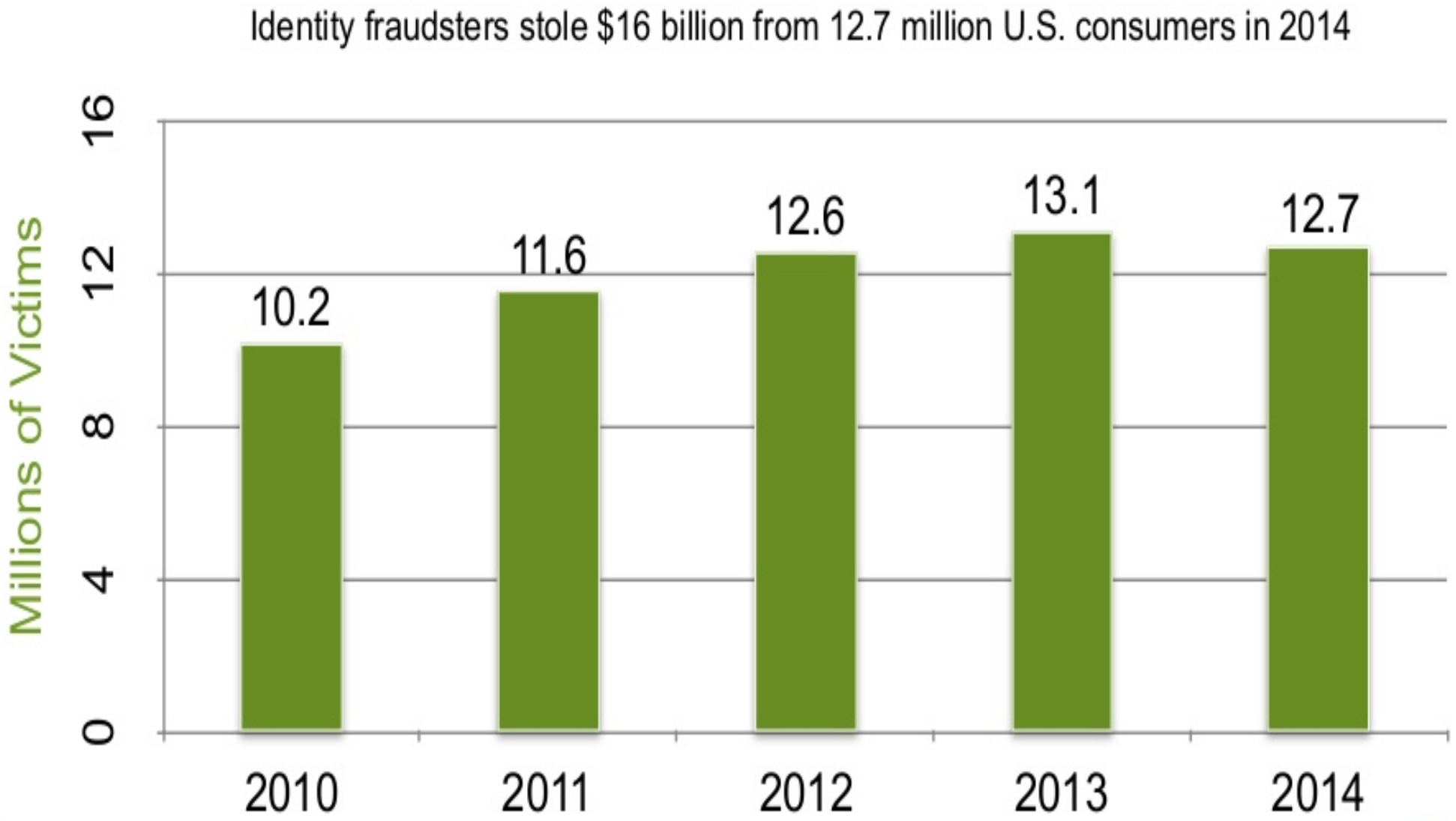
Credit Fraud Prevention with Spark and Graph Analysis

Chris D'Agostino
VP Technology, Capital One
@chrisdagostino

Credit Card Fraud Costs Billions…

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

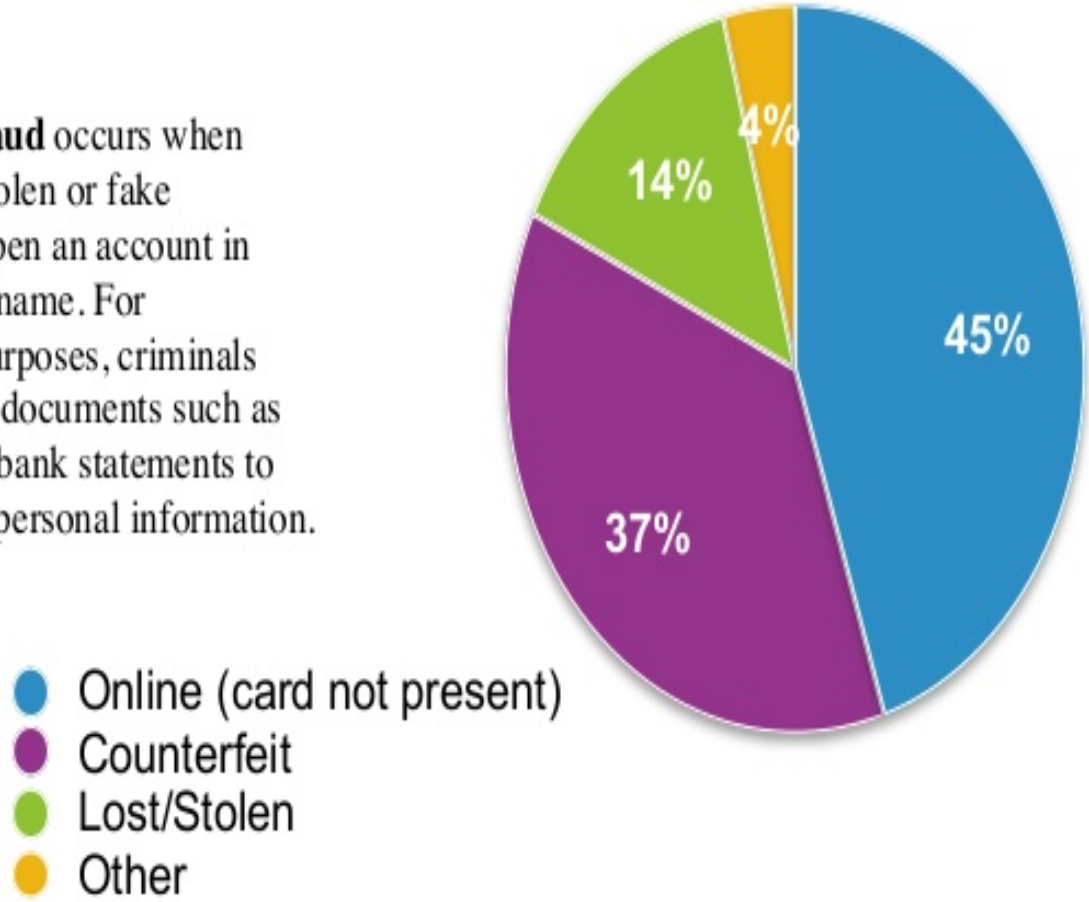
…and Affects Many People



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

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.



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.

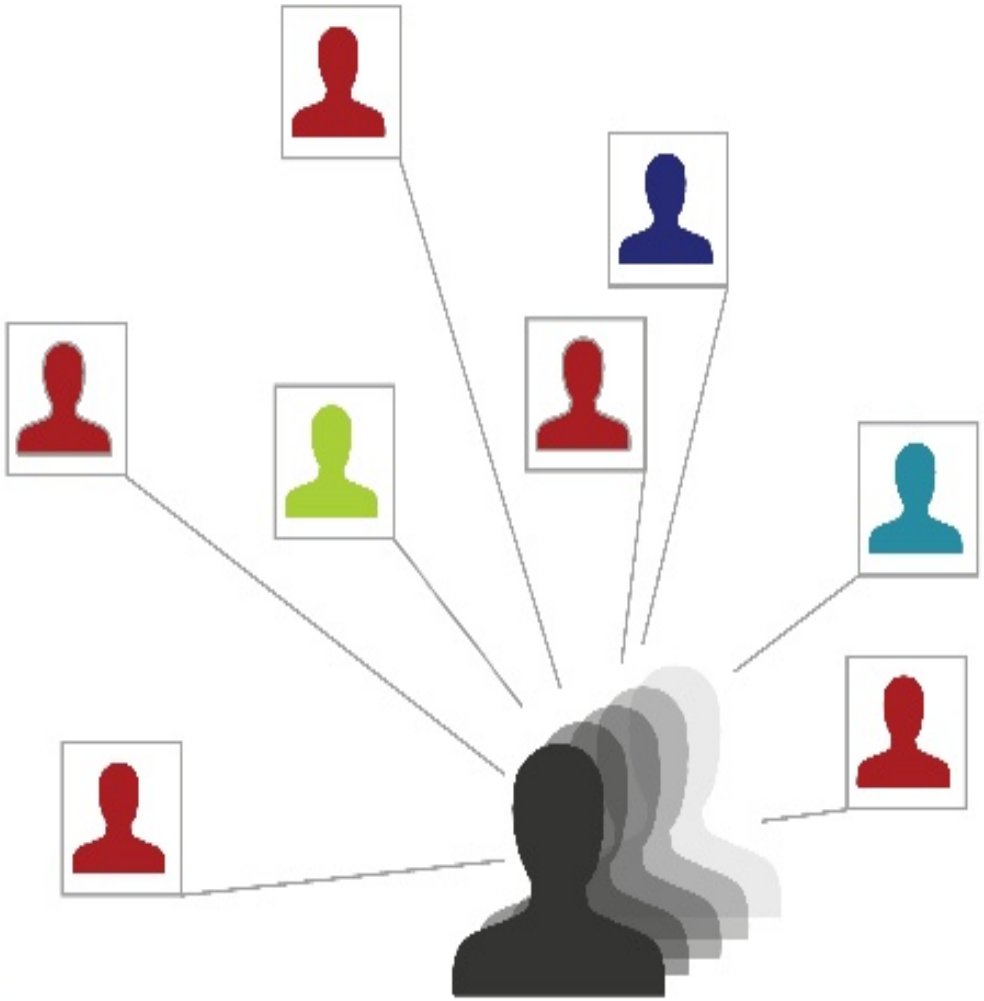
The criminal then arranges for funds to be transferred out of the account, or will change the address on the account and ask for new or replacement cards to be sent which is then used fraudulently.

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



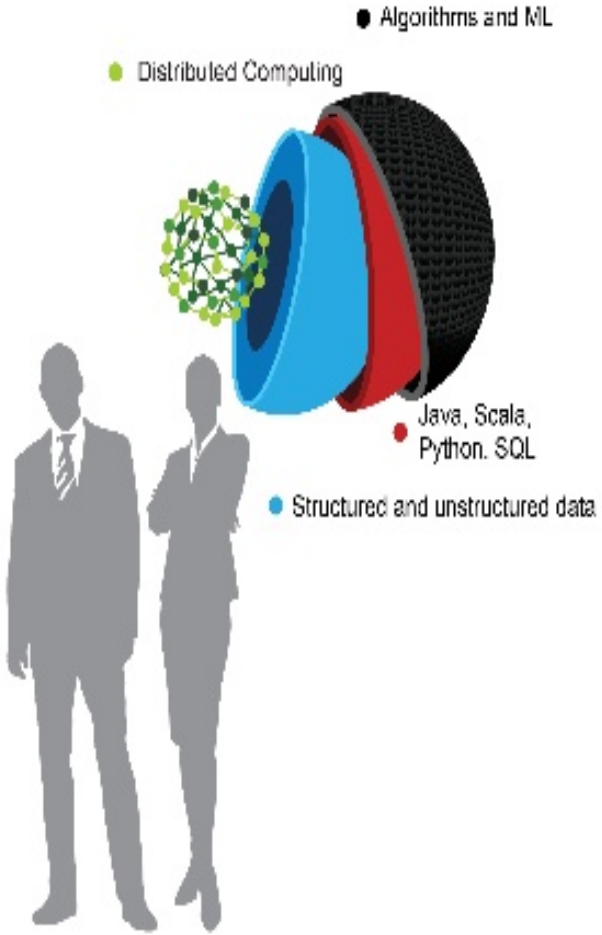
Key Defenses

- 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

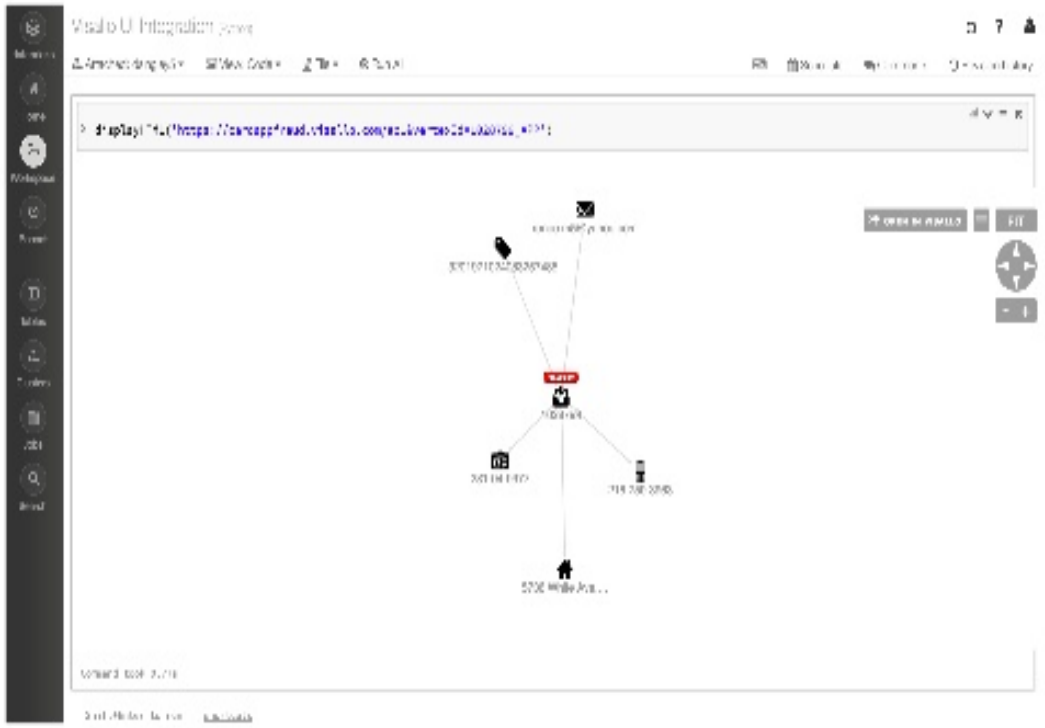
databricks™

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"

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

VISALLO

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



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 Pipeline

Review Databricks Notebook

Read historical credit card applications

Create vertex and edge DataFrames for the graph –> GraphFrame

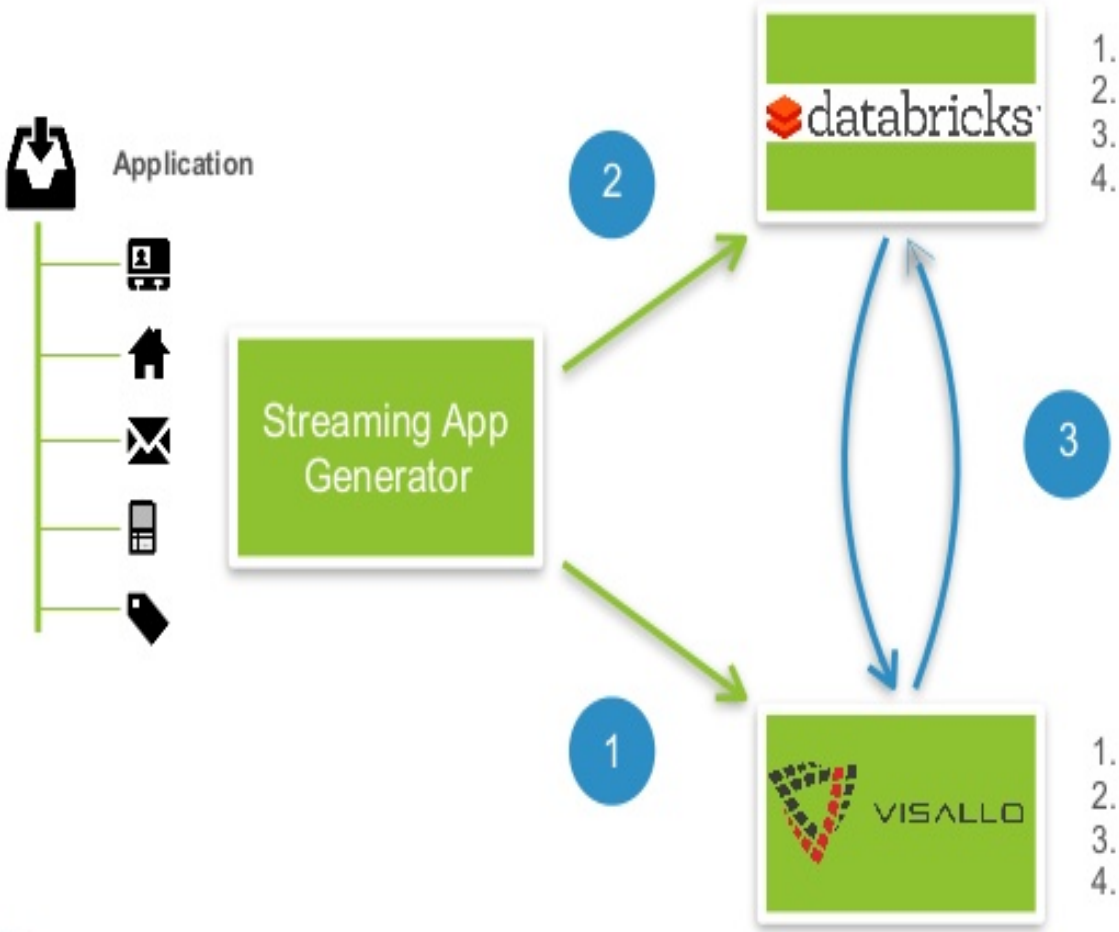
Compute features for every subgraph

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

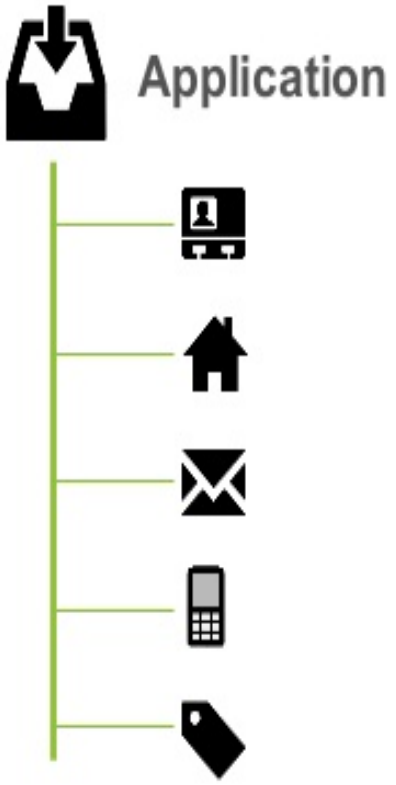


1. Stream into Spark cluster
2. Invoke Visallo connector to determine "connectedness" in real-time
3. Score application using trained logistic regression model (earlier step)
4. Identify potentially fraudulent applications and boost network count

Databricks makes it easy to integrate with their notebook via APIs and inline frames

1. Insert into Graph Database and index in Elasticsearch
2. Attach to one or more existing network based on features
3. Compute "connectedness" and expose via RESTful API
4. Provide exploratory analytics and out-of-the-box visualizations

Performance Metrics



	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

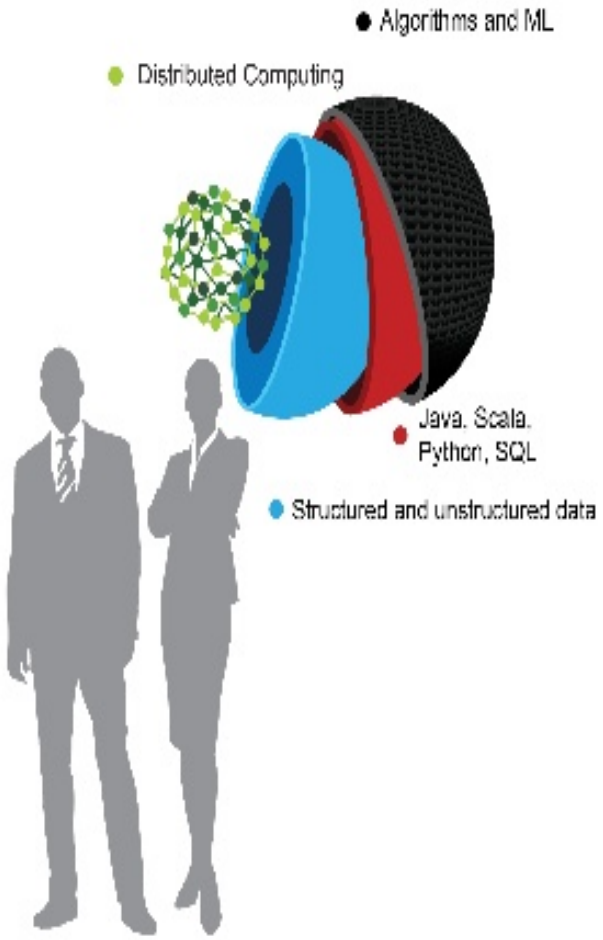
DEMO

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


Capital One
THANK YOU.
chris.dagostino@capitalone.com
@chrisdagostino
Spark summit
SPARK SUMMIT 2016
DATA SCIENCE AND ENGINEERING AT SCALE