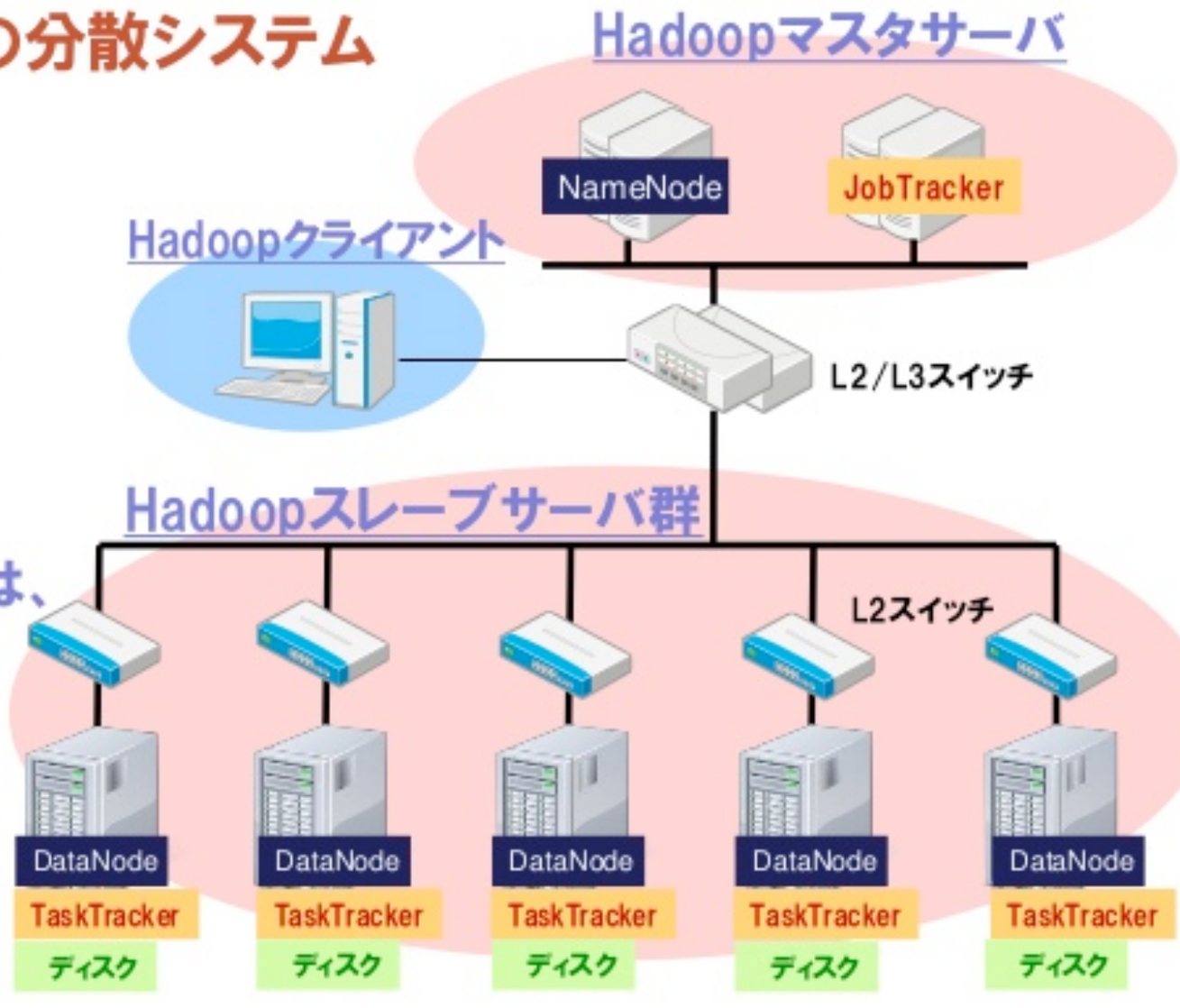


Hadoop
大規模分散処理を容易に扱えるミドルウェア

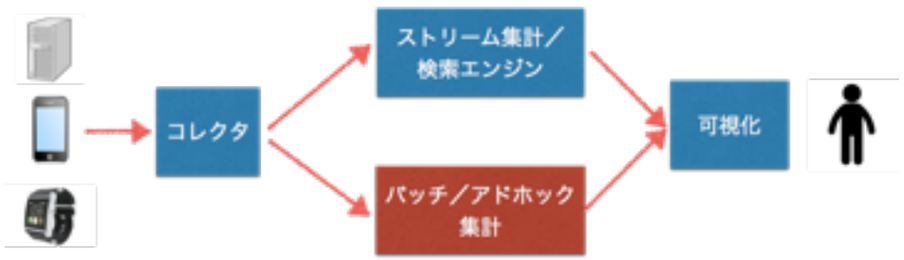
Hadoopの構成（従来）

■ 集中管理型の分散システム

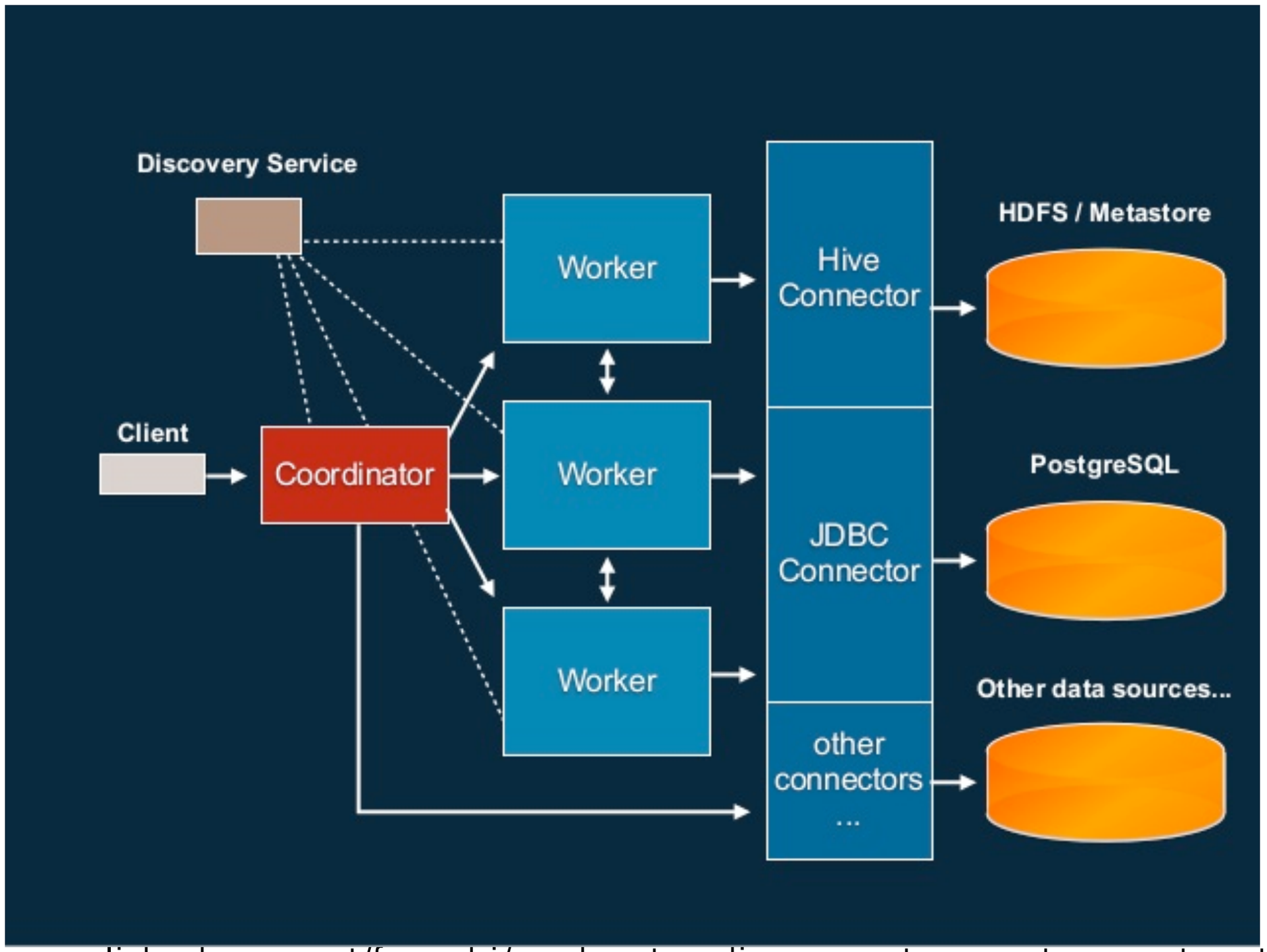
- データ管理や分散処理ジョブの管理はマスタサーバが実施
- スレーブサーバは、分散処理の実行やデータの実体を保存



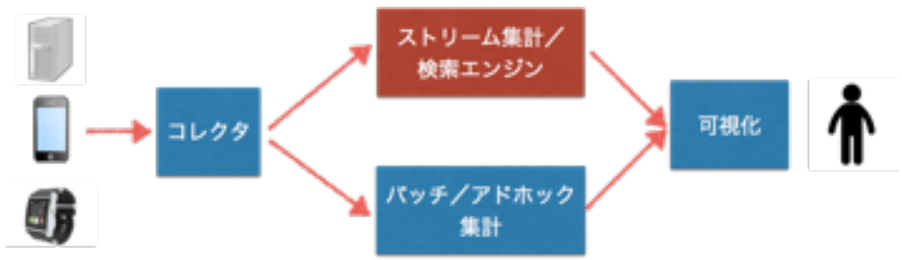
引用：http://www.slideshare.net/hamaken/hadoop-cloudera-world-tokyo-2014



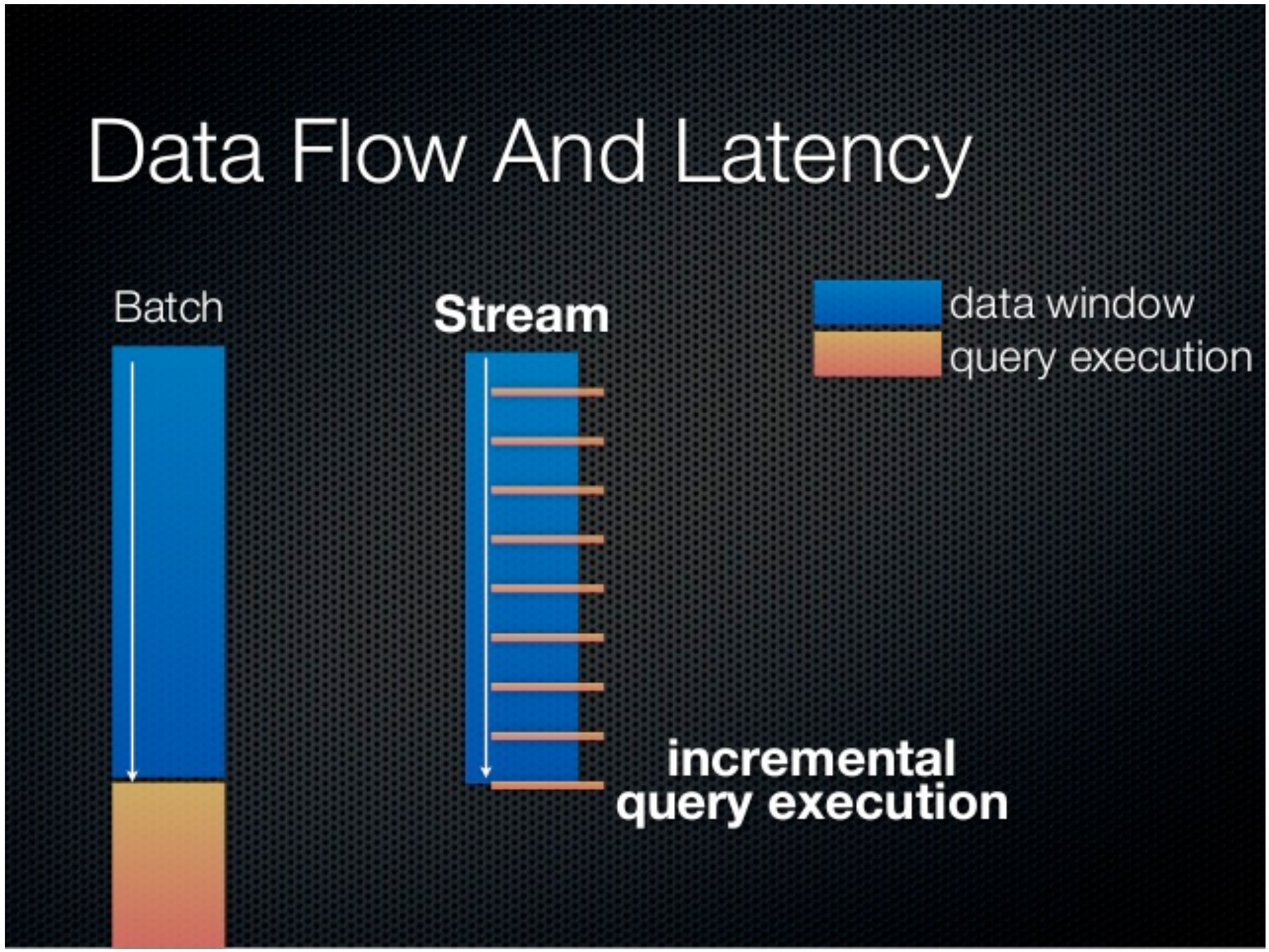
Presto
MPP型クエリエンジン



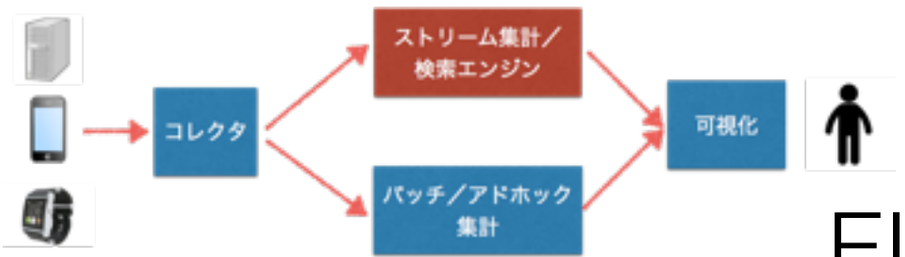
http://www.slideshare.net/frsyuki/understanding-presto-presto-meetup-tokyo-1



Norikra
ストリーミングクエリエンジン



引用：http://www.slideshare.net/tagomoris/norikra-stream-processing-with-sql

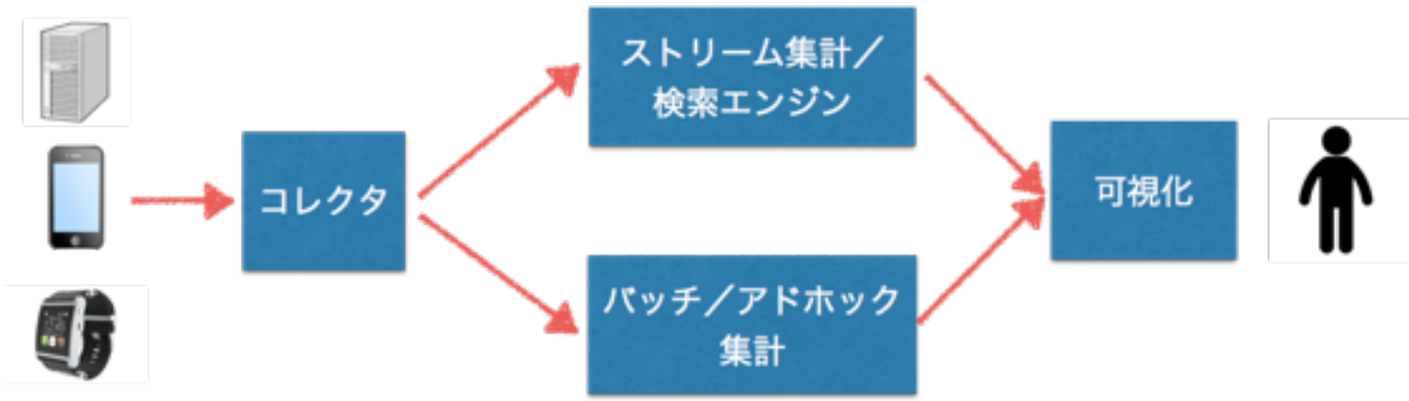


Elasticsearch (+ Kibana)
全文検索エンジン(+可視化)

引用：http://www.elasticsearch.org/blog/kibana-4-literally/

ビッグデータ分析基盤を支えるOSSたち
まとめ

- 多種多様で大量のデータの分析と利益の獲得が目的
  - データ収集
  - データ保存と分析
  - データ可視化



    - 紹介していないけど、他にもモニタリング、ジョブ管理、etc…

- 様々なOSSが課題の解決に向けて、開発が活発に行われている
- 各OSS自体について知り、特長を組み合わせることでそれぞの良さを引き出していくことが大事

終わりに。

- 今回はビッグデータ分析基盤に関わるOSSについて紹介をしました。
- 今回紹介できなかった素晴らしいOSSが世の中にはいっぱいあります。
- 素晴らしいOSSは、時間とともにさらに素晴らしくなっていきます。
- そんな素晴らしいOSSの最先端の今をOSCで学んでいきましょう！