



# Credit Fraud Prevention with Spark and Graph Analysis

Chris D'Agostino

VP Technology, Capital One

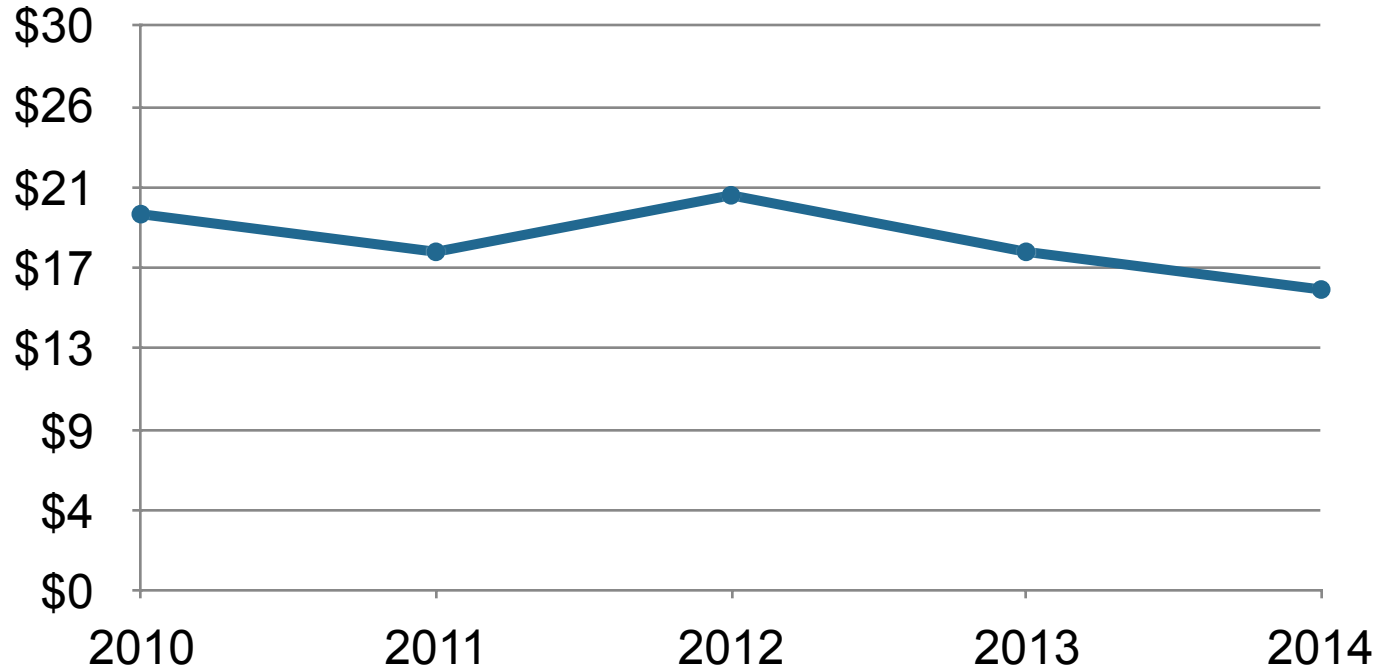
@chrisdagostino



SPARK SUMMIT 2016  
DATA SCIENCE AND ENGINEERING AT SCALE  
JUNE 6-8, 2016 SAN FRANCISCO

# Credit Card Fraud Costs Billions...

Total One Year Fraud Amount (Billions)



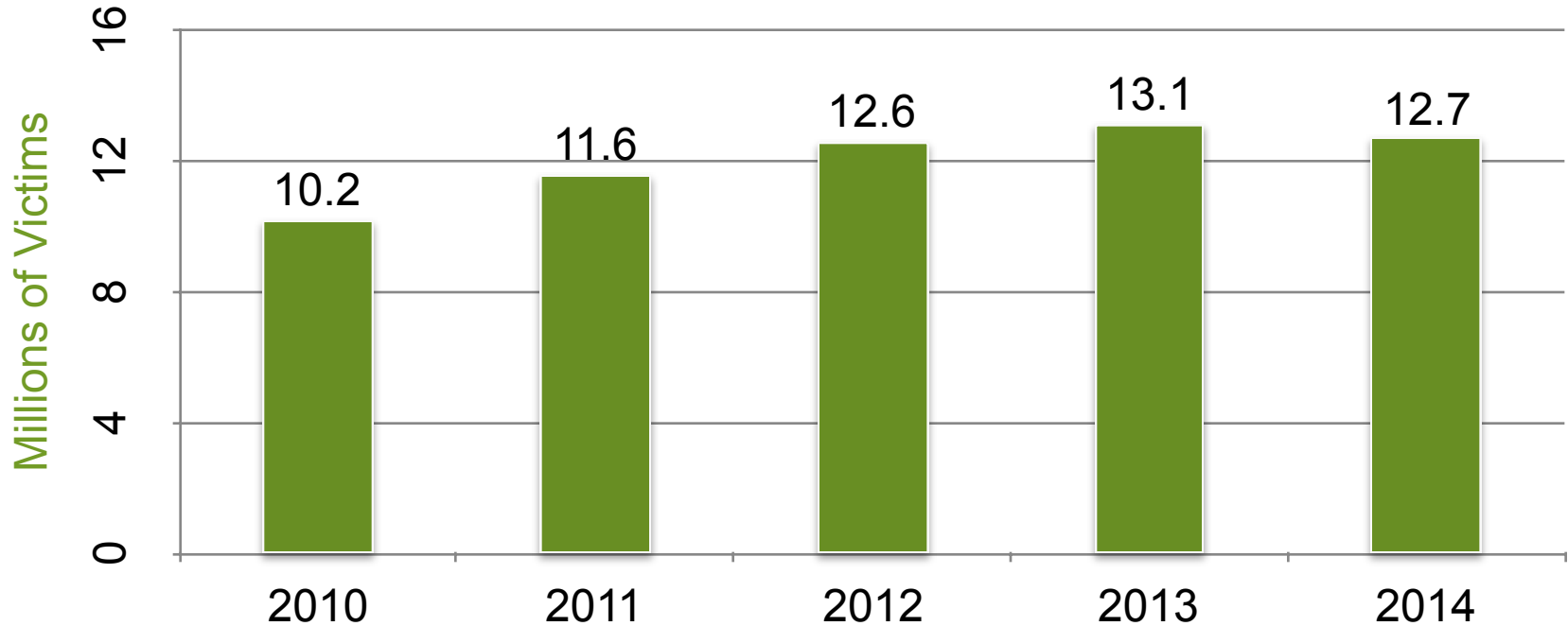
SPARK SUMMIT 2016

<http://www.nasdaq.com/article/credit-card-fraud-and-id-theft-statistics-cm520388>



# ...and Affects Many People

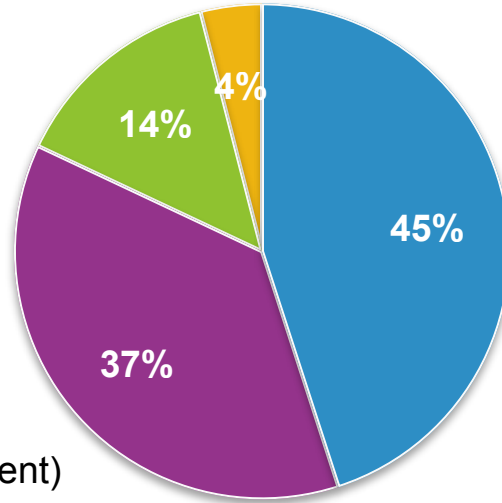
Identity fraudsters stole \$16 billion from 12.7 million U.S. consumers in 2014



# U.S. Card Fraud By Type, 2014

**Application fraud** occurs when criminals use stolen or fake documents to open an account in someone else's name. For identification purposes, criminals may try to steal documents such as utility bills and bank statements to build up useful personal information.

- Online (card not present)
- Counterfeit
- Lost/Stolen
- Other



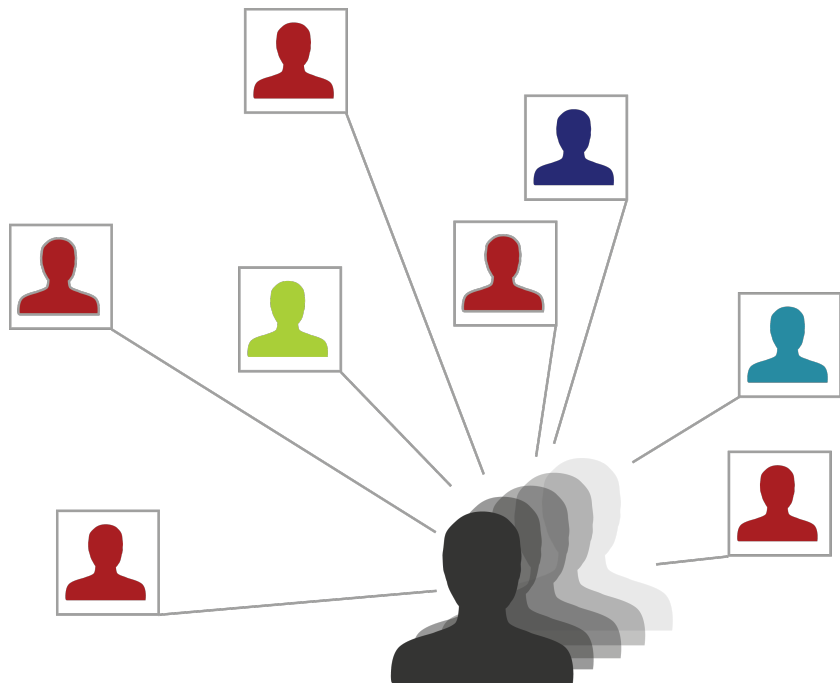
**Account Takeover** involves a criminal fraudulently using another person's bank, credit or debit card account, first by gathering information about the intended victim, then contacting their bank or credit card issuer to masquerade as the genuine account or card.

The criminal then arranges for funds to be transferred out of the account, or will change the address on the account and ask for new or replacement cards to be sent which is then used fraudulently.

Sources:

FRAUD THE FACTS 2016 FINANCIAL FRAUD ACTION UK  
<http://www.nasdaq.com/article/credit-card-fraud-and-id-theft-statistics-cm520388>

# Developing world-class defenses



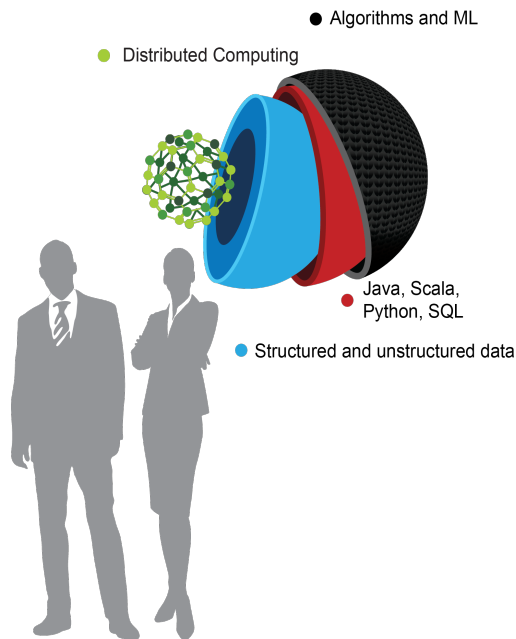
## Key Defenses

- Fraud
- Anti-Money Laundering (AML)
- Know Your Customer (KYC)

## Key Attack Vectors

- Stolen IDs
- Synthetic IDs
- Hijacked accounts

# Real-time Defenses with Spark and Graph Database



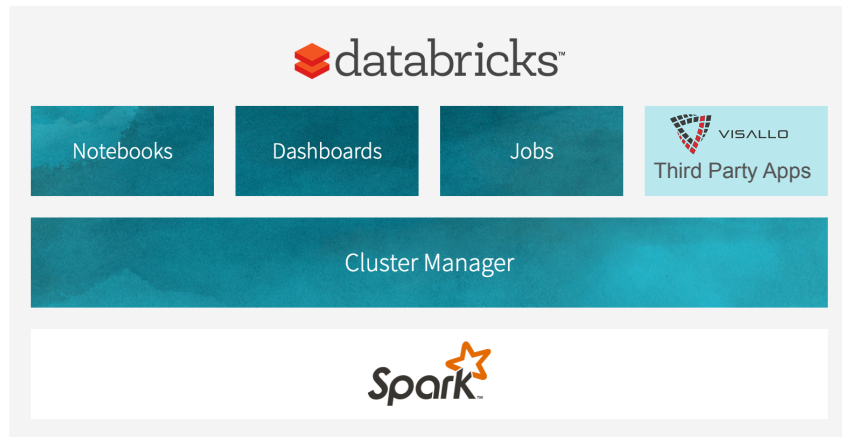
## Goals:

- Minimize financial losses, investigative costs and help customers avoid identity theft
- Combine more data sources than ever before: applicant provided, 3rd party and internal data sources to score the application for fraud — as fast as possible
- Provide a unified platform for business analysts, data scientists and data engineers to analyze data
- BDFD - Big Data (“data at rest”) & Fast Data (“streaming data”)

# Databricks provides the building blocks for supporting BDFD use cases



1. Scalable compute environment that supports batch and streaming data to train fraud models
2. Support for multiple programming languages with interactive notebooks and self-service infrastructure
3. SQL queries, graph queries and machine learning
4. Easily integrate BI tools and other 3rd party apps
5. Able to score applications quickly once we determine their “connectedness”



SPARK SUMMIT 2016

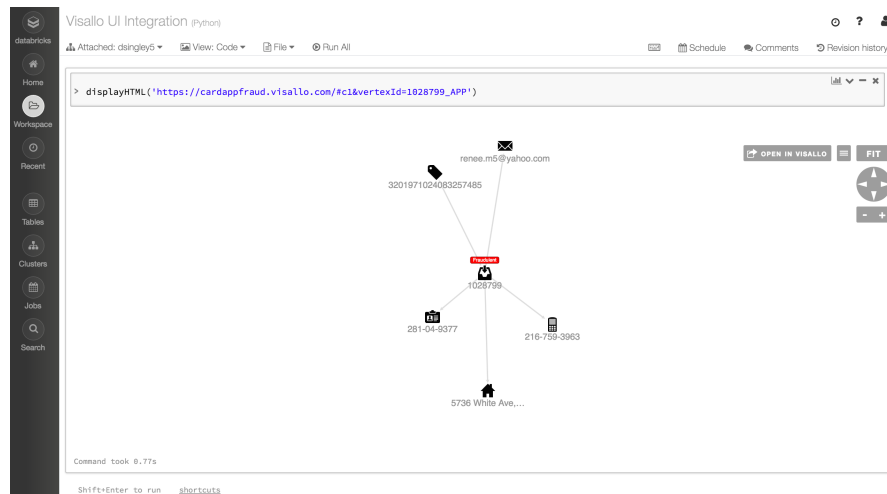
<http://www.databricks.com>



# Visallo is an open-source Graph Database for exploratory analytics and visualization



1. Provides a scalable graph database on top of Hadoop, Accumulo and Elasticsearch
2. Full CRUD operations with fine-grained access controls — attribute-level on vertices and edges
3. Integrates with Spark and Databricks — export graph/sub-graph as RDDs for computation and render sub-graphs in-notebook
4. Serves as the mutable System of Record (SOR) and case management solution



SPARK SUMMIT 2016

<http://www.visallo.org>





# Sample Data and a Simple Fraud Model

*All sample data is 100% machine generated using public datasets*



## Application



SSN — generate valid social security numbers that conform to the SSA rules



Mailing Address — generate valid street addresses and distribute applications across geographies based on U.S. Census data and income distributions



Names & Email Address — generate names and realistic emails using U.S. Census data and popular email providers



Phone number



Visitor ID

Any resemblance between the data in this prototype and any persons, living, dead or undead, is a miracle



SPARK SUMMIT 2016



Read historical credit card applications



Create vertex and edge DataFrames for the graph -> GraphFrame



Compute features for every subgraph



Get account status information for every historical application

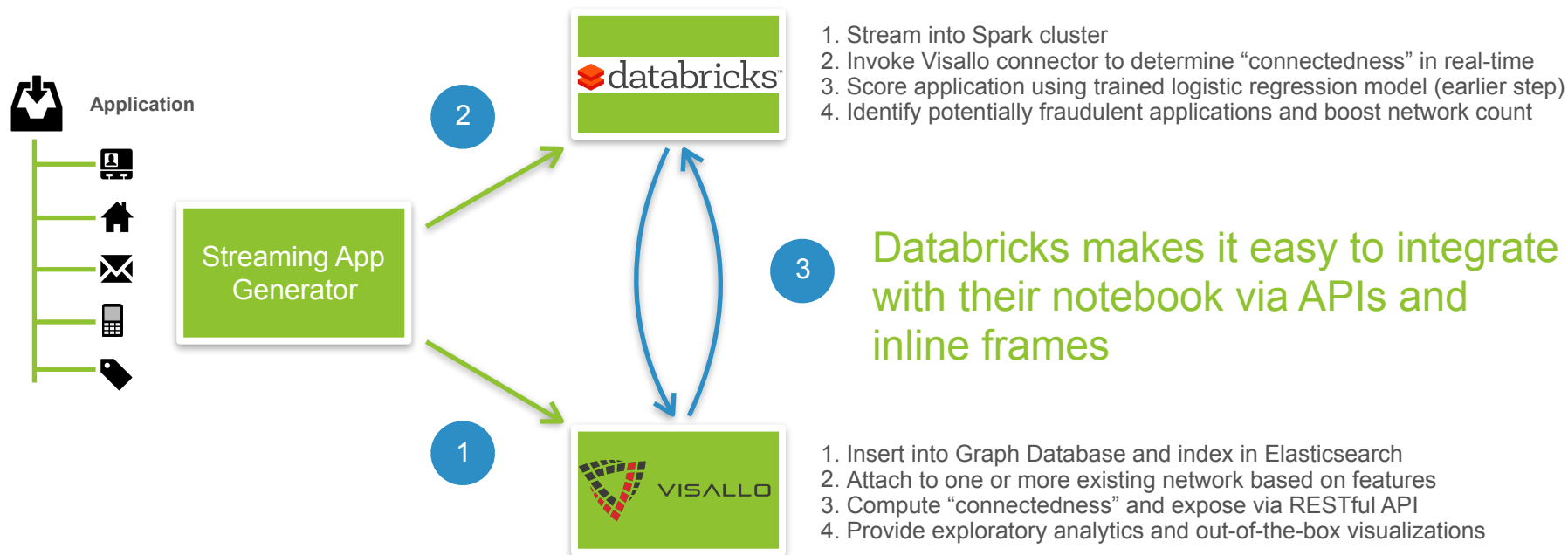


Create the model training DataFrame



Train and save the model -> `org.apache.spark.ml.classification.LogisticRegression`

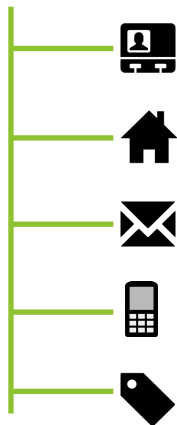
# Overview of Real-Time Architecture



# Performance Metrics



**Application**



|                       | Databricks   | Visallo   |
|-----------------------|--|---|
| Cluster Size and Type | 8 x (30GB RAM, 4 cores)  | 5 x (30GB RAM, 16 cores, 320 GB local storage)  |
| Number of Records     | 6M + 25 new applications   | 6M + 25 new applications  |
| Ingest Time           | Ingesting 6M records using Parquet files and building the immutable, in-memory Graph takes ~5mins. | Ingesting 6M records from scratch takes ~2 hours. Ingesting new apps takes ~200ms/app |
| Model Training Time   | Takes ~5 mins to train the model using 6M records  | Computing the feature vector for the new application takes ~80ms                      |
| Scoring Time          | Takes ~100ms to score a new application fraud/not fraud  | N/A   |



SPARK SUMMIT 2016





# DEMO

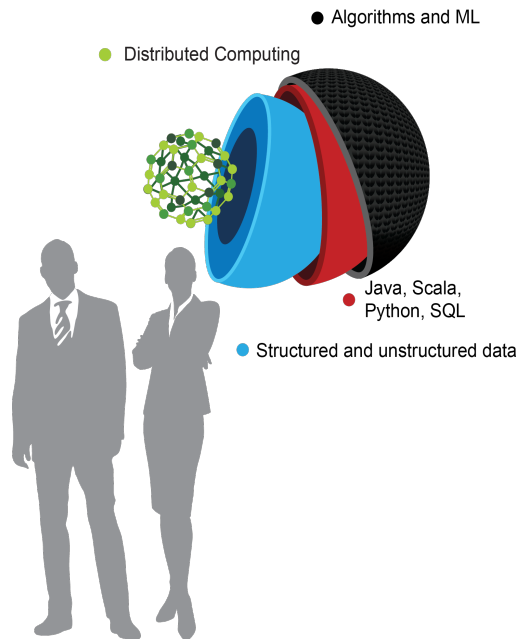
<https://dbc-ce3883ee-8add.cloud.databricks.com/#notebook/34911>

<https://dbc-ce3883ee-8add.cloud.databricks.com/#notebook/34946>

<https://cardappfraud.visallo.com>



# A Big “Thank You” to Databricks and the Team



- Richard Garris — Databricks
- Saurabh Gupte — Capital One
- Matt Wizeman — Visallo



# THANK YOU.



chris.dagostino@capitalone.com



@chrisdagostino

