Fraud Prevention with Machine Learning

November 29th, 2016

João António
joao.antonio@feedzai.com

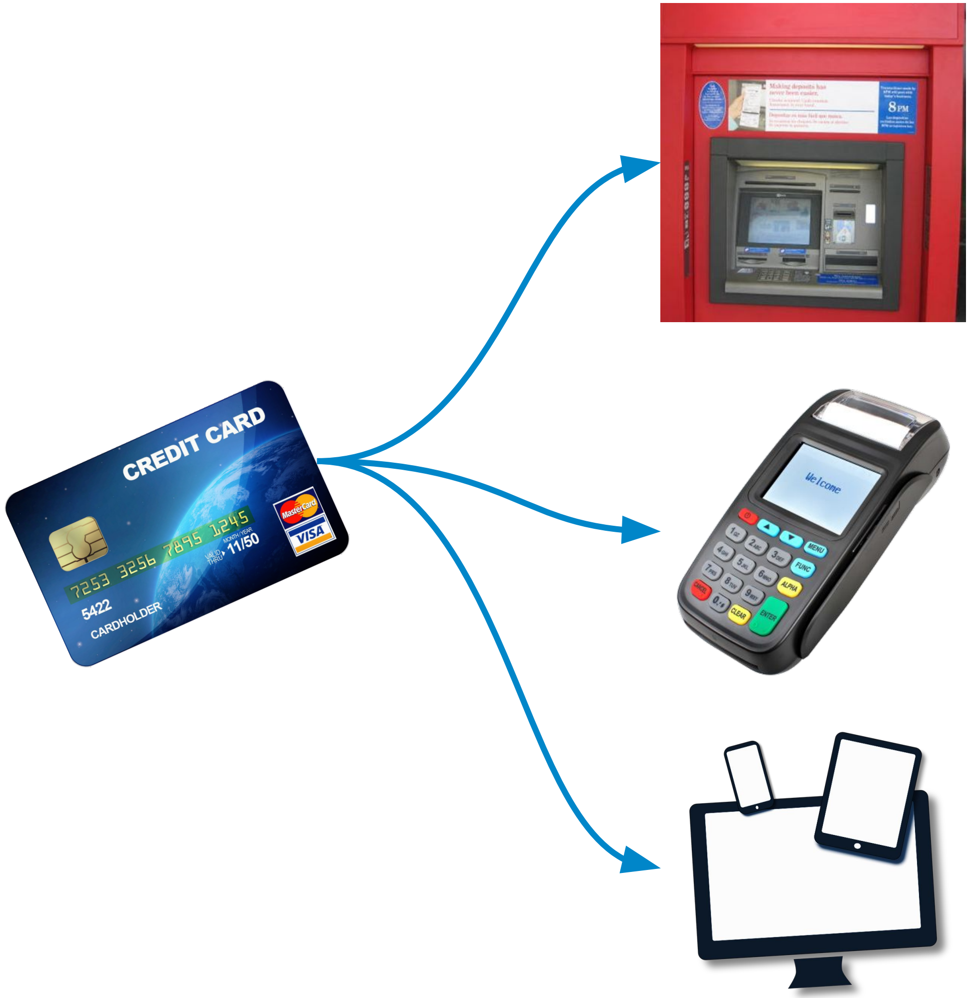
feedzai
research

Outline

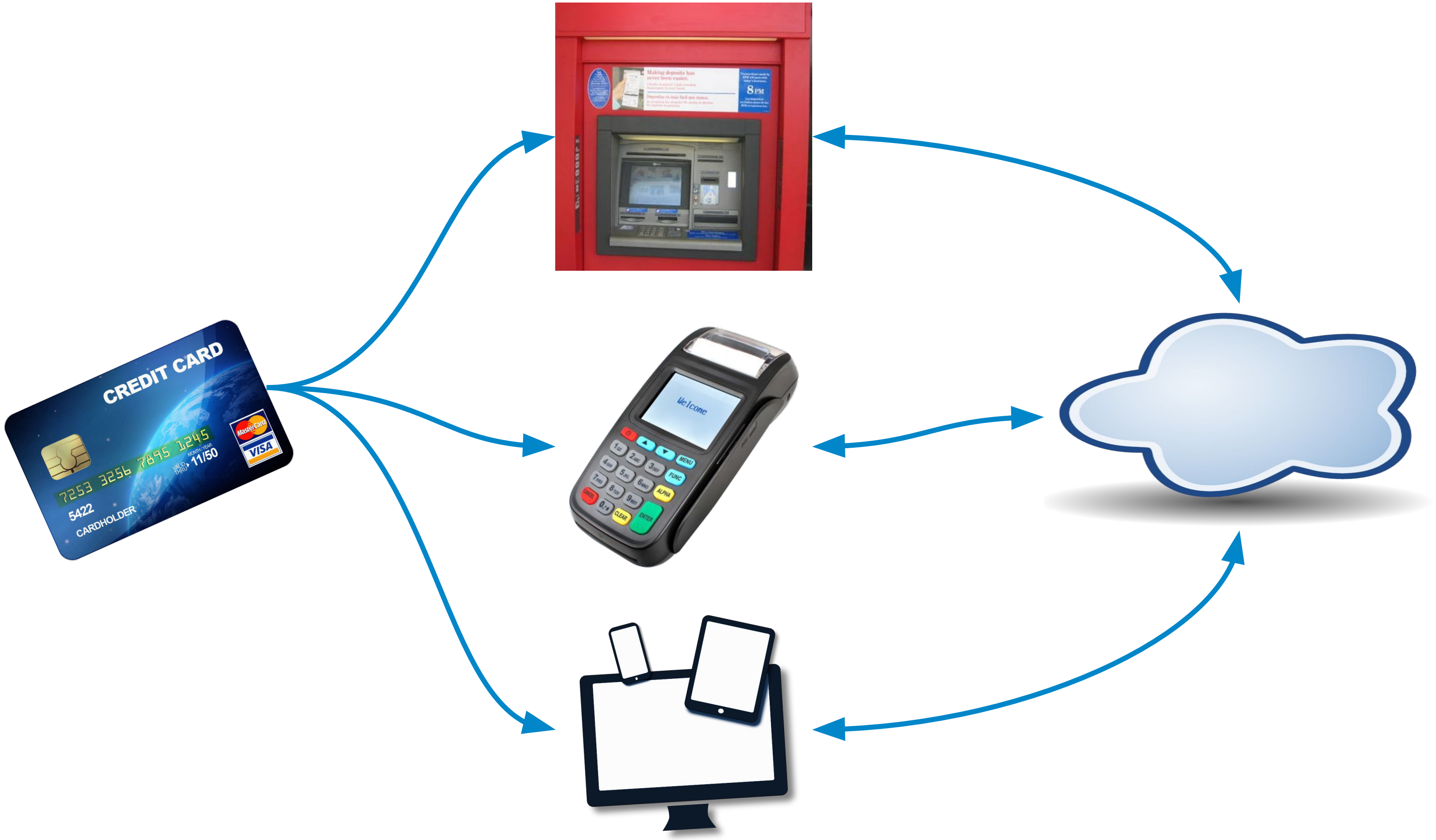
- Primer on Credit Card Fraud
- Data Science
    - Challenges
    - Human Intuition
    - Machine Learning
- Data Engineering
    - Real-time Metrics
    - Batch Metrics
    - Transaction Lifecycle

Primer on Credit Card Fraud

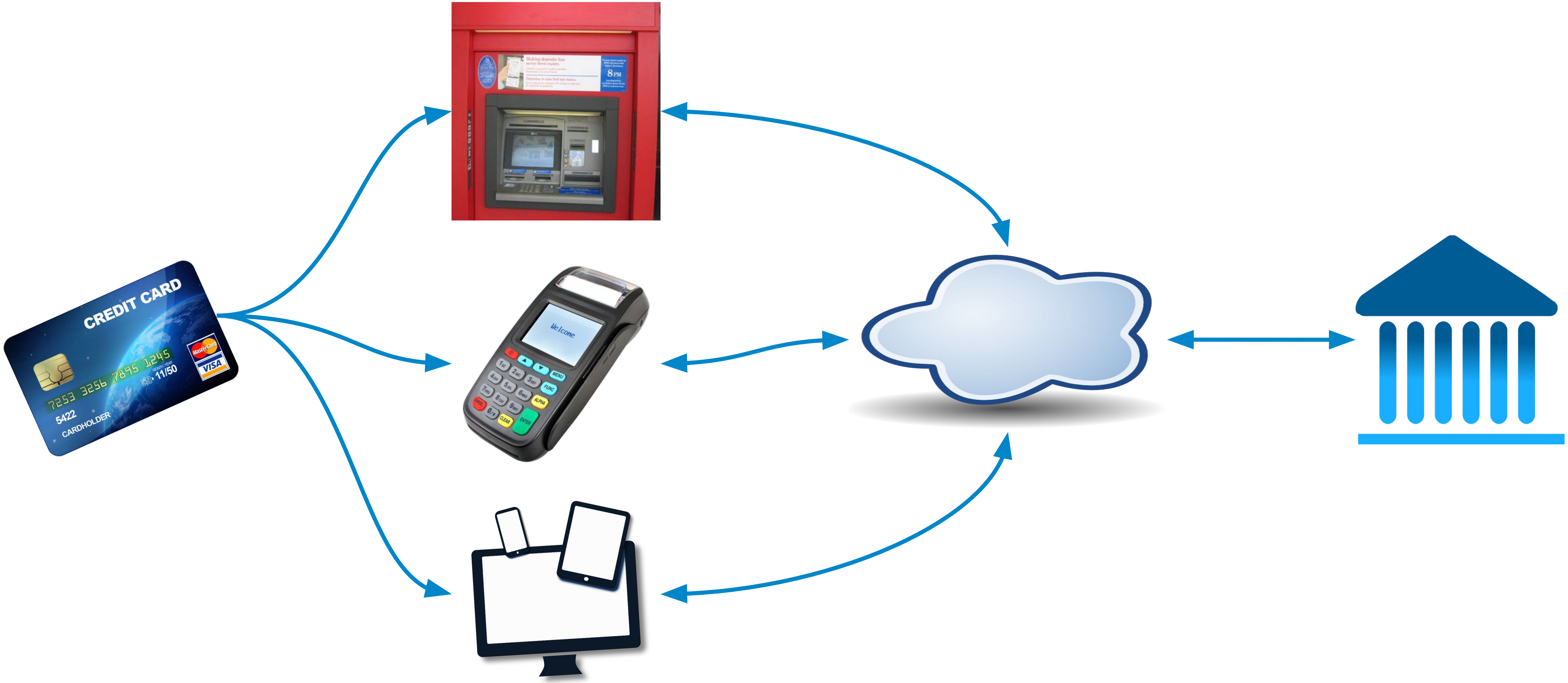
Primer on Credit Card Fraud



Primer on Credit Card Fraud



Primer on Credit Card Fraud



How do you prevent fraud?

Data Science

Challenges

- Data Volume
  - 20M+ transactions per day
  - 200+ added features

Challenges

- Data Volume
    - 20M+ transactions per day
    - 200+ added features

- (Really) Unbalanced Problem
    - Positive Class: 0.1% ~ 1.0%
    - Stratified Sampling
    - Undersampling

Challenges

- Data Volume
    - 20M+ transactions per day
    - 200+ added features

- (Really) Unbalanced Problem
    - Positive Class: 0.1% ~ 1.0%
    - Stratified Sampling
    - Undersampling

- Time Series Problem

Human Intuition

- Real-time Features
  - Distance between consecutive transactions
  - Number of failed attempts in last hour
  - Total money spent in last hour by merchant

Human Intuition

- Real-time Features
    - Distance between consecutive transactions
    - Number of failed attempts in last hour
    - Total money spent in last hour by merchant
- Batch Features
    - Total money spent in last month
    - Average transaction money by merchant in last month
    - Percentage of transactions abroad in last month

Human Intuition

- Real-time Features
  - Distance between consecutive transactions
  - Number of failed attempts in last hour
  - Total money spent in last hour by merchant
- Batch Features
  - Total money spent in last month
  - Average transaction money by merchant in last month
  - Percentage of transactions abroad in last month
- Enrichment Features

Machine Learning

- Literature Algorithms
    - Random Forests
    - XGBoost

Machine Learning

- Literature Algorithms
  - Random Forests
  - XGBoost

- In-house Algorithms
  - Points Of Compromise
  - Cost Sensitive Scoring

Machine Learning

- Literature Algorithms
    - Random Forests
    - XGBoost
- In-house Algorithms
    - Points Of Compromise
    - Cost Sensitive Scoring
- Others

Data Engineering

Real-time Metrics

- Event Stream Processing
  - Data Studio
  - PQL
  - PKernel

Real-time Metrics

- Event Stream Processing
    - Data Studio
    - PQL
    - PKernel

- Mission-critical Java
    - Zing® JVM (by Azul Systems®)

Batch Metrics

- (Replayed) Event Stream Processing
    - Data Studio
    - PQL
    - PKernel
    - Cassandra

Batch Metrics

- (Replayed) Event Stream Processing
    - Data Studio
    - PQL
    - PKernel
    - Cassandra

- Yarn / Spark for Job Scheduling

Batch Metrics

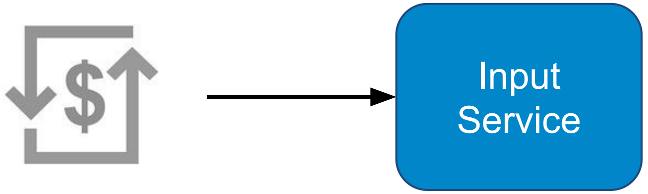
- (Replayed) Event Stream Processing
  - Data Studio
  - PQL
  - PKernel
  - Cassandra

- Yarn / Spark for Job Scheduling

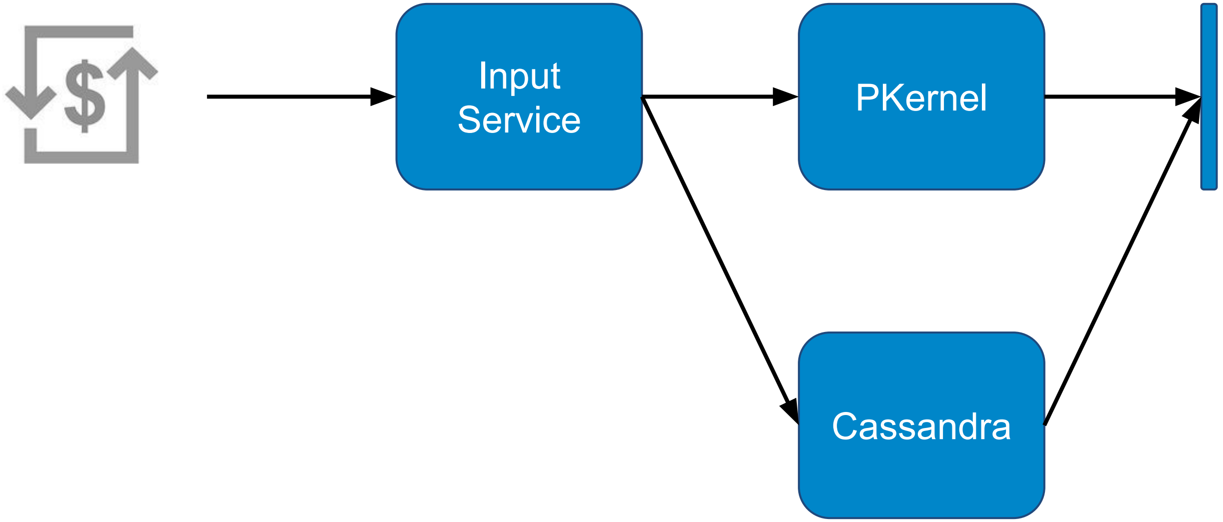
- Mission-critical Java
  - Zing® JVM (by Azul Systems®)

Transaction Lifecycle

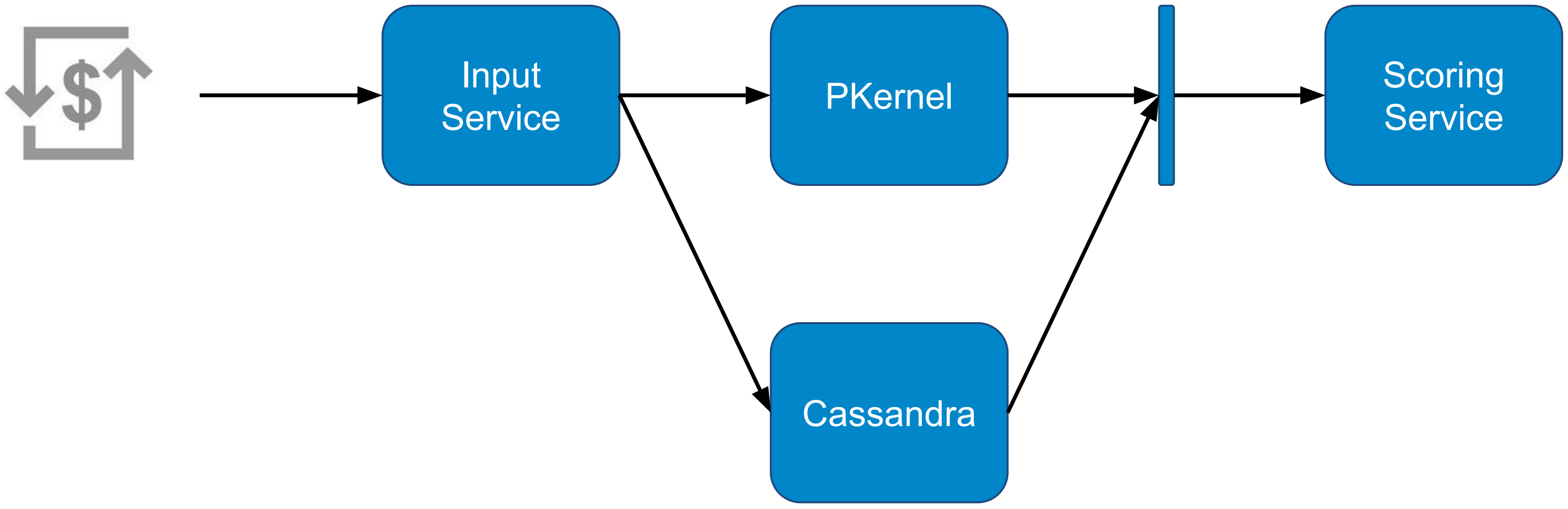
Transaction Lifecycle



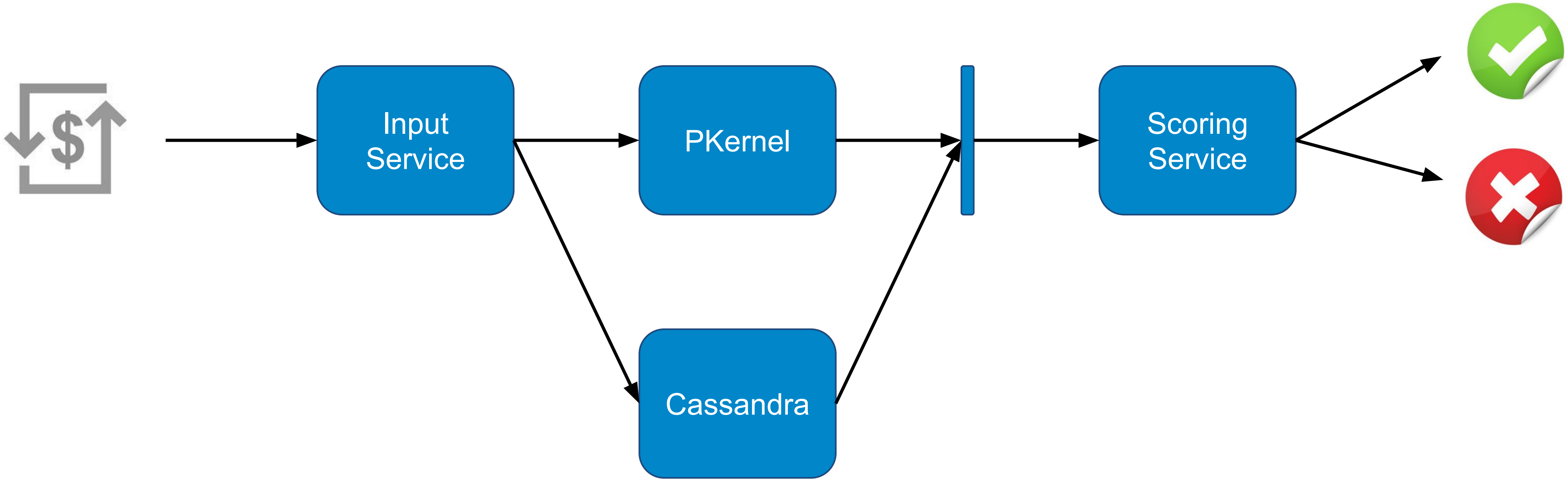
Transaction Lifecycle



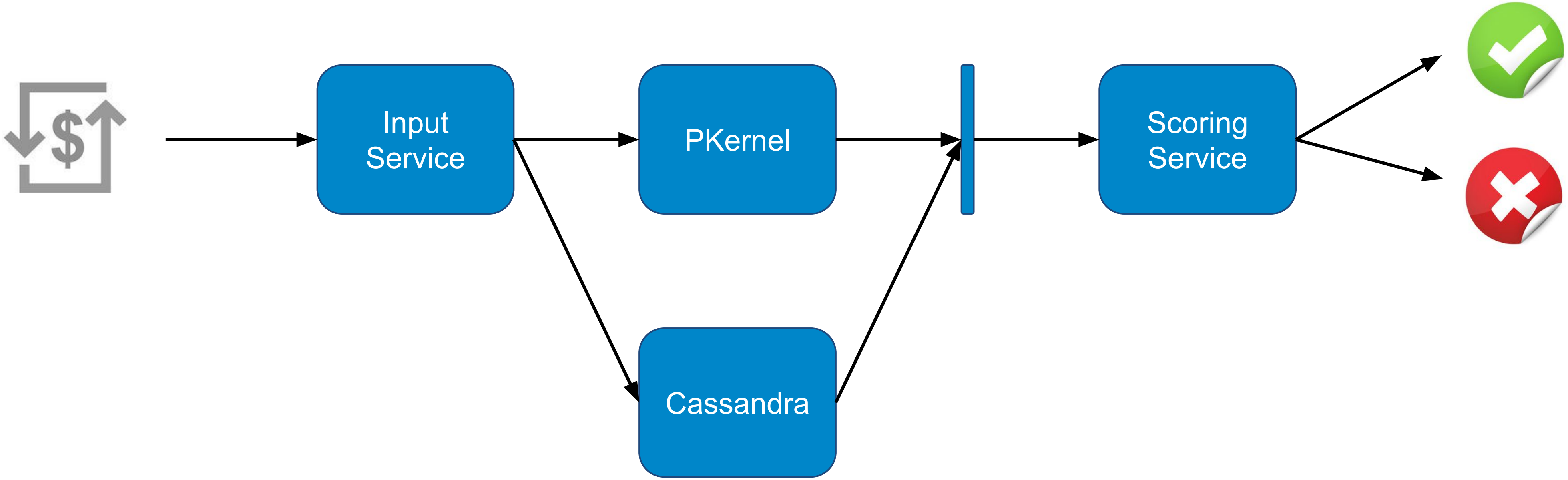
Transaction Lifecycle



Transaction Lifecycle



Transaction Lifecycle



Processing time ~ 100 ms on 99.999% percentile

Thank you!