# Embulk - Getting Started -

shibataka000

embulk

# Agenda

1. What's Embulk?
2. Getting started
3. 所感

# What's Embulk?

# What's Embulk?

- Treasure Data製のバルクデータローダ
  （大量のデータを一括登録するアプリケーション）

# Fluentd vs Embulk

## Fluentd

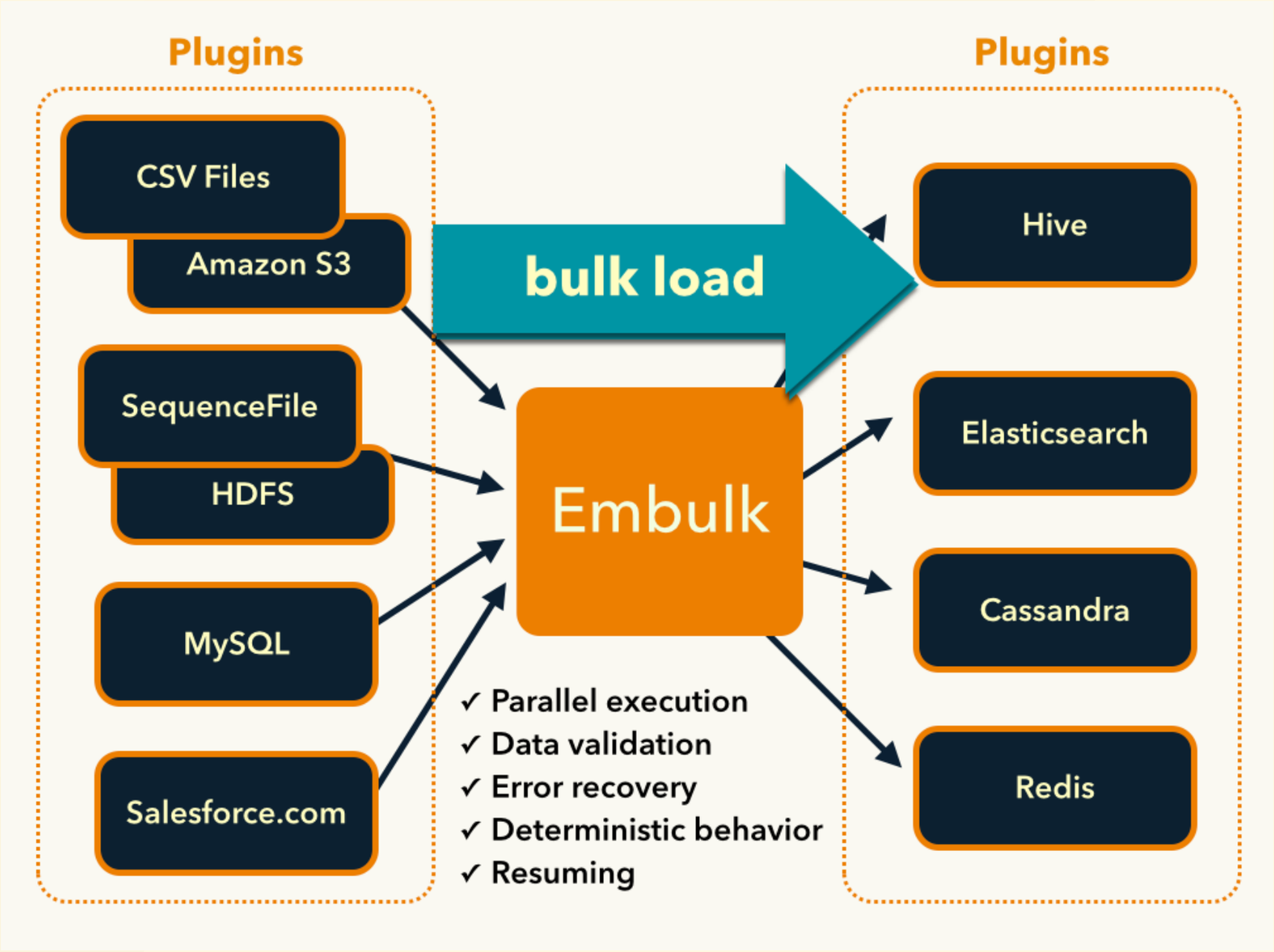
- リアルタイム性の高いケース
- ログ収集など

## Embulk

- 日次でのマスタデータのコピーなど

# Embulk supports

- 入力ファイルフォーマットの自動推定
- 巨大データセットの並列＆分散実行
- トランザクション制御
- プラグインによる入力/出力方式の拡張
- 再開


Plugins
Plugins
CSV Files
Amazon S3
SequenceFile
HDFS
MySQL
Salesforce.com
bulk load
Embulk
Hive
Elasticsearch
Cassandra
Redis
✓ Parallel execution
✓ Data validation
✓ Error recovery
✓ Deterministic behavior
✓ Resuming

# Plugins (In/Out)

List of Plugins by Category より

- mysql, postgresql, redis, influxdb
- bigquery, redshift, dynamodb, s3, azure_blob_storage
- elasticsearch, hdfs

など

# 注意点

- 本番稼働中のサーバで実行しないこと
  - 実行時の負荷が高いため

# Getting started

# Getting started

## Scenario

- CSVファイルをElasticsearchに転送する

## Data

```
id,account,time,purchase,comment
1,32864,2015-01-27 19:23:49,20150127,embulk
2,14824,2015-01-27 19:01:23,20150127,embulk jruby
3,27559,2015-01-28 02:20:02,20150128,"Embulk ""csv"" parser plugin"
4,11270,2015-01-29 11:54:36,20150129,NULL
```

# Getting started

## 1. Setup Embulk

```
curl --create-dirs -o ~/.embulk/bin/embulk -L "http://dl.embulk.org/embulk-late
chmod +x ~/.embulk/bin/embulk
echo 'export PATH="$HOME/.embulk/bin:$PATH"' >> ~/.bashrc
source ~/.bashrc
```

# Getting started

## 2. Install Elasticsearch plugin

```
embulk gem install embulk-output-elasticsearch
```

# Getting started

## 3. Guess input file formats

```
embulk guess ./mydata/csv -o config.yml
```

# Getting started

## 4. config.yml (in)

```
in:
  type: file
  path_prefix: ./mydata/csv/
  decoders:
  - {type: gzip}
  parser:
    type: csv
    #（中略）
    columns:
    - {name: id, type: long}
    - {name: account, type: long}
    - {name: time, type: timestamp, format: '%Y-%m-%d %H:%M:%S'}
    - {name: purchase, type: timestamp, format: '%Y%m%d'}
    - {name: comment, type: string}
```

# Getting started

## 5. config.yml (out)

```
out:
  type: elasticsearch
  index: embulk
  index_type: embulk
  nodes:
  - {host: localhost}
```

# Getting started

## 6. Run the bulk loading

```
embulk run config.yml -c state.yml
```

`state.yml` には最後に転送したファイル名などが記録される

# 所感

# 所感

- Embulkは簡単に使える
  - 実行はjarを実行するだけ
  - pluginによる多様な入出力形式のサポート
- FluentdでなくEmbulkを使うメリットがいまいちわからない
  - 特にデータ解析班の場合

# 未検証

- 実際にバルクデータを扱った場合の性能や安定性はどうか？

# まとめ

- Embulkとは
  - バルクデータローダ
- Embulkの主な特徴
  - 入力ファイルフォーマットの自動推定
  - プラグインによる入力/出力方式の拡張